Análise das variáveis Saresp Questionário pais - moda por escola

Série 3EM

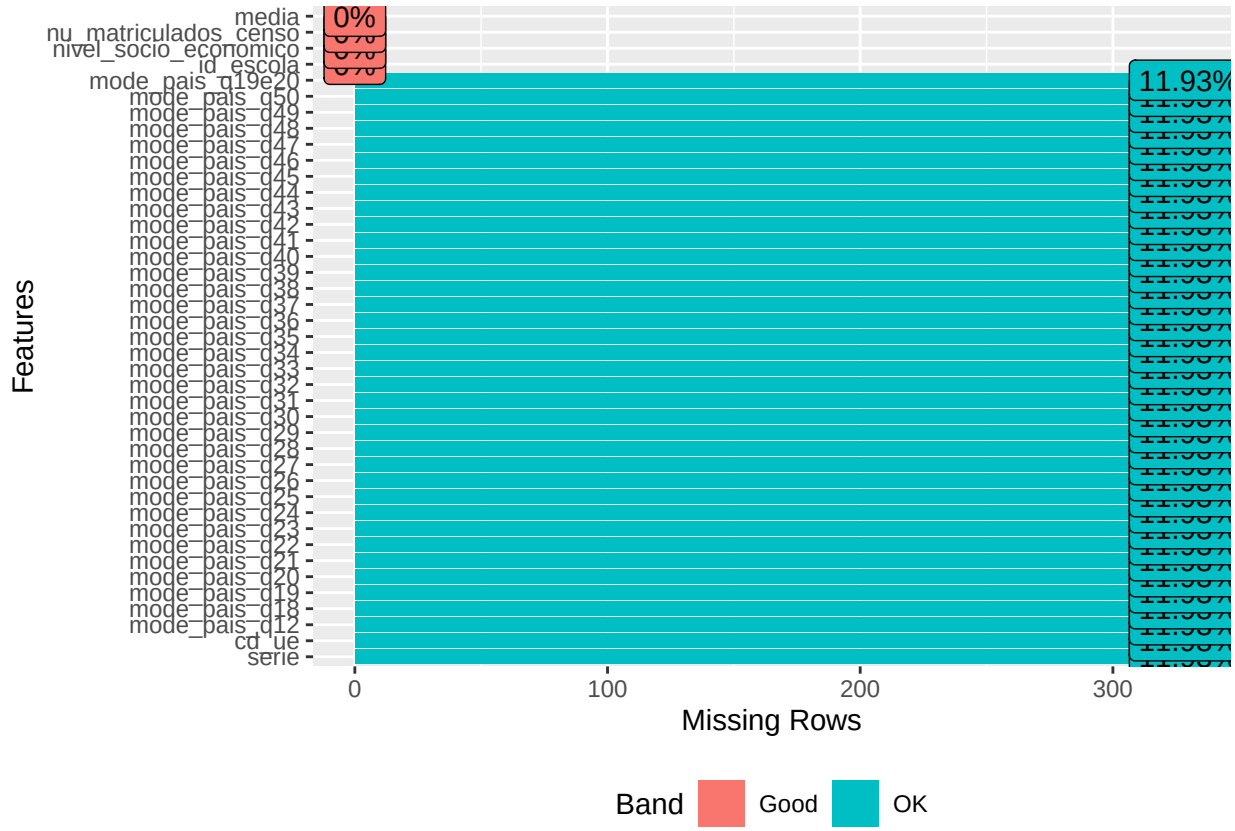
Livia Kobayashi

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(params$book)
```

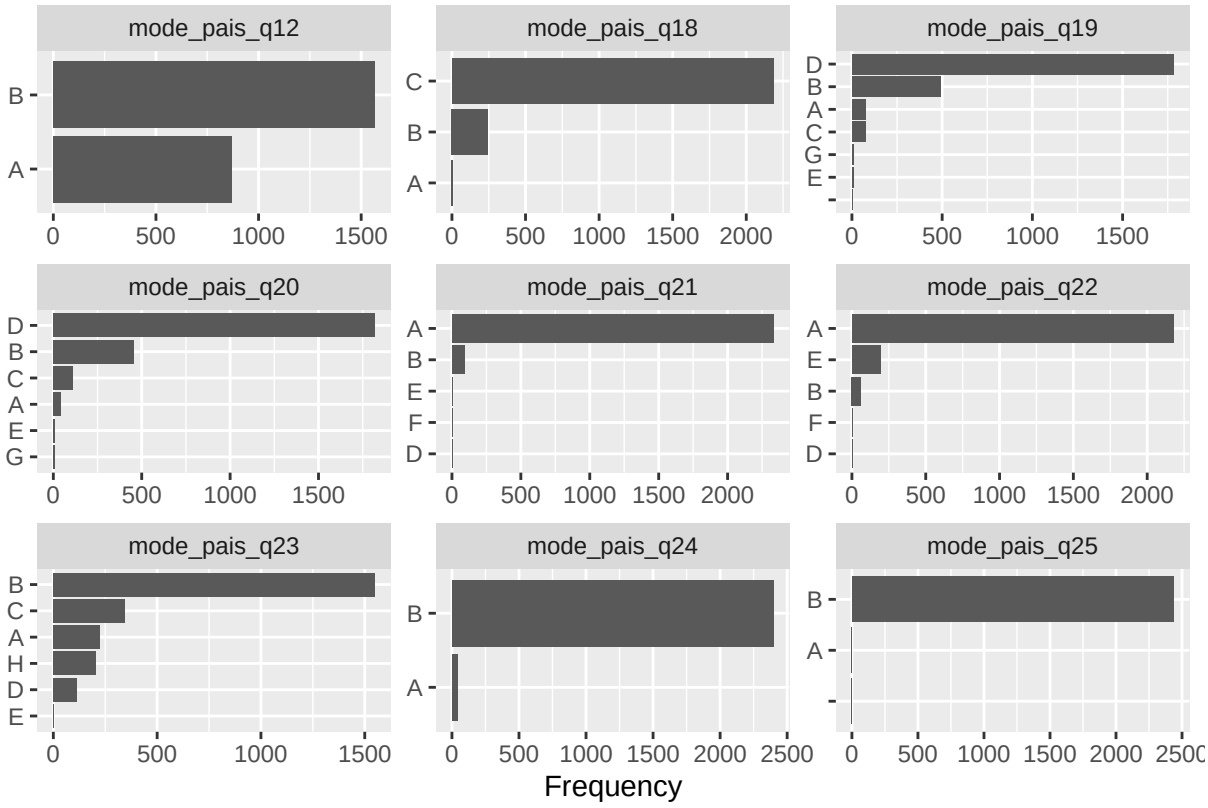
Missing

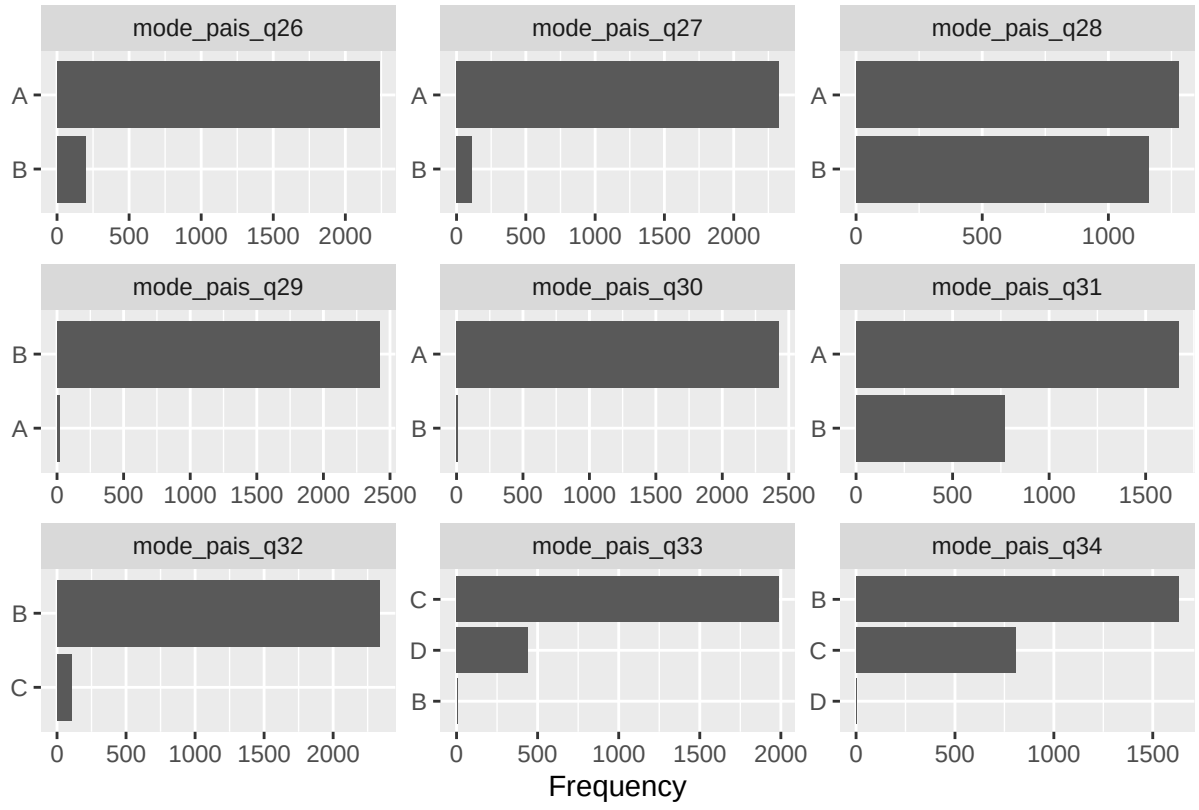
```
plot_missing(df)
```

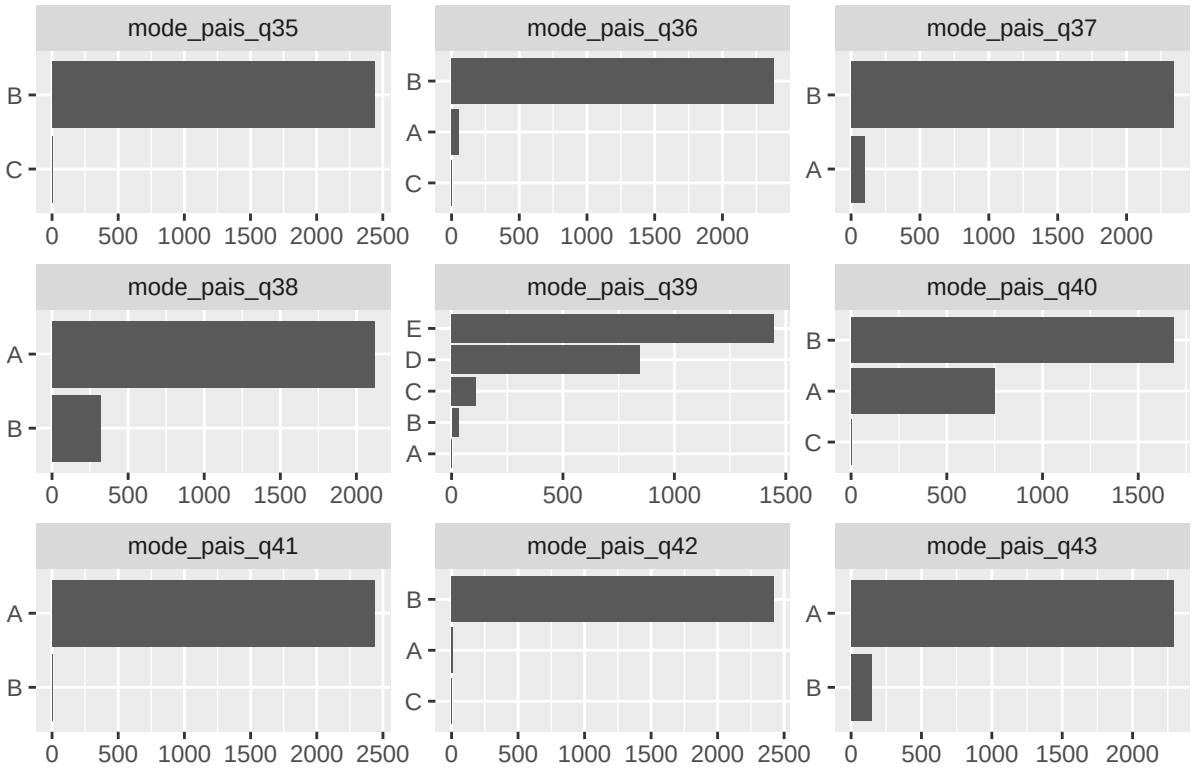


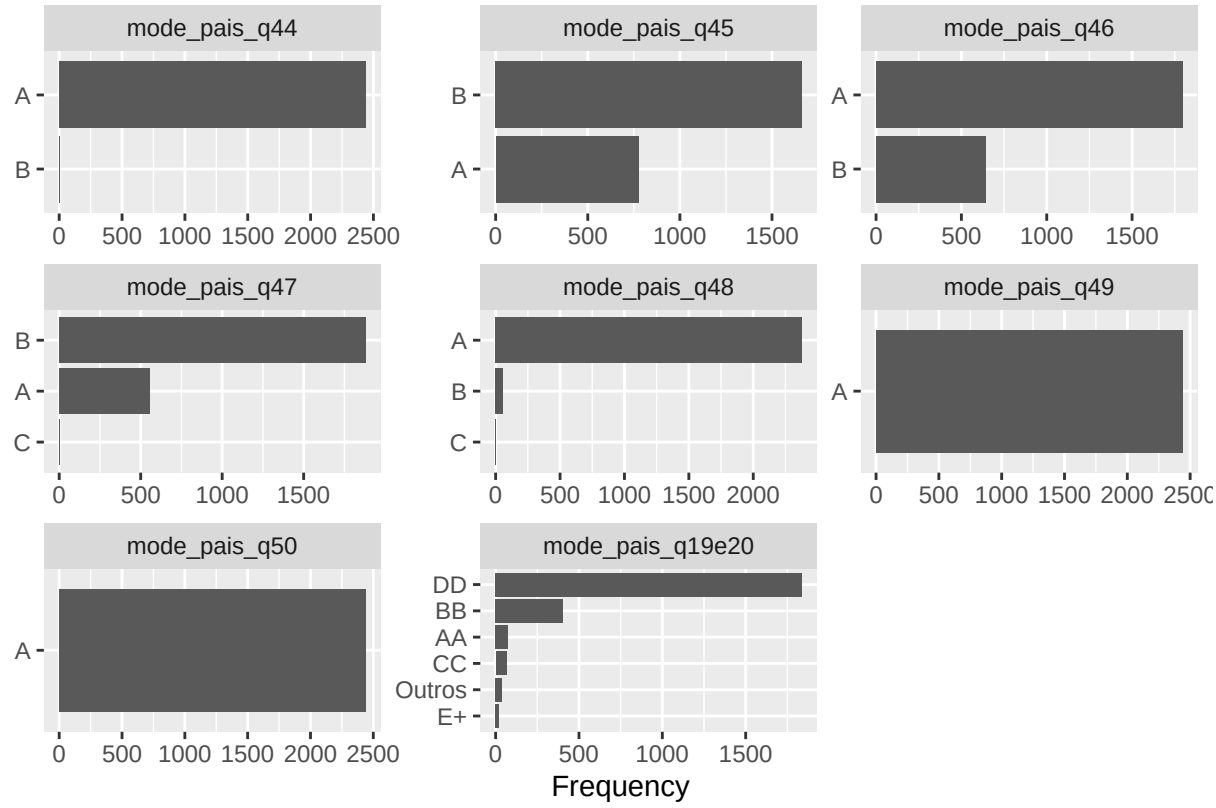
Volume

```
plot_bar(final_data)
```


mode_pais_q12
B
A
0
500
1000
1500
mode_pais_q18
C
B
A
0
500
1000
1500
2000
mode_pais_q19
D
B
A
C
G
