

# Introdução ao software R

---

## OBJETIVOS

1. Entender a linguagem de programação do R;
2. Visão geral da estrutura do RStudio;
3. Funções básicas do R.
4. Leitura de uma base de dados no formato .csv
5. Realizar algumas análises, construir gráficos e tabelas

# Introdução ao software R

---

## Sobre o “R” e “RStudio”

- R é um ambiente de software livre para análise gráfica e estatística. É de código aberto e, portanto, disponível gratuitamente.
- O RStudio pode ser instalado no Windows, Mac e Linux.

# RStudio

Janela com a Sintaxe

Janela que ficarão os arquivos Gerados na sintaxe

Janela com os Resultados da Sintaxe (Console)

Janela com os pacotes, gráficos, ajuda, Arquivos da pasta,...

The screenshot shows the RStudio desktop environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The main editor window displays a script file named 'SCRIPT\_TCT.R' with R code for installing packages and setting the Java environment. The bottom-left panel is the Console, showing the R startup message and the current working directory. The bottom-right panel is the Environment pane, which is currently empty. The top-right panel is the Files pane, showing the project directory structure and a list of files including 'ENEM\_CH\_2009.csv', 'ENEM\_CH\_2009\_GABARITO.csv', 'SCRIPT\_TCT.R', and 'SCRIPT\_TCT\_FUNCTIONS.R'.

```
1 ##### Script para gerar resultado através da Análise Clássica dos Testes - TCT #####
2
3 # Java 64 bits atualizado e instalado na máquina
4
5 ##### Etapa 00 #####
6
7 ## Instalar os seguintes pacotes
8 # install.packages("data.table")
9 # install.packages("intoyou")
10 # install.packages("stats")
11 # install.packages("psych")
12 # install.packages("dplyr")
13 # install.packages("plotly")
14 # install.packages("webshot")
15 # webshot::install_phantomjs()
16 # install.packages("plyr")
17 # install.packages("devtools")
18 # devtools::install_github("jorgealmeida/tct")
19
20 # Sys.setenv(JAVA_HOME="C:\\Program Files\\Java\\jre1.8.0_144")
21 # jre1.8.0_144 é a última versão do java 64 bits
22 #
23
24 ##### Etapa 00 #####
```

Console

Platform: x86\_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

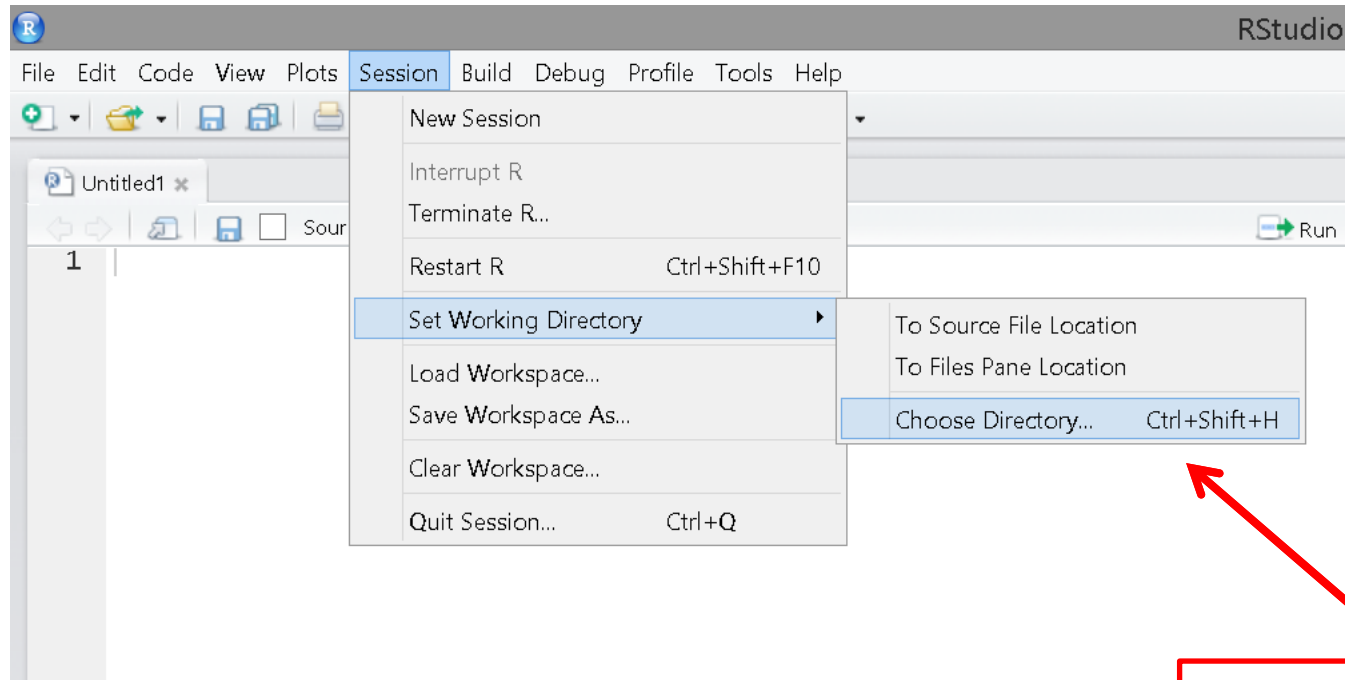
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> setwd("C:/Adriano/MGA/TRI1/2018-MGA-TRI1/SemanaI/Aula2/ENEM\_CH\_2009")  
> |

