Introdução ao R

Maurício Bueno

2021-10-21

Table of Contents

Esta disciplina tem como objetivo aprender a realizar análises estatísticas mais empregadas em pesquisas em psicologia, no ambiente de programação R. Inicialmente, vamos carregar (install.packages()) e ativar (library()) os pacotes que iremos utilizar na disciplina.

# Carregamento dos pacotes

if(!require(tidyverse)) install.packages("tidyverse")  
if(!require(psych)) install.packages("psych")  
if(!require(knitr)) install.packages("knitr")  
if(!require(kableExtra)) install.packages("kableExtra")  
if(!require(expss)) install.packages("expss")  
if(!require(janitor)) install.packages("janitor")  
if(!require(pander)) install.packages('pander')  
if(!require(arsenal)) install.packages('arsenal')  
if(!require(moments)) install.packages('moments')  
if(!require(readr)) install.packages('moments')  
if(!require(readxl)) install.packages('readxl')  
if(!require(descr)) install.packages("descr")  
# Ativação dos pacotes ====  
  
library(tidyverse)  
library(psych)  
library(knitr)  
library(kableExtra)  
library(expss)  
library(janitor)  
library(pander)  
library(arsenal)  
library(readxl)  
library(readr)

# Importar banco de dados

Inicialmente vamos trabalhar com o banco de dados dataset\_mapfre.csv, clique [aqui](https://docs.google.com/spreadsheets/d/1G0lQKeU0atMk3Q4wO5zz4HdvReRxKS68vx_Q0WbB37A/edit?usp=sharing) para baixá-lo. Esse banco de dados se refere ao artigo “[Sintomas de depressão e ansiedade em uma amostra representativa de universitários espanhóis, portugueses e brasileiros](https://drive.google.com/file/d/1qa0iCDm2DRvh3M6a2k7ENGahceDAlPIg/view?usp=sharing)”. Sugere-se a leitura desse artigo para a compreensão dos dados.

Dataset <- read.csv("dataset\_mapfre.csv",encoding="UTF-8",stringsAsFactors=TRUE)  
  
