Análise das variáveis Saresp Questionário pais - moda por escola

Série 5EF

Livia Kobayashi

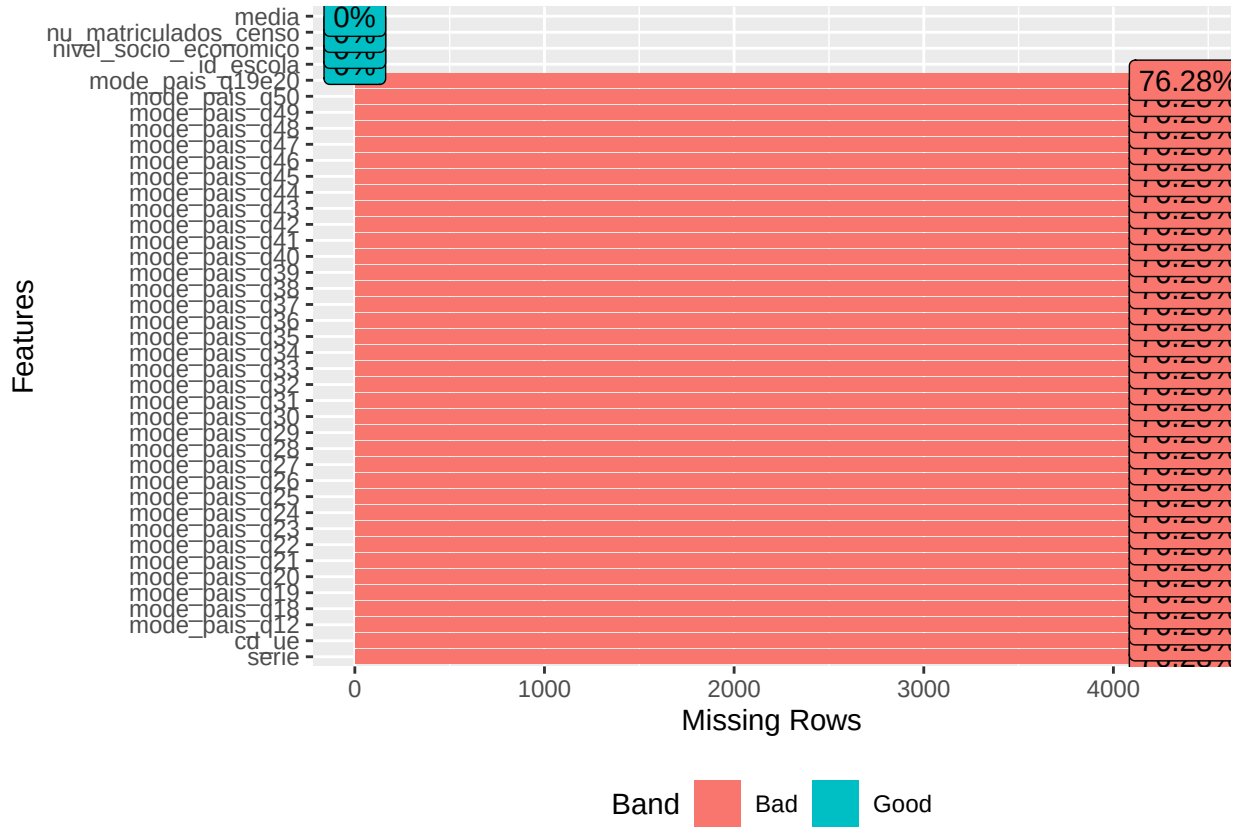
09 junho 2021

```
library(tidyverse)
library(DataExplorer)
library(gridExtra)
library(grid)
library(caret)
library(ggcorrplot)
library(vcd)
df_publico <- read.csv2("../books/df_publico.csv")
#book <- read.csv2("../books/saresp5ef_pais_mode.csv")
book <- read.csv2(book)
```

