CEFET/RJ - CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA
Campus Nova Friburgo
Bacharelado em Sistemas de Informação
5º Período

Gestão do Conhecimento e da Informação

Compilação de Materiais

Livro North, K., Kumta, G. Knowledge Management – Value Creation Through Organization Learning 2nd ed. Springer (2018). Disponível em https://doi.org/10.1007/978-3-319-59978-6

Projeto de Banco de Dados e Inteligência de Negócios Operacional
Universidade da Califórnia - Irvine

Análise de Dados
Universidade Wesleyana

Datawarehouse
Universidade do Colorado

Índice

1. Gestão do Conhecimento

1.1. Rumo a uma sociedade do conhecimento habilitada digitalmente

Klaus North[1] e Gita Kumta[2]

(1) Wiesbaden Business School, Hochschule RheinMain, Wiesbaden, Alemanha
(2) School of Business Management, SVKM's Narsee Monj. Inst. de Estudos de Administração, Mumbai, Maharashtra, Índia

"Em uma economia onde a única certeza é a incerteza, a única fonte segura de vantagem competitiva duradoura é o conhecimento"

Ikujiro Nonaka

Resultados de aprendizagem

Depois de concluir este capítulo
- Você terá adquirido uma compreensão da criação de valor na economia do conhecimento habilitada digitalmente,
- Você conhecerá os desafios e abordagens para gerenciar organizações intensivas em conhecimento;
- Será capaz de avaliar a «aptidão» de uma organização para a competição baseada no conhecimento;
- Você pode administrar um café do conhecimento.

1.1.1. Conhecimento: um recurso para criar riqueza

Sociedades e Economias do Conhecimento

O **conhecimento** como recurso para a criação de riqueza está a ganhar uma importância crescente a nível global ao nível das nações, regiões, organizações, equipas e indivíduos. As sociedades do conhecimento emergentes desenvolvem suas capacidades para identificar, produzir, processar, transformar, disseminar e usar a informação para construir e aplicar o conhecimento para o desenvolvimento humano. Eles exigem uma visão social empoderadora que englobe pluralidade, inclusão, solidariedade e participação (UNESCO 2005). Nas sociedades do conhecimento, os valores e práticas de criatividade e inovação desempenham um papel importante para sustentar a vantagem competitiva. A criatividade e a inovação também levam à promoção de novos tipos de processos colaborativos (UNESCO 2005), cada vez mais habilitados digitalmente.

Devemos observar, no entanto, que toda sociedade tem seus próprios ativos de conhecimento desenvolvidos muitas vezes ao longo dos séculos. É necessário trabalhar no sentido de conectar as formas de conhecimento que as sociedades já possuem e as novas formas de desenvolvimento, aquisição e disseminação do conhecimento valorizadas pelo modelo de economia do conhecimento (UNESCO 2005). As sociedades do conhecimento são dominadas por especialistas profissionais e seus métodos científicos. As economias do conhecimento são marcadas pela expansão das ocupações produtoras ou disseminadoras de conhecimento (Burke 2000).

Peter Drucker usou o termo ***sociedade do conhecimento*** já em 1969 no seu livro "***The Age of Discontinuity***". Em seu estudo seminal "***The Production and Distribution of Knowledge in the United States***", Fritz Machlup (1962) concentrou sua pesquisa no sistema de patentes, mas percebeu que as patentes eram simplesmente uma parte de uma ***economia do conhecimento*** muito maior que ele analisava. Na década de 1990, estudos detalhados sobre a transformação do trabalho, propriedade e conhecimento foram conduzidos (Stehr 1994; Mansell e When 1998; Adolf e Stehr 2017; Kornienko 2015).

Três Forças Motrizes

A crescente importância do conhecimento como um recurso pode ser rastreada até três forças motrizes interdependentes (Fig. 1.1):

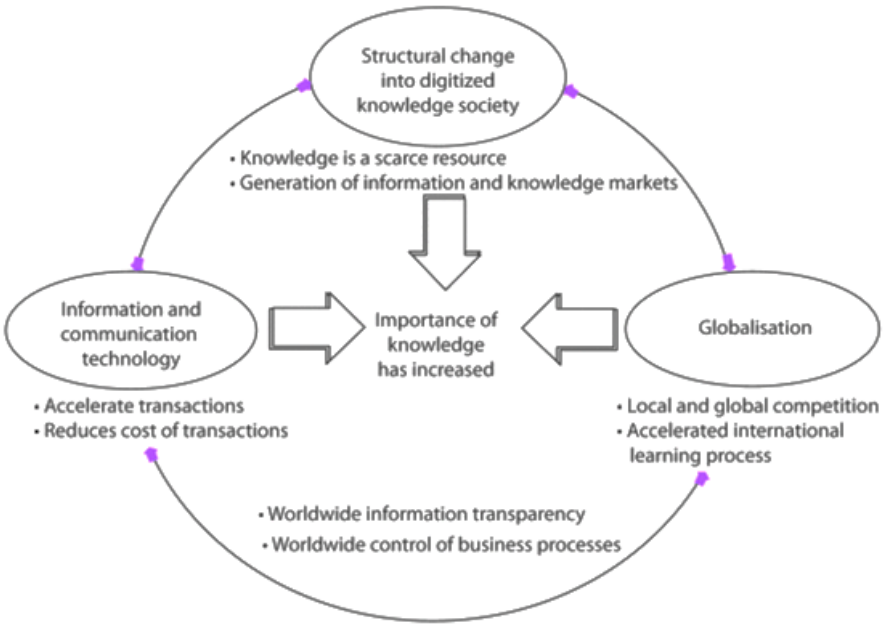


Fig 1.1 Três Forças Motrizes aumentando a importância do conhecimento como um fator de competitividade

- **Mudança estrutural:** Passar de atividades intensivas em trabalho e capital para atividades intensivas em informação e conhecimento significa que as empresas vendem cada vez mais informações, conhecimento ou produtos e serviços inteligentes. O trabalho e o capital são substituídos pelo conhecimento como um recurso escasso. Essa mudança estrutural resulta em formas alteradas de organização e transação dentro e entre as empresas, bem como em uma mudança no papel da administração e dos funcionários.

- **Globalização:** A globalização da economia mudou a divisão internacional do trabalho. Os países conhecidos como nações industriais estão agora se tornando nações do conhecimento. Os processos de aprendizagem internacional estão ganhando ritmo de tal forma que novos concorrentes estão surgindo no mercado mundial em um curto espaço de tempo devido aos rápidos ciclos de aprendizagem. A digitalização permite a entrega internacional de serviços.

- **Tecnologias de Informação e Comunicação (TIC):** As TIC permitem lidar com big data, conectar-se facilmente, colaborar e interagir com baixos custos de transação e trazer transparência de informações em todo o mundo. Assim, com «informação perfeita» podemos chegar um passo mais perto da concorrência ideal. Isso resulta em mudanças rápidas no mercado e uma maior taxa de inovação que se reflete em reduções de preços, ciclos de vida de produtos mais curtos, personalização dos requisitos do cliente e surgimento de novas áreas de negócios. Um novo mercado global de informações é estabelecido. A transformação digital acelera a mudança estrutural e a globalização.

Rumo a ativos digitalizados e intangíveis

Atualmente, assistimos a uma evolução para sociedades do conhecimento digitalizadas à escala global. O que isto significa? A mudança para um mundo cada vez mais digital está mudando rapidamente a maneira como as pessoas e organizações criam, usam e compartilham dados, informações e conhecimento. Uma definição comum de "transformação digital" é a cunhada por Bounfour (2016), ou seja, "a mudança associada à aplicação da tecnologia digital em todos os aspectos da sociedade humana".

Séc 16 - 17	Séc 18 - 19	Séc. 20	Séc. 21
Era da Razão	Sociedade Industrial	Sociedade do Conhecimento e da Informação	Sociedade do Conhecimento Digital
• Exploração científica da natureza (Rousseau, Galilei, Newton…) • Desenvolvimento do Método Científico - metodologia sistemática para a apropriação de novo conhecimento • Interação entre acadêmicos e artesãos, Surgimento das instituições do conhecimento (universidades)	• Produção de Conhecimento permeia todas as áreas da vida • Revolução Industrial - Separação do conhecimento (planejamento e projeto) e execução (conhecimento incorporado em máquinas) • Profissionalização dos produtores de conhecimento (engenheiros e doutores)	• Conhecimento se torna o fator de produção dominante • Surgimento da Computação, Internet, Inteligência Artificial, Algoritmos para rotinas • Domínio de profissionais especialistas e seus métodos científicos	• Digitalização da vida cotidiana e da criação de valor • Sistemas cognitivos, sociais, colaborativos e interconectados - Inteligência Aumentada • Penetração digital de profissionais e educação
Conhecimento 1.0	Conhecimento 2.0	Conhecimento 3.0	Conhecimento 4.0

A Figura 1.2 mostra esse desenvolvimento em uma perspectiva histórica (van Doren 1991; Burke 2000) começando com a ***Idade da Razão*** (Conhecimento 1.0). Embora na antiguidade tenham existido escolas de filósofos que refletiam sobre o conhecimento, pelo menos na Europa, o século XVI é considerado o início de uma exploração científica sistemática da natureza e o desenvolvimento de um método científico mais amplamente aceito. A partir de 1700, tornou-se possível seguir uma carreira intelectual não apenas como professor ou escritor, mas também como membro assalariado de certas organizações dedicadas à acumulação de conhecimento, notadamente as academias de ciências (van Doren 1991).

Os *insights* obtidos na "Era da Razão" permitiram o desenvolvimento de uma "Sociedade Industrial" (Conhecimento 2.0) no século XVIII. O conhecimento foi cada vez mais incorporado em máquinas e sistemas de produção. A criação do conhecimento foi profissionalizada.

O século XX testemunhou o surgimento de uma "Sociedade da Informação e do Conhecimento" (Conhecimento 3.0), onde a informação e o conhecimento se tornaram fatores de produção dominantes. Nos Estados Unidos e na Europa, já por volta do ano 2000 mais de 30% da população economicamente ativa trabalhava em profissões intensivas em conhecimento e criativas como engenharia, ciências, ensino, consultoria, bancos, administração, jornalismo, prática médica, direito e arte; nas profissões sociais; ou no setor de informação e comunicação, para citar apenas alguns (Florida 2002). A mudança estrutural para uma sociedade da informação e do conhecimento também envolve mudanças nas relações de trabalho, onde o status de emprego formal e em tempo integral é cada vez mais complementado por trabalho autônomo, trabalho autônomo e atividade empreendedora (North e Gueldenberg 2011).

Nesta economia, os investimentos intangíveis em produtos, desenvolvimento, educação e formação em software bem como no aumento da eficácia dos processos de gestão e fornecimento de informação acabam por ser os indicadores decisivos para o desempenho futuro da economia. O valor de uma empresa é, portanto, determinado cada vez mais por seu "capital intelectual" e menos com base no valor contábil, ou seja, os ativos físicos de uma empresa (Sveiby 1997). Assim, desde o início da década de 1980, assistimos a uma evolução divergente do valor contabilístico e do valor de mercado das empresas, onde algumas empresas são avaliadas no mercado de ações em dez vezes ou mais o seu valor real contabilístico. O termo "ativos intangíveis" foi cunhado para explicar a diferença entre esses dois valores. Os elementos desses ativos intangíveis que são tradicionalmente chamados de "goodwill" (durante a venda da empresa) incluem nomes de marcas, base de

clientes e fornecedores, o conhecimento de mercado relacionado, a competência individual dos funcionários, bem como a "competência coletiva de resolução de problemas" que é representada por funcionários, tecnologias, software, processos de produção, patentes, etc. (Sveiby 1997). Portanto, não é surpreendente que, além das empresas de software, mesmo as empresas de marca e fabricantes de produtos intensivos em conhecimento, como medicamentos, apresentem um grau particularmente alto de ativos intangíveis (Fig. 1.3).

Conhecimento 4.0 refere-se a um estágio social em que as aplicações de tecnologias digitais são difundidas na vida cotidiana, levando a uma "ubiquidade digital" (Iansiti e Lakhani 2014) e também contribuem com uma parcela significativa para a criação de valor. Pesquisadores descobrem que produtos inteligentes e conectados com suas quatro capacidades de monitoramento, controle, otimização e autonomia transformam a competição na economia do conhecimento habilitada digitalmente (Porter e Heppelmann 2014). Assim, a expertise profissional é cada vez mais alavancada ou "aumentada" Davenport e Kirby (2016) por sistemas cognitivos e em rede. Por exemplo, a McKinsey prevê um impacto econômico potencial de cinco a sete trilhões de dólares americanos por meio da automação do trabalho do conhecimento até 2025 (Manyika *et al.* 2013).

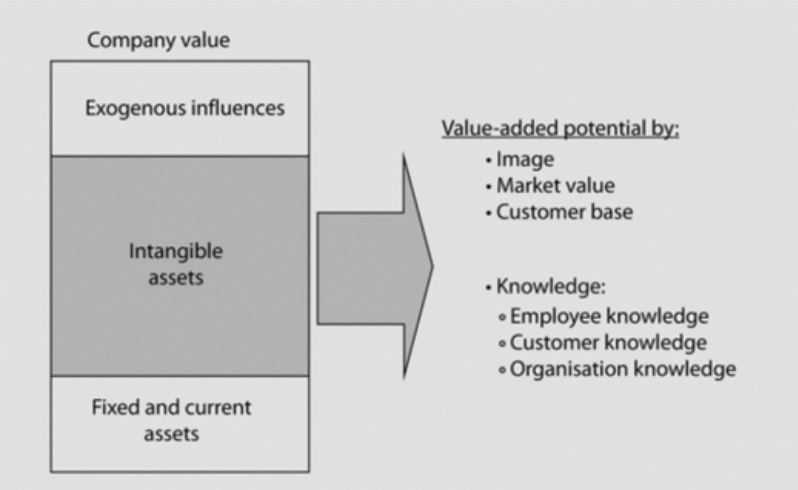


Fig 1.3 O valor de uma companhia está sendo cada vez mais determinado com base em seus ativos intangíveis

Na "sociedade do conhecimento digitalizado" (Conhecimento 4.0), as estratégias de transformação digital assumem uma perspectiva diferente e perseguem objetivos diferentes (North *et al.* 2018). De uma perspectiva centrada nos negócios, eles se concentram na transformação de produtos, processos, modelos de negócios e aspectos organizacionais devido às novas tecnologias, como big data, análise de negócios, computação em nuvem, sistemas cognitivos, robôs, software social e Internet das Coisas. De uma perspectiva centrada no ser humano, o foco da gestão do conhecimento em coleções de conhecimento (documentado) foi estendido para incluir conexões entre pessoas (Kaschig *et al.* 2016) e para abraçar as relações sociais com seu suporte tecnológico correspondente, também chamado de ambientes de conhecimento social (Pawlowski *et al.* 2014).

Pesquisadores associaram as capacidades de análise de big data a um "capitalismo de dados" que está "lucrando com nossa privacidade" (Thornhill 2017). Nessa visão, os dados se tornaram uma importante fonte de monetização, pois permitem a análise das preferências do cliente e fornecem publicidade, produtos e serviços otimizados para o usuário e para desenvolvê-los ainda mais. Algoritmos são cada vez mais difundidos em muitos campos (Ausiello e Petreschi 2013).

Seja nos negócios ou no dia a dia, as estratégias de transformação digital têm alguns elementos em comum. Esses elementos podem ser atribuídos a quatro dimensões: uso de tecnologias, mudanças na criação de valor, mudanças estruturais e aspectos financeiros (Matt *et al*. 2015). A transformação de ativos análogos em representações eletrônicas está associada a novas formas de cognição.

1.1.2. Divisão Internacional do Trabalho Baseada em Ativos Intangíveis

A disponibilidade mundial de informações, bem como as facilidades de comunicação eficientes e de baixo custo, levaram a um aumento explosivo do comércio internacional e dos investimentos estrangeiros diretos por meio da participação de mais e mais países.

Em uma geração, a proporção do produto interno bruto dos EUA em relação ao produto bruto mundial caiu de aproximadamente 50% para cerca de 20%. Novos concorrentes se lançam no mercado mundial e aprendem rápido. A ACER, por exemplo, empresa eletrônica fundada em Taiwan em 1976 com 11 funcionários, aprendeu rapidamente por meio de *joint ventures* e alianças. Hoje, é um dos principais fabricantes internacionais de computadores e semicondutores.

Na nova divisão internacional do trabalho, a «venda» de informação e conhecimento embalados em produtos e serviços tem vindo a ganhar cada vez mais importância face à mera exploração das diferenças de custos e puras «economias de escala» que caracterizaram a divisão internacional do trabalho na quarta vaga de Kondratiev (Huws 2005). Em particular, o comércio de serviços intensivos em conhecimento e royalties internacionais e pagamentos de taxas de licença (como uma medida para a venda de propriedade intelectual) cresceram significativamente. pesquisa de propriedade intelectual para pedidos de patente; pesquisa de negócios e de mercado, serviços jurídicos e médicos; formação, consultoria e investigação e desenvolvimento (Mehrotra 2005; Agarwal e Nisa 2009; Contractor *et al.* 2010). A transformação digital permite a terceirização avançada de serviços.

As economias avançadas transformam-se cada vez mais em «nações do conhecimento». As suas empresas têm conhecimento dos mercados mundiais, desenvolvem conceitos de produtos, organizam processos de produção a nível internacional, bem como controlam a logística internacional da «cadeia de abastecimento». A produção física e, até certo ponto, até mesmo o desenvolvimento de componentes de produtos ocorrem nas novas nações industriais ou mercados emergentes. Chamamos isso de conceito de empresário de divisão internacional do trabalho, conforme descrito no estudo de caso abaixo (North 1997).

A disponibilidade de conhecimento também é um critério para decisões relativas à localização das atividades de negócios. Isso envolve não apenas a criação de conhecimento do mercado local, mas também a disponibilidade de funcionários e fornecedores qualificados correspondentes. As empresas visam pesquisar, desenvolver ou produzir em um local onde se possa aprender mais. Não é difícil prever que, no futuro, o uso da vantagem comparativa de custo será menos importante do que o uso da vantagem comparativa de conhecimento.

A criação e a transferência de conhecimento desempenham um papel importante até mesmo na gestão operacional de empresas internacionais. Isso envolve decisões sobre "qual conhecimento é criado onde" e "como o conhecimento pode ser transferido de forma eficiente". As empresas multinacionais estão se transformando em redes mundiais de conhecimento com seus clientes e fornecedores.

Estudo de caso

Empresários de produção: orquestrando redes internacionais de manufatura

«Como garantir a presença no mercado global e minimizar o investimento próprio?» é o desafio para as empresas globais. Uma solução é tornar-se um «empresário de produção» em vez de um fabricante com alta «integração» vertical. Um «empresário de produção» desenvolve o conceito do produto, encomenda os módulos do produto aos fornecedores do sistema, coordena a produção e montagem de peças numa rede internacional de produção e encarrega-se da venda e distribuição dos produtos. O poder do empresário de produção reside em seu conhecimento dos mercados mundiais, tecnologia e inovações. Para organizar o desenvolvimento, a produção e as vendas de produtos, o empresário deve estar em condições de transferir informações relevantes ao longo das cadeias de valor, ou seja, deve ser capaz de controlar o processo de aprendizagem internacional e oferecer suporte logístico. O conceito de empresário de produção se consolidou principalmente na indústria automobilística, têxtil e eletrônica global.

Assim, por exemplo, o conceito multidoméstico de um fabricante líder de caminhões é baseado no conhecimento de que os mercados, especialmente nos países em desenvolvimento, não podem ser conquistados com veículos de alta tecnologia produzidos em países de alto custo, mas vendidos em dólares. Somente os caminhões adaptados ao poder aquisitivo e às condições de uso desses países e que possivelmente contenham muitas peças de produção local são adequados para esses mercados em crescimento. A ideia básica é a seguinte: o fabricante de caminhões rompe com os riscos de investimento e produção própria com o objetivo de assumir cada vez mais o papel de fornecedor de *know-how*, desenvolvedor e especialista em logística mundial. Desta forma, a empresa se livra do risco e se torna mais ágil ao passar os problemas dos pools de custos fixos para outras pessoas envolvidas na produção.

A Benetton também opera como empresário de produção. Até 2000 a Benetton fazia parte de sua produção em fábricas próprias e através de uma ampla rede de subcontratados nacionais, principalmente especializados em costura. Agora a Benetton mudou drasticamente para uma nova estratégia, abandonando a Itália e organizando a produção em torno de uma cadeia de suprimentos dupla: locais próximos (Europa Oriental e Norte da África) para produção rápida e locais distantes (Ásia) para produtos mais padronizados. Isso leva a uma redefinição de competências para o distrito de roupas de Treviso, onde os subcontratados tradicionais da Benetton estiveram em poucos anos, drasticamente reduzidos. A reestruturação da Benetton marca a transição para uma nova rede de competências entre os agentes do distrito. A rede de vendas é organizada através de um sistema de *franchising* multinível. Aproximadamente 70 empresas independentes trabalham como revendedores regionais do grupo. Mais de 3.000 pontos de venda em todo o mundo são operados por empresas independentes como parceiras de franquia da Benetton. A Benetton é responsável mundial pelo marketing e tem representantes regionais. Assim, com vendas relevantes e dados de mercado, ela está em posição de aumentar rapidamente seu baixo patrimônio usando o conceito de franquia (Crestanello e Tatara 2009; Fornengo Pent 1992; North 1997).

1.1.3. Concorrência Acelerada: Melhorando Mais Rápido e Tornando-se Diferente

Repensar as definições tradicionais de economia, criação de riqueza, modelos de negócios e organizações e estruturas institucionais também tem consequências sobre como as empresas competem e as instituições agem em ambientes cada vez mais habilitados digitalmente.

Enquanto o desejo de «melhorar mais rápido» visa aumentar a eficiência, isso apenas traz um alívio de curto prazo para manter a liderança competitiva. Tomemos o exemplo de uma empresa líder em eletrônicos que vê uma erosão anual de 15% no preço de seus produtos. A transferência de melhores práticas pode levar a um aumento na produtividade, mas não é um remédio duradouro. Para evitar essa queda de preço, os parâmetros de concorrência devem ser alterados por meio da inovação de produtos, processos ou modelos de negócios. Esforços devem ser feitos para trazer produtos e serviços únicos e inimitáveis ao mercado.

Assim, a gestão orientada para o conhecimento não significa apenas «melhorar mais rapidamente», mas também «tornar-se diferente, gradualmente».

Diferente, porque torna-se impossível ou muito difícil imitar a empresa que adquire uma nova configuração de recursos em decorrência de uma mudança em sua cultura. Aos poucos, porque na maioria dos casos isso significa uma mudança para uma nova cultura empresarial baseada na inovação que é resultado de um processo altamente complexo. Tal mudança deve ser iniciada, organizada e sustentada com muita paciência.

Nesse sentido, a inovação pode ser definida como uma nova configuração de conhecimento que resulta em processos, produtos ou modelos de negócios novos ou aprimorados.

Os produtos podem ser imitados a curto ou longo prazo, dependendo de sua complexidade. É muito difícil, porém, imitar a capacidade que está organizada e fixada em uma empresa de criar, combinar, transferir e armazenar conhecimento e gerar soluções a partir do conhecimento para as necessidades presentes e futuras dos clientes. Portanto, é uma fonte de vantagem competitiva duradoura. A competição do conhecimento recompensa

a habilidade de jogar com um número infinito de opções para encontrar novas e melhores maneiras de fazer as coisas (Romer 1986). Para isso as empresas precisam desenvolver «capacidades dinâmicas».

Por que essa nova "evolução do conhecimento" não pode levar ao desenvolvimento de uma qualidade totalmente nova de competição dentro e entre as empresas? Podemos tomar o exemplo análogo dos processos de desenvolvimento da vida, que envolve o surgimento de formas superiores a partir de uma interação construtiva das diferentes formas primitivas, através de um «Jogo de Soma Mais» em que a vantagem de uma forma está ligada à vantagem simultânea da outra. A partilha de conhecimento dentro e entre as organizações é um «Jogo de Soma Mais» em que a soma do que é ganho por todos os jogadores é maior do que a soma combinada do que os jogadores entraram no jogo (conceito de coopetição).

Outro fator que contribui para novas formas de interação e competição é que os limites clássicos das empresas mudam e até desaparecem às vezes, o que, por exemplo, se aplica ao conceito de inovação aberta (Chesbrough *et al.* 2006).

As empresas estão cada vez mais sendo consideradas como entidades virtuais que revisam os conceitos tradicionais de negócios: de competição-rival para apreciação cooperativa da concorrência, de uma mera organização baseada em tarefas para uma organização orientada a processos que é direcionada para a criação de valor, de gestão de aliança baseada em desconfiança para gestão de aliança baseada em confiança. Todos na organização estão envolvidos «num processo ininterrupto de auto-renovação pessoal e organizacional. Todo mundo é um trabalhador do conhecimento - ou seja, um empreendedor» (Nonaka e Takeuchi 1995). O empreendedorismo corporativo pode, portanto, ser caracterizado por três dimensões: inovação de produto, propensão a assumir riscos e proatividade na busca de novas oportunidades (Barringer e Bluedorn, 1999).

No entanto, problemas significativos de implementação confrontam o reconhecido potencial da gestão do conhecimento em uma empresa. Apesar da superioridade da tecnologia da informação, bancos de dados, troca de experiências, grupos de trabalho, comitês diretores, etc., muitas empresas conseguem parcialmente ou falham totalmente em trazer transparência ao conhecimento e no aproveitamento de sinergias. Acabam assim por «reinventar a roda». Em muitos casos, os funcionários não estão cientes dos desenvolvimentos que ocorrem em alguma outra área da mesma organização. Quando trabalhar em conjunto dentro de uma área de negócio é um desafio em si, é ainda mais difícil cooperar em um segmento de negócio com o objetivo de converter todo o conhecimento disponível de forma rápida e eficiente em soluções para os problemas dos clientes.

Isso pode ser visto como resultado de um mal-entendido sobre o processo de criação do conhecimento. Embora uma visão seja restrita ao processamento de informações, a abordagem mais bem-sucedida é ver a criação do conhecimento como um processo que permite à empresa responder rapidamente aos clientes, criar novos mercados e desenvolver rapidamente novos produtos e serviços. O processamento de informações apenas cria conhecimento formal em termos de dados, procedimentos e princípios codificados e é medido usando métricas como maior eficiência, custos mais baixos e melhor retorno sobre os investimentos (Nonaka e Takeuchi 1995).

A forma multidivisional de organização encontrada em uma série de grandes empresas muitas vezes atrapalha o fluxo suave de conhecimento entre os segmentos. Portanto, há um argumento de que uma criação e transferência eficiente de conhecimento dentro da estrutura de uma organização hierárquica e multidivisional é difícil (Hedlund 1994). Além da estrutura organizacional mencionada acima, até mesmo os valores que são praticados na organização podem criar restrições, pois conhecimento é poder e é mantido em sigilo. A síndrome do «não foi inventado aqui» dificulta a transferência de conhecimento. Frequentemente, os sistemas de recompensa e avaliação que têm uma orientação individualista oferecem muito pouco incentivo para criar e distribuir conhecimento (ver Fig. 1.4).

Fig 1.4 Quais são as barreiras para a criação e transferência de conhecimento?

No entanto, é cada vez maior a consciência de que «A criação e troca de conhecimento é muito importante para o nosso negócio e leva-nos para a frente». Essa crescente conscientização entre a administração e os funcionários é um bom ponto de partida para a mudança para uma nova qualidade de competição. Os gerentes entrevistados por nós resumiram os problemas e o potencial da gestão do conhecimento da seguinte forma:

> *Se soubéssemos o que nossa empresa sabe, poderíamos atender melhor aos requisitos do cliente, oferecer produtos inovadores mais cedo, reagir mais rapidamente às "mudanças do mercado e aumentar nossa produtividade. Em suma, poderíamos melhorar em um ritmo mais rápido.*

Estudo de caso

Engenharia K&P: aprendendo rápido

A K&P Engineering realiza análises estruturais para edifícios complexos (por exemplo, pontes) em dois escritórios com aproximadamente 30 funcionários, a maioria engenheiros. Somente os engenheiros que lidam com projetos com eficiência e aprendem rapidamente com seus erros, bem como aqueles que se destacam como especialistas em uma área específica, são bem-sucedidos neste negócio. O cérebro desses funcionários contém conhecimento altamente especializado sobre soluções e erros recorrentes na construção. Como esta informação pode ser armazenada, disponibilizada a todos e utilizada para formação e melhoria contínua dos colaboradores mais jovens?

Na K&P, erros recorrentes de construção e boas soluções são documentados em um banco de dados estruturado de acordo com os tipos de edifícios. Se um funcionário tiver que realizar uma análise estrutural para um novo objeto, ele pode se atualizar com os defeitos de construção recorrentes consultando o banco de dados, detectá-los rapidamente, evitá-los se possível em sua obra e aprender os elementos de uma «boa solução». Isso gera um conhecimento coletivo comumente acessível da empresa de engenharia.

Embora seja fácil usar o banco de dados da solução, nem sempre é fácil convencer os funcionários a alimentar suas informações no sistema. Eles cometem erros, pois trabalham sob alta pressão e não gostariam de ser vinculados a erros documentando-os. Além disso, eles possivelmente sentem que o valor de sua experiência diminuirá se outros também tiverem acesso à sua experiência. Até agora, a K&P conseguiu motivar seus funcionários a fornecer informações, comunicando-se com eles e convencendo-os. Com o aumento do conteúdo da base de dados, há um aumento de sua utilização pelos funcionários. Assim, uma cultura de aprender com os erros começa a se estabelecer.

Alguns problemas típicos de conhecimento nas organizações

- Os funcionários não conseguem encontrar informações críticas existentes quando necessário. Isso resulta em funcionários usando informações incompletas ou reinventando a roda. Informações sobre um estudo realizado em uma determinada área, se encontradas facilmente, ajudarão a reduzir o tempo para iniciar um estudo em outra área semelhante e estimar o esforço de forma mais realista. O conhecimento é de pouco valor se não puder ser encontrado quando necessário.
- As lições são aprendidas, mas não compartilhadas. O conhecimento adquirido por meio do fracasso costuma ser subestimado. Os eventos que atrasaram a conclusão do projeto ou afetaram negativamente as vendas são frequentemente esquecidos. Tende-se a repetir os erros do passado por falta de conhecimento ou pela inacessibilidade das lições aprendidas com os fracassos.
- Muitas vezes, as organizações não sabem o que já sabem. Na economia baseada no conhecimento, a sobrevivência depende da melhor resposta possível a uma multiplicidade de desafios, usando principalmente o conhecimento adquirido através da experiência passada. Devido à falta de cultura de compartilhamento e facilitação, as melhores práticas de um grupo não são incorporadas aos procedimentos da organização.
- Muitas vezes, os indivíduos que possuem informações valiosas não são rastreados na organização e esse conhecimento se move com eles sem nenhum benefício para a organização.

1.1.4. O que é Gestão do Conhecimento?

Desde meados da década de 1990, há um intenso discurso acadêmico e experimentação prática sobre modelos e práticas de gestão do conhecimento nas organizações. A pesquisa se concentrou em "A empresa criadora de conhecimento" (Nonaka e Takeuchi 1995) e na "Nova Riqueza Organizacional" (Sveiby 1997) e "Capital Intelectual" (Stewart 1997). Os dois últimos discutindo novas formas de medir e gerenciar recursos baseados em conhecimento. Abordagens multidisciplinares para a gestão do conhecimento surgiram em áreas como ciência da informação e da computação, biblioteconomia, administração de empresas, psicologia, sociologia, educação, engenharia, filosofia e outras áreas científicas (Heisig 2015).

Zack e*t al.* (2009) postulam que a gestão do conhecimento (GC) progrediu de um conceito emergente para uma função cada vez mais comum nas organizações empresariais. O caminho para uma empresa inteligente e orientada para o conhecimento começa inicialmente com cinco perguntas básicas:

1. Qual a importância do conhecimento em relação aos ativos físicos para o sucesso do nosso negócio?
2. Quais objetivos estratégicos queremos apoiar com a gestão do conhecimento?
3. Que conhecimentos/competências possuímos e que conhecimentos/competências necessitamos no futuro para garantir uma competitividade duradoura?
4. Como gerimos o recurso «conhecimento» na empresa?
5. Como devemos organizar e desenvolver nossa empresa para que possamos enfrentar a competição atual e futura baseada no conhecimento?

A Figura 1.5 mostra que, ao comparar o conhecimento atual e o futuro, as organizações podem desenvolver respostas para as perguntas acima.

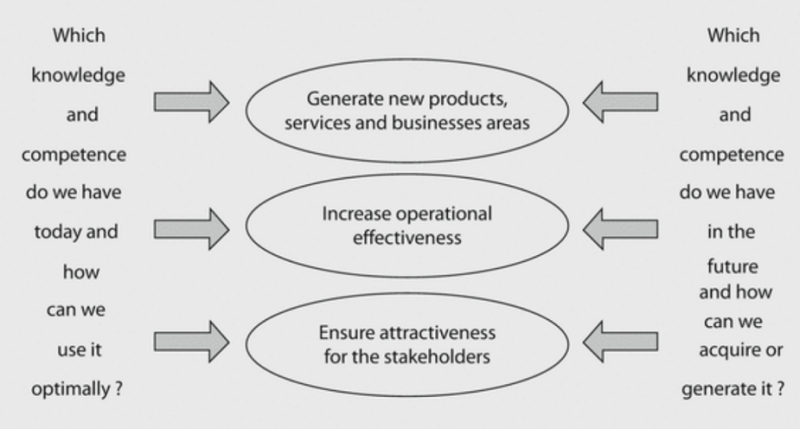


Fig 1.5 Questões Básicas para a Gestão do Conhecimento nas Organizações

Antes de fornecermos nossa definição de Gestão do Conhecimento, gostaríamos de deixar claro qual é o nosso entendimento de gestão. O papel da administração em uma organização que aprende foi bem formulado por Drucker:

Gerenciamento significa:

1. Fazendo os pontos fortes das pessoas eficazes e suas fraquezas irrelevantes
2. Entendendo a capacidade das pessoas de contribuir,
3. Integração de pessoas em um empreendimento comum pensando, definindo e exemplificando os objetivos, valores e objetivos organizacionais
4. Habilitando a empresa e seus membros a crescer e se desenvolver através de treinamento, desenvolvimento e ensino
5. Assegurando que todos sabem o que precisa ser realizado, o que eles podem esperar de você e o que é esperado deles. A gerência nos permite coordenar centenas ou milhares de pessoas com diferentes habilidades e conhecimentos para alcançar objetivos comuns.

Com base nesse entendimento, definimos GC da seguinte forma (uma pesquisa na literatura revela muitas definições de GC que contêm elementos semelhantes):

Definição:

A gestão do conhecimento permite que indivíduos, equipes e organizações inteiras, bem como redes, regiões e nações criem, compartilhem e apliquem coletiva e sistematicamente o conhecimento para atingir seus objetivos estratégicos e operacionais. A gestão do conhecimento contribui para aumentar a eficiência e eficácia das operações, por um lado, e para mudar a qualidade da concorrência (inovação), por outro, desenvolvendo uma organização de aprendizagem.

Papel da Gestão do Conhecimento em Ambientes "VUCA"

No passado, as organizações envolviam-se principalmente em práticas de gestão do conhecimento (GC) que se concentravam na gestão do conhecimento atual e experiências passadas com forte ênfase na documentação (Pawlowsky *et al*. 2011, Bolisani e Handzic 2015). Hoje, um ambiente "VUCA" hipercompetitivo (volátil, incerto, complexo, ambíguo), mudou os comportamentos de comunicação e a evolução para o trabalho do conhecimento 4.0 definiu o cenário para o gerenciamento do conhecimento dentro e entre as organizações na sociedade digitalizada.3

Em analogia ao conceito de "ambidestria" (Tushman e O'Reilly 1996), GC tem que suportar uma série de atividades de conhecimento conflitantes, como "exploração" e "exploração" ou "compartilhamento" e "proteção" ao mesmo tempo em tais configurações VUCA. À luz do conflito resultante entre estabilidade e flexibilidade, a GC estabiliza as capacidades da organização em um modo de proteção e exploração, por um

lado, e simultaneamente suporta capacidades dinâmicas em um modo de exploração e compartilhamento para aumentar a agilidade e a renovação. A capacidade de uma organização de gerenciar esses processos e práticas aparentemente contraditórios ganha cada vez mais importância com a transformação digital. Vamos examinar com mais detalhes essas duas funções de GC.

GC operacional como estabilizador

Também no futuro, o GC operacional continuará a ter como objetivo disponibilizar o conhecimento certo no momento e no local certo para apoiar os funcionários de uma organização, além das partes interessadas relevantes no ambiente da organização para as operações do dia a dia.cotidianas. Os meios e formas de atingir esse ambicioso objetivo, entretanto, mudarão sob a perspectiva do GC 4.0. As organizações podem se envolver nas seguintes atividades para estabilizar o portfólio de competências em uma organização:"

1. **Facilitar fluxos de conhecimento onipresentes e com curadoria**: O acesso rápido, fácil e onipresente à base de conhecimento da organização e em todas as organizações ganha importância e pode ser caracterizado por repositórios descentralizados e cada vez mais conectados em rede, aumentados por inteligência de máquina em rápida evolução. Murray e Wheaton (2016) argumentam que há uma necessidade de "curadoria de conhecimento", pois até mesmo tecnologias avançadas, como ontologias legíveis por máquina, ainda não chegaram perto de extrair um significado profundo ou organizar com precisão o conteúdo em categorias contextuais adequadas. A curadoria estabelece, mantém e agrega valor aos repositórios de conhecimento e ajuda a mantê-los relevantes e atualizados. Na prática, a curadoria pode significar que um especialista compila uma seleção de links e os compartilha, acrescentando uma explicação clara dos critérios de seleção usados para compilar a lista, bem como breves introduções explicando por que cada link é relevante. No entanto, as decisões necessárias em tal processo também podem ser aumentadas por inteligência de máquina, por uma equipe ou multidão engajada no domínio que é curado pelo especialista.

2. **Habilitar colaboração**: a ênfase da GC mudou do suporte para coletar para conectar atividades de conhecimento (Kaschig *et al.* 2016) que ajudam a fazer a colaboração funcionar. As atividades de conexão do conhecimento são vistas de forma abrangente para incluir conexões entre pessoas, ou seja, criação, compartilhamento e aquisição conjunta de conhecimento, e conexões de conhecimento tanto de forma abstrata quanto manifesta - a integração do conhecimento de diversas fontes, sejam pessoas, documentos ou algoritmos. A GC precisa ajudar as pessoas a desenvolver as competências necessárias para o trabalho 4.0, entre as quais se destacam as competências para colaboração mediada por tecnologia e colaboração com máquinas como "colegas de equipe".

3. **Monitore e controle o aprendizado aumentado e a tomada de decisões:** À medida que as organizações desenvolvem e implantam cada vez mais algoritmos para automatizar tarefas e decisões rotineiras de conhecimento, além de fornecer suporte à decisão em situações conhecidas, esse comportamento de conhecimento automatizado precisa ser monitorado e controlado para ser não apenas eficiente, mas também compatível com o sistema regulatório interno e externo de uma organização. As correspondentes experiências feitas precisam ser sistematicamente refletidas e interpretadas a esse respeito, a GC terá que garantir a transparência das tecnologias cognitivas, para que os usuários estejam sempre cientes de como os sistemas cognitivos "pensam" e agem. Um desafio específico aqui é identificar e alavancar o conhecimento tácito de especialistas ou comunidades no assunto e fornecer os meios para que os humanos se mantenham atualizados com o crescimento exponencial de oportunidades criadas por sistemas de autoaprendizagem.

GC Estratégico como Catalisador

Num ambiente cada vez mais turbulento e complexo, cabe à GC examinar criticamente os conhecimentos e competências da organização, rede ou negócio, rede ou ecossistema empresarial e identificar os seus "pontos cegos". Aqui, GC assume o papel de inovador e "irrita o sistema" ao questionar aprendizados passados, comportamentos e práticas estabelecidas. GC deve ter sucesso em apoiar o desenvolvimento de "capacidades dinâmicas" das organizações para reconfigurar, realinhar e integrar competências essenciais com a ajuda de recursos externos. As organizações podem se envolver nas seguintes atividades para promover produtivamente o crescimento de capacidades para melhorar o desempenho organizacional sob condições ambientais em constante mudança:

1. **Identifique o conhecimento crítico**: GC precisa fornecer uma visão profunda dos ativos de conhecimento crítico necessários para embarcar na jornada de aprendizado envolvida nas atividades para perseguir os objetivos organizacionais futuros. Portanto, GC também questiona as competências essenciais atuais, direitos de propriedade intelectual, compreensão do mercado e da indústria e compreensão e expectativas do cliente (MacMillan *et al.* 2017). GC deve identificar os bolsões e ilhas de criação de conhecimento dentro e além dos limites organizacionais que podem ser conectados para adquirir novas competências essenciais que podem ser apropriadas pela organização. Portanto, as organizações precisam integrar conhecimentos isolados e visões do ambiente para dar sentido à informação como base para aproveitar novas oportunidades e transformar a organização. O mapeamento de conhecimento estratégico ajuda a descobrir e ter uma visão integral de ativos de conhecimento críticos, fornecendo o contexto para descobrir as estratégias de digitalização mais promissoras (MacMillan *et al.* 2017).

2. **Facilitar o *sensemaking*[1] e o entendimento compartilhado como base para agir**: descrever o *sensemaking* como uma forma de entender conexões entre pessoas, lugares e eventos que ocorrem agora ou ocorreram no passado, a fim de antecipar trajetórias futuras e agir de acordo. A capacidade de enquadrar (contextualizar) e reformular problemas e observações é particularmente importante quando a análise de big data parece fornecer respostas sem o conhecimento adequado do contexto (Madsbjerg 2017). *Insights* profundos e entendimentos compartilhados emergem através de múltiplos discursos de pessoas. Os mecanismos subjacentes de criação de significado podem ser vistos como a essência da colaboração e destacam que os processos de negociação são interativos, recíprocos e que o significado reside na esfera social e pode se manifestar em sistemas sócio-técnicos. *Sensemaking* é uma atividade compartilhada e comunal que produz conhecimento apropriado para a ação, mas altamente tendenciosa com base nos indivíduos que fazem o *sensemaking* - isto é, cada grupo de pessoas que têm as várias conversas de *sensemaking* "falarão à existência" um conjunto muito diferente de situações, organizações e ambientes (Weick *et al.* 2005). Nessa visão, o *sensemaking* é um processo altamente colaborativo, eficaz para o crescimento e planejamento organizacional a curto e longo prazo e altamente dependente de interpretação.

   A crescente complexidade das tarefas de trabalho intensifica a demanda por colaboração, que por sua vez requer GC para apoiar a criação de entendimento compartilhado entre grupos de trabalho (Bittner, Leimeister 2014). No nível organizacional, o entendimento compartilhado entre organizações que colaboram em ecossistemas de negócios é vital para a criação eficiente de conhecimento em tais ecossistemas. Os pesquisadores descobriram que, no início da formação do ecossistema de negócios, as organizações precisam compartilhar suas capacidades, expertise e conhecimento e, em particular, tornar explícito o conhecimento tácito para aumentar a integração.

[1] Sensemaking significa "dar sentido", significação

3. **Incentivar a renovação, o aprendizado ágil e a reflexão**: Para garantir a renovação em um ambiente em constante mudança e muitas vezes disruptivo, as empresas precisam aprender a desenvolver sistematicamente novos modelos de negócios e as organizações sem fins lucrativos precisam ser capazes de redesenhar suas missões de maneira acelerada (Kotter 2014). A GC pode desempenhar um papel fundamental nessas questões acima descritas relacionadas a tornar as organizações mais dinâmicas no futuro. Em um ambiente caracterizado pela imprevisibilidade e várias crises imprevistas, a GC deve apoiar a rápida resolução de problemas, incentivar a experimentação constante, promover o aprendizado colaborativo e facilitar a reflexão profissional para aprender com os erros. Por exemplo, GC pode ser responsável por desenvolver um processo de "próximas práticas" em uma organização. Os desenvolvimentos futuros em uma área de negócios ou tecnologia, ou em um modelo de negócios, podem ser explorados em workshops interdepartamentais que incluem uma variedade de partes interessadas, como clientes e a comunidade científica.

4. **Crie plataformas para engajamento:** em uma era de sobrecarga de informações, a atenção humana é um recurso escasso. Para atrair conhecimentos heterogêneos e inesperados é de importância estratégica construir plataformas que engajem os membros dentro e fora dos limites organizacionais. Ghazawneh, Henfridsson (2010) apontam para a importância de governar o desenvolvimento de terceiros por meio de conhecimentos específicos que eles chamam de "recursos de fronteira de plataforma". Isso inclui o design de recursos de limites técnicos, como kits de desenvolvimento de software e interfaces de programação de aplicativos, e recursos de limites sociais, como incentivos, direitos de propriedade intelectual e sistemas de controle. O papel da GC é construir plataformas que atraiam o engajamento de uma comunidade mais ampla para o desenvolvimento estratégico de competências organizacionais, produtos e serviços.

Estudo de caso

A ascensão do mercado do conhecimento

Hoje, testemunhamos o surgimento de mercados de conhecimento on-line onde você pode vender seu conhecimento pessoal. Você pode ver suas raízes na tendência de perguntas e respostas de origem coletiva que gerou sites como Quora, Aardvark, Stackoverflow ou Ask.com e onde você pode obter suas perguntas respondidas gratuitamente.

A start-up sueca www.Mancx.com está provando o sucesso de seu conceito de um mercado de conhecimento online para troca de informações pessoais por dinheiro. Mancx é um mercado de conhecimento totalmente transacional com recursos globais de pagamento/pagamento. Para compradores de informações, Mancx é o lugar certo para obter respostas às questões de negócios que eles enfrentam diariamente. Para os vendedores de informações, a Mancx oferece uma maneira de capitalizar o conhecimento acumulado e construir seu perfil de marca pessoal como fontes de informações valiosas. A Mancx fornece um ambiente seguro e anonimato para negociar e intermediar um negócio de venda de conhecimento, recebendo uma comissão de 20% em cada transação concluída.

Esta é a mesma filosofia que www.Acabiz.com tem em relação à informação. Acabiz é uma empresa italiana financiada por investidores privados e o braço financeiro do órgão governamental da Lombardia. Acabiz teve a ideia de um mercado de conhecimento a partir do desejo de criar uma plataforma para acadêmicos se conectarem com empresas, governos e ONGs. Proporciona assim uma ligação direta entre o consumidor final e o fornecedor de conhecimento especializado e dispensa intermediários ou consultores.

"Acessar conhecimento de nicho ou especializado é fundamental para qualquer atividade comercial bem-sucedida e direcionada hoje em dia", disse Guido Uglietti, sócio fundador da Acabiz. "Todos reconhecem a

importância da academia para as transferências de conhecimento empresarial, mas não existe uma ferramenta de plataforma global para facilitar e promover a transferência de conhecimento de forma simples e escalável". Acabiz criou uma plataforma para acadêmicos, que eles chamam de detentores do conhecimento, para se conectar com empresas, conhecidas como caçadores de conhecimento, que estão interessadas em sua experiência ou conhecimento específico de pesquisa. A plataforma Acabiz permite que as empresas acessem fácil e diretamente a rede de conhecimento de milhares de acadêmicos em todo o mundo, todos com conhecimento altamente especializado em áreas como arquitetura, engenharia, direito, medicina, ciência, finanças, economia e outras áreas.

Fonte: Adaptado de Jeniffer Hicks: The Rise of the Knowledge Market. Disponível em <http://www.forbes.com/sites/jenniferhicks/2011/06/27/the-rise-of-the-knowledge-market/>>

1.2. Como as Organizações Aprendem

Competir em um ambiente em constante mudança exige que as organizações aprendam. Como isso acontece? O subcapítulo a seguir é uma adaptação da excelente revisão de literatura de Brenda Barker Scott sobre aprendizagem organizacional.

O que é aprender?

A questão de saber se a aprendizagem é um processo cognitivo, bem como um processo comportamental, tem implicações práticas e teóricas.

Os teóricos que aderem a uma perspectiva puramente cognitiva veem a aprendizagem como o desenvolvimento de novos *insights* por meio da revisão de suposições, mapas causais ou esquemas interpretativos. Uma organização aprendeu «se alguma das suas unidades adquire conhecimentos que reconhece como potencialmente úteis para a organização».

Os teóricos que defendem uma abordagem cognitivo-comportamental dupla sugerem que, embora o desenvolvimento cognitivo seja necessário, a ação também é necessária para o aprendizado pleno e completo. Aqui, diz-se que a aprendizagem ocorre à medida que novos *insights*, suposições e mapas causais levam a um novo comportamento ou, inversamente, um novo comportamento leva a novos insights. Apontando para a íntima relação que a aprendizagem tem com a ação, Argyris (1999) sugere: "Pode-se dizer que uma organização aprende na medida em que identifica e corrige erros".

Os teóricos do conhecimento organizacional (OK) também notaram a distinção cognitivo-comportamental, mas do ponto de vista do produto da aprendizagem; seja o desenvolvimento do *know what* ou do *know how*.

Central para a questão cognição-comportamento é a noção de que a aprendizagem é uma função do pensamento consciente. A aprendizagem potencial, no entanto, é bloqueada quando os membros carecem do aparato cognitivo apropriado para perceber ou experimentar uma "necessidade de aprendizagem" e para a criação de sentido. O ***sensemaking*** também tem sido associado aos níveis de desenvolvimento cognitivo, em que o aprendizado de rotina está associado ao aprendizado de loop único e ao aprendizado de loop duplo com ajuste cognitivo mais profundo. Aqueles que exploram a interação entre cognição e ação investigaram como a ação surge ou leva a uma cognição mais profunda por meio de processos reflexivos, como aprendizado de ação e revisão após a ação. Como o conhecimento é altamente situacional, suas lições não podem ser facilmente codificadas e transferidas em protocolos e manuais de treinamento. Em vez disso, o conhecimento desenvolvido pelo praticante deve ser absorvido por meio da interação por meio da improvisação, aprendizado, conversação e narrativa.

As organizações podem aprender?

Enquanto alguns acadêmicos sustentam que a aprendizagem organizacional é simplesmente a soma do que os indivíduos nas organizações aprendem, outros afirmam que a aprendizagem organizacional é um reflexo das ideias coletivas, atividades, processos, sistemas e estruturas da organização. Nonaka (1991), descreve uma empresa como um organismo vivo com um senso coletivo de identidade e um propósito fundamental, que por sua vez influencia o comprometimento de cada membro em aprender e compartilhar conhecimento. "Independentemente dos benefícios para a aprendizagem individual, a interação social e as experiências comuns também desempenham um papel importante no desenvolvimento e na transferência do conhecimento do grupo.

Aqueles que exploram a aprendizagem em nível de grupo identificaram como os processos sociais permitem a troca, a síntese e a ampliação do conhecimento de membros individuais no conhecimento sinérgico que reside entre o grupo. Aqui, os acadêmicos estudaram os muitos processos e condições associados a interações produtivas de aprendizagem por meio de princípios de conversação e interação e experiências comuns de trabalho em aprendizado.

Para esse fim, teóricos práticos desenvolveram tecnologias sociais como conversas em cafés, processos de mudança de sistemas inteiros e teoria (Scharmer 2007) para oferecer inquilinos filosóficos, processuais e logísticos para a facilitação, foco, ritmo e fluxo de experiências produtivas de aprendizagem entre e entre grupos e comunidades.

A Quinta Disciplina – As organizações que aprendem são organizações…

- onde as pessoas expandem continuamente sua capacidade de criar as coisas que realmente desejam,
- onde novos e expansivos padrões de pensamento são nutridos,
- onde a aspiração coletiva é liberada e onde as pessoas estão continuamente aprendendo a ver o todo juntas.

Os elementos:

1. Domínio pessoal
2. Modelos mentais
3. Construindo uma visão compartilhada
4. Aprendizado em equipe
5. Pensamento sistêmico

Fonte: Senge (1990).

Recursos organizacionais que promovem o aprendizado

Outros, principalmente aqueles que trabalham nas organizações que podem aprender, sugerem que a capacidade de uma organização para aprender depende de uma série de características organizacionais. Em resposta ao apelo por organizações adaptáveis e responsivas, nas quais a aprendizagem é a norma, não a exceção, os estudiosos identificaram uma série de características pertinentes, incluindo a intenção de aprendizagem de uma empresa, estratégias de apoio à inovação ou desenvolvimento de capacidades, liderança esclarecida e autoridade distribuída, normas e sistemas de crença de apoio à aprendizagem, o uso de sistemas completos de planejamento e fóruns de tomada de decisão, processos e ferramentas que permitem o fluxo ou transferência de conhecimento entre indivíduos e grupos, e apoio e legitimidade da aprendizagem orientada para o profissional.

A capacidade de uma organização de explorar novos conhecimentos tem sido atribuída a quão bem ela é capaz de agir sobre novos *insights* (flexibilidade e velocidade), quão amplamente ela é capaz de espalhar novos *insights* para outras partes da organização (respiração) e o grau em que incorpora o aprendizado em recursos organizacionais, como normas, protocolos, produtos, processos e estruturas (profundidade).

Alternativamente, descrevendo as organizações como sistemas interpretativos, os notáveis teóricos Richard Daft e Karl Weick (1984) atribuíram esquemas interpretativos às organizações que, por sua vez, influenciam como os tomadores de decisão organizacionais percebem, atendem e interpretam os sinais em seus ambientes. Por sua vez, diferentes interpretações levam a diferentes respostas organizacionais, que acabam por moldar a estratégia, as normas, a forma e os protocolos de aprendizagem.

O relato de Daft e Weick (1984) sobre organizações descobridoras *versus* atuantes fornece uma lente útil para explorar como diferentes esquemas interpretativos influenciam a natureza e o tipo de aprendizagem organizacional. Em uma organização descobridora, os gerentes assumem que o ambiente é previsível e analisável. Depois disso, os gerentes tentam se adaptar e aprender definindo metas de desempenho previsíveis para esforços de melhoria contínua. Por outro lado, os gerentes em uma organização ativa assumem que o ambiente é imprevisível e maleável e, portanto, inovam e aprendem por meio da experimentação por tentativa e erro. Aqui, os gerentes entendem que, à medida que aprendem e aplicam seus aprendizados, eles, por sua vez, cocriam ou promovem um ambiente enriquecido. O mundo se transforma à medida que eles se transformam.

