

CURSO R - 2024

Professora Dra. Naiara Sandi de Almeida Alcantara



TÓPICOS

1-INTRODUÇÃO AO R

2-ANÁLISE EXPLORATÓRIA E
MANIPULAÇÃO DOS DADOS

3-SALVAMENTO E ABERTURA

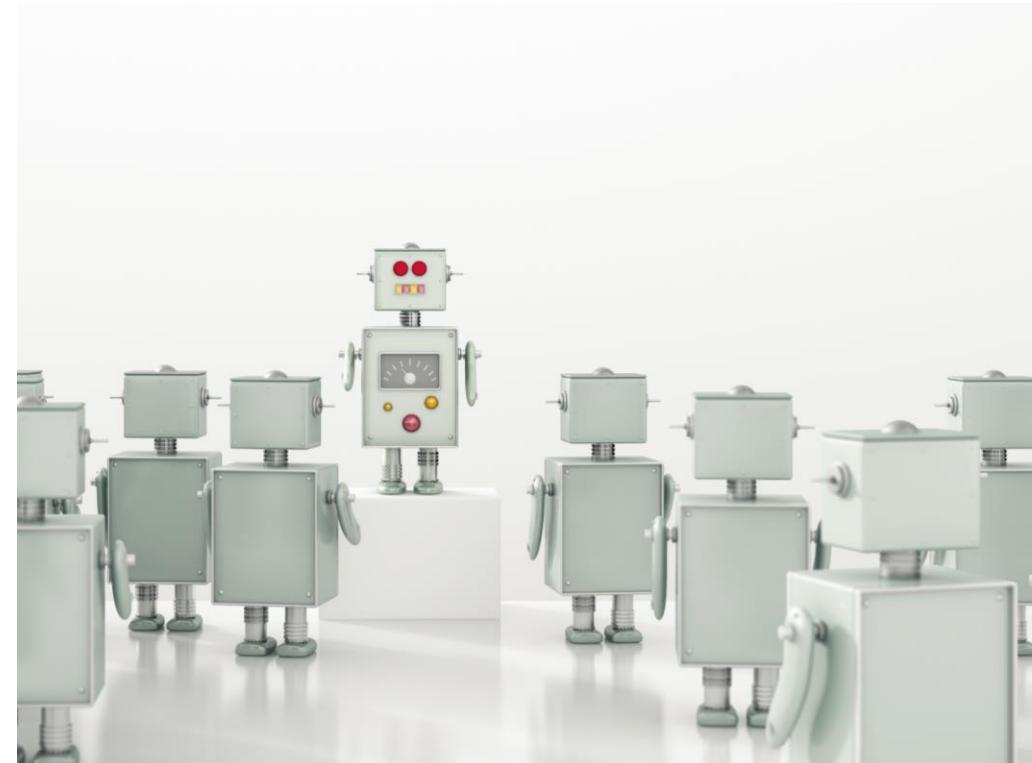
4-ANÁLISES DESCRIPTIVAS

5-APRESENTAÇÃO GRÁFICA

6- PROCESSAMENTO DE DADOS

7- ANÁLISES INFERENCIAIS

- **Download do cran e rstudio**
- **Apresentação do R**
- **Funções básicas**
- **Criação de data frame**



INSTALAÇÃO DO R

Vídeo demonstrando a instalação*

É possível que uma parte do vídeo esteja desatualizada, nesse caso, siga o passo a passo abaixo:

<https://www.youtube.com/watch?v=SwEizTHk57A>

(este vídeo foi produzido pelo aluno do mestrado em Ciência Política da UFPR, Renan Arnon de Souza, e disponibilizado para ser utilizado nessa oficina)

para utilizar o r é necessário seguir os 2 passos que estão disponíveis no site: <https://posit.co/download/rstudio-desktop/>

É preciso baixar o cran e depois o r studio

obs.: geralmente o próprio r já entende qual é o seu sistema operacional e indica o cran corretamente

The screenshot shows a web browser displaying the CRAN (Comprehensive R Archive Network) website at cran.r-project.org. The page title is "R-4.2.2 for Windows". On the left, there is a large blue "R" logo with "CRAN" written vertically next to it. Below the logo are links for "Mirrors", "What's new?", "Search", and "CRAN Team". In the center, there is a call-to-action button with the text "Download R-4.2.2 for Windows (76 megabytes, 64 bit)". Below this button are links for "RFADME on the Windows binary distribution" and "New features in this version". At the bottom of the page, there is a note about UCRT requirements and a link to a "Frequently asked questions" section.

ENTENDENDO A INTERFACE DO R



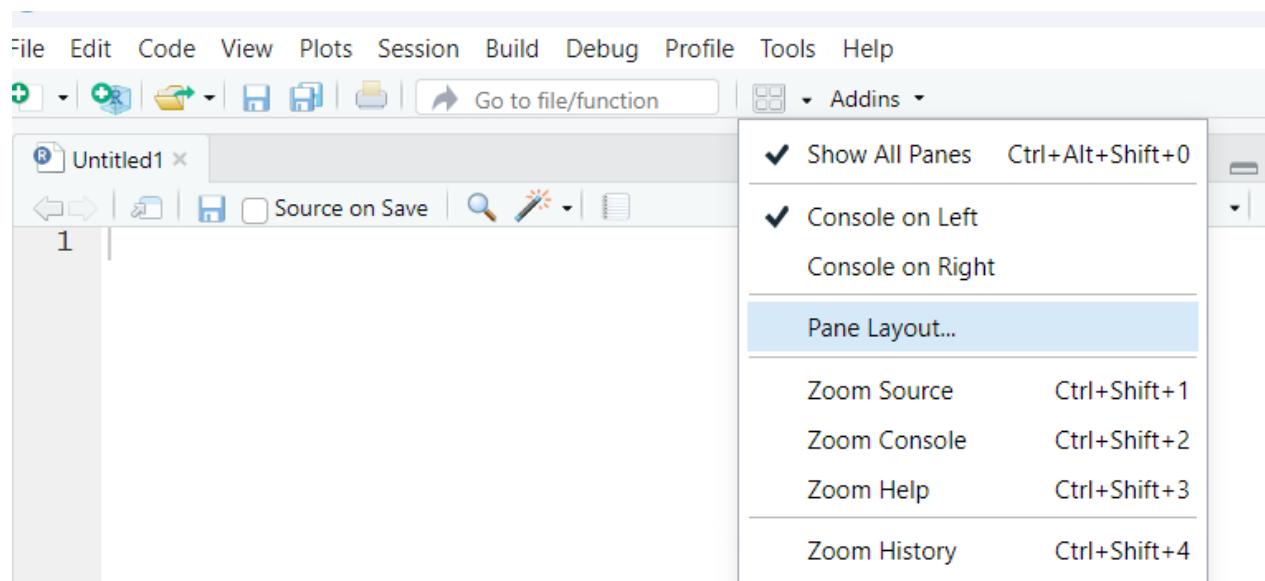
The screenshot displays the RStudio interface with several yellow callout boxes highlighting different components:

- Script**: Points to the top-left pane where an R script file named "Untitled1" is open.
- Environment/History/Connections**: Points to the top-right pane containing tabs for Environment, History, and Connections, along with a Global Environment viewer.
- Console**: Points to the bottom-left pane showing the R console output, including the R version information and a welcome message in Portuguese.
- Files/Plots/Packages/Help/Viewer**: Points to the bottom-right pane which includes tabs for Files, Plots, Packages, Help, and Viewer, along with a Zoom and Export tool.

Other visible elements include the RStudio menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help) and various toolbars and status bars throughout the interface.

BOAS PRÁTICAS E MUDANÇA DE VISUAL

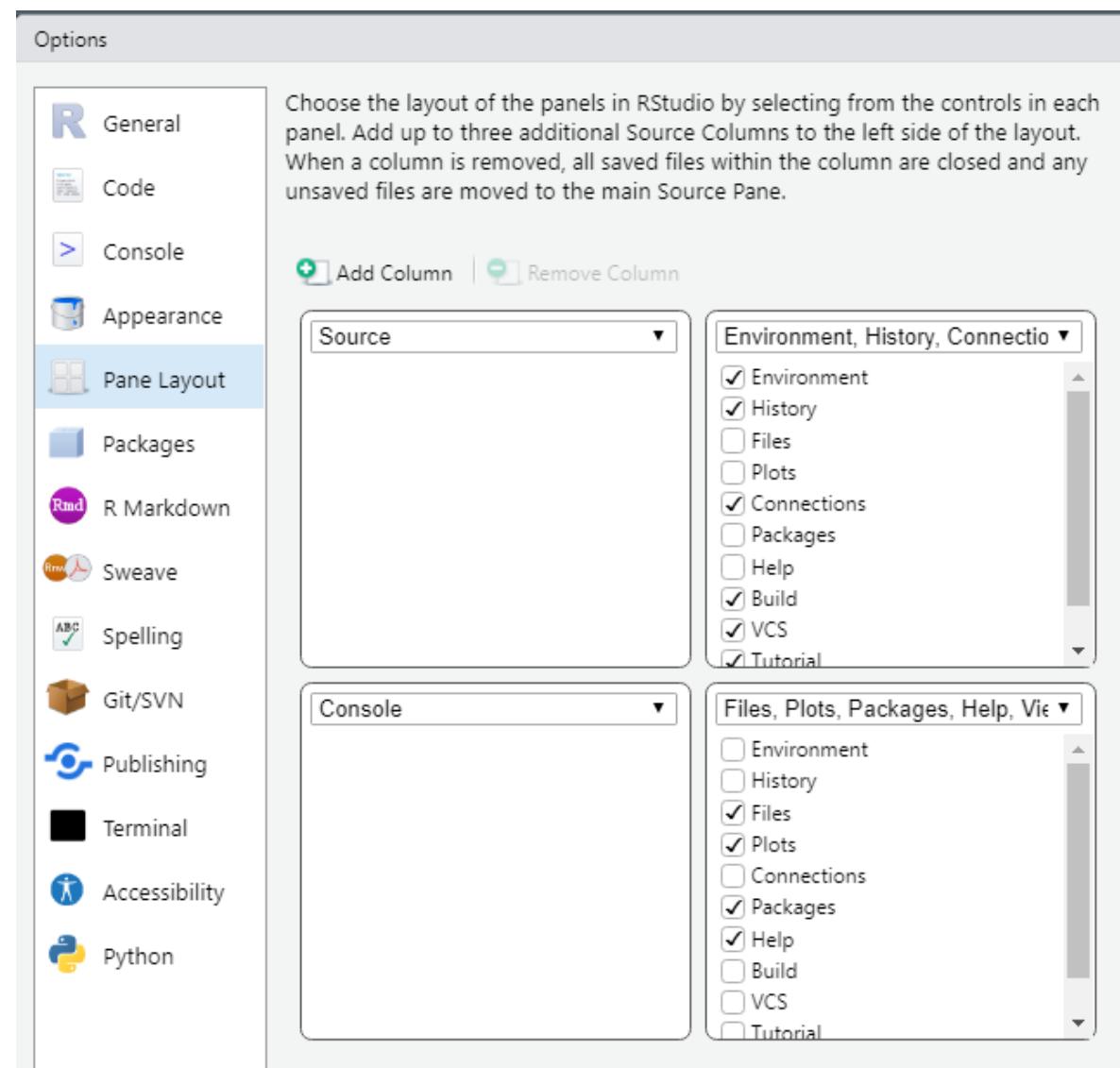
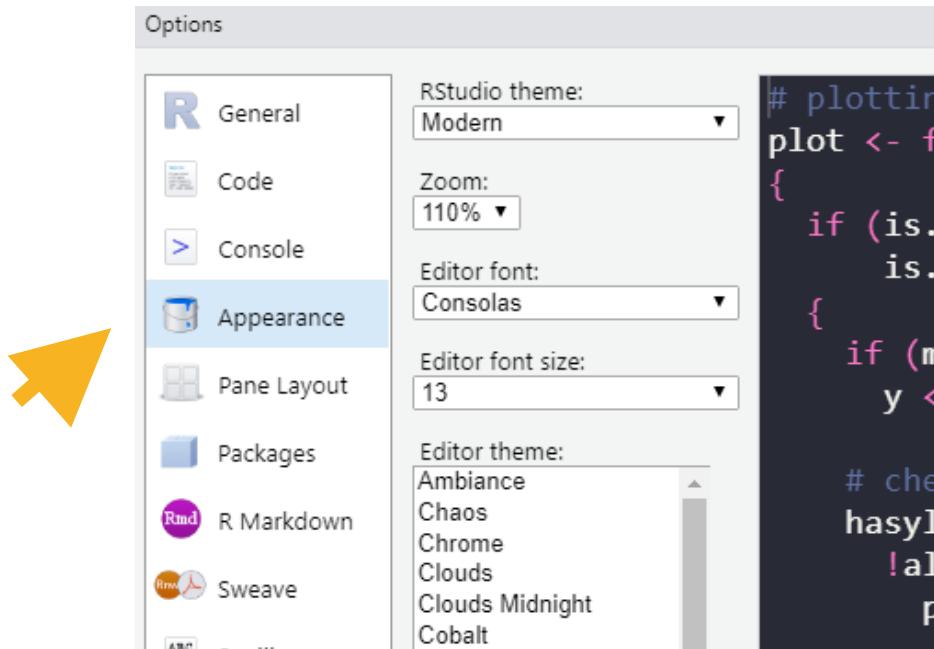
MUDANÇA DE VISUAL DO R:
1º WORKSPACE PANE > 2º PANE LAYOUT ... >



BOAS PRÁTICAS E MUDANÇA DE VISUAL

3º ABRIRÁ UMA CAIXA DE DIÁLOGO CHAMADA “OPTION”

4º CLIQUE EM APPEARANCE
5º ESCOLHA A FONTE, O TAMANHO, O ZOOM E ETC.

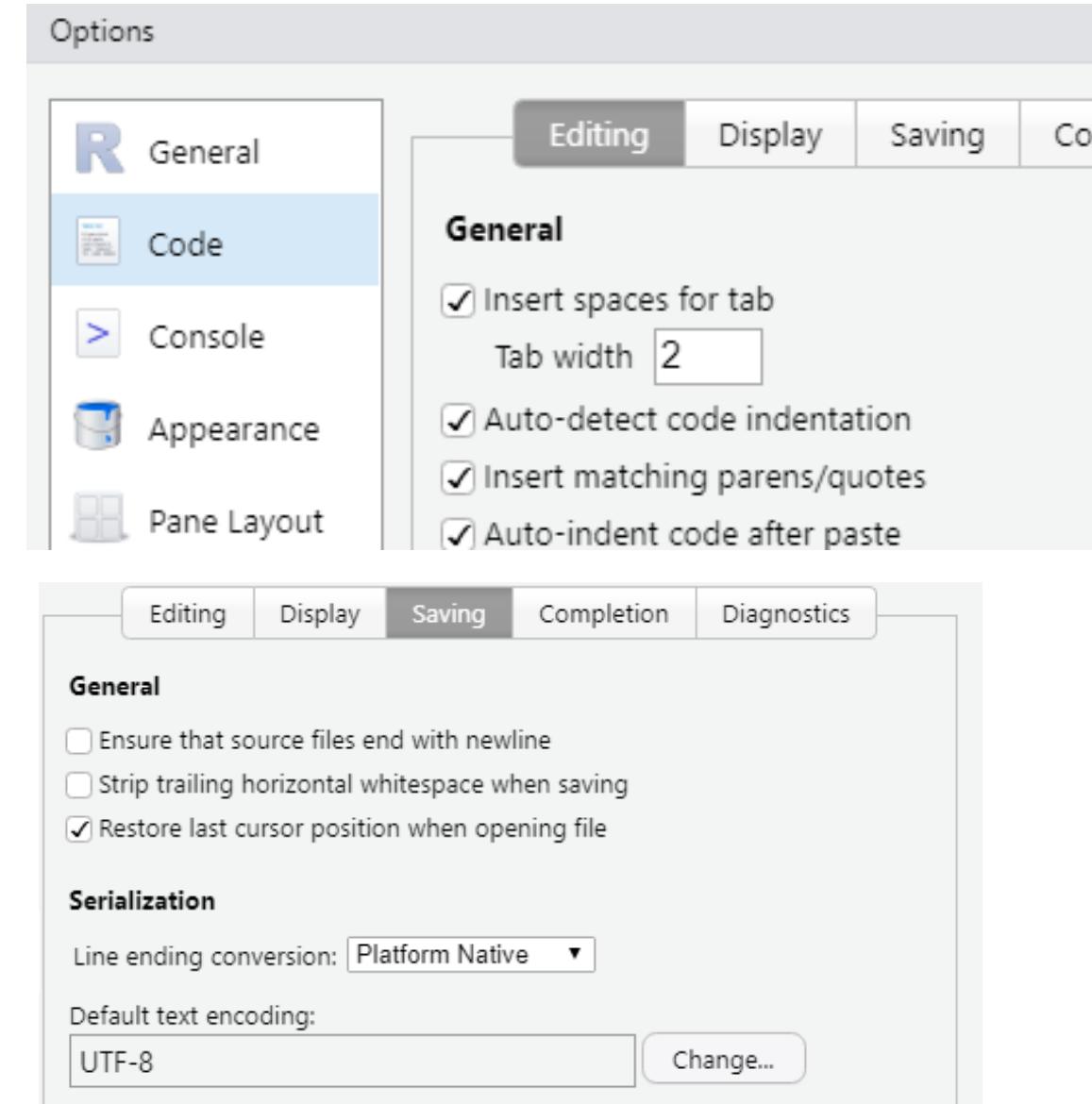


BOAS PRÁTICAS E MUDANÇA DE VISUAL

6º MUDANÇA DO CÓDIGO DE SALVAMENTO

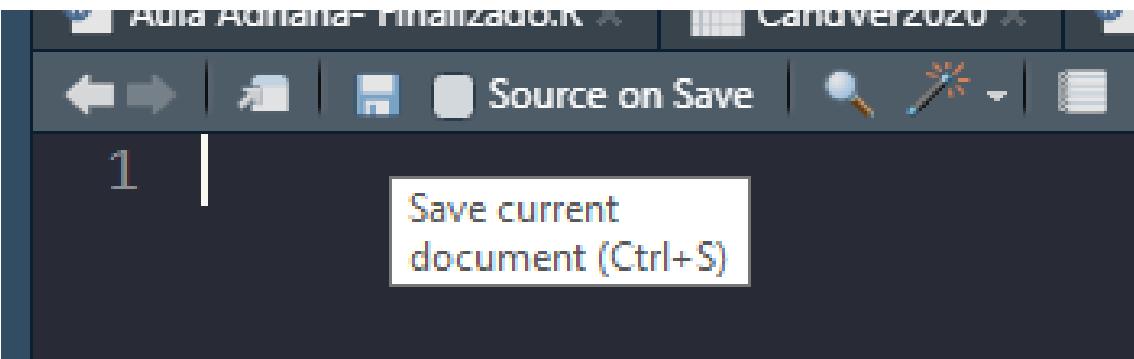
LINGUAGEM UNIVERSAL

- É ACONSELHÁVEL SALVAR OS SCRIPTS PARA QUE DEPOIS A ATIVIDADE DESENVOLVIDA POSSA SER RETOMADA, COMPARTILHADA E REPRODUZIDA.
- SALVE OS SCRIPTS EM UM FORMATO UNIVERSAL, POIS DESSA FORMA AO SER COMPARTILHADO O FORMATO NÃO SOFRERÁ MODIFICAÇÕES.
- ENTÃO NA MESMA CAIXA DE DIÁLOGO CLIQUE EM CODE
- ESCOLHA O TIPO DE SALVAMENTO CLICANDO EM CHANGE > OPTE SEMPRE PELO TIPO UNIVERSAL UTF- 8
- SALVE A LINGUAGEM DE ESCRITA EM PORTUGUÊS



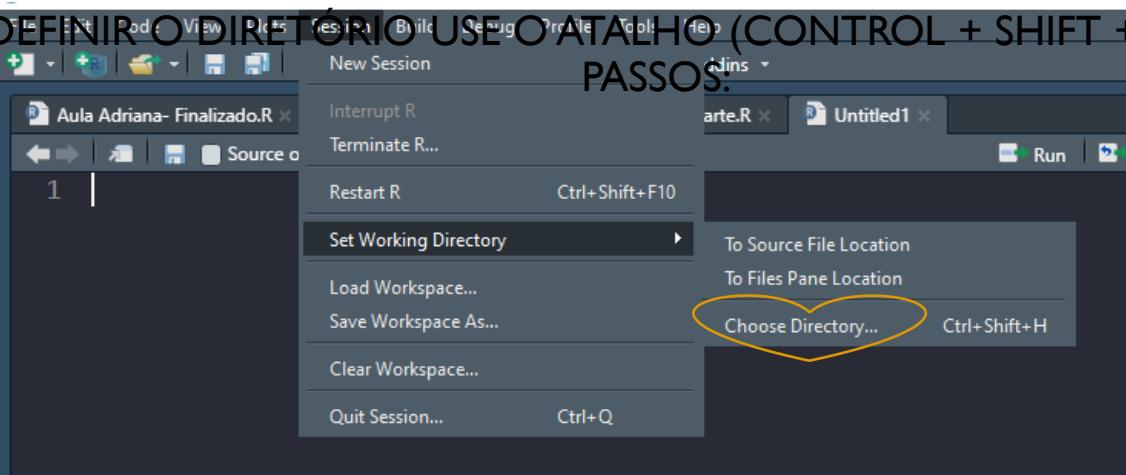
BOAS PRÁTICAS E MUDANÇA DE VISUAL

7º ORGANIZAÇÃO DO SUMÁRIO E SALVAR O SCRIPT



8º ENQUANTO ESCREVE O SCRIPT DE MANEIRA ORGANIZADA, SALVE EM SEU DIRETÓRIO DE TRABALHO,

PARA DEFINIR O DIRETÓRIO USE O ATALHO (CONTROL + SHIFT + H) OU SIGA OS PASSOS:



FUNÇÕES BÁSICAS DO R

Funções Básicas do R

1. Estrutura das Funções

- Definição: As funções em R seguem a estrutura nome_da_função(argumentos).
- Exemplos Comuns:

- `mean(x)`: Calcula a média de um vetor x.
- `sum(x)`: Soma todos os elementos de x.
- `length(x)`: Retorna o número de elementos em x.
- `sd(x)`: Calcula o desvio padrão de x.
- `min(x)` e `max(x)`: Retornam o menor e o maior valor de x, respectivamente.

2. Funções de Manipulação de Dados

- `subset()`: Extrai subconjuntos de dados.
 - Exemplo: `subset(dados, idade > 30)`.
- `apply()`: Aplica uma função a elementos de um array ou matriz.
 - Exemplo: `apply(matriz, 1, mean)` (calcula a média de cada linha).
- `tapply()`: Aplica uma função por grupo.
 - Exemplo: `tapply(vendas, grupo, sum)` (soma as vendas por grupo).
- `merge()`: Junta duas bases de dados por uma chave.
 - Exemplo: `merge(df1, df2, by = "ID")`.

FUNÇÕES BÁSICAS DO R

Objeto de atribuição

Atalho Alt - (menos)

`<-`

Pipi

Atalho Ctrl + shift + m

`%>%`

- Sempre que atribuímos valor a alguma coisa no R, essa coisa se chamará objeto.
- Tudo no R é um objeto.
- Nunca haverá dois objetos com o mesmo. Mas o mesmo valor poderá ter dois nomes diferentes.

Operadores	Descrição
<code><</code>	Menor que
<code><=</code>	Menor ou igual a
<code>></code>	Maior que
<code>>=</code>	Maior ou igual a
<code>==</code>	Igual a
<code>!=</code>	Diferente de
<code>!x</code>	Não x
<code>x y</code>	x OU y
<code>x & y</code>	x E y

Operadores de comparação

Operadores lógicos

FUNÇÕES BÁSICAS DO R

2. Operadores Lógicos

- E lógico: & (compara elemento por elemento) ou && (compara o primeiro elemento).
 - Exemplo: $(x > 5) \& (y < 10)$ (verifica se ambas as condições são verdadeiras para cada elemento).
- OU lógico: | (elemento por elemento) ou || (primeiro elemento).
 - Exemplo: $(x > 5) | (y < 10)$ (verifica se ao menos uma das condições é verdadeira).
- Não lógico: !
 - Exemplo: !($x == 5$) (nega a condição, ou seja, verifica se x não é igual a 5).

3. Funções Lógicas

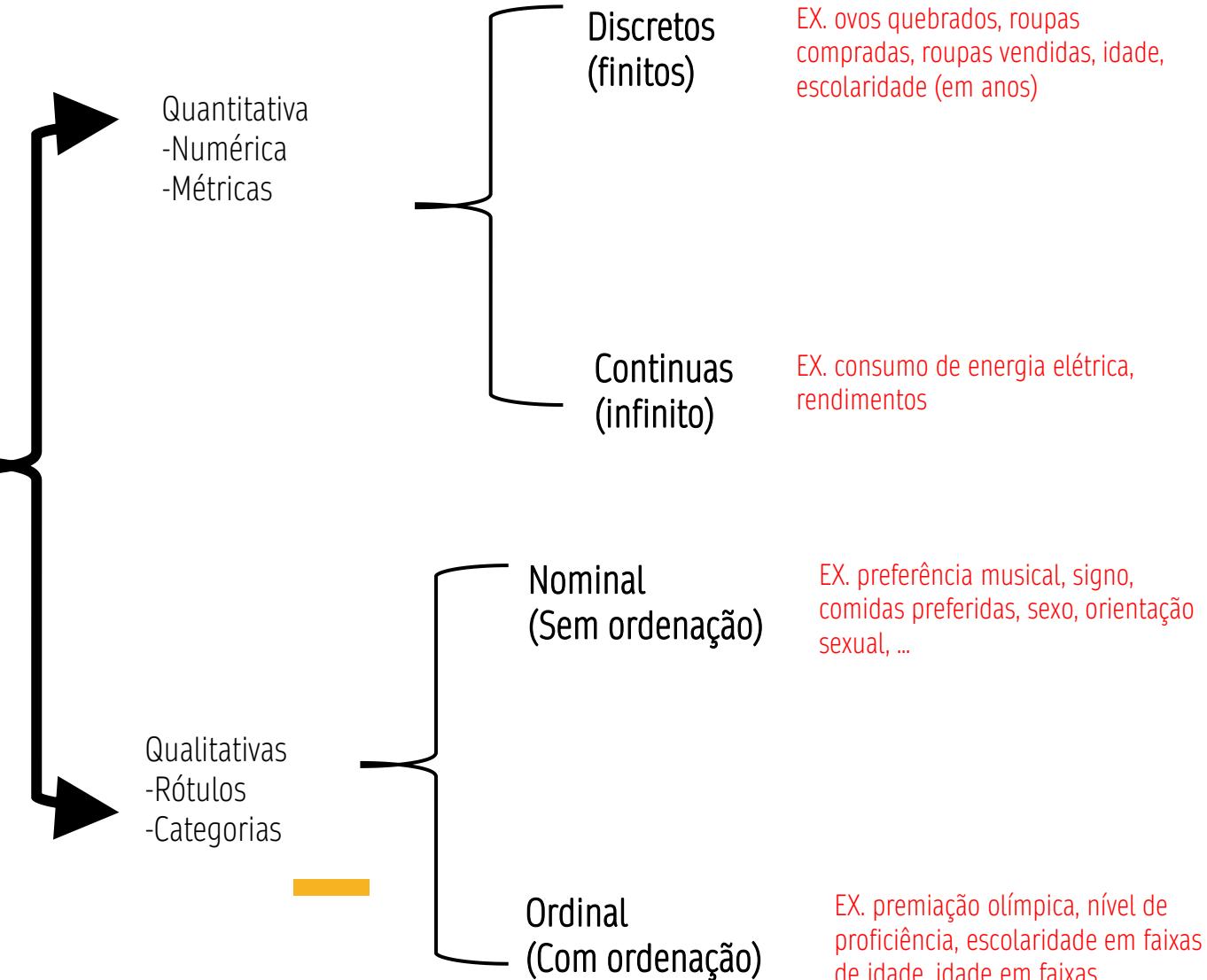
- all(): Verifica se todos os valores são verdadeiros.
 - Exemplo: `all(c(TRUE, TRUE, FALSE))` retorna FALSE.
- any(): Verifica se ao menos um valor é verdadeiro.
 - Exemplo: `any(c(FALSE, FALSE, TRUE))` retorna TRUE.
- is.na(): Verifica se há valores ausentes (NA).
 - Exemplo: `is.na(c(1, NA, 3))` retorna TRUE para o segundo elemento.

VARIÁVEIS

Antes da criação de um data frame é importante entender quais são os tipos de variáveis existentes, para:

- ! 1-Entender como inserir a variável na base
- 2-Escolher os testes mais adequados

Variável
1- Características
2-Atributos



CRIAÇÃO DE UM DATA FRAME

Primeiro criam-se os vetores > Depois juntamos os vetores através da função data.frame

```
#Criar data frame

Id ← c(1:4)
Nomes ← c("Ana", "Gilberto", "Rodrigo", "Marcela")
Peso ← c(75.6, 99, 62.8, 102)
Idades ← c(25, 18, 44, 23)
Escolaridade ← c("Graduação", "Mestrado", "Primário", "Graduação")
Exerc_Recomend ← c("Natação", "Pilates", "Musculação", "Corrida")
Comidas_preferidas ← c("Chocolate", "Sorvete", "Milho", "Pão")

Ficha_Pacientes ← data.frame(Id, Nomes, Peso,
                               Idades, Escolaridade,
                               Exerc_Recomend, Comidas_preferidas)
```



Lembre-se de salvar tanto o script, quanto a base de dados que você criou no diretório de trabalho correto.

```
#Função para ver o data frame
View(Ficha_Pacientes)

#Função para identificar o caminho em que o arquivo de script será salvo
getwd()
#[1] "C:/Users/nayar/OneDrive/8. AMBIENTE DE PROGRAMAÇÃO R/1. CURSO 2023"

#Função manual para selecionar o diretório de trabalho
setwd("C:/Users/nayar/OneDrive/8. AMBIENTE DE PROGRAMAÇÃO R/1. CURSO 2023")

#Função para salvar o data frame
save(Ficha_Pacientes, file = "Ficha_Pacientes.RData")
```

CRIAÇÃO DE UM DATA FRAME

*Criamos um
data frame,
mas o que é um
data frame?*

Id	Nomes	Peso	Idades	Escolaridade	Exerc_Recomend	Comidas_preferidas	Escolaridade2
1	Ana	75.6	25	Graduação	Natação	Chocolate	Graduação
2	Gilberto	99.0	18	Mestrado	Pilates	Sorvete	Mestrado
3	Rodrigo	62.8	44	Primário	Musculação	Milho	Primário
4	Marcela	102.0	23	Graduação	Corrida	Pão	Graduação

É o mesmo que base de dados, trata-se de um conjunto de informações em que todas as colunas possuem o mesmo tamanho* e podem ser abertas/criadas/modificadas em qualquer ambiente de programação (python, stata, sql) ou software de análise de dados, como o Excel, Spss e afins.

Cada linha do data frame representa um caso/observação e cada coluna representa uma variável/questão. O cabeçalho das bases de dados devem ser compostos pelos nomes das variáveis.

Variável é nome atribuído as questões estudadas, é um "objeto capaz de reter e representar um valor ou expressão". No ambiente de programação do R, **É importante que seja um nome curto, um resumo e que represente a questão, como "Nome", "Peso", ...**

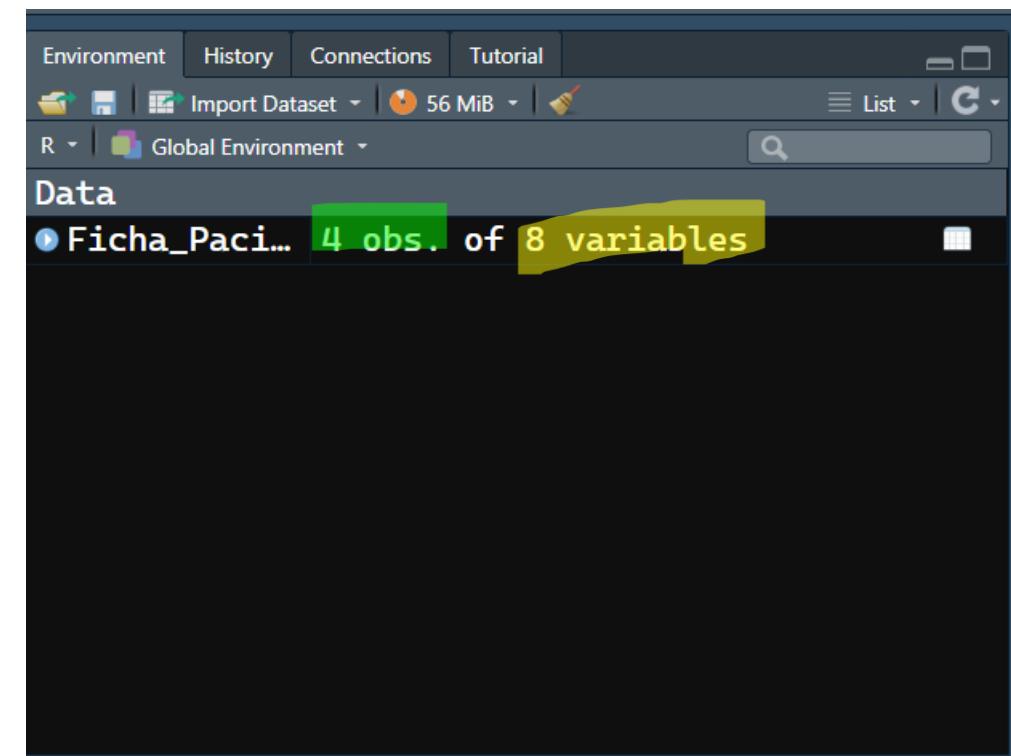
A principal característica de uma variável é sua capacidade de variar, portanto não existe variável com valor único. Se não varia, não é variável, é constante.

