2 O processo de ciência de dados

Este capítulo cobre

ÿ Compreender o fluxo de um processo de ciência de dados

ÿ Discutir as etapas de um processo de ciência de dados

O objetivo deste capítulo é fornecer uma visão geral do processo de ciência de dados sem mergulhar ainda no big data. Você aprenderá como trabalhar com conjuntos de big data, dados de streaming e dados de texto nos capítulos subsequentes.

2.1 Visão geral do processo de ciência de dados

Seguir uma abordagem estruturada para ciência de dados ajuda você a maximizar suas chances de sucesso em um projeto de ciência de dados com o menor custo. Também torna possível assumir um projeto em equipe, com cada membro da equipe se concentrando no que faz de melhor. Porém, tome cuidado: essa abordagem pode não ser adequada para todos os tipos de projeto ou ser a única maneira de fazer uma boa ciência de dados.

O processo típico de ciência de dados consiste em seis etapas pelas quais você iterará comeu, conforme mostrado na figura 2.1.

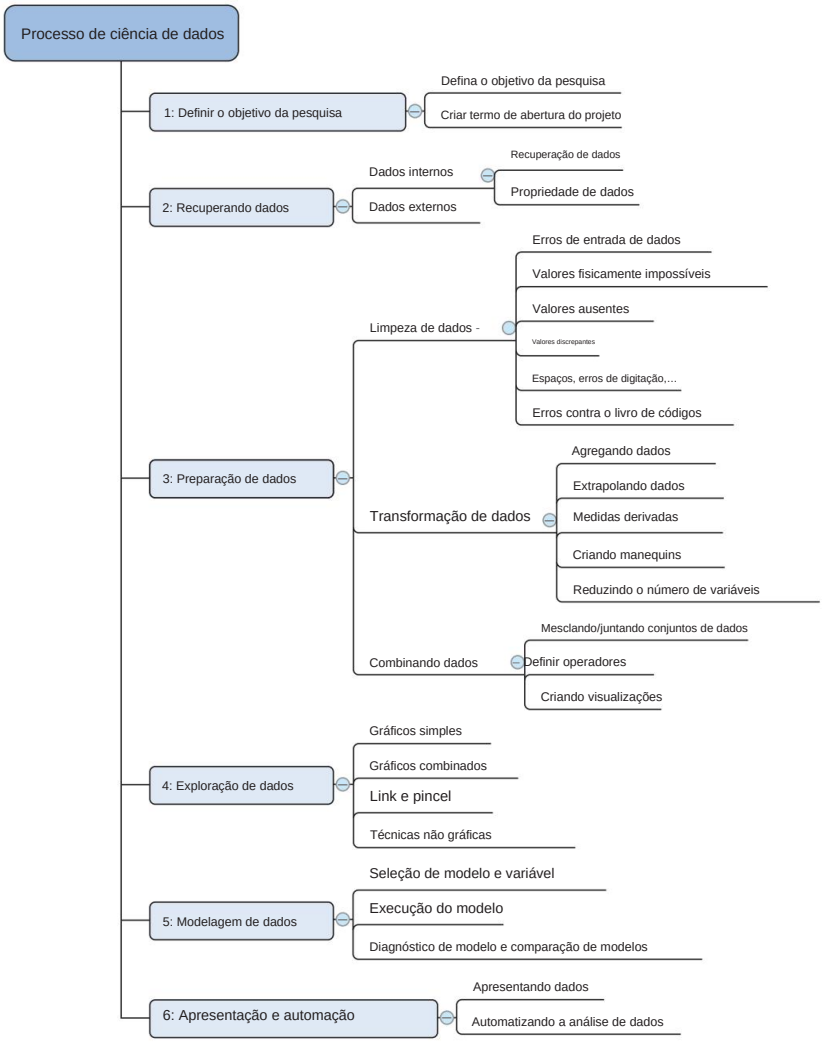


Figura 2.1 As seis etapas do processo de ciência de dados

A Figura 2.1 resume o processo de ciência de dados e mostra as principais etapas e ações que você executará durante um projeto. A lista a seguir é uma breve introdução; cada uma das etapas será discutida com maior profundidade ao longo deste capítulo.

1 O primeiro passo deste processo é definir um *objetivo de pesquisa*. O principal objetivo aqui é garantir que todas as partes interessadas entendam o *quê, como* e *por quê* do projeto. Em todo projeto sério isso resultará em um termo de abertura do projeto.

2 A segunda fase é *a recuperação de dados.* Você deseja ter dados disponíveis para análise, então esta etapa inclui encontrar dados adequados e obter acesso aos dados do

proprietário dos dados. O resultado são dados em sua forma bruta, que provavelmente precisam de polimento e transformação antes de se tornar utilizável.

3 Agora que você tem os dados brutos, é hora de *prepará* -los. Isso inclui transformar os dados de um formato bruto em dados que podem ser usados diretamente em seus modelos. Para Para conseguir isso, você detectará e corrigirá diferentes tipos de erros nos dados, combinará dados de diferentes fontes de dados e os transformará. Se você tiver sucesso Concluída esta etapa, você poderá avançar para a visualização e modelagem de dados.

4 A quarta etapa é *a exploração de dados*. O objetivo desta etapa é obter uma compreensão profunda dos dados. Você procurará padrões, correlações e desvios com base em técnicas visuais e descritivas. Os insights que você obtém nesta fase serão permitir que você comece a modelar.

5 Finalmente, chegamos à parte mais interessante: *a construção de modelos* (muitas vezes referida como "modelagem de dados" neste livro). É agora que você tenta obter os insights ou faça as previsões indicadas no termo de abertura do projeto. Agora é a hora de trazer armas pesadas, mas lembre-se que a pesquisa nos ensinou que muitas vezes (mas não sempre) uma combinação de modelos simples tende a superar um complicado modelo. Se você executou esta fase corretamente, está quase terminando.

6 A última etapa do modelo de ciência de dados é *apresentar seus resultados e automatizar o análise,* se necessário. Um objetivo de um projeto é mudar um processo e/ou fazer melhores decisões. Talvez você ainda precise convencer a empresa de que suas descobertas de fato mudará o processo de negócios conforme o esperado. É aqui que você pode brilhe em seu papel de influenciador. A importância desta etapa é mais evidente em projetos em nível estratégico e tático. Certos projetos exigem que você execute o processo de negócios repetidamente, portanto, automatizar o projeto irá economizar tempo.

Na realidade, você não progredirá de forma linear do passo 1 ao passo 6. Freqüentemente, você regredirá e iterar entre as diferentes fases.

Seguir estas seis etapas compensa em termos de uma maior taxa de sucesso do projeto e aumento do impacto dos resultados da investigação. Este processo garante que você tenha um plano bem definido plano de pesquisa, uma boa compreensão da questão do negócio e resultados claros antes mesmo de começar a analisar os dados. As primeiras etapas do seu processo se concentram em obter dados de alta qualidade como entrada para seus modelos. Desta forma, seus modelos terão melhor desempenho mais tarde. Na ciência de dados há um ditado bem conhecido: *Lixo que entra é igual a lixo que sai.*

Outro benefício de seguir uma abordagem estruturada é que você trabalha mais no *modo protótipo* enquanto procura o melhor modelo. Ao construir um *protótipo,* você provavelmente tentará vários modelos e não se concentrará muito em questões como programa velocidade ou escrever código de acordo com os padrões. Isso permite que você se concentre em agregar valor ao negócio.

Nem todo projeto é iniciado pela própria empresa. Os insights aprendidos durante a análise ou a chegada de novos dados podem gerar novos projetos. Quando a equipe de ciência de dados gera uma ideia, já foi feito trabalho para fazer uma proposta e encontrar uma patrocinador empresarial.

Dividir um projeto em etapas menores também permite que os funcionários trabalhem juntos como um equipe. É impossível ser especialista em tudo. Você precisaria saber como carregue todos os dados para todos os bancos de dados diferentes, encontre um esquema de dados ideal que funciona não apenas para sua aplicação, mas também para outros projetos dentro de sua empresa, e depois acompanhar todas as técnicas estatísticas e de mineração de dados, ao mesmo tempo que um especialista em ferramentas de apresentação e política empresarial. Essa é uma tarefa difícil, e é por isso cada vez mais empresas contam com uma equipe de especialistas em vez de tentar encontrar uma pessoa que pode fazer tudo.

O processo que descrevemos nesta seção é mais adequado para um projeto de ciência de dados que contém apenas alguns modelos. Não é adequado para todos os tipos de projeto. Por exemplo, um projeto que contém milhões de modelos em tempo real precisaria de uma abordagem diferente do que o fluxo que descrevemos aqui. No entanto, um cientista de dados iniciante deve percorrer um longo caminho seguindo essa maneira de trabalhar.

2.1.1 Não seja escravo do processo

Nem todo projeto seguirá esse modelo, porque seu processo está sujeito às preferências do cientista de dados, da empresa e à natureza do projeto em que você trabalha.
Algumas empresas podem exigir que você siga um protocolo rígido, enquanto outras têm um forma mais informal de trabalhar. Em geral, você precisará de uma abordagem estruturada quando você trabalha em um projeto complexo ou quando muitas pessoas ou recursos estão envolvidos.

O modelo *ágil* de projeto é uma alternativa a um processo sequencial com iterações. Como essa metodologia ganha mais espaço no departamento de TI e em toda a empresa, também está sendo adotada pela comunidade de ciência de dados. Embora a metodologia ágil seja adequada para um projeto de ciência de dados, muitas políticas da empresa favorecerão uma abordagem mais abordagem rígida em relação à ciência de dados.

Planejar antecipadamente todos os detalhes do processo de ciência de dados nem sempre é possível, e na maioria das vezes você irá iterar entre as diferentes etapas do processo.
Por exemplo, após o briefing você inicia seu fluxo normal até chegar à fase exploratória de análise de dados. Seus gráficos mostram uma distinção no comportamento entre dois grupos – homens e mulheres, talvez? Você não tem certeza porque não tem uma variável que indique se o cliente é homem ou mulher. Você precisa recuperar um conjunto de dados extra para confirmar isso. Para isso você precisa passar pelo processo de aprovação, o que indica que você (ou a empresa) precisa fornecer uma espécie de termo de abertura do projeto. Nas grandes empresas, obter todos os dados necessários para finalizar seu projeto pode ser uma provação.

2.2 Passo 1: Definir objetivos de pesquisa e criar uma carta de projeto

Um projeto começa com a compreensão do *quê,* do *porquê* e do *como* do seu projeto (figura 2.2). O que a empresa espera que você faça? E por que a gestão coloca tanto valor em sua pesquisa? É parte de um quadro estratégico mais amplo ou de um “lobo solitário” projeto originado de uma oportunidade que alguém detectou? Respondendo a esses três

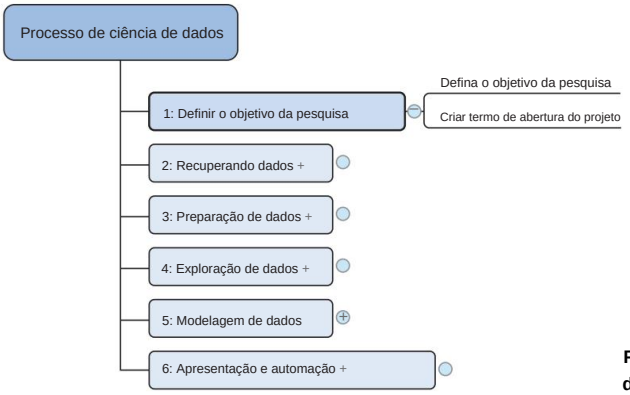


Figura 2.2 Etapa 1: Definição do objetivo da pesquisa

perguntas (o que, por que, como) é o objetivo da primeira fase, para que todos saibam o que fazer e podem chegar a acordo sobre o melhor curso de ação.

O resultado deve ser um objectivo de investigação claro, uma boa compreensão do contexto, resultados bem definidos e um plano de acção com um calendário. Essa informação é então melhor colocado em um termo de abertura do projeto. A duração e a formalidade podem, é claro, diferir entre projetos e empresas. Nesta fase inicial do projeto, as habilidades das pessoas e perspicácia empresarial são mais importantes do que grande habilidade técnica, e é por isso que este parte será frequentemente orientada por pessoal mais experiente.