Independentemente de como uma empresa define suas características, é amplamente reconhecido que esses fatores contextuais moldam a aprendizagem individual e em grupo.

Em um estudo exploratório, Chawla e Joshi (2011) analisaram o impacto da gestão do conhecimento nas práticas de organização de aprendizagem (LO) na Índia e, com base em uma pequena amostra de empresas, concluíram que as empresas de TI e os serviços habilitados para TI pontuam mais alto na maioria das dimensões LO. O teste de suas hipóteses revelou que a maioria das dimensões GC teve um impacto positivo no LO. O tipo de indústria, no entanto, não teve impacto diferencial estatístico sobre as dimensões do LO na maioria dos casos.

1.3. A Empresa do Conhecimento: Uma Avaliação Rápida

Uma empresa baseada no conhecimento é caracterizada por sua capacidade de aprender e, assim, gerar conhecimento relevante para obter sucesso comercial a partir desse recurso. O sucesso econômico dessas empresas é atribuído às suas capacidades relacionadas ao conhecimento, que variam de acordo com o tipo de negócio. Uma categoria específica são empresas ou organizações intensivas em conhecimento, como empresas de auditoria, consultorias, empresas de engenharia, laboratórios de pesquisa, escolas ou universidades que vendem "conjunto de conhecimentos" de especialistas altamente qualificados ou organizam processos de aprendizagem. Para uma empresa franqueada como o McDonalds, a criação e transferência de conhecimento significa treinar com eficiência funcionários com poucas qualificações para atingir um nível de competência necessário para expandir os processos padronizados e replicáveis e as operações padronizadas de preparação de um «BigMac®» em todo o mundo. As grandes empresas indianas de TI, Infosys Technologies e Wipro, incubaram com sucesso «serviços de aprendizagem» e estão vendendo-os para clientes globais que lutam com mudanças tecnológicas e de processo em suas empresas, bem como mudanças demográficas na força de trabalho. Enquanto a Infosys integrou o serviço em seu *Enterprise Solutions Group* em 2010, a Wipro alavancou sua capacidade no espaço de aprendizado para estendê-lo como um serviço aos clientes em termos de gerenciamento de conteúdo de aprendizado, entrega de aprendizado e hospedagem e gerenciamento de plataformas de aprendizado (Das 2010).

Dimensões da Intensidade do Conhecimento

Até agora falamos de «firma baseada em conhecimento» ou de «empresa intensiva em conhecimento» sem explicar o que significa intensidade de conhecimento. A intensidade do conhecimento tem duas dimensões – intensidade do conhecimento do processo e intensidade do conhecimento do produto/serviço. Distinguimos quatro campos no portfólio de intensidade de conhecimento (ver Fig. 1.6):

- **Inteligência do produto:** Produtos e serviços variam no grau de conhecimento embutido neles. Um indicador de «inteligência do produto» é o esforço de pesquisa e desenvolvimento (P&D) como uma

porcentagem do custo total ou das vendas. A inteligência do produto é alta no caso de produtos de software, máquinas-ferramentas que identificam seus próprios erros, produtos farmacêuticos, etc.

- **Inteligência de processo:** Refere-se à complexidade dos processos e ao conhecimento embutido neles. O volume de investimentos em P&D em desenvolvimento e melhoria de processos, bem como o nível de qualificação das pessoas empregadas na produção são indicadores da inteligência de processos. A alta inteligência de processo pode ser encontrada na «Customização em massa» (Pine 1993), em que produtos feitos sob medida são produzidos com mais de milhões de variações. Os produtos resultantes, como uma bicicleta ou um fato feito à medida, não são particularmente inteligentes em si, mas a inteligência reside na conceptualização e execução do processo. Algoritmos cada vez mais sofisticados governam processos, por exemplo, na indústria financeira («FinTechs»).

- **Produto e processo inteligente** combina ambos os fenômenos descritos. Um exemplo prático é uma empresa que fabrica balanças de alta precisão em uma produção voltada para o cliente.

- **Valor agregado pelo trabalho físico:** A baixa intensidade de conhecimento na cadeia de valor agregado e no desempenho é evidente na venda do trabalho físico (até o boxe dá dinheiro!).

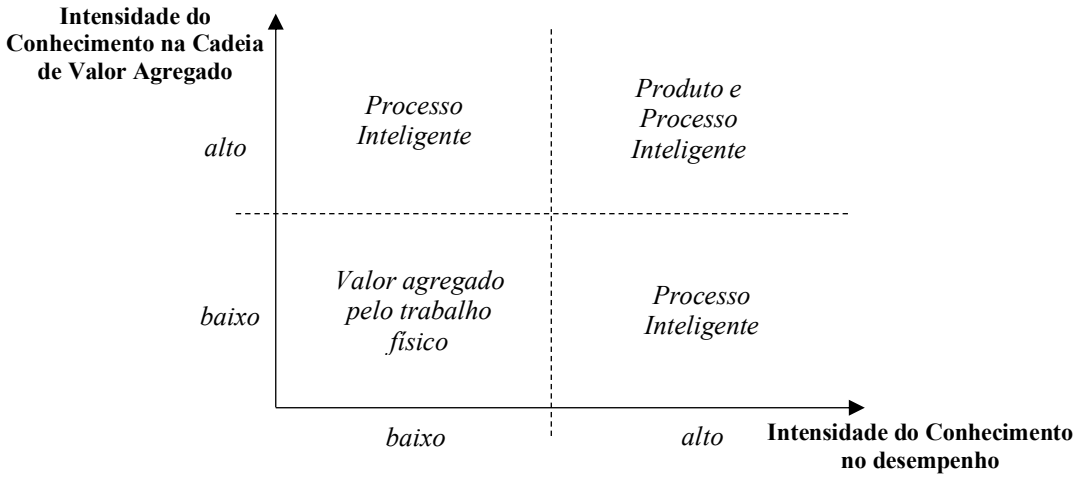


Fig 1.6 Matriz da Intensidade do Conhecimento (Porter e Millar, 1985)

Um bom indicador de intensidade de conhecimento é o valor agregado ao produto/serviço. Isto reflete o valor gerado pela transformação de insumos (matérias primas, componentes e informações) em um produto apreciado pelo consumidor. O conhecimento mais especializado e único é incorporado no processo de transformação de alto valor agregado (Porter e Millar, 1985)

O que torna uma empresa de conhecimento?

Quais são as características de uma empresa que converte conhecimento em vantagem competitiva sustentável? As empresas orientadas para o conhecimento podem ser distinguidas por uma série de características que são descritas aqui brevemente. No final deste capítulo, o leitor tem a opção de avaliar se a sua empresa é uma «empresa insensível ao conhecimento» ou uma «empresa orientada para o conhecimento». Esta breve análise permite a conscientização sobre o assunto e dar os primeiros passos para a criação de uma empresa de conhecimento. No entanto, isso não significa que toda empresa deva se tornar uma empresa do conhecimento, pois uma empresa insensível ao conhecimento também pode ser bem-sucedida (mas até quando?). Recomendamos ao leitor a leitura da breve análise ao final deste capítulo. O texto a seguir explica as seções individuais da análise subsequente.

As empresas se desenvolverão especificamente em uma empresa de conhecimento quando os requisitos do cliente forem altamente diferenciados e exigirem produtos feitos sob medida. As empresas de conhecimento irão combater uma queda no preço de produtos e serviços padrão ou "me-too", oferecendo soluções integradas complexas. Isso, por exemplo, se aplica à indústria fornecedora para a oferta de módulos e sistemas em oposição à produção de peças ou componentes individuais. Mesmo em uma consultoria, a implantação de produtos padrão é menos valorizada pelos clientes do que projetos turn-key[2] ou pacotes de soluções completas que exigem significativamente mais conhecimento e, portanto, são mais bem pagos. Mercados com alta velocidade de inovação e ciclos de vida curtos de produtos exigem criação e transferência de conhecimento rápidas. Uma empresa de conhecimento oferece soluções para problemas de clientes, que são menos intensivos em termos de trabalho e capital e cada vez mais intensivos em conhecimento. É difícil imitá-los e substituí-los, pois eles se valem de conhecimentos e habilidades complexos. Mesmo a capacidade de imitar com eficiência sob o lema «Somos imbatíveis na imitação» pode ser uma estratégia de negócios bem-sucedida.

Estudo de caso

Dabbawalas de Mumbai – Um modelo de simplicidade gerencial e organizacional

O caso da Dabbawalas de Mumbai demonstra como uma simples ideia de negócio que oferece soluções para os problemas do cliente pode se tornar um modelo de negócio de sucesso, difícil de imitar quando executado com disciplina e dedicação. Ganhou reconhecimento mundial por seu serviço e operação e nas palavras do Prof. C. K. Prahlad, é «Um modelo de simplicidade gerencial e organizacional».

«Dabbawalas», é um grupo de pessoas em Mumbai, na Índia, cujo trabalho é transportar e entregar comida caseira em lancheiras para funcionários de escritório. «Dabba» significa lancheira ou tiffin. Diariamente, nas ruas de Mumbai, 5.000 dabbawalas rotineiramente entregam almoços caseiros em transportadores de tiffin para 200.000 trabalhadores em toda a cidade.

Eles estão no mercado há mais de 100 anos e, em 1998, a revista Forbes Global conduziu uma análise e deu a eles uma classificação Seis Sigma quanto à eficiência. No mesmo ano, dois cineastas holandeses, Jascha De Wilde e Chris Relleke, realizaram um documentário chamado «Dabbawalas, o exclusivo serviço de almoço de Mumbai».

O sistema que os dabbawalas desenvolveram ao longo dos anos gira em torno de um forte trabalho em equipe e uma gestão rigorosa do tempo. Todas as manhãs, às 9h, as refeições caseiras são recolhidas em caixas especiais, que são carregadas em carrinhos e empurradas para uma estação ferroviária. Eles então seguem de trem para uma estação de descarga. As caixas são rearranjadas para que as que vão para destinos semelhantes, indicados por um sistema de letras coloridas, fiquem no mesmo carrinho. Um sistema de codificação de cores simples funciona como um sistema de identificação para o destino e o destinatário. As refeições são então entregues – 99,9999% das vezes no endereço certo. A organização depende inteiramente do esforço humano na forma de elos na extensa cadeia de entrega sem tecnologia. O sucesso do sistema depende, assim, do trabalho em equipe, de uma atitude de colaboração competitiva e de uma excelente gestão do tempo. A sinergia e a cooperação são muito altas, pois todos eles vêm de uma única seita de aldeias remotas ao redor de Mumbai.

A capacidade de combinar o conhecimento de diferentes áreas de negócios para inovar está ganhando importância, e o mesmo se aplica à velocidade de geração de novas áreas de negócios e desenvolvimento de produtos com mais eficácia do que os concorrentes.

Os investidores nas empresas de conhecimento estão interessados em um aumento duradouro no valor de uma empresa, especialmente aquelas que lidam com bens intangíveis.

---

[2] A expressão Turn Key vem do inglês e significa “Entrega de Chave”. Ela serve para descrever um tipo de contrato em que apenas um fornecedor é contratado para realizar o trabalho do início ao fim, do projeto a entrega da chave.

As empresas tradicionais costumam tratar o conhecimento como uma mercadoria, como a informação, que pode ser dividida e armazenada («comida congelada»). Mas as empresas de conhecimento estão cientes de que a criação e transferência de conhecimento é um processo de aprendizado individual e coletivo que não pode ser dominado e controlado completamente. Os funcionários de tal empresa podem discernir corretamente que aprendemos rápido com outras empresas, transferimos conhecimento de forma eficaz dentro da empresa e de/para nossos clientes, fornecedores, parceiros e competidores.

As empresas tradicionais costumam tratar o conhecimento como uma mercadoria, como a informação, que pode ser dividida e armazenada («comida congelada»). Mas as empresas de conhecimento estão cientes de que a criação e transferência de conhecimento é um processo de aprendizado individual e coletivo que não pode ser dominado e controlado completamente. Os funcionários de tal empresa podem discernir corretamente que aprendemos rápido com outras empresas, transferimos conhecimento de forma eficaz dentro da empresa e de/para nossos clientes, fornecedores, parceiros e competidores.

A empresa do conhecimento caracteriza-se principalmente por valores, processos e estruturas, a «ecologia» organizacional, que permite ao «plantar conhecimento» crescer e prosperar numa empresa. A este respeito, podemos também falar de uma «ecologia do conhecimento». Os valores básicos praticados por tal organização são confiança, abertura a novos conceitos e autenticidade. O termo autenticidade indica que os funcionários são apoiados no uso de soluções não convencionais, desfrutam de liberdade em seu comportamento e na "organização de seu trabalho e podem ser eles mesmos. Em empresas de conhecimento, boas ideias são implementadas independentemente de quem as discute. Por exemplo, especialistas de software bem pagos que muitas vezes vivem em ambientes de escritório não convencionais e podem pagar por seus "tiques" porque são criativos e incentivam as liberdades por meio de sua criatividade. O Google é um bom exemplo de empresa que sabe como estimular a criatividade e o comprometimento.

A visão e missão corporativa enfatizam a importância do conhecimento para o sucesso dos negócios. A liderança e os incentivos devem ser organizados de forma a recompensar tanto o desempenho individual quanto a contribuição para o sucesso geral da empresa. Isso gera o interesse em gerar bom desempenho não só para a própria unidade, mas também para ajudar outras unidades, clientes e fornecedores a melhorar.

Embora não haja indicadores-chave de desempenho (KPI) para a criação e transferência de conhecimento na empresa tradicional, a empresa de conhecimento mede ambos com base nas metas de negócios. A criação de conhecimento não faz sentido se estiver isolada desses objetivos. Esses indicadores são parte integrante do sistema de relatórios que mostram como o conhecimento é convertido em sucesso do negócio. Os indicadores não financeiros que se referem a clientes, colaboradores e processos ganham importância em relação aos indicadores financeiros tradicionais.

Em uma empresa de conhecimento, uma mudança significativa em relação às empresas hierárquicas tradicionais é que a posição da administração e dos especialistas é valorizada igualmente. Em uma empresa tradicional, exige-se a responsabilidade por um determinado número de funcionários ou a responsabilidade por um determinado orçamento para escalar uma posição de gerente de departamento ou chefe de departamento. Mas em uma empresa de conhecimento, a pessoa alcança sua posição na empresa pelo conhecimento que possui, pelo conhecimento que transmite aos outros, pela capacidade de treinar outros funcionários, pela capacidade de aprender coisas novas e de demonstrar expertise. A pessoa que está na posição de especialista deve desenvolver-se continuamente. As empresas de conhecimento desenvolvem «mercados de conhecimento» onde a procura e a oferta são decisivas para a criação e troca de conhecimento. Uma empresa de conhecimento alcança transparência sobre «quem sabe o quê» dentro e fora da empresa e a transferência e desenvolvimento de conhecimento são baseados em interesses comuns. As melhores práticas e competências são enfatizadas na empresa, oferecendo assim um estímulo permanente para a implementação de boas práticas. As empresas do conhecimento superaram a síndrome do «conhecimento é poder»; agora «partilha de conhecimento é poder».

Vários agentes, processos e mídias suportam as tarefas operativas em nossa visão de empresa do conhecimento. Em tal empresa, os processos de transferência de conhecimento são definidos, bem como a estrutura de desenvolvimento de novos campos de negócios, produtos e processos. Um gerente de topo promove a criação e transferência de conhecimento como «Coordenador Foco no Cliente» ou «Diretor de Gestão do Conhecimento». No entanto, esses treinadores não gerenciam o conhecimento da mesma forma que gerenciam os recursos financeiros. Em vez disso, eles garantem que a «ecologia do conhecimento» seja correta e que as regras dos mercados do conhecimento sejam seguidas. Eles promovem o crescimento dos funcionários neste novo tipo de empresa.

O conhecimento estrategicamente importante de uma organização é agrupado em redes de competências que também são responsáveis pela distribuição e proteção desse conhecimento. Os colaboradores trocam conhecimentos em «comunidades de prática». Numa empresa do conhecimento, vários projetos cooperativos promovem o trabalho em equipa entre as funções e áreas de negócio num «comportamento sem fronteiras».

Uma empresa de conhecimento pratica *benchmarking* intensivo tanto interna quanto externamente. Ele descobre as melhores práticas, as distribui, questiona sinceramente se tais práticas podem ser usadas nas unidades individuais e, se não, procura os motivos. Vários grupos de solução de problemas fornecem todas as informações disponíveis de seus funcionários. A síndrome do «não inventado aqui» é substituída por «implementar boas ideias venham de onde vierem».

O treinamento e o desenvolvimento de competências são uma alta prioridade. Os processos de aprendizagem individual e coletiva são baseados na demanda e a aprendizagem conjunta acontece em equipes próximas às situações de trabalho e unidades de negócios. Os colaboradores já não são «enviados» para formação. Em vez disso, eles mesmos controlam ativamente seu próprio processo de aprendizagem.

Enquanto os contatos informais não são apreciados na empresa hierárquica tradicional – «preferem não falar com os nossos colegas em Deli porque podem arrebatar-nos o negócio» – o trabalho de equipe e os contatos informais são promovidos na empresa do conhecimento através de feiras de conhecimento, intermediação de conhecimento, cantinas atrativas, *lounges*, *coffee corners* e outras opções de reuniões informais. Mas nem todas as opções de comunicação eletrônica são implementadas para permitir que os colegas se conheçam por meio de reuniões pessoais. Em tal empresa, o layout do escritório e a estrutura geral do local de trabalho e dos espaços sociais favorecem a interação entre os funcionários.

A tecnologia da informação e comunicação é um componente importante de uma empresa de conhecimento. Ele conecta todos os funcionários da organização, bem como clientes relevantes, fornecedores e outros especialistas externos em know-how. A mídia eletrônica é utilizada intensamente para discussão e transferência de conhecimento. As bases de dados e outras fontes de informação estão disponíveis para um acesso atualizado, completo e integrado a informação relevante que está para além dos limites das unidades funcionais e de negócio. Esses bancos de dados e fontes constroem a memória coletiva da organização. A mídia é amigável, fácil de aprender, adaptável ao método de trabalho de um indivíduo e permite contribuições fáceis (por exemplo, wikis, blogs).

O leitor bem-informado argumentará que tal empresa descrita acima não existe na realidade ou que essa utopia também não encontrará aplicação prática no futuro. Esse argumento pode ser contestado porque já existem muitas empresas que se aproximam dos critérios aqui mencionados, o que nos aproxima dessa visão. Uma dessas empresas de sucesso é a General Electric, que já avançou muito nas suas «reinvenções» para uma empresa orientada para o conhecimento e é mencionada várias vezes neste livro. A Phonak (Suíça) e a Oticon (Dinamarca), ambas fabricantes de aparelhos auditivos, exibem muitas das características de uma empresa de conhecimento mencionada aqui. A lista continua com Buckman Laboratories e Sequent Computers nos EUA, KaO no Japão, Semco no Brasil, os serviços financeiros MLP na Alemanha, Infosys, Wipro, Tata Steel, Eureka Forbes e Tata Chemicals na Índia, etc.

Para os funcionários e a administração, uma mudança para uma empresa do conhecimento significa uma mudança no método de trabalho e nos papéis descritos pelos principais representantes da aprendizagem organizacional (Argyris e Schön 1978; Senge 1990; Flood 2009). Os colaboradores deste novo contexto empresarial devem ser capazes de «aprender a aprender». Além de sua competência específica de campo, eles devem ter a capacidade básica de lidar com as novas tecnologias de informação e comunicação para obter informações o mais cedo possível e convertê-las em conhecimento. Espera-se que os funcionários tenham uma competência de comunicação distinta e a habilidade de autogerenciamento, bem como a capacidade de serem criativos e resolver problemas por conta própria. A competência social ou «capacidade de trabalhar em equipe» envolve a consulta dentro do grupo, resolução de conflitos, lidar com o stress e comportamentos inesperados dos outros. A administração é a principal responsável por organizar as condições estruturais acima mencionadas «ecologia», bem como por determinar as metas e medir o alcance das metas de acordo com os critérios estendidos de uma "empresa do conhecimento. A própria gestão é especialista – seja para um tema específico, seja para treinar os outros a aprender, ou seja, para comunicar os valores e objetivos.

Análise Breve: Adequação para Competição de Conhecimento

Avalie a forma como avalia a posição da sua empresa na competição do conhecimento entre uma «empresa orientada para o conhecimento» e uma «empresa insensível ao conhecimento». Os alunos podem fazer o mesmo com sua universidade, departamento ou trabalho em equipe com seus colegas. Uma boa abordagem para a sensibilização é copiar e distribuir o questionário fornecido abaixo entre os colegas para que os resultados possam ser discutidos posteriormente em pontos como o quão diferente a categorização acabou sendo? Onde estava a diferença máxima na classificação? Onde vemos os maiores obstáculos no caminho para uma empresa de conhecimento e quais medidas podem dar o máximo de resultados com menos esforço? Como cada um de nós pode contribuir para a distribuição do conhecimento na empresa?

1.4. Principais pontos do capítulo

- Conhecimento é um recurso e a capacidade de aprender se tornou o principal ingrediente para competitividade sustentável
- Por todo mundo, percebemos mudanças estruturais impulsionadas por uma economia e sociedade do conhecimento resultando em mudanças em sistemas educacionais, novas formas de aprendizado e valorização de talentos e competências
- A cada vez mais os ativos intangíveis vêm determinando o valor das organizações
- Autoavaliações proveem importantes insights se uma organização pode ser considerada como uma ***empresa de conhecimento***

1.5. Questões

1. Quais são as características de uma economia do conhecimento?
2. Quais são as forças que direcionam a competição baseada em conhecimento?
3. Qual é a influência dos ativos intangíveis no valor das organizações?
4. Como você definiria Gestão do Conhecimento? Descreve pelo menos cinco fatores que determinam o sucesso de gestão baseada em conhecimento.
5. Quais são os objetivos e questões básicas para a gestão baseada em conhecimento?
6. Quais são as barreiras para a criação e transferência de conhecimento dentro e através das organizações?
7. Quais são as características de empresas de conhecimento?

Empresa Insensível ao Conhecimento	1	2	3	4	5	Empresa Orientada ao Conhecimento
Nosso Mercado						
Baixa diferenciação						Requisitos de Clientes são altamente diferenciados, produtos e serviços feitos sob demanda
Produtos padrão de demanda						Valoriza produtos/serviços personalizados e de alto valor
Baixa velocidade de inovação e longos ciclos de vida						Alta velocidade de inovação e curtos ciclos de vida
Nossas soluções para os problemas dos clientes						
Intensivo em trabalho ou capital						Intensiva em Conhecimento
Pode ser facilmente imitado						São difíceis de serem imitadas
Pode ser substituído						Não pode ser substituída atualmente
A empresa enfrenta dificuldades para gerar novas áreas de negócio						A geração de novas áreas de negócio e produtos é mais efetiva que os competidores
Nossos financiadores						
Estão interessados em resultados imediatos						Estão interessados em um aumento duradouro no valor da companhia
Conhecimento e Aprendizado						
Conseguimos poucas ideias a partir dos empregados						Boas ideias são implementadas independente de suas origens
Aprendemos lentamente (de outras empresas)						Aprendemos rapidamente (a partir de outras empresas)
Não sabemos “quem sabe o quê”						Sabemos como localizar nosso conhecimento
Não empenhamos muitos esforços para proteger nosso conhecimento						Nos protegemos sistematicamente contra perda de conhecimento
Há receio em enfatizar as melhores práticas e habilidades						Enfatizamos as melhores práticas e habilidades
Treinamento não conduzem a um processo coletivo de aprendizado						Empregados controlam ativamente seus próprios processos de aprendizado
Não há institucionalização de GC						Processos de GC e regras estão implementadas

1.6. Tarefa

Empresa orientada ao conhecimento

1. Dê exemplos de empresas que apresentam características de empresas orientadas ao conhecimento, de acordo com os critérios descritos no teste acima.

Definições de Gestão do Conhecimento

2. Faça uma pesquisa na internet sobre definições de Gestão do Conhecimento e compare-as.

1.7. KM-Tool: Knowledge Café

O que é um Knowledge Café?

Um Knowledge Café é um meio de reunir um grupo de pessoas para uma conversa aberta e criativa sobre um tema de interesse mútuo, para trazer à tona seu conhecimento coletivo, compartilhar ideias e *insights* e obter uma compreensão mais profunda do assunto e das questões envolvidas.

Por que usá-lo?

Um Knowledge Café oferece um espaço para as pessoas se encontrarem, discutirem e refletirem. Em última análise, isso leva à ação na forma de uma melhor tomada de decisão e inovação e, portanto, resultados de negócios tangíveis.

Como executá-lo?

Uma sessão simples pode ser mais ou menos assim:

1. O facilitador «Dono da casa de café» dá as boas-vindas às pessoas no café e explica o que são os cafés do conhecimento e o papel da conversa na vida empresarial (máx. 15 min).

2. O facilitador gasta de 10 a 15 minutos descrevendo o assunto ou tema do café e faz uma única pergunta aberta. Por exemplo, se o tema for compartilhamento de conhecimento, a pergunta para o grupo pode ser "quais são as barreiras ao compartilhamento de conhecimento em uma organização e como você as supera?"

3. O grupo se divide em pequenos grupos de cerca de cinco pessoas cada e discute as questões por cerca de 30 a 45 minutos e, em seguida, voltamos juntos como um grupo inteiro para os 30 a 45 minutos finais, onde os grupos individuais compartilham seus pensamentos.

4. Opcionalmente, nas sessões de pequenos grupos, as pessoas mudam de mesa a cada 15 minutos para ampliar o número de pessoas com quem podem interagir e, assim, as diferentes perspectivas do grupo. Normalmente, nenhuma tentativa é feita para capturar a conversa, pois isso tende a destruí-la. O valor do café está na própria conversa e na aprendizagem que cada um leva. Em algumas circunstâncias, porém, faz sentido capturar coisas do café dependendo de sua finalidade e existem maneiras de fazer isso que interferem minimamente na dinâmica da conversa. Uma boa ideia é ter uma toalha de mesa de papel e mesas de café nas quais os participantes possam escrever, desenhar e fazer mapas mentais.

Para mais informações consulte:

- http://www.gurteen.com/gurteen/gurteen.nsf/id/run-kcafe
- http://en.wikipedia.org/wiki/Knowledge_Cafe
- www.youtube.com/watch?v=NTZ0vf0Tmi4

2. Conhecimento nas Organizações

Klaus North[1] e Gita Kumta[2]

(1) Wiesbaden Business School, Hochschule RheinMain, Wiesbaden, Alemanha
(2) School of Business Management, SVKM's Narsee Monj. Inst. de Estudos de Administração, Mumbai, Maharashtra, Índia

Resultados de aprendizagem
Depois de concluir este capítulo

- Você saberá a diferença entre informação, conhecimento e competência,
- Você será capaz de aplicar o modelo SECI de conversão de conhecimento explícito/tácito a organizações reais;
- Você será capaz de explicar a vantagem competitiva pela visão baseada em recursos usando o conceito «VRIN» e a construção de «capacidades dinâmicas»;
- Você aprenderá abordagens para estruturar o conhecimento organizacional e avaliar o valor dos recursos de conhecimento;
- Você será capaz de realizar um concurso de ideias.

2.1. Criação de Valor Baseada no Conhecimento

2.1.1. A «Escada do Conhecimento»: Informação, Conhecimento e Competência

O conhecimento nas organizações assume muitas formas. Inclui as competências e capacidades dos funcionários, o conhecimento de uma empresa sobre clientes e fornecedores, know-how para entregar processos específicos, propriedade intelectual na forma de patentes, licenças e direitos autorais, sistemas para alavancar a força inovadora da empresa e assim por diante. O conhecimento é o produto da aprendizagem individual e coletiva e se materializa em produtos, serviços e sistemas. O conhecimento está relacionado às experiências das pessoas nas organizações e na sociedade, mas apenas uma pequena parte do conhecimento é explicitado. O conhecimento tácito determina em grande parte como as pessoas se comportam e agem.

Para as empresas, o conhecimento é um recurso, um ativo intangível e faz parte do chamado capital intelectual de uma organização. Para possibilitar a criação de valor baseada no conhecimento, a administração precisa entender o que é conhecimento e como o conhecimento está relacionado à competitividade. A seguir, explicaremos a terminologia subjacente à criação de valor baseada no conhecimento, primeiro por meio de um breve estudo de caso e, posteriormente, sistematizando o relacionamento por meio da escada do conhecimento (Fig. 2.1).

Dados e Informações: Matéria-Prima para Criação de Valor

Vamos começar na parte inferior da escada de competência. As pessoas se comunicam por meio de símbolos; estes podem ser letras, números ou sinais. Esses símbolos podem ser interpretados apenas se houver regras claras de compreensão. Essas regras são chamadas de sintaxe. Símbolos mais sintaxe tornam-se dados. A combinação dos números 1, 3, 5 e os símbolos de unidade para graus Celsius mais um ponto para 13,5 °C transforma símbolos em dados. Esses dados só podem ser interpretados se for dado um significado exato. Torna-se informação se adicionarmos aos dados se falamos sobre a temperatura do ar, a hora e o local precisos dessa temperatura.

Informações são dados organizados que agregam significado a uma mensagem. Esta informação é interpretada de forma diferente dependendo do contexto, experiência e expectativas das pessoas.

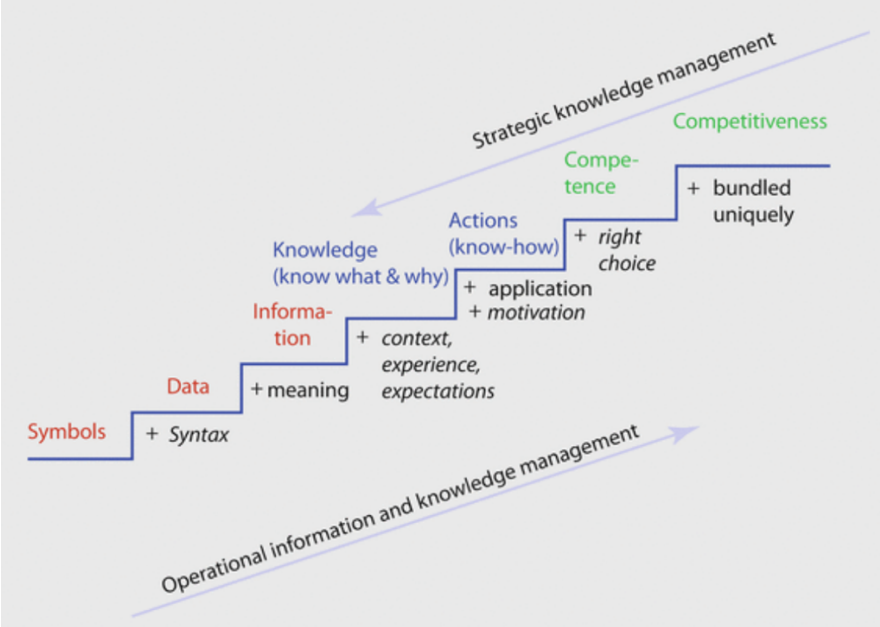


Fig. 1 Escada do Conhecimento

Conhecimento: Criando um Entendimento como Base para AgirNo desenvolvimento do conhecimento, distinguimos entre diferentes níveis. A primeira, «saber o quê», resulta da interiorização da informação. Por exemplo a leitura de um livro que só cria valor para uma organização se a pessoa souber aplicar essa informação ou seja, o «saber o quê», se transforma em «saber fazer» por via da aplicação. O quão difícil pode ser esta transferência de «saber o quê» para «saber como» é sentida por muitas pessoas que leem as instruções de operação de um telefone celular, por exemplo, e desejam aplicar as informações para programar funções específicas. Como os modelos mentais de quem escreveu as instruções de operação e de quem aplica as instruções de operação são diferentes, o usuário pode não ser capaz de interpretar as instruções corretamente. Uma solução pode ser fazer com que os usuários em potencial escrevam as instruções de operação.

O conhecimento refere-se ao entendimento tácito ou explícito das pessoas sobre as relações entre os fenômenos. Está incorporado em rotinas para o desempenho de atividades, em estruturas e processos organizacionais e em crenças e comportamentos incorporados. O conhecimento implica uma capacidade de relacionar entradas e saídas, observar regularidades na informação, codificar, explicar e, finalmente, prever (Carnegie Bosch Institute [CBI] 1995).

O conhecimento nas organizações é explícito apenas em pequena extensão. Usando a metáfora do iceberg podemos dizer que apenas a pequena parte visível acima da água é conhecimento explícito e a grande parte escondida sob a água é conhecimento tácito. De acordo com Polanyi (1966), o conhecimento tácito é pessoal, específico do contexto, muitas vezes inconsciente e, portanto, difícil de formalizar e comunicar. Conhecimento explícito ou codificado refere-se ao conhecimento que é transmissível em linguagem formal e sistemática. Polanyi diz «que podemos saber mais do que podemos contar». Veremos a seguir como a transformação do conhecimento explícito em conhecimento tácito e vice-versa é um importante processo de criação e distribuição de conhecimento.

Competência: A Ação Certa na Hora Certa

A capacidade de aplicar o conhecimento é baseada em motivos específicos («saber por quê»). As pessoas só agem se estiverem motivadas. Portanto, uma importante tarefa de gerenciamento para aumentar a criação de valor baseada no conhecimento é garantir a configuração motivacional correta para que os trabalhadores do conhecimento desenvolvam, compartilhem e apliquem seus conhecimentos de acordo com o objetivo da empresa. O valor é criado quando o conhecimento certo é aplicado no momento certo para resolver um problema

específico ou para explorar uma nova "oportunidade" de negócios. A escolha certa de conhecimento no momento certo é chamada de competência. Com von Krogh e Roos (1996) «vemos a competência como um evento, em vez de um ativo». Isso significa simplesmente que as competências não existem da mesma forma que um carro; eles existem apenas quando o conhecimento (e habilidade) cumpre a tarefa.

O termo competência (ou competência) de uma pessoa ou grupo descreve a relação entre as tarefas atribuídas ou assumidas pela pessoa ou grupo e sua capacidade e potencial para entregar um desempenho desejado. As pessoas mobilizam conhecimentos, competências e comportamentos para «fazer a coisa certa no momento certo».

A interação de um ator com um público, a habilidade de venda de um vendedor de sucesso ou a adaptação de estratégias por um consultor de experiência para atender às necessidades do momento do cliente refletem uma competência que muitas vezes também é chamada de expertise.

Competitividade: agrupar competências para exclusividade. se agruparmos as competências de pessoas ou organizações exclusivamente para que não sejam igualadas por outras organizações, então falamos de competitividade. As competências essenciais de uma organização são consideradas particularmente relevantes para a competição.

As competências essenciais (Hamel e Prahalad 1994; Rumelt 1994) são uma combinação de habilidades e tecnologias que agregam valor ao cliente. Essa combinação é baseada em conhecimento explícito e oculto e é caracterizada pela estabilidade temporal e influência nos produtos. Competências essenciais:

1. Fornecer acesso potencial a uma ampla variedade de mercados.

2. Fazer uma contribuição significativa para os benefícios do produto final percebidos pelo cliente.

3. São difíceis de imitar pelos concorrentes.

Estão em sinergia com outras competências e tornam a empresa única e melhor que as outras. Nessa visão, as competências essenciais representam a base da competitividade.

Voltando à escada do conhecimento, podemos formular o objetivo da gestão baseada no conhecimento como a transformação da informação em conhecimento e competência para criar valor mensurável de forma sustentável.

Para isso, precisamos construir cada degrau da escada do conhecimento. Como em uma escada real você não pode dizer que a escada de cima é mais importante que a escada de baixo, você tem que construir todas elas. A visão de baixo para cima reflete os processos operacionais de gestão da informação e do conhecimento, enquanto a visão de cima para baixo reflete a visão estratégica de definição das competências de uma organização e de seus membros que eventualmente levarão à competitividade.

Estudo de caso

Transferência de melhores práticas (serviços de fabricação eletrônica)

Pela manhã, a gerente da fábrica, Janya Gupta, clicou na caixa de entrada na tela do computador. Um flash de notícias mostrou a ela que os resultados do benchmarking periódico em torno das 50 unidades de fabricação eletrônica da empresa haviam sido inseridos diretamente no banco de dados de melhores práticas. Ela clicou nas notícias e obteve uma visão geral das informações formatadas graficamente. Na comparação de benchmarking, sua fábrica foi colocada na metade superior. Por correio de voz, solicitou à equipa de boas práticas da sua fábrica que analisasse a informação e estudasse a possibilidade de adotar as «melhores práticas» de outras fábricas de forma a aumentar a produtividade e assim compensar a queda constante dos preços dos componentes eletrônicos. Ela conheceu a equipe de melhores práticas à tarde e mais uma vez verificou os dados de sua fábrica que foram relatados ao banco de dados de melhores práticas. Tudo estava ok. A equipe de melhores práticas desenvolveu conhecimento sobre as diferenças estabelecendo uma relação entre as

informações de benchmarking de sua própria fábrica e das fábricas comparáveis. Através de uma videoconferência agendada em cima da hora com os membros das equipas de boas práticas de duas «fábricas irmãs» aprenderam e receberam o know-how. A equipe recebeu dicas de como alterar a configuração para montagem em sua fábrica. Os insights os motivaram a agir. Os resultados foram mensuráveis apenas 3 dias depois. A equipe de melhores práticas dos serviços de manufatura eletrônica demonstrou sua competência coletiva na solução de problemas. A gerente da fábrica, Janya Gupta, está satisfeita e destaca que, para ela, a capacidade de aprender mais rápido que a concorrência é uma vantagem competitiva duradoura.

Estudo de caso

Experiência mental: o conhecimento é uma "crença verdadeira justificada"?

O professor de filosofia Edmund Gettier questionou a teoria do conhecimento dominante entre os filósofos por milhares de anos quando definiu o conhecimento como "crença verdadeira justificada". Segundo Gettier, há certas circunstâncias em que não se tem conhecimento, mesmo quando todas as condições acima são atendidas. Gettier propôs dois experimentos mentais, que vieram a ser conhecidos como "casos Gettier", como contra-exemplos para a explicação clássica do conhecimento. Um dos casos envolve dois homens, Smith e Jones, que aguardam os resultados de suas candidaturas para o mesmo emprego. Cada homem tem dez moedas no bolso. Smith tem excelentes razões para acreditar que Jones conseguirá o emprego e, além disso, sabe que Jones tem dez moedas no bolso (ele as contou recentemente). Disto Smith infere, «o homem que vai conseguir o emprego tem dez moedas no bolso». No entanto, Smith não sabe que também tem dez moedas em seu próprio bolso. Além disso, Smith, não Jones, vai conseguir o emprego. Embora Smith tenha fortes evidências para acreditar que Jones conseguirá o emprego, ele está errado. Smith tem uma crença verdadeira justificada de que um homem com dez moedas no bolso conseguirá o emprego; no entanto, de acordo com Gettier, Smith não sabe que um homem com dez moedas no bolso conseguirá o emprego, porque a crença de Smith é «... verdadeira em virtude do número de moedas no bolso de Jones, enquanto Smith não sabe quantas moedas há no bolso de Smith e baseia sua crença... (Gettier 1963). Esses casos falham em ser conhecimento porque a crença do sujeito é justificada, mas só passa a ser verdadeira em virtude da sorte. Em outras palavras, ele fez a escolha correta (neste caso, prevendo um resultado) pelas razões erradas. Este exemplo é semelhante àqueles frequentemente dados ao discutir crença e verdade, em que a crença de uma pessoa sobre o que acontecerá pode coincidentemente estar correta sem que ela tenha o conhecimento real para baseá-la.

Fonte: Gettier (1963) citado de acordo com http://en.wikipedia.org/wiki/Epistemology

2.1.2. Campos de Atuação da Gestão do Conhecimento

A gestão do conhecimento de uma organização significa organizar todas as etapas da escada do conhecimento. Se um determinado degrau da escada não for construído (por exemplo, falta de compatibilidade de dados, disponibilidade incompleta de informações, falta de motivação para ações), a pessoa «tropeça» ao subir e descer a escada. A implementação de estratégias de negócios ou o negócio operacional é dificultado. Da escada do conhecimento deduzem-se três campos de atuação da «gestão da informação e do conhecimento»:

A gestão estratégica do conhecimento percorre a escada do conhecimento de cima para baixo para responder a perguntas sobre «quais as competências necessárias para ser competitivo», deduzindo assim que conhecimento e know-how são necessários. Os objetivos do conhecimento devem ser deduzidos dos objetivos da empresa. Além disso, a gestão estratégica do conhecimento deve desenvolver um modelo de empresa que conceitue as estruturas e processos motivacionais e organizacionais que tornam a empresa apta para a competição baseada no conhecimento.

A gestão operacional do conhecimento envolve particularmente a interligação da informação ao conhecimento, know-how e ações. A forma de organizar o processo de transferência do conhecimento individual para o conhecimento coletivo e vice-versa é decisiva para o sucesso da gestão baseada no conhecimento. Aqui, a conversão do conhecimento tácito em conhecimento explícito e vice-versa é de vital importância. No entanto,

esse processo não ocorre sem incentivos efetivos. Assim, a gestão operativa do conhecimento implica também o estabelecimento de condições facilitadoras que sirvam como estimulantes para a criação, distribuição e uso do conhecimento.

A gestão da informação e dos dados (Digitalização) é a base da gestão do conhecimento. Se olharmos para a escada do conhecimento, notamos que o fornecimento, o armazenamento e a distribuição da informação são pré-requisitos para a criação e transferência do conhecimento. A partir de pesquisas, pudemos constatar que muitas empresas começam a caminhar para a gestão do conhecimento com medidas de gestão de informações e dados, mas acabam percebendo que a tecnologia de informação e comunicação não pode ser usada de forma otimizada sem condições organizacionais e motivacionais adequadas.

2.1.3. Avaliação de Maturidade GC

As organizações variam no grau de maturidade de sua gestão baseada em conhecimento (ver Fig. 2.2). A consciência da importância de gerir os recursos de conhecimento é um processo de aprendizagem e depende da «maturidade» das organizações. A mudança para uma organização baseada no conhecimento «inteligente» é um esforço progressivo que envolve algumas «tentativas e erros».

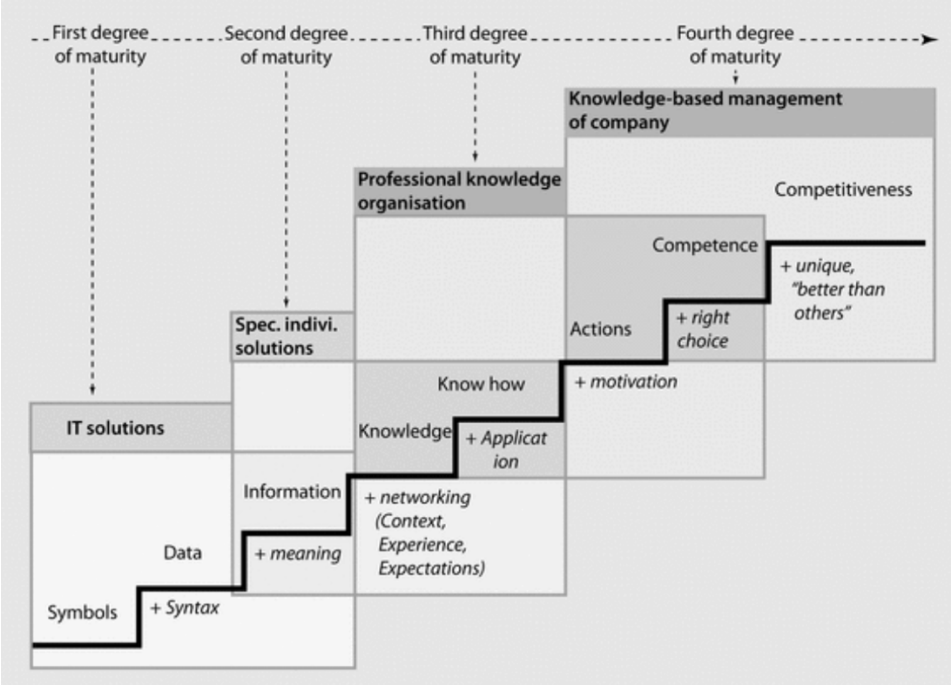


Fig 2.2 Grau de maturidade de Gestão baseada em Conhecimento de uma Organização

Para avaliar o estado atual de desenvolvimento e fornecer orientação para maior evolução em direção a uma organização baseada no conhecimento, vários modelos de maturidade foram desenvolvidos. Em geral, um modelo de maturidade descreve o desenvolvimento de uma entidade ao longo do tempo.

Um modelo de maturidade de gestão do conhecimento pode ser considerado como uma abordagem estruturada para a implementação da gestão do conhecimento. Um modelo de maturidade também pode fornecer um entendimento comum das terminologias envolvidas na implementação da gestão do conhecimento para várias partes interessadas.

Com base em estudos empíricos, identificamos quatro níveis de maturidade na forma como as organizações gerenciam seu conhecimento (North e Schmidt 2004):

O primeiro nível: gerenciamento de informações

As empresas no primeiro estágio de maturidade concentram-se na gestão da informação. Eles implementam uma infra-estrutura de informação e comunicação para permitir acesso específico a bancos de dados e documentos. As medidas organizacionais de acompanhamento para promover o intercâmbio de conhecimento ainda não foram estabelecidas ou foram estabelecidas apenas em certa medida. Os esforços estão concentrados em tecnologia da informação e comunicação. Nesse nível de maturidade, as organizações obtêm maior transparência e agilidade nos processos, evitam a duplicação de trabalho e abreviam os períodos de treinamento de novos entrantes, o que resulta em um aumento geral da qualidade dos produtos e serviços. Exemplos de sistemas de GC de primeiro nível: Implementação de uma intranet, desenvolvimento de plataformas comunitárias.

O segundo nível: soluções «Ilha»

As organizações que intencionalmente implementam iniciativas de gestão do conhecimento em áreas ou unidades de negócios específicas representam o segundo estágio de maturidade. Eles perceberam que a tecnologia de informação e comunicação por si só não é suficiente para uma gestão baseada no conhecimento. Em vez disso, eles entenderam que um "caso de negócios" é necessário para demonstrar que a gestão do conhecimento produz benefícios claros. Assim, são desenvolvidas soluções específicas em áreas específicas, por ex. conhecimento do serviço, conhecimento do pessoal e conhecimento do cliente. As soluções de GC contribuem para acelerações de processos (resposta rápida, por exemplo, às consultas dos clientes), aumento da reutilização do conhecimento interno (nem sempre a roda se reinventa), bem como melhoria do trabalho em equipe e aumento da qualidade. Mesmo esta abordagem pode levar a ganhos rápidos onde são criadas «ilhas GC» que podem ser difíceis de integrar em uma estratégia abrangente de GC posterior.

Exemplos de sistemas de GC de segundo nível são o estabelecimento de Sistemas de Gestão de Relacionamento com Clientes integrados na gestão de vendas ou um portal com «dicas e truques» para técnicos de serviço para os quais os técnicos de serviço contribuem ativamente.

O terceiro nível: Organização do conhecimento profissional

As organizações no terceiro estágio de maturidade são aquelas que implementaram uma organização de conhecimento profissional em todos os departamentos e unidades de negócios e exibem as seguintes características:

- A infraestrutura de informação e comunicação garante fácil disponibilidade de informações relevantes.
- Os funcionários são motivados e recompensados por compartilhar conhecimento.
- Integração da gestão do conhecimento nos objetivos de negócios, processos e organização do projeto.
- A troca de conhecimento é apoiada por Comunidades de Prática (CoPs) e centros de competência.
- Os benefícios da gestão do conhecimento são medidos.

Uma distribuição equilibrada de benefícios resultando em processos aprimorados, maior motivação dos funcionários e satisfação do cliente é uma característica típica da organização do conhecimento profissional.

Exemplos de sistemas GC de terceiro nível são o estabelecimento de papéis e responsabilidades GC em níveis centralizados/descentralizados de uma organização. Os funcionários são treinados regularmente sobre como usar as ferramentas GC.

Nesse terceiro nível, a GC é vista como um conjunto de regras e ferramentas para melhorar o desempenho. No entanto, ainda não está totalmente integrado nas mentes e no comportamento das pessoas.

O quarto nível: cultura do conhecimento

O quarto nível de maturidade representa uma condição ideal que foi alcançada apenas por algumas organizações até agora. Esse nível de maturidade é caracterizado por valores profundamente compartilhados,

trabalho em equipe, troca ativa de conhecimento além das fronteiras dos departamentos e além da empresa, busca ativa por inovação e uma cultura aberta e confiável que é preenchida e vivida pela administração e funcionários de forma consistente. Um componente importante dessa cultura é aprender de fora (por exemplo, mercados, tecnologias, rivais, fornecedores, clientes etc.) e de dentro. A cultura da empresa é apoiada por um sistema maduro de informação e comunicação e mídia como CoPs, centros de competência e work-outs. A colaboração, o compartilhamento do conhecimento e a busca contínua pela inovação fazem parte dessa cultura do conhecimento. Valores compartilhados, não ferramentas, impulsionam a criação, transferência e proteção do conhecimento. Essas empresas atingem níveis gerais de excelência. Estariam no nível 5 da autoavaliação de GC proposta por Collison e Parcell (2004) em seu guia prático «Learning to fly».

Estudo de caso

Evolução da gestão do conhecimento na Eureka Forbes Ltd.

O caso da Eureka Forbes Ltd., uma corporação de múltiplos produtos e canais de USD 250 milhões e líder em sistemas domésticos e industriais de purificação de água, limpeza a vácuo e soluções de purificação de ar na Índia, demonstra como uma abordagem em fases ajuda a obter vantagem competitiva. É pioneira em vendas diretas na Índia e é a maior organização de vendas diretas da Ásia. Sua força de vendas diretas de 7.000 pessoas atinge cerca de 1,5 milhão de lares indianos, adicionando 1.500 clientes diariamente. Tem operações em mais de 135 cidades e 500 vilas em toda a Índia. «Uma função formal de GC existe na empresa há mais de sete anos e passou por diferentes fases. A gestão do conhecimento evoluiu de ser vista como um trabalho adicional para ser reconhecida como uma vantagem estratégica, impactando significativamente tanto a linha superior quanto a linha inferior», diz Shubha Ashraf, gerente de conhecimento da Eureka. A primeira fase foi o período inicial de estabelecimento do capital intelectual estrutural como função e processos de GC para facilitar que as pessoas conheçam e sejam capazes de apreciar que isso ajuda um indivíduo a ter um desempenho mais rápido e melhor. A fase seguinte foi a «acrescentar valor» ao capital estruturante através da criação de um portal que permite diversos canais e funcionalidades de captação de pessoas. O foco mudou de ser uma plataforma de contato para ser uma plataforma de capacitação para os clientes internos, melhorando assim o capital intelectual humano. A terceira fase concentra-se na melhoria do capital intelectual social, alavancando o conhecimento reunido para melhorar a capacidade de resposta do mercado, a satisfação do cliente e do funcionário. A Eureka Forbes Ltd. ganhou o prêmio MAKE e em janeiro de 2010 foi reconhecida e distinguida por três Prêmios UNESCO-Water Digest de Melhor P&D e avanço tecnológico para um novo produto.

2.2. Dimensões do Conhecimento

Para «gerir» o conhecimento nas organizações precisamos de perceber com que tipo de «espécies» estamos a lidar. Vamos, portanto, olhar mais de perto as seguintes três dimensões do termo «conhecimento».

- «Natureza» do conhecimento: O que é o conhecimento? É considerado um objeto, um resultado que pode ser partilhado, duplicado e transportado como «comida congelada» ou é um processo individual difícil de controlar?
- «Disponibilidade» do conhecimento: Em que formas o conhecimento se torna disponível e acessível dentro e entre as organizações? Aqui, trataremos particularmente da diferença entre conhecimento individual versus conhecimento coletivo e conhecimento tácito versus conhecimento explícito.
- «Valor» do conhecimento: Qual é o valor do conhecimento? Frequentemente, o conhecimento é também identificado como componente de ativos intangíveis ou como «Capital Intelectual». Conhecimento é capital. A questão é como o conhecimento pode ser medido?