E
0
500
1000
1500
mode_pais_q20
D
B
C
A
E
G
0
500
1000
1500
mode_pais_q21
A
B
E
F
D
0
500
1000
1500
2000
mode_pais_q22
A
E
B
F
D
0
500
1000
1500
2000
mode_pais_q23
B
C
A
H
D
E
0
500
1000
1500
mode_pais_q24
B
A
0
500
1000
1500
2000
2500
mode_pais_q25
B
A
0
500
1000
1500
2000
2500
Frequency

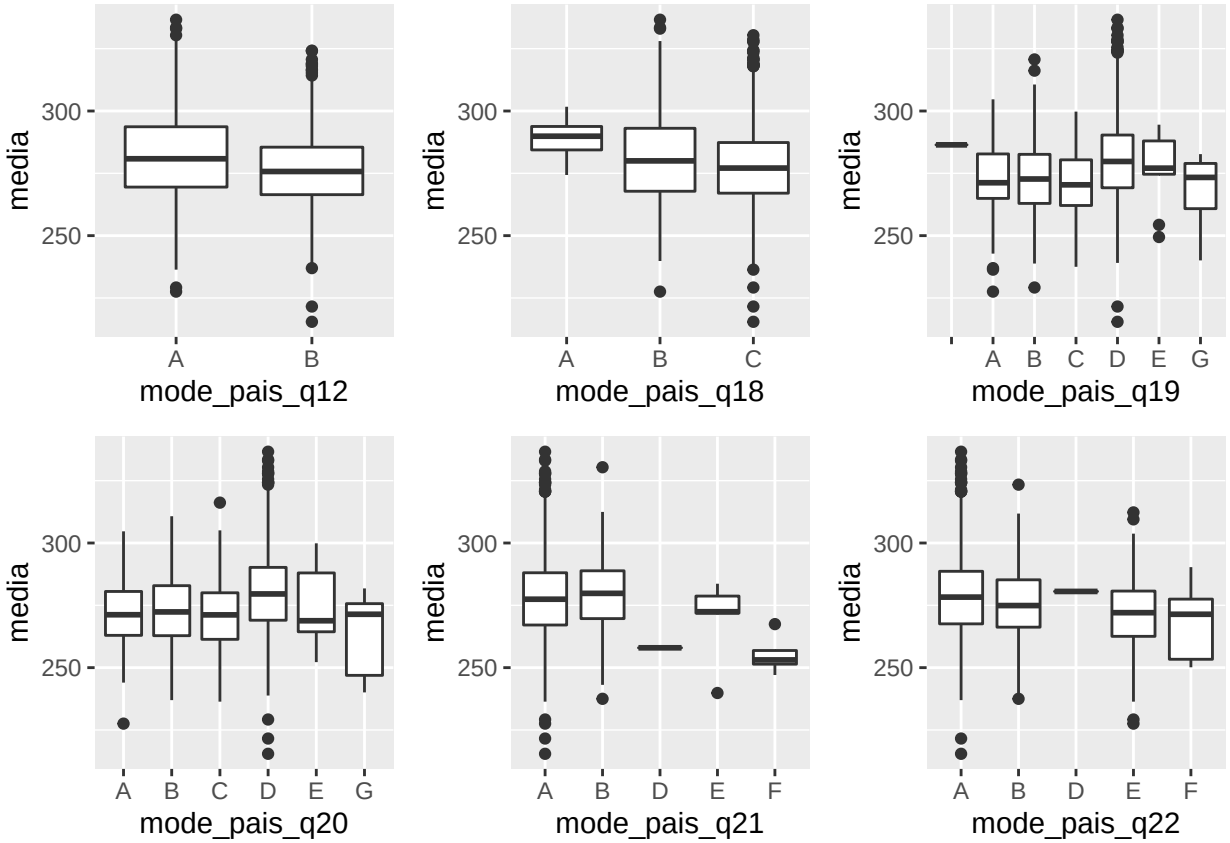

mode_pais_q26
mode_pais_q27
mode_pais_q28
mode_pais_q29
mode_pais_q30
mode_pais_q31
mode_pais_q32
mode_pais_q33
mode_pais_q34
Frequency

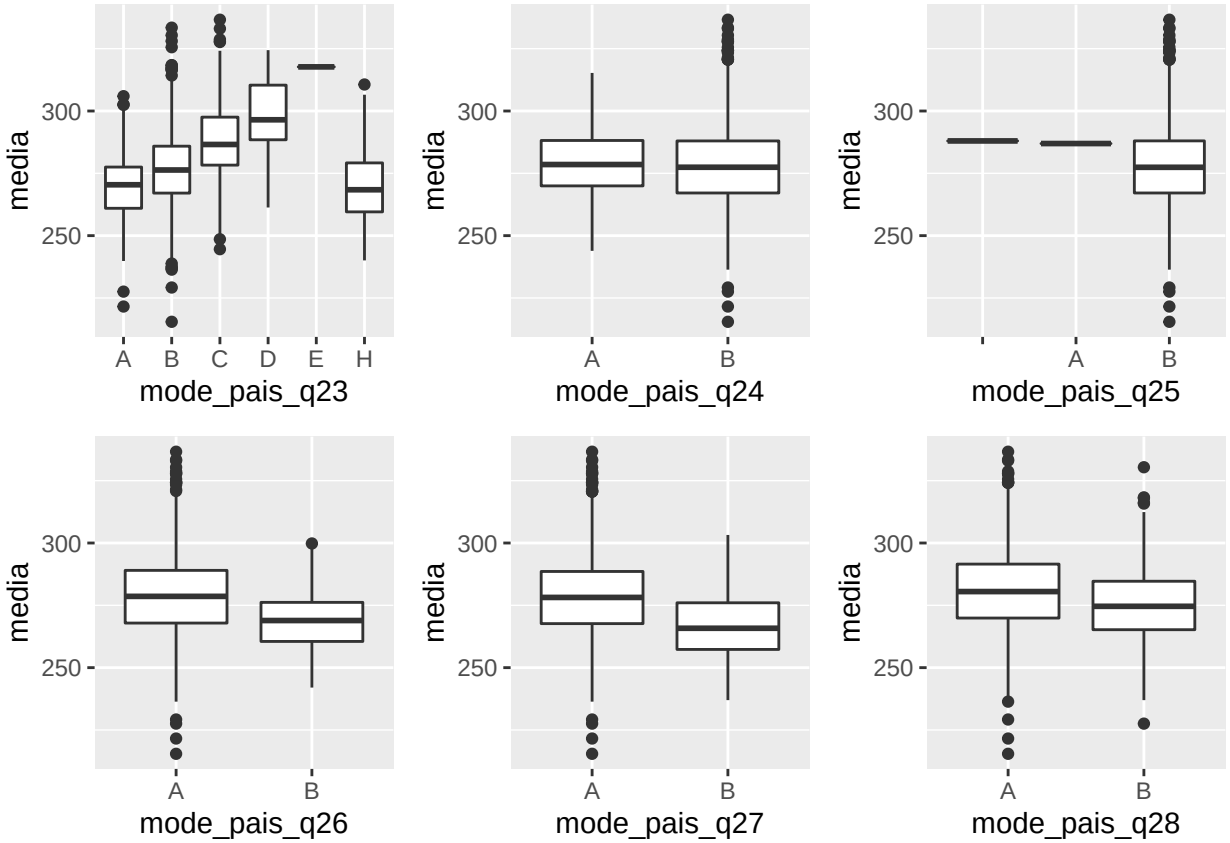

mode_pais_q35
B
C
0 500 1000 1500 2000 2500