CRIAÇÃO DE UM DATA FRAME

📊 Quando criamos um data frame suas informações aparecem no Environment

📊 Apesar da base criada aparecer, isso não significa que ela está salva, por isso é importante salvá-la a cada modificação, utilizando a função save.

📊 Sobre salvamentos e aberturas de base de dados trataremos mais detalhadamente na seção sobre esse assunto.



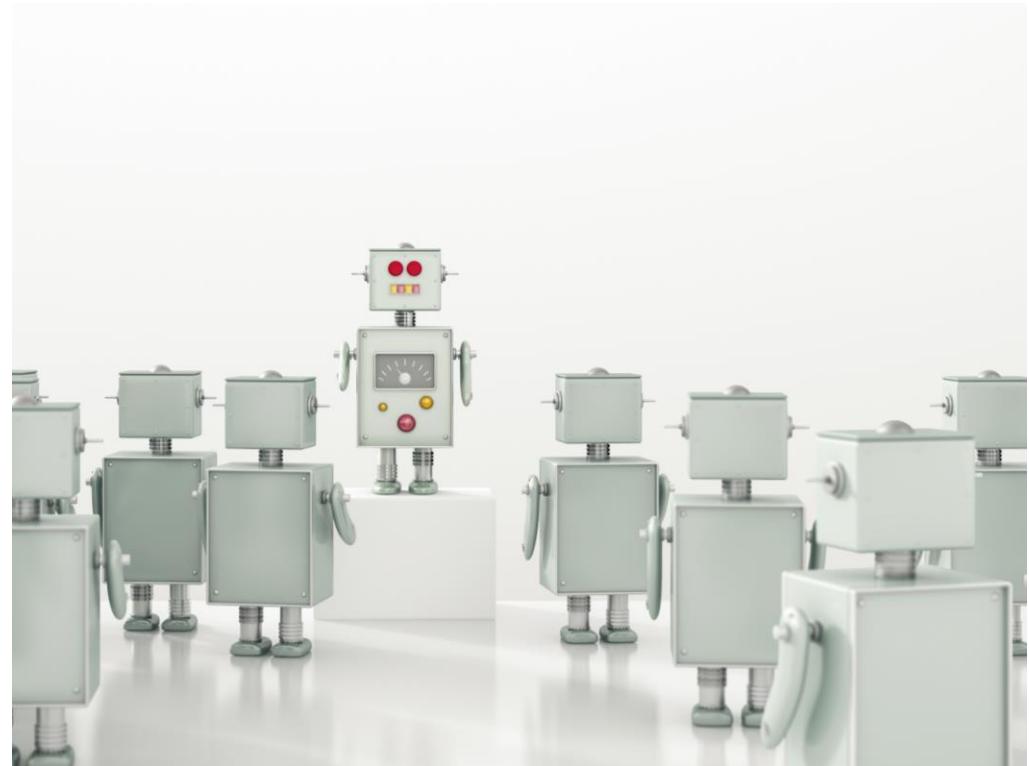


ANÁLISE EXPLORATÓRIA E ORGANIZAÇÃO DOS DADOS

TÓPICOS

- 1-INTRODUÇÃO AO R
- 2-ANÁLISE EXPLORATÓRIA E MANIPULAÇÃO DOS DADOS
- 3-SALVAMENTO E ABERTURA
- 4-ANÁLISES DESCRIPTIVAS
- 5-APRESENTAÇÃO GRÁFICA
- 6- PROCESSAMENTO DE DADOS
- 7- ANÁLISES INFERENCIAIS

- **Funções para observar a estrutura dos dados**
- **Seleção**
- **Exclusões**
- **Inserções e filtros**
- **Alterações**
- **Tipos de variáveis**



ANÁLISE DO DATA FRAME

Estrutura dos dados

```
#Função para análise das classes de cada variável
str(Ficha_Pacientes)
# 'data.frame': 4 obs. of 7 variables:
#   $ Id           : int 1 2 3 4
#   $ Nomes        : chr "Ana" "Gilberto" "Rodrigo" "Marcela"
#   $ Peso          : num 75.6 99 62.8 102
#   $ Idades        : num 25 18 44 23
#   $ Escolaridade  : chr "Graduação" "Mestrado" "Primário" "Graduação"
#   $ Exerc_Recomend: chr "Natação" "Pilates" "Musculação" "Corrida"
#   $ Comidas_preferidas: chr "Chocolate" "Sorvete" "Milho" "Pão"
```

- número, valor real, *numeric, double*
- texto, string, *character*, caracteres
- lógico, *logical*, booleano, valor TRUE/FALSE

Int são os números inteiros

- O R identifica o tipo de variável e diz qual seu tipo.
- É possível que o tipo não esteja adequado, nesse caso é necessário alterar o tipo usando a função “as.” com o tipo que queremos que seja inserido na variável, e depois para verificar se o tipo está correto usaremos a função “is.”. Falaremos mais sobre isso, em uma etapa mais avançada do curso

```
#A base de dados deverá ter um nome de coluna e linhas
#Para saber esses nomes use as funções rownames e colnames
rownames(Ficha_Pacientes)
#[1] "1" "2" "3" "4"
colnames(Ficha_Pacientes) #Usando a função names também
#Conseguimos ver todas as colunas
# [1] "Id"           "Nomes"         "Peso"          "Idades"
# [5] "Escolaridade" "Exerc_Recomend" "Comidas_preferidas"
```

ANÁLISE DO DATA FRAME - SELEÇÕES

```
#Análise das médias
print(mean(Ficha_Pacientes$Idades))
#[1] 27.5
summary(Ficha_Pacientes$Idades)
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 18.00 21.75 24.00 27.50 29.75 44.00
summary(Ficha_Pacientes$Exerc_Recomend)
# Length Class Mode
# 4 character character
table(Ficha_Pacientes$Escolaridade)
# Graduação Mestrado Primário
# 2 1 1
```



O cifrão é um caractere importante, pois permite que selecionemos um objeto dentro de outro.

Permite chamar uma variável dentro de uma base específica

```
#Selação dos dados indicando 1º a linha e 2º a coluna
Ficha_Pacientes[2,5]
#[1] "Mestrado"
Ficha_Pacientes[2,5, drop=F]
# Escolaridade
# 2 Mestrado
Ficha_Pacientes[2,c("Peso", "Idades", "Nomes")]
# Peso Idades Nomes
# 2 99 18 Gilberto
```

MANIPULAÇÃO DO DATA FRAME - EXCLUSÕES

```
#Exclusão da primeira linha  
Ficha_Pacientes[-1, ]  
#Exclusão da primeira linha e última coluna  
Ficha_Pacientes[ -1 , -7]
```



Id	Nomes	Peso	Idades	Escolaridade	Exerc_Recomend	Comidas_preferidas
2	Gilberto	99.0	18	Mestrado	Pilates	Sorvete
3	Rodrigo	62.8	44	Primário	Musculação	Milho
4	Marcela	102.0	23	Graduação	Corrida	Pão

! Lembre-se que os resultados somente estão impressos na base, mas não estão salvos

```
#Exclusões de colunas  
## deleta quant_filhos  
Ficha_Pacientes$Quant_filhos ← NULL  
  
## deleta colunas específicas,  
## mesmo pode ser feito para linhas  
Ficha_Pacientes ← Ficha_Pacientes[, c(-4, -6)]
```

! Todas essas inclusões e exclusões serão muito importante na manipulação real de dados.

MANIPULAÇÃO DO DATA FRAME –INSERÇÕES E FILTROS

```
#Inserir novos dados em toda a base  
Ficha_Pacientes$Sexo ← c("F", "M", "M", "F")  
Ficha_Pacientes$Quant_filhos ← c(4:7)
```

```
#Uma outra forma de inserção de dados  
#é utilizando uma função muito importante no rbase  
# a função cbind
```

```
Ficha_Pacientes ← cbind(Ficha_Pacientes,  
                         Prim_emprego =  
                         c("sim", "nao", "nao", "sim"))
```

O subset é uma forma de selecionar
e criar bases menores, inserindo
somente suas variáveis de interesse.
Rbase

FILTRAR USANDO O SUBSET = SUBCONJUNTO

```
#Subset:  
#Quero uma base de dados com apenas 5 variáveis  
subset(Ficha_Pacientes, select = c("Id", "Nomes", "Peso",  
                                   "Idades", "Sexo"))  
  
Base_menor ← subset(Ficha_Pacientes, select =  
                     c("Id", "Nomes", "Peso",  
                       "Idades", "Sexo"))
```

```
#Mas eu queria apenas pessoas do sexo masculino  
Base_menor ← subset(Ficha_Pacientes,  
                     Sexo == "M",  
                     select = c("Id", "Nomes", "Peso", "Idades", "Sexo"))  
  
#Mas eu tbém queria pessoas com idades maiores ou  
#igual a 20 anos  
Base_menor ← subset(Ficha_Pacientes,  
                     Idades ≥ 20,  
                     select = c("Id", "Nomes", "Peso", "Idades", "Sexo"))
```

MANIPULAÇÃO DO DATA FRAME –FILTROS E ALTERAÇÕES

```
#Filtro usando o filter
# seleciona apenas colunas numéricas
Filter(is.numeric, Ficha_Pacientes)
# Id Peso Idades
# 1 1 75.6    25
# 2 2 99.0    18
# 3 3 62.8    44
# 4 4 102.0   23

# seleciona apenas colunas de texto
Filter(is.character, Ficha_Pacientes)
#      Nomes Escolaridade Exerc_Recomend Comidas_preferidas
# 1 Ana     Graduação       Natação           Chocolate
# 2 Gilberto Mestrado       Pilates          Sorvete
# 3 Rodrigo Primário       Musculação        Milho
# 4 Marcela Graduação       Corrida          Pão
#Separando dados das colunas#####
library(stringr)

#Inserindo uma variável com 2 informações
Ficha_Pacientes$Nome_Mãe ←
c('John, Mae', 'Maude, Lebowski', 'Mia, Amy', 'Andy, James')

#Separando
str_split_fixed(Ficha_Pacientes$Nome_Mãe, ", ", 2)

#salvando na base
Ficha_Pacientes$Nomes ←
  str_split_fixed(Ficha_Pacientes$Nome_Mãe, ", ", 2)
```

MANIPULAÇÃO DO DATA FRAME –ALTERAÇÕES

Alterando valores individuais

```
#Vamos alterar a idade da Marcela de 23 para 53  
Ficha_Pacientes$Idades[4] ← 53  
#Seguindo essa mesma lógica podemos alterar o nome das variáveis  
#Ou então criar uma cópia com um nome diferente|
```

Alterando grandes conjuntos de dados

```
library(memisc)  
  
#Quero transformar a idade, fazer com que ela deixe de  
#Ser numérica e se torne categorica (faixas de idade)  
Ficha_Pacientes$Idade_Faixas ← recode(Ficha_Pacientes$Idades,  
"Jovens" ← c(18:25), "Idoso" ← 44)  
#Assim podemos alterar qualquer variável, mas é importante  
#lembiar que tipo de variável estamos alterando|
```

! Antes de usar o
recode pense no tipo
de variável que está
sendo modificada

MANIPULAÇÃO DO DATA FRAME – TIPOS DE VARIÁVEIS NO AMBIENTE DE PROGRAMAÇÃO

- No slide 14 nós vimos que existem tipos diferentes de variáveis, essa informação é essencial em várias etapas da análise dos dados, desde a construção do data frame, até a utilização de testes mais robustos estatisticamente.
- Nesse momento veremos a importância dos tipos das variáveis para análise dos dados e recodificação
- No ambiente de programação do r, os tipos de variáveis recebem nomes em inglês, os principais são:

FACTOR são um tipo de dado especial usado principalmente para representar variáveis categóricas. Quando você especifica um objeto como um fator, automaticamente o R irá atribuir uma ordem para esses dados

The diagram illustrates the classification of R variables into four main types: Character, Numeric, Integer, and Factor. Character variables are represented by a grid of names (Nomes) and values (Peso, Idades). Numeric variables are represented by a grid of numerical values (Peso, Idades). Integer variables are represented by a grid of numerical values (Idades). Factor variables are represented by a grid of categorical values (Escolaridade).

Nomes	Peso	Idades	Escolaridade
Gilberto	99.0	18	Mestrado
Marcela	102.0	23	Graduação
Ana	75.6	25	Graduação
Rodrigo	62.8	44	Primário

MANIPULAÇÃO DO DATA FRAME – TIPOS DE VARIÁVEIS NO AMBIENTE DE PROGRAMAÇÃO

```
#Podemos alterar qualquer variável, mas é importante  
#lembrar que tipo de variável estamos alterando  
#e entender como essa variável estará após as alterações
```

```
Ficha_Pacientes$Escolaridade ←  
  as.factor(Ficha_Pacientes$Escolaridade)
```

```
#Olhar a estrutura dos dados  
str(Ficha_Pacientes)
```

```
#Verificar os níveis  
Ficha_Pacientes$Escolaridade  
#Levels: Graduação Mestrado Primário
```

```
#Se eu não concordar com a ordem atribuída pelo R,  
#Se essa ordem não fizer sentido para as análises estatísticas  
#então eu posso alterar a ordem usando a função levels  
Ficha_Pacientes$Escolaridade2 ←  
  factor(Ficha_Pacientes$Escolaridade,  
levels = c("Primário", "Graduação", "Mestrado"))  
  
#as.numeric/as.character####  
#Em geral o ambiente de programação entende quais são  
#os tipos das variáveis, porém pode ser que ele entenda errado  
#ou que em razão da abertura da base de dados seja necessário  
#fazer alguns ajustes, dizendo para o ambiente o que é numérico  
#e o que é caracter, nesses casos usamos as funções as. =como
```

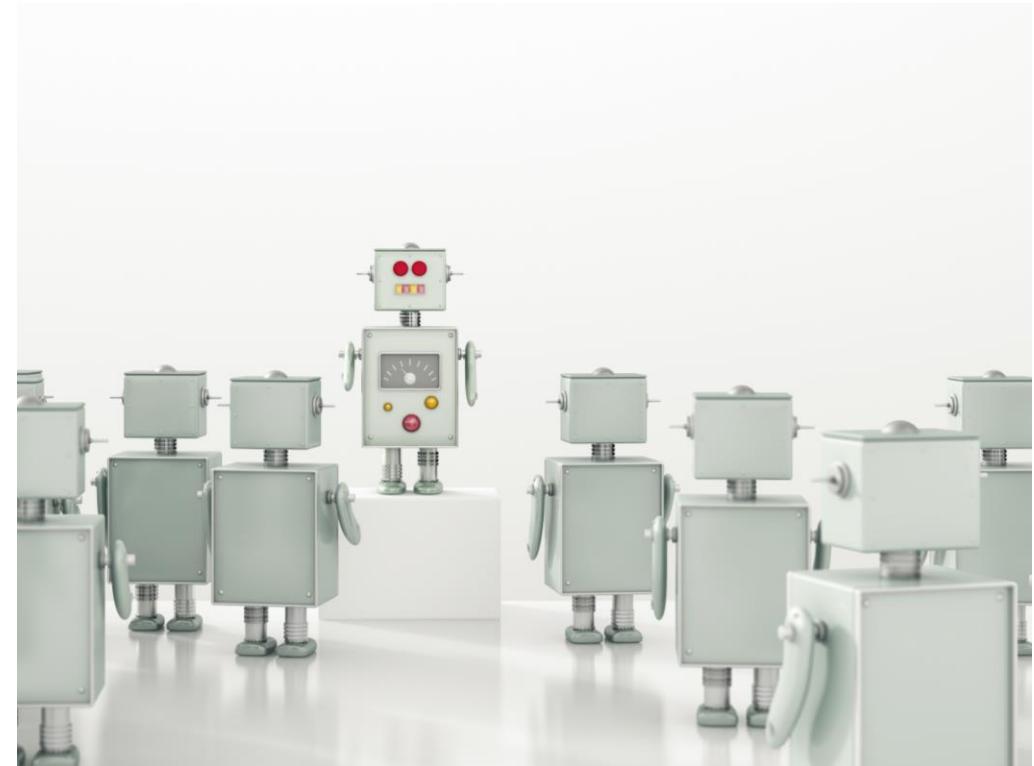


as.character
as.numeric

TÓPICOS

- 1-INTRODUÇÃO AO R
- 2-ANÁLISE EXPLORATÓRIA E MANIPULAÇÃO DOS DADOS
- 3-SALVAMENTO E ABERTURA
- 4-ANÁLISES DESCRIPTIVAS
- 5-APRESENTAÇÃO GRÁFICA
- 6- PROCESSAMENTO DE DADOS
- 7- ANÁLISES INFERENCIAIS

- **CSV**
- **XLSX**
- **STATA**
- **SPP**
- **Abertura clicável**





SALVAMENTOS E ABERTURA DA BASE DE DADOS

SALVAMENTO E ABERTURA DA BASE DE DADOS

Na seção inicial vimos uma forma de salvamento dos dados e abertura da base. Nessa seção veremos diversas outras formas.

A forma vista para salvamento é mais comum dentro do ambiente do R, “.RData”

A abertura da base ocorreu através da função “view”, porém essa função somente irá funcionar se a base de dados já estiver carregada dentro do R, então ela serve somente para ver o que já está lá. Portanto, agora veremos formas de abrir bases de dados salvas anteriormente no computador

Iº CSV

```
#CSV salvamento#####
write.csv2(Ficha_Pacientes, file = "Ficha2.csv",
           quote = F, #Dividir por ponto e vírgulas
           row.names = F,#remover a primeira coluna com id do sistema
           fileEncoding = "latin1")#função pra definir a linguagem
#CSV abertura#####
#Precisaremos do pacote readr
library(readr)
#abertura
Ficha2 ← read_delim("Ficha2.csv", delim = ";",
                      locale = locale(encoding = "Latin1"))
```



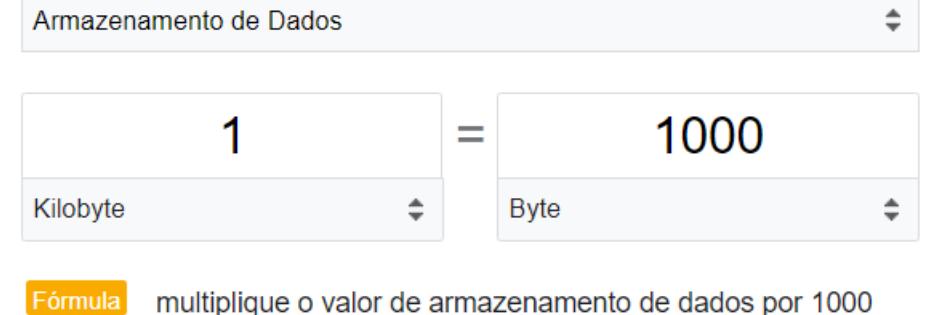
Lembre-se de sempre definir o seu diretório de trabalho antes de tentar abrir ou salvar qualquer base.

SALVAMENTO E ABERTURA DA BASE DE DADOS

2º XLSX

```
#Precisaremos do pacote  
library(writexl)  
  
#salvando  
writexl::write_xlsx(Ficha_Pacientes, path = "Ficha3.xlsx")  
  
#Xlsx Abertura####  
#Pacote:  
library(readxl)  
  
Ficha3 ← read_excel("Ficha3.xlsx")  
  
View(Ficha3)
```

! Atenção para o tamanho dos documentos



	Name	Size
...	..	428 B
	[REDACTED]	12.6 KB
	Antigo	37.8 MB
	Conselho	
	Documentos	
	Ficha_Pacientes.RData	379 B
	Ficha.csv	316 B
	Ficha2.csv	292 B
	Ficha3.xlsx	5.2 KB

SALVAMENTO E ABERTURA DA BASE DE DADOS

3º Abertura em stata e

Brazil 2014 Elizabeth Zechmeister

Data Files and Notes

Brazil 2014 Tech Info	Brazil 2014 Stata	Brazil 2014 SPSS
-----------------------	-------------------	------------------

Brazil 2014 Questionnaire - Portuguese

To download the files, click on the links shown or right-click on the file name and save it. You will not be able to open STATA or SPSS files unless your computer has the corresponding software available. The Questionnaires and Technical Information documents are saved as .pdf files which can be opened with pdf reader software available for free online such as Adobe Acrobat Reader get.adobe.com/reader/ and Foxit Reader www.foxitsoftware.com. Stata files contain labels in both English and Spanish. To change the language, use the commands: label language en OR label language es.

Change Log:329664568Brazil AmericasBarometer 2014 Change Log v3.0_W.pdf

Site do lapop:

<http://datasets.americasbarometer.org/database/index.php?freeUser=true>

Atualmente os dados do LAPOP
são disponibilizados somente em
Stata, porém até 2014 eram
disponibilizados em:

- Stata
- Spss
- e Tech Info

Brincadeirinha! Não vá
confundir as informações
técnicas com formato da
base rsrs



SALVAMENTO E ABERTURA DA BASE DE DADOS



Baixe em seu diretório de trabalho todo aquele material: base de dados, questionário e informações técnicas.



Para quem comumente utiliza SPSS, sabe que o formato de salvamento é sav, já do STATA é dta

Name	Size	Modified
...		
265338482LAOPBra14-v15.2.5.1-Por-140316_W.pdf	1.1 MB	Jan 25, 2023, 5:50 PM
379249517Brazil_Tech_Info_2014_W_112114.pdf	296.5 KB	Jan 25, 2023, 5:49 PM
636339374Brazil LAPOP AmericasBarometer 2014 v3.0_W.dta	707.2 KB	Jan 25, 2023, 5:49 PM
863896541Brazil LAPOP AmericasBarometer 2014 Espanol v3.0_W.sav	762.4 KB	Jan 25, 2023, 5:49 PM

SALVAMENTO E ABERTURA DA BASE DE DADOS

💡 Vamos abrir e salvar bases em **sav**

```
#Abertura de dados reais#####
#Abertura em SAV#####

#Pacote:
library(haven)

#Como o nome da base é bastante grande, vamos chamar por um
#nome mais resumido
Brasil2014 ← read_sav("LAPOP 2014/863896541Brazil_LAPOP_AmericasBarometer 2014_Espanol_v3.0_W.sav")
View(X863896541Brazil_LAPOP_AmericasBarometer_2014_Espanol_v3_0_W)

#Após realizar as alterações na base
#se quiser salvar no mesmo formato, use:
write_sav(data = Brasil2014, path = "Brasil2014.sav")
```

💡 Em dta o processo é semelhante veja:

```
#Abertura em DTA#####
#Use o mesmo pacote haven
Brasil2014dta ←
  read_dta("LAPOP 2014/636339374Brazil_LAPOP_AmericasBarometer 2014_v3.0_W.dta")

#Após realizar as alterações na base
#se quiser salvar no mesmo formato, use:
write_dta(data = Brasil2014dta, path = "Brasil2014dta.sav")
```

Em seu diretório de trabalho as pastas devem aparecer dessa forma:

- Brasil2014.sav
- Brasil2014dta.sav

SALVAMENTO E ABERTURA DA BASE DE DADOS



Até aqui vimos algumas formas de abertura de base de dados, essas são as mais utilizadas, contudo é importante que você esteja ciente o **ambiente de programação do R consegue ler bases nos mais diferentes formatos**, por isso agora vamos aprender uma forma manual de abrir. Isso poderá lhe auxiliar com outros formatos.

Files Plots Packages Help Viewer Presentation
New Folder New Blank File Delete Rename More
C: > Users > nayar > OneDrive > 8. AMBIENTE DE PROGRAMAÇÃO R > 1. CURSO 2023

	Name	Size	Modified
Brasil2014.sav	530.8 KB	Jan 25, 2023, 6:23 PM	
Brasil2014dta.sav	2.7 MB	Jan 25, 2023, 6:28 PM	
CURSO R - 2023.pptx	38 MB	Jan 25, 2023, 6:34 PM	
DISCIPLINA LUCAS			
Ficha_Pacientes.RData	379 B	Jan 15, 2023, 9:20 PM	
Ficha.csv	316 B	Jan 25, 2023, 4:58 PM	
Ficha2.csv	292 B	Jan 25, 2023, 5:01 PM	
Ficha3.xlsx	5.2 KB	Jan 25, 2023, 5:30 PM	
Ficha4.csv	1.2 KB	Jan 25, 2023, 6:19 PM	
GRÁFICOS			
LAPORATÓRIOS			
Script Curso 2023.R	18.2 KB	Jan 25, 2023, 6:21 PM	

Clique com o botão direito na base e peça para “Import Dataset...”



Alguns formatos como o csv você poderá configurar a base antes de abrir. Para isso essa janela de comunicação abrirá:

Import Text Data
File/URL: C:/Users/nayar/OneDrive/8. AMBIENTE DE PROGRAMAÇÃO R/1. CURSO 2023/Ficha2.csv
Data Preview:
Id;Nomes;Peso;iades;Escolaridade;Exerc_Recomend;Comidas_preferidas;Escolaridade2
(character)
1:Ana;75;6:25;Gradua♦♦oNata♦♦oChocolate;Gradua♦♦o
2:Gilberto;99;18;Mestrado;Pilates;Sorvete;Mestrado
3:Rodrigo;62;8:44;Prim♦rio;Muscula♦♦o;Milho;Prim♦rio
4:Marcela;102;23;Gradua♦♦oCorrida;P♦o;Gradua♦♦o
Previewing first 50 entries. 2 parsing errors.
Import Options:
Name: Ficha2
Skip: 0
First Row as Names
Delimiter: Comma
Escape: None
Trim Spaces
Quotes: Default
Comment: Default
Open Data Viewer
Locale: Configure...
NA: Default
Code Preview:
library(readr)
Ficha2 <- read_csv("Ficha2.csv")
View(Ficha2)
Import Cancel

SALVAMENTO E ABERTURA DA BASE DE DADOS

Import Text Data

File/URL:

C:/Users/nayar/OneDrive/8. AMBIENTE DE PROGRAMAÇÃO R/1. CURSO 2023/Ficha2.csv Update

Data Preview:

	Id;Nomes;Idades;Escolaridade;Exerc_Recomend;Comidas_preferidas;Escolaridade2 (character)
1	Ana;75,6;25;Graduado;Natação;Chocolate;Graduado
2	Gilberto;99;18;Mestrado;Pilates;Sorvete;Mestrado
3	Rodrigo;62,8;44;Primário;Musculação;Milho;Primário
4	Marcela;102;23;Graduado;Corrida;Pão;Graduado

Iº defina como será a divisão dos dados aqui: Delimiter: Semicolon

Escolha ponto e vírgula

Previewing first 50 entries. 2 parsing errors.

Import Options:

Name: Ficha2	<input checked="" type="checkbox"/> First Row as Names	Delimiter: Comma	Escape: None
Skip: 0	<input checked="" type="checkbox"/> Trim Spaces	Quotes: Default	Comment: Default
	<input checked="" type="checkbox"/> Open Data Viewer	Locale: Configure...	NA: Default

Code Preview:

```
library(readr)
Ficha2 ← read_csv("Ficha2.csv")
View(Ficha2)
```

② Reading rectangular data using readr Import Cancel

SALVAMENTO E ABERTURA DA BASE DE DADOS

Import Text Data

File/URL:
C:/Users/nayar/OneDrive/8. AMBIENTE DE PROGRAMAÇÃO R/1. CURSO 2023/Ficha2.csv

Data Preview:

Id (double)	Nomes (character)	Peso (double)	Idades (double)	Escolaridade (character)	Exerc_Recomend (character)	Corrida (character)
1	Ana	756	25	Graduação	Natação	Choque
2	Gilberto	99	18	Mestrado	Pilates	Sorvete
3	Rodrigo	628	44	Primário	Musculação	Milho
4	Marcela	102	23	Graduação	Corrida	Pão

Previewing first 50 entries.

Import Options:

Name: Ficha2 First Row as Names: Delimiter: Semicolon
Skip: 0 Trim Spaces: Quotes: Default Escape: None
Open Data Viewer: Locale: Configure... Comment: Default
NA: Default

Configure Locale

Date Name: en Encoding: UTF-8
Date Format: %AD Time Format: UTC
Decimal Mark: . Grouping Mark:
Time Zone: UTC

Locales in reader

Config

Code Preview:

```
library(readr)
Ficha2 ← read_delim("Ficha2.csv", delim = ";",
                      escape_double = FALSE, trim_ws = TRUE)
View(Ficha2)
```

Import Cancel

2º Configure a base para a linguagem que está sendo utilizada, no caso Latin1, isso fará com que a base não fique com caracteres estranhos no lugares dos acentos. Depois de escrever Latin1 dê ok.

Encoding Identifier

Please enter an encoding identifier. For a list of valid encodings run iconvlist().

Latin1

OK Cancel

SALVAMENTO E ABERTURA DA BASE DE DADOS

Import Text Data

File/URL:
C:/Users/nayar/OneDrive/8. AMBIENTE DE PROGRAMAÇÃO R/1. CURSO 2023/Ficha2.csv Update

Data Preview:

Id (double)	Nomes (character)	Peso (double)	Idades (double)	Escolaridade (character)	Exerc_Recomend (character)	Comidas_preferidas (character)	Escolaridade2 (character)
1	Ana	756	25	Graduação	Natação	Chocolate	Graduação
2	Gilberto	99	18	Mestrado	Pilates	Sorvete	Mestrado
3	Rodrigo	628	44	Primário	Musculação	Milho	Primário
4	Marcela	102	23	Graduação	Corrida	Pão	Graduação

Previewing first 50 entries.

Import Options:

Name: Ficha2 First Row as Names Delimiter: Semicolon Escape: None
Skip: 0 Trim Spaces Quotes: Default Comment: Default
 Open Data Viewer Locale: Configure... NA: Default

Code Preview:

```
library(readr)
Ficha2 ← read_delim("Ficha2.csv", delim = ";",
                      escape_double = FALSE, locale = locale(encoding =
"Latin1"),
                      trim_ws = TRUE)
```

? Reading rectangular data using readr Import Cancel



ANÁLISE DESCRIPTIVA

TÓPICOS

1-INTRODUÇÃO AO R

2-ANÁLISE EXPLORATÓRIA E
MANIPULAÇÃO DOS DADOS

3-SALVAMENTO E ABERTURA

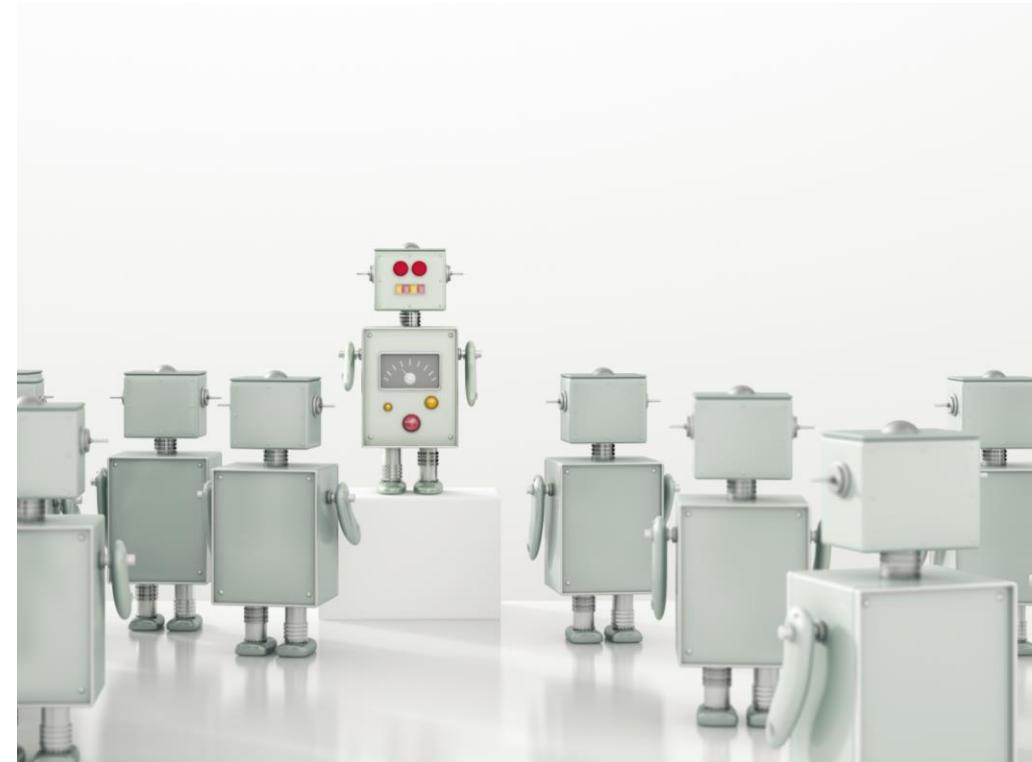
4-ANÁLISES DESCRIPTIVAS

5-APRESENTAÇÃO GRÁFICA

6- PROCESSAMENTO DE DADOS

7- ANÁLISES INFERENCIAIS

- **Introdução a análise descritiva**
- **Organização de uma base de dados real: seleção dos dados; recategorização; análise das medidas centrais (média, mediana e moda);**
- **Análises cruzadas**
- **Análise de frequências absolutas e relativas**
- **Introdução a análise gráfica**



ANÁLISE DESCRIPTIVA

Base de dados

- Para realizar a análise descritiva vamos utilizar a base de dados do LAPOP 2014 que baixamos anteriormente do site do LAPOP
- Quem não baixou, basta entrar nesse link: <http://datasets.americasbarometer.org/database/index.php?freeUser=true>
- Escrever “Brazil” e baixar a base em stata ou sav do ano de 2014

O que é uma análise descritiva.