2.2.1 Passe algum tempo entendendo os objetivos e o contexto de sua pesquisa

Um resultado essencial é o objetivo da pesquisa que declara o propósito do seu trabalho de forma clara e focada. Compreender os objetivos e o contexto do negócio é fundamental para o sucesso do projeto. Continue fazendo perguntas e criando exemplos até que você compreender as expectativas exatas do negócio, identificar como seu projeto se encaixa no cenário geral, avaliar como sua pesquisa mudará o negócio e entender como eles usarão seus resultados. Nada é mais frustrante do que passar meses pesquisando algo até ter aquele momento de brilho e resolver o problema, mas quando você relata suas descobertas à organização, todos percebem imediatamente que você entendeu mal a pergunta. Não ignore esta fase levemente. Muitos cientistas de dados falham aqui: apesar de sua inteligência matemática e brilhantismo científico, eles nunca parecem compreender os objetivos e o contexto do negócio.

2.2.2 Criar um termo de abertura do projeto

Os clientes gostam de saber antecipadamente quanto estão pagando; portanto, depois de compreender bem o problema do negócio, tente obter um acordo formal sobre as entregas.

Todas essas informações são melhor coletadas em um termo de abertura do projeto. Para qualquer projeto significativo isso seria obrigatório.

Um termo de abertura do projeto requer trabalho em equipe e sua contribuição abrange pelo menos o seguinte:

ÿ Um objetivo de pesquisa claro
ÿ A missão e o contexto do projeto ÿ Como você realizará sua análise ÿ Que recursos você espera usar ÿ Prova de que é um projeto alcançável ou prova de conceitos ÿ Resultados e uma medida de sucesso

ÿ Uma linha do tempo

Seu cliente pode usar essas informações para fazer uma estimativa dos custos do projeto e dos dados e pessoas necessárias para que seu projeto se torne um sucesso.

2.3 Etapa 2: Recuperando dados

O próximo passo na ciência de dados é recuperar os dados necessários (figura 2.3). Às vezes você precisa ir a campo e projetar você mesmo um processo de coleta de dados, mas na maioria das vezes você não estará envolvido nesta etapa. Muitas empresas já terão coletado e armazenado os dados para você, e o que elas não possuem pode muitas vezes ser comprado de terceiros. Não tenha medo de procurar dados fora de sua organização, porque cada vez mais organizações estão disponibilizando gratuitamente até mesmo dados de alta qualidade para uso público e comercial.

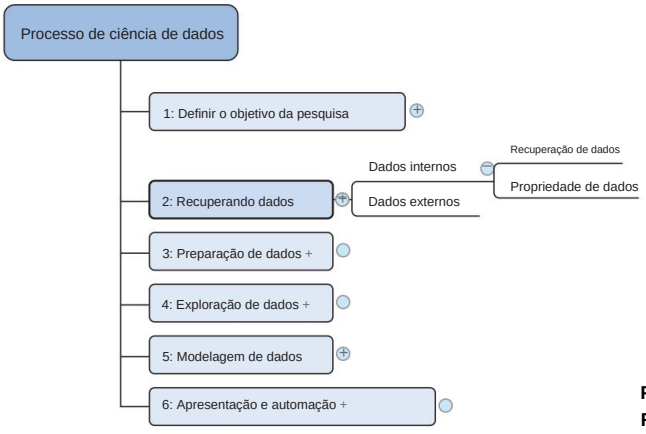


Figura 2.3 Etapa 2: Recuperando dados

Os dados podem ser armazenados de várias formas, desde simples arquivos de texto até tabelas em um banco de dados. O objetivo agora é adquirir todos os dados que você precisa. Isso pode ser difícil e, mesmo que você tenha sucesso, os dados geralmente são como um diamante bruto: eles precisam de polimento para serem úteis para você.

2.3.1 Comece com dados armazenados dentro da empresa

Seu primeiro ato deve ser avaliar a relevância e a qualidade dos dados que estão prontamente disponíveis em sua empresa. A maioria das empresas possui um programa para manutenção de dados importantes, portanto, grande parte do trabalho de limpeza já pode ter sido feito. Esses dados podem ser armazenados em repositórios de dados oficiais, como *bancos de dados, data marts, data warehouses* e *data lakes* mantidos por uma equipe de profissionais de TI . O objetivo principal de um banco de dados é o armazenamento de dados, enquanto um data warehouse é projetado para ler e analisar esses dados. Um data mart é um subconjunto do data warehouse voltado para atender uma unidade de negócios específica. Embora os data warehouses e data marts abriguem dados pré-processados, os data lakes contêm dados em seu formato natural ou bruto. Mas existe a possibilidade de que seus dados ainda residam em arquivos Excel na área de trabalho de um especialista no domínio.

Encontrar dados mesmo dentro da sua própria empresa às vezes pode ser um desafio. À medida que as empresas crescem, seus dados ficam espalhados por muitos lugares. O conhecimento dos dados pode ser disperso à medida que as pessoas mudam de cargo e saem da empresa. Documentação e metadados nem sempre são a principal prioridade de um gerente de entrega, então é possível que você precise desenvolver algumas habilidades semelhantes às de Sherlock Holmes para encontrar todos os bits perdidos.

Obter acesso aos dados é outra tarefa difícil. As organizações entendem o valor e a sensibilidade dos dados e muitas vezes têm políticas em vigor para que todos tenham acesso ao que precisam e nada mais. Estas políticas traduzem-se em barreiras físicas e digitais chamadas *Muralhas da China.* Estas “paredes” são obrigatórias e bem regulamentadas para dados de clientes na maioria dos países. Isto também ocorre por boas razões; imagine todos em uma empresa de cartão de crédito tendo acesso aos seus hábitos de consumo. Obter acesso aos dados pode levar tempo e envolver políticas da empresa.

2.3.2 Não tenha medo de comprar

Se os dados não estiverem disponíveis dentro da sua organização, procure fora dos muros da sua organização. Muitas empresas se especializam na coleta de informações valiosas. Por exemplo, a Nielsen e a GFK são bem conhecidas por isto na indústria retalhista. Outras empresas fornecem dados para que você, por sua vez, possa enriquecer seus serviços e ecossistema. Esse é o caso do Twitter, LinkedIn e Facebook.

Embora os dados sejam considerados um activo mais valioso do que o petróleo por algumas empresas, cada vez mais governos e organizações partilham os seus dados gratuitamente com o mundo. Esses dados podem ser de excelente qualidade; depende da instituição que o cria e administra. As informações que partilham abrangem uma vasta gama de tópicos, tais como o número de acidentes ou a quantidade de abuso de drogas numa determinada região e a sua demografia. Esses dados são úteis quando você deseja enriquecer dados proprietários, mas também são convenientes ao treinar suas habilidades em ciência de dados em casa. A Tabela 2.1 mostra apenas uma pequena seleção do número crescente de fornecedores de dados abertos.

Tabela 2.1 Uma lista de provedores de dados abertos que devem ajudar você a começar

Site de dados abertos	Descrição
Dados.gov	A casa dos dados abertos do governo dos EUA
https://open-data.europa.eu/	A casa dos dados abertos da Comissão Europeia
Freebase.org	Um banco de dados aberto que recupera suas informações de sites como Wikipédia, a enciclopédia livre
Data.worldbank.org	Iniciativa de dados abertos do Banco Mundial
Aiddata.org	Dados abertos para o desenvolvimento internacional
Open.fda.gov	Dados abertos da Food and Drug Administration dos EUA

2.3.3 Faça verificações de qualidade dos dados agora para evitar problemas mais tarde

Espere gastar boa parte do tempo do seu projeto fazendo correção e limpeza de dados, às vezes até 80%. A recuperação de dados é a primeira vez que você inspecionará os dados no processo de ciência de dados. A maioria dos erros que você encontrará durante a fase de coleta de dados são fáceis de detectar, mas ser muito descuidado fará com que você gaste muito horas resolvendo problemas de dados que poderiam ter sido evitados durante a importação de dados.

Você investigará os dados durante a importação, preparação de dados e exploração fases. A diferença está no objetivo e na profundidade da investigação. Durante *os dados recuperação,* você verifica se os dados são iguais aos dados no documento de origem e veja se você tem os tipos de dados corretos. Isso não deve demorar muito; quando você tiver evidências suficientes de que os dados são semelhantes aos dados encontrados no documento de origem, você para. Com *a preparação dos dados,* você faz uma verificação mais elaborada. Se você fez um bom trabalho durante a fase anterior, os erros que você encontra agora também estão presentes na documento Fonte. O foco está no conteúdo das variáveis: você quer se livrar delas erros de digitação e outros erros de entrada de dados e trazer os dados para um padrão comum entre os conjuntos de dados. Por exemplo, você pode corrigir USQ para EUA e Reino Unido para Reino Unido. Durante a *fase exploratória,* seu foco muda para o que você pode aprender com os dados. Agora você assume que os dados estão limpos e observa as propriedades estatísticas, como distribuições, correlações e valores discrepantes. Freqüentemente, você iterará nessas fases. Por exemplo, quando você descobre valores discrepantes na fase exploratória, eles podem apontar para uma entrada de dados erro. Agora que você entende como a qualidade dos dados é melhorada durante o processo, examinaremos mais profundamente a etapa de preparação de dados.

2.4 Etapa 3: Limpeza, integração e transformação de dados

Os dados recebidos na fase de recuperação de dados provavelmente serão “um diamante no duro." Sua tarefa agora é higienizá-lo e prepará-lo para uso na fase de modelagem e geração de relatórios. Fazer isso é extremamente importante porque seus modelos terão um desempenho melhor e você perderá menos tempo tentando consertar resultados estranhos. Não pode ser mencionado quase vezes suficientes: lixo que entra é igual a lixo que sai. Seu modelo precisa dos dados em um determinado

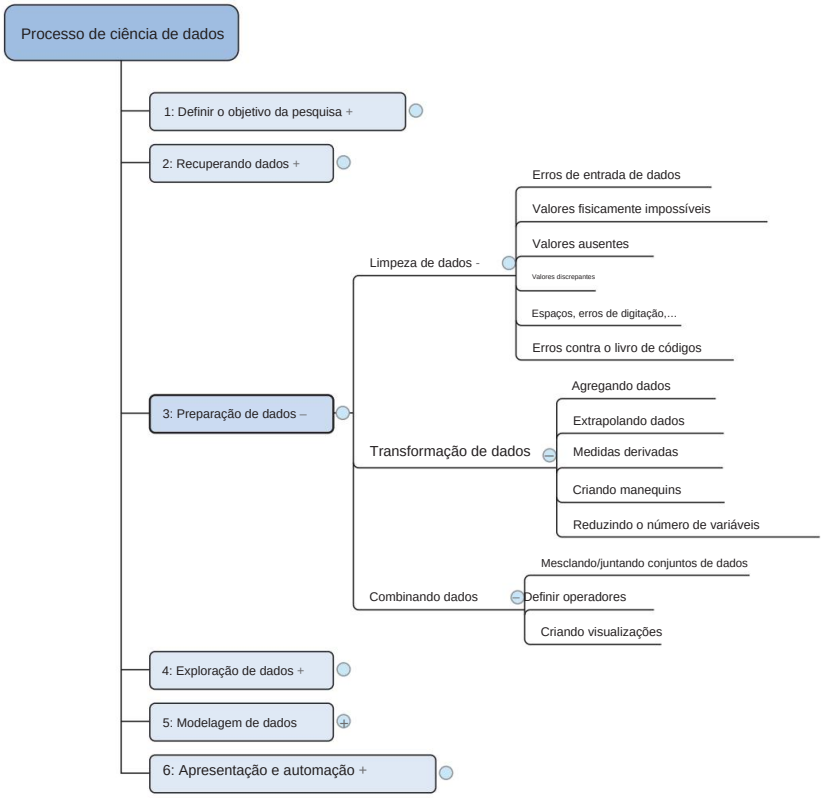


Figura 2.4 Etapa 3: Preparação de dados

formato, então a transformação de dados sempre entrará em jogo. É um bom hábito corrigir erros de dados o mais cedo possível no processo. No entanto, isso nem sempre é possível em um cenário realista, portanto você precisará tomar ações corretivas em seu programa.