# OBS: Para que este comando funcione, é necessário que o arquivo csv esteja no diretório de trabalho do R.  
  
# Observar o banco de dados  
glimpse(Dataset)

## Rows: 1,957  
## Columns: 94  
## $ country <fct> SPAIN, SPAIN, SPAIN, SPAIN, SPAIN, SPAIN, SPAIN, SPAIN, ~  
## $ sex <fct> NA, NA, NA, NA, NA, NA, NA, M, M, M, M, M, M, M, M, M, M~  
## $ age <int> 18, NA, 26, NA, 20, NA, NA, 19, 21, 21, 23, 21, 22, 20, ~  
## $ grado <fct> "Educ.Infantil ~  
## $ curso\_ou\_ano <int> 1, 2, 4, 3, 1, NA, 3, 1, 3, 3, 4, 4, 4, 2, 4, 3, 3, 1, 1~  
## $ ceri\_1 <int> 1, 1, 1, 2, 1, 2, 1, 1, 1, 2, 3, 2, 1, 1, 2, 1, 2, 2, 1,~  
## $ ceri\_2 <int> 1, 2, 1, 2, 3, 2, 3, 1, 1, 3, 2, 3, 2, 2, 2, 1, 3, 3, 3,~  
## $ ceri\_3 <int> 2, 2, 2, 1, 2, 2, 3, 2, 1, 2, 1, 3, 2, 3, 2, 1, 3, 1, 2,~  
## $ ceri\_4 <int> 1, 1, 1, 2, 1, 1, 2, 1, 1, 3, 3, 3, 1, 2, 3, 1, 2, 1, 2,~  
## $ ceri\_5 <int> NA, 1, 2, 1, NA, 2, 2, 1, 1, 3, 1, 2, 2, 1, 2, 1, 2, 1, ~  
## $ ceri\_6 <int> 2, 2, 1, 2, 1, 1, 2, 1, 2, 1, 4, 3, 3, 2, 3, 2, 2, 2, 1,~  
## $ ceri\_7 <int> 1, 1, 1, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 1,~  
## $ ceri\_8 <int> 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1, 1,~  
## $ ceri\_9 <int> 1, 2, 2, 2, 3, 3, 4, 2, 1, 4, 4, 4, 3, 3, 2, 3, 3, 2, 2,~  
## $ ceri\_10 <int> 1, 1, 1, 1, 2, 2, 1, 1, 1, 4, 2, 3, 2, 1, 2, 1, 2, 1, 1,~  
## $ cerm\_1 <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1, 3, 1, 1, 1, 1,~  
## $ cerm\_2 <int> 2, 2, 1, 1, 3, 2, 2, 1, 1, 3, 1, 3, 2, 3, 2, 1, 1, 1, 1,~  
## $ cerm\_3 <int> 1, 1, 1, 1, 1, 2, 4, 1, 1, 2, 2, 2, 1, 1, 2, 1, 1, 1, 2,~  
## $ cerm\_4 <int> 1, 1, 1, 1, 2, 1, 4, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1,~  
## $ cerm\_5 <int> 1, 1, 1, 2, 2, 1, 2, 1, 1, 2, 1, 1, 1, 1, 3, 1, 1, 2, 1,~  
## $ cerm\_6 <int> 2, 1, 1, 2, 1, 1, 2, 1, 2, 2, 3, 3, 2, 2, 1, 2, 1, 2, 1,~  
## $ cerm\_7 <int> 1, 1, 1, 1, 1, 2, 2, 1, 1, 3, 2, 1, 2, 2, 2, 1, 1, 1, 1,~  
## $ cerm\_8 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1,~  
## $ cerm\_9 <int> 3, 3, 2, 2, 4, 3, 3, 2, 3, 4, 4, 4, 3, 3, 2, 2, 3, 4, 2,~  
## $ cerm\_10 <int> 2, 1, 1, 1, 1, 2, 1, 1, 1, 4, 2, 3, 3, 1, 2, 1, 1, 1, 2,~  
## $ cybvic\_1 <int> 1, 1, 1, 2, 1, 1, 2, 2, 2, 2, 1, 1, NA, 1, 1, 1, 1, 1, 1~  
## $ cybvic\_2 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_3 <int> 1, 1, 2, 2, 1, 2, 1, 3, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_4 <int> 1, 2, 3, 2, 2, 1, 3, 2, 1, 2, 3, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_5 <int> 1, 1, 1, 2, 2, 2, 2, 2, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_6 <int> 1, 1, 1, 2, 1, 3, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_7 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_8 <int> 1, 1, 1, 2, 1, 2, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybvic\_9 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 1, 1,~  
## $ cybvic\_10 <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,~  
## $ cybagress\_1 <int> 1, 1, 1, 1, 1, 1, 2, 1, 2, 3, 1, 1, NA, 1, 1, 1, 1, 1, 1~  
## $ cybagress\_2 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_3 <int> 1, 1, 1, 1, 2, 3, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_4 <int> 1, 1, 1, 2, 1, 1, 2, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_5 <int> 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_6 <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 4, 1, 1, 2, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_7 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 1, 1, 1, 1, 1, 1, 1, 1~  
## $ cybagress\_8 <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_9 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ cybagress\_10 <int> 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~  
## $ bai\_1 <int> 0, 0, 1, 0, 0, 1, 1, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_2 <int> 0, 0, 1, 1, 1, 0, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_3 <int> 0, 0, 0, 0, 0, 0, 2, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_4 <int> 0, 0, 0, 1, 0, 2, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_5 <int> 0, 1, 0, 1, 0, 2, 2, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_6 <int> 0, 0, 0, 0, 0, 0, 0, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_7 <int> 0, 1, 0, 0, 0, 1, 0, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_8 <int> 0, 0, 0, 1, 0, 1, 1, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_9 <int> 0, 0, 0, 0, 0, 0, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_10 <int> 0, 0, 1, 1, 1, 1, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_11 <int> 0, 0, 0, 0, 0, 0, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_12 <int> 0, 0, 0, 1, 0, 2, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_13 <int> 0, 0, 0, 1, 0, 1, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_14 <int> 0, NA, 0, 1, 0, 0, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0~  
## $ bai\_15 <int> 0, 0, 0, 0, 0, 0, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_16 <int> 0, 0, 0, 0, 2, 0, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_17 <int> 0, 0, 0, 0, 2, 1, 3, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_18 <int> 0, 0, 0, 1, 2, 0, 1, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_19 <int> 0, 0, 0, 0, 2, 0, 2, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_20 <int> 0, 0, 0, 0, 2, 1, 1, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_21 <int> 0, 0, 0, 0, 1, 1, 1, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bdi\_1 <int> 0, 0, 0, 0, 0, 0, 0, 1, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_2 <int> 0, 0, 1, 0, 1, 0, 1, 1, NA, NA, 1, 1, 0, 0, 0, 0, 1, 0, ~  
## $ bdi\_3 <int> 0, 0, 1, 0, 0, 0, 0, 1, NA, NA, 1, 0, 0, 0, 0, 0, 1, 0, ~  
## $ bdi\_4 <int> 0, 0, 1, 0, 0, 0, 0, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bdi\_5 <int> 0, 0, 1, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0, 1, 0, 0, 0, 0, ~  
## $ bdi\_6 <int> 0, 0, 0, 0, 0, 0, 0, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bdi\_7 <int> 0, 0, 2, 0, 0, 0, 0, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bdi\_8 <int> 0, 0, 3, 1, 1, 0, 1, 2, NA, NA, 1, 1, 0, 1, 1, 0, 0, 0, ~  
## $ bdi\_9 <int> 0, 0, 0, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_10 <int> 0, 0, 0, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0, 0, 1, 0, 0, 0, ~  
## $ bdi\_11 <int> 0, 0, 0, 0, 1, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_12 <int> 0, 0, 1, 1, 0, 1, 0, 3, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_13 <int> 0, 0, 0, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_14 <int> 0, 0, 0, NA, 0, 0, 0, 1, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bdi\_15 <int> 0, 0, 1, 1, 1, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 1, 0, 0, ~  
## $ bdi\_16 <int> 0, 0, 1, 0, 1, 0, 0, 0, NA, NA, 1, 0, 0, 0, 1, 1, 0, 1, ~  
## $ bdi\_17 <int> 0, 0, 0, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_18 <int> 0, 0, 1, 0, 1, 0, 0, 0, NA, NA, 1, 0, 0, 0, 0, 1, 0, 0, ~  
## $ bdi\_19 <int> 1, 2, 1, 0, 2, 0, 1, 0, NA, NA, 0, 1, 0, 0, 1, 1, 0, 0, ~  
## $ bdi\_20 <int> 0, 0, 1, 0, 1, 1, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_21 <int> 0, 0, 1, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~  
## $ bdi\_sum <int> 1, 2, 16, 3, 9, 2, 3, 9, NA, NA, 5, 3, 0, 2, 4, 4, 2, 1,~  
## $ number\_na\_bdi <int> 0, 0, 0, 1, 0, 0, 0, 3, 21, 21, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ bdi\_class <fct> minima, minima, leve, minima, minima, minima, minima, mi~  
## $ bai\_sum <int> 0, 2, 3, 9, 13, 14, 44, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0,~  
## $ number\_na\_bai <int> 0, 1, 0, 0, 0, 0, 0, 21, 21, 21, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ bai\_class <fct> minima, minima, minima, minima, leve, leve, grave, NA, N~

# Estatística descritiva

## Frequências

# usando a função tabyl do janitor  
# contar quantas pessoas de cada país.  
Dataset %>% # pegar o dataframe Dataset  
 tabyl(country) %>% # use a função tabyl para tabular as categorias de país  
 adorn\_totals() %>% # acrescente uma linha com os totais  
 pander() # utilize o pander para fazer a tabela

|  |  |  |
| --- | --- | --- |
| country | n | percent |
| BRAZIL | 315 | 0.161 |
| PORTUGAL | 426 | 0.2177 |
| SPAIN | 1216 | 0.6214 |
| Total | 1957 | 1 |

# contar por gênero e país (retirando NAs)  
Dataset %>%   
 filter(!is.na(sex)) %>%   
 tabyl(sex,country) %>%   
 adorn\_totals() %>%   
 pander()

|  |  |  |  |
| --- | --- | --- | --- |
| sex | BRAZIL | PORTUGAL | SPAIN |
| F | 166 | 223 | 825 |
| M | 149 | 203 | 384 |
| Total | 315 | 426 | 1209 |

# contar por gênero e país, acrescentando funções  
Dataset %>%   
 filter(!is.na(sex)) %>%   
 tabyl(country,sex) %>% # função de tabulação do janitor  
 adorn\_totals(c("row","col")) %>% # adiciona o N  
 adorn\_percentages("row") %>% # adiciona porcentagens  
 adorn\_pct\_formatting(rounding = "half up", digits = 0) %>% # multiplica por 100   
 adorn\_ns() %>% # mostra N e % juntas  
 pander() # melhora a visualização dos dados.

|  |  |  |  |
| --- | --- | --- | --- |
| country | F | M | Total |
| BRAZIL | 53% (166) | 47% (149) | 100% (315) |
| PORTUGAL | 52% (223) | 48% (203) | 100% (426) |
| SPAIN | 68% (825) | 32% (384) | 100% (1209) |
| Total | 62% (1214) | 38% (736) | 100% (1950) |

# adição de um qui-quadrado para a distribuição  
# para rodar essa análise tem que tirar os totais.  
  
# Dataset %>%   
 # filter(!is.na(sex)) %>%   
 # tabyl(country,sex) %>% # função de tabulação do janitor  
 # adorn\_totals(c("row","col")) %>% # adiciona o N  
 # chisq.test()  
 # pander() # melhora a visualização dos dados.

