Escola Politécnica de Pernambuco
Especialização em Ciência de Dados e Analytics

Estatística Computacional

Aula 1.2 – Aplicações Computacionais da Estatística – PARTE II

Prof. Dr. Rodrigo Lins Rodrigues

rodrigolins.rodrigues@ufrpe.br

Importação de Bases de dados

Importação de dados

- Saber importar dados para a ferramenta de análise é um dos **passos mais importantes**;
- **Dificilmente** você fará todo o trabalho de análise **em uma só ferramenta**;
- Os dados podem estar **armazenados em diversos sistemas** e em diversos formatos.

Importação de dados

- De onde vem as bases que devem ser importadas?

...

Importação de dados

- Importando arquivos CSV

```
5   # Usando o pacote readr
6   install.packages("readr")
7   library(readr)
8
9   # Abre o promt para escolher o arquivo
10  meu_arquivo <- read_csv(file.choose())
11  meu_arquivo <- read_delim(file.choose(), sep = "|")
```

```
13  # Importando arquivos
14  df1 <- read_table("temperaturas.txt", col_names = c("DAY","MONTH","YEAR","TEMP"))
```

Importação de dados

- Importando arquivos Excel

```
# Instalando pacotes
install.packages("xlsx")
install.packages("readxl")
library(xlsx)
library(readxl)

# Lista as worksheet no arquivo Excel
excel_sheets("UrbanPop.xlsx")

# Importando com o pacote readxl
df <- read_excel("UrbanPop.xlsx", sheet = 3)
head(df)

# Importando com o pacote xlsx
df2 <- read.xlsx("UrbanPop.xlsx", sheetIndex = 1)
head(df2)
```

Importação de dados

- Outros pacotes de importação de dados

```
3  # package para banco de dados SQLite
4  install.packages("RSQLite")
5  # package para banco de dados Mysql
6  install.packages("RMySQL")
7  # package para dados do SPSS
8  install.packages("SPSStoR")
9  # package para dados em formato xml
10 install.packages("XML")
11 # package para banco de dados mongo db
12 install.packages("rmongodb")
```

Limpeza dos dados

- Problemas que devem ser tratados na fase de limpeza dos dados:

  - ✓Os cabeçalhos das colunas são valores e não nomes das variáveis;
  - ✓Diversas variáveis são armazenadas em uma coluna;
  - ✓As variáveis são distribuídas em diversas tabelas relacionais;
  - ✓As variáveis tem grandes variabilidades;
  - ✓Alta presença de valores nulos ou faltantes;
  - ✓Abreviações preenchidas de diversas formas por usuários;
  - ✓Etc..

Limpeza dos dados

Você vai receber os dados assim:

Você vai deixá-los assim:

Limpeza dos dados

- Pacote - *dplyr.R*

  ✓É um dos principais pacotes para o processo de limpeza de dados;

  ✓É ideal para manipulação de dados;

  ✓Ele não faz parte do pacote básico e precisa ser instalado;

Limpeza dos dados

- Pacote - *dplyr.R*

```r
install.packages("readr")
install.packages("dplyr")
library(readr)
library(dplyr)

# Carregando o dataset
df_sono <- read_csv("Base_de_dados/sono.csv")
head(df_sono)
str(df_sono)
```

Limpeza dos dados

- Pacote - *dplyr.R*

```
14  #contando a quantidade de vezes que a cidade aparece
15  count(df_sono, cidade)
16
17  # Mostrar a base com x linhas
18  sample_n(df_sono, size = 10)
19
20
21  # Filtrando de acordo com uma variável - filter()
22  filter(df_sono, sono_total >= 16)
23  filter(df_sono, sono_total >= 16, peso >= 80)
24  filter(df_sono, cidade %in% c("Recife", "Curitiba"))
```

Limpeza dos dados

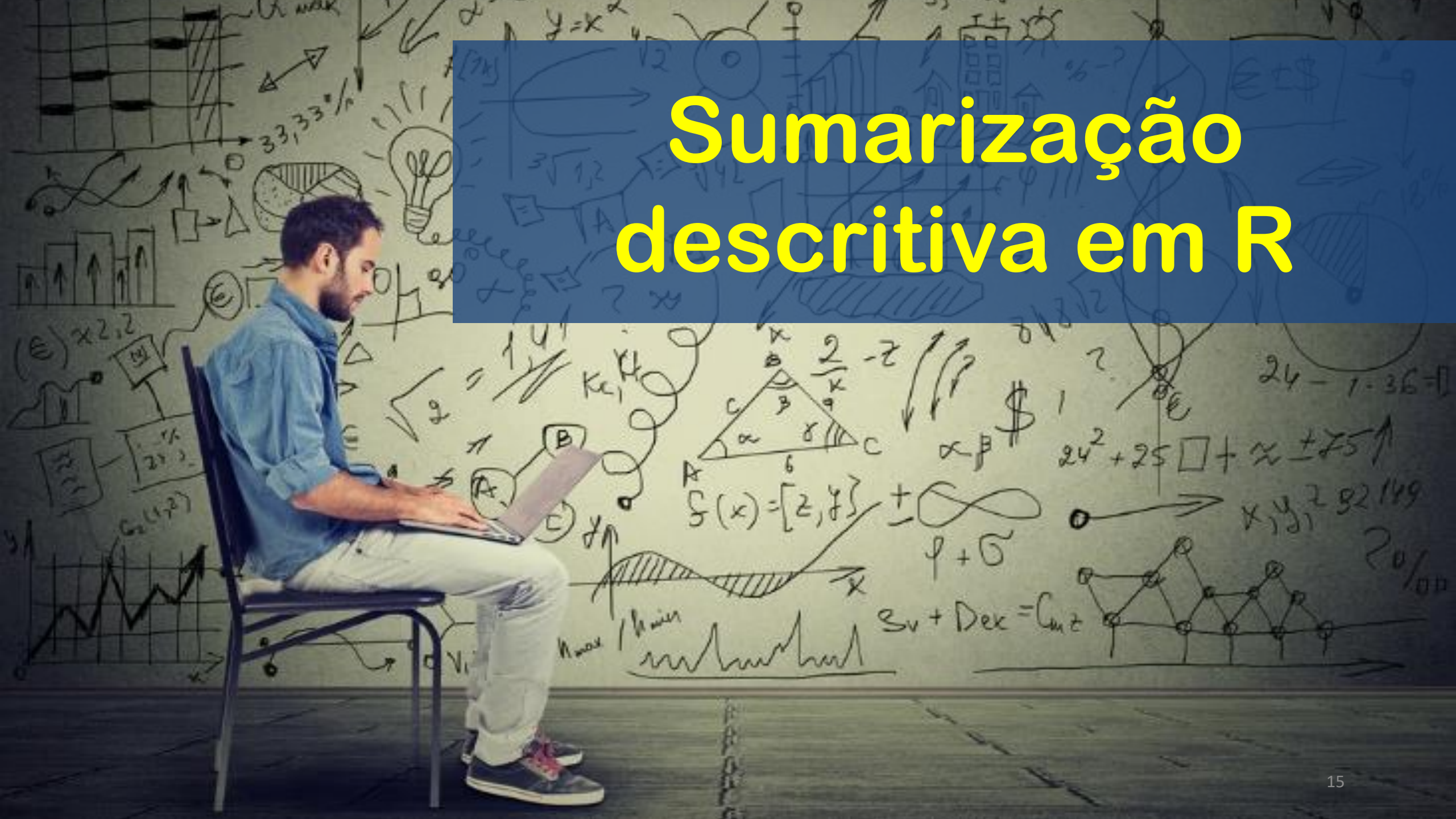
- Pacote - *dplyr.R*

```
27  # arrange()
28  df_sono %>% arrange(cidade) %>% head
29
30  df_sono %>%
31    select(nome, cidade, sono_total) %>%
32    arrange(cidade, sono_total) %>%
33    head
34
35  df_sono %>%
36    select(nome, cidade, sono_total) %>%
37    arrange(cidade, sono_total) %>%
38    filter(sono_total >= 16)
39
40  df_sono %>%
41    select(nome, cidade, sono_total) %>%
42    arrange(cidade, desc(sono_total)) %>%
43    filter(sono_total >= 16)
```

