Formação Cientista de Dados

Dia 01 - Manhã - Introdução à Data Science

Vítor Wilher

Cientista de Dados | Mestre em Economia

Plano de Voo

- Sobre o Professor
- Objetivo do nosso Curso
- Período e Sistemática
- Introdução
- Coleta e Tratamento
- Exploração
- Modelagem
- Comunicação
- Tipos de Cientistas de Dados
- Conhecimentos necessários
- Tipos de Questões
- DS e Telecom
- Um exemplo simples

Sobre o Professor

- Análise Macro Founder
- BS and MSc in Economics
- DS Specialization
- R Specialist

Objetivo do nosso Curso

Capacitar as equipes nas análises estatísticas com alto volume de dados e múltiplas fontes, permitindo que análises complexas sejam feitas em um curto espaço de tempo, beneficiando as tomadas de decisões operacionais e executivas.

Período e Sistemática

Nos encontraremos aqui dos dias 11/03 a 15/05, das 9h às 18h.

- 9h às 11h
- Intervalo de 15min
- 11h15 às 13h
- Almoço
- 14h às 16h
- Intervalo de 15min
- 16h15 às 18h

Introdução

O avanço da informática e das telecomunicações possibilitou o armazenamento e a distribuição de conjuntos de dados cada vez mais complexos. Lidar com essas bases de dados exigiu a sistematização de diversas técnicas de coleta, tratamento, análise e apresentação de dados.

Essa sistematização de técnicas deu origem ao que hoje chamamos de **data science**, cujo objetivo principal é extrair informações úteis de conjuntos de dados aparentemente confusos.

Introdução

Aplicações interessantes:

- Identificar mensagens indesejáveis em um e-mail (spam);
- Segmentação do comportamento de consumidores para propagandas direcionadas;
- Redução de fraudes em transações de cartão de crédito;
- Predição de eleições;
- Otimização do uso de energia em casas ou prédios;
- etc, etc, etc...

Introdução

Ao coletar dados, introduzimos em uma plataforma de Análise de Dados (como o R) informações coletadas no mundo real. Seja vindas de base de dados prontas ou adquiridas minerando dados abertos em sites. Depois tratamos as informações para que sejam devidamente legíveis para um computador e bem formatadas para nosso próprio entendimento. Então começamos a análise propriamente dita. Visualizamos os dados, procurando perguntas interessantes e padrões, depois avaliamos nossas intuições com modelos estatísticos. Por fim, comunicamos nossos achados - afinal de pouco importa achar algo somente para si.

Introdução

Com base em Grolemund and Wickham [2017], podemos sistematizar o processo de compreensão dos dados conforme a figura abaixo.