	Estratégia de GC	Comportamentos de Liderança	Networking	Aprendendo antes, durante e depois	Capturando Conhecimento
Nível 5	Ativos intelectuais claramente Identificada Estratégia de GC está incorporada na estratégia de negócio Framework e ferramentas permitem aprendizado antes, durante e depois.	Líderes reconhecem a ligação entre GC e desempenho As atitudes corretas existem para compartilhar e usar know-how de outros Líderes reforçam os comportamentos corretos e ações como regras de modelo	Papéis e responsabilidades claramente definidos Redes e CoPs têm um propósito claro, alguns têm entregas claras, outros desenvolvem capacidade na organização. As redes se reúnem anualmente	Prompts para aprendizado incorporados aos processos de negócios. As pessoas rotineiramente descobrem quem sabe e conversam com eles. Linguagem, modelos e diretrizes comuns levam a um compartilhamento eficaz nos processos de negócios. As pessoas rotineiramente descobrem quem sabe e conversam com eles. Linguagem, modelos e diretrizes comuns levam a um compartilhamento eficaz.	O conhecimento é fácil de obter, fácil de recuperar. Conhecimento relevante é empurrado para você. É constantemente atualizado e destilado. As redes atuam como guardiãs do conhecimento.
Nível 4	Discussões em andamento sobre os ativos intelectuais da organização. Existe uma estratégia de GC, mas não está vinculada aos resultados do negócio. Uma estrutura clara e um conjunto de ferramentas para o aprendizado são amplamente comunicados e compreendidos.	KM é responsabilidade de todos; alguns trabalhos são dedicados à gestão do conhecimento. «Compartilhar conhecimento é poder.» Os líderes estabelecem expectativas «fazendo as perguntas certas» e recompensando os comportamentos certos.	As redes são organizadas em torno das necessidades de negócios. As redes têm um documento de governança claro. A tecnologia de suporte está em vigor e é bem utilizada.	Aprender antes, durante e depois é a forma como fazemos as coisas por aqui. «Clientes» e parceiros participam em sessões de avaliação.	O conhecimento just-in-time é atual e facilmente acessível. Um indivíduo destila e atualiza, embora muitos contribuam. Esse indivíduo age como o proprietário.
Nível 3	Não há estrutura ou estratégia articulada de GC. Algumas descrições de trabalho incluem captura, compartilhamento e destilação de conhecimento. As pessoas estão usando várias ferramentas para ajudar no aprendizado e no compartilhamento.	A KM é vista como responsabilidade de uma equipe especializada. Alguns líderes falam por falar, mas nem sempre cumprem o que prometem!	As pessoas estão em rede para obter resultados. As redes são criadas	As pessoas podem descobrir facilmente o que a empresa sabe. Exemplos de compartilhamento e uso são reconhecidos. Os colegas estão ajudando os colegas além dos limites organizacionais.	As redes assumem a responsabilidade pelo conhecimento, reúnem o conhecimento de seus sujeitos em um só lugar em um formato comum. Pesquisar antes de fazer é encorajado. Pouca ou nenhuma destilação.
Nível 2	A maioria das pessoas diz que compartilhar know-how é importante para o sucesso das organizações. As pessoas estão usando algumas ferramentas para ajudar no aprendizado e no compartilhamento	Alguns gerentes dão às pessoas tempo para compartilhar e aprender, mas há pouco apoio visível do topo.	Rede ad hoc para ajudar indivíduos que se conhecem.	As pessoas aprendem antes de fazer e programar sessões de revisão. Eles capturam o que aprendem para que outros possam acessar. Na prática, poucos o acessam.	As equipes capturam as lições aprendidas após um projeto. As equipes buscam conhecimento antes de iniciar um projeto. Acesso a muito conhecimento, mas não resumido.
Nível 1	Algumas pessoas expressam que o know-how é importante para a organização. Pessoas isoladas e apaixonadas por KM começam a falar e compartilhar como é difícil.	KM visto como uma moda passageira de gestão. Os líderes são céticos quanto aos benefícios. Os líderes acham que o networking leva à falta de responsabilidade. "conhecimento é poder"	Os acumuladores de conhecimento parecem ser recompensados.	As pessoas estão conscientes da necessidade de aprender com o que fazem, mas raramente têm tempo. Compartilhar é para o benefício da equipe.	Algumas pessoas reservam um tempo para registrar suas lições em vários armários e bancos de dados. Eles raramente são atualizados, poucos contribuem, menos ainda pesquisam.

2.2.1. Natureza do Conhecimento

Von Krogh e Roos (1996, p. 334) contrastam três epistemologias com três perspectivas de conhecimento em uma empresa:

A epistemologia do processamento da informação assume que conhecimento e informação são aproximadamente a mesma coisa. Neste caso é natural investir na velocidade de processamento da informação. Nessa perspectiva, o aumento da capacidade de processamento da informação leva ao aumento do desenvolvimento do conhecimento também na empresa. As organizações que apostam nesta epistemologia irão investir em sistemas de informação e comunicação como o relançamento ou otimização da sua intranet.

A epistemologia da rede assume que o conhecimento é resultado da interação das pessoas nas redes. Assim, a empresa deve investir para aproximar os funcionários da organização. Consequentemente, quanto maior for o número de oportunidades para as pessoas se encontrarem e trocarem, maior será o desenvolvimento do conhecimento. As organizações que se concentram nesta epistemologia promoverão comunidades de prática e outras redes sociais, criarão zonas de encontro e oportunidades para as pessoas se encontrarem (por exemplo, brown bag lunch).

A epistemologia autorreferencial assume que o conhecimento é um processo particular dependente da história dentro de cada um de nós. O conhecimento de uma pessoa é um mero dado bruto para outra. Cada pessoa compartilha conhecimento organizacional com outra. Assim, é necessário encontrar um contexto que estimule o diálogo contínuo na organização. As empresas que se concentram nesta epistemologia promoverão pequenas equipes e forças-tarefa, criarão grupos de solução de problemas do tipo "work-out" e fornecerão aos especialistas ambientes estimulantes (veja, por exemplo, o design e o layout do escritório do Google Zurich).

Von Krogh e Roos preferem a última perspectiva da criação do conhecimento. No entanto, eles enfatizam que toda organização funciona de acordo com as três epistemologias em diferentes momentos e para diferentes funções. Portanto, o conhecimento pode pertencer tanto à posição extrema viz. «conhecimento é objeto» e «conhecimento é processo» dependendo da situação. Por exemplo, se o vendedor souber o número de seus clientes classe A, trata-se de uma informação com as características de um objeto. No entanto, o conhecimento apresenta mais características de um processo se a informação disponível sobre o cliente for melhor utilizada para fechar negócios. Gardner (1995) descreveu esses diferentes aspectos com os termos «saber o quê», «saber como», «saber por quê», «saber onde» e «saber quando». Polanyi (1966) enfatizou a perspectiva do processo com a seguinte declaração:

O conhecimento é uma atividade melhor descrita como um processo de conhecimento.

A perspectiva extrema de «conhecimento é objeto» e «conhecimento é processo» talvez fique mais clara se dividirmos a nova palavra «capital do conhecimento» em seus dois componentes, viz. conhecimento e capital e descubra a diferença entre esses dois termos (consulte a Fig. 2.3). Sveiby (1997) argumenta que a analogia entre conhecimento e capital não ajuda na criação e transferência de conhecimento porque leva a uma falsa compreensão do conhecimento (ver Fig. 2.3).

Capital	Conhecimento
• Independente de pessoa • Diminui quando compartilhado • É amortizado em investimento • Estático (objeto) • Fácil de medir	• Depende da pessoa • Cresce quando distribuído/compartilhado • Amplia em valor quando usado • Dinâmico (processo) • Difícil de medir

Para criar uma organização baseada no conhecimento – uma perspectiva de processo de conhecimento deve ser adotada. Consequentemente, é necessário desenvolver condições facilitadoras que estimulem a criação e transferência de conhecimento.

Para além destas diferentes perspetivas – «conhecimento é objeto» e «conhecimento é processo» – a natureza do conhecimento é determinada por duas características. O conhecimento pode ser privado e individual para uns e público e coletivo para outros. Além disso, o conhecimento pode estar presente em formas tácitas e explícitas. Esses aspectos determinam a disponibilidade do conhecimento.

Estudo de caso

Integração de conhecimento: Assumindo uma empresa estrangeira

O problema:

Uma empresa alemã adquire uma empresa francesa com aproximadamente 500 funcionários para obter rapidamente know-how adicional. Do lado alemão, as negociações de aquisição são conduzidas pelo departamento de «Fusões e Aquisições» (M&A). Após a conclusão do contrato, uma unidade de negócios operacional assume a tarefa de integrar a nova subsidiária francesa na empresa sem ter experiência anterior. Embora a M&A conheça a empresa francesa, ela só se envolve informalmente em uma maior integração após a conclusão do contrato.

Os especialistas franceses se opõem à fusão. O valor da aquisição seria reduzido devido ao atrito. O conhecimento é documentado rudimentarmente. O comprador alemão tem apenas alguns funcionários francófonos que podem preencher a lacuna para a nova subsidiária ou integrar os funcionários franceses em suas equipes. Há muita diferença entre a cultura da empresa alemã e a da empresa francesa de médio porte. A nova empresa-mãe alemã envia uma equipa de gestão de alto nível para assumir a gestão da subsidiária francesa. É quando os problemas começam.

Elementos da solução:

Como o processo de integração pode ser organizado de forma mais eficaz? O valor da aquisição é decidido pelo know-how dos funcionários. Portanto, é útil não apenas alertar as Fusões e Aquisições em um estágio inicial, mas também tomar ações que edifiquem a fé, por exemplo. encorajar os colaboradores de ambas as empresas a conhecerem-se, identificando os importantes detentores de conhecimento e/ou equipas e influenciando positivamente a sua atitude perante a fusão. Após a conclusão das negociações, especialistas experientes do departamento de M&A devem iniciar o coaching do processo de integração. Além disso, a estruturação contínua de um processo de M&A e o processo de integração são úteis. Para garantir o sucesso, é fundamental que o conhecimento e os portadores do conhecimento não sejam vistos como objetos de uso livre mediante a assinatura de uma compra.

Atribuição:

Identificar fusões transfronteiriças ou transregionais. Quais foram as razões para o fracasso ou sucesso?

2.2.2. Disponibilidade e Conversão do Conhecimento: Modelo SEICI

A «disponibilidade» do conhecimento é afetada pela forma, tempo e lugar. A forma não envolve apenas o aspecto «conhecimento individual versus coletivo», mas também o aspecto «conhecimento tácito versus conhecimento explícito». Ambos os aspectos estão estreitamente interligados (Hedlund e Nonaka 1993; Nonaka e Takeuchi 1995).

A forma de organizar a transferência do conhecimento individual para o conhecimento coletivo e vice-versa é decisiva para o sucesso da gestão baseada no conhecimento. «Uma empresa é um lugar onde o conhecimento individual e a inteligência individual convergem para formar uma inteligência coletiva e criativa que pode ser colocada em uso empresarial» (Morin, 1997).

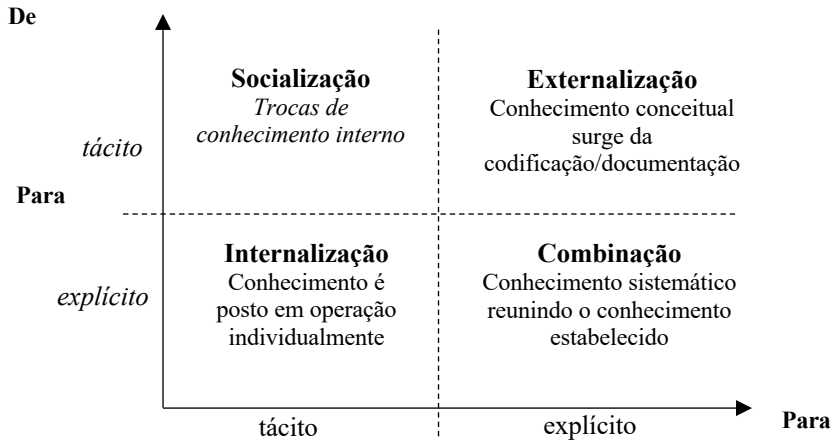
Existem dois tipos de conhecimento para descrever esse processo: conhecimento explícito e conhecimento tácito.

O conhecimento tácito representa o conhecimento pessoal de um indivíduo. Baseia-se na educação, ideais, valores e sentimentos de cada pessoa. Os insights subjetivos e a intuição incorporam o conhecimento tácito que está profundamente enraizado nas ações e experiências da pessoa em particular. O termo «conhecimento tácito» foi introduzido pela primeira vez na filosofia por Michael Polanyi, observando que «podemos saber mais do que podemos dizer» (Polanyi 1966). Essa forma de conhecimento é muito difícil de formular e transmitir porque está incorporada nos indivíduos. O conhecimento tácito é transmitido, entre outras coisas, durante nossa educação, na qual assumimos os padrões de comportamento dos pais sem saber.

Ao contrário do conhecimento tácito, o conhecimento explícito é metódico e sistemático e está presente de forma articulada. Ele é armazenado na mídia fora do cérebro (conhecimento desencarnado) de um indivíduo e pode ser transferido e armazenado por meio de tecnologia de informação e comunicação. Exemplos de conhecimento explícito são descrições detalhadas de processos, patentes, árvores organizacionais, documentos de qualidade, etc.

Nonaka e Takeuchi expressaram que a conversão do conhecimento tácito em conhecimento explícito é o problema básico da gestão do conhecimento. A razão é que o conhecimento é útil para uma empresa e pode ser usado por indivíduos ou grupos apenas se estiver presente de forma explícita. Assim, sob esse ponto de vista, cabe à gestão do conhecimento organizar e dirigir um processo de geração de conhecimento organizacional. Nonaka e Takeuchi formularam isso da seguinte forma: «Por criação de conhecimento organizacional, entendemos a capacidade de uma empresa como um todo para criar novos conhecimentos, distribuídos por toda a organização e incorporados em produtos, serviços e sistemas» (Geração de conhecimento organizacional significa a capacidade de uma empresa gerar conhecimento completamente novo, distribuí-lo dentro da organização e incorporá-lo em produtos, serviços e sistemas) (Nonaka e Takeuchi 1995; von Krogh et al. 2000).

Nonaka e Takeuchi (1995) assumem que o conhecimento é criado por meio da interação entre conhecimento tácito e explícito por quatro modos diferentes de conversão, conforme mostrado na Fig. 2.4. Explicaremos todas as quatro formas de conversão do conhecimento, pois são a base para a criação de valor.



Socialização: do conhecimento tácito ao conhecimento tácito

A conversão do conhecimento tácito de uma pessoa para o conhecimento tácito de outra pessoa é chamada de socialização. É um processo de compartilhar experiências e, assim, criar conhecimento tácito, como modelos mentais compartilhados e habilidades técnicas. A socialização ocorre quando um aprendiz observa um mestre, quando um consultor recém-contratado é integrado a um grupo de projeto e aprende por observação, imitação e prática. A experiência partilhada é a chave da socialização e da criação de valor nas organizações baseadas no

conhecimento. Muitas vezes, a mera transferência de informações fará pouco sentido se for abstraída das emoções associadas e dos contextos específicos nos quais as experiências compartilhadas estão inseridas.

Externalização: do tácito ao explícito

A externalização é o processo de articulação do conhecimento tácito em conceitos explícitos. A externalização acontece quando descrevemos um processo de fabricação para fins de certificação ISO 9000. Na consultoria de gestão, por exemplo, a externalização ocorre quando um perfil de projeto é escrito para fornecer informações específicas sobre o desenvolvimento do projeto e as lições aprendidas como base para futuros projetos similares. Muitas empresas têm esse tipo de lições aprendidas em bancos de dados. Como a externalização revela apenas uma parte do conhecimento tácito, é bom não "confiar exclusivamente nessas declarações escritas, mas permitir, por exemplo, consultores que precisam planejar um novo projeto para obter um contato pessoal com aqueles que já realizaram projetos semelhantes antes. Da mesma forma, um processo real sempre será diferente da descrição formal do projeto. A externalização é a base para refletir experiências, para processos de aprendizagem formalizados e, finalmente, para padronização e melhoria de processos.

Combinação: do conhecimento explícito ao conhecimento explícito

A combinação refere-se à conversão de conhecimento explícito em conhecimento explícito. Os indivíduos trocam e combinam conhecimentos por meio de documentos, reuniões, redes de comunicação. Eles reconfiguram as informações existentes por meio da classificação, adição, combinação e categorização do conhecimento explícito que pode levar a novas informações. Na consultoria, por exemplo, diferentes apresentações são combinadas e reconfiguradas com o propósito de uma apresentação de vendas para um novo cliente. A combinação de conhecimento explícito com conhecimento explícito geralmente segue uma economia de reutilização e também é a base para uma estratégia inovadora cumulativa em que os produtos e processos são aprimorados de forma incremental.

Internalização: do conhecimento explícito ao conhecimento tácito

Internalização é o processo de incorporação do conhecimento explícito em conhecimento tácito. Está intimamente relacionado com aprender fazendo. Um engenheiro de serviço, por exemplo, lê um manual de operação para programar equipamentos eletrônicos. Grande parte dos nossos processos formais de aprendizagem acontece por internalização. De acordo com o modelo de Nonaka e Takeuchi, a criação do conhecimento é uma interação contínua e dinâmica entre conhecimento tácito e explícito que acontece no nível do indivíduo, do grupo, da organização e entre organizações.

É, portanto, uma importante tarefa gerencial criar oportunidades de interações entre esses níveis para que a conversão do conhecimento possa acontecer. De acordo com Nonaka e Takeuchi, as condições de habilitação são:

Intenção

O elemento mais crítico da estratégia corporativa é conceituar uma visão sobre que tipo de conhecimento deve ser desenvolvido e torná-lo operacional em um sistema de gestão para implementação.

Autonomia

No nível individual, todos os membros de uma organização devem ser autorizados a agir de forma autônoma na medida em que as circunstâncias permitirem. Isso pode aumentar a chance de introduzir ideias inesperadas e oportunidades tácitas.

Flutuação e caos criativo

Isso significa adotar uma atitude aberta em relação aos sinais ambientais, explorar a ambigüidade, a redundância desses sinais e usar a flutuação para quebrar rotinas, hábitos ou estruturas cognitivas.

Redundância

Nas organizações empresariais, a redundância refere-se à sobreposição intencional de informações sobre atividades comerciais, responsabilidades de gerenciamento e a empresa como um todo. O compartilhamento de

informações redundantes promove o compartilhamento do conhecimento tácito e, assim, acelera o processo de criação do conhecimento.

Variedade Necessária

Com base na suposição de que a diversidade interna de uma organização deve corresponder à variedade e complexidade do ambiente para lidar com os desafios impostos pelo ambiente, todos na organização devem ter acesso rápido às informações e conhecimentos necessários. Quando existem diferenciais de informação dentro da organização, os membros da organização não podem interagir em igualdade de condições; isso dificulta a busca por diferentes interpretações de novas informações.

Nonaka e Takeuchi assumiram um modelo de «espiral do conhecimento» para transformar conhecimento tácito em conhecimento explícito e para transferir conhecimento de um indivíduo para um grupo ou organização. O ponto de partida da espiral é o funcionário individual e sua capacidade de criar conhecimento. Ao se comunicar com os funcionários de um grupo, o funcionário individual cede seu próprio conhecimento (externalização) e o transfere para outros. Por outro lado, o indivíduo internaliza o histórico de experiências de todo o grupo (internalização). A contínua exteriorização e interiorização do conhecimento entre colaboradores e equipas dentro e fora da organização conduz à oferta de conhecimento a estes vários níveis e resulta no crescimento do conhecimento da organização. A comunicação pessoal entre os funcionários e o uso de tecnologia de informação e comunicação é um pré-requisito para todo esse processo. A espiral do conhecimento passa por quatro fases, conforme mostrado na Fig. 2.5.

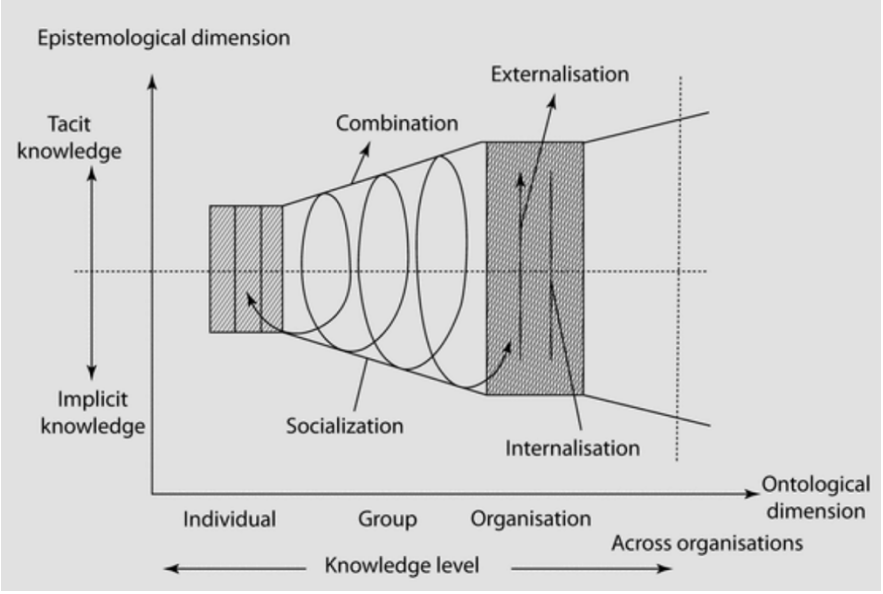


Fig 2.5 A Espiral de Criação e Transferência de Conhecimento Organizacional (Nonaka and Takeuchi 1995)

- Na fase de socialização (troca de conhecimento tácito), o conhecimento interno, por ex. modelo mental ou habilidades técnicas são gerados.
- A fase de externalização do conhecimento (do tácito ao explícito) produz o conhecimento conceitual e o novo.
- A fase de combinação (combinação de conhecimento explícito) desenvolve conhecimento sistemático que se manifesta em protótipos, novos métodos ou novas ideias de negócios.
- A fase de internalização do conhecimento (do explícito ao tácito) gera o conhecimento operativo.

O estudo de caso a seguir, «O melhor pão de Osaka», explica as fases acima individualmente.

No entanto, o referido modelo que descreve a conversão do conhecimento de privado para coletivo e implícito para explícito não considera a distribuição desigual do conhecimento na empresa causada por barreiras

estruturais ou motivacionais na organização. Por outro lado, o conhecimento existente não está disponível no local desejado na hora desejada.

A gestão do conhecimento, portanto, não deve se restringir apenas ao processo de aprendizagem individual e organizacional como tal, mas também remover os obstáculos na informação e na comunicação. Em termos positivos, a gestão deve criar condições que promovam o compartilhamento do conhecimento, garantir a interação dos processos de aprendizagem individual e organizacional. von Hipple (1994) e Szulanski (1966) usam o termo «aderência» para descrever o fato de que o conhecimento está disponível gratuitamente apenas até certo limite. O conhecimento tem tendência a «grudar». Deve ser posto à tona por medidas adequadas de design organizacional. Veremos mais de perto esse tópico nas seções sobre transferência de conhecimento e mercado de conhecimento.

A disponibilidade de conhecimento está ligada ao fator tempo e lugar. Os profissionais não estão disponíveis 24 horas por dia em todo o mundo, especialmente em empresas que operam globalmente. Um problema de software que aparece em uma subsidiária na Europa às vezes pode não ser resolvido porque o especialista na Índia não está disponível ou porque está de férias. Em uma indústria que depende de respostas rápidas, por ex. empresas de consultoria, a McKinsey montou uma rede de resposta rápida em seu centro de prática. Os «consultores de plantão» desta rede garantem uma resposta qualificada em 24 horas a uma questão específica de um dos cerca de 60 escritórios em 28 países (Peters 1994). Em Eureka Forbes, o vendedor no campo exige uma resposta rápida às suas perguntas. Ele usa seu celular para se conectar e obter suas respostas dos consultores de plantão que têm acesso ao repositório de conhecimento.

A tomada de decisão requer um conjunto completo de informações e conhecimentos atualizados. Hoje, em muitas empresas, há um atraso considerável na disponibilização de informações atualizadas e, assim, as decisões de hoje para as ações de amanhã são baseadas em conhecimentos obsoletos. A disponibilidade de conhecimento e informações atualizados em toda a empresa é de vital importância, especialmente para empresas que estão cercadas por um ambiente de mercado em rápida mudança.

Além disso, a disponibilidade de conhecimento é afetada pelo local onde o conhecimento se origina ou onde um indivíduo busca conhecimento. Apesar da mídia eletrônica, conhecer as pessoas pessoalmente e a confiança resultante são necessários para a troca de conhecimento. É difícil construir essa confiança em grandes distâncias geográficas sem encontrar as pessoas pessoalmente regularmente. Para além destes aspetos mais motivacionais, a criação de centros de conhecimento locais e globais e a sua interligação é uma importante tarefa estratégica das empresas internacionais (Bartlett e Ghoshal 1989; Doz 1997).

A gestão do conhecimento, portanto, não deve se restringir apenas ao processo de aprendizagem individual e organizacional como tal, mas também remover os obstáculos na informação e na comunicação. Em termos positivos, a gestão deve criar condições que promovam o compartilhamento do conhecimento, garantir a interação dos processos de aprendizagem individual e organizacional. von Hipple (1994) e Szulanski (1966) usam o termo «aderência» para descrever o fato de que o conhecimento está disponível gratuitamente apenas até certo limite. O conhecimento tem tendência a «grudar». Deve ser posto à tona por medidas adequadas de design organizacional. Veremos mais de perto esse tópico nas seções sobre transferência de conhecimento e mercado de conhecimento.

A disponibilidade de conhecimento está ligada ao fator tempo e lugar. Os profissionais não estão disponíveis 24 horas por dia em todo o mundo, especialmente em empresas que operam globalmente. Um problema de software que aparece em uma subsidiária na Europa às vezes pode não ser resolvido porque o especialista na Índia não está disponível ou porque está de férias. Em uma indústria que depende de respostas rápidas, por ex. empresas de consultoria, a McKinsey montou uma rede de resposta rápida em seu centro de prática. Os «consultores de plantão» desta rede garantem uma resposta qualificada em 24 horas a uma questão específica de um dos cerca de 60 escritórios em 28 países (Peters 1994). Em Eureka Forbes, o vendedor no

campo exige uma resposta rápida às suas perguntas. Ele usa seu celular para se conectar e obter suas respostas dos consultores de plantão que têm acesso ao repositório de conhecimento.

A tomada de decisão requer um conjunto completo de informações e conhecimentos atualizados. Hoje, em muitas empresas, há um atraso considerável na disponibilização de informações atualizadas e, assim, as decisões de hoje para as ações de amanhã são baseadas em conhecimentos obsoletos. A disponibilidade de conhecimento e informações atualizados em toda a empresa é de vital importância, especialmente para empresas que estão cercadas por um ambiente de mercado em rápida mudança.

Além disso, a disponibilidade de conhecimento é afetada pelo local onde o conhecimento se origina ou onde um indivíduo busca conhecimento. Apesar da mídia eletrônica, conhecer as pessoas pessoalmente e a confiança resultante são necessários para a troca de conhecimento. É difícil construir essa confiança em grandes distâncias geográficas sem encontrar as pessoas pessoalmente regularmente. Para além destes aspetos mais motivacionais, a criação de centros de conhecimento locais e globais e a sua interligação é uma importante tarefa estratégica das empresas internacionais (Bartlett e Ghoshal 1989; Doz 1997).

2.3. Memória Organizacional

Nas organizações, o conhecimento não está embutido apenas em documentos ou repositórios, mas também em rotinas, processos, práticas e normas organizacionais. Esta observação descreve de forma sucinta um dos grandes problemas das organizações atuais.

Em um contexto dinâmico, que prioriza a produtividade, muitas das vezes as organizações chegam a reconhecer o valor da criação e disseminação do conhecimento, mas não percebem que seus objetivos, valores e estratégias impedem a aplicação da Gestão do Conhecimento. Estudos indicam a importância de um ambiente que permita o compartilhamento de informações, conhecimentos e vivências como apoio à solução de problemas cada vez mais complexos e que possibilite a reutilização de conhecimentos, artefatos e produtos anteriormente gerados.

Mais que disseminar o conhecimento gerado, as organizações necessitam de redes sociais, partindo de um mapeamento das competências. Conhecendo as habilidades da organização, o segundo passo é saber como abordá-las, envolvê-las nas resoluções dos problemas. Este envolvimento, desde que assumido por ambas as partes, possibilita à organização respostas mais rápidas às mudanças e problemas, além de prover o aprendizado organizacional e estimular a criatividade dos envolvidos.

Embora nos últimos anos tenha crescido o investimento no apoio à Gestão do Conhecimento através de sistemas de informação e novas metodologias, a aplicação deste ferramental ainda é problemática. Ente as dificuldades apresentadas estão o impacto sobre a produtividade dos envolvidos, a dificuldade das organizações em integrar suas estratégias às práticas de Gestão do Conhecimento. A falta de esclarecimento sobre a forma como estas estratégias serão atendidas pelos sistemas de Gestão do Conhecimento, e a falta de materiais que mostrem como se aplicam, na prática, os conceitos, também interferem na aplicação de tais sistemas nas organizações.

As organizações intensivas em conhecimento, como as empresas de desenvolvimento de software, merecem uma profunda análise por sofrerem constantes mudanças ligadas ao negócio, envolvendo diferentes pessoas e afetando às diversas fases do desenvolvimento e sobretudo, envolvendo diferentes objetivos e tipos de conhecimento. A Gestão do Conhecimento nestas organizações levaria à redução de tempo, de custos, e ao aumento da qualidade, reduzindo erros e o retrabalho, alcançando o sucesso a partir de informações dos projetos anteriores.

Entretanto, no cenário atual, observa-se que as equipes de TI adquirem valores individuais, não compartilham e repetem os mesmos erros de projetos anteriores. As abordagens utilizadas para melhoria de

processo como o CMM, não definem explicitamente como implantar efetivamente a gestão do conhecimento, embora sugira esta atividade como uma boa prática. O aprendizado através da prática, sem se ater a experiências anteriores, pode ser um risco para o projeto levando a erros e retrabalho. Entretanto, isso é comum entre estas organizações. A falta de um histórico do projeto, o fato de participantes do projeto possuírem conhecimento individual e não o explicitarem também são problemáticos.

Para obter bons resultados, as organizações devem fazer uma análise profunda em seus problemas, estabelecer seus objetivos e estratégias antes de implementar seu sistema de Gestão do Conhecimento.

A Memória Organizacional (MO) pode ser classificada como o registro de dados, informações e conhecimento úteis para a organização, de forma que ela possa reaproveitá-los. O valor estrategicamente elevado da Memória Organizacional se explica pelo aumento do grau de importância do conhecimento na sociedade atual, criando a necessidade de existência de uma memória institucional que o armazene, de maneira organizada.

A Memória Organizacional pode ser definida como o registro de uma organização representado por seus documentos e artefatos, desprezando explicitamente a memória das pessoas. A Memória Organizacional integra técnicas básicas em um sistema computacional que continuamente coleta, atualiza e estrutura conhecimento e informação, e os provê em diferentes atividades operacionais de forma sensível ao contexto, intencionada e ativa.

A Memória Organizacional pode estar retida na cultura organizacional, nas transformações organizacionais, na ecologia organizacional, nos arquivos externos, em manuais corporativos, nas bases de dados, nas histórias organizacionais e nos indivíduos.

De fato, não se pode afirmar que todo o conhecimento da organização está registrado em documentos, sejam eles de papel ou eletrônicos. A experiência dos membros da organização e suas ideias, valores e decisões não podem ficar à margem da Memória Organizacional. No entanto, um dos maiores desafios da Gestão do Conhecimento é exatamente capturar esse conhecimento individual, tão arraigado à pessoa e de difícil registro.

2.4. Culturas e valores organizacionais

Antes de mostrarmos, exatamente, o que esta dimensão nos traz, vamos esmiuçar os termos que dela fazem parte. Assim, entendendo suas partes, será mais fácil compreender o seu todo.

O que podemos entender por cultura? De acordo com o Dicionário Aurélio Cultura significa: Ato, efeito ou modo de cultivar; desenvolvimento intelectual; saber; utilização industrial de certos produtos naturais; estudo; elegância; esmero (sociológico) sistema de atitudes e modelos de agir, costumes e instruções de um povo. Conhecimento Geral. Muitos significados para uma só palavra, contudo tendo em vista a lógica a qual estamos sujeitando esta palavra e o contexto no qual a estamos estudando, podemos considerar o seu significado sociológico o mais adequado, uma vez que estamos lidando com pessoas presentes em uma organização.

Valor organizacional, por sua vez, pode ser definido como o conjunto de princípios que direciona as políticas e práticas utilizadas pela organização no seu cotidiano, os parâmetros para a tomada de decisões e para a hierarquização das significâncias entre os processos e o total de atenção dispensada durante o trabalho e a condução gerencial.

Feitas estas definições podemos dar continuidade ao entendimento desta dimensão que, por ora, anda desvalorizada pelos gestores por relacionar-se à vertente considerada soft do mundo empresarial. No entanto, consideramos que sob uma visão de gestão pelo conhecimento, tratar dimensão é de suma importância para o corpo da empresa.

Terra (2000, p.102) define acultura organizacional como sendo o conjunto de normas e valores que ajudam a nortear eventos e avaliar o que é realmente apropriado ou não para o empreendimento. Estas normas e valores, contudo, podem ser vistas, ainda, como um sistema de controle capaz de atingir grande eficácia, e porque não eficiência, uma vez que levam a um alto grau de conformação, ao mesmo tempo em que incitam uma elevada sensação de autonomia. Sob esta lógica, nos contrapomos aos modelos antigos que, no lugar de desenvolverem a sensação de autonomia, impunham restrição constante aos seus funcionários.

É importante conceber que a construção desta cultura organizacional, baseada na cultura e valores fomentados pela equipe, uma vez bem direcionada, torna-se um canal de comunicação e consenso coletivo, que sob uma ótica criativa e inovadora, transforma o ambiente em mais ameno e propulsor do crescimento.

Todavia, existe uma relação importante neste percurso, que podemos chamar de chefes-subordinados. Caso a chefia não saiba direcionar bem o processo, pode ocorrer a perda da identidade e a massificação de uma equipe. Para que vocês, futuros gestores e empreendedores, não caiam nesta armadilha gerada pela lógica diretiva, disponibilizaremos um quadro (Tabela 6.3) com fatores impeditivos e meios para contorná-los, elaborado por Duailibi & Simonsen no livro Criatividade e Marketing

Tabela 6.3 - Estimulando Ambientes Criativos

Fatores Impeditivos criatividade	Sugestões para os gerentes superarem barreiras à criatividade
- Pressão para se conformar - Atitudes e meio excessivamente autoritários - Medo do ridículo - Intolerância para com as atitudes mais joviais - Excesso de ênfase nas recompensas e nos sucessos imediatos - A busca excessiva de certezas - Hostilidade para com a personalidade divergente - Falta de tempo para pensar - Rigidez da organização	- Condições para um aprendizado autogerador, isto é, para que as pessoas criativas a empresa obtenham estímulos em si mesmas; - Tome cuidado para que o meio não seja autoritário em excesso - Pressione para o subordinado *superaprender* - Na medida do possível, postergue os seus julgamentos, mesmo quando experiências, sem ciúme profissional nem superioridade - Estimule a flexibilidade intelectual encarando a solução de qualquer problema sob várias formas; - Encoraje a auto-realização do processo individual, permitindo que o próprio subordinado analise o seu trabalho e o seu desenvolvimento - Ajude seu pessoal a se tornar mais sensível; - Propicie oportunidades para que todos exercitem sua criatividade - Auxilie cada subordinado a compreender, aceitar e superar os seus fracassos - Insista para que os problemas sejam abordados como um todo

FONTE: Duailibi & Simonsen apud (Terra, 2000)

Como bem definido na Tabela 6.3, o que possibilita uma gestão eficiente quando o assunto é cultura e valores organizacionais é o espaço concedido pelo gerente para que sua equipe inove e crie, contanto que ao final o produto seja o almejado e desejado por todos.

Não devemos perder de vista, quando tratamos da gestão do conhecimento, alguns aspectos fundamentais para que tanto a criatividade quanto a inovação ocorram no ambiente empresarial. O primeiro aspecto considerado é o tempo. O tempo é fator preponderante, pois se sabe que trabalhar de forma criativa é extremamente fatigante

Conforme Von Fag e Terra (2000) a atividade criativa precisa ser intercalada com rotineiras. Outro fator relevante quando se trata de tempo os prazos exigidos. O gestor precisar estar em consonância com as atividades requeridas para que suas datas sejam coerentes com a exigência feita, caso contrário, o nível de energia física e mental cairá a tal ponto que as expectativas lançadas sob a tarefa não renderão. Para tal adequação o planejamento participativo é uma arma infalível.

O outro aspecto chave, que mencionamos anteriormente, refere-se ao espaço de trabalho, que, ao contrário do que se achava, está diretamente ligado ao processo de aprendizado organizacional, à criatividade e ao **clima organizacional**3 que propicia a inovação empresarial. Sob a nova tendência de gestão, os espaços antes tidos como fechados e as ideias geradas para a organização hierárquica perdem espaço para os espaços abertos e não hierárquicos, de acordo com Terra (2000). Para Quinn Terra (2000) a abordagem atual baseia-se nos **skunk works**[4] na intenção de emular o ambiente inovativo de pequenas empresas.

Para que a criatividade e inovação ganhem espaço na gestão voltada para o conhecimento, aspectos básicos como os **fatores higiênicos**[5] devem ser contemplados a fim de que um ambiente sadio favoreça o desabrochar da criatividade e os princípios inovadores tomem corpo no desenvolvimento tanto individual quanto coletivo, tendo sempre como norte os princípios culturais a o estabelecimento dos valores organizacionais.

Em geral, ao tentarmos condensar todas estas "receitas" em práticas contextualizadas, emerge um estilo democrático, que nega o pré-julgamento de ideias, que possibilita às pessoas testarem suas ideias e, de forma geral, possibilita que elas convivam bem com o erro (Terra, 2000).

Estudo de caso

O melhor pão em Osaka

Em 1985, os desenvolvedores de produtos da Matsushita Electric Company em Osaka ponderaram sobre a construção de uma máquina de fazer pão para uso doméstico. Mas o protótipo não conseguia amassar a massa adequadamente e assá-la completamente. Apesar de todos os esforços, a crosta externa queimou enquanto o pão permaneceu cru por dentro. Foi quando o desenvolvedor de software, Ikuko Tanaka, teve uma ideia brilhante. O Osaka International Hotel desfrutou da glória de fazer o melhor pão de Osaka. Tanaka pensou em usar isso em benefício da empresa. Ela foi ao mestre padeiro do hotel para observar sua técnica de amassar e viu como o mestre padeiro esticava a massa de uma maneira particular. Após um ano de experimentos em estreita colaboração com os engenheiros do projeto, Tanaka finalmente mudou as características de construção da máquina (adicionando nervuras especiais dentro da caixa) de forma que o dispositivo imitasse efetivamente a técnica de amassar do padeiro e assasse a massa da maneira que Tanaka aprendera no hotel. O resultado foi o "método de amassar" exclusivo da Matsushita e um produto que quebrou todos os recordes de vendas para novos dispositivos de cozimento apenas no primeiro ano. Assim, Tanaka converteu o conhecimento tácito do padeiro em conhecimento explícito na forma de uma especificação clara para a máquina de fazer pão. Ikuko Tanaka primeiro adquiriu o conhecimento interno do mestre padeiro do hotel (socialização). Ela então converteu esses segredos em conhecimento explícito que poderia passar para os membros de sua equipe e outros na Matsushita (externalização). A partir daí, a equipe padronizou esse conhecimento, fundiu-o em um guia e um manual de instruções e deixou o produto moldar de acordo (combinação). Finalmente, as experiências de Tanaka

---

[3] O clima organizacional constitui o meio interno, a atmosfera psicológica característica de cada organização e está ligado ao moral e à satisfação das necessidades dos participantes e pode ser saudável ou doentio, negativo ou positivo, satisfatório ou insatisfatório, dependendo de como os participantes se sentem em relação à organização.

[4] Skunk Works são espaços propositadamente informais e desconectados do ambiente coorporativo, que propiciam maior identidade entre o colaborador e seu espaço de trabalho.

[5] Fatores Higiênicos são considerados como os fatores mínimos que uma pessoa deve receber de sua empresa, para que ela se esforce na realização de suas atividades. Dentre estes estão: condições de trabalho e conforto; salário; benefícios; segurança no cargo e políticas de organização e administração. Este conceito é difundido por Frederic Hersberg e tais fatores incidem diretamente na motivação do funcionário, contudo, ele acredita que uma vez que eles sejam saciados, aumentá-los não gerará maior motivação.

e dos membros da equipe durante a construção do novo produto aumentaram sua própria base de conhecimento tácito (internalização).

Fonte: O estudo de caso segue a descrição em Nonaka 1991

2.4.1. A Dimensão de Valor do Conhecimento

A década de 1980 presenciou o início de um pensamento baseado na constatação de que o valor de mercado das empresas aumentava em relação ao seu valor contábil. Os peritos interrogaram-se sobre como é que se explicava este gap – denominado «goodwill» – e concluíram que a diferença entre o valor de mercado e o valor contabilístico pode ser atribuída ao valor dos ativos intangíveis, que é definido na Norma Internacional de Contabilidade (IAS 38) da seguinte forma (http://www.iasplus.com/pt/standards/standard38):

Definição

Um ativo intangível é um ativo não monetário identificável sem substância física. Um ativo é um recurso controlado pela entidade como resultado de eventos passados (por exemplo, compra ou autocriação) e do qual se esperam benefícios econômicos futuros (entradas de caixa ou outros ativos). [IAS 38.8] Assim, os três atributos críticos de um ativo intangível são identificabilidade, controle (poder de obter benefícios do ativo) e benefícios econômicos futuros (como receitas ou custos futuros reduzidos).

A seguradora sueca Skandia e o Canadian Imperial Bank of Commerce foram as primeiras empresas que desenvolveram uma nova estrutura de capital empresarial. Na sua abordagem, o capital financeiro era complementado pelo «capital intelectual».

Definição

O capital intelectual é definido como conhecimento que pode ser convertido em valor (Edvinsson e Sullivan 1996; Edvinsson e Malone 1997) ou como recurso utilizado na criação de valor futuro sem uma incorporação física (OCDE 2008).

O conhecimento é considerado parte dos ativos intangíveis. Esta integra a gestão do conhecimento na atual lógica de gestão dos recursos financeiros e físicos e ajuda a estruturar e a medir o tipo de "conhecimento" disponível nas organizações.

A analogia do «conhecimento é capital» é intrigante. No entanto, tende a ignorar o caráter do conhecimento como um processo, como já discutimos em «Natureza do conhecimento».

O termo «ativos intangíveis» abrange outros recursos para criação de valor que não estão no cerne do «capital intelectual». Assim, a base de clientes, a imagem de uma empresa ou o valor das marcas é apenas em certa medida «conhecimento convertido em valor». No entanto, esses elementos podem ser adicionados ao valor dos ativos intangíveis.

O conhecimento de e sobre os clientes que é acessível à empresa, bem como o conhecimento dos funcionários sobre os clientes, processos, tecnologias, etc., fazem parte do capital intelectual. Funcionários e clientes não pertencem à empresa da mesma forma que os ativos tangíveis – o controle é restrito. É por isso que o valor dos empregados não é contabilizado no balanço (ver Fig. 2.6).

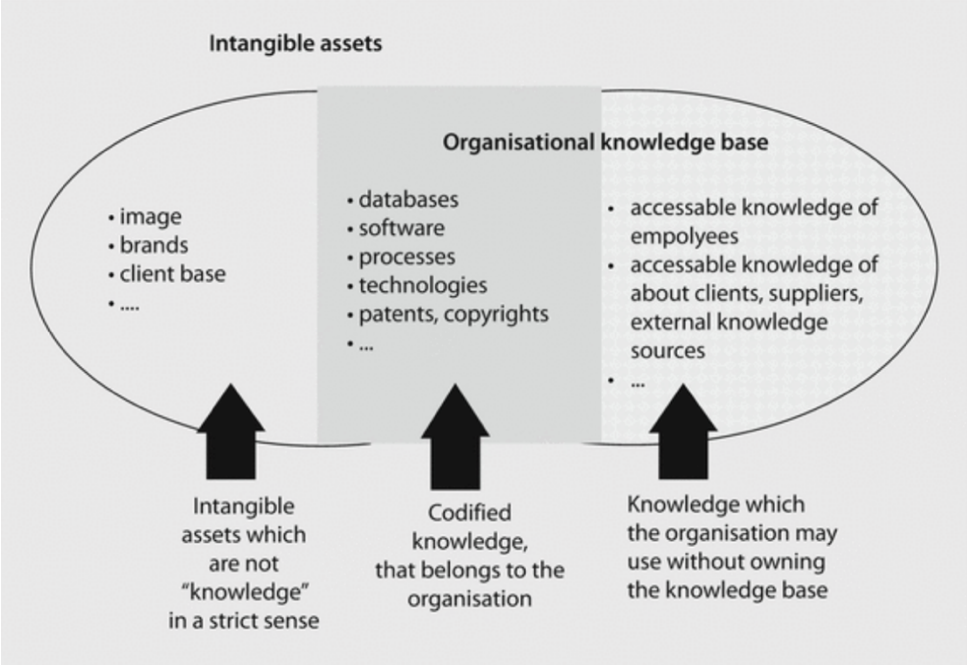


Fig. 2.6 A base de conhecimento organizacional faz parte dos ativos intangíveis

Como o conhecimento pode ser estruturado do ponto de vista do capital intelectual e quais fatores determinam o valor do conhecimento?

Seguindo os passos da Skandia, ao estruturar o capital da empresa, o valor de mercado de uma empresa é descrito pelo capital financeiro e pelo capital intelectual (Skandia 1998). O capital intelectual, por sua vez, é dividido em capital humano, capital do cliente e capital organizacional.

O capital humano é composto pelas competências da força de trabalho, sua motivação, bem como relações e valores. Em suma, podemos dizer: Capital humano = competência × motivação.

O capital do cliente representa o valor do relacionamento da empresa com o cliente. Saint-Onge define o capital do cliente como a profundidade (penetração), largura (cobertura) e o vínculo (lealdade) da base de clientes (Bontis 1996). Os exemplos de capital do cliente são pacientes de um médico, base de clientes de uma empresa de vendas por correspondência, redes de agências de um banco e seus relacionamentos com clientes. Sveiby enfatizou que as relações entre fornecedores e distribuidores também devem ser incluídas nesta categoria de capital (Sveiby 1997).

A terceira categoria de capital intelectual é o capital organizacional ou estrutural. A Skandia dividiu o capital organizacional em capital de inovação, capital de processo e cultura. O valor combinado dos processos de criação de valor é registrado no capital do processo. Isso inclui, por exemplo, o valor do processo de pedido do cliente ou o valor do processo de aquisição. O valor do processo de compras baseia-se no conhecimento dos colaboradores do departamento de compras sobre os mercados fornecedores, na sua capacidade de negociação com os fornecedores, na estruturação do ciclo do processo desde os pedidos de compra até à procura de um fornecedor e gestão das relações com os fornecedores. O conhecimento está vinculado aos bancos de dados, software, bem como valores e definição de metas dos funcionários do departamento de compras.

Costuma-se dizer que o capital estrutural é o capital «que sobra quando os empregados vão para casa». Temos que observar, no entanto, que esse capital ganha vida e tem valor apenas com os funcionários. Embora as informações codificadas nos bancos de dados, o software e o processo garantam as operações diárias, são inúteis em grande medida se houver uma fuga maciça de cérebros.

O capital de inovação, o segundo pilar do capital estrutural, é definido pela Skandia como a força de renovação de uma empresa e é evidente na propriedade intelectual protegida como patentes, licenças ou marcas e virtudes intangíveis que permitem fluxos de caixa futuros. Isso contém, por exemplo, a valorização da criatividade. A estrutura do capital organizacional da Skandia é ilustrada na Fig. 2.7.

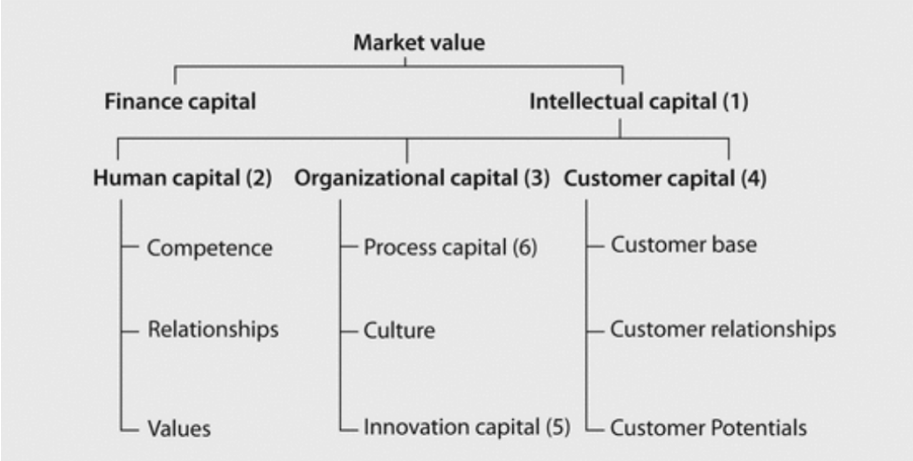


Fig. 2.7 Estrutura do Capital Organizacional da Skandia (Skandia 1998)

Critérios para avaliar o valor do conhecimento

Acima, explicamos como dividir o conhecimento em componentes que podem receber um valor sob certas condições. A seguir, trataremos da questão de como um valor pode ser atribuído ao conhecimento e quais critérios influenciam isso.

O valor do conhecimento é medido principalmente com base na escassez e no potencial de criação de valor desse recurso. Muitas vezes é difícil para ambos – o «vendedor» e o «comprador» avaliar o potencial de criação de valor do conhecimento (por exemplo, qual é o valor de uma patente?, quanto estou disposto a pagar por um consultor de tecnologia?)

Ao avaliar o conhecimento, o «vendedor» do conhecimento pode tomar como primeira orientação os esforços envolvidos na aquisição do conhecimento. «Tenho investido tanto tempo e dinheiro na aquisição deste conhecimento. Agora, quero vendê-lo por um valor maior, se possível.

Os custos internos irrecuperáveis incorridos, por exemplo, no treinamento de funcionários ou na formação de uma equipe operacional no desenvolvimento de software são significativos apenas até certo ponto ao determinar o valor do recurso de conhecimento. Isto porque, em primeiro lugar, as despesas efectuadas pela empresa nem sempre podem ser apuradas em termos de custo. Em segundo lugar, as despesas podem ter aumentado devido a uma formação ineficaz e medidas de formação avançada, ou o conhecimento obtido pode já não ter qualquer valor devido às rápidas mudanças no mercado. Deste ponto de vista, a valoração dos recursos de conhecimento com base na despesa é inadequada. Por outro lado, o «comprador» do conhecimento não tem a certeza do valor potencial que pode ser acrescentado pelo conhecimento transferido. Esse é um problema básico das consultorias porque o cliente – principalmente no caso de consultoria orientada a processos – compra processos de aprendizagem sem um resultado garantido.

Uma orientação melhor pode ser considerar o custo de substituição de um ativo intelectual: o que vai me custar para construir uma equipe de pesquisa eficaz» e relacionar isso com o potencial de criação de valor da equipe.

As seguintes questões-chave introduzem os «vendedores» de conhecimento, os «compradores» de conhecimento e os investidores na avaliação do conhecimento:

- **Usuários do conhecimento:** Para qual finalidade utilizo o conhecimento e qual é o «potencial de agregação de valor» relacionado a esse conhecimento?
- **«Vendedores» de conhecimento:** Qual foi o meu custo para adquirir este conhecimento e como posso torná-lo valioso no mercado? Investidor: Como o conhecimento desta empresa contribuirá para seu sucesso no mercado? Qual é a relação entre o valor de mercado e o custo de reposição?

Os vendedores de conhecimento, compradores de conhecimento e investidores avaliarão o conhecimento implicitamente por meio de uma série de critérios que discutiremos a seguir:

Especificidade

Assumimos que quanto mais específico o conhecimento, maior o seu valor. Os usuários valorizam soluções prontas e sob medida para seus problemas. O conhecimento que contribui para isso será mais valorizado do que os princípios gerais. Isso leva, por exemplo, a uma discussão estratégica em empresas de consultoria sobre o valor de metodologias padronizadas versus assessoria individualizada.

A **validade do conhecimento** pode ser vista de uma perspectiva de conteúdo e de uma perspectiva de tempo. A perspectiva do conteúdo refere-se à maneira como o conhecimento é criado e validado.

- Conhecimento cientificamente aceito que tem validade universal sob condições precisamente definidas
- Julgamentos e avaliações que podem ser rastreados objetivamente
- Experiências individuais ou coletivas e potenciais de atuação derivados de tais experiências.

Argumenta-se que o custo de aquisição do conhecimento – e em certo sentido, o valor – é menor para o conhecimento aceito e maior para o conhecimento potencial. Os pesquisadores farmacêuticos podem comprar conhecimento aceito na forma de um banco de dados científico a um preço relativamente baixo, mas o custo da modelagem molecular ou da aquisição de conselhos de especialistas experientes será muito maior. Portanto, o valor de uma equipe de pesquisa ou de uma aliança estratégica com um laboratório deve ser estimado como significativamente maior do que o acúmulo de conhecimento aceito.

A **validade temporal do conhecimento** refere-se ao seu «prazo de validade». Uma base de conhecimento tecnológico geral tem uma validade mais longa do que o conhecimento de mercado, que pode cair para valor zero em apenas alguns dias ou semanas.

Ainda outro critério de avaliação do conhecimento é sua singularidade ou seu valor de escassez. No entanto, deve haver uma demanda correspondente quando o conhecimento é avaliado dessa maneira. Um especialista pode ser a única pessoa com conhecimento sobre um assunto específico sem que haja qualquer demanda por seu conhecimento. Igualmente importante é a velocidade com que esse conhecimento pode ser imitado ou substituído.

Todas essas perspectivas são consideradas na valorização do conhecimento.

Estudo de caso

O valor do conhecimento

O trator de um fazendeiro parou de funcionar. Todos os esforços do fazendeiro e seus amigos para consertar o trator foram em vão. Por fim, o fazendeiro decidiu chamar um mecânico. O mecânico deu uma olhada no trator, ligou o motor de partida, levantou o capô do motor e verificou cada detalhe. Finalmente, o mecânico pegou seu martelo. Com um único golpe do martelo em um determinado local, o trator voltou a funcionar. O motor funcionou como se nunca tivesse quebrado. Quando o mecânico entregou uma fatura ao fazendeiro, o fazendeiro ficou completamente chocado e zangado e disse: «O quê? Você quer cinquenta Toman por um golpe de martelo! O mecânico disse: «Meu caro amigo, cobro apenas um Toman pelo golpe do martelo. Mas tenho que cobrar quarenta e nove Toman por saber onde atacar.

2.5. Conhecimento como Fator Competitivo

2.5.1. Teoria da Firma Baseada no Conhecimento

Morin reconhece a empresa como um lugar onde o conhecimento individual e a inteligência individual convergem para formar uma inteligência coletiva e criativa que pode ser colocada em uso empresarial. Deste ponto de vista, as empresas existem porque estão em condições de converter o conhecimento individual em conhecimento coletivo e empregá-lo para fins empresariais. Assim, o negócio é bem-sucedido:

- Se os indivíduos disponibilizarem seu conhecimento e experiência relevantes para a operação da firma e
- Se houver um processo efetivo de transformação do conhecimento do nível individual para o coletivo e
- Se as atividades estiverem alinhadas com o espírito empreendedor para atingir os objetivos da empresa.

