Análise das variáveis Saresp Questionário - moda por escola

Série 5EF

Livia Kobayashi

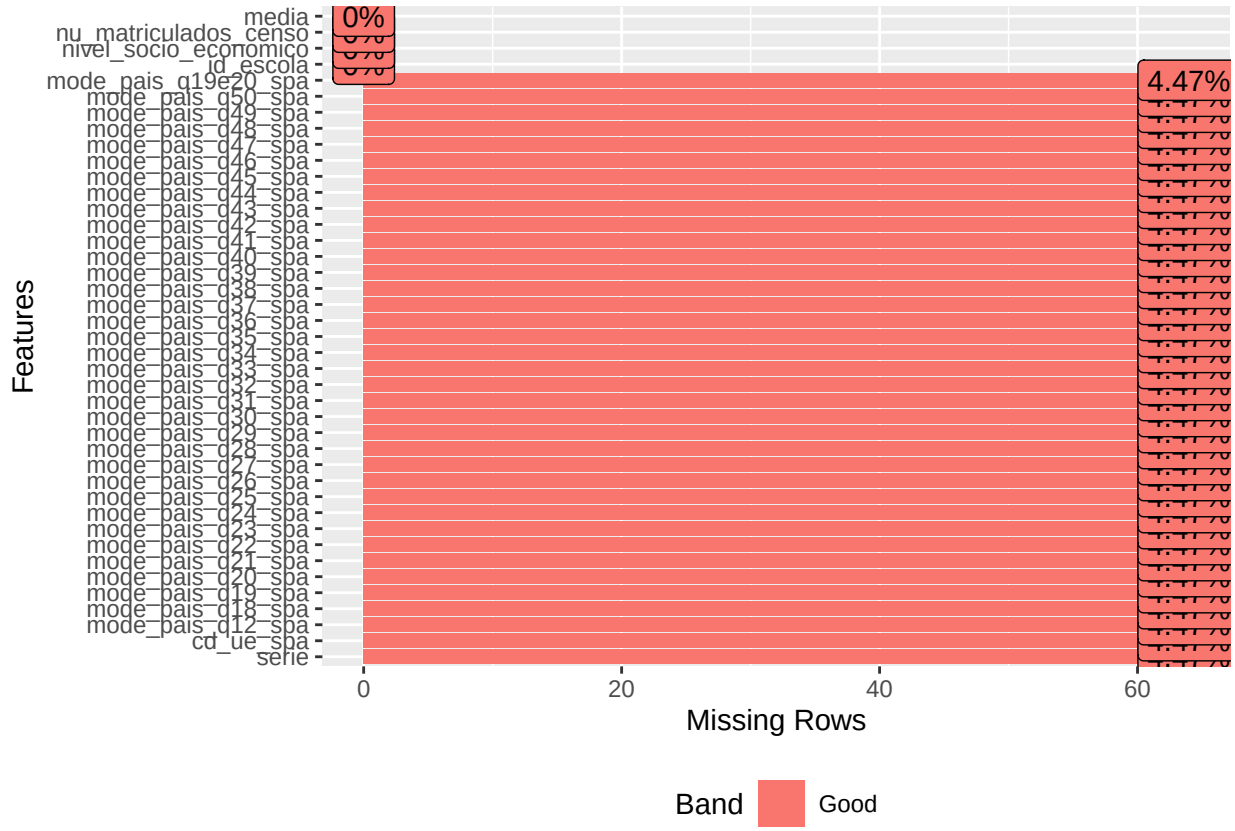
14 junho 2021

```
library(tidyverse)
library(DataExplorer)
library(gridExtra)
library(grid)
library(caret)
library(ggcorrplot)
library(vcd)
df_publico <- read.csv2("../output/books/df_publico.csv")
book <- read.csv2(params$book)
```

```
##   id_serie
## 1      5EF
```