Limpeza dos dados

- Pacote - *dplyr.R*

```
46    # summarize()
47  df_sono %>%
48    summarise(media_sono = mean(sono_total))
49
50  df_sono %>%
51    summarise(media_sono = mean(sono_total),
52              min_sono = min(sono_total),
53              max_ssono = max(sono_total),
54              total = n())
55
```

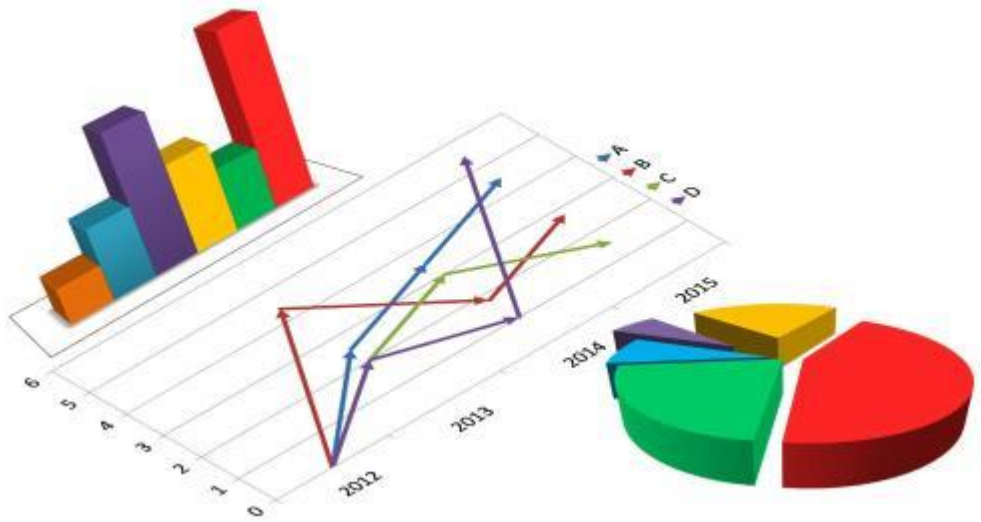


Sumarização descritiva em R

Sumarização Descritiva

- É importante sempre aplicar medidas descritivas antes de qualquer análise:

  ✓Medidas de tendência central;
  ✓Medidas de dispersão;
  ✓ Amplitude de variáveis;
  ✓Construção de tabelas;



Sumarização Descritiva

```r
# Carregando o dataset
carros <- read.csv2("Base_de_dados/carros-usados.csv", head=T, sep=",")

# Resumo dos dados
head(carros)
str(carros)

# Medidas de Tendencia Central
summary(carros$ano)
summary(carros[c('preco', 'kilometragem')])

mean(carros$preco)
median(carros$preco)
```

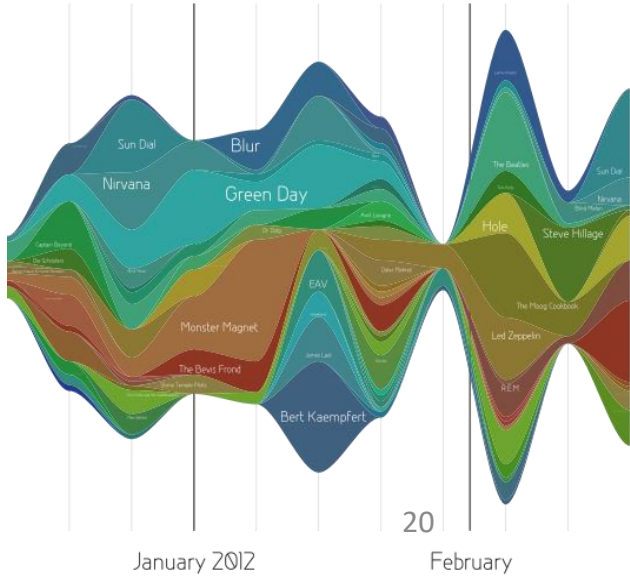
Sumarização Descritiva

```
19  # verificando a variabilidade dos dados
20  range(carros$preco)
21  diff(range(carros$preco))
22  var((carros$preco))
23  sd((carros$preco))
24
25  # Criando tabelas de contingencia
26  ?table
27  str(carros)
28  table(carros$cor)
29  table(carros$modelo)
30  str(carros)
31
32  # Calculando a proporcao de cada categoria
33  model_table <- table(carros$modelo)
34  prop.table(model_table)
```

Plotando gráficos em R

Plotando gráficos

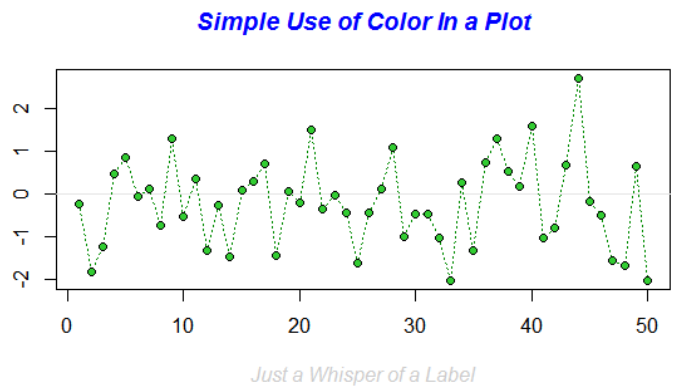
- Representação gráfica construída a partir de dados;
- *O R* é uma das melhores ferramentas para construção de gráficos;
- A biblioteca básica do R para construção de gráficos é a ***graphics***.