No entanto, essa descrição de empresa do ponto de vista do conhecimento não explica a existência da empresa. Os indivíduos podem se reunir para compartilhar seus conhecimentos, criar conhecimento coletivo e, assim, realizar negócios (Spender 1996; Grant 1996; Tsoukas 1996; Kogut e Zander 1992). Segundo Grant (Grant 1996), a existência de uma empresa decorre da capacidade restrita do cérebro humano de adquirir, armazenar e processar conhecimento. Isso dá origem à especialização individual em diversas áreas do conhecimento. No entanto, oferecer soluções complexas para problemas requer esforços coordenados de vários especialistas. Os mercados sozinhos são incapazes de assumir o papel dessa coordenação porque não podem mobilizar conhecimento tácito e não podem responder ao risco de roubo de propriedade intelectual (no caso de conhecimento explícito) por um potencial comprador de conhecimento. Assim, as empresas existem porque são capazes de criar condições que favoreçam a produção de bens e serviços e permitam aos indivíduos integrar os seus saberes especializados. Assim, uma tarefa importante da gestão baseada no conhecimento de uma empresa é criar condições para que funcionários com conhecimento específico estejam em condições de criar conhecimento coletivo e implementá-lo para garantir o sucesso do negócio.

2.5.2. Conhecimento como Fator Competitivo Estratégico

Mas como garantir o sucesso empresarial em um ambiente competitivo? Nesse sentido, o conhecimento é cada vez mais considerado como um fator competitivo estratégico. Isso formou pontos de vista complementares - a visão baseada no mercado (Porter 1985) e a visão baseada em recursos (Penrose 1959; Hamel e Heene 1994) posteriormente desenvolvida pela teoria das «capacidades dinâmicas» (Teece 2009; Teece et al. 1997, 2000).

A visão relacionada ao ambiente ou baseada no mercado (Porter 1985) assume que a vantagem competitiva surge da distribuição desigual de informação e conhecimento entre as empresas e o posicionamento resultante das empresas em seu setor. Como as empresas individuais estão à frente dos concorrentes em termos de informação e conhecimento, elas reconhecem as oportunidades de mercado antes da concorrência. Por terem as competências correspondentes, convertem essas oportunidades em negócios. Nessa perspectiva, o empreendedorismo envolve a detecção de diferenças relevantes de informação e conhecimento, bem como a conversão dessa diferença em negócios. Mas isso resulta em uma competição dinâmica onde as ações da empresa de sucesso são imitadas e assim as vantagens competitivas são continuamente perdidas e torna-se necessário identificar novos desenvolvimentos em informação e conhecimento, bem como implementá-los nas atividades empreendedoras. Assim, esse tipo de competição exige que a empresa seja mais rápida que seus concorrentes, ao mesmo tempo em que é difícil construir uma vantagem competitiva duradoura.

Na visão baseada em recursos (Penrose 1959; Nelson e Winter 1982), as empresas obtêm vantagens competitivas sendo e agindo de forma diferente de seus concorrentes. Ao contrário da abordagem orientada para o ambiente, esta abordagem permite uma diferenciação contínua entre as empresas. Essas diferenciações são difíceis de imitar. Considerando o potencial dos recursos para alcançar vantagem competitiva contínua, Barney (1992) reviu-os em quatro critérios que são frequentemente abreviados como «VRIN»:

- Valioso (para o cliente)

- Raro em comparação com os rivais
- Imitável imperfeitamente devido a condições históricas únicas, ambiguidade causal e complexidade social
- Não substituível

Os dois últimos critérios são vistos como particularmente relevantes para alcançar vantagens competitivas contínuas. Os obstáculos à imitação surgem em primeiro lugar porque o conhecimento é codificado, mas legalmente protegido, por ex. marcas ou patentes. Em segundo lugar, porque o conhecimento existe de forma tácita e por meio de fatos, mesmo o conhecimento explícito está relacionado a pessoas e grupos de pessoas. Os obstáculos na imitação estão ligados direta ou indiretamente ao conhecimento ou ao desenvolvimento do conhecimento. Além disso, argumenta-se que os ativos intangíveis são a verdadeira fonte de força competitiva e fatores-chave na adaptabilidade da empresa devido a três razões a seguir: os ativos intangíveis são difíceis de acumular, podem ser usados várias vezes simultaneamente e são entradas e saídas das atividades empresariais (Itami e Roehl 1987).

Isso também vale para ambientes de negócios em movimento rápido, abertos à competição global e caracterizados pela dispersão nas fontes geográficas e organizacionais de inovação e manufatura? Teece (2009, p. 4) argumenta que a vantagem sustentável requer mais do que a propriedade de ativos difíceis de replicar (conhecimento). De acordo com Teece, isso também requer as chamadas «capacidades dinâmicas?» únicas e difíceis de replicar. Esses recursos podem ser aproveitados para criar, estender, atualizar, proteger e manter continuamente relevante a base de ativos exclusiva da empresa.

Definição

As capacidades dinâmicas são a capacidade de reconfigurar, redirecionar, transformar e moldar e integrar adequadamente as competências centrais existentes com recursos externos e ativos estratégicos e complementares para enfrentar os desafios de um mundo schumpeteriano de competição e imitação pressionado pelo tempo e em rápida mudança (Teece et al. 2000).

Para fins analíticos, as capacidades dinâmicas podem ser desagregadas na capacidade (1) de perceber e moldar oportunidades e ameaças, (2) aproveitar oportunidades e (3) manter a competitividade por meio do aprimoramento, combinação, proteção e, quando necessário, reconfiguração dos ativos intangíveis e tangíveis da empresa. Os recursos dinâmicos incluem recursos corporativos difíceis de replicar, necessários para se adaptar às mudanças nas oportunidades tecnológicas e dos clientes. Eles também abrangem a capacidade da empresa de moldar o ecossistema que ocupa, desenvolver novos produtos e processos e projetar e implementar modelos de negócios viáveis. (Teece 2009)

Como essas vantagens competitivas são desenvolvidas a partir de fatores de produção que podem ser comprados no mercado? Consideremos o seguinte exemplo:

> Um laboratório recruta graduados (fator de produção) no mercado de trabalho e os integra em uma equipe de funcionários experientes em P&D, a fim de desenvolver um grupo especializado e inovador de desenvolvedores. A equipe torna-se um recurso difícil de imitar devido aos valores compartilhados e ao entendimento tácito. O laboratório estabeleceu rotinas e processos de tecnologia e gerenciamento de projetos ao longo dos anos, através dos quais as habilidades e competências individuais das equipes de P&D são organizadas para oferecer serviços de desenvolvimento únicos e difíceis de imitar. O conteúdo e o tipo de trabalho de desenvolvimento são continuamente refletidos em um diálogo estratégico com os principais institutos de pesquisa e clientes. Com base nisso, novas áreas de conhecimento são integradas e, assim, é assegurado um enriquecimento e um enriquecimento das competências essenciais existentes. As capacidades dinâmicas são desenvolvidas para sustentar a singularidade.

Fica claro neste capítulo que, comparado aos recursos físicos, o conhecimento é um recurso empresarial mais difícil de imitar e mais raro, que oferece um potencial muito alto de geração de valor. O conhecimento é

cada vez mais considerado como «uma justificação da existência»; como fator determinante para existência e tamanho de uma empresa.

A análise do que são as organizações deve ser fundamentada na compreensão do que elas sabem fazer (Kogut e Zander 1992).

2.5.3. Impacto das Práticas de Gestão do Conhecimento no Desempenho

«Quais são os benefícios da Gestão do Conhecimento?» é uma pergunta frequentemente feita pela administração.

Vários estudos relacionando processos, práticas e casos de negócios de GC ao desempenho organizacional fornecem a resposta. Estudos que fornecem uma visão geral dessas relações foram publicados por Zack *et al.* 2009, Andreeva e Kianto 2012, Inkinen 2016).

A evidência empírica sobre a associação entre GC e desempenho da empresa é baseada em três linhas de pesquisa. Inkinen (2016) descreve os dois primeiros:

Em primeiro lugar, estudos que investigam como os processos de conhecimento (aquisição, compartilhamento e utilização) que normalmente ocorrem nas empresas, mesmo sem intervenção gerencial sistemática, estão relacionados com vários resultados de desempenho da empresa. Em segundo lugar, os estudos se concentram no impacto de práticas organizacionais e gerenciais conscientes, com a intenção de atingir os objetivos organizacionais por meio da gestão eficiente e eficaz dos recursos de conhecimento da empresa (Andreeva e Kianto 2012). Um terceiro tipo de estudos analisa casos de negócios concretos e estabelece uma relação entre intervenção e resultado (North e Hornung 2003; North *et al.* 2004).

O impacto dos processos, práticas, casos de negócios e desempenho organizacional da GC são bem resumidos por Andreeva e Kianto (2012). Embora se argumente que a GC pode trazer benefícios econômicos diretos para a empresa por meio de economia ou ganho de dinheiro, uma visão mais comum parece ser a de que o impacto no desempenho financeiro da empresa é indireto. Zack *et al.* (2009) constataram que as práticas de GC estão diretamente relacionadas a vários resultados intermediários do desempenho organizacional estratégico e que esses resultados intermediários estão associados ao desempenho financeiro. GC oferece benefícios econômicos para a empresa de várias maneiras, como aceleração da inovação e agilidade estrutural; redução do tempo de ciclo na produção e falhas de programa; criar uma cultura saudável e amiga do conhecimento; atrair e manter uma força de trabalho de conhecimento de alta qualidade; e melhorando os níveis de reutilização do conhecimento e da memória corporativa. GC também tem sido conectado com liderança de produto, intimidade com o cliente e excelência operacional, inovação, criatividade organizacional.

Definição

As práticas de GC são as práticas organizacionais e gerenciais conscientes destinadas a atingir os objetivos organizacionais por meio do gerenciamento eficiente e eficaz dos recursos de conhecimento da empresa. (Inkinen 2016).

Vamos agora analisar quais práticas de GC produzem quais resultados em relação ao desempenho da empresa. A seguir, resumimos as principais descobertas da revisão da literatura de Inkinen (2016).

Práticas de gestão de recursos humanos baseadas no conhecimento

A literatura sugere fortemente que as práticas de gestão de recursos humanos estão altamente associadas à inovação. Alguns estudos apontam que a utilização de práticas de GRH aumenta os processos de conhecimento, como aquisição, compartilhamento e criação, que impactam na capacidade de inovação. Além disso, as práticas de gestão de recursos humanos parecem aumentar as inovações e melhorar a capacidade de inovação, influenciando positivamente o comprometimento afetivo dos funcionários e a confiança impessoal.

Práticas de liderança em gestão do conhecimento

O apoio da alta direção está associado ao aumento dos processos de conhecimento, que resultam em maior aprendizado organizacional e na capacidade de desenvolver novos produtos ou serviços, prever negócios ou riscos e lidar com novas informações sobre mercados. A liderança orientada para o conhecimento, em termos de capacitar os funcionários e promover a confiança e o aprendizado, aumenta o efeito que as práticas de exploração e exploração do conhecimento têm sobre as inovações. Um modo transformacional de liderança, incluindo influência idealizada, estimulação intelectual, motivação inspiradora e consideração individualizada, aumenta o desempenho relativo da empresa em comparação com seus concorrentes por meio de aquisição de conhecimento aprimorada e desempenho financeiro por meio de aprendizado e inovação.

Práticas de gestão do conhecimento orientadas para a tecnologia

O suporte de TI para colaboração, comunicação, busca de informações, aprendizado em tempo real, simulação e previsão está associado à capacidade de inovação de uma empresa. Os pesquisadores também observam que o suporte de TI é o principal facilitador da aquisição, criação e compartilhamento de conhecimento, o que leva as empresas a melhorar o desempenho por meio de inovações e agilidade organizacional.

Práticas de gestão do conhecimento orientadas para a organização

Um estudo descobriu que o estabelecimento de uma unidade especial responsável pela GC está significativamente associado ao desempenho da empresa em uma perspectiva de aprendizado e crescimento, uma perspectiva de processo interno e uma perspectiva do cliente.

Análise de casos de negócios

Para aprofundar a compreensão da relação entre desempenho e iniciativas de GC, foram avaliadas as candidaturas de 48 empresas alemãs ao prémio «Knowledge Manager of the year 2002 and 2003». 240 declarações de impacto foram agrupadas de acordo com as dimensões do Balanced Scorecard (North e Hornung 2003; North *et al.* 2004). As empresas colheram principalmente benefícios relacionados à melhoria de processos e ao desempenho dos funcionários. Relativamente poucas declarações se referiram ao impacto das iniciativas de GC nos resultados financeiros e na Inovação. Na dimensão do processo, os benefícios foram percebidos principalmente na área de aceleração do processo, redução do trabalho duplo e reutilização do conhecimento interno. Olhando para a dimensão das empresas verifica-se que as pequenas empresas apostam sobretudo na reutilização do conhecimento interno disponível e na redução de erros e os grandes players elegem a «poupança de tempo» e a «transparência dos processos» como os principais benefícios nesta categoria. Em relação aos funcionários, os argumentos dominantes são: aumento da motivação, aumento da base de conhecimento pessoal e menor tempo de integração para novos funcionários.

Para as pequenas empresas, o desenvolvimento de competências representa um benefício significativo, enquanto as grandes empresas mencionam a melhoria do trabalho em equipe como o principal benefício nessa dimensão. Em relação aos clientes, as empresas argumentam que as atividades de GC levaram a um aumento na qualidade dos produtos e serviços. Isso se aplica independentemente do tamanho da empresa. Os benefícios na área dos «resultados financeiros» referem-se a um aumento do volume de negócios, a uma melhor gestão do risco e a uma redução dos custos administrativos. Algumas empresas apresentaram cálculos de como uma melhor disponibilidade de informações reduz os tempos de busca e qual é o potencial de economia de custos relacionado. Os efeitos das iniciativas de GC na inovação foram a criação de novos produtos e serviços, seguidos – mencionados principalmente pelas grandes empresas – da aplicação de novas tecnologias. Com base nas descobertas, a Fig. 2.8 fornece uma visão geral do impacto no desempenho operacional que as empresas podem esperar dos casos de negócios de GC.

2.6. Principais percepções do Capítulo 2

O conhecimento em uma organização pode ser classificado de diversas formas e pode ser avaliado. O tratamento da informação é afetado pela perspectiva «O que é conhecimento e qual a sua importância para a nossa organização.

A escada do conhecimento descreve a criação de valor ligando informação, conhecimento, competência e competitividade

Existem pelo menos três epistemologias do conhecimento. Dependendo da situação, o conhecimento pode ser visto como um objeto ou um processo. A perspectiva do processo de conhecimento é explicada neste livro.

O modelo SEICI descreve a transformação do conhecimento de individual para coletivo e de tácito para explícito.

O conhecimento é visto como um componente dos ativos intangíveis ou «capital intelectual». O valor do conhecimento é baseado em sua escassez e potencial para agregar valor.

O conhecimento é considerado um fator de produção, um fator competitivo estratégico e a base da existência de uma empresa. O conhecimento pode ser imitado e substituído – esses dois aspectos do conhecimento são os critérios decisivos para uma vantagem competitiva sustentável.

Fig. 2.8 Aprimoramento de Desempenho atribuído à Iniciativas de GC, North et al. 2004)

2.7. Perguntas

1. Explique a diferença entre informação e conhecimento e conhecimento e competência.
2. Qual a diferença entre conhecimento tácito e explícito? O conhecimento explícito é apenas «informação»?
3. Como você interpretaria a Maturidade do Conhecimento em uma organização?
4. Avalie o valor de uma equipe de pesquisa e desenvolvimento de cinco pessoas. Quais critérios você usaria?
5. Quais são os critérios para avaliar as competências essenciais?
6. Cite tecnologias existentes que influenciam a disseminação do conhecimento nas organizações contemporâneas.
7. Cite ferramentas (computacionais ou não) de transferência de conhecimento.
8. Dê exemplos de transmissão de conhecimento por socialização.
9. Dê exemplos de transmissão de conhecimento por combinação.
10. Dê exemplos de transmissão de conhecimento por externalização.
11. Dê exemplos de transmissão de conhecimento por internalização.
12. Por que os processos de socialização e internalização têm relação direta com a cultura organizacional?
13. Por que os processos de combinação e externalização têm relação direta com as tecnologias adotadas pela organização?

2.8. Atribuições

1. Transferindo práticas de vendas bem-sucedidas

Em sua empresa, vários representantes de vendas experientes estão prestes a se aposentar. Vários novos representantes de vendas. foram recrutados.

Você é solicitado a propor como estruturar uma transferência de conhecimento eficaz entre antigos e novos representantes de vendas. Você se lembra do modelo SEICI de Nonaka e Takeuchi e acha que isso pode ser uma boa base para desenvolver uma proposta.

2. Análise de competências essenciais

A Apple é frequentemente citada como uma empresa inovadora e bem-sucedida. Analise as principais competências da Apple.

Ferramenta GC: Concurso de Ideias

O que é um concurso de ideias?

Aproveitar a imaginação criativa dos funcionários em conjunto com a emoção da competição é uma maneira poderosa de obter ideias atraentes e bem articuladas.

Uma competição de ideias é uma maneira bem focada de acessar ideias e soluções inovadoras de funcionários, usuários e clientes em potencial. A qualidade das ideias aumenta exponencialmente quando os participantes recebem uma pergunta de desafio clara e focada.

Os concursos de ideias baseiam-se na natureza da competição como um meio de incentivar a participação em um processo de inovação aberta, inspirar sua criatividade e aumentar a qualidade e o foco das inscrições. Quando o concurso termina, os envios são avaliados por um painel de especialistas. Aqueles cujos envios pontuam mais geralmente recebem um bônus ou um prêmio.

Por que usar concursos de ideias?

1. Em muitas organizações, os esquemas de sugestão funcionam ou funcionam bem. As pessoas não apresentam suas ideias por causa dos trâmites burocráticos. Os concursos de ideias abrem uma mudança para uma coleção de ideias focada, oportuna e simples.

2. Explorar ideias da «multidão» de utilizadores ou outras pessoas fora da organização tem um enorme potencial de criação de valor
3. Competições de ideias criam um espírito de interação e desafiam práticas e sabedoria atuais

Como organizar concursos de ideias?

Prepare um processo claro e transparente:

As competições de ideias envolvem vários participantes, incluindo patrocinadores, administradores, competidores e juízes. As responsabilidades dos administradores incluem:

1. **Projeto:** Antes de lançar uma competição, é importante definir as regras, projetar a estrutura, selecionar prêmios e incentivos e determinar o cronograma.
2. **Planejamento:** É essencial planear cuidadosamente, antecipar o número de submissões e definir os vários papéis e responsabilidades durante as várias fases do processo.
3. **Priorização:** Se centenas de ideias forem enviadas, é importante filtrar com eficiência as submissões para identificar rapidamente as melhores ideias.
4. **Proporcionar uma experiência agradável:** Cada participante deve se sentir energizado para participar e sentir que o processo da competição é divertido e fácil de se envolver.
5. **Transparência:** responda aos participantes em tempo hábil e torne as informações acessíveis para reduzir os gargalos administrativos e fazê-los se sentirem importantes.
6. **Avaliação justa:** O julgamento uniforme é fundamental para uma competição justa. Os juízes devem receber um cartão de pontuação e critérios de avaliação para avaliar de forma justa cada plano/ideia conceitual.
7. **Gerenciamento de escala:** Devido à natureza viral das competições online, os administradores devem estar preparados para lidar com centenas ou talvez milhares de inscrições. O uso de um sistema robusto e comprovado baseado na Web evitará que a administração seja sobrecarregada.

Assegure a participação e prepare-se para resultados de alta qualidade.

Como um processo de competição de ideias dirigido por funcionários pode ser projetado para oferecer ideias melhores? Algumas orientações importantes são as seguintes:

1. **Patrocínio de nível executivo:** Tenha um executivo sênior patrocinando a competição, desempenhe um papel na definição do foco estratégico e comunique a importância do esforço no apoio à estratégia corporativa.
2. **Seção de participantes:** Recrute participantes criativos e apaixonados com conjuntos de habilidades e perspectivas complementares (marketing, percepções do consumidor, P&D, canal de vendas, produção etc.) e reúna-os em equipes. Envolver as principais partes interessadas no processo de inovação promove conversas que levam a ideias de maior qualidade. Também cria a propriedade que acelera o processo de tomada de decisão e constrói a adesão necessária para a implementação.
3. **Preparação dos participantes:** trate as competições de ideias (e qualquer esforço de inovação) como um processo – não como um evento. Esperar que os participantes inovem sem nenhuma preparação significativa, contexto ou inspiração normalmente leva a “ideias no vácuo” irrelevantes.
4. **Percepção do consumidor:** Certifique-se de que os participantes tenham uma visão sobre as necessidades do consumidor – tanto articuladas quanto não articuladas. Vá além dos dados históricos do consumidor e dos padrões de uso e procure entender a voz, o coração e a mente dos consumidores. No mínimo, aumente a conscientização dos participantes sobre problemas conhecidos que os consumidores têm com os produtos, serviços e soluções atuais, mas, para obter melhores resultados, crie um módulo de "experiência do consumidor" (como uma viagem de campo) em que os participantes observem os consumidores usando o produto ou serviço atual.

5. **Previsão da Indústria:** Criar uma orientação para o pensamento orientado para o futuro. Ajude os participantes a identificar tendências emergentes em várias dimensões, por exemplo: tecnologias de ponta, mudanças antecipadas no cenário competitivo, modelos de negócios incomuns, hipóteses sobre tendências sociais, mudanças regulatórias antecipadas, canais de vendas emergentes, novas práticas de fabricação, etc. Esteja ciente de que é fácil focar em dados históricos e tendências estabelecidas, mas normalmente limita a produção a ideias incrementais mais próximas, como extensões de linha. A maioria das empresas está familiarizada com os dados de tendências históricas, mas não se sente à vontade para pensar em "tendências emergentes" – e, no entanto, isso é crítico.
6. **Pensamento estratégico e imaginativo:** Incentive os participantes a romper com os modos de pensamento tradicionais e desafiar suas próprias suposições. Peça-lhes que procurem lições e análogos de outras indústrias. No mínimo, introduza estímulos interativos (vídeos, anúncios, «cenários de usuários», depoimentos de clientes, etc.). Velhos hábitos e padrões de pensamento são difíceis de quebrar – estender o pensamento dos participantes a níveis totalmente novos exige uma abordagem radicalmente diferente.

3. A Natureza dos dados e o Projeto de Banco de Dados Relacionais

3.1. Business Intelligence, Business Analytics e Data Science

Decifrando a confusão de nomes

Inteligência de negócios, análise de negócios e ciência de dados são todos usados como termos abrangentes para campos relacionados, e essas semelhanças geralmente podem levar à confusão ao tentar entender o que significam. Embora esses conceitos estejam de fato relacionados, eles também são distintamente diferentes.

- **Inteligência de negócios:** Business Intelligence (BI) é um processo bem definido de análise e processamento de dados para fins de visualização e aplicação de informações acionáveis. O conceito de business intelligence evoluiu ao longo de várias décadas e é frequentemente usado como um termo abrangente. Em última análise, a inteligência de negócios adiciona "contexto" aos dados para produzir informações acionáveis, ou seja, aquelas que auxiliam no suporte à decisão. Um dos principais objetivos do BI é colocar o poder da visualização nas mãos dos usuários finais e permitir a tomada de decisões orientada por dados. Existem muitas ferramentas e aplicativos no mercado atual para dar suporte ao BI e impulsionar as soluções de negócios.
- **Análise de negócios:** A análise de negócios usa matemática e estatística para analisar os dados de uma organização. A análise de negócios oferece suporte direto ao BI para permitir a tomada de decisões orientada por dados e obter insights para suporte à decisão. Os principais componentes da análise de negócios são qualidade de dados, análise precisa e profunda, aplicação eficiente de ferramentas e modelos preditivos e automação. Os dados podem ser coletados de muitas fontes diferentes, incluindo sistemas transacionais, data warehouses e até mesmo fontes de dados não estruturadas. A análise de negócios geralmente é categorizada como descritiva, preditiva ou prescritiva, e essas categorias aumentam em valor de negócios (e complexidade) à medida que você passa de uma para outra.
    - **A análise descritiva** é usada para rastrear os principais indicadores de desempenho (KPIs) e para entender e descrever o estado atual. A inteligência de negócios tradicional usa análises descritivas para analisar as operações de negócios existentes e gerar uma imagem atual dos negócios.
    - **A análise preditiva** é usada para realizar análises de tendências e tentar identificar resultados futuros.
    - **A análise prescritiva** usa dados de desempenho anteriores para gerar recomendações para situações futuras com entradas semelhantes.
- **Ciência de dados:** A ciência de dados é um campo avançado que abrange áreas como mineração de dados, aprendizado de máquina e estatística. Essas áreas geralmente exigem níveis profundos de codificação personalizada para explorar perguntas abertas. Os cientistas de dados empregam métodos estatísticos avançados para explorar e descobrir padrões e novos insights por meio de análises. Os objetivos da ciência de dados incluem aumentar a eficiência operacional, encontrar oportunidades e fornecer vantagens competitivas. A ciência de dados também é essencial para alavancar o poder de processamento computacional para suporte a decisões, modelagem preditiva, simulação avançada e muitos outros aplicativos de negócios.

Ter uma melhor compreensão das distinções desses termos (inteligência de negócios, análise de negócios e ciência de dados) nos ajudará a explorar outros conceitos relacionados neste e nos módulos futuros.

Sistemas de Suporte a Decisão

Um sistema de suporte à decisão (DSS) é um sistema de informação que permite e suporta diretamente a tomada de decisões orientada por dados. Os gerentes e líderes organizacionais tradicionalmente empregam esses sistemas para fornecer uma imagem de "verdade básica" de uma determinada situação. O DSS permite a análise rápida de grandes quantidades de dados para resolver desafios complexos. O poder do DSS vem por meio de

relatórios em tempo real, que fornecem dados constantemente atualizados para dar suporte a decisões críticas em um ambiente de negócios complexo. Um exemplo bem conhecido, mas direto de um DSS é o planejamento de destino/rota usando GPS. O sistema de informação GPS gera várias rotas disponíveis para o usuário e recomenda uma rota com base em variáveis como tráfego, interdições de estradas, pedágios, etc. dados relacionados e fazer recomendações.

Atores de BI e Análise

Existem várias e amplas preocupações que impulsionam a necessidade de análise de negócios. Alguns dos fatores mais comuns incluem o enorme volume de dados coletados, os requisitos de disponibilidade e segurança de dados e a necessidade de tomar decisões de negócios melhores e mais rápidas. À medida que as organizações coletam mais volumes de dados em velocidades cada vez maiores, a necessidade de organizar e analisar esses dados com eficiência também aumenta. Além disso, a natureza móvel dos negócios exige disponibilidade consistente de dados para dar suporte à tomada de decisões em tempo real, independentemente da localização.

Embora a disponibilidade de dados seja fundamental para a implementação eficaz do BI, a segurança dos dados também é um foco principal e continuaremos a discutir ao longo deste programa. E, finalmente, o ambiente de negócios em rápida mudança de hoje exige decisões melhores e mais rápidas, e a análise de dados pode capacitar e apoiar os líderes na tomada de decisões orientadas por dados. À medida que grandes volumes de dados são coletados, é crucial ter uma estratégia de dados clara para uma análise adequada e esforços de implementação. O foco precisa estar na conversão de dados em informações acionáveis.

Uma taxonomia simples para análise

Desenvolver uma taxonomia simples e aceitável para análise de negócios é essencial, pois os conceitos e as tecnologias mudam tão rapidamente. As partes interessadas podem maximizar o valor e garantir clareza se puderem falar a partir de um contexto compartilhado e entender a terminologia de análise de negócios. Várias empresas e instituições acadêmicas tentaram alinhar o contexto, a compreensão e a terminologia, e seu trabalho acabou produzindo uma versão da taxonomia vista na tabela de análise de negócios, vinculada aqui (adaptado de Delen, 2020).

Referências

Delen, D. (2020). Prescriptive analytics: The final frontier for evidence-based management and optimal decision making. << https://www.pearson.com/us/higher-education/program/Delen-Prescriptive-Analytics-The-Final-Frontier-for-Evidence-Based-Management-and-Optimal-Decision-Making/PGM239919.html >>

3.2. OLTP versus OLAP

OLTP e OLAP são ambos sistemas de processamento online. A distinção entre esses sistemas está no que está sendo processado, ou seja, transações ou consultas analíticas.

- OLTP = Processamento de Transações Online
- OLAP = Processamento analítico online

Processamento de transações on-line (OLTP)

O OLTP é utilizado para processar sistemas transacionais e normalmente envolve a modificação de um sistema de banco de dados online. Um exemplo simples é um site de comércio eletrônico. Cada vez que um pedido é feito, um banco de dados (ou vários bancos de dados) são modificados para armazenar detalhes do cliente e do pedido (entre outros dados). Essa transação é processada pelo OLTP, que lida com inserções, atualizações e exclusões. Os bancos de dados OLTP são atualizados com frequência e geralmente são chamados de sistemas transacionais ou operacionais.

Processamento analítico online (OLAP)

O OLAP lida com a consulta de um sistema de banco de dados online. Os bancos de dados OLAP armazenam dados históricos para relatórios e análises em suporte direto à tomada de decisões orientada por dados. O mesmo site de comércio eletrônico pode relatar contagens de estoque atuais ou gerar relatórios de vendas. Nesse caso, o OLAP extrai dados do sistema de banco de dados para suporte à decisão.

3.3. Data Warehousing para BI

Qual é a razão histórica para o desenvolvimento de DW não envolver primeiramente a tecnologia? Esta questão é muito relevante hoje em dia, porque o sucesso da implantação do DW depende desta capacidade organizacional.

Para melhorar seu entendimento sobre bancos de dados operacionais e data warehouses, você estará apto a explicar várias diferenças entre os dois tipos de bancos de dados. Os bancos de dados de suporte às tomadas de decisão nas organizações. A hierarquia tradicional de tomada de decisão representa os níveis de gestão e os volumes de decisão em cada nível. As empresas acreditavam que os bandos de dados operacionais dariam suporte à tomada de decisão nos três níveis. Bancos de dados operacionais foram desenhados para processar de modo eficiente as transações, e para dar suporte ao nível operacional de decisões como resolver atrasos dos pedidos. Entretanto, as empresas encontraram grande dificuldade no uso de bancos de dados operacionais para níveis mais altos de tomada de decisão.

Tal dificuldade estimulou o desenvolvimento da tecnologia e implantação de data warehouses iniciados em meados da década de 1990. O fracasso dos bancos de dados operacionais em dar suporte ao nível mais alto na tomada de decisão se deveu a uma combinação de inadequação da tecnologia do banco de dados, com as limitações na implantação dos bancos de dados. As empresas fornecedoras de SGBDs descobriram que um único banco de dados não poderia ser configurado para atingir o desempenho adequado para ambos os processamentos de transações e os de inteligência de negócios ao mesmo tempo.

As empresas descobriram que a falta de integração entre os bancos de dados operacionais impedia a tomada de decisões nos níveis mais altos. A falta de integração não era uma falha no projeto. Os bancos de dados operacionais objetivavam, primeiramente, dar suporte ao processamento de transações, não ao processamento de inteligência de negócios. As empresas perceberam que retroagir, realizando a integração dos bancos de dados operacionais seria difícil. Os fornecedores de produtos descobriram que a falta de características chaves para suportar a síntese de dados e cálculos analíticos era vital ao processamento de inteligência de negócios.

A cláusula "group by" do SQL era inapropriada para especificar consultas SQL envolvendo somatório de dados. O comando "select" em SQL não tinha nenhuma facilidade para cálculos analíticos tais como médias móveis. Os métodos de otimização de armazenamento não eram adequados para as consultas que envolviam dados sintéticos. Gradualmente, as soluções para estes problemas apareceram nas empresas e nos fornecedores de produtos. Limitações de desempenho demandaram que os data warehouses se separassem dos bancos de dados operacionais.

A falta de integração exigiu um foco intenso na agregação de valor para as fontes de dados via transformações na integração entre os bancos operacionais e os DWs. A falta de utilidades alavancou o desenvolvimento de um conjunto de novas utilidades para representação, para a armazenar as manipulações, e para o processamento de cálculos analíticos e dados sintetizados (somatórios, resumos). Os data warehouses se tornaram, então, uma parte essencial da infraestrutura das empresas para dar suporte à inteligência de negócios.

Data warehouse, um termo cunhado por William Inmon em 1990, se refere a um repositório de dados centralizado logicamente, onde os dados dos bancos de dados operacionais e das fontes de dados externas são integrados, tratados (limpos), e padronizados para dar suporte à inteligência de negócios. As atividades de transformação, limpeza, mesclagem (merging), e padronização, são essenciais para que os dados tenham valor para a inteligência de negócios. Os data warehouses são otimizados para relatórios, sempre envolvendo

sumarização de grandes quantidades de dados. Bem como processamento periódico para integrar e transformar os dados da fonte de origem. O processamento de transações usa dados primários advindos de grandes volumes de transações para dar suporte às operações diárias e à tomada de decisão de curto prazo das empresas. Em contraste, o processamento de inteligência de negócios usa dados secundários, transformados, para dar suporte às tomadas de decisão de médio e longo prazos.

Um data warehouse gera valor para as tomadas de decisão de longo prazo através das transformações e da integração com os bancos de dados operacionais e com as fontes de dados de origem externas. Por conta do suporte aos tipos distintos de processamento, os bancos de dados operacionais diferem muito dos data warehouses. Bancos de dados operacionais largamente contém dados correntes de nível individual, enquanto os data warehouses têm dados históricos em ambos os níveis: individual e sintetizado. O nível de dados individual fornece flexibilidade para responder a grande necessidade de inteligência de negócios enquanto os dados sumarizados fornecem respostas rápidas às consultas repetitivas. Por exemplo, um banco de dados operacional para dar suporte ao processamento de pedidos requer dados sobre clientes individuais, sobre as ordens dele, e sobre itens individuais de inventário.

Por outro lado, um aplicativo de inteligência de negócios pode usar vendas mensais sobre um período de vários anos. Bancos de dados operacionais, portanto, têm uma orientação a processos, por exemplo, todos os dados relevantes de um processo de negócios em particular, comparados com uma orientação ao assunto, de um data warehouse. Por exemplo, todos os dados do cliente ou dados de um pedido. Uma transação tipicamente atualiza apenas uns poucos registros em um banco de dados operacional, ou em um aplicativo de inteligência de negócios pode consultar entre milhares até milhões de registros de um data warehouse.

A normalização é menos importante para os data warehouses, porque o foco deles está nos relatórios ao invés de estar no processamento das transações. Bancos de dados operacionais são altamente voláteis, processando grandes volumes de transações, conquanto que data warehouses são não-voláteis, com renovação periódica ao integrar novas fontes de dados. Bancos de dados operacionais usam, primeiramente, modelo relacional de dados, enquanto os data warehouses usam padrões de esquema em estrela das tabelas bem como um modelo de dados multidimensional. Os padrões de esquemas tipicamente diferem entre bancos de dados operacionais e data warehouses.

Num banco de dados operacional, para dar suporte ao processamento de pedidos, relacionamentos m para n, vários para vários, ou tabelas associativas equivalentes são usados geralmente para representar o cabeçalho do pedido e os detalhes. Diferentemente, um data warehouse irá tipicamente apenas exibir um nível de detalhe e nenhum relacionamento de m para n. Além do mais, relacionamentos especializados como auto-referenciados e dependência de identificação são menos comumente usados em modelagens de DW.

O data warehousing (DW) emprega um processo de extração, transformação e carregamento (ETL) para coletar dados de sistemas transacionais distintos (OLTP) e armazenar esses dados para fins históricos, analíticos e de relatórios. Os data warehouses são imutáveis, integrados, granulares e históricos por natureza. Eles são frequentemente considerados a “fonte única da verdade” devido à sua natureza imutável; ou seja, uma vez que os dados passaram pelo processo de ETL, eles não são alterados novamente.

O processo ETL limpa, normaliza, alinha e carrega dados no data warehouse para permitir análises e relatórios eficientes e eficazes por meio do data warehouse (OLAP). O DW fornece contexto histórico e um conjunto de dados normalizado a partir do qual relatórios e análises podem ser conduzidos. O data warehouse geralmente agrega e calcula cálculos comuns normalmente incluídos nos relatórios e visualizações organizacionais. Essas etapas reduzirão o tempo de processamento computacional ao executar análises e gerar relatórios *ad hoc*.

Exercício

1. Business Intelligence (BI) adiciona _____ aos dados para produzir informações acionáveis.

a) Visualizações
b) Tecnologia
c) Contexto

2. Quais dos seguintes são objetivos do BI?

a) Coloque o poder da visualização nas mãos dos usuários finais
b) Habilite a tomada de decisões orientada por dados
c) Todas essas opções estão corretas.

3._____ inclui qualidade de dados, análise precisa e profunda, aplicação eficiente de ferramentas e modelos preditivos e automação.

a) Analista de negócios
b) Inteligência de negócios
c) Ciência de dados

4. Que tipo de análise de negócios é usada para conduzir a análise de tendências?

a) Descritivo
b) Preditivo
c) Prescritivo

5. A ciência de dados mergulha profundamente em _____.

a) Mineração de dados
b) Aprendizado de máquina
c) Todas essas opções estão corretas.

6. _____ é um sistema de informação que suporta e permite a tomada de decisões orientada por dados, fornecendo uma imagem da verdade.

a) Sistema de Informação Geoespacial
b) Sistema de Apoio à Decisão
c) Sistema de Gestão de Relacionamento com o Cliente

7. Os drivers de análise reduzem a clareza das implementações de BI e causam confusão sobre dados críticos.

a) Verdadeiro
b) Falso

8. Em qual categoria de análise de negócios normalmente pertence o Business Intelligence?

a) Descritivo
b) Prescritivo
c) Preditivo

9. Qual das opções a seguir lida com a consulta de um sistema de banco de dados online?

a) OLTP
b) OLAP
c) Todas essas opções estão corretas.

10. O processo ETL limpa, normaliza, alinha e carrega dados no data warehouse.

a) Verdadeiro
b) Falso

Leituras recomendadas:

Article: Yellowfin Team. (nd). Business Intelligence: Drivers, Challenges, Benefits and ROI. *(5 min)* *<<https://www.yellowfinbi.com/blog/2011/04/yfcommunitynews-business-intelligence-drivers-challenges-benefits-and-roi-103783>>*

Article: Glen, S. (2020). Business Intelligence vs Business Analytics. *(5 min)* *<<https://www.datasciencecentral.com/profiles/blogs/business-intelligence-vs-business-analytics-vs-data-analytics>>*

3.4. Definindo Bancos de Dados Relacionais

Um banco de dados relacional é uma coleção de dados relacionados armazenados em um local ou repositório centralizado. Os dados armazenados são organizados em tabelas que abrigam informações sobre vários objetos armazenados no banco de dados. Os bancos de dados relacionais fornecem uma maneira eficiente, flexível e escalável de armazenar e acessar informações estruturadas.

Os bancos de dados relacionais geralmente são hospedados e gerenciados usando um sistema de gerenciamento de banco de dados relacional (RDBMS). O RDBMS emprega Structured Query Language (SQL) para permitir a recuperação e interação com dados em várias tabelas. Esses sistemas também geralmente implantam autenticação, autorização, ajuste de desempenho e muitos outros recursos.

Bancos de dados relacionais são organizados por agrupamentos de objetos que possuem um identificador único ou chave primária. A chave primária identifica a linha em uma tabela que corresponde a um registro individual e seus dados associados. A chave primária também pode ser usada como chave estrangeira em outra tabela para indicar relacionamento. Chaves estrangeiras criam conexões lógicas entre tabelas e estabelecem relacionamentos.

3.4.1. Diagrama Entidade-Relacionamento (ERD)

Um diagrama entidade-relacionamento (ERD) é uma representação gráfica de um projeto de banco de dados. Os diagramas de exemplo abaixo (Figuras 1-3) ilustram um ERD simples que descreve o design geral e estabelece a base e os requisitos para implementação em um RDBMS. O ERD também estabelece relacionamentos entre objetos e serve como documentação para o sistema de banco de dados.

O Processo de Projetar Bancos de Dados

A modelagem de dados é o processo de projetar bancos de dados e existem três modelos de dados: dados conceituais, dados lógicos e dados físicos.

Projeto conceitual

O projeto conceitual estabelece entidades, atributos e relacionamentos. O objetivo de um modelo de dados conceitual é apresentar uma imagem de alto nível do sistema a ser implementado com foco nos objetos de negócios envolvidos no sistema. As tabelas de banco de dados não são projetadas no nível conceitual.

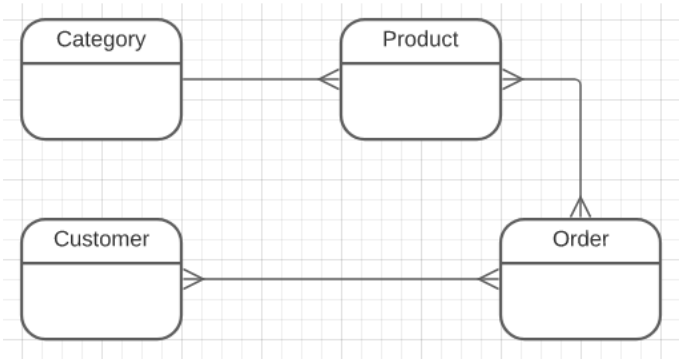


Figura 1 Objetos de Negócio da Entidade (design conceitual)

A Figura 1 descreve os objetos de negócios da entidade que interagem ou fazem parte de um sistema de informações. Neste exemplo, temos clientes solicitando produtos. As relações de base são identificadas usando a notação pé de galinha. Uma única linha indica um único relacionamento (ou seja, um produto só pode estar em uma categoria), e um pé de galinha de três linhas indica um relacionamento do tipo “muitos” (ou seja, uma categoria pode ter muitos produtos).

Projeto Lógico

O design lógico define a estrutura dos elementos de dados e estabelece relacionamentos entre os elementos de dados. O modelo de dados lógico adiciona uma camada de detalhes ao projeto conceitual, definindo as colunas de dados que precisam ser incluídas em cada entidade, como visto na Figura 2. Nesta fase do projeto,

ainda não há consideração por um sistema de banco de dados específico já que o foco está na estrutura e no relacionamento.

Projeto conceitual de objetos de negócios de entidade com atributos.

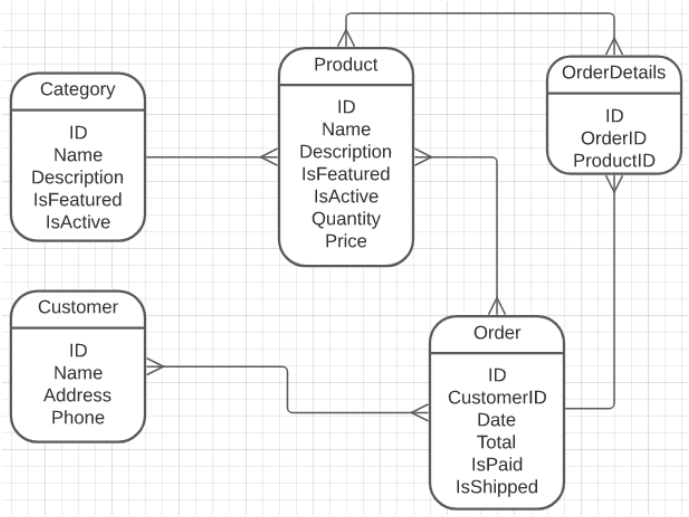


Figura 2 Objetos de Negócio da Entidade (design conceitual com atributos)

Cada objeto de negócios ou entidade agora inclui atributos ou colunas que descreverão registros individuais dentro da eventual tabela do banco de dados. Esses atributos começam a detalhar as informações que compõem um único registro (ou linha) dentro de uma eventual tabela de banco de dados.

Projeto Físico

O design físico descreve detalhes de implementação específicos do banco de dados e fornece um plano para o banco de dados relacional. O modelo de dados físico inclui detalhes adicionais sobre cada coluna dentro de uma entidade. Nesta fase do projeto, é importante operar dentro das construções de um RDBMS específico, pois as estruturas, convenções e restrições podem variar.

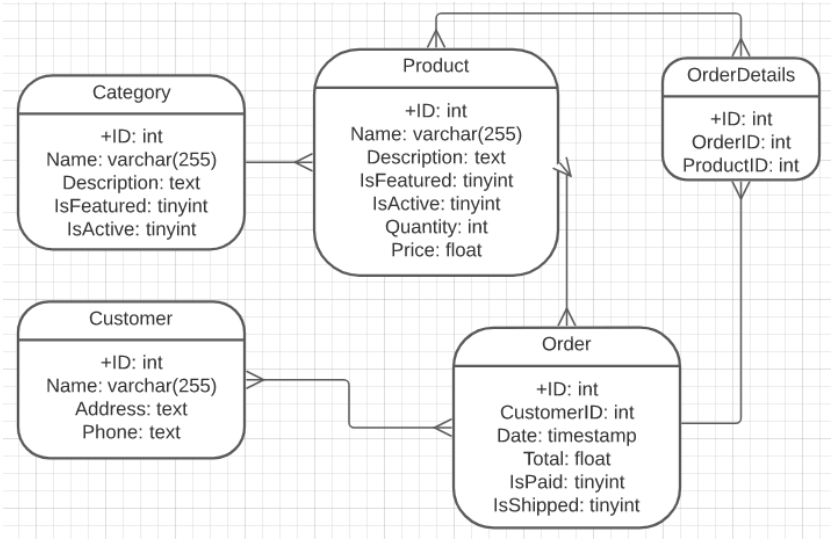


Figura 3 Objetos de Negócio da Entidade (modelo de dados físico)

Como mostra a Figura 3, agora temos um design de banco de dados totalmente definido que está pronto para implementação em nosso RDBMS selecionado. As chaves primárias para cada tabela são marcadas com

um símbolo “+”, e os tipos de dados para cada coluna são identificados e seguem os tipos de dados aceitáveis para o MySQL RDBMS.

3.4.2. Normalização e Desnormalização

Os conceitos de normalização e desnormalização descrevem a organização do conteúdo de um banco de dados. A normalização envolve a separação de dados em objetos bem definidos para limitar a redundância de dados. Na normalização, há um grande foco nos relacionamentos entre tabelas, e cada tabela contém informações exclusivas que são necessárias para descrever um registro ou entidade individual. Para recuperar todos os dados associados sobre um determinado registro, o usuário precisaria executar muitas junções (exploradas mais adiante), o que pode causar problemas de desempenho.

A desnormalização combina dados em uma única tabela para remover relacionamentos e dependências externas. Embora essa abordagem possa acelerar as consultas SQL, muitas vezes também resulta em dados redundantes ou duplicados em todo o banco de dados. A tabela de normalização e desnormalização vinculada aqui contém mais detalhes sobre cada um desses conceitos.

Aplicação do Diagrama Entidade-Relacionamento

Selecione um sistema e crie um ERD com progressão do projeto conceitual para o lógico e o físico.

Material de apoio

A seguir está uma lista de recursos opcionais que você pode achar úteis para melhorar sua compreensão dos tópicos deste módulo.

Vídeo: Lucidchart. (2018). The Basics of Relational Database Design. *(5 min)*

Vídeo: CBT Nuggets. (2019). How to Normalize Databases. *(7 min)*

Artigo: Guru99. (n.d.). What is Normalization? 1NF, 2NF, 3NF, BCNF Database Example. *(10 min)*

A seguir está uma lista de recursos opcionais que você pode achar úteis para melhorar sua compreensão sobre SQL.

Video: Guru99. (2013). What is Database & SQL? *(6 min)*

Video: Socratica. (2019). SQL SELECT Tutorial ||| SQL Tutorial ||| SQL for Beginners. *(9 min)*

Article: Menshov, S. (2019). Tutorial on SQL (DDL, DML) on the Example of MS SQL Server Dialect. *(30 min)*

Article: W3Schools. (n.d.). SQL Tutorial. *(30 min)*

Exercício

1. Um _____ é uma coleção de dados relacionados armazenados em um local ou repositório centralizado.

a) Sistema de gerenciamento de banco de dados relacional
b) Banco de dados relacional
c) Diagrama de Entidade-Relacionamento

2. O que é SQL?

a) Structured Query Language
b) Simple Question Location
c) Simplified Query Language

3. Um _____ identifica a linha em uma tabela que corresponde a um registro individual e seus dados associados.

a) Chave primária
b) ERD
c) Chave estrangeira

4. _____ criar conexões lógicas entre tabelas e estabelecer relacionamento.

a) Chaves primárias
b) Chaves estrangeiras
c) Nenhuma dessas opções está correta.

5. _____ é o processo de projetar bancos de dados.

a) ERD
b) Linguagem de consulta estruturada
c) Modelagem de dados

6. _____ estabelece entidades, atributos e relacionamentos.

a) Projeto conceitual
b) Projeto físico
c) Projeto lógico

7. O modelo _____ adiciona uma camada de detalhes ao projeto conceitual definindo as colunas de dados que precisam ser incluídas em cada entidade.

a) Projeto físico
b) Projeto conceptual
c) Projeto lógico

8. O modelo de dados _____ inclui detalhes adicionais sobre cada coluna dentro de uma entidade.

a) Lógico
b) Físico
c) Conceptual

9. O _____ estabelece relacionamentos entre objetos e serve como documentação para o sistema de banco de dados.

a) Modelo de dados conceitual
b) ERD
c) RDBMS

10. Qual das opções a seguir reduz a redundância e a inconsistência de dados?

a) Desnormalização
b) Normalização
c) Modelagem de dados

Exercício de SQL

1. _____ é uma linguagem de programação de banco de dados que permite interagir com um banco de dados para executar operações como SELECT, INSERT, UPDATE e DELETE.

a) PHP
b) SQL
c) RDBMS

2. Qual dos seguintes não faz parte do DDL?

a) SELECT
b) ALTER
c) CREATE

3. Qual dos seguintes não faz parte da DML?

a) DELETE
b) ALTER
c) INSERT

4. DCL inclui todos os itens a seguir, exceto:

a) REVOKE
b) GRANT
c) Todas essas opções estão corretas

5. Quais dos seguintes não fazem parte do TCL?

a) COMMIT
b) Todas essas opções fazem parte do TCL
c) ROLLBACK

6. _____ retorna linhas e nos permite coletar dados de tabelas normalizadas.

a) Subqueries
b) Inserts
c) Joins

7. Um uso comum para um _____ pode ser calcular o total de todos os produtos em nosso pedido ou um preço médio de nossos produtos.

a) DDL
b) Join
c) Subquery

8. Qual palavra-chave do MySQL gerencia a atribuição de um valor de chave primária sem intervenção do usuário?

a) AUTO_INCREMENT
b) PRIMARY KRY
c) NOT NULL

9. Qual tipo de dados MySQL permite que uma coluna não contenha mais de 255 caracteres?

a) FLOAT
b) TEXT
c) VARCHAR(255)

10. Qual palavra-chave do MySQL define um valor padrão para uma coluna da tabela de banco de dados quando o usuário não fornece um valor?

a) INSERT
b) NOT NULL
c) DEFAULT

4. Data Warehousing e Business Intelligence

4.1. Necessidade de armazenamento de dados

Um data warehouse (DW) é um repositório que armazena dados relacionais organizados, limpos e padronizados para uso corporativo. Um data warehouse é organizado por bancos de dados orientados a assunto e não é volátil no suporte direto à funcionalidade do sistema de suporte à decisão (DSS). Ao fazer isso, um data warehouse inclui dados estrategicamente selecionados que são importantes para uma empresa para rastreamento histórico, relatórios e análises.

Um data warehouse tem as seguintes características:

- **Orientado a assunto:** os dados são baseados em tema ou objeto (ou seja, cliente, produto, vendas, etc.)
- **Integrado:** dados díspares são combinados e normalizados a partir de sistemas de origem
- **Variante de tempo:** os dados são organizados por vários intervalos de tempo para relatórios históricos e preservação (ou seja, semana, mês, trimestre, ano)
- **Não volátil:** os dados nunca são alterados ou excluídos; os dados são somente leitura e atualizados em intervalos de tempo bem definidos
- **Resumido:** os dados geralmente são agregados para otimização dos relatórios

Um data warehouse deve incluir metadados, que são "dados que descrevem dados". Metadados geralmente incluem localização de dados, estrutura de dados e parâmetros de valores válidos. Essencialmente, os metadados atuam como "um dicionário vivo" e documentação para o data warehouse.

A necessidade de armazenamento de dados (data warehousing) torna-se evidente quando entendemos que os dados estão em toda parte. Muitas organizações utilizam meios e sistemas diferentes para coletar dados. Um data warehouse extrai dados desses sistemas de origem díspares, que podem incluir ponto de venda (SPT), planejamento de recursos empresariais (ERP), gerenciamento de relacionamento com o cliente (CRM), etc. O processo de extração, transformação, carregamento (ETL), que será discutido posteriormente neste módulo, prepara e normaliza os dados extraídos para análise e relatório. Além disso, um data warehouse permite o rastreamento e a manutenção de informações históricas e fornece uma única fonte de verdade.

4.1.1. Arquiteturas de armazenamento de dados

Os data warehouses podem ser arquitetados usando abordagens variadas. Existem duas abordagens principais: a abordagem dimensional (popularizada por Ralph Kimball) e a abordagem normalizada (popularizada por Bill Inmon).

Abordagem Dimensional

A abordagem de Kimball descreve um data warehouse por meio de um modelo dimensional (esquema em estrela ou floco de neve). A abordagem dimensional usa um design bottom-up "de baixo para cima", no qual data marts individuais são criados em nível departamental ou organizacional (ou seja, vendas, recursos humanos, finanças, etc.) e construído para um armazém de dados corporativo (Enterprise Data Warehouse - EDW). Hoje, a abordagem de Kimball é mais popular porque os usuários de negócios podem rapidamente ganhar utilidade com ela.

Abordagem Normalizada

A Inmon, por outro lado, utilizou uma abordagem Top-Down "de cima para baixo" para normalizar um data warehouse. O modelo de dados corporativos normalizado cria um repositório central ou data warehouse

corporativo. Data marts dimensionais para departamentos ou unidades organizacionais específicas podem ser criados a partir do data warehouse corporativo mestre.