A Figura 2.4 mostra as ações mais comuns a serem tomadas durante a limpeza de dados, integração fase de integração e transformação.

Este mapa mental pode parecer um pouco abstrato por enquanto, mas trataremos de todos esses pontos com mais detalhes nas próximas seções. Você verá uma grande semelhança entre todas essas ações.

2.4.1 Limpeza de dados

A limpeza de dados é um subprocesso do processo de ciência de dados que se concentra na remoção de erros em seus dados para que eles se tornem uma representação verdadeira e consistente dos processos de onde se originam.

Por “representação verdadeira e consistente” implicamos que existem pelo menos dois tipos de erros. O primeiro tipo é o *erro de interpretação,* como quando você pega o valor em seu

dados como garantidos, como dizer que a idade de uma pessoa é superior a 300 anos. O segundo tipo de erro aponta para *inconsistências* entre fontes de dados ou contra a sua empresa valores padronizados. Um exemplo desta classe de erros é colocar “Female” em uma tabela e “F” em outro quando representam a mesma coisa: que a pessoa é do sexo feminino. Outro exemplo é usar Libras em uma tabela e Dólares em outra. Também existem muitos erros possíveis para que esta lista seja exaustiva, mas a tabela 2.2 mostra uma visão geral dos tipos de erros que podem ser detectados com verificações fáceis – o “fruto mais fácil de alcançar”, por assim dizer.

Tabela 2.2 Uma visão geral dos erros comuns

Solução geral	
Tente resolver o problema no início da cadeia de aquisição de dados ou então corrija-o no programa.	
Descrição de erro	**Solução possível**
Erros que apontam para valores falsos em um conjunto de dados	
Erros durante a entrada de dados	Anulações manuais
Espaço em branco redundante	Usar funções de string
Valores impossíveis	Anulações manuais
Valores ausentes	Remover observação ou valor
Valores discrepantes	Validar e, se errado, tratar como valor faltante (remover ou inserir)
Erros que apontam para inconsistências entre conjuntos de dados	
Desvios de um livro de códigos	Combine as teclas ou use anulações manuais
Diferentes unidades de medida	Recalcular
Diferentes níveis de agregação	Trazer para o mesmo nível de medição por agregação ou extrapolação

Às vezes você usará métodos mais avançados, como modelagem simples, para encontrar e identificar erros de dados; gráficos de diagnóstico podem ser especialmente esclarecedores. Por exemplo, na figura 2.5 utilizamos uma medida para identificar pontos de dados que parecem deslocados. Nós fazemos um regressão para conhecer os dados e detectar a influência de observações na linha de regressão. Quando uma única observação tem muita influência, isso pode apontar para um erro nos dados, mas também pode ser um ponto válido. Nos dados fase de limpeza, estes métodos avançados são, no entanto, raramente aplicados e muitas vezes considerado por certos cientistas de dados como um exagero.

Agora que demos uma visão geral, é hora de explicar esses erros com mais detalhes.

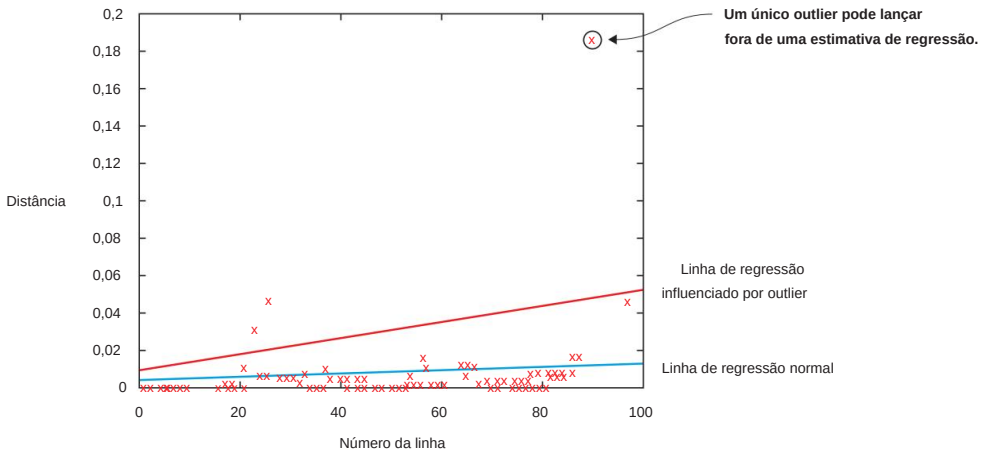


Figura 2.5 O ponto circundado influencia fortemente o modelo e vale a pena investigar porque pode apontar para uma região onde não há dados suficientes ou pode indicar um erro nos dados, mas também pode ser um ponto de dados válido.

ERROS DE ENTRADA DE DADOS

A coleta e entrada de dados são processos sujeitos a erros. Eles muitas vezes exigem humanos intervenção, e porque os humanos são apenas humanos, cometem erros de digitação ou perdem a concentração por um segundo e introduzem um erro na cadeia. Mas os dados coletados por máquinas ou computadores também não estão isentos de erros. Erros podem surgir de humanos negligência, enquanto outros são devidos a falhas de máquinas ou hardware. Exemplos de erros originados de máquinas são erros de transmissão ou bugs na fase de extração, transformação e carregamento (ETL).

Para pequenos conjuntos de dados, você pode verificar cada valor manualmente. Detectando erros de dados quando as variáveis que você estuda não possuem muitas classes pode ser feito tabulando os dados com contagens. Quando você tem uma variável que pode assumir apenas dois valores: “Bom” e “Ruim”, você pode criar uma tabela de frequência e ver se esses são realmente os dois únicos valores presente. Na tabela 2.3, os valores “Godo” e “Bade” indicam que algo deu errado em pelo menos 16 casos.

Tabela 2.3 Detecção de valores discrepantes em variáveis simples com uma tabela de frequência

Valor	Contar
Bom	1598647
Ruim	1354468
Bom	15
Bad	1

A maioria dos erros desse tipo são fáceis de corrigir com instruções de atribuição simples e regras if-then-else:

```
se x == “Godô”:
        x = “Bom”
se x == “Bade”:
        x = “ruim”
```

ESPAÇOS EM BRANCO

REDUNDANTES Espaços em branco tendem a ser difíceis de detectar, mas causam erros como outros caracteres redundantes causariam. Quem não perdeu alguns dias em um projeto por causa de um bug causado por espaços em branco no final de uma string? Você pede ao programa para unir duas chaves e percebe que faltam observações no arquivo de saída. Depois de procurar o código por dias, você finalmente encontra o bug. Depois vem a parte mais difícil: explicar o atraso às partes interessadas do projeto. A limpeza durante a fase ETL não foi bem executada e as chaves de uma tabela continham um espaço em branco no final de uma string. Isso causou uma incompatibilidade de chaves como “FR ” – “FR”, eliminando as observações que não puderam ser correspondidas.

Se você souber tomar cuidado com eles, consertar espaços em branco redundantes é felizmente fácil na maioria das linguagens de programação. Todos eles fornecem funções de string que removerão os espaços em branco iniciais e finais. Por exemplo, em Python você pode usar a função strip() para remover espaços iniciais e finais.

CORRIGINDO INCIDÊNCIAS DE LETRAS MAIÚSCULAS Incompatibilidades de letras maiúsculas são comuns. A maioria das linguagens de programação faz uma distinção entre “Brasil” e “brasil”. Neste caso você pode resolver o problema aplicando uma função que retorne ambas as strings em letras minúsculas, como .lower() em Python. “Brasil”.lower() == “brazil”.lower() deve resultar em verdadeiro.

VALORES IMPOSSÍVEIS E VERIFICAÇÕES DE SANIDADE

As verificações de sanidade são outro tipo valioso de verificação de dados. Aqui você compara o valor com valores fisicamente ou teoricamente impossíveis, como pessoas com mais de 3 metros de altura ou alguém com 299 anos de idade. As verificações de sanidade podem ser expressas diretamente com regras:

```
verificar = 0 <= idade <= 120
```

OUTLIERS

Um outlier é uma observação que parece estar distante de outras observações ou, mais especificamente, uma observação que segue uma lógica ou processo generativo diferente das outras observações. A maneira mais fácil de encontrar valores discrepantes é usar um gráfico ou tabela com os valores mínimo e máximo. Um exemplo é mostrado na figura 2.6.

O gráfico na parte superior não mostra valores discrepantes, enquanto o gráfico na parte inferior mostra possíveis valores discrepantes no lado superior quando uma distribuição normal é esperada. A distribuição normal, ou distribuição gaussiana, é a distribuição mais comum nas ciências naturais.

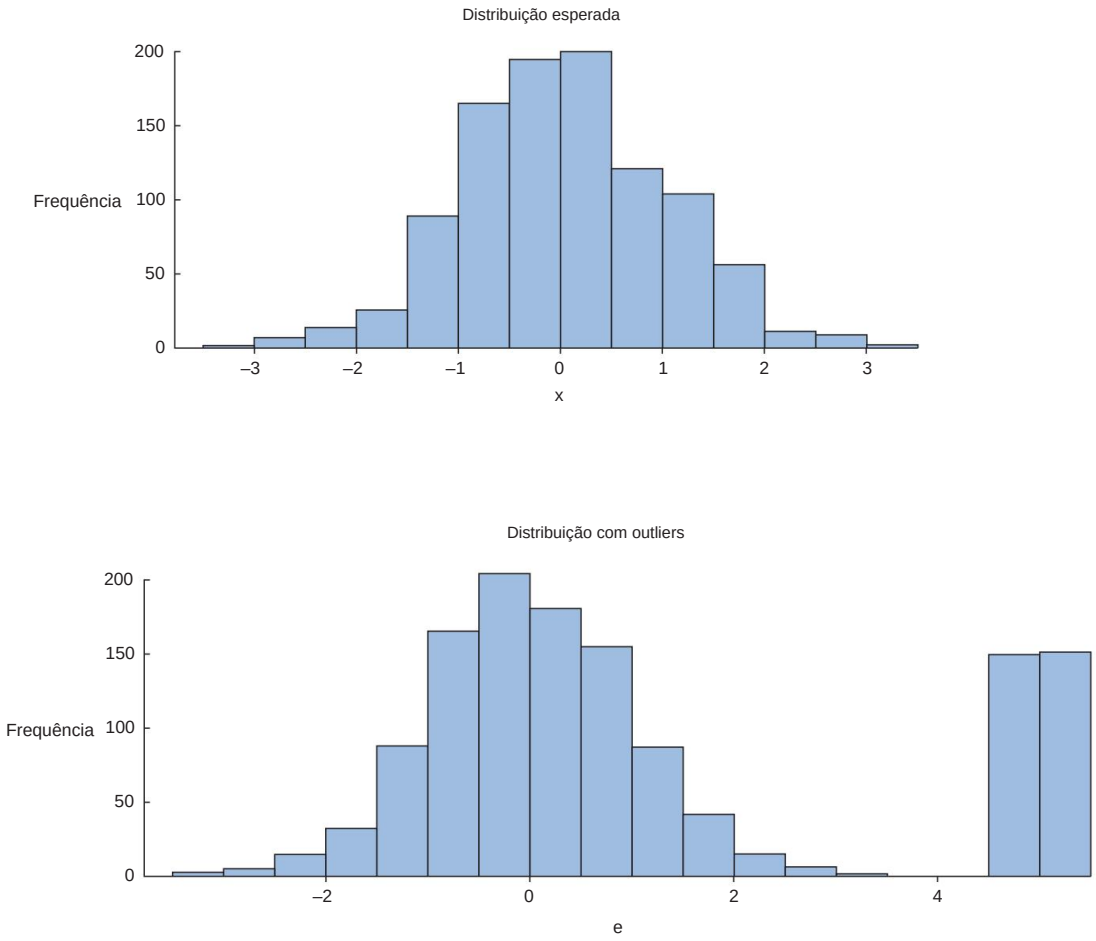


Figura 2.6 Os gráficos de distribuição são úteis para detectar valores discrepantes e ajudar você a entender a variável.

