

MAE0399 – Análise de Dados e Simulação: Introdução ao R para análise exploratória de dados

Prof^a: Márcia D'Elia Branco

Monitor PAE: Rafael Oliveira Silva

IME-USP

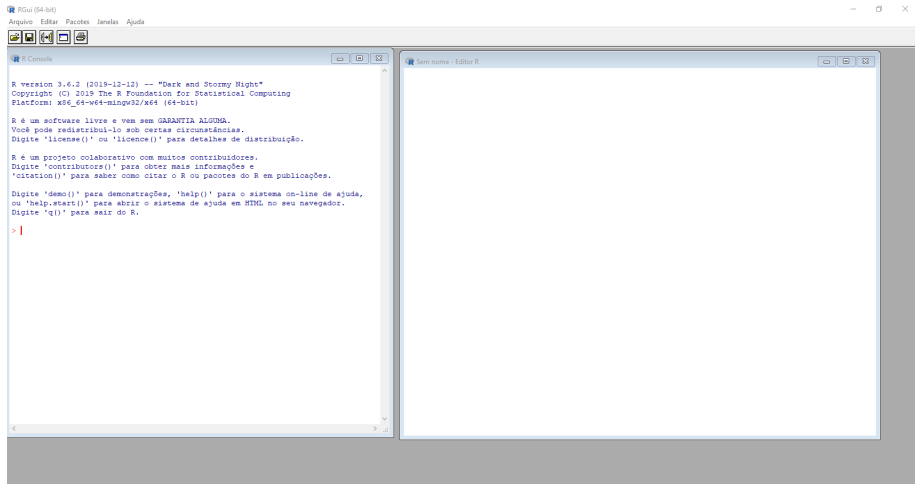
Apresentando o R

- **O que é o R ?** O R é uma linguagem e ambiente voltados para estatística computacional e gráficos.
- O R foi desenvolvido por Ross Ihaka e Robert Gentleman em 1990.

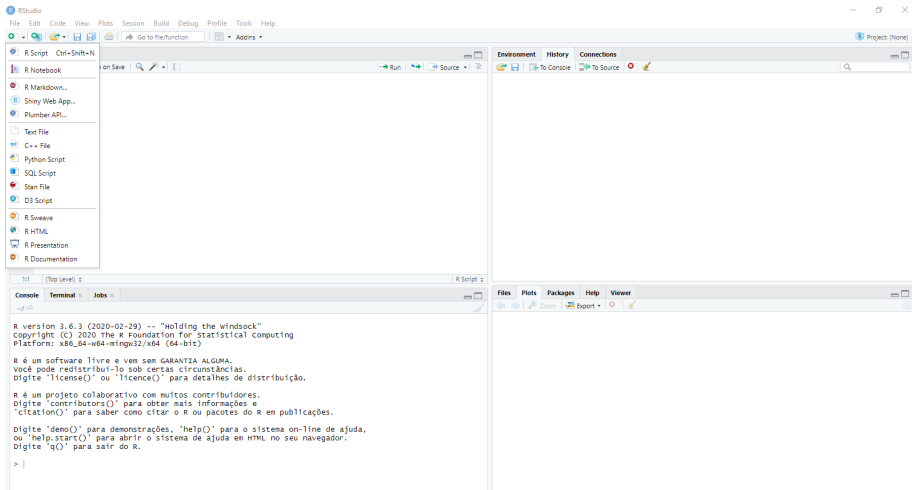


- Aberto e gratuito, é compatível com Windows, Linux e Mac.
- Atualmente o R está na versão **3.6.3**.
- RStudio é um Ambiente de Desenvolvimento Integrado (IDE - Integrated Development Environment) para criar e rodar o código R.
- Onde podemos baixar o R e o RStudio ?
 - <https://cran.r-project.org/>
 - <https://rstudio.com/products/rstudio/download/>