4.1.2. Extração, transformação e carga (ETL)

Extração, transformação e carga (ETL) é o processo de integração de dados de sistemas operacionais ou transacionais de origem para combinar dados diferentes em um único formato em um repositório central. Os dados de origem são extraídos de sistemas transacionais; transformado para normalização, formatação e correção de erros; e carregado no data warehouse para análise e relatórios (como visto na Figura 4).

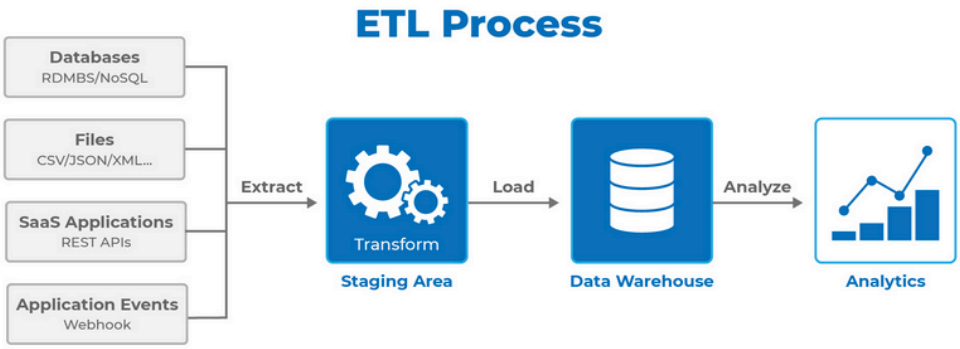


Figura 4 O processo ETL

4.1.3. Data Marts

Um data mart é um subconjunto de um data warehouse corporativo e geralmente é chamado de "data warehouse departamental". Um data mart contém o mesmo tipo de informação que existe em um data warehouse corporativo, mas os dados são organizados e otimizados para um departamento específico ou unidade organizacional. O diagrama na Figura 5 fornece uma arquitetura de alto nível de data warehousing e mostra como os data marts se encaixam nessa arquitetura.

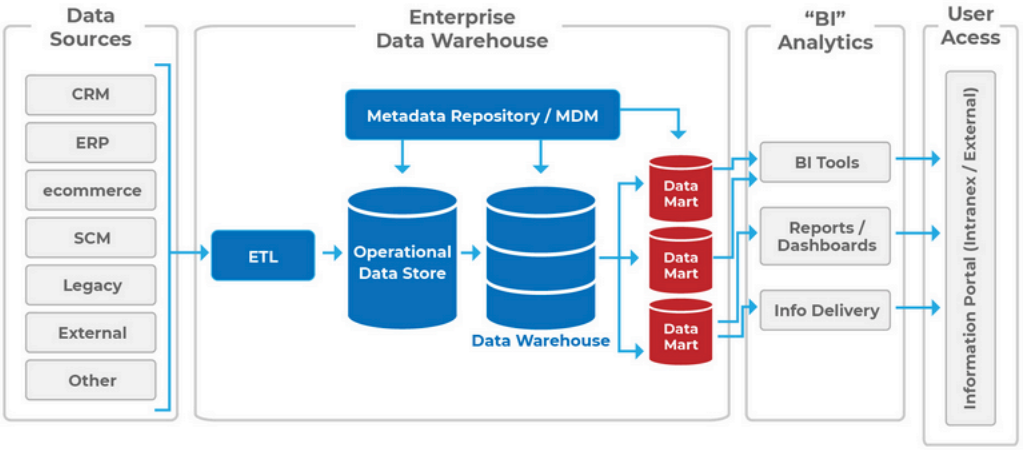


Figura 5 Data marts em uma arquitetura de data warehousing

4.1.4. Armazenamentos de dados operacionais

Um armazenamento de dados operacionais (ODS) utiliza snapshots de dados de sistemas operacionais ou transacionais para fornecer relatórios operacionais de negócios. O ODS difere de um data warehouse porque os dados são acessados diretamente dos bancos de dados do sistema transacional e o armazenamento de dados operacional pode gravar dados de volta nos sistemas de origem. Um objetivo principal de um armazenamento de dados operacional é lidar com as complexidades de manter dados atualizados no data warehouse. Assim, o ODS pode ser visto como uma abordagem menos dispendiosa para relatórios de dados em tempo real.

4.1.5. Armazenamento de dados na nuvem

Os data warehouses tradicionalmente existem dentro da infraestrutura local de uma organização (on-premises), onde a responsabilidade pela configuração e manutenção recai exclusivamente sobre a equipe de tecnologia da informação (TI) da organização. O armazenamento de dados na nuvem transfere grande parte da responsabilidade de hardware, rede, segurança e manutenção para terceiros, o que permite que a organização se concentre mais nas metas e objetivos de negócios. Essa abordagem também permite aos usuários (que geralmente são remotos ou móveis) um nível mais alto e mais consistente de disponibilidade de data warehouse.

Exercício

1. Um _____ é um repositório que armazena dados relacionais organizados, limpos e padronizados para uso corporativo.
a) Base de dados
b) Sistema de gerenciamento de banco de dados
c) Data Warehouse

2. Qual das seguintes características descreve um DW como sendo organizado por intervalos de tempo?
a) Não volátil
b) Tempo variável
c) Integrado

3. Metadados são dados sobre dados.
a) Falso
b) Verdadeiro

4. Qual abordagem de arquitetura de data warehousing utiliza um design de bottom-up?
a) Desnormalizado
b) Dimensional
c) Normalizado

5. A abordagem top-down de Inmon para a arquitetura DW cria um repositório central normalizado ou _____.
a) Armazenamento de dados operacionais
b) Enterprise Data Warehouse
c) Data Mart

6. O processo _____ combina dados díspares em um repositório central.
a) Extração
b) Transformação
c) Extrair, transformar, carregar (ETL)

7. Qual das opções a seguir é um subconjunto de um data warehouse e geralmente é focado no departamento?
a) Data Mart
b) Armazenamento de dados operacionais
c) Banco de dados transacional

8. Qual dos seguintes usa instantâneos de sistemas transacionais para fornecer relatórios operacionais de negócios?
a) Armazenamento de dados operacionais (ODS)
b) Data Mart
c) Enterprise Data Warehouse

9. Qual das opções a seguir é um exemplo de uma fonte de dados transacional?
a) CRM
b) ERP
c) Todas essas opções estão corretas

10. O data warehouse baseado em nuvem transfere grande parte da responsabilidade de hardware, rede, segurança e manutenção para terceiros.
a) Falso
b) Verdadeiro

Material Complementar

https://www.youtube.com/watch?v=Tff34jj_V-0

4.2. Modelagem de dados para Data Warehouse

Anteriormente, vimos a importância da modelagem de dados no projeto e implementação de banco de dados. Isso também se aplica ao Data Warehouse. O processo de modelagem de dados permanece o mesmo, sendo o objetivo “a organização e armazenamento de dados de longo prazo para análise e relatórios”. O modelo de dados precisa suportar as características básicas de um data warehouse, ou seja, ser orientado por assunto, integrado, variante no tempo, não volátil e resumido. O processo de modelagem de dados para data warehousing

ainda segue o processo de design - do conceitual ao lógico e aos ERDs físicos (diagramas de entidade-relacionamento). Outra área a ser considerada é a arquitetura de data warehouse selecionada (ou seja, dimensional ou normalizada) e se os data marts serão incorporados à arquitetura.

4.2.1. Modelagem de dados multidimensionais

Os modelos de dados multidimensionais representam estruturas de dados complexas (geralmente em formato de cubo) em oposição a uma única dimensão (geralmente representada por uma lista). Modelos de dados bidimensionais e tridimensionais são frequentemente utilizados em data warehouse. Esses modelos permitem uma estrutura e organização de dados bem definida. As etapas gerais na construção de um modelo de dados multidimensional incluem:

- Coletando os requisitos do usuário
- Categorizando os módulos do sistema
- Identificando dimensões para organizar dados em torno de objetos e funções
- Esboçar as dimensões em tempo real e as propriedades correspondentes
- Descobrindo os fatos a partir das dimensões e suas propriedades
- Construindo o esquema para armazenamento de dados

Esquema em estrela (Star Schema)

Um esquema em estrela é um modelo que descreve dados em uma forma semelhante à de uma estrela. Uma tabela de **fatos** existe no centro da estrela e contém chaves primárias e estrangeiras para tabelas de **dimensões associadas**, bem como dados agregados dos sistemas operacionais ou transacionais. As tabelas de dimensão descrevem os dados e são incluídas com base nas necessidades de negócios. Um esquema em estrela não é normalizado e fornece modelagem simples sem a necessidade de junções complexas.

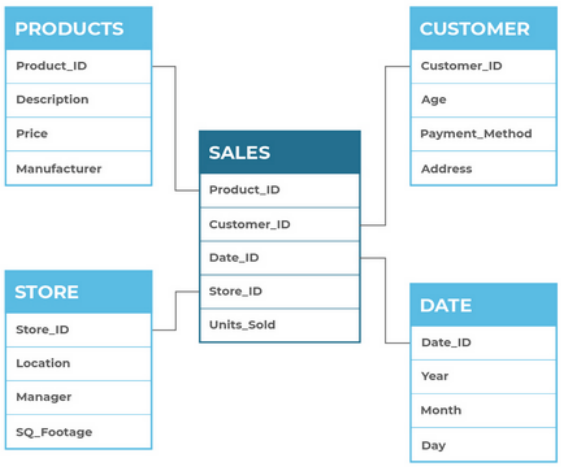


Figura 6 Exemplo de um esquema estrela

Esquema de floco de neve (Snowflake schema)

O design do esquema floco de neve contém os mesmos dados que existiriam em um esquema em estrela, e a tabela de fatos e as tabelas de dimensões têm a mesma aparência. A principal diferença entre os dois é que o esquema floco de neve é normalizado. O processo de normalização do projeto é conhecido como floco de neve. O esquema floco de neve também requer menos trabalho para adicionar mais dados às dimensões existentes e requer menos armazenamento devido à falta de redundância no processo de normalização. A Figura 7 exibe um exemplo de um esquema de floco de neve.

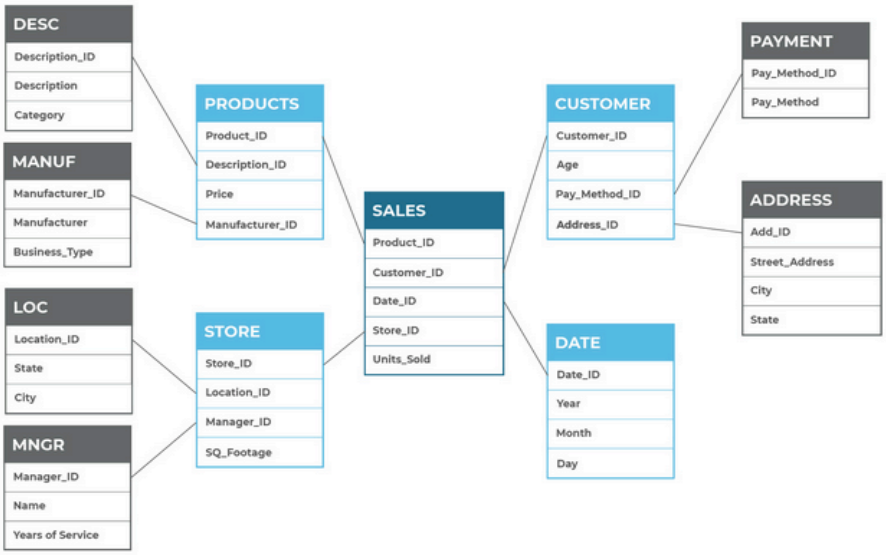


Figura 7 Exemplo de um esquema floco de neve (snowflake)

4.2.2. NoSQL, Big Data, Data Lakes e Data Warehousing

Ao contrário da abordagem tradicional de banco de dados relacional para armazenamento de dados, o NoSQL é uma abordagem alternativa que utiliza bancos de dados não relacionais e não estruturados. O NoSQL pode armazenar dados de qualquer forma porque não é limitado pelas estruturas estritamente definidas dos bancos de dados relacionais. Devido à falta de clareza e requisitos em torno da estrutura dos dados, muitas vezes não é possível desenvolver um esquema. Assim, os bancos de dados NoSQL permitem a flexibilidade de armazenar e consultar dados não estruturados. Isso é realizado por meio de uma organização orientada a documentos, em vez da organização orientada a tabelas de bancos de dados SQL estruturados. No entanto, é importante observar que esse tipo de armazenamento de dados também requer processamento e armazenamento adicionais.

Big data é um conceito para lidar com grandes quantidades de dados brutos e não estruturados em vários tipos e formatos. Torna-se rapidamente difícil para um data warehouse gerenciar esse tipo de estratégia de dados e o modelo de big data tenta resolver o problema. Devido ao tamanho, complexidade e natureza dinâmica do big data, os dados geralmente são transformados durante a análise e requerem poder de processamento significativo.

O conceito relativamente novo de data lakes oferece uma abordagem descentralizada para armazenamento e análise de dados, em vez da abordagem centralizada empregada por data warehouses tradicionais. Um data lake prefere ter repositórios de dados brutos de sistemas operacionais ou transacionais de origem disponíveis para analistas e cientistas de dados, em vez de transformar e carregar todos os dados em um repositório centralizado. Esse conceito fornece uma estratégia de armazenamento de dados e limita o pré-processamento e a governança rígida, o que certamente pode trazer benefícios e desafios para a organização. Após a análise e processamento de dados, os dados em um data lake podem ser incorporados a um data warehouse para armazenamento de longo prazo e análise futura, embora um data lake não seja necessariamente um substituto para um data warehouse.

4.3. O Processo de Preparação de Dados

A preparação de dados garante a prontidão de um conjunto de dados para análise. Em geral, esse processo consiste em preparar dados brutos para ingestão em uma ferramenta ou serviço de análise de dados. Como consideramos brevemente no módulo anterior, os dados devem passar por um processo definido chamado extrair, transformar, carregar (ETL).

- **Extração**: os dados são extraídos de sistemas de origem, repositórios e ferramentas.
- **Transformação**: os dados são limpos, normalizados e agregados para facilitar a análise.
- **Carga (load)**: os dados são carregados em um banco de dados comum, data warehouse, etc. para facilitar o acesso comum e uma única fonte de verdade para análise.

Embora o ETL descreva o processo geral, há muitas etapas detalhadas que geralmente estão envolvidas nas fases preparatórias da análise. Isso inclui agregação, combinação ou separação de campos, normalização do formato de um ponto de dados, codificação, transcrição, tratamento de valores nulos ou ausentes, verificação de erros, etc.

4.4. Representação do Cubo de Dados

Por que dois são usados dois modelos distintos para representar um DW?

Analistas de negócios tipicamente pensam sobre os problemas a partir de uma perspectiva de fatores e e as variáveis resultantes. Um fator é geralmente uma variável qualitativa, tal como localidade, impactando a variável calculada, como o turnover de empregados. Os analistas de negócios sempre usarão um diagrama para representar relacionamentos entre os fatores e as variáveis resultantes. Um diagrama pode mostrar a direção dos relacionamentos, se um impacto é positivo ou negativo, e influências diretas e indiretas.

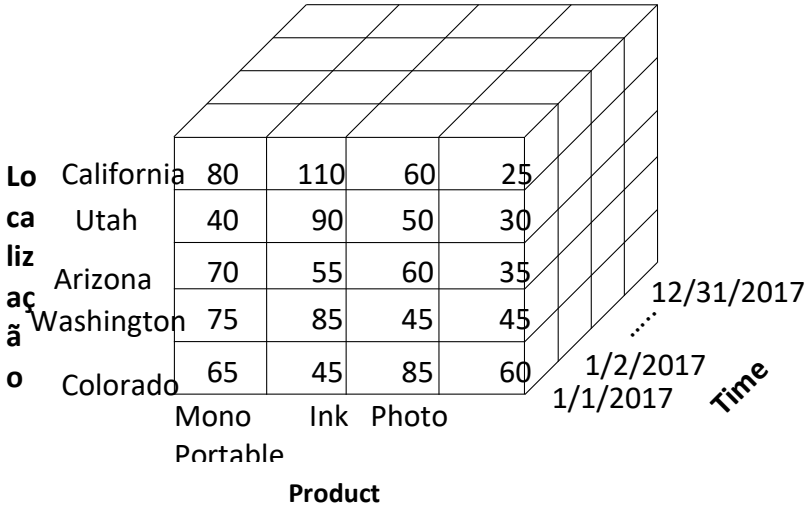
Esta variação de um diagrama de espinha de peixe mostra quatro fatores: gestão, localidade, mercado, e remuneração, os quais influenciam diretamente o turnover dos empregados. A perspectiva do analista de negócios dá uma ideia para a representação do DW. Uma representação de um DW deveria suportar este tipo de raciocínio sobre os problemas dos negócios. Os primeiros desenvolvedores de software para DW desenvolveram um modelo que suporta diretamente este tipo de raciocínio.

Um cubo de dados suporta esta perspectiva de análise de negócios. Um cubo de dados fornece uma disposição multidimensional de fatores como dimensões e variáveis quantitativas nas células dos cubos de dados. Uma dimensão é um item nomeado de uma linha ou coluna. Por exemplo, uma dimensão pode ser o tamanho da cidade ou tipo de plano de saúde ofertado. Um cubo de dados é multidimensional. Ele não se limita a duas ou três dimensões.

Uma métrica é uma variável quantitativa de interesse armazenada nas células de um cubo de dados. Por exemplo, uma métrica pode ser o turnover dos empregados, uma métrica importante sobre o custo do emprego. Uma célula pode conter múltiplas métricas visando a flexibilidade. Em duas ou três dimensões, os cubos de dados podem ser facilmente visualizados. Este cubo de dados tridimensional sobe as vendas contém localidade, produto e tempo como suas dimensões. A dimensão localidade, gravada nas linhas contém os estados dos EUA, tais como Califórnia e Utah. A dimensão do produto, que está nas colunas, mostra os tipos de impressoras tais como laser monocromática, jato de tinta... A dimensão de tempo, na profundidade, ou se preferir, no eixo z, mostra as datas.

Uma célula contém as vendas em milhares de dólares americanos, para uma combinação de estado, tipo de impressora, e data. Por exemplo, as vendas de impressoras laser monocromáticas no estado do Colorado, em 1º de janeiro de 2013, totalizaram US$65000, já que a unidade das células está em milhares de dólares americanos. A visualização não é simples para cubos de dados com mais de três dimensões. Outras aulas neste módulo mostrarão visualizações fornecidas em programas de software para cubos de dados com mais de duas dimensões.

O cubo de dados das vendas dá uma ideia sobre a extensão da representação dos cubos de dados. Uma melhoria importante é a necessidade da representação hierárquica de algumas dimensões. Por exemplo, a dimensão localidade pode conter região, país, província ou estado, cidade e CEP. O cubo de dados das vendas mostra apenas o nível dos estados dos EUA, mas, claramente a localidade tem uma estrutura hierárquica. Em muitos tipos de análises de negócios, raciocinar sobre a estrutura hierárquica de uma dada dimensão é importante.



A dispersão ou o fenômeno das células vazias é algo comum nos cubos de dados. O cubo de dados das vendas não mostra as células vazias, já que apenas a face mais externa do cubo está sendo exibida. É comum que algumas combinações entre estados, tipos de impressoras e datas não tenham nenhuma venda. Isto é, vendas iguais a zero. A dispersão aumenta à medida que o detalhe granular aumenta, tal como dos estados para as cidades, e o número de dimensões aumenta, tal como de três para dez dimensões. Para cubos de dados enormes, a maioria das células pode estar vazia. A dispersão impacta a visualização e a necessidade de espaço de armazenamento. Duas extensões importantes para as células são as métricas múltiplas e métricas derivadas (calculadas). Tipicamente, uma organização tem um conjunto de métricas que são importantes de serem acompanhadas em uma determinada área. Por exemplo, para vendas no varejo, o "número de transações", o "número de unidades" e as "vendas brutas" são importantes métricas.

Métricas derivadas, ou calculadas, tais como as vendas por transação também são importantes. A propriedade de agregação indica a disponibilidade de operações de totalização das métricas. Métricas aditivas podem ser totalizadas em todas as dimensões, usando a adição. Métricas aditivas comuns são: vendas, custos e lucro. Métricas semi-aditivas podem ser sumarizadas em algumas dimensões, mas não em todas elas, tipicamente não nas dimensões de tempo. Métricas periódicas como as de saldos contábeis e de níveis de inventário são semi-aditivas. Não-aditivas são as métricas que não podem ser totalizadas em quaisquer dimensões.

Fatos históricos envolvendo entidades individuais, como um preço unitário, são métricas não-aditivas. Algumas métricas não-aditivas podem ser convertidas em aditivas ou semi-aditivas. Por exemplo, preço estendido, que é o preço unitário vezes a quantidade - é aditiva embora o preço unitário seja uma métrica não-

aditiva. Um analista de negócios que não entende as operações permitidas para uma dada métrica pode realizar operações que não tenham nenhum significado. Portanto, compreender a agregação de métricas é importante para o desenho de um DW e um DW. Consideremos um cubo de dados com várias dimensões hierárquicas: curso, aluno, Curso, aluno e tempo. e quatro medidas: horas de crédito, as notas, o curso, e a receita dos cursos.

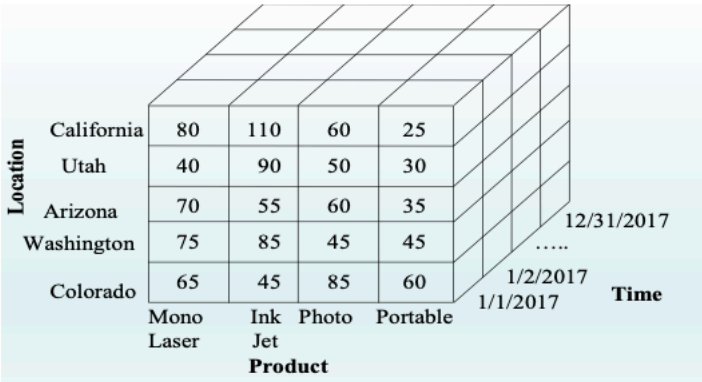
Respondendo à pergunta inicial, dois modelos são importantes para DWs. Para representar a perspectiva do analista de negócios, os cubos de dados são perfeitos. Os primeiros softwares de DW usavam a representação de cubos de dados para dar suporte aos analistas de negócios. Com o crescimento do uso de DW, no entanto, as limitações da representação dos cubos de dados se tornaram aparentes. Em particular, dispersão e falta de integração com um SGBD relacional se tornaram problemas principais. Fornecedores de SGBDs logo perceberam o potencial do mercado para DWs, e desenvolveram produtos com funcionalidades para suportar grandes DWs.

4.4.1. Operações com o Cubo de Dados

Qual o significado genérico para a locução verbal "ser pivô"? O que pivô significa para um cubo de dados?

Analistas de negócios geralmente querem pegar um subconjunto de um cubo de dados, já que tais cubos de dados com mais do que cinco dimensões são comuns em DWs. Um analista de negócios pode querer focar em um subconjunto de dimensões, tal como nos produtos e localidades de um estado dos EUA em particular, ou um subconjunto de valores membros para uma ou mais dimensões, tal como os estados do oeste na dimensão de localidade. Os analistas de negócios também querem alterar o nível de detalhamento nas dimensões hierárquicas, tais como passar dos estados dos EUA para cidades, ou das datas específicas para semanas. Os valores medidos nas células recalculados assim que os níveis de detalhamento são alterados.

Os analistas de negócios podem querer alterar a aparência de um cubo de dados, rodando as dimensões, como mudar a posição das dimensões de localidade e de produtos. Fatiar é uma das operações de subconjuntos. Usar um operador fatiador (slice), um analista de negócios pode focar no subconjunto das dimensões, trocando a dimensão por um valor único. Por exemplo, esta operação de fatiar troca a dimensão com a data pontual de 1º de Janeiro de 2013. Esta operação de fatiar apenas mostra a face frontal do cubo de dados, com o primeiro valor de profundidade, ou eixo z, que é 01/janeiro/2017.



Uma variação do operador que fatia permite ao tomador de decisões totalizar todos os membros, ao invés de focar apenas num único membro. O operador totalizador da fatia substitui uma ou mais dimensões pelos cálculos de totalização. O cálculo de totalização geralmente indica um valor total para todos os membros ou uma tendência central da dimensão, tal como um valor da média ou como um valor da mediana. Este exemplo mostra o resultado de uma operação de totalização da fatia com uma dimensão produto que é substituída pela soma das vendas de todos os produtos. Uma nova coluna, chamada total das vendas, pode ser adicionada para armazenar o geral das vendas dos produtos do ano todo. Dado que as dimensões individuais podem conter um grande número de membros, os usuários precisam focar em um

(Location × Product Slice for Time = 1/1/2017)

Location	Product			
	Mono Laser	*Ink Jet*	*Photo*	*Portable*
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

Location	Product			
	Mono Laser	*Ink Jet*	*Photo*	*Portable*
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

(Utah, Colorado, Arizona Dice)

Location	Product			
	Mono Laser	*Ink Jet*	*Photo*	*Portable*
Utah	40	90	50	30
Arizona	70	55	60	35
Colorado	65	45	85	60

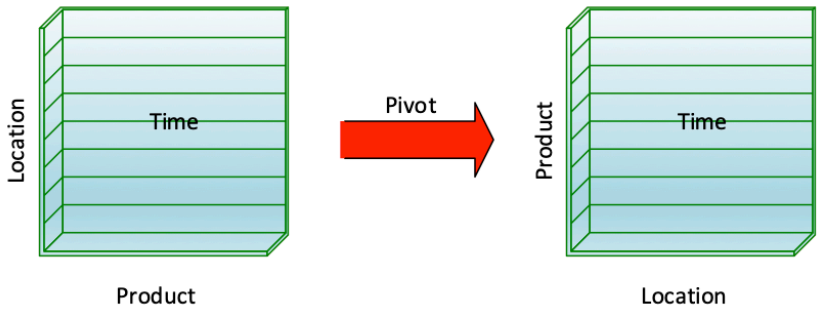
Exemplo de Dice Operator

subconjunto de membros para conseguirem compreendê-los. O operador de sub-cubos (*dice operator*) substitui a dimensão por um subconjunto de valores da dimensão. Este exemplo mostra o resultado da operação de sub-cubo para exibir as vendas dos estados norte-americanos de Utah, Arizona, e Colorado, de 01 de janeiro de 2013. Uma operação de sub-cubo tipicamente permite uma operação de fatiar e retorna um subconjunto de células que foram exibidas na operação que anteriormente fatiou o cubo. Os analistas de negócios geralmente desejam navegar entre os níveis das dimensões hierárquicas.

O operador de "drill-down" permite aos analistas navegarem de um nível mais genérico para um nível mais específico, mais detalhado, como navegar de grupos de produtos para produtos individuais. A operação de "roll-up" é oposta à de drill-down. Roll-up envolve mover de um nível mais detalhado para um nível mais abrangente de uma dimensão hierárquica. Por exemplo, um analista pode fazer roll-up das vendas de diárias para trimestrais para as necessidades dos relatórios de final de trimestre. Este exemplo mostra uma operação de drill-down no estado de Utah na dimensão de localidade. O sinal de menos em Utah indica que ocorreu uma operação de drill-down. Note que o valor das vendas em Utah, 40, estão distribuídas em três cidades: Salt Lake, Park City e Ogden. Uma operação de roll-up é o inverso. Para fazer roll-up, o sinal de menos muda para um sinal de mais, para eliminar os detalhes das cidades e somar as métricas de valores nas cidades. O operador pivô suporta rearranjos nas dimensões em um cubo de dados. Por exemplo, as posições das dimensões produto e localidade podem ser invertidas no cubo de dados das vendas de modo que o produto apareça nas colunas e a localidade, nas linhas. O operador pivô permite que as dimensões sejam apresentadas na ordem visual mais apropriada. Esta tabela mostra um resumo conveniente dos operadores de cubos de dados mais comuns. Foram sugeridos muitos outros operadores, mas não são de uso comum.

Drill-down Example

Location	Product			
	Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
- Utah				
Salt Lake	20	20	10	15
Park City	5	30	10	5
Ogden	15	40	30	10
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60



Respondendo às perguntas iniciais, o significado comum da locução verbal "ser pivô" é rotacionar um objeto ao redor de um ponto, ser o eixo pivô. No basquete, pivô significa manter um pé no lugar enquanto

segurando a bola e rodando o outro pé. No basquete, pivô significa mudar a direção usando a base como um pivô para fazer o arremesso final e completar a jogada dupla. O operador pivô realiza a rotação no cubo de dados. Em ferramentas de software de cubos de dados, entretanto, pivô envolve rearranjar as dimensões, ao invés de rodá-las. Em cubos de dados que são maiores do que duas dimensões, múltiplas dimensões aparecem na área das linhas e das colunas porque mais de duas dimensões não podem ser exibidas de outra forma. Por exemplo, para exibir um cudo de dados com localidade, produto e tempo (datas), a dimensão de tempo pode ser exibida na área da linha, dentro da dimensão de localidade. Uma operação pivô poderia rearranjar o cubo de dados de modo que a dimensão de localidade exiba dentro dela a dimensão de tempo.

Operação	Objetivo	Descrição
Slice	Dar atenção no subconjunto de dimensões	Substitui uma dimensão com um número de valores ou um sumário de suas medidas
Dice	Dar atenção a um subconjunto de valores membros	Substitui dimensões com um subconjunto de membros
Drill-down	Obter mais detalhes sobre uma dimensão	Navegar de um nível mais geral para um mais específico
Roll-up	Sumarizar detalhes sobre uma dimensão	Navegar de um nível mais específico para uma mais geral
Pivot	Apresentar dados em uma ordem diferente	Rearrumar as dimensões em um cubo de dados

4.5. Metodologias de Projeto de Data Warehouse

Qual metodologia de modelagem você prefere, e por quê?

A metodologia dá suporte ao trabalho de modelagem de DWs complexos, envolvendo muitas fontes de dados e partes de uma empresa. Uma metodologia de modelagem é uma ferramenta vital no desenvolvimento de um DW, na combinação das fases, no trabalho de automação, e no gerenciamento do projeto. Sem uma metodologia apropriada, os melhores esforços tenderão ao fracasso na produção de um DW com alto valor para uma empresa. Uma metodologia de modelagem envolve fases para criar artefatos num sistema de trabalho. Artefatos da modelagem de um DW são modelos dimensionais, a modelagem de um esquema usando padrões apresentados nas outras aulas, procedimentos de integração de dados e nos data marts que darão suporte às análises do negócio. Ambos os processos humanos e automatizados são usados numa metodologia de modelagem. Habilidades em gerenciamento de projetos são necessárias para coordenar as atividades e para avaliar a qualidade e a completude dos artefatos.

Metodologias de modelagem de DWs diferem na ênfase da demanda da inteligência de negócios, do fornecimento de fontes de dados, e num possível nível de automação no processo de desenvolvimento. A demanda da inteligência de negócios envolve exigências de relatórios e análises. O fornecimento de fontes de dados envolve fontes de dados internas e externas e qualidade dos dados. A automação no processo de modelagem pode reduzir o esforço necessário na sua confecção. A automação pode ter um papel importante, porque as fontes de dados já existem como matéria prima da modelagem de um DW.

A metodologia de modelagem de um DW "guiada pela demanda", também conhecida como abordagem "guiada pelos requisitos", proposta por Kimball em 1988, é uma das primeiras metodologias de modelagem de DW. A metodologia "guiada pela demanda" possui três fases:

1. identificar os data marts, de acordo com os requisitos do usuário;
2. construir uma matriz com os data marts relacionados e uma matriz das dimensões;
3. modelar as tabelas fato.

A Metodologia Guiada pela Demanda enfatiza a identificação dos data marts para capturar a intenção de uso do DW. Após identificar os data marts, listar as possíveis dimensões de cada um deles. As dimensões, padronizadas entre os data marts, são conhecidas como dimensões conformadas. Uma matriz relacionando as dimensões conformadas dos data marts é desenvolvida para refinar a especificação inicial do data mart. O passo final envolve a especificação das tabelas fato, com uma ênfase na granularidade das tabelas fato. As granularidades típicas são transações individuais, snapshots, que são os pontos no tempo, e as linhas, cada registro, dos itens contidos nos documentos.

A granularidade é geralmente determinada pelas dimensões primárias. Após especificar a granularidade de uma tabela fato, os detalhes da dimensão são especificados, incluindo os níveis hierárquicos. Na última parte, as métricas para cada tabela fato são especificadas, incluindo as propriedades das medidas, tais como agregação e derivabilidade.

A metodologia guiada pelo Fornecimento enfatiza a análise das fontes de dados existentes. As entidades, nos Diagramas de Entidade x Relacionamento, das fontes de dados existentes são analisadas para dar um ponto de partida para a modelagem do DW. A metodologia guiada pelo fornecimento possui três fases:

1. Classificar Entidades,
2. Refinar as Dimensões e
3. Refinar o Esquema.

A Metodologia Guiada pelo Fornecimento parece gostar da automação, embora ferramentas automatizadas para dar suporte à metodologia não tenham sido relatadas. No primeiro passo, a abordagem guiada pelo fornecimento classifica os tipos de entidades que existem nos DERs. Entidades tipo que contenham dados de eventos em um determinado tempo são classificadas como entidades tipo transação. Entidades tipo evento tipicamente irão se tornar uma tabela fato num esquema estrela. Entidades tipo relacionadas a eventos em relacionamentos 1:m são classificadas como entidades tipo componente. Entidades tipo componente geralmente se tornam tabelas dimensão num esquema estrela. Concluindo, o primeiro passo fornece um conjunto inicial de esquemas estrela ou de um esquema constelação se contiver(em) dimensões conformadas.

O segundo passo da metodologia guiada pelo fornecimento refina as dimensões. Entidades tipo relacionadas às entidades tipo componente são marcadas como entidades tipo de classificação. Hierarquias de dimensão são formadas por entidades tipo classificação e componente. Cada sequência de entidade tipo classificação e componente que realiza um "join", uma junção de relacionamento 1:m, na mesma direção se torna uma hierarquia dimensão.

O terceiro passo da metodologia refina um esquema estrela usando dois operadores. O operador usado para compactar, "collapse", desnormaliza tipos ID das dimensões para evitar produzir flocos. O operador de agregação torna o grão mais grosso nas entidades tipo transação. A agregação de uma tabela fato pode exigir modificações na tabela dimensão primária para fazer com que as tabelas dimensão sejam consistentes com o grão da tabela fato. A metodologia de modelagem de um DW híbrido proposta em 2001 combina as metodologias de demanda e suprimento.

A Metodologia Híbrida envolve um estágio guiado pela demanda, um estágio guiado pelo fornecimento, e então, um terceiro estágio para integrar a demanda em estágios guiados pelo fornecimento. Os estágios de

demanda e de fornecimento poderiam ocorrer independentemente, como mostra este diagrama. A ênfase geral na abordagem híbrida é balancear os aspectos de demanda e de fornecimento na modelagem do DW, possivelmente auxiliado por ferramentas de automação. O estágio guiado pela demanda coleta os requisitos usando os objetivos, questões e métricas, ou seja, a abordagem GQM. A abordagem GQM fornece algumas linhas gerais informais para se definir as medidas e as dimensões dos objetivos. O segundo passo na metodologia híbrida envolve análise dos DER existentes.

A metodologia fornece as linhas gerais para identificar tabelas fato e tabelas dimensão e os DER existentes. Tabelas fato em potencial, são identificadas baseadas no número de atributos aditivos. Tabelas dimensão estão envolvidas em relacionamentos 1:m, um para vários, com as tabelas fato.

O terceiro passo da metodologia híbrida integra o modelo dimensional no estágio guiado pela demanda e o esquema estrela no estágio guiado pelo fornecimento. A metodologia provê linhas gerais para se converter ambos os modelos em um vocabulário comum usando a análise da terminologia. Após a conversão para um vocabulário comum, a metodologia fornece um processo para relacionar os modelos de demanda e fornecimento.

Respondendo à questão inicial, você deve considerar cada metodologia especialmente se tiver uma oportunidade para liderar o projeto de modelagem de um DW. Eu acho a abordagem híbrida a de maior apelo, já que ela foi desenvolvida para solucionar os contratempos das outras duas abordagens: demanda e fornecimento. A abordagem híbrida tem uma certa estrutura para a abordagem GQM na análise dos DERs existentes. Um apelo maior da abordagem guiada pela demanda é a ênfase na determinação da granularidade. Grãos das tabelas fato influenciam a flexibilidade de uso e os requisitos de capacidade de armazenamento, logo, os grãos devem ser cuidadosamente determinados.

4.6. Integração de dados

Qual processo de integração de dados resultou em falha em muitos projetos de data warehouse e por quê?

O principal objetivo de integração de dados é fornecer uma única fonte confiável para a tomada de decisão. Integrar fontes de dados envolve desafios de grandes volumes de dados, muitos formatos variáveis, e unidades de medidas, distintas frequências de atualização, dados perdidos, e falta de identificadores comuns.

Integração de dados é um fator crítico para o sucesso de projetos de data warehouse. Muitos projetos falham por conta de dificuldades inesperadas ao povoar e dar manutenção ao data warehouse. Organizações devem realizar investimentos substanciais de esforço, equipamentos, e software para vencer os desafios da integração de dados. O processo de renovação envolve fontes de dados internas e externas. Fontes de dados internas geram mudanças em ambas as tabelas de fatos e dimensões. Inserção de eventos concluídos em registro de tabelas de fatos como pedidos de compra, envios, e compras com ligações para as dimensões relacionadas. Ambas as atualizações de inserções devem ser efetuadas para tabelas de dimensões. Por exemplo, o processamento de renovação deve atualizar um registro de dimensão de cliente após o endereço dele ter sido modificado e inserir novos registros após clientes serem adicionados às fontes de dados internas.

Fontes de dados externas primeiramente envolvem alterações de dimensão para entidades seguidas por outras organizações. Gestão das diferenças de tempo entre a atualização das fontes de dados e os objetos relacionados ao warehouse é imperativo no processamento da renovação. Atraso válido é a diferença da ocorrência do evento no mundo real, o qual tem uma hora válida no armazenamento do evento em um banco de dados operacional conhecido como hora da transação. Atraso da carga é a diferença entre a hora da transação e a hora de armazenamento do evento no data warehouse, conhecido como hora da carga. Para fontes de dados internas, o processo de renovação tem algum controle sobre o atraso válido. Para fontes de dados externas, o processo de renovação geralmente não tem controle sobre o atraso válido. Portanto, um administrador de data warehouse tem mais controle sobre o atraso da carga. Além disso, um administrador de data warehouse deve

administrar o atraso da carga separadamente para fontes de dados internas e para as externas. Este diagrama mostra as fases comuns de processamento de renovação e as tarefas em cada fase. Este diagrama é genérico, então ele deve ser customizado para cada processo de renovação.

A Fase de Preparação manipula as alterações de dados a partir de cada sistema de origem. A Extração retira os dados de fontes de dados individuais. O Transporte move os dados extraídos para uma área intermediária. A Limpeza envolve uma variedade de tarefas para padronizar e melhorar a qualidade dos dados extraídos. Registros de auditoria resultam do processo de limpeza, perfazendo a completude e verificação de razoabilidade e tratando as exceções.

A Fase de Integração une fontes limpas que estão separadas em uma única fonte. Esta fusão pode envolver a remoção de inconsistências que existem nos dados de origem. Registros de auditoria resultam do processo de fusão, perfazendo verificações de completude e de razoabilidade, e manipulando as exceções. A fase de atualização envolve a propagação das alterações dos dados integrados para várias partes do data warehouse. Após a propagação, uma notificação pode ser enviada aos grupos de usuários e administradores. Além da renovação periódica, a integração de dados envolve uma carga inicial de um data warehouse. Este processo de carga inicial é menos limitado do que o processo de renovação. Requisitos de tempo para descobrir e resolver problemas de qualidade de dados pode ser difícil de estimar.

Ferramentas de perfis podem facilitar a descoberta de problemas na qualidade de dados. Problemas de qualidade de dados são geralmente resolvidos através de procedimentos de integração de dados. Se os donos da fonte de dados cooperarem, a resolução pode envolver alterações no sistema de origem dos dados. O processo de carga inicial, deve ser executado a cada grande expansão de um data warehouse. O objetivo principal, ao gerenciar o processo de renovação, é determinar a frequência de renovação para cada fonte de dados, e estabelecer agendamentos detalhados para estas renovações enquanto satisfaz restrições importantes.

O valor dos dados em relação à linha do tempo dependerá da sensitividade para tomar uma decisão baseada nos dados correntes. Algumas decisões são muito sensíveis ao tempo, como decisões de inventário para o mix de produtos em lojas. Outras decisões não são tão sensíveis ao tempo, como decisões de localização das lojas. O custo de renovação de um data warehouse inclui ambos os recursos computacionais e recursos humanos.

Recursos computacionais são necessários para todas as tarefas no fluxo de manutenção. Recursos humanos podem ser necessários em tarefas de auditoria durante a preparação nas fases de integração. O nível da qualidade de dados e da fonte de dados também afeta o nível de recursos humanos necessários. Além de somar o valor do tempo contra o custo da renovação. O administrador de data warehouse deve satisfazer as restrições do processo de renovação, restrições ou no data warehouse, ou no sistema de origem podem restringir a frequência da renovação. Restrições de acesso aos dados de origem podem ser devidas à tecnologia do legado com restrição de escalabilidade para fontes de dados internas, ou problemas de coordenação de fontes de dados externas.

Restrições de integração geralmente envolvem identificação de entidades comuns como clientes e transações ao longo dos sistemas de origem. Restrições de consistência envolvem o uso no mesmo período de tempo que o dado estiver sendo atualizado. Restrições de completude envolvem a inclusão de dados alterados em cada fonte de origem dos dados. A disponibilidade do data warehouse sempre envolve conflitos entre disponibilidade online e a carga do warehouse.

O processamento de renovação na carga inicial de um data warehouse requer investimentos substanciais em tecnologia e esforço. Ferramentas de integração de dados são importantes para aumentar a produtividade no desenvolvimento de procedimentos de integração de dados. Respondendo à questão inicial, o processo de carga inicial tem levado à falha em muitos projetos de data warehouse. O esforço em custo neste processo é difícil de estimar por conta do desconhecimento do nível da qualidade de dados nos dados de origem e falta de ferramentas que facilitem a descoberta e a resolução. Fornecedores de software têm respondido a essas necessidades desenvolvendo ferramentas robustas para determinar o perfil dos dados e para fluxos de integração de dados.

Mesmo com melhores ferramentas, o processo de carga inicial permanece o mais difícil em muitos dos projetos de data warehouse.

4.6.1. Mudança no Conceito de Dados

Qual o relacionamento entre problemas de qualidade de dados e o tipo de alteração de dados usado nos procedimentos de integração de dados?

Alteração de dados, derivada de fontes de dados internas e externas, são a entrada para povoar e renovar um data warehouse. A alteração de dados mais comum envolve inserções de novos fatos. Inserções de novas dimensões e atualizações de dimensões são menos comuns, mas ainda assim, são importantes para a captura. Exclusão de fatos e dimensões só são necessárias para corrigir os dados que não deveriam ter sido inseridos em um data warehouse. A fonte de dados traz desafios na manipulação com uma variedade de formatos e restrições nos sistemas de origem. Sistemas fonte externos geralmente não podem ser alterados.

Sistemas fonte internos podem ser alterados se os recursos estiverem disponíveis e o desempenho não for impactado. Fonte de dados armazenada em formato legado geralmente impedem a obtenção de dados usando linguagens não procedurais tais como SQL. A menos que armazenados com dados descritivos, dados do legado e páginas web podem ser difíceis de serem decompostos em partes menores. Descritivo ou metadado usualmente envolve dados XML junto a um esquema XML para fornecer interpretação do dado XML.

Alteração de dados pode ser classificado pelo nível de processamento e requerimentos do sistema de origem. Ela pode ser vista neste espaço de duas dimensões. Requerimentos do sistema de origem envolvem modificações nos sistemas de origem para receberem os dados alterados. Alterações típicas no sistema de origem são: novas colunas, tais como data e hora obrigatórias para alteração de dados que podem ser consultados, e código disparador obrigatório para dados alterados cooperativos. Uma vez que os sistemas origem são difíceis de serem alterados, dados alterados possíveis de se consultar e dados alterados cooperativos podem não estar disponíveis.

Nível de processamento envolve consumo de recurso e desenvolvimento necessários para procedimentos de integração de dados. Registros e salvas instantâneas de alteração de dados envolvem processamento substancial. A quantidade de processamento para registrar uma alteração varia, então seus requisitos de processamento podem ser maiores do que uma salva instantânea dos dados alterados. Se um sistema fonte ainda não gera nenhum registro de alteração, é pouco provável que um registro de alteração de dados esteja disponível.

Alteração de dados via cooperativa envolvem notificação a partir de um sistema fonte sobre as alterações. A notificação ocorre tipicamente durante o tempo da transação usando um gatilho. Um gatilho é uma regra executada por um SGBD quando um evento ocorre, por exemplo, quando inserimos uma nova linha. Um gatilho envolve desenvolvimento de software e execução como parte de um sistema fonte. Alteração de dados cooperativa, podem ser gravadas imediatamente no DW. ou colocadas numa fila ou área de teste para posterior processamento, possivelmente com outras alterações. Dado que a alteração cooperativa de dados requer modificações no sistema de origem, ela tem tradicionalmente sido a menos comum dos formatos de alteração de dados. Entretanto, à medida que os projetos de dw amadurecem, e os sistemas legado são desenvolvidos novamente, alteração cooperativa de dados vão se tornando mais comuns. Alteração de dados registradas envolvem arquivos de log que gravam as alterações ou outras atividades do usuário. Por exemplo, um log de transação contém cada alteração que a transação efetuou, e um log da web contém histórias de acesso à página chamados de registros de clique dos visitantes da internet.

Registrar no log as alterações de dados não envolve nenhuma modificação nos sistemas de origem, já que os logs já estão prontamente disponíveis na maioria dos sistemas de origem. Este diagrama mostra um exemplo de um log da internet. Processamento substancial durante a integração dos dados é necessário para logs da internet para decompor um texto já que os logs da internet seguem vários formatos de padrões. Além disso, logs

da internet gravam visitas às páginas, então, um processamento substancial se faz necessário para ligar os registros de log relacionados. Como o nome diz, alteração de dados que pode ser consultada vem diretamente de uma fonte de dados via uma consulta.

Alteração de dados que pode ser consultada exige selo de data e hora nos dados de origem. Dado que poucas fontes de dados contêm selo de data e hora para todos os dados, alteração de dados que pode ser consultada geralmente são aumentadas com outros tipos de alteração de dados. Alteração de dados que pode ser consultada é mais comumente aplicável às tabelas "fato" usando colunas como data do pedido, data de envio e data da contratação, as quais são gravadas nos bancos de dados de origem operacionais. Uma salva instantânea dos dados alterados envolve salvas periódicas de dados do banco de origem. Para obter dados alterados, uma operação de diferença usa as duas salvas instantâneas mais recentes.

O resultado de uma operação de diferença é chamado de delta. Gerar um delta envolve comparar arquivos fonte para identificar novas linhas, linhas alteradas e linhas excluídas. Salvas instantâneas são a única forma de dados alterados sem requisitos no sistema de origem. Salvas instantâneas são usadas principalmente para sistemas legado em fontes de dados externas. Dado que recuperar dados em arquivos fonte pode consumir muitos recursos pode haver restrições sobre o tempo e a frequência de recuperar uma salva instantânea. Problemas de qualidade de dados podem ocorrer em todos os tipos de dados alterados, mas são mais comuns em sistemas legado.

Problemas de qualidade de dados devem ser endereçados em procedimentos de integração de dados, a menos que as alterações possam ser feitas nos sistemas de origem. Esses são problemas típicos de qualidade de dados encontrados na alteração de dados. Múltiplos identificadores. Algumas fontes de dados usam chaves primárias distintas para a mesma entidade, tais como números distintos para o código do cliente. Unidades distintas. Unidades de medida distintas e granularidades para medidas podem ser usadas em fontes de dados.

Valores ausentes. Dados podem não existir em algumas fontes de dados e valores default podem variar em cada fonte de dados distinta. Dados texto não padronizados. Fontes de dados podem combinar múltiplos dados em uma única coluna de tipo texto, tal como o endereço que poderia conter múltiplos componentes: rua, número, cep, cidade, tudo em única coluna. Além disso, o formato dos componentes do endereço pode variar em cada fonte de dados distinta. Dados conflitantes. Algumas fontes de dados podem ter dados conflitantes, tais como endereços distintos do mesmo cliente. Hora de atualização distinta. Algumas fontes de dados podem realizar atualizações em intervalos de tempo distintos.

Respondendo à pergunta inicial, alteração de dados de sistemas legado tipicamente têm mais problemas de qualidade de dados do que os sistemas modernos. Sistemas legado geralmente não têm acesso SQL, nem dados descritivos e nem restrições de integridade. Grandes quantidades de recursos podem ser necessárias para incluir nos sistemas legado tais características padrões. Para resolver os problemas de qualidade de dados nos sistemas legado, vários níveis de manipulação manual de exceções podem ser necessários.

4.6.2. Atividades de Limpeza de Dados

As abordagens para valores ausentes apresentadas nesta aula são proativas ou reativas?

A decomposição de objetos complexos, usando texto, em suas partes constituídas. Para integração de dados, a decomposição é importante para decompor dados em texto de múltiplos propósitos em campos individuais. Por exemplo, decompor um endereço físico, números de telefone e endereços de e-mail são transformações típicas de data warehouses de marketing. Para facilitar as análises alvo de marketing, estes campos constituintes devem ser decompostos em partes padronizadas.

A decomposição tem sido estudada na ciência da computação há várias décadas. A ferramenta padrão para decomposição de livre contexto é uma expressão regular. De livre contexto francamente quer dizer que o significado de um símbolo não depende de sua relação com outros símbolos ou textos.

O processamento da linguagem natural nasceu da decomposição e do entendimento do texto de linguagem natural que é dependente do contexto. Fontes de dados que contém endereços em um único campo tipicamente requerem uma decomposição em componentes padronizados, tais como nome da rua, número, cidade, estado, país e CEP. Este exemplo demonstra a decomposição do nome do cliente e de seu endereço em campos componentes. Algumas decomposições são baseadas na posição, com cada nova linha fornecendo diferentes grupos de campos. Corrigir os valores envolve a resolução de valores ausentes e conflitantes.

Para valores ausentes, a resolução depende do significado de um valor ausente. Valores ausentes inaplicáveis a uma entidade podem geralmente ser resolvidos com valores default. Por exemplo, valores ausentes de um pedido sem um empregado podem ser trocados com um valor default indicando que é um pedido que veio da internet. Valores ausentes que são desconhecidos ao invés de inaplicáveis são mais difíceis de serem resolvidos. Por exemplo, ausência de data de nascimento, de partes de um endereço e de médias das notas são mais difíceis de serem solucionados.

Uma abordagem para valores desconhecidos envolve valores típicos. Para valores numéricos, uma mediana ou mesmo um valor médio podem ser usados. Para valores desconhecidos não numéricos, a moda, que é o valor mais frequente, pode ser usada. Uma abordagem mais complexa para valores desconhecidos é prever valores usando relacionamentos com outros campos.

Abordagens mais complexas irão prever os valores ausentes usando algoritmos de mineração de dados. Para valores conflitantes, abordagens simples como a do valor mais recente podem ser usadas. Determinar um valor mais confiável geralmente envolve uma investigação por um perito no domínio. Investigações detalhadas, possivelmente conduzidas por servidores de pesquisa, podem solucionar alguns casos de valores desconhecidos e de valores conflitantes. Este exemplo demonstra o resultado de uma investigação para determinar os componentes ausentes do endereço em um registro de um empregado. Um mapa e o conhecimento sobre a localização do prédio puderam ser usados para se obter os componentes ausentes do endereço.

A padronização envolve regras de conversão para transformar valores em representações preferenciais. Regras de conversão são geralmente desenvolvidas para unidades de medida e abreviações. Ambos padrões e regras customizadas podem ser desenvolvidos. Além disso, serviços de padronização de dados podem ser comprados para nomes, e detalhes de produtos, mesmo assim, uma customização pode ser necessária. Este exemplo acrescenta uma padronização ao exemplo previamente corrigido. A função, a empresa, a rua, e o estado foram padronizados usando um dicionário de padrão de nomes. Tal dicionário contém o valor completo dos valores que são tipicamente abreviados.

Respondendo à questão inicial, as abordagens apresentadas nesta aula são reativas, elas tentam solucionar problemas que ocorrem em fontes de dados existentes. Se os sistemas fontes não podem ser alterados, abordagens reativas são a única escolha. Abordagens proativas podem ser de baixo custo se alterações nos procedimentos de coleta de dados puderem ser feitas em partes distintas de uma organização. Padrões podem ser facilitados por esquemas XML com regras claras sobre o intercâmbio de dados. Talvez seja possível aplicar padrões a fontes de dados externas e os usuários externos serão beneficiados com um data warehouse.

4.6.3. Identificação de Padrões com Expressões Regulares

Como você desenvolve expressões regulares para padrões complexos em endereços, endereços url da internet, cartões de crédito e números de telefones?

As expressões regulares especificam padrões de validação de campos tipo texto com múltiplos componentes, comuns em tarefas de integração de dados. As ferramentas de expressões regulares são largamente suportadas em ferramentas de integração de dados, em SGBDs, nas interfaces de aplicativos de programação que testam os web sites. Uma expressão regular, ou REGEX para abreviar, contém literais, meta-caracteres e sequências de caracteres escape. Um literal é um caractere de identificação exata. Meta-caracteres,

ou caracteres de padrão identificável, dão significado especial dentro de uma expressão de busca, dando força às expressões regulares. Sequências de escape removem o significado especial dos meta-caracteres para tratá-los como literais comuns. O meta-caractere barra invertida "\" posicionado antes de outro meta-caractere remove o significado especial do meta-caractere. Para realizar a identificação de padrões, o usuário fornece uma expressão regular conhecida como expressão de busca em uma string alvo.