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

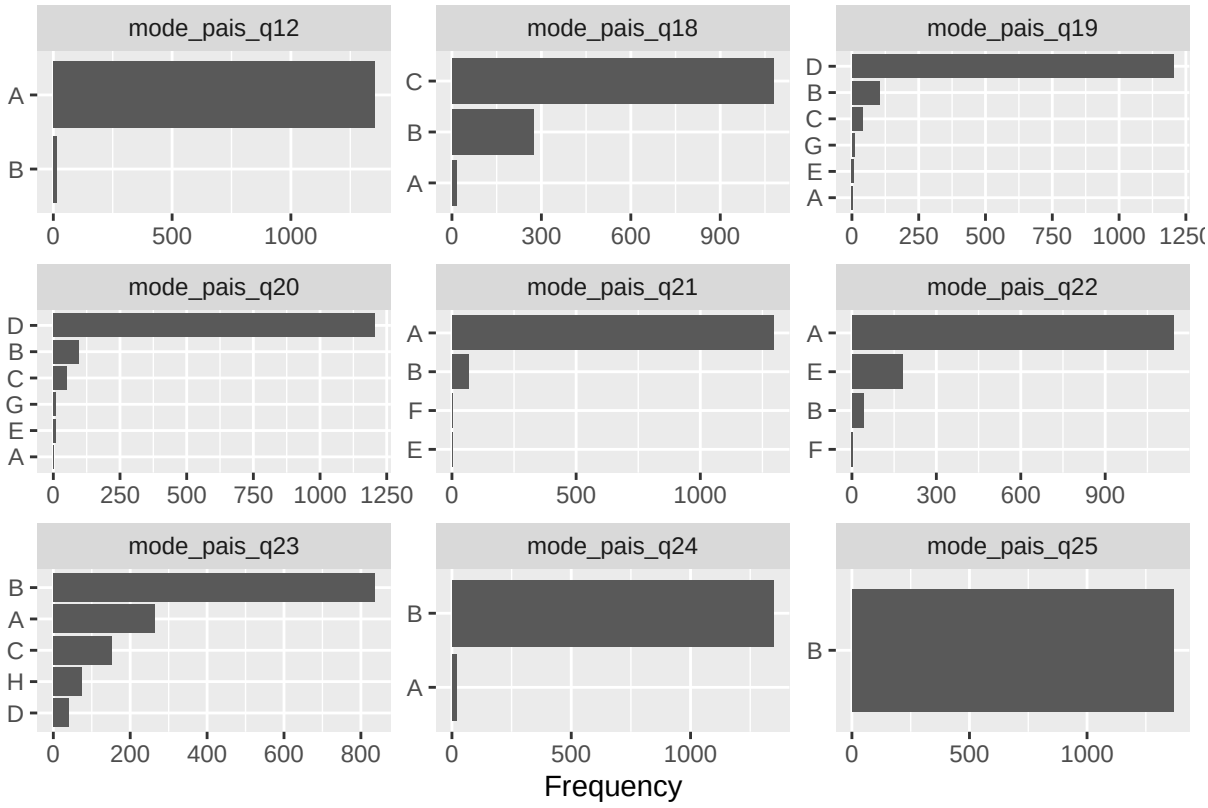
Missing

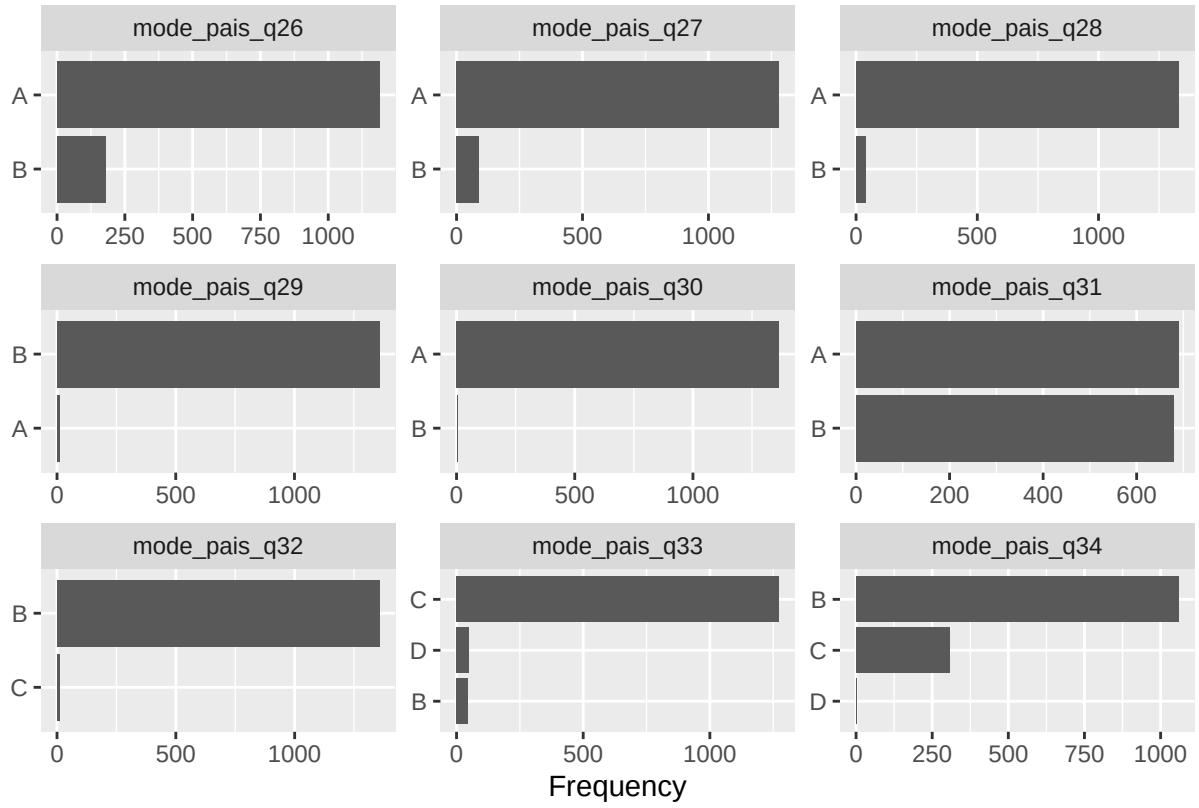
```
plot_missing(df)
```



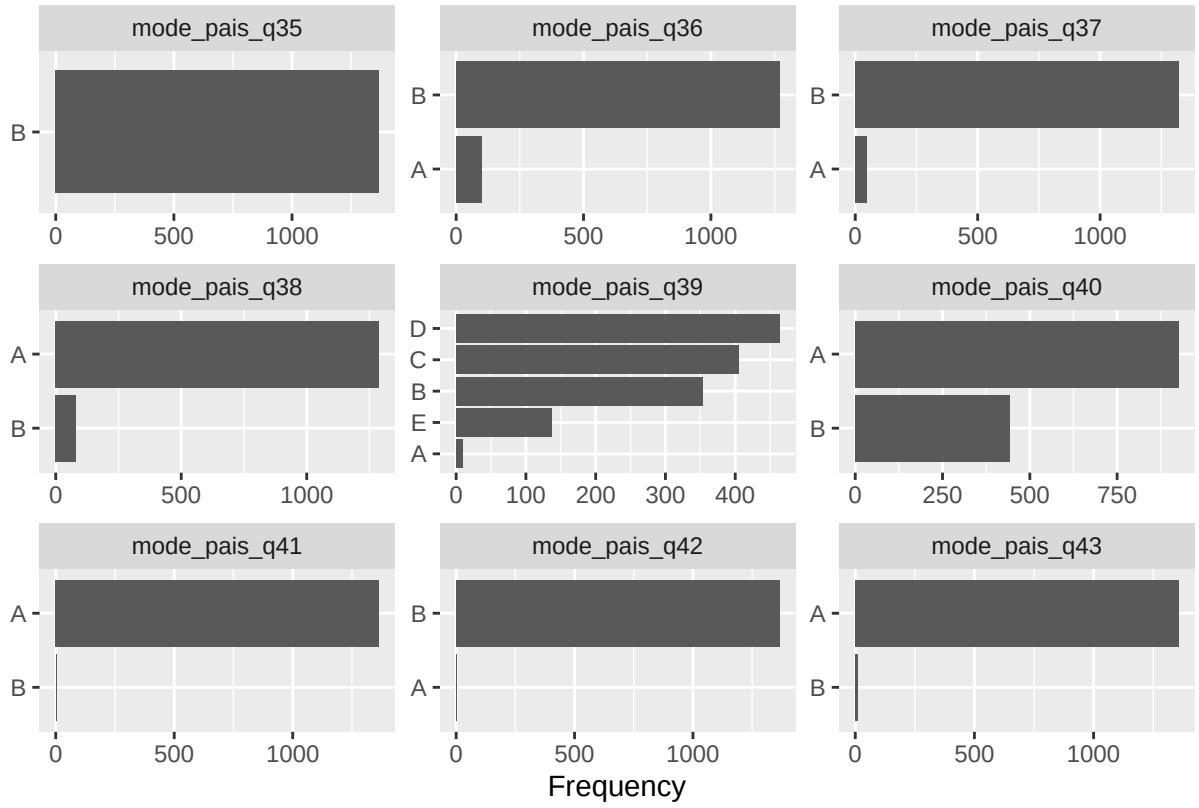
Volume

```
plot_bar(final_data)
```

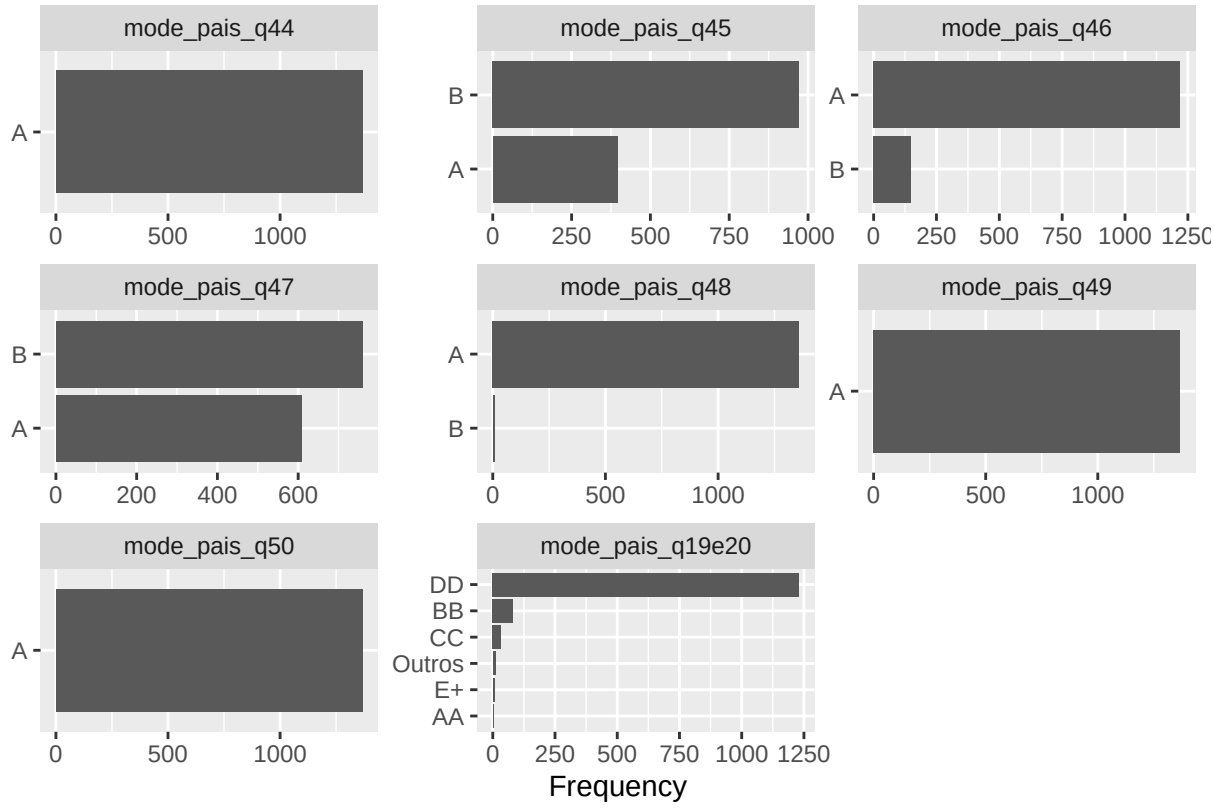

mode_pais_q12
A
B
0
500
1000
mode_pais_q18
C
B
A
0
300
600
900
mode_pais_q19
D
B
C
G
E
A
0
250
500
750
1000
mode_pais_q20
D
B
C
G
E
A
0
250
500
750
1000
1250
mode_pais_q21
A
B
F
E
0
500
1000
mode_pais_q22
A
E
B
F
0
300
600
900
mode_pais_q23
B
A
C
H
D
0
200
400
600
800
mode_pais_q24
B
A
0
500
1000
mode_pais_q25
B
0
500
1000
Frequency


mode_pais_q26
A
B
0
250
500
750
1000
mode_pais_q27
A
B
0
500
1000
mode_pais_q28
A
B
0
500
1000
mode_pais_q29
B
A
0
500
1000
mode_pais_q30
A
B
0
500
1000
mode_pais_q31
A
B
0
200
400
600
mode_pais_q32
B
C
0
500
1000
mode_pais_q33
C
D
B
0
500
1000
mode_pais_q34
B
C
D
0
250
500
750
1000
Frequency