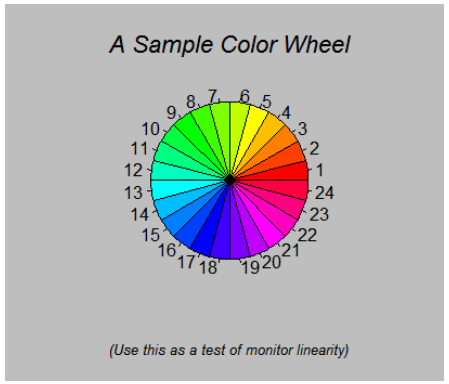


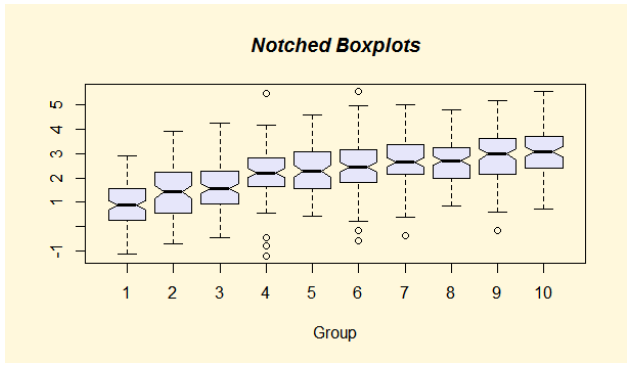
Plotando gráficos

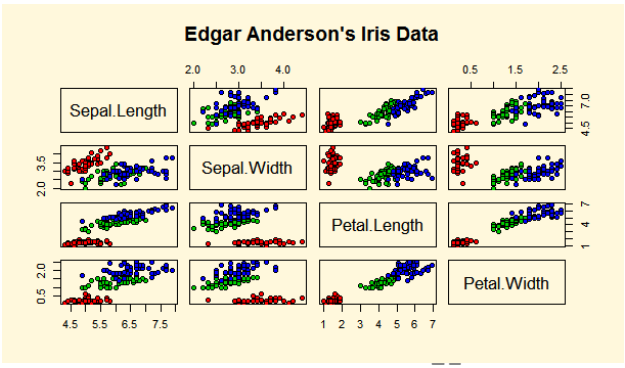
- Usando a função abaixo, vamos ver exemplos de gráficos com o pacote *graphics* do R:

```
4    demo("graphics")
```









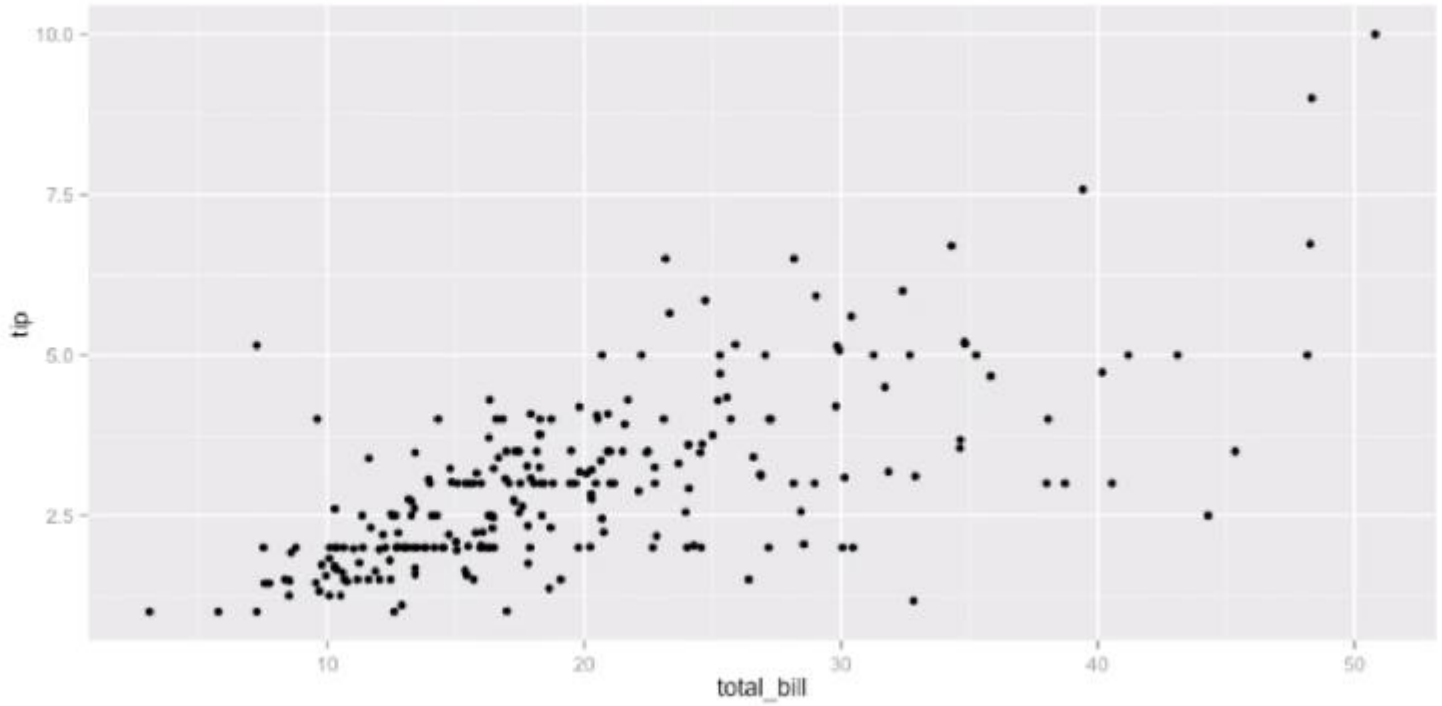
Plotando gráficos

- Gramática dos gráficos;
  - ✓ É usada para descrever as características que fundamenta a construção de gráficos;

Elemento	Descrição
Dados	Conjunto de dados a ser analisado
Estética	A escala em que nós mapeamos os dados
Geometria	Os elementos visuais usados para representar os dados
Facets	Visualizar os gráficos em porções menores
Estatística	Representação e análise dos dados
Coordenadas	A área na qual o gráfico será construído
Temas	Visão geral do gráfico