Mostra a maioria dos casos ocorrendo em torno da média da distribuição e as ocorrências diminuem quando mais distantes dela. Os valores altos no gráfico inferior podem
apontam para valores discrepantes ao assumir uma distribuição normal. Como vimos anteriormente com o exemplo de regressão, valores discrepantes podem influenciar gravemente sua modelagem de dados, então investigue eles primeiro.

LIDANDO COM VALORES FALTANTES

Os valores ausentes não são necessariamente errados, mas você ainda precisa tratá-los separadamente; certas técnicas de modelagem não conseguem lidar com valores ausentes. Eles podem ser um indicador que algo deu errado em sua coleta de dados ou que ocorreu um erro no
Processo ETL . As técnicas comuns usadas pelos cientistas de dados estão listadas na tabela 2.4.

Tabela 2.4 Uma visão geral das técnicas para lidar com dados faltantes

Técnica	Vantagem	Desvantagem
Omitir os valores	Fácil de executar	Você perde as informações de uma observação
Definir valor como nulo	Fácil de executar	Nem toda técnica e/ou implementação de modelagem pode lidar com valores nulos
Imputar um valor estático como 0 ou a média	Fácil de executar Você não perde informações das outras variáveis na observação	Pode levar a estimativas falsas de um modelo
Imputar um valor de uma distribuição estimada ou teórica	Não atrapalha tanto o modelo	Mais difícil de executar Você faz suposições de dados
Modelando o valor (não dependente)	Não atrapalha muito o modelo	Pode levar a muita confiança no modelo Pode aumentar artificialmente a dependência entre as variáveis Mais difícil de executar Você faz suposições de dados

Qual técnica usar e em que momento depende do seu caso específico. Se, por
Por exemplo, você não tem observações de sobra, omitir uma observação provavelmente é
não é uma opção. Se a variável puder ser descrita por uma distribuição estável, você poderia
imputar com base nisso. No entanto, talvez um valor ausente realmente signifique “zero”? Isso pode
ser o caso das vendas, por exemplo: se nenhuma promoção for aplicada a uma cesta de clientes, isso
a promoção do cliente está faltando, mas provavelmente também é 0, sem redução de preço.

DESVIOS DE UM LIVRO DE CÓDIGOS

Detectando erros em conjuntos de dados maiores em um livro de códigos ou em valores padronizados
pode ser feito com a ajuda de operações definidas. Um livro de códigos é uma descrição dos seus dados,
uma forma de metadados. Ele contém coisas como o número de variáveis por observação, o número de observações
e o que significa cada codificação dentro de uma variável.
(Por exemplo, “0” é igual a “negativo”, “5” significa “muito positivo”.) Um livro de códigos também
informa o tipo de dados que você está vendo: é hierárquico, gráfico ou outra coisa?

Você olha para os valores que estão presentes no conjunto A, mas não no conjunto B. Esses são valores
isso deveria ser corrigido. Não é por acaso que *os conjuntos* são a estrutura de dados que iremos
usar quando estivermos trabalhando em código. É um bom hábito pensar mais sobre suas estruturas de dados; pode
economizar trabalho e melhorar o desempenho do seu programa.

Se você tiver vários valores para verificar, é melhor colocá-los no livro de códigos
em uma tabela e use um operador de diferença para verificar a discrepância entre ambos
tabelas. Dessa forma, você pode lucrar diretamente com o poder de um banco de dados. Mais sobre isso em
capítulo 5.

DIFERENTES UNIDADES DE MEDIDA

Ao integrar dois conjuntos de dados, você deve prestar atenção às suas respectivas unidades de medição. Um exemplo disso seria quando você estuda os preços da gasolina em o mundo. Para fazer isso, você coleta dados de diferentes provedores de dados. Conjuntos de dados podem conter preços por galão e outros podem conter preços por litro. Uma simples conversão irá faça o truque neste caso.

DIFERENTES NÍVEIS DE AGREGAÇÃO

Ter diferentes níveis de agregação é semelhante a ter diferentes tipos de medição. Um exemplo disso seria um conjunto de dados contendo dados por semana versus um contendo dados por semana de trabalho. Esse tipo de erro geralmente é fácil de detectar, e *resumir* (ou o inverso, *expandir)* os conjuntos de dados irá corrigi-lo.

Depois de limpar os erros de dados, você combina informações de diferentes dados fontes. Mas antes de abordarmos este tópico, faremos um pequeno desvio e enfatizaremos a importância de limpar os dados o mais cedo possível.

2.4.2 Corrija os erros o mais cedo possível

Uma boa prática é mediar erros de dados o mais cedo possível na coleta de dados cadeia e consertar o mínimo possível dentro do seu programa enquanto conserta a origem do problema. Recuperar dados é uma tarefa difícil e as organizações gastam milhões de dólares nisso na esperança de tomar decisões melhores. O processo de coleta de dados está sujeito a erros e, em uma grande organização, envolve muitas etapas e equipes.

Os dados devem ser limpos quando adquiridos por vários motivos:

- ÿ Nem todos detectam anomalias nos dados. Os tomadores de decisão podem cometer erros dispendiosos em informações baseadas em dados incorretos de aplicativos que não conseguem corrigir os dados defeituosos.
- ÿ Se os erros não forem corrigidos no início do processo, a limpeza terá que ser feito para cada projeto que usa esses dados.
- ÿ Erros de dados podem indicar um processo de negócios que não está funcionando conforme planejado. Para Por exemplo, ambos os autores trabalharam em um varejista no passado e criaram um sistema de cupons para atrair mais pessoas e obter lucros maiores. Durante um dado projeto de ciências, descobrimos clientes que abusaram do sistema de cupons e ganhou dinheiro comprando mantimentos. O objetivo do sistema de cupons era estimular a venda cruzada e não distribuir produtos gratuitamente. Essa falha custou o dinheiro da empresa e ninguém na empresa sabia disso. Nesse caso os dados não estavam tecnicamente errados, mas trouxeram resultados inesperados.
- ÿ Erros de dados podem indicar equipamentos defeituosos, como transmissão quebrada linhas e sensores defeituosos.
- ÿ Erros de dados podem apontar para bugs no software ou na integração de software que pode ser crítico para a empresa. Ao realizar um pequeno projeto em um banco, descobrimos que dois aplicativos de software usavam configurações locais diferentes. Isso causou problemas com números maiores que 1.000. Para um aplicativo, o número 1.000 significava um, e para o outro significava mil.

Corrigir os dados assim que são capturados é bom em um mundo perfeito. Infelizmente, um cientista de dados nem sempre tem voz na coleta de dados e simplesmente dizer ao departamento de TI para corrigir certas coisas pode não ser o suficiente. Se você não conseguir corrigir os dados na origem, precisará lidar com isso dentro do seu código. A manipulação de dados não termina com a correção de erros; você ainda precisa combinar os dados recebidos.

Como observação final: guarde sempre uma cópia dos seus dados originais (se possível). Às vezes, você começa a limpar os dados, mas comete erros: imputa variáveis de maneira errada, exclui valores discrepantes que continham informações adicionais interessantes ou altera dados como resultado de uma má interpretação inicial. Se você mantiver uma cópia, poderá tentar novamente. Para “dados fluidos” que são manipulados no momento da chegada, isso nem sempre é possível e você terá aceitado um período de ajustes antes de poder usar os dados que está capturando. Uma das coisas mais difíceis não é a limpeza de conjuntos de dados individuais; no entanto, é combinar diferentes fontes em um todo que faz mais sentido.

2.4.3 Combinando dados de diferentes fontes de dados

Seus dados vêm de vários lugares diferentes e nesta subetapa nos concentramos na integração dessas diferentes fontes. Os dados variam em tamanho, tipo e estrutura, desde bancos de dados e arquivos Excel até documentos de texto.

Nós nos concentramos nos dados em estruturas de tabelas neste capítulo por uma questão de brevidade. É fácil preencher livros inteiros apenas sobre esse tópico, e optamos por focar no processo de ciência de dados em vez de apresentar cenários para cada tipo de dados. Mas tenha em mente que existem outros tipos de fontes de dados, como armazenamentos de valores-chave, armazenamentos de documentos e assim por diante, que trataremos em locais mais apropriados no livro.

AS DIFERENTES FORMAS DE COMBINAR DADOS

Você pode realizar duas operações para combinar informações de diferentes conjuntos de dados. A primeira operação é *a junção:* enriquecer uma observação de uma tabela com informações de outra tabela. A segunda operação é *anexar* ou *empilhar:* adicionar as observações de uma tabela às de outra tabela.

Ao combinar dados, você tem a opção de criar uma nova tabela física ou virtual criando uma visualização. A vantagem de uma visualização é que ela não consome mais espaço em disco. Vamos elaborar um pouco sobre esses métodos.

JUNTAR TABELAS

A união de tabelas permite combinar as informações de uma observação encontrada em uma tabela com as informações encontradas em outra tabela. O foco está em enriquecer uma única observação. Digamos que a primeira tabela contenha informações sobre as compras de um cliente e a outra tabela contenha informações sobre a região onde seu cliente mora. A união das tabelas permite combinar as informações para que você possa utilizá-las em seu modelo, conforme mostra a figura 2.7.

Para unir tabelas, você usa variáveis que representam o mesmo objeto em ambas as tabelas, como uma data, um nome de país ou um número de Seguro Social. Esses campos comuns são conhecidos como chaves. Quando essas chaves também definem exclusivamente os registros na tabela, elas

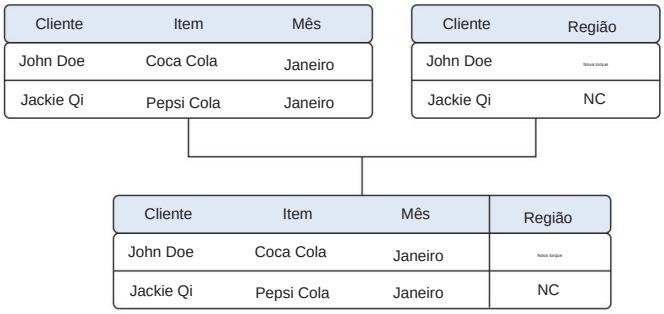


Figura 2.7 Unindo duas tabelas nas chaves Item e Região

são chamadas de *chaves primárias.* Uma tabela pode conter comportamento de compra e a outra tabela pode conter informações demográficas de uma pessoa. Na figura 2.7 ambas as tabelas contêm o nome do cliente, o que facilita o enriquecimento dos gastos do cliente com a região do cliente. Pessoas que estão familiarizadas com o Excel notarão a semelhança com o uso de uma função de pesquisa.

O número de linhas resultantes na tabela de saída depende do tipo exato de junção que você usa. Apresentaremos os diferentes tipos de junções posteriormente neste livro.

ANEXOS DE TABELAS

Acrescentar ou empilhar tabelas é efetivamente adicionar observações de uma tabela a outra. A Figura 2.8 mostra um exemplo de anexação de tabelas. Uma tabela contém as observações do mês de janeiro e a segunda tabela contém as observações do mês de fevereiro. O resultado da anexação destas tabelas é maior, com as observações de Janeiro e Fevereiro. A operação equivalente na teoria dos conjuntos seria a união, e este também é o comando em SQL, a linguagem comum dos bancos de dados relacionais. Outros operadores de conjunto também são usados na ciência de dados, como diferença de conjunto e interseção.

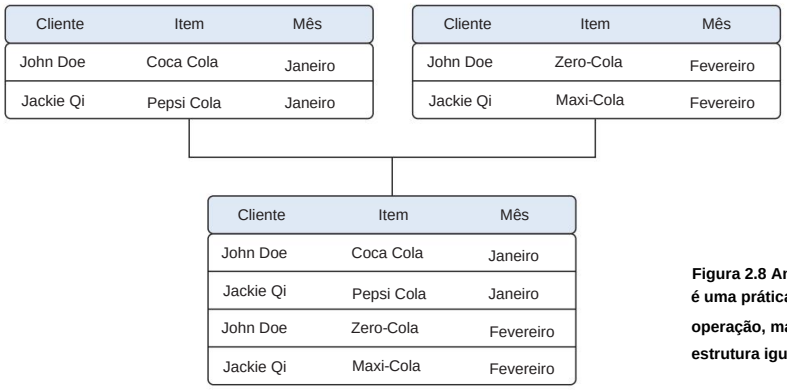


Figura 2.8 Anexar dados de tabelas é uma prática comum operação, mas requer uma estrutura igual nas tabelas anexadas.