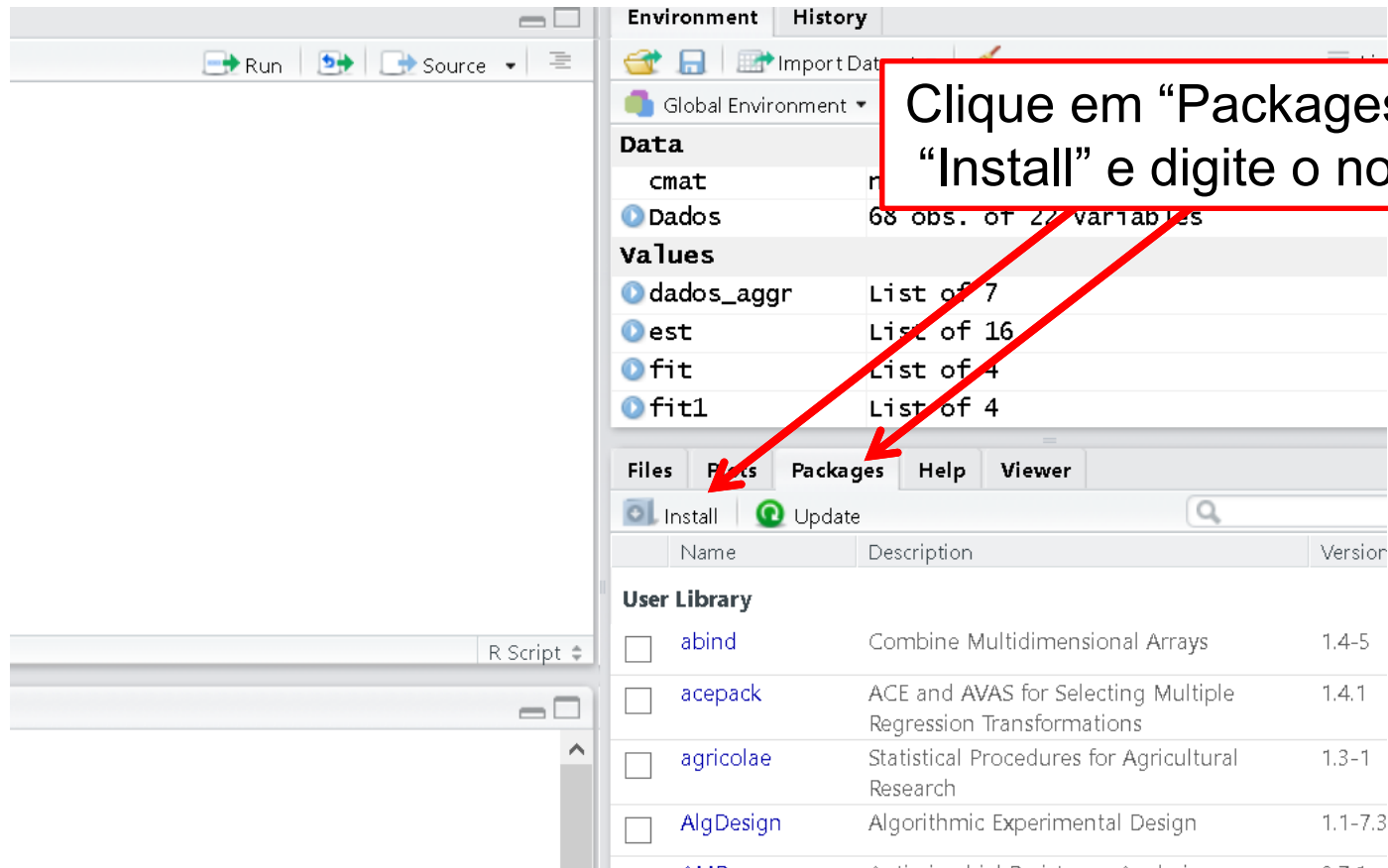
| Name                      | Size     | Modified               |
|---------------------------|----------|------------------------|
| ENEM_CH_2009.csv          | 936.6 KB | Jun 10, 2018, 10:04 PM |
| ENEM_CH_2009_GABARITO.csv | 264 B    | May 30, 2018, 3:16 PM  |
| SCRIPT_TCT.R              | 2.6 KB   | Jun 26, 2018, 9:32 PM  |
| SCRIPT_TCT_FUNCTIONS.R    | 34.2 KB  | Jun 20, 2018, 8:14 PM  |

