



**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

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	4
1.1.	A Sociedade do Conhecimento	4
1.2.	A Importância da Gestão do Conhecimento	6
1.3.	Dado, Informação e Conhecimento.....	7
1.4.	Conhecimento Tácito e Conhecimento Explícito	8
1.5.	Os Três Pilares da Gestão do Conhecimento	8
1.6.	Capital Intelectual	11
1.6.1.	A criação do Conhecimento na Empresa.....	12
1.6.2.	Disseminação do Conhecimento	13
1.6.3.	Marketing do Conhecimento	14
1.7.	Memória Organizacional.....	17
1.8.	Culturas e valores organizacionais.....	19
1.8.1.	Gestão de Conteúdo.....	21
1.8.2.	A Gestão de Conteúdos no Contexto da Gestão do Conhecimento	22
1.8.3.	Comunidades de Práticas Compartilhadas	23
1.8.4.	Portais Corporativos	26
2.	A Natureza dos dados e o Projeto de Banco de Dados Relacionais	29
2.1.	Business Intelligence, Business Analytics e Data Science	29
2.2.	OLTP versus OLAP	30
2.3.	Data Warehousing para BI	31
2.4.	Definindo Bancos de Dados Relacionais	33
2.4.1.	Diagrama Entidade-Relacionamento (ERD)	34
2.4.2.	Normalização e Desnormalização	36
3.	Data Warehousing e Business Intelligence	39
3.1.	Necessidade de armazenamento de dados.....	39
3.1.1.	Arquiteturas de armazenamento de dados.....	39
3.1.2.	Extração, transformação e carga (ETL).....	40
3.1.3.	Data Marts	40
3.1.4.	Armazenamentos de dados operacionais.....	40
3.1.5.	Armazenamento de dados na nuvem.....	41
3.2.	Modelagem de dados para Data Warehouse	41
3.2.1.	Modelagem de dados multidimensionais	42
3.2.2.	NoSQL, Big Data, Data Lakes e Data Warehousing.....	43
3.3.	O Processo de Preparação de Dados	43
3.4.	Representação do Cubo de Dados	44
3.4.1.	Operações com o Cubo de Dados.....	46
3.5.	Metodologias de Projeto de Data Warehouse	48
3.6.	Integração de dados	50
3.6.1.	Mudança no Conceito de Dados.....	52
3.6.2.	Atividades de Limpeza de Dados	53
3.6.3.	Identificação de Padrões com Expressões Regulares	54
3.6.4.	Correspondência e Consolidação	57
3.6.5.	<i>Quasi</i> -Identificadores e Funções de Distância para Correspondência de Entidades.....	59
3.7.	Pentaho Data Integration - PDI	61
4.	A Natureza dos Dados	65
4.1.	Análise de dados	65
4.2.	Dados e Tipos de Dados.....	66
4.3.	Datasets e Codebooks	68
4.4.	Desenvolvendo uma questão de pesquisa	69
5.	Estatística	71
5.1.	Estatística descritiva.....	71
5.1.1.	Análise Exploratória de Dados.....	71

5.1.2.	Examinando a distribuição de frequência.....	72
5.1.3.	Plotando as distribuições	72
5.1.4.	Medidas de Centralidade e Dispersão	77
5.2.	Estatística inferencial	81
5.2.1.	Da amostra à população.....	82
5.2.2.	Teste de Hipótese	84
5.2.3.	Valor-p e Intervalo de Confiança	86
5.2.4.	Escolhendo testes estatísticos	87
5.2.5.	Análise de Variância - ANOVA.....	88
5.2.6.	Teste de Independência Qui-Quadrado	93
5.2.7.	Teste de Correlação de Pearson.....	96
6.	Business Intelligence e Visual Analytics.....	99
6.1.	Visualização e Análise de Dados	99
6.2.	Qual visualização é boa para que propósito?	99
6.3.	Uma visão geral do Public Tableau ou Power BI???	101

1. Gestão do Conhecimento

1.1. A Sociedade do Conhecimento

As últimas duas décadas do século XX e os anos seguintes experimentaram uma transformação em nível mundial em muitos setores da sociedade. Avanços científicos e tecnológicos foram obtidos, os meios de transporte e de comunicação evoluíram e a consequente diminuição das distâncias consolidou o fenômeno da globalização. Hoje, é possível conversar com um amigo, seja ele um vizinho, seja ele um morador do hemisfério oposto. Os efeitos de uma crise econômica na Rússia sensibilizam os cinco continentes. Os produtos são vendidos em uma cidade e entregues em outra.

As mudanças ocorridas no período são facilmente percebidas, entretanto, a principal modificação do panorama social, cultural, político e econômico na virada do século XXI relaciona-se à importância e à dimensão que o conhecimento atingiu, fato que é causa e consequência do desenvolvimento da tecnologia da informação, formando uma espécie de círculo vicioso.

Por um lado, a flexibilidade e o poderio das novas tecnologias da informação (TI) tornam o conhecimento uma das prioridades estratégicas da organização. Os produtos das novas indústrias de TI são dispositivos de processamento de informação ou o próprio processamento. O avanço da TI, assim, permite que a informação se transforme no centro do processo produtivo .

A globalização também gerou um novo contexto político e econômico, em que as organizações competem, cooperam e se comunicam em escala mundial, amplificando a concorrência e as possibilidades além dos limites geográficos. A inovação contínua tornou-se um aspecto vital para o sucesso e a manutenção da empresa. Logo, o maior desafio das organizações é adquirir a competência necessária para transformar a informação e o conhecimento em um recurso econômico estratégico, através de inovações que devem ser apresentadas em intervalos de tempo cada vez menores.

O círculo vicioso se fecha quando o conhecimento, centro estratégico das organizações, justifica o aprimoramento das tecnologias. As redes funcionam como canais de transmissão de informações. Os limites de volume dos sistemas de armazenamento aumentam rapidamente. Os bancos de dados incorporam novas funcionalidades e adquirem maior robustez e segurança ao longo do tempo. A evolução técnica precisa acompanhar a necessidade da geração, manutenção e disseminação do conhecimento organizacional. Não é à toa que atualmente computadores e redes disponibilizam vários recursos, como o correio eletrônico, groupware, intranets e a internet, aptos a indicar pessoas com conhecimento e interligar indivíduos que precisem compartilhar conhecimento à distância.

Há vários indicadores econômicos de que a sociedade, hoje, pode ser classificada como uma “Sociedade do Conhecimento”:

- **A importância da inovação tecnológica para o crescimento econômico e a competitividade empresarial.** Nos anos 90, a inovação tecnológica tem sido responsável por cerca de 70% do crescimento econômico e por entre 80 e 90% dos ganhos de produtividade. No fim da década, o setor de alta tecnologia foi responsável por 35% do crescimento do PIB norte-americano.
- **Queda de preços e da participação na economia dos Recursos Naturais e Agricultura.** Estima-se que os preços dos recursos naturais tenham caído quase 60% entre os anos 70 e os 90 e cerca de 0,6% ao ano desde 1800 até os dias de hoje. Nos países desenvolvidos, a participação da agricultura gira em torno de 1 a 3% do PIB.
- **Evolução dos setores de informática e telecomunicações.** Em 1998, havia mais de 200 milhões de computadores e mais de 15 bilhões de chips instalados em diversos tipos de máquinas e equipamentos. Em 1981, nos EUA, os investimentos em telecomunicações e informática correspondiam à metade do investido em setores tradicionais. Dez anos depois, os valores se igualaram. Em 1997, o panorama inverteu-se: os

investimentos nos setores de informática chegaram a cerca de 225 bilhões de dólares, correspondendo ao dobro dos investimentos nos setores tradicionais.

- **Os impactos econômicos e sociais dos níveis de educação e qualificação profissional.** Em 1900, os operários de produção dos EUA representavam 73,4% dos trabalhadores contra 17,6%, que eram administrativos, técnicos e profissionais liberais, onde se enquadram os funcionários cujo resultado de seu trabalho tem valor segundo seu conteúdo informativo. Em 1980, esta categoria representava 52,1% dos trabalhadores americanos, contra 34,2% que eram operários. Estima-se que em 2006, metade da população americana esteja empregada na indústria da informática ou em setores que utilizam intensamente a informática.

Um outro importante indicador da importância do conhecimento para o desenvolvimento de produtos e sucesso organizacional pode ser percebido através do modelo de vantagem competitiva baseado em cinco “objetivos de desempenho” básicos que orientam o gestor na adoção de ações, contribuindo para o aumento da competitividade da empresa.

Fazendo uma análise histórica da exploração desses objetivos de desempenho, percebemos que há um padrão em torno das grandes ondas de competitividade ao longo do tempo. Em plena era do desenvolvimento industrial, durante a década de 1970, as empresas conseguiam ser competitivas quando focavam sua atenção no custo da produção. Reduzir custos produtivos – mão-de-obra, processos, matéria-prima – significava redução do preço final ao cliente, aumentando suas vantagens sobre a concorrência. Nessa época, popularizaram-se as principais técnicas de gestão de custos e otimização da linha de produção, com controles do processo, gestão de estoques e materiais e fluxos operacionais.

Com o aumento da oferta e a ampla disseminação das técnicas de produção, as empresas concluíram, na década seguinte, que apenas reduzir custos não as tornavam mais competitivas. Os ganhos de custos, propiciados por altos investimentos na busca da “produção enxuta”, já não representavam melhorias significativas no preço final ao consumidor. Quando o cliente tem mais opções com pouca variação de custo, ele busca um diferencial, que geralmente está na qualidade do produto ou do serviço. Nesse período, capitaneadas pelas indústrias japonesas, o “boom” da Qualidade teve seu momento de apogeu, sobretudo na figura de grandes técnicas e metodologias de melhoria da qualidade, a busca da excelência, os controles estatísticos, a política de zero erro, a padronização de processos preconizada por normas como a ISO 9000 e similares, melhoraram não só a eficácia dos processos internos, mas principalmente propiciaram melhoria e confiabilidade aos produtos e serviços de forma perceptível ao consumidor final.

Logo, a Qualidade se tornou uma prática constante e comum em todas as empresas, perdendo então o seu caráter diferenciador para se tornar condição mínima para se ingressar em determinados mercados altamente competitivos. Os consumidores já não se contentavam em ter o melhor preço e a melhor qualidade, o diferencial competitivo exigia outra dimensão que não poderia mais se ater aos aspectos produtivos. O advento da tecnologia, sobretudo de comunicação e de informação, fomentou uma nova possibilidade de exploração de diferencial competitivo: o Tempo. As empresas que buscavam se diferenciar da concorrência passaram a se valer da tecnologia para atrair clientes que, tendo à disposição produtos e serviços com o mesmo padrão de preços e níveis de qualidade aceitáveis, escolhiam aquelas empresas que pudessem entregar, executar, desenvolver e implementar mais rápido. A tecnologia permitiu a redução dos ciclos produtivos, do tempo entre a criação e o lançamento de produtos no mercado, e o ciclo geral entre a produção e o consumo. Essa foi a tônica que movimentou as empresas durante a década de 1990. Muitas empresas puderam reduzir ou eliminar intermediários, permitindo o acesso cada vez mais direto ao consumidor. A informação se tornou, então, forte elemento de competitividade, tanto para melhorar a eficácia interna quanto para promover novos serviços para o cliente.

Com o tempo, o acesso à tecnologia foi se tornando cada vez mais fácil, neutralizando a vantagem do tempo das empresas de ponta. Mas ela não só continuou tendo papel fundamental como elemento de

transformação das organizações, como trouxe uma nova forma de exploração que vivenciou o seu ápice com o advento e o crescimento da Internet que, por sua vez, propiciou o crescimento de empresas, com ou sem base tecnológica, focadas em serviços. Assim, a década de 2000 vem se caracterizando como a década dos serviços. Grandes organizações estão aprendendo a atender às necessidades específicas de seus clientes por meio da prestação de serviços especializados. Esta é a década da Flexibilidade, da capacidade de adequar a organização à realidade mutável e dinâmica do mercado e de usar a tecnologia para conhecer melhor cada cliente ou segmento específico, de forma a atender às suas necessidades particulares e únicas. Com a tecnologia de CRM5, por exemplo, pode-se praticar o chamado marketing one-to-one específicas. Ser competitivo em Flexibilidade ainda significa ser competitivo em tecnologia.

O que esperamos para a próxima década? Partindo do pressuposto que a tecnologia não diferenciará mais as organizações, pois rapidamente ela está se tornando uma commodity, e que o mercado terá suas demandas satisfeitas, será competitiva a empresa que puder se antecipar às necessidades do mercado e, se possível, criar as necessidades futuras de seus consumidores. Para isso, o fundamento básico não está mais na tecnologia, nas finanças ou na produção. A capacidade criativa e inovadora está nas pessoas. A próxima onda, que já começou, encontrará nas pessoas o grande elemento de competitividade. As habilidades de imaginar o que não existe, criar o que ninguém tem, encontrar e explorar nichos de oportunidade, a ousadia em propor novos paradigmas, a determinação em estabelecer novos padrões e a visão para detectar tendências são características inerentemente humanas.

Portanto, as organizações não podem mais esperar que os produtos e práticas que fizeram sucesso no passado possam mantê-las viáveis no futuro. O ciclo de desenvolvimento de produtos e sua introdução no mercado dura cada vez menos tempo. Hoje, atividades baseadas no conhecimento, como o incremento da qualidade, valor, bom atendimento, inovação e velocidade de chegada no mercado, são muito mais importantes que o trabalho de produção em si.

São muitos os sinais de como o conhecimento vem ganhando importância econômica e estratégica no mundo globalizado. Logo, as práticas de Gestão do Conhecimento são fundamentais para o aumento da competitividade e da produtividade das organizações. Entretanto, antes de apresentar alguns conceitos relacionados à Gestão do Conhecimento, convém definir, com mais clareza, o que é conhecimento.

1.2. A Importância da Gestão do Conhecimento

A crescente importância do conhecimento no dia-a-dia das organizações levou as empresas a desenvolverem recursos que facilitem e estimulem a gestão desse insumo. Assim, passou a ser adotado em larga escala o termo Gestão do Conhecimento (GC). Embora sua função seja intuitiva, isto é, administrar o conhecimento da organização, na prática o conjunto de atividades que compreende a GC é tão complexa que se encontram, na literatura, diversas definições para a expressão.

Conforme relatado, um chefe da área de Inteligência Artificial de uma instituição norte-americana comentou através de e-mail escrito em 2001, a forma pela qual seu grupo começou a levar em conta a questão do conhecimento:

Começamos a pensar em termos de criação, aprendizado, compartilhamento (transferência), e o uso ou a alavancagem do conhecimento como um conjunto de processos e dinâmicas sociais que precisava ser administrado [...]. Depois, vimos que não poderíamos descrever tudo aquilo de uma maneira melhor que Gestão do Conhecimento (SVEIBY, 2001, p.5).

Desta forma, pode-se dizer que são atividades básicas de GC a identificação, aquisição, desenvolvimento, disseminação, utilização e preservação do conhecimento.

Das definições existentes na literatura sobre GC, todas abarcam, sempre, a disseminação do conhecimento por toda a organização, o que pressupõe o envolvimento dos processos de captura e de armazenamento. Assim, uma definição satisfatória para a GC, que reflete bem o foco deste trabalho, é esta:

Gestão do Conhecimento é a disciplina que administra o conjunto de procedimentos, práticas e ferramentas que visam capturar, armazenar e disseminar o conhecimento dentre os funcionários da organização, aproveitando os recursos tecnológicos existentes.

A disseminação do conhecimento deve ocorrer por entre os diferentes níveis hierárquicos da organização. Este conhecimento que permeia a organização encontra-se sob duas formas distintas: conhecimento tácito e conhecimento explícito.

1.3. Dado, Informação e Conhecimento

Em conversas informais, é muito comum usar expressões, como dado, informação e sabedoria, como sinônimos de conhecimento. No entanto, vários autores distinguem o significado de cada um desses termos.

De todos os termos com que trabalham, Davenport e Prusak (1998) afirmam que conhecimento pode incluir os conceitos de sabedoria e insight, mas diferencia-se de dados e informações. Dados são um conjunto de fatos distintos e objetivos, relativos a eventos. O exemplo dos autores baseia-se no abastecimento de um automóvel feito por um consumidor. A data da compra, os litros de gasolina consumidos e o valor pago são dados relativos à venda.

As informações são dados dotados de relevância e propósito, capazes de exercer algum impacto sobre o julgamento ou comportamento de quem as recebe. Em resumo, dados tornam-se informações quando o seu criador lhes acrescenta significado. No exemplo anterior, uma tabela com a quantidade de litros de gasolina vendidas num ano, dividida mês a mês, informa qual o período em que o consumo de combustível foi mais alto e aquele em que foi mais baixo.