Como o nome já diz, é um processo de descrição dos dados ou então é a fase inicial de um “processo de estudo dos dados coletados. Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.”*

Após baixar a base de dados e o questionário:

- Vamos olhar a base de dados no ambiente de programação do R, abrindo através do files ou usando a linha de comando que aprendemos na seção anterior

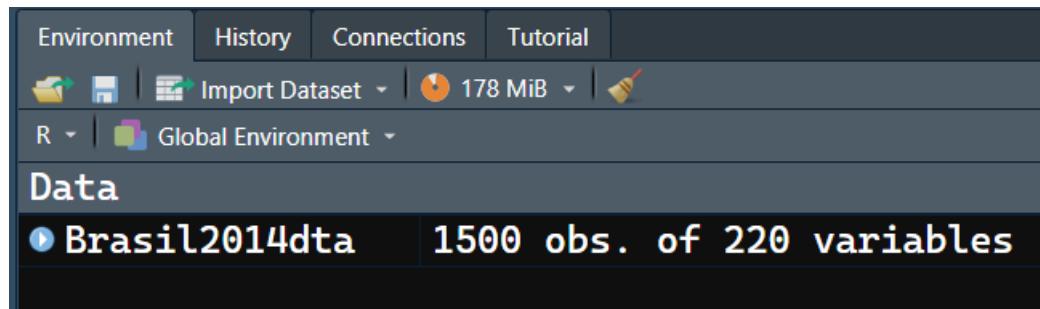
```
#Análise descritiva#####
library(haven)

Brasil2014dta ←
  read_dta("636339374Brazil LAPOP AmericasBarometer 2014 v3.0_W.dta")
```

Para mais informações acessar Reis, E.A., Reis I.A. (2002) Análise Descritiva de Dados. Relatório Técnico do Departamento de Estatística da UFMG. Disponível em: www.est.ufmg.br/portal/arquivos/rts/rte0202.pdf

ANÁLISE DESCRIPTIVA

- Após carregar a base de dados no ambiente do R suas informações devem aparecer em que lugar?
- Isso mesmo, se você respondeu no Environment você acertou !



Temos uma base de dados com 1500 observações e 220 variáveis.

O número de observações está ótimo. Mas será que precisaremos analisar todas essas 220 variáveis em nossa pesquisa acadêmica? Teremos que fazer gráficos e testes usando todo esse material? Acho que não, então que tal selecionar o que será útil e trabalhar com uma base menor?

- Pode analisar dentro do R a base de dados, através da função View (ou clicando na base), mas usando o questionário ou um dicionário de códigos fica mais fácil, assim saberemos o que cada código significa e quais são suas categorias.
`View(Brasil2014dta)`
- Selecionei do questionário todas as questões que acredito serem suficientes para as análises nesse momento. Agora vamos fazer o mesmo procedimento no R.

ANÁLISE DESCRIPTIVA

```
#Seleção das variáveis#####
colnames(Brasil2014dta)
Brasil2014dta$q11n

# ls3 = Satisfação com a vida (SatVida) - inverter a escala
# q2 = Idade
# q1 = Sexo
# q10new = Renda familiar (RendFam)
# q10d = Percepção sobre a renda (PercRend) - Inverter a escala
# q11n = Estado civil (EstCiv)
# q12c = Quantidade de moradores no mesmo domicílio (QuantMor)
# ocup4a = Ocupação (Ocupacao)
# q12= Quantos filhos tem (Filhos)

#Após definir todas as variáveis que iremos querer basta
#utilizar a função subset que aprendemos na seção anterior

Bra2014Menor ← subset(Brasil2014dta, select =
                         c("ls3", "q2", "q1",
                           "q10new", "q10d", "q10d",
                           "q11n", "q12c", "ocup4a",
                           "q12"))
```

ANÁLISE DESCRIPTIVA

- Agora veja como está nossa base de dados:

The screenshot shows the RStudio environment. The top menu bar includes 'Environment', 'History', 'Connections', and 'Tutorial'. Below the menu is a toolbar with icons for file operations like 'Import Dataset' (188 MiB), a search bar, and a 'List' dropdown. The main area is titled 'Data' and shows two entries: 'Bra2014Menor' (1500 obs. of 9 variables) and 'Brasil2014d...' (1500 obs. of 220 variables). The 'Global Environment' tab is selected.

- **Mas porque a base de dados com 220 variáveis ainda aparece?**

Porque criamos uma outra base de dados, através de uma cópia de algumas variáveis existentes na base original.

Fique atento a esse procedimento.

Cuidado para NÃO substituir a base original e, perder dados que você poderá precisar posteriormente.



ANÁLISE DESCRIPTIVA

- Antes de iniciar as análises, temos que entender como estão organizadas as variáveis, fazer as recodificações e reorganizações, isso representa uma grande parte do trabalho de análise.
- Para essa etapa usaremos as funções **table** e **summary** que irão nos auxiliar mostrando o descritivo de cada variável
- Depois usaremos a função **recode** do pacote **memisc** para recategorizar as variáveis.

Agora a base está organizada, mas notem que ao longo de processo parte da análise descritiva já estava sendo feita, através, por exemplo:

- Distribuição da frequência, é comum querer entender e comparar a quantidade total de casos de cada categoria, por exemplo: i) total de homens em relação ao total de mulheres entrevistados i) total de pessoas casadas em relação a solteira, iii) total de pessoas que se sente satisfeita com salários em relação aos insatisfeitos.

Além das frequências absolutas e percentuais, as medidas centrais são também muito utilizadas para auxiliar nas descrições. As mais comuns são:

- Moda: representa o valor/categoria mais comum em uma distribuição

- Mediana: representa a medida central ou então a divisão das medidas centrais (quando são duas) $M = \frac{N+1}{2}$ **N= número de observações**

- Média: representa o valor médio dos casos $\bar{x} = \frac{\sum x}{N}$ **\bar{x} = soma de todas as observações/pelo N**

ANÁLISE DESCRIPTIVA

💡 Tanto a média, quanto a mediada nós já vimos que é possível obter através da função summary do rbase

```
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
0.000 0.000 2.000 1.888 3.000 16.000 2
```

Em média as famílias possuem
2 filhos
A mediana (medida central) é 2

💡 Já a moda é possível obter através da função table para variáveis categóricas e também para as numéricas, o problema é que você tem olhar os dados para encontrar a moda. *É um problema se a quantidade de dados foi

Número de filhos	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Totais	453	290	319	214	83	50	35	23	11	9	1	4	2	1	1	1	1

Ocupação	Emprego_Rem	Emprego_N_Rem	Estudante	Aposentado	Nao_Empregado
	812	200	125	180	183

💡 Vamos entender outros meios de obter essas medidas

```
#Funções específicas para média e mediana  
mean(Bra2014Menor$Filhos, na.rm = T)  
median(Bra2014Menor$Filhos, na.rm = T)
```

```
> mean(Bra2014Menor$Filhos, na.rm = T)  
[1] 1.88785  
> median(Bra2014Menor$Filhos, na.rm = T)  
[1] 2
```

ANÁLISE DESCRIPTIVA

```
506 #Para o cálculo da moda:http://www.dma.uem.br/kit/outros-arquivos/moda.pdf
507 #Explicações do help
508 moda <- function(x) {
509   modal <- unique(x)
510   modal[which.max(tabulate(match(x, modal)))]
511 }
512
513 #unique returns a vector or data frame like x
514 #but with duplicate elements/rows removed.
515
516 #which.max: Determines the location, i.e.,
517 #index of the (first) minimum or maximum of
518 #a numeric (or logical) vector.
519
520 #tabulate:counts the number of times each integer occurs in it.
521
522 #match returns a vector
523 #of the positions of its first argument in its second.
524
525 # Agora podemos rodar a função moda:
526 moda(Bra2014Menor$Filhos)
527 #[1] 0
```

ANÁLISE DESCRIPTIVA

Até esse momento aprendemos:

- 1- Organizar uma base de dados**
- 2- Observar medidas de frequência absolutas**
- 3- Observar medidas de tendência central**

Isso não significa que esgotamos todo o assunto desses tópicos acima, pelo contrário, existem uma diversidade de funções que podem ser utilizadas para organizar e observar uma base de dados.

Muitas dessas questões vocês irão se deparar em situações práticas, por isso é importante entender as lógicas do que estamos fazendo, com esse conhecimento você será capaz de buscar soluções para outras questões.

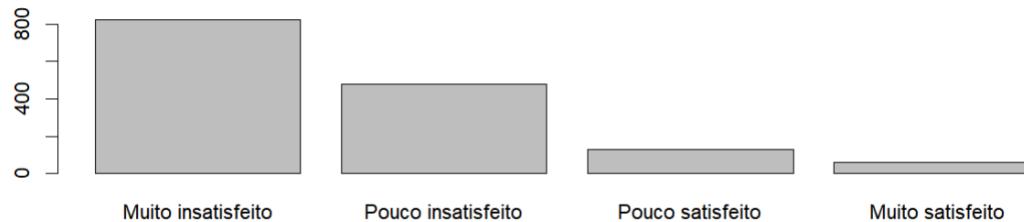
Agora veremos mais algumas formas de analisar as variáveis, especialmente envolvendo mais de uma questão. Isso nos auxiliará no futuro, quando chegarmos as análise bivariadas

ANÁLISE DESCRIPTIVA

Frequência absoluta e percentual

```
#Frequência  
#Pacote  
library(descr)  
freq(Bra2014Menor$SatVida)
```

Bra2014Menor\$SatVida	Frequência	Percentual
Muito insatisfeito	825	55.0000
Pouco insatisfeito	480	32.0000
Pouco satisfeito	131	8.7333
Muito satisfeito	62	4.1333
NA's	2	0.1333
Total	1500	100.0000



Com a função freq. além do percentual temos na saída dos plots um gráfico que é de ? _____

Só não teremos na saída um gráfico, se inserirmos uma função plot=F

Esse é o primeiro contato com os gráficos, ainda veremos alguns nesse tópico, porém não apresentaremos nenhuma elaboração gráfica.

O tópico seguinte será somente sobre GGPLOT2

ANÁLISE DESCRIPTIVA

Análise cruzada - table

```
541 table(Bra2014Menor$SatVida, Bra2014Menor$Sexo)
542 #                                     Homem Mulher
543 # Muito insatisfeito    433     392
544 # Pouco insatisfeito    236     244
545 # Pouco satisfeito      51      80
546 # Muito satisfeito      28      34
547
548
549 #Vamos salvar o table que criamos dentro de um
550 #objeto e depois vamos chamar ele usando
551 #Uma outra função.
552 Obj ← table(Bra2014Menor$SatVida, Bra2014Menor$Sexo)
553
554 #Vamos utilizar a função: prop.table()
555 #Para apresentar os valores percentuais, contudo
556 #Teremos que multiplicar por 100, por isso salvamos
557 #tudo dentro de um novo objeto
558 Obj2 ← prop.table(Obj, margin = 1)
559 #Apresenta o percentual na linha
560 Obj2*100
561
562 Obj3 ← prop.table(Obj, margin = 2 )
563 #Apresenta o percentual na coluna
564 Obj3 * 100
```

	Homem	Mulher
Muito insatisfeito	52	48
Pouco insatisfeito	49	51
Pouco satisfeito	39	61
Muito satisfeito	45	55

	Homem	Mulher
Muito insatisfeito	58	52
Pouco insatisfeito	32	33
Pouco satisfeito	7	11
Muito satisfeito	4	5

ANÁLISE DESCRIPTIVA

```
570 ##  
571 plot(Bra2014Menor$Idade)  
572 plot(Bra2014Menor$Ocupacao)  
573 #  
574 hist(Bra2014Menor$Idade)
```

Gráfico de dispersão

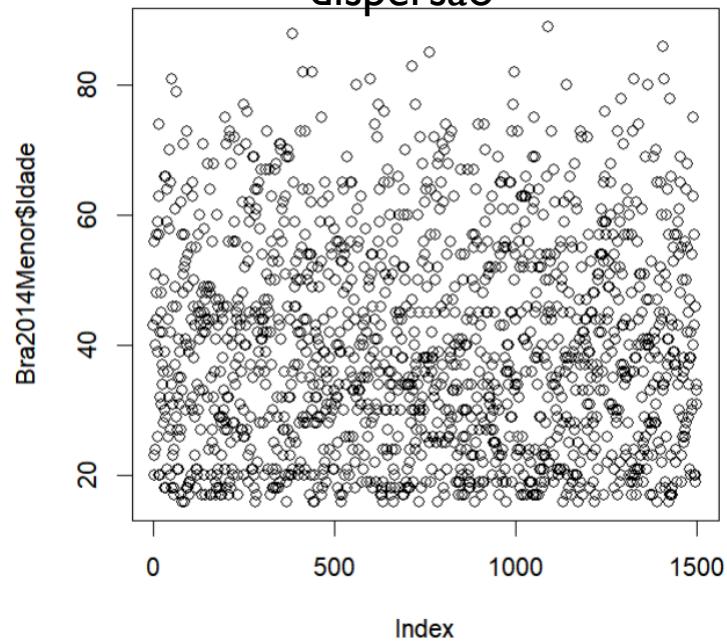
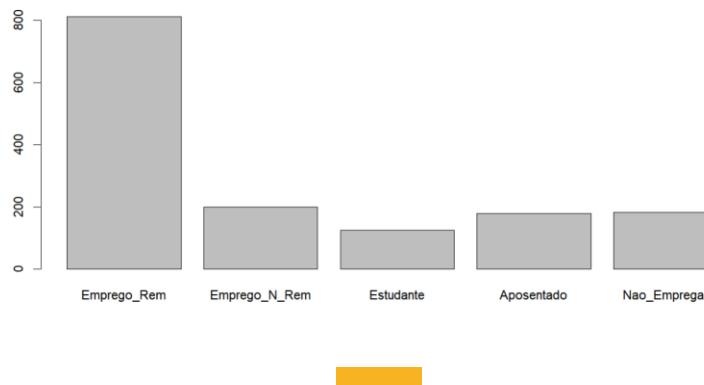
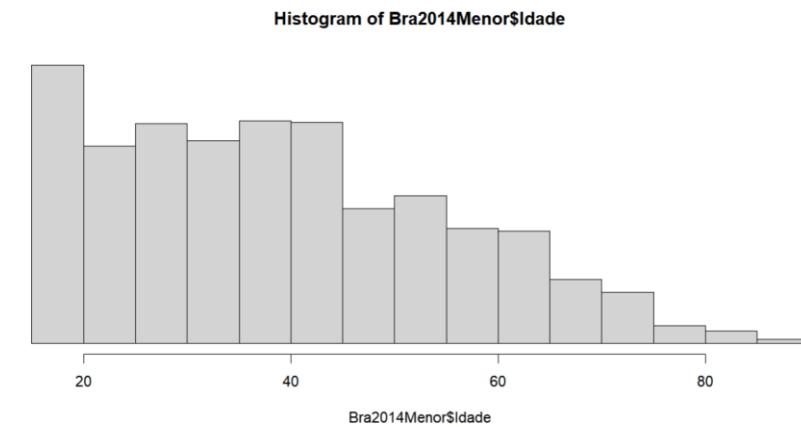


Gráfico de barras

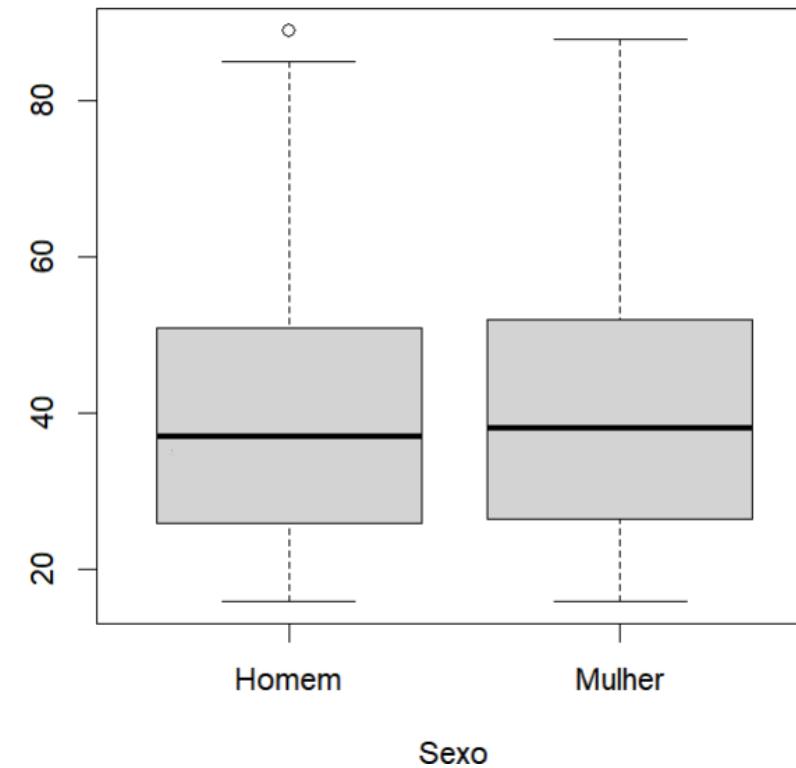
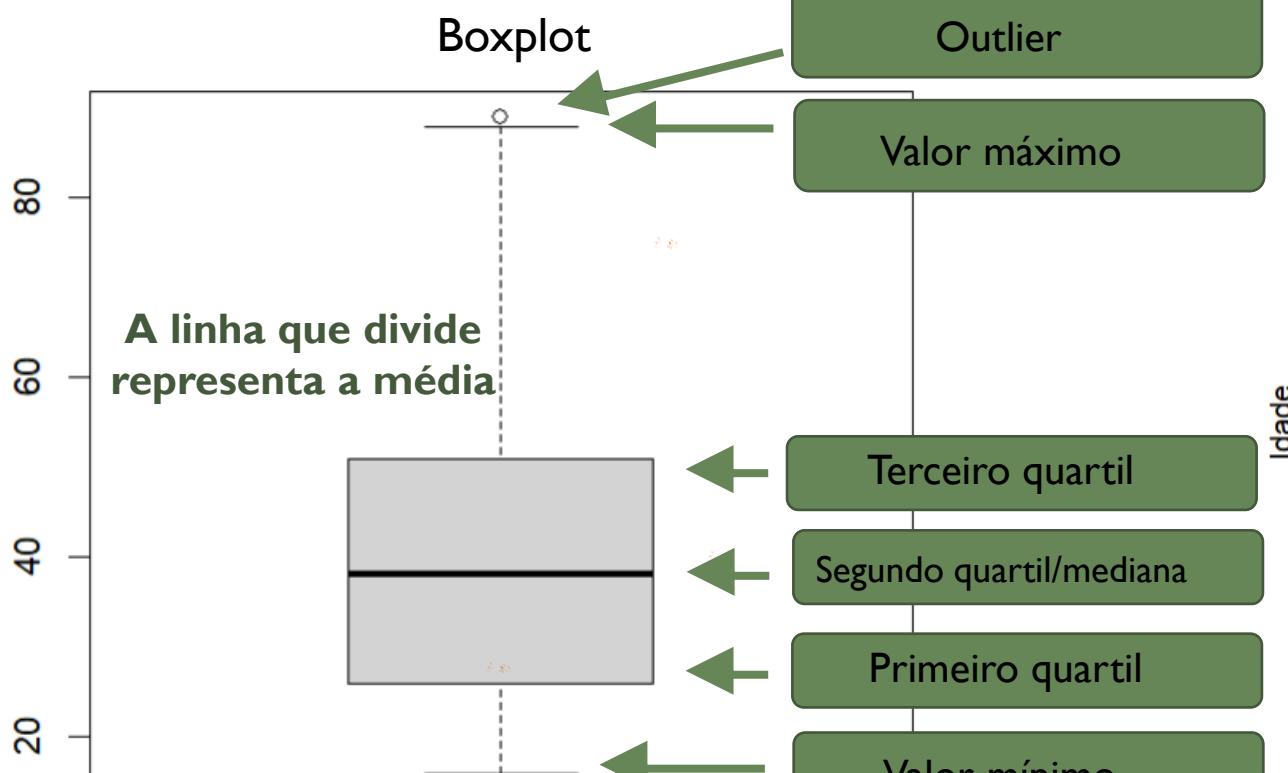
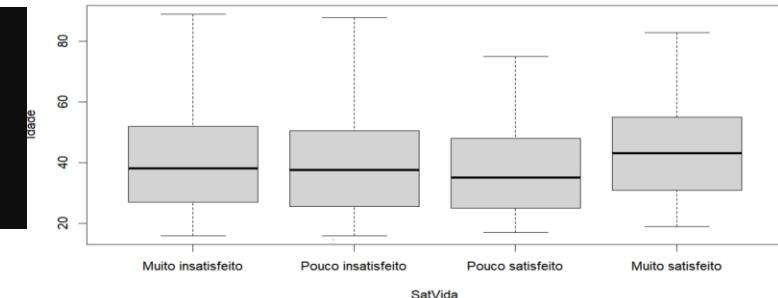


Histograma



ANÁLISE DESCRIPTIVA

```
575 boxplot(Bra2014Menor$Idade)
576 #
577 boxplot(Idade ~ Sexo, data= Bra2014Menor)
578 #
579 boxplot(Idade ~ SatVida, data= Bra2014Menor)
580
```



Nós vimos alguns tipos de gráficos, agora veremos esses mesmos gráficos de maneiras mais elaboradas, bem como, outros tipos



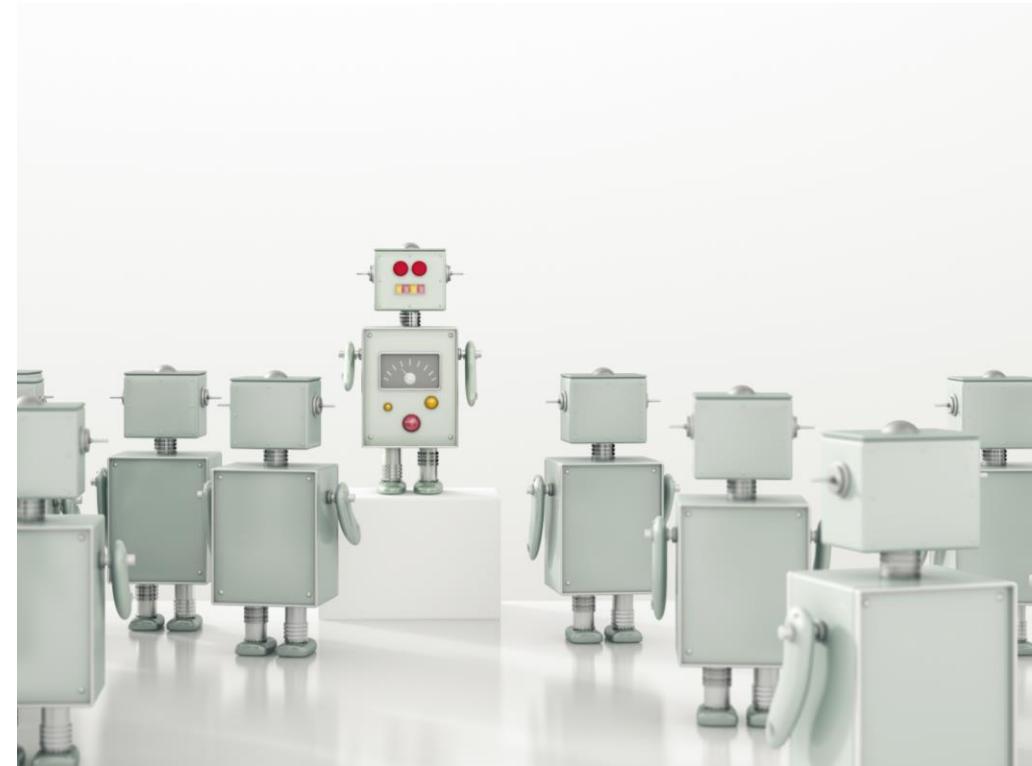
APRESENTAÇÃO GRÁFICA

**Nos slides serão apresentadas as versões
dos gráficos.
As linhas de construção estão nos scripts**

TÓPICOS

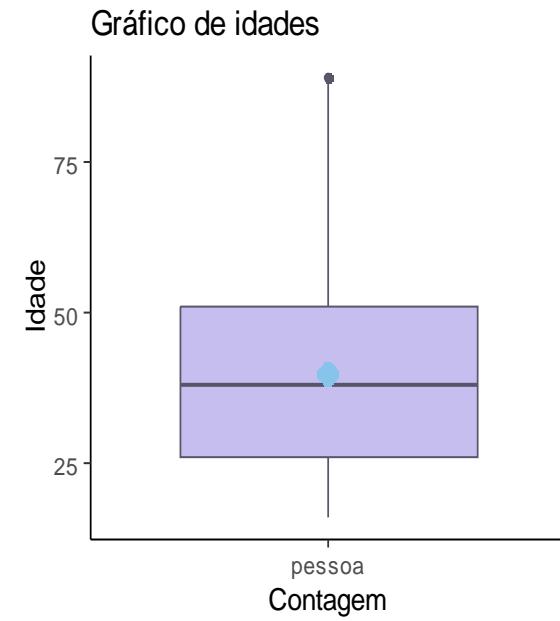
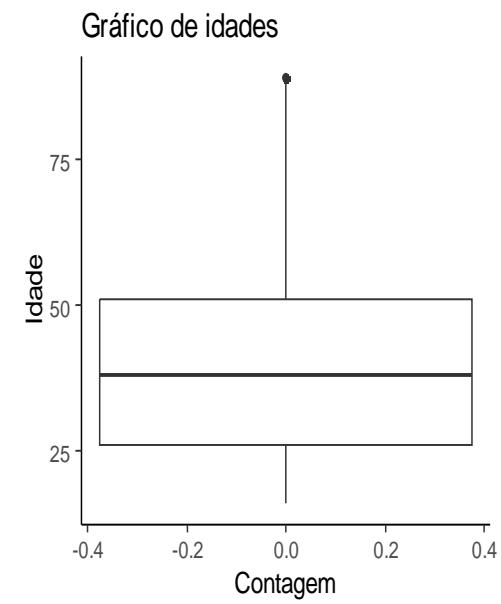
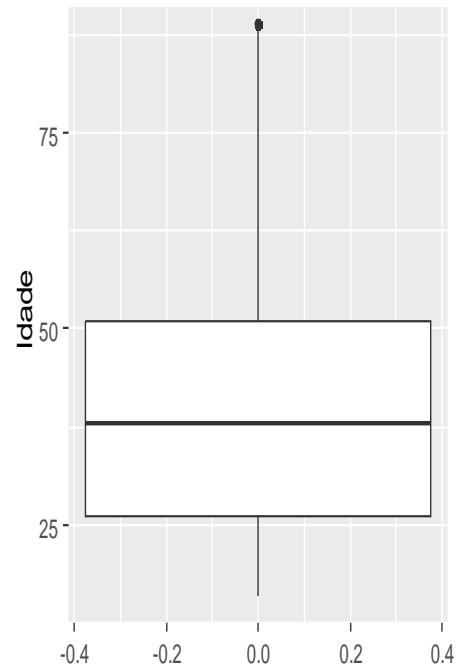
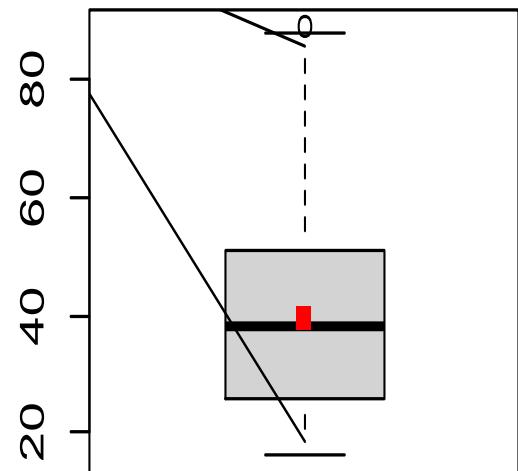
- 1-INTRODUÇÃO AO R**
- 2-ANÁLISE EXPLORATÓRIA E MANIPULAÇÃO DOS DADOS**
- 3-SALVAMENTO E ABERTURA**
- 4-ANÁLISES DESCRIPTIVAS**
- 5-APRESENTAÇÃO GRÁFICA**
- 6- PROCESSAMENTO DE DADOS**
- 7- ANÁLISES INFERENCIAIS**

- 1-Boxplot**
- 2-Histograma**
- 3-Gráfico de barras**
- 4-Gráfico de barras empilhadas**
- 5-Gráfico de pizza e seção**
- 6-Gráfico de dispersão**
- 7-Gráfico de linhas**



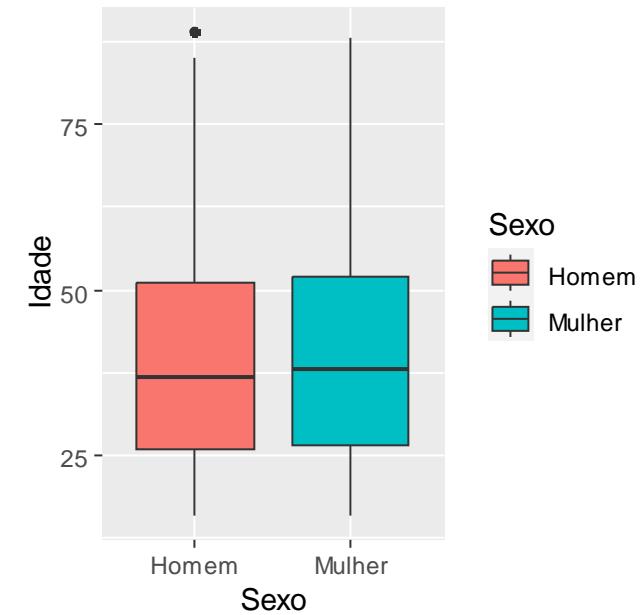
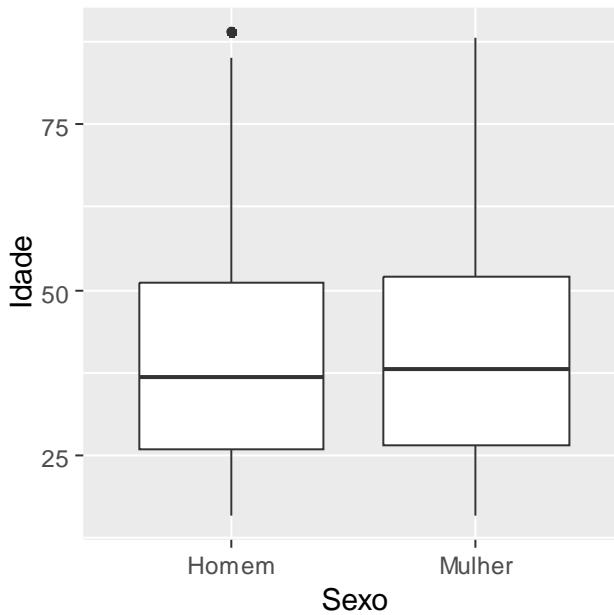
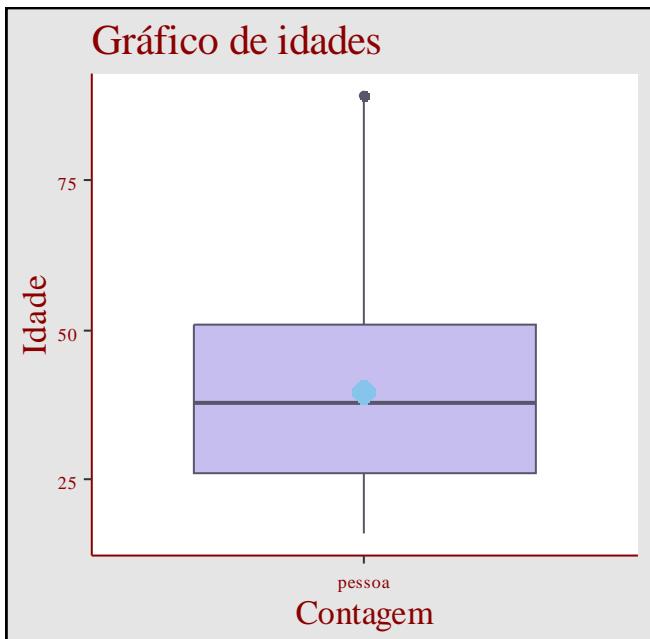
ANÁLISE DESCRIPTIVA