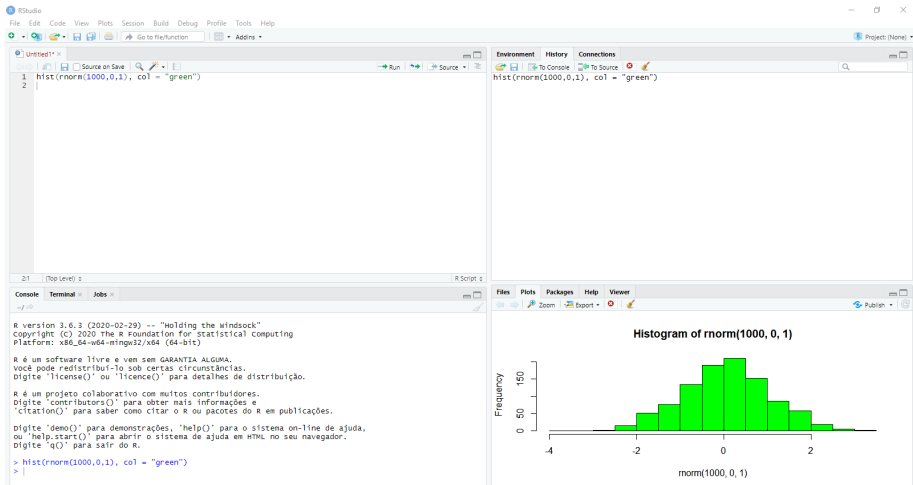
Apresentando o R



Apresentando o RStudio



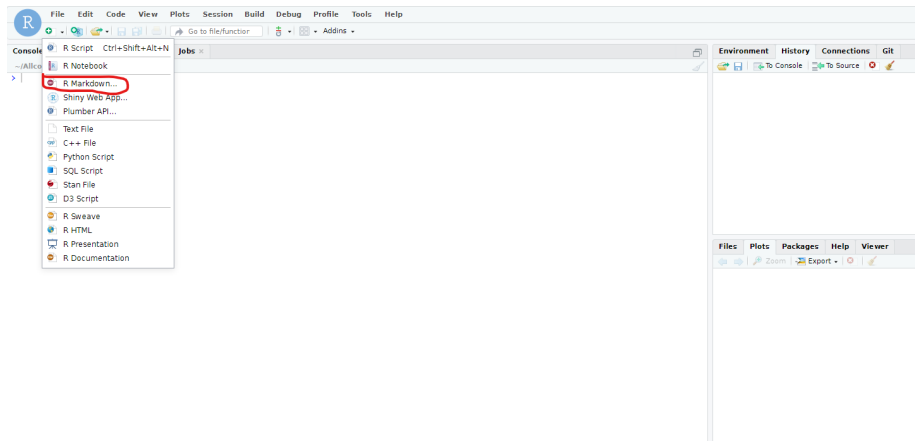
Apresentando o RStudio



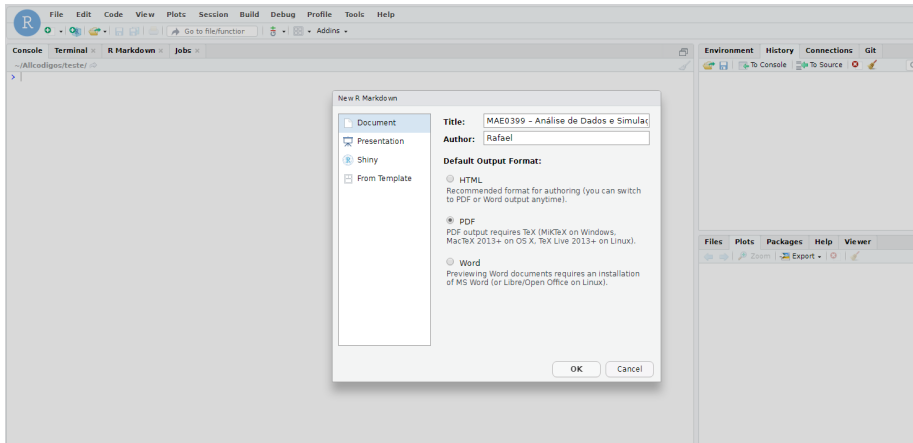
The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains the R script `hist(rnorm(1000,0,1), col = "green")`.
- Environment/History:** Shows the executed command `hist(rnorm(1000,0,1), col = "green")`.
- Console:** Displays the R version (3.6.3), copyright notice, platform information, and a series of help messages for various functions like `license()`, `contributors()`, `demon()`, `help()`, `help.start()`, and `q()`. The final command executed is `> hist(rnorm(1000,0,1), col = "green")`.
- Plots:** A histogram titled "Histogram of rnorm(1000, 0, 1)" is shown. The x-axis is labeled `rnorm(1000, 0, 1)` and ranges from -4 to 2. The y-axis is labeled "Frequency" and ranges from 0 to 150. The histogram bars are green.