Já o conhecimento,

é uma mistura fluida de experiência condensada, valores, informação contextual e insight experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações. Ele tem origem e é aplicado na mente dos conhecedores. Nas organizações, ele costuma estar embutido não só em documentos ou repositórios, mas também em rotinas, processos, práticas e normas organizacionais (DAVENPORT; PRUSAK, 1998, p.5).

O conhecimento é uma informação transformada através de quatro processos:

- **Comparação:** de que forma as informações relativas a esta situação se compararam a outras situações conhecidas?
- **Consequências:** que implicações estas informações trazem para as decisões e tomadas de ação?
- **Conexões:** quais as relações deste novo conhecimento com o conhecimento já acumulado?
- **Conversação:** o que as outras pessoas pensam desta informação?

Os conceitos de dados, informação e conhecimento estão em uma espécie de “escala do saber”: o dado seria o elemento unitário; a informação é um dado com relevância e propósito; e o conhecimento é um juízo de valor formado em conjunto pela informação e pela experiência do indivíduo.

Dados informação e conhecimento também podem ser visualizados em uma “hierarquia do entendimento”, cuja ordem é semelhante àquela definida pela “escala do saber”, mas com a presença de uma nova categoria, a sabedoria, acima do conhecimento. Essa hierarquia foi desenvolvida depois que foram analisadas outras

conceituações de conhecimento, também muito similares entre si, principalmente no que tange à diferenciação entre dados, informação e conhecimento.

Sob a análise da sabedoria envolvida, dados são símbolos sem significado e sem contexto; informações são dados com significado estabelecido por conexões relacionais entre si; o conhecimento é um conjunto de informações organizado sob padrões de significado, definidos por processos repetidos, sugerindo previsibilidade e, por isso, pode fornecer respostas a perguntas sobre como realizar determinados procedimentos; por fim, a sabedoria é representada pelos princípios fundamentais aos quais estão enraizados os padrões que definem o conhecimento, exemplifica-se pelos valores, moral e insights do indivíduo, envolve a criatividade e o poder inovativo e difere-se dos outros elementos da “hierarquia do entendimento” por ser mais eficaz, ou seja, ao ser usada, subentende-se que a decisão tomada é a melhor coisa a se fazer, enquanto a utilização de dados, informações e conhecimento implica apenas uma ajuda à decisão correta.

A escala da “hierarquia do entendimento” pode ser analisada sob três aspectos:

- **Percepção** e discernimento: quanto maior o nível na hierarquia, mais estruturado é o entendimento necessário para reconhecê-lo.
- **Contexto**: quanto maior o nível na hierarquia, menor a independência do contexto para o entendimento (dados descontextualizados não têm o menor significado ou importância).
- **Detecção de ruído**: quanto maior o nível na hierarquia, mais difícil é a detecção de ruído (dados são muito menos complexos que informações, conhecimento e sabedoria, daí a facilidade de se corrigirem erros e falhas).

1.4. Conhecimento Tácito e Conhecimento Explícito

O conhecimento explícito é fácil de se articular e de se expressar em termos claros. Pode ser representado através de documentos, textos, relatórios, tabelas, programas de computador e, portanto, é de simples compartilhamento.

Em contrapartida, o conhecimento tácito consiste de habilidades técnicas: o tipo de destreza informal e de difícil especificação, incorporada ao termo know-how e também abrange modelos mentais, crenças e perspectivas tão arraigadas que são tidas como algo certo, não sujeitas a fácil manifestação. Refere-se à experiência individual de se lidar com certas situações e é difícil de ser compartilhado. É aprendido por períodos extensos de experiências e cumprimento de tarefas, durante os quais o indivíduo desenvolve um sentido e uma capacidade de fazer julgamentos intuitivos.

O conhecimento tácito corresponde claramente à “mistura fluida de experiência condensada, valores, informação contextual e insight experimentado”, enquanto o explícito refere-se aos “documentos”, “repositórios” e, no caso de instituições, “rotinas, processos, práticas e normas organizacionais”.

1.5. Os Três Pilares da Gestão do Conhecimento

Gestão do conhecimento tem três pilares, ou como costumo falar três C's que compreendem Consultar, Compartilhar e Colaborar. Esses três pilares atuam de maneira transversal, exigindo a atuação em três dimensões: Ferramentas (ou mecanismos), Cultura e Capital Humano.

Informação é um bem dinâmico que possui um valor associado. Toda informação possui um ciclo de vida desde o instante em que foi gerada, passando por sua organização, armazenamento, distribuição e utilização, até o instante no qual, eventualmente, perde seu valor e pode ser descartada, quando então se finaliza o ciclo. Um fator crítico para o sucesso de empresas é sua habilidade de manipular e utilizar todo artefato de informação disponível. De acordo com pesquisa do Gartner Group, “Enterprise Content Management (ECM) will be one of the key application software areas during the next five years” [Austin 2005]. Além disso, há uma tendência das

empresas dotarem o ambiente de trabalho de elevado desempenho, i.e. High-Performance Workplace (HPC), permitindo os profissionais de informação (PI) explorarem dados, desenvolverem processos e produtos inovadores, e atenderem a solicitações e demandas de clientes e fornecedores de modo eficiente. Este tipo de solução possibilita os PI's localizarem de maneira efetiva conteúdo, artefatos e pessoas, bem como disporem de mecanismos de comunicação e colaboração efetivos. Aliado a isto está a necessidade de incorporar mecanismos de integração de aplicações às implementações de gestão de conhecimento. Nesse sentido, as funcionalidades da gestão do conhecimento ou KM (Knowledge Management) podem ser providas por meio de web services numa arquitetura orientada a serviços.

Cabe destacar que um diferencial de gestão é alcançado quando os gestores de uma empresa dispõem de mecanismos de acesso a qualquer artefato de informação de maneira contínua e customizada num curto intervalo de tempo, assegurando o uso efetivo de informações pertinentes a web organizacional (sistemas de informação na intranet da organização) e web global. Além disso, a instituição pode prover diferentes níveis de acesso e visibilidade às informações, dependendo das necessidades do usuário e em conformidade com a hierarquia de acesso a informação da instituição.

Note que a capacidade de compartilhar o entendimento ou consciência, criar conhecimento promovendo a aprendizagem organizacional, e prover suporte à colaboração permite transformar informação em vantagem operacional para empresa num mercado competitivo. Nesse sentido, há uma constante preocupação em transformar dados em informação e conhecimento de modo a promover um entendimento ou consciência geral de uma instituição, bem como disponibilizar o resultado deste processo aos gestores por meio, por exemplo, de um portal corporativo, permitindo a gestão de toda informação organizacional.

Vale ressaltar que a administração de uma instituição pode fazer uso da tecnologia da informação no suporte a criação e compartilhamento de conhecimento, possibilitando tomada de decisão de forma eficiente e segura. Esta necessidade tem se tornado num desafio devido ao crescimento contínuo do volume de artefatos de informação. Além disso, grande parte das empresas atua de forma centrada no conhecimento com os PI's necessitando ter acesso a uma ampla variedade de conteúdo. Essa constante busca por informação se justifica pela demanda por otimização de recursos e agilidade da gestão. Nesse sentido, um ambiente de gestão do conhecimento (que inclui cultura e ferramentas) deve prover suporte às atividades de gestão do conhecimento de maneira sistemática, além da integração de aplicações, permitindo identificar, gerenciar e compartilhar todos os artefatos de informação. Isto inclui bancos de dados, documentos, procedimentos e políticas, bem como qualquer outro conteúdo (código, artefato, etc.).

É importante observar a convergência de gestão de conteúdo, portais e ambientes colaborativos resultantes de um ambiente de trabalho que requer interatividade centrada em tecnologia. Nesse sentido, o conhecimento de uma instituição pode ser encontrado em grandes massas de informações não estruturadas. Considera-se que a informação não estruturada pode ser empregada para observação de eventos (tendências / anomalias) nos dados de uma variedade de aplicações. Aqui, a gestão do conhecimento pode ser utilizada para buscar, organizar e extrair informação de múltiplas fontes. Há uma tendência de unificar os esforços de Business Intelligence (BI) e KM, onde se pode ter análise de dados e texto ocorrendo de maneira indistinta. A ênfase em KM, contudo, leva em conta os dados estruturados, bem como os dados não estruturados que compõem mais de 70% das informações existentes.

Muito além das questões tecnológicas está a importância do dado capturado. É fundamental assegurar a procedência das informações e, por meio de um processo de gestão de pessoas, identificar quem são os colaboradores – internos e/ou externos – mais adequados para responder às demandas da organização. É preciso identificar e valorizar as pessoas que compõem o capital intelectual da organização.

Existem diversas estratégias propostas para a realização da Gestão do Conhecimento, tanto abordagens teóricas como estudos baseados em casos práticos. Serão apresentadas duas formas dada a sua relevância. A

primeira faz a distinção entre estratégias de codificação versus personalização. A segunda, baseada na Gestão do Capital Intelectual, concatena duas abordagens bastante similares e apresenta propostas para o nível de alta gerência. Por fim, apresenta-se uma proposta mais teórica, que sugere a conversão do conhecimento tácito em explícito e vice-versa seguindo um processo em espiral.

Codificação X Personalização

Duas estratégias para a questão de Gestão do Conhecimento (Knowledge Management - KM) podem ser pontuadas segundo estudos de empresas de consultoria:

1. **Codificação** – onde o conhecimento é cuidadosamente codificado e armazenado em bancos de dados onde pode ser acessado e utilizado por qualquer pessoa dentro da empresa; O conhecimento é extraído dos indivíduos que o desenvolveram, representado e reutilizado com vários propósitos. Esta abordagem permite aos indivíduos a busca e a recuperação de conhecimento codificado sem que estes tenham contato com a pessoa que o desenvolveu originalmente. A gestão de conteúdos se enquadra no papel de gerenciar o conhecimento codificado, apoiando todo o processo de armazenamento e recuperação do conhecimento.
2. **Personalização** – o conhecimento está fortemente ligado à pessoa que o desenvolveu e é compartilhado principalmente através de interação person-to- person. O principal propósito da utilização de Tecnologia de Informação neste caso é facilitar a disseminação do conhecimento entre as pessoas, não armazená-lo. O foco é o diálogo entre os indivíduos e não os objetos de conhecimento localizados em repositórios de informações. Esta abordagem também pode lançar mão da gestão de conteúdos com o objetivo de recuperar informações sobre quem faz o quê. Geralmente, nas organizações, estas duas abordagens convivem, mas uma predomina sobre a outra.

No primeiro caso, o conhecimento é codificado utilizando-se a abordagem de people-to-documents, ou seja, o conhecimento é “extraído” da pessoa que o desenvolveu, separado do contexto em que foi criado e reutilizado para outros propósitos. Isto facilita a recuperação do conhecimento codificado sem a necessidade de contato com a pessoa que o criou.

Como exemplo, é citada a empresa de consultoria Ernst & Young, que remove o contexto da informação, desenvolvendo knowledge objects , que utiliza uma abordagem de Gestão do Conhecimento com quatro fases:

1. Geração e armazenando eletronicamente em um repositório compartilhado.
2. Representação
3. Codificação
4. Aplicação

A Accenture define Gestão do Conhecimento como um processo de seis fases:

1. Obter
2. Criar
3. Sintetizar
4. Compartilhar
5. Usar para atingir metas organizacionais
6. Estabelecer um ambiente que estimule o compartilhamento do conhecimento

A Segunda estratégia proposta foca no diálogo entre indivíduos, não em knowledge objects em um banco de dados. O conhecimento não é codificado (e provavelmente não poderia ser), é transferido em sessões de brainstormings e em conversas individuais.

Para por em prática esta estratégia, empresas como a Bain investem pesadamente em construir redes de pessoas. O conhecimento é compartilhado não apenas em reuniões face a face, mas via e-mail e videoconferências.

Pode-se constatar que empresas que utilizam estratégias de Gestão do Conhecimento de forma eficaz estão focadas predominantemente em uma das estratégias e utilizam a outra para dar apoio à principal. Empresas que tentam seguir as duas estratégias ao mesmo tempo acabam por enfrentar problemas como ter de lidar com um mix de pessoas que tanto engloba pessoas altamente criativas como aquelas com grande capacidade técnica, o que acaba por criar conflitos de interesse dentro da empresa.

A Estratégia Competitiva deve guiar a estratégia de Gestão do Conhecimento. Empresas que fornecem alta qualidade, confiabilidade e velocidade de implementação de soluções devem seguir a estratégia de codificação e reutilização de conhecimento. Em oposição, para companhias que possuem como diferenciais a criatividade, Inovação e soluções de alto nível de complexidade, é mais indicada a segunda opção, ou seja, a estratégia de personalização.

Para que seja eficaz a estratégia de Gestão do Conhecimento, é necessário criar incentivos reais que possam estimular a contribuição das pessoas. Na abordagem de codificação, os gerentes precisam desenvolver um modelo que vise a estimular as pessoas a externalizar seu conhecimento e compartilhá-lo em um repositório comum. Na abordagem de personalização, é necessário estimular as pessoas a interagirem e compartilharem conhecimento umas com as outras.

1.6. Capital Intelectual

Um aspecto fundamental para a gestão do conhecimento é o capital intelectual. Capital intelectual é a posse de conhecimento, experiência aplicada, tecnologia organizacional, relacionamento com clientes e habilidades profissionais que proporcionem à organização uma vantagem competitiva no mercado. O capital intelectual pode ser decomposto em dois grandes grupos: o capital humano e o capital estrutural. Algumas definições trazem o capital do cliente como um outro grupo. Outras o trazem como integrante do capital estrutural.

Capital intelectual ou conhecimento é qualquer coisa valorizada pela organização que esteja contida nas pessoas, ou seja, derivada de processos, de sistemas e da cultura organizacional – conhecimento e habilidades individuais, normas e valores, bases de dados, metodologias, software, know-how, licenças, marcas e segredos comerciais, para citar alguns.

Uma distinção importante entre capital intelectual e o que tradicionalmente tem sido considerado gerador de valor nas organizações – os ativos físicos “tangíveis” – é que ele nem sempre é propriedade da organização. Isso significa que obter benefícios do conhecimento ou do capital intelectual pode ser considerado como alugado, arrendado ou emprestado. Para o presidente e COO da American Skandia, uma companhia de seguros sediada na Suécia, chega mesmo a sugerir que grande parte do capital intelectual de uma organização é obtido voluntariamente no dia-a-dia. O valor gerado resulta de atos arbitrários por parte dos indivíduos; assim, o processo de extrair esse valor deve ser gerido muito diferentemente do processo dos ativos tangíveis.

O capital humano inclui toda capacidade, conhecimento, habilidade e experiência individuais dos empregados e gerentes, além da criatividade e inovação organizacionais. Já o capital estrutural é o arcabouço e a infra-estrutura que apoiam o capital humano. Inclui-se aí toda estrutura tecnológica utilizada como suporte ao conhecimento intelectual. Parte do que pertence à categoria de capital estrutural está relacionada aos direitos legais de propriedade: tecnologias, invenções, dadas e publicações.

O capital estrutural inclui, ainda, a imagem da empresa, os conceitos organizacionais e a documentação. Os sistemas de gestão do conteúdo compõem o capital estrutural, ao passo que oferecem a estrutura para a gerência de outros tipos de capital estrutural como, por exemplo, documentos, manuais, código-fonte, etc. Finalmente, o capital de cliente é o valor de uma empresa para as pessoas com as quais faz negócios, incluindo o valor dos relacionamentos com os fornecedores.

1.6.1. A criação do Conhecimento na Empresa

São apontadas dez estratégias para Gestão do Conhecimento para aumentar a agregação de valor para a organização. Esta busca pela melhoria da transferência de conhecimento entre as três famílias de recursos intangíveis tem o intuito de aumentar a capacidade de agir das pessoas de dentro e fora da organização. Estas três famílias de recursos intangíveis são: estrutura externa, estrutura interna e competência individual. O capital intelectual da empresa pode ser dividido em três áreas: capital humano, capital do cliente e capital estrutural.

A estrutura externa consiste em um conjunto de relações intangíveis com consumidores e fornecedores que compõem a imagem da empresa. O capital do cliente é o valor das relações com pessoas e organizações para as quais se vende. Este recurso é representado por alguns ativos que apesar de intangíveis, podem ser valorados com mais facilidade como, por exemplo, marcas e fatias de mercado.

