MBA Business Analytics e Big Data

Análise Exploratória de Dados

Prof. Dr. João Rafael Dias

2º semestre - 2022

Agenda

Na aula de hoje...

Apresentação do curso
Introdução ao R/Rstudio
Estrutura de dados
Comando básicos
Leitura e escrita de dados
Prática no RStudio

Manipulação de *strings* com `stringr`
Manipulação de datas com `lubridate`
Manipulação de *dataframes* com `dplyr`
Prática no RStudio

Tipos de variáveis
Medidas de Centralidade
Medidas de Dispersão
Medidas de Associação
Prática no RStudio

Técnicas de visualização
Tipos de gráficos
Análise univariada
Análise bivariada
Prática no RStudio

Trabalhos práticos
Aplicação do conteúdo

Técnicas de visualização dos dados

Prof. Dr. João Rafael Dias Pinto

Visualização dos dados

Para começar...

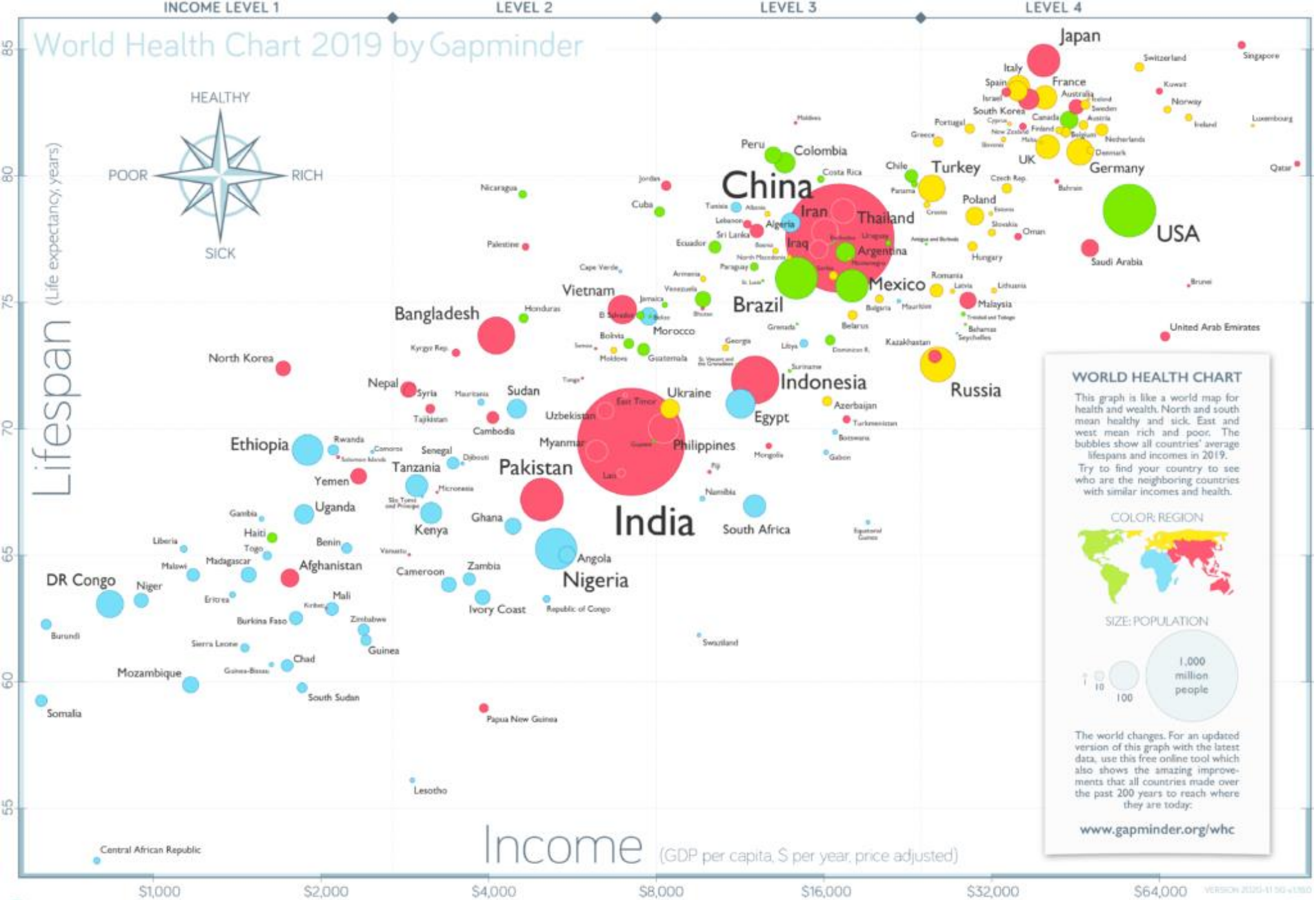
Ronald Coase
(1910 – 2013)

"Torture the data, and it will confess to **anything**"

Visualização dos dados

Para começar...

https://medium.com/code-heroku/introduction-to-exploratory-data-analysis-eda-c0257f888676



https://www.gapminder.org/fw/world-health-chart/whc2019/

Gráficos podem transmitir informações poderosas

Visualização dos dados

Overview

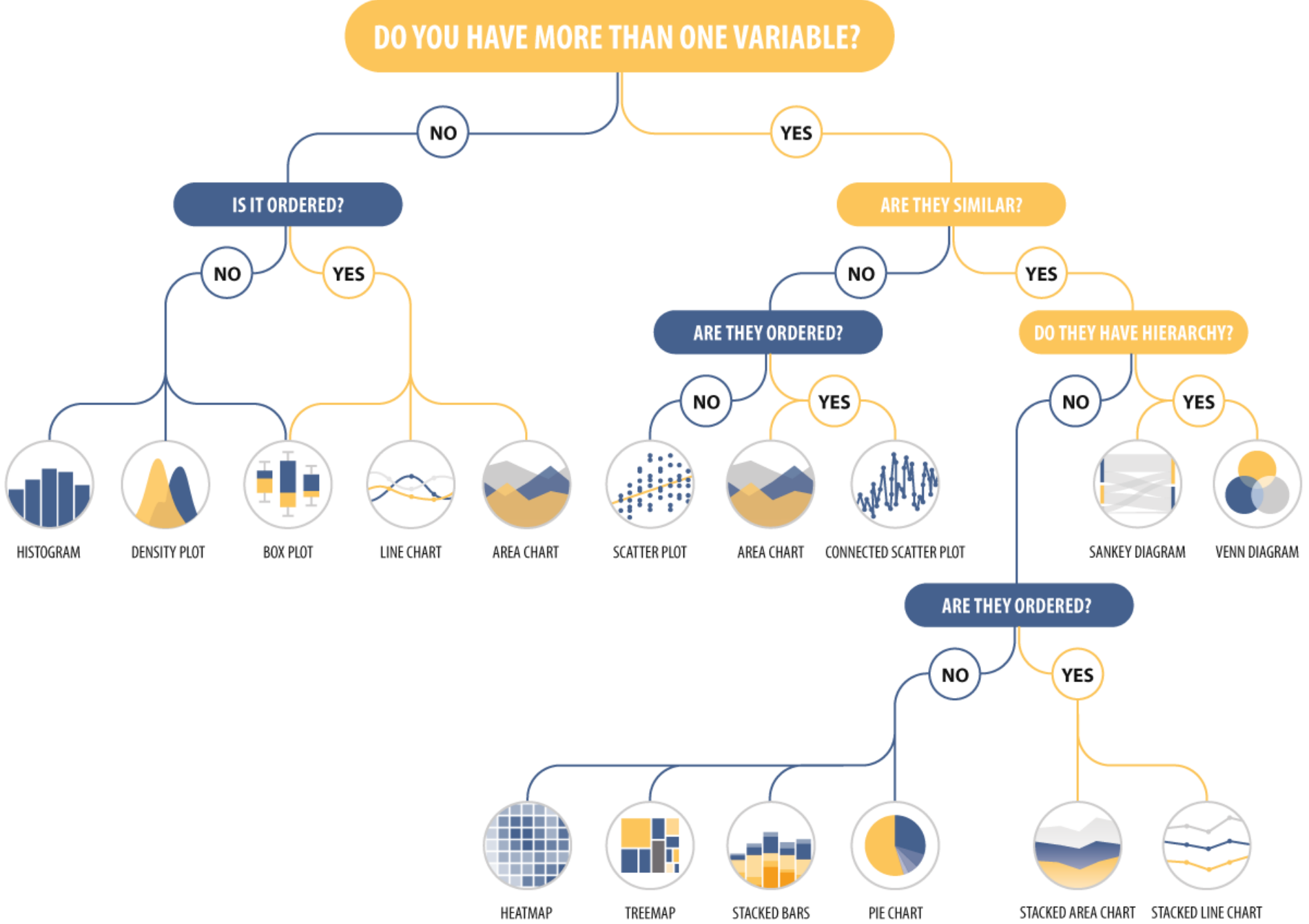
- A visualização de dados é a arte na qual números tornam-se em informação e, posteriormente, conhecimento útil.
- Ela possui dois papéis fundamentais:

  - comunicação dos resultados de forma clara para uma audiência e;

  - organizar uma exibição dos dados de forma que sugira uma nova hipótese ou uma próxima etapa no projeto,
- A linguagem R possui diversas funções próprias para plotagem (base R) como também pacotes específicos para visualização de dados. Aqui vamos focar no base R.

Tipo de análise pretendida	Tipo de variáveis envolvidas	Tipo de gráfico apropriado
• Distribuição? • Comparação? • Relação?	• Qualitativa? • Quantitativa?	• Histograma? • Gráfico de barras/linhas? • Gráfico de dispersão • Boxplot?

- Seria melhor uma tabela? 



https://medium.com/activewizards-machine-learning-company/how-to-choose-the-right-chart-type-infographic-86ca7c7fdf70

Existe um cardápio variado de formas de apresentação dos dados...

Visualização dos dados

Análise univariada vs bivariada

- Existem duas formas distintas que serão usadas para a análise dos dados:

Univariada

Analisa cada variável individualmente sem verificar relações ou criar comparações entre elas

medidas descritivas

histogramas

boxplots

diagramas de barras

VS

Bivariada

Analisa a relação de duas (ou mais) variáveis de um conjunto de dados, procurando entender suas distribuições e variações

correlações

diagramas de dispersão

medidas descritivas em cada grupo

boxplot

tabelas de contingência

Visualização dos dados

Arquivo de trabalho

Leitura do arquivo

```
> path <- 'D:/AULAS FGV/3. INFERENCIA ESTATISTICA/0. Desenvolvimento/DADOS/DADOS v2/'
>
> data <- read.csv(str_c(path,'Car Prices/car_prices.csv', sep = '', collapse = ''),
+                  sep = ',', dec = '.',
+                  stringsAsFactors = T,
+                  header = T, row.names = NULL)
```

- O R permite visualizar tanto "a cara" do banco de dados e a sua estrutura básica.