#### Exercício

Obter uma tabela de frequências dos cursos (grado) dos participantes.  
Obter uma tabela de frequencias dos cursos dos participantes por sexo, com totais.

Dataset %>%   
 tabyl(grado) %>%   
 adorn\_totals() %>%  
 pander()

|  |  |  |
| --- | --- | --- |
| grado | n | percent |
| Ade-Economía | 17 | 0.008687 |
| Adm | 24 | 0.01226 |
| Arquitetura | 10 | 0.00511 |
| Biologia | 2 | 0.001022 |
| Ciencias Ambientales | 15 | 0.007665 |
| Ciências e Tecnologia | 153 | 0.07818 |
| Ciências sociais | 1 | 0.000511 |
| Comunicação | 1 | 0.000511 |
| Design | 36 | 0.0184 |
| Direito | 17 | 0.008687 |
| Doble Ade-Eco | 1 | 0.000511 |
| Economia | 66 | 0.03373 |
| Economia & Direito | 52 | 0.02657 |
| Edificación | 21 | 0.01073 |
| Educ.Infantil | 493 | 0.2519 |
| Enfermería | 75 | 0.03832 |
| Engenharia | 112 | 0.05723 |
| Enología | 2 | 0.001022 |
| Filosofia | 5 | 0.002555 |
| Física | 2 | 0.001022 |
| Gestión.Pública | 1 | 0.000511 |
| Historia y Patrimonio Histórico | 96 | 0.04905 |
| Ingenieria Civil | 113 | 0.05774 |
| Ingenieria Informática | 105 | 0.05365 |
| Letras | 85 | 0.04343 |
| Matemática | 6 | 0.003066 |
| Medicina, Farmácia | 47 | 0.02402 |
| Pedagogia | 23 | 0.01175 |
| Psicologia | 29 | 0.01482 |
| Psicologia, Educação, SSocial, Desporto | 89 | 0.04548 |
| Química | 65 | 0.03321 |
| Relações internacionais | 7 | 0.003577 |
| Teologia | 16 | 0.008176 |
| Terapia ocupacional | 170 | 0.08687 |
| Total | 1957 | 1 |

Dataset %>%   
 filter(!is.na(sex)) %>%   
 tabyl(grado,sex) %>%   
 adorn\_totals(c("row","col")) %>%   
 adorn\_percentages("row") %>%   
 adorn\_pct\_formatting(rounding = "half up", digits = 0) %>%   
 adorn\_ns() %>%  
 pander()

|  |  |  |  |
| --- | --- | --- | --- |
| grado | F | M | Total |
| Ade-Economía | 71% (12) | 29% (5) | 100% (17) |
| Adm | 38% (9) | 63% (15) | 100% (24) |
| Arquitetura | 60% (6) | 40% (4) | 100% (10) |
| Biologia | 50% (1) | 50% (1) | 100% (2) |
| Ciencias Ambientales | 47% (7) | 53% (8) | 100% (15) |
| Ciências e Tecnologia | 29% (44) | 71% (109) | 100% (153) |
| Ciências sociais | 100% (1) | 0% (0) | 100% (1) |
| Comunicação | 100% (1) | 0% (0) | 100% (1) |
| Design | 75% (27) | 25% (9) | 100% (36) |
| Direito | 76% (13) | 24% (4) | 100% (17) |
| Doble Ade-Eco | 100% (1) | 0% (0) | 100% (1) |
| Economia | 42% (28) | 58% (38) | 100% (66) |
| Economia & Direito | 54% (28) | 46% (24) | 100% (52) |
| Edificación | 48% (10) | 52% (11) | 100% (21) |
| Educ.Infantil | 93% (455) | 7% (33) | 100% (488) |
| Enfermería | 81% (61) | 19% (14) | 100% (75) |
| Engenharia | 26% (29) | 74% (83) | 100% (112) |
| Enología | 100% (2) | 0% (0) | 100% (2) |
| Filosofia | 60% (3) | 40% (2) | 100% (5) |
| Física | 0% (0) | 100% (2) | 100% (2) |
| Gestión.Pública | 100% (1) | 0% (0) | 100% (1) |
| Historia y Patrimonio Histórico | 36% (34) | 64% (61) | 100% (95) |
| Ingenieria Civil | 27% (30) | 73% (82) | 100% (112) |
| Ingenieria Informática | 17% (18) | 83% (87) | 100% (105) |
| Letras | 36% (31) | 64% (54) | 100% (85) |
| Matemática | 50% (3) | 50% (3) | 100% (6) |
| Medicina, Farmácia | 81% (38) | 19% (9) | 100% (47) |
| Pedagogia | 100% (23) | 0% (0) | 100% (23) |
| Psicologia | 90% (26) | 10% (3) | 100% (29) |
| Psicologia, Educação, SSocial, Desporto | 92% (82) | 8% (7) | 100% (89) |
| Química | 54% (35) | 46% (30) | 100% (65) |
| Relações internacionais | 100% (7) | 0% (0) | 100% (7) |
| Teologia | 63% (10) | 38% (6) | 100% (16) |
| Terapia ocupacional | 81% (138) | 19% (32) | 100% (170) |
| Total | 62% (1214) | 38% (736) | 100% (1950) |

## Medidas de tendência central e de dispersão

A descrição dos dados geralmente é realizada pelas seguintes estatísticas: Média, desvio padrão, mediana, amplitude, mínimo, máximo, assimetria e curtose. Para calculá-las podemos usar funções específicas para cada uma ou funções que realizam um conjunto de análises descritivas.