A estrutura interna é composta pelas patentes obtidas, conceitos, modelos, sistemas e outros processos e ferramentas mais ou menos explícitos. Esta estrutura é criada pelos empregados e é de propriedade da organização. Entretanto, a organização pode possuir legalmente apenas uma pequena parte destes recursos. As redes de relacionamentos e a cultura organizacional são alguns exemplos de peças de conhecimento sobre os quais a organização não possui um controle formal. O capital estrutural da empresa como resultante do conhecimento retido e que pode ser transformado em propriedade pela organização.

A competência individual consiste na capacidade do corpo técnico e administrativo da organização. O capital humano pode ser definido como o conhecimento que reside na mente dos empregados e que é relevante para os objetivos da organização. É formado e desenvolvido quando as pessoas que trabalham na organização devotam tempo e talento a atividades que resultam em Inovação. A competência do funcionário pode ser desenvolvida através da sua interação com outros funcionários, o que sugere que este ativo tenha origem nas relações sociais dentro e fora da empresa. O trabalho rotineiro que pode ser facilmente automatizado não é fonte de capital humano, pois não supre seu principal objetivo, a Inovação.

As dez estratégias propostas são:

1. Transferência de conhecimento entre indivíduos
2. Transferência de conhecimento dos indivíduos para a estrutura externa
3. Transferência de conhecimento da estrutura externa para os indivíduos
4. Transferência de conhecimento da competência individual para a estrutura interna
5. Transferência de conhecimento da estrutura interna para a competência individual
6. Transferência de conhecimento na estrutura externa
7. Transferência de conhecimento da estrutura externa para a interna
8. Transferência de conhecimento da estrutura interna para a externa
9. Transferência de conhecimento na estrutura interna
10. Maximizar a agregação de valor - visão global

É essencial para a Gestão do capital intelectual que se defina qual sua utilidade e como transformá-lo em valor real para a empresa e apresenta alguns princípios para a Gestão do Conhecimento, dos quais podem ser destacados os seguintes:

- Somente após reconhecer que os capitais humanos e do cliente possuem uma natureza compartilhada, as empresas podem gerenciar e lucrar com estes ativos;
- Para gerar capital humano, a empresa precisa estimular o trabalho em grupo, comunidades de prática e outras formas sociais de aprendizado;
- As organizações crescem em torno de talentos e habilidades que são escassos e inerentes às pessoas. Para gerenciar e desenvolver o capital humano, empresas devem reconhecer que pessoas com estas características são ativos a serem mantidos;

- O capital estrutural é o mais fácil de ser gerenciado, mas é o que menos importa para os consumidores;
- Os capitais humano, estrutural e do cliente trabalham juntos. Não é suficiente investir em pessoas, em sistemas ou nos clientes separadamente.

1.6.2. Disseminação do Conhecimento

Um dos principais desafios da Gestão do Conhecimento é exatamente capturar o conhecimento tácito. Por não ser lógica e estruturalmente documentado, é difícil sua manutenção. Ele deve ser disseminado entre os membros da organização porque representa ideias que foram planejadas, decisões que foram tomadas, motivos que levaram ao voto de outras decisões, opções e escolhas diante de impasses em determinados projetos e outras informações essenciais. É mais do que necessário identificar quais partes do conhecimento tácito dos indivíduos possuem maior significado estratégico, para que gradativamente sejam criados e ajustados os processos adequados para capturá-los e incorporados à Memória Organizacional. Uma das soluções possíveis para este problema é transformar o conhecimento tácito em explícito, de modo que ele possa ser registrado e documentado facilmente.

Porém, não é somente essa transformação de conhecimento tácito em explícito que é possível de ocorrer. A existência de duas classes de conhecimento sugere quatro modalidades de transmissão e criação de conhecimento também definidas por Nonaka e Takeuchi (1997), apresentadas a seguir:

- **Socialização** (tácito para tácito): consiste na transmissão direta de conhecimentos tácitos. Ocorre, em geral, através de observação, imitação e prática. São compartilhados modelos mentais, pontos de vista, experiências. Exemplo: um aprendiz que observa o trabalho do chefe durante meses e depois atua de forma semelhante à de seu mestre.
- **Combinação** (explícito para explícito): ocorre quando há combinação de vários conhecimentos explícitos para a formação de um novo todo. Exemplo: gerente de controladoria que coleta informações da organização e apresenta-as sob a forma de relatórios financeiros.
- **Externalização** (tácito para explícito): padrão em que novo conhecimento explícito é gerado através da formalização de um conhecimento tácito apresentado de forma categorizada e contextualizada. Exemplo: o mesmo gerente de controladoria do item anterior, mas que, em vez de compilar um relatório financeiro, desenvolve uma documentação inovadora sobre controle orçamentário, baseada em sua experiência conquistada após anos de trabalho. Outro exemplo: as melhores práticas são selecionadas dentro do universo informal da organização e, então, documentadas.
- **Internalização** (explícito para tácito): consiste na utilização de conhecimentos explícitos para a ampliação, extensão e reformulação do conhecimento tácito dos indivíduos. Exemplo: um aluno sempre incorpora o que aprende em documentos, manuais e treinamentos ao seu conhecimento tácito, isto é, à sua experiência de vida, possibilitando o surgimento de insights capazes de gerar novos conhecimentos.

Numa organização criadora de conhecimento, os quatro processos interagem dinamicamente, formando o que Nonaka e Takeuchi (1997) chamam de espiral do conhecimento, uma espécie de ciclo de vida em que o conhecimento passa por todos os quatro padrões de conversão seguindo o esquema mostrado abaixo:

Independente de qualquer categorização, mais importante é o modo como a organização promove e estimula seus funcionários a construir e compartilhar conhecimento. A gestão do conhecimento é um recurso valioso para as empresas. Por isso, deve-se priorizar a criação e a implementação de processos que organizem e sistematizem a capacidade da companhia de capturar, armazenar, gerar, criar, analisar, traduzir, compartilhar e fornecer a informação exata de maneira rápida e precisa.

Estes mecanismos e processos devem ser elaborados para assegurar que a empresa aumente o valor de um dos mais importantes recursos competitivos nos dias de hoje – o conhecimento. O conhecimento é um bem intangível. É uma combinação de dados que, tratados e contextualizados, fornecem soluções essenciais no processo de tomada de decisões em todos os níveis de uma corporação. As informações navegam nas áreas de vendas, marketing, publicidade, serviços, operações e administração das companhias, colaborando com ações estratégicas e criando oportunidades de negócios.



Imagine a área de uma empresa onde o conhecimento é proveniente única e exclusivamente da mente de seu gestor. Digamos que este colaborador deixe a companhia. Provavelmente você já viu isso de perto ou ouviu dizer que aconteceu, correto? Levando em consideração que vivemos em uma economia baseada na informação e o sucesso das corporações depende muito da capacidade de gestão do conhecimento e da manutenção da inteligência organizacional, é essencial que as empresas busquem a criação de um mecanismo para gerenciar este conhecimento organizacional.

Caso contrário, como no exemplo acima, a falta de uma única peça irá comprometer os resultados e gerar um impacto negativo no desempenho de determinada área ou até da companhia inteira.

1.6.3. Marketing do Conhecimento

Como resposta às ameaças e necessidades competitivas das empresas, durante os últimos quinze anos vimos evoluir e expandir largamente as atividades de marketing. Ao longo dos últimos anos vimos o marketing absorver novas funções, conceitos, tecnologias e práticas, incorporando seguidas inovações.

Com todos os desafios de apoiar as estratégias competitivas, auxiliar a adaptação da companhia ao ambiente cada vez mais globalizado e extremamente competitivo e atender as demandas crescentes por conhecimento era de se esperar que a gestão do conhecimento provocasse mudanças em todos os níveis da empresa, contudo, muitos de nós ainda se perguntam: Em que medida o marketing será afetado pela gestão do conhecimento? Como as funções de marketing poderiam apoiar as estratégias de conhecimento das corporações? Qual a expectativa da empresa do conhecimento em relação ao marketing e qual os novos desafios que a gestão do conhecimento impõe a esta área?

Nestes últimos anos falamos sobre marketing de relacionamento, marketing de massa, marketing de produto, marketing de marca, marketing de segmentação, marketing de clientes, marketing de afinidade, marketing direto, one to one marketing e tantos outros. Com mais frequência gestores admitem o apoio da gestão

conhecimento ao marketing. No entanto, poucos se atrevem a reconhecer que a gestão do conhecimento possa se beneficiar das atividades de marketing, sendo as discussões em torno do que o marketing pode fazer para apoiar a estratégia de gestão do conhecimento ainda rara e incomum.

Confirmando que a área de marketing além de se beneficiar da gestão do conhecimento pode também ajudá-la, tenho percebido sinais sutis de um novo marketing surgir. Um marketing não só de produto, mas também de conhecimento, caracterizando uma tendência incipiente, mas crescente. Tal como outras tendências do passado, um novo desafio aos profissionais, o marketing do conhecimento também será incorporado pela área de marketing, que passará a se envolver e comprometer-se com a gestão do conhecimento.

Esta ideia emergiu de reflexões e pode ser confirmada a partir da definição de Philip Kotler sobre marketing. Philip Kotler afirma que o marketing é a arte de descobrir oportunidades, desenvolvê-las e lucrar com elas. A frase adaptada aos propósitos da gestão do conhecimento, pouco alterada em sua essência, poderia ficar assim: O marketing é a arte de descobrir oportunidades para aplicação e uso do conhecimento, desenvolvê-las e lucrar com elas. Neste contexto surge a importância deste “novo marketing” e a necessidade de novas funções que irão suportá-lo. Esta afirmativa adaptada facilitou sensivelmente a conclusão de que a área de marketing poderia auxiliar as estratégias de gestão do conhecimento nas empresas. Desde então, iniciou-se um trabalho de relacionamento de alguns trabalhos relevantes à gestão do conhecimento, como a gestão de demanda e oferta do conhecimento, estruturar e organizar as entregas de conhecimento aos clientes, elaborar estratégias mercadológicas para o conhecimento, tornar visível a diferenciação por competências e capital intelectual, entre outras, aos serviços executados pela área de marketing. Os resultados relacionados à atividade de marketing, como a descoberta de necessidades não atendidas, oferta e comunicação dos diferenciais do conhecimento, medição e avaliação de oportunidades resultados a partir do conhecimento e etc..., quando obtidas e entregues às equipes de gestão do conhecimento são extremamente importantes ao sucesso dos programas. Conclui também que a execução de muitas destas atribuições, inerentes ao marketing, poderiam firmar a área de marketing como um importante aliado da gestão do conhecimento nas empresas. A este grande conjunto de atividades executadas pela área de marketing de apoio à gestão do conhecimento, aqui chamada de marketing do conhecimento ou marketing de competências (como preferir). Esta ideia está acima de tudo, baseada no princípio de que a gestão do conhecimento gerará demandas (algumas ainda desconhecidas) só atendidas por uma função do marketing. Acredita-se que muito em breve, antes mesmo de nos especializarmos no assunto, estaremos ouvindo falar em nossas empresas sobre o marketing do conhecimento (mesmo que ele conquiste outro nome).

Mas o que vem a ser marketing do conhecimento?

O marketing do conhecimento é atividade que considera a venda não só do resultado do saber (produtos e serviços), mas a "venda" do "próprio" conhecimento e do modo como ele é criado, alavancado, produzido, compartilhado, utilizado, embalado e entregue pelas empresas, extraíndo proveito dos possíveis diferenciais que cercam sua gestão e do modo como eles são percebidos e valorizados pelo mercado.

Este marketing de competências (se preferir) acima de tudo seguirá a tendência de valor do saber. Ao contrário dos produtos e serviços, o marketing do conhecimento em seus trabalhos considerará e enfatizará o conhecimento e o capital intelectual e se empenhará em torná-los evidentes e transparentes ao público alvo de interesse.

Dentro desses vários casos, ainda que sutis, pôde-se identificar a síntese desta tendência que desabrocha. A Renault, por exemplo, em uma de suas campanhas recentes veiculou em algumas revistas o seguinte anúncio: 3 fábricas em 3 anos. Não é só nas estradas que o desempenho da Renault impressiona. (O anúncio exibe um carro Renault em perspectiva com alguns robôs da fábrica.) A United, enfatizando a vasta experiência de seus pilotos, veiculou, também recentemente, a seguinte mensagem: Pilotos com muita quilometragem, aviões com

pouca. A Semp Toshiba possui vários exemplos: Notebooks Toshiba. O único lançamento nosso que a concorrência consegue pôr no bolso. Ou Os nossos japoneses são melhores que os da concorrência.

A maioria das empresas aplica o marketing focado em seus produtos e serviços. Um novo mix de marketing deve emergir focando também o conhecimento e o capital intelectual. Este mix de marketing, até então composto por quatro Ps (Produto, Preço, Praça, Promoção) poderia incorporar um novo P. O "P" de Potencial Intelectual. Fazendo ou não parte do famoso mix, este quinto "P" conquistará a atenção do marketing nos próximos anos e sobre ele concentrarão as promessas de sucesso.

Cada vez mais, observam-se anúncios das empresas na tentativa de identificar sinais ou evidências do marketing do conhecimento. Embora tudo que tenha conseguido sejam iniciativas isoladas, sutis e embrionárias, elas revelam uma certa espontaneidade em direção ao tema. Ainda que sutilmente, intencional ou não, empresas como a HP, Toshiba, Brastemp, Renault, Semp Toshiba e algumas outras, através de suas agências de publicidade, se antecipam, revelando o marketing do conhecimento como uma tendência.

O marketing do conhecimento deverá assim, ainda que lentamente, se concentrar em tornar transparente e revelar através de campanhas e iniciativas, os conhecimentos de valor e o potencial intelectual da empresa, atraindo clientes, investidores, parceiros, talentos, investidores e progresso como consequência.

Figueiredo (2009) aponta algumas funções do marketing do conhecimento:

1. Analisar e monitorar constantemente o mercado em busca de oportunidades para aplicação do conhecimento

O marketing do conhecimento, baseado em seus ativos intelectuais e na sua capacidade de atuação e de entrega do seu potencial intelectual, deve analisar e monitorar constantemente o mercado em busca de oportunidades para aplicação e uso de suas competências e de seus ativos intangíveis. A disponibilização desses ativos ao mercado, através dos produtos e serviços criados a partir deles, depende fundamentalmente dessa capacidade de análise e identificação das oportunidades, que também é uma função do marketing.

Quando as oportunidades encontradas estão vinculadas à geração de receitas e são coerentes com as competências essenciais da companhia, com sua proposição de valor e visão, a empresa entra em ação e se dispõe a combinar uma série de ativos intelectuais específicos (conhecimentos), para atender a estas demandas estratégicas, frequentemente vinculadas à mobilização, embalagem e entrega de conhecimentos.

2. Classificar e qualificar os ativos intangíveis

Qualquer que seja o ativo intangível identificado, ele precisa ser classificado, qualificado e avaliado à luz da estratégia da organização. A empresa, entre outras coisas, deve avaliar se os ativos intangíveis somam e agregam valor à produção de bens, serviços e produtos, se apóiam a proposição de valor da companhia, se são positivos à atuação empresarial, se são percebidos e valorizados pelos clientes, se geram ou poderiam gerar novos produtos e resultados, se contribuem ou poderiam contribuir efetivamente aos resultados, etc. A partir desta análise, a empresa descobre a relevância dos ativos intangíveis, promove seu alinhamento estratégico e planeja uma ação efetiva para sua utilização em benefício dos negócios e da empresa.

3. Alinhamento mercadológico do capital intelectual

A criação apenas dos ativos intangíveis pode não ser suficiente para gerar os retornos esperados. Em ambientes de hipercompetitividade como o nosso, é necessário dar-lhes foco e posicionamento mercadológico para que sejam percebidos, reconhecidos e valorizados externamente, o que chamo de alinhamento mercadológico do capital intelectual. A empresa tem, assim, que descobrir a melhor maneira de identificar e comunicar a contribuição que se tem a oferecer ao público e mercado alvos, a partir dos seus ativos intangíveis,

produzindo coerência entre a oferta de sua capacidade intelectual e o que é procurado ou valorizado pelo mercado (demanda).

4. Posicionamento dos ativos intelectuais para a entrega de produtos e serviços inovadores

A materialização resultante dos ativos intangíveis é caracterizada pela produção de valor aos consumidores e resulta na entrega de produtos e serviços inovadores. A empresa deve posicionar adequadamente seus ativos e diferenciais intelectuais no mercado, comunicando ao mundo externo todo o potencial associado a eles e evidenciando como estes ativos a tornam uma empresa diferentemente exclusiva. A comunicação com o mercado é, portanto, fundamental, existindo muitas maneiras diferentes de dizer o que sabe, de comunicar o potencial intelectual, de compartilhar a visão de futuro, de informar o compromisso com a vanguarda e toda capacidade intelectual capaz de criar valor ao mercado.