Page 2

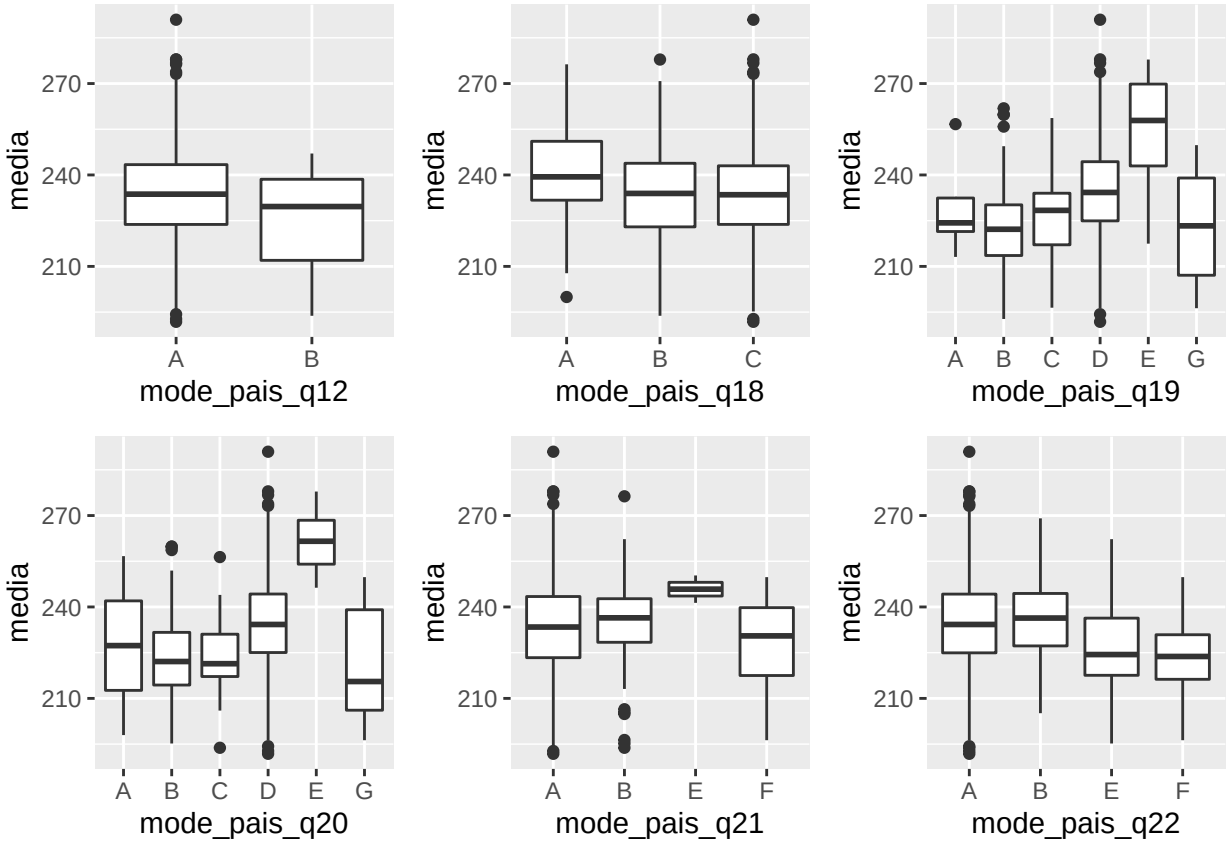

mode_pais_q35
B
0
500
1000
mode_pais_q36
B
A
0
500
1000
mode_pais_q37
B
A
0
500
1000
mode_pais_q38
A
B
0
500
1000
mode_pais_q39
D
C
B
E
A
0
100
200
300
400
mode_pais_q40
A
B
0
250
500
750
mode_pais_q41
A
B
0
500
1000
mode_pais_q42
B
A
0
500
1000
mode_pais_q43
A
B
0
500
1000
Frequency