Missing: 4,7% de dados faltantes

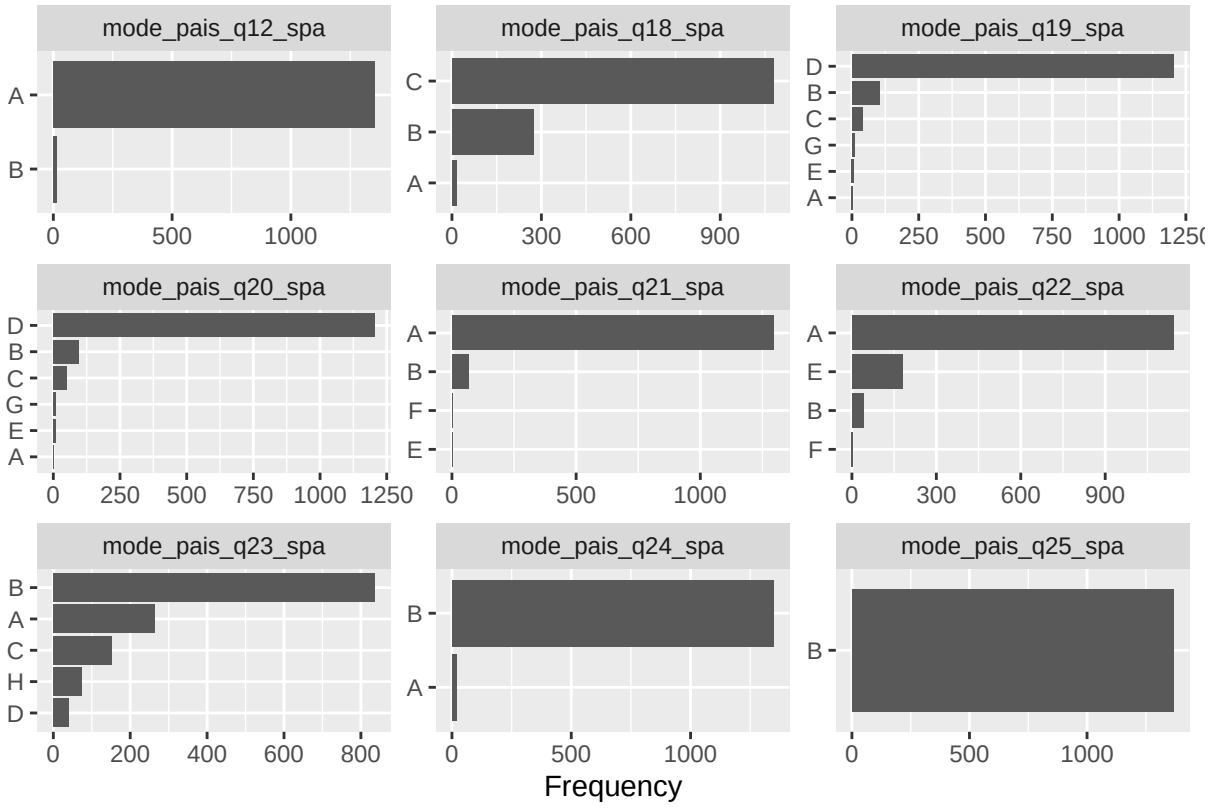
```
plot_missing(df)
```

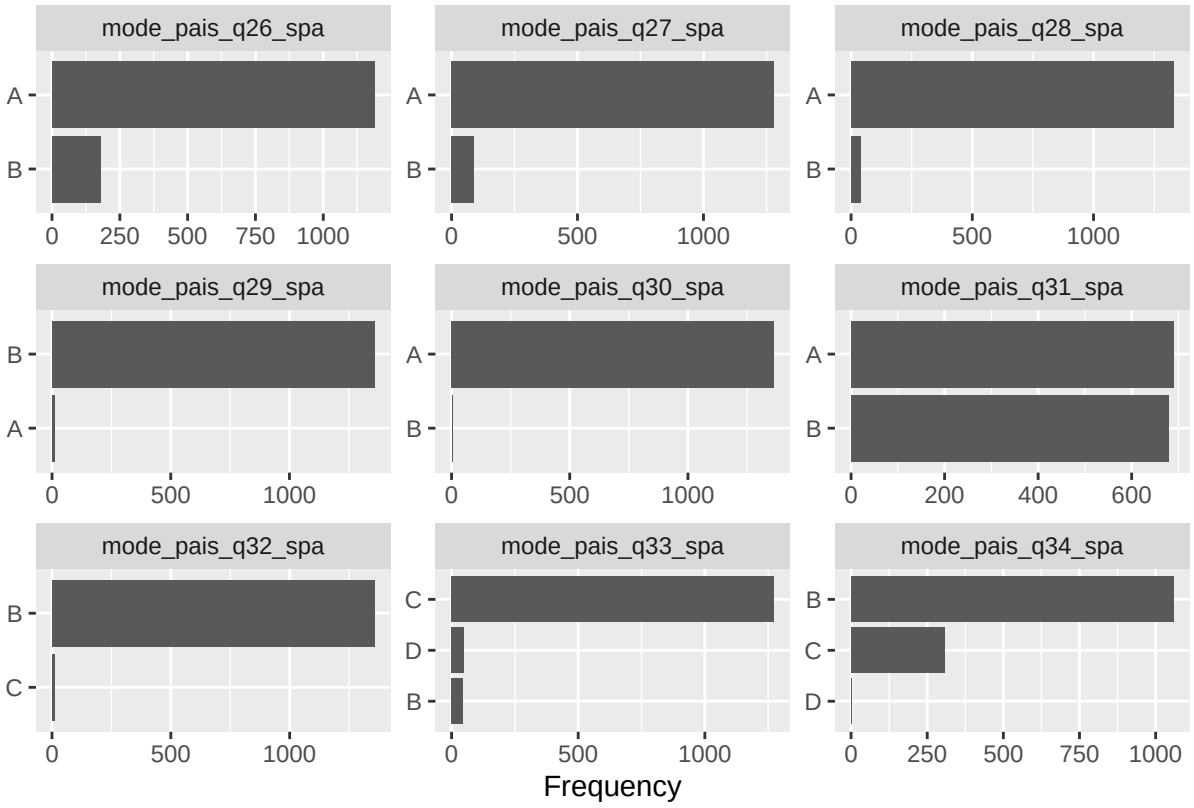


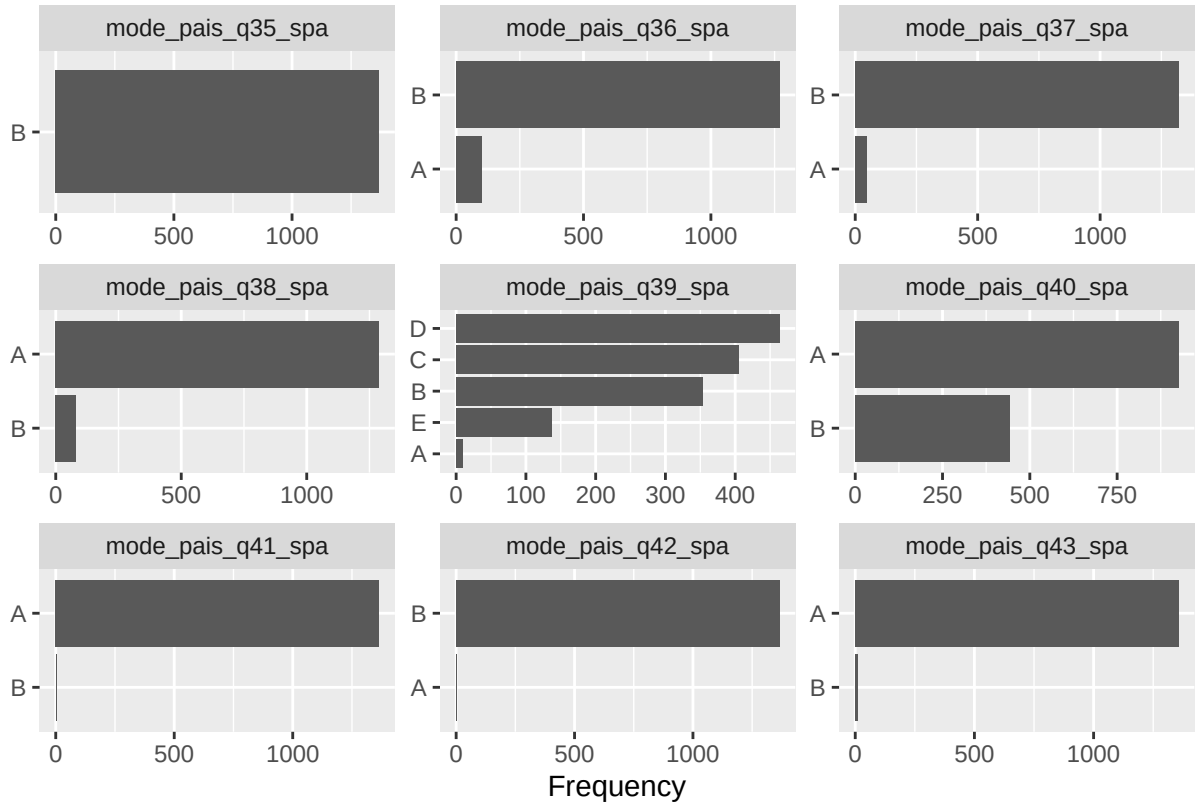
Volume: VAriáveis com bom volume

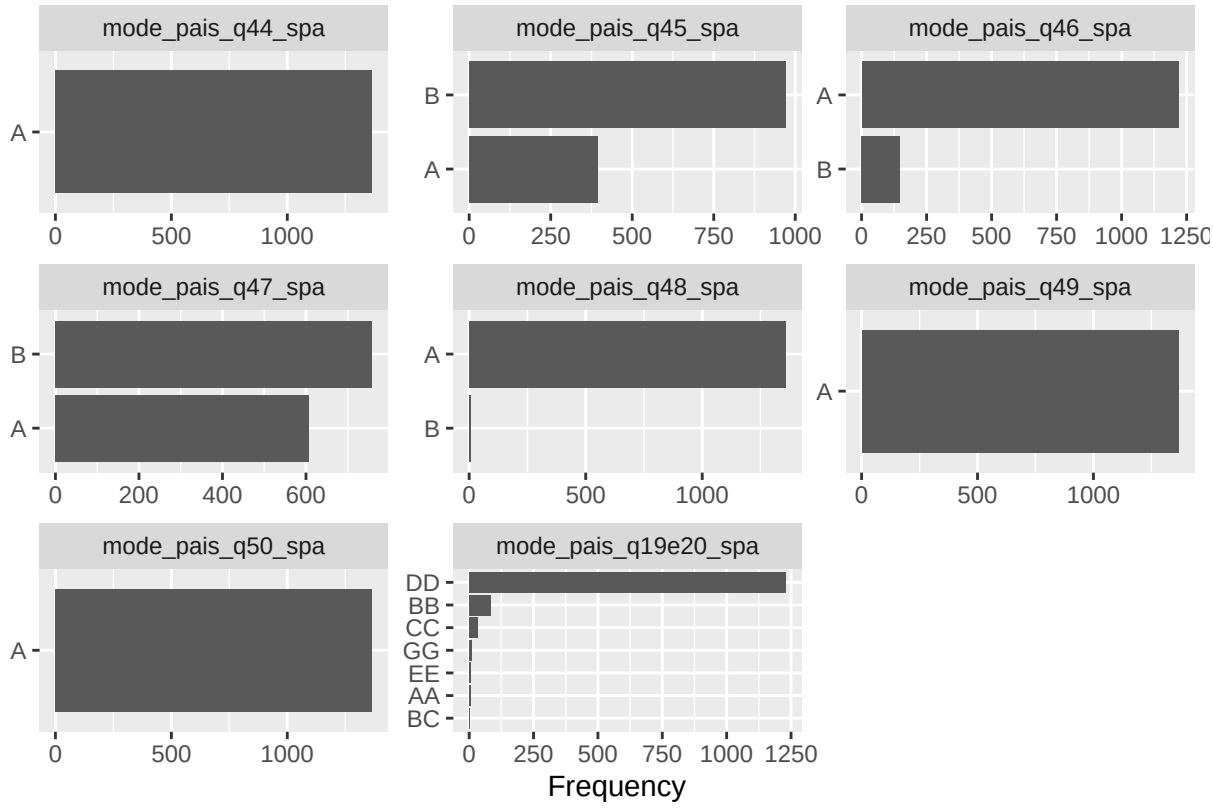
- mode_pais_q18
- mode_pais_q22
- mode_pais_q23
- mode_pais_q26
- mode_pais_q31
- mode_pais_q34
- mode_pais_q39
- mode_pais_q40
- mode_pais_q45
- mode_pais_q46
- mode_pais_q47

```
plot_bar(final_data)
```


mode_pais_q12_spa
A
B
0
500
1000
mode_pais_q18_spa
C
B
A
0
300
600
900
mode_pais_q19_spa
D
B
C
G
E
A
0
250
500
750
1000
mode_pais_q20_spa
D
B
C
G
E
A
0
250
500
750
1000
1250
mode_pais_q21_spa
A
B
F
E
0
500
1000
mode_pais_q22_spa
A
E
B
F
0
300
600
900
mode_pais_q23_spa
B
A
C
H
D
0
200
400
600
800
mode_pais_q24_spa
B
A
0
500
1000
mode_pais_q25_spa
B
0
500
1000
Frequency


mode_pais_q26_spa
A
B
0
250
500
750
1000
mode_pais_q27_spa
A
B
0
500
1000
mode_pais_q28_spa
A
B
0
500
1000
mode_pais_q29_spa
B
A
0
500
1000
mode_pais_q30_spa
A
B
0
500
1000
mode_pais_q31_spa
A
B
0
200
400
600
mode_pais_q32_spa
B
C
0
500
1000
mode_pais_q33_spa
C
D
B
0
500
1000
mode_pais_q34_spa
B
C
D
0
250
500
750
1000
Frequency


mode_pais_q35_spa
B
0
500
1000
mode_pais_q36_spa
B
A
0
500
1000
mode_pais_q37_spa
B
A
0
500
1000
mode_pais_q38_spa
A
B
0
500
1000
mode_pais_q39_spa
D
C
B
E
A
0
100
200
300
400
mode_pais_q40_spa
A
B
0
250
500
750
mode_pais_q41_spa
A
B
0
500
1000
mode_pais_q42_spa
B
A
0
500
1000
mode_pais_q43_spa
A
B
0
500
1000
Frequency