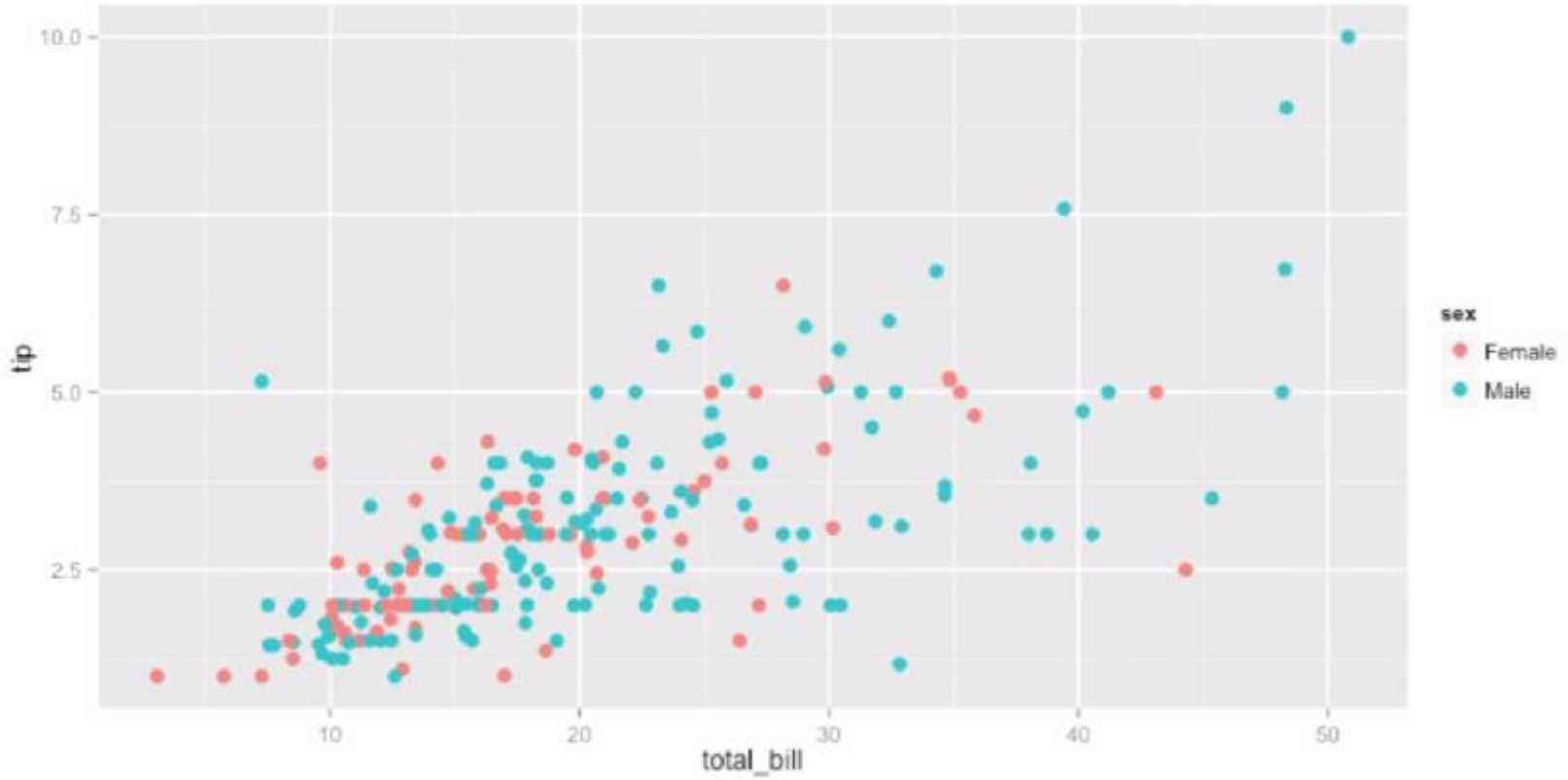
Plotando gráficos

Camada 1



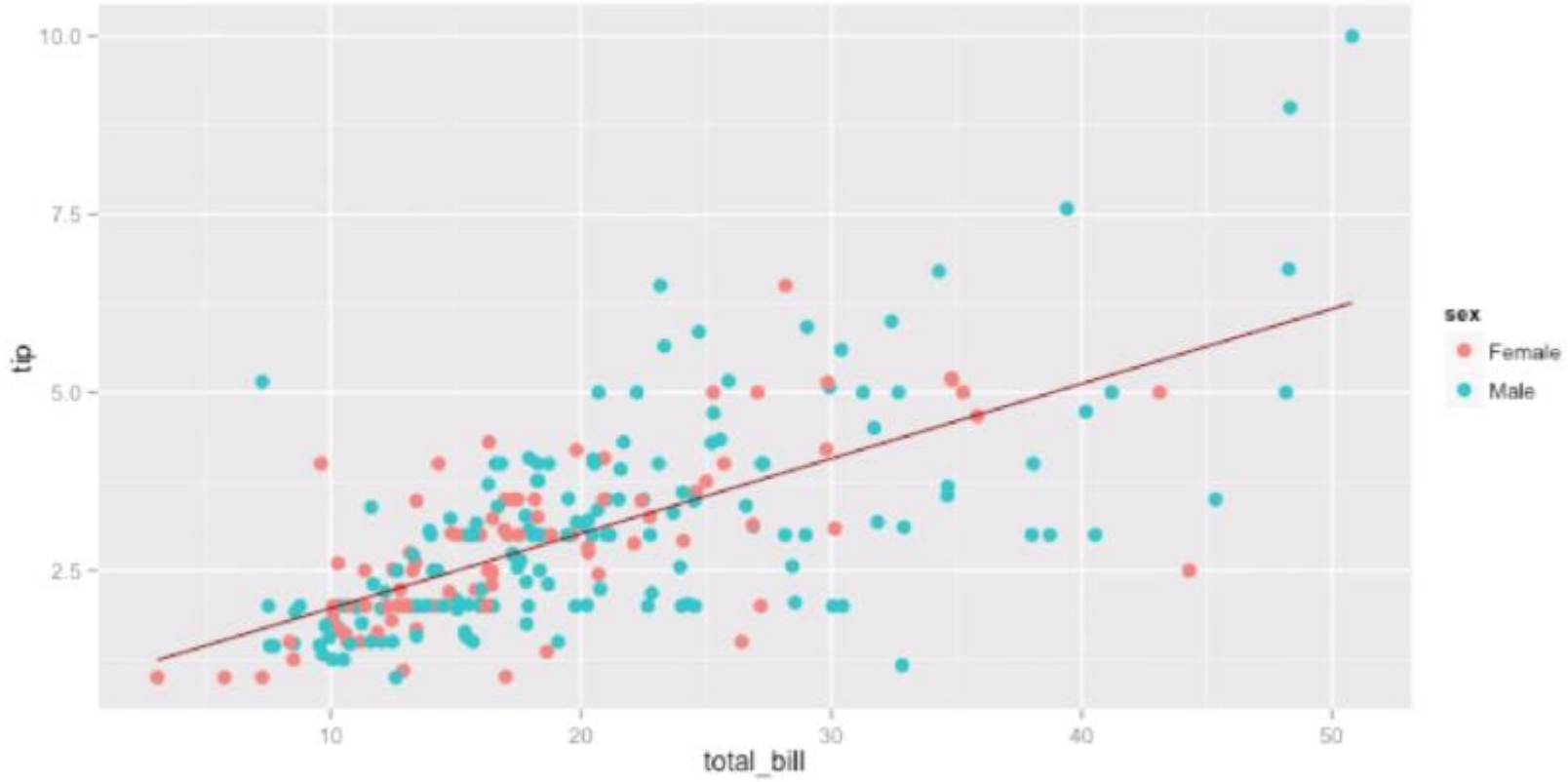
Plotando gráficos

Camada 2



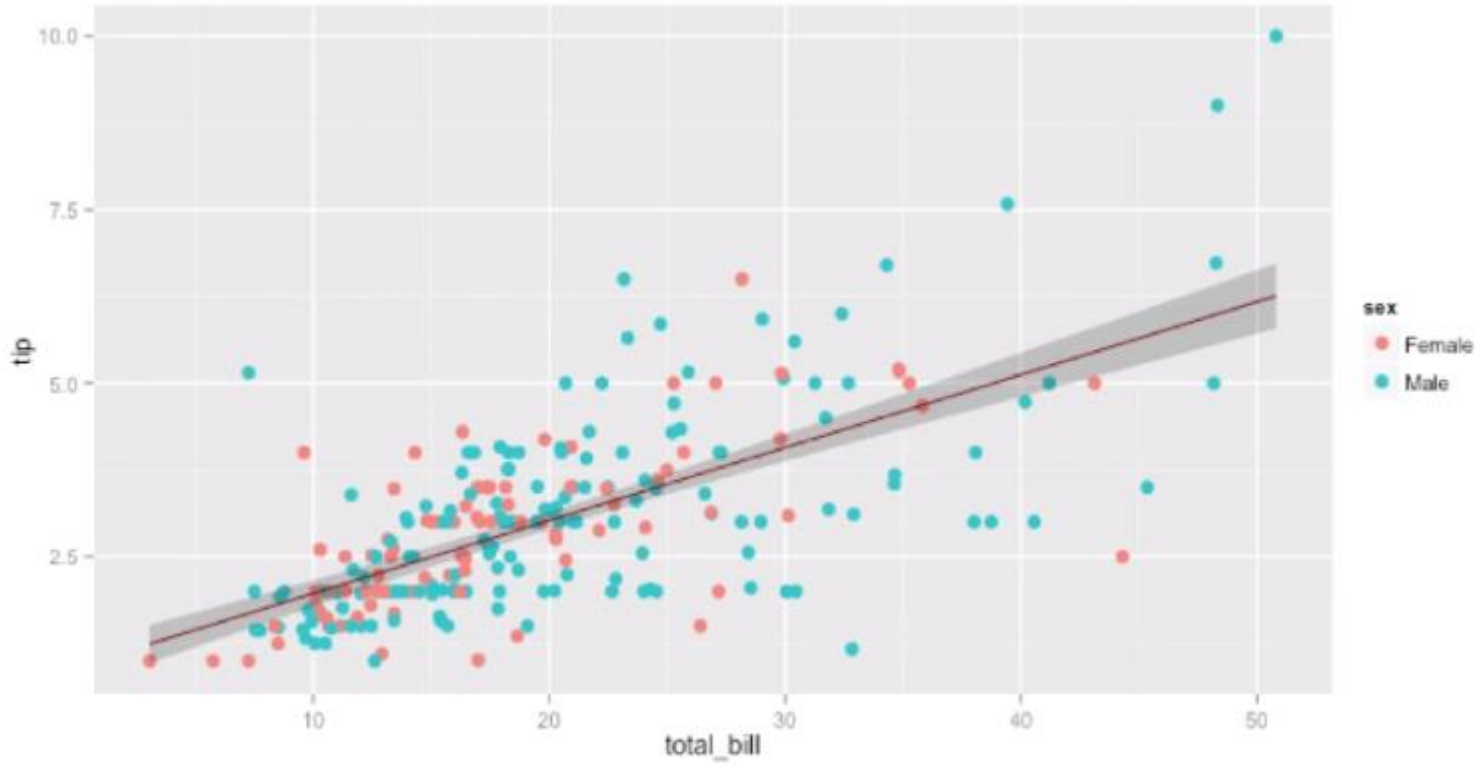
Plotando gráficos

Camada 3



Plotando gráficos

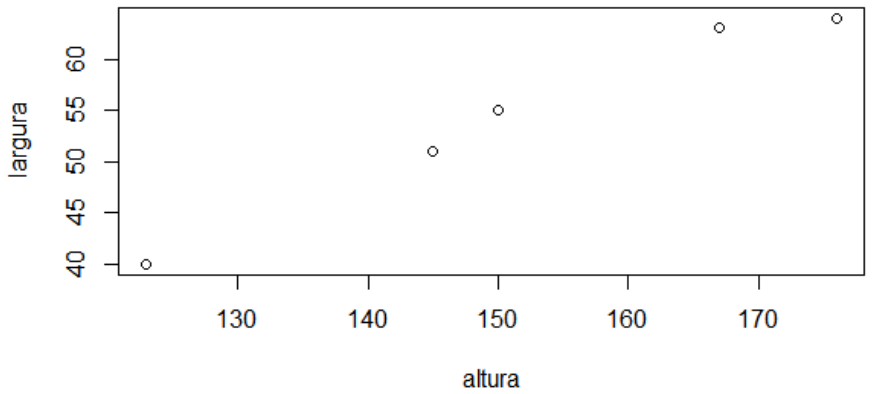
Camada 4



Plotando gráficos

- Plotando um gráfico básico

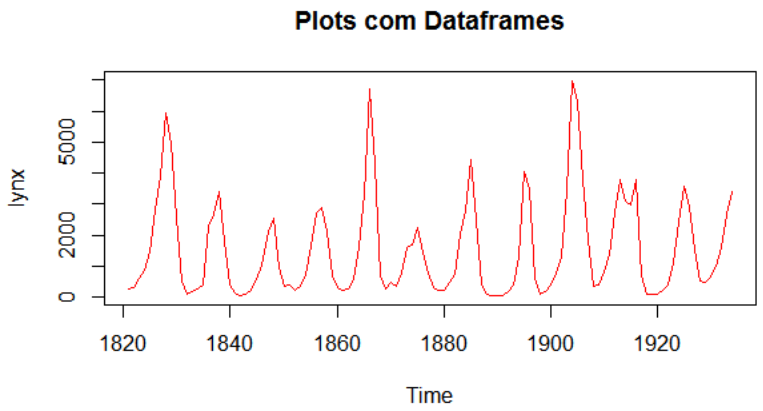
```
 9  # Plot Basico
10  altura <- c(145, 167, 176, 123, 150)
11  largura <- c(51, 63, 64, 40, 55)
12
13  plot(altura, largura)
```



Plotando gráficos

- Personalizando gráfico