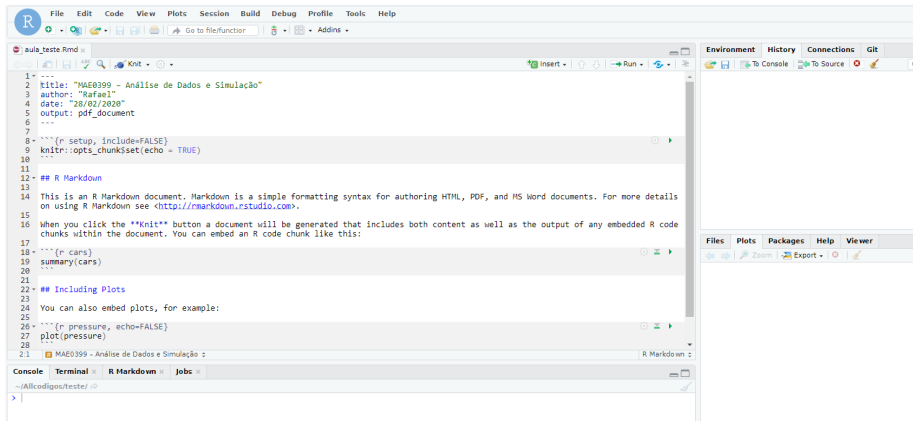
Apresentando o R



Apresentando o RStudio



Apresentando o RStudio



Apresentando o RStudio

The screenshot displays the RStudio environment with a file named `sula_teste.Rmd` open in the editor. The document is an R Markdown file titled "MAE0399 – Análise de Dados e Simulação" by Rafael, dated 28/02/2020, set to output as a PDF document. The code includes an R setup chunk, an R Markdown section explaining the format, an R code chunk using `summary(cars)`, and an R plot chunk using `plot(pressure)`.

The right pane shows the rendered PDF output. The title is "MAE0399 – Análise de Dados e Simulação" by Rafael, dated 28/02/2020. The content includes an R Markdown section, a summary of the `cars` dataset, and a scatter plot of pressure versus speed.

Summary of cars dataset:

	speed	dist
## Min.	4.0	2.00
## 1st Q.	12.0	26.00
## Median	15.0	36.00
## Mean	19.6	42.98
## 3rd Q.	19.0	56.00
## Max.	25.0	120.00

Scatter plot of pressure vs speed:

The plot shows a clear upward trend, with pressure values starting near zero for low speeds and reaching approximately 800 for speeds above 300.

Operações Básicas no R

#Adição

2+2

[1] 4

#Subtração

2-2

[1] 0

#Multiplicação

2*2

[1] 4

#Divisão

2/2

[1] 1

#Potenciação

2^2

[1] 4

Funções Básicas no R

```
#Raiz quadrada  
sqrt(4)
```

```
## [1] 2
```

```
#Logaritmo na base 10  
log(2,10)
```

```
## [1] 0.30103
```

```
#Logaritmo na base e  
log(2)
```

```
## [1] 0.6931472
```

```
#Exponencial  
exp(2)
```

```
## [1] 7.389056
```

```
#Fatorial  
factorial(2)
```

```
## [1] 2
```

Funções Básicas no R

```
#Valor absoluto  
abs(-2)
```

```
## [1] 2
```

```
#Arredondando  
round(3.141516,2)
```

```
## [1] 3.14
```

```
#Função Piso  
floor(3.141516)
```

```
## [1] 3
```

```
#Função Teto  
ceiling(3.141516)
```

```
## [1] 4
```

Objetos do R: vetor

• Vetor

```
x <- c(3,1,5,4)
#Selecionando o segundo valor do vetor
x[2]
```

```
## [1] 1
#Apagando um elemento do vetor
x[-2]
```

```
## [1] 3 5 4
#Sequência
y <- 1:10
y
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
z <- seq(1,10,length = 10)
z
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
```