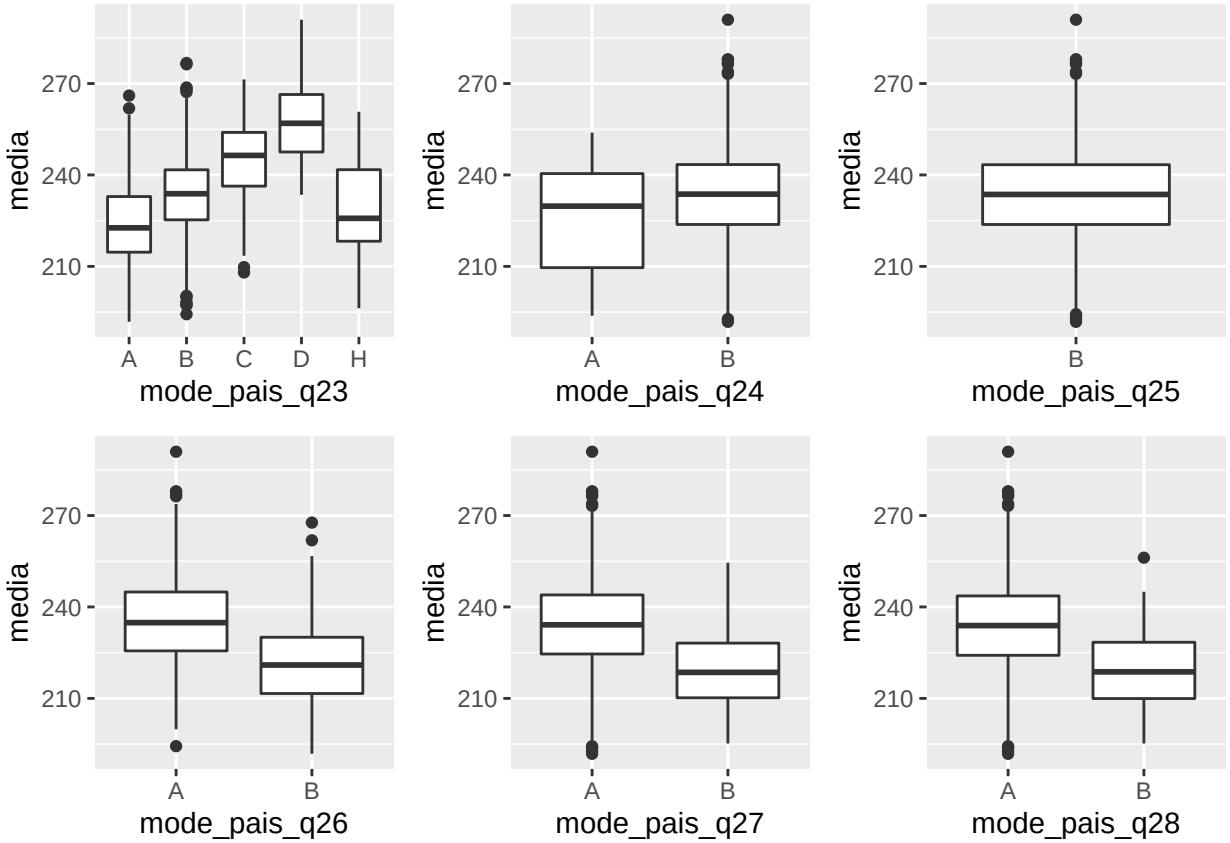
Page 3

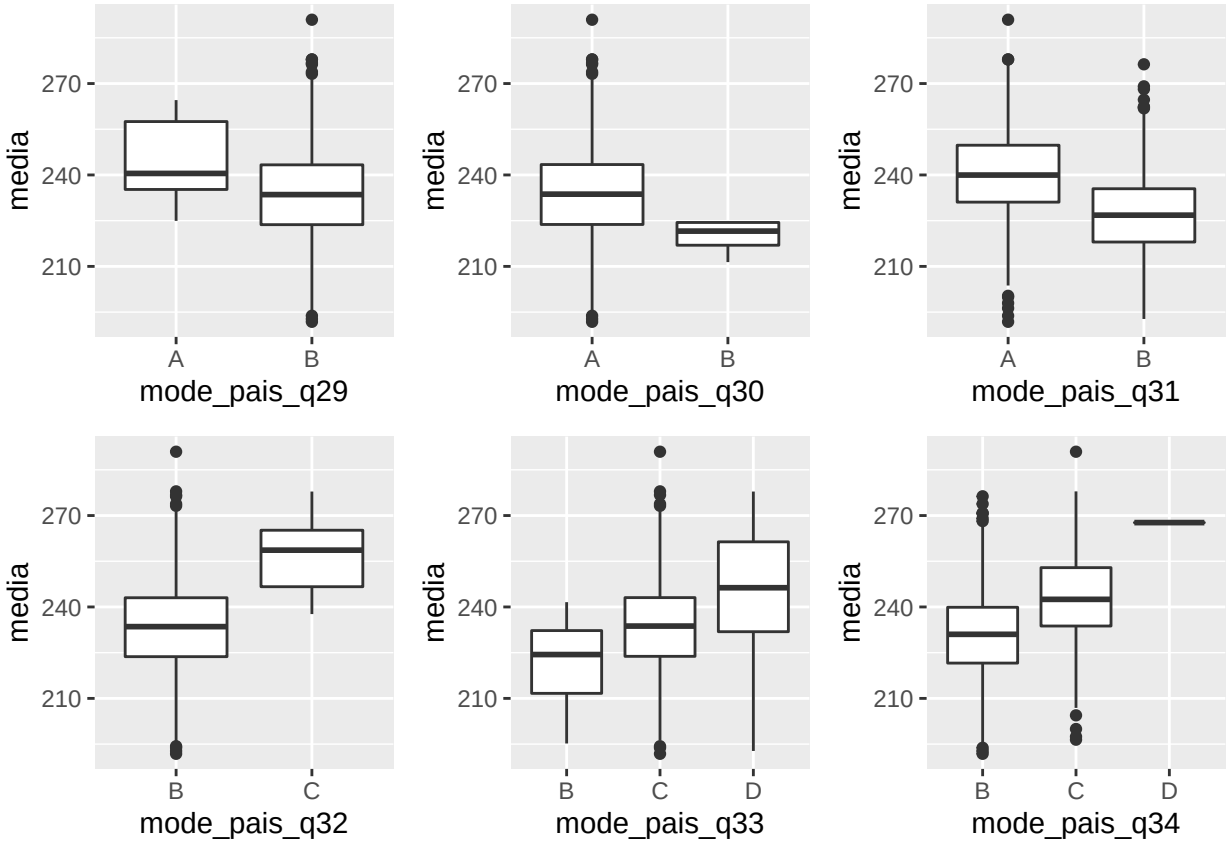


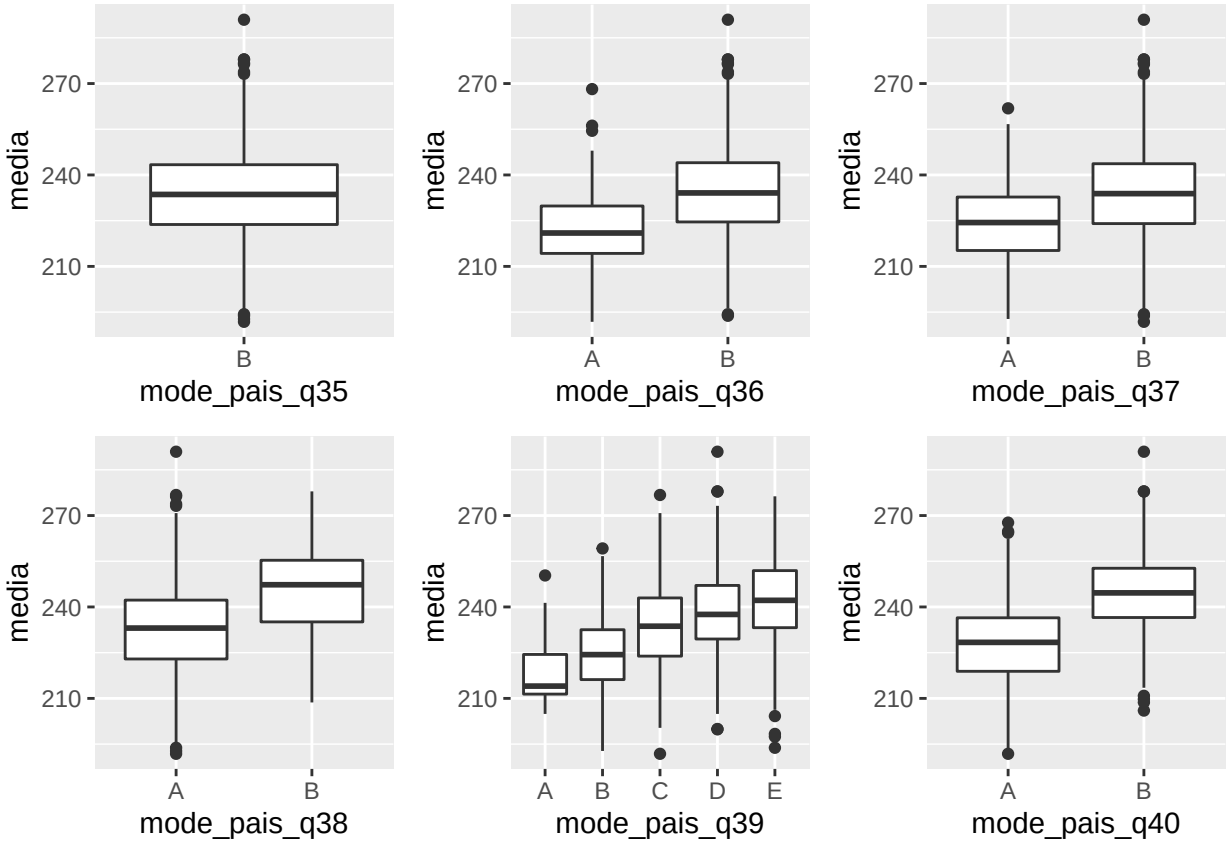