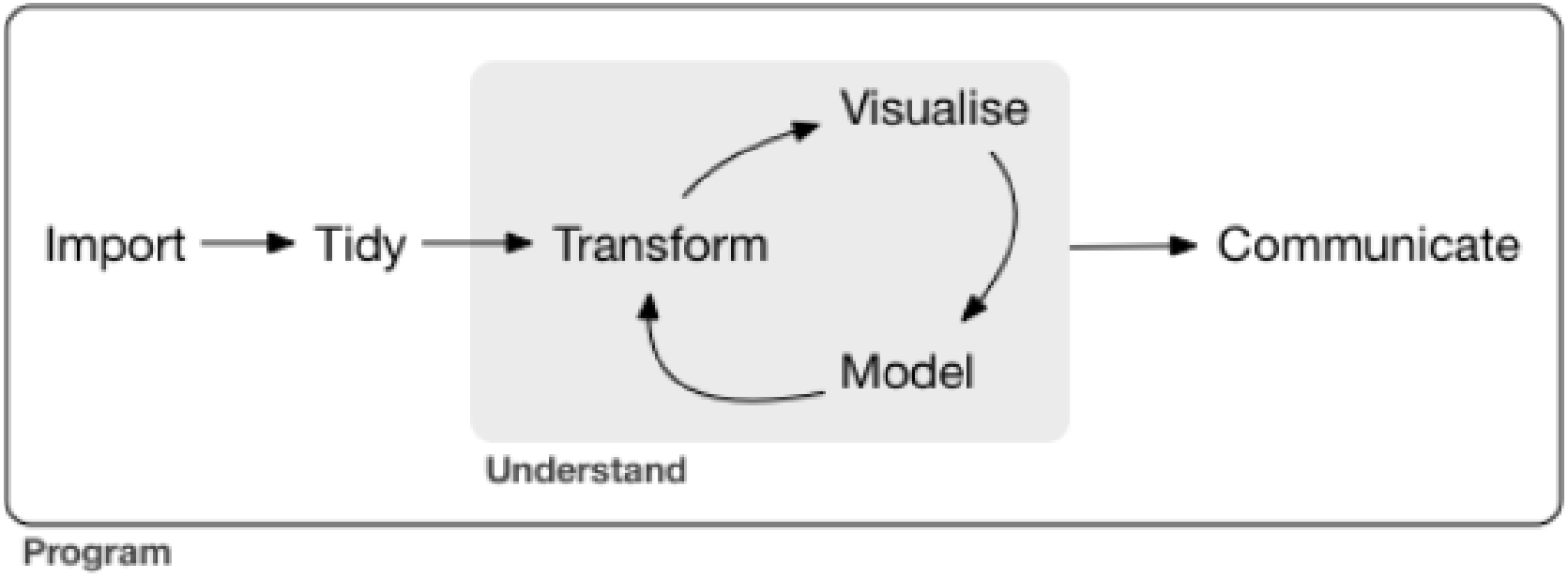


Figure 1: O processo de compreensão dos dados

Coleta e Tratamento

Dados podem estar dispostos em diferentes formatos:

- Excel;
- XML;
- JSON;
- txt;
- HTML;
- MySQL;
- Formatos proprietários (Weka, Stata, Minitab, Octave, SPSS, SAS, etc).

Coleta e Tratamento

Dados precisam ser tratados:

- Limpeza de dados;
- Tratamento de *missing values*;
- Construção de números índices;
- Deflacionar valores correntes;
- Obtenção de taxas de crescimento, a partir de comparações mensais, interanuais, acumuladas em 12 meses, etc;
- Tratando tendências;
- Dessazonalização;
- Obtendo subconjuntos (*subsetting*) relevantes;
- Classificando dados de acordo com algum critério;
- Transformando dados de acordo com alguma operação.

Exploração

A exploração de dados é a arte de analisar seus dados, gerando hipóteses rapidamente, testando-os rapidamente, repetindo-os várias vezes. O objetivo da exploração de dados é gerar muitos leads promissores que você poderá explorar mais tarde com mais profundidade.

Modelagem

O objetivo da modelagem é capturar a essência de um conjunto de dados. Alguns exemplos de modelos que podem ser utilizados para isso:

- Modelos ARIMA;
- Regressão linear;
- Árvores de regressão;
- Neural Network;
- Support Vector Machine;
- Naive Bayes;
- etc, etc, etc.

Comunicação

Os resultados encontrados devem ser compartilhados com gestores, clientes ou colaboradores. Para os primeiros grupos, não se faz necessário mostrar o `código` utilizado, enquanto para o segundo isso deve ser uma característica importante. Por isso, é preciso encontrar uma plataforma que consiga integrar as quatro etapas da análise de dados, gerando como `produto` um {documento reprodutíve}, que unam **código** e **texto**. Seja o código visível ou não para a ponta final.

Tipos de Cientistas de Dados

Com a difusão da área de *data science*, é muito comum que as empresas estejam atrás de profissionais multidisciplinares, que consigam cumprir aquelas quatro etapas do processo. Como começamos a ver, contudo, essas quatro etapas envolvem diversos conhecimentos que são difíceis de serem encontrados em um único profissional. Não é que eles não existam, de fato, há alguns *unicórnios* por aí, mas são raros de serem encontrados.