# Pasta de Trabalho do RStudio



Clique aqui e selecione  
a pasta que estão  
localizados os arquivos

# Instalação de Pacotes

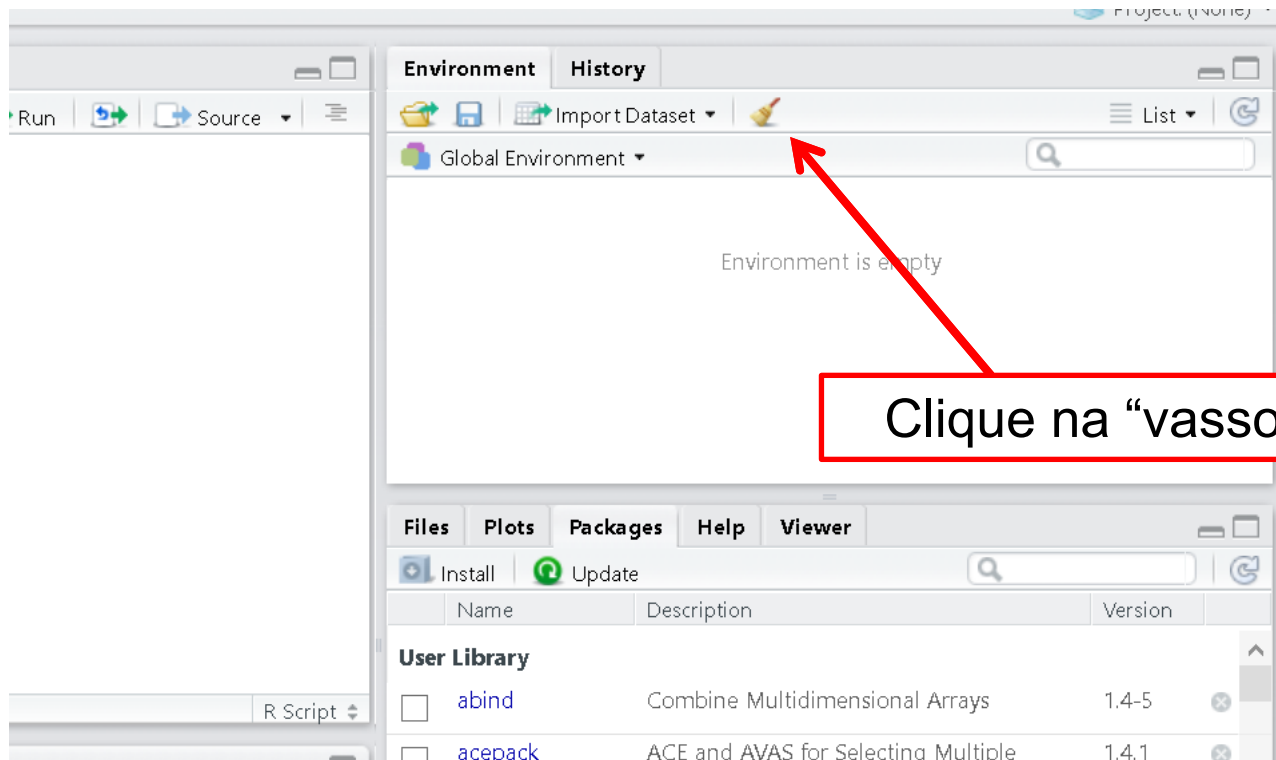


Clique em “Packages” e depois em “Install” e digite o nome do pacote

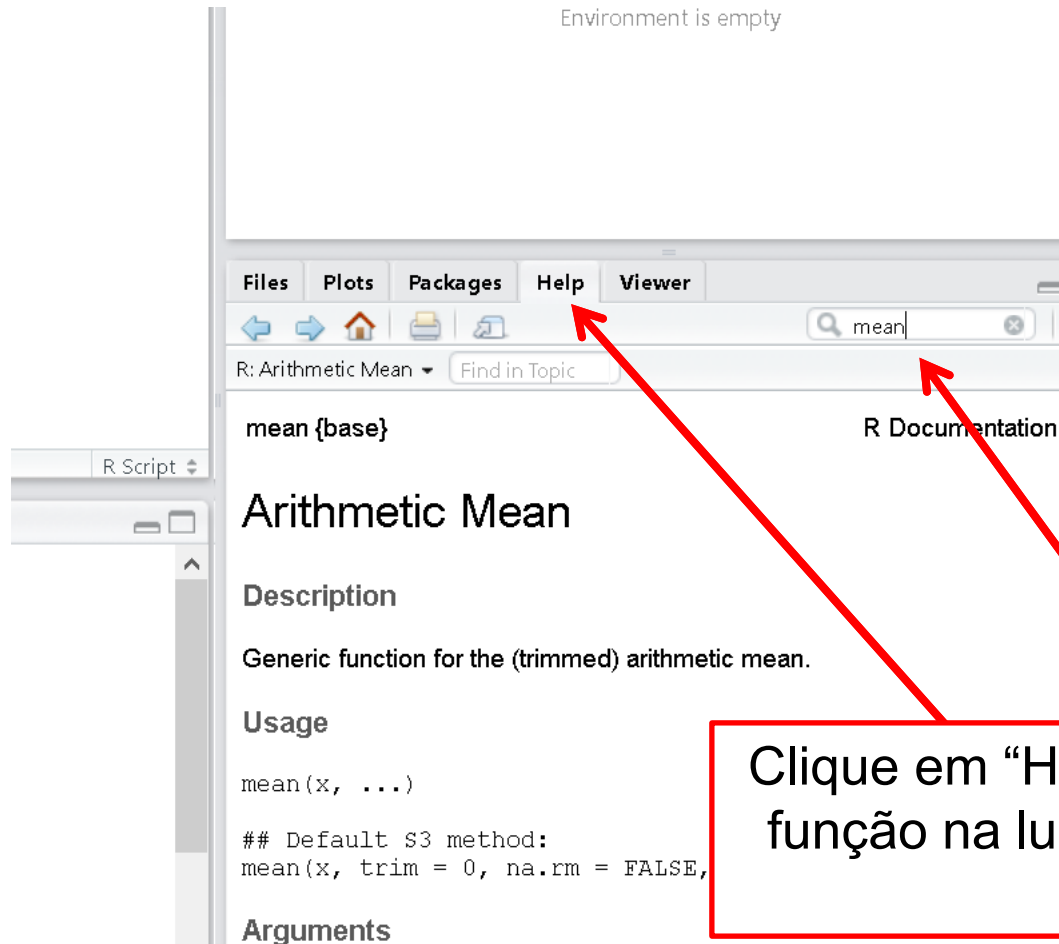
The screenshot shows the RStudio interface with the 'Packages' tab selected. The 'User Library' section lists several installed packages:

| Name                               | Description  | Version |
|------------------------------------|--|---------|
| <input type="checkbox"/> abind     | Combine Multidimensional Arrays                                | 1.4-5   |
| <input type="checkbox"/> acepack   | ACE and AVAS for Selecting Multiple Regression Transformations | 1.4.1   |
| <input type="checkbox"/> agricolae | Statistical Procedures for Agricultural Research               | 1.3-1   |
| <input type="checkbox"/> AlgDesign | Algorithmic Experimental Design                                | 1.1-7.3 |

# Limpar Arquivos



# Ajuda das Funções



Clique em “Help” e depois digite a função na lupinha. Por exemplo, “mean”

# R como calculadora

1+2+3

2+3\*4

3/2+1

3\*\*3

sqrt(2)

abs(-2\*2)

| Nome              | Operação   |
|-------------------|--|
| sqrt              | <i>raiz quadrada</i>                                 |
| abs               | <i>valor absoluto (positivo)</i>                     |
| sin cos tan       | <i>funções trigonométricas</i>                       |
| asin acos atan    | <i>funções trigonométricas inversas</i>              |
| sinh cosh tanh    | <i>funções hiperbólicas</i>                          |
| asinh acosh atanh | <i>funções hiperbólicas inversas</i>                 |
| exp log           | <i>exponencial e logaritmo natural</i>               |
| log10             | <i>logaritmo base-10</i>                             |
| gamma lgamma      | <i>função gamma function e seu logaritmo natural</i> |



# Operadores Lógicos

---

1 == 2

1 != 2

1 <= 2

1 < 2

1 > 2

1 >= 2

# Variáveis (Objetos)

```
x <- 2*3
```

```
x
```

```
x = 2*3
```

```
y <- sqrt(5)
```

```
z <- y+x
```

```
z
```

```
xsq = x**2 + y**2
```

```
xsq
```

# Estrutura de Dados

- **Vetores:** Podemos definir os vetores como uma sequência de valores alfanuméricos.

```
idade <- c(25, 32, 27, 33, 42, 21, 35, 45, 33, 25)
```

- **Fatores:** Podemos definir os fatores como uma sequência de valores, definido por níveis.

```
sexo <- c("Masc", "Fem", "Fem", "Fem", "Masc",  
"Fem", "Masc", "Masc", "Fem", "Fem")
```

# Estrutura de Dados

- **Dataframe:** A forma como os dados estão estruturados pode ser determinante para se conseguir realizar determinada análise. O objeto do tipo *dataframe* pode ser a melhor forma de armazenar os dados, pois ele pode conter vetores alfanuméricos e fatores.

```
df <- data.frame(idade, sexo)
```

# Estrutura de Dados

---

- **Listas:** Objetos da classe lista são muito úteis, pois são estruturas capazes de conter objetos de diversos tipos de classes.

```
lista <- list(idade, sexo, df)
```

# Leitura de uma base de dados

---

- Base de dados no formato “csv”
- Instalar o pacote “data.table”
- Carregar o pacote
- Leitura da base