### Funções específicas

## mean()  
mean(Dataset$age, na.rm = TRUE)

[1] 21.46034

## sd()  
sd(Dataset$age, na.rm = TRUE)

[1] 3.828075

## median()  
median(Dataset$age, na.rm = TRUE)

[1] 21

## min()  
min(Dataset$age, na.rm = TRUE)

[1] 17

## max()  
max(Dataset$age, na.rm = TRUE)

[1] 68

## range()  
range(Dataset$age, na.rm = TRUE)

[1] 17 68

## skewness()  
moments::skewness(Dataset$age, na.rm = TRUE)

[1] 5.018834

## kurtosis()  
moments::kurtosis(Dataset$age, na.rm = TRUE)

[1] 44.35922

### Funções que sumariam dados

# describe() {psych}  
describe(Dataset$age, na.rm = TRUE) %>% pander()

Table continues below

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | vars | n | mean | sd | median | trimmed | mad | min | max |
| **X1** | 1 | 1929 | 21.46 | 3.828 | 21 | 20.94 | 2.965 | 17 | 68 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | range | skew | kurtosis | se |
| **X1** | 51 | 5.015 | 41.31 | 0.08716 |

describe(Dataset$age, na.rm = TRUE) %>% kable()

vars

n

mean

sd

median

trimmed

mad

min

max

range

skew

kurtosis

se

X1

1

1929

21.46034

3.828075

21

20.93528

2.9652

17

68

51

5.014931

41.31324

0.0871594

# summary() {base}  
Dataset %>% select(age) %>% summary()

## age   
## Min. :17.00   
## 1st Qu.:19.00   
## Median :21.00   
## Mean :21.46   
## 3rd Qu.:22.00   
## Max. :68.00   
## NA's :28

# tableby() {arsenal}  
tableby(country ~ bdi\_sum + bai\_sum, # calcule as descritivas de bdi e bai por country   
 test = FALSE, # se FALSE, não faz teste de significância   
 data = Dataset) %>% # especificar base de dados  
 summary(text=TRUE) # tem que colocar text=TRUE senão aparecem uns códigos estranhos porque o R entende o espaço como código. Assim a tabela fica bacaninha.

##   
##   
## | | BRAZIL (N=315) | PORTUGAL (N=426) | SPAIN (N=1216) | Total (N=1957) |  
## |:------------|:--------------:|:----------------:|:--------------:|:--------------:|  
## |bdi\_sum | | | | |  
## |- N-Miss | 1 | 3 | 4 | 8 |  
## |- Mean (SD) | 10.895 (8.294) | 9.054 (7.727) | 8.859 (7.537) | 9.229 (7.736) |  
## |- Range | 0.000 - 41.000 | 0.000 - 53.000 | 0.000 - 57.000 | 0.000 - 57.000 |  
## |bai\_sum | | | | |  
## |- N-Miss | 0 | 2 | 3 | 5 |  
## |- Mean (SD) | 9.013 (8.403) | 7.915 (8.042) | 8.547 (8.057) | 8.485 (8.114) |  
## |- Range | 0.000 - 46.000 | 0.000 - 45.000 | 0.000 - 48.000 | 0.000 - 48.000 |

## mesma função mas para a variável sexo.  
tableby(sex ~ bdi\_sum + bai\_sum,  
 test = FALSE,  
 data = Dataset) %>%   
 summary(text=TRUE)

##   
##   
## | | F (N=1214) | M (N=736) | Total (N=1950) |  
## |:------------|:--------------:|:--------------:|:--------------:|  
## |bdi\_sum | | | |  
## |- N-Miss | 4 | 4 | 8 |  
## |- Mean (SD) | 9.588 (7.725) | 8.676 (7.737) | 9.244 (7.740) |  
## |- Range | 0.000 - 53.000 | 0.000 - 57.000 | 0.000 - 57.000 |  
## |bai\_sum | | | |  
## |- N-Miss | 0 | 5 | 5 |  
## |- Mean (SD) | 9.470 (8.446) | 6.813 (7.142) | 8.471 (8.082) |  
## |- Range | 0.000 - 48.000 | 0.000 - 44.000 | 0.000 - 48.000 |

## Pode descrever apenas uma variável em função de sexo  
tableby(sex ~ age, data = Dataset) %>% summary(text = TRUE)

##   
##   
## | | F (N=1214) | M (N=736) | Total (N=1950) | p value|  
## |:------------|:---------------:|:---------------:|:---------------:|-------:|  
## |age | | | | 0.155|  
## |- N-Miss | 8 | 16 | 24 | |  
## |- Mean (SD) | 21.556 (4.106) | 21.300 (3.308) | 21.461 (3.829) | |  
## |- Range | 17.000 - 68.000 | 17.000 - 48.000 | 17.000 - 68.000 | |

## Pode descrever somenta a variável contínua sem ser em função de alguma outra variável.  
tableby( ~ age + bdi\_sum + bai\_sum, data = Dataset, strata = country) %>% summary(text = TRUE)

##   
##   
## |country | | Overall (N=1957) |  
## |:--------|:------------|:----------------:|  
## |BRAZIL |age | |  
## | |- N-Miss | 22 |  
## | |- Mean (SD) | 22.826 (7.184) |  
## | |- Range | 17.000 - 68.000 |  
## | |bdi\_sum | |  
## | |- N-Miss | 1 |  
## | |- Mean (SD) | 10.895 (8.294) |  
## | |- Range | 0.000 - 41.000 |  
## | |bai\_sum | |  
## | |- Mean (SD) | 9.013 (8.403) |  
## | |- Range | 0.000 - 46.000 |  
## |PORTUGAL |age | |  
## | |- Mean (SD) | 20.434 (1.664) |  
## | |- Range | 18.000 - 23.000 |  
## | |bdi\_sum | |  
## | |- N-Miss | 3 |  
## | |- Mean (SD) | 9.054 (7.727) |  
## | |- Range | 0.000 - 53.000 |  
## | |bai\_sum | |  
## | |- N-Miss | 2 |  
## | |- Mean (SD) | 7.915 (8.042) |  
## | |- Range | 0.000 - 45.000 |  
## |SPAIN |age | |  
## | |- N-Miss | 6 |  
## | |- Mean (SD) | 21.491 (3.018) |  
## | |- Range | 18.000 - 48.000 |  
## | |bdi\_sum | |  
## | |- N-Miss | 4 |  
## | |- Mean (SD) | 8.859 (7.537) |  
## | |- Range | 0.000 - 57.000 |  
## | |bai\_sum | |  
## | |- N-Miss | 3 |  
## | |- Mean (SD) | 8.547 (8.057) |  
## | |- Range | 0.000 - 48.000 |

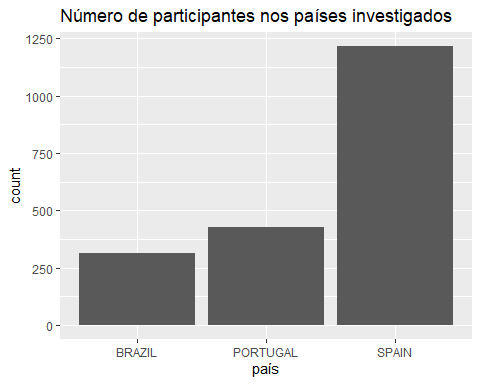
## Interação entre variáveis nominais  
  
tableby(interaction(sex, country) ~ bdi\_sum + bai\_sum,  
 test = FALSE,  
 data = Dataset) %>%   
 summary(text = TRUE)