Tipos de Cientistas de Dados

De maneira geral, há *tipos* de cientistas de dados. Podemos diferenciá-lo da seguinte forma:

- *Analistas de BI*: raramente codificam (podem até utilizar GUIs para acessar bancos de dados, para que nem sequer escrevam consultas SQL - a ferramenta faz isso para eles; no entanto, eles precisam entender o esquema do banco de dados.), são os responsáveis por definir métricas e trabalhar com o gerenciamento para identificar fontes de dados ou para criar dados. Eles também trabalham na criação de *dashboards* de dados com vários usuários finais em mente, desde segurança, finanças, vendas, marketing até executivos. Muitos têm um diploma de MBA;

Tipos de Cientistas de Dados

- *Engenheiros de Dados*: obtêm os requisitos desses analistas de BI para configurar os pipelines de dados e têm o fluxo de dados em toda a empresa e fora dele, com pequenos pedaços (normalmente dados resumidos) terminando em vários laptops de funcionários para análise ou relatório. Eles trabalham com administradores de sistema para configurar o acesso a dados, personalizado para cada tipo de usuário. Eles estão familiarizados com o data warehousing, os diferentes tipos de infraestrutura em nuvem (interna, externa, híbrida) e sobre como otimizar as transferências e o armazenamento de dados, equilibrando a velocidade com o custo e a segurança. Eles estão muito familiarizados com o funcionamento da Internet, bem como com a integração e padronização de dados. Eles são bons em programar e implantar sistemas projetados pelo terceiro tipo de cientistas de dados, descrito abaixo. Às vezes, particularmente para funções seniores, eles são chamados de arquitetos de dados;

Tipos de Cientistas de Dados

- *ML Data Scientists*: Os cientistas de dados de aprendizado de máquina projetam e monitoram sistemas preditivos e de pontuação, têm um grau avançado, são especialistas em todos os tipos de dados (grandes, pequenos, em tempo real, não-estruturados etc.) Eles executam muitos algoritmos, testes, ajustes e manutenção. Eles sabem como selecionar / comparar ferramentas e fornecedores, e como decidir entre o aprendizado de máquina caseiro ou ferramentas (fornecedor ou código aberto). Eles geralmente desenvolvem protótipos ou provas de conceitos, que acabam sendo implementados no modo de produção por engenheiros de dados. Suas linguagens de programação de escolha são Python e R;

Tipos de Cientistas de Dados

- *Analistas de Dados*: Os analistas de dados são cientistas juniores de dados que fazem muita análise de números, limpeza de dados e trabalho em análises únicas e geralmente projetos de curto prazo. Eles interagem e suportam cientistas de dados de BI ou ML. Eles às vezes usam técnicas de modelagem estatística mais avançadas.

Conhecimentos necessários

Uma pergunta muito frequente associada à data science é justamente o que é preciso saber para se tornar um cientista de dados. Como vimos acima, nem todos os profissionais que trabalham com ciência de dados são iguais. Logo, não há uma cesta de conhecimento única para se trabalhar na área. Analistas de BI, por exemplo, vão estar muito mais preocupados com a criação e apresentação de indicadores de desempenho; enquanto engenheiros de dados estarão mais frequentemente envolvidos com a infraestrutura necessária para armazenar e distribuir conjuntos de dados entre diferentes perfis de usuários.

Conhecimentos necessários

De maneira geral, cientistas de dados são altamente qualificados, tendo mestrado ou doutorado em áreas como matemática, estatística, economia, ciência da computação, etc, de modo que disciplinas como cálculo, estatística descritiva e inferência estatística são conhecidas. Ademais, têm conhecimento de uma ou mais linguagens de programação, como R e Python. Outros conhecimentos específicos, como Hadoop, SQL, Apache Spark ou algoritmos de machine learning vão depender do tipo de tarefa e/ou estrutura de dados a que o profissional está submetido.