## Exemplo: BASE SALÁRIO DE FUNCIONÁRIOS QUE TRABALHAM COM DATA SCIENCE

### DADOS

| A    | B           | C       | D          | E          | F         | G         | H         | I           |  |
|------|-------------|---------|------------|------------|-----------|-----------|-----------|-------------|--|
| ano  | experiencia | emprego | cargo      | salario_US | pais_empr | trab_remo | pais_empr | tam_empresa |  |
| 2020 | MI          | FT      | Data Scien | 79833      | DE        | 0         | DE        | L           |  |
| 2020 | SE          | FT      | Machine L  | 260000     | JP        | 0         | JP        | S           |  |
| 2020 | SE          | FT      | Big Data E | 109024     | GB        | 50        | GB        | M           |  |
| 2020 | MI          | FT      | Product Da | 20000      | HN        | 0         | HN        | S           |  |
| 2020 | SE          | FT      | Machine L  | 150000     | US        | 50        | US        | L           |  |
| 2020 | EN          | FT      | Data Analy | 72000      | US        | 100       | US        | L           |  |
| 2020 | SE          | FT      | Lead Data  | 190000     | US        | 100       | US        | S           |  |
| 2020 | MI          | FT      | Data Scien | 35735      | HU        | 50        | HU        | L           |  |
| 2020 | MI          | FT      | Business D | 135000     | US        | 100       | US        | L           |  |
| 2020 | SE          | FT      | Lead Data  | 125000     | NZ        | 50        | NZ        | S           |  |
| 2020 | EN          | FT      | Data Scien | 51321      | FR        | 0         | FR        | S           |  |
| 2020 | MI          | FT      | Data Scien | 40481      | IN        | 0         | IN        | L           |  |
| 2020 | EN          | FT      | Data Scien | 39916      | FR        | 0         | FR        | M           |  |
| 2020 | MI          | FT      | Lead Data  | 87000      | US        | 100       | US        | L           |  |
| 2020 | MI          | FT      | Data Analy | 85000      | US        | 100       | US        | L           |  |
| 2020 | MI          | FT      | Data Analy | 8000       | PK        | 50        | PK        | L           |  |
| 2020 | EN          | FT      | Data Engin | 41689      | JP        | 100       | JP        | S           |  |

salarios.csv

<https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>

### Descrição dos Dados

| Variável     | Descrição  |
|--------------|--|
| ano          | O ano em que o salário foi pago.   |
| experiencia  | O nível de experiência no cargo durante o ano com os seguintes valores possíveis: <b>EN</b> (Nível básico / Junior), <b>MI</b> (Nível médio / Intermediário), <b>SE</b> (Nível sênior / Expert), <b>EX</b> (Nível executivo / Diretor) |
| emprego      | O tipo de emprego para a função: <b>PT</b> (Part-time), <b>FT</b> (Full-time), <b>CT</b> (Contract), <b>FL</b> (Freelance)   |
| cargo        | A função exercida durante o ano  |
| salario_USD  | O salário em USD (taxa de câmbio dividida pela taxa média em USD para o respectivo ano via <a href="https://fxdata.foorilla.com">fxdata.foorilla.com</a> ).  |
| pais_empreg  | O país de residência do funcionário durante o ano de trabalho como um código de país ISO 3166.   |
| trab_remoto  | O tempo total de trabalho feito remotamente, os valores possíveis são os seguintes: 0 Nenhum trabalho remoto (menos de 20%), 50 Parcialmente remoto, 100 Totalmente remoto (mais de 80%)   |
| pais_empresa | O país da sede do empregador ou da filial contratante como um código de país ISO 3166.   |
| tam_empresa  | O número médio de pessoas que trabalharam para a empresa durante o ano: S menos de 50 funcionários (pequeno), M 50 a 250 funcionários (médio), L mais de 250 funcionários (grande)   |



```
# INSTALAR O PACOTE data.table  
install.packages("data.table")
```

```
# CARREGAR O PACOTE  
library(data.table)
```

```
# LEITURA DA BASE  
dados <- fread(input = "salarios.csv", header = T, na.strings = "NA",  
data.table = FALSE, dec=",")
```

## SINTAXE DE VERIFICAÇÃO DA LEITURA DOS DADOS

```
class(dados)
dim(dados)
names(dados)
str(dados)
head(dados)
tail(dados)
sapply(dados, function(x)(sum(is.na(x)))) # contagem de dados faltantes

mean(dados$salario_USD) # salário médio da população N = 607
sd(dados$salario_USD)   # desvio padrão
```

Exemplo 1: Selecionar os salários do ano de pagamento 2000

```
dados1 <- dados[dados$ano==2020,]
```

Exemplo 2: Selecionar os funcionários com contrato de tempo integral

```
dados2 <- dados[dados$emprego=="FT",]
```

## Tamanho da amostra para estimar uma média no R

# Exemplo para o cálculo do tamanho de uma amostra para estimar o salário médio dos profissionais de Data Science

# N = 607

# n = ?

# d = 10.000 (erro estipulado pelo pesquisador)

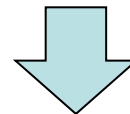
# nível de confiança desejado (95%)

$$n_0 = \frac{z^2 \sigma^2}{d^2}$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

```
install.packages("samplingbook")
library(samplingbook)
# o pacote samplingbook utiliza as fórmulas de tamanho de
amostra apresentadas na aula 2

# tamanho da amostra para média
sample.size.mean(e=10000, S = 71000, N = 607, level =
0.95)
# o "e" equivale ao "d" das fórmulas apresentadas na aula 2
```



sample size needed: 147

## Extraindo uma amostra para estimar uma média no R

```
# AMOSTRA SIMPLES AO ACASO
```

```
asa147 <- dados[sample(nrow(dados), size=147),]
```

```
mean(asa147$salario_USD) # salário médio da amostra n =147
```

```
mean(dados$salario_USD) # salário médio da população N = 607
```

Salário médio da  
população: \$ 112.297,90

Salário médio da  
amostra: \$ 111.066,20

```
# intervalo de confiança
```

```
a = mean(asa147$salario_USD)
```

```
b = mean(dados$salario_USD)
```

```
a
```

```
b
```

```
e = 10000 # erro definido pelo pesquisador
```

```
li = a - e # limite inferior do IC
```

```
ls = a + e # limite superior do IC
```

```
li
```

```
ls
```

```
b # verdadeiro salário médio
```

Intervalo de confiança de  
95% para o verdadeiro  
salário médio  
[101.066,20 ; 121.066,20]

# Tabelas

## Variáveis Qualitativas (uma variável qualitativa)

### # Construção de tabela para uma variável qualitativa

```
# carregando pacotes
```

```
rm(list=ls(all=TRUE))
```

```
library(data.table)
```

```
library(RcmdrMisc) ## Para usar as funções do Rcmdr (ex. Recode)
```

```
library(dplyr)
```

```
# leitura da base
```

```
base <- fread(input = "salarios.csv", header = T, na.strings = "NA", data.table = FALSE,  
dec = ",")
```

```
str(base$trab_remoto)
```

```
base$trab_remoto <- as.character(base$trab_remoto)
```

```
str(base$trab_remoto)
```