##   
##   
## | | F.BRAZIL (N=166) | M.BRAZIL (N=149) | F.PORTUGAL (N=223) | M.PORTUGAL (N=203) | F.SPAIN (N=825) | M.SPAIN (N=384) | Total (N=1950) |  
## |:------------|:----------------:|:----------------:|:------------------:|:------------------:|:---------------:|:---------------:|:--------------:|  
## |bdi\_sum | | | | | | | |  
## |- N-Miss | 1 | 0 | 1 | 2 | 2 | 2 | 8 |  
## |- Mean (SD) | 12.158 (8.534) | 9.497 (7.812) | 9.464 (7.421) | 8.602 (8.047) | 9.106 (7.542) | 8.395 (7.537) | 9.244 (7.740) |  
## |- Range | 0.000 - 41.000 | 0.000 - 38.000 | 0.000 - 53.000 | 0.000 - 52.000 | 0.000 - 53.000 | 0.000 - 57.000 | 0.000 - 57.000 |  
## |bai\_sum | | | | | | | |  
## |- N-Miss | 0 | 0 | 0 | 2 | 0 | 3 | 5 |  
## |- Mean (SD) | 9.663 (8.708) | 8.289 (8.017) | 9.498 (8.449) | 6.159 (7.188) | 9.424 (8.402) | 6.580 (6.680) | 8.471 (8.082) |  
## |- Range | 0.000 - 46.000 | 0.000 - 44.000 | 0.000 - 45.000 | 0.000 - 42.000 | 0.000 - 48.000 | 0.000 - 42.000 | 0.000 - 48.000 |

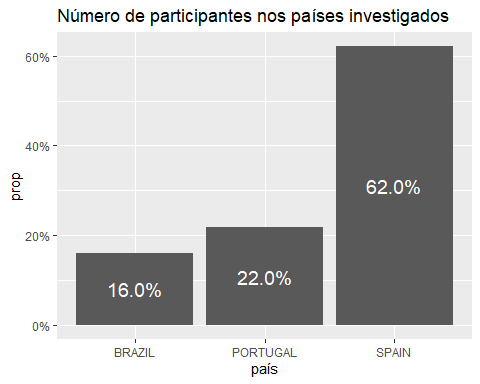
## Representações gráficas

### Gráficos de barras

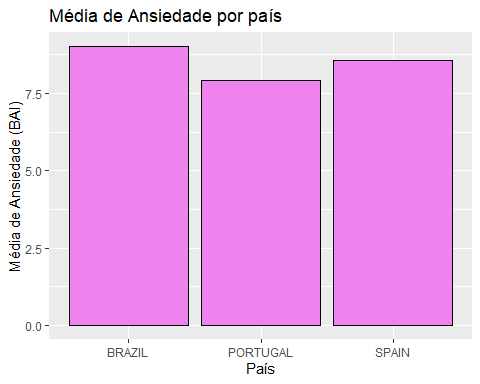
ggplot(Dataset, aes(x = country)) +   
 geom\_bar() +   
 labs(x = "país",  
 title = "Número de participantes nos países investigados")



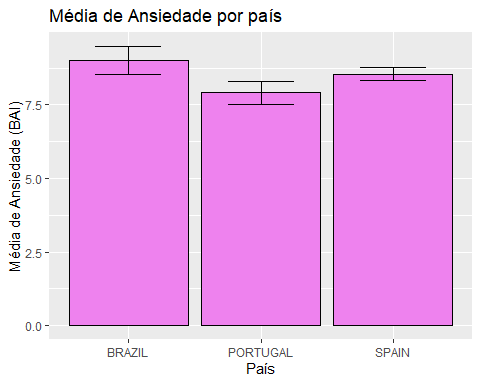
# alterando o eixo y para ser a % ao invés do N  
ggplot(Dataset, aes(x = country, y = ..prop.., group = 1)) +   
 geom\_bar(stat = "count") +  
 geom\_text(aes(label = scales::percent(round(..prop..,2)),  
 y = ..prop..),  
 stat = "count", color = "white",  
 size = 5, position = position\_stack(vjust = 0.5)) +  
 scale\_y\_continuous(labels = scales::percent\_format()) +  
 labs(x = "país",  
 title = "Número de participantes nos países investigados")



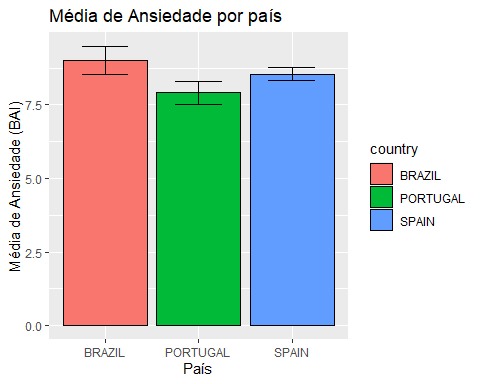
# Alterando argumentos  
ggplot(Dataset,   
 aes(x=country, y=bai\_sum)) +  
 geom\_bar(stat = "summary",   
 fun = mean,   
 fill = "violet",   
 color = "black",  
 na.rm = TRUE) +   
 labs(title = "Média de Ansiedade por país",  
 x = "País", y = "Média de Ansiedade (BAI)")



# adicionando outros elementos ao gráfico (barras de erro)  
ggplot(Dataset,   
 aes(x=country, y=bai\_sum)) +  
 geom\_bar(stat = "summary",   
 fun = mean,   
 fill = "violet",   
 color = "black",  
 na.rm = TRUE) +   
 stat\_summary(geom = "errorbar", fun.data = mean\_se,width = 0.5) +  
 labs(title = "Média de Ansiedade por país",  
 x = "País", y = "Média de Ansiedade (BAI)")

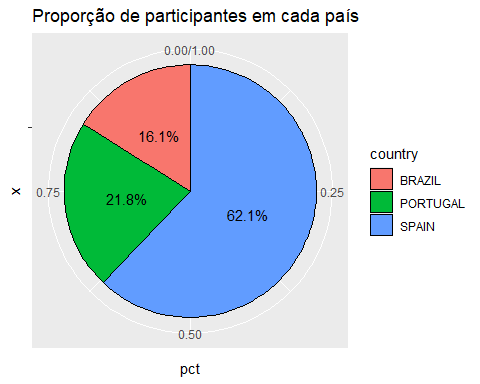


# se modificar a posição do argumento fill, para dentro de aes e com o critério country... o gráfico fica colorido por país.  
ggplot(Dataset,   
 aes(x=country, y=bai\_sum,fill = country,)) +  
 geom\_bar(stat = "summary",   
 fun = mean,   
 color = "black",  
 na.rm = TRUE) +   
 stat\_summary(geom = "errorbar", fun.data = mean\_se,width = 0.5) +  
 labs(title = "Média de Ansiedade por país",  
 x = "País", y = "Média de Ansiedade (BAI)")