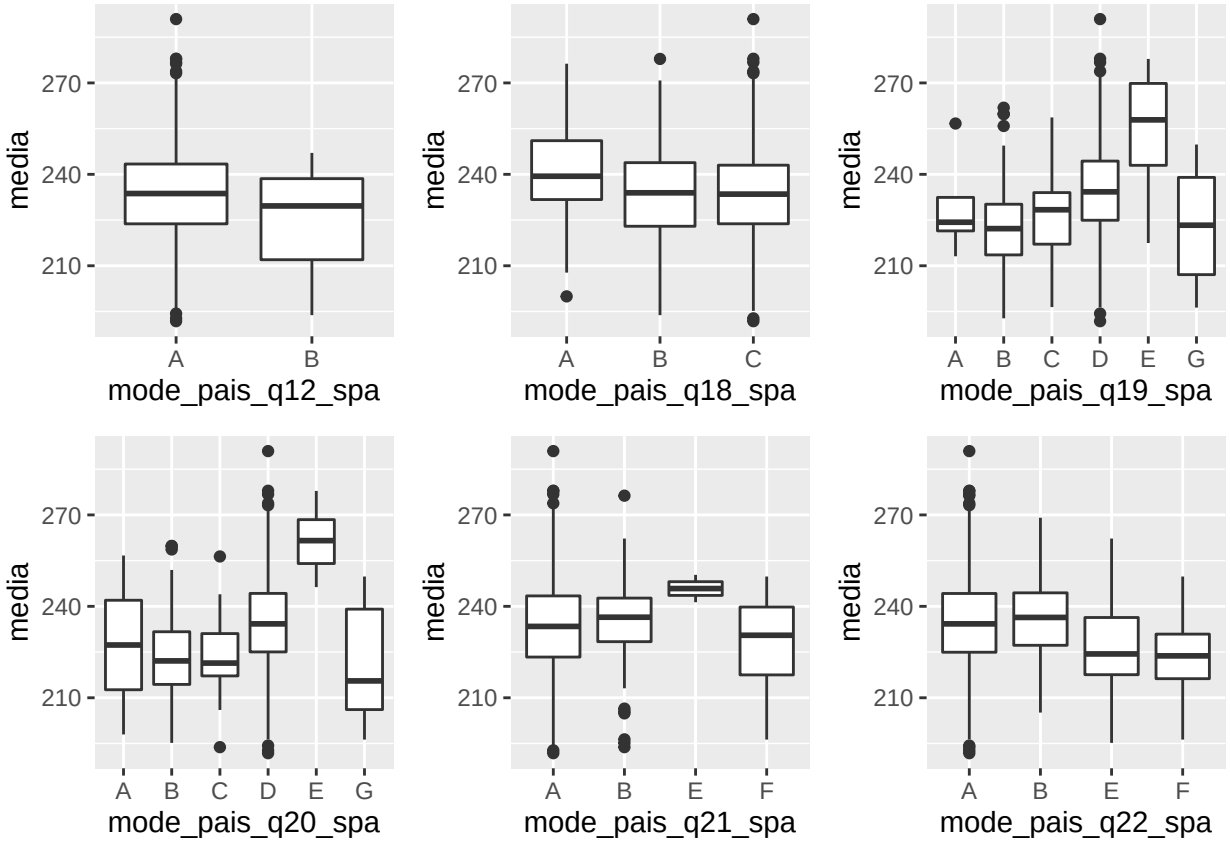


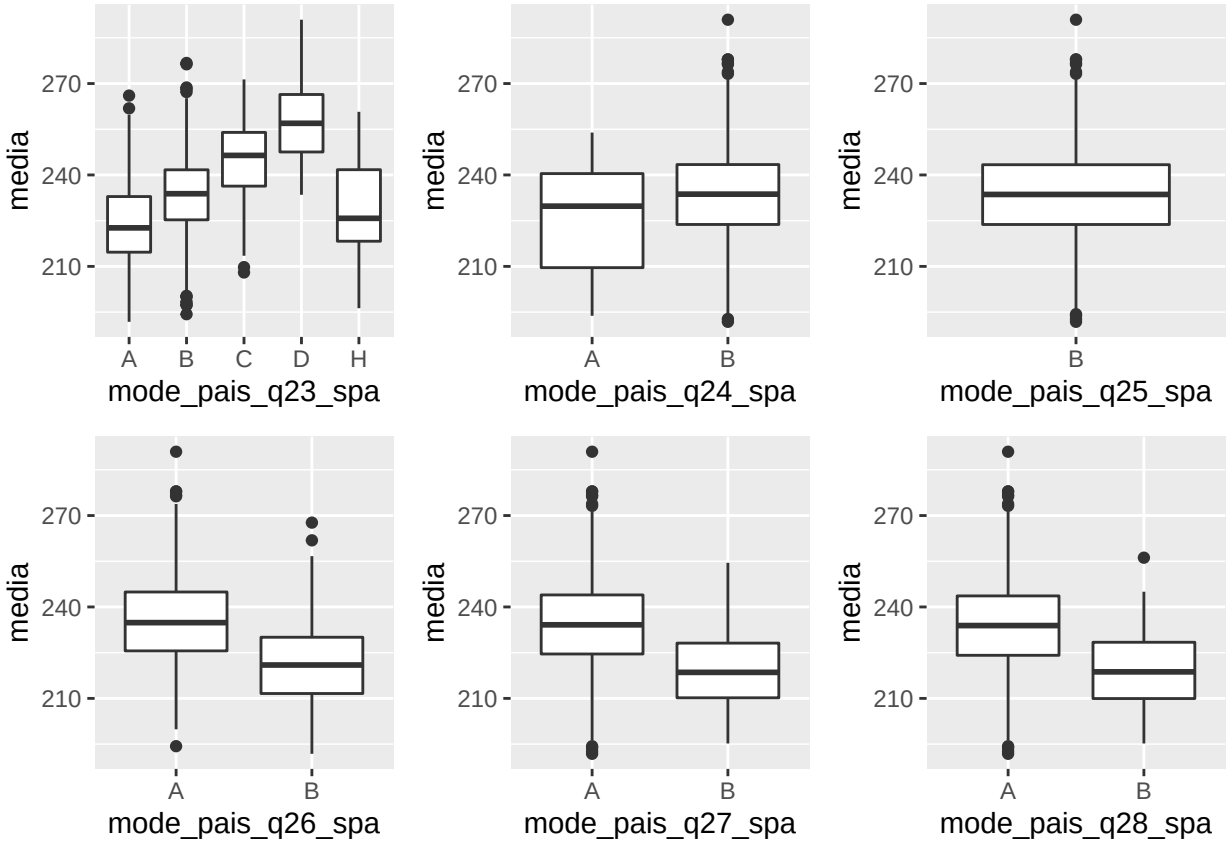
Boxplot: Variáveis com bom volume e variância