USANDO VISUALIZAÇÕES PARA SIMULAR JUNÇÕES E ANEXOS DE DADOS

Para evitar a duplicação de dados, você combina virtualmente dados com visualizações. No exemplo anterior pegamos os dados mensais e combinamos em uma nova tabela física. O problema é que duplicamos os dados e, portanto, precisávamos de mais espaço de armazenamento. No exemplo com o qual estamos trabalhando, isso pode não causar problemas, mas imagine que cada tabela consiste em terabytes de dados; então torna-se problemático duplicar os dados.

Por esta razão, o conceito de visão foi inventado. Uma visualização se comporta como se você estivesse trabalhando em uma tabela, mas essa tabela nada mais é do que uma camada virtual que combina as tabelas para você. A Figura 2.9 mostra como os dados de vendas dos diferentes meses são combinados virtualmente em uma tabela de vendas anual, em vez de duplicar os dados. As visualizações apresentam uma desvantagem, no entanto. Embora uma junção de tabela seja executada apenas uma vez, a junção que cria a visualização é recriada sempre que é consultada, usando mais poder de processamento do que uma tabela pré-calculada teria.

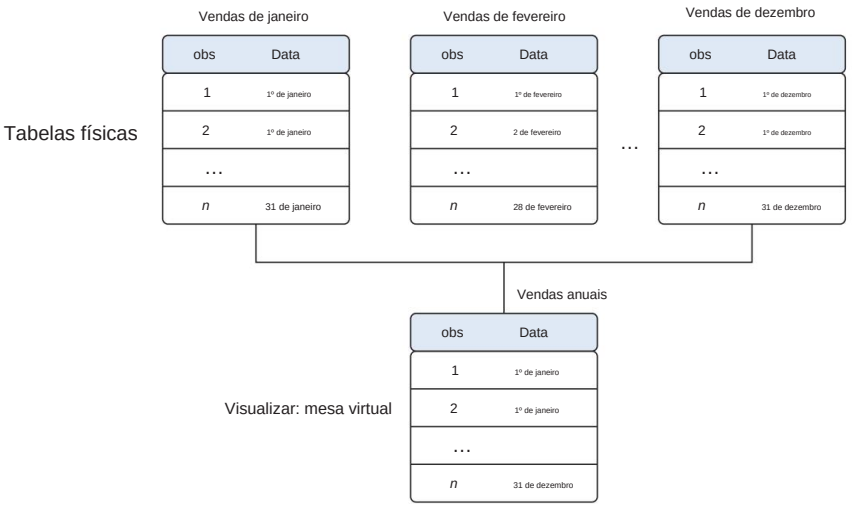


Figura 2.9 Uma visualização ajuda a combinar dados sem replicação.

ENRIQUECENDO MEDIDAS AGREGADAS O

enriquecimento dos dados também pode ser feito adicionando informações calculadas à tabela, como o número total de vendas ou qual percentual do estoque total foi vendido em uma determinada região (figura 2.10).

Medidas extras como essas podem adicionar perspectiva. Olhando para a figura 2.10, temos agora um conjunto de dados agregados, que por sua vez pode ser utilizado para calcular a participação de cada produto dentro da sua categoria. Isso pode ser útil durante a exploração de dados, mas ainda mais ao criar modelos de dados. Como sempre, isso depende do caso exato, mas pela nossa experiência, modelos com “medidas relativas”, como % de vendas (quantidade de

Classe de produto	produtos	Vendas em $	Vendas t-1 em $	Crescimento	Vendas por classe de produto	Classificar vendas
A	B	X	E	(XY) / Y	MACHADO	NX
Esporte	Esporte 1	95	98	–3,06%	215	2
Esporte	Esporte 2	120	132	–9,09%	215	1
Sapato	Sapatos 1	10	6	66,67%	10	3

Figura 2.10 Crescimento, vendas por classe de produto e vendas classificadas são exemplos de medidas derivadas e agregadas.

produto vendido/quantidade total vendida) tendem a superar os modelos que usam os números brutos (quantidade vendida) como insumo.

2.4.4 Transformando dados

Certos modelos exigem que seus dados tenham um determinado formato. Agora que você limpou e integrou os dados, esta é a próxima tarefa que você executará: transformar seus dados para que assume uma forma adequada para modelagem de dados.

TRANSFORMANDO DADOS

As relações entre uma variável de entrada e uma variável de saída nem sempre são lineares. Tomemos, por exemplo, uma relação da forma *y = ae bx.* Tomar o logaritmo das variáveis independentes simplifica dramaticamente o problema de estimativa. A Figura 2.11 mostra como

x	1	2	3	4	5	6	7	8	9	10
registro( ) *x*	0,00	0,43	0,68	0,86	1,00	1.11	1.21	1,29	1,37	1,43
e	0,00	0,44	0,69	0,87	1.02	1.11	1,24	1,32	1,38	1,46

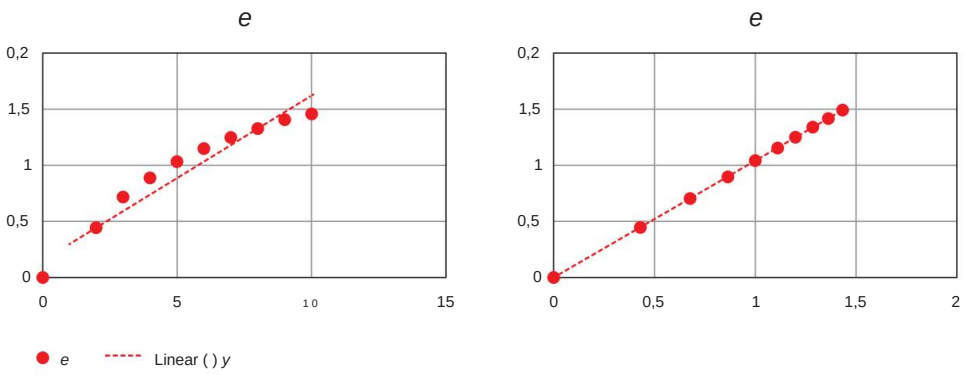


Figura 2.11 Transformar x em log x torna a relação entre x e y linear (direita), em comparação com o não-log x (esquerda).

transformar as variáveis de entrada simplifica muito o problema de estimativa. Outro vezes você pode querer combinar duas variáveis em uma nova variável.

REDUZINDO O NÚMERO DE VARIÁVEIS

Às vezes você tem muitas variáveis e precisa reduzir o número porque elas não adicione novas informações ao modelo. Ter muitas variáveis em seu modelo torna o modelo difícil de manusear e certas técnicas não funcionam bem quando você os sobrecarrega com muitas variáveis de entrada. Por exemplo, todas as técnicas com base em uma distância euclidiana, apresentam bom desempenho apenas até 10 variáveis.

Distância euclidiana

A distância euclidiana ou distância "comum" é uma extensão de uma das primeiras coisas qualquer um aprende matemática sobre triângulos (trigonometria): teorema da perna de Pitágoras. Se você souber o comprimento dos dois lados próximos ao ângulo de 90° de um ângulo reto triângulo, você pode facilmente derivar o comprimento do lado restante (hipotenusa). A fórmula para isso é hipotenusa = ÿ ÿ lado1 + lado2 2 $\sqrt{}$. A distância euclidiana entre dois pontos em um plano bidimensional é calculado usando uma fórmula semelhante: distância = $\sqrt{\text{ÿ ÿ x1 x2 – ÿ }^{2} + \text{ ÿ y1 y2 – ÿ }^{2}}$ ÿ . Se você quiser expandir este cálculo de distância para mais dimensões, adicione as coordenadas do ponto dentro dessas dimensões superiores à fórmula. Para três dimensões obtemos distância = ÿ y1 y2 – ÿ $\sqrt{\text{ÿ ÿ x1 x2 – ÿ }^{2} + \text{ }^{2} + \text{ ÿ z1 z2 – ÿ }^{2}}$ ÿ .

Os cientistas de dados usam métodos especiais para reduzir o número de variáveis, mas mantêm o quantidade máxima de dados. Discutiremos vários desses métodos no capítulo 3. A Figura 2.12 mostra como a redução do número de variáveis facilita a compreensão do

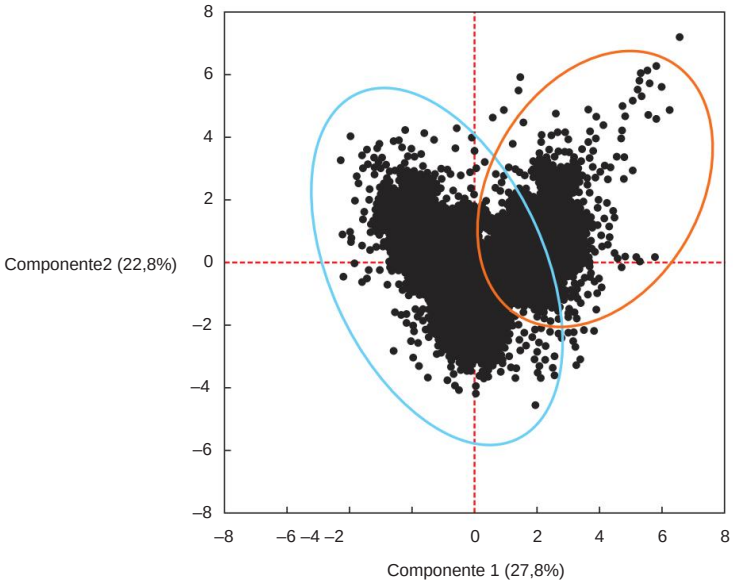


Figura 2.12 A redução de variáveis permite reduzir o número de variáveis enquanto mantém o máximo de informações possível.

valores-chave. Também mostra como duas variáveis respondem por 50,6% da variação do conjunto de dados (componente1 = 27,8% + componente2 = 22,8%). Essas variáveis, chamadas "componente1" e "componente2", são combinações das variáveis originais.
Eles são os *principais componentes* da estrutura de dados subjacente. Se não estiver tão claro neste ponto, não se preocupe, a análise de componentes principais (PCA) será explicada mais detalhadamente no capítulo 3. O que você também pode ver é a presença de uma terceira variável (desconhecida) que divide o grupo de observações em dois.

TRANSFORMANDO VARIÁVEIS EM DUMMIES

As variáveis podem ser transformadas em variáveis dummies (figura 2.13). *Variáveis fictícias* só podem assumir dois valores: verdadeiro(1) ou falso(0). São usados para indicar a ausência de um efeito categórico que possa explicar a observação. Neste caso você criará colunas separadas para as classes armazenadas em uma variável e indicará com 1 se a classe for

presente e 0 caso contrário. Um exemplo é transformar uma coluna chamada Dias da semana nas colunas de segunda a domingo. Você usa um indicador para mostrar se a observação foi numa segunda-feira; você coloca 1 na segunda-feira e 0 em outro lugar. Transformar variáveis em manequins é uma técnica usada em modelagem e é popular, mas não exclusiva, dos economistas.

Nesta seção, apresentamos a terceira etapa do processo de ciência de dados – limpeza, transformação e integração de dados – que transforma seus dados brutos em entradas utilizáveis para a fase de modelagem. A próxima etapa no processo de ciência de dados é obter uma melhor

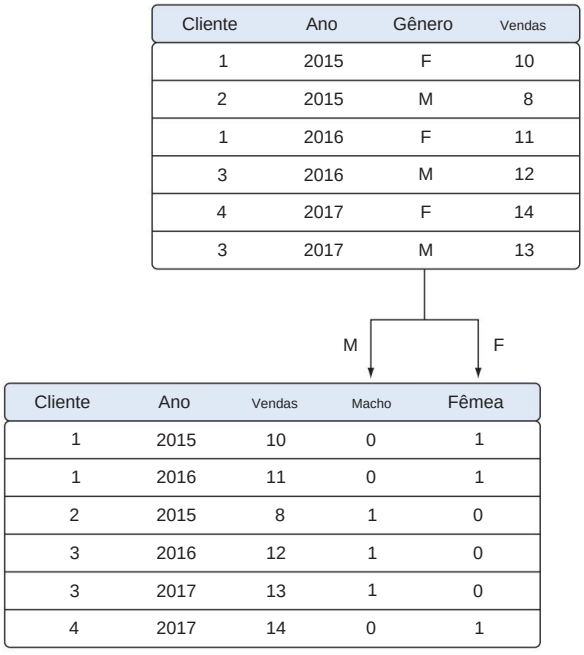


Figura 2.13 Transformar variáveis em dummies é uma transformação de dados que divide uma variável que possui múltiplas classes em múltiplas variáveis, cada uma tendo apenas dois valores possíveis: 0 ou 1.

compreensão do conteúdo dos dados e das relações entre as variáveis e as observações; exploraremos isso na próxima seção.