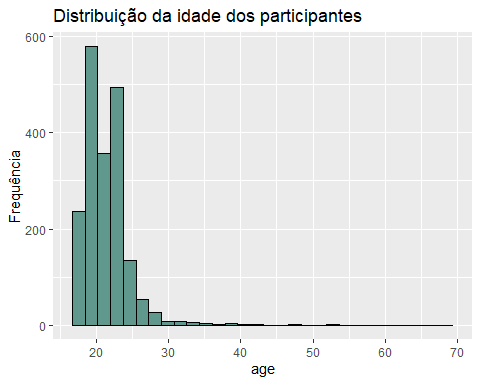
### Gráfico de setor (polar, pizza ou torta)

Dataset %>%   
 count(country) %>%   
 mutate(pct = n/sum(n)) %>%   
 ggplot(.,aes(x = "", y = pct, fill = country)) +   
 geom\_col(color = "black") +   
 geom\_text(aes(label = scales::percent(round(pct,3))),  
 position = position\_stack(vjust = 0.5)) +   
 coord\_polar(theta = "y") +   
 labs(title = "Proporção de participantes em cada país")

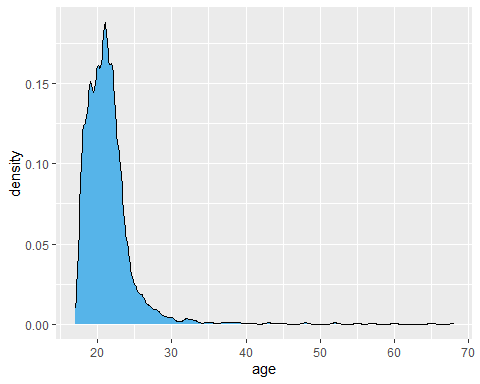


### Gráficos para descrição de distribuição ou variabilidade

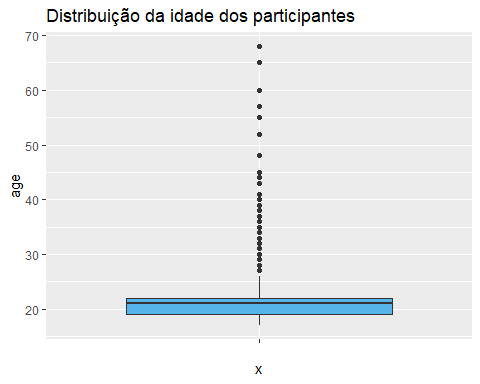
# Histograma  
ggplot(data = Dataset, aes(x = age)) +   
 geom\_histogram(bins = 30,color = "black", fill = "#61988E") +   
 labs(y = "Frequência",  
 title = "Distribuição da idade dos participantes")



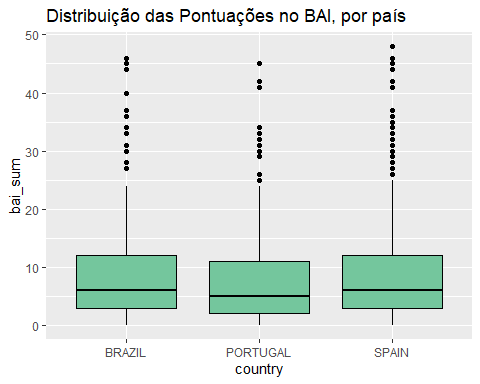
# Densidade  
ggplot(data = Dataset, aes(age)) +   
 geom\_density(fill = "#56b4e9") +   
 labs("Distribuição da idade dos participantes")



# Boxplot (diagrama de caixa e bigode)  
## Uma variável continua  
ggplot(data = Dataset, aes(y=age,x="")) +  
 geom\_boxplot(fill = "#56b4e9") +   
 labs(title = "Distribuição da idade dos participantes")



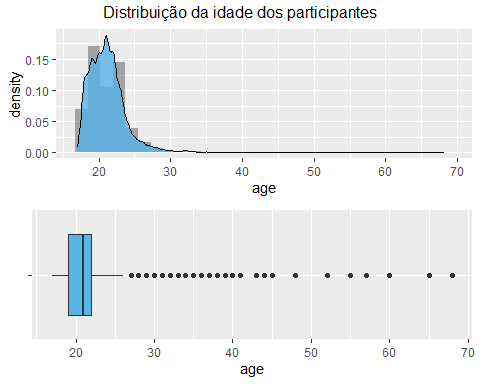
## uma variável continua e uma categórica (discreta)  
## no caso, bai\_sum por país.  
ggplot(Dataset,aes(x=country, y=bai\_sum)) +  
 geom\_boxplot(color = "black", fill = "#74c69d") +   
 labs(title = "Distribuição das Pontuações no BAI, por país")



names(Dataset)

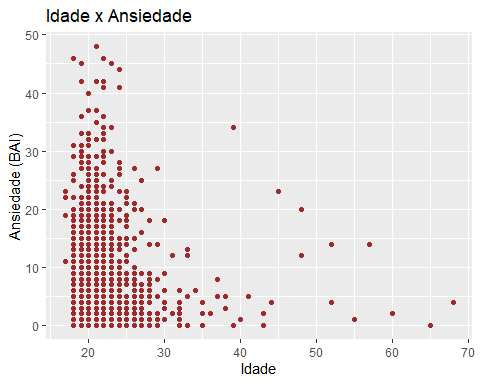
## [1] "country" "sex" "age" "grado"   
## [5] "curso\_ou\_ano" "ceri\_1" "ceri\_2" "ceri\_3"   
## [9] "ceri\_4" "ceri\_5" "ceri\_6" "ceri\_7"   
## [13] "ceri\_8" "ceri\_9" "ceri\_10" "cerm\_1"   
## [17] "cerm\_2" "cerm\_3" "cerm\_4" "cerm\_5"   
## [21] "cerm\_6" "cerm\_7" "cerm\_8" "cerm\_9"   
## [25] "cerm\_10" "cybvic\_1" "cybvic\_2" "cybvic\_3"   
## [29] "cybvic\_4" "cybvic\_5" "cybvic\_6" "cybvic\_7"   
## [33] "cybvic\_8" "cybvic\_9" "cybvic\_10" "cybagress\_1"   
## [37] "cybagress\_2" "cybagress\_3" "cybagress\_4" "cybagress\_5"   
## [41] "cybagress\_6" "cybagress\_7" "cybagress\_8" "cybagress\_9"   
## [45] "cybagress\_10" "bai\_1" "bai\_2" "bai\_3"   
## [49] "bai\_4" "bai\_5" "bai\_6" "bai\_7"   
## [53] "bai\_8" "bai\_9" "bai\_10" "bai\_11"   
## [57] "bai\_12" "bai\_13" "bai\_14" "bai\_15"   
## [61] "bai\_16" "bai\_17" "bai\_18" "bai\_19"   
## [65] "bai\_20" "bai\_21" "bdi\_1" "bdi\_2"   
## [69] "bdi\_3" "bdi\_4" "bdi\_5" "bdi\_6"   
## [73] "bdi\_7" "bdi\_8" "bdi\_9" "bdi\_10"   
## [77] "bdi\_11" "bdi\_12" "bdi\_13" "bdi\_14"   
## [81] "bdi\_15" "bdi\_16" "bdi\_17" "bdi\_18"   
## [85] "bdi\_19" "bdi\_20" "bdi\_21" "id"   
## [89] "bdi\_sum" "number\_na\_bdi" "bdi\_class" "bai\_sum"   
## [93] "number\_na\_bai" "bai\_class"