Boxplot

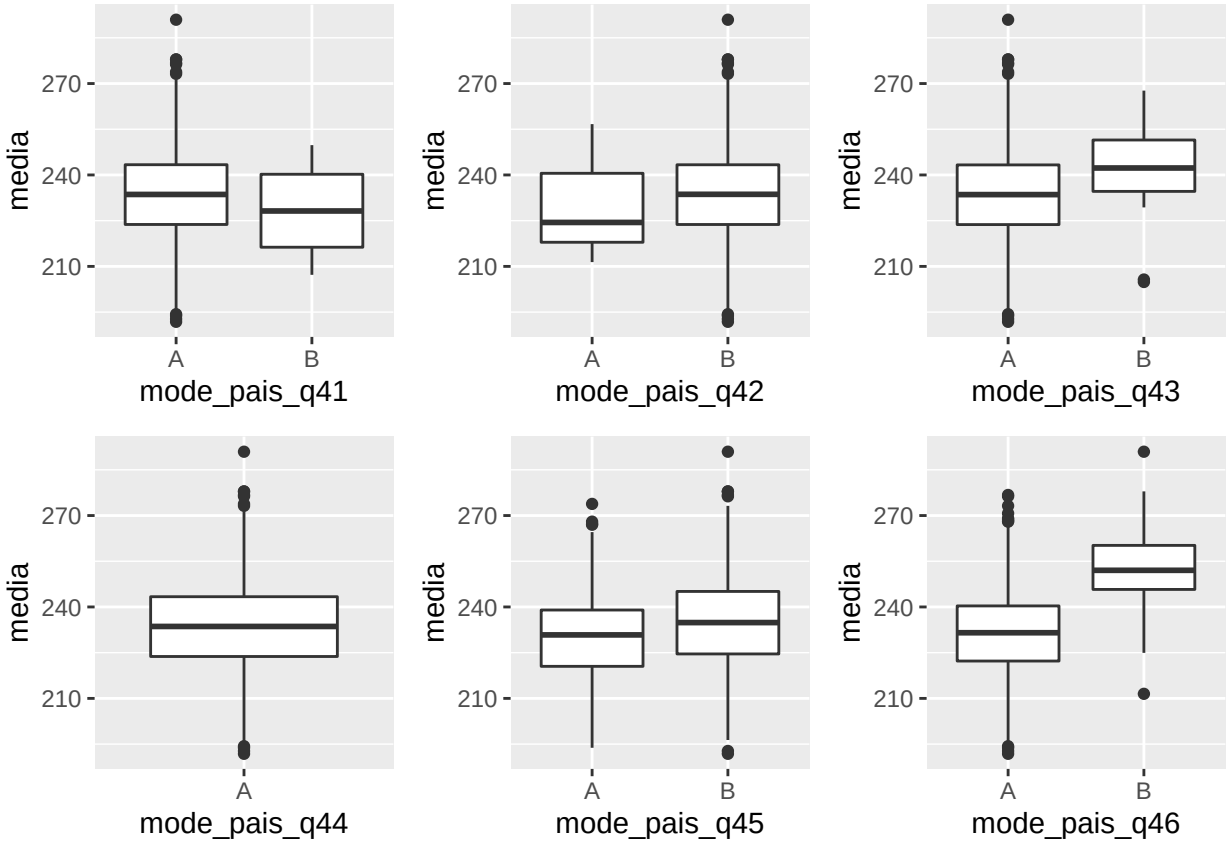
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