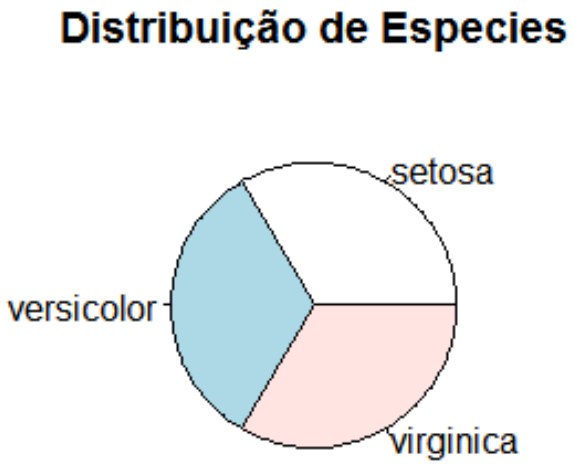
```
16  # Plotando um Dataframe
17  ?lynx
18  plot(lynx)
19  plot(lynx, ylab = "Plots com Dataframes", xlab = "")
20  plot(lynx, ylab = "Plots com Dataframes", xlab = "Observações")
21  plot(lynx, main = "Plots com Dataframes", col = 'red')
22
```



Plotando gráficos

- Plotando gráfico de setores ou pizza

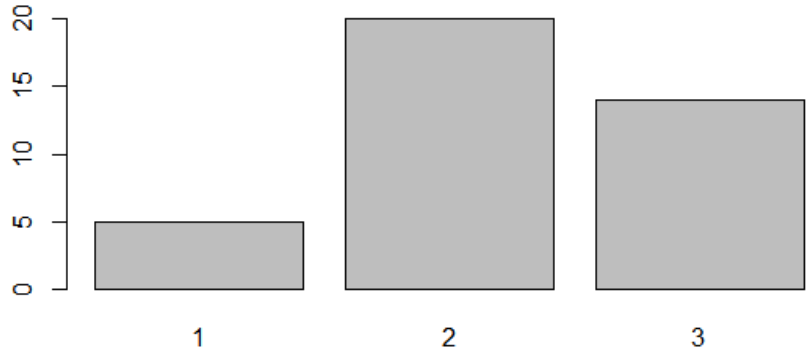
```
23  # plotando um gráfico de pizza
24  ?iris
25  attach(iris)
26  Values = table(Species)
27  labels = paste(names(Values))
28  pie(Values, labels = labels,
29  main = "Distribuição de Especies")