# Tabelas

## Variáveis Qualitativas (uma variável qualitativa)

### # Construção de tabela para uma variável qualitativa

```
# Alterando a base de dados
```

```
# Recodificar a variável trab_remoto com a função ifelse
```

```
base$trab_remoto <-
```

```
ifelse(base$trab_remoto=="0","1:Não",ifelse(base$trab_remoto=="50","2:Parcial","3:Total"))
```

```
table(base$trab_remoto)
```

```
str(base$experiencia)
```

```
table(base$experiencia)
```

```
# Recodificar a variável experiencia sem usar função
```

```
base$experiencia[base$experiencia == "EN"] = "1:EN"
```

```
base$experiencia[base$experiencia == "MI"] = "2:MI"
```

```
base$experiencia[base$experiencia == "SE"] = "3:SE"
```

```
base$experiencia[base$experiencia == "EX"] = "4:EX"
```

```
table(base$experiencia)
```

```
write.csv2(base,"base_modif.csv", row.names=FALSE)
```

# Tabelas

## Variáveis Qualitativas (uma variável qualitativa)

# Construção de tabela para uma variável qualitativa

# TABELAS

# UMA VARIÁVEL QUALITATIVA: Tabela para **trabalho remoto**

```
local({  
  .Table <- with(base, table(trab_remoto))  
  cat("\ncounts:\n")  
  print(.Table)  
  cat("\npercentages:\n")  
  print(round(100*.Table/sum(.Table), 2))  
})
```



## Tabela para variável trab\_remoto

trab\_remoto

| 1:Não | 2:Parcial | 3:Total |
|-------|-----------|---------|
| 127   | 99        | 381     |

percentages:

trab\_remoto

| 1:Não | 2:Parcial | 3:Total |
|-------|-----------|---------|
| 20.92 | 16.31     | 62.77   |

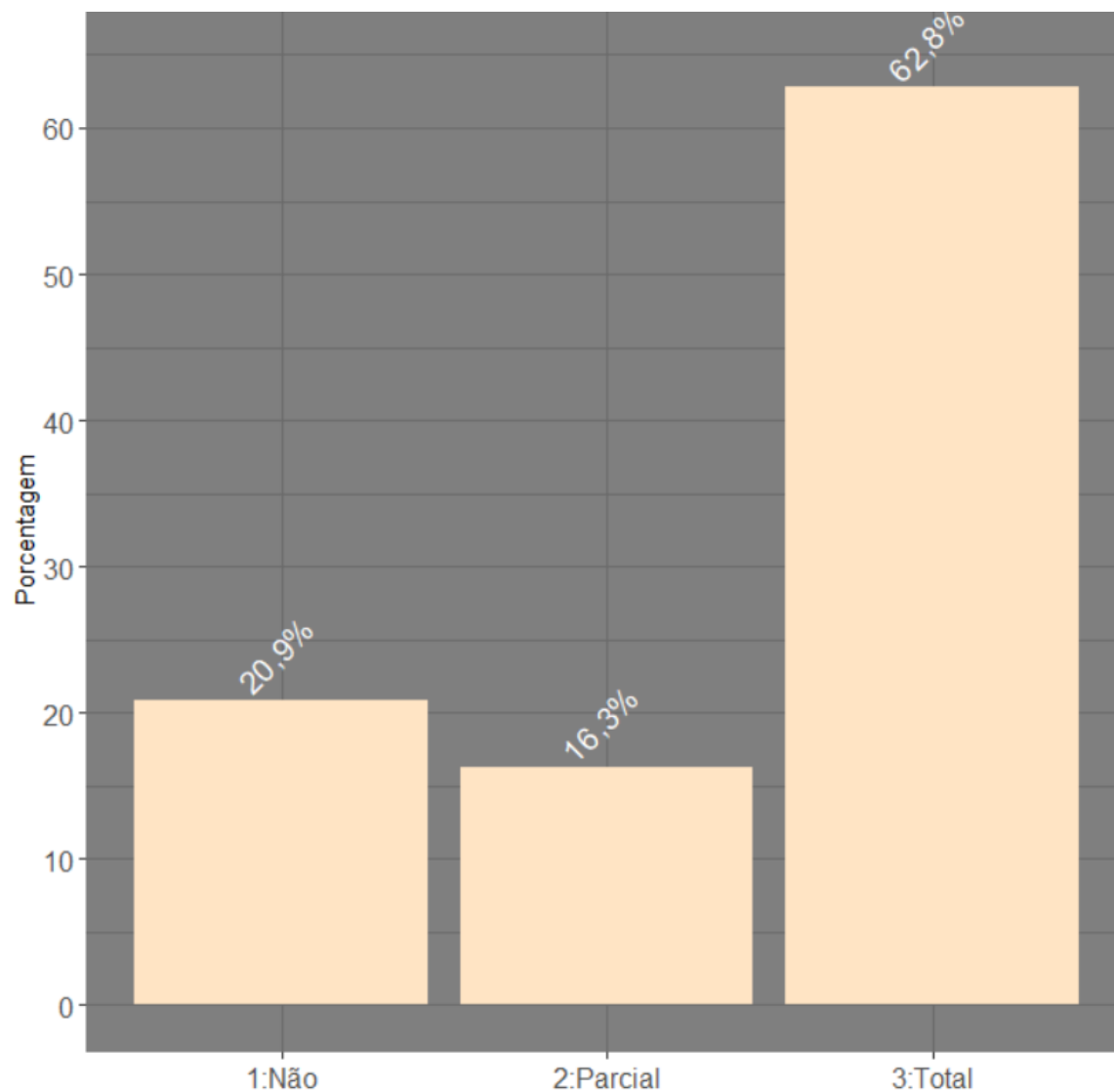
# Gráficos

## Variável Qualitativa – gráfico para trabalho remoto

```
library(ggplot2)
library(stringr)

format.args = list(decimal.mark = ",", big.mark = ".")
Freq = data.frame(table(base$trab_remoto))
Freq$Percentual=round(100*Freq[,2]/sum(Freq[,2]),1)
names(Freq)=c("Resposta","Frequência","Porcentagem")
Graf1 = ggplot(Freq, aes(y = Porcentagem, x = Resposta, ymax=65)) + # Ajustar ymax=65
  geom_bar(stat="identity", position="dodge", fill="bisque1") +
  geom_text(aes(label=scales::percent(Porcentagem/100, decimal.mark = ",",
accuracy=0.1)),vjust=-1.0, hjust=0.2,
            size=5.0, position = position_dodge(0.9), angle=45, colour="white") +
  xlab("") +
  theme_dark() +
  theme(legend.text = element_text(size=12), axis.text=element_text(size=12),
legend.position="bottom") +
  scale_y_continuous(breaks = c(0,10,20,30,40,50,60)) #Corrigido em função do tamanho das
colunas
print(Graf1)
```