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
}
```

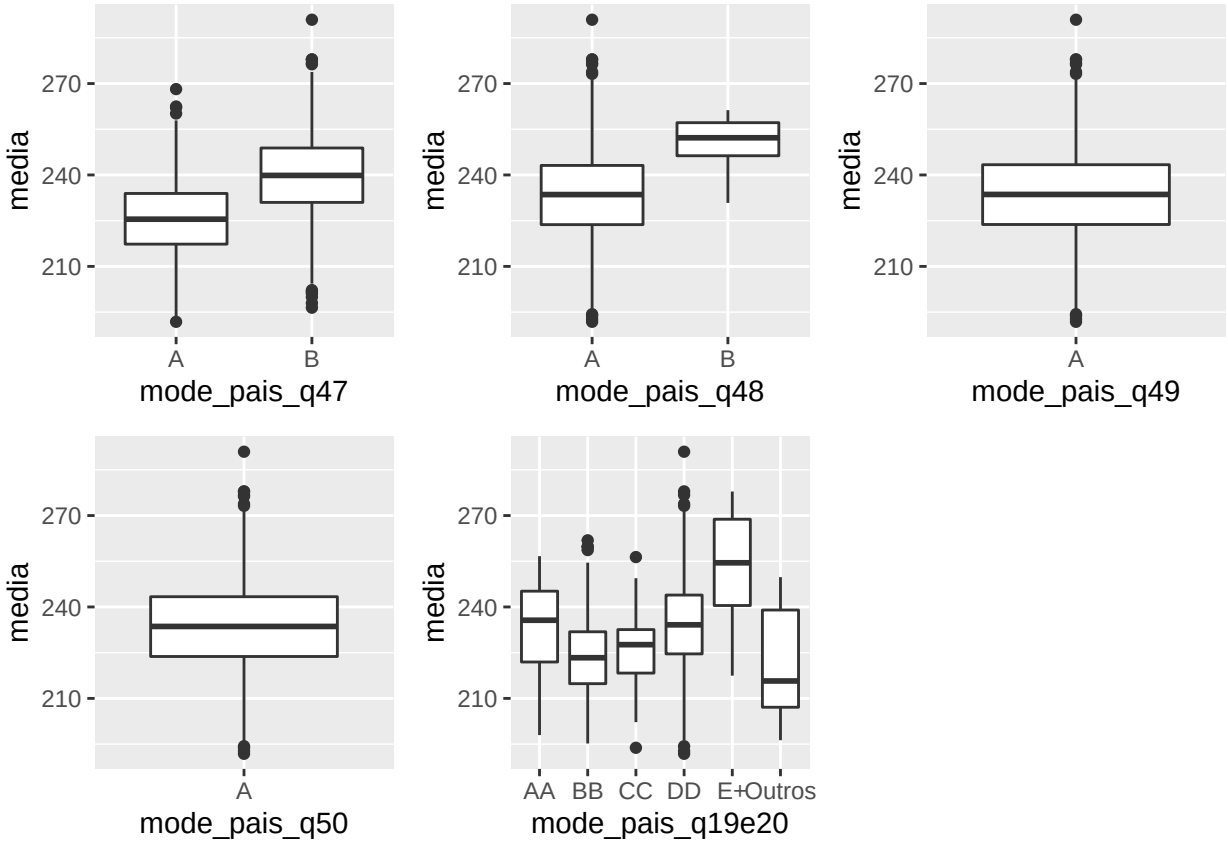

media
270
240
210
A
B
mode_pais_q12
media
270
240
210
A
B
C
mode_pais_q18
media
270
240
210
A
B
C
D
E
G
mode_pais_q19
media
270
240
210
A
B
C
D
E
G
mode_pais_q20
media
270
240
210
A
B
E
F
mode_pais_q21
media
270
240
210
A
B
E
F
mode_pais_q22


media
270
240
210
A
B
C
D
H
mode_pais_q23
media
270
240
210
A
B
mode_pais_q24
media
270
240
210
B
mode_pais_q25
media
270
240
210
A
B
mode_pais_q26
media
270
240
210
A
B
mode_pais_q27
media
270
240
210
A
B
mode_pais_q28


media
270
240
210
A
B
mode_pais_q29
media
270
240
210
A
B
mode_pais_q30
media
270
240
210
A
B
mode_pais_q31
media
270
240
210
B
C
mode_pais_q32
media
270
240
210
B
C
D
mode_pais_q33
media
270
240
210
B
C
D
mode_pais_q34


media
270
240
210
B
mode_pais_q35
media
270
240
210
A
B
mode_pais_q36
media
270
240
210
A
B
mode_pais_q37
media
270
240
210
A
B
mode_pais_q38
media
270
240
210
A
B
C
D
E
mode_pais_q39
media
270
240
210
A
B
mode_pais_q40


media
270
240
210
A
B
mode_pais_q41
mode_pais_q42
mode_pais_q43
mode_pais_q44
mode_pais_q45
mode_pais_q46



Análise Univariada