Boxplot



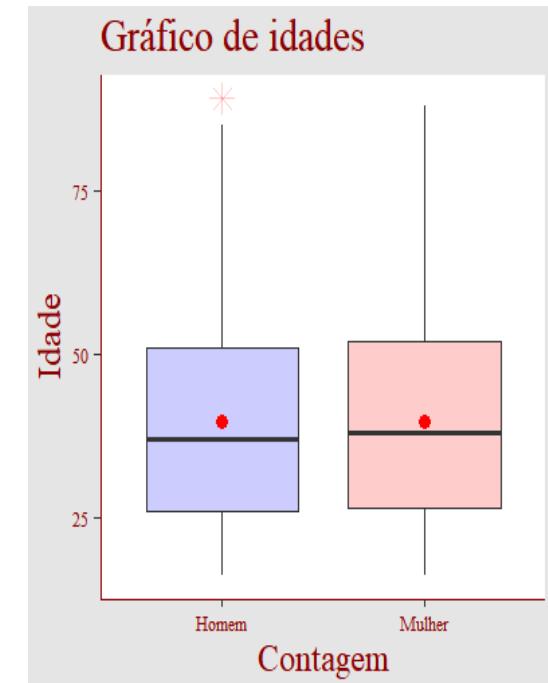
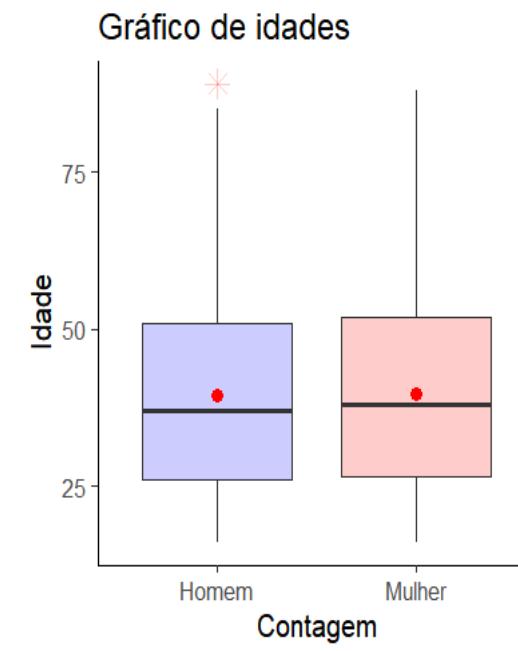
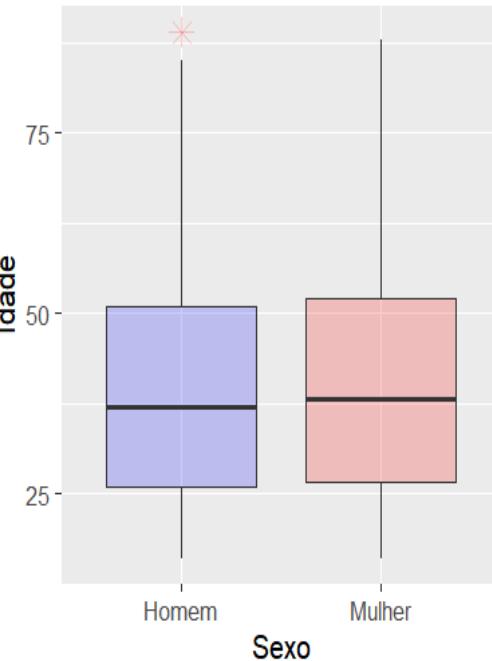
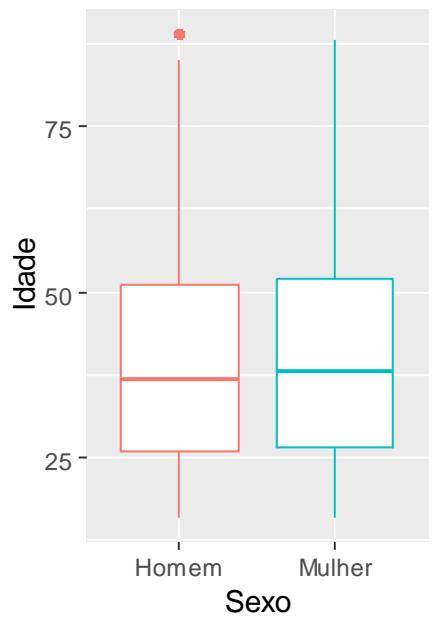
ANÁLISE DESCRIPTIVA

Boxplot



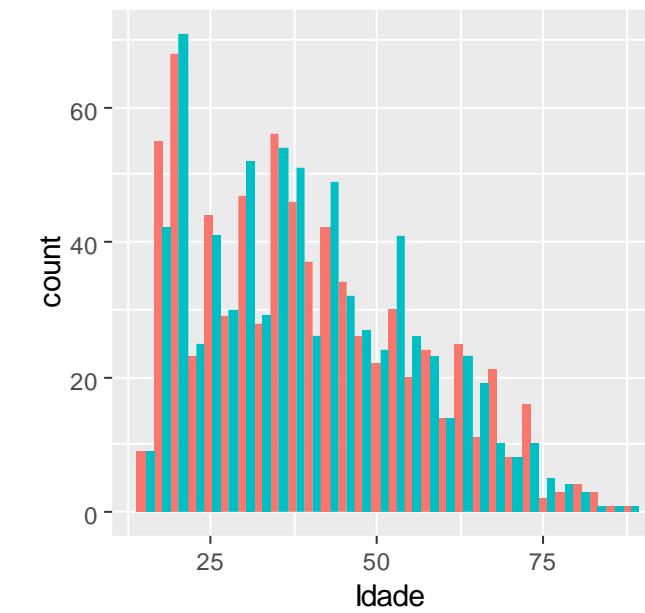
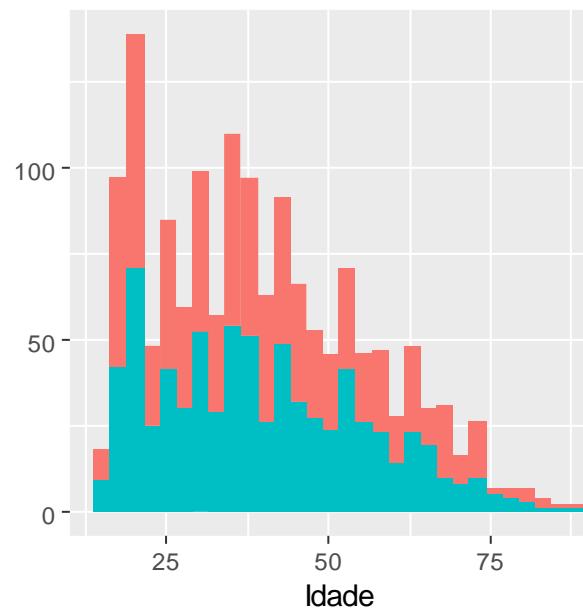
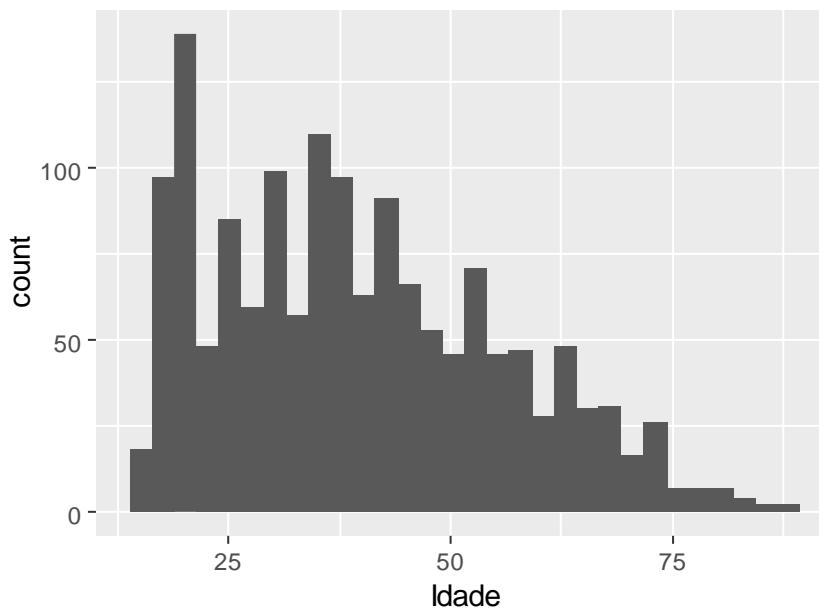
ANÁLISE DESCRIPTIVA

Boxplot



ANÁLISE DESCRIPTIVA

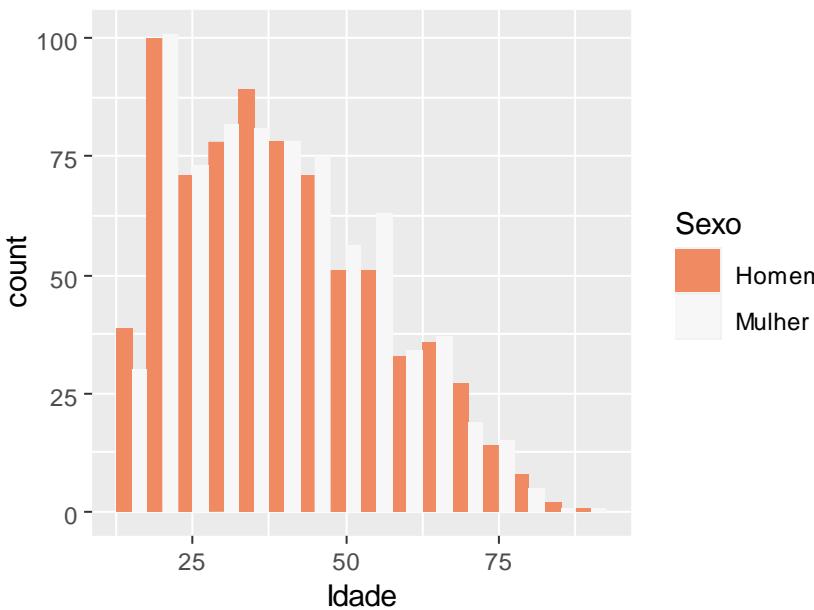
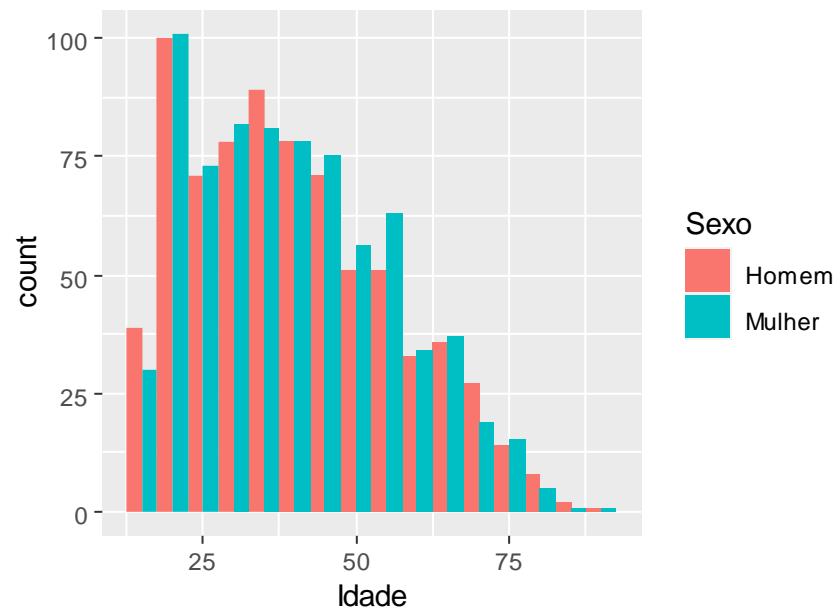
Histograma



Sexo
Homem
Mulher

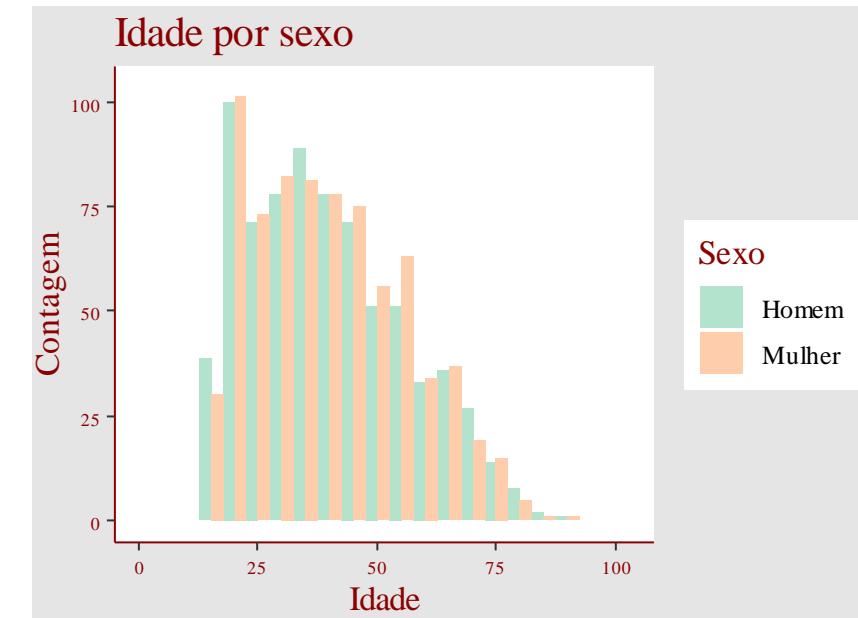
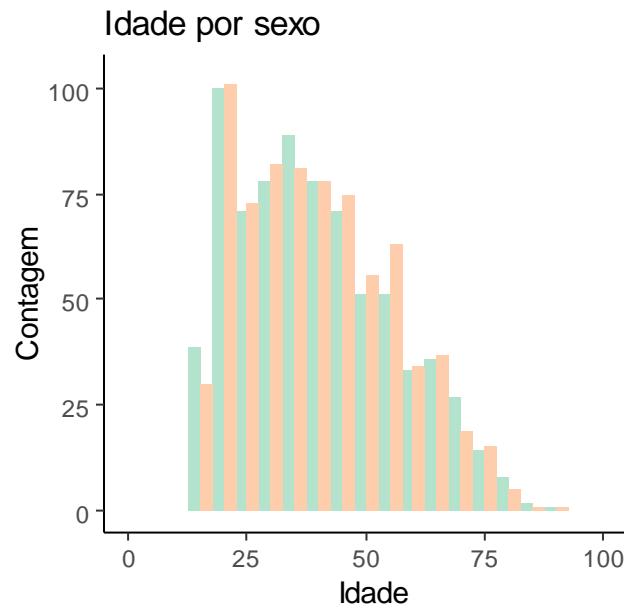
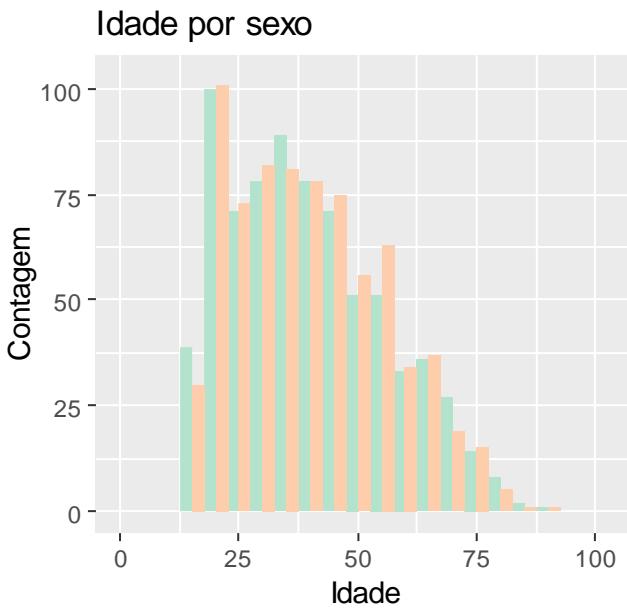
ANÁLISE DESCRIPTIVA

Histograma



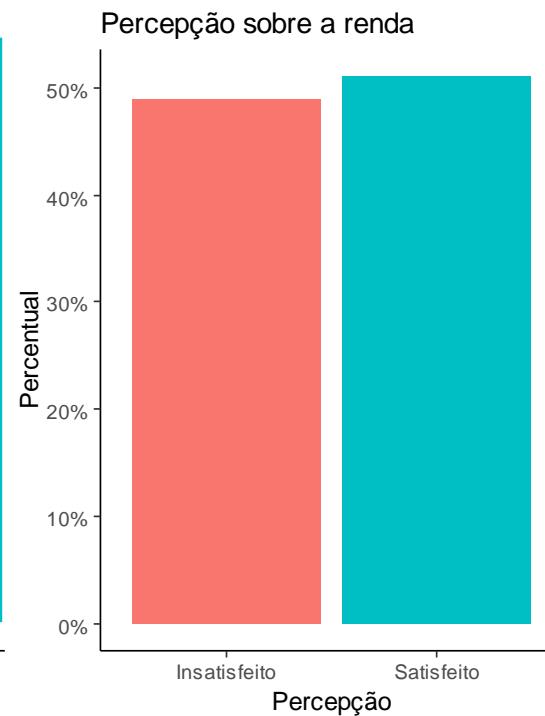
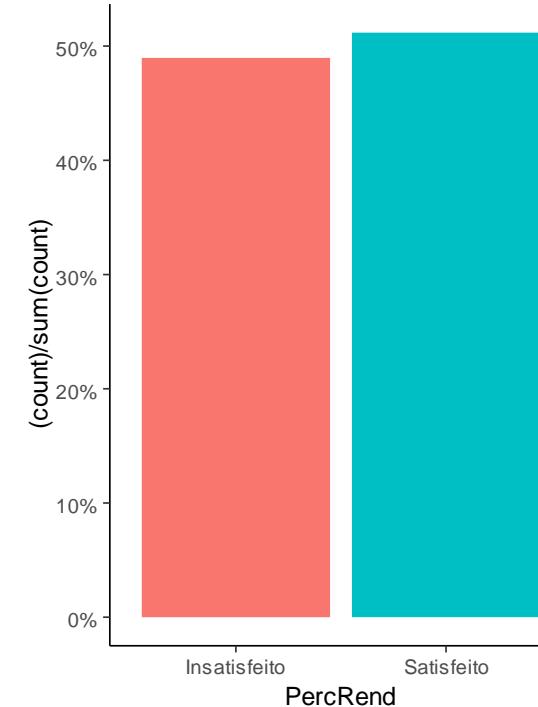
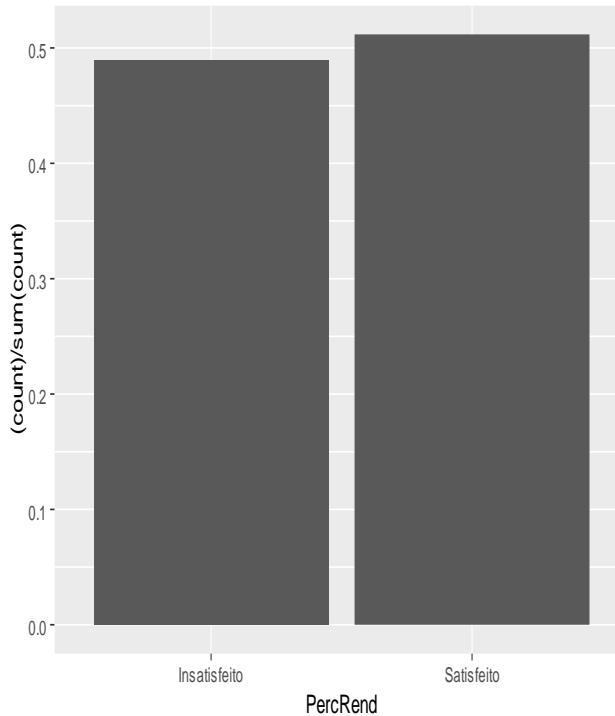
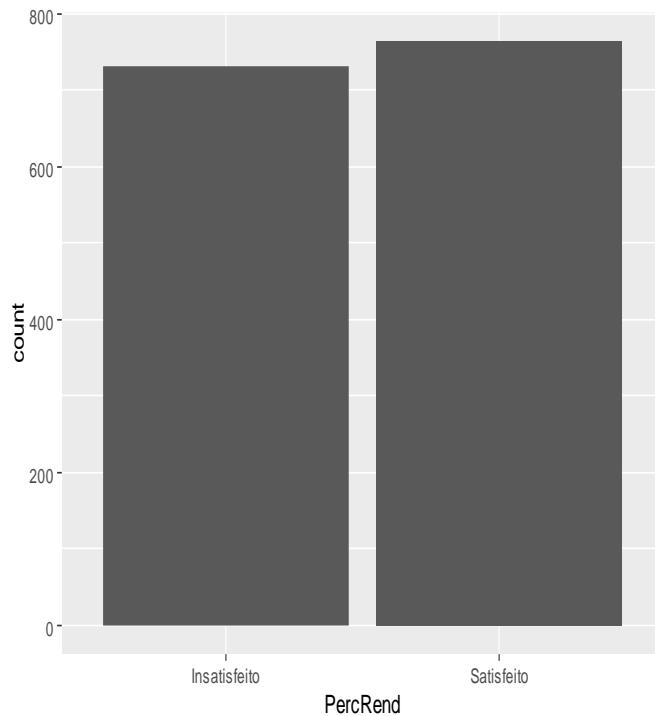
ANÁLISE DESCRIPTIVA

histograma



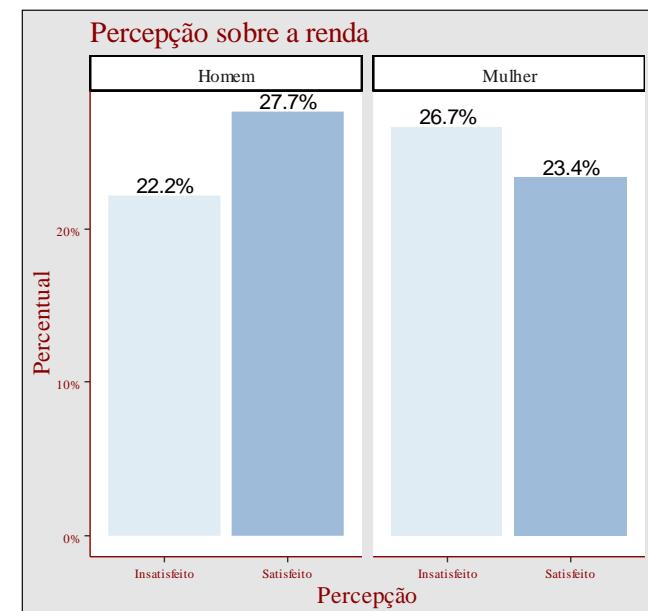
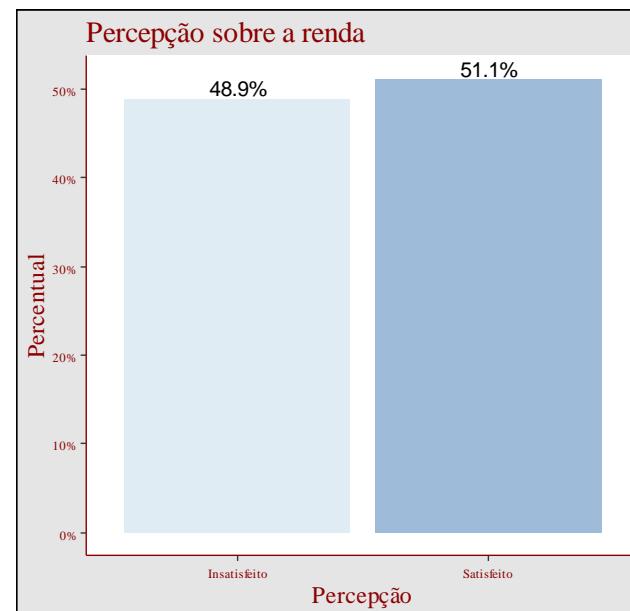
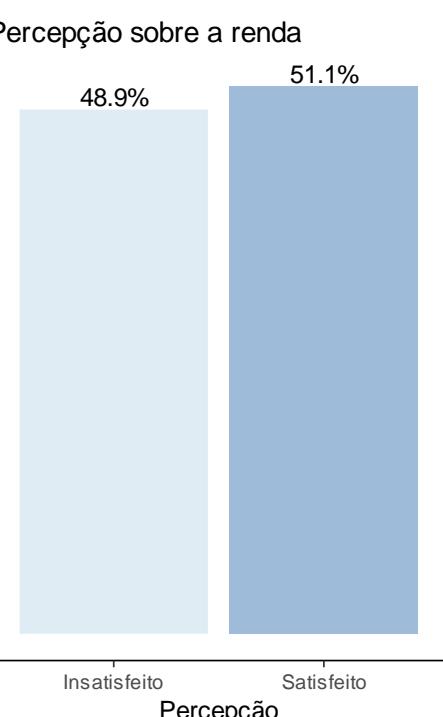
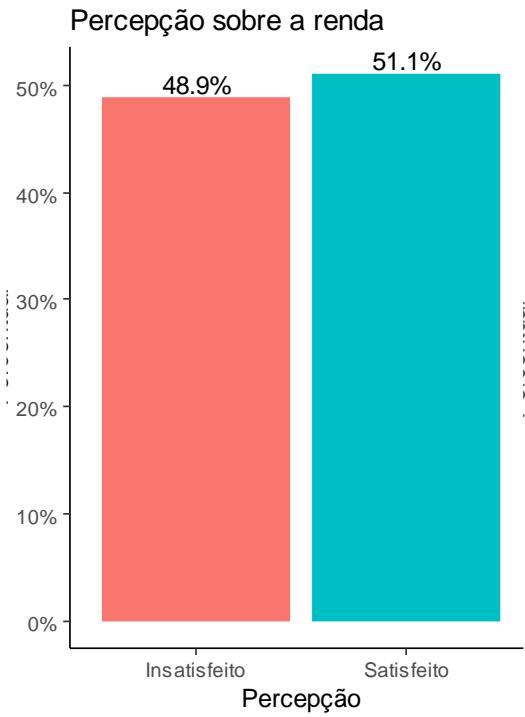
ANÁLISE DESCRIPTIVA

Gráfico de barras



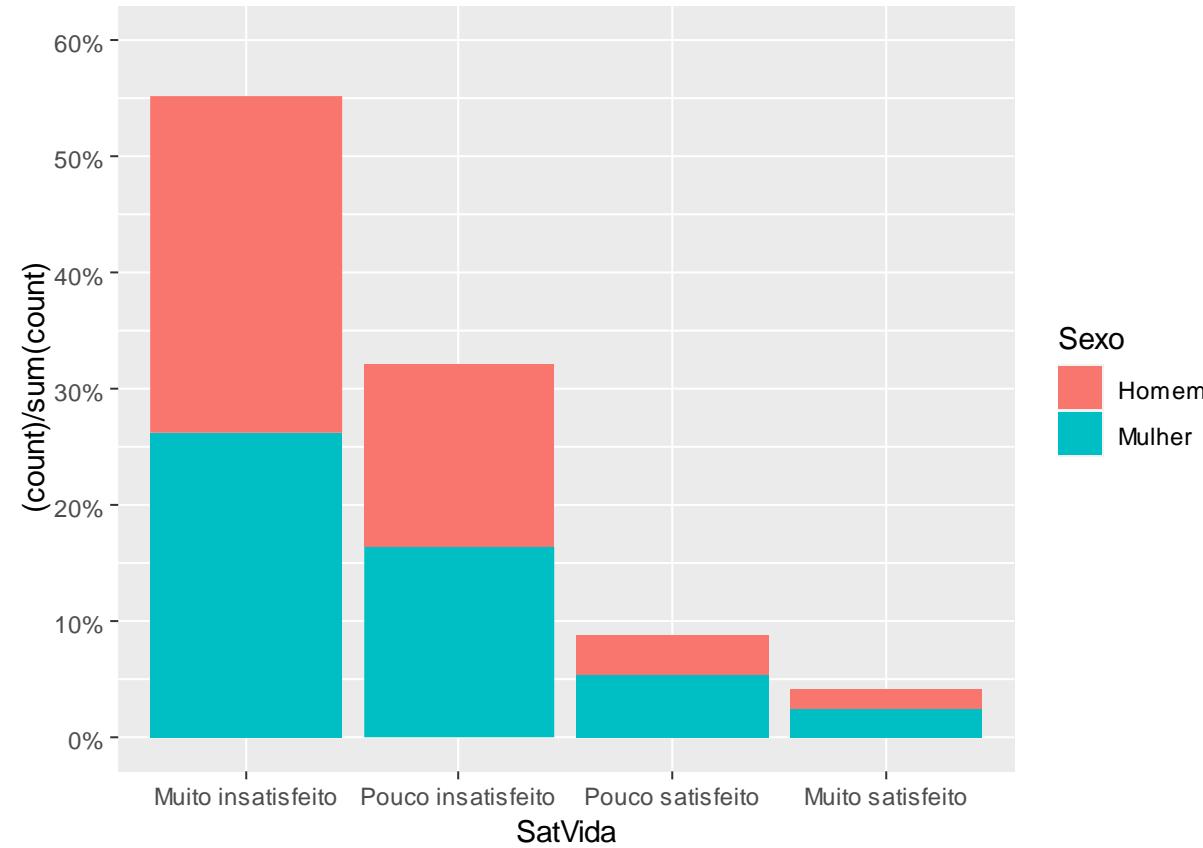
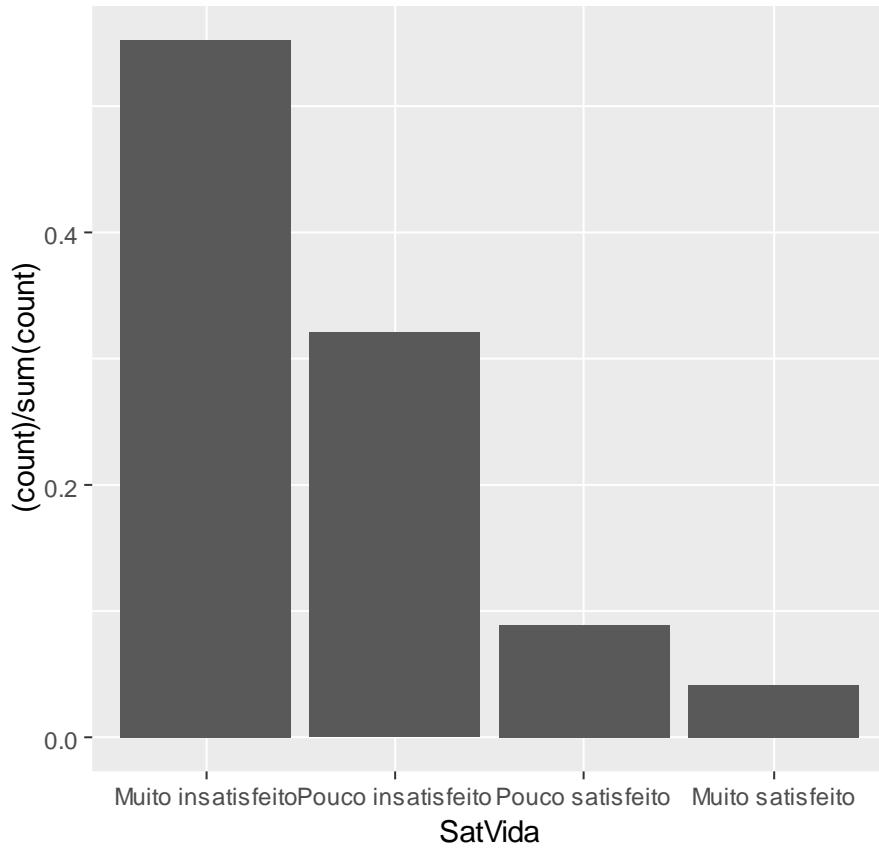
ANÁLISE DESCRIPTIVA