# Gráfico para variável trab\_remoto



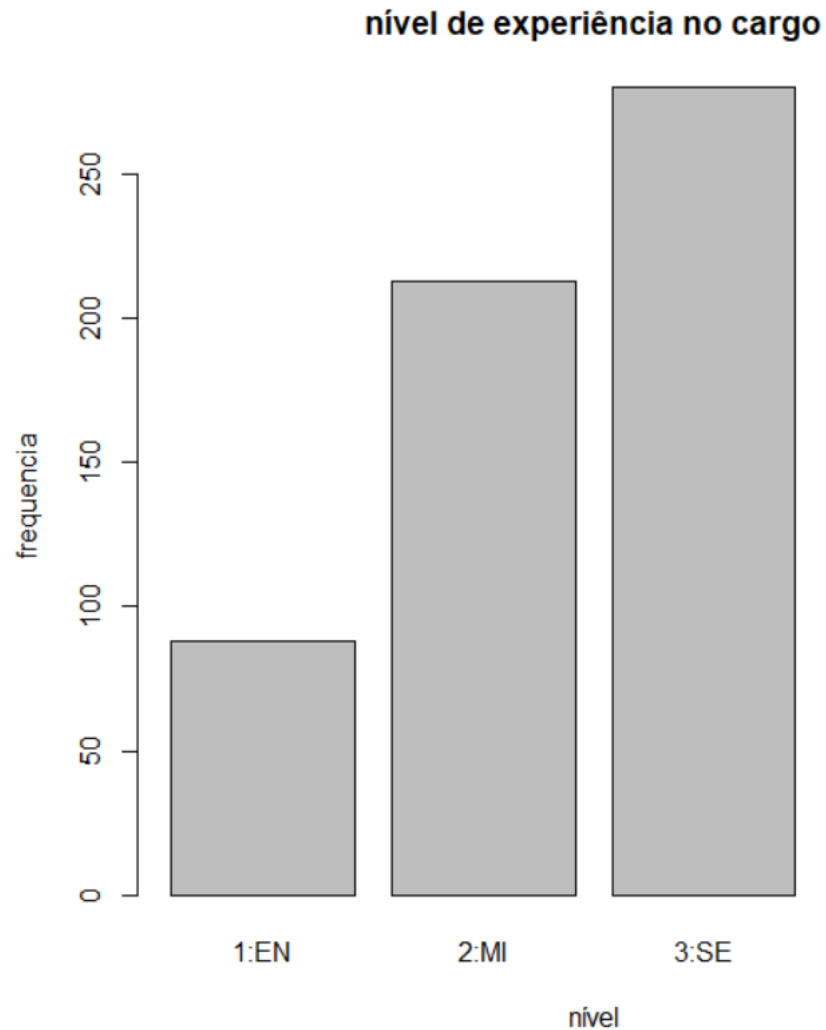
# Gráficos

## Variável Qualitativa – gráfico para experiência

# Gráfico para experiência

```
install.packages("vcd")  
library(vcd)  
counts <- table(base$experiencia)  
counts  
barplot(counts,  
        main = "nível de experiência no cargo",  
        xlab = "nível", ylab = "frequencia")
```

# Gráfico para variável experiência no cargo



# Tabelas

Variáveis: Qualitativa vs Qualitativa (tabela cruzada)

## # Construção de tabela para duas variáveis qualitativas

```
.Table <- xtabs(~experiencia+trab_remoto, data=base)  
rowPercents(.Table)
```

|             | trab_remoto |           |         |       |       |
|-------------|-------------|-----------|---------|-------|-------|
| experiencia | 1:Não       | 2:Parcial | 3:Total | Total | Count |
| 1:EN        | 15.9        | 28.4      | 55.7    | 100.0 | 88    |
| 2:MI        | 26.3        | 19.7      | 54.0    | 100.0 | 213   |
| 3:SE        | 19.3        | 9.6       | 71.1    | 100.0 | 280   |
| 4:EX        | 11.5        | 19.2      | 69.2    | 99.9  | 26    |

# Gráficos

## Duas Variáveis Qualitativas (experiência x trabalho remoto)

### OPÇÃO 1

```
freq.tabela <- table(base$experiencia,base$trab_remoto, useNA = "ifany")
```

```
freq.tabela
```

```
porc.tabelaL <- round(prop.table(freq.tabela,1)*100,1)
```

```
porc.tabelaL
```

```
tabela <- data.frame(table(base$experiencia,base$trab_remoto))
```

```
colnames(tabela) <- c("Experiencia","Trab_Remoto","Freq")
```

```
ggplot(tabela, aes(fill=Trab_Remoto, y=Freq, x=Experiencia)) +
```

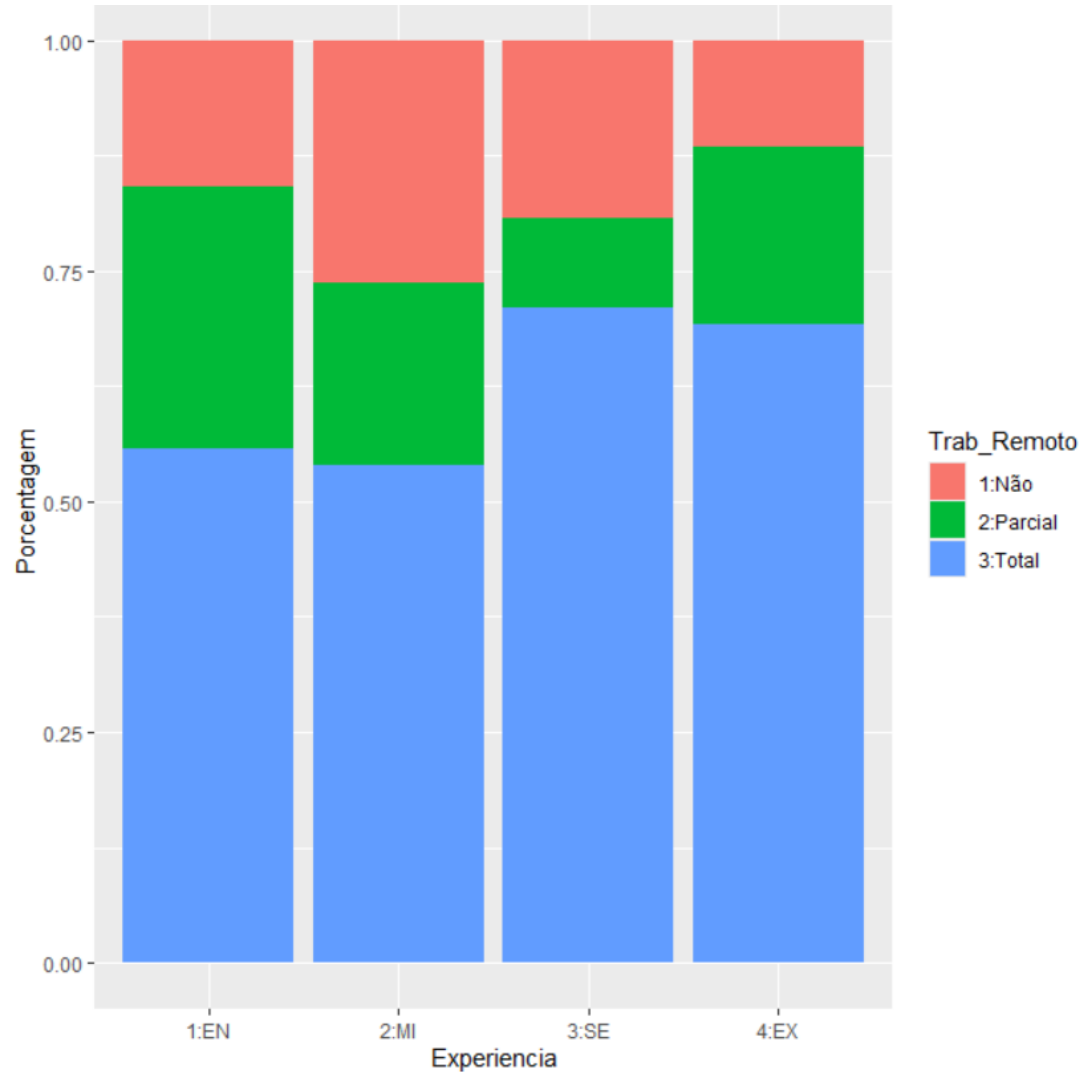
```
  geom_bar(position="fill", stat="identity") +
```

```
  ylab("Porcentagem")
```