# Combinação de gráficos usando o grid.arrange do pacote gridExtra  
  
gridExtra::grid.arrange(  
   
# Gráfico 1  
ggplot(data = Dataset, aes(age)) +   
 geom\_histogram(aes(y=..density..), alpha = 0.5, position = "identity") +  
 geom\_density(alpha = 0.8, fill = "#56b4e9") +   
 labs("Distribuição da idade dos participantes"),  
# Gráfico 2  
ggplot(data = Dataset, aes(y=age,x="")) +  
 geom\_boxplot(fill = "#56b4e9") +   
 labs(x = "") +  
 coord\_flip(),  
top = "Distribuição da idade dos participantes"  
)

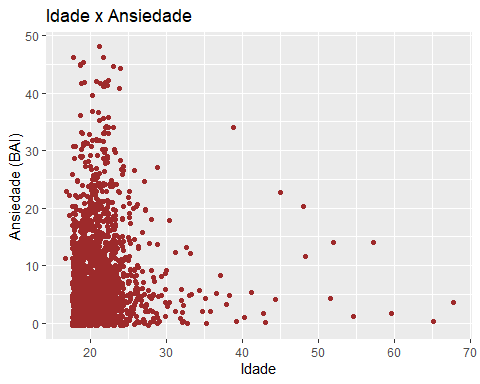


### Gráficos de pontos

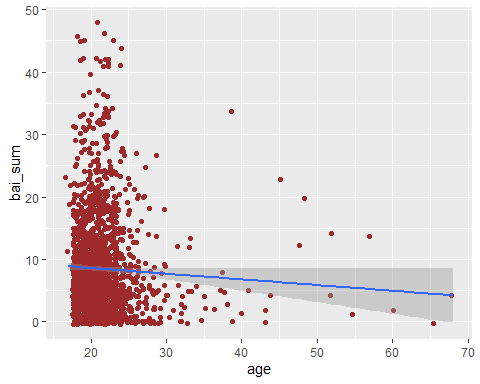
ggplot(Dataset, aes(x=age, y=bai\_sum)) +  
 geom\_point(color = "#9e2a2b") +  
 labs(title = "Idade x Ansiedade",  
 x = "Idade", y = "Ansiedade (BAI)")



ggplot(Dataset, aes(x=age, y=bai\_sum)) +  
 geom\_jitter(color = "#9e2a2b") +  
 labs(title = "Idade x Ansiedade",  
 x = "Idade", y = "Ansiedade (BAI)")



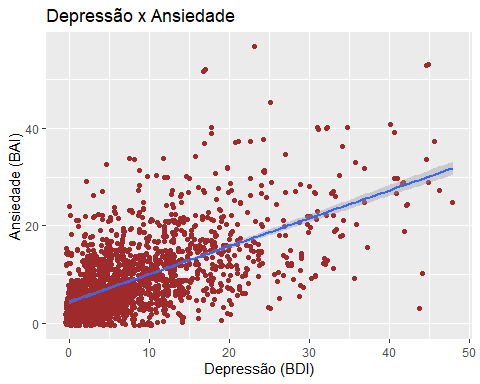
# com a reta de regressão  
ggplot(Dataset, aes(x=age, y=bai\_sum)) +  
 geom\_jitter(color = "#9e2a2b") +  
 geom\_smooth(method = "lm")



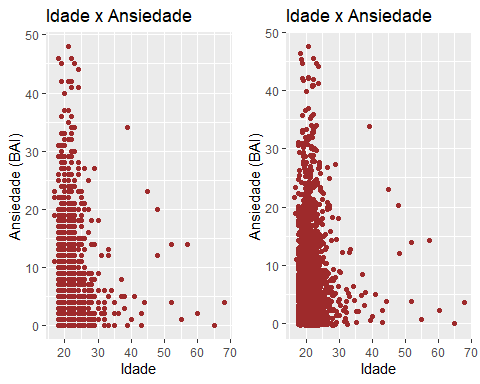
labs(title = "Idade x Ansiedade",  
 x = "Idade", y = "Ansiedade (BAI)")

## $x  
## [1] "Idade"  
##   
## $y  
## [1] "Ansiedade (BAI)"  
##   
## $title  
## [1] "Idade x Ansiedade"  
##   
## attr(,"class")  
## [1] "labels"

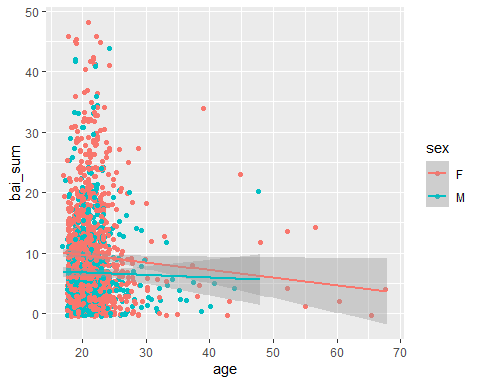
# ansiedade x depressão  
   
ggplot(Dataset, aes(x=bai\_sum, y=bdi\_sum)) +  
 geom\_jitter(color = "#9e2a2b") +  
 geom\_smooth(method = "lm") +  
 labs(title = "Depressão x Ansiedade",  
 x = "Depressão (BDI)", y = "Ansiedade (BAI)")



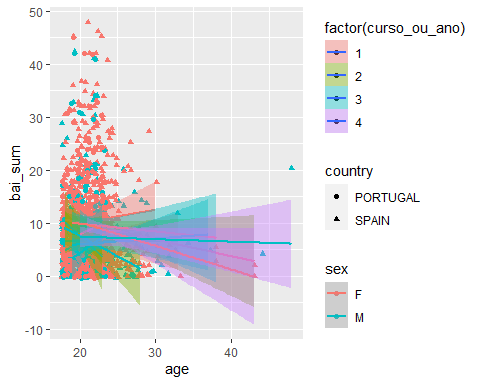
# colocar os dois gráficos juntos na mesma imagem  
  
gridExtra::grid.arrange(  
   
 #Gráfico 1  
   
 ggplot(Dataset, aes(x=age, y=bai\_sum)) +  
 geom\_point(color = "#9e2a2b") +  
 labs(title = "Idade x Ansiedade",  
 x = "Idade", y = "Ansiedade (BAI)"),  
  