A expressão de busca especifica que padrão deve ser procurado na string alvo. Neste exemplo, a expressão de busca contém sete meta-caracteres. O circunflexo, o abre-colchetes, o fecha-colchetes, o sinal de mais, o sinal de menos ou hífen, a barra invertida e o sinal de cifrão, ou dólar. Seis caracteres literais: as letras minúsculas a, z, c, o, m mais o sinal de ponto final. E uma sequência escape, a barra invertida e um ponto, para desativar o significado especial do símbolo de ponto final.

```
^[a-z]+\.com$
```

O resultado de identificação, em "match result" mostra a parte da string alvo, que corresponde à expressão de busca. Meta-caracteres, ou caracteres de padrão identificável, dão força às expressões regulares. Esse diagrama exibe os meta-caracteres mais usados. Os meta-caracteres de iteração, ou de quantificação são: interrogação, asterisco, sinal de adição, e as chavetas (abre e fecha), estas dão suporte à identificação de caracteres consecutivos. As expressões de busca usam o sinal de adição para identificar um ou mais caracteres iguais aos que precedem este sinal. A posição de um meta-caractere é âncora. Os sinais de ponto final, circunflexo e o cifrão, dão suporte à identificação em posições especificadas de uma string. A expressão de busca usa o sinal de circunflexo para localizar o início, e o sinal de cifrão para identificar o final de uma string alvo.

Metacaracter	Tipo	Meaning
?	Iteração	Corresponde à ocorrência do caracter 0 ou 1 vez
*	Iteração	Corresponde à ocorrência do caracter 0 ou mais vezes
+	Iteração	Corresponde à ocorrência do caracter 1 ou mais vezes
{n}	Iteração	Corresponde à ocorrência do caracter exatamente n vezes
{n,m}	Iteração	Corresponde à ocorrência do character pelo menos n vezes e no máximo m vezes
[]	Intervalo	Corresponde a um conjunto de caracteres
^	Posição	Corresponde a o início da string alvo; só tem sentido como primeiro caracter em uma expressão regular
^	Intervalo	Negação de um padrão de pesquisa se o ^ estiver dentro de []. Hífen dentro de [] define um interval de caracteres.
$	Posição	Corresponde à ocorrência no fim de uma string alvoo; só tem sentido no fim de uma expressão regular.
.	Posição	Corresponde a qualquer caractere exceto um caractere de nova linha apenas na posição especificada
|	Alteração	Corresponde a qualquer padrão à esquerda ou à direita do |

Na outra categoria, os meta-caracteres de faixa de valores vão dentro de abre e fecha colchetes e identificam um único caractere dentro de uma faixa de caracteres especificada. A expressão de busca usa a faixa de letras

minúsculas, de "a" até "z", que está especificada aqui dentro dos colchetes. Observe que o sinal de adição,"+", se aplica à faixa de letras minúsculas. O sinal de barra invertida desativa o significado dos meta-caracteres que virão em seguida. A expressão de busca usa a barra invertida para desativar o significado do símbolo de ponto final. A alteração de um meta-caractere, a barra vertical, suporta partes opcionais de padrões de busca. Esta tabela mostra um resumo conveniente de meta-caracteres comuns.

Para entender as expressões de busca, você precisa trabalhar com vários exemplos. Esta tabela mostra seis exemplos com múltiplas strings alvo em cada um deles. Aqui estão algumas breves anotações sobre estes exemplos. No exemplo um, a interrogação identifica o caractere precedente zero vezes na primeira string alvo. No exemplo dois, o asterisco identifica o caractere precedente zero vezes na terceira string alvo. No terceiro exemplo, o meta-caractere sinal de adição não identifica a terceira string alvo porque o terceiro caractere é "o", e não "e". No quarto exemplo, a expressão de busca não identifica a terceira string alvo, porque ela não contém nenhuma das letras dentro dos colchetes. Os dois últimos exemplos são de meta-caracteres de iteração que especificam o número de identificações. No exemplo cinco, a primeira faixa deve ser identificada três vezes, e a segunda faixa, quatro vezes. No último exemplo, o caractere precedente "a", deve ser identificado entre duas e três vezes. Esta tabela mostra as expressões de busca usando meta-caracteres de posição, iteração e de alteração. Aqui estão algumas breves anotações sobre estes exemplos.

No exemplo um, a expressão de busca não identifica a primeira string alvo porque "win" não aparece no início da string alvo. No exemplo dois, a expressão de busca não identifica a segunda string alvo porque "win" não aparece ao final da string alvo. No exemplo três, o circunflexo dentro dos colchetes nega a sequência de caracteres de 0-9 identificando assim, quaisquer não-dígitos, ou seja, apenas letras. No exemplo quatro, o ponto final, que é um meta-caractere posicional na expressão de busca exige um caractere após "abc", logo, a expressão de busca não identifica a primeira string alvo. No exemplo cinco, os meta-caracteres de alteração, ou seja, barra vertical, identificam todas as três strings alvo, já que cada uma contém uma das escolhas: "dog", "cat" ou "frog". Este exemplo mostra expressões de busca mais complexas.

Os últimos três exemplos contêm grupos de identificação de partes da string alvo, delimitados entre parênteses. Devido a complexidade destes exemplos, eu recomendo que você use um website regular de teste de expressões para provar cada uma delas. Aqui, brevemente, use **regex101.com** para testar o primeiro exemplo para nomes de usuários simplificados. Após copiar a expressão de busca para o campo de expressão regular o testador fornece uma explicação detalhada do lado direito da tela. A explicação contém quatro componentes da expressão de busca, o circunflexo, as quatro classes de caracteres, as minúsculas de "a" a "z", zero a nove, sublinhado, e um hífen dentro dos colchetes. O quantificador de números 3 e 16 dentro das chavetas e o sinal de cifrão. Vou testar agora várias strings. Depois de digitar em minúsculas "joe", o testador indica uma identificação em quatro passos. A identificação ainda ocorre em quatro passos, para joe_123, e para joe-7890. A identificação não ocorre para a string "jo", nenhuma identificação em dois passos. Joe com o J maiúsculo, nenhuma identificação em dois passos. Joe com uma exclamação, "Joe!", nenhuma identificação em três passos. E para username_too_long, nenhuma identificação em 16 passos.

Respondendo à questão inicial, você deve desenvolver com muito cuidado as expressões regulares complexas. Existem muitos outros meta-caracteres, notavelmente agrupados para identificação de partes de uma string alvo. Você deve se lembrar que expressões regulares apenas se aplicam ao contexto de validação livre, não de uma validação de linguagem natural. Para campos comuns, tais como endereços físicos, endereços url da internet, números de cartão de crédito, e números de telefone, você pode pesquisar expressões válidas em bibliotecas de expressões regulares. Tais campos são complexos de serem validados e de praticar com eles, então, expressões regulares para eles são difíceis de serem escritas e depuradas do zero.

Search Expression	Target Strings	Evaluation
“colou?r”	“color”, “colour”	Corresponde a ambas as strings alvos
“tre*”	“tree”, “tread”, “trough”	Corresponde a todas as três strings alvo; corresponde ao caracter anterior 0 vezes na terceira string
“tre+”	“tree”, “tread”, “trough”	Não corresponde à terceira string
“[abcd]”	“dog”, “fond” , “pen”	Encontra as duas primeiras strings mas não a terceira
“[0-9]{3}-[0-9]{4}”	“123-4567”, “1234-567”	Encontra a primeira string mas não a segunda
“ba{2,3}b”	“baab”, “baaab”, “bab”, “baaaab”	Encontra as primeiras duas strins mas não as duas últimas

4.6.4. Correspondência e Consolidação

Qual erro custa mais na correspondência de entidades: uma falsa correspondência de duas entidades distintas, ou uma falsa não correspondência entre duas entidades idênticas?

A correspondência de entidades identifica registros duplicados em duas ou mais fontes de dados quando nenhum identificador comum confiável existir. A aplicação clássica envolve a identificação de clientes duplicados e fontes de dados de diferentes empresas. Por não existir um identificador comum, as duplicidades devem ser identificadas a partir de outros campos comuns como nomes, componentes de endereço, números de telefone e idades. Por tais campos comuns advirem de distintas fontes de dados, inconsistências e representações não padronizadas podem existir, complicando o processo de correspondência.

O processo de correspondência de entidades tem sido estudado como um problema de mineração de dados, 'data mining', há décadas na ciência da computação, em sistemas de informação e em estatística. Vários nomes foram atribuídos a este problema, tais como: ligação de registros, identificação de entidades, e resolução de entidades. Muitas abordagens têm sido desenvolvidas, mas nenhuma abordagem dominante apareceu. Além disso, serviços comerciais de customização para requisitos de fontes de dados individuais podem corresponder as entidades, mas, geralmente, com um custo relativamente alto. Para melhorar os resultados de correspondência das entidades, a empresa deve considerar investimentos na melhoria da consistência e na completude nas fontes de dados de origem.

Source 1	
First name	Aimee
Middle name	Christina
Last name	Parker
Job title	Product Manager
Firm	Microsoft Corporation
Street	15580 NE 31st Street
City	Redmond
State	WA
Postal Code	98052
Country	USA

Source 2	
First name	Aimee
Middle name	C.
Last name	Parker-Lewis
Job title	Prod. Mgr.
Firm	Microsoft
Street	16517 78th Place NE
City	Bothell
State	WA
Postal Code	98020
Country	USA

Este exemplo simples descreve as dificuldades de correspondência de duas entidades. As fontes de dados não têm um identificador comum para confiavelmente realizarem a correspondência, então campos não-únicos devem ser usados. O texto em vermelho indica conflito entre os dois casos. A fonte de dados um contém o nome de solteira, anterior ao casamento, e o endereço comercial. A fonte de dados dois possui o nome de casada e o endereço residencial. O nome do meio, a função exercida, e a empresa também possuem valores distintos. A experiência indica que tais registros são praticamente correspondentes.

	Target
First name	Aimee
Middle name	Christina
Last name	Parker-Lewis
Job title	Product Manager
Firm	Microsoft Corporation
Street	16517 78th Place NE
City	Bothell
State	WA
Postal Code	98020
Country	USA

Dada a proximidade de Bothell e Redmond no estado de Washington, a correspondência do primeiro nome com a mesma grafia incomum, parte do último sobrenome e a correspondência da função exercida ao padronizar a empresa e as funções exercidas. A diferença no último sobrenome pode ser explicada combinando-se os sobrenomes após o casamento. Um algoritmo de correspondência de entidades, sem este especialista no assunto, pode levar a uma correspondência não conclusiva, ao invés de determinar que são o mesmo dado. Uma investigação custosa feita por um especialista pode ser necessária para solucionar esta correspondência não conclusiva.

Este exemplo mostra campos comuns entre duas fontes de dados. A correspondência é mais complexa se as fontes de dados têm dados não estruturados, tais como textos, imagens e eventos nos campos das estruturas comuns. No que tange a dificuldade de correspondência das entidades, isso é importante em muitos aplicativos.

O Marketing é uma área proeminente, já que as empresas frequentemente estão interessadas em expandir suas bases de clientes. A fusão de empresas, tipicamente dispara um esforço maior de correspondência dos clientes. Agências aplicadoras da lei precisam ligar crimes e suspeitos, e combinar nomes e apelidos num único suspeito.

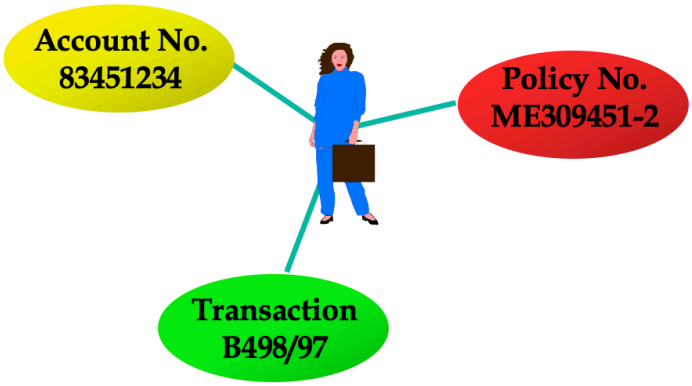
A detecção de fraude deve resolver indivíduos que reclamam benefícios usando identificadores distintos, quando o indivíduo for a mesma pessoa. Por exemplo, a mesma pessoa pode, de modo fraudulento, submeter várias solicitações de devolução de impostos para receber créditos. Analistas de negócios nos sistemas de saúde, constantemente precisam combinar registros de consultas médicas de indivíduos tratados por distintos hospitais e clínicas médicas. Há muitas outras aplicações de correspondência de entidades nos negócios e no governo.

Para obter um conhecimento mais preciso de correspondência de entidades, é necessário entender os resultados da comparação de dois casos. Nesta matriz, as linhas representam previsões, e as colunas representam os resultados reais da correspondência de duas entidades. Uma correspondência verdadeira, 'true match', envolve uma correspondência prevista e uma correspondência real, permitindo que as duas entidades sejam combinadas corretamente. Uma correspondência fala, 'false match', envolve uma correspondência prevista, mas, nenhuma correspondência real, resultando na combinação de duas entidades que deveriam permanecerem separadas.

Uma correspondência falsa envolve uma previsão de não correspondência, mas com uma correspondência real resultando em duas entidades mantidas separadas, mas que deveriam terem sido combinadas. Uma não correspondência verdadeira, 'true non match', envolve uma previsão de não correspondência, e não correspondência real, resultando em duas entidades separadas. As situações de possíveis não-correspondências envolvem previsões sem um grau de certeza suficiente para indicar uma correspondência ou uma não correspondência. Uma investigação pode ser necessária para resolver os casos inconclusivos. Entidades correspondidas podem ser unificadas, 'merged', ou ligadas, 'linked'. Se intercalar duas entidades, às vezes, dados

antigos de uma das fontes são descartados. Além disso, novos campos podem ser adicionados visando obter dados únicos de cada fonte de dados.

A ligação mantém as entidades separadas, mas estabelece um relacionamento entre elas. Para casas, a ligação combina indivíduos com família e outros relacionamentos sociais. Para transações, a ligação associa transações, como políticas de seguro distintas ou crimes, com o mesmo indivíduo, ou conjunto de indivíduos. Este exemplo mostra um resultado possível ao unir registros vistos no exemplo anterior de entidades correspondentes. No registro resultante, o endereço profissional foi excluído e o último sobrenome de casada, Parker-Lewis sobrepôs o sobrenome de solteira, Parker. Além disso, usamos valores por extenso do nome do meio, da função exercida e da empresa. A consolidação da casa envolve registros de ligação dos indivíduos que vivem na mesma casa. Esta prática é, às vezes, conhecida como 'householding', ou união familiar. Na ligação de transações, todas as contas e transações são associadas à mesma pessoa. Frequentemente, detalhes de transações distintas são armazenados em bancos de dados operacionais distintos antes que um DW seja construído. Um benefício importante do esforço de integração de dados é ligar as transações ao mesmo indivíduo por entre os bancos de dados operacionais e as fontes de dados externas.



Respondendo à questão inicial, os procedimentos de correspondência de entidades deveriam calcular os benefícios de listas de entidades unificadas, isto é, de correspondências verdadeiras e de não correspondências verdadeiras, contra o custo de ações incorretas, ou seja, falsas correspondências e falsas não correspondências, mais os custos de investigação. O custo de falsas correspondências geralmente é o maior, já que uma falsa correspondência elimina uma entidade potencial, como um cliente, em um DW. Calcular níveis de incerteza, que levam a custos de investigação, pode ser importante. Custos de investigação podem ser trabalho intensivo dos funcionários, então os custos, em certos casos, são maiores do que os custos de falsas não correspondências.

4.6.5. *Quasi*-Identificadores e Funções de Distância para Correspondência de Entidades

Por que existe uma distância na edição, geralmente usada para quantidades de texto relativamente pequenas, como correções ortográficas de cada palavra?

Os algoritmos de correspondência de entidades usam os quasi-identificadores para compensar a falta de identificadores comuns. Os quasi-identificadores podem ser quase únicos quando combinados. Num estudo publicado em 2000, Sweeney demonstrou que 87% da população dos EUA podia ser identificada por uma combinação de gênero, data de nascimento e CEP. Outros exemplos de quasi-identificadores são nomes de componentes, local dos componentes, profissão e raça~.

Antes do algoritmo de correspondência de entidades poder ser aplicado, é preciso determinar quais são os quasi-identificadores comuns. A baixa qualidade dos dados, como a ausência de valores e A baixa qualidade dos dados, como a ausência de valores e horários desconhecidos das atualizações complicam as escolhas dos quase-identificadores. As abordagens de correspondência de entidades usam funções de distância para

determinar se os quasi-identificadores de duas entidades indicam uma mesma entidade. No sentido geométrico, a distância é a quantidade de espaço entre dois pontos.

Para correspondência de entidades, um ponto é uma combinação de valores, um valor para cada quase-identificador. Quasi-identificadores numéricos são fáceis de se comparar, mas quase-identificadores textuais podem ser difíceis de serem comparados. Quase-identificadores texto tais como nome e local dos componentes. Eles se diferenciam na grafia, no tamanho e no contexto.

As funções de distância para textos podem ser usadas para comparar quasi-identificadores com estas diferenças. As funções de distância têm muita aplicabilidade além da correspondência de entidades por exemplo, na correção ortográfica. A distância de edição é uma função comum para comparar valores de textos relativamente curtos que ocorrem nos aplicativos de correspondência de entidades. A ideia básica é contar o número de caracteres usados nas operações de adição para transformar o valor de um texto fonte no valor de um texto destino. Uma operação pode deletar um caractere, inserir um caractere ou substituir um caractere por outro caractere.

A distância de edição é definida como o menor número de operações para transformar o valor de um texto fonte no valor de um texto destino. Determinar este menor número de operações envolve um algoritmo de otimização que está além do escopo desta aula. Portanto, o foco aqui é contar o número de operações e os exemplos. Os exemplos esclarecem as operações de edição que transformam valores texto, e determinam a solução mínima.

Neste exemplo, a distância adicionada para transformar "Saturday" em "Sunday", é de três operações. Este exemplo aqui mostra duas sequências de operações: A primeira envolve duas deleções, de "a" e de "t", que são seguidas pela substituição de "n" por "r". A segunda sequência envolve duas substituições, "u" pela "a", seguida de "n" pela letra "r", e duas exclusões, de "u" e de "r". A primeira sequência é a preferida, porque ela contém menos operações.

1. Sturday (delete “a”)
2. Surday (delete “t”)
3. Sunday (substitute “n” for “r”)

1. Suturday (substitute “u” for “a”)
2. Sunurday (substitute “n” for “t”)
3. Sunrday (delete “u”)
4. Sunday (delete “r”)

Este exemplo tem apenas duas sequências de operações, logo, identificar o número mínimo de operações é fácil! Para valores de textos mais complexos, um número maior de sequências precisa ser avaliado, até se encontrar a solução mínima. A distância fonética tem grande uso na aplicação das leis, para contar diferentes grafias de nomes, que possuam fonemas semelhantes. As palavras com mesmos fonemas, devem ter o mesmo valor fonético. A distância fonética basicamente codifica palavras em sons de consoantes padrões.

Duas funções de distâncias fonéticas, a Soundex e a Metaphone, têm sido largamente implementadas nos SGBDs e nas ferramentas de integração de dados. Tais funções primeiramente se distinguem pelo número de sons de consoantes utilizado. A Metaphone, com mais sons de consoantes, foi desenvolvida como uma evolução da Soundex. A Metaphone foi melhorada em duas variações: a Double Metaphone e a Metaphone 3, que incrementou as codificações fonéticas.

Estes exemplos dão uma amostra das funções de distância Soundex e Metaphone. A Soundex converte "assistance" e "assistants" no mesmo código com a primeira letra seguida pelos mesmos três sons de consoantes.

As codificações Soundex sempre têm o tamanho igual a quatro. A Metaphone converte "assistance" e "assistants" em dois códigos ligeiramente distintos. A Metaphone insere outro som de consoante em "assistants", para o "t" que está no final da palavra.

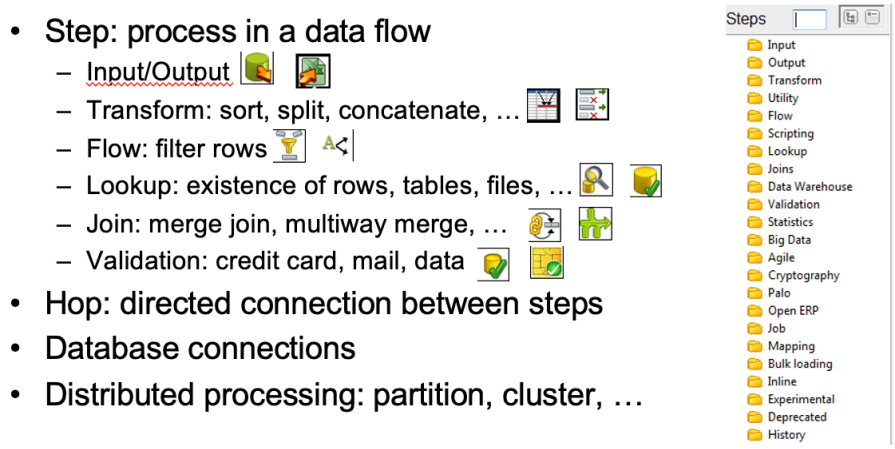
Nas funções de distância para correspondência inexata de valores de texto em quase-identificadores. Funções de distância de edição e de distância fonética são largamente implementadas em ferramentas de mineração de dados e de integração de dados, bem como em SGBDs. Respondendo à questão inicial, a função de distância de edição é geralmente limitada à comparação de palavras. A distância de edição consome muitos recursos se usada em grandes quantidades de texto, por conta da necessidade de minimização. Os valores de textos devem ser alinhados para determinar um número mínimo de operações. Embora a distância de edição use um algoritmo eficiente, o algoritmo ainda consome muitos recursos para comparar grandes quantidades de texto.

4.7. Pentaho Data Integration - PDI

Para estender sua experiência de aprendizagem, você deve instalar o Pentaho e usá-lo para fazer o exercício de prática e atribuição graduada. Uma lição de demonstração de software e documento tutorial detalhado estão disponíveis para aumentar a visão geral desta lição.

Pentaho fornece uma plataforma unificada para integração de dados, análise de negócios, e big data. Oo Pentaho usa o modelo de núcleo aberto, com uma edição comunitária de código aberto e extensões proprietárias e adições comerciais. Oferece produtos comerciais para integração de dados, análise de negócios e análise de big data.

O Pentaho Data Integration também é conhecido como Kettle, disponível no site da Sourceforge, em vez de uma edição comercial, disponível no site da Pentaho. O conceito básico de Pentaho abordado aqui é a transformação. Uma transformação Pentaho suporta fluxo de dados entre etapas e salta para conectar etapas.

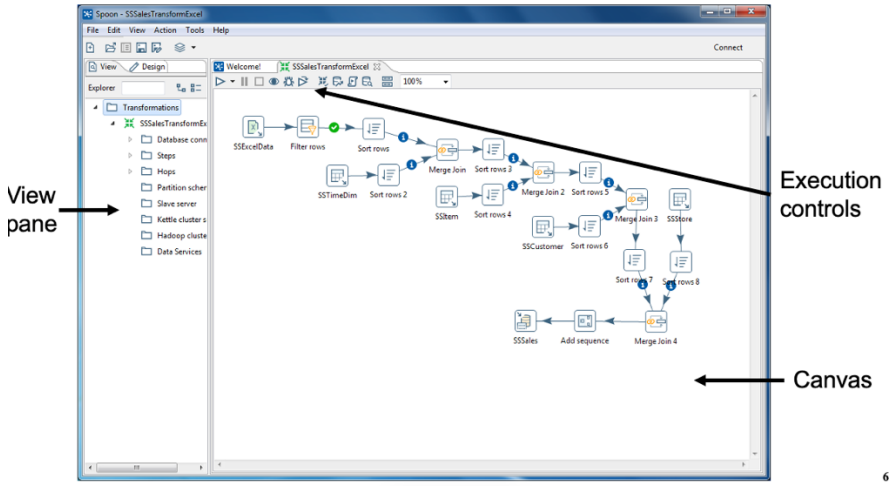


Um trabalho é um fluxo de dados de nível superior entre transformações e entidades externas. Kettle contém três componentes, Spoon fornece design gráfico de transformações e trabalhos, Pan executa transformações, enquanto Kitchen executa trabalhos. Uma transformação envolve etapas, saltos, conexões de banco de dados e recursos de processamento distribuídos.

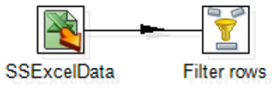
Pentaho fornece uma biblioteca de tipos de etapas, como mostrado na lista de pastas de etapas. As etapas de entrada e saída envolvem operações de arquivo, como leitura de texto e arquivos Excel. As etapas de transformação processam uma fonte de dados, como classificação, divisão, concatenação e seleção de valores.

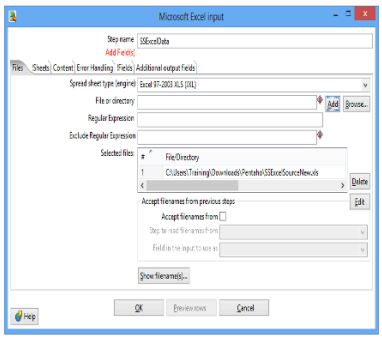
As etapas de fluxo reduzem sua fonte de dados aumentada, como filtrar linhas. As etapas de pesquisa testam a existência de linhas, tabelas, arquivos e outros objetos. As etapas de junção combinam fontes, como uma junção de mesclagem e mesclagem multiway.

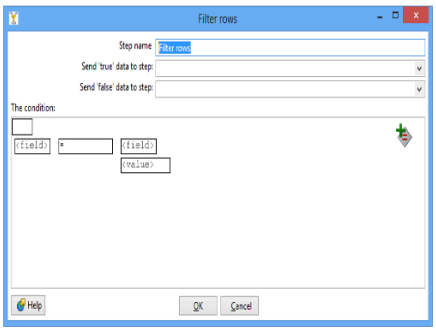
Spoon IDE



As etapas de validação executam verificações de qualidade de dados padrão, como validação de cartão de crédito e por e-mail. Os saltos fornecem conexões direcionadas entre as etapas. As etapas podem ter várias conexões de entrada e saída especificadas em saltos.

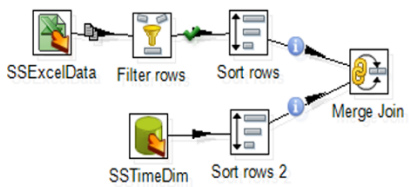


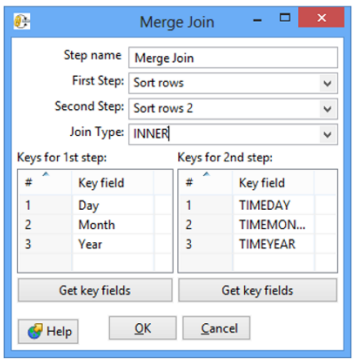




Pentaho também suporta especificação de conexões de banco de dados em recursos de processamento distribuídos, como partições e clusters. O ambiente de desenvolvimento integrado Spoon suporta visualizar componentes e transformações, projetar transformações e executar transformações. A guia Exibir mostra etapas, saltos e outros componentes, e a transformação é exibida na tela.

Merge Join Step





A guia Design contém pastas de tipos de etapa. Um analista arrasta uma etapa de uma pasta aberta na guia Design e a coloca na tela de desenho. Os controles de execução aparecem em uma barra de ferramentas acima da tela de desenho. Esses instantâneos retratam uma transformação simples para filtrar um arquivo do Microsoft Excel. A exibição gráfica na transformação contém duas etapas, uma etapa de entrada para o arquivo do Microsoft Excel e uma etapa de linha de filtro.

O salto indica o fluxo de dados da etapa de arquivo do Excel para a etapa de linha do filtro. Um analista usa uma janela de especificação para fornecer valores de propriedade para a etapa. Esta janela de especificação para o arquivo do Excel indica o local do arquivo, a planilha, os campos na planilha e outros detalhes.

A janela de especificação para a etapa da linha de filtro indica as condições para a parte inferior e as próximas etapas são executadas para passar e não especificar condições. Essa transformação estende uma transformação anterior com mais etapas e saltos. Essa transformação mescla a entrada de arquivo do Excel com uma tabela TimeTM.

As etapas da linha de classificação são necessárias porque uma etapa de junção de mesclagem requer fontes de dados classificadas nos mesmos critérios. A janela de especificação para a etapa de junção de mesclagem indica duas etapas de entrada, tipo de junção e campos de chave para mesclagem.

A etapa de mesclagem usa três campos de data, dia, mês e ano, do arquivo do Excel, e três colunas de data, dia de hora, mês e ano horário, da tabela TimeTM. Este exemplo de etapa de mesclagem indica a natureza tediosa de algumas transformações na arquitetura ETL.

Os compiladores de banco de dados manipulam detalhes sobre algoritmos de junção e ordem de junção para SQL SELECT instruções. Na arquitetura ETL do Pentaho, as transformações indicam alguns detalhes manipulados pelos compiladores de banco de dados na abordagem ELT.

Material Complementar

Article: The Kimball Group. (2016). Dimensional Modeling Techniques. *(20 min)* << *https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/* >>

Exercício

1. _____ é uma etapa crítica para data warehousing e produz projetos em suporte às características de DW.
a) Análise de dados
b) ETL
c) Modelagem de dados

2. As considerações da arquitetura de data warehouse devem ser incluídas na fase de projeto de modelagem de dados.
a) Falso
b) Verdadeiro

3. _____ são usados para representar estruturas de dados complexas, geralmente em formato de cubo.
a) Diagramas de Relacionamento de Entidade (ERDs)
b) Modelos de dados multidimensionais
c) Modelos de dados unidimensionais

4. Qual das alternativas a seguir não é uma etapa na construção de um modelo de dados multidimensional?
a) Coletando os requisitos do usuário
b) Identificando dimensões para organizar dados em torno de objetos e funções
c) Todas essas opções estão corretas

5. Um _____ é um modelo que descreve os dados em uma forma semelhante a uma estrela.
a) Modelo de dados multidimensional
b) Star Schema
c) Snowflake Schema

6. O esquema floco de neve fornece um design normalizado.
a) Verdadeiro
b) Falso

7. As tabelas _____ contêm chaves primárias e estrangeiras para atributos associados de um modelo de dados.

a) Base de dados
b) Dimensão
c) Fato

8. O esquema _____ requer menos armazenamento e tem menos _____ no processo de normalização.

a) Snowflake; Redundância
b) Estrela; Redundância
c) Snowflake; Confiança

9. _____ é uma abordagem de banco de dados alternativa que utiliza bancos de dados não relacionais e não estruturados.

a) PSQL
b) NoSQL
c) MySQL

10. _____ fornece uma abordagem descentralizada para armazenamento e análise de dados.

a) Data Warehouse
b) Data Lake
c) Big Dat

5. A Natureza dos Dados

5.1. Análise de dados

A estatística desempenha um papel significativo nas ciências físicas e sociais. E é sem dúvida o ponto mais saliente da interseção entre diversas disciplinas. É a linguagem comum da ciência. Cientistas usam estatísticas para converter dados em informações úteis. Estatística existe para um processo, em que estamos coletando dados, resumindo dados e interpretando dados. O processo de estatística começa quando identificamos qual grupo queremos estudar ou aprender algo. Chamamos este grupo de população.

A palavra população não é apenas usada para se referir a pessoas. É usado em um sentido estatístico mais amplo. Onde a população se refere a um grupo inteiro no qual você deseja se concentrar. Pode ser um grupo inteiro de pessoas ou animais ou insetos, ou objetos inanimados como prédios de apartamentos ou crateras em Marte. Por exemplo, podemos estar interessados nas opiniões da população adulta dos EUA sobre a pena de morte. Como a população de ratos reage a um determinado produto químico. O preço médio da população de todos os apartamentos de um quarto em uma determinada cidade. População, então, é todo o grupo que é o alvo de nosso interesse. Na maioria dos casos, a população é tão grande, que por mais que quisermos, não há absolutamente nenhuma maneira de podermos estudar tudo isso.

Uma abordagem mais prática seria examinar e coletar dados apenas de um subgrupo da população, que chamamos de amostra. Chamamos este primeiro passo que envolve a escolha de uma amostra e coleta de dados dele, produzindo dados. Uma vez que, por razões práticas, precisamos comprometer e examinar apenas um subgrupo de a população em vez de toda a população, devemos fazer um esforço para escolher uma amostra de tal forma que ela representará a população também.

James Thompson tem estudado o sucesso da polinização de uma flor obscura que floresce em altas altitudes, o lírio glaciado. Estamos olhando para uma espécie de alta altitude. E como as pessoas interessadas em mudanças globais e climas perceberam que é aqui que podemos primeiro ver coisas como espécies fazendo mal porque as situações mudaram, e, de fato, há muito foco na pesquisa de alta elevação em o contexto geral da mudança climática. Existem milhões destas flores em toda a Rockys e o professor Thompson não pode estudá-las todas, ele tem que escolher uma amostra. Se eu estou dizendo alguma coisa precisa esses dados têm que realmente refletir o que está acontecendo aqui. Então, por exemplo, quando eu olho para uma amostra de flores eu tenho que estar pensando o tempo todo sobre se eu estou selecionando um conjunto de flores que é adequadamente representativo de todo o que eu quero falar.

Isto é verdade se estamos estudando flores, crateras em Marte, ou as opiniões de adultos norte-americanos sobre a pena de morte. Nossa amostra não representaria adultos dos EUA, se perguntássemos apenas aos republicanos ou apenas perguntassem aos democratas. Tal amostra não representaria a população. Os conjuntos de dados podem ser muito diferentes dependendo do que está sendo estudado. Estes dados podem assumir a forma de respostas a perguntas de pesquisa, tabelas de números, como detalhes da cratera, ou, no caso de lírios glaciares, observações coletadas ao longo de muitos anos. Essencialmente, eu fiz uma pergunta muito simples. O que faz com que uma flor seja um sucesso? Será que ele é polinizado? Será que ela define uma fruta? Faz sementes? Quantas sementes faz?

Para dar sentido a esses dados, eles precisam ser resumidos de forma significativa. Isso é chamado de análise exploratória de dados. Análise exploratória de dados muitas vezes revela novas maneiras de pensar sobre os dados. Como acontece frequentemente na ciência, quanto mais cuidadosamente eu olhava, mais coisas eu via para me interessar. A análise exploratória de dados ajuda os cientistas a refinar suas perguntas. E às vezes até revelam perguntas inteiramente novas.

Se os climas estão mudando, as relações entre plantas e polinizadores e outras relações mutuamente benéficas, essas relações podem ser interrompidas. Cientistas que estudam a mudança climática têm muitas

vezes se perguntado que efeito um clima de aquecimento terá sobre a relação entre plantas e animais. É possível que pequenas mudanças no clima possam ter um grande impacto nessas relações? Análise exploratória de dados sugere que essa é uma pergunta que o Professor Thompson pode ser capaz de responder usando 30 anos de dados.

Isso leva até a etapa final, a inferência. O que podemos inferir sobre a população como um todo a partir dos dados em nossa amostra? Lembre-se, após análise exploratória de dados, somos capazes de fazer perguntas específicas sobre nossos dados. Inferência é aonde chegamos círculo completo com a esperança de revelar novos conhecimentos sobre a população.

Então, o que o Professor Thompson pode inferir sobre Lírios Glaciares? O que seus dados revelaram é que os lírios glaciares e as abelhas que os polinizam estão se separando no tempo. À medida que o clima aquece, os lírios florescem mais cedo antes das abelhas chegarem:

> "Meu artigo sobre lírios glaciares, tanto quanto posso dizer, é a primeira demonstração dele ou a primeira demonstração mesmo plausível dele. Não é uma coisa fácil de mostrar. É uma coisa fácil de dizer, ei, isso pode acontecer. Meus conjuntos de dados de longo prazo me permitiram fazer é dizer, sim, e parece que aconteceu."

James Thompson foi capaz de explorar seus dados para mostrar como mudanças climáticas faz com que plantas e animais se desconectem no tempo. Você também estará olhando para grandes conjuntos de dados e fazendo novas perguntas de interesse para você. Você não criará novos dados, mas criará novos conhecimentos através da análise de dados exploratória e análise de dados inferenciais. A educação estatística é mais frequentemente conduzida dentro de um contexto específico de disciplina ou como treinamento matemático genérico.

5.2. Dados e Tipos de Dados

O que realmente queremos dizer com dados? Simplificando, dados são pedaços de informação sobre indivíduos organizados em variáveis. Por indivíduo, queremos dizer uma unidade de observação. Uma observação ou unidade de observação refere-se a uma determinada pessoa ou um objeto específico, qualquer unidade específica de observação dentro de sua amostra de estudo. Os dados fornecem a base para inteligência de negócios, análise de negócios e ciência de dados. Como tal, é importante entender os vários tipos de dados que podem ser coletados, explorados, analisados e visualizados.

Por uma variável precisamos de uma característica particular da unidade de observação. No nível da pessoa, podemos coletar dados sobre Altura, Peso, Sexo, Corrida etc. Se estamos coletando dados em uma amostra de carros, podemos medir variáveis como Cor, Tamanho do Pneu, Quilometragem, Modelo e Número de assentos etc. Se nossa amostra incluir cidades, podemos medir variáveis como Tamanho da população, Receita Fiscal, Consumo de Energia, Número de Hospitais e assim por diante.

Um conjunto de dados é composto de observações e variáveis individuais. Os conjuntos de dados são normalmente exibidos em tabelas nas quais as linhas representam indivíduos, ou unidades de observação, e as colunas representam variáveis. Aqui está um conjunto de dados que mostra registros médicos de uma pesquisa. Neste exemplo, as unidades de observação são pacientes e as variáveis são Sexo, Idade, Altura, Peso, Fumar e Raça. Cada linha nos dá todas as informações sobre uma observação específica. Neste caso, um paciente. E cada coluna nos dá informações sobre uma característica particular de todos os pacientes.

Dados estruturados vs. não estruturados

Dados estruturados são dados bem definidos com padrões facilmente identificáveis. Alguns exemplos familiares são números de telefone e endereços de correspondência. Você pode discernir facilmente as partes

dessas informações (dados) porque entende seu padrão e formato distintos. A natureza organizada e estruturada desses dados também os torna facilmente pesquisáveis. Os dados estruturados normalmente estão presentes em um sistema de gerenciamento de banco de dados (relacional) (RDBMS ou DBMS), que discutiremos mais no próximo módulo.

Dados não estruturados são entendidos como "todo o resto", ou seja, dados em que os padrões não surgem facilmente e nem sempre podem se encaixar em um formato padrão. Exemplos típicos incluem arquivos de áudio, arquivos de vídeo e postagens de mídia social. Embora os dados não estruturados possam ser armazenados em vários formatos em um RDBMS, geralmente é mais comum encontrar dados não estruturados em um banco de dados não relacional ou armazenado em um sistema de arquivos.

A análise de dados estruturados é um processo bem definido e maduro, enquanto a análise de dados não estruturados é fortemente investida em pesquisa e desenvolvimento e na descoberta de novas tecnologias para analisar tipos de dados complexos com mais eficiência. Devido às complexidades inerentes à análise de dados não estruturados, essa análise requer muito mais tempo e poder de processamento. Consulte a tabela de comparação de dados estruturados versus não estruturados vinculada aqui para obter detalhes adicionais.

Dados Estruturados e Dados Não-estruturados		
	Estruturado	**Não-estruturado**
Características	• Modelo de dados pré-definidos • Tipicamente textual • Facilmente pesquisável • Facilmente identificável por padrões	• Modelo de dados não estabelecido • Pode ser texto, imagem, som, vídeo, etc • Difícil de pesquisar • Difícil de identificar padrão
Reside em	• Bancos de dados relacionais • Data Warehouses	• Aplicações • Bancos de dados NoSQL • Data Warehouse • Data Lakes
Exemplos	• Número de telefone • Endereço de e-mail • Número do CPF • Informação de transação	• Imagens • Audio • Vídeo • Web e mídia social

Dados Quantitativos x Qualitativos

Agora, vamos considerar algumas das diferenças entre dados quantitativos e qualitativos.

Variáveis também podem ser classificadas em um dos dois tipos, Quantitativo ou Categórico (ou qualitativos). Variáveis quantitativas tomam valores numéricos e representam algum tipo de medição. Variáveis categóricas, por outro lado, tomam valores de categoria ou e colocam uma observação ou indivíduo em um dos vários grupos. Neste exemplo, existem várias variáveis de cada tipo. Idade, peso e altura são variáveis quantitativas. Raça, Sexo e Fumar são variáveis categóricas.

Quantitativo: Discreto e Contínuo

Os dados quantitativos são estruturados e estatísticos e, portanto, podem ser contados, medidos e expressos usando números e cálculos. Esse requisito permite a facilidade de computação, agregação e análise.

Dois tipos principais de dados quantitativos são dados discretos e contínuos.

- **Dados discretos** são dados que não podem ser divididos em partes menores. Portanto, existe um conjunto finito de valores que podem ser aplicados. Dados discretos normalmente incluem números inteiros ou inteiros.
- Os **dados contínuos** podem ser divididos em partes menores e têm o potencial de flutuar continuamente.

Qualitativo: Nominal & Ordinal

Os dados qualitativos (ou categóricos) são de natureza descritiva e conceitual. É não estatístico e normalmente não estruturado ou semiestruturado. Os dados qualitativos são frequentemente categorizados usando traços e características. Geralmente é aberto e pode ajudar a responder à pergunta "Por quê?" No entanto, para fins de análise, os valores qualitativos geralmente precisam ser convertidos ou mapeados em dados numéricos.

Dois tipos principais de dados qualitativos são dados nominais e ordinais.

- Os **dados nominais** consistem em valores que não possuem ordem natural. Por exemplo, o gênero de uma pessoa não pode ser classificado como superior ou inferior a qualquer outro gênero.
- Os **dados ordinais** têm uma ordem natural e podem ser categorizados por agrupamentos de ordem. Os tamanhos das camisas são um ótimo exemplo de ordem em que grande > médio > pequeno.

Observe que os valores da variável categórica FUMANTE podem ser codificados como zero ou um. É bastante comum codificar os valores de uma variável categórica como números. Mas você deve sempre lembrar que estes são apenas códigos. Muitas vezes referido como *Códigos Dummy* (códigos fictícios) porque eles não têm significado aritmético. Ou seja, não faz sentido adicioná-los, subtraí-los, multiplicá-los ou dividi-los. Ou até mesmo comparar a magnitude desses valores.

IDs

Finalmente, um identificador exclusivo é uma variável que se destina a distinguir cada uma das unidades de observação do seu conjunto de dados. Exemplos podem incluir números de série para dados sobre um determinado produto, números de segurança social para dados sobre uma pessoa individual. Ou talvez números aleatórios gerados para qualquer tipo de observação. Para nos ajudar a organizar nossos dados, cada conjunto de dados deve ter uma variável que identifique exclusivamente as observações. Esta variável é particularmente útil se você precisar mesclar informações em diferentes conjuntos de dados.

5.3. Datasets e Codebooks

Alguns dos conjuntos de dados disponíveis para o curso incluem o Estudo Longitudinal Nacional de Saúde de Adolescentes, comumente conhecido como Add Health. Esta é uma pesquisa nacional representativa baseada na escola. A onda um da pesquisa incluiu adolescência nos graus 7 a 12 em os Estados Unidos. O Add Health inclui dados de pesquisa sobre bem-estar social, econômico, psicológico e de adolescentes. Em seguida é o estudo das crateras de Marte. Como você deve saber, o planeta Marte tem terreno fortemente craterizado. Estas crateras foram criadas há cerca de 4 bilhões de anos durante um período de bombardeamento pesado de asteroides, protoplanetas e cometas. Disponibilizado por pesquisadores da Universidade do Colorado Boulder, este conjunto de dados inclui características de mais de 350.000 dessas crateras de Marte. Também está disponível uma parte do Wave 1, Estudo Epidemiológico Nacional de Álcool e Condições Relacionadas, comumente conhecido como NESARC. Esta é uma amostra representativa da população adulta dos EUA com idade igual ou superior a 18 anos. E inclui dados sobre saúde mental e distúrbios do uso de substâncias que são experimentados por adultos. Outro conjunto de dados é o conjunto de dados GapMinder, que é disponibilizado por gapminder.org. Inclui numerosas medidas de 195 países. Os dados foram coletados de várias fontes, incluindo a Organização Mundial de Saúde, a Agência Internacional para Research on Cancer, as Nações Unidas e o Banco Mundial.

Para ajudá-lo a aprender mais sobre esses conjuntos de dados e em qual deles você está mais interessado, você estará revisando os códigos disponíveis desses conjuntos de dados. Às vezes chamados de dicionários de dados, os codebooks geralmente oferecem informações completas sobre o conjunto de dados. Isso é tópicos

gerais abordados, perguntas e/ou medidas usadas para registrar cada uma das variáveis. E em alguns casos, a frequência de respostas ou valores de cada uma das variáveis.

Rever um livro de códigos é sempre o primeiro passo na pesquisa com base em dados existentes. Primeiro de tudo, os livros de código podem ser usados para gerar perguntas de pesquisa. Em segundo lugar, os dados são muitas vezes inúteis e completamente impossíveis de interpretá-los sem eles.

O livro de códigos descreve como os dados são organizados no arquivo do computador. O que significam os vários números e letras, e quaisquer instruções especiais sobre como usar os dados corretamente. Como qualquer outro livro, alguns codebooks são melhores do que outros. No livro de códigos Add Health, cada variável tem uma descrição do que é medido. Neste caso, é a questão de qual nota você está.

Um livro de código também incluirá as várias opções de medição ou resposta. Para esta variável, possíveis opções de resposta incluem 7ª a 12ª série, recusa em responder à pergunta, um salto legítimo para aqueles que não estão na escola, não sabem e a escola não tem os níveis da série, ou a pergunta não é aplicável. Além de incluir uma listagem ou descrição das opções de resposta para a variável, o livro de códigos também incluirá valores correspondentes que podem ser encontrados no conjunto de dados.

Como vimos anteriormente com o exemplo do conjunto de dados de registros médicos, conjuntos de dados normalmente incluem números em vez de palavras. Assim, para variáveis categóricas, como nível de grau, cada uma das opções de resposta tem um valor numérico correspondente. E é esse valor numérico que pode ser encontrado no conjunto de dados. Você pode ver que os alunos da 7ª a 12ª série são logicamente codificados como os números 7 a 12.

- 96 indica que o adolescente **se recusou a responder**.
- 97 indica um salto legítimo para os adolescentes que **não estão atualmente na escola**.
- 98 indica que **não sei**.
- 99 é gravado em um caso em que **a escola não tem níveis de série**.

Estes valores numéricos são conhecidos como códigos fictícios, como estão incluídos no conjunto de dados, mas não têm significado numérico direto. No livro de código das crateras de Marte, encontramos uma descrição das variáveis para nome, latitude, longitude e diâmetro da cratera. Também a profundidade da borda da cratera, bem como o nome da variável no conjunto de dados. Como a maioria dessas variáveis são quantitativas, em vez de listar uma opção de resposta, o livro de códigos inclui uma descrição de como a variável é medida. Por exemplo, a latitude é medida em graus decimais Norte, longitude é medida em graus decimais Leste e o diâmetro e a profundidade são medidos em quilômetros.

Do dataset Gapminder, vemos uma aparência ligeiramente diferente do livro de código, mas características muito semelhantes. Você pode ver que a coluna do meio descreve cada uma das variáveis. A coluna à esquerda indica o nome da variável usada no conjunto de dados. E, finalmente, a coluna da direita lista a fonte de dados. Novamente, estas são variáveis quantitativas. O livro de códigos inclui informações sobre como cada uma dessas variáveis foi medida. Você pode ver que a renda por pessoa é medida em dólares americanos. O consumo de álcool é medido em litros de álcool puro. Forças de trabalho é medida como a porcentagem da força de trabalho total, e taxa de câncer de mama é medida como novos casos por 100.000 mulheres.

5.4. Desenvolvendo uma questão de pesquisa

Uma vez que você tenha uma compreensão geral acerca dos conjuntos de dados, tipos de variáveis e livros de código, o próximo passo é selecionar um conjunto de dados. Selecione um conjunto de dados que inclua variáveis em uma área que lhe interessa.

Depois de selecionar os conjuntos de dados, identifique um tópico específico de interesse e imprima as páginas do livro de códigos que incluem a variável ou as variáveis que medem o tópico selecionado. Note que

muitos livros de código são muito grandes para imprimir, por isso é muito importante criar o seu próprio livro de código pessoal com apenas as páginas que incluem as variáveis que você gostaria de examinar.

Nosso exemplo vem do conjunto de dados NESARC, e nosso tópico escolhido é a dependência da nicotina. Existem várias variáveis relacionadas à Dependência de Nicotina, e podemos ver 2 aqui: dependência de nicotina ao longo da vida e dependência de nicotina nos últimos 12 meses. Um valor de zero para estas variáveis indica que não há Dependência de Nicotina, e um valor de 1 indica a presença de Dependência de Nicotina. O nome dessas variáveis são TAB12MDX e TABLIFEDX. Usaremos esses nomes de variáveis quando começarmos a trabalhar com os dados.

Não estamos sugerindo que este tópico seja mais ou menos interessante, ou mais ou menos importante do que qualquer outro. O que é importante é que você escolha um tópico que é de seu interesse. Escolhemos analisar a dependência da nicotina.

Depois de ter um tópico e ter impresso as páginas do livro de código que medem esse tópico, é hora de criar uma pergunta de pesquisa. Uma das perguntas de pesquisa mais simples que podem ser feitas é se dois tópicos estão associados um ao outro. Por exemplo, a procura de tratamento médico está associada à renda? A profundidade da cratera está associada ao diâmetro da cratera? A fluoração da água está associada ao número de cavidades durante visitas ao dentista? Esses conjuntos de dados são vastos, portanto, há muitas associações potenciais para explorar. Vamos olhar para o nosso exemplo escolhido: dependência de nicotina.

Primeiro eu preciso determinar o que é sobre a dependência de nicotina que me interessa. Parece-me que amigos e conhecidos que eu conheci ao longo dos anos, que ficou viciado em cigarros o fizeram em períodos muito diferentes de tempo. Alguns pareciam ser dependentes de fumar fortemente logo após sua primeira experiência com um cigarro, e outros depois de muitos anos de comportamento geralmente irregular de fumar.

Decidimos que estamos mais interessados em explorar a associação entre o comportamento do tabagismo e a dependência da nicotina. Acreditamos que eles estão positivamente associados. Ou seja, quanto mais um indivíduo fuma, mais provável é que seja dependente da nicotina. Também estamos nos perguntando o quanto uma pessoa precisa fumar para ser dependente da nicotina.

Continuamos a ler o livro de códigos NESARC e descobrimos que o comportamento de fumar também foi medido nesta amostra. Então, em seguida, eu dou um passo semelhante a um que eu acabei de tomar ao escolher a dependência de nicotina. Ou seja, identifique as variáveis que medem o segundo tópico, comportamento de tabagismo, no meu conjunto de dados. As variáveis que escolho incluem status de tabagismo, frequência usual, e quantidade usual.

Durante sua segunda revisão do livro de códigos para o conjunto de dados que você selecionou, você também deve identificar um segundo tópico que você gostaria de explorar em termos de associação com seu tópico original. E, novamente, imprima as páginas do livro de códigos que incluem a variável, ou variáveis, que medem o segundo tópico selecionado.

6. Estatística

Em sua essência, a estatística é uma análise técnica de dados baseada em matemática usando vários testes e análises. Embora não possamos aprofundar muito neste curso, é importante considerarmos as metodologias estatísticas que são usadas na análise de dados. Para os propósitos deste curso, exploraremos brevemente dois métodos principais: estatística descritiva e estatística inferencial.

6.1. Estatística descritiva

A estatística descritiva permite a sumarização e a representação gráfica de um conjunto de dados. A natureza descritiva das informações resultantes permite que um analista descreva uma amostra da população do conjunto de dados. (Observe que isso não nos permite generalizar uma população inteira ou inferir atributos ou propriedades da população.)

Geralmente usamos estatística descritiva para explorar:

- **Tendência central** (média): média, mediana ou moda para explicar as médias de um ponto de dados
- **Dispersão**: intervalo e desvio padrão para descrever a distância da média ou distância entre os valores de dados mais altos e mais baixos
- **Skewness (distorção)**: descreve a natureza simétrica ou assimétrica do conjunto de dados
- **Correlação**: explora as relações entre as variáveis no conjunto de dados de amostra

6.1.1. Análise Exploratória de Dados

Os dados brutos consistem em longas listas de números e rótulos que não parecem ser muito informativos. Dados brutos carece de contexto. Análise exploratória de dados é o que você usa para entender os dados. Você faz isso convertendo dados de sua forma bruta, em um formulário que faz sentido, que tem contexto, que conta a história que você quer contar.

Basicamente, a análise exploratória de dados consiste em organizar e resumindo dados brutos, procurando características e padrões importantes em os dados, procurando quaisquer desvios marcantes desses padrões, e interpretando suas descobertas no contexto do problema ou questão de pesquisa. Começaremos a análise exploratória de dados analisando uma variável de cada vez, também chamada de análise univariada.

Para converter dados brutos em informações úteis, precisamos resumir e, em seguida, examinar a distribuição de quaisquer variáveis de interesse. Por distribuição de uma variável, queremos dizer quais valores a variável toma, e com que frequência a variável leva esses valores.

Se estivéssemos estudando um pequeno número de observações, poderíamos fazer isso com um lápis e papel, uma calculadora, ou mesmo em nossas cabeças. Os conjuntos de dados com os quais você está trabalhando, muitas vezes têm milhares de observações. Trabalhar com amostras tão grandes só é possível se usarmos software estatístico. Esses programas de software exigem o uso de sintaxe ou código formal para recuperar, analisar e manipular dados. Aprender a escrever código, aprender o uso adequado da sintaxe pode realmente expandir sua capacidade de se envolver em aplicativos estatísticos. Essa habilidade também expandirá muito sua capacidade de se engajar em níveis mais profundos de raciocínio quantitativo sobre dados.