# Gráfico para variáveis experiência x trab\_remoto

## Duas Variáveis Qualitativas

OPÇÃO 1





# Gráficos

## Duas Variáveis Qualitativas (experiência x trabalho remoto) OPÇÃO 2

```
library(vcd)

counts <- table(base$experiencia, base$trab_remoto)

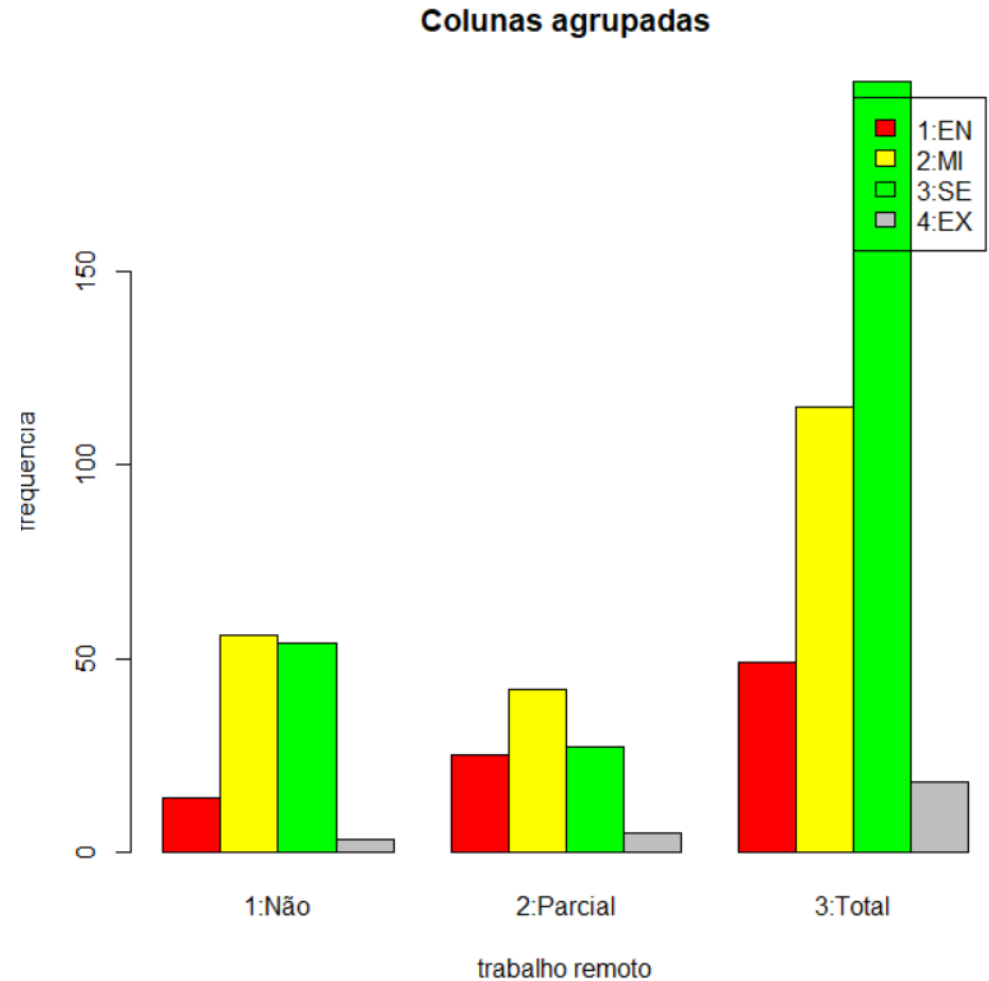
counts

barplot(counts,
        main = "Colunas agrupadas",
        xlab = "trabalho remoto", ylab = "frequência",
        col = c("red", "yellow","green","gray"),
        legend=rownames(counts), beside=TRUE)
```

# Gráfico para variáveis experiência x trab\_remoto

## Duas Variáveis Qualitativas

OPÇÃO 2



# Gráficos

## Duas Variáveis Qualitativas (experiência x trabalho remoto) OPÇÃO 3

```
# opção 3
```

```
counts <- table(base$trab_remoto, base$experiencia)
```

```
counts
```

```
barplot(counts,
```

```
  main = "Colunas agrupadas",
```

```
  xlab = "experiência", ylab = "frequência",
```