E mais:

- Tornar transparentes ao público-alvo os ativos intelectuais mais valorizados pelo mercado e clientes ou que fariam a diferença entre eles;
- Divulgar os diferenciais competitivos baseados na competência e no conhecimento dos talentos da empresa;
- Publicar diferenciais intelectuais que poderiam atrair talentos, investidores, clientes, parceiros e despertar atenção do mercado;
- Tornar visível ao mercado o valor associado às competências essenciais da empresa;
- Enfatizar o conhecimento coletivo organizacional admirável e de alto valor requerido pelo mercado;
- Evidenciar ao mercado a contribuição do capital intelectual das pessoas ao sucesso da empresa;
- Vincular e comunicar ao público externo a real contribuição dos ativos intangíveis no alcance dos resultados financeiros, na obtenção das melhorias de processos, na satisfação e fidelização de clientes, no desenvolvimento de soluções de vanguarda e da inovação, na obtenção da criatividade, na retenção e atração dos recursos humanos, etc;
- Mostrar aos compradores potenciais, investidores e talentos que se querem atrair, indicadores relacionados à qualidade do capital intelectual da empresa e como eles resultam em eficiência, competência, crescimento, alcance de resultados, capacidade de invenção e criatividade, satisfação dos clientes, transferência e uso do conhecimento, qualidade das soluções, intensidade de relacionamentos, registro de patentes, intensidade de pesquisas e aplicação, eficácia, desenvolvimento técnico, adaptação, qualidade de produtos e serviços, capacidade de aprendizado e inovação, criação de novos produtos, valorização da companhia no mercado, capacidade lucrativa e de negócios, entre outros;
- Divulgar os diferenciais baseados nos ativos intelectuais e sua proposição de valor; e
- Construir campanhas que evidenciem a qualidade dos ativos intelectuais intangíveis mais estratégicos da empresa, considerando as pretensões competitivas, de inovação, de evolução, participação no mercado, produção e criação de produtos, prestação de serviços, etc.

1.7. 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. Entre 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 estratégicamente 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.

1.8. 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
<ul style="list-style-type: none"> - 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 - Rígidez da organização 	<ul style="list-style-type: none"> - 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 <i>superaprender</i> - 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 renderá. 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**¹ 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

¹ 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.

hierárquicos, de acordo com Terra (2000). Para Quinn Terra (2000) a abordagem atual baseia-se nos **skunk works**² 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**³ 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 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 idéias, que possibilita às pessoas testarem suas idéias e, de forma geral, possibilita que elas convivam bem com o erro (Terra, 2000).

1.8.1. Gestão de Conteúdo

Hoje, em vez de serem forçadas a agir com base em pouca ou nenhuma informação, as pessoas tendem a achar que o desafio confrontado por elas é depurar pilhas de informações irrelevantes para obter a “pepita” indispensável para as suas necessidades. A tecnologia da informação é tanto o infrator quanto o guarda de trânsito para esse passo do processo de gestão do conhecimento. Por um lado, a TI abriu as comportas da informação e a enviou através das organizações. Por outro lado, a cada dia, os sistemas de TI tornam-se mais inteligentes e capazes de direcionar as pessoas por labirintos de irrelevâncias, até a informação que elas necessitam.

Em face desse desafio da busca por informação relevante, modelos de gestão de conteúdos surgiram em decorrência da explosão do conteúdo digital. Alguns modelos oferecem a estrutura necessária para a criação, o gerenciamento e a publicação de conteúdo eletrônico, seja na Internet, em intranets ou em outros sistemas de informações.

A gestão de conteúdos é uma combinação de tecnologia e processos organizacionais: a tecnologia facilita a criação, o armazenamento e a disponibilidade do conteúdo; e os fluxos de trabalho e os processos organizacionais são a essência para o sucesso da implementação tecnológica. A explosão da quantidade de informações cria a necessidade de sistemas que reduzam o caos criado por este contexto. O aumento no tempo de busca por informações é um sintoma deste problema. A criação de relações semânticas entre documentos, indivíduos e processos agrupa valor ao negócio, facilitando a tarefa de buscar os documentos relevantes.

O volume de informações e dados criados para uso interno ou externo em organizações de médio e grande porte é assustador. Relatórios são criados, websites são publicados, documentos são produzidos, etc. Questões como onde tudo isso será armazenado ou como estas informações serão recuperadas são latentes nas organizações que precisam lidar com esta revolução. A implantação de um sistema de gestão de conteúdos facilita a tarefa de administrar os repositórios de conhecimento. Documentos criados em um sistema de gestão de conteúdos podem ser armazenados em uma base de dados central, representados por metadados, e recuperados através de palavras-chave. A implementação de níveis de segurança garante que documentos considerados importantes não sejam acessados por indivíduos não autorizados.

² Skunk Works são espaços propositadamente informais e desconectados do ambiente corporativo, que propiciam maior identidade entre o colaborador e seu espaço de trabalho.

³ 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.

1.8.2. A Gestão de Conteúdos no Contexto da Gestão do Conhecimento

A gestão do conteúdo se identifica com a dimensão explícita do conhecimento, uma vez que gerencia os objetos portadores de conhecimento explícito. Observa-se, também, o papel da gestão de conteúdos em três dos quatro modos de conversão do conhecimento propostos por Nonaka e Takeuchi e ilustrados na Figura 7. A conversão do tipo socialização não envolve o apoio por um sistema de gestão de conteúdos, visto que este processo baseia-se na troca de experiências entre os indivíduos, dispensando a linguagem em alguns casos.

Na externalização, o conhecimento tácito é articulado em conceitos explícitos. É o momento em que o indivíduo cria suas representações na forma de metáforas, analogias, conceitos, hipóteses ou modelos. Estas representações geralmente tomam, entre outras, a forma de documentos, que são instrumentos de entrada em um sistema de gestão de conteúdos. Dentre os quatro modos de conversão do conhecimento, a externalização é a chave para a criação do conhecimento, pois cria conceitos novos e explícitos a partir do conhecimento tácito.

Ao sistematizar os conceitos em um sistema de conhecimento dá-se origem à combinação. “A reconfiguração das informações existentes através da classificação, do acréscimo, da combinação e da categorização do conhecimento explícito pode levar a novos conhecimentos”. Os sistemas de gestão de conteúdos oferecem suporte à combinação de diferentes conjuntos de conhecimento explícito, por meio da criação de relações semânticas entre os documentos, artefatos, etc., conforme ilustrado na Figura 8.

Na internalização o conhecimento explícito é incorporado ao conhecimento tácito. Para tanto, é necessárias a verbalização e diagramação do conhecimento sob a forma de documentos, manuais ou histórias orais. A documentação ajuda os indivíduos a internalizarem suas experiências, aumentando seu conhecimento. Os documentos são peças chaves na transferência do conhecimento explícito para outros indivíduos.

Sistemas de Gestão de Conteúdo (Content Management System)

As ferramentas de gestão de conteúdo Web são atualmente objeto de forte interesse. Poucos projetos Internet de expressão são encarados hoje sem recurso à uma solução que integre o maior número possível de funcionalidades de gestão de conteúdo.

Um sistema de gestão de conteúdo é geralmente composto de módulos que fornecem funcionalidades básicas sobre as quais desenvolvem-se as aplicações mais próximas do usuário final. As funcionalidades essenciais dentre muitas outras, que caracterizam o conceito e que se desenvolvem à medida que novos produtos de mercado chegam à maturidade são:

- Gestão de usuários e dos seus direitos (autenticação, autorização, auditoria);
- Criação, edição e armazenamento de conteúdo em formatos diversos (html, doc, pdf, etc);
- Uso intenso de metadados (ou propriedades que descrevem o conteúdo);
- Controle da qualidade de informação (com fluxo/trâmite de documentos ou workflow);
- Classificação, indexação e busca de conteúdo (recuperação da informação com mecanismos de busca);
- Gestão da interface com os usuários (atenção à usabilidade, arquitetura da informação);
- Sindicalização (syndication, disponibilização de informações em formatos XML visando seu agrupamento ou agregação de diferentes fontes);
- Gestão de configuração (gestão de versões);
- Gravação das ações executadas sobre o conteúdo para efeitos de auditoria e possibilidade de desfazê-las em caso de necessidade.

Esse é o número de funcionalidades minimamente necessárias à gestão de conteúdos e existentes nos diversos produtos hoje, a preocupação foi meramente enumerar as mais significativas. Além disso, quando se fala em conteúdos, informação e conhecimento, cada organização é única e exige adequação própria à sua realidade. Em geral, os fornecedores de software de gestão de conteúdo não têm uma solução universal, completa, que integre de forma consistente todas as funcionalidades requeridas para cada organização.

Frequentemente, as soluções são especializadas apenas em certos aspectos, fazendo-se necessário na maioria das vezes integrar ou associar produtos. Dessa forma a instância de decisão de implantação na organização alvo acaba, quase sempre, cumprindo um papel de integradora.

Do exposto acima se pode concluir que é preferível que uma solução de gestão de conteúdo forneça os "tijolos" (o mais atônicos possível para o nível de abstração requerido) para que o projetista possa montar a seu modo, com flexibilidade a solução mais adequada às necessidades peculiares de cada organização. O projetista está sujeito, muito frequentemente, a ter que realizar escolhas quem envolvem compromissos. Geralmente, quanto mais genérica é a solução (ou seja, responde a um grande número de necessidades) mais é complicado configurá-la para atender às necessidades específicas. Contrariamente, quanto mais específica é a solução, mais especializada ela se torna e menos oferece flexibilidade.

As áreas de aplicação

A gestão de conteúdo vem permitir a industrialização da construção e do funcionamento de sites web que têm restrições críticas: vasta audiência, atualização freqüente, segurança, tempo de carregamento de páginas, conteúdo multimídia, transações comerciais, etc.

São inúmeros e variados os tipos de sites web, bem como as maneiras como estes se relacionam com sua comunidade de usuários (pela Internet, intranet ou extranet). Não existe consenso entre os especialistas sobre uma tipologia ou taxonomia universal. Entre as aplicações típicas de um sistema de gestão de conteúdo, pode-se citar três categorias básicas: sítios editoriais, comunidades em linha, e portais corporativos. Óbvio que a lista não é exaustiva, ela serve apenas para ilustrar algumas áreas de aplicação da gestão de conteúdos.

Sites Editoriais

É talvez o tipo de site mais comum hoje na web, que assume natureza de mídia de comunicação. Um sítio deste tipo permite a um indivíduo ou a um grupo posicionar- se como fonte de informação, infomediário, ou veilleur sobre assuntos específicos. Apresenta-se sob diferentes formas de acordo com o modelo econômico, o objetivo visado, e a tendência do momento.

Os sites editoriais mais encontrados são os portais de informação verticais, os jornais e revistas em linha, e os mais recentes weblogs. Esses últimos, sob restrições diferentes das do mundo editorial, pela extrema facilidade de manipulação dos conteúdos, são atualmente muito populares no mundo dos sítios pessoais e do jornalismo amador. Uma indicação prática da importância crescente desses sites foi a aquisição recente pelo Google do bloguer.com, que no início de 2003 possuía em torno de 1.1 milhões de usuários registrados.

Os sites do tipo weblogs são os primeiros a utilizarem amplamente características rudimentares de gestão de conteúdos, orientadas a usuários não especialistas.

1.8.3. Comunidades de Práticas Compartilhadas

O termo Comunidade de Prática (Community of Practice – CoP) surgiu a partir de um estudo sobre modelos de aprendizagem. Uma Comunidade de Prática é um grupo de pessoas que compartilham um interesse, um problema que enfrentam regularmente, e que se unem para desenvolver conhecimento de forma a criar uma prática em torno desse tópico.

Uma comunidade de práticas compartilhadas reúne pessoas que compartilham centros de interesse de ordem geral ou profissional, não se resumindo obviamente a códigos de programas. Os aplicativos de suporte oferecem a possibilidade de contribuir com informações na forma de artigos, notícias etc., e alertar a comunidade para informações disponíveis noutros lugares da web.

Geralmente espaços de wikis, fórum e listas de discussão, etc. permitem aos membros da comunidade reagir e compartilhar seus conhecimentos, dando pareceres sobre documentos publicados ou adicionando contribuições ou complementos diretamente aos documentos publicados.

Para uma comunidade ser considerada uma “Comunidade de Práticas”, a mesma deve possuir os seguintes três elementos, obrigatoriamente:

- **O domínio:** representa uma identidade, um domínio de interesse. Os membros de uma comunidade possuem um compromisso com o domínio e compartilham uma competência que os distingue de outras pessoas que não são membros da comunidade.
- **A comunidade:** Para atender aos interesses relacionados a seu domínio, os membros participam de atividades e discussões em conjunto, ajudam uns aos outros e compartilham informação. São construídos relacionamentos que permitem que um membro aprenda com o outro. Além disso, não é necessário que todos os membros se conheçam diretamente ou trabalhem diariamente juntos para participar de uma mesma comunidade.
- **A prática:** Uma comunidade de práticas não é meramente uma comunidade de interessados sobre um assunto e sim de pessoas que trabalham com um assunto. Por exemplo, não é uma comunidade de pessoas que gostam de pintura e sim de pintores. Desta forma diz-se que os membros de uma Comunidade de Prática são os praticantes de um determinado assunto. Consequentemente, eles trocam histórias, experiências, ferramentas, formas de resolução de problemas, e outros recursos viabilizando a geração de práticas compartilhadas.

O desenvolvimento destes três itens em paralelo é o que mantém uma comunidade ativa e sustentável. Segundo estudos realizados, as Comunidades de Práticas estão assumindo um papel-chave na gestão do conhecimento das organizações. Isto ocorre por elas transcederem fronteiras criadas por estruturas hierárquicas, funções, aspectos geográficos e tempo. Nas organizações modernas, globais e baseadas no conhecimento, as comunidades viabilizam um ambiente de geração de conhecimento que flui por todas estas fronteiras.

A criação de uma Comunidade de Práticas viabiliza um meio através do qual é proporcionado à empresa a retenção do conhecimento, a troca de conhecimentos entre os mais experientes e mais jovens, o fortalecimento das redes entre os profissionais. Com isto, a empresa passa a ter economia de tempo e recursos, além de possibilitar uma melhoria contínua na qualidade dos seus processos.

Apesar da quantificação dos benefícios relacionados à utilização de Comunidades de Práticas não ser realizada na maior parte das empresas, são descritos a seguir alguns resultados positivos reportados por empresas do setor de Óleo & Gás:

- Chevron Corporation reportou que a utilização de Comunidades de Práticas resultou em uma redução de US\$ 2 bilhões em custos operacionais;
- Schlumberger utilizou uma combinação de Comunidades de Práticas e obteve uma economia de US\$ 10 milhões em um ano de atividade;

É importante observar que as comunidades de práticas sempre fizeram parte das organizações, só que de maneira informal. Um dos aspectos que reforça a necessidade desta formalização de conhecimento é o fato de hoje em dia haver uma alta rotatividade de recursos humanos nas empresas e de que, quando uma pessoa sai de uma empresa (por motivos diversos, desde sua aposentadoria até uma busca de oportunidade em outra organização), o conhecimento muitas vezes é perdido.

Como o tema relacionado à Comunidade de Práticas é muito extenso para um post, optamos por dividir o assunto por partes, a saber:

- Implantação de uma Comunidade de Práticas
- Fatores Críticos de Sucesso na Implantação de uma Comunidade de Práticas

Existem quatro tipos de Comunidades de Práticas, segundo a American Productivity & Quality Center (APQC, 2002):

- **Comunidades de Ajuda (Helping Communities):** focam na conexão entre seus membros, de forma que os mesmos possam facilmente solicitar ajuda para resolução de problemas específicos e espontaneamente compartilhem ideias;
- **Comunidades de Melhores Práticas (Best-practice Communities):** focam no desenvolvimento, validação e disseminação de práticas que são consideradas por um grupo de validadores como melhores práticas. Estas são armazenadas em uma Base de Conhecimento;
- **Comunidades de Administração de Conhecimento (Knowledge-Stewarding Communities):** visam organizar e administrar o conhecimento coletivo da Comunidade, incluindo o material que seus membros utilizam no dia-a-dia;
- **Comunidades de Inovação (Innovation Communities):** possuem como objetivo primário desenvolver ideias e práticas inovadoras.