```



Plotando gráficos

- Plotando gráfico de barras

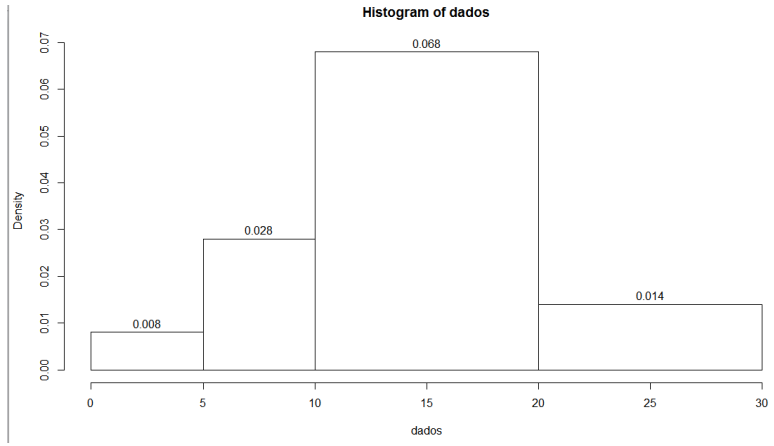
```
32  ?barplot
33
34  dados <-read.csv2("Base_de_dados/dados.csv", head=T)
35  sexo<-table(dados$Sexo)
36  barplot(sexo)
37
38  escolaridade<-table(dados$Escolaridade)
39  barplot(escolaridade)
```



Plotando gráficos

- Plotando gráfico do tipo histograma

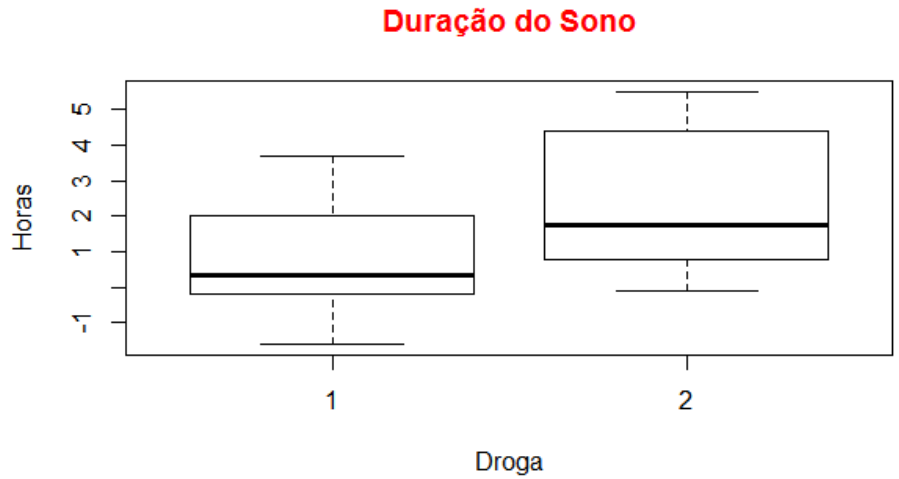
```
40  # plotando um gráfico histograma
41  dados = cars$speed
42  hist(dados)
43  hist(dados, breaks = 10)
44  hist(dados, labels = T, breaks=c(0,5,10,20,30))
```