Exemplo

```
> head(data, n = 4)
   maker car_age mileage engine_power transmission door_count seat_count price_eur
1 nissan      10  149465          121         auto          5           5   4811.25
2   ford       7   99713           74          man          5           5   6476.68
3 toyota       0       5           97          man          5           5  14985.20
4 toyota       0       5           51          man          5           4   8878.57
>
> str(data)
'data.frame':   111726 obs. of  8 variables:
 $ maker       : Factor w/ 9 levels "audi","bmw","citroen",..: 8 5 9 9 3 2 6 3 3 4 ...
 $ car_age     : int  10 7 0 0 9 7 5 8 9 11 ...
 $ mileage     : int  149465 99713 5 5 91960 107763 29934 52542 99039 63588 ...
 $ engine_power: int  121 74 97 51 65 225 57 92 65 76 ...
 $ transmission: Factor w/ 2 levels "auto","man": 1 2 2 2 2 1 2 2 2 2 ...
 $ door_count  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ seat_count  : int  5 5 5 4 5 4 5 5 5 5 ...
 $ price_eur   : num  4811 6477 14985 8879 3886 ...
```

Cabeçalho da base de dados (4 primeiras linhas). Para ver o final basta usar `tail()`

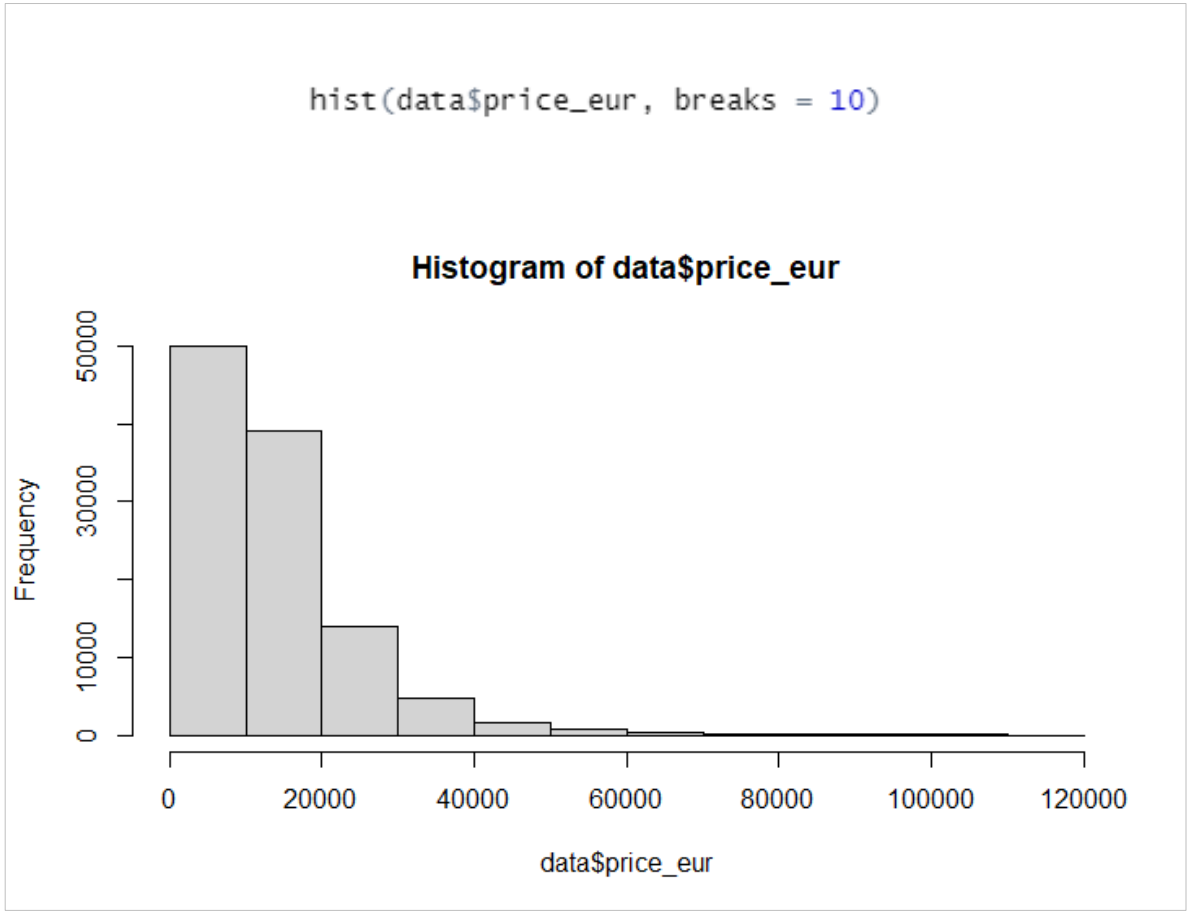
Estrutura do *data frame* mostrando todas as variáveis, formatos, etc.

Visualização dos dados

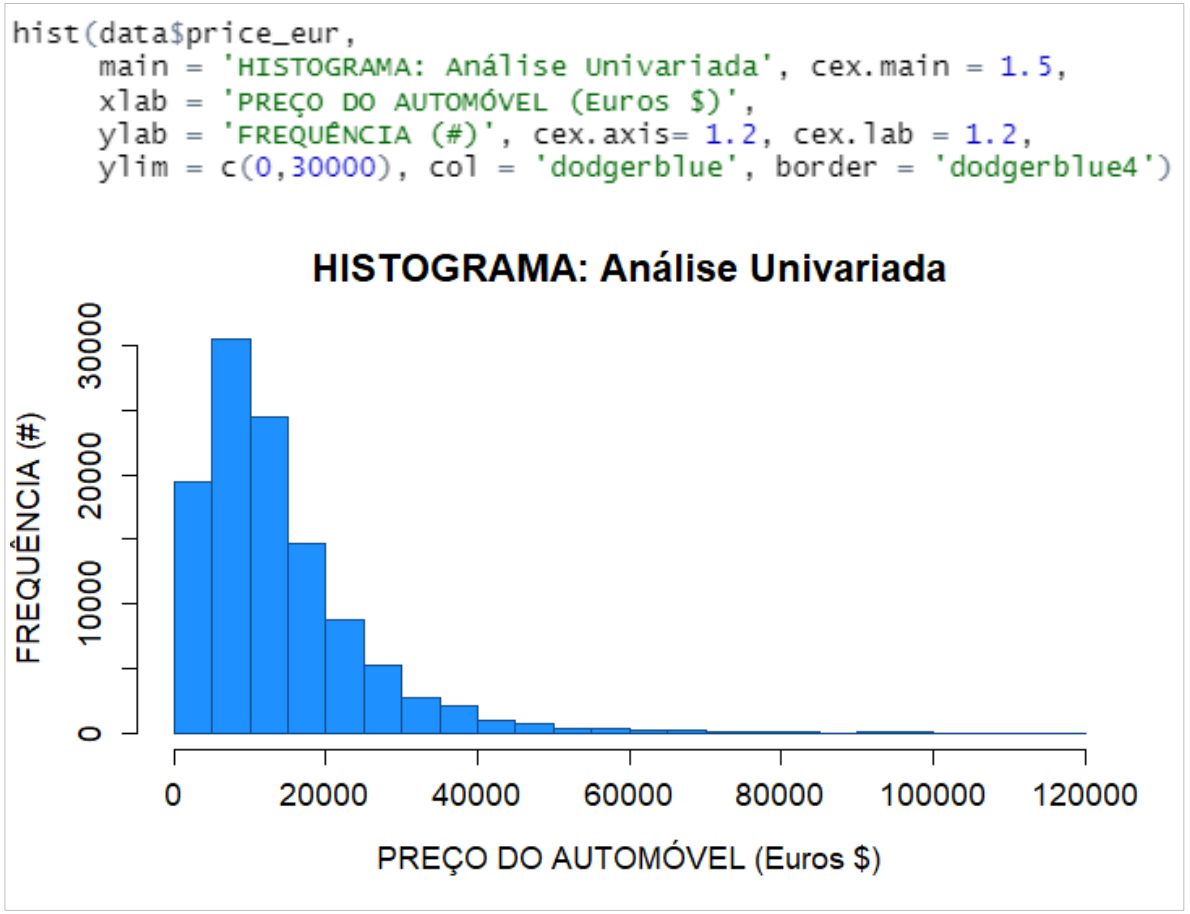
Histograma

- Histograma é usado para plotar variáveis contínuas. Ele particiona os dados em faixas e mostra a frequência de distribuição.

```
hist(data$price_eur, breaks = 10)
```




```
hist(data$price_eur,
     main = 'HISTOGRAMA: Análise Univariada', cex.main = 1.5,
     xlab = 'PREÇO DO AUTOMÓVEL (Euros $)',
     ylab = 'FREQUÊNCIA (#)', cex.axis= 1.2, cex.lab = 1.2,
     ylim = c(0,30000), col = 'dodgerblue', border = 'dodgerblue4')
```