2.5 Etapa 4: Análise exploratória de dados

Durante a análise exploratória de dados, você se aprofunda nos dados (veja a figura 2.14). As informações se tornam muito mais fáceis de entender quando mostradas em uma imagem, portanto você usa principalmente técnicas gráficas para entender seus dados e as interações entre as variáveis. Esta fase trata da exploração de dados, portanto, manter a mente aberta e os olhos bem abertos é essencial durante a fase de análise exploratória de dados. O objetivo não é limpar os dados, mas é comum que você ainda descubra anomalias que não percebeu antes, forçando-o a dar um passo atrás e corrigi-las.

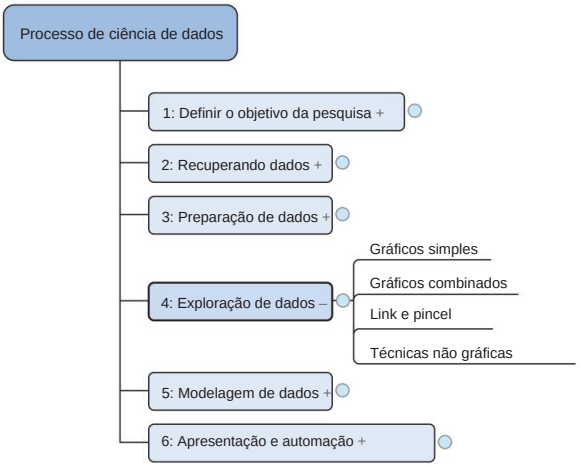


Figura 2.14 Etapa 4: Exploração de dados

As técnicas de visualização usadas nesta fase variam desde simples gráficos de linha ou histogramas, como mostrado na figura 2.15, até diagramas mais complexos, como Sankey e gráficos de rede. Às vezes é útil compor um gráfico composto a partir de gráficos simples para obter ainda mais informações sobre os dados. Outras vezes, os gráficos podem ser animados ou interativos para torná-lo mais fácil e, admitamos, muito mais divertido. Um exemplo de diagrama Sankey interativo pode ser encontrado em http://bost.ocks.org/mike/sankey/ .

Mike Bostock tem exemplos interativos de quase qualquer tipo de gráfico. Vale a pena gastar tempo em seu site, embora a maioria de seus exemplos sejam mais úteis para apresentação de dados do que para exploração de dados.

Figura 2.15 De cima para baixo, um gráfico de barras, um gráfico de linhas e uma distribuição são alguns dos gráficos utilizados na análise exploratória.

Estes gráficos podem ser combinados para fornecer ainda mais informações, como mostra a figura 2.16.

A sobreposição de vários gráficos é uma prática comum. Na figura 2.17 combinamos simples gráficos em um diagrama de Pareto ou diagrama 80-20.

A Figura 2.18 mostra outra técnica: *escovação e ligação.* Com escovação e vinculação, você combina e vincula diferentes gráficos e tabelas (ou visualizações) para que as alterações sejam feitas em um. gráfico são transferidos automaticamente para os outros gráficos. Um exemplo elaborado disso pode ser encontrado no capítulo 9. Esta exploração interativa de dados facilita a descoberta de novos insights.

A Figura 2.18 mostra a pontuação média por país nas perguntas. Não só isso indicam uma alta correlação entre as respostas, mas é fácil perceber que quando você selecionar vários pontos em uma subparcela, os pontos corresponderão a pontos semelhantes na outros gráficos. Neste caso os pontos selecionados no gráfico esquerdo correspondem a pontos nos gráficos do meio e da direita, embora correspondam melhor nos gráficos do meio e da direita gráficos certos.

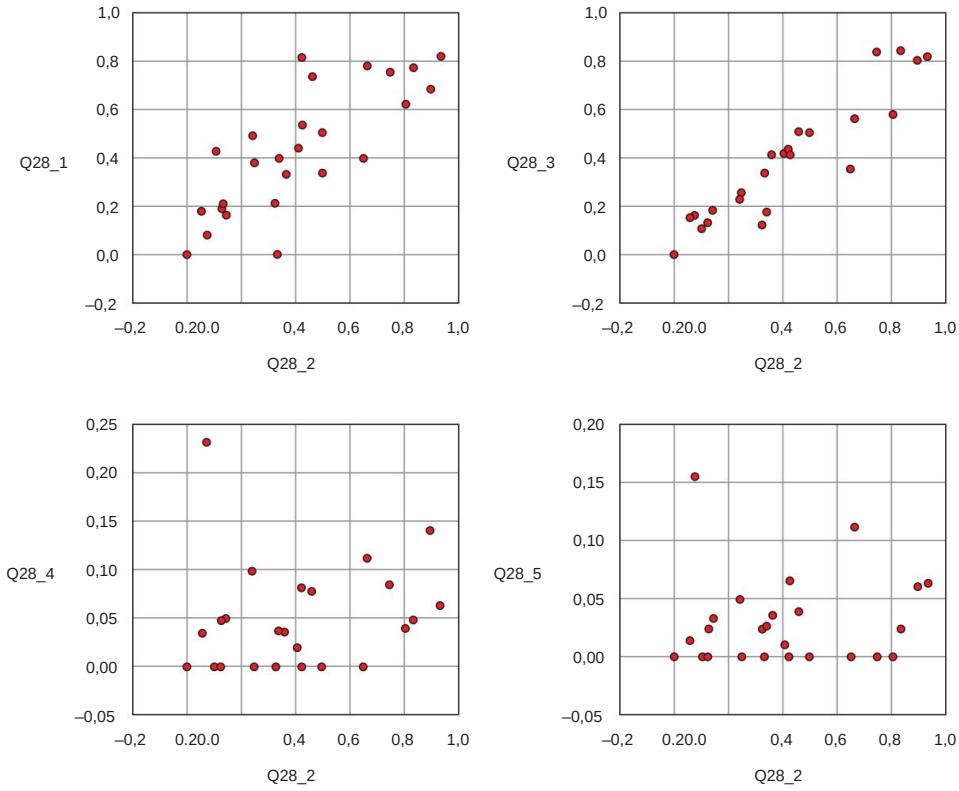


Figura 2.16 Desenhar vários gráficos juntos pode ajudá-lo a entender a estrutura dos seus dados sobre múltiplas variáveis.

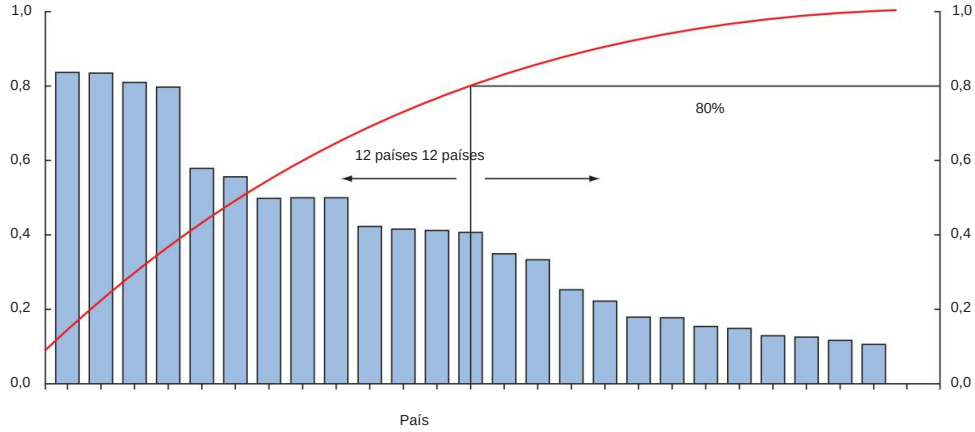


Figura 2.17 Um diagrama de Pareto é uma combinação dos valores e uma distribuição cumulativa. É fácil ver neste diagrama que os primeiros 50% dos países contêm pouco menos de 80% do montante total. Se este gráfico representasse o poder de compra do cliente e vendessemos produtos caros, provavelmente não precisaríamos gastar nosso orçamento de marketing em todos os países; poderíamos começar com os primeiros 50%.

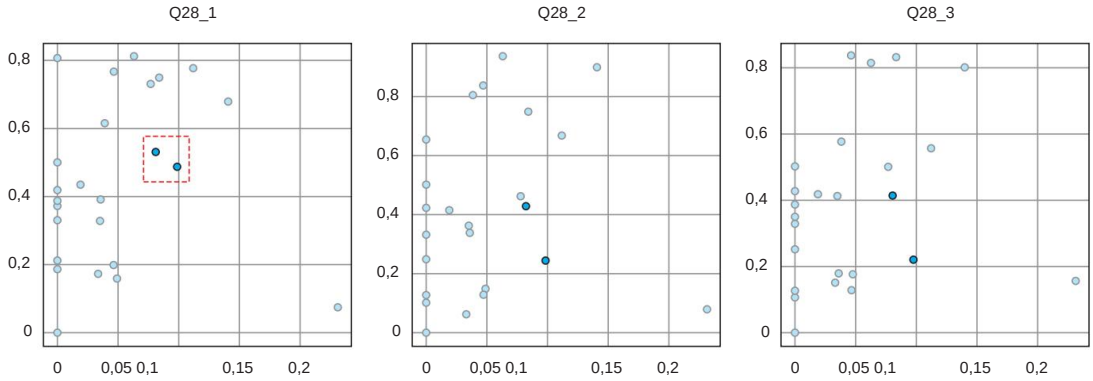


Figura 2.18 Link e pincel permitem selecionar observações em uma parcela e destacar as mesmas observações nas outras parcelas.

Dois outros gráficos importantes são o histograma mostrado na figura 2.19 e o boxplot mostrado na figura 2.20.

Num histograma, uma variável é cortada em categorias discretas e o número de ocorrências em cada categoria é somado e mostrado no gráfico. O boxplot, por outro lado, não mostra quantas observações estão presentes, mas oferece uma impressão da distribuição dentro das categorias. Pode mostrar o máximo, o mínimo, a mediana e outras medidas caracterizadoras ao mesmo tempo.

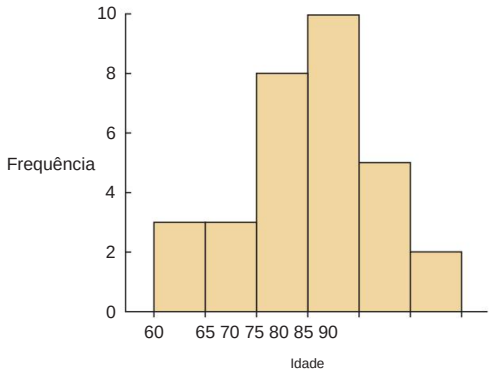


Figura 2.19 Exemplo de histograma: o número de pessoas nas faixas etárias de intervalos de 5 anos

As técnicas que descrevemos nesta fase são principalmente visuais, mas na prática certamente não se limitam a técnicas de visualização. Tabulação, agrupamento e outras técnicas de modelagem também podem fazer parte da análise exploratória. Até mesmo a construção de modelos simples pode fazer parte desta etapa.

Agora que você concluiu a fase de exploração de dados e ganhou uma boa compreensão dos seus dados, é hora de passar para a próxima fase: construir modelos.