Gráfico de barras



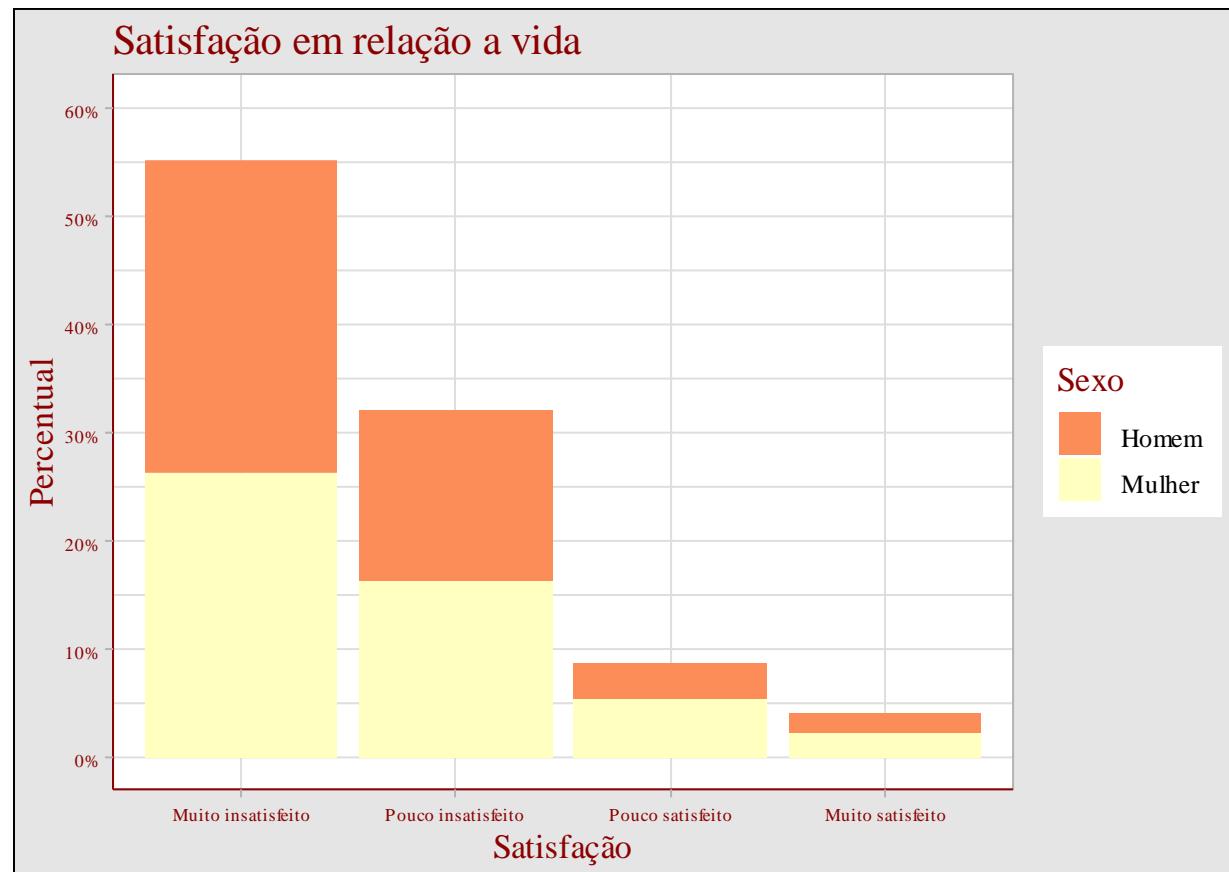
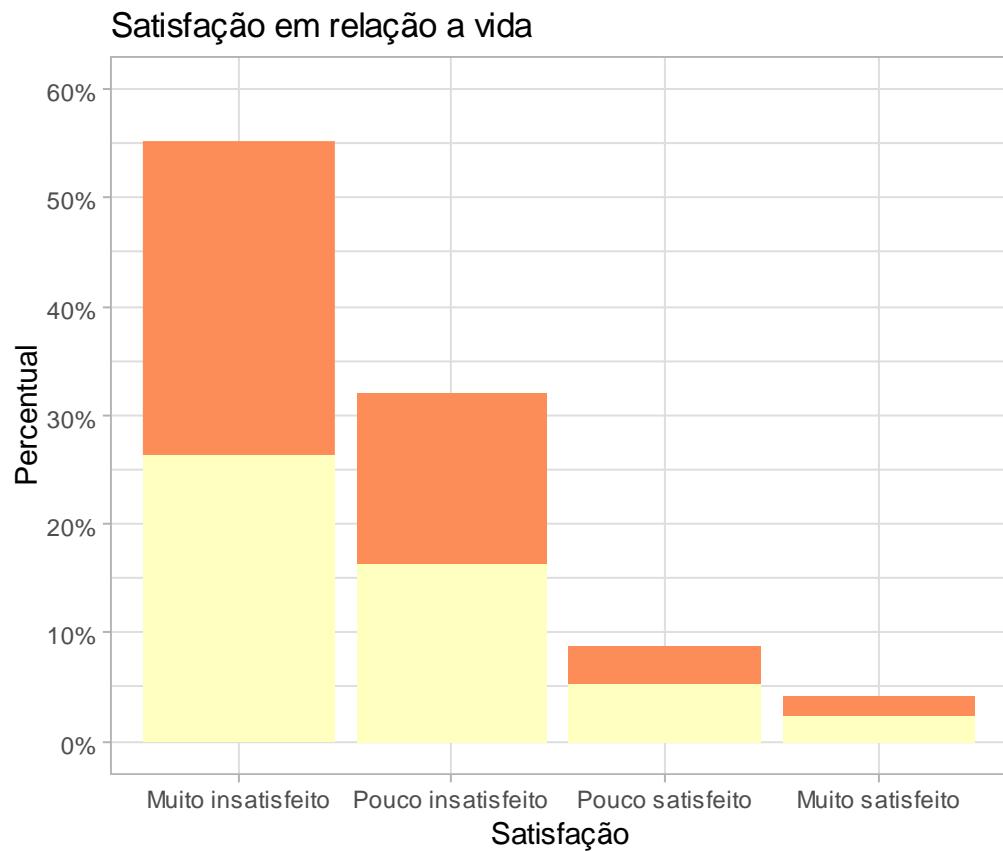
ANÁLISE DESCRIPTIVA

Gráfico de barras empilhadas



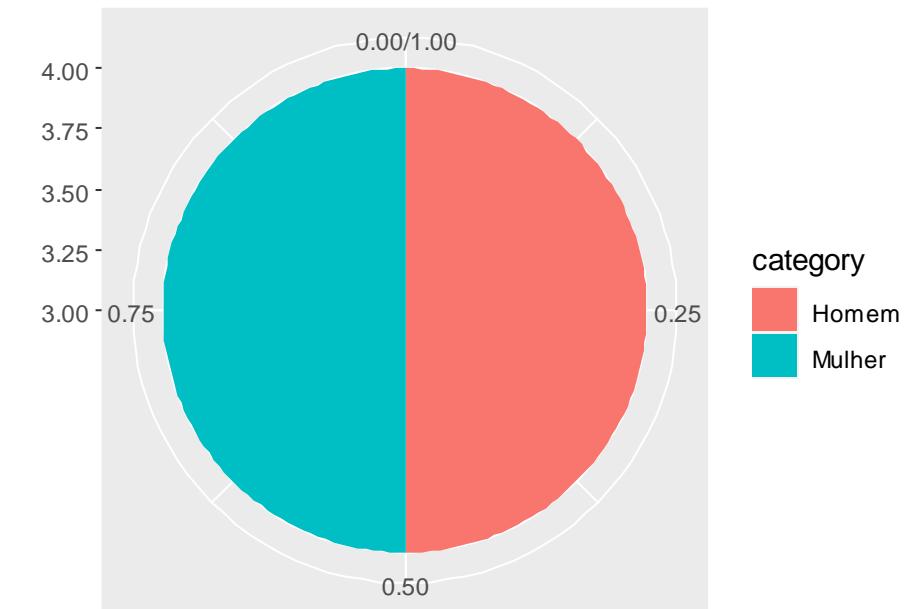
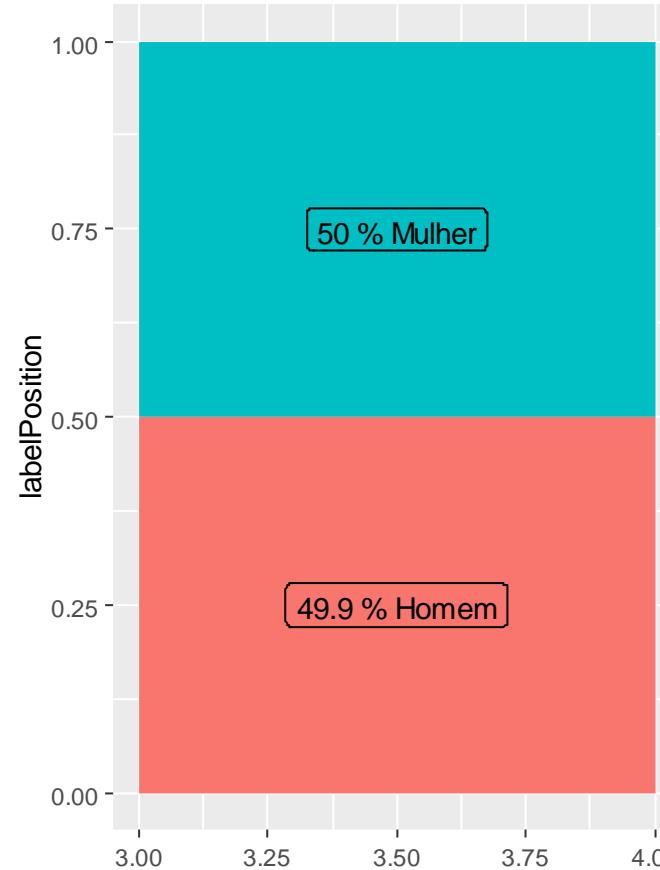
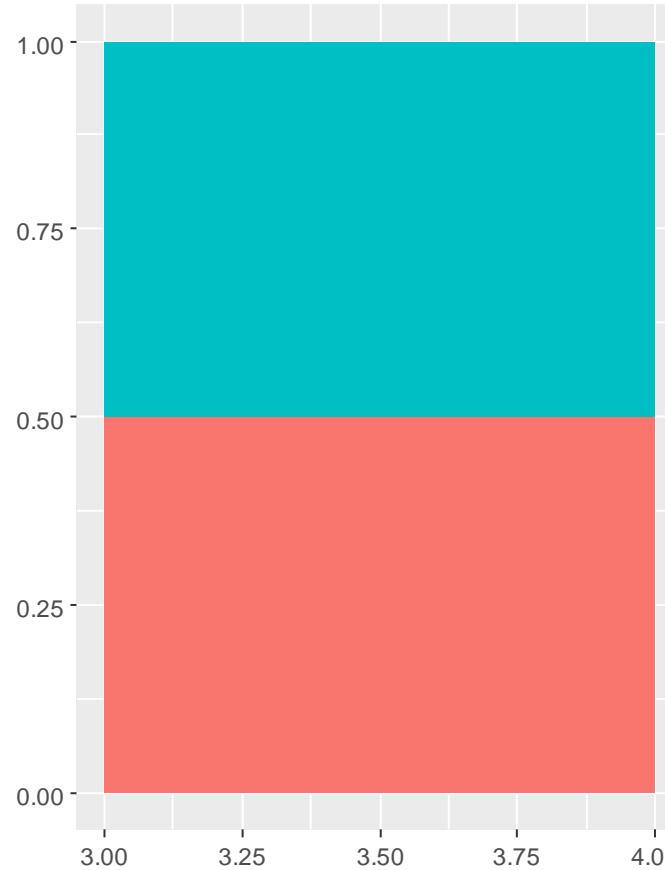
ANÁLISE DESCRIPTIVA

Gráfico de barras empilhadas



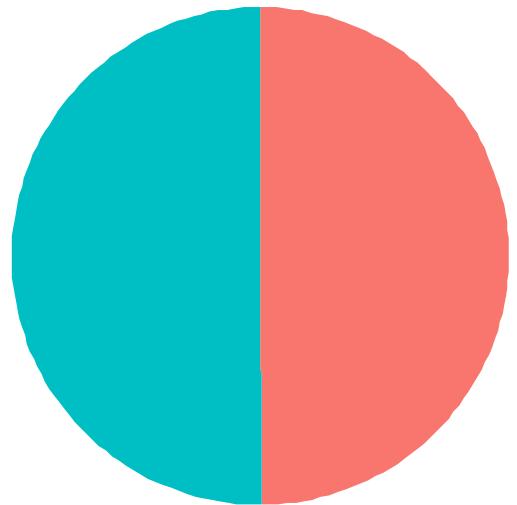
ANÁLISE DESCRIPTIVA

Gráfico de pizza e de seção

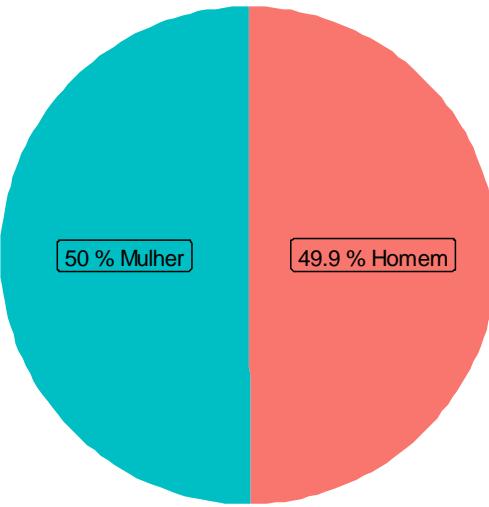


ANÁLISE DESCRIPTIVA

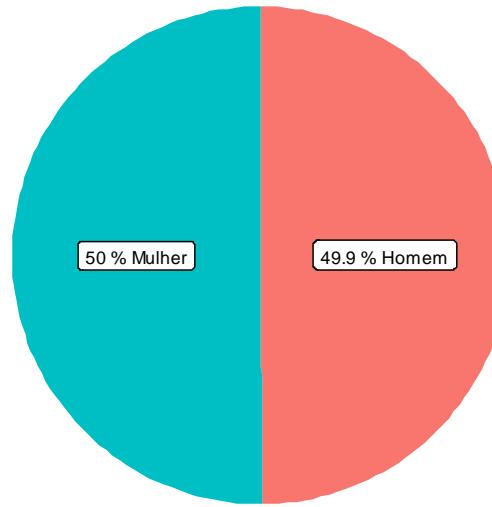
Gráfico de pizza e de seção



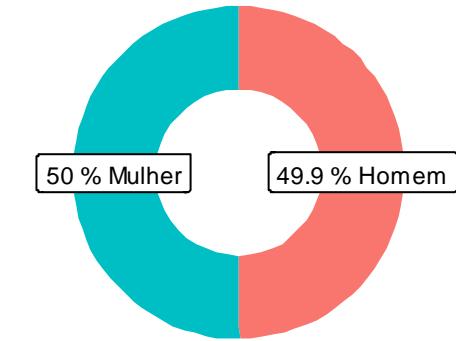
category Homem Mulher



category Homem Mulher



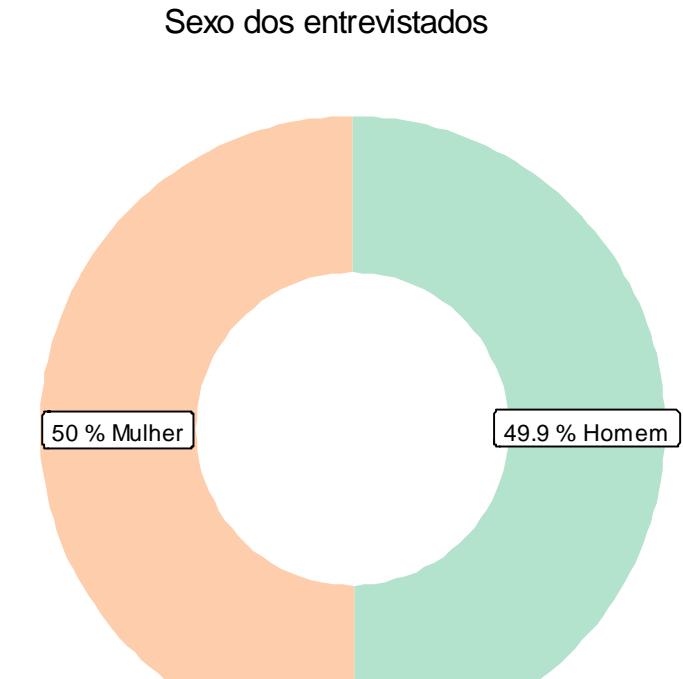
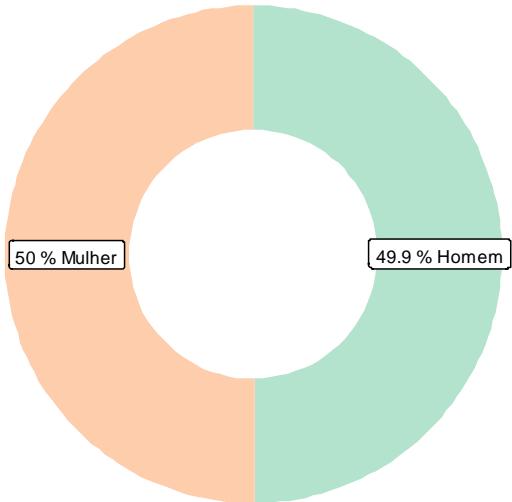
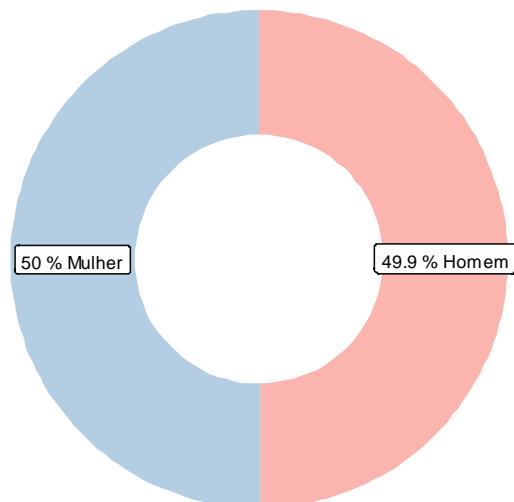
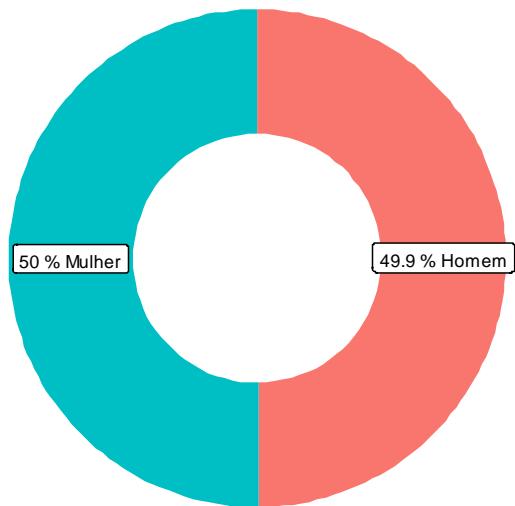
category Homem Mulher



category Homem Mulher

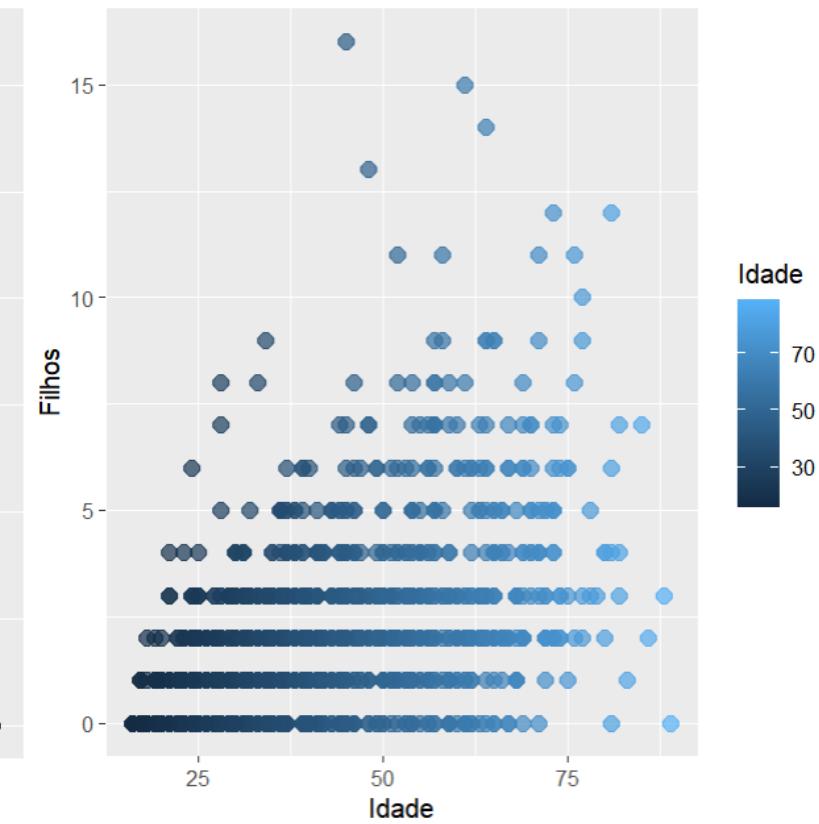
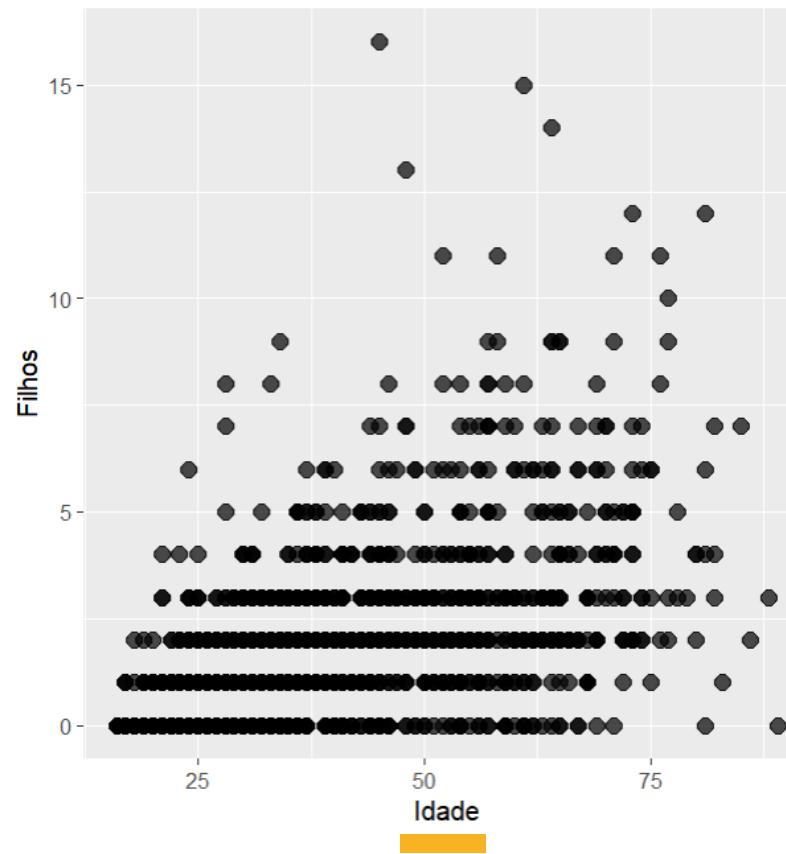
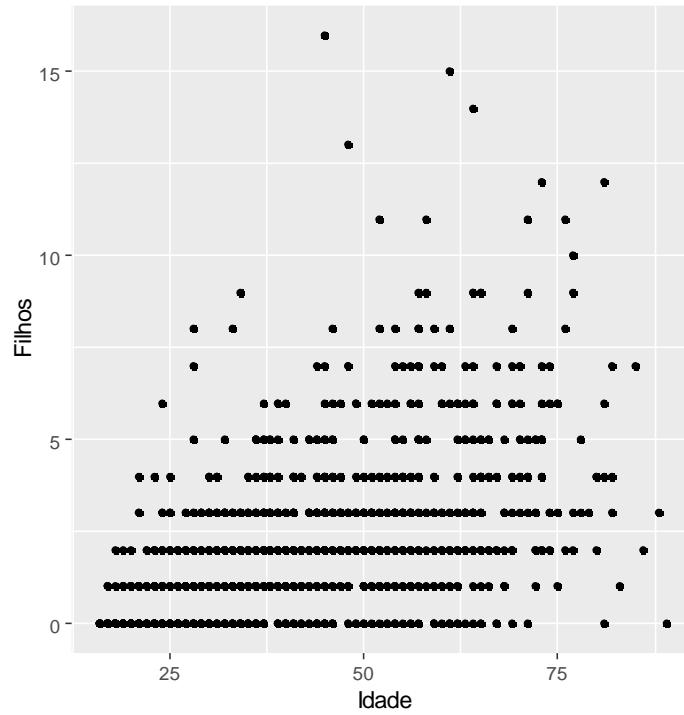
ANÁLISE DESCRIPTIVA

Gráfico de pizza e de seção



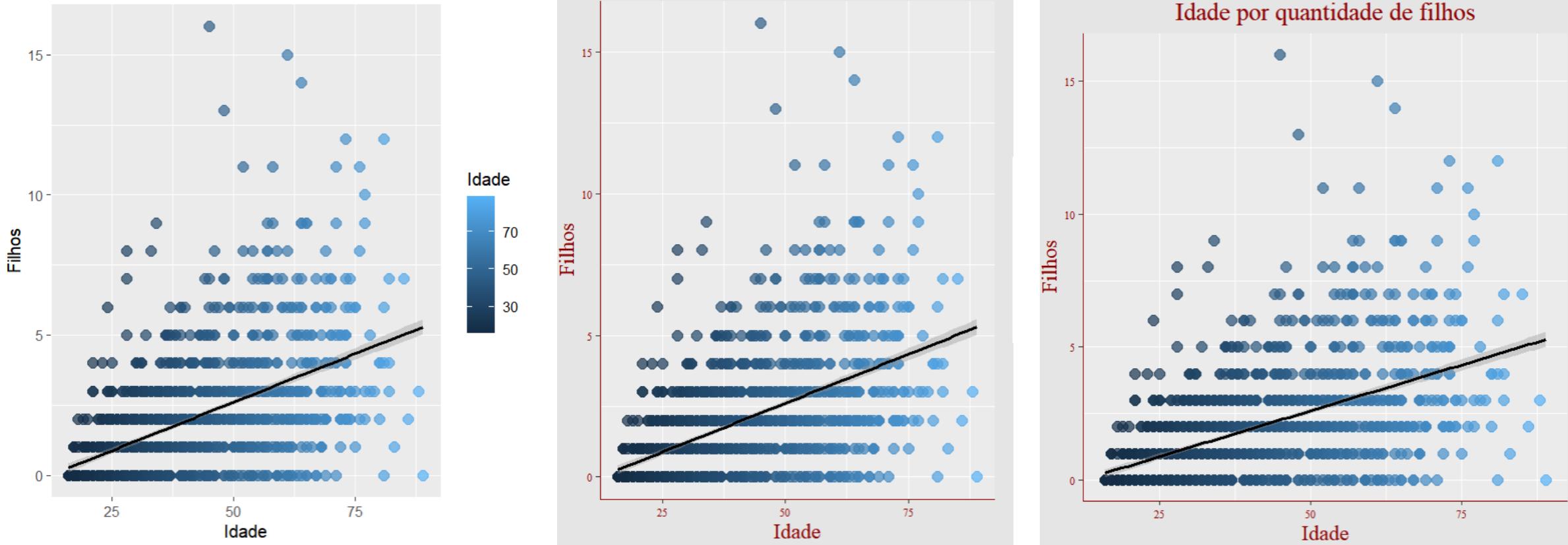
ANÁLISE DESCRIPTIVA

Gráfico de dispersão



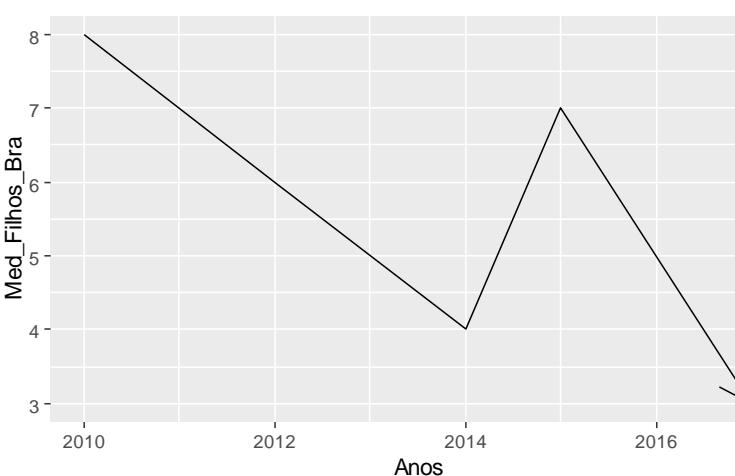
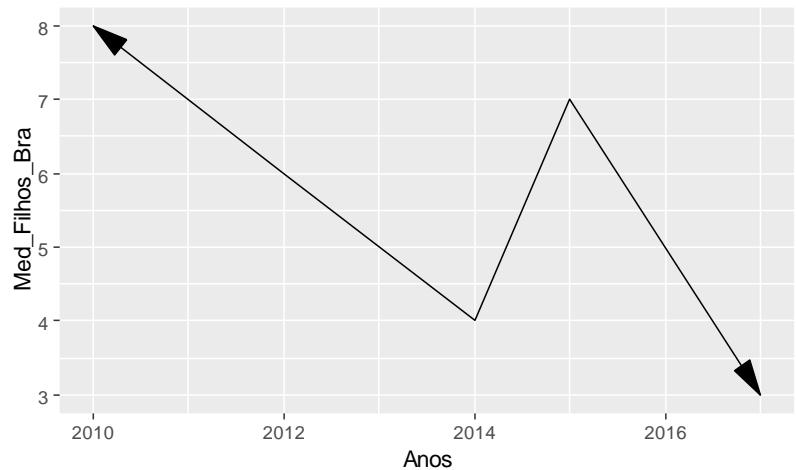
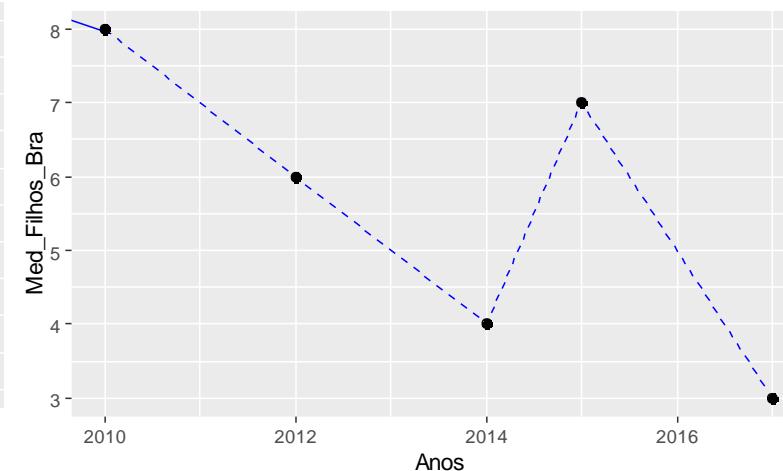
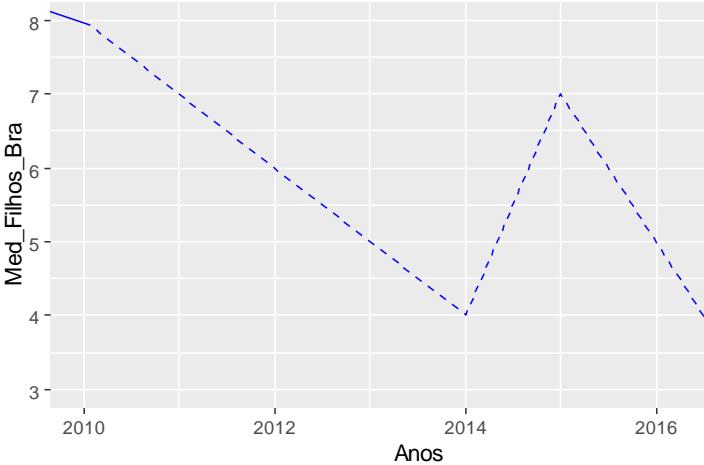
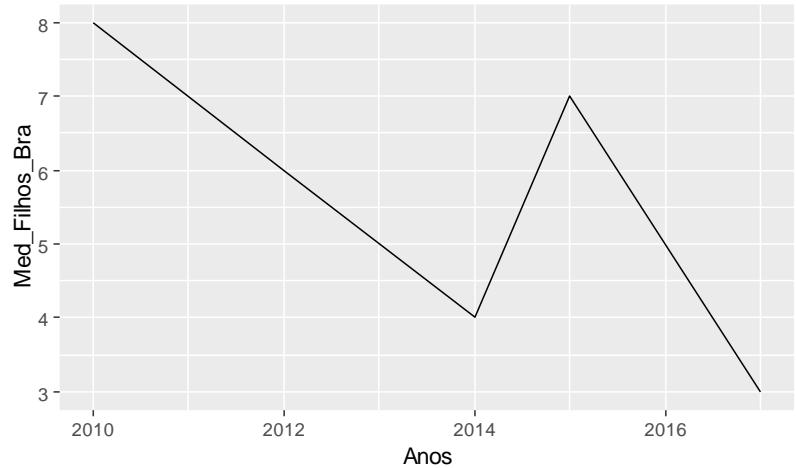
ANÁLISE DESCRIPTIVA