O primeiro passo para a implementação de uma Comunidade de Práticas, independente do tipo, é a definição de sua identidade: qual a expectativa da organização com relação a esta comunidade, qual o tema que ela abordará, quem será seu grupo-alvo e quem será seu patrocinador.

Na sequência da implementação, deve ser estruturado um modelo de governança que irá depender do tipo de CoP, se a mesma terá ferramentas associadas e do porte da organização. Esta governança definirá: a dinâmica de funcionamento, as atividades de gestão da comunidade e os papéis e responsabilidades de cada um dos membros desta governança. Um modelo simplificado de governança conta com três tipos básicos de membros:

- **Coordenadores**, que de maneira geral, são responsáveis pela definição da estratégia de implantação da comunidade e zeladores dos seus processos básicos;
- **Multiplicadores**, que atuam diretamente na motivação dos usuários em manter a movimentação na comunidade;
- **Usuários**, que farão uso do conhecimento registrado e contribuirão com o acréscimo de novos conteúdos.

O próximo passo, após a implantação da governança, é a estruturação do ambiente tecnológico em que a comunidade será desenvolvida e a criação dos processos básicos de funcionamento da mesma (publicação, validação, cadastramento de membros, etc.).

Com a comunidade desenhada inicia-se o planejamento dos treinamentos, um específico para cada perfil. Primeiramente, estes são realizados para os coordenadores e multiplicadores considerando suas atribuições específicas. Em seguida, o treinamento dos usuários da comunidade dá continuidade ao processo, tendo como objetivos fundamentais: garantir a equalização dos usuários na comunidade, incentivar a sua participação efetiva nas atividades da comunidade e obter o feedback da percepção deles em relação a comunidade buscando um comprometimento dos mesmos.

Em linhas gerais, o processo de implantação de qualquer comunidade de práticas: define a sua identidade, cria estrutura de governança, estabelece processos básicos, define ambiente tecnológico e realiza treinamento de membros. No entanto, a forma como cada Comunidade de Práticas é definida e estruturada em uma organização, de acordo com a APQC, pode variar muito, dependendo da filosofia da organização, do que ela espera da comunidade e das circunstâncias na qual ela é criada. Independente destas especificidades, a implantação da comunidade será tão melhor sucedida quanto for a participação de cada membro dentro de suas atribuições.

Fatores Críticos de Sucesso na Implantação de uma Comunidade de Práticas

A implantação de uma Comunidade de Práticas envolve diversos fatores críticos de sucesso, dos quais os mais significativos estão descritos a seguir:

- **Mudança de cultura.** Incluir o hábito de participar da Comunidade de Práticas na rotina da força de trabalho envolve uma mudança de cultura, que se torna mais crítica no que se refere à troca de experiências

negativas. Dedicar um tempo para consultar e utilizar práticas criadas por outras pessoas também envolve uma mudança de processos organizacionais;

- **Seleção de pessoas com perfil adequado para compor os papéis-chave da estruturação e implantação de uma Comunidade de Práticas.** Estas pessoas deverão ser reconhecidas corporativamente pelo excelente relacionamento inter-pessoal com seus pares e equipes, e pelo papel de formadores de opinião que exercem na sua área de atuação. Caso a Comunidade envolva atividades de validação de práticas, estas deverão ser realizadas por membros notoriamente reconhecidos na comunidade e que tenham experiência no assunto;
- **Necessidade de criação e manutenção de uma estrutura para o planejamento, implantação e continuidade da Comunidade.** É fundamental que esta estrutura envolva pessoas desde a alta gerência até os responsáveis pelo treinamento e disseminação. Esta estrutura deverá ter a representatividade de diversas áreas da empresa, a saber: Gestão de Conhecimento, Processos, Recursos Humanos, Tecnologia de Informação e as áreas técnicas que serão o foco da Comunidade de Práticas;
- **Obtenção do Patrocínio e comprometimento da alta e média gerência.** Uma vez que a utilização da Comunidade de Práticas envolve mudança de cultura, criação e manutenção de uma infra-estrutura de suporte e dedicação parcial da força de trabalho da organização, a obtenção e manutenção do patrocínio é um fator crítico de sucesso;
- **Transformação do conhecimento implícito em conhecimento tácito,** para que este possa ser registrado de forma a conter informação suficiente para possibilitar a disseminação dos benefícios na organização;
- **Treinamento adequado.** O treinamento não costuma merecer a atenção adequada na maior parte das organizações. Muitas vezes erra-se em focar o treinamento apenas na ferramenta a ser utilizada como suporte à metodologia, sem considerar o treinamento no processo e nos aspectos humanos que envolvem a gestão do conhecimento. Outras vezes, há a preocupação de se treinar apenas os membros da Comunidade, sem se preocupar com os demais perfis da governança;
- **Comunicação com os componentes da governança de forma eficiente e com qualidade.** Existem diversas formas que podem ser utilizadas para divulgar uma comunidade ou obter feedbacks sobre sua utilização. Contudo, deve-se tomar cuidado com o excesso de informação para que os membros não percam o interesse.

A transferência de conhecimento é um processo que não envolve apenas a aplicação de técnicas ou utilização de ferramentas. Envolve também uma mudança de cultura organizacional, na qual as experiências geradas sejam elas positivas ou negativas se convertem em capital da empresa. A criação de uma Comunidade de Práticas viabiliza um meio através do qual é proporcionando à empresa a retenção do conhecimento e a troca de conhecimentos entre os mais experientes e mais jovens.

Como toda técnica, ferramenta ou processo que se deseja utilizar em uma organização, a utilização de Comunidades de Práticas deve considerar como aspecto essencial o foco em Pessoas, uma vez que todos os fatores críticos de sucesso listados anteriormente estão relacionados, de uma forma ou outra a pessoas.

1.8.4. Portais Corporativos

São aplicações que funcionam em intranets ou extranets, mas podem também ser acessadas pela Internet. Dentro vários benefícios, essas aplicações permitem capitalizar a informação, o conhecimento e a competência das organizações: ideias estruturadas ou não, documentação, procedimentos administrativos, técnicos, de marketing etc. De preferência essa capitalização deve ser feita de maneira estruturada e coerente, garantindo segurança no acesso às informações públicas e privadas. Esses últimos são papéis importantes cumpridos pela ferramenta de gestão de conteúdo que é parte de todo portal corporativo.

Assim, esse tipo de aplicação requer a utilização de tecnologia e padrões universais avançados, capazes de permitir uma gestão adequada da informação, tanto estruturada quanto não estruturada. É importante também que estas tecnologias estejam a serviço de todos os funcionários e demais colaboradores que são a principal origem do capital intelectual capitaneado pelo portal.

As ferramentas de gestão desenvolvidas precisam ser flexíveis, de utilização simples e extremamente conveniente, facilitando o compartilhamento das idéias no momento exato em que elas surgem e estão prontas a serem explicitadas em um texto. Por isso diz-se comumente que essas ações de compartilhamento ou são triviais ou são impossíveis, simplesmente não acontecem. Com efeito, sem muita conveniência o portal não acontece. Claro que os incentivos e o reconhecimento pela cultura da organização são fatores que motivam uma certa disciplina mínima necessária a qualquer explicitação de conhecimentos.

Principais benefícios da gestão de conteúdo

A gestão de conteúdo visa dar respostas aos seguintes problemas principais:

- Gargalos diversos que estrangulam a produção de conteúdos para a Web;
- Falta de comprometimento ou implicação dos usuários, devido a dificuldades técnicas de publicação e uso. Excluindo-se questões motivacionais que a gestão de conteúdo, embora não tenha respostas diretas, pode apoiar com instrumentos;
- Falta de organização mais elaborada do conteúdo, que apresentem por exemplo os itens informacionais e suas relações na forma de links;
- Riscos de erros diversos e informação de baixa qualidade;
- Interfaces rígidas misturadas ao conteúdo, não personalizáveis ou não configuráveis.

Eliminar os gargalos que estrangulam a produção web

Tradicionalmente, a manutenção e a atualização do conteúdo são responsabilidade de um ou dois indivíduos, encarregados da administração do sítio e únicos a disporem das competências técnicas requeridas para tal. Com a expansão do sítio, estes encontram-se rapidamente sobrecarregados e a atualização do sítio fica atrasada. A correta gestão de conteúdo deve permitir à qualquer membro de uma organização ou de uma comunidade colocar em linha informação sem dificuldade técnica.

Para isso, a pessoa necessita apenas coletar conteúdos, e disponibilizá-los com a ajuda de um formulário eletrônico no próprio navegador web. O potencial de um sistema de gestão de conteúdo eficaz é tal que os diferentes membros intervenientes em um projeto ou atividade intelectual, para serem produtivos, não precisam se encontrar fisicamente ou nem mesmo se conhecerem pessoalmente.

Facilitar e motivar a produção de conteúdo e seu uso

A publicação de conteúdo não deve exigir mais do que o uso de um simples navegador web. O uso de programas clientes específicos acaba impedindo a colaboração quando o funcionário não está usando o seu computador. Assim o produtor de conteúdo pode publicar suas informações no site de qualquer lugar e à qualquer momento. Esta conveniência é importante e não deve ser subestimada. Ela passa a ser indispensável no caso de empresas onde as equipes estão espalhadas geograficamente, e trabalham a distância e/ou com horários deslocados. A ferramenta de publicação do conteúdo deve, por outro lado, permitir inserir os documentos produzidos com os instrumentos de uso diário no escritório, mesmo que para isso sejam necessárias algumas conversões de formato automatizadas.

O conteúdo produzido por um usuário é armazenado numa base de dados. É assim acessível e passível de sofrer alteração pelos usuários autorizados. Além de permanecer manipulável por todos os meios de tratamento informático, e poder ser distribuído a outros servidores por algum mecanismo de replicação, ou sindicalização (syndication).

Organizar a produção de conteúdo

A complexidade da produção de conteúdo cria não somente gargalos de estrangulamento que prejudicam a empresa, mas igualmente desencoraja e desmotiva a implicação dos empregados, clientes, e parceiros, fornecedores potenciais de conteúdo que agregam valor ao negócio. Com a correta gestão de conteúdo, qualquer

colaborador da empresa, detentor de informação, pode, dentro do seu perímetro de responsabilidade, produzir o seu conteúdo no site, sem intrometer-se no trabalho de colegas ou parceiros. Todos os atores da organização participam assim na vida da "empresa virtual" com certo grau de autonomia. Daí resulta a valorização do trabalho dos colaboradores e, por conseguinte, os lucros em produtividade e oportunidades comerciais para a empresa.

Gerir a qualidade da informação

A produção manual de documentos HTML pelo webmaster, que não conhece todos os aspectos dos ofícios ligados ao conteúdo, aumenta os riscos de erros. Da mesma maneira que, numa empresa, a divulgação de um relatório não validado, ou validado por pessoas não habilitadas pode conduzir a catástrofes.

A gestão de conteúdo permite que as informações postas em linha sigam um circuito de validação que reduz os riscos de erros de publicação (trâmite documental, ou workflow). Pode-se assim rejeitar um conteúdo, que uma vez corrigido, está novamente sujeito à validação, tudo com bastante agilidade. Torna-se também possível comentar um conteúdo, bem como acrescentar informações complementares ou expandir o seu contexto.

Outros benefícios importantes ligados à melhoria da qualidade da informação no site são: a normalização dos gabaritos de páginas; o acompanhamento da vida dos documentos no tempo; a possibilidade de volta a trás nas alterações realizadas, e a arquivagem automática. Estas funções garantem a melhor experiência do usuário com o site, beneficiando o incremento de sua audiência.

Interface de usuário configurável

A presença e a importância dos sistemas web na sociedade atual é tal que desenvolveu-se um novo ofício, o de especialista em "usabilidade" (usability), que cobre os aspectos gráficos e ergonômicos do sítio ou aplicação.

A gestão de conteúdo considera este importante aspecto dos sítios Web modernos, ou seja, a maneira como as funcionalidades são dispostas e apresentadas aos usuários e como a navegação funciona. Se o sítio se dirige a um público largo e diversificado, freqüentemente internacional, o funcionamento da disposição gráfica e a gestão da interface com o usuário não podem mais ser subestimadas.

Exercícios

1. Cite tecnologias existentes que influenciam a disseminação do conhecimento nas organizações contemporâneas.
2. Cite ferramentas (computacionais ou não) de transferência de conhecimento.
3. Dê exemplos de transmissão de conhecimento por socialização.
4. Dê exemplos de transmissão de conhecimento por combinação.
5. Dê exemplos de transmissão de conhecimento por externalização.
6. Dê exemplos de transmissão de conhecimento por internalização.
7. Por que os processos de socialização e internalização têm relação direta com a cultura organizacional?
8. Por que os processos de combinação e externalização têm relação direta com as tecnologias adotadas pela organização?

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

2.1. 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 açãoáveis.

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> >>

2.2. 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.

2.3. 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 bancos 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 que 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 que 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, enquanto 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 que 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 agrupa 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.

Leituras recomendadas:

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

2.4. 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.

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

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.

2.4.1. 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.

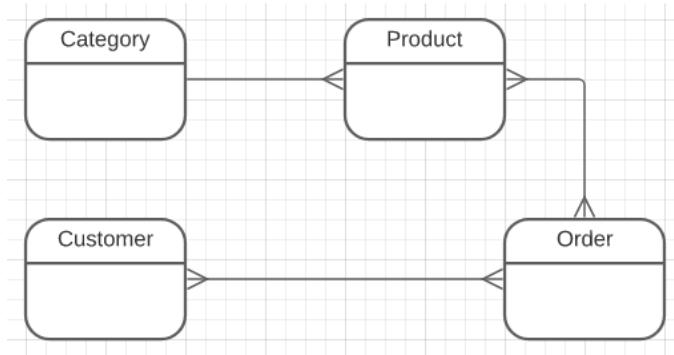


Figura 1 Objetos de Negócio da Entidade (design conceitual)

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.

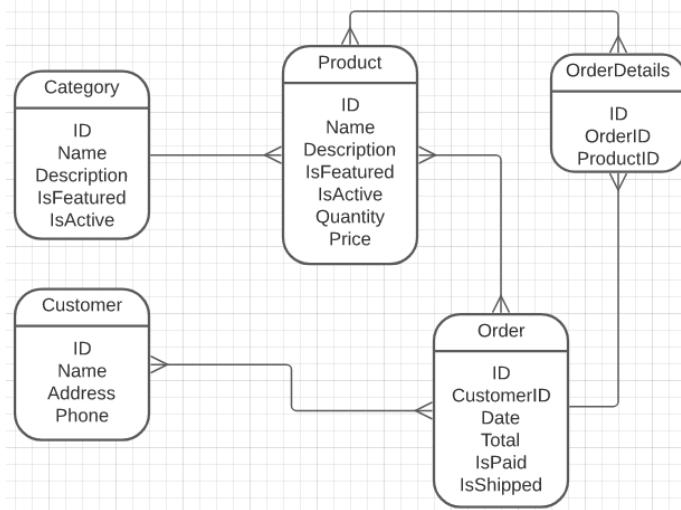


Figura 2 Objetos de Negócio da Entidade (design conceitual com atributos)

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.

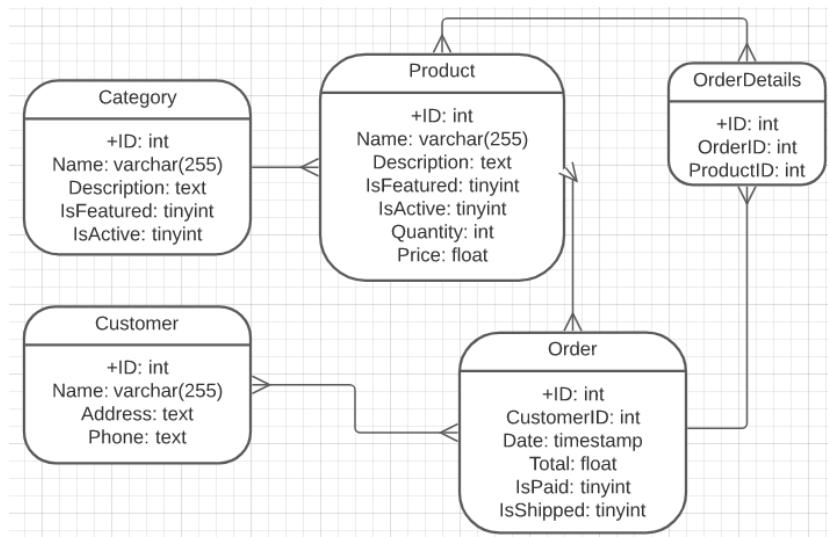


Figura 3 Objetos de Negócio da Entidade (modelo de dados físico)

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.