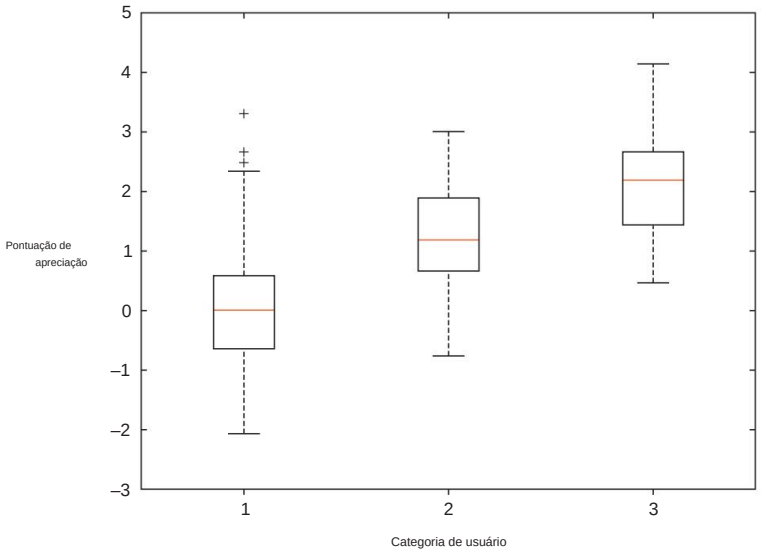


Figura 2.20 Exemplo de boxplot: cada categoria de usuário tem uma distribuição da apreciação que cada um tem por uma determinada imagem em um site de fotografia.

2.6 Etapa 5: Construa os modelos

Com dados limpos e um bom entendimento do conteúdo, você está pronto para construir modelos com o objetivo de fazer melhores previsões, classificar objetos ou obter uma compreensão do sistema que você está considerando. re modelagem. Esta fase é muito mais focada do que a etapa de análise exploratória, porque você sabe o que está procurando e qual será o resultado. A Figura 2.21 mostra os componentes da construção do modelo.

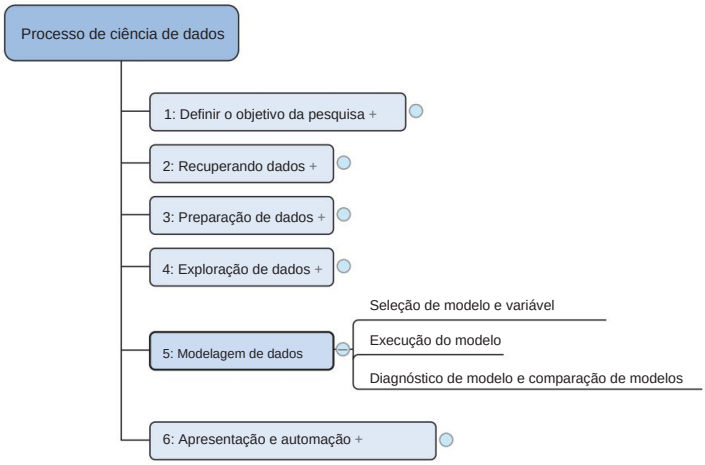


Figura 2.21 Etapa 5: Modelagem de dados

As técnicas que você usará agora são emprestadas do campo de aprendizado de máquina, mineração de dados e/ou estatística. Neste capítulo exploramos apenas a ponta do iceberg das técnicas existentes, enquanto o capítulo 3 as apresenta adequadamente. Está além do escopo deste livro fornecer mais do que uma introdução conceitual, mas é o suficiente para você começar; 20% das técnicas irão ajudá-lo em 80% dos casos porque as técnicas se sobrepõem naquilo que tentam realizar. Freqüentemente, eles alcançam seus objetivos de maneiras semelhantes, mas ligeiramente diferentes.

Construir um modelo é um processo iterativo. A maneira como você constrói seu modelo depende se você segue a estatística clássica ou a escola de aprendizado de máquina um pouco mais recente e o tipo de técnica que deseja usar. De qualquer forma, a maioria dos modelos consiste nas seguintes etapas principais:

1 Seleção de uma técnica de modelagem e variáveis para inserir no modelo
2 Execução do modelo
3 Diagnóstico e comparação de modelos

2.6.1 Modelo e seleção de variáveis

Você precisará selecionar as variáveis que deseja incluir em seu modelo e uma técnica de modelagem. As suas conclusões da análise exploratória já devem dar uma ideia razoável

de quais variáveis ajudarão você a construir um bom modelo. Muitas técnicas de modelagem estão disponíveis, e escolher o modelo certo para um problema requer julgamento sobre o seu papel. Você precisará considerar o desempenho do modelo e se o seu projeto atende a todos os requisitos para usar seu modelo, bem como outros fatores:

- ÿ O modelo deve ser movido para um ambiente de produção e, em caso afirmativo, seria fácil de implementar?
- ÿ Quão difícil é a manutenção do modelo: por quanto tempo ele permanecerá relevante se deixado intocado?
- ÿ O modelo precisa ser fácil de explicar?

Quando o pensamento estiver concluído, é hora de agir.

2.6.2 Execução do modelo

Depois de escolher um modelo, você precisará implementá-lo em código.

> **OBSERVAÇÃO** Esta é a primeira vez que entraremos na execução real do código Python, então certifique-se de ter um ambiente virtual instalado e funcionando. Saber como configurar isso é necessário conhecimento, mas se for sua primeira vez, confira o apêndice *D*.
>
> Todo o código deste capítulo pode ser baixado em https://www.manning .com/books/Introducing-data-science. Este capítulo vem com um ipython (.ipynb) notebook e arquivo Python (.py).

Felizmente, a maioria das linguagens de programação, como Python, já possui bibliotecas como StatsModels ou Scikit-learn. Esses pacotes usam diversas das técnicas mais populares. Codificar um modelo não é uma tarefa trivial na maioria dos casos, portanto, ter essas bibliotecas disponível pode acelerar o processo. Como você pode ver no código a seguir, é bastante fácil usar regressão linear (figura 2.22) com StatsModels ou Scikit-learn. Fazer isso sozinho exigiria muito mais esforço, mesmo para técnicas simples. A seguir a listagem mostra a execução de um modelo de previsão linear.

Listagem 2.1 Executando um modelo de predição linear em dados semi-aleatórios

```
importar statsmodels.api como sm importar
numpy como preditores np
= np.random.random(1000).reshape(500,2)
alvo = preditores.dot(np.array([0,4, 0,6])) + np.random.random(500) lmRegModel = sm.OLS(target,preditores)
resultado = lmRegModel.fit()

Resumo de Resultados()
```

Importa módulos Python necessários.

Cria dados aleatórios para preditores (valores x) e dados semi-aleatórios para o alvo (valores y) do modelo. Usamos preditores como entrada para criar o destino, então inferimos uma correlação aqui.

Adapta-se linearmente **regressão nos dados.**

Mostra estatísticas de ajuste do modelo.

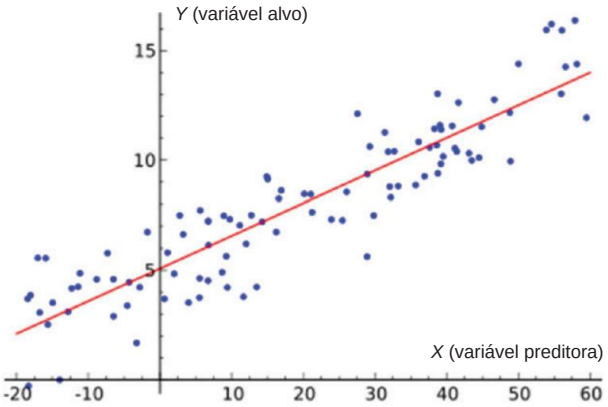


Figura 2.22 A regressão linear tenta ajustar uma linha enquanto minimiza a distância até cada ponto

Ok, nós trapaceamos aqui, bastante. Criamos valores preditores que visam prever como as variáveis de destino se comportam. Para uma regressão linear, uma "relação linear" entre cada variável x (preditor) e y (alvo) é assumida, conforme mostrado na figura 2.22.

No entanto, criamos a variável alvo, com base no preditor, adicionando um pouco de aleatoriedade. Não deveria ser surpresa que isso nos dê um modelo bem ajustado. O results.summary() gera a tabela da figura 2.23. Veja bem, o resultado exato depende das variáveis aleatórias que você obteve.

Dep. Variable:	y	**R-squared:**	0.893
Model:	OLS	**Adj. R-squared:**	0.893
Method:	Least Squares	**F-statistic:**	2088.
Date:	Fri, 30 Oct 2015	**Prob (F-statistic):**	7.13e-243
Time:	12:44:31	**Log-Likelihood:**	-176.74
No. Observations:	500	**AIC:**	357.5
Df Residuals:	498	**BIC:**	365.9
Df Model:	2		
Covariance Type:	nonrobust		

Ajuste do modelo: maior é melhor, mas muito alto é suspeito.

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
x1	0.7658	0.040	19.130	0.000	0.687 0.844
x2	1.1252	0.039	28.603	0.000	1.048 1.202

valor p para mostrar se uma variável preditora tem uma influência significativa o alvo. Menor é melhor e <0,05 é frequentemente considerado "significativo".

Omnibus:	34.269	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13.480
Skew:	-0.125	**Prob(JB):**	0.00118
Kurtosis:	2.235	**Cond. No.**	2.51

Coeficientes de equações lineares.
y = 0,7658xl + 1,1252x2.

Figura 2.23 Saída de informações do modelo de regressão linear

Vamos ignorar a maior parte dos resultados que obtivemos aqui e nos concentrar nas partes mais importantes:

- ÿ Ajuste *do modelo* —Para isso é usado o R ao quadrado ou o R ao quadrado ajustado. Esta medida é uma indicação da quantidade de variação nos dados que são capturados pelo modelo. A diferença entre o R-quadrado ajustado e o R-quadrado é mínima aqui porque o ajustado é o normal + uma penalidade pela complexidade do modelo. Um modelo fica complexo quando muitas variáveis (ou recursos) são introduzidas. Você não precisa de um modelo complexo se um modelo simples estiver disponível, então o R-quadrado ajustado pune você por complicar demais. De qualquer forma, 0,893 é um valor alto, e deveria ser porque trapaceamos. Existem regras práticas, mas para modelos empresariais, modelos acima de 0,85 são frequentemente considerados bons. Se você quer vencer uma competição, você precisa estar na casa dos 90. Para a pesquisa, no entanto, muitas vezes são encontrados ajustes de modelo muito baixos (<0,2 mesmo). O que é mais importante é a influência das variáveis preditoras introduzidas. ÿ *As variáveis preditoras possuem um coeficiente – para* um modelo linear isso é fácil de interpretar. Em nosso exemplo, se você adicionar

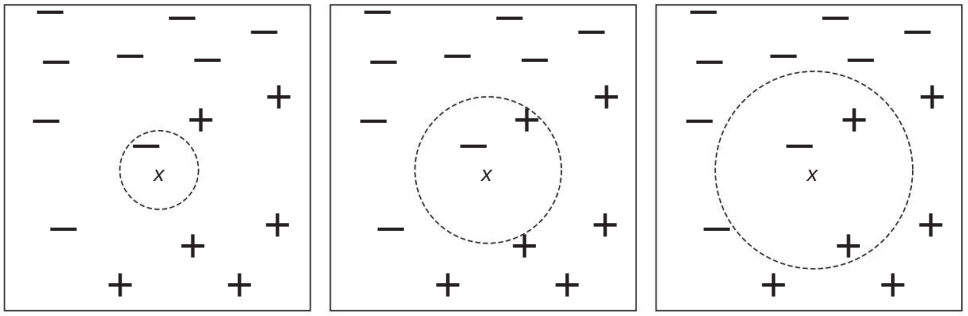
"1" a x1, y será alterado por "0,7658". É fácil ver como encontrar um bom preditor pode ser o caminho para o Prêmio Nobel, mesmo que seu modelo como um todo seja uma porcaria. Se, por exemplo, determinarmos que um determinado gene é significativo como causa do cancro, este é um conhecimento importante, mesmo que esse gene por si só não determine se uma pessoa irá ter cancro. O exemplo aqui é classificação, não regressão, mas a questão permanece a mesma: detectar influências é mais importante em estudos científicos do que modelos perfeitamente ajustados (para não dizer mais realistas). Mas quando sabemos que um gene tem esse impacto?

Isso é chamado de significância.