Para este curso, você estará usando Python. Python é uma linguagem de uso geral amplamente utilizada que é projetado para ser mais legível. Ou seja, o código é mais fácil de ler e escrever do que em outras linguagens de uso geral, como C++ ou Java. Embora o Python não tenha sido desenvolvido especificamente para análise de dados pandas e outras bibliotecas fornecem ferramentas de análise de dados para uso com a linguagem Python. Olhando para todas as janelas, opções, menus e recursos embora, pode ser bastante assustador. Portanto, é importante para você perceber, este curso irá apresentá-lo ao básico. Você aprenderá o que precisa saber para começar a perguntar e respondendo perguntas interessantes sobre dados. >> No início, você pode sentir que está

aprendendo outro idioma. Basicamente, é. À medida que você trabalha em seu projeto, você deve começar a se sentir mais confortável implementando as várias decisões que você vai tomar sobre os dados.

6.1.2. Examinando a distribuição de frequência

A análise exploratória de dados começa olhando em uma variável de cada vez. Isso é chamado de univariado ou análise descritiva. Para converter dados brutos em informações úteis, precisamos resumir e examinar a distribuição de qualquer variável de interesse. As variáveis de interesse são as variáveis de interesse para você, pesquisador. Ao responder suas perguntas de pesquisa, abordando seu problema de pesquisa e contando a história que você deseja contar com sua pesquisa. Por distribuição de uma variável, queremos dizer quais valores a variável leva e com que frequência a variável leva esses valores.

Aqui está um exemplo. Em uma amostra aleatória de 1.200 estudantes universitários dos EUA convidados a responder as seguintes perguntas como parte de uma pesquisa maior: Qual é a sua percepção do seu próprio corpo? Você sente isso você está acima do peso? Razoável? Ou abaixo do peso? Esta tabela mostra parte dos dados, cinco das 1.200 observações.

Informações que seriam interessantes obter a partir desses dados inclui que porcentagem dos alunos da amostra se enquadram em cada categoria ou como os alunos são divididos ao longo dos três tipos categorias de imagem? Eles estão igualmente divididos? Se não, faça as percentagens seguir algum tipo de padrão? Não há como responder a essas perguntas por olhando para os dados brutos, que estão na forma de uma longa lista de 1.200 respostas.

Isso não é muito útil. No entanto, todas essas perguntas serão facilmente respondidas quando resumirmos e observarmos a distribuição de frequência da imagem corporal variável. Isto é, uma vez que resumimos com que frequência cada uma das categorias ocorre. Para resumir a distribuição de uma variável categórica, primeiro criamos uma tabela dos diferentes valores ou categorias que a variável assume.

Quantas vezes cada ocorre a variável, que é a contagem, e, mais importante, com que frequência cada variável ocorre, o que é expresso convertendo as contagens em percentagens. Agora que resumimos a distribuição da variável de imagem corporal, vamos voltar e interpretar os resultados no contexto das perguntas que postamos.

Qual porcentagem de os alunos da amostra se enquadram em cada categoria? Como os alunos são divididos em três corpos categorias de imagens, e elas estão igualmente divididas? Você pode ver isso a maioria das amostras, ou seja, 71,3% sentida que seu peso estava quase certo e que uma pequena porcentagem sentiu-se abaixo do peso em 9,2%. A categoria sobrepeso foi de 19,6%.

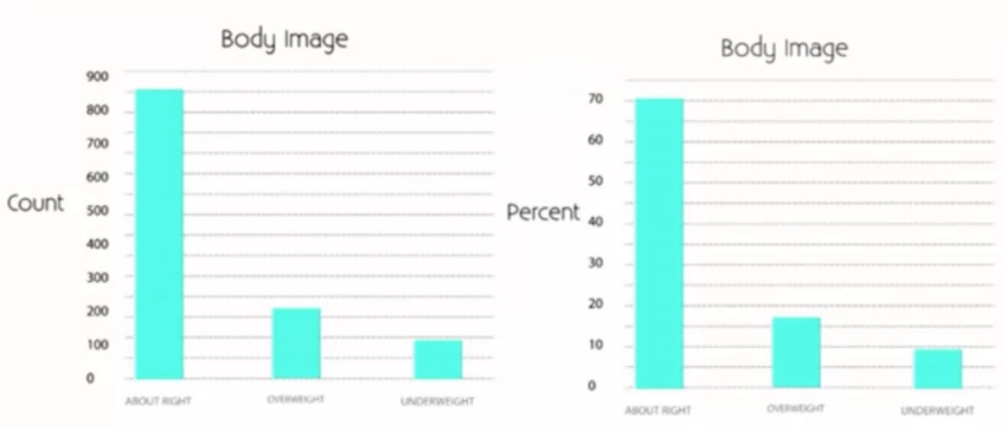
6.1.3. Plotando as distribuições

Ferramentas de visualização são importantes meios para ampliar a compreensão acerca do comportamento dos dados. Para começar a visualizar nossas variáveis com gráfico, iniciaremos com gráficos com uma variável de cada vez, usaremos isso como um trampolim para visualizar várias variáveis simultaneamente com gráficos internos.

Acompanhe o exemplo abaixo através do script disponibilizado "**plotando_distrib**".

Os gráficos de barras são mais comumente usados examinar a distribuição de variáveis individuais. Considere uma distribuição para a amostra aleatória de 1.200 estudantes universitários americanos que foram questionados sobre o que é a sua percepção do seu próprio corpo. Neste gráfico de barras, o eixo X ou horizontal inclui as três categorias de resposta. Abaixo do peso, acima do peso e quase certo. No primeiro gráfico de barras, a altura das barras é medida no eixo Y, ou vertical, como o número ou contagem de estudantes universitários dando cada resposta. O segundo gráfico de barras mostra os mesmos dados, mas como uma porcentagem da

amostra total. Um gráfico de barras nos ajuda a exibir a distribuição de uma variável categórica, por exemplo, porcentagem de observações em cada categoria.

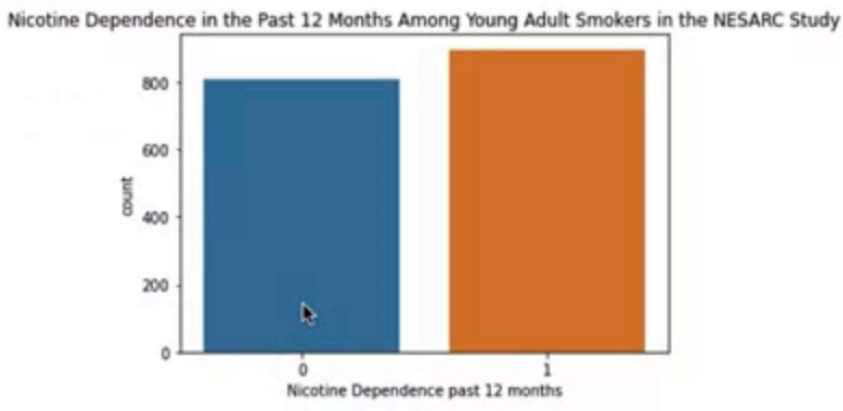


Para ilustrar, usaremos o dataset NESARC, buscando interpretar as relações entre a dependência de nicotina nos últimos meses (representado no dataset pela variável TAB12MDX) e a estimativa de cigarros fumados por mês (representado por NUMCIG_EST). Vamos executar distribuições de frequência para cada uma dessas variáveis, incluindo contagens e percentagens. Vou usar a função groupby para isso que também apresentamos quando introduzindo distribuições de frequência. Além das distribuições de frequência, também queremos examinar os gráficos de barras correspondentes para essas duas variáveis também. O gráfico de barras é uma das mais visualizações gráficas frequentemente usadas.

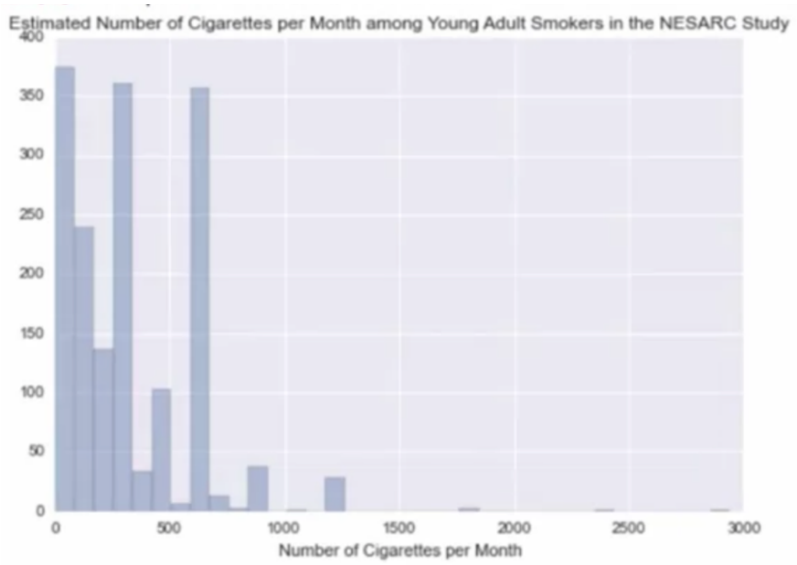
Ao visualizar dados em Python, precisaremos importar bibliotecas em nosso programa. Primeiro, vamos importar o seaborn pacote com a sintaxe **import seaborn**. Também precisamos importar a biblioteca **matplotlib.pyplot** porque o pacote seaborn é dependente neste pacote para criar gráficos. Porque o nome deste pacote é tão longo, daremos a ele o apelido **plt**, que pode ser usado no lugar de o nome completo do pacote quando escrevemos o código chamando isso pacote em nosso programa. Nós vamos mantê-lo simples. Usaremos o código Python para gerar gráficos que nos ajudam a aprender mais sobre nossos dados e a tomar decisões sobre próximos passos de nossa pesquisa.

Estamos focando na função de visualizações gráficas em vez de produzir imagens polidas e prontas para apresentações gráficas neste momento. Variáveis categóricas podem ser visualizadas um de cada vez com os gráficos univariados, ou seja, com gráficos de barras de variável única. Em primeiro lugar, a fim de categórico variáveis sejam ordenadas corretamente no eixo horizontal ou X de uma variável univariada gráfico, você deve converter suas variáveis categóricas, que geralmente são formatados como variáveis numéricas, em um formato que o Python reconhece como categórico. Aqui está o código. Aqui estou usando o **astype** função para converter TAB12MDX em uma variável categórica, mantendo o nome da variável original como está.

O código básico para um gráfico univariado de uma variável categórica é a seguinte. Com a função de gráfico de contagem, nomeamos a variável categórica para o eixo X e para encontrar o quadro de dados aqui, sub2. Com a função **xlabel**, podemos rotular o eixo X, e com a função **title**, fornecer o gráfico de barras com um título. Aqui está o código do gráfico de barras univariado inserido em nosso programa de exemplo, e salvamos e executamos o programa para gerar o gráfico de barras solicitado. Podemos visualizar o gráfico clicando em a guia de plotagens para abrir o painel de plotagens. Isto mostrará o número de jovens adultos fumantes com dependência de nicotina, 896, indicado por um código de resposta de 1. E aqueles sem dependência de nicotina, 810, indicado por um 0.



Agora vamos exibir graficamente a distribuição de frequência para uma de nossas variáveis de tabagismo gerenciadas por dados, ou seja, o número estimado de cigarros fumava por mês, **NUMCIGMO_EST**. Porque **NUMCIGMO_EST** é na verdade uma variável quantitativa, a sintaxe que usamos no Python programa é um pouco diferente. Para visualizar uma variável quantitativa, você usaria a seguinte sintaxe. Com a função de plotagem de distribuição, ou **distplot**, nomeamos a variável quantitativa para o eixo X e peça ao Python para descartar os dados ausentes. Isso é as NaNs. Também incluímos a opção **kde=False**. Novamente da biblioteca matplotlib.pyploy, que estamos chamando de plt, usamos o rótulo X para rotular o eixo X com direito a fornecer o gráfico com o título. Ao executar isso, você verá que o programa gera uma distribuição gráfica da variável quantitativa. Gera um histograma. Em um histograma, intervalos de valores são plotados no eixo X em vez de valores discretos ou separados. Das barras aqui, você pode ver que o que é exibido é o ponto médio dos intervalos.



Vamos olhar para um exemplo mais básico de como um histograma pode ser construído, e então usar isso como um trampolim para falar sobre estatísticas descritivas adicionais que podem ser geradas para variáveis quantitativas. Neste exemplo, temos as notas de exame de 15 alunos.

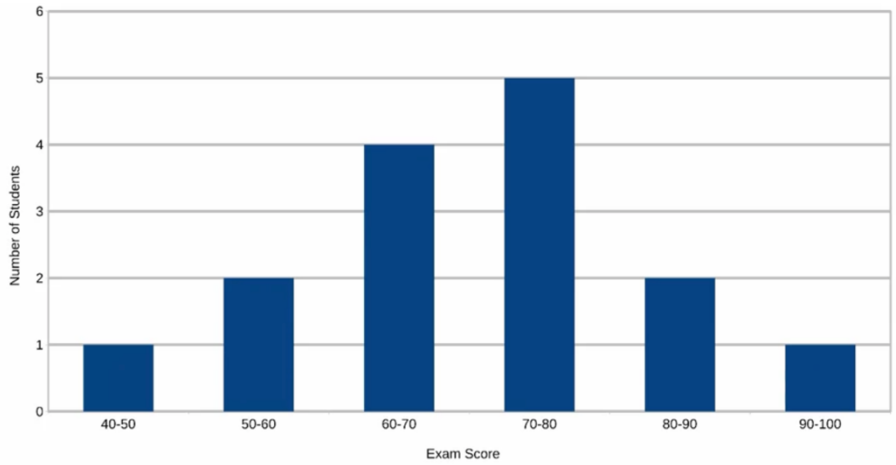
88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

Primeiro precisamos dividir o intervalo de valores em intervalos. Também chamado de compartimentos, grupos ou classes. Neste caso, uma vez que o nosso conjunto de dados consiste em pontuações de exames, fará sentido escolher intervalos que normalmente correspondam ao intervalo de notas de letra. Então dez pontos de largura, 40 a 50, 50 a 60, etc. Ao contar quantos das 15 observações caem em cada um dos intervalos, obtemos esta tabela.

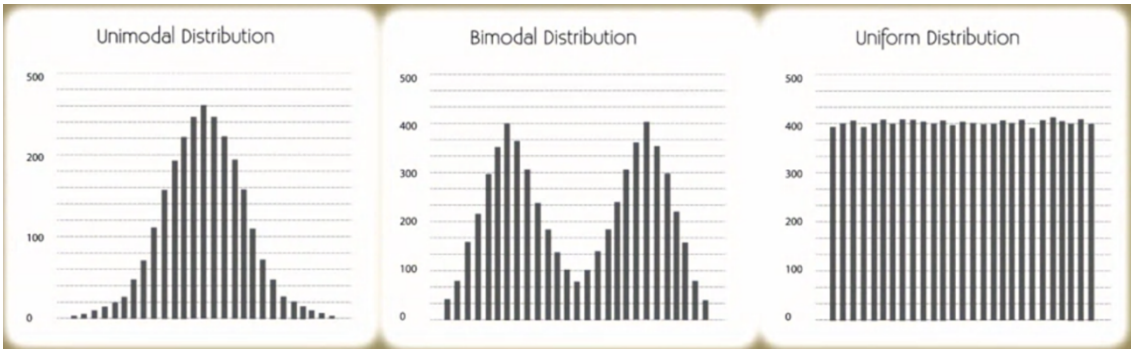
Para construir o histograma a partir desta tabela, os intervalos são plotados no eixo X e mostram o número de observações em cada intervalo, ou a porcentagem de observações em cada intervalo no acesso Y, que é representada pela altura da barra localizada acima do intervalo.

Pontuação	Ocorrências
40-50	1
50-60	2
60-70	4
70-80	5
80-90	2
90-100	1

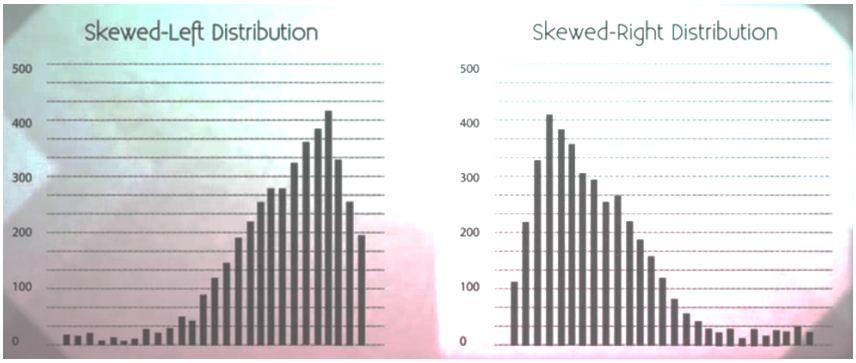
Uma vez que a distribuição tenha sido exibida graficamente como um histograma, podemos descrever o padrão geral da distribuição e mencionar quaisquer desvios marcantes desse padrão. Mais especificamente, devemos considerar os seguintes recursos. Teremos uma noção do padrão geral dos dados do centro dos histogramas, da dispersão e da forma, enquanto os outliers destacarão desvios desse padrão.



Ao descrever a forma de uma distribuição, devemos considerar simetria ou assimetria da distribuição e pico ou modalidade. Ou seja, o número de picos ou modos que a distribuição tem. Aqui, todas as três distribuições seriam referidas como simétricas. Mas eles são diferentes em sua modalidade ou pico. A primeira distribuição é unimodal. Ele tem um modo, aproximadamente em 10, em torno do qual as observações estão concentradas. A segunda distribuição é bimodal. Tem dois modos, aproximadamente em 10 e 20, em torno dos quais as observações estão concentradas. A terceira distribuição é tipo plana ou uniforme. A distribuição não tem modos, ou nenhum valor em torno do qual as observações estão concentradas. Em vez disso, as observações são distribuídas aproximadamente uniformemente entre os diferentes valores.

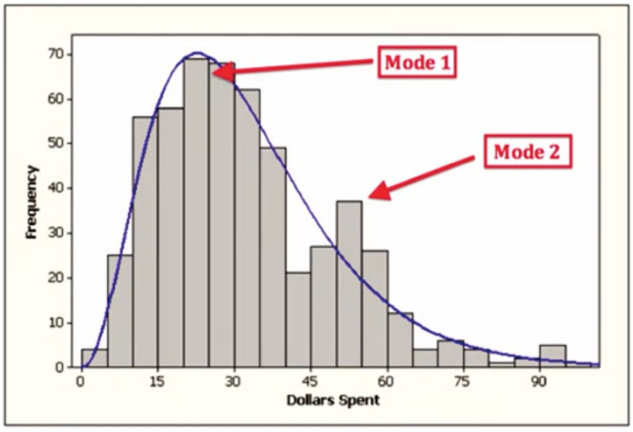


Uma distribuição é chamada skewed-right (distorcida à direita). Com a cauda direita, os valores maiores são muito mais longos do que a cauda esquerda, ou valores menores. Note que em uma distribuição assimétrica à direita, como você pode ver aqui à direita. A maior parte das observações é pequena a média, com algumas observações que são muito maiores do que o resto. Um exemplo de uma variável da vida real que tem uma distribuição assimétrica à direita é o salário. A maioria das pessoas ganha na faixa baixa a média de salários com algumas exceções, como CEOs, atletas profissionais, etc. Que são distribuídos ao longo de uma ampla gama, que é a longa cauda de valores mais elevados.



Uma distribuição é chamada skewed-left (distorcida à esquerda) se a cauda esquerda ou valores menores forem muito maiores do que a cauda direita ou valores maiores. Não que em uma distribuição assimétrica à esquerda, a maior parte das observações seja de média a grande, com algumas observações que são muito menores do que as restantes. Um exemplo de uma variável da vida real que tem uma distribuição distorcida à esquerda é a idade da morte por causas naturais. A maioria das mortes por causas naturais ocorre em idades mais velhas, com menos casos acontecendo em idades mais jovens.

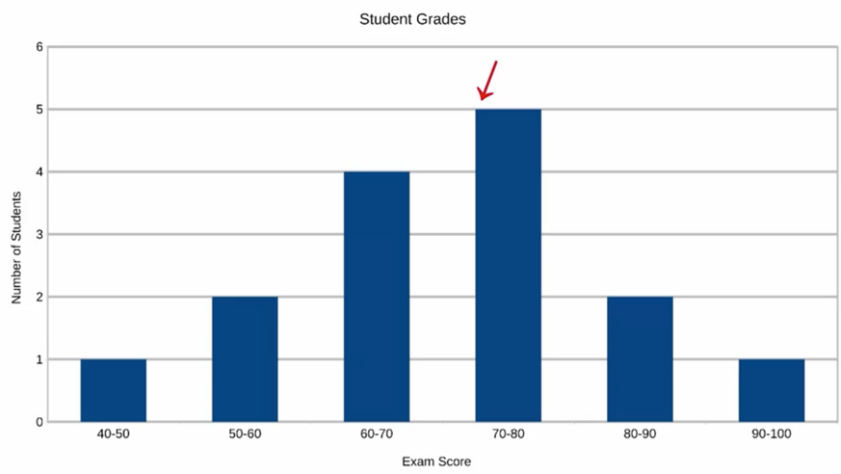
Distribuições distorcidas também podem ser bimodais. Aqui está um exemplo, um bairro de tamanho médio 24 horas loja de conveniência coletou dados de 537 clientes sobre a quantidade de dinheiro que gastaram em uma única visita à loja. Observe o histograma abaixo. Você pode ver que a quantidade de dinheiro gasto é concentrada em torno de US $20, e, em seguida, concentrado novamente em torno de US $50.



A moda de uma variável são os valores que ocorrem com mais frequência. E saber disso pode ajudá-lo a tomar melhores decisões. A moda, por exemplo, tem aplicativos na publicação de livros. Não surpreendentemente, é importante para a editora imprimir mais dos livros mais populares, porque imprimir livros diferentes em números iguais causaria uma escassez de alguns livros e um excesso de oferta de outros. Da mesma forma, o modo tem aplicações na fabricação. Por exemplo, também é importante fabricar mais dos sapatos e tamanhos de sapato mais populares.

A moda nem sempre está no centro. O centro da distribuição é o seu ponto médio, o valor que divide as distribuições de modo que aproximadamente metade das observações leve valores menores e aproximadamente

metade leva valores maiores. Como você pode ver no histograma, o centro da distribuição de graus é aproximadamente 70. Podemos obter apenas uma estimativa aproximada para o centro da distribuição. Sete alunos marcaram menos de 70, e oito alunos pontuaram acima de 70. Estimativas geralmente podem ser feitas a partir do exame de um histograma.



Então, e quanto a dispersão? A dispersão da distribuição, também chamada de variabilidade, pode ser descrita pelo intervalo aproximado coberto pelos dados. De olhar para o histograma, podemos aproximar a menor observação, ou mínimo, e a maior observação, ou máximo, e assim aproximar o intervalo. Em nosso exemplo de pontuação de exame, você pode ver que o mínimo aproximado é 45, que é o meio do menor intervalo de pontuações. O máximo aproximado é 95, o meio do maior intervalo de pontuações. Então, nosso alcance aproximado é de cerca de 50 pontos. 95 menos 45. O padrão geral da distribuição da variável quantitativa é descrito por sua forma, centro e dispersão. Ao inspecionar o histograma, podemos descrever a forma da distribuição, mas como vimos, só podemos obter uma estimativa aproximada do centro e propagação.

6.1.4. Medidas de Centralidade e Dispersão

Para descrever a distribuição de uma variável quantitativa, você também precisa de descrições numéricas precisas do centro e da dispersão. A moda é um tipo de média. Há três tipos de média e cada um nos diz algo diferente. Portanto, precisamos ter certeza de que entendemos o que cada média significa. Quando usamos o termo média, queremos dizer uma das três coisas geralmente, ou queremos dizer a média aritmética, a moda ou mediana.

É muito fácil entender a diferença entre estes, especialmente se você já jogou dardos antes. Depois de dois lotes de três dardos e no meu sexto lançamento marquei um 2, 3, 3, 12 e 13. Agora vamos ver se podemos descobrir a média aritmética, a mediana e a moda. Primeiro de tudo a média aritmética. Tomamos o total de todas as seis pontuações e dividimos pelo número de observações, e essa é a média. Se quisemos a pontuação modal simplesmente procuramos a pontuação mais comum, o número mais comum de observações. Se quisermos a pontuação mediana, escrevemos as pontuações em ordem crescente e, em seguida, procuramos o valor do meio.

Há um pequeno problema aqui que temos um número par de observações, então pegamos os dois valores médios, e calculamos a média desses dois. Então, para o meu dardo não muito bom jogando, as pontuações foram 2, 3, 3, 3, 12, 13. A média é 2+3+3+12+13 dividido por 6, 36/6 = 6. A moda é 3. A mediana desde que temos um número par de observações, é 3 + 3, o meio duas observações, dividido por 2, que é igual a 3.

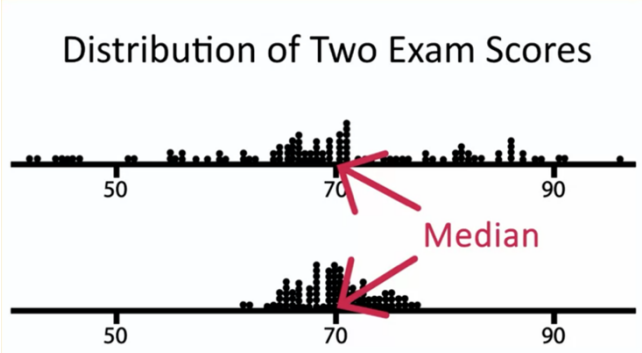
Aviso, se o jogador de dardo tivesse marcado, 19 em vez de 13. A média aumenta para 7, mas a moda e a pontuação mediana permanecem inalterados.

Então vamos rever brevemente as medidas numéricas centrais. Intuitivamente falando, a medida aritmética do centro está nos dizendo o que é um valor típico da distribuição de uma variável. As três principais medidas numéricas do centro da distribuição são a moda, a mediana e a média aritmética. Até agora, quando olhamos para a forma da distribuição, identificamos a moda como o valor em que a distribuição tem um pico. E vimos exemplos quando as distribuições têm uma moda, que é uma distribuição unimodal, ou duas modas, uma distribuição bimodal. Em outras palavras, até agora identificamos a moda visualmente a partir do histograma. Olhando para os nossos histogramas novamente, podemos facilmente ver a moda. É o valor que ocorre mais comum na distribuição.

A mediana, que é o ponto médio da distribuição, é o número tal que metade das observações cai acima e metade cai abaixo? Encontramos a mediana ordenando os dados do menor para o maior. Considere quando N, o número de observações é par ou ímpar. Se N for ímpar a mediana é a observação central na lista ordenada, quando o número de observações é mesmo a mediana é a média ou média do valor das duas observações centrais.

A média, é claro, pode ser calculada adicionando os valores para todas as observações e dividindo pelo número de observações para gerar uma média aritmética. Nosso objetivo aqui é descrever a distribuição. Como você descreveria essas duas distribuições de escores de exames? Ambas as distribuições estão centradas em 70. A média de ambas as distribuições é de aproximadamente 70. Mas as distribuições são realmente muito diferentes. A primeira distribuição tem variabilidade muito maior e pontuações em comparação com a segunda.

Para descrever uma distribuição, precisamos complementar a exibição gráfica, não só com a medida do centro, mas também com a medida da variabilidade ou dispersão da distribuição. Existem várias maneiras de descrever a dispersão. Uma medida comumente usada é o desvio padrão. A ideia por trás do desvio padrão é quantificar a propagação de a distribuição medindo o quão longe as observações estão de sua média.



O desvio padrão dá a média ou distância típica entre um ponto de dados e a média. Para entender melhor o desvio padrão, seria útil ver um exemplo de como ele é calculado. Na prática, é claro, o software estará fazendo esses cálculos para nós.

Empresas de serviços médicos de emergência gostariam de estimar quantas tripulações de ambulância devem manter em espera. Aqui está o número de chamadas de ambulância durante um período de oito horas.

7, 9, 5, 13, 3, 11, 15, 9

Média => $\bar{X} = (7 + 9 + 5 + 13 + 3 + 11 + 15 + 9) / 8 = 9$

Para encontrar o desvio padrão do número de chamadas por hora, primeiro encontraríamos a média dos nossos dados. Em seguida, precisaríamos encontrar os desvios da média. Essa é a diferença entre cada observação na média. Como nossa média é 9, subtrairíamos 9 de cada uma de nossas observações.

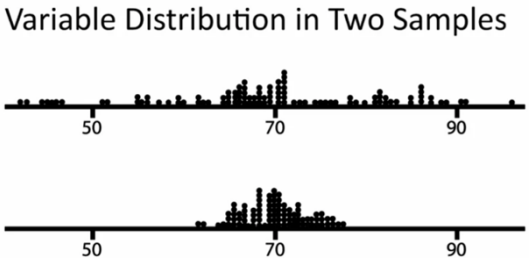
7	9	5	13	3	11	15	9
-	-	-	-	-	-	-	-
9	9	9	9	9	9	9	9
-2	0	-4	4	-6	2	6	0

Como um terceiro passo, nós elevaríamos ao quadrado cada um desses desvios. Em seguida, medimos os desvios quadrados adicionando-os e dividindo-os por N-1, que é um a menos do que o tamanho amostral, esta média dos desvios quadrados é chamada de variância. O desvio padrão de sua variável é a raiz quadrada dessa variância.

$$\frac{(4 + 0 + 16 + 16 + 36 + 4 + 36 + 0)}{(8 - 1)} = \frac{112}{7} = \sqrt{16} = 4$$

Então, por que tomamos raiz quadrada? Note que 16 é a média dos desvios quadrados e, portanto, tem diferentes unidades de medida. Neste caso, 16 é medido em número quadrado de chamadas de ambulância, que obviamente não pode ser interpretado. Nós, portanto, tomamos a raiz quadrada para compensar o fato de que nós temos quadrado todos os nossos desvios e também para voltar para a unidade de medida original. Lembre-se de que o número médio de chamadas de emergência em uma hora é 9. A interpretação do desvio padrão igual a 4 é que, em média, o número real de chamadas de emergência a cada hora está a 4 de distância de 9. Outra maneira de dizer isso é que há uma média de chamadas de ambulância/hora = 9 ± 4.

Desde que estamos trabalhando com um grande número de observações cálculos de mão de desvio padrão realmente não são viáveis. Python fará todos esses cálculos para você, mas é importante saber como calcular desvios padrão para que você possa entender sua variabilidade. Por exemplo, olhando para uma distribuição de variáveis em duas amostras diferentes, você deve ser capaz de dizer qual tem maior variabilidade, ou seja, um desvio padrão maior. Para calcular o desvio padrão e gerar outras estatísticas descritivas para uma variável quantitativa, muitas vezes usamos a função de descrição do Python.



Aqui está a sintaxe para descrever NUMCIGMO_EST como a variável quantitativa.

```
61 seaborn.distplot(sub2["NUMCIGMO_EST"].dropna(), kde=False);
62 plt.xlabel('Number of Cigarettes per Month')
63 plt.title('Estimated Number of Cigarettes per Month among Young Adult Smokers in the NESARC Study')
64
65 # standard deviation and other descriptive statistics for quantitative variables
66 print('describe number of cigarettes smoked per month')
67 desc1 = sub2['NUMCIGMO_EST'].describe()
68 print(desc1)
69
```

desc1 é o nome dado ao objeto que armazenará esses cálculos, igual a NUMCIGMO_EST. Antes, há um título da saída e então mandamos Python imprimir os resultados. Isso fornece uma contagem, média, desvio padrão, valores mínimos e máximos e os valores de percentil 25, 50 e 70. Então você pode ver que descrever é

```
describe number of cigarettes smoked per month
count   1706.000000
mean       0.525205
std         .499751
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000
Name: NUMCIGMO_EST, dtype: float64
```

extremamente útil na melhor compreensão das características importantes desta variável cigarros fumados por mês.

Sabemos agora que os jovens fumantes adultos em nossa amostra fumam em média 320 cigarros por mês. Em que o desvio padrão é de cerca de 274, podemos dizer que, em média, jovens fumantes adultos fumaram 320 por mês ± 274 cigarros. Então, como você pode ver, há uma gama extremamente grande em termos de cigarros fumados, e muita variabilidade nesta variável. Código muito semelhante pode ser usado para calcular muitas dessas estatísticas individualmente ou para gerar estatísticas descritivas adicionais. Aqui está o código adicional para gerar a média, desvio padrão, mínimo e máximo, mediana e moda de uma variável quantitativa.

```
print('mean')
mean1 = sub2['NUMCIGMO_EST'].mean()
print(mean1)

print('std')
std1 = sub2['NUMCIGMO_EST'].std()
print(std1)

print('min')
min1 = sub2['NUMCIGMO_EST'].min()
print(min1)

print('max')
max1 = sub2['NUMCIGMO_EST'].max()
print(max1)

print('median')
median1 = sub2['NUMCIGMO_EST'].median()
print(median1)

print('mode')
mode1 = sub2['NUMCIGMO_EST'].mode()
print(mode1)
```

Note que a contagem para esta variável é 1697 em vez de o tamanho da nossa amostra de jovens fumantes adultos que foi 1706. Isso ocorre porque o Python não inclui os casos com dados ausentes ou NaN nesses cálculos. Mas e se incluirmos uma variável categórica ao empregar a função describe? Como definimos anteriormente TAB12MDX, nossa variável de dependência de nicotina é categórica. Adicionando a sintaxe de descrição nos fornece estatísticas descritivas apropriadas para dados categóricos. Isto é count, número de valores exclusivos, o valor superior ou mais alto e a frequência desse valor superior.

```
70 print('describe nicotine dependence')
71 desc2 = sub2['TAB12MDX'].describe()
72 print(desc2)
```

```
describe nicotine dependence
count     1706
unique       2
top          1
freq       896
Name: TAB12MDX, dtype: int64
```

Se você não tivesse descrito esta variável como categórica, Python ainda geraria estatísticas descritivas. No entanto, muitos não fariam nenhum sentido. Se você se lembrar da variável de dependência de nicotina representada com códigos fictícios. Ou seja, sim é indicado com um 1 e não indicado com um 0. Como você

```
describe number of cigarettes smoked per month
count   1697.000000
mean     320.304361
std      274.436777
min        1.000000
25%       90.000000
50%      300.000000
75%      600.000000
max     2940.000000
Name: NUMCIGMO_EST, dtype: float64
```

pode ver aqui temos um desvio padrão baseado em códigos fictícios de 1 e 0. Além disso, os percentis são listados representando sim e não em vez de quantidades reais.

Então, novamente, é muito importante lembrar de usar as estatísticas descritivas apropriadas para variáveis quantitativas e categóricas. Para variáveis quantitativas, é melhor examinar histogramas e, em seguida, complementá-los com medidas exatas de forma, centro e propagação. Variáveis categóricas podem ser descritas com frequência distribuições ou com um gráfico de barras.

Tarefa - Criar gráficos sobre seus dados

Há uma variedade de maneiras convencionais de visualizar dados - tabelas, histogramas, gráficos de barras, etc. Agora que seus dados foram gerenciados, é hora de representar graficamente suas variáveis. Essa parte do

projeto é vital, pois fornecerá aos leitores representações visuais de seus dados e ajudará você a exibir melhor suas descobertas.

Pontuação

Sua avaliação será baseada nas evidências fornecidas por você de que concluiu todas as etapas. Quando relevante, a pontuação deverá recompensar a clareza (por exemplo, você receberá um ponto por enviar gráficos que não representam seus dados com precisão, mas dois pontos se os dados forem representados com precisão).

Você será avaliado igualmente em sua descrição de suas distribuições de frequência. Os itens específicos e seus valores de pontos são os seguintes:

1. Foi criado um gráfico univariado para cada uma das variáveis selecionadas? (2 pontos)
2. Foi criado um gráfico bivariado para as variáveis selecionadas? (2 pontos)
3. O resumo descreveu o que os gráficos revelaram em termos de variáveis individuais e a relação entre elas? (2 pontos)

Instruções

Continue com o programa que você executou com sucesso.

PASSO 1: Crie gráficos de suas variáveis uma de cada vez (gráficos univariados). Examine as medidas centrais e a dispersão.

PASSO 2: Crie um gráfico mostrando a associação entre suas variáveis explicativas e de resposta (gráfico bivariado). Sua saída deve ser interpretável (ou seja, organizada e rotulada).

O QUE APRESENTAR: Depois de escrever um programa bem-sucedido que cria gráficos univariados e bivariados, crie uma entrada de blog onde você publica seu programa e os gráficos que criou. Escreva algumas frases descrevendo o que seus gráficos revelam em termos de suas variáveis individuais e a relação entre elas.

6.2. Estatística inferencial

Até agora, demos os primeiros passos em um quadro maior de pesquisa estatística. Você identificou um conjunto de dados e usou a análise exploratória de dados para organizar e resumir os dados brutos de forma significativa e informativa. As ferramentas de análise exploratória de dados, incluindo avaliação de frequência de distribuição, representações gráficas de suas variáveis de interesse, e cálculos centrais e de dispersão, que nos ajudam a descobrir características importantes e padrões nos dados e quaisquer desvios marcantes desses padrões. Tudo isso se enquadra em Estatística Descritiva. A Estatística Descritiva visa descrever quantitativamente ou resumir uma amostra de dados.

Agora você será apresentado às Estatísticas Inferenciais, que é o nosso objetivo final. A estatística inferencial é usada para fazer inferências sobre uma população a partir da análise de uma amostra dessa população. Isso tem o objetivo expresso de chegar a conclusões generalizadas que se aplicam a toda a população. Normalmente, uma amostra aleatória da população é selecionada com base na média. Algumas das análises mais comuns em estatística inferencial incluem testes de hipóteses, intervalos de confiança e análise de regressão.

O teste de hipóteses é uma das ferramentas inferenciais mais importantes na aplicação de estatísticas para problemas da vida real. É usado quando precisamos tomar decisões sobre populações, com base em apenas uma amostra. Teste de Hipótese Estatística é definido como a avaliação de evidências fornecidas pelos dados a favor ou contra cada hipótese sobre a população. O teste de hipóteses usa métodos estatísticos para gerar evidências e tirar conclusões sobre populações inteiras. Esse teste usa teorias mutuamente exclusivas dentro do conjunto de dados da amostra, operando dentro da taxa de erro da amostra, para determinar qual hipótese tem o suporte dos dados. Os intervalos de confiança (ICs) incorporam incerteza e taxas de erro de amostra para criar uma faixa

viável de valores para um valor desconhecido em toda a população. Já a análise de regressão explica a relação entre várias variáveis independentes e uma variável dependente. Os modelos de regressão permitem que os analistas façam previsões com base nos valores presentes em um conjunto de dados de amostra.

Para realmente entender como a inferência funciona, primeiro precisamos falar sobre Probabilidade. Porque é a base subjacente de todos os métodos estatísticos. Aqui está a ideia básica. Como você sabe, as estatísticas usam uma amostra para aprender sobre a população maior da qual a amostra foi desenhada. Idealmente, a amostra deve ser aleatória para que possa representar melhor toda a população. É muito importante reconhecer embora que isso não significa que todas as amostras aleatórias são ideais. Nenhuma amostra aleatória será exatamente a mesma que qualquer outra.

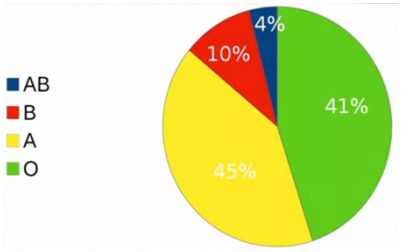
Uma amostra aleatória pode ser uma representação bastante precisa da população maior, enquanto outra amostra aleatória pode não ser precisa, puramente devido ao acaso. Infelizmente, ao olhar para uma amostra aleatória específica, que é o que acontece nas estatísticas, nunca saberemos o quanto que amostra aleatória difere da população. Esta incerteza é onde a probabilidade entra na imagem. Usamos probabilidade para quantificar o quanto esperamos que amostras aleatórias variem. Isso nos dá uma maneira de tirar conclusões sobre a população em face da incerteza que é gerada pelo uso de uma amostra aleatória.

Como exemplo, vamos supor que estamos interessados em estimar a porcentagem de adultos norte-americanos que favorecem a pena de morte. Para fazer isso, escolhemos uma amostra aleatória de 1.200 adultos norte-americanos e pedir sua opinião a favor ou contra a pena de morte. Descobrimos que 744 dos 1200, ou 62% são a favor. Aqui está uma imagem que ilustra o que fizemos e encontramos em nosso exemplo. Nosso objetivo aqui é inferir, tirar conclusões sobre as opiniões de toda a população de adultos norte-americanos sobre a pena de morte, com base nas opiniões de apenas 1200 deles, podemos concluir absolutamente que 62% da população favorece a pena de morte?

Outra amostra aleatória poderia dar um resultado muito diferente, então estamos incertos. Como nossa amostra é aleatória, sabemos que nossa incerteza é devido ao acaso. Não se deve a problemas de como a amostra foi coletada. Portanto, podemos usar a probabilidade para descrever a probabilidade de que nossa amostra esteja dentro de um nível desejado de precisão. Por exemplo, probabilidade pode responder à pergunta, quão provável é que nossa estimativa amostral esteja dentro de 3% da porcentagem REAL de TODOS os adultos norte-americanos que são a favor da pena de morte. A resposta a esta pergunta, que encontramos usando a probabilidade obviamente terá um impacto importante na confiança que podemos anexar ao passo de inferência. Em particular, se acharmos bastante improvável que a porcentagem da amostra seja muito diferente da porcentagem da população, então temos boa confiança de que podemos tirar conclusões sobre a população com base na amostra. Então vamos definir probabilidade um pouco mais cuidadosamente.

6.2.1. Da amostra à população

Para entender melhor a relação entre amostra e população, vamos considerar dois exemplos simples. Aqui estão as distribuições de tipos sanguíneos na população dos EUA. Você pode ver os tipos de sangue comuns incluem Tipo A e Tipo O, com tipos de sangue menos comuns, incluindo AB e B. Vamos supor agora que tomamos uma amostra de 500 pessoas nos Estados Unidos, registramos seu tipo sanguíneo e exibir os resultados da amostra.

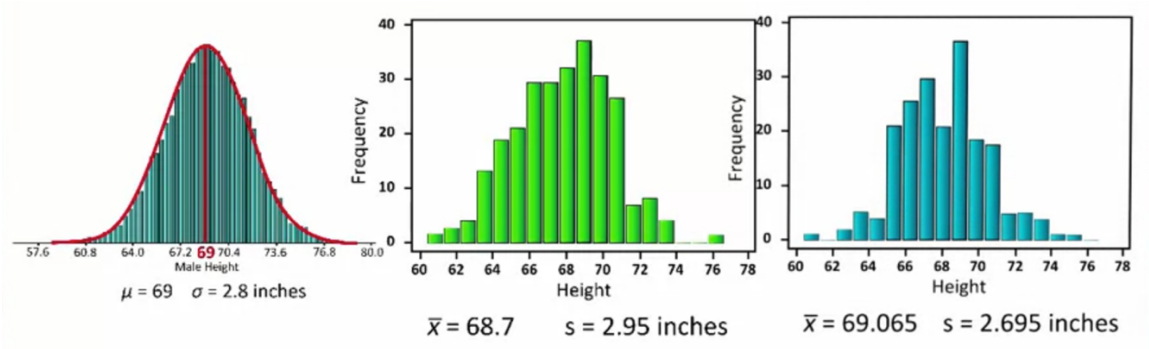


Se você olhar com cuidado, notará que as percentagens de cada tipo sanguíneo de nossa amostra são ligeiramente diferentes das percentagens da população. Mas tenho certeza de que isso não te surpreende, certo? Quero dizer, já que pegamos uma amostra de apenas 500 indivíduos, não podemos esperar que nossa amostra se comporte exatamente como a população. Mas se a amostra é aleatória, e este foi, esperamos obter resultados que não são tão diferentes dos resultados de toda a população e isso é o que encontramos. Mais uma amostra aleatória de 500 indivíduos, revela resultados que são ligeiramente diferentes das figuras populacionais e também de que temos na primeira amostra.

Esta ideia muito intuitiva de que os resultados da amostra mudam de amostra para amostra, é chamada de variabilidade de amostragem. Aqui está outro exemplo para ajudar a entender melhor a relação entre a população de amostragem. Este exemplo é baseado nas alturas entre a população dos EUA de **todos** os homens adultos. Como você pode ver, segue uma distribuição normal com uma média de 69 polegadas e um desvio padrão de 2,8 polegadas.

Digamos que uma amostra de 200 homens foi escolhida e suas alturas foram registradas. Estes são os resultados da amostra 2. A média da amostra é de 68,7 polegadas, e o desvio padrão da amostra é de 2,95 polegadas. Novamente, observe que os resultados da amostra são ligeiramente diferentes dos resultados da população.

O histograma que criamos para a primeira amostra, se assemelha à distribuição normal da população. No entanto, a média da amostra no desvio padrão é ligeiramente diferente da média da população no desvio padrão. Vamos tirar outra amostra de duzentos homens exibidos aqui na amostra dois. A média da amostra é de 69,065 polegadas e o desvio padrão da amostra é de 2,659 polegadas. Este exemplo, novamente, demonstra a variabilidade da amostragem. Embora os resultados da amostra estejam muito próximos dos resultados da população, eles são ligeiramente diferentes dos resultados encontrados na primeira amostra.



Em ambos os exemplos, temos números que descrevem a população e números que descrevem a amostra.

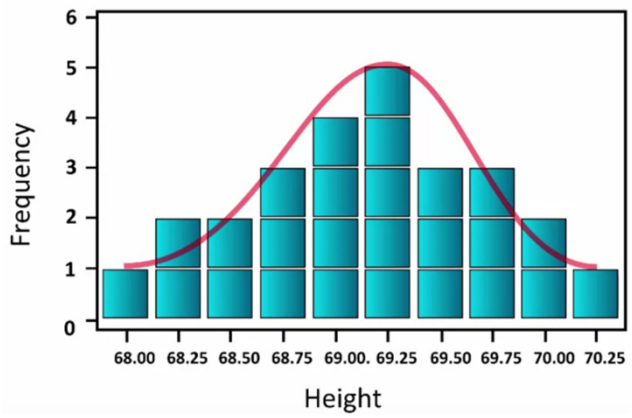
Um *parâmetro* é um número que descreve a população e uma ***estatística* é um número calculado a partir de uma amostra**. Os parâmetros são tipicamente desconhecidos, porque é impraticável ou até mesmo impossível saber exatamente quais valores uma variável leva para cada membro de uma população muito grande. As estatísticas são calculadas a partir de amostras, e cada amostra de uma população terá estatísticas

diferentes. As estatísticas de diferentes amostras de uma população variam. Isto é devido à variabilidade da amostragem.

Até agora, temos feito distribuições baseadas em variáveis individuais. Teoricamente, podemos criar distribuições a partir de médias ou proporções tiradas de várias amostras aleatórias extraídas de uma população. Esta é a grande ideia por trás das estatísticas inferenciais.

Como exemplo, suponha que selecionamos 30 amostras aleatórias separadas em vez de apenas duas. E cada uma das 30 amostras aleatórias tem 500 indivíduos retirados da população de adultos norte-americanos. A primeira amostra tem uma altura média de 69 polegadas. Poderíamos criar um gráfico de barras e plotar essa média para nossa primeira amostra no gráfico. Se nossa segunda amostra tivesse uma altura média de 68,5 polegadas, adicionaríamos isso ao gráfico. À medida que continuamos a traçar a altura média de cada amostra aleatória, um padrão começaria a surgir. Observe como há mais meios de amostra a 69,25 polegadas do que em qualquer outro comprimento.

Observe também como, à medida que o comprimento se torna maior ou menor, há cada vez menos meios de amostra. Esta é uma característica da distribuição amostral, se estamos medindo a média de uma variável quantitativa ou a proporção de variável categórica ou qualquer outra estatística amostral. Ou seja, à medida que desenhamos mais e mais amostras, a distribuição da amostra estatística se tornará cada vez mais normalmente distribuída.



Este resultado é conhecido como o **Teorema do Limite Central**, que afirma que, enquanto amostras adequadamente grandes e um número suficientemente grande de amostras são extraídas de uma população, a distribuição das estatísticas das amostras, seja de média, proporção, desvio padrão ou qualquer outra estatística, será normalmente distribuída. Nossos projetos dependem de apenas uma amostra. No entanto, se essa amostra é representativa de uma população maior, os testes estatísticos inferenciais nos permitem estimar com diferentes níveis de parâmetros de certeza para toda a população. Esta ideia é a base para cada uma das ferramentas inferenciais que você usará para responder à sua pergunta de pesquisa.

6.2.2. Teste de Hipótese

Teste de hipóteses é uma das ferramentas inferenciais mais importantes quando se trata de para a aplicação de estatísticas para problemas da vida real. Teste de hipóteses é usado quando precisamos tomar decisões sobre populações com base apenas em informações de amostra. Uma variedade de testes estatísticos é usada para ajudar a chegar a essas decisões. Por exemplo, a análise do teste de variância, ANOVA. E o Qui Quadrado Teste da Independência, para citar alguns. Mas todos eles incluem os mesmos passos básicos.

Passos envolvidos no teste de hipóteses, incluem especificar a hipótese nula H_0, e a hipótese alternativa, H_a. Escolhendo uma amostra, avaliando a evidência e tirando conclusões. Teste de hipóteses estatísticas é definido como a avaliação de evidências fornecidas pelos dados a favor ou contra cada hipótese sobre a população.

Para fornecer um exemplo de teste de hipótese, vamos usar o conjunto de dados NESARC. Uma amostra representativa de 43.093 adultos nos Estados Unidos. Vamos avaliar se existe ou não uma associação entre um diagnóstico de depressão maior e o quanto uma pessoa fuma. Vamos trabalhar através do exemplo usando as quatro etapas.

1. Especificar a hipótese nula e alternativa,
2. Escolher uma amostra
3. Avaliar a evidência e
4. Tirar conclusões

Primeiro, há duas hipóteses opostas para questionar. A hipótese nula, comumente mostrada como H_0, é que não há diferença na quantidade de tabagismo entre pessoas com e sem depressão. A hipótese alternativa, mostrada como H_a ou às vezes mostrado como H_1, é que existe uma diferença na quantidade de tabagismo entre pessoas com e sem depressão.

A hipótese nula basicamente, diz que nada de especial está acontecendo entre depressão e tabagismo. Em outras palavras, que eles não estão relacionados uns com os outros. A hipótese alternativa diz que existe uma relação e permite que a diferença no tabagismo naqueles indivíduos com e sem depressão possa ser positiva ou negativa. Ou seja, indivíduos com depressão podem fumar mais do que indivíduos sem depressão, ou podem fumar menos.

Depois de declarar a hipótese nula e alternativa, precisamos escolher uma amostra. Nós vamos usar o conjunto de dados NESARC, e nós só vamos avaliar essas hipóteses entre indivíduos que são fumantes e que são mais jovens, em vez de adultos mais velhos. Restringimos os dados NESARC para indivíduos que são: 1. fumantes diários atuais, ou seja, eles fumaram todos os dias no mês anterior ao questionário. E, 2. tem idades entre 18 a 25 anos.

Esta amostra, N = 1320, mostrou o seguinte. Jovens adultos fumantes diários com depressão fumavam uma média de 13,9 cigarros por dia com um desvio padrão de 9,2 cigarros. Jovens adultos fumantes diários sem depressão fumavam uma média de 13,2 cigarros por dia com um desvio padrão de 8,5 cigarros. Embora seja verdade que 13,9 cigarros por dia são mais de 13,2 cigarros por dia, não é de todo claro que este é uma diferença grande o suficiente para rejeitar a hipótese nula. Ou dizer que os fumantes com depressão fumam significativamente mais do que os fumantes sem depressão.

Embora seja verdade que 13,9 cigarros por dia são mais de 13,2 cigarros por dia, não é de todo claro que esta é uma diferença grande o suficiente para rejeitar a hipótese nula. Ou dizer que os fumantes com depressão fumam significativamente mais do que os fumantes sem depressão. Portanto, precisamos avaliar a evidência, a fim de determinar se os dados fornecem evidência forte o suficiente contra a hipótese nula. Ou seja, contra a alegação de que não há relação entre fumar e depressão. Nós realmente precisamos nos perguntar, quão surpreendente ou raro é para obter uma diferença de 0,7 cigarros fumaça por dia entre nossos dois grupos? Ou seja, aqueles com depressão, e aqueles sem, assumindo que a hipótese nula é verdadeira, que não há relação entre fumar e depressão.

Esta é uma etapa onde calculamos a probabilidade de obter dados como este quando a hipótese nula é verdadeira. Em certo sentido, este é realmente o coração do processo, uma vez que tiramos nossas conclusões com base na estimativa de probabilidade. A hipótese nula é geralmente assumida como verdadeira até que a evidência indique o contrário. A probabilidade de obtermos uma diferença desse tamanho no número médio de cigarros fumados em uma amostra aleatória de 1.320 participantes é de aproximadamente 0,17 ou 17%.

Vamos falar sobre como isso é calculado para os diferentes testes estatísticos mais tarde. O ponto importante nesta fase é que é esse tipo de evidência que teremos considerando cada vez que decidirmos aceitar ou rejeitar a hipótese nula. Então, como exatamente usamos essa probabilidade para chegar a uma conclusão

sobre a hipótese nula? Lembre-se, se a hipótese nula for verdadeira, não há associação. Há uma probabilidade de 0,17 ou 17% de observar esse tamanho de diferença entre fumantes com e sem depressão.

A tradução desta probabilidade de 17% é que se tirássemos 100 amostras aleatórias de nossa população, estaríamos errados 17 de 100 vezes se rejeitássemos a hipótese nula e dissemos que havia uma diferença na quantidade de tabagismo para fumantes com e sem depressão. Agora temos que decidir se ou não isso é algo que nos sentimos confortáveis. Importa-se de cometer um erro e dizendo que há uma diferença na quantidade de fumar 17 em cada 100 vezes?

Essa probabilidade de 0,17 torna o que estamos observando raro o suficiente para nos fazer sentir confiantes em rejeitar a hipótese nula?