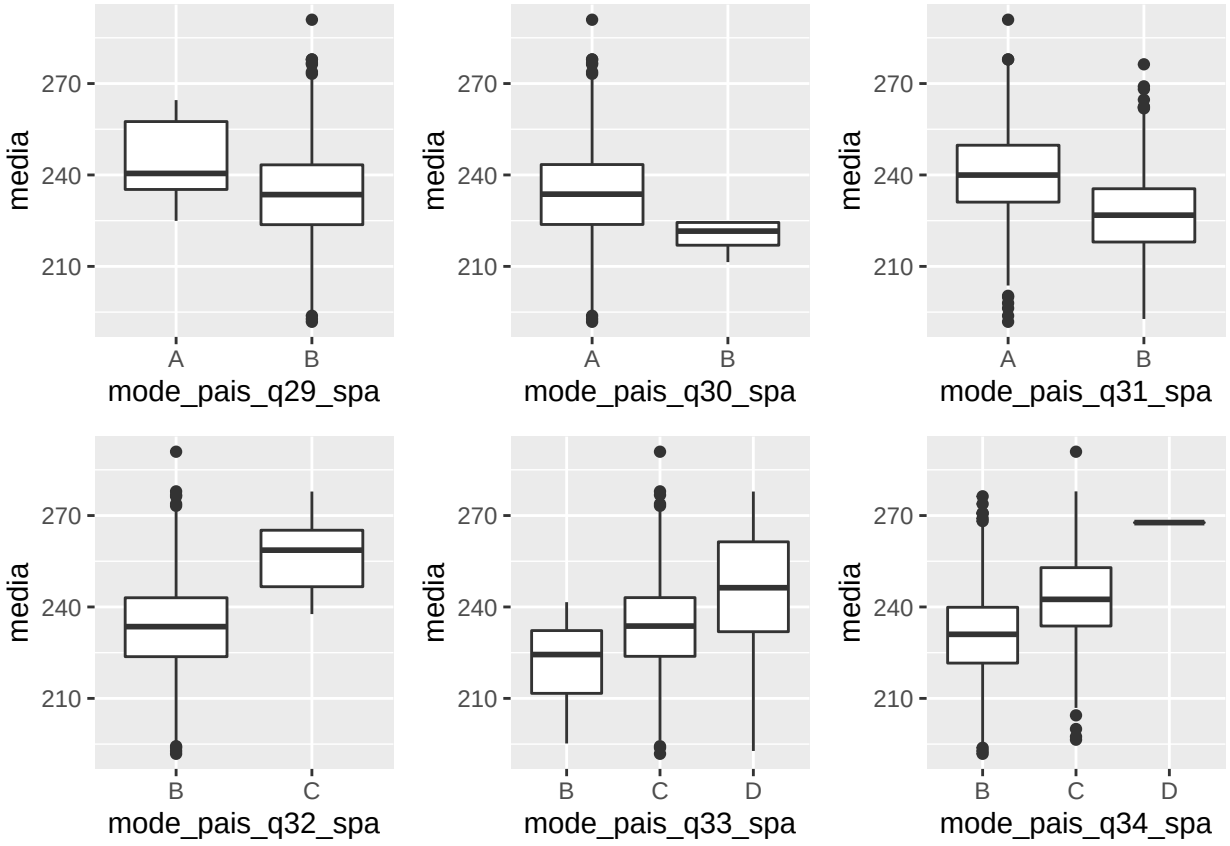
- mode_pais_q22: notas < para mães desempregadas ou que não querem responder
- mode_pais_q23: renda maior=> notas >
- mode_pais_q26: Dicionário => Notas >
- mode_pais_q31: TV assinatura => Notas >
- mode_pais_q34: #TV => Notas > (B x C)
- mode_pais_q39: #Celular => Notas >
- mode_pais_q40: COmputador => Notas > (A x B)
- mode_pais_q46: Aspirador => Notas > (A x B)
- mode_pais_q47: Carro => Notas > (A x B)

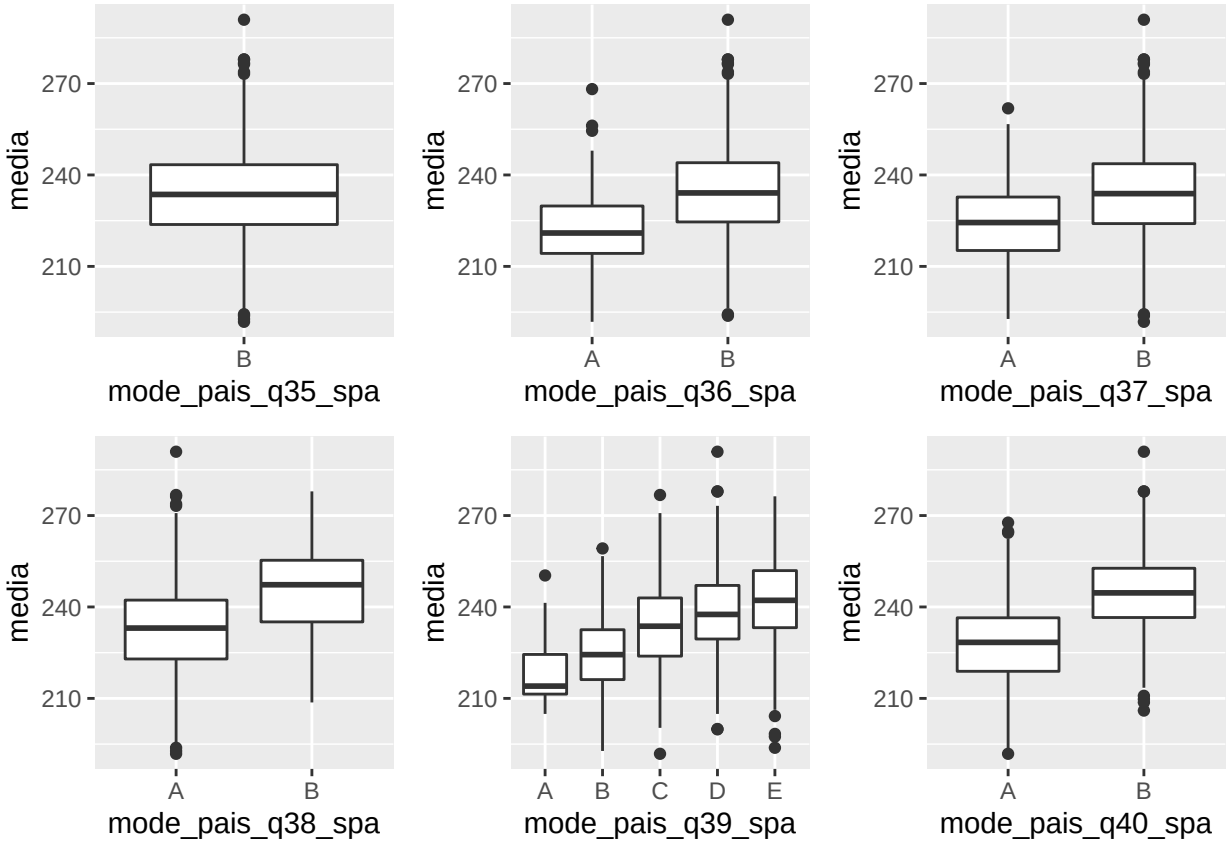
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