Vetor

```
#Repetindo o valor 6 quatro vezes  
rep(6,4)
```

```
## [1] 6 6 6 6
```

```
#Repetindo o vetor quatro vezes  
rep(c(1,2,3),4)
```

```
## [1] 1 2 3 1 2 3 1 2 3 1 2 3
```

```
#Repetindo cada elemento do vetor quatro vezes  
rep(c(1,2,3),each = 4)
```

```
## [1] 1 1 1 1 2 2 2 2 3 3 3 3
```

```
#Soma dos elementos de um vetor  
sum(c(1,2,3))
```

```
## [1] 6
```

```
#Produto dos elementos de um vetor  
prod(c(1,2,3))
```

```
## [1] 6
```

Vetor

```
#Comprimento de um vetor  
length(c(1,2,3))
```

```
## [1] 3
```

```
#Ordenando os elementos de um vetor de forma crescente  
sort(c(4,1,6),decreasing = FALSE)
```

```
## [1] 1 4 6
```

```
#Esta função retorna a posição dos elementos do vetor  
#ordenados conforme os valores do vetor  
order(c(4,1,6),decreasing = FALSE)
```

```
## [1] 2 1 3
```

```
#Esta função retorna o máximo  
max(c(1,1,2,3))
```

```
## [1] 3
```

```
##Esta função retorna o mínimo  
min(c(0,1,2,3))
```

```
## [1] 0
```

Matriz

```
M <- matrix(c(1,2,3,4),ncol = 2,nrow = 2)
```

```
M
```

```
##      [,1] [,2]
```

```
## [1,]    1    3
```

```
## [2,]    2    4
```

```
#Selecionando os valores de uma matriz
```

```
M[1,1]
```

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

```
M[,1]
```

```
## [1] 1 2
```

```
M[2,]
```

```
## [1] 2 4
```


Matriz

```
M <- matrix(0,ncol = 2,nrow = 2)
```

```
M
```

```
##      [,1] [,2]
```

```
## [1,]    0    0
```

```
## [2,]    0    0
```

```
# Inserindo os valores em uma matriz
```

```
M[1,1] <- 1
```

```
M[2,] <- c(2,2)
```

```
M
```

```
##      [,1] [,2]
```

```
## [1,]    1    0
```

```
## [2,]    2    2
```

Matriz

```
#Matriz diagonal  
M <- diag(c(1,2))  
M
```

```
##      [,1] [,2]  
## [1,]    1    0  
## [2,]    0    2
```

```
#Multiplicando uma matriz por outra matriz  
M%*%M
```

```
##      [,1] [,2]  
## [1,]    1    0  
## [2,]    0    4
```

```
#Multiplicando um matriz por um escalar  
2*M
```

```
##      [,1] [,2]  
## [1,]    2    0  
## [2,]    0    4
```

Matriz

```
D <- matrix(c(1,2,3,4),ncol = 2,nrow = 2)
D
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

```
#Matriz transposta
t(D)
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

```
#Matriz inversa
solve(D)
```

```
##      [,1] [,2]
## [1,]   -2  1.5
## [2,]    1 -0.5
```

Funções

Criando uma função:

nome da função <- function(argumentos){ comandos da função }

Exemplo: Considere uma função que retorna a média de um vetor.

```
z <- c(1,3,2,5,3,6)
```

```
zbar <- function(x){ sum(x)/length(x)}  
zbar(z)
```

```
## [1] 3.333333
```

```
mean(z)
```

```
## [1] 3.333333
```

Funções do R

Operadores de comparação

- Igualdade: `==`
- Diferente: `!=`
- Menor: `<`
- Maior: `>`
- Menor ou igual: `<=`
- Maior ou igual: `>=`

Operadores lógicos

- E: `&`
- Ou: `||`

Condicionais

`if (Expressão teste) { declaração } else declaração`

```
x <- 3
if(x >= 1 & x <= 6){print("Sim")} else {print("Não")}
```

```
## [1] "Sim"
```

Funções do R - Ciclos

While

```
x <- 3  
while(x < 5){ x = x + 1 ; print(x)}
```

```
## [1] 4
```

```
## [1] 5
```