Provavelmente todos concordamos que uma probabilidade de 0,50 certamente não nos daria confiança suficiente para rejeitar a hipótese nula. Porque 0,50, ou 50%, significa que estaríamos certos 50 em 100 vezes, e errado 50 em 100 vezes. Não é melhor do que tomar decisões baseadas no lançar de uma moeda.

Estar errado 17 de 100 vezes nos faria muito menos propensos a ser errados ao rejeitar a hipótese nula, mas ainda estaríamos menos certos do que se a probabilidade fosse ainda menor, digamos 10 ou até 5%. Basicamente, esta é a nossa decisão ao testar hipóteses. Para tomar essa decisão, seria bom ter algum tipo de diretriz ou padrão. Que probabilidade nos daria confiança em rejeitar uma hipótese nula?

6.2.3. Valor-p e Intervalo de Confiança

A razão para usar um Teste Inferencial é obter um valor de probabilidade, comumente chamado valor-p. O valor de p fornece uma estimativa de quantas vezes nós iríamos obter o resultado obtido por acaso se de fato, a hipótese nula é verdadeira. Em estatística, um resultado é chamado de estatisticamente significativo se é improvável que tenha ocorrido apenas por acaso.

O padrão ou corte mais comumente usado é 0,05 ou 5%. Porque este padrão, ou corte é tão importante que tem um nome especial. É chamado de nível de significância de um teste, e é geralmente denotado pela letra grega alfa, então alfa é igual a 0,05.

Se o valor de p for pequeno, menor que 0,05, isso sugere que é mais de 95% provável que a associação de interesse esteja presente após amostras repetidas tiradas da população, em outras palavras, uma distribuição de amostragem. Se o valor de p for menor que alfa, que geralmente é 0,05, então os dados que obtivemos são considerados raros ou surpreendentes o suficiente quando a hipótese nula, H_0 é verdadeira. E dizemos, que os dados fornecem evidências significativas contra a hipótese nula. Então, rejeitamos a hipótese nula e aceitamos a hipótese alternativa, H_a.

Se o valor-p for maior que alfa, então os dados não são considerados surpreendentes o suficiente quando a hipótese nula é verdadeira. E dizemos, que nossos dados não fornecem evidências suficientes para rejeitar a hipótese nula. Ou equivalentemente, que os dados não fornecem evidências suficientes para aceitar a hipótese alternativa.

Assim, encontrar um valor de p menor que ou igual a 0,05 significa que o achado é estatisticamente significativo, e podemos rejeitar a hipótese nula e aceitar a hipótese alternativa. Este valor-p também é conhecido como **Taxa de Erro do Tipo Um**, uma vez que denota o número de vezes que estaríamos errados ao rejeitar a hipótese nula quando era verdadeira.

Rejeitar a hipótese nula quando é verdadeira também é chamado de Erro do Tipo Um.

Olhando para o valor de p em nosso exemplo, vemos que não há evidência adequada para rejeitar a hipótese nula porque o valor de p foi 0,17, que é definitivamente maior que 0,05. Em outras palavras, não foi rejeitada a hipótese nula de que não há associação entre depressão e número de cigarros fumados entre jovens fumantes diários. Aceitamos a hipótese nula. Não há associação entre tabagismo e depressão, porque os dados não

fornecem evidências suficientes para aceitar a hipótese alternativa, de que existe associação entre tabagismo e depressão.

Vamos mudar ligeiramente a questão da pesquisa para demonstrar que as decisões que você toma sobre sua amostra e suas variáveis podem afetar suas descobertas e as conclusões que você tira. Usando nosso exemplo, ainda estamos interessados na associação entre depressão e tabagismo. No entanto, decidimos não nos limitar a considerar apenas indivíduos que fumam diariamente. Vamos olhar para uma população mais ampla de jovens adultos, e considerar aqueles que já fumaram no ano passado, seja diariamente ou mais irregularmente.

O tamanho da amostra no conjunto de dados NESARC é 1.706. Com esta amostra, descobrimos que jovens adultos com depressão fumavam uma média de 351,7 cigarros por mês com um desvio padrão de 300 cigarros. Jovens adultos sem depressão fumavam uma média de 313,5 cigarros por mês, com um desvio padrão de 268,2 cigarros. Assim, a diferença entre a quantidade de cigarros fumados entre os jovens adultos que fumou no ano passado com e sem depressão é de 38,2 cigarros por mês, quase 2 pacotes.

O valor de p deste cenário revisado é 0,0285, obviamente inferior a 0,05. Isso significa que a probabilidade de obtermos uma diferença desse tamanho em o número médio de cigarros fumados em uma amostra aleatória de 1.706 participantes é menor que 3%, que é um valor-p inferior a 0,05. Então, neste caso, podemos rejeitar a hipótese nula, e dizem que jovens fumantes adultos com depressão fumam significativamente mais cigarros por mês do que jovens fumantes adultos sem depressão.

Se olharmos novamente para a linha numérica de probabilidades, podemos traduzir esta descoberta da seguinte maneira. Se rejeitarmos a hipótese nula e dissermos que há uma diferença entre o número médio de cigarros fumados por mês entre os jovens adultos, com e sem depressão, estaríamos errados menos de 3 em cada 100 vezes. Estaríamos corretos mais de 97% do tempo. Baseado nos padrões da ciência, este é um nível de certeza que nos dá confiança em dizer que há uma associação significativa entre fumar e depressão entre jovens fumantes adultos atuais.

6.2.4. Escolhendo testes estatísticos

Você foi apresentado ao processo geral de testes de hipóteses. É hora de aprender a testar sua própria hipótese. Você sempre estará interpretando valores p, independentemente do teste inferencial que você usa.

O teste estatístico específico que você usa para avaliar suas hipóteses, dependerá do tipo de variáveis explicativas e de resposta que você escolheu.

- Se você tiver uma variável explicativa categórica e uma variável de resposta quantitativa, você usaria uma Análise de Variância, ANOVA como teste inferencial.
- Se você tem uma variável explicativa categórica, e sua variável de resposta também é uma variável categórica, você usaria o Teste de Independência Qui-Quadrado como seu teste inferencial.
- Se ambas as variáveis explicativas e de resposta forem quantitativas, você usaria um coeficiente de correlação como teste inferencial.
- Se sua variável explicativa for quantitativa e sua variável de resposta for categórica, você categorizaria sua variável explicativa com apenas dois níveis e, em seguida, use o Teste Qui-Quadrado da Independência como seu teste inferencial.

		Resposta (dependente)	
		Categórica	Quantitativa
Explanatória (independente)	Categórica	C -> C **Teste de Independência Qui-quadrado**	C -> Q **Análise de Variância (ANOVA)**
	Quantitativa	Q -> C **Qui-quadrado ajustado**	Q -> Q **Correlação de Pearson**

6.2.5. Análise de Variância - ANOVA

Então, finalmente, estamos prontos para começar a testar nossas perguntas de pesquisa estatisticamente. Embora tenhamos demorado algum tempo para chegar aqui, nossos passos anteriores nunca devem ser evitados. Ou seja, não importa o quão sofisticado você possa se tornar como um pesquisador quantitativo, você sempre precisará examinar seu livro de códigos, gerenciar seus dados e examinar estatísticas descritivas para as variáveis de interesse.

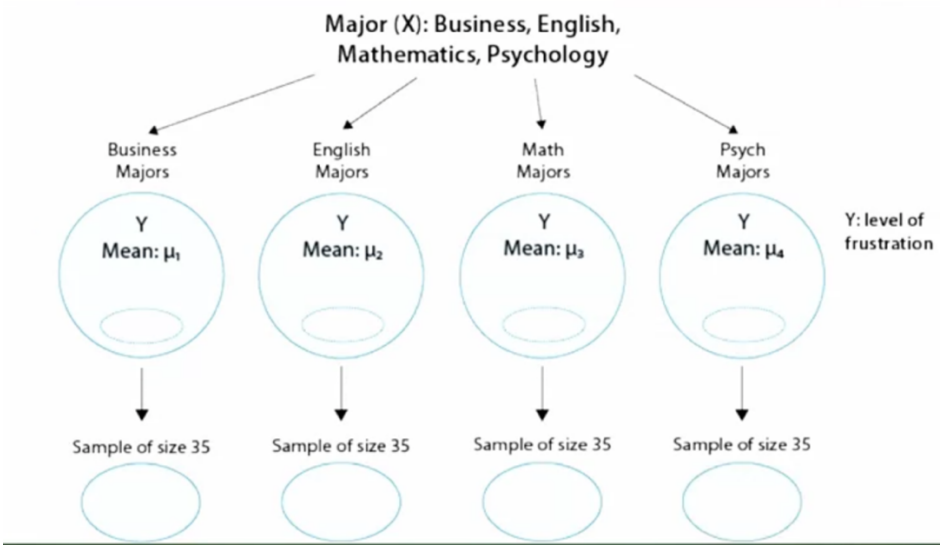
Na descrição do teste de hipóteses, quando analisamos a associação entre depressão e tabagismo, estávamos trabalhando com uma variável explicativa categórica, a presença ou ausência de depressão, e uma variável de resposta quantitativa, o número de cigarros fumados por mês. Quando você está testando hipótese com a variável explicativa categórica e uma variável de resposta quantitativa, a ferramenta que você deve usar é Análise de Variância, também chamada ANOVA.

Agora que você entende em quais situações você usaria ANOVA, estamos prontos para aprender como ela funciona ou mais especificamente o que a ideia está por trás da comparação de médias. O teste que você usará chama-se ANOVA F-test. Então vamos usar outra questão de pesquisa categórica para quantitativa.

A frustração acadêmica está relacionada à área cursada?

Neste exemplo, um reitor da faculdade acredita que estudantes com diferentes cursos podem experimentar diferentes níveis de frustração acadêmica. Amostras aleatórias de 35 indivíduos, cada um dos cursos de Negócios, Inglês, Matemática, e Psicologia foram convidados a avaliar seu nível de frustração acadêmica, em uma escala de um, o mais baixo, para vinte, o mais alto.

Esta figura destaca que estaremos examinando a relação entre major, nossa variável explicativa ou X, e o nível de frustração, nossa resposta, ou variável Y para comparar os diferentes meios de níveis de frustração entre os quatro principais definidos por X.



As alegações de hipótese nula que não há relação entre as variáveis de resposta explicativa e, x e y. Uma vez que a relação é examinada comparando as médias de y nas populações, definidas pelos valores de x, nenhuma relação significaria que todas as médias são iguais. Portanto, a hipótese nula do teste f é média da população 1 igual à média da população 2 é igual a média da população 3 igual à média da população 4.

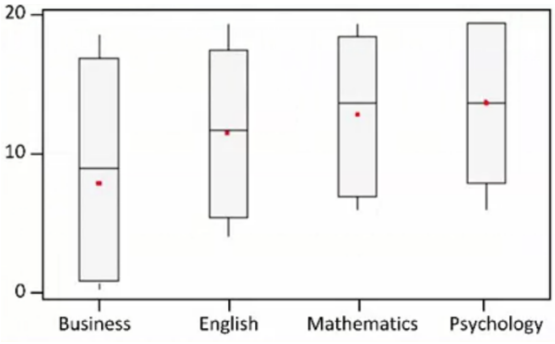
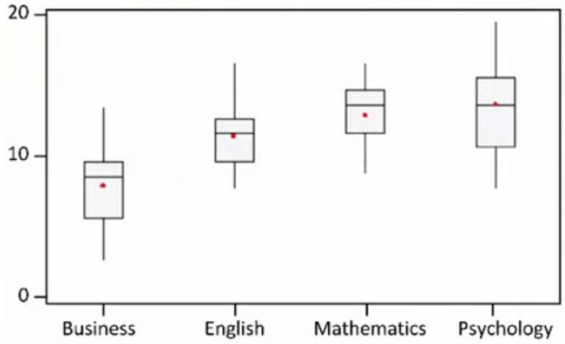
Aqui temos apenas uma hipótese alternativa que afirma que há uma relação entre x e y. A variável independente e dependente. Em termos dos meios, ele simplesmente diz o contrário, que nem todos os meios são iguais e simplesmente escrevemos h1, nem todas as médias da população são iguais. Há muitas maneiras para a população significa não ser igual. Falaremos sobre isso mais tarde.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a: \textit{not all the } \mu \textit{ are equal}$$

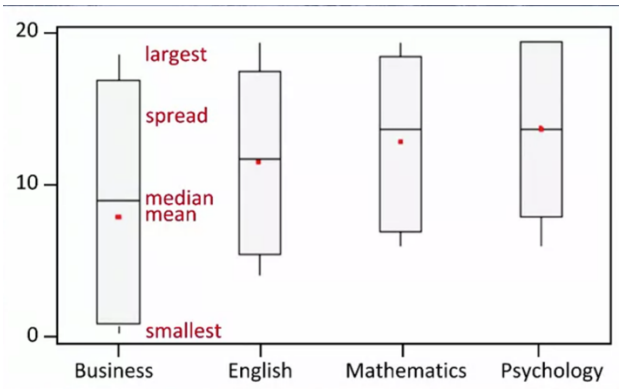
Por enquanto, vamos pensar sobre como iríamos testar se a população significa são iguais. Poderíamos calcular o nível médio de frustração para cada major e ver quão distantes essas médias de amostra estão. Ou, em outras palavras, meça a variação entre as médias da amostra. Se descobrirmos que as quatro médias da amostra não estão todas juntas, diremos que temos evidências contra a hipótese nula. E caso contrário, se eles estão próximos, diremos que não temos evidências contra a hipótese nula. Isso parece bastante simples, mas isso é suficiente?

- A pontuação média de frustração da amostra dos 35 alunos da área de negócios é: y1 = 7.3.
- A pontuação média de frustração da amostra para os 35 alunos de inglês é: y2 = 11,8.
- A pontuação média de frustração da amostra para os 35 alunos de matemática é: y3 = 13,2.
- E a pontuação média de frustração da amostra para os 35 alunos de Psicologia é: y4 = 14,0.

Aqui está uma representação gráfica de dois conjuntos de dados hipotéticos tirados de duas diferentes populações. Por exemplo, estudantes no País Um e estudantes no País Dois. Em nossas amostras hipotéticas, os meios são os mesmos, mas eles aparecem neste boxplot de forma muito diferente.

País Um	**País Dois**
	

Um boxplot é uma maneira conveniente de descrever graficamente grupos de dados numéricos incluindo informações descritivas como a menor observação do grupo, a média e a mediana, a maior observação e a dispersão ou variabilidade dos valores. A parte superior da linha que se destaca do topo do gráfico de caixa e a parte inferior da linha que se destaca da parte inferior do gráfico de caixa são os valores mais altos e mais baixos. O ponto vermelho é a média. A linha horizontal do meio é a mediana.



Você pode ver que cada conjunto de dados tem o mesmo conjunto de médias e, portanto, as mesmas diferenças entre eles. Ou seja, estudantes no País Um e estudantes no País Dois. Ambos mostram dados para

quatro grupos com uma média de amostra de 7.3, 11.8, 13.2 e 14.0 indicada com marcas vermelhas. A diferença importante entre os dois conjuntos de dados é que o primeiro representa os dados com uma grande quantidade de variação dentro de cada um dos quatro grupos. O segundo representa dados com uma pequena quantidade de variação dentro de cada os quatro grupos.

Boxplots para País Um mostram muita sobreposição entre os quatro grupos devido à grande quantidade de variação nas pontuações de frustração dentro dos grupos. Pode-se imaginar os dados decorrentes de quatro amostras aleatórias tiradas de quatro populações, todas com a mesma média de cerca de 11 ou 12. O primeiro grupo de valores pode ter sido um pouco no lado baixo e os outros três um pouco no lado alto. Mas tais diferenças poderiam ter surgido por acaso. Este seria o caso se a hipótese nula alegando que médias de população iguais fossem verdadeiras.

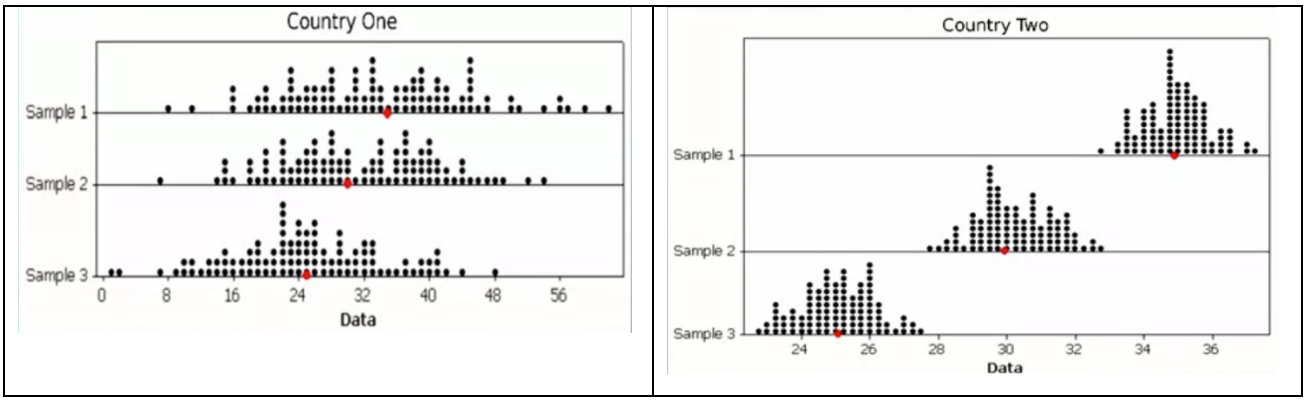
Boxplots para o País Dois mostram muito pouca sobreposição por causa da pequena quantidade de pontuação de variação e frustração dentro dos grupos. Seria muito difícil acreditar que estamos amostrando de quatro grupos que têm necessidades populacionais iguais. Este caso é um exemplo de quando a hipótese nula alegando que a população igual precisa seria falsa.

A pergunta que precisamos responder com o Teste ANOVA F é, as diferenças entre as médias da amostra devido a verdadeiras diferenças entre as médias da população, ou meramente devido à variabilidade amostral? Para responder a esta pergunta, usando nossos dados, obviamente precisamos olhar para a variação entre as médias de amostra. Mas isso não é suficiente. Também precisamos olhar para a variação entre as médias de amostra em relação à variação dentro dos grupos.

Então F é a variação entre as médias da amostra dividida pela variação dentro dos grupos. Em outras palavras, precisamos olhar para a quantidade, variação entre as médias de amostra, dividido por variação dentro de grupos. Que mede até que ponto a diferença entre os grupos amostrais, significa, domina sobre a variação usual dentro dos grupos amostrais. Que reflete diferenças em indivíduos que são típicos em amostras aleatórias.

$$F = \frac{\text{Variação entre as médias das amostras}}{\text{Variação dentro dos Grupos}}$$

Quando a variação dentro dos grupos é grande, como no País Um, as diferenças ou variação entre as médias da amostra podem se tornar insignificantes. E os dados forneceriam muito pouca evidência contra a hipótese nula. Quando a variação dentro de grupos é pequena, como no País Dois, a variação entre as médias da amostra domina. E os dados têm evidências mais fortes contra a hipótese nula. Olhando para a proporção de variações é a ideia por trás das comparações e significa, portanto, a análise do nome da variância.



Aqui estão os resultados da análise de variância para o País Dois. Testando a relação entre pontuação maior e frustração. A estatística F circulada em vermelho é 46,60. Como sabemos que esta é a variabilidade entre as médias de amostra divididas pela variabilidade dentro dos grupos, esse grande número sugere que a variabilidade entre as médias amostrais é muito maior do que a dos grupos amostrais.

O valor P do Teste ANOVA F é a probabilidade de obter uma estatística F como maior que obtivemos ou mesmo maior se a hipótese nula fosse verdadeira. Ou seja, se a população significa ser igual. Em outras palavras, ele nos diz como é surpreendente encontrar dados como os observados, assumindo que não há diferença entre os meios populacionais. Este valor P é praticamente 0, dizendo-nos que seria quase impossível obter dados como aqueles observados se o nível médio de frustração dos quatro cursos fosse o mesmo que as alegações de hipótese nula.

One-Way ANOVA: Frustration Score Versus Major

Source	DF	SS	MS	F	P
Major	3	939.85	313.28	46.60	0.0001
Error	136	914.29	6.72		
Total	139	1854.14			

S = 2.593 R-Sq = 50.69% R-Sq =(adj) = 49.60%

Level	N	Mean	StDev
Business	35	7.314	2.898
English	35	11.771	2.088
Mathematics	35	13.200	2.153
Psychology	35	14.029	3.080

O valor P 0,0001 sugere que vamos rejeitar incorretamente a hipótese nula uma em dez mil vezes. E estaremos corretos em aceitar a hipótese alternativa 9999 vezes em 10.000 vezes. Assim, podemos concluir com confiança que os meios de nível de frustração dos quatro cursos não são todos iguais. Ou em outras palavras, há uma associação significativa entre nível de frustração e maior. Então aceitamos a hipótese alternativa e rejeitamos a hipótese nula. Agora que você tem uma sensação de análise de variância, vamos executar o teste usando SAS. Usaremos um exemplo descrito pela primeira vez no teste de hipóteses.

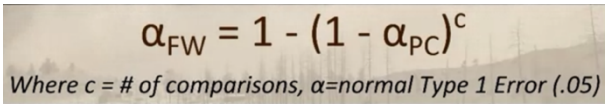
Teste de Post Hoc com Anova

Quando a variável explicativa (independente) representa mais de dois grupos, um teste ANOVA significativo não nos diz quais grupos são diferentes dos outros. Para determinar quais grupos são diferentes dos outros, precisaríamos realizar um teste post hoc. Um teste post hoc conduz comparações emparelhadas post hoc. Post hoc significa depois do fato. E essas comparações emparelhadas post hoc devem ser conduzidas de uma maneira específica, a fim de evitar erros excessivos do tipo 1.

Erro do Tipo 1, ocorre quando você toma uma decisão incorreta sobre a hipótese nula. Ou seja, você rejeita a hipótese nula quando a hipótese nula for verdadeira. Por que não podemos simplesmente executar vários ANOVAs? Ou seja, por que não podemos apenas subdefinir nossas observações e levar duas de cada vez?

Como você sabe, aceitamos significância e rejeitamos a hipótese nula em p menor ou igual a 0,05. Uma chance de 5% de estarmos errados e cometermos um erro de tipo 1. Na verdade, há 5% de chance de fazer um erro de tipo 1 para cada análise de variância que realizamos nesta questão. Portanto, realizar vários testes significa que nossa chance geral de cometer erro tipo 1, pode ser muito maior do que 5%. Veja como funciona.

# Tests	Comparison α	Family-wise α
1	.05	.05
3	.05	.14
6	.05	.26
10	.05	.40
15	.05	.54



Usando a fórmula exibida sob esta tabela, você pode ver que, enquanto um teste tem uma Taxa de Erro Tipo 1 de 0,05, no momento em que realizamos dez testes sobre esta questão, nossa chance de rejeitar a hipótese nula quando a hipótese nula for verdadeira é de até 40%. Este aumento na taxa de erro Tipo 1 é chamado de taxa de erro familiar e é a taxa de erro para o grupo de comparação de pares.

Os testes post-hoc são projetados para avaliar a diferença entre pares de médias enquanto protegem contra a inflação de erros de Tipo 1. E há muitos testes post hoc para escolher, quando se trata de análise de variância. Há o Sidak, o teste T Holm. e Teste de diferença menos significativa de Fisher. Teste de diferença honestamente significativa de Tukey, teste de Scheffe, teste de Newman-Keuls, teste de Comparação Múltipla de Dunnett, teste de alcance múltiplo Duncan e o Procedimento Bonferroni. É o suficiente para fazer sua cabeça nadar.

Embora haja certamente diferenças em quão conservador cada teste é em termos de proteção contra erro do tipo um, em muitos casos é muito menos importante qual teste post hoc você conduz e muito mais importante que você conduza um.

Para realizar comparações emparelhadas post hoc no contexto da minha ANOVA, examinando a associação entre etnia e número de cigarros fumados por mês, vou usar o Tukey HSDT, ou Honestamente Significativa Diferença Test. Para fazer isso, vou primeiro adicionar uma instrução import para a biblioteca statsmodels.stats.multicomp no meu script python como multi, o termo que usarei para me referir à biblioteca mais tarde no meu programa. Em seguida, adicionarei o seguinte código ao final do meu programa. Estou chamando o objeto que irá armazenar minhas múltiplas comparações MC1 e usar a função multicomparação da biblioteca multicomp de estatísticas de modelos multicomp, que eu importei como multi acima. Depois, incluo nesta declaração a variável de resposta quantitativa e a variável explicativa categórica entre parênteses. **res1** é o nome que estou dando ao objeto que armazenará meus resultados post hoc. Em seguida, eu defini que igual ao meu objeto de comparações múltiplas, e eu solicito o teste hsd tukey. Finalmente, peço ao Python para imprimir esses resultados com a função de resumo.

```
70 mc1 = multi.MultiComparison(sub3['NUMCIGMO_EST'], sub3['ETHRACE2A'])
71 res1 = mc1.tukeyhsd()
72 print(res1.summary())
73
```

Aqui vemos uma tabela exibindo as comparações emparelhadas post hoc Tukey. Ou seja, diferenças na quantidade de tabagismo para cada par de grupos étnicos. Na primeira linha da tabela, vemos a comparação entre o grupo étnico um e dois. Indivíduos endossando etnia branca versus aqueles que endossam etnia negra. Assim como diferenças médias no número de cigarros fumados entre estes dois grupos. Python calculou um valor P, embora não seja exibido, que leva as múltiplas comparações em consideração e nos protege de inflar nosso erro tipo 1 e rejeitar a hipótese nula quando a hipótese nula é verdadeira. Na última coluna, podemos determinar quais grupos étnicos fumam significativamente diferente do número médio de cigarros que os outros identificando as comparações nas quais podemos rejeitar a hipótese nula, isto é, em que rejeitar é igual a verdadeiro. Assim, podemos ver que o grupo étnico um é significativamente diferente dos grupos étnicos dois, quatro e cinco. E quando examinamos novamente meios de grupo, podemos dizer que indivíduos endossando branco etnia, grupo um, fumam significativamente mais cigarros por mês, do que indivíduos endossando etnia negra, asiática e hispânica. Grupos dois, quatro e cinco.

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
=================================================
group1 group2  meandiff   lower     upper   reject
-------------------------------------------------
  1      2    -109.5127 -164.6441  -54.3814  True
  1      3     -57.7984 -172.5914  56.9945  False
  1      4    -124.5279 -222.9229  -26.1329  True
  1      5    -149.0283  -194.89  -103.1665  True
  2      3     51.7143   -71.6021  175.0307 False
  2      4     -15.0152  -123.233  93.2026  False
  2      5     -39.5156 -103.8025  24.7714  False
  3      4     -66.7295 -214.5437  81.0848  False
  3      5     -91.2298 -210.6902  28.2305  False
  4      5     -24.5004 -128.3027   79.302  False
-------------------------------------------------
```

6.2.6. Teste de Independência Qui-Quadrado

A análise de variância envolveu examinar a relação entre uma variável explicativa categórica e a variável Resposta quantitativa. Em seguida, vamos considerar inferências sobre as relações entre duas variáveis categóricas. O teste estatístico que responderá a esta pergunta é chamado de Teste de Independência Qui-Quadrado. Chi é uma letra grega que se parece com um grande X. Então, às vezes, você verá este teste denotado com um X ao quadrado.

Para esta ferramenta estatística, vamos começar com um novo exemplo. No início da década de 1970, um jovem desafiou uma Lei Estadual de Oklahoma que proibiu a venda de 3,2 cerveja, homens com menos de 21 anos de idade. Mas permitiu que fosse vendida a mulheres na mesma faixa etária. O caso foi finalmente ouvido pelo Supremo Tribunal dos EUA. A principal justificativa fornecida por Oklahoma para a lei era a segurança do trânsito. Uma das três principais peças de dados apresentadas ao tribunal foi o resultado de uma pesquisa aleatória na estrada que registrou informações sobre gênero. E se o motorista estava ou não bebendo álcool em nas duas horas anteriores. Houve um total de 619 motoristas com menos de 20 anos de idade incluídos na pesquisa. Abaixo representamos uma tabela bidirecional resumindo os relatos observados na pesquisa na estrada.

Gênero	Sim	Não	Total
Masculino	77	404	481
Feminino	16	122	138
Total	93	526	619

Nossa tarefa é abordar se esses resultados fornecem evidências de uma significativa ou estatisticamente significativa relação entre gênero e direção embriagada. Ambas as variáveis são duas variáveis categóricas valorizadas e, portanto, nossa tabela de duas vias de contagens observadas é um dois por dois.

O procedimento Qui-Quadrado não se limita a duas situações. Ele também pode ser usado para um número maior de categorias explicativas. A chave para relatar resumos apropriados para uma tabela bidirecional é decidir qual das duas variáveis categóricas desempenha o papel da variável explicativa. E, em seguida, calculando as percentagens condicionais separadamente. Ou seja, as percentagens da variável de resposta para cada valor da variável explicativa.

Neste caso, uma vez que a variável explicativa é gênero, calculamos a porcentagem de motoristas que beberam e não beberam álcool para machos e para fêmeas separadamente. Aqui está a tabela das percentagens condicionais. Para os 619 motoristas da amostra, verificou-se que um percentual maior de homens era embriagado do que as mulheres, 16% versus 11,6%. Nossos dados em outras palavras, fornece algumas evidências de que a condução embriagada está relacionada ao gênero. No entanto, isso por si só não é suficiente para concluir que tal relação existe em uma população maior de motoristas com menos de 20 anos.

Precisamos investigar mais os dados e decidir entre os dois pontos de vista a seguir. Que não há diferença na taxa de condução embriagada entre homens e mulheres com menos de 20 anos, nossa hipótese nula. Ou que há uma diferença na taxa de condução embriagada entre homens e mulheres com menos de 20 anos, nossa hipótese alternativa. Em outras palavras, é a evidência fornecida pela pesquisa na estrada, 16% versus 11,6%, forte o suficiente para concluir além de uma dúvida razoável que deve ser devido a uma relação entre dirigir bêbado e gênero na população de motoristas menores de 20 anos. Ou a evidência fornecida pelo inquérito à beira da estrada não é suficientemente forte para chegar a essa conclusão? E isso poderia ter acontecido por acaso?

Isso se deve à variabilidade da amostragem e não necessariamente porque existe uma relação na população. Estas são as hipóteses alternativas nulas e para o teste de independência qui-quadrado. Aqui estão outras maneiras que a hipótese nula e alternativa pode ser declarada para um teste qui-quadrado de independência. Não há relação entre as duas variáveis categóricas. Eles são independentes. Ou, há uma relação entre as duas variáveis categóricas. Eles não são independentes.

Algebricamente, a independência entre gênero e dirigir bêbado equivale a ter proporções iguais de quem bebeu ou não bebeu para homens versus mulheres. Na verdade, a hipótese nula e alternativa poderia ser reformulada, já que a proporção de motoristas homens bêbados é igual à proporção de motoristas mulheres bêbadas. Ou a proporção de motoristas bêbados masculinos não é igual à proporção de mulheres motoristas bêbadas.

A ideia por trás do teste de independência qui-quadrado, muito parecido com a análise da variância é medir o quão longe os dados estão do que é reivindicado na hipótese nula. Quanto mais longe os dados estiverem da hipótese nula, mais evidências os dados apresentam contra ela. Aqui, os dados de gênero e condução embriagada são representados pelas contagens observadas. Para representar a hipótese nula, vamos calcular outro conjunto de contagens. As contagens que esperaríamos ver, em vez das observadas.

Se dirigir bêbado e sexo eram realmente independentes. Ou seja, se a hipótese nula fosse verdadeira. Por exemplo, nós realmente observamos 77 homens que dirigiam bêbados. Se dirigir bêbado e sexo fossem realmente independentes, se a hipótese nula fosse verdadeira, quantos motoristas bêbados do sexo masculino esperaríamos ver em vez de 77?

Também faremos o mesmo tipo de pergunta sobre as outras três células em nossa tabela. Se a hipótese nula fosse verdadeira, quantas motoristas bêbadas esperaríamos ver em vez de 16? Quantos não bêbados dirigindo machos esperaríamos ver em vez de 404? Quantas mulheres dirigindo não bêbadas esperaríamos ver em vez de 122?

Em outras palavras, teremos dois conjuntos de contagens. As contagens observadas, que são os dados. E as Contagens Esperadas, se a hipótese nula fosse verdadeira. Vamos medir o quão longe estão as contagens observadas das esperadas. Basearemos nossa decisão no tamanho da discrepância entre o que observamos e o que esperaríamos observar, se a hipótese nula fosse verdadeira. Como as contagens esperadas foram calculadas?

Se os eventos A e B forem independentes, a probabilidade de A e B é igual à probabilidade de A vezes a probabilidade de B. Usamos esta regra para calcular contagens esperadas uma célula de cada vez. Aplicando a regra à primeira célula superior esquerda. Se dirigir bêbado e gênero são independentes, então a probabilidade de um homem ter bebido é igual à probabilidade de ser bêbado vezes a probabilidade de ser homem. Ao dividir as contagens em nossa tabela, vemos que a probabilidade de ter bebido é igual a 93 dividido por 619. E a probabilidade de ser homem é 481 dividida por 619. Então a probabilidade de estar bêbado e ser homem é 93 dividido por 619 vezes 481 dividido por 619. Portanto, uma vez que há um total de 619 motoristas. Se a

condução bêbada e o sexo fossem independentes, a contagem de motoristas bêbados do sexo masculino que esperaríamos ver são os seguintes.

$$P(A \text{ AND } B) = P(A) * P(B)$$

$$P(\text{DRUNK AND MALE}) = P(\text{DRUNK}) * P(\text{MALE})$$

$$P(\text{DRUNK}) = 93/619$$

$$P(\text{MALE}) = 481/619$$

$$P(\text{DRUNK AND MALE}) = (93/619) * (481/619)$$

Portanto, a fórmula para calcular Contagens Esperadas é Total da Coluna vezes Total da Linha dividido pelo Total da Tabela. Seguindo esta fórmula, aqui estão as tabelas completas de Contagens Esperadas e Observadas.

Drank Alcohol in the Last 2 Hours			
Gender (x)	Yes	No	Total
Male	77 / 72.3	404 / 408.7	481
Female	16 / 20.7	122 / 117.3	138
Total	93	526	619

Importante, o único número que resume a diferença geral entre Observadas e Contagens Esperadas é a estatística qui-quadrado denotada como chi ou X^2. O que nos diz de forma padronizada, o quão longe o que observamos, que é os dados são. Do que esperaríamos observar, se a hipótese nula fosse verdadeira. Aqui está a fórmula.

$$\chi 2 = \sum_{\text{all cells}} \frac{(\textit{Observed Count} - \textit{Expected Count})^2}{\text{Expected Count}}$$

Para cada célula, tomamos a Contagem Observada, subtraímos a Contagem Esperada e elevamos ao quadrado esse valor. Este valor é dividido pela contagem esperada e, em seguida, este número é somado para todas as células na tabela. Uma vez que a estatística qui-quadrado tenha sido calculada, podemos ter uma sensação de seu tamanho. No nosso caso, o valor de $\boldsymbol{X^2 = 1{,}62}$. Existe uma diferença relativamente grande entre o que observamos e o que a hipótese nula afirma? Ou relativamente pequeno? Acontece que para dois casos como o nosso, estamos inclinados a chamar a estatística qui-quadrado grande se for maior que 3,84. Portanto, nossa estatística de teste não é grande, indicando que os dados não são diferentes o suficiente da hipótese nula para nós rejeitá-lo. Para casos diferentes de dois por dois, há cortes diferentes para o que é considerado grande, que são determinados pela distribuição nula nesse caso. Assim, vamos confiar apenas no valor p para conclusões.

Mesmo que não possamos realmente usar a estatística qui-quadrado, foi importante aprender sobre isso, já que engloba a ideia por trás do teste. O valor de p para o teste de independência do qui-quadrado é a probabilidade de obter contagens como as observadas, assumindo que as duas variáveis não estão relacionadas. Que é o que é reivindicado pela hipótese nula. Quanto menor o valor p, mais surpreendente seria obter contagens como fizemos, se a hipótese nula fosse verdadeira. Tecnicamente, o valor p é a probabilidade de observar um qui-quadrado pelo menos tão grande quanto o observado. Usando nosso software estatístico, descobriremos que o valor de p para este teste é 0,201. O valor de p de 0,201 não é pequeno.

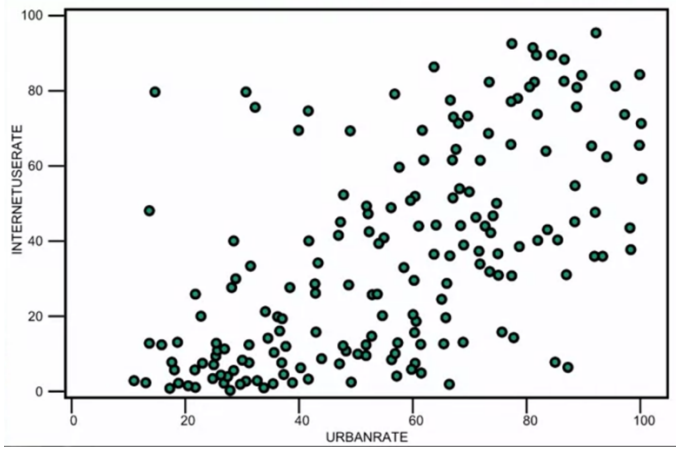
Não há evidência estatística convincente para rejeitar a hipótese nula. E assim continuaremos a assumir que pode ser verdade. Gênero e condução embriagada podem ser independentes. E assim os dados sugerem que

uma lei que proíbe a venda de 3,2% de cerveja a homens e permite às mulheres é injustificada. Na verdade, o Supremo Tribunal, por um voto de sete a dois maioria derrubou a Lei de Oklahoma como discriminatória e injustificada.

6.2.7. Teste de Correlação de Pearson

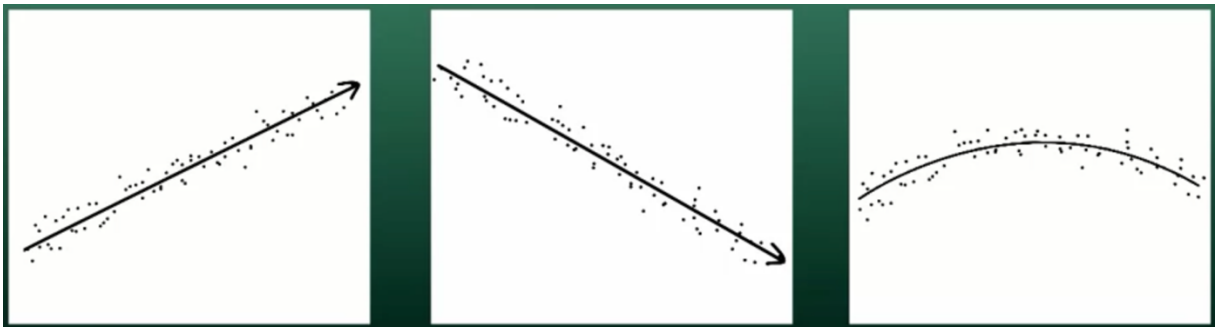
Análise de Variância examina a relação entre uma variável explicativa categórica e uma variável de resposta quantitativa, na qual analisamos a primeira ferramenta inferencial. O teste de independência Qui Quadrado é uma ferramenta inferencial que examina a relação entre dois valores categóricos. Se você tem uma variável explicativa quantitativa e uma variável de resposta categórica, para o propósito deste curso eu encorajo você a categorizar a variável explicativa quantitativa e usar este teste de independência de qui quadrado para examinar este tipo de associação.

A próxima ferramenta inferencial que vamos olhar é usada para examinar a associação entre duas variáveis quantitativas. A Correlação de Pearson. Já discutimos anteriormente que um gráfico de dispersão é a maneira apropriada de gráfico ou visualizar duas variáveis quantitativas quando você deseja examinar a relação entre elas. Vamos primeiro rever brevemente Scatterplots e como interpretá-los. Para criar um gráfico de dispersão, cada par de valores é plotado de modo que que o valor da variável explicativa x, seja plotado no eixo horizontal e o valor da variável de resposta y, seja plotado no eixo vertical. Em outras palavras, cada indivíduo aparece no gráfico de dispersão como um único ponto cuja coordenada x é o valor da variável explicativa para esse indivíduo, e cuja coordenada y é o valor da variável de resposta. Ao descrever o padrão geral do relacionamento, vamos olhar para sua direção, forma e força.

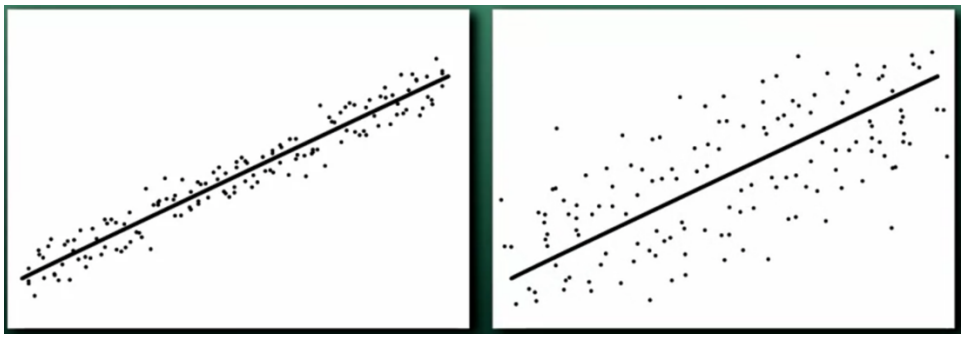


A direção do relacionamento pode ser positiva, negativa ou nenhuma delas. Um relacionamento positivo, ou crescente, significa que um aumento em uma das variáveis está associado a um aumento na outra. Um negativo, ou diminuição no relacionamento significa que um aumento em uma das variáveis está associado a uma diminuição na outra. Nem todos os relacionamentos podem ser classificados como positivos ou negativos. A forma da relação é a sua forma geral. Ao identificar o formulário, tentamos encontrar a maneira mais simples de descrever a forma do gráfico de dispersão. Existem muitas formas possíveis. Aqui estão alguns que são bastante comuns.

Relacionamentos com uma forma linear são mais simplesmente descritos como pontos espalhados sobre uma linha. Relacionamentos com uma forma curvilínea são mais simplesmente descritos como pontos dispersos em torno da mesma linha curva. Por definição, o coeficiente de correlação mede uma relação linear entre duas variáveis quantitativas. Portanto, neste momento, não nos preocuparemos com curvilíneos ou quaisquer outras formas possíveis que um gráfico de dispersão possa tomar. A força do relacionamento é determinada pela proximidade com que os dados seguem a forma do relacionamento.



Esses dois gráficos de dispersão abaixo exibem relações lineares positivas. A força do relacionamento é determinada pela proximidade com que os pontos de dados seguem o formulário. Pontos de dados no gráfico de dispersão à esquerda seguem o padrão linear bastante de perto. Este é um exemplo de uma relação forte. Pontos de dados no gráfico de dispersão à direita também seguem o padrão linear, mas muito menos de perto. Portanto, podemos dizer que o relacionamento é mais fraco em geral. Embora avaliar a força de um relacionamento apenas olhando para o gráfico de dispersão é bastante problemático. Precisamos de uma medida numérica para nos ajudar com isso.



A medida numérica que mede a força de uma relação linear entre duas variáveis quantitativas é chamada de coeficiente de correlação. E é denotado por um **r** minúsculo. O valor de **r** varia de -1 a +1. Não surpreendentemente valores negativos de **r** indicam uma direção negativa para uma relação linear entre as duas variáveis. E valores positivos indicam uma direção positiva para a relação linear. Valores próximos a 0, sejam negativos ou positivos. Indique uma relação linear fraca. E valores próximos a -1 ou próximos a +1 indicam uma forte relação linear. Negativo ou positivo.

Material complementar

Blog: Mehta, A. (2019). Descriptive Vs Inferential Statistics: Which Is Better & Why. *(7 min)*
https://www.digitalvidya.com/blog/descriptive-vs-inferential-statistics/

Article: Laerd Statistics. (n.d.). Descriptive and Inferential Statistics. *(5 min)*
https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php

Video: The Organic Chemistry Tutor. (2019). Descriptive Statistics vs Inferential Statistics. *(7 min)*
https://www.youtube.com/watch?v=VHYOuWu9jQI

Exercício

1. Os dados _____ incluem dados bem definidos com padrões facilmente identificáveis.

a) Não estruturado
b) Estruturada
c) Semi-estruturado

2. Qual das opções a seguir é um exemplo de dados não estruturados?

a) Arquivos de imagem
b) Informações da transação
c) Números de segurança social

3. Qual dos seguintes é considerado dados quantitativos?

a) Nominal
b) Discreto
c) Ordinal

4. Qual dos seguintes é considerado dados qualitativos?

a) Contínuo
b) Discreto
c) Ordinal

5. A preparação de dados inclui todas as etapas a seguir, exceto:

a) Extrair
b) Transformar
c) Carregar
d) Todas essas opções são etapas adequadas na preparação de dados

6. As estatísticas _____ permitem a sumarização e a representação gráfica de um conjunto de dados.

a) Descritivo
b) Inferencial
c) Regressão

7. _____ é usado para explicar as médias de um ponto de dados.

a) distorção
b) Correlação
c) Tendência Central

8. A estatística descritiva permite que um analista generalize os resultados de uma amostra para uma população inteira.

a) Verdadeiro
b) Falso

9. O objetivo da estatística _____ é inferir e generalizar conclusões de uma amostra para uma população inteira.

a) Descritivo
b) Regressão
c) Inferencial

10. Qual dos seguintes é usado para fazer previsões com base em valores dentro de um conjunto de dados de amostra?

a) Teste de hipóteses
b) Correlação
c) Modelos de regressão

7. Business Intelligence e Visual Analytics

7.1. Visualização e Análise de Dados

A necessidade de visualização de dados para relatórios de negócios

Existe um ditado bem estabelecido: "Uma imagem vale mais que mil palavras". Agora, imagine que você está percorrendo milhares de linhas de dados tabulares para coletar informações pertinentes aos negócios para tomar uma decisão. Esta tarefa cansativa pode levar horas! E então o que você faz quando os dados são atualizados? Recomeçar? Este é um exemplo de porque a visualização de dados pode ser tão importante e impactante.

A visualização de dados fornece uma imagem que descreve os dados, permitindo que você tome decisões mais rápidas e precisas. Padrões claros geralmente surgem e podem ser reconhecidos mais facilmente por meio da visualização de dados, e muitas vezes você pode reter e explicar melhor a saída por meio de uma visualização pictórica das informações. A visualização de dados permite criar uma representação visual (ou imagem) de informações para um conjunto de dados ou coleção de fontes de dados. Esse processo traz clareza, envolvimento do usuário, insights eficazes e tomada de decisão informada aos dados. Muitas ferramentas de inteligência de negócios existem hoje para aprimorar e permitir a criação eficiente de visualizações de dados.

Tipos de visualização de dados

A visualização eficaz de dados é tanto "arte" quanto "ciência". Existem inúmeras visualizações que são usadas para representar dados. Um dos principais desafios é selecionar o tipo adequado de visualização para comunicar efetivamente a história que está sendo contada por meio dos dados.

As visualizações de dados geralmente podem ser categorizadas em sete tipos:

- **Linear** (1-dimensional): listas de itens
- **Planar** (2-dimensional): mapas geoespaciais
- **Volumétrico** (3-dimensional): renderizações de superfície e volume, simulações de computador, modelos 3D, etc.
- **Temporais**: linhas do tempo, gráficos de séries temporais, gráficos de Gantt, etc.
- **Multidimensional**: gráficos de pizza, histogramas, nuvens de tags, gráficos de barras, gráficos de linhas, gráficos de dispersão, etc.
- **Hierárquico**: árvores
- **Rede**: matrizes, diagramas nó-link, etc.

Cada tipo de visualização de dados tem sua finalidade exclusiva e caso de melhor uso. Algumas categorias, como temporal e multidimensional, são muito mais comumente encontradas em visualizações de dados, painéis e infográficos hoje. Outros são bastante complexos e normalmente usados em domínios altamente científicos.

7.2. Qual visualização é boa para que propósito?

Com tantos tipos de visualizações disponíveis, como você decide qual visualização é uma representação eficaz para seu propósito específico? Embora existam ferramentas disponíveis para ajudar a orientar sua seleção, parte da decisão dependerá de sua experiência, dos requisitos do trabalho e até de tentativa e erro. É aí que entra a "arte" da visualização de dados. É crucial selecionar uma visualização que contribua para a história dos dados, seja rapidamente compreendida pelo público e mostre uma imagem clara e precisa dos dados.

Uma ferramenta eficaz que ajuda a determinar o tipo de gráfico, gráfico ou visualização a ser exibido é o Catálogo de Visualização de Dados[6]. Esta ferramenta organiza visualizações por várias funções, por exemplo,

[6] Catálogo de Visualização de Dados - https://datavizcatalogue.com/search.html

comparação, relacionamento, distribuição, dados ao longo do tempo, etc. Por exemplo, se você precisa visualizar diferenças ou semelhanças entre valores em um conjunto de dados, você pode selecionar a função Comparações na Visualização de Dados Catálogo. Fazê-lo apresenta dois grupos de visualizações: "Com eixo" ou "Sem eixo". Se você deseja visualizar dados quantitativos em um período de tempo, pode selecionar a opção Gráfico de linhas na categoria "Com um eixo". Uma vez selecionada, a ferramenta fornece informações descritivas sobre o gráfico/tipo de gráfico selecionado, bem como seleções de gráficos adicionais que podem ser adequadas para seus dados.

Visão geral de Análise Visual

A análise visual emprega visualizações de dados para apoiar o raciocínio analítico e o desenvolvimento de ferramentas e processos para analisar conjuntos de dados. A análise visual geralmente produz padrões e insights que podem não surgir tão facilmente por outros meios analíticos. As visualizações de dados geralmente respondem a perguntas sobre "o quê", enquanto a análise visual mergulha no "porquê" mais profundo da exploração de dados. Essa abordagem se presta ao aprendizado profundo sobre o conjunto de dados e à compreensão dos padrões emergentes, anomalias e relacionamentos intrincados entre os pontos de dados. A análise visual agrega valor ao permitir que o usuário altere parâmetros rapidamente, explore visualizações de dados para explorar por que um gráfico se parece com ele ou forneça visualizações alternativas de visualizações de dados com o mínimo de esforço. A análise visual é poderosa por causa de sua flexibilidade, capacidade de atualizações em tempo real e eficiência na exploração de dados, o que permite descobrir padrões inesperados que impulsionam os "porquês" por trás dos dados.

O cenário das ferramentas de análise visual

Muitas ferramentas de análise visual existem hoje no mercado, e a demanda por essas ferramentas está aumentando exponencialmente à medida que as organizações descobrem a necessidade de explorar e aprender profundamente com os dados que coletam há muitos anos. Ferramentas poderosas que anteriormente exigiam investimentos significativos em hardware, redes e infraestrutura de TI agora estão disponíveis por meio de soluções baseadas em nuvem, navegadores da Web e dispositivos móveis. Toda essa inovação ajuda a aliviar o fardo da análise, colocando o poder da análise visual nas mãos de cada usuário e levando a soluções mais fáceis de usar e econômicas.

Uma das plataformas de análise visual mais populares e poderosas do mercado atualmente é o SAS Visual Analytics e o SAS Viya[7]. Essa solução baseada em nuvem fornece análises visuais ao usuário, ajudando-o a criar insights poderosos sobre os dados, recursos de relatório de dados e ferramentas de exploração de dados. Passaremos o restante deste curso nos familiarizando com essa plataforma e ganhando experiência prática no desenvolvimento de soluções analíticas.

Atividade: Seleção de visualização do conjunto de dados

Usando uma ferramenta como o kaggle, selecione um conjunto de dados de seu interesse e descreva uma visualização eficaz que derivaria valor e insights com base nos pontos de dados disponíveis. Certifique-se de aplicar as leituras do módulo à sua avaliação e escolha de seleção.[8]

Exercício

1. _____ fornece uma imagem que descreve os dados, permitindo que você tome decisões mais rápidas e precisas.

a) Data Analysis
b) Data Visualization
c) Statistics

2. Qual das opções a seguir é um benefício da visualização de dados?

a) Insights eficazes
b) Tomada de decisão informada

[7] SAS Visual Analytics and SAS Viya - https://www.sas.com/en_us/software/visual-analytics.html
[8] Kaggle - https://kaggle.com/datasets

c) Todas essas opções estão corretas

3. Um dos principais desafios da visualização de dados é selecionar o tipo adequado de visualização para comunicar efetivamente _____.

a) Histórias
b) Finanças
c) Erros nos dados

4. Qual tipo de visualização de dados inclui histogramas e gráficos de dispersão?

a) Multidimensional
b) Planar
c) Temporal

5. A seleção da visualização mais eficaz depende significativamente da experiência, requisitos do trabalho e tentativa e erro.

a) Falso
b) Verdadeiro

6. A seleção de uma visualização que contribui para a história dos dados diminui a probabilidade de o público entender rapidamente os dados.

a) Verdadeiro
b) Falso

7. _____ emprega visualizações de dados para apoiar o raciocínio analítico e o desenvolvimento de ferramentas e processos para analisar conjuntos de dados.

a) Visual Analytics
b) Business Intelligence
c) Data Visualization

8. Um objetivo principal da análise visual é descobrir _____ e intrincadas relações entre os pontos de dados.

a) Falso-positivo
b) Padrões emergentes
c) Dados incorretos

9. Muitas ferramentas de análise visual estão agora disponíveis como soluções baseadas em nuvem para aumentar a disponibilidade e oferecer poder adicional aos usuários finais.

a) Verdadeiro
b) Falso

10. O SAS Viya é uma plataforma em nuvem que fornece _____ aos usuários, ajudando-os a criar insights poderosos sobre os dados, recursos de relatórios de dados e ferramentas de exploração de dados.

a) Visual Analytics
b) Data visualizations
c) Data Warehousing

Material complementar

Book: SAS Institute. (2019). Exploring SAS Viya: Visual Analytics, Statistics, and Investigations. (1 hour) << https://support.sas.com/content/dam/SAS/support/en/books/free-books/exploring-sas-viya-va-statistics-investigations.pdf >>