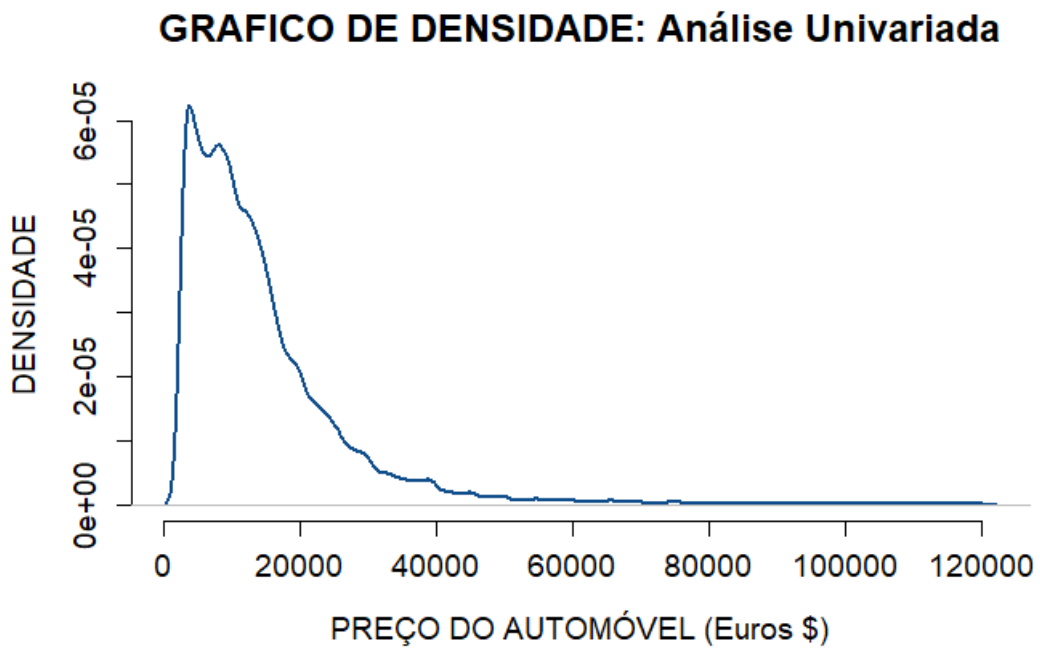


Visualização dos dados

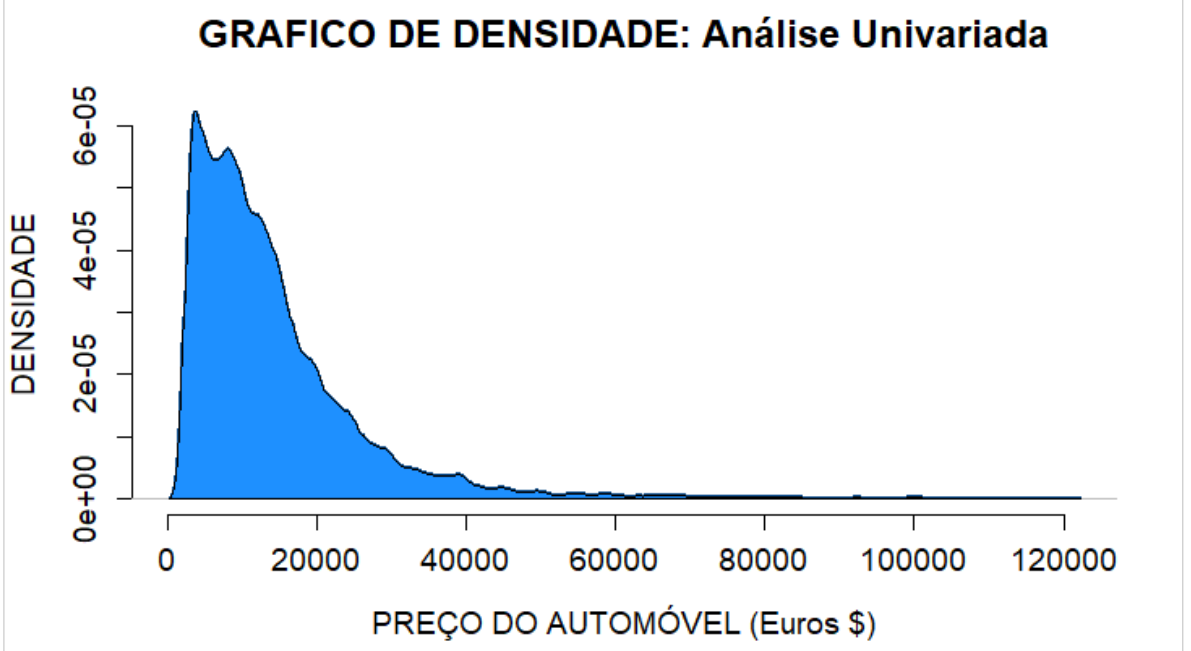
Density plot (univariado)

- *Density plot* são usados para plotar a representação da distribuição de uma variável quantitativa. Ela utiliza um kernel de estimativa de densidade para mostrar a função de densidade de probabilidade da variável.

```
plot(density(data$price_eur), frame = FALSE,
     main = 'GRAFICO DE DENSIDADE: Análise Univariada', cex.main = 1.5,
     xlab = 'PREÇO DO AUTOMÓVEL (Euros $)',
     ylab = 'DENSIDADE', cex.axis= 1.2, cex.lab = 1.2,
     lwd = 2, col = 'dodgerblue4')
```




```
plot(density(data$price_eur),
     main = 'GRAFICO DE DENSIDADE: Análise Univariada', cex.main = 1.5,
     xlab = 'PREÇO DO AUTOMÓVEL (Euros $)',
     ylab = 'DENSIDADE', cex.axis= 1.2, cex.lab = 1.2,
     lwd = 2, col = 'dodgerblue4')
polygon(density(data$price_eur), col = "dodgerblue")
```



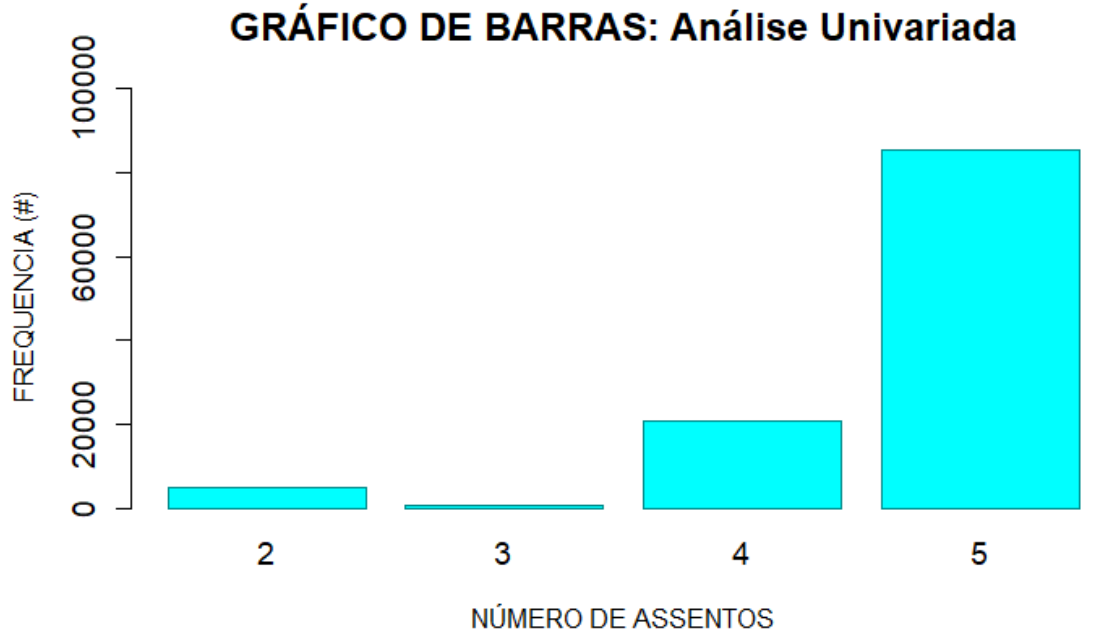
https://r-coder.com/density-plot-r/

Visualização dos dados

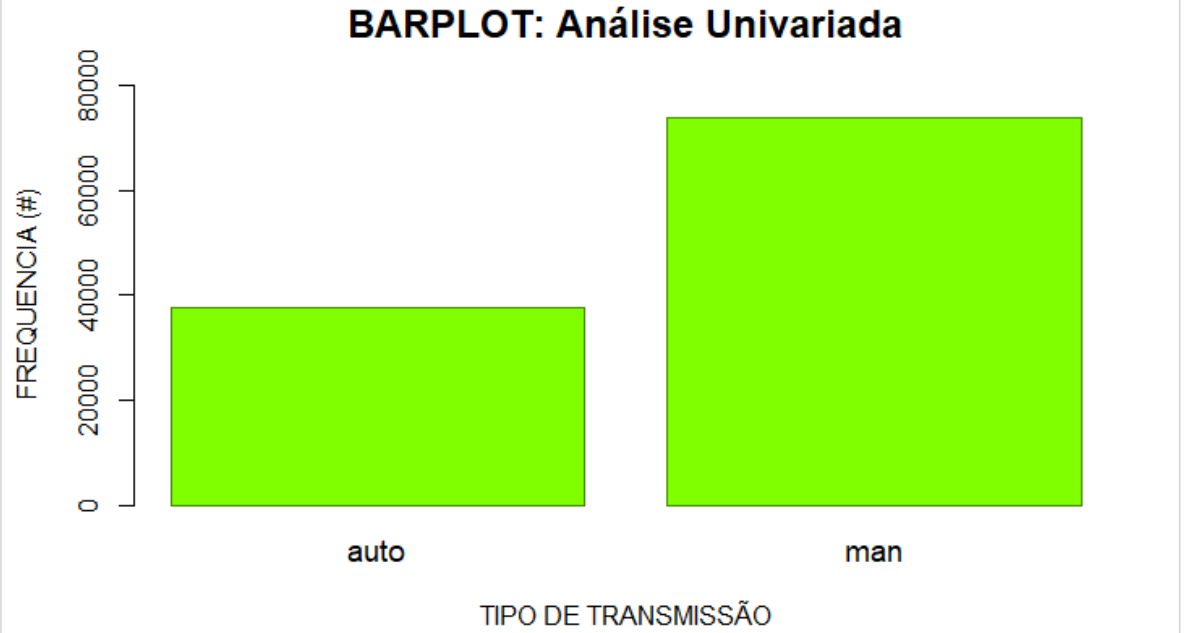
Diagrama de barras (univariado)

- Diagramas de barra são recomendados quando se quer plotar variáveis categóricas, variáveis quantitativas discretas (quando há poucos valores).

```
barplot(table(data$seat_count),
        main = 'GRÁFICO DE BARRAS: Análise Univariada', cex.main = 1.5,
        cex.names = 1.2,
        xlab = 'NÚMERO DE ASSENTOS',
        ylab = 'FREQUENCIA (#)', cex.axis = 1.2,
        ylim = c(0,100000), col = 'cyan', border = 'darkcyan')
```




```
barplot(table(data$transmission),
        main = 'BARPLOT: Análise Univariada', cex.main= 1.5,
        cex.names = 1.2,
        xlab = 'TIPO DE TRANSMISSÃO',
        ylab = 'FREQUENCIA (#)', cex.axis = 1.0,
        ylim = c(0,80000), col = 'chartreuse',border = 'chartreuse4')
```