For

```
x <- c()  
for(i in 0:4){x[i+1] <- 2*i+1}
```

```
x
```

```
## [1] 1 3 5 7 9
```

Pacotes no R

Como instalar um pacote no R ?

```
install.packages("nome do pacote")
```

Como carregar um pacote ?

```
require("nome do pacote") ou library("nome do pacote")
```

Função Sample()

```
x <- c("Cara", "Coroa")  
#lançando uma moeda  
resultados <- sample(x,5,replace = T)  
  
resultados  
  
## [1] "Cara" "Cara" "Cara" "Coroa" "Cara"  
table(resultados)
```

```
## resultados  
## Cara Coroa  
##      4      1
```

```
#Função replicate()  
replicate(3,sample(x,5,replace = T))
```

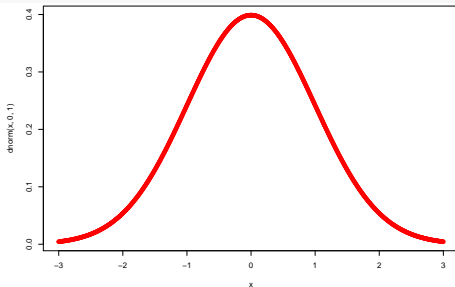
```
##      [,1] [,2] [,3]  
## [1,] "Cara" "Coroa" "Coroa"  
## [2,] "Coroa" "Cara" "Cara"  
## [3,] "Cara" "Coroa" "Coroa"  
## [4,] "Cara" "Cara" "Cara"  
## [5,] "Cara" "Coroa" "Coroa"
```


Distribuições de Probabilidade no R

Fazendo $Z \sim N(0,1)$, teremos que:

- Densidade $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

```
x <- seq(-3,3,length = 1000)
plot(x,dnorm(x,0,1), col = "red")
```



- $P(Z < 0) = 0.5$

```
pnorm(0,0,1)
```

```
## [1] 0.5
```

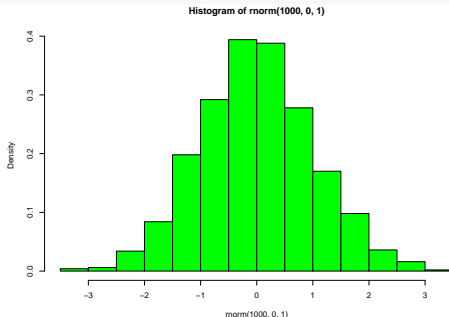
- Mediana

```
qnorm(0.5,0,1)
```

```
## [1] 0
```

- gerando valores aleatórios

```
hist(rnorm(1000,0,1),
     freq = F, col = "green")
```



Distribuições de Probabilidade no R

- `p` para probabilidade acumulada;
- `q` para quantil;
- `d` para densidade;
- `r` gerar uma amostra de uma distribuição.

Simulando variáveis aleatórias no R

- `rnorm(n, mu, sigma)` gerando n valores da distribuição Normal(μ, σ);
- `runif(n, a, b)` gerando n valores da distribuição Uniforme(a, b);
- `rexp(n, lambda)` gerando n valores da distribuição Exponencial(λ);
- `rpois(n, lambda)` gerando n valores da distribuição Poisson(λ);
- `rbinom(n, K, p)` gerando n valores da distribuição Binomial(K, p);
- etc.

Análise descritiva no R

- Iremos trabalhar com dados simulados:

```
n <- 100
x <- seq(-2,10,len = n)
y <- 10 + 2*x + rnorm(n,0,2)
z <- sample(c("s","n"), n, replace = T, prob = c(0.5,0.5))
```

```
Dados <- data.frame(x=x,y=y,z=z)
```

```
str(Dados)
```

```
## 'data.frame':    100 obs. of  3 variables:
## $ x: num  -2 -1.88 -1.76 -1.64 -1.52 ...
## $ y: num  4.96 9.14 11.05 8.35 5.52 ...
## $ z: Factor w/ 2 levels "n","s": 1 1 1 2 2 1 2 2 1 1 ...
```

```
Dados[1:3,]
```

```
##           x           y z
## 1 -2.000000  4.964303 n
## 2 -1.878788  9.141300 n
## 3 -1.757576 11.046834 n
```