Plotando gráficos

- Plotando gráfico de boxplot

```
46  # plotando um boxplot
47  sleep <- data.frame(sleep)
48  attach(sleep)
49  sleepboxplot = boxplot(data = sleep,
50  extra ~ group,main = "Duração do Sono",
51  col.main = "red", ylab="Horas", xlab="Droga")
```



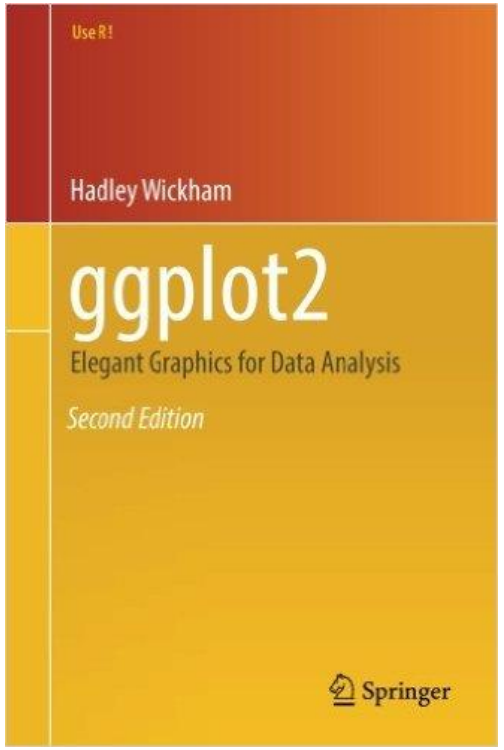
Plotando gráficos

- Conhecendo a pacote ***ggplot2***

  ✓É um das principais bibliotecas para construção de gráficos em R;

  ✓A documentação completa encontra-se em: http://ggplot2.org/;

  ✓Através da documentação é possível entender diversos exemplos.



Plotando gráficos

ggplot2 2.1.0 Index

Help topics

Geoms

Geoms, short for geometric objects, describe the type of plot you will produce.

- `geom_abline` (geom_hline, geom_vline)
  Lines: horizontal, vertical, and specified by slope and intercept.
- `geom_bar` (stat_count)
  Bars, rectangles with bases on x-axis
- `geom_bin2d` (stat_bin2d, stat_bin_2d)
  Add heatmap of 2d bin counts.
- `geom_blank`
  Blank, draws nothing.
- `geom_boxplot` (stat_boxplot)
  Box and whiskers plot.
- `geom_contour` (stat_contour)
  Display contours of a 3d surface in 2d.
- `geom_count` (stat_sum)
  Count the number of observations at each location.

Vignettes

- Extending ggplot2
- Aesthetic specifications

Dependencies

- **Imports**: digest, grid, gtable, MASS, plyr, reshape2, scales, stats
- **Suggests**: covr, ggplot2movies, hexbin, Hmisc, lattice, mapproj, maps, maptools, mgcv, multcomp, nlme, testthat, quantreg, knitr, rpart, rmarkdown, svglite

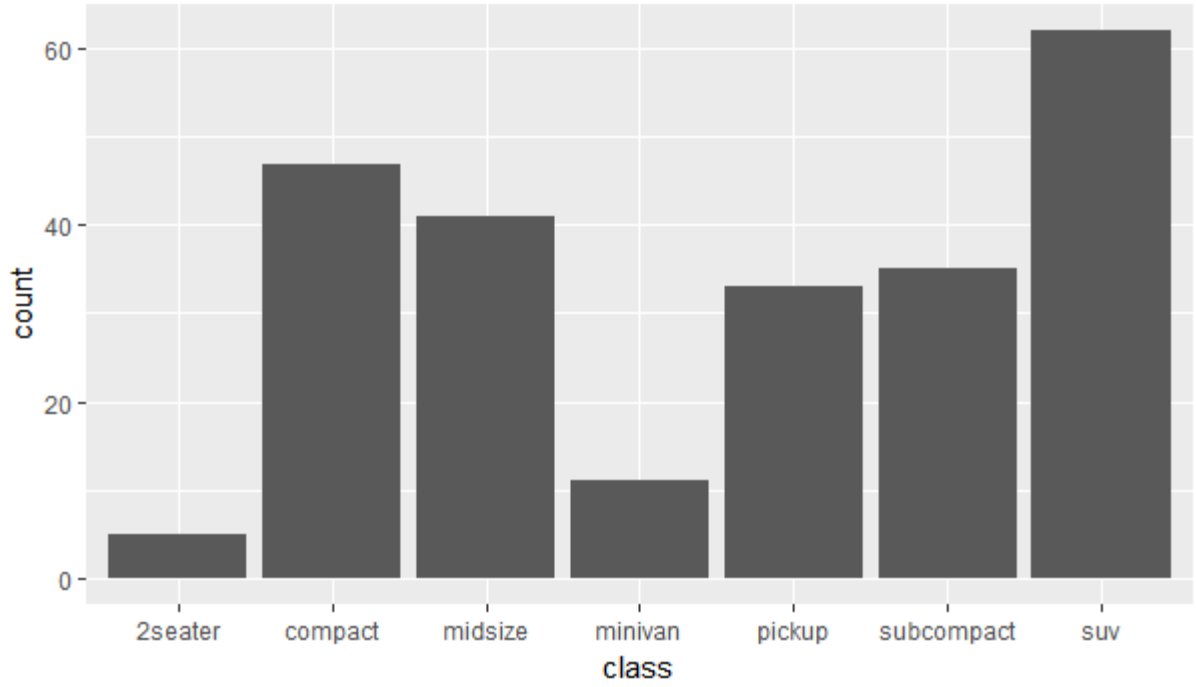
Authors

- Hadley Wickham [aut, cre]
- Winston Chang [aut]
- RStudio [cph]