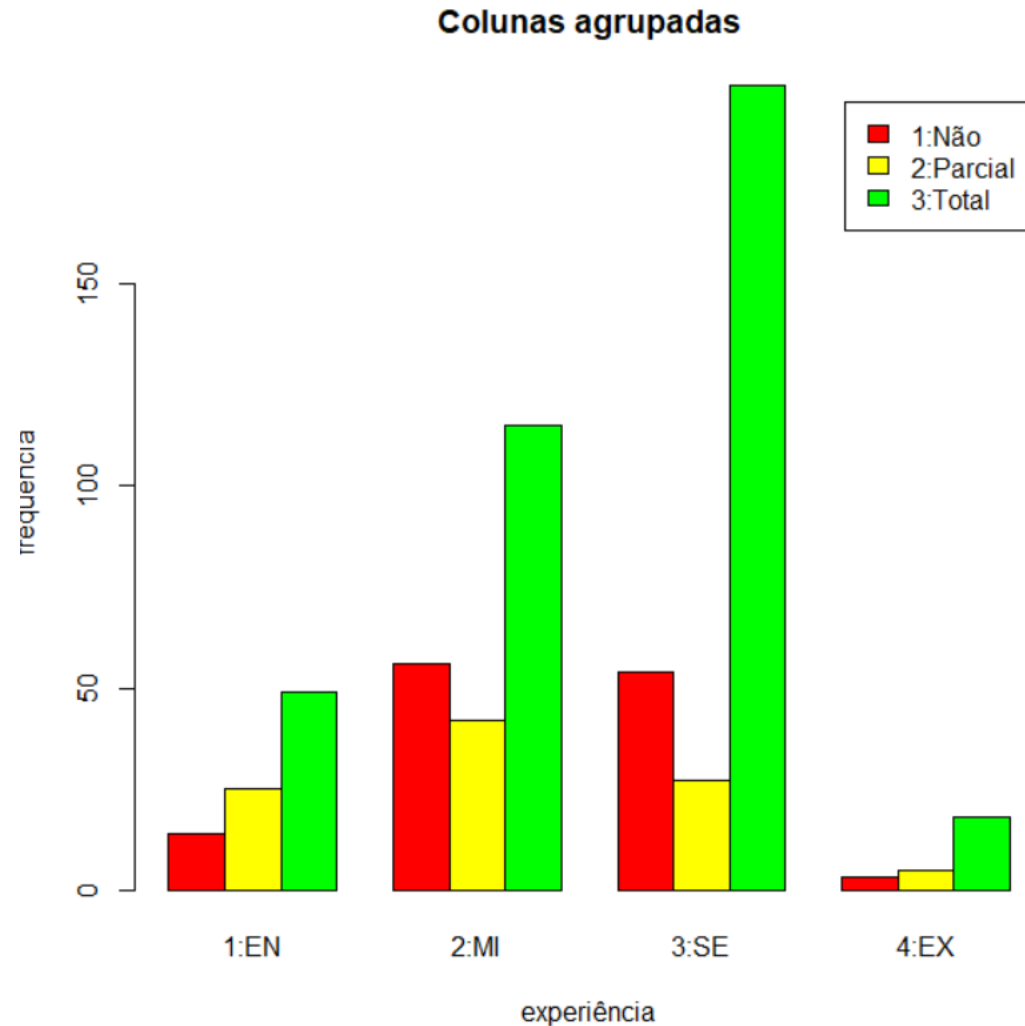
```
  col = c("red", "yellow", "green"),
```

```
  legend=rownames(counts), beside=TRUE)
```

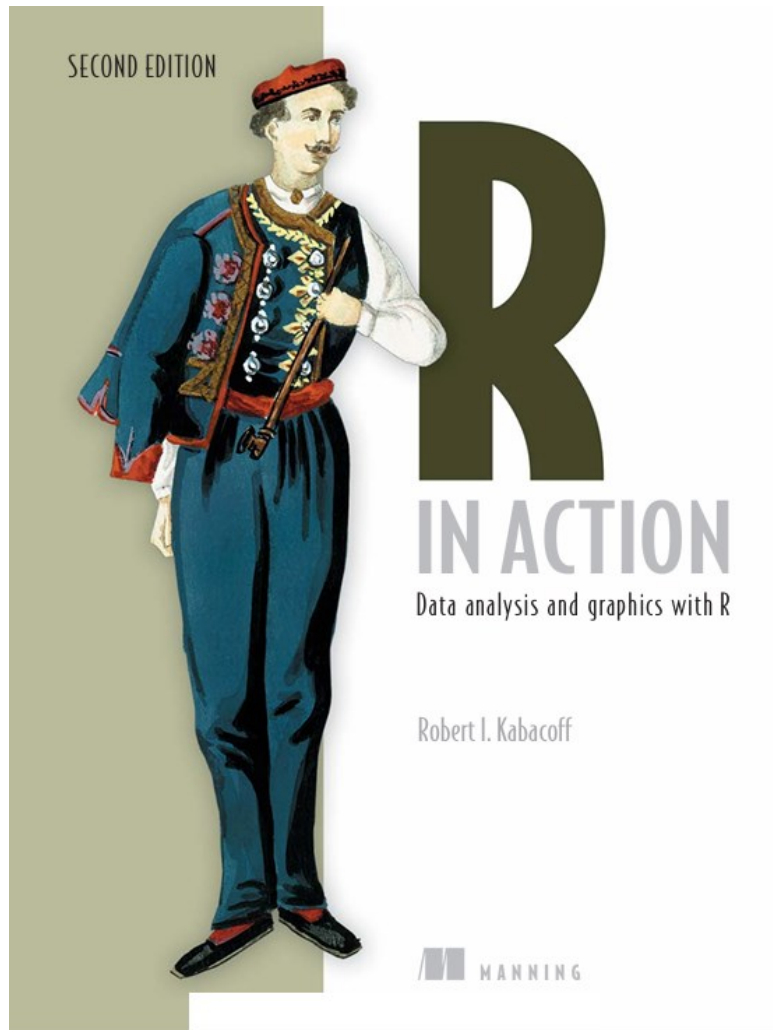
# Gráfico para variáveis experiência x trab\_remoto

## Duas Variáveis Qualitativas

OPÇÃO 3



# Sugestão para estudar o R



## Capítulos 1 a 7

### *brief contents*

|               |                               |            |
|---------------|-------------------------------|------------|
| <b>PART 1</b> | <b>GETTING STARTED .....</b>  | <b>1</b>   |
| 1             | ▪ Introduction to R           | 3          |
| 2             | ▪ Creating a dataset          | 20         |
| 3             | ▪ Getting started with graphs | 46         |
| 4             | ▪ Basic data management       | 71         |
| 5             | ▪ Advanced data management    | 89         |
| <b>PART 2</b> | <b>BASIC METHODS .....</b>    | <b>115</b> |
| 6             | ▪ Basic graphs                | 117        |
| 7             | ▪ Basic statistics            | 137        |

[https://drive.google.com/file/d/1uXjKdm3Vo3h\\_54h22byOe5K9P1ATBdeV/view?usp=drive\\_link](https://drive.google.com/file/d/1uXjKdm3Vo3h_54h22byOe5K9P1ATBdeV/view?usp=drive_link)