2.4.2. 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 KEY
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

3. Data Warehousing e Business Intelligence

3.1. 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.

3.1.1. 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.

3.1.2. 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).

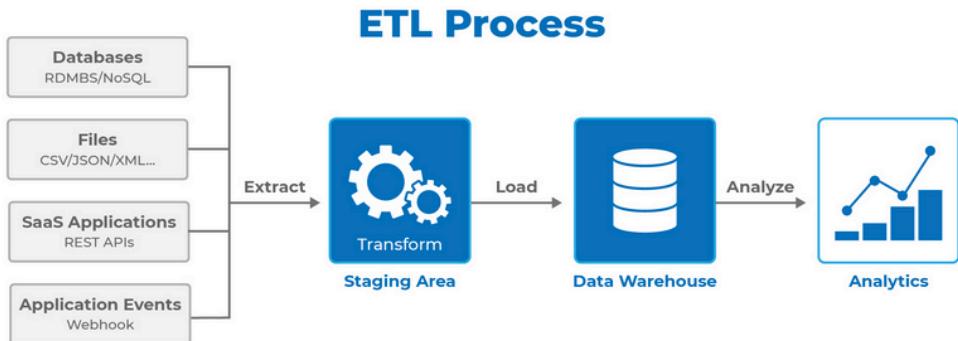


Figura 4 O processo ETL

3.1.3. 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.

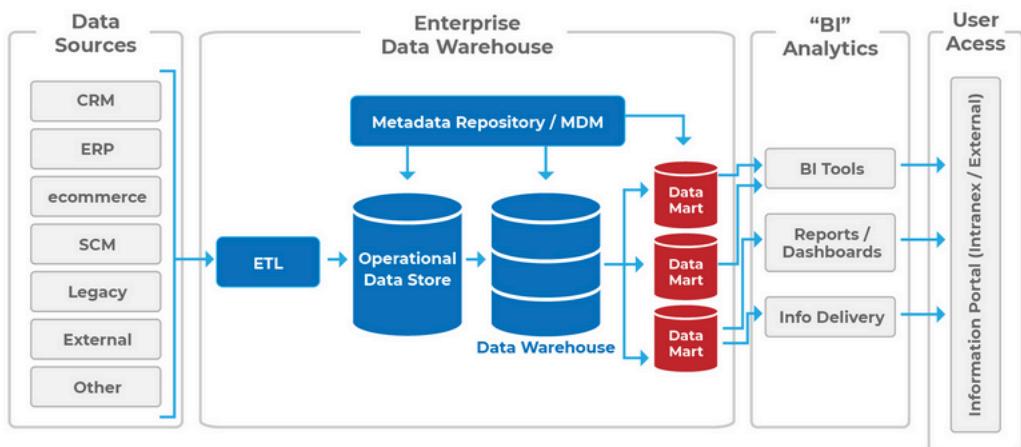


Figura 5 Data marts em uma arquitetura de data warehousing

3.1.4. 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.

3.1.5. 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

3.2. 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.

3.2.1. 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.

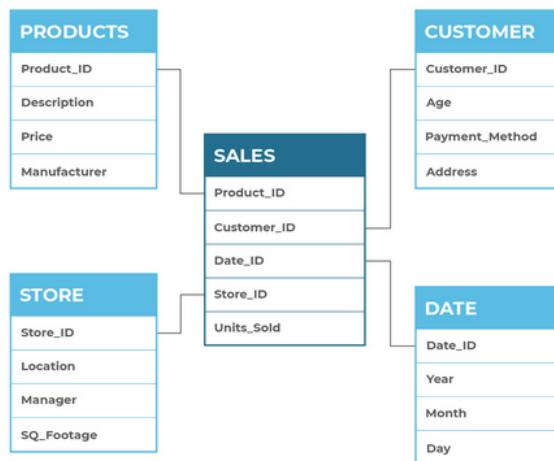


Figura 6 Exemplo de um esquema estrela

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.

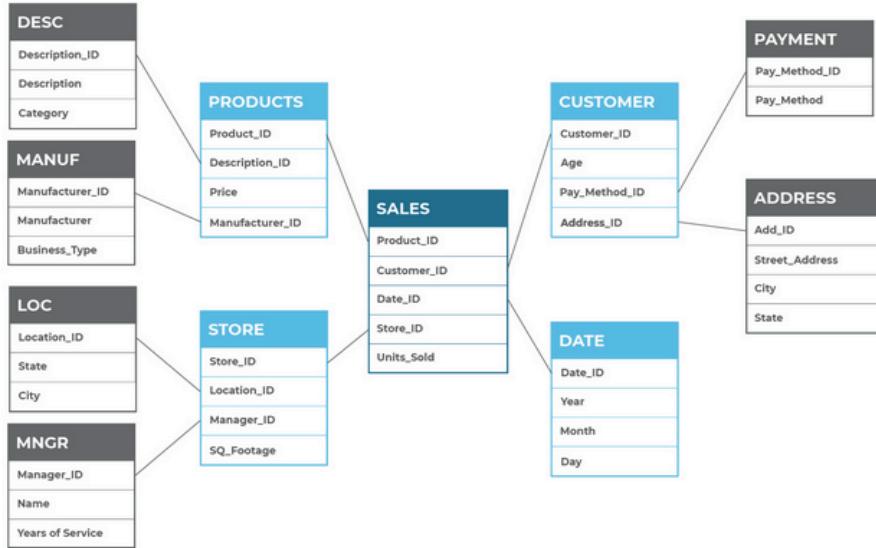


Figura 7 Exemplo de um esquema floco de neve (snowflake)

3.2.2. 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.

3.3. 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.

3.4. 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 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.



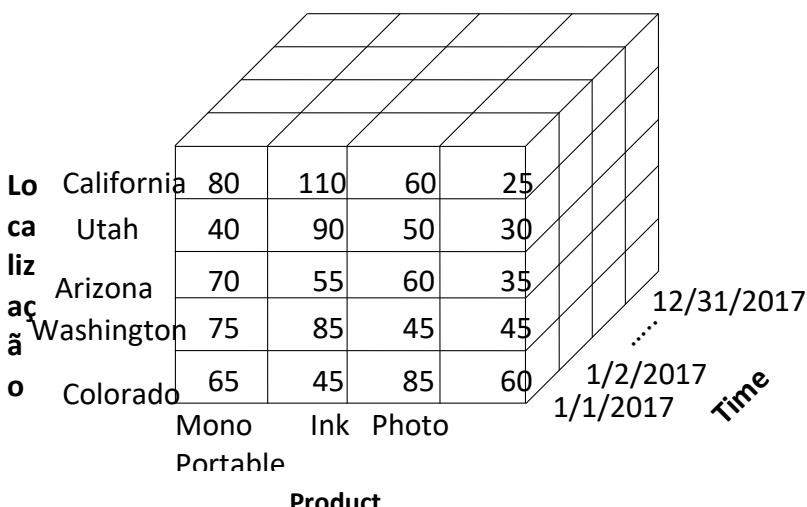
Esta variação de um diagrama de espinha de peixe mostra quatro fatores: gestão, localidade, mercado, e remuneração, os quais influenciam diretamente o turnover dos empregados. A perspectiva do analista de negócios dá uma ideia para a representação do DW. Uma representação de um DW deveria suportar este 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 tamanho da cidade ou tipo de plano de saúde oferecido. 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 Utha. 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.

3.4.1. 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				Time
	Mono Laser	Ink Jet	Photo	Portable	
California	80	110	60	25	12/31/2017
Utah	40	90	50	30
Arizona	70	55	60	35	1/2/2017
Washington	75	85	45	45	1/1/2017
Colorado	65	45	85	60	
	Mono Laser	Ink Jet	Photo	Portable	



(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

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

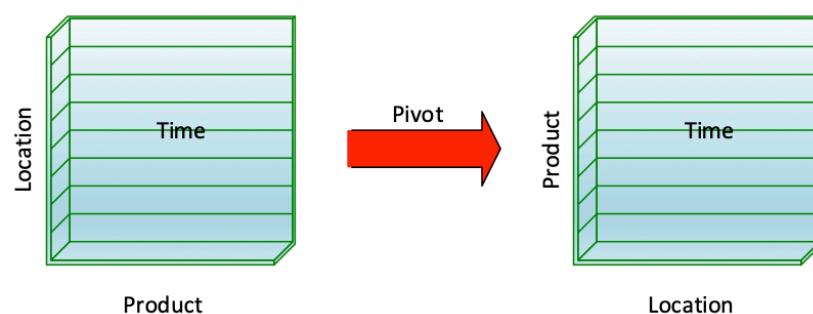
para um sinal de mais, para eliminar os detalhes das cidades e somar as métricas de valores nas cidades. O operador pivô suporta rearranjos nas dimensões em um cubo de dados. Por exemplo, as posições das dimensões produto e localidade podem ser invertidas no cubo de dados das vendas de modo que o produto 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.

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 navearem de um nível mais genérico para um nível mais específico, mais detalhado, como

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

3.5. 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

demandas 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.

3.6. 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 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 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.

3.6.1. 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 sim, 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.

3.6.2. 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 se usar. Uma abordagem mais complexa para valores desconhecidos é prever valores usando relacionamentos com outros campos.

Abordagens mais complexas irão prever os valores ausentes usando algorítimos 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 pró-ativas 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.

3.6.3. 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 cífrão, 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.

$^{\text{[a-z]}} + \backslash\text{.com\$}$

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 cífrão, 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 cífrão para identificar o final de uma string alvo.

Metacaracter	Tipo	Meaning
?	Iteração	Corresponde à ocorrência do caractere 0 ou 1 vez
*	Iteração	Corresponde à ocorrência do caractere 0 ou mais vezes
+	Iteração	Corresponde à ocorrência do caractere 1 ou mais vezes
{n}	Iteração	Corresponde à ocorrência do caractere exatamente n vezes
{n,m}	Iteração	Corresponde à ocorrência do caractere 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 caractere em uma expressão regular
^	Intervalo	Negação de um padrão de pesquisa se o ^ estiver dentro de []. Hífen dentro de [] define um intervalo de caracteres.
\$	Posição	Corresponde à ocorrência no fim de uma string alvo; 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 dificeis 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

3.6.4. 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		Source 2	
First name	Aimee	First name	Aimee
Middle name	Christina	Middle name	C.
Last name	Parker	Last name	Parker-Lewis
Job title	Product Manager	Job title	Prod. Mgr.
Firm	Microsoft Corporation	Firm	Microsoft
Street	15580 NE 31st Street	Street	16517 78 th Place NE
City	Redmond	City	Bothell
State	WA	State	WA
Postal Code	98052	Postal Code	98020
Country	USA	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.

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.

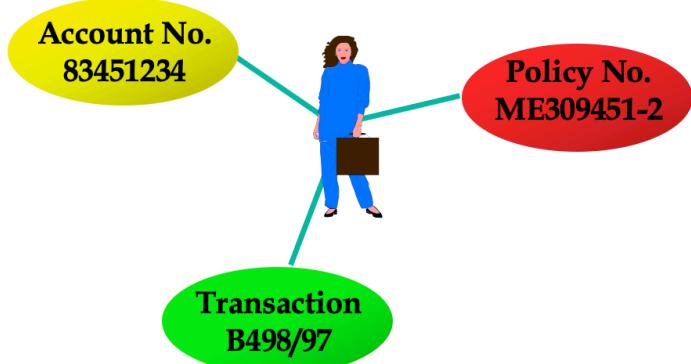
Para obter um conhecimento mais preciso de correspondência de entidades, é necessário entender os resultados da comparação de dois casos. Nesta matriz, as linhas representam previsões, e as colunas representam os resultados reais da correspondência de duas entidades. Uma correspondência verdadeira, 'true 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

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

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 sobrepondo 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.

3.6.5. 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

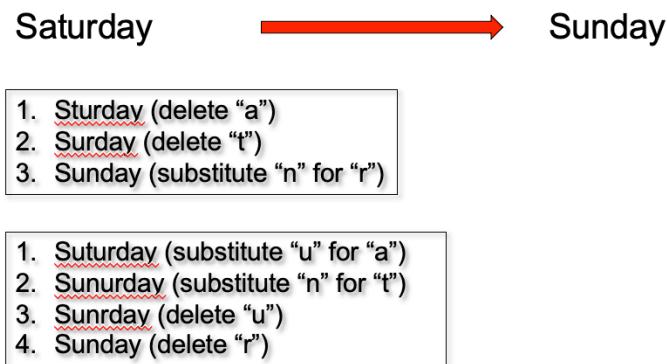
determinar se os quasi-identificadores de duas entidades indicam uma mesma entidade. No sentido geométrico, a distância é a quantidade de espaço entre dois pontos.

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 texto 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 quase-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.



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.

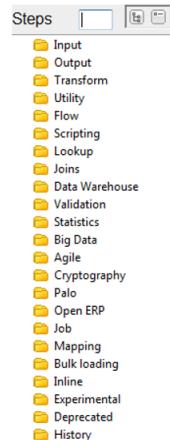
3.7. Pentaho Data Integration - PDI

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. O 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 conhecida 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.

- Step: process in a data flow
 - Input/Output  
 - Transform: sort, split, concatenate, ...  
 - Flow: filter rows  
 - Lookup: existence of rows, tables, files, ...  
 - Join: merge join, multiway merge, ...  
 - Validation: credit card, mail, data  
- Hop: directed connection between steps
- Database connections
- Distributed processing: partition, cluster, ...

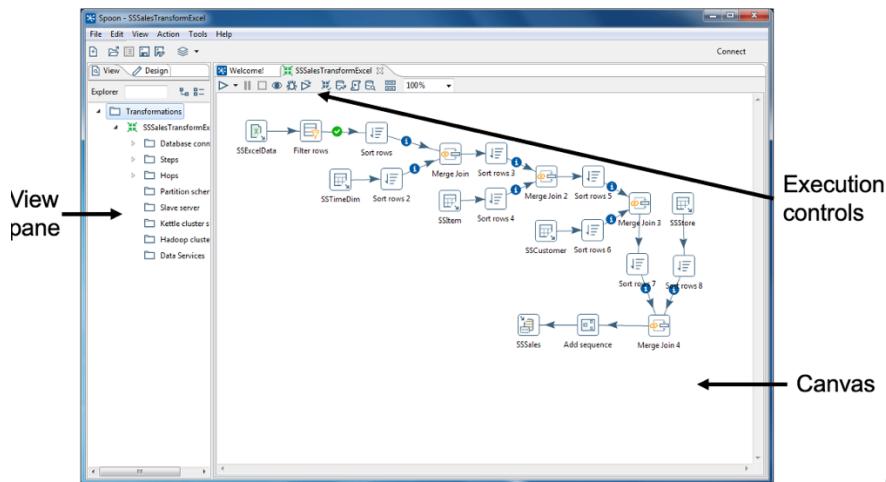


Um trabalho é um fluxo de dados de nível superior entre transformações e entidades externas. Kettle contém três componentes, Spoon fornece design gráfico de transformações e trabalhos, Pan executa transformações, enquanto Kitchen executa trabalhos. Uma transformação envolve etapas, saltos, conexões de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

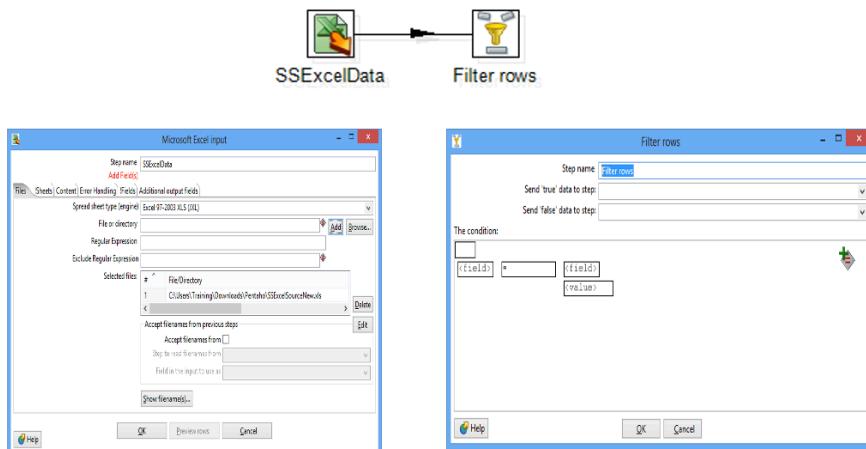
As etapas de fluxo reduzem sua fonte de dados aumentada, como filtrar linhas. As etapas de pesquisa testam a existência de linhas, tabelas, arquivos e outros objetos. As etapas de junção combinam fontes, como uma junção de mesclagem e mesclagem multiway.