mode_pais_q36
B
A
C
0 500 1000 1500 2000
mode_pais_q37
B
A
0 500 1000 1500 2000
mode_pais_q38
A
B
0 500 1000 1500 2000
mode_pais_q39
E
D
C
B
A
0 500 1000 1500
mode_pais_q40
B
A
C
0 500 1000 1500
mode_pais_q41
A
B
0 500 1000 1500 2000 2500
mode_pais_q42
B
A
C
0 500 1000 1500 2000 2500
mode_pais_q43
A
B
0 500 1000 1500 2000
Frequency

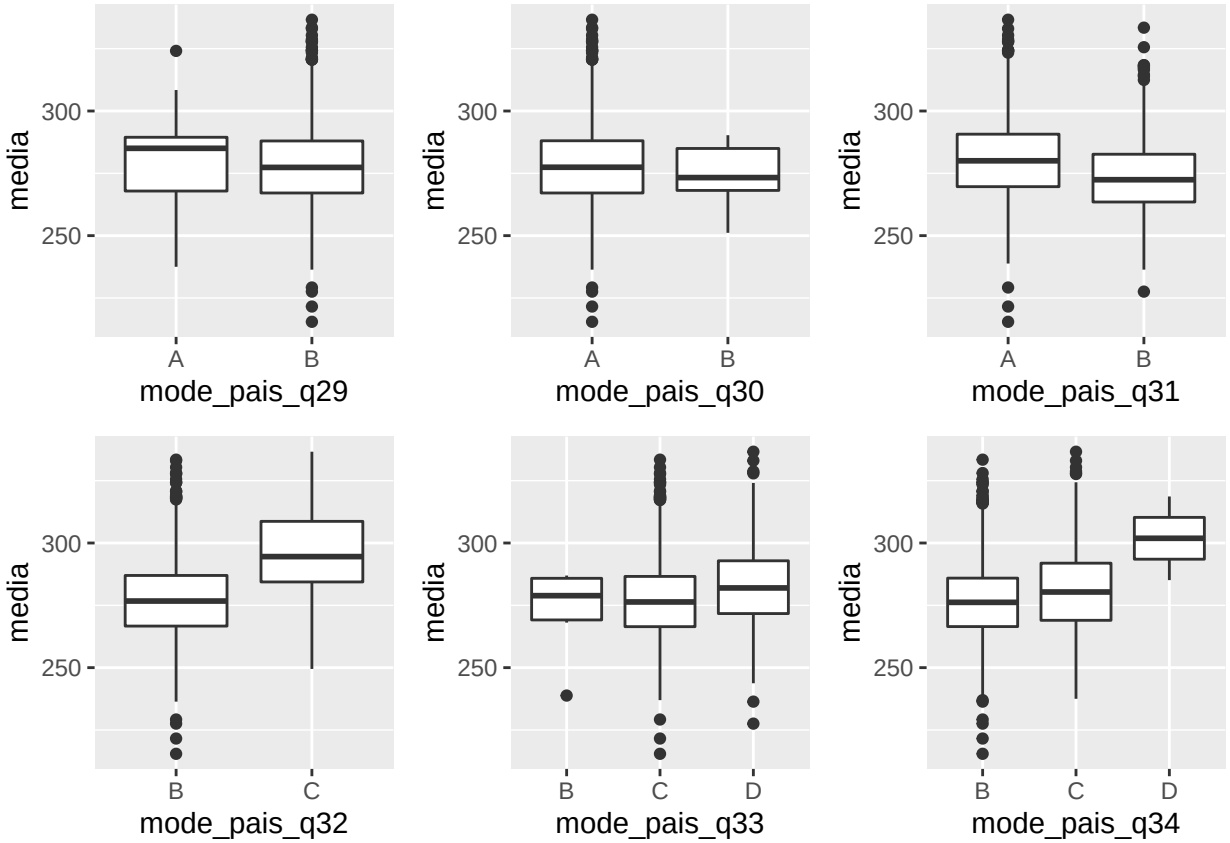


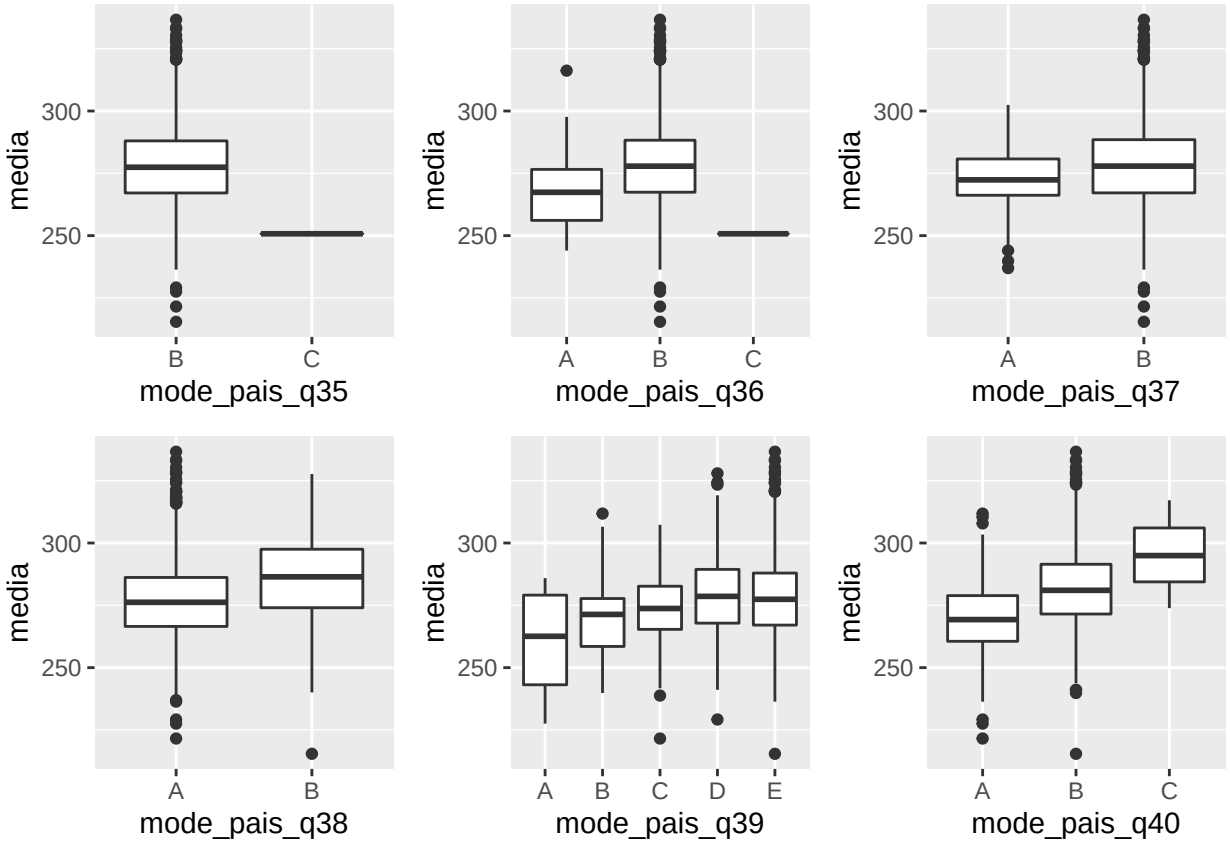
Boxplot

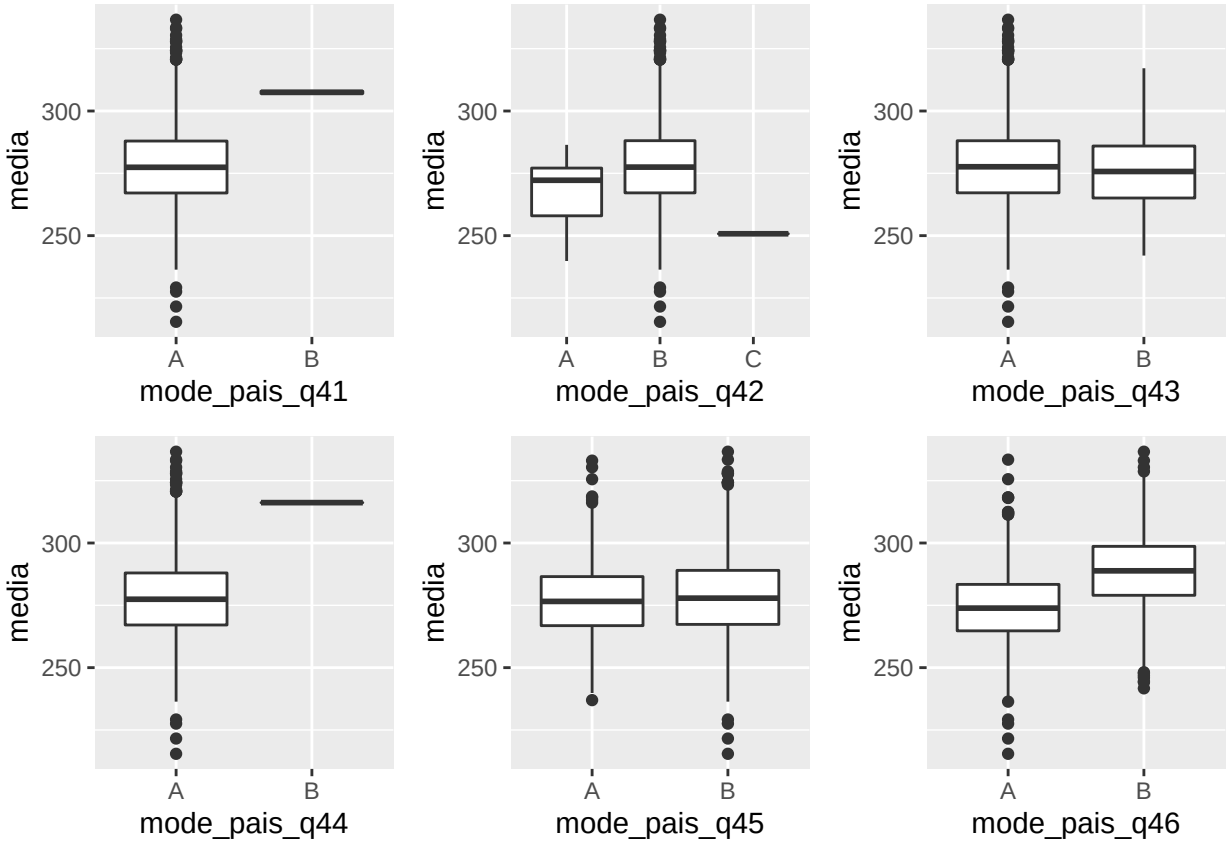
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