Plotando gráficos

- Um exemplo utilizando a biblioteca ***ggplot2***

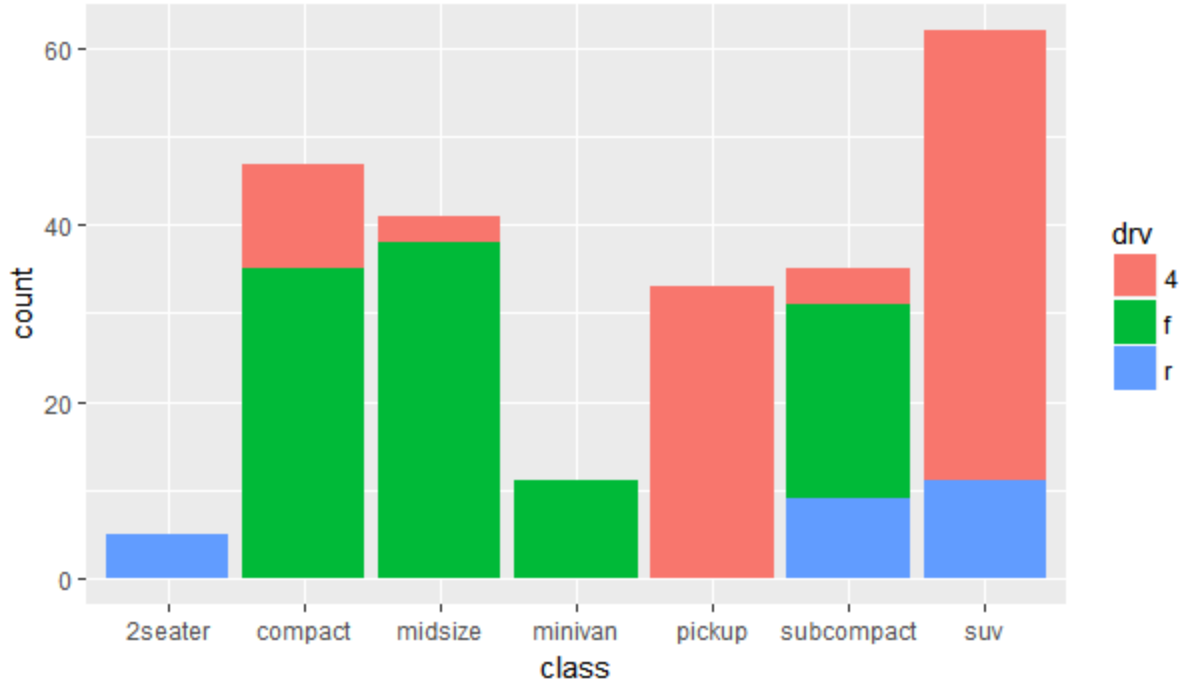
```
53  # Utilizando a biblioteca ggplot2
54
55  library(ggplot2)
56
57  # primeira versão do gráfico
58  mpg<-data.frame(mpg)
59  g <- ggplot(mpg, aes(class))
60  g + geom_bar()
```



Plotando gráficos

- Um exemplo utilizando a biblioteca ***ggplot2***

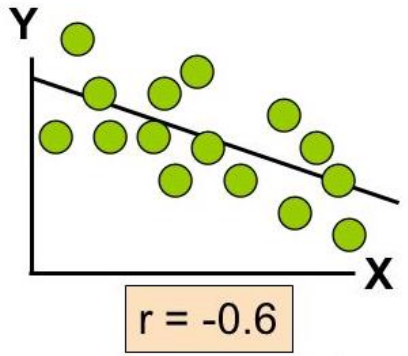
```
62  # segunda versão do gráfico
63  g <- ggplot(mpg, aes(class))
64  g + geom_bar(aes(fill = drv))
```



Correlação em R

Correlação em R

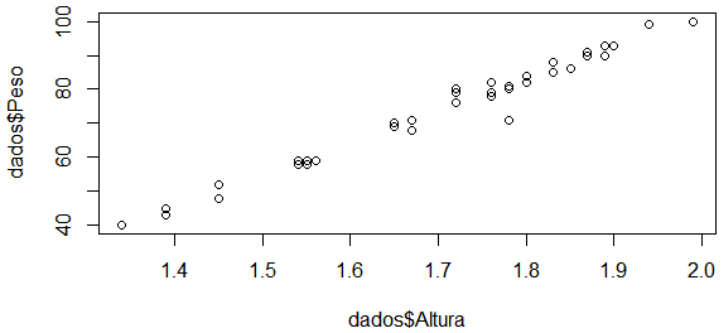
...recapitulando o conceito de Correlação !



Correlação em R

- Calculando correlações

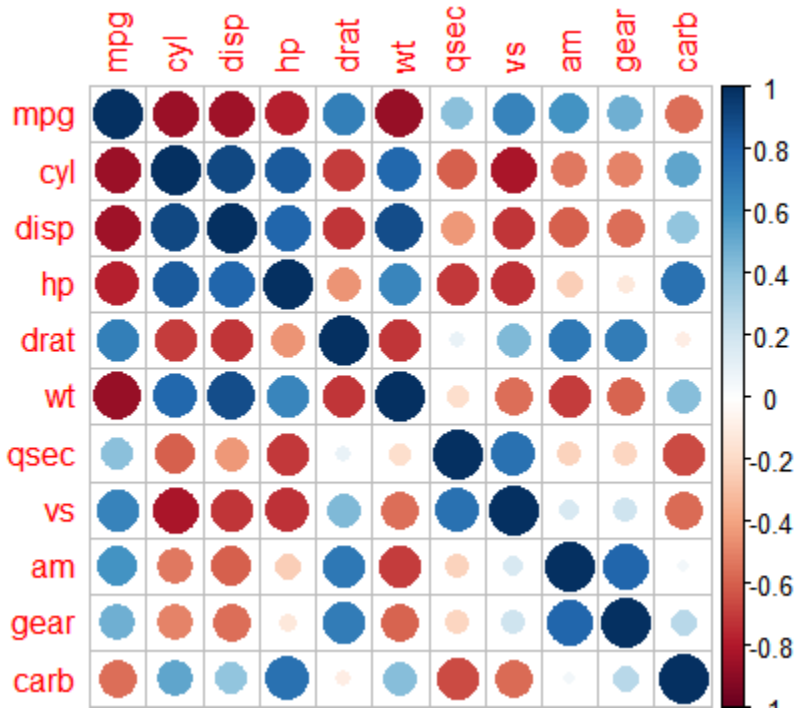
```
 9  # Importando base de dados
10
11  dados <- read.csv2("Base_de_dados/dados.csv", head=T)
12  dados
13
14  # Verificando valores de correlação de pearson
15
16  cor(dados$Idade, dados$Peso)
17  plot(dados$Idade, dados$Peso)
18
19  cor(dados$Altura, dados$Peso)
20  plot(dados$Altura, dados$Peso)
```



Correlação em R

- Plotando gráfico de correlações

```
4  # Gráfico de correlação
5
6  library(corrplot)
7  M <- cor(mtcars)
8  corrplot(M, method="circle")
```



Dúvidas

- **Contatos:**
  - ✓Email: rodrigo.linsrodrigues@ufrpe.br
  - ✓Facebook: /rodrigomuribec