n <- length(plots)
```

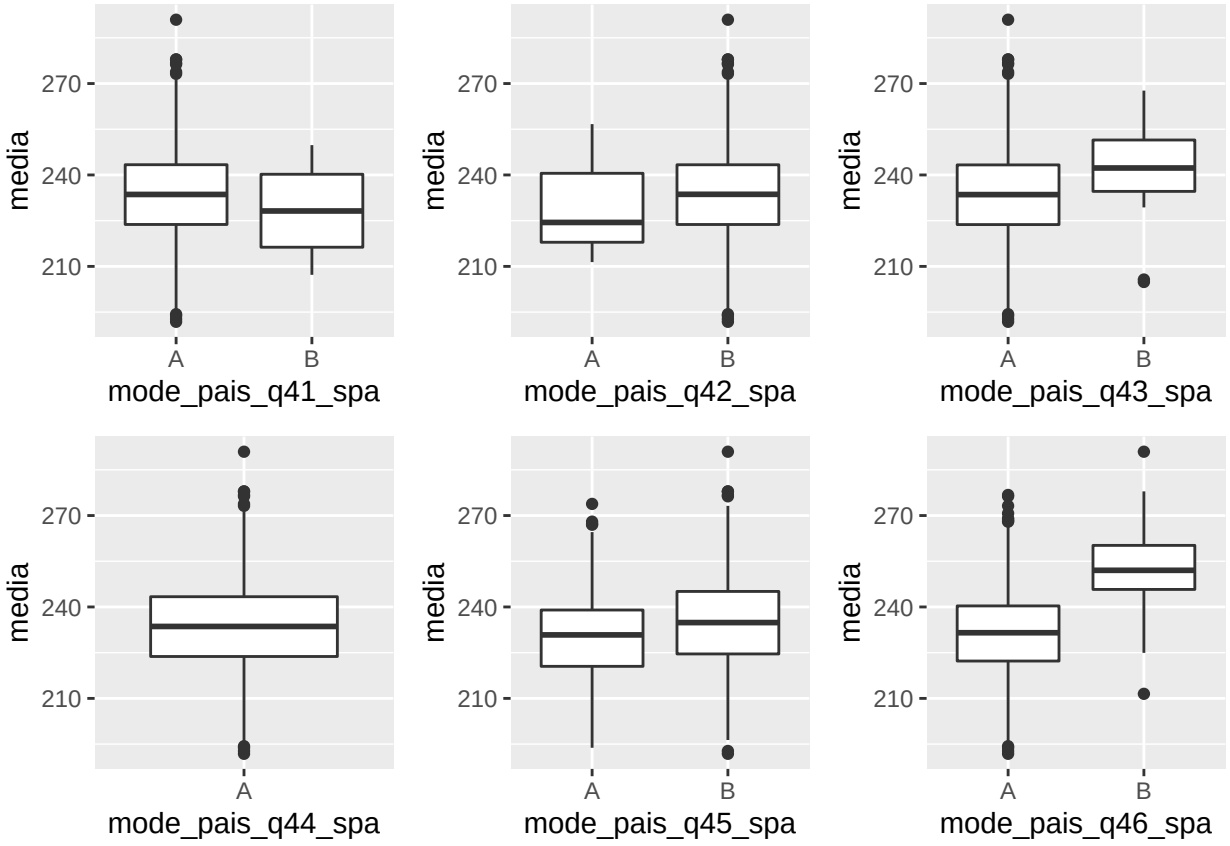
```
i <- 1
while (i <= n) {
  do.call("grid.arrange", c(plots[i:(min(i+5, n))], ncol=3, nrow = 2))
  i <- i + 6
}
```

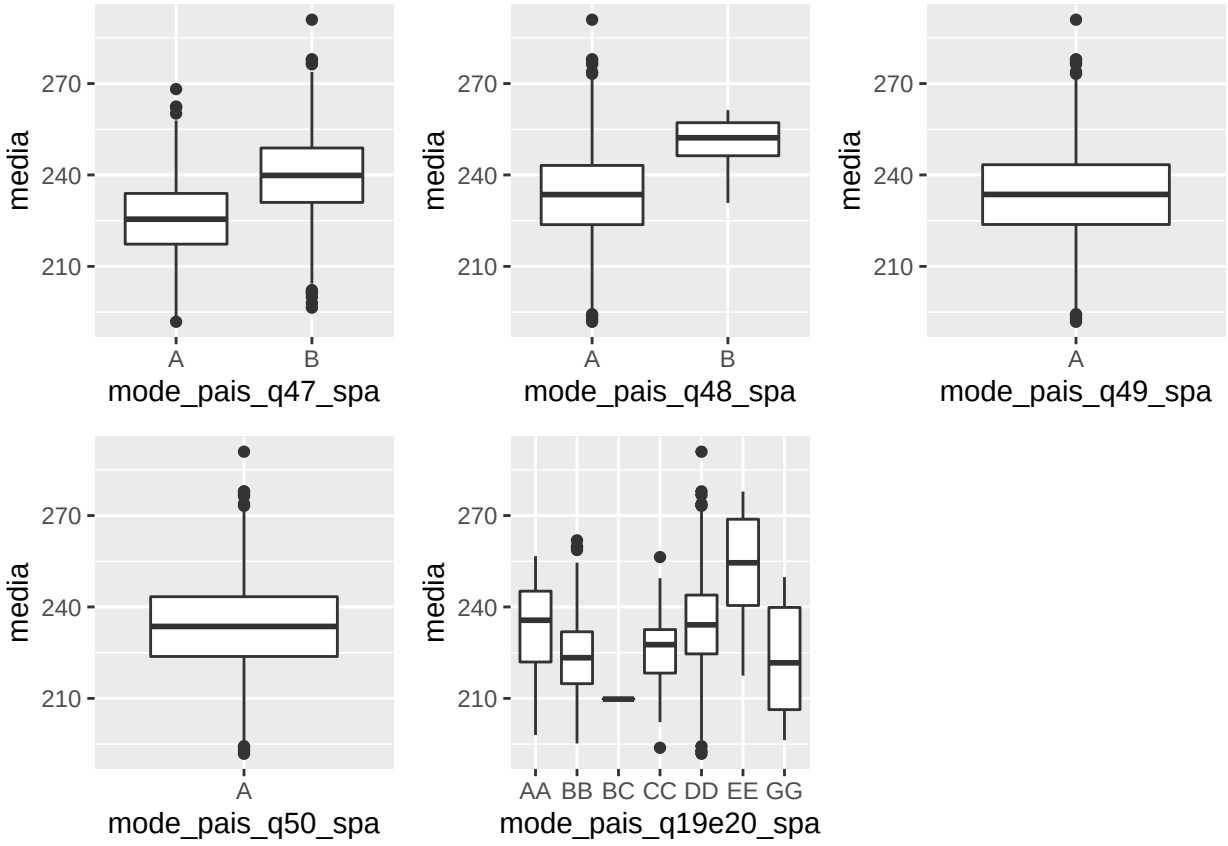



media
270
240
210
A
B
C
D
H
mode_pais_q23_spa
media
270
240
210
A
B
mode_pais_q24_spa
media
270
240
210
B
mode_pais_q25_spa
media
270
240
210
A
B
mode_pais_q26_spa
media
270
240
210
A
B
mode_pais_q27_spa
media
270
240
210
A
B
mode_pais_q28_spa


media
270
240
210
A
B
mode_pais_q29_spa
media
270
240
210
A
B
mode_pais_q30_spa
media
270
240
210
A
B
mode_pais_q31_spa
media
270
240
210
B
C
mode_pais_q32_spa
media
270
240
210
B
C
D
mode_pais_q33_spa
media
270
240
210
B
C
D
mode_pais_q34_spa


media
270
240
210
B
mode_pais_q35_spa
media
270
240
210
A
B
mode_pais_q36_spa
media
270
240
210
A
B
mode_pais_q37_spa
media
270
240
210
A
B
mode_pais_q38_spa
media
270
240
210
A
B
C
D
E
mode_pais_q39_spa
media
270
240
210
A
B
mode_pais_q40_spa


media
270
240
210
A
B
mode_pais_q41_spa
mode_pais_q42_spa
mode_pais_q43_spa
mode_pais_q44_spa
mode_pais_q45_spa
mode_pais_q46_spa



Análise Univariada

- mode_pais_q40: COmputador => Notas > (A x B)
- mode_pais_q23: renda maior=> notas >
- mode_pais_q47: Carro => Notas > (A x B)
- mode_pais_q46: Aspirador => Notas > (A x B)
- mode_pais_q31: TV assinatura => Notas >