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
}
```

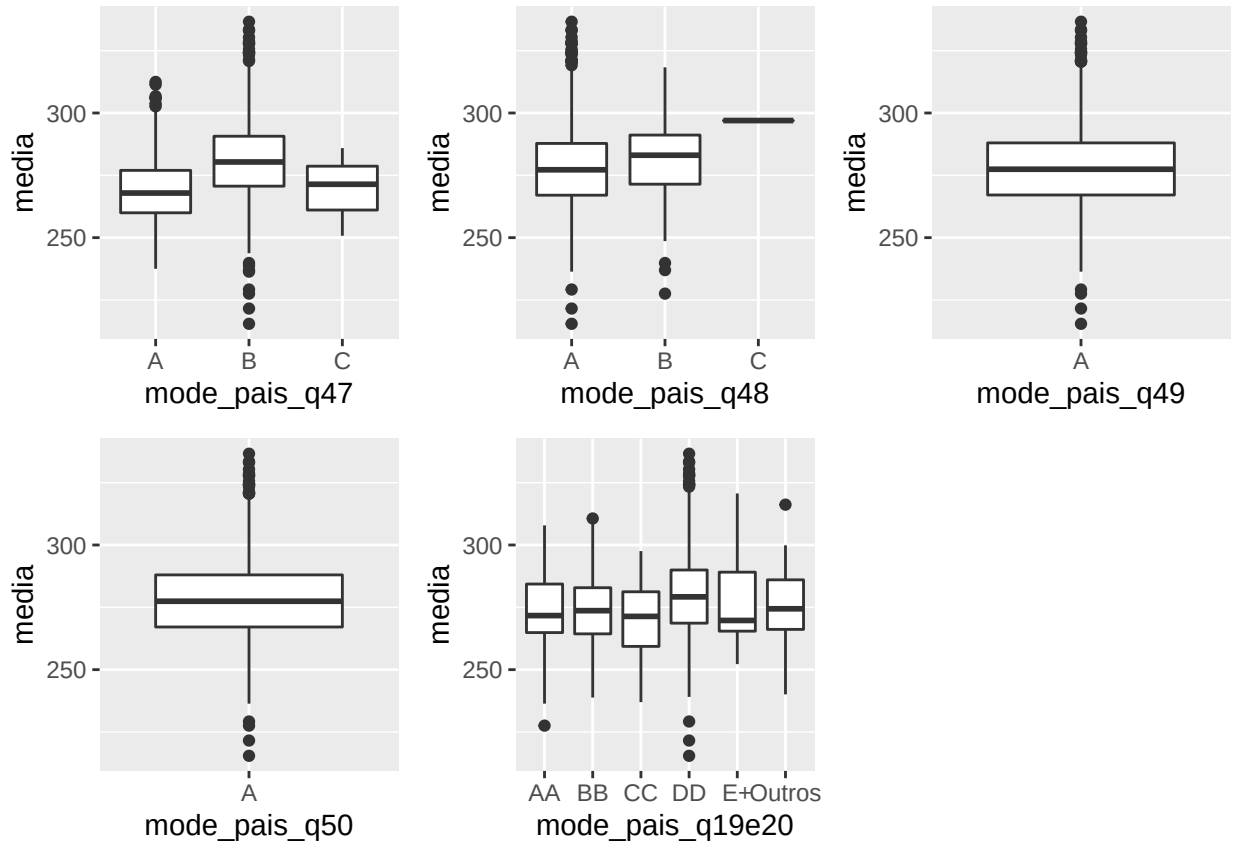

media
300
250
A
B
mode_pais_q12
media
300
250
A
B
C
mode_pais_q18
media
300
250
A
B
C
D
E
G
mode_pais_q19
media
300
250
A
B
C
D
E
G
mode_pais_q20
media
300
250
A
B
D
E
F
mode_pais_q21
media
300
250
A
B
D
E
F
mode_pais_q22


media
300
250
A B C D E H
mode_pais_q23
media
300
250
A B
mode_pais_q24
media
300
250
A B
mode_pais_q25
media
300
250
A B
mode_pais_q26
media
300
250
A B
mode_pais_q27
media
300
250
A B
mode_pais_q28


media
300
250
A
B
mode_pais_q29