- ÿ Significância *do preditor* —Os coeficientes são excelentes, mas às vezes não existem evidências suficientes para mostrar que a influência existe. É disso que se trata o valor p. Uma longa explicação sobre erros do tipo 1 e tipo 2 é possível aqui, mas as explicações curtas seriam: se o valor p for inferior a 0,05, a variável é considerada significativa para a maioria das pessoas. Na verdade, este é um número arbitrário. Isso significa que há 5% de chance de o preditor não ter qualquer influência. Você aceita essa chance de 5% de estar errado? Isso é contigo. Várias pessoas introduziram os limites extremamente significativos (p<0,01) e marginalmente significativos (p<0,1).

A regressão linear funciona se você quiser prever um valor, mas e se quiser classificar algo? Depois você vai para os modelos de classificação, sendo os mais conhecidos entre eles os k-vizinhos mais próximos.

Conforme mostrado na figura 2.24, k-vizinhos mais próximos olham para pontos rotulados próximos a um ponto não rotulado e, com base nisso, faz uma previsão de qual deveria ser o rótulo.



(a) 1 vizinho mais próximo (b) 2 vizinhos mais próximos (c) 3 vizinhos mais próximos

Figura 2.24 As técnicas de K-vizinho mais próximo observam o k-ponto mais próximo para fazer uma previsão.

Vamos tentar isso em código Python usando a biblioteca de aprendizado Scikit, como na próxima listagem.

Listagem 2.2 Executando classificação de k-vizinhos mais próximos em dados semi-aleatórios

```
de sklearn importar preditores de vizinhos
= np.random.random(1000).reshape(500,2)
alvo = np.around(predictors.dot(np.array([0,4, 0,6])) + np.random.random(500))

clf = vizinhos.KNeighborsClassifier(n_neighbors=10) knn = clf.fit(preditores,alvo)

knn.score(preditores, alvo)
```

Importa módulos.

Cria dados de preditores aleatórios e dados de destino semi-aleatórios com base em dados de preditores.

Adapta-se ao modelo de 10 vizinhos mais próximos.

Obtém a pontuação de ajuste do modelo: que porcentagem da classificação estava correta?

Como antes, construímos dados correlacionados aleatoriamente e surpresa, surpresa obtemos 85% de casos corretamente classificados. Se quisermos olhar em profundidade, precisamos pontuar o modelo. Não se deixe enganar por knn.score() ; ele retorna a precisão do modelo, mas ao “marcar uma modelo”, muitas vezes queremos dizer aplicá-lo em dados para fazer uma previsão.

```
previsão = knn.predict(preditores)
```

Agora podemos usar a previsão e compará-la com a realidade usando uma confusão matriz.

```
métricas.confusion_matrix(alvo,predição)
```

Obtemos uma matriz 3 por 3 conforme mostrado na figura 2.25.

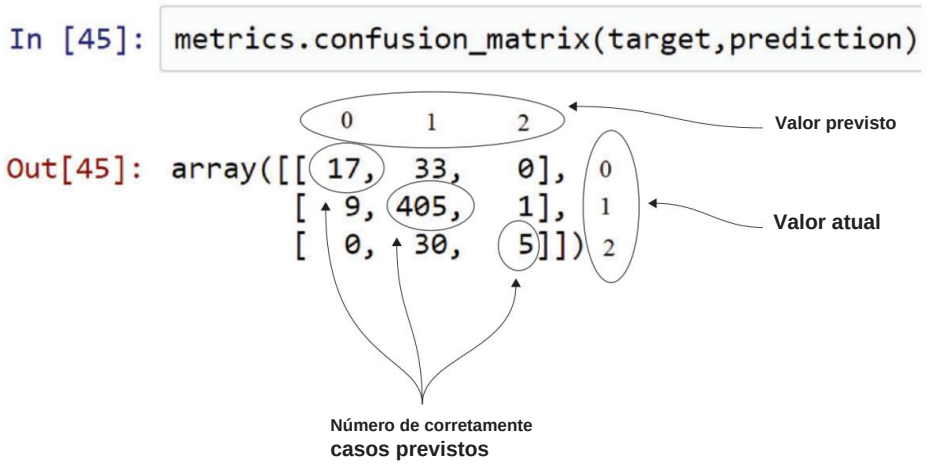


Figura 2.25 Matriz de confusão: mostra quantos casos foram classificados corretamente e classificados incorretamente comparando a previsão com os valores reais. Observação: as classes (0,1,2) foram adicionadas na figura para esclarecimento.

A matriz de confusão mostra que previmos corretamente 17+405+5 casos, então isso é bom. Mas é realmente uma surpresa? Não, pelos seguintes motivos:

- ÿ Por um lado, o classificador tinha apenas três opções; marcando a diferença com o último time np.around() arredondará os dados para o número inteiro mais próximo. Neste caso é ou 0, 1 ou 2. Com apenas 3 opções, você não pode acertar muito pior do que 33% em 500 palpites, mesmo para uma distribuição aleatória real, como jogar uma moeda.
- ÿ Segundo, trapaceamos novamente, correlacionando a variável resposta com os preditores. Devido à maneira como fizemos isso, obtivemos a maioria das observações sendo “1”. Por adivinhando “1” para cada caso já teríamos um resultado semelhante.
- ÿ Comparamos a previsão com os valores reais, é verdade, mas nunca previmos com base em dados recentes. A previsão foi feita usando os mesmos dados dos dados usado para construir o modelo. Está tudo bem e elegante para você se sentir bem, mas não dá nenhuma indicação se o seu modelo funcionará quando encontrar dados realmente novos. Para isso precisamos de uma amostra de validação, como será discutido em a próxima seção.

Não se deixe enganar. Digitar este código não fará milagres por si só. Pode demorar um pouco para acertar a peça de modelagem e todos os seus parâmetros.

Para ser honesto, apenas algumas técnicas têm implementações prontas para a indústria em Python. Mas é bastante fácil usar modelos disponíveis em R dentro do Python com a ajuda da biblioteca RPy . RPy fornece uma interface de Python para R. *R* é um ambiente de software livre, amplamente utilizado para computação estatística. Se você ainda não já vale pelo menos dar uma olhada, pois em 2014 ainda era um dos mais populares

(se não a mais popular) linguagens de programação para ciência de dados. Para obter mais informações, consulte http://www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html .

2.6.3 Diagnóstico de modelo e comparação de modelos

Você construirá vários modelos dos quais poderá escolher o melhor com base em vários critérios. Trabalhar com uma amostra de validação ajuda você a escolher a amostra de melhor desempenho modelo. Uma amostra de validação é uma parte dos dados que você deixa de fora da construção do modelo, para que pode ser usado para avaliar o modelo posteriormente. O princípio aqui é simples: o modelo deve funcionar com dados invisíveis. Você usa apenas uma fração dos seus dados para estimar o modelo e a outra parte, a amostra de validação, é mantida fora da equação. O modelo é então liberado sobre os dados invisíveis e medidas de erro são calculadas para avaliá-los.

Múltiplas medidas de erro estão disponíveis e na figura 2.26 mostramos a ideia geral sobre comparando modelos. A medida de erro usada no exemplo é o erro quadrático médio.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Yi})^2$$

Figura 2.26 Fórmula para erro quadrático médio

O erro quadrático médio é uma medida simples: verifique para cada previsão a que distância ela estava a verdade, eleve ao quadrado esse erro e some o erro de cada previsão.

A Figura 2.27 compara o desempenho de dois modelos para prever o tamanho do pedido do preço. O primeiro modelo tem *tamanho = 3 * preço* e o segundo modelo tem *tamanho = 10.* Para

	n	Tamanho	Preço	Previsto modelo 1	Previsto modelo 2	Erro modelo 1	Erro modelo 2
	1	10	3				
	2	15	5				
	3	18	6				
	4	14	5				
	…	…					
80% treina	800	9	3				
	801	12	4	12	10	0	2
	802	13	4	12	10	1	3
	…						
	999	21	7	21	10	0	11
Teste de 20%	1000	10	4	12	10	–2	0
				Total		5861	110225

Figura 2.27 Uma amostra de validação ajuda a comparar modelos e garante que você possa generalizar os resultados para dados que o modelo ainda não viu.

estimar os modelos, usamos 800 observações escolhidas aleatoriamente entre 1.000 (ou 80%), sem mostrar os outros 20% dos dados ao modelo. Uma vez treinado o modelo, prever os valores para os outros 20% das variáveis com base naquelas para as quais já conhece o valor verdadeiro e calcula o erro do modelo com uma medida de erro. Em seguida escolhemos o modelo com menor erro. Neste exemplo escolhemos o modelo 1 porque tem o menor erro total.

Muitos modelos fazem suposições fortes, como a independência dos insumos, e você deve verificar se essas suposições são realmente atendidas. Isso é chamado de *diagnóstico de modelo.*

Esta seção forneceu uma breve introdução às etapas necessárias para construir um modelo válido. Depois de ter um modelo funcional, você estará pronto para ir para a última etapa.

2.7 Etapa 6: Apresentar descobertas e construir aplicativos com base nelas

Depois de analisar os dados com êxito e criar um modelo com bom desempenho, você estará pronto para apresentar suas descobertas ao mundo (figura 2.28). Esta é uma parte emocionante; todos suas horas de trabalho duro valeram a pena e você pode explicar o que descobriu para o partes interessadas.

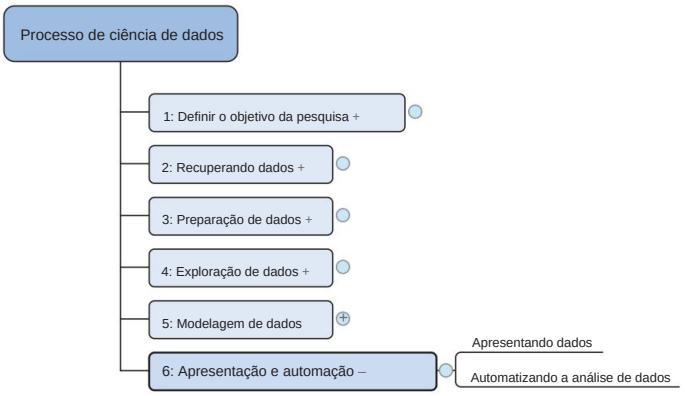


Figura 2.28 Etapa 6: Apresentação e automação

Às vezes as pessoas ficam tão entusiasmadas com o seu trabalho que você precisa repeti-lo repetidamente porque valorizam as previsões dos seus modelos ou os insights que você produziu. Por esse motivo, você precisa automatizar seus modelos. Isso nem sempre significa que você terá que refazer todas as suas análises o tempo todo. Às vezes é suficiente que você implemente apenas a pontuação do modelo; outras vezes, você pode criar um aplicativo que atualize automaticamente relatórios, planilhas do Excel ou apresentações do PowerPoint. A última etapa do processo de ciência de dados é onde suas *habilidades interpessoais* serão mais úteis e, sim, são extremamente importantes. Na verdade, recomendamos que você encontre livros dedicados e outras informações sobre o assunto e trabalhe com eles, porque por que se preocupar em fazer todo esse trabalho duro se ninguém escuta o que você tem a dizer?

Se você fez isso direito, agora você tem um modelo funcional e partes interessadas satisfeitas, então podemos concluir este capítulo aqui.

2.8 Resumo

Neste capítulo você aprendeu que o processo de ciência de dados consiste em seis etapas:

- *Definir o* objetivo da pesquisa – Definir o quê, o porquê e o como do seu projeto em um termo de abertura do projeto.
- *Recuperação* de dados — Encontrar e obter acesso aos dados necessários ao seu projeto. Esses dados são encontrados dentro da empresa ou recuperados de terceiros.
- *Preparação de dados — verifica* e corrige erros de dados, enriquecendo os dados com dados de outras fontes de dados e transformando-os em um formato adequado para seus modelos.
- *Exploração de dados — Aprofunde* -se em seus dados usando estatísticas descritivas e técnicas visuais.
- *Modelagem de dados — Uso de* aprendizado de máquina e técnicas estatísticas para atingir o objetivo do projeto.
- *Apresentação e* automação — Apresentando seus resultados às partes interessadas e industrializando seu processo de análise para reutilização repetitiva e integração com outras ferramentas.