Gráfico de dispersão



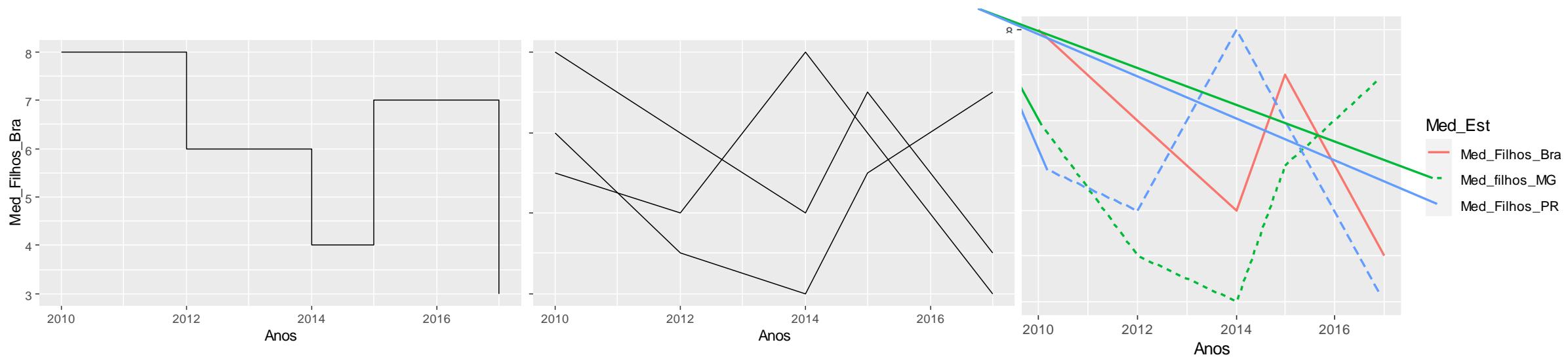
ANÁLISE DESCRIPTIVA

Gráfico de linhas



ANÁLISE DESCRIPTIVA

Gráfico de linhas



ANÁLISE DESCRIPTIVA

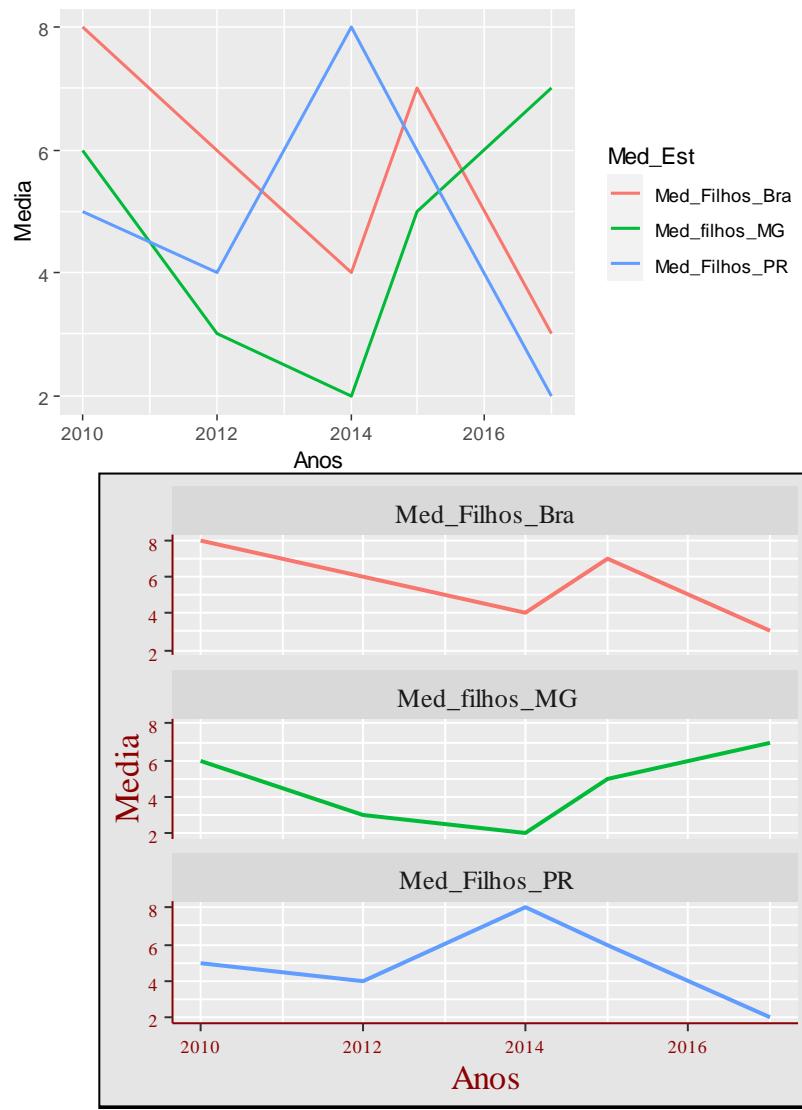
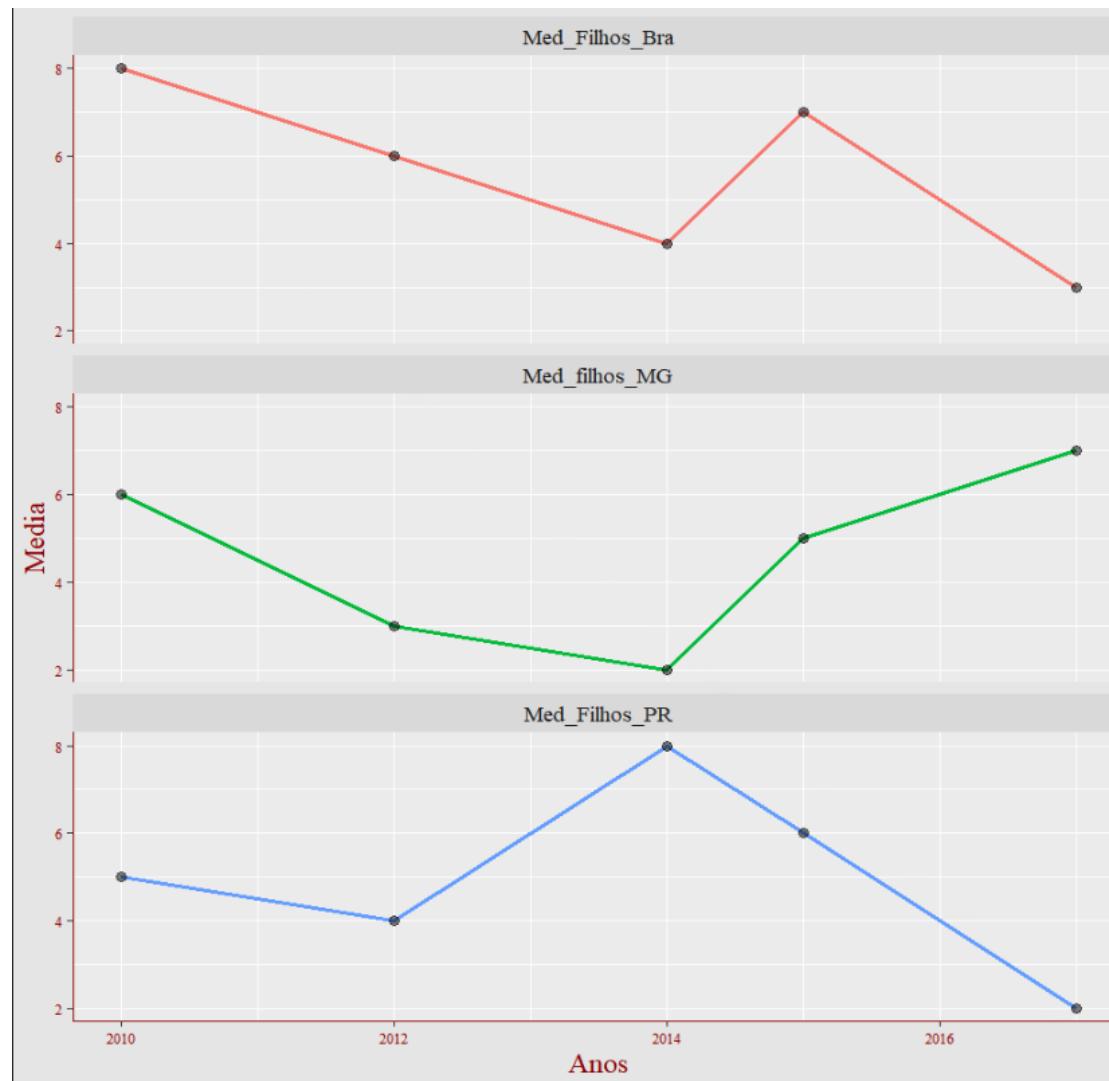


Gráfico de linhas





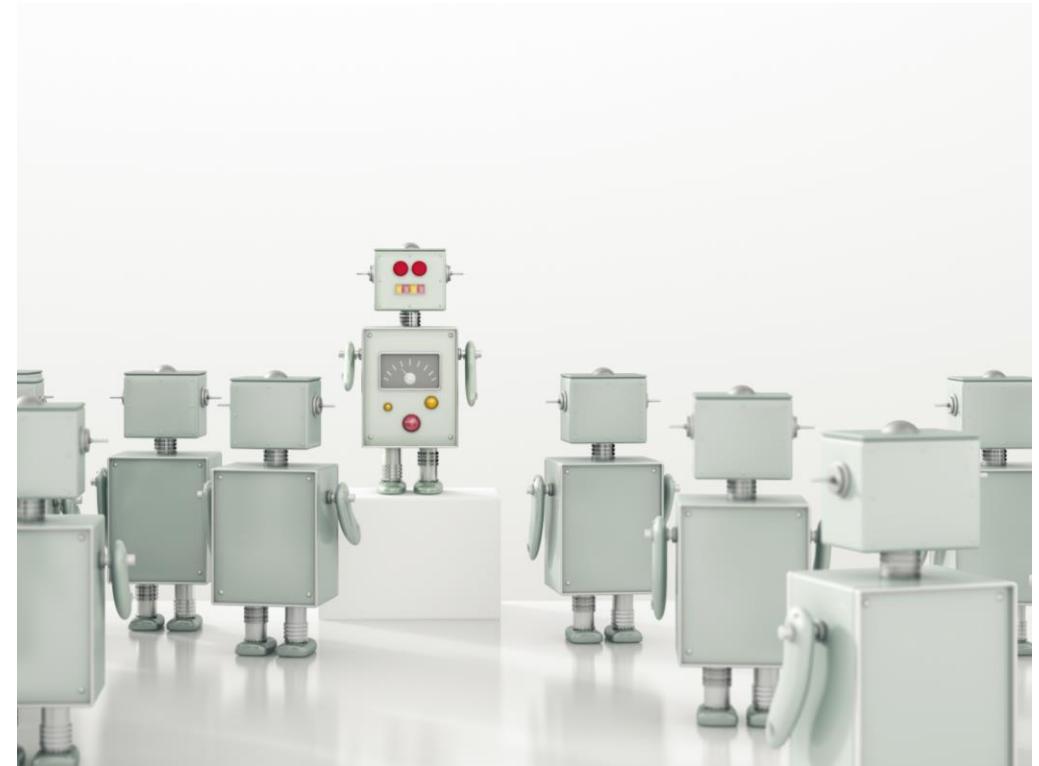
—

PROCESSAMENTO DOS DADOS

TÓPICOS

- 1-INTRODUÇÃO AO R
- 2-ANÁLISE EXPLORATÓRIA E MANIPULAÇÃO DOS DADOS
- 3-SALVAMENTO E ABERTURA
- 4-ANÁLISES DESCRIPTIVAS
- 5-APRESENTAÇÃO GRÁFICA
- 6- PROCESSAMENTO DE DADOS
- 7- ANÁLISES INFERENCIAIS

Análise descritiva com output composto por valores absolutos e relativos (percentuais), pode ser aplicado a grandes bases de dados.



PROCESSAMENTO DOS DADOS

Além da análise descritiva de questão por questão, podemos ter a necessidade de processar de uma única vez uma quantidade grande de dados, a fim de criar um relatório executivo por exemplo;

Através do pacote: expss, conseguimos fazer isso.

Inicialmente vamos criar uma base de dados com diversos tipos de variáveis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	NOME	SEXO	IDADE	CIDADE	ESTADO	RENDIMENTO	Q1	Q1_1	Q1_2	Q2	Q2_1	Q3	Q11	Peso
2	Beatriz	F	29	Abadia dos Dourados	MG	1	LARANJA		GOIABA	FRIO	CALOR	10	10	0,59
3	Elisa	F	65	Abaeté	MG	4		LIMÃO			CALOR	6	6	0,59
4	Pedro	M	28	Abre Campo	MG	6	LARANJA		GOIABA	FRIO	CALOR	4	4	1,59
5	Benjamin	M	35	Acaíaca	MG	5		LIMÃO	GOIABA	FRIO		9	9	1,59
6	Isaac	M	28	Açucena	MG	1	LARANJA				CALOR	1	1	1,59
7	Camila	F	39	Abadia dos Dourados	MG	4	LARANJA	LIMÃO		FRIO	CALOR	8	8	0,59
8	Victor	M	34	Abaeté	MG	6	LARANJA		GOIABA		CALOR	7	7	1,59
9	Aarão	M	39	Abre Campo	MG	5				FRIO	CALOR	11	11	1,59
10	Vinicius	M	40	Acaíaca	MG	8	LARANJA		GOIABA		CALOR	6	6	1,59
11	César	M	50	Açucena	MG	7			GOIABA		CALOR	7	7	1,59
12	Karen	F	46	Abadia dos Dourados	MG	4	LARANJA	LIMÃO			CALOR	6	6	0,59

Essa base, assim como as utilizadas nos demais tópicos estarão disponíveis no GitHub



PROCESSAMENTO DOS DADOS

```
1126 #Tópico 6#####
1127 #Processamento dos dados#####
1128
1129 ##### para gerar o output como porcentagem
1130 add_percent = function(x, digits = get_expss_digits(), ...){
1131   UseMethod("add_percent")
1132 }
1133
1134 add_percent.default = function(x, digits = get_expss_digits(), ...){
1135   res = formatC(x, digits = digits, format = "f")
1136   nas = is.na(x)
1137   res[nas] = ""
1138   res[!nas] = paste0(res[!nas], "%")
1139   res
1140 }
1141
1142 add_percent.etable = function(x, digits = get_expss_digits(),
1143                               excluded_rows = "#", ...){
1144   included_rows = !grepl(excluded_rows, x[[1]], perl = TRUE)
1145   for(i in seq_along(x)[-1]){
1146     if(!is.character(x[[i]])){
1147       x[[i]][included_rows] = add_percent(x[[i]][included_rows])}}x|}
```

PROCESSAMENTO DOS DADOS

```
1149 - #####
1150
1151 library(tidyverse)
1152
1153 #Para realizar o processamento precisamos
1154 #dos seguintes pacotes: expss e openxlsx
1155 install.packages("expss")
1156 library(expss)
1157
1158 install.packages("openxlsx")
1159 library(openxlsx)
1160
1161
1162 #Por gentileza, todos usando o default de salvamento em UTF-8
1163
1164 options(OutDec = ",", digit = 0)
1165
1166 #Como em nossos relatórios utilizamos valores percentuais
1167 #com o símbolo do %, então teremos que criar um objeto
1168 #chamado "add_percent" que irá adicionar percentual a nossa
1169 #planilha.
1170 #Então olhem no sumário, clique em add_percent,
1171 #selecionem até o final e rodem. Depois voltem aqui
1172 #para retornarmos o passo a passo.
1173
1174 my
```

PROCESSAMENTO DOS DADOS

```
1174 #Vamos inicialmente importar a base de dados que coletamos
1175 #Chamando ela de base, utilizando o pacote rio e a função
1176 #import
1177
1178 install.packages("rio")
1179 library(rio)
1180
1181 base = rio::import("BANCOTREI.xlsx", which = 1)
1182
1183 head(base)
1184
1185
1186 #Queremos analisar o que ?
1187 #-Cidade, com base no sexo e rendimentos..
1188
1189 #Agora transforme todos os labels em caracteres,
1190 #usando essas linhas:
1191
1192 #install.packages("Hmisc")#Para quem já ativou o tidyverse não precisa
1193 #library(Hmisc)
1194
1195 var.labels = as.character(names(base))
1196
1197 for(i in seq_along(base)){
1198   Hmisc::label(base[, i]) ← var.labels[i]
1199 }
```

PROCESSAMENTO DOS DADOS

```
1200  
1201 #RU####  
1202 #PROCESSAMENTO DE RESPOSTA ÚNICA  
1203  
1204 #Agora é preciso que você escolha quais variáveis  
1205 #irão compor a sua lista de variáveis de respostas únicas,  
1206 #anotando o número das colunas, que pode ser verificado na  
1207 #na própria base de dados  
1208 vars_ps = list(base[,c(3:4)], base[,c(5)]) #falar da 5  
1209  
1210  
1211 #Agora escolhemos o que desejamos cruzar com as variáveis  
1212 #escolhidas anteriormente  
1213 vars_cruz = with(base, list(total(), SEXO))  
1214  
1215 #Feito isso, criamos a base, que no caso, vamos chamar de  
1216 #RU, se referindo a resposta única  
1217 ru = base %>%  
1218   calculate(cro_cpct(cell_vars = vars_ps,  
1219                 col_vars = vars_cruz)) %>%  
1220   tab_sort_desc %>%  
1221   set_caption("RU") %>%  
1222   add_percent(digit = 0)  
1223
```

PROCESSAMENTO DOS DADOS

```
1225 #criamos uma pasta de trabalho
1226 wb = createWorkbook()
1227 #Adicionamos nesse documento criado a aba RU
1228 sh = addWorksheet(wb, "RU")
1229 #Gravamos todas tabelas em um documento
1230 xl_write(ru, wb, sh)
1231
1232
1233 #valor absoluto + percentual#####
1234
1235 #Achado do arthur para análise dos valores absolutos também
1236 # Absoluto = base %>%
1237 #   tab_cells(vars_ps) %>%
1238 #   tab_cols(total(),vars_cruz) %>%
1239 #   tab_stat_cases(label = "N", total_label = "") %>%
1240 #   tab_stat_cpct(label = "%", total_statistic = "w_cpct",
1241 #                   total_label = "") %>%
1242 #   tab_pivot(stat_position = "outside_rows") %>%
1243 #   set_caption("Absoluto")
1244
1245 #Argumentos que podem ser utilizados para organização
1246 #dos valores percentuais:
1247 #“outside_rows”, “inside_rows”,
1248 #“outside_columns”, “inside_columns”
1249
```

PROCESSAMENTO DOS DADOS

```
1251 Absoluto = base %>%
1252   tab_cells(vars_ps) %>%
1253   tab_cols(total(),vars_cruz) %>%
1254   tab_stat_cases(label = "N", total_label = "") %>%
1255   tab_pivot(stat_position = "outside_rows") %>%
1256   set_caption("Absoluto")
1257
1258 sh2 = addWorksheet(wb, "Absoluto")
1259 xl_write(Absoluto, wb, sh2)
```

PROCESSAMENTO DOS DADOS

```
1261 #PESO#####
1262 base$Peso ← as.numeric(base$Peso)
1263
1264
1265 ANALIPESO = base %>%
1266   calculate(cro_cpct(cell_vars = vars_ps,
1267                       col_vars = vars_cruz,
1268                       total_statistic = "w_cases",
1269                       weight = base$Peso)) %>%
1270   tab_sort_desc %>%
1271   set_caption("PESO")%>%
1272   add_percent(digits = 0)
1273
1274
1275 sh1 = addWorksheet(wb, "PESO")
1276 xl_write(ANALIPESO, wb, sh1)
1277
```

PROCESSAMENTO DOS DADOS

```
1278 #RM#####
1279 #Agora a criação da aba com o processamento de respostas múltiplas
1280 #Utilizamos a função mrset_p para analisar mais de uma coluna
1281 rm = base %>%
1282   calculate(cross_cpct(base, cell_vars = list(mrset_p("Q1"),
1283                           mrset_p("Q2")),
1284                           col_vars = vars_cruz)) %>%
1285   tab_sort_desc %>% set_caption("RM")%>%
1286   add_percent(digits = 0)
1287
1288 sh2 = addWorksheet(wb, "RM")
1289 xl_write(rm, wb, sh2)
1290 #MÉDIA#####
1291 vmedias = list(base[,c(12)])
1292
1293 medias_cruz = with(base, list(total(), SEXO))
1294
1295 MEDIA = base %>%
1296   calculate(cro_mean(cell_vars = vmedias,
1297                       col_vars = medias_cruz)) %>%
1298   tab_sort_desc %>% set_caption("MEDIA")
1299
1300 sh3 = addWorksheet(wb, "MEDIA")
1301 xl_write(MEDIA, wb, sh3)
1302
```

PROCESSAMENTO DOS DADOS

```
1304 #criar arquivo xls  
1305 saveWorkbook(wb, "Output3.xlsx", overwrite = TRUE)  
1306
```

	A	B	C	D	E	F	G
1	RU						
2			#Total	SEXO			
3					F	M	
4	IDADE	18		10,0%	12,5%	8,3%	
5		28		10,0%		16,7%	
6		39		10,0%	12,5%	8,3%	
7		51		10,0%	25,0%		
8		23		5,0%		8,3%	
9		29		5,0%	12,5%		
10		34		5,0%		8,3%	
11		35		5,0%		8,3%	
12		40		5,0%		8,3%	
13		46		5,0%	12,5%		
14		50		5,0%		8,3%	
15		54		5,0%		8,3%	
16		62		5,0%		8,3%	
17		64		5,0%		8,3%	
18		65		5,0%	12,5%		
19		70		5,0%	12,5%		
20		#Total cases		20	8	12	
21	CIDADE	Abadia dos D		20,0%	50,0%		
22		Abaeté		20,0%	37,5%	8,3%	
23		Abre Campo		20,0%	12,5%	25,0%	
24		Acaiaca		20,0%		33,3%	
25		Açucena		20,0%		33,3%	
26		#Total cases		20	8	12	
27	ESTADO	MG		100,0%	100,0%	100,0%	
28		#Total cases		20	8	12	



7- ANÁLISES INFERENCIAIS

1º Análise bivariada

Testes paramétricos e não paramétricos

NAIARA ALCANTARA

TÓPICOS

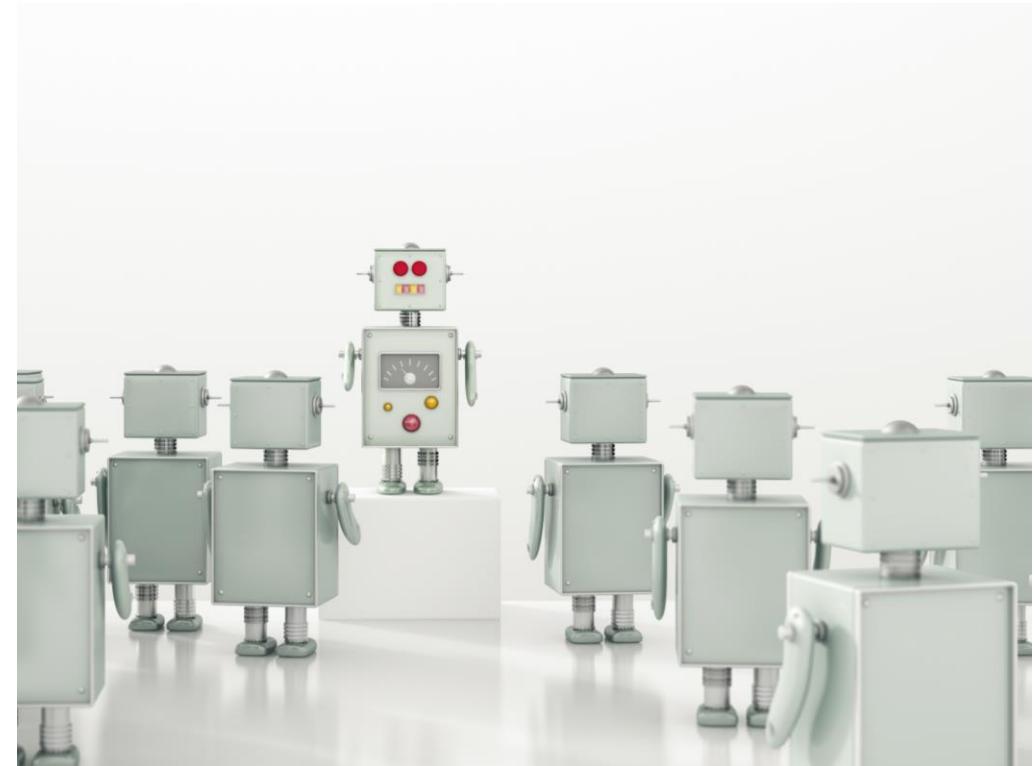
- 1-INTRODUÇÃO AO R**
- 2-ANÁLISE EXPLORATÓRIA E MANIPULAÇÃO DOS DADOS**
- 3-SALVAMENTO E ABERTURA**
- 4-ANÁLISES DESCRIPTIVAS**
- 5-APRESENTAÇÃO GRÁFICA**
- 6- PROCESSAMENTO DE DADOS**
- 7- ANÁLISES INFERENCIAIS**

Análise bivariada

Teste de correlação

Teste de regressão linear simples e múltiplo

Teste de regressão logística simples e múltiplo



ANÁLISES INFERENCIAIS: BIVARIADA

- A primeira atividade que faremos, antes mesmo de entender o que é uma análise de dados bivariada, é aprender a usar mais um tipo de base de dados (dados eleitorais).
- Então já aprendemos a baixar dados sobre opinião pública lá do site do LAPOP, no tópico 3, e trabalhamos com esses dados nos tópicos seguintes (descrevendo e analisando graficamente).
- Agora usaremos dados do TSE. Dessa forma, não precisaremos justificar teoricamente nossas escolhas a todo momento.
- Isso é muito importante porque a escolha do material empírico deve ser feita com base na teoria.
- Usaremos a base de dados para senador, cuja eleições ocorrem a cada 4 anos.
- Quem não se lembra quais são os tipos de variáveis é indicado de busque os tópicos iniciais do curso.

ANÁLISES INFERENCIAIS: BIVARIADA

- O que é uma amostra ?
É uma pequena fração do universo



Amostra NÃO é:
Mínimo de 10% da população

**Amostragem
probabilística**

- Amostra Aleatória simples
- Amostra Sistemática
- Amostra Estratificada
- Amostra por Conglomerado



**Amostragem
não-
probabilísticas**

- Amostra por julgamento
- Amostra por cotas
- Amostra bola de neve
- Amostra desproporcional

Para entender mais sobre amostra, ler: Métodos quantitativos para iniciantes

<https://cpop.ufpr.br/publicacoes-cpop/>

ANÁLISES INFERENCIAIS: BIVARIADA

- Para realizar uma análise inferencial nós temos que utilizar amostras que sejam estatisticamente representativas da população amostrada/Universo.
- Não é possível realizar análises e fazer inferências sobre amostras que não podem ser extrapoladas para o Universo.
- Em uma pesquisa sempre devemos partir de uma hipótese de pesquisa (H_1)-hipótese nula-de que não existe uma relação estatística entre as questões que estamos estudando. Se for possível rejeitar essa hipótese nula, poderemos confirmar sua hipótese contrária (H_1).

Exemplo:

- Minha hipótese de pesquisa é de anos de escolaridade influenciam no aumento médio da renda da população.
- Se eu confirmar essa hipótese de pesquisa (H_1), estarei rejeitando a hipótese de que anos de escolaridade não aumentam a média de renda da população (H_0). 

ANÁLISES INFERENCIAIS: BIVARIADA

- Para saber se a hipótese nula deve ser confirmada ou rejeitada, devemos realizar testes inferenciais.
- Esses testes irão fornecer um valor de probabilidade que irá indicar se a H_0 deve ou não ser rejeitada.
- ~~Em geral devemos rejeitar a H_0 , quando p for < 0.05~~

Probabilidade de se obter um resultado igual (ou mais extremo) que o obtido, dado que a hipótese nula é verdadeira

$P < 0.001$ Altamente significante
 $p < 0.01$ Razoavelmente significante
 $p < = 0.05$ Pouco significante
 $p > 0.05$ pouca evidência de existência de significância.



ANÁLISES INFERENCIAIS: BIVARIADA

Teste qui- quadrado

Apropriado para variáveis qualitativas/categóricas não ordenadas, que também podem ser dicotômicas, isto é, com apenas 2 respostas válidas (exemplo de variáveis que podem ser utilizadas nesse teste: sexo, gosto musical, cor dos olhos, preferencias,...)

Assim como para qualquer teste estatístico inferencial é interessante que as amostras sejam relativamente grandes.

Esse teste somente indica se existe um relacionamento estatisticamente significativo entre as variáveis, portanto não é possível saber a direção da associação.

Teste não paramétrico: não depende de parâmetros populacionais (média e variância)



ANÁLISES INFERENCIAIS: BIVARIADA

```
##Teste de qui-quadrado#####
#Diferença de média entre eleitos e não eleitos por cor e dps sexo|  
Teste1 <- table(BASE_SEN_2022$DS_SIT_TOT_TURNO,  
                  BASE_SEN_2022$DS_COR_RACA)|  
  
chisq.test(Teste1)  
  
#  
# Pearson's Chi-squared test  
#  
# data: Teste1  
# X-squared = 15.701, df = 5, p-value = 0.00775
```



Apesar das variáveis escolhidas cumprirem o requisito para o teste, essa base de dados não é a mais adequada para o qui-quadrado.

Por que?

Porque fere o pressuposto de terem amostras grandes e semelhantes.

Explicar melhor.

```
Teste2 <- table(BASE_SEN_2022$DS_SIT_TOT_TURNO,  
                  BASE_SEN_2022$DS_GENERO)  
  
chisq.test(Teste2)  
  
# Pearson's Chi-squared test with Yates' continuity correction  
#  
# data: Teste2  
# X-squared = 1.0059, df = 1, p-value = 0.3159
```

- o título do teste,
- quais variáveis foram usadas dentro do objeto criado
- os graus de liberdade e o p valor do teste.

O teste qui-quadrado permite apenas verificar se há relação estatística entre as variáveis, através da rejeição ou confirmação da H_0

ANÁLISES INFERENCIAIS: BIVARIADA

Teste lambda (λ)

A estatística de Goodman-Kruskal Lambda é uma medida de associação entre variáveis categóricas, especialmente quando uma das variáveis é ordinal (ou seja, tem uma ordem específica). Mas também pode ser utilizado para variáveis sem ordenação.

Essa estatística varia entre 0 e 1, onde 0 indica que não há associação, e 1 indica uma associação perfeita.

Apresenta um pouco mais de informação que o qui-quadrado, porque permite analisar também a direção da relação entre as variáveis testadas.

Sempre uma das variáveis será considerada dependente, ainda que ainda não tenha sido feito essa distinção.