Spoon IDE



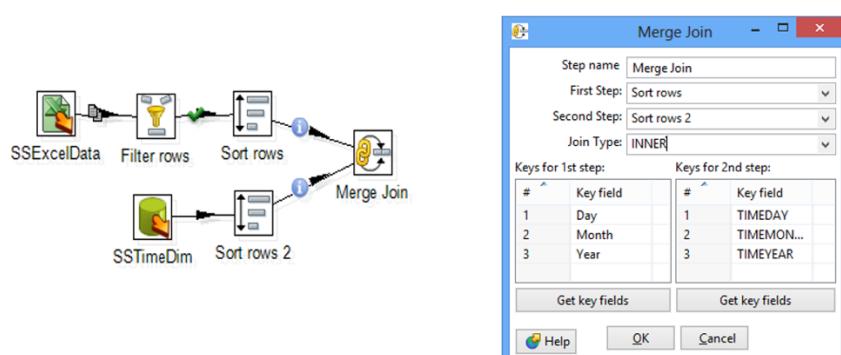
6

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 em 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/>>](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

4. A Natureza dos Dados

4.1. Análise de dados

A estatística desempenha um papel significativo nas ciências físicas e sociais. 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, onde 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 da 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 perguntassemos 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 onde 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.

4.2. 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	<ul style="list-style-type: none"> Modelo de dados pré-definidos Tipicamente textual Facilmente pesquisável Facilmente identificável por padrões 	<ul style="list-style-type: none"> Modelo de dados não estabelecido Pode ser texto, imagem, som, vídeo, etc Difícil de pesquisar Difícil de identificar padrão
Reside em	<ul style="list-style-type: none"> Bancos de dados relacionais Data Warehouses 	<ul style="list-style-type: none"> Aplicações Bancos de dados NoSQL Data Warehouse Data Lakes
Exemplos	<ul style="list-style-type: none"> Número de telefone Endereço de e-mail Número do CPF Informação de transação 	<ul style="list-style-type: none"> Imagens Áudio 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.

4.3. 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 impossível 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 7^a a 12^a 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. É esse valor numérico que pode ser encontrado no conjunto de dados. Você pode ver que os alunos da 7^a a 12^a 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.

4.4. 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.

5. 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.

5.1. 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

5.1.1. 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.

5.1.2. 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 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 de imagens? 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 um 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 sobre peso foi de 19,6%.

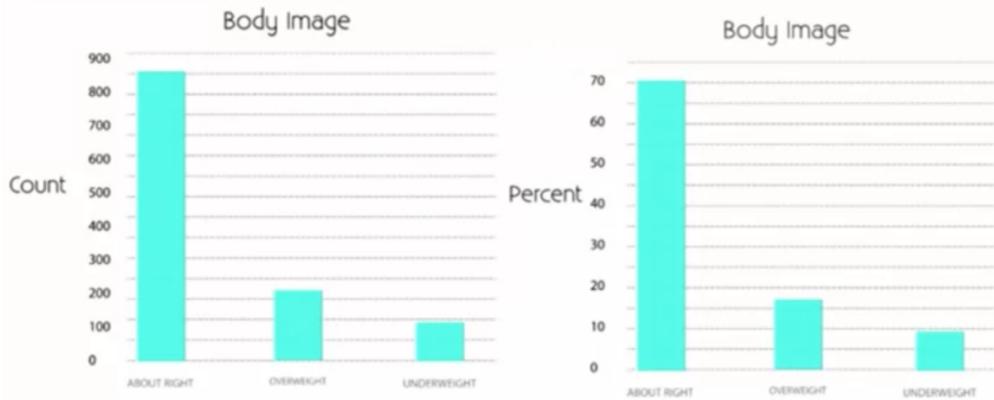
5.1.3. Plotando as distribuições

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.

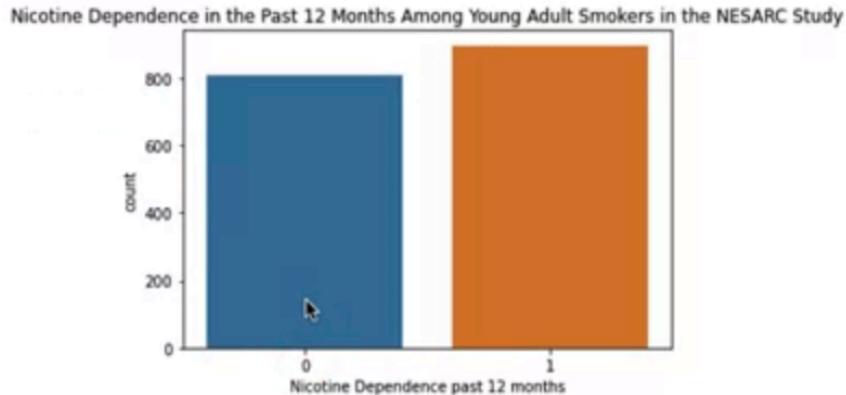


Para ilustrar, usaremos o dataset NESARC, buscando interpretar as relações entre a dependência de nicotina nos últimos meses (representado no dataset pela variável TAB12MDX) e a estimativa de cigarros fumados por mês (representado por NUMCIG_EST). Vamos executar distribuições de frequência para cada 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.

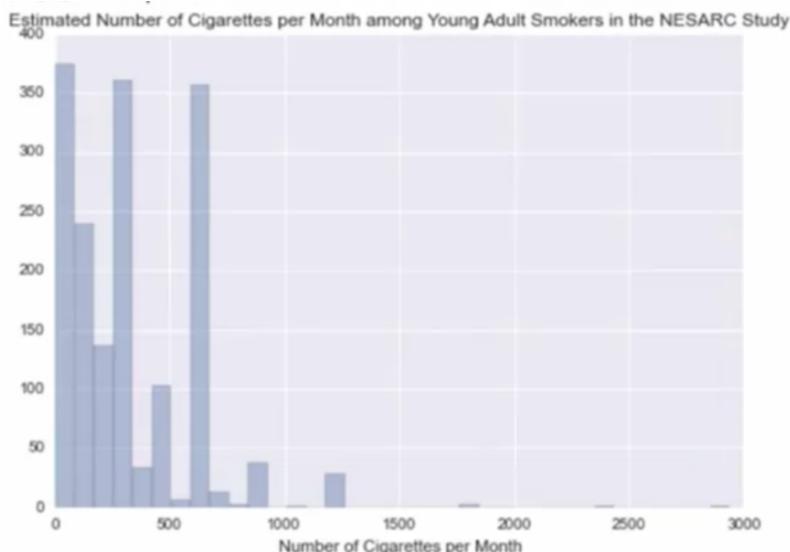
Ao visualizar dados em Python, precisaremos importar bibliotecas em nosso programa. Primeiro, vamos importar o seaborn pacote com a sintaxe **import seaborn**. Também precisamos importar a biblioteca **matplotlib.pyplot** porque o pacote seaborn é dependente neste pacote para criar gráficos. Porque o 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.pyplot, 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.

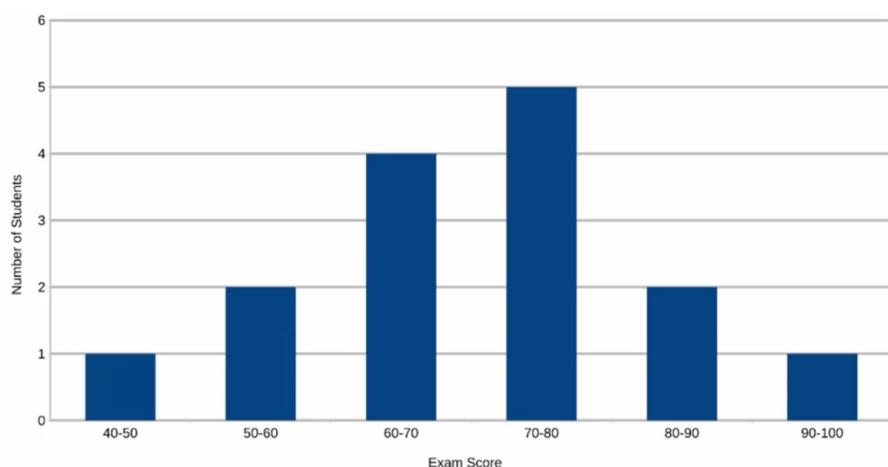
88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

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.

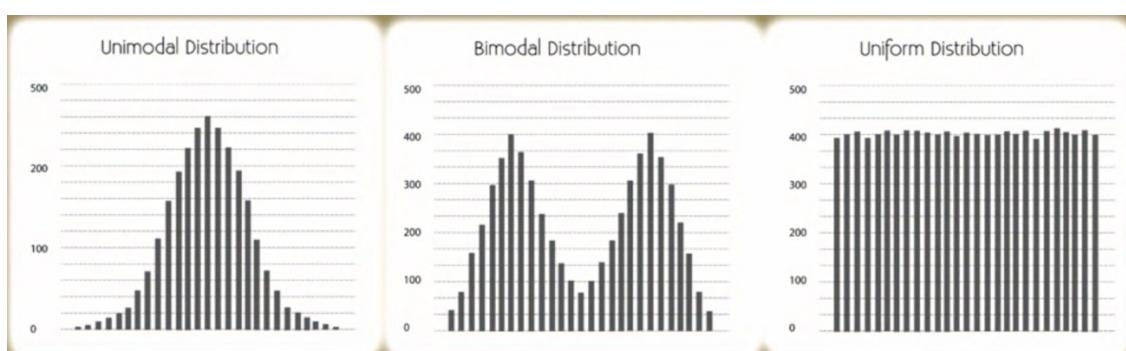
Para construir o histograma a partir desta tabela, os intervalos são plotados no eixo X e mostram o número de observações em cada intervalo, ou a porcentagem de observações em cada intervalo no eixo Y, que é representada pela altura da barra localizada acima do intervalo.

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

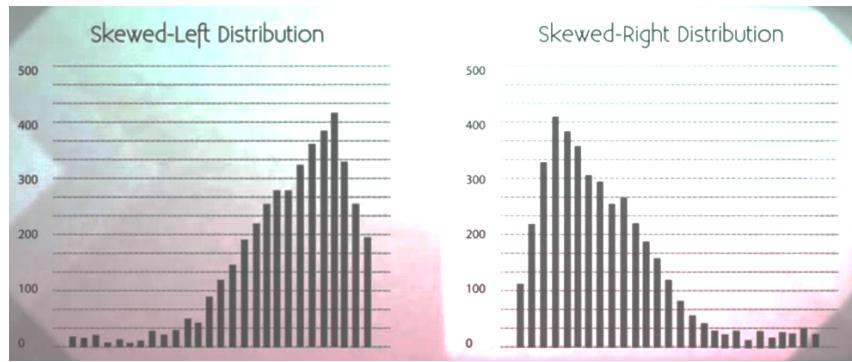
Uma vez que a distribuição tenha sido exibida graficamente como um histograma, podemos descrever o padrão geral da distribuição e mencionar quaisquer desvios marcantes desse padrão. Mais especificamente, devemos considerar os seguintes recursos. Teremos uma noção do padrão geral dos dados do centro dos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

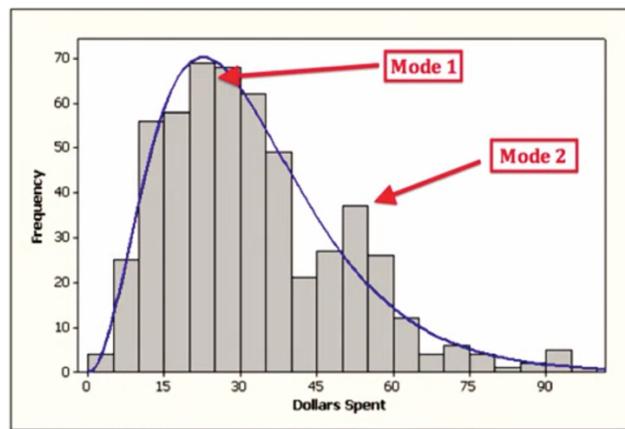


Uma distribuição é chamada skewed-right (distorcida à direita). Com a cauda direita, os valores maiores são muito mais longos do que a cauda esquerda, ou valores menores. Note que em uma distribuição assimétrica à direita, como você pode ver aqui à direita. A maior parte das observações é pequena a média, 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.

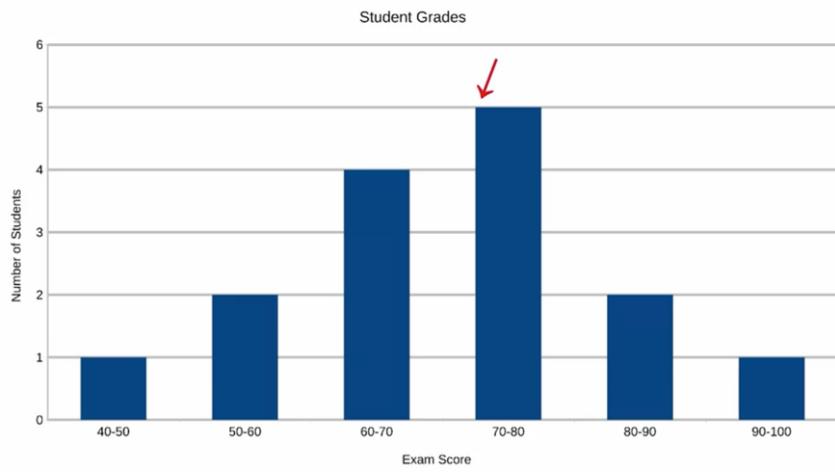
Distribuições distorcidas também podem ser bimodais. Aqui está um exemplo, um bairro de tamanho médio 24 horas loja de conveniência coletou dados de 537 clientes sobre a quantidade de dinheiro que gastaram em uma única visita à loja. Observe o histograma abaixo. Você pode ver que a quantidade de dinheiro 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 dispersão.

5.1.4. 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 quisermos 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.

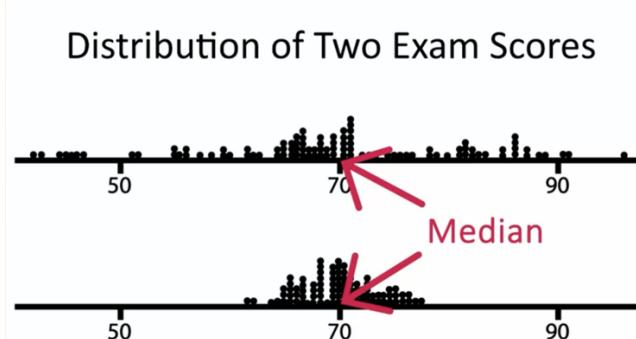
Aviso, se o jogador de dardo tivesse marcado, 19 em vez de 13. A média aumenta para 7, mas a moda e a pontuação mediana permanecem inalterados.

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. Nossa 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 quanto 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

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

Para encontrar o desvio padrão do número de chamadas por hora, primeiro encontrariam 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, subtraíramos 9 de cada uma de nossas observações.

$$\begin{array}{r}
 7 \quad 9 \quad 5 \quad 13 \quad 3 \quad 11 \quad 15 \quad 9 \\
 - \quad - \\
 \frac{9}{-2} \quad \frac{9}{0} \quad \frac{9}{-4} \quad \frac{9}{4} \quad \frac{9}{-6} \quad \frac{9}{2} \quad \frac{9}{6} \quad \frac{9}{0}
 \end{array}$$

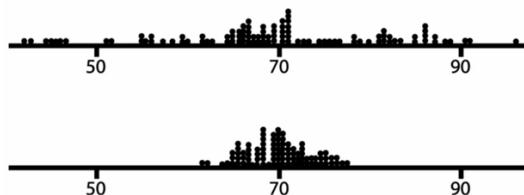
Como um terceiro passo, nós elevaríamos ao quadrado cada um desses desvios. Em seguida, medimos os desvios quadrados adicionando-os e dividindo-os por N-1, que é um a menos do que o tamanho amostral, esta média dos desvios quadrados é chamada de variância. O desvio padrão de sua variável é a raiz quadrada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

Variable Distribution in Two Samples



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

```

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

```

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 \pm 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.

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 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.

```

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)

```

```

69
70 print('describe nicotine dependence')
71 desc2 = sub2['TAB12MDX'].describe()
72 print(desc2)
73

```

	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 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.

```

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

```

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.

5.2. 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.