Análise descritiva no R

```
summary(Dados$y)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.964  11.621  18.427  17.966  23.592  31.639
```

```
sd(Dados$y)
```

```
## [1] 7.173522
```

```
var(Dados$y)
```

```
## [1] 51.45941
```

```
table(Dados$z)
```

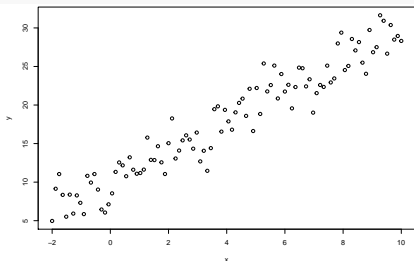
```
##
```

```
##  n  s
```

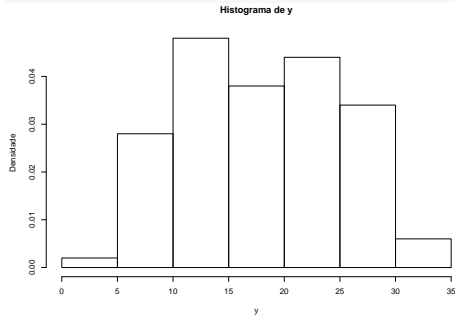
```
## 52 48
```

Análise descritiva no R

```
plot(Dados$x,Dados$y,  
xlab = "x",ylab = "y")
```



```
hist(Dados$y, freq = F,  
main = "Histograma de y",  
xlab = "y",ylab = "Densidade")
```

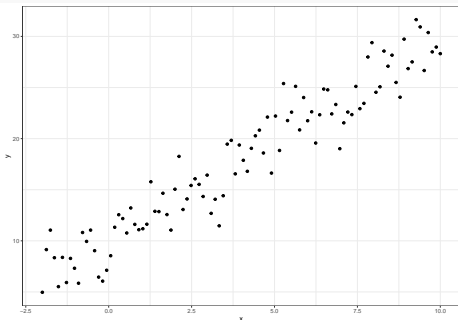


Análise descritiva no R

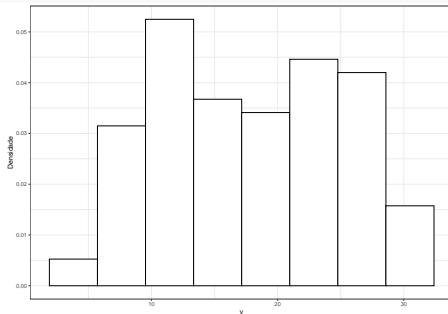
```
require(ggplot2)

## Loading required package: ggplot2

ggplot(Dados, aes(x=x,y=y)) +
  theme_bw() +
  geom_point()
```

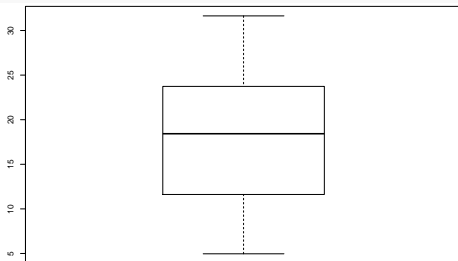


```
ggplot(Dados, aes(x=y)) +
  theme_bw() +
  geom_histogram(aes(y=..density..),
    colour="black", fill="white", bins = 8) +
  xlab("y") +
  ylab("Densidade")
```