ANÁLISES INFERENCIAIS: BIVARIADA

```
#Antes de rodar o teste, vamos atribuir os levels  
BASE_SEN_2022$DS_COR_RACA ←  
  factor(BASE_SEN_2022$DS_COR_RACA,  
         levels = c("PRETA", "PARDA", "INDÍGENA",  
                  "NÃO INFORMADO", "AMARELA", "BRANCA"))  
  
levels(BASE_SEN_2022$DS_COR_RACA)  
  
BASE_SEN_2022$DS_GENERO ←  
  factor(BASE_SEN_2022$DS_GENERO,  
         levels = c("FEMININO", "MASCULINO"))  
  
levels(BASE_SEN_2022$DS_GENERO)
```

- Não apresenta o valor de p, por isso deve ser feito após o qui-quadrado
- Resultado positivo na coluna, isto é, existe uma associação entre a cor e o sucesso eleitoral que é positivo.

```
lambda.test(Teste1)  
# $row  
# [1] 0  
#  
# $col  
# [1] 0.07407407  
  
lambda.test(Teste2)  
# $row  
# [1] 0  
#  
# $col  
# [1] 0
```

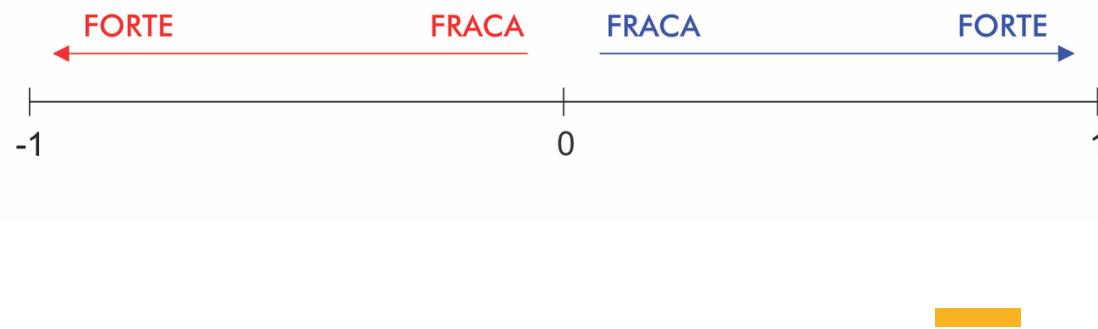
ANÁLISES INFERENCIAIS: BIVARIADA

Gamma (γ)

Apropriado para variáveis qualitativas e ordenadas

Também considera que uma das variáveis é dependente e a outra independente

Apresenta o valor de p , e a direção da associação que varia entre:



ANÁLISES INFERENCIAIS: BIVARIADA

```
#Teste de Gamma (Y)
install.packages("vcdExtra")
library(vcdExtra)

#Vamos criar uma variável chamada
#Satisfação com a vida, somente para rodar o teste
#Vamos criá-la, a partir do estdo civil

# table(BASE_SEN_2022$DS_ESTADO_CIVIL)
# CASADO(A)          DIVORCIADO(A)    SEPARADO(A) JUDICIALMENTE
# 133                  36                 4
# SOLTEIRO(A)          VIÚVO(A)
# 27                   5
```

```
#1= será o totalmente insatisfiito e 5=totalmente satisfeito
library(memisc)
BASE_SEN_2022$SatVida ← recode(BASE_SEN_2022$DS_ESTADO_CIVIL,
  "Totalmente insatisfiito" ← "VIÚVO(A)",
  "Insatisfiito" ← "SEPARADO(A) JUDICIALMENTE",
  "Meio termo" ← "DIVORCIADO(A)",
  "Satisfiito" ← "SOLTEIRO(A)",
  "Totalmente satisfiito" ← "CASADO(A)")

table(BASE_SEN_2022$SatVida)
BASE_SEN_2022$SatVida ← as.character(BASE_SEN_2022$SatVida)
DS_GRAU_INSTRUCAO ← as.character(DS_GRAU_INSTRUCAO)

tab2 ← table(BASE_SEN_2022$SatVida,
  BASE_SEN_2022$DS_GRAU_INSTRUCAO)
GKgamma(tab2)
# gamma      : -0.371
# std. error  : 0.041
# CI         : -0.451 -0.291
```

```
chisq.test(tab2, simulate.p.value = T) #Apenas para simular o valor de p
# Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)
#
# data: tab2
# X-squared = 26.001, df = NA, p-value = 0.3208
```

Interpretação:

1º Verificar se o p é < 0,05

Como não é, a hipótese nula não pode ser rejeitada.

2º Com a não rejeição da hipótese nula, não iremos interpretar o resultado do teste.

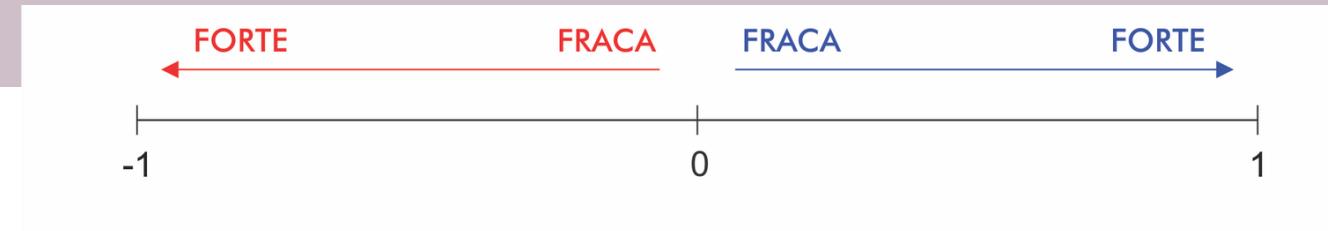
Caso a H₀ fosse <0,05, então interpretaríamos o valor do teste para saber se a relação é positiva ou negativa

ANÁLISES INFERENCIAIS: BIVARIADA

Kendall Tau

O teste de Kendall Tau, também conhecido como coeficiente de concordância de Kendall (ou apenas Kendall's tau), é uma medida estatística utilizada para avaliar o grau de concordância ou associação entre classificações ou rankings de duas variáveis. Especificamente, ele mede a correlação entre as ordens dos pares de observações em duas variáveis.

Bastante parecido com o teste de Gamma, isto é, apropriado para variáveis qualitativas ordenadas
Porém apresenta o valor do qui-quadrado junto com teste
Sua estimativa também varia entre:



ANÁLISES INFERENCIAIS: BIVARIADA

```
#Teste de Kendall#####
install.packages("Kendall")
library(Kendall)

#Utilização da variável criada para realização do teste de gamma

library(memisc)
BASE_SEN_2022$SatVida ← recode(BASE_SEN_2022$DS_ESTADO_CIVIL,
                                     "Totalmente insatisfeito" ← "VIÚVO(A)",
                                     "Insatisfeito" ← "SEPARADO(A) JUDICIALMENTE",
                                     "Meio termo" ← "DIVORCIADO(A)",
                                     "Satisfeito" ← "SOLTEIRO(A)",
                                     "Totalmente satisfeito" ← "CASADO(A)")

table(BASE_SEN_2022$SatVida)
BASE_SEN_2022$SatVida ← as.factor(BASE_SEN_2022$SatVida)
BASE_SEN_2022$DS_GRAU_INSTRUCAO ← as.factor(BASE_SEN_2022$DS_GRAU_INSTRUCAO)

Kendall(BASE_SEN_2022$SatVida,
        BASE_SEN_2022$DS_GRAU_INSTRUCAO)
# tau = 0.0245, 2-sided pvalue = 0.7069
```

Interpretação:

1º Tau = 0.0245 - sugere uma correlação muito fraca entre as duas variáveis testadas.

2º p-value = 0.7069 é bastante alto, o que indica que não há evidência estatisticamente significativa para rejeitar a hipótese nula de que não há correlação entre as variáveis.

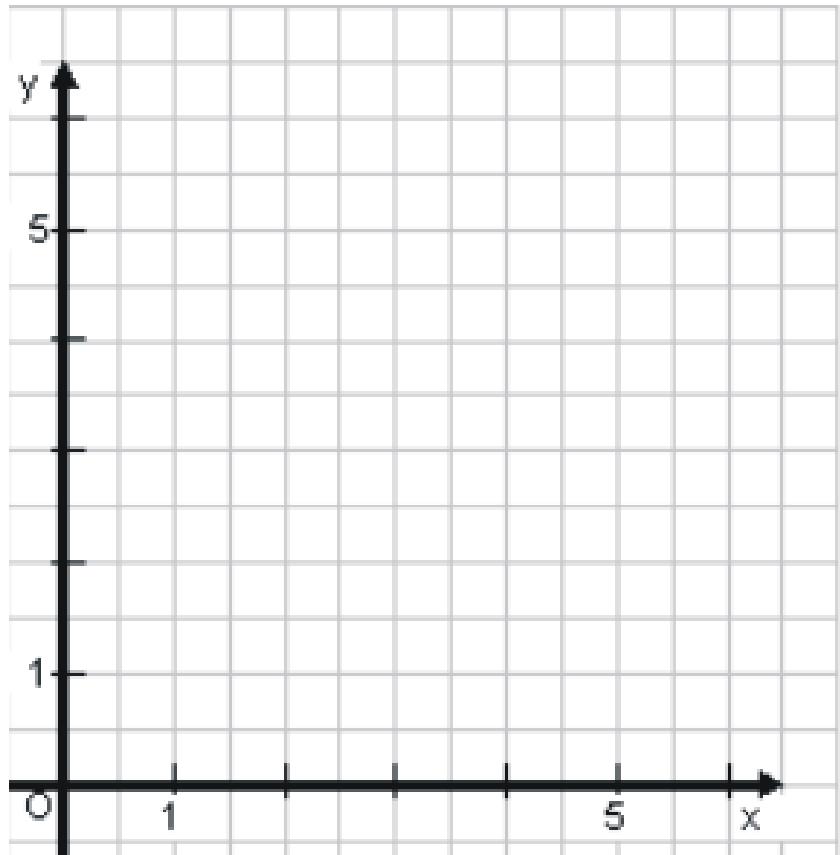


ANÁLISES INFERENCIAIS: BIVARIADA

Todos os testes de aprendemos até agora não dependiam de uma distribuição normal dos dados.

Porque os testes de normalidade se aplicam somente a variáveis quantitativas

! NORMALIDADE DOS DADOS



- Quando analisamos variáveis qualitativas ou quantitativas que não possuem distribuição normal, utilizamos os testes não paramétricos, como os que eu ensinei ou então: teste de Mann-Whitney, Kruskal-Wallis, Teste de Friedman.
- Quando analisamos variáveis quantitativas com distribuição normal podemos utilizar o teste t independente, teste t pareado, teste de anova

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T

O teste t de Student (ou simplesmente teste t) compara duas médias e mostra se as diferenças entre essas médias são significativas

Permitindo que você avalie se essas diferenças ocorreram por um mero acaso ou não.

Exemplos de quando é interessante comparar médias:

Para amostras diferentes (independentes)

- Quero saber se existe diferença de médias de votos em homens e mulheres
- Quero saber se existe diferença de média entre as notas de crianças de escolas públicas e privadas
- Quero saber se existe diferença de média entre desenvolvimento de doenças entre fumantes e não fumantes

Para mesma amostra (pareadas)

- Quero saber se existe diferença de média entre as notas dos alunos no 1º semestre e 3º semestre
- Quero saber se existe diferença de média para o exame de uma doença x, antes e depois de tomar uma determinada substância
- Quero saber se existe diferença de média entre um grupo antes e depois de assistirem uma palestra

Muitas vezes a gente acha que existe uma diferença óbvia, mas como saber se existe mesmo? Se não é somente uma percepção construída por influência social?

Através de um teste estatístico

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T não pareado –APROPRIADO PARA AMOSTRAS INDEPENDENTES

```
1257 #Teste T#####
1258
1259 options(scipen = 999, digits = 1)
1260
1261 HOMEM ← subset(BASE_SEN_2022, CD_GENERO == 2)
1262 summary(HOMEM$TOTAL_VOTOS)
1263
1264 MULHER ← subset(BASE_SEN_2022, CD_GENERO == 4)
1265 summary(MULHER$TOTAL_VOTOS)
1266
1267
1268 t.test(MULHER$TOTAL_VOTOS,HOMEM$TOTAL_VOTOS)
1269 #
1270 # Welch Two Sample t-test
1271 #
1272 # data: MULHER$TOTAL_VOTOS and HOMEM$TOTAL_VOTOS
1273 # t = -2, df = 203, p-value = 0.02
1274 # alternative hypothesis: true difference in means is not equal to 0
1275 # 95 percent confidence interval:
1276 # -506729 -35829
1277 # sample estimates:
1278 # mean of x mean of y
1279 # 280697 551975
```

Primeira linha: Nome do teste que foi feito
O teste-t de Welch é uma adaptação do teste t de Student, que é mais confiável quando as duas amostras têm variâncias desiguais e tamanhos de amostra desiguais.

A segunda linha informa de onde foram extraídos os dados para o teste:
Total de votos de homens e mulheres.

A terceira apresenta: o valor da estatística t: $t = -2$ /os graus de liberdade da curva de distribuição t: $df = 203$ / o valor de p: $p\text{-value} = 0.02$

A quarta linha informa qual a hipótese alternativa do teste: $\text{true difference in means is not equal to } 0$
A hipótese alternativa é que as médias da amostras são diferentes. Podemos então inferir que a hipótese nula do teste é que as médias das amostra são iguais. Essa hipótese foi rejeitada.

A quinta e a sexta linha informam o intervalo de confiança do teste.

A sétima, oitava e nona linha informam os dados da amostra: a média de cada amostra 280.697 e 551975

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T pareado –APROPRIADO PARA AMOSTRAS DEPENDENTES

Como nossa base de dados do senado não tem nenhuma variável que seja comparada entre o tempo em relação ao mesmo grupo, iremos criar uma pequena base de dados com dados inventados sobre a média da quantidade de mulheres candidatas ao senado.

Teremos uma média para o ano de 2020 e uma para o ano 2050, são dados inventados, por isso as datas também são inventadas.

```
#Criação da base de dados para a realização do teste t
BaseHipSen ← data.frame(
  "Regiões" = c("Norte", "Nordeste", "Centro-Oeste", "Sudeste",
               "Sul", "Argentina"),
  "Can_2020" = c(5, 10, 7, 6, 11, 8),
  "Can_2050" = c(25, 15, 4, 12, 30, 17)
)
View(BaseHipSen)
```

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T pareado

-APROPRIADO PARA AMOSTRAS DEPENDENTES

```
#Realização do teste  
t.test(BaseHipSen$Can_2020, BaseHipSen$Can_2050, paired = T)  
  
#saída do teste  
  
# Paired t-test  
#  
# data: BaseHipSen$Can_2020 and BaseHipSen$Can_2050  
# t = -2.5908, df = 5, p-value = 0.04879  
# alternative hypothesis: true mean  
#difference is not equal to 0  
# 95 percent confidence interval:  
# -18.59377288 -0.07289378  
# sample estimates:  
# mean difference  
# -9.333333
```

A interpretação é muito semelhante ao do teste não pareado
Primeira linha: Nome do teste que foi feito, indicando que é pareado

A segunda linha informa de onde foram extraídos os dados para o teste

A terceira apresenta: o valor da estatística t: $t = -2$ /os graus de liberdade da curva de distribuição t: $df = 5$ / o valor de p: $p\text{-value} = 0.04$

A quarta linha informa qual a hipótese alternativa do teste: true difference in means is not equal to 0

A hipótese alternativa é que as médias das amostras são diferentes. Podemos então inferir que a hipótese nula do teste é que as médias das amostras são iguais. Essa hipótese foi rejeitada.

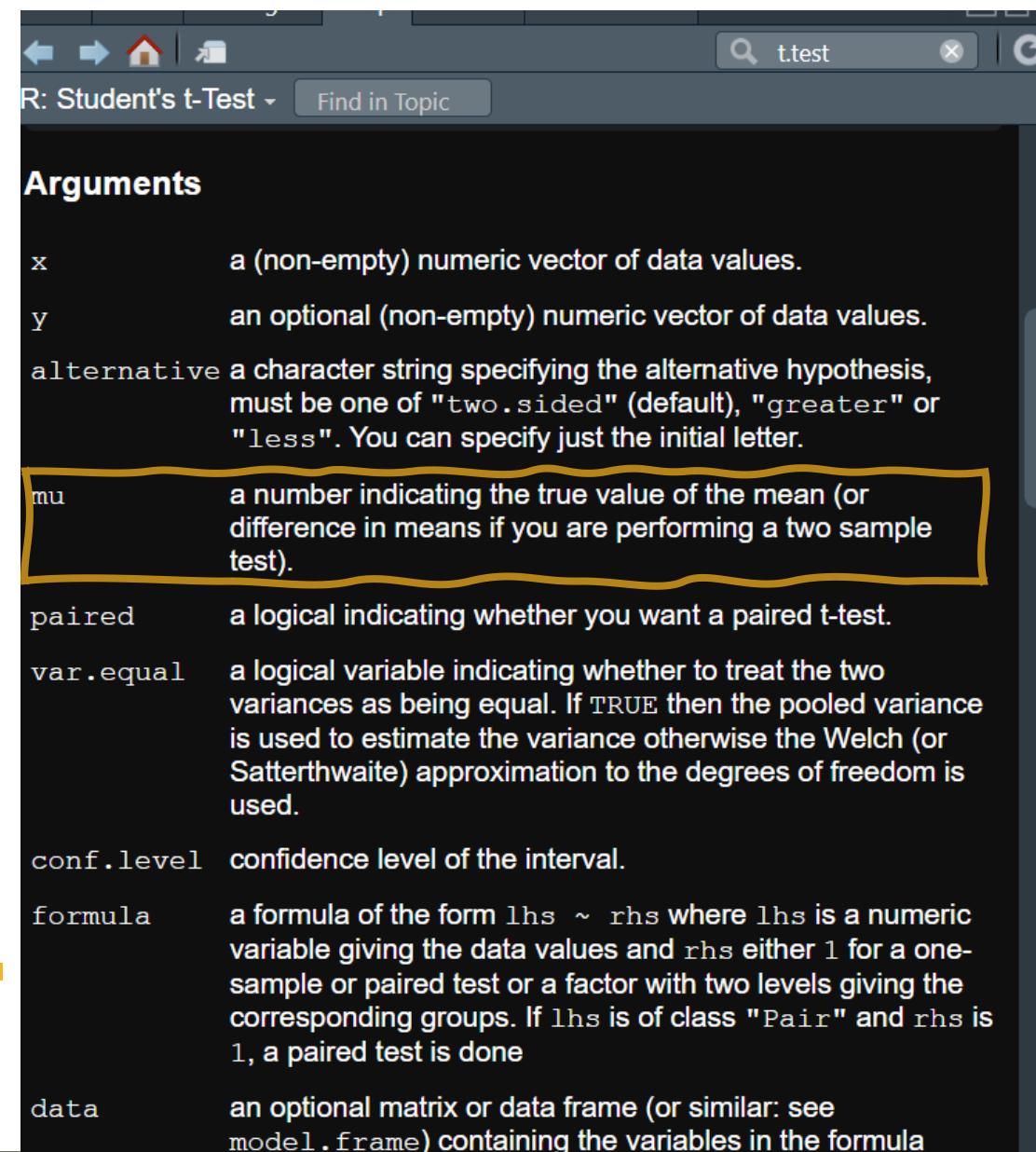
A quinta e a sexta linha informam o intervalo de confiança do teste.

A sétima, oitava e nona linha informam a diferença de média entre uma amostra e outra que é de -9

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T – Argumentos do teste

Para verificar todos os argumentos que podemos inserir em um teste t basta escrever no help “t.test” que ele irá mostrar um modelo do teste e uma lista com todos os argumentos que podemos inserir.



ANÁLISES INFERENCIAIS: BIVARIADA

Teste T para uma amostra

- Vamos supor que queremos entender se a média de mulheres candidatas ao senado em 2050 é superior ou inferir que a média na Argentina.
- Nessa suposição nós sabemos qual é a média de candidatas na Argentina em 2050.

#Rodamos o teste:

```
t.test(BaseMenor$Can_2050, mu=17)
```

Primeira linha: Nome do teste que foi feito

A segunda linha informa de onde foram extraídos os dados para o teste

A terceira apresenta: o valor da estatística t: $t = 0.04$ /os graus de liberdade da curva de distribuição t: $df = 4$ / o valor de p: $p\text{-value} = 0.9$

A quarta linha informa qual a hipótese alternativa do teste: true difference in means is not equal to 0

A hipótese alternativa é que as médias da amostras são diferentes. Podemos então inferir que a hipótese nula do teste é que as médias das amostra são iguais. Essa hipótese não pode ser rejeitada

A quinta e a sexta linha informam o intervalo de confiança do teste.

A sétima, oitava e nona linha informam os dados da amostra: a média estimada de x

! **Nesse caso não teria como dar significativo, por que ?**

```
One Sample t-test

data: BaseMenor$Can_2050
t = 0.043093, df = 4, p-value = 0.9677
alternative hypothesis: true mean is not equal to 17
95 percent confidence interval:
 4.314184 30.085816
sample estimates:
mean of x
17.2
```



ANÁLISES INFERENCIAIS

2º Correlação

ANÁLISES INFERENCIAIS: BIVARIADA

I. O que é correlação?

A correlação é uma medida estatística que indica a força e a direção da relação entre duas variáveis. Ela varia de -1 a 1, onde:

- 1 significa uma correlação perfeita positiva: à medida que uma variável aumenta, a outra também aumenta proporcionalmente.
- -1 significa uma correlação perfeita negativa: quando uma variável aumenta, a outra diminui proporcionalmente.
- 0 indica que não há relação linear entre as variáveis.

No entanto, a correlação não implica causalidade, ou seja, mesmo que duas variáveis estejam correlacionadas, isso não significa que uma causa a outra.

Mostrar no quadro

ANÁLISES INFERENCIAIS: BIVARIADA

Uma das formas de mensurar a relação entre duas variáveis são os testes de correlação

Correção de Pearson técnica para medir se duas variáveis estão relacionadas de maneira linear

Correção de Spearman é uma medida não paramétrica da dependência dos postos das variáveis

Correção de Kendall



A correlação é uma ferramenta essencial para entender relações entre variáveis, mas é importante lembrar que ela não indica causalidade. No R, temos funções simples e poderosas para calcular e visualizar essas correlações, permitindo análises rápidas e eficazes de conjuntos de dados.

ANÁLISES INFERENCIAIS: BIVARIADA

2. Tipos de correlação

- **Correlação de Pearson:** Mede a correlação linear entre duas variáveis numéricas contínuas.
- **Correlação de Spearman:** Utilizada para variáveis ordinais ou quando as variáveis não atendem aos pressupostos da correlação de Pearson (não-linearidade ou dados com outliers).
- **Correlação de Kendall:** Similar à correlação de Spearman, mas com um método ligeiramente diferente de calcular.

ANÁLISES INFERENCIAIS: BIVARIADA

I. Coeficiente de Correlação de Pearson

O coeficiente de Pearson é calculado como a razão entre a covariância das variáveis e o produto de seus desvios padrão:

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Onde:

- $cov(X, Y)$ é a covariância entre X e Y ,
- σ_X e σ_Y são os desvios padrão de X e Y .



ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

```
#Correlação de Pearson#####
# Aqui, simulamos um conjunto de dados com variáveis
# relacionadas ao nível de educação e à renda anual,
# temas centrais nas análises socioeconômicas.

#Criar uma base de dados
set.seed(123)
educ ← sample(8:20, 100, replace = TRUE)
renda ← educ * 5000 + rnorm(100, mean = 30000, sd = 10000)

dados_educ_renda ← data.frame(educ, renda)

# Visualizar as primeiras linhas da base de dados
head(dados_educ_renda)
```

	educ	renda
1	10	83035.29
2	10	84482.10
3	17	115530.04
4	9	84222.67
5	13	115500.85
6	18	115089.69
7	12	66908.31
8	11	95057.39
9	13	87907.99
10	16	103119.91
11	17	125255.71
12	18	117152.27
13	12	77792.82
14	10	81813.03
15	18	118611.09
16	16	110057.64

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Calcular a correlação de Pearson entre educação e renda  
cor(dados_educ_renda$educ, dados_educ_renda$renda)  
#OU  
cor(dados_educ_renda$educ, dados_educ_renda$renda, method = "pearson")
```

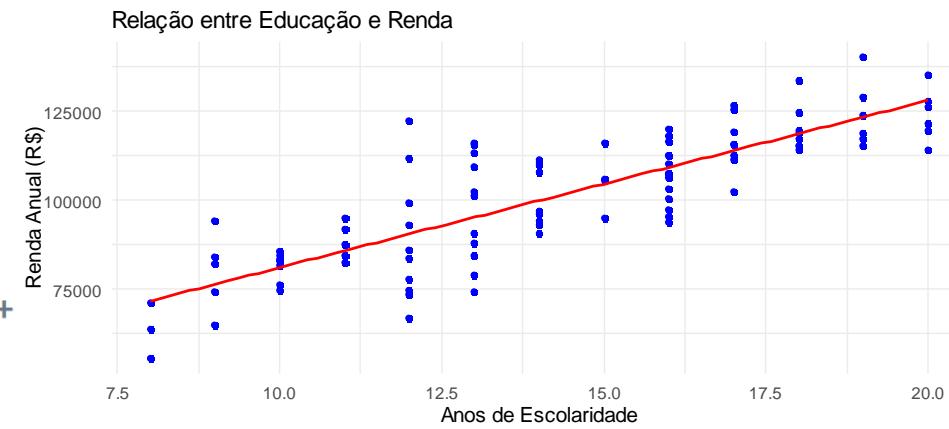
[1] 0.8450715

```
#Apresentação gráfica####
```

```
# Gráfico de Dispersão (Scatter Plot) com Linha de Tendência  
# O gráfico de dispersão é uma das maneiras mais comuns de  
# visualizar a correlação entre duas variáveis.  
# Podemos adicionar uma linha de regressão para destacar a  
# relação entre educação e renda.
```

```
# Carregar pacotes necessários  
library(ggplot2)
```

```
# Criar o gráfico de dispersão com linha de tendência  
ggplot(dados_educ_renda, aes(x = educ, y = renda)) +  
  geom_point(color = "blue") + # Adiciona pontos  
  geom_smooth(method = "lm", color = "red", se = FALSE) +  
  # Adiciona linha de regressão linear  
  labs(title = "Relação entre Educação e Renda",  
       x = "Anos de Escolaridade",  
       y = "Renda Anual (R$)") +  
  theme_minimal()
```



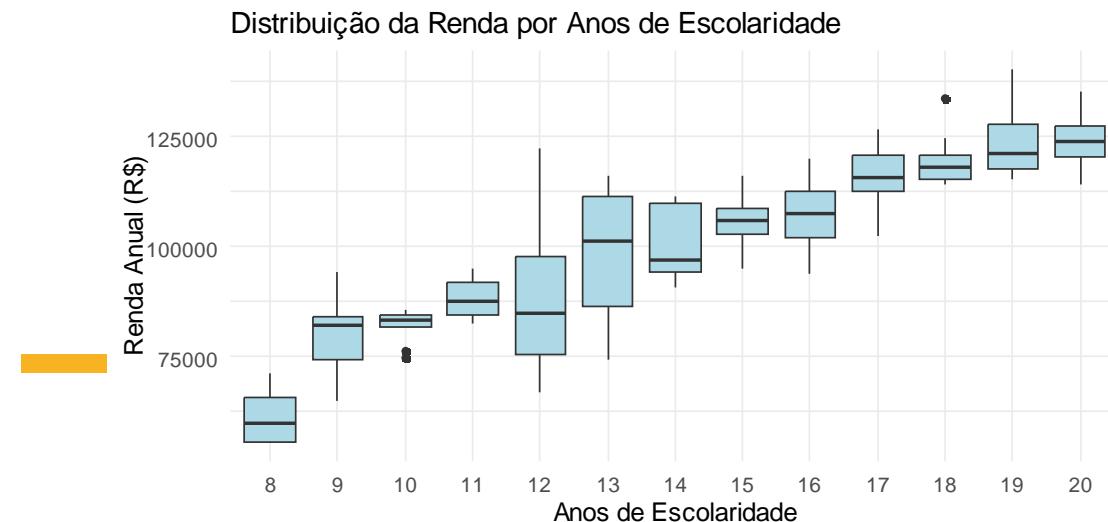
ANÁLISES INFERENCIAIS: BIVARIADA

```
# Esse exemplo simula a relação entre o nível educacional  
# (anos de escolaridade) e a renda anual. Em pesquisas sociais,  
# esperamos encontrar uma correlação positiva, pois mais anos de educação  
# geralmente estão associados a maiores rendimentos.
```

```
#Boxplot: Mostra a variação da renda em cada nível educacional,  
#incluindo valores atípicos.
```

```
# Gráfico de Boxplot
```

```
ggplot(dados_educ_renda, aes(x = factor(educ), y = renda)) +  
  geom_boxplot(fill = "lightblue") +  
  labs(title = "Distribuição da Renda por Anos de Escolaridade",  
       x = "Anos de Escolaridade",  
       y = "Renda Anual (R$)") +  
  theme_minimal()
```



ANÁLISES INFERENCIAIS: BIVARIADA

4. Correlação de Spearman

A correlação de Spearman é uma medida não-paramétrica que avalia a força da associação entre duas variáveis classificando-as e calculando a correlação de Pearson entre esses postos.

A fórmula da correlação de Spearman é:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Onde:

- d_i é a diferença entre os postos das duas variáveis para cada observação,
- n é o número de observações.

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

```

# Correlação de Spearman#####
# Primeiro, vamos carregar o conjunto de dados
# e calcular a correlação de Pearson entre duas
# variáveis: mpg (milhas por galão) e
# wt (peso do carro).

# Carregar conjunto de dados
data(mtcars)

# Visualizar as primeiras linhas dos dados
head(mtcars)

```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

Essas variáveis vêm do conjunto de dados mtcars, que contém especificações técnicas de diferentes modelos de carros. CODEBOOK

- **mpg**: *Miles per gallon* (milhas por galão) — Representa a eficiência de combustível do carro, ou seja, quantas milhas ele pode percorrer com um galão de combustível. Quanto maior o valor, mais eficiente o carro em termos de consumo de combustível.
 - **cyl**: *Cylinders* (cilindros) — Refere-se ao número de cilindros no motor do carro. Carros com mais cilindros tendem a ter motores mais potentes, mas também consomem mais combustível.
 - **disp**: *Displacement* (deslocamento) — O volume total deslocado por todos os cilindros do motor, medido em polegadas cúbicas. Está relacionado à capacidade do motor e é uma medida de seu tamanho.
 - **hp**: *Horsepower* (potência) — A potência do motor, medida em cavalos de potência (HP). Uma medida de quão rápido e potente o carro pode ser.
 - **drat**: *Rear axle ratio* (relação do eixo traseiro) — Refere-se à relação entre as rotações do eixo de saída do motor e as rotações das rodas traseiras. Influencia a aceleração e a economia de combustível.
 - **wt**: *Weight* (peso) — O peso do carro em milhares de libras. Carros mais pesados tendem a ser menos eficientes em termos de consumo de combustível.
 - **qsec**: *1/4 mile time* (tempo de 1/4 de milha) — O tempo que o carro leva para percorrer um quarto de milha, medido em segundos. Quanto menor o valor, mais rápido o carro é.
 - **vs**: *Engine shape* (forma do motor) — Um indicador binário (0 ou 1) que refere-se ao tipo de motor. 0 representa um motor em V, e 1 representa um motor em linha.
 - **am**: *Transmission* (transmissão) — Tipo de transmissão do carro. 0 representa transmissão automática, e 1 representa transmissão manual.
 - **gear**: *Gears* (marchas) — Número de marchas na transmissão do carro.
 - **carb**: *Carburetors* (carburadores) — Número de carburadores no carro. Um carburador mistura ar e combustível para alimentar o motor.
- Essas variáveis são usadas frequentemente em análises para estudar a eficiência, desempenho e outras características dos carros, muitas vezes em contextos de consumo, desempenho ambiental ou inovação automotiva.

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Calcular correlação de Pearson
cor(mtcars$mpg, mtcars$wt)

# Neste caso, estamos calculando a correlação
# entre o consumo de combustível e o peso do carro.
# Espera-se uma correlação negativa, ou seja,
# quanto maior o peso, menor a
# eficiência do carro em termos de consumo.

# Se suspeitarmos que a relação não é linear,
# podemos utilizar a correlação de Spearman,
# que não exige linearidade entre as variáveis.

# Calcular correlação de Spearman
cor(mtcars$mpg, mtcars$wt, method = "spearman")
> cor(mtcars$mpg, mtcars$wt, method = "spearman")
[1] -0.886422
```

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Matriz de correlação
# Podemos calcular a correlação entre várias
# variáveis de uma vez, gerando uma matriz de
# correlação.

# Matriz de correlação entre algumas variáveis
#do conjunto de dados
cor(mtcars[, c("mpg", "wt", "hp", "qsec")])

# Para facilitar a interpretação, é possível visualizar a correlação
# utilizando um gráfico de calor.

# Instalar e carregar a biblioteca necessária para visualização
install.packages("corrplot")
library(corrplot)

# Criar a matriz de correlação
matriz_cor ← cor(mtcars)                                > cor(mtcars[, c("mpg", "wt", "hp", "qsec")])
                                                               mpg          wt          hp          qsec
mpg      1.00000000 -0.8676594 -0.7761684  0.4186840
wt       -0.8676594  1.0000000  0.6587479 -0.1747159
hp       -0.7761684  0.6587479  1.0000000 -0.7082234
qsec     0.4186840 -0.1747159 -0.7082234  1.0000000
|
```

ANÁLISES INFERENCIAIS: BIVARIADA

Para facilitar a interpretação, é possível visualizar a correlação
utilizando um gráfico de calor.