Tipos de Questões

- in order of difficulty: ***Descriptive*** $\rightarrow$ ***Exploratory*** $\rightarrow$ ***Inferential*** $\rightarrow$ ***Predictive*** $\rightarrow$ ***Causal*** $\rightarrow$ ***Mechanistic***

- **Análise Descritiva** = describe set of data, interpret what you see (census, Google Ngram)
- **Análise Exploratória** = discovering connections (correlation does not = causation)
- **Análise Inferencial** = use data conclusions from smaller population for the broader group
- **Análise Preditiva** = use data on one object to predict values for another (if X predicts Y, does not = X cause Y)
- **Análise Causal** = how does changing one variable affect another, using randomized studies, Strong assumptions, golden standard for statistical analysis
- **Análise Mecanicista** = understand exact changes in variables in other variables, modeled by empirical equations (engineering/physics)

DS e Telecom

Não importa como você olhe, o setor de telecomunicações está sob muita pressão. Startups podem fornecer serviços ágeis e on-demand, como as de telecomunicações, os players OTT mudaram a forma como as pessoas consomem o conteúdo, novos modelos de negócios e tecnologias estão constantemente atrapalhando o mercado.

DS e Telecom

Para acompanhar essa pressão, as empresas de telecomunicações estão investindo pesadamente em áreas fora do seu core business, agregando novo valor através de seus serviços e criando fluxos de receita adicionais para ficar ao lado de seu comércio tradicional. Para não ficar para trás e se tornar apenas um mediador de serviços inovadores, eles estão se adaptando e adotando setores emergentes e inovadores.