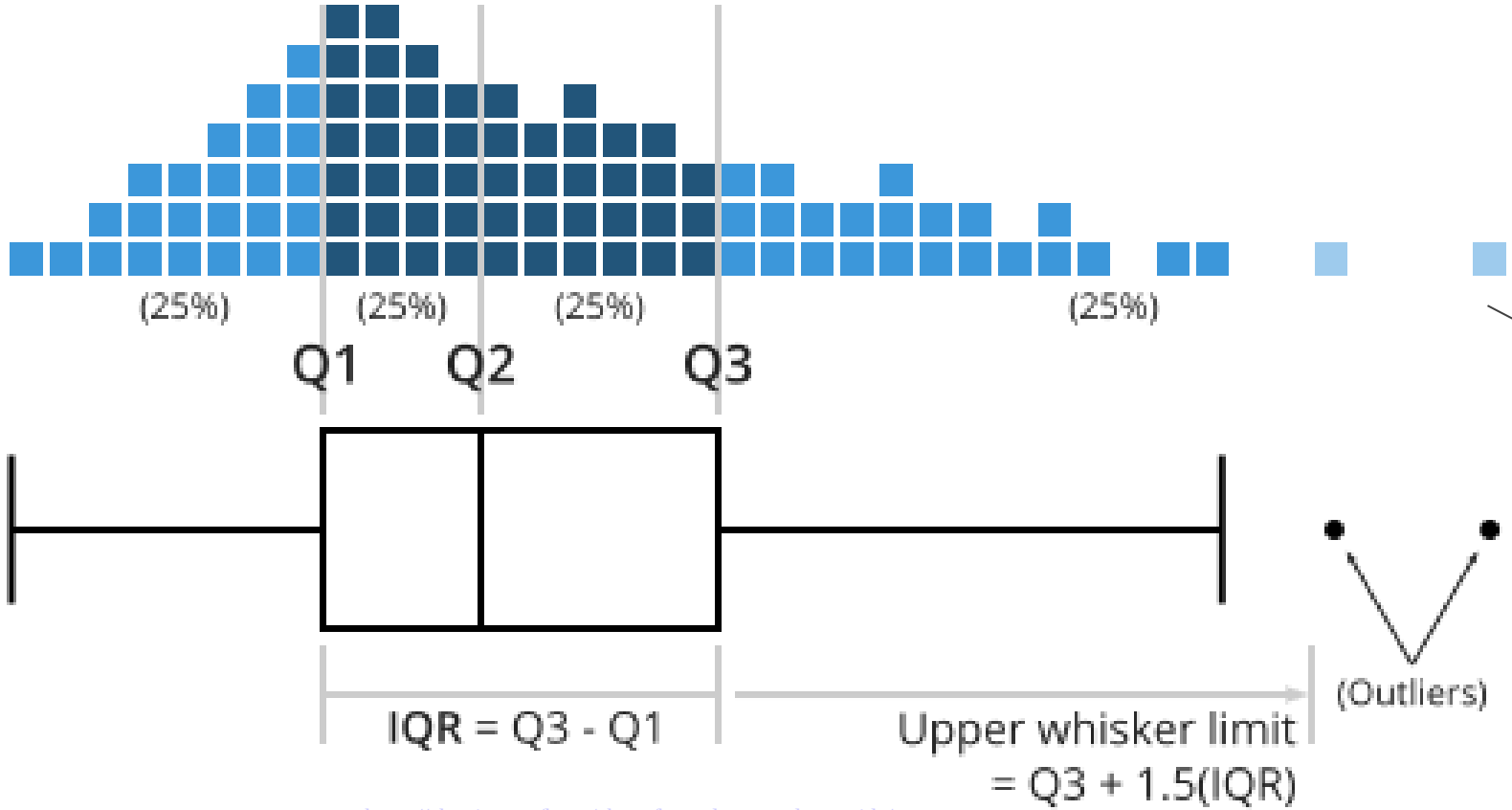


Visualização dos dados

Boxplots (univariado)

- *Boxplots* são usados descrever a distribuição de um dado através de algumas quantidades. Esse tipo de plot é útil para a visualização do espalhamento dos dados e detecção de *outliers.*

Estrutura



https://chartio.com/learn/charts/box-plot-complete-guide/

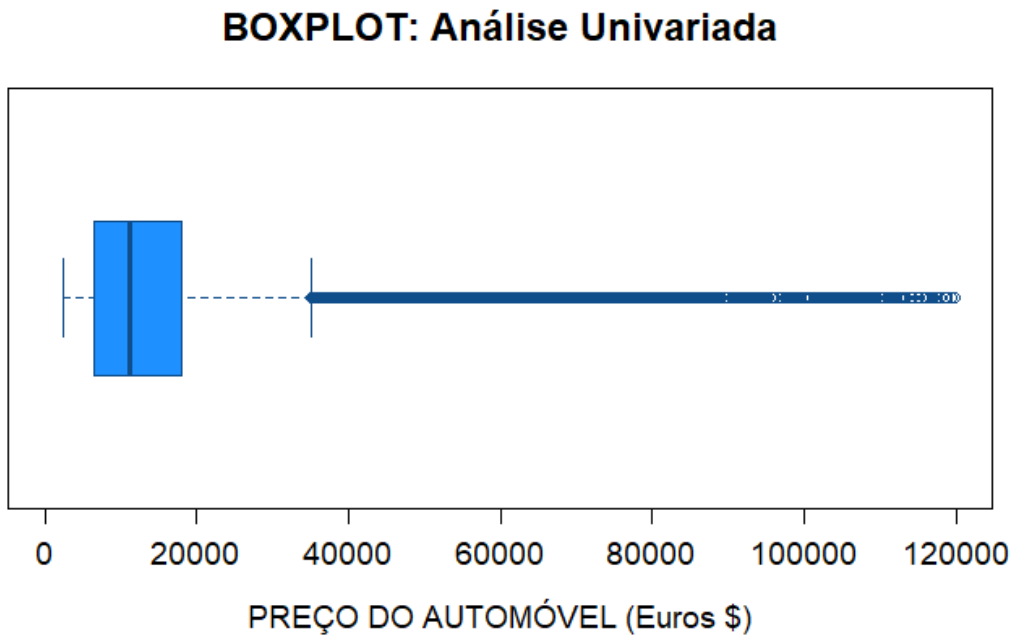
Cuidado com distribuições que são assimétricas!!! Nem tudo é *outlier*.

Visualização dos dados

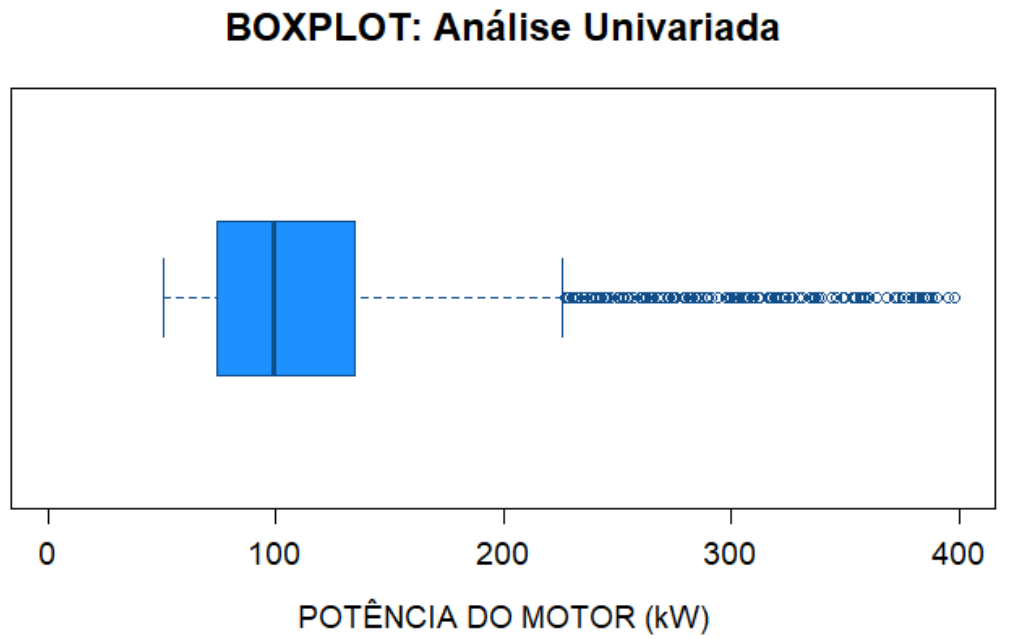
Boxplots (univariado)

- *Boxplots* são usados descrever a distribuição de um dado através de algumas quantidades. Esse tipo de plot é útil para a visualização do espalhamento dos dados e detecção de *outliers.*

```
boxplot(data$price_eur,
        main = 'BOXPLOT: Análise Univariada', cex.main = 1.5,
        xlab = 'PREÇO DO AUTOMÓVEL (Euros $)', cex.axis = 1.2,
        horizontal = T, cex.lab = 1.2,
        ylim = c(0,120000), col = 'dodgerblue', border = 'dodgerblue4')
```




```
boxplot(data$engine_power,
        main = 'BOXPLOT: Análise Univariada', cex.main = 1.5,
        xlab = 'POTÊNCIA DO MOTOR (kW)', cex.axis = 1.2,
        horizontal = T, cex.lab = 1.2,
        ylim = c(0,400), col = 'dodgerblue', border = 'dodgerblue4')
```



Visualização dos dados

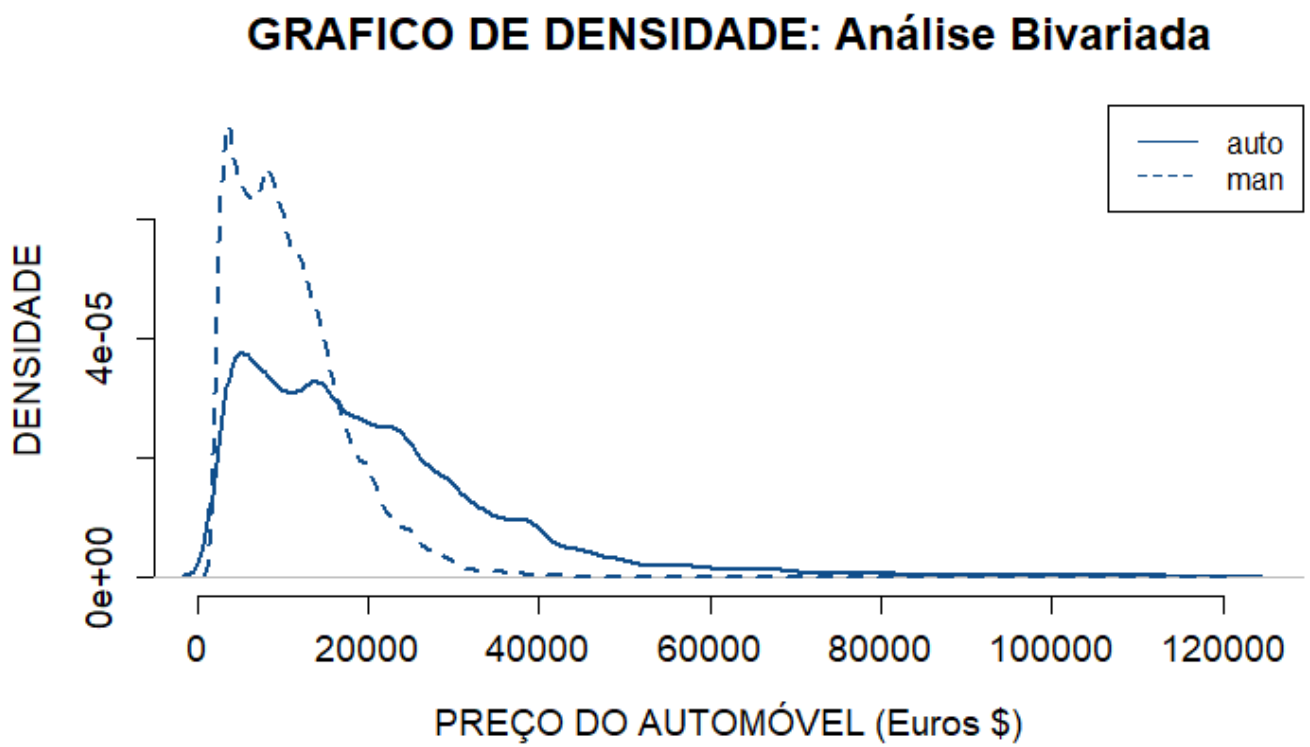
Density plots (bivariado)