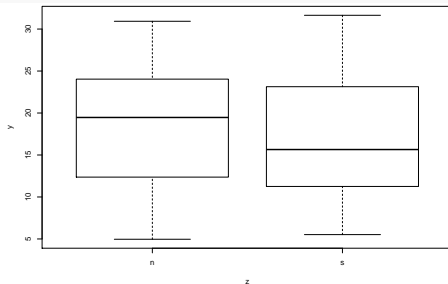


Análise descritiva no R

```
boxplot(Dados$y)
```

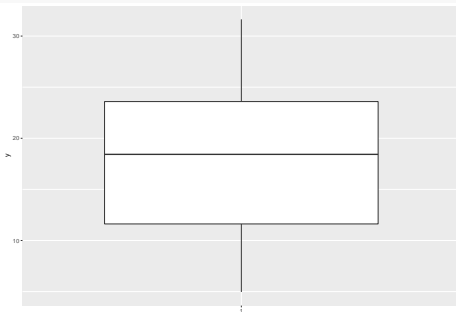


```
boxplot(y~z, data=Dados)
```

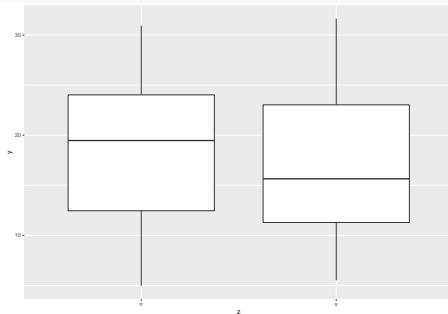


Análise descritiva no R

```
ggplot(Dados) +  
  geom_boxplot(aes(x = factor(1), y = y)) +  
  xlab(" ")
```



```
ggplot(Dados, aes(x=z, y=y)) +  
  geom_boxplot()
```

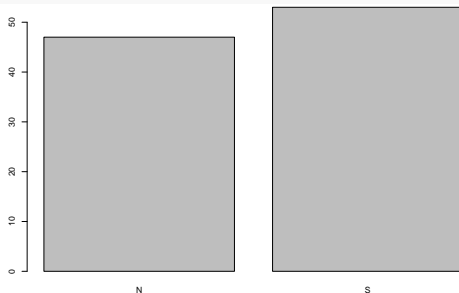


Análise descritiva no R

```
dados <- data.frame(x = sample(c("S","N"),  
  100,replace = T),y = sample(  
  c("baixo","Médio","Alto"),  
  100,replace = T))
```

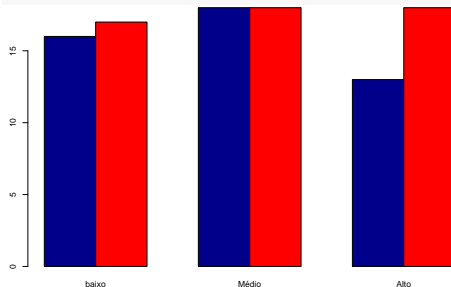
```
Tabela1 <- table(dados$x)
```

```
barplot(Tabela1)
```



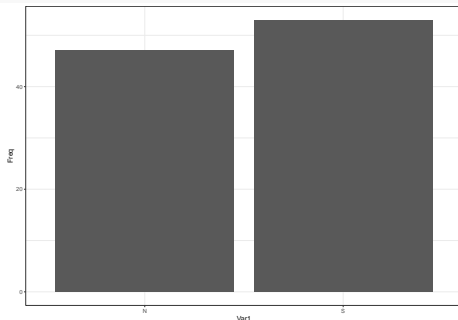
```
Tabela2 <- table(dados)
```

```
barplot(Tabela2,col=c("darkblue","red"),  
  names.arg = c("baixo","Médio","Alto"),  
  beside=TRUE)
```

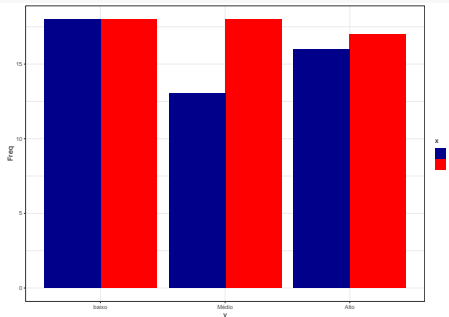


Análise descritiva no R

```
ggplot(data = data.frame(Tabela1),  
  aes(x=Var1, y=Freq)) +  
  theme_bw() +  
  geom_bar(stat = "identity")
```



```
Dado1 <- data.frame(Tabela2)  
Dado1$y <- factor(Dado1$y, levels =  
  c("baixo", "Médio", "Alto"))  
ggplot(data = Dado1 ,  
  aes(x=y, y=Freq, fill=x)) +  
  theme_bw() +  
  scale_fill_manual(values=  
  c("darkblue", "red")) +  
  geom_bar(stat = "identity",  
  position=position_dodge())
```



Ajustando um modelo de regressão no R

A sintaxe básica para ajustar um modelo de regressão linear no R é **lm(y~x,data)**, em que y é a variável resposta e a variável x é a variável explicativa.