 # Gráfico 2  
ggplot(Dataset, aes(x=age, y=bai\_sum)) +  
 geom\_jitter(color = "#9e2a2b") +  
 labs(title = "Idade x Ansiedade",  
 x = "Idade", y = "Ansiedade (BAI)"),  
nrow = 1 # sem esse comando, os gráficos ficam 2 linhas  
)



Dataset %>%   
filter(!is.na(sex)) %>%   
 ggplot(.,aes(x = age, y = bai\_sum, color = sex)) +  
 geom\_jitter() +  
 geom\_smooth(method = "lm")



Dataset %>%   
 filter(!is.na(sex), !is.na(curso\_ou\_ano)) %>%   
 ggplot(., aes(x = age, y=bai\_sum,  
 fill = factor(curso\_ou\_ano),  
 color = sex,  
 shape = country)) +   
 geom\_jitter() +  
 geom\_smooth(method = "lm")



# Testes de Qui-Quadrado

Existem três tipos de testes de qui-quadrado:

* de aderência: quando se deseja verificar as distribuições de probabilidades de cada categoria de uma variável em relação a um valor teórico esperado.
* de homogeneirdade: quando se deseja verificar se as distribuições das categorias são as mesmas para diferentes subpopulações de interesse.
* de independência: verficiar se duas variáveis categóricas são independentes

Para estas análises vamos usar o banco de dados Base CSV - ADHD 2020 after processing.csv:

ds\_selected <- read.csv("Base CSV - ADHD 2020 after processing.csv", stringsAsFactors=TRUE)  
glimpse(ds\_selected)

## Rows: 7,114  
## Columns: 19  
## $ id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~  
## $ public\_School <fct> public, public, public, public, public, public, pu~  
## $ city\_size <fct> medium, medium, medium, medium, big, big, medium, ~  
## $ region <fct> NE, NE, NE, NE, NE, NE, NE, NE, NO, NE, NE, NE, NE~  
## $ age <int> 9, 9, 8, 12, 9, 10, 8, 10, 12, 9, 5, 11, 10, 14, 1~  
## $ sex\_male <fct> female, male, male, male, male, female, male, male~  
## $ race\_white <fct> other, white, other, other, white, other, white, o~  
## $ married <fct> married, married, divorced, married, married, marr~  
## $ schooling <fct> primary, high\_or\_above, high\_or\_above, illiteracy,~  
## $ economic\_status <fct> DE, AB, C, C, DE, AB, C, AB, AB, AB, AB, C, C, C, ~  
## $ smoking <fct> no, no, no, no, no, no, no, no, no, no, no, no, no~  
## $ alcohol <fct> no, no, no, no, no, yes, yes, no, no, no, no, no, ~  
## $ scholar\_achievement <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~  
## $ snap\_parents\_only <fct> no, no, yes, yes, no, no, no, no, no, no, no, no, ~  
## $ age\_group <int> 1, 1, 1, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 3, 2, 1, 2,~  
## $ snap\_teachers\_only <fct> no, no, no, no, no, no, no, no, no, no, no, no, no~  
## $ adhd\_parent <fct> yes, no, yes, yes, yes, no, no, no, no, no, no, no~  
## $ adhd\_risk <fct> yes, yes, no, no, no, no, no, no, no, no, no, no, ~  
## $ psychostimulant <fct> no, no, no, no, no, no, no, no, no, no, no, no, no~

O primeiro passo para executar uma análise de qui-quadrado é montar uma tabela de referência cruzada. Para essa finalidade usaremos o pacote descr. Podemos visualizar a tabela ou simplesmente executar a análise.

# Para não ter que ativar o pacote "descr", podemos chamar apenas a função que desejamos usar, nesse caso a função CrossTable()  
# Para isso, usamos o comando descr::CrossTable()  
  
  
## Tabela de Referência cruzada  
descr::CrossTable(ds\_selected$sex\_male,ds\_selected$adhd\_parent, # linhas e colunas da tabela  
 digits = list(expected = 1, prop = 1, percent = 1, others = 1), # especificar o número de casas decimais de cada informação  
 expected = TRUE, # valor esperado  
 prop.c = TRUE, # proporção nas colunas  
 prop.chisq = FALSE, # proporção qui-quadrado  
 prop.r = TRUE, # proporção nas linhas  
 prop.t = FALSE) %>% # proporção dos totais (linhas e colunas)  
 pander(digits = 1,style = "grid") # quando se usa o pander, o digits tem que ir dentro dele.

|  |  |  |  |
| --- | --- | --- | --- |
| ds\_selected$sex\_male | ds\_selected$adhd\_parent no | yes | Total |
| \*\*female\*\*\  N\ Expected N\  Row(%)\  Column(%) | &nbsp;\  3379\  3309.1\  94.9%\  51.1% | 183 252.9 5.1% 36.2% | 3562  50.1% |
| \*\*male\*\*\  N\ Expected N\  Row(%)\  Column(%) | &nbsp;\  3230\  3299.9\  90.9%\  48.9% | 322 252.1 9.1% 63.8% | 3552  49.9% |
| Total\ | 6609\  92.9% | 505 7.1% | 7114 |

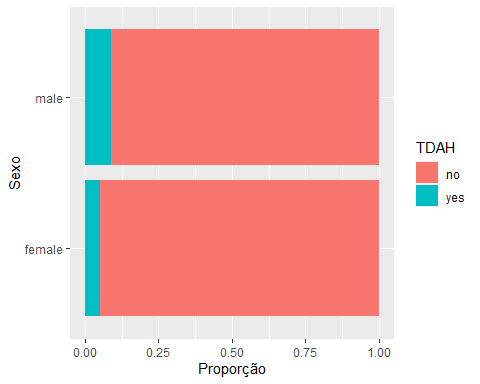
## Análise de qui-quadrado.  
descr::CrossTable(ds\_selected$sex\_male,ds\_selected$adhd\_parent,chisq = TRUE)$CST

##   
## Pearson's Chi-squared test  
##   
## data: tab  
## X-squared = 41.605, df = 1, p-value = 1.117e-10

# digite panderOptions() para lista de argumentos para o pander.

## Representação gráfica da tabela de referência cruzada

ggplot(ds\_selected, aes(x = sex\_male, fill = adhd\_parent)) +  
 geom\_bar(position = "fill") +   
 coord\_flip() +  
 labs(x = "Sexo", y = "Proporção", fill = "TDAH")



Nesse caso, o valor de p foi menor que 0,05, indicando que há associação entre o sexo do participantes e o diagnóstico de TDAH. Essa associação diz que ser do sexo masculino aumenta as chances de receber uma diagnóstico de TDAH.