- *Density plot* podem ser usados também para plotar as densidades de probabilidade de vários grupos ao mesmo tempo criando efeito de comparação entre as curvas.

```
g1 <- density(data$price_eur[data$transmission == 'auto'])
g2 <- density(data$price_eur[data$transmission == 'man'])

plot(g1, frame = FALSE,
     main = 'GRAFICO DE DENSIDADE: Análise Bivariada',
     cex.main = 1.5,
     xlab = 'PREÇO DO AUTOMÓVEL (Euros $)',
     xlim = c(0, max(g1$x, g2$x)),
     ylab = 'DENSIDADE', ylim = c(0, max(g1$y, g2$y)),
     cex.axis= 1.2, cex.lab = 1.2,
     lwd = 2, col = 'dodgerblue4')
lines(g2, lwd = 2, col = 'dodgerblue4', lty = 2)

legend("topright", levels(data$transmission),
       col = 'dodgerblue4', lty = 1:2)
```

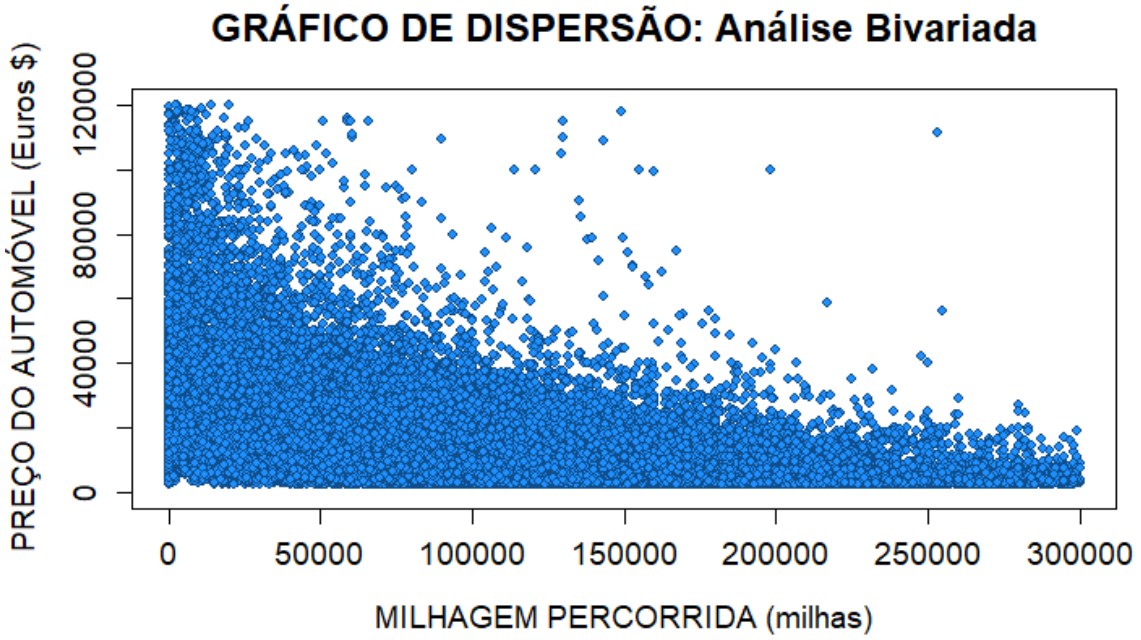


Visualização dos dados

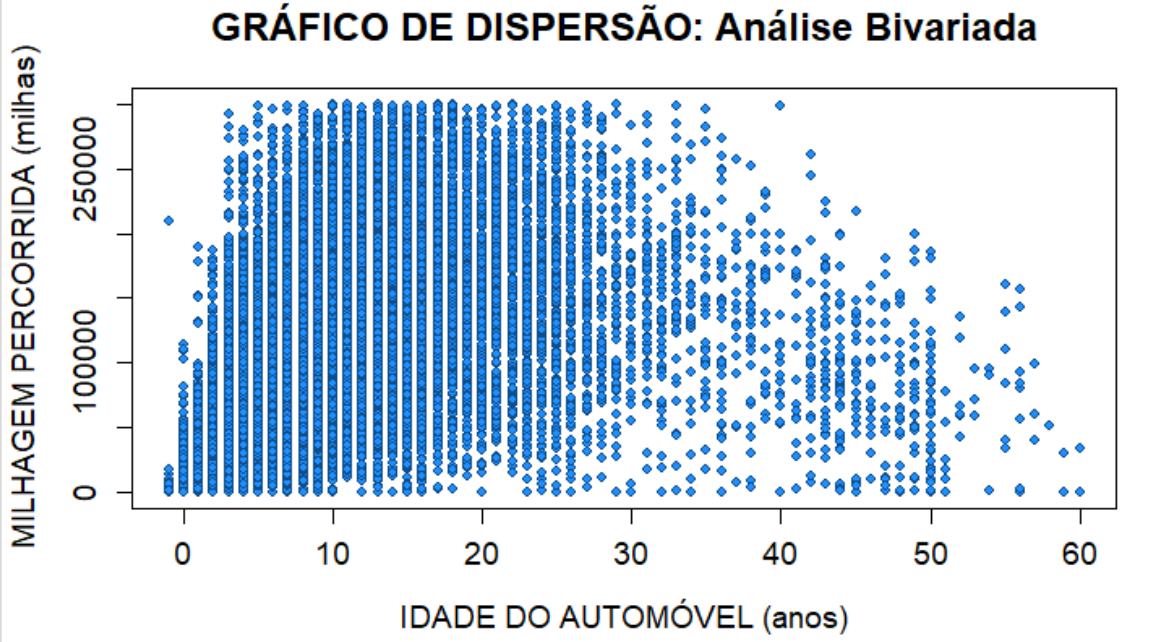
Diagrama de dispersão (bivariado)

- Diagramas de dispersão (*scatterplots*) é usado para ver a relação entre duas variáveis contínuas em um plano cartesiano (x, y).

```
plot(data$mileage, data$price_eur,
     main = 'GRÁFICO DE DISPERSÃO: Análise Bivariada', cex.main = 1.5,
     xlab = 'MILHAGEM PERCORRIDA (milhas)',
     ylab = 'PREÇO DO AUTOMÓVEL (Euros $)',
     cex.axis = 1.2, cex.lab = 1.2, pch = 21, cex = 0.8,
     col = 'dodgerblue4', bg = 'dodgerblue', ylim = c(0,120000))
```




```
plot(data$car_age, data$mileage,
     main = 'GRÁFICO DE DISPERSÃO: Análise Bivariada', cex.main = 1.5,
     xlab = 'IDADE DO AUTOMÓVEL (anos)',
     ylab = 'MILHAGEM PERCORRIDA (milhas)',
     cex.axis = 1.2, cex.lab = 1.2, pch = 21, cex = 0.8,
     col = 'dodgerblue4', bg = 'dodgerblue', ylim = c(0,300000))
```