Assim, considerando

```
n <- 40
x <- seq(1,10, len = n)
y <- 2 + 0.5*x + rnorm(n,0,1)

Dados <- data.frame(x=x,y=y)

Ajuste <- lm(y ~ x,data = Dados )
```

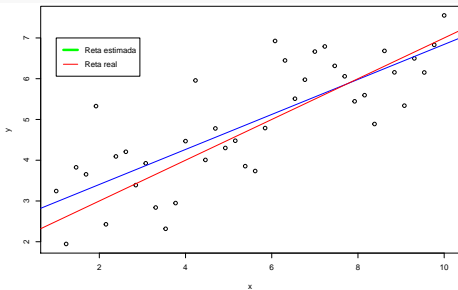
Ajustando um modelo de regressão no R

```
summary(Ajuste)
```

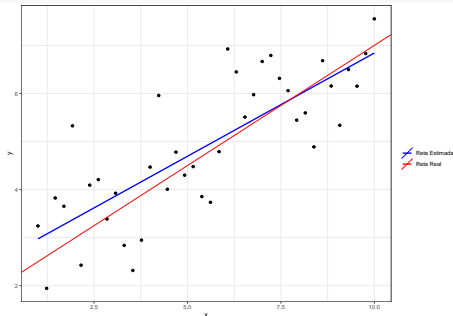
```
##
## Call:
## lm(formula = y ~ x, data = Dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74821 -0.49410  0.07231  0.52675  1.95566
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.54561     0.32746   7.774 2.26e-09 ***
## x             0.42962     0.05358   8.018 1.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9028 on 38 degrees of freedom
## Multiple R-squared:  0.6285, Adjusted R-squared:  0.6187
## F-statistic: 64.28 on 1 and 38 DF,  p-value: 1.08e-09
```

Ajustando um modelo de regressão no R

```
plot(Dados$x ,Dados$y,  
xlab = "x", ylab = "y")  
abline(Ajuste,col = "blue")  
abline(c(2,0.5),col = "red")  
legend(1,7,c("Reta estimada","Reta real"),  
lwd=c(5,2), col=c("green","red"),  
y.intersp=1.5)
```



```
ggplot(Dados,aes(x=x,y=y))+  
theme_bw()+ geom_point() +  
geom_smooth(method = "lm", se = FALSE,  
aes(colour="Reta Estimada") )+  
geom_abline(aes(slope = 0.5,  
intercept = 2, colour = "Reta Real")) +  
scale_colour_manual(name="",  
values=c("blue", "red"))
```



Ajustando um modelo de regressão no R

```
n <- 50; x <- seq(1,10, len = n) ; z <- seq(20,40, len = n)
y <- 0.5*x + 2*z + rnorm(n,0,2)
```

```
Dados2 <- data.frame(x=x,y=y,z=z)
```

```
Ajuste2 <- lm(y ~ x + z -1,data = Dados2 )
```

```
summary(Ajuste2)
```

```
##
## Call:
## lm(formula = y ~ x + z - 1, data = Dados2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.1159 -1.7737 -0.3742  1.1250  5.4839
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## x    0.5478      0.2133   2.568  0.0134 *
## z    1.9781      0.0426  46.433  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Referências

- COOKBOOK FOR R. Cookbook for R. Disponível em: <http://www.cookbook-r.com/>. CURSO R. Material das aulas do curso de R. Disponível em: <http://curso-r.github.io/>.
- KABACOFF, R.I. Quick-R: Descriptives. Disponível em: <http://www.statmethods.net/stats/descriptives.html>.
- KABACOFF, R.I. Quick-R: Frequencies. Disponível em: <http://www.statmethods.net/stats/frequencies.html>.
- KABACOFF, R.I. Quick-R: Pie Charts. Disponível em: <http://www.statmethods.net/graphs/pie.html>.
- IQSS-HARVARD. Introduction to R Graphics with ggplot2. Disponível em: < <http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html> >.
- STACK OVERFLOW. Frequent 'r' Questions – Stack Overflow. Disponível em: < <https://stackoverflow.com/questions/tagged/r> >.