media
300
250
A
B
mode_pais_q30
media
300
250
A
B
mode_pais_q31
media
300
250
B
C
mode_pais_q32
media
300
250
B
C
D
mode_pais_q33
media
300
250
B
C
D
mode_pais_q34


media
300
250
B
C
mode_pais_q35
media
300
250
A
B
C
mode_pais_q36
media
300
250
A
B
mode_pais_q37
media
300
250
A
B
mode_pais_q38
media
300
250
A
B
C
D
E
mode_pais_q39
media
300
250
A
B
C
mode_pais_q40


media
300
250
A
B
mode_pais_q41
media
300
250
A
B
C
mode_pais_q42
media
300
250
A
B
mode_pais_q43
media
300
250
A
B
mode_pais_q44
media
300
250
A
B
mode_pais_q45
media
300
250
A
B
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
```

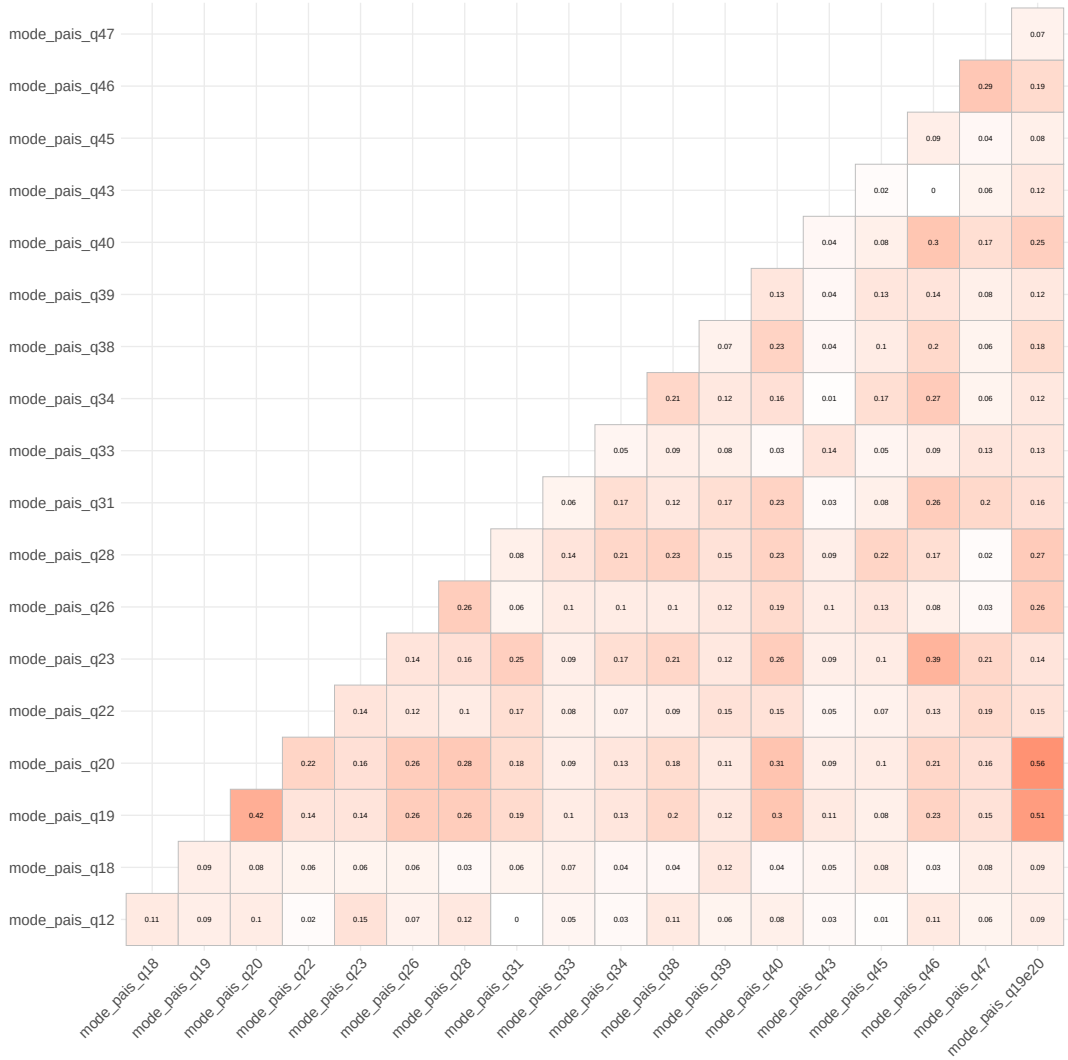
```
##  [1] "mode_pais_q12"    "mode_pais_q18"    "mode_pais_q19"    "mode_pais_q20"    "mode_pais_q22"
##  [6] "mode_pais_q23"    "mode_pais_q26"    "mode_pais_q28"    "mode_pais_q31"    "mode_pais_q33"
## [11] "mode_pais_q34"    "mode_pais_q38"    "mode_pais_q39"    "mode_pais_q40"    "mode_pais_q43"
## [16] "mode_pais_q45"    "mode_pais_q46"    "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_q12 0.0305360025
## 2     