DS e Telecom

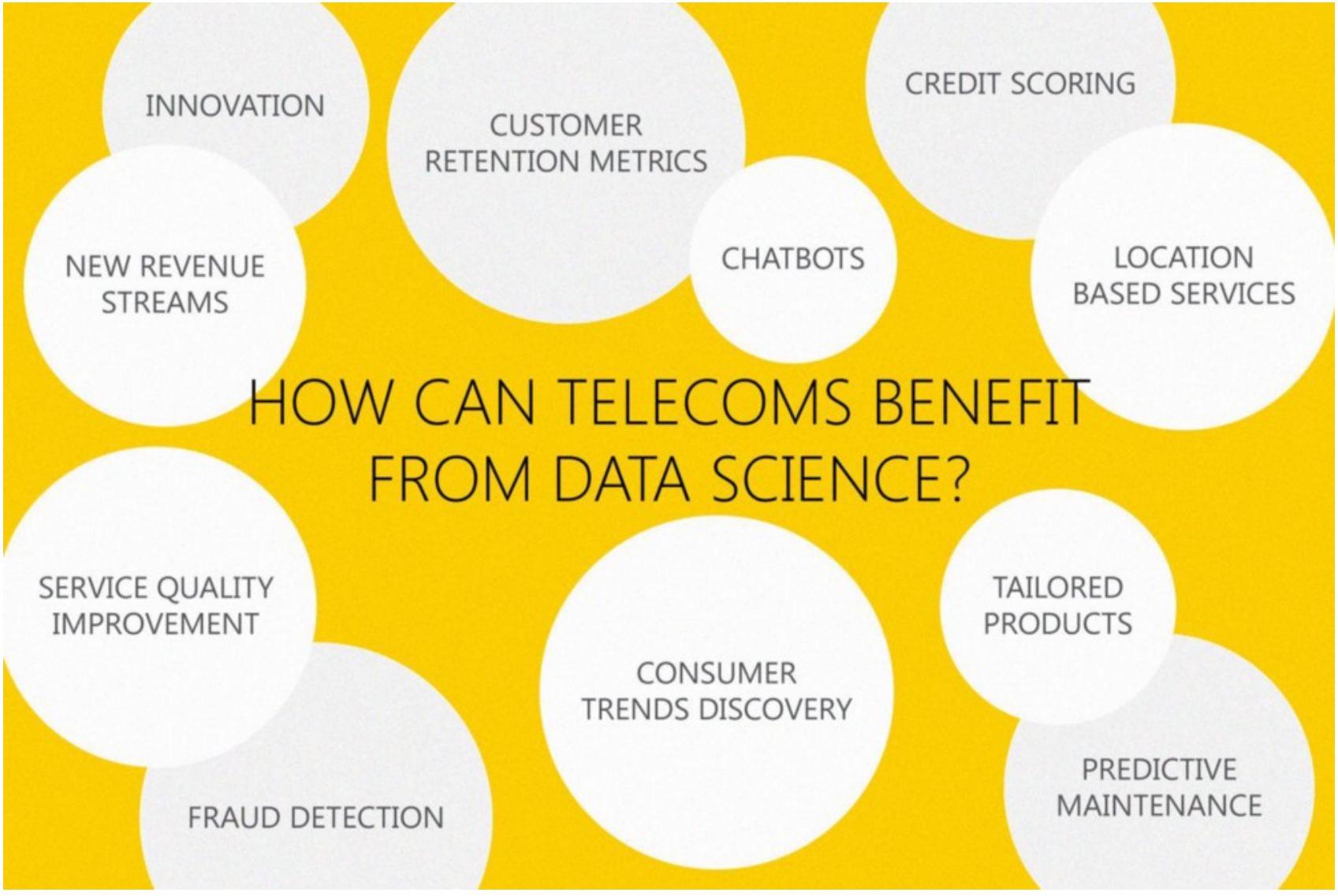


Figure 2: DS e Telecomunicações

Um exemplo simples

Cerca de 80% do trabalho de um cientista de dados envolvem a *exploração* e o *entendimento* dos dados. Para ilustrar, vamos ver um exemplo envolvendo uma base de dados simples de **carros usados**.

```
## data exploration example using used car data
usedcars <- read.csv("usedcars.csv", stringsAsFactors = FALSE)
```

Um exemplo simples

```
# get structure of used car data
str(usedcars)
```

```
## 'data.frame':    150 obs. of  6 variables:
##  $ year        : int  2011 2011 2011 2011 2012 2010 2011 2010 2011 2010 ...
##  $ model       : chr  "SEL" "SEL" "SEL" "SEL" ...
##  $ price       : int  21992 20995 19995 17809 17500 17495 17000 16995 16995 16995 ...
##  $ mileage     : int  7413 10926 7351 11613 8367 25125 27393 21026 32655 36116 ...
##  $ color       : chr  "Yellow" "Gray" "Silver" "Gray" ...
##  $ transmission: chr  "AUTO" "AUTO" "AUTO" "AUTO" ...
```

Um exemplo simples

```
## Exploring numeric variables -----
# summarize numeric variables
summary(usedcars$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    2008    2009    2009    2010    2012
```

```
summary(usedcars[c("price", "mileage")])
```

```
##      price          mileage
##  Min.   : 3800   Min.   :  4867
##  1st Qu.:10995   1st Qu.: 27200
##  Median :13592   Median : 36385
##  Mean   :12962   Mean   : 44261
##  3rd Qu.:14904   3rd Qu.: 55125
##  Max.   :21992   Max.   :151479
```

Um exemplo simples

```r
# the min/max of used car prices
range(usedcars$price)
```

```
## [1]  3800 21992
```

```r
# the difference of the range
diff(range(usedcars$price))
```

```
## [1] 18192
```

```r
# IQR for used car prices
IQR(usedcars$price)
```

```
## [1] 3909.5
```

```r
# use quantile to calculate five-number summary
quantile(usedcars$price)
```

```
##      0%     25%     50%     75%    100%
##  3800.0 10995.0 13591.5 14904.5 21992.0
```

Um exemplo simples

```r
# the 99th percentile
quantile(usedcars$price, probs = c(0.01, 0.99))
```