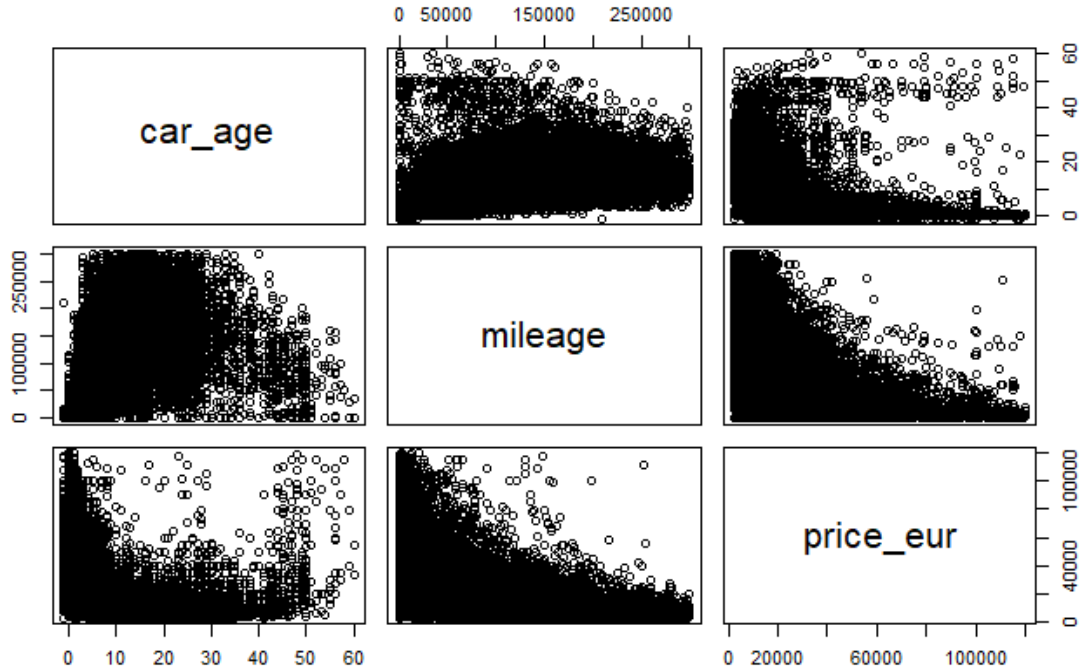


Visualização dos dados

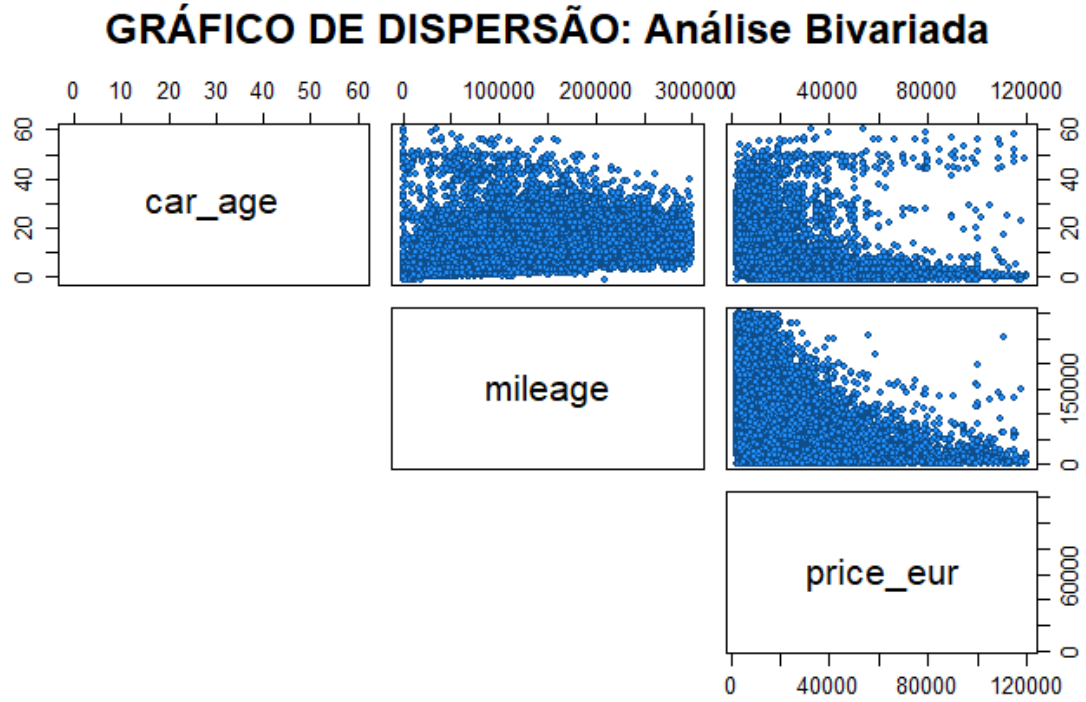
Pair plot (bivariado)

- A função do R de *pair plot* retorna uma matriz de diagramas de dispersão para cada combinação de variável quantitativa do *dataframe.*

```
pairs(~ car_age + mileage + price_eur , data = data)
```




```
pairs(~ car_age + mileage + price_eur , data = data, lower.panel = NULL,
      main = 'GRÁFICO DE DISPERSÃO: Análise Bivariada', cex.main = 1.5,
      cex.axis = 1.2, cex.lab = 1.2, pch = 21, cex = 0.8,
      col = 'dodgerblue4', bg = 'dodgerblue')
```



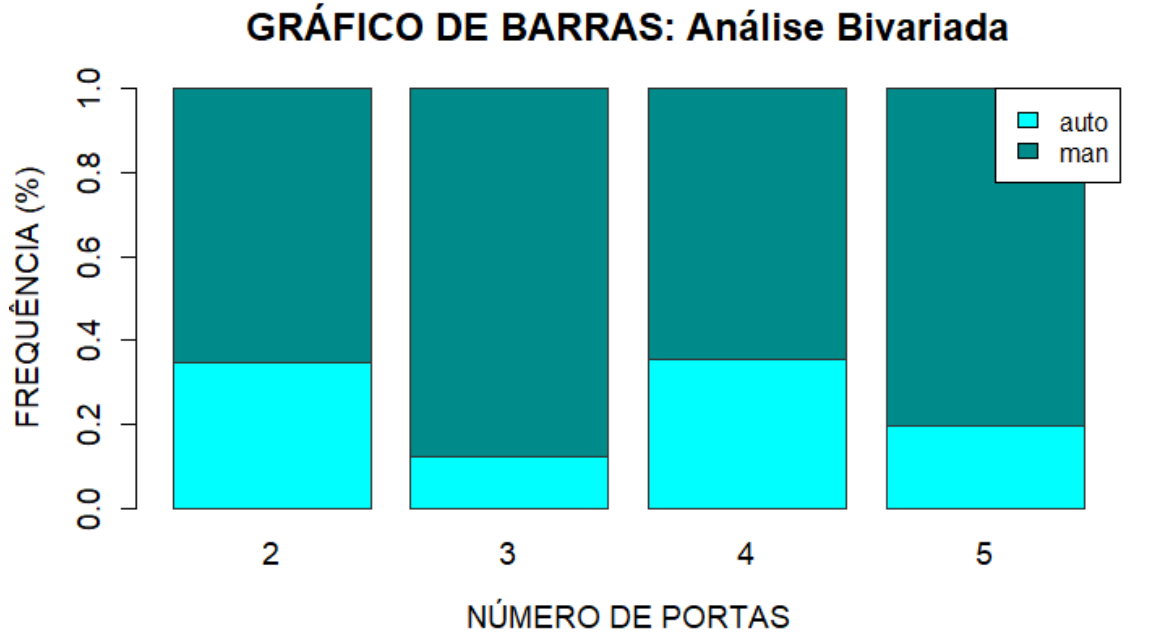
http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

Visualização dos dados

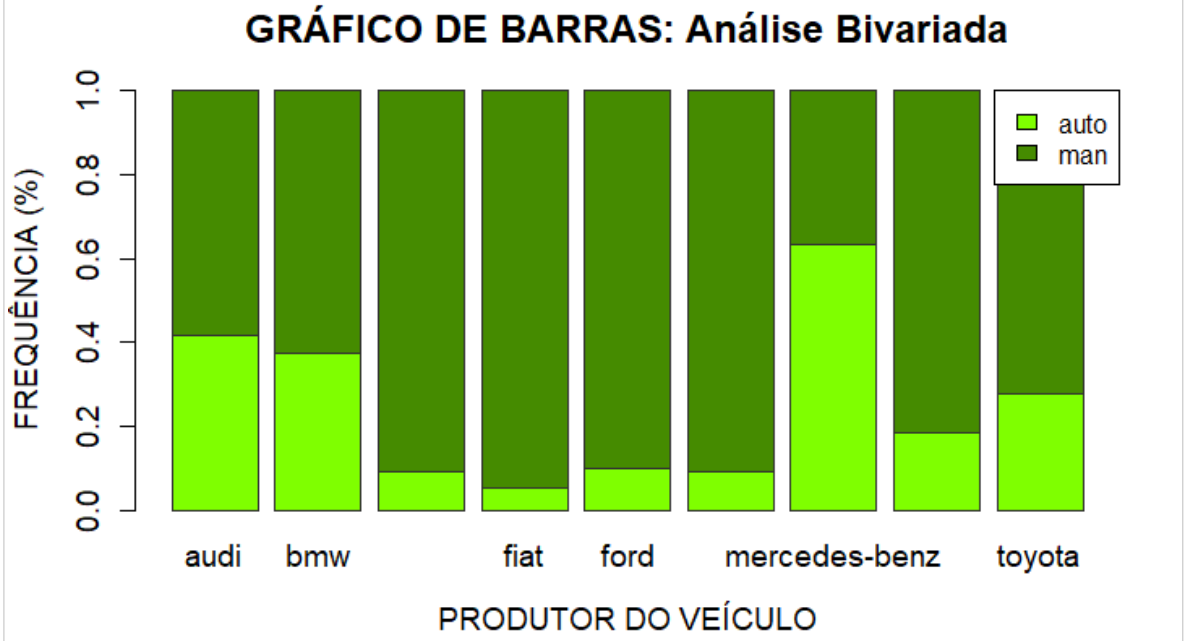
Diagrama de barras (bivariado)

- Diagramas de barra podem ser combinados para mostrar a proporção das categorias de uma variável versus outra (categórica ou quantitativa discreta). As barras podem somar 100% ou totais relativos.

```
tbl <- prop.table(table(data$transmission,data$door_count),margin = 2)
barplot(tbl,
        main = 'GRÁFICO DE BARRAS: Análise Bivariada', cex.main= 1.5,
        cex.names = 1.2, cex.lab = 1.2, cex.axis = 1.2, ylim = c(0,1),
        xlab = 'NÚMERO DE PORTAS', ylab = 'FREQUÊNCIA (%)',
        col = c('cyan','darkcyan'), border = 'gray20')
legend("topright", legend = row.names(tbl), fill = c('cyan','darkcyan'))
```




```
tbl <- prop.table(table(data$transmission,data$maker),margin = 2)
barplot(tbl,
        main = 'GRÁFICO DE BARRAS: Análise Bivariada', cex.main= 1.5,
        cex.names = 1.2, cex.lab = 1.2, cex.axis = 1.2, ylim = c(0,1),
        xlab = 'PRODUTOR DO VEÍCULO', ylab = 'FREQUÊNCIA (%)',
        col = c('chartreuse','chartreuse4'), border = 'gray20')
legend("topright", legend = row.names(tbl), fill = c('chartreuse','chartreuse4'))
```