```
# Instalar e carregar a biblioteca necessária para visualização
```

```
install.packages("corrplot")
```

```
library(corrplot)
```

```
# Criar a matriz de correlação
```

```
matriz_cor ← cor(mtcars)
```

```
# Plotar o gráfico de correlação
```

```
corrplot(matriz_cor, method = "color",
          type = "upper", tl.col = "black")
```

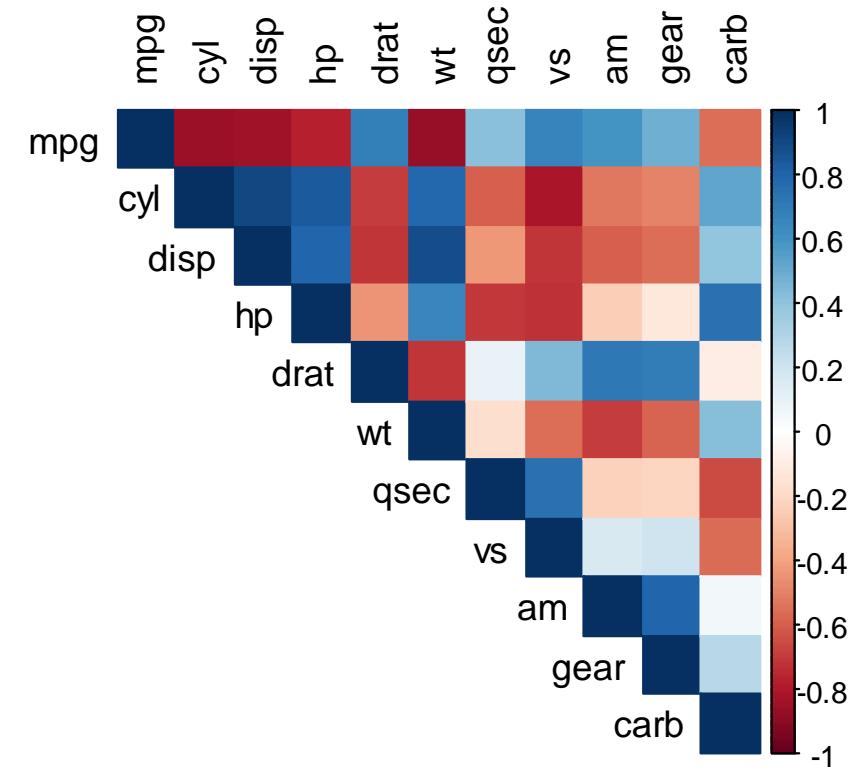
Esse gráfico mostra a força e a direção

da correlação entre as variáveis.

As cores indicam a intensidade da correlação,

com azul representando correlação positiva e

vermelho correlação negativa.



ANÁLISES INFERENCIAIS: BIVARIADA

5. Correlação de Kendall (Tau-b)

A correlação de Kendall mede a associação entre duas variáveis ao comparar o número de pares concordantes e discordantes. É uma outra medida não-paramétrica.

A fórmula é:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_0 - n_t)(n_0 - n_u)}}$$

Onde:

- n_c é o número de pares concordantes,
- n_d é o número de pares discordantes,
- n_0 é o número total de pares possíveis,
- n_t e n_u são correções para empates nas variáveis.

Essas fórmulas são especialmente úteis quando os dados não seguem uma relação linear ou apresentam valores discrepantes (outliers). A correlação de Spearman e a de Kendall são mais robustas para esses

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

```
#Correlação de Kendall#####
#
# (Confiança Social e Participação Religiosa)
# Vamos agora criar um exemplo onde avaliamos a correlação
# entre a confiança nas instituições sociais e a frequência à igreja,
# medidos em escalas ordinais.

# Criar uma base de dados fictícia
set.seed(456)
dados_conf_religiao ← data.frame(
  confiança_instituições = sample(1:5, 100, replace = TRUE),
  # 1 = Nenhuma confiança, 5 = Muita confiança
  freq_igreja = sample(1:5, 100, replace = TRUE)
  # 1 = Nunca, 5 = Todo fim de semana
)
# Visualizar as primeiras linhas da base de dados
head(dados_conf_religiao)
```

	confiança_instituições	freq_igreja
1	5	5
2	5	5
3	3	2
4	5	1
5	4	2
6	3	4

> |

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Calcular a correlação de Kendall entre confiança nas
#instituições e frequência à igreja
cor(dados_conf_religiao$confiança_instituições,
    dados_conf_religiao$freq_igreja, method = "kendall")
> cor(dados_conf_religiao$confiança_instituições,
+      dados_conf_religiao$freq_igreja, method = "kendall")
[1] -0.05724389
```

Neste exemplo fictício, estamos analisando a correlação
entre a confiança nas instituições sociais e a frequência
à igreja. Esses dados são frequentemente utilizados para
investigar a relação
entre crenças religiosas e confiança na sociedade.



ANÁLISES INFERENCIAIS

3º Regressão linear

REGRESSÃO LINEAR



- O modelo de regressão linear simples explica uma variável (y) com base em modificações em outra variável (x).
- Ou seja, é usado para avaliar a relação entre duas variáveis.

*Estamos interessados em explicar y , através da observação de x .
Ou seja, queremos estudar alterações em Y com base nas variações de x .

Exemplos:

X= Anos de escolaridade / X= Tipo de escola frequentada / X=Networkin

ar?

X= Denominação religioso/ X=Frequência religiosa/ X=Literalismo bíblico

RESUMO



INTRODUÇÃO A REGRESSÃO

1. Gráfico de dispersão
2. Regressão linear simples
3. Análise de resíduos



REGRESSÃO LINEAR MÚLTIPLA

1. Teste
2. Condicionantes
3. Apresentação gráfica

REGRESSÃO LINEAR SIMPLES

- A mensuração do teste ocorre através de uma estimativa, que assim como outros testes, tem uma variação que vai de -1 a +1, passando por 0.
- Quanto mais próximo de 0,00 menor é a evidência de correlação. Valores próximos de -1 indicam correlação negativa e próximos de 1 revelam correlação positiva.
- Na prática esse é um modelo estatístico que busca explorar a relação entre uma variável dependente e uma variável independente*
 - **Em modelos lineares a variável dependente sempre terá que ser numérica**

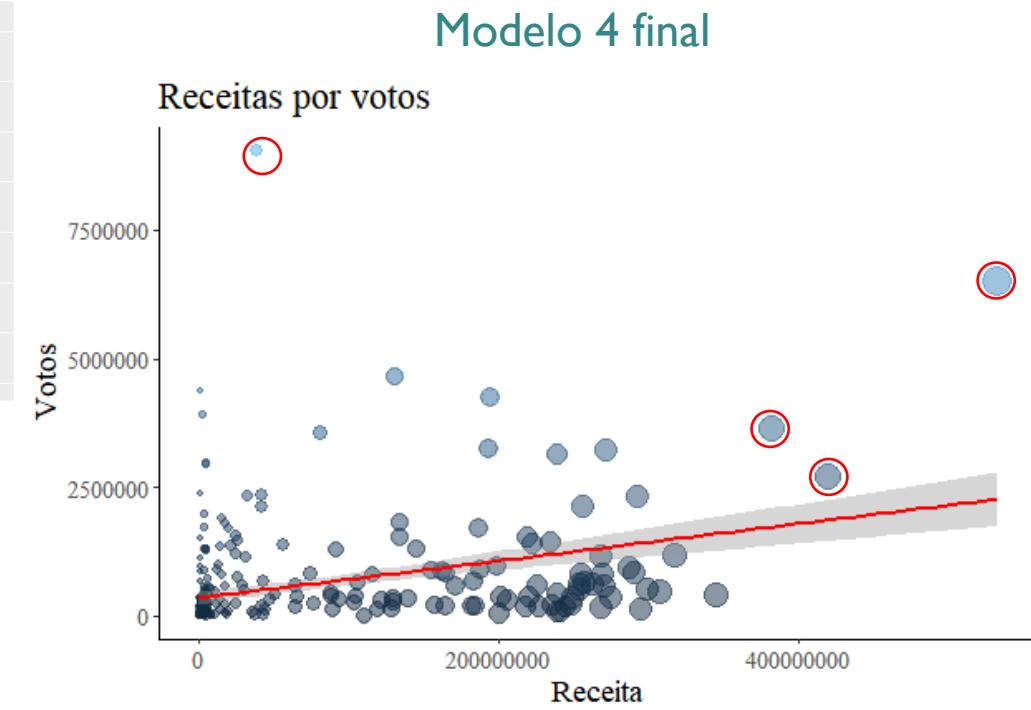
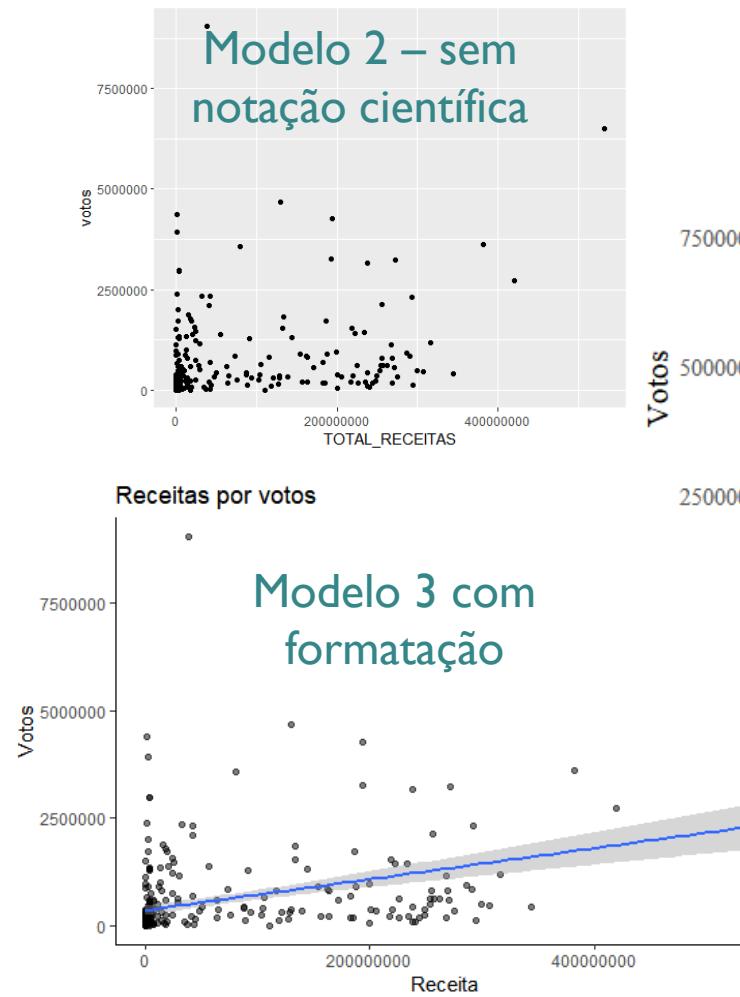
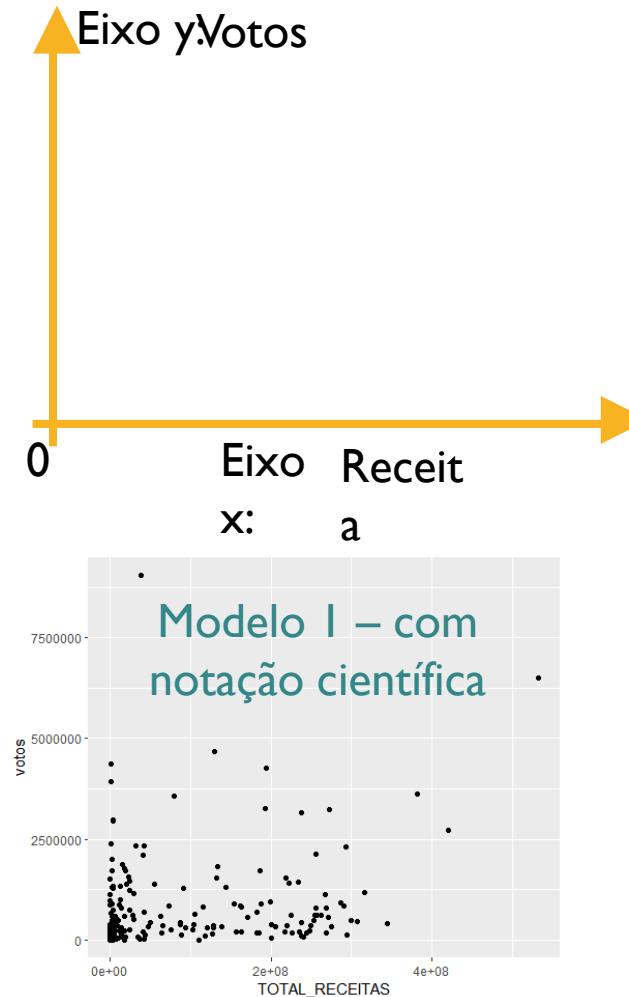
y	x
Variável Dependente	Variável Independente
Variável Explicada	Variável Explicativa
Variável de Resposta	Variável de Controle
Variável Prevista	Variável Previsora

! Existem alguns pressupostos que teremos que responder, para saber se o modelo de regressão linear é adequado para os dados.

- A melhor forma de entender se há relação entre as variáveis é através de um gráfico de dispersão.

REGRESSÃO LINEAR SIMPLES – NA PRÁTICA

As análises serão realizadas utilizando o **banco** de candidatos eleitos e não eleitos para o senado no ano de 2018. Uma boa forma de avaliar se existe relação entre as variáveis é através de um gráfico de dispersão.



REGRESSÃO LINEAR SIMPLES – NA PRÁTICA

Saída do modelo usando a função

`summary`

```
Call:  
lm(formula = log.votos ~ log.receitas, data = sen2018)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-3.9274 -0.8677 -0.0294  0.9192  3.7049  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 7.43396   0.34037  21.84 <0.000000000000002 ***  
log.receitas 0.30823   0.02228  13.84 <0.000000000000002 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 1.366 on 300 degrees of freedom  
(9 observations deleted due to missingness)  
Multiple R-squared:  0.3896, Adjusted R-squared:  0.3876  
F-statistic: 191.5 on 1 and 300 DF, p-value: < 0.0000000000000022
```

Resíduos: servem para indicar pontos influentes (distantes) e a tendência geral de acerto do modelo

Coeficientes: O intercepto de -2,29 é estatisticamente significativo ($p < 0.001$) e indica que um candidato com log de gasto 0 teria o log de votação de 7,43. O coeficiente de log.vgasto é estatisticamente significativo e indica que a cada ponto adicionado nesse medida ocorre a elevação de 0,30 na votação.

Legenda dos níveis de significância das estrelinhas

Estatísticas de ajuste do modelo: o R-quadrado indica a proporção de variação na variável dependente que pode ser explicada pela(s) variável(es) independente(s). O R-quadrado tem como valor máximo 1, portanto **cerca de 38% dos votos são explicados pela variação das receitas.**

Saída do modelo usando a função

`tab_model`

```
tab_model(Model1.1, show.ci = F, auto.label = T, show.se = T,  
          collapse.se = T, wrap.labels = 60, p.style = "stars")
```

Predictors	log.votos
(Intercept)	7.43 *** (0.34)
log receitas	0.31 *** (0.02)
Observations	302
R ² / R ² adjusted	0.390 / 0.388

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

ITENS QUE COMPÕE O TAB_MODEL

- `show.se` = (mostra o desvio padrão) se não colocarmos ele não aparece
- `collapse.se` = (esse item vai juntar “colapsar” o desvio padrão Std e a estimativa, isto é, o valor do desvio irá aparecer logo em baixo do valor da estimativa)
- `auto.label` = (A única função desse item é separar a variável da categoria que ela está testando. Por exemplo, a variável é faixa etária e está testando para idoso, na saída irá aparecer Faixa etária [idoso]).
- `Show.ci` = (Mostra o intervalo de confiança (Ci)) devemos pedir sempre que não apareça, usando o `F`
- Quando pedimos os coeficientes padronizados de beta, podemos colocar no `tab_model` essa função (`show.std = T`), para verificar o desvio padrão do beta
- Por fim, os estilos podem ser os seguintes: “numeric”, “stars”, “numeric_stars”, “scientific”, “scientific_stars”

DIAGNÓSTICOS DE RESÍDUOS

- A análise de regressão pode ser influenciada por pontos influentes (outliers) que podem fazer com que os resíduos se elevem bastante;
- Além disso, algumas exigências quanto à distribuição residual precisam ser confirmadas
- Isso pode ser feito construindo uma planilha com os valores preditos e os erros para cada caso a partir do nosso modelo
- Comando cbind que combina linhas e colunas em uma base de dados

```
#Análise de resíduos
resid ← (cbind(sen2018$log.votos, predict(Model1.1),
                 residuals(Model1.1)))
view(resid)
```

The screenshot shows the RStudio interface with three data frames:

- resid**: A data frame with columns V1, V2, and V3. The first row (index 252) has values 11.349700, 12.585256, and -3.9274315. The value 252 is circled in green.
- candidato**: A data frame with columns SQ_CANDIDATO, SIGLA_UF, and NOME_CANDIDATO. The first row (index 252) has values 230000601676, RR, and ROMERO JUC \diamond FILHO.
- resultados**: A data frame with columns votos, TOTAL_RECEITAS, and DESC_SIT_TOT_TURNO. The first row (index 84940) has values 2250000 and N \diamond O ELEITO.

R\$2.250.000

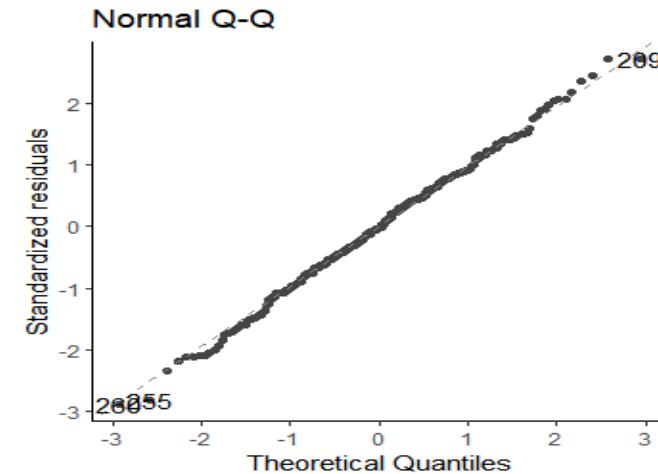
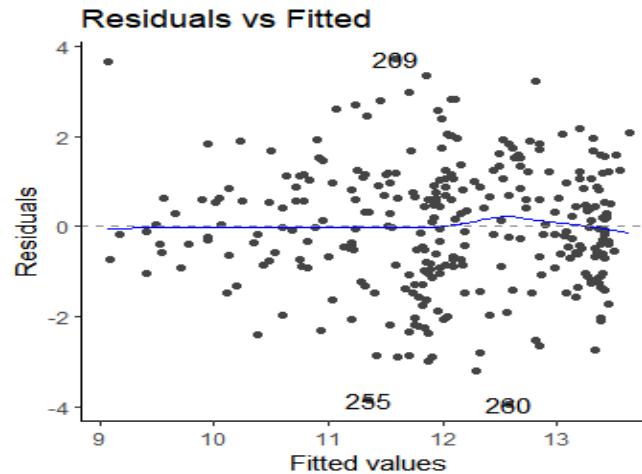
SLIDE OCULTO DE QM GANHOU

SQ_CANDIDATO	SIGLA_UF	NOME_CANDIDATO	NUMERO_PARTIDO	SIGLA_PARTIDO	DESCRICAO_COR_RACA	DESCRICAO_OCUPACAO	DESC_SIT_TOT_TURNO	SITUACAO_RELEICAO	DES_SITUACAO_CANDIDATURA	votos	TOTAL_RECEITAS
2,3E+11	RR	FRANCISCO DE ASSIS RODRIGUES	25	DEM	PARDA	ENGENHEIRO	ELEITO	N	APTO	111466	891508,3
2,3E+11	RR	ANTÔNIO MECIAS PEREIRA DE JESUS	10	PRB	PRETA	ADMINISTRADOR	ELEITO	N	APTO	85366	2380209

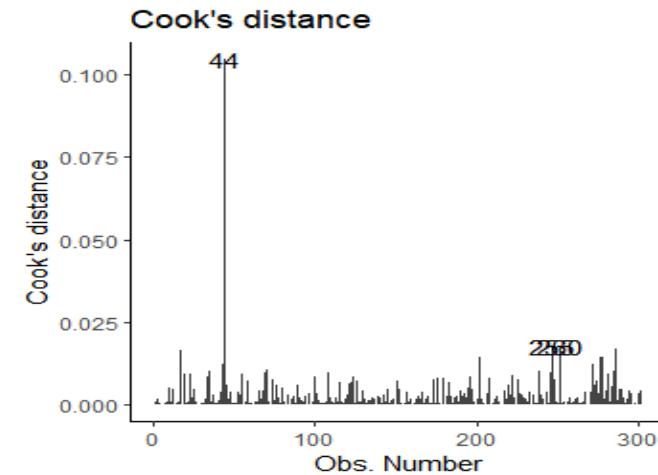
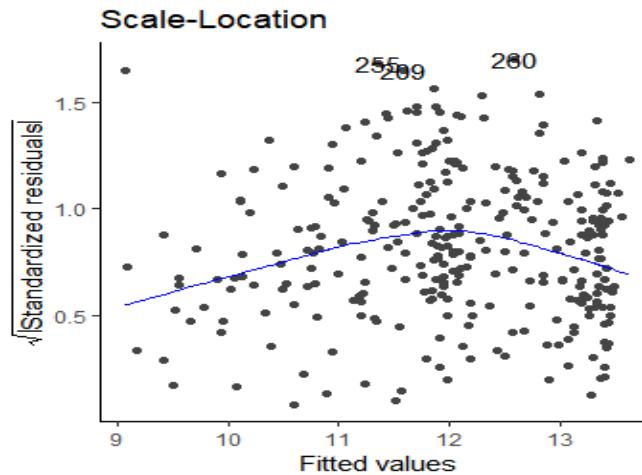
ANÁLISE DOS RESÍDUOS

Outra forma de olhar para os resíduos é através de gráficos.

1º resíduos pelos valores ajustados,
linearidade,
aproximadamente horizontal útil para testarmos a independência entre valores preditos e resíduos



3º Homocedasticidade -
não pode ter padrão triangular



2º Distribuição normal,
Os pontos dem estar deve estar em cima da linha
O Normal QQ plot nos ajuda a verificar essa exigência ao exibir no eixo horizontal a distribuição esperada em uma distribuição normal e no vertical os resíduos padronizados

4º Mostra pra gente se existem outliers

REGRESSÃO LINEAR MÚLTIPLA – TEORIA E PRÁTICA

Análise multivariada: É muito semelhante ao modelo simples, porém com a inserção de mais variáveis explicativas.



Geralmente uma questão social é explicada apenas por um fator?

Raça?

Predictors	log.votos	
	Estimates	p
(Intercept)	5.62 *** (1.67)	0.001
log receitas	0.28 *** (0.02)	<0.001
Idade	0.00 (0.01)	0.563
Sexo [MASCULINO]	-0.05 (0.21)	0.792
InSTRUÇÃO [Superior]	0.51 (0.92)	0.581
Raca [AMARELA]	0.66 (1.84)	0.720
Raca [BRANCA]	1.66 (1.30)	0.202
Raca [PARDA]	1.12 (1.31)	0.393
Raca [PRETA]	0.59 (1.32)	0.658
Observations	272	
R ² / R ² adjusted	0.450 / 0.433	

* p<0.05 ** p<0.01 *** p<0.001

Interpretação
I - As variáveis que acreditávamos ser importante foram de fato?

2- A capacidade preditiva do teste mudou?

CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

Iº Colinearidade

Definição: a perfeita colinearidade ocorre quando uma das variáveis independentes possui uma relação linear perfeita com outra (s) variável(s) independentes.

Problemas:

- Se as duas variáveis se relacionam intimamente provavelmente são redundantes, ou seja, medem a mesma coisa, sendo impossível distinguir o efeito de uma e outra.

Diagnóstico de Colinearidade

Iº Variance Inflation Factors

2ª Condition Index

CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

Iº Variance Inflation Factors

- Mede a inflação na estimativa de um parâmetro derivada da colinearidade existente entre os preditores
- VIF de 1 indica que não há correlação entre um determinado preditor e os demais incluídos no modelo, logo, não ocorre a inflação do seu coeficiente estimado
- Valores próximos de 4 inspiram cuidado e próximos de 10 revelam grave problema

```
> ols_vif_tol(Model2)
    Variables   Tolerance      VIF
1       log.receitas 0.93141736  1.073633
2             Idade 0.90301139  1.107406
3  SexoMASCULINO 0.96812327  1.032926
4 InstruçãoSuperior 0.99366053  1.006380
5     RacaAMARELA 0.50128312  1.994881
6     RacaBRANCA 0.01698194 58.886088
7     RacaPARDA 0.02266780 44.115445
8     RacaPRETA 0.03725117 26.844794
```

2º Condition Index

- Uma alternativa ao VIF é o Condition Index que se baseia no cálculo dos eigenvalues para cada variável, que basicamente indicam o quanto da sua variabilidade está relacionada a outra variável do modelo
- Eigenvalues próximos de “0” indicam forte colinearidade e o condition index é basicamente a raiz quadrada do maior eigenvalue para cada variável
- Condition index com valor maior do que 30 alertam para

```
> ols_eigen_cindex(Model2)
            Eigenvalue Condition Index
1 5.764162399 1.000000
2 1.002405328 2.397985
3 1.001597459 2.398952
4 1.000023916 2.400838
5 0.153974478 6.118483
6 0.041607740 11.770124
7 0.029058883 14.084085
8 0.005684059 31.844822
9 0.001485739 62.286920
```

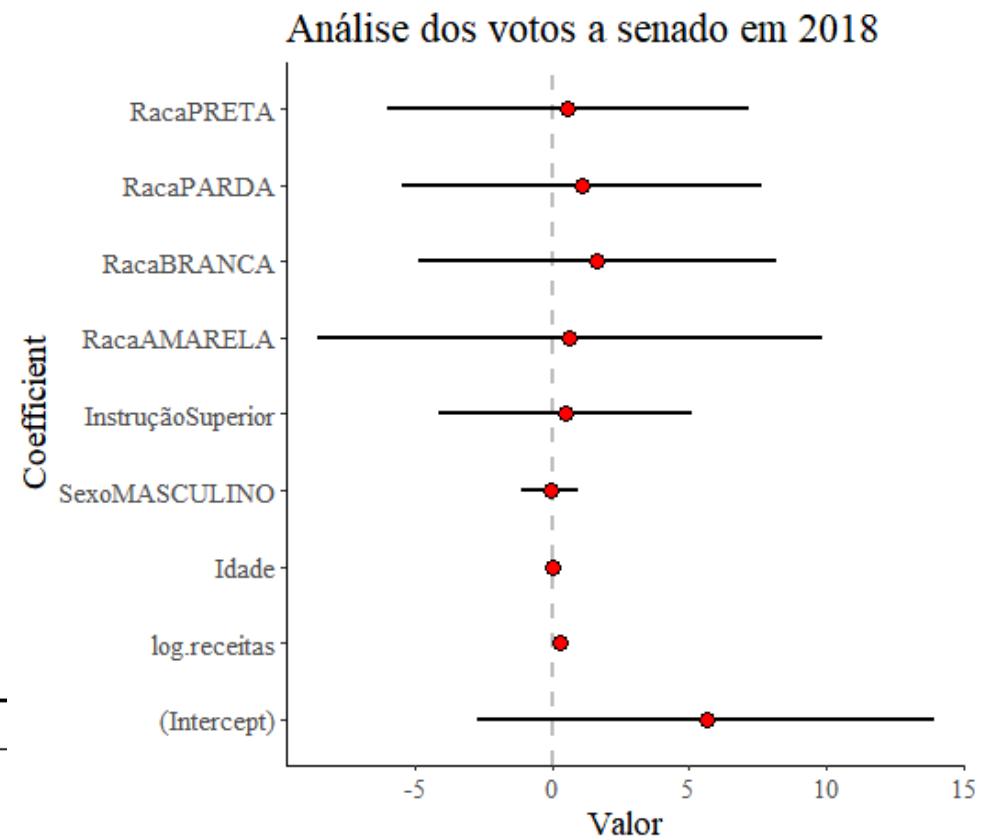
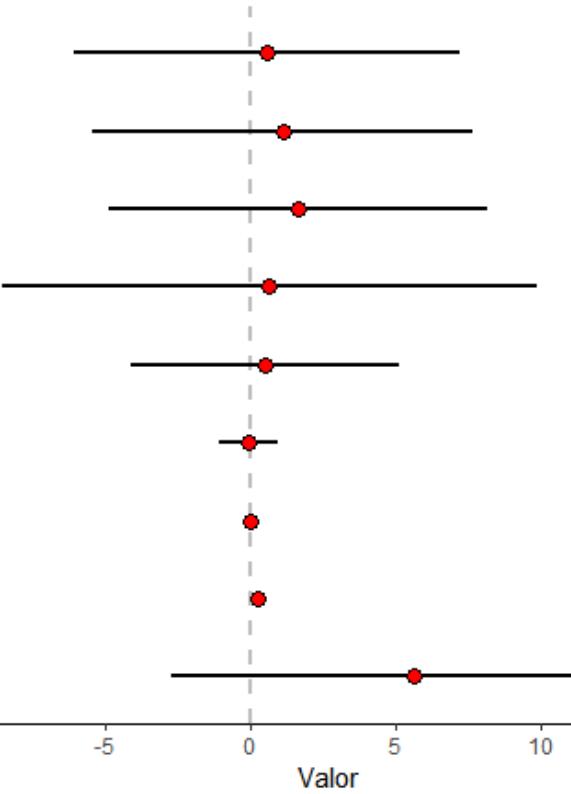
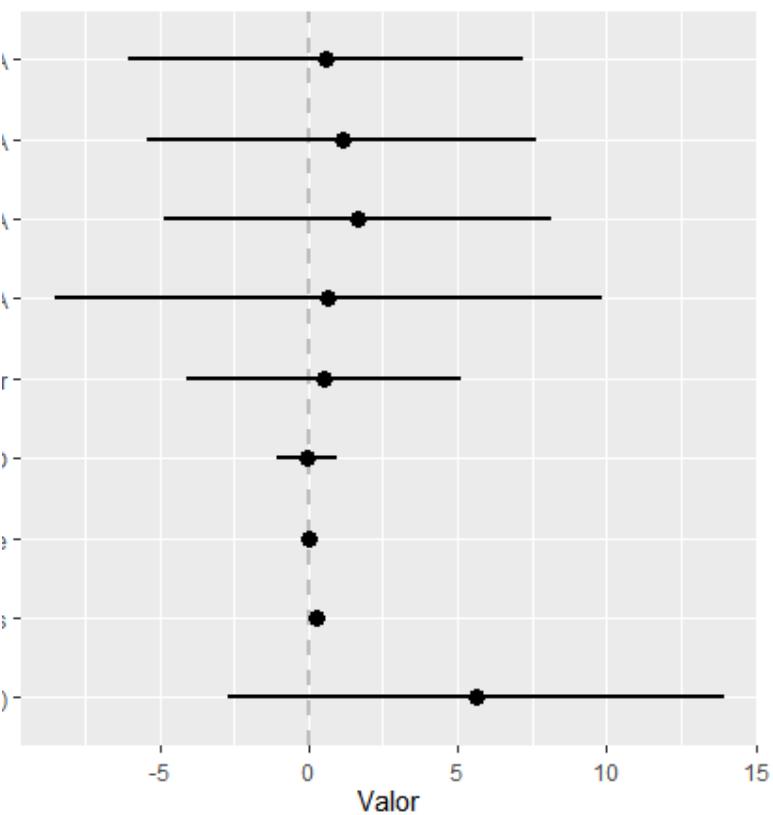
CONDICIONANTES PARA O TESTE DE REGRESSÃO LINEAR MÚLTIPLA

Diante da constatação de que existe multicolinearidade entre os preditores algumas estratégias podem ser adotadas.

- Aumentar o tamanho da amostra nos casos em que existem poucos casos e muitas variáveis no modelo.
- Combinar preditores em alguma espécie de indicador ou índice.
- Combinar as raças em uma única ou em duas como branca e parda
- Excluir variáveis redundantes.

APRESENTAÇÃO GRÁFICA DO TESTE DE REGRESSÃO

O Coefplot é um tipo de gráfico que pode ser utilizado para apresentar graficamente os resultados de um teste de regressão



COMO IDENTIFICAR ERROS MAIS FACILMENTE NO R?

1- O ambiente de programação sempre ajuda! Porém nem sempre o erro que aparece deve ser o buscado!

2-Mas se você não buscar o erro indicado, como vai identificar o problema? Seguindo um passo a passo:

- Trabalhe apenas com os dados que você precisa para realizar o teste. Se você não precisa de 250 variáveis, então limpe a base deixando apenas o que for necessário.
- Siga o passo acima, especialmente se for iniciante na programação
- Revise mentalmente (ou anotando) o teste que você precisa e as condicionantes necessárias
- Confira todas as variáveis e suas estruturas

3- Se depois de seguir todos os passos acima, ainda não conseguir resolver o erro você pode:

- Buscar o erro na internet
- Buscar outra forma de resolver seu problema.