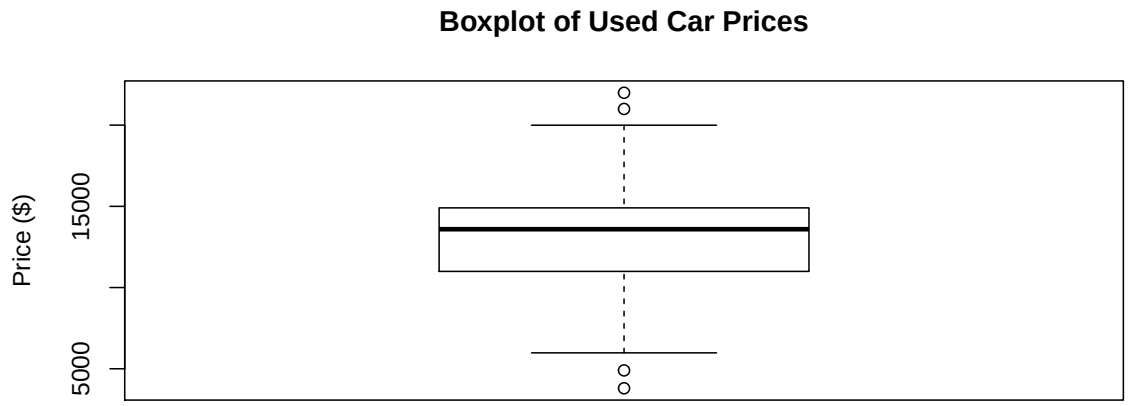
```
##       1%      99%
##  5428.69 20505.00
```

```r
# quintiles
quantile(usedcars$price, seq(from = 0, to = 1, by = 0.20))
```

```
##      0%     20%     40%     60%     80%    100%
##  3800.0 10759.4 12993.8 13992.0 14999.0 21992.0
```

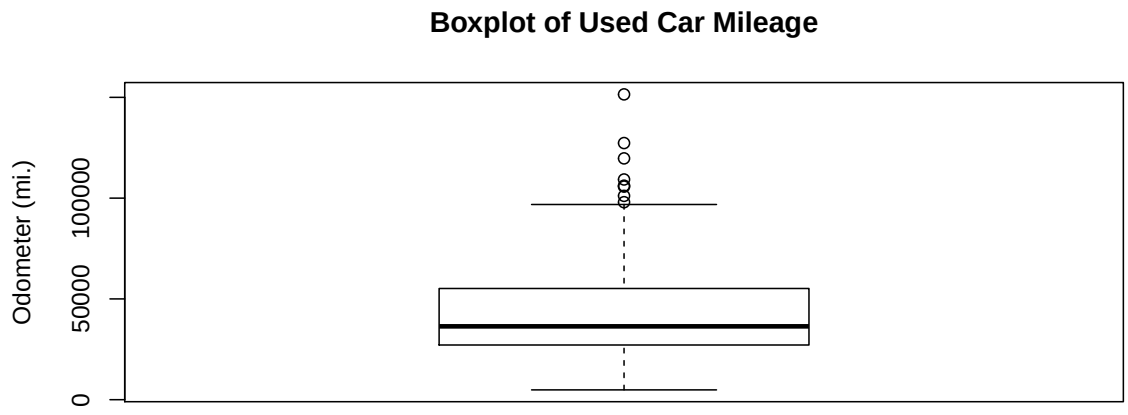
Um exemplo simples

```r
# boxplot of used car prices and mileage
boxplot(usedcars$price, main="Boxplot of Used Car Prices",
      ylab="Price ($)")
```