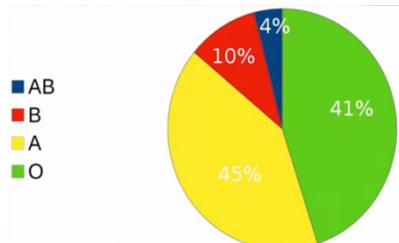
Uma amostra aleatória pode ser uma representação bastante precisa da população maior, enquanto outra amostra aleatória pode não ser precisa, puramente devido ao acaso. Infelizmente, ao olhar para uma amostra aleatória específica, que é o que acontece nas estatísticas, nunca saberemos o quanto que amostra 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.

5.2.1. 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?

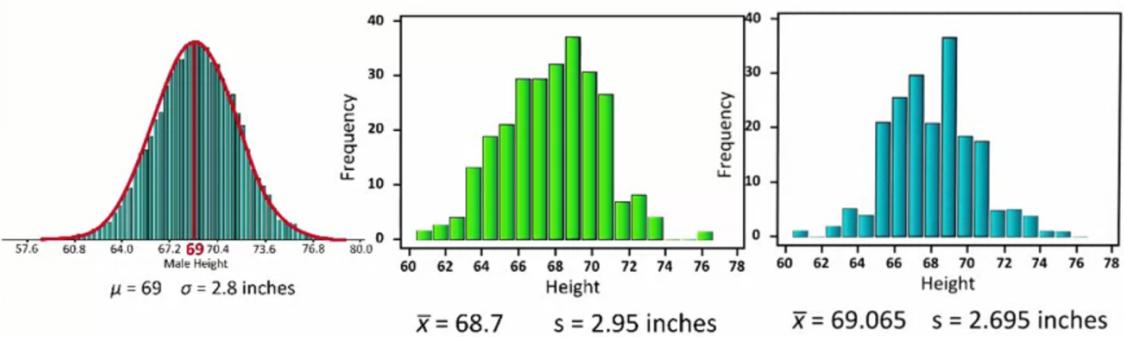
Quero dizer, já que pegamos uma amostra de apenas 500 indivíduos, não podemos esperar que nossa amostra se comporte exatamente como a população. Mas se a amostra é aleatória, e este foi, esperamos obter resultados que não são tão diferentes dos resultados de toda a população e isso é o que encontramos. 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.



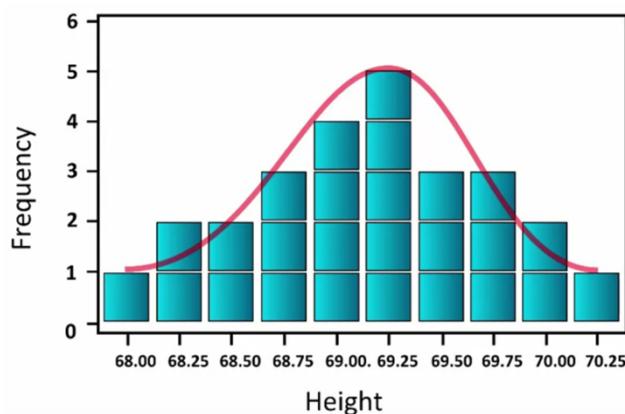
Em ambos os exemplos, temos números que descrevem a população e números que descrevem a amostra.

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.

5.2.2. 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 seremos 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, estariam 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 estariam 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 estariam 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?

5.2.3. 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 estariam 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, estariámos errados menos de 3 em cada 100 vezes. Estariámos 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.

5.2.4. 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

5.2.5. 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.

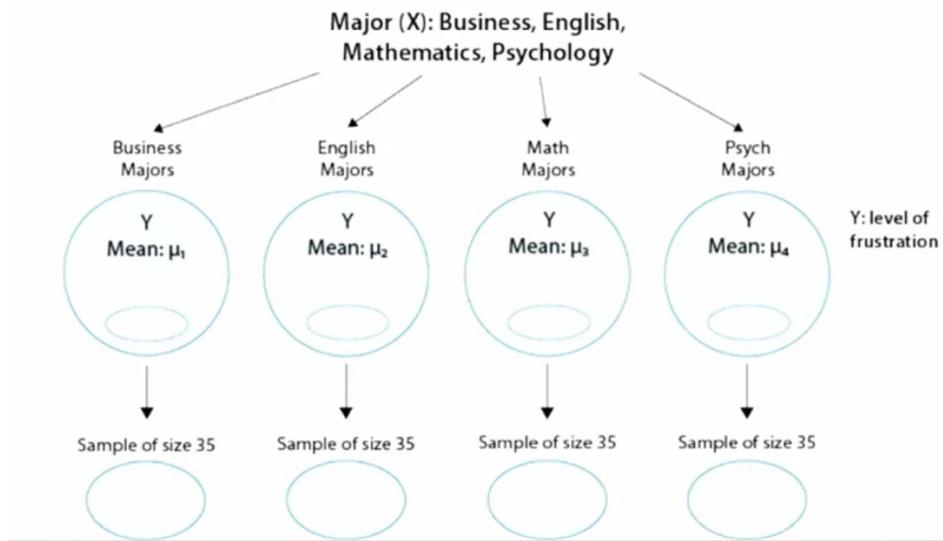
Na descrição do teste de hipóteses, quando analisamos a associação entre depressão e tabagismo, estávamos trabalhando com uma variável explicativa categórica, a presença ou ausência de depressão, e uma variável de resposta quantitativa, o número de cigarros fumados por mês. Quando você está testando 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 significar não ser igual. Falaremos sobre isso mais tarde.

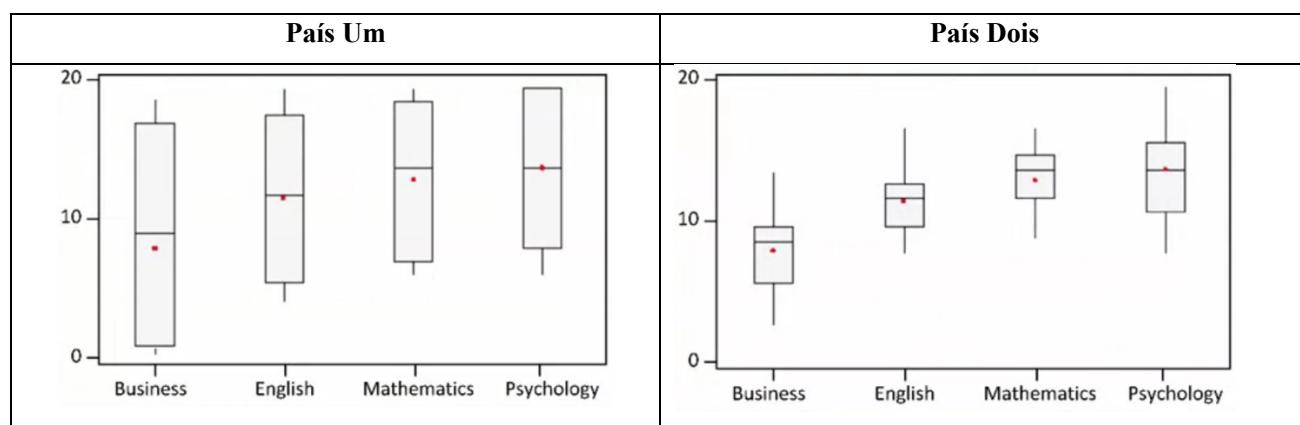
$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a: \text{not all the } \mu \text{ are equal}$$

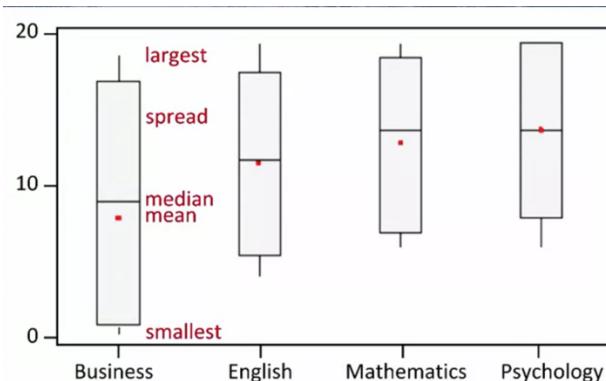
Por enquanto, vamos pensar sobre como iríamos testar se a população significa são iguais. Poderíamos calcular o nível médio de frustração para cada major e ver quão distantes essas médias de amostra estão. Ou, em outras palavras, meça a variação entre as médias da amostra. Se descobrirmos que as quatro 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: $y_1 = 7,3$.
- A pontuação média de frustração da amostra para os 35 alunos de inglês é: $y_2 = 11,8$.
- A pontuação média de frustração da amostra para os 35 alunos de matemática é: $y_3 = 13,2$.
- E a pontuação média de frustração da amostra para os 35 alunos de Psicologia é: $y_4 = 14,0$.

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.



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 um dos 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 ser falsa.

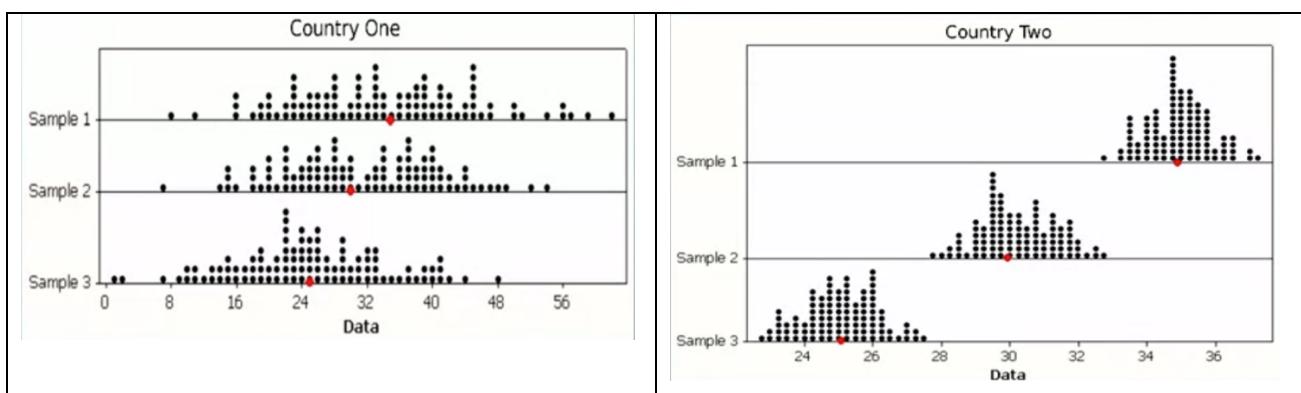
A pergunta que precisamos responder com o Teste ANOVA F é, as diferenças entre as médias da amostra devido a verdadeiras diferenças entre as médias da população, ou meramente devido à variabilidade amostral? Para responder a esta pergunta, usando nossos dados, obviamente precisamos olhar para a variação 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 = Variação entre as médias das amostras

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 fornecem 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 da 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			
$S = 2.593 \quad R-Sq = 50.69\% \quad R-Sq = (adj) = 49.60\%$					
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

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^c$$

Where c = # of comparisons, α =normal Type 1 Error (.05)

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 quanto 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 multicomp 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.

```
03
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 etnia branca, 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	

5.2.6. 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 proibia a venda de 3,2 cerveja, homens com menos de 21 anos de idade. Mas permitiu que fosse venda 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 quanto 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 quanto 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	404	408.7	481
Female	16	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 quanto longe o que observamos, que é os dados são. Do que esperaríamos observar, se a hipótese nula fosse verdadeira. Aqui está a fórmula.

$$\chi^2 = \sum_{\text{all cells}} \frac{(Observed Count - Expected Count)^2}{Expected Count}$$

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 $X^2 = 1,62$. Existe uma diferença relativamente grande entre o que observamos e o que a hipótese nula afirma? Ou relativamente pequena? 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-la. 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.

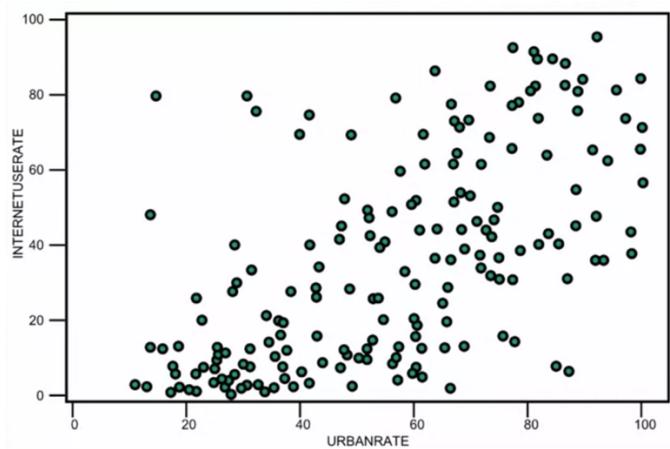
Não há evidência estatística convincente para rejeitar a hipótese nula. E assim continuaremos a assumir que pode ser verdade. Gênero e condução embriagada podem ser independentes. E assim os dados sugerem que

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.

5.2.7. Teste de Correlação de Pearson

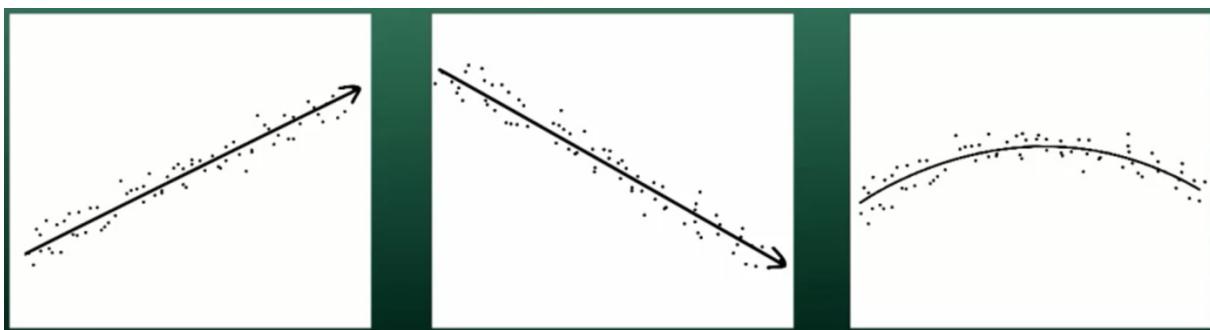
Análise de Variância examina a relação entre uma variável explicativa categórica e uma variável de resposta quantitativa, na qual analisamos a primeira ferramenta inferencial. O teste de independência Qui Quadrado é uma ferramenta inferencial que examina a relação entre dois valores categóricos. Se você 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 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.

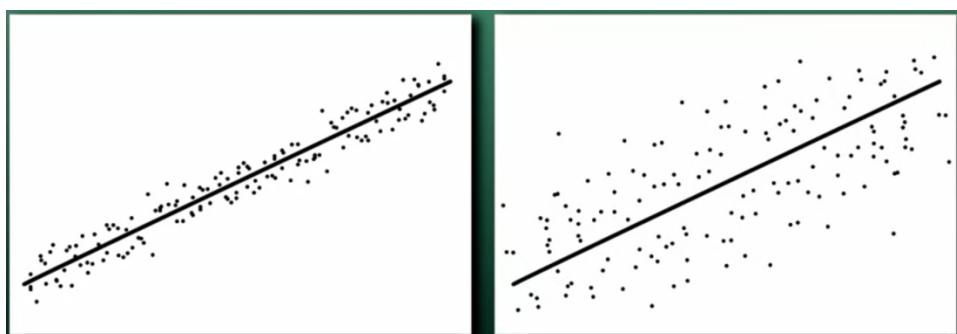


A direção do relacionamento pode ser positiva, negativa ou nenhuma delas. Um relacionamento positivo, ou crescente, significa que um aumento em uma das variáveis está associado a um aumento na outra. Um negativo, ou diminuição no relacionamento significa que um aumento em uma das variáveis está associado 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.](https://www.digitalvidya.com/blog/descriptive-vs-inferential-statistics/) (7 min)

<https://www.digitalvidya.com/blog/descriptive-vs-inferential-statistics/>

Article: Laerd Statistics. (n.d.). [Descriptive and Inferential Statistics.](https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php) (5 min)

<https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php>

Video: The Organic Chemistry Tutor. (2019). [Descriptive Statistics vs Inferential Statistics.](https://www.youtube.com/watch?v=VHYOuWu9jQI) (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

6. Business Intelligence e Visual Analytics

6.1. 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.

6.2. 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 Dados⁴. Esta ferramenta organiza visualizações por várias funções, por exemplo,

⁴ 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 Viya⁵. 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.⁶

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

⁵ SAS Visual Analytics and SAS Viya - https://www.sas.com/en_us/software/visual-analytics.html

⁶ 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> >>

6.3. Uma visão geral do Public Tableau ou Power BI???