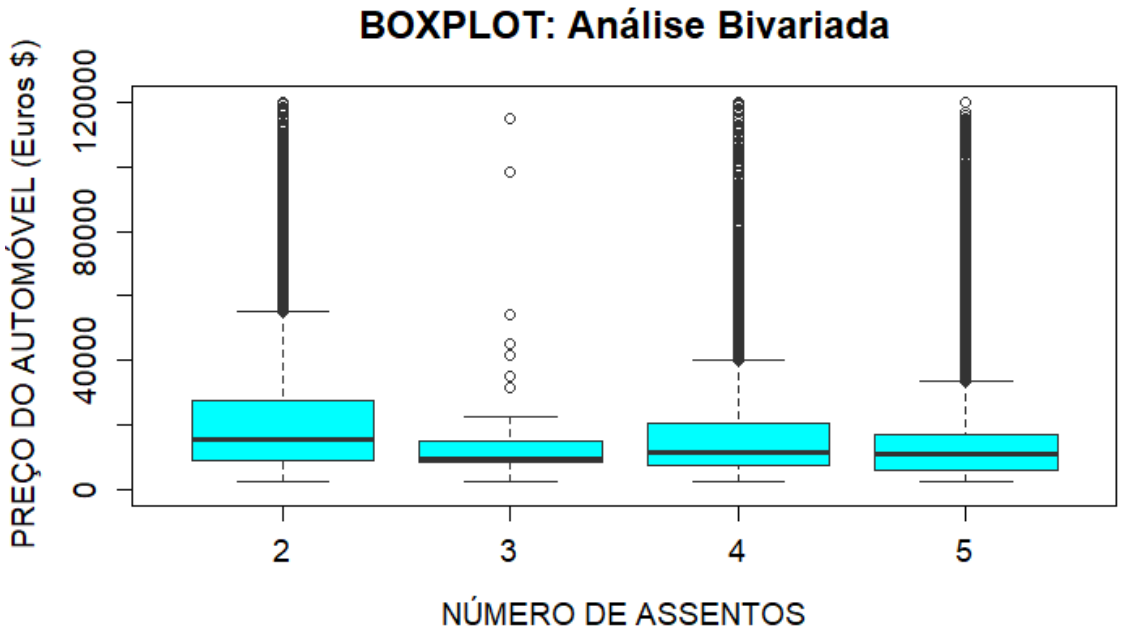


Visualização dos dados

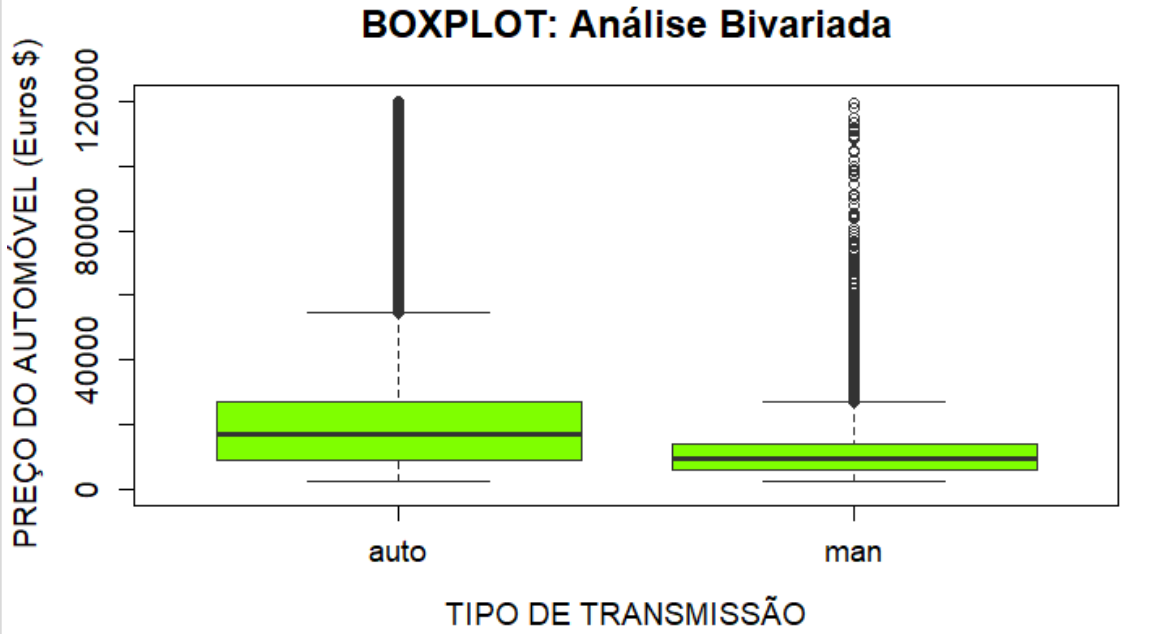
Boxplots (bivariado)

- *Boxplots* podem ser usados para plotar a distribuição de uma variável quantitativa em relação a outra variável, seja ela quantitativa discreta ou categórica. Com isso é possível fazer comparativos.

```
boxplot(price_eur ~ seat_count, data = data,
        main = 'BOXPLOT: Análise Bivariada', cex.main = 1.5,
        xlab = 'NÚMERO DE ASSENTOS',
        ylab = 'PREÇO DO AUTOMÓVEL (Euros $)', cex.axis = 1.2,
        horizontal = F, cex.lab = 1.2,
        ylim = c(0,120000), col = 'cyan', border = 'gray20')
```




```
boxplot(price_eur ~ transmission, data = data,
        main = 'BOXPLOT: Análise Bivariada', cex.main = 1.5,
        xlab = 'TIPO DE TRANSMISSÃO',
        ylab = 'PREÇO DO AUTOMÓVEL (Euros $)', cex.axis = 1.2,
        horizontal = F, cex.lab = 1.2,
        ylim = c(0,120000), col = 'chartreuse', border = 'gray20')
```

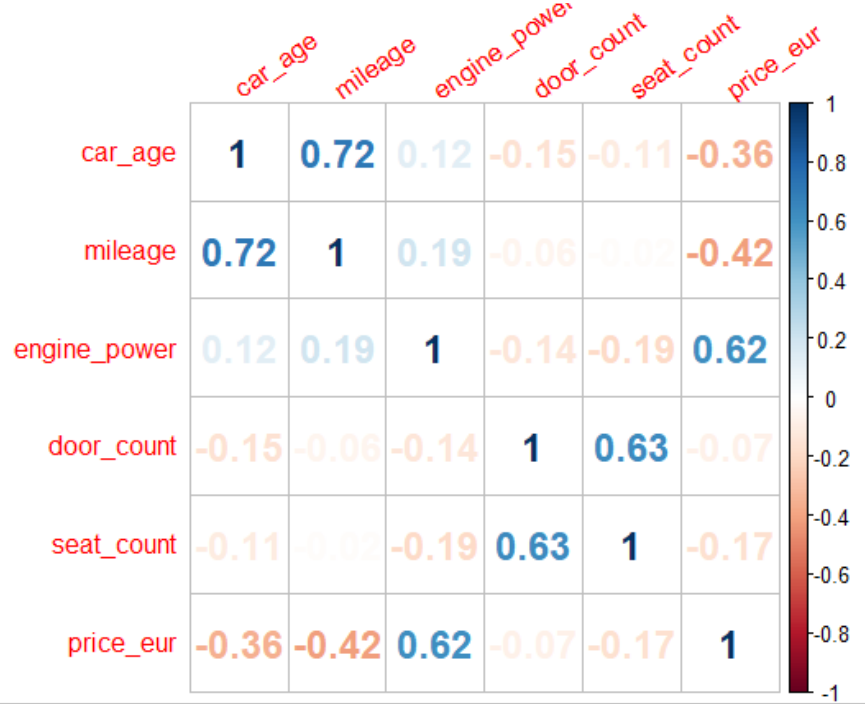


Visualização dos dados

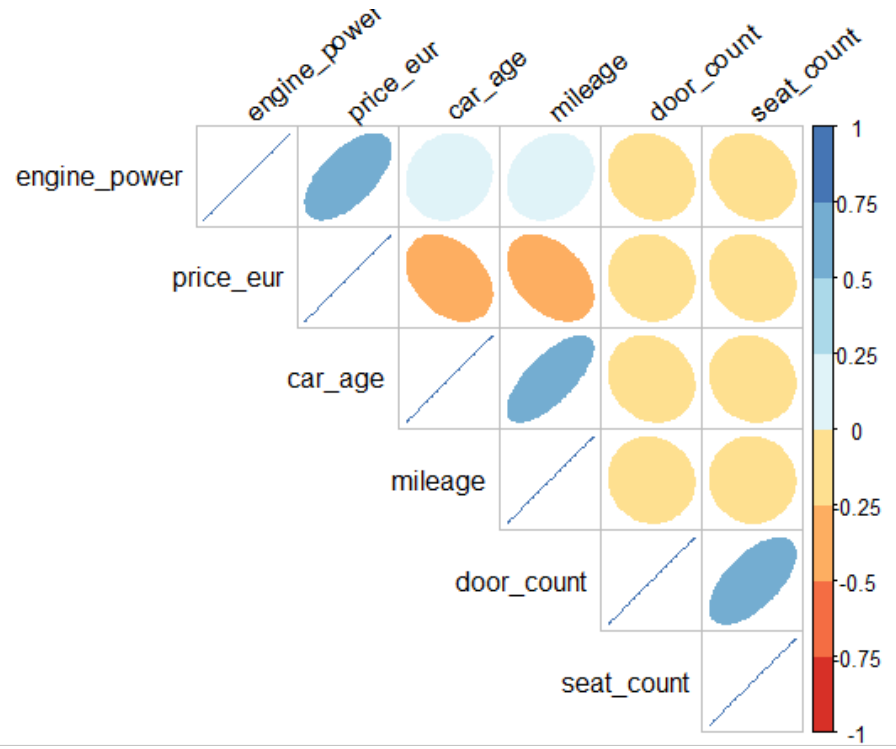
Correlograma

- São uma forma de destacar as correlações par a par entre as variáveis ao apresentar uma gráfico na forma de matriz de correlação. Pode-se colorir em função dos ranges ou ordenar em função de similaridade.

```
library(corrplot)
library(RColorBrewer)
M <- cor(data[,c(2,3,4,6,7,8)])
corrplot(M, method = "number", tl.srt = 35,
         number.cex = 1.5)
```




```
M <- cor(data[,c(2,3,4,6,7,8)])
corrplot(M, type="upper", order="hclust", method = 'ellipse',
         tl.col="black", number.cex = 1.5, tl.srt = 40,
         col = brewer.pal(n=8, name="RdYlBu"))
```



http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram

Visualização dos dados

Tabelas de frequência e contingência

- São formas de registrar as observações de variáveis categóricas, sejam elas de forma univariada ou confrontando outra variável. Aqui observam-se as volumetrias (e %) das observações em cada classe.