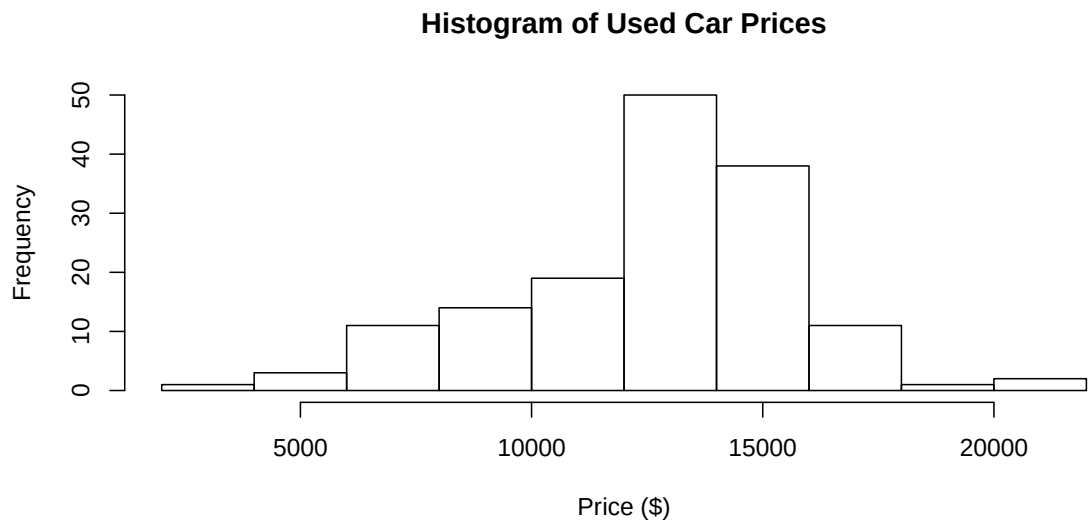
Um exemplo simples

```
boxplot(usedcars$mileage, main="Boxplot of Used Car Mileage",
      ylab="Odometer (mi.)")
```



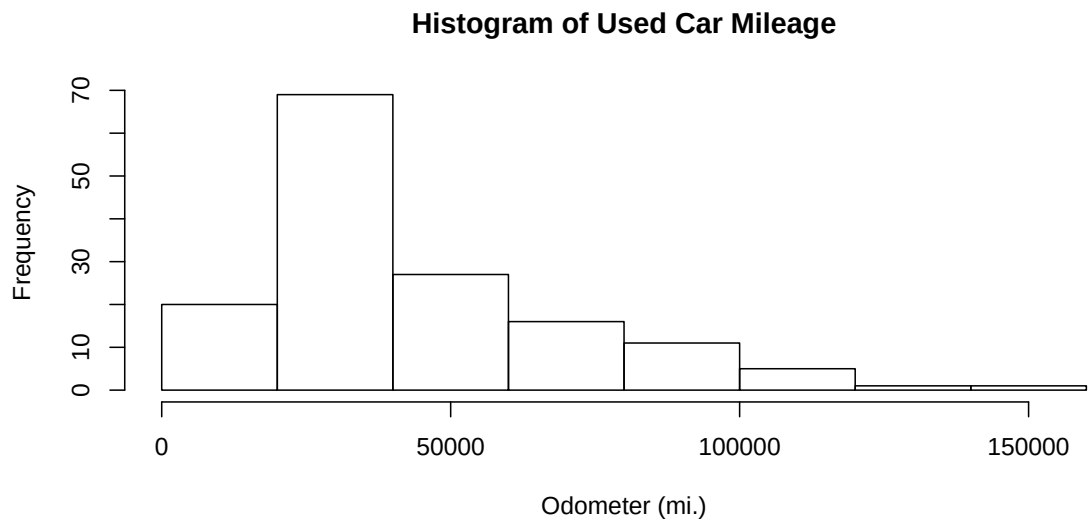
Um exemplo simples

```
# histograms of used car prices and mileage
hist(usedcars$price, main = "Histogram of Used Car Prices",
     xlab = "Price ($)")
```



Um exemplo simples

```
hist(usedcars$mileage, main = "Histogram of Used Car Mileage",
     xlab = "Odometer (mi.)")
```