```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
y_resp <- "media"

remove_cols <- nearZeroVar(df, names = TRUE)
final_cols <- setdiff(vars, remove_cols)
final_cols
```

```
##  [1] "mode_pais_q18_spa"    "mode_pais_q19_spa"    "mode_pais_q20_spa"
##  [4] "mode_pais_q22_spa"    "mode_pais_q23_spa"    "mode_pais_q26_spa"
##  [7] "mode_pais_q27_spa"    "mode_pais_q31_spa"    "mode_pais_q34_spa"
## [10] "mode_pais_q36_spa"    "mode_pais_q38_spa"    "mode_pais_q39_spa"
## [13] "mode_pais_q40_spa"    "mode_pais_q45_spa"    "mode_pais_q46_spa"
## [16] "mode_pais_q47_spa"    "mode_pais_q19e20_spa"
```

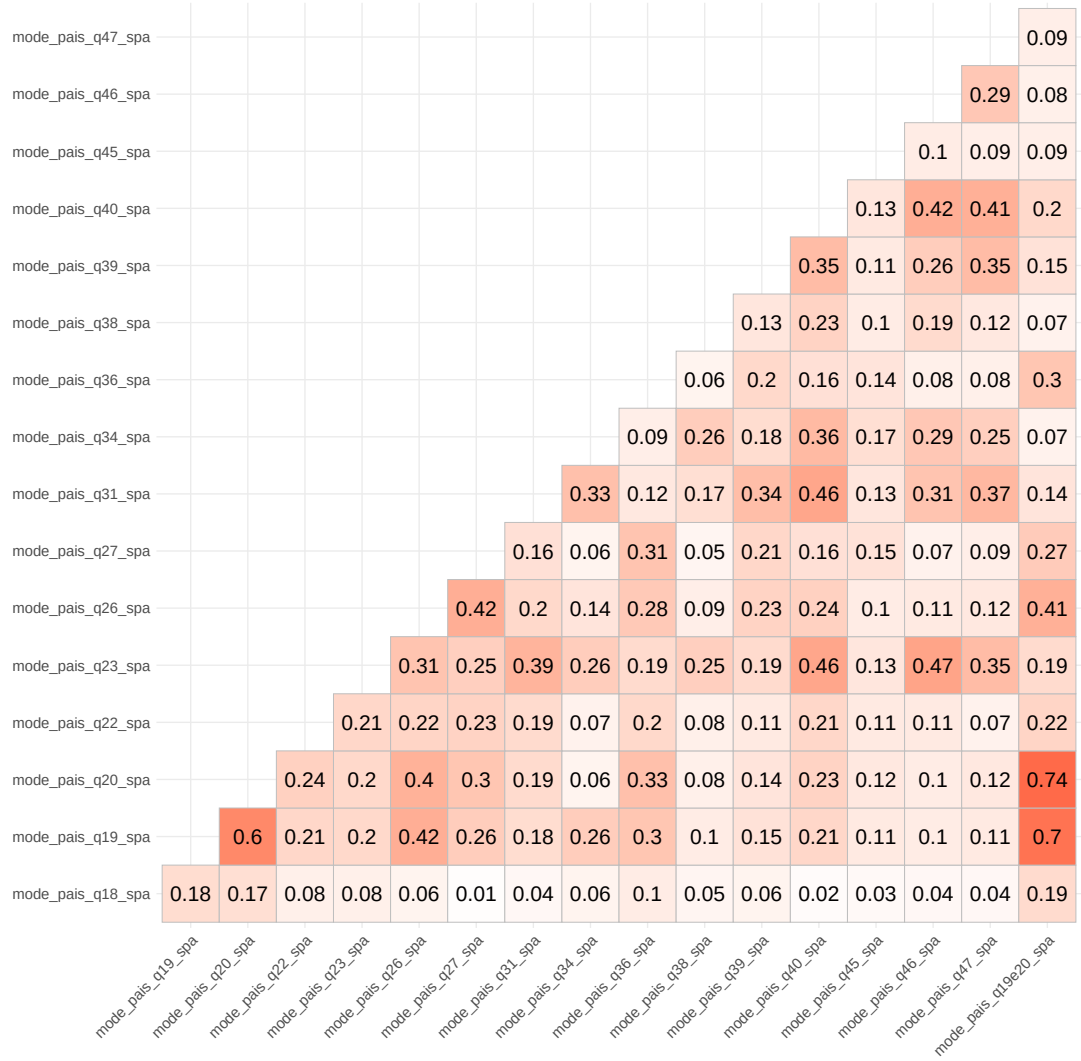
```r
tb_r2 <- data.frame(var = final_cols)