Frequencies

data$maker

Type: Factor

maker	Freq	%	% Cum.
audi	20280	18.15	18.15
bmw	21165	18.94	37.10
citroen	4243	3.80	40.89
fiat	7711	6.90	47.79
ford	15345	13.73	61.53
hyundai	5405	4.84	66.37
mercedes-benz	24209	21.67	88.04
nissan	5093	4.56	92.59
toyota	8275	7.41	100.00
Total	111726	100.00	100.00

Cross-Tabulation, Row Proportions

maker * transmission

Data Frame: data

	transmission		
maker	**auto**	**man**	**Total**
audi	8457 (41.7%)	11823 (58.3%)	20280 (100.0%)
bmw	7934 (37.5%)	13231 (62.5%)	21165 (100.0%)
citroen	387 (9.1%)	3856 (90.9%)	4243 (100.0%)
fiat	413 (5.4%)	7298 (94.6%)	7711 (100.0%)
ford	1520 (9.9%)	13825 (90.1%)	15345 (100.0%)
hyundai	500 (9.3%)	4905 (90.7%)	5405 (100.0%)
mercedes-benz	15335 (63.3%)	8874 (36.7%)	24209 (100.0%)
nissan	937 (18.4%)	4156 (81.6%)	5093 (100.0%)
toyota	2287 (27.6%)	5988 (72.4%)	8275 (100.0%)
Total	37770 (33.8%)	73956 (66.2%)	111726 (100.0%)

```
library(summarytools)
view(freq(data$maker, report.nas = F,
          style = 'rmarkdown'))
```

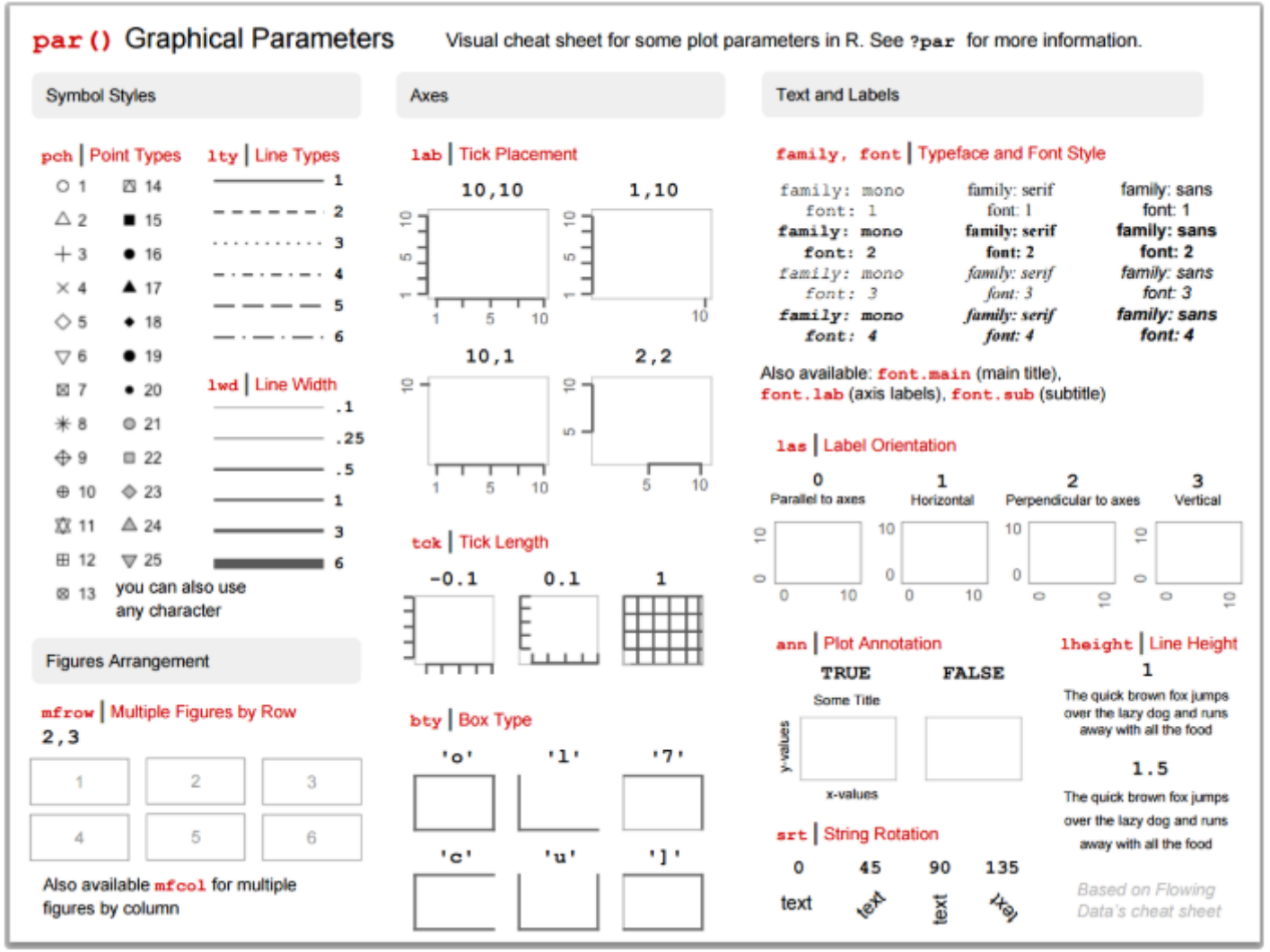
```
library(summarytools)
view(ctable(x = data$maker,
            y = data$transmission, prop = "r"))
```

Visualização dos dados

Parâmetros de plotagem [extra]

- O base R possui diversos parâmetros que podem ser usados para modificar o tipo de linhas, as formas, margens e fontes nos gráficos.

```
# Exemplo de plotagem com linhas
plot(endiv$total,
     bty=  "n", #frame do plot: "o" "l", "7", "c", "u", or "]"
     type= "o", #tipo de plot: "p", "l" "b" "c" "o" "h" "n"
     lwd= 2,
     lty=  "solid", #tipo de linha (pode ser numero, 0-6):  "blank",
     #                                                      "solid",
     #                                                      "dashed",
     #                                                      "dotted",
     #                                                      "dotdash",
     #                                                      "longdash",
     #                                                      "twodash"
     pch= 19,   #tipo de solido: 19: solid circle,
     #                           20: bullet, 21: filled circle,
     #                           22: filled square,23: filled diamond,
     #                           24: filled triangle point-up,
     #                           25: filled triangle point down
     col=  "blue",#tipo de cor da linha
     cex= 1,    #tamanho da forma geométrica que conecta a linha
     ylim= c(40,50), #limite do eixo y, c(min, max)
     las= 0,    # lables do eixo y vertical ou horizontal
     cex.axis= 1.2, #tamanho da fonte dos numeros dos eixos
     cex.lab= 1.2,  #tamanho da fonte dos lables dos eixos
     col.lab= "black",# cor da fonte dos labels dos eixos
     cex.main= 1.7,
     col.main= "black",
     yaxt="n", #retira o label do eixo y
     ylab= "TOTAL (%)",
     xlab= "TEMPO (meses)",
     main= "ENDIVIDAMENTO DAS FAMILIAS",
     col.main= "black",
     font.main= 1 #1=plain, 2=bold, 3=italic, 4=bold italic, 5=symbol
     )
```



http://www.gastonsanchez.com/visually-enforced/resources/2015/09/22/R-cheat-sheet-graphical-parameters/

Visualização dos dados

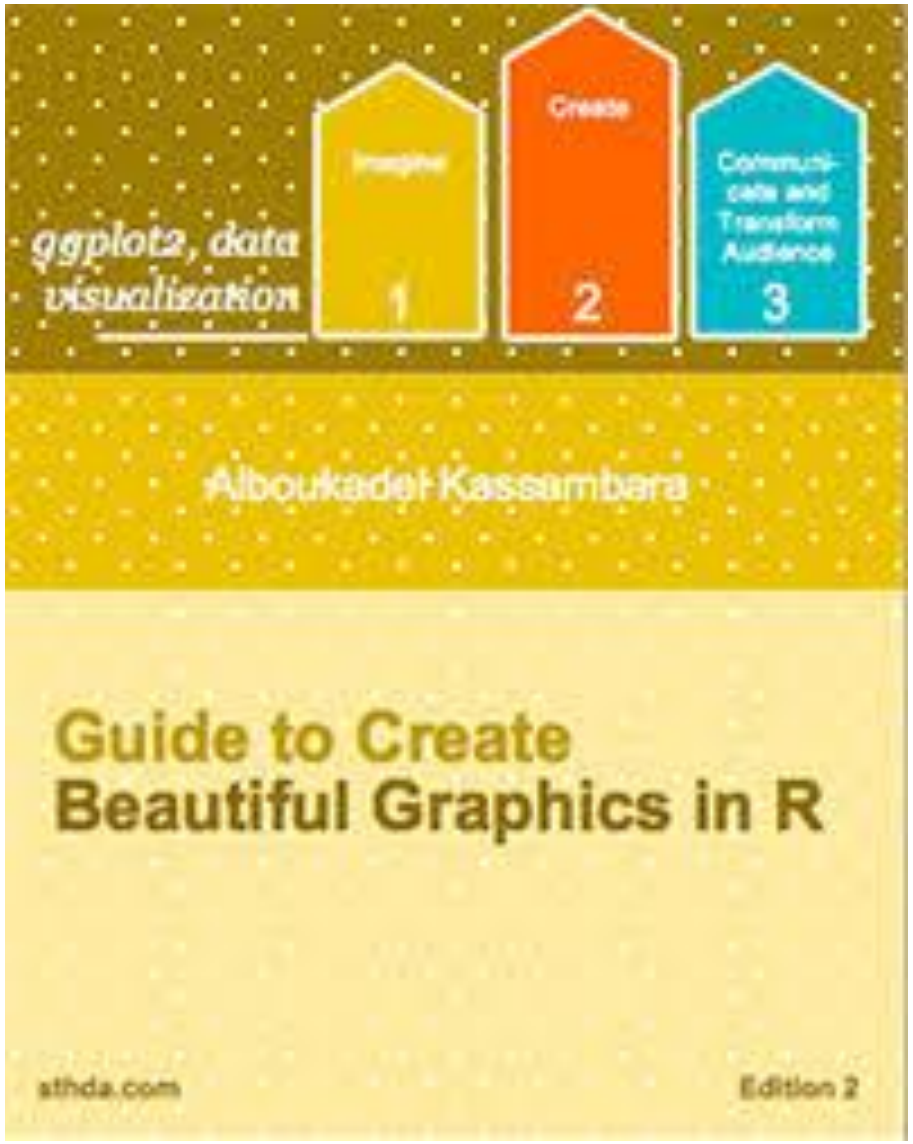
ggplot2 [extra]

- Além dos comandos básicos do base R, existem outros pacotes que permitem a visualização dos dados de forma clara e com apelo estético.
- Apresentação a linguagem dos gráficos como uma gramática, dividindo-os em componentes semânticos.
- O pacote ggplot2 é muito usado e fornece recursos extras que não existem no base R (documentação em: https://cloud.r-project.org/web/packages/ggplot2/ggplot2.pdf)

Acesso

```
# instalando o ggplot2
install.packages('ggplot2')

# carregando ggplot2
library(ggplot2)
```



http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-visualization

Prática no RStudio

Prof. Dr. João Rafael Dias Pinto

...foco de hoje

- **CASE 4: Dados censitários de domicílios na região de Ilocos, Filipinas**

  Explorando as variáveis da base de dados, criando plots univariados e bivariados adequados para as variáveis. Entendendo a conexão entre as grandezas.

Agenda

Na aula de hoje...

Apresentação do curso
Introdução ao R/Rstudio
Estrutura de dados
Comando básicos
Leitura e escrita de dados
Prática no RStudio

Manipulação de *strings* com `stringr`
Manipulação de datas com `lubridate`
Manipulação de *dataframes* com `dplyr`
Prática no RStudio

Tipos de variáveis
Medidas de Centralidade
Medidas de Dispersão
Medidas de Associação
Prática no RStudio

Técnicas de visualização
Tipos de gráficos
Análise univariada
Análise bivariada
Prática no RStudio

Trabalhos práticos
Aplicação do conteúdo