Um exemplo simples

```
# variance and standard deviation of the used car data
var(usedcars$price)
```

```
## [1] 9749892
```

```
sd(usedcars$price)
```

```
## [1] 3122.482
```

```
var(usedcars$mileage)
```

```
## [1] 728033954
```

```
sd(usedcars$mileage)
```

```
## [1] 26982.1
```

Um exemplo simples

```
## Exploring numeric variables -----
# one-way tables for the used car data
table(usedcars$year)
```

```
##
## 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
##    3    1    1    1    3    2    6   11   14   42   49   16    1
```

```
table(usedcars$model)
```

```
##
##  SE SEL SES
##  78  23  49
```

```
table(usedcars$color)
```

```
##
##  Black   Blue   Gold   Gray  Green    Red Silver  White Yellow
##     35     17      1     16      5     25     32     16      3
```

Um exemplo simples

```
# compute table proportions
model_table <- table(usedcars$model)
prop.table(model_table)
```

```
##
##        SE       SEL       SES
## 0.5200000 0.1533333 0.3266667
```

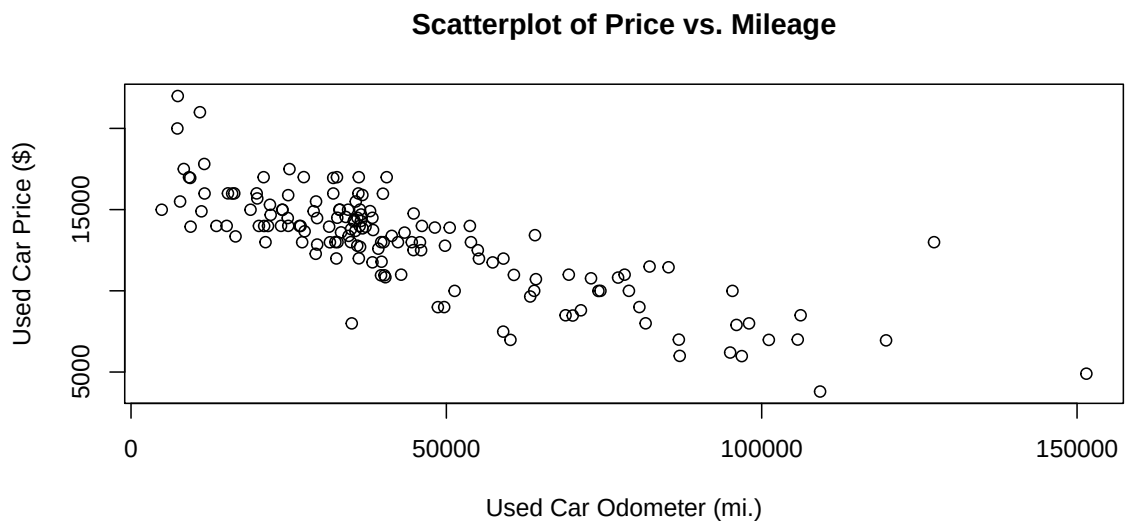
Um exemplo simples

```
# round the data
color_table <- table(usedcars$color)
color_pct <- prop.table(color_table) * 100
round(color_pct, digits = 1)
```

```
##
##  Black   Blue   Gold   Gray  Green    Red Silver  White Yellow
##   23.3   11.3    0.7   10.7    3.3   16.7   21.3   10.7    2.0
```

Um exemplo simples

```
## Exploring relationships between variables -----
# scatterplot of price vs. mileage
plot(x = usedcars$mileage, y = usedcars$price,
     main = "Scatterplot of Price vs. Mileage",
     xlab = "Used Car Odometer (mi.)",
     ylab = "Used Car Price ($)")
```



Um exemplo simples

```
# new variable indicating conservative colors
usedcars$conservative <-
  usedcars$color %in% c("Black", "Gray", "Silver", "White")

# checking our variable
table(usedcars$conservative)
```

FALSE TRUE 51 99

G. Grolemund and H. Wickham. *R for Data Science*. O'Reilly Media, 2017.