rsquared <- c()
for (variable in final_cols) {
  lm_formula <- as.formula(str_glue("{y_resp} ~ {variable}"))
  model_lm <- lm(lm_formula, df)
  rsquared <- append(rsquared, summary(model_lm)$r.squared)
}

tb_r2$rsquared <- rsquared
tb_r2 %>% head(nrow(tb_r2))
```

```
##                     var    rsquared
## 1     mode_pais_q18_spa 0.001278939
## 2     mode_pais_q19_spa 0.067813271
## 3     mode_pais_q20_spa 0.075943421
## 4     mode_pais_q22_spa 0.034688144
## 5     mode_pais_q23_spa 0.227685862
## 6     mode_pais_q26_spa 0.098311229
## 7     mode_pais_q27_spa 0.056846510
## 8     mode_pais_q31_spa 0.186641489
## 9     mode_pais_q34_spa 0.106198064
## 10    mode_pais_q36_spa 0.043615100
## 11    mode_pais_q38_spa 0.045762133
## 12    mode_pais_q39_spa 0.158987288
## 13    mode_pais_q40_spa 0.263138879
## 14    mode_pais_q45_spa 0.019305677
## 15    mode_pais_q46_spa 0.189081532
## 16    mode_pais_q47_spa 0.216287414
## 17 mode_pais_q19e20_spa 0.046970393
```

Matriz de correlação

- q19 x q20: Escolaridade da mãe e do pai: tentar combinar (apenas unir as duas não deu certo)

```r
catcorrm <- function(vars, dat) sapply(vars, function(y) sapply(vars, function(x) assocstats(table(dat[

matriz <- catcorrm(final_cols, data_corr)

ggcorrplot(matriz, show.diag = F, type="lower", lab=TRUE, lab_size=6, show.legend = F)
```


mode_pais_q47_spa
mode_pais_q46_spa
mode_pais_q45_spa
mode_pais_q40_spa
mode_pais_q39_spa
mode_pais_q38_spa
mode_pais_q36_spa
mode_pais_q34_spa
mode_pais_q31_spa
mode_pais_q27_spa
mode_pais_q26_spa
mode_pais_q23_spa
mode_pais_q22_spa
mode_pais_q20_spa
mode_pais_q19_spa
mode_pais_q18_spa
0.09
0.29 0.08
0.1 0.09 0.09
0.13 0.42 0.41 0.2
0.35 0.11 0.26 0.35 0.15
0.13 0.23 0.1 0.19 0.12 0.07
0.06 0.2 0.16 0.14 0.08 0.08 0.3
0.09 0.26 0.18 0.36 0.17 0.29 0.25 0.07
0.33 0.12 0.17 0.34 0.46 0.13 0.31 0.37 0.14
0.16 0.06 0.31 0.05 0.21 0.16 0.15 0.07 0.09 0.27
0.42 0.2 0.14 0.28 0.09 0.23 0.24 0.1 0.11 0.12 0.41
0.31 0.25 0.39 0.26 0.19 0.25 0.19 0.46 0.13 0.47 0.35 0.19
0.21 0.22 0.23 0.19 0.07 0.2 0.08 0.11 0.21 0.11 0.11 0.07 0.22
0.24 0.2 0.4 0.3 0.19 0.06 0.33 0.08 0.14 0.23 0.12 0.1 0.12 0.74
0.6 0.21 0.2 0.42 0.26 0.18 0.26 0.3 0.1 0.15 0.21 0.11 0.1 0.11 0.7
0.18 0.17 0.08 0.08 0.06 0.01 0.04 0.06 0.1 0.05 0.06 0.02 0.03 0.04 0.04 0.19
mode_pais_q19_spa
mode_pais_q20_spa
mode_pais_q22_spa
mode_pais_q23_spa
mode_pais_q26_spa
mode_pais_q27_spa
mode_pais_q31_spa
mode_pais_q34_spa
mode_pais_q36_spa
mode_pais_q38_spa
mode_pais_q39_spa
mode_pais_q40_spa
mode_pais_q45_spa
mode_pais_q46_spa
mode_pais_q47_spa
mode_pais_q19e20_spa