```
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
```

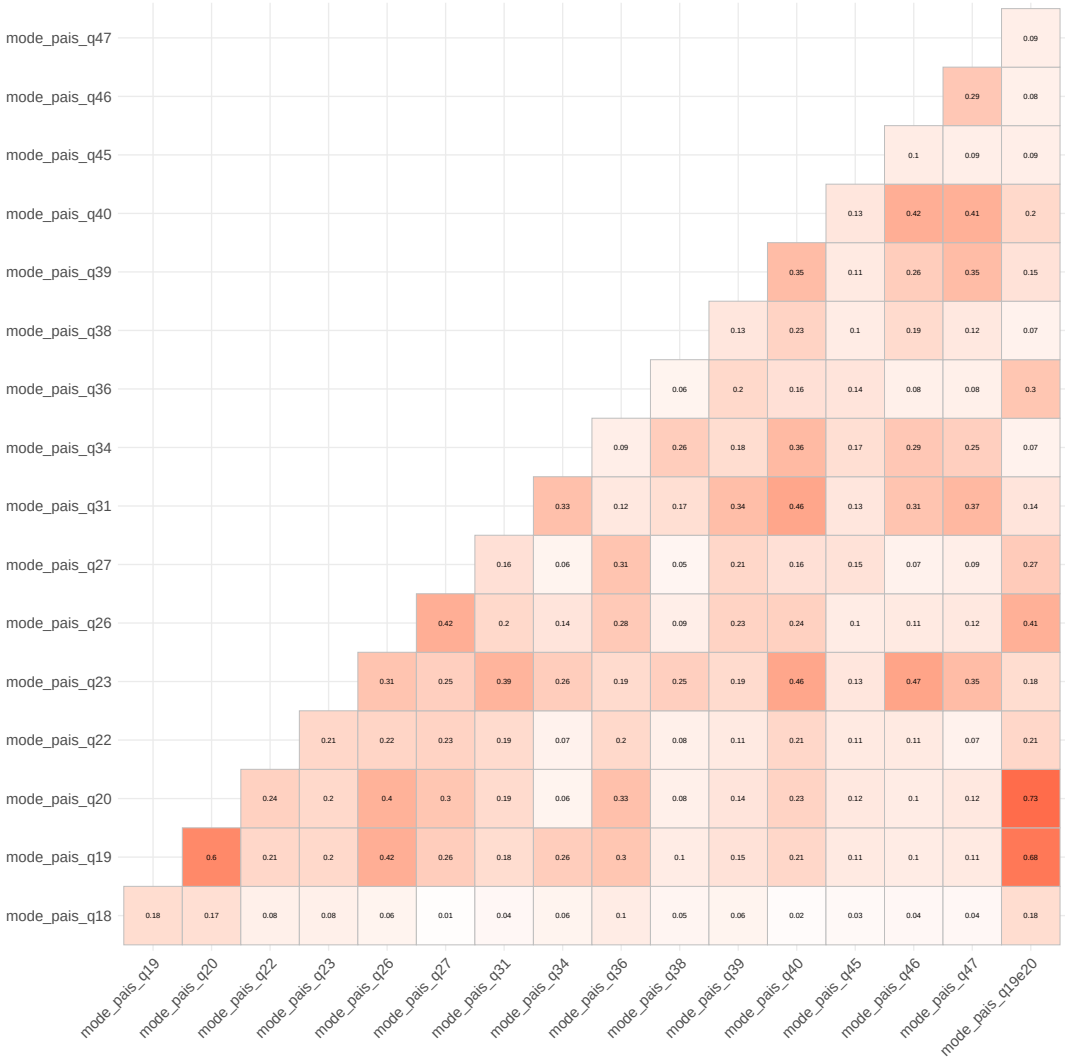
```
##  [1] "mode_pais_q18"    "mode_pais_q19"    "mode_pais_q20"    "mode_pais_q22"    "mode_pais_q23"
##  [6] "mode_pais_q26"    "mode_pais_q27"    "mode_pais_q31"    "mode_pais_q34"    "mode_pais_q36"
## [11] "mode_pais_q38"    "mode_pais_q39"    "mode_pais_q40"    "mode_pais_q45"    "mode_pais_q46"
## [16] "mode_pais_q47"    "mode_pais_q19e20"
```

```
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 0.001278939
## 2     mode_pais_q19 0.067813271
## 3     mode_pais_q20 0.075943421
## 4     mode_pais_q22 0.034688144
## 5     mode_pais_q23 0.227685862
## 6     mode_pais_q26 0.098311229
## 7     mode_pais_q27 0.056846510
## 8     mode_pais_q31 0.186641489
## 9     mode_pais_q34 0.106198064
## 10    mode_pais_q36 0.043615100
## 11    mode_pais_q38 0.045762133
## 12    mode_pais_q39 0.158987288
## 13    mode_pais_q40 0.263138879
## 14    mode_pais_q45 0.019305677
## 15    mode_pais_q46 0.189081532
## 16    mode_pais_q47 0.216287414
## 17 mode_pais_q19e20 0.046421942
```

```
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=2, show.legend = F)
```


mode_pais_q47
mode_pais_q46
mode_pais_q45
mode_pais_q40
mode_pais_q39
mode_pais_q38
mode_pais_q36
mode_pais_q34
mode_pais_q31
mode_pais_q27
mode_pais_q26
mode_pais_q23
mode_pais_q22
mode_pais_q20
mode_pais_q19
mode_pais_q18
mode_pais_q19
mode_pais_q20
mode_pais_q22
mode_pais_q23
mode_pais_q26
mode_pais_q27
mode_pais_q31
mode_pais_q34
mode_pais_q36
mode_pais_q38
mode_pais_q39
mode_pais_q40
mode_pais_q45
mode_pais_q46
mode_pais_q47
mode_pais_q19e20