mode_pais_q18 0.0063650511
## 3     mode_pais_q19 0.0511221223
## 4     mode_pais_q20 0.0489465496
## 5     mode_pais_q22 0.0139599360
## 6     mode_pais_q23 0.1747604750
## 7     mode_pais_q26 0.0300737222
## 8     mode_pais_q28 0.0366676140
## 9     mode_pais_q31 0.0417335441
## 10    mode_pais_q33 0.0181893585
## 11    mode_pais_q34 0.0190818420
## 12    mode_pais_q38 0.0354549924
## 13    mode_pais_q39 0.0102577151
## 14    mode_pais_q40 0.1282199753
## 15    mode_pais_q43 0.0008581503
## 16    mode_pais_q45 0.0023292533
## 17    mode_pais_q46 0.1682761862
## 18    mode_pais_q47 0.0956628970
## 19 mode_pais_q19e20 0.0348846672
```

```
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_q43
mode_pais_q40
mode_pais_q39
mode_pais_q38
mode_pais_q34
mode_pais_q33
mode_pais_q31
mode_pais_q28
mode_pais_q26
mode_pais_q23
mode_pais_q22
mode_pais_q20
mode_pais_q19
mode_pais_q18
mode_pais_q12
mode_pais_q18
mode_pais_q19
mode_pais_q20
mode_pais_q22
mode_pais_q23
mode_pais_q26
mode_pais_q28
mode_pais_q31
mode_pais_q33
mode_pais_q34
mode_pais_q38
mode_pais_q39
mode_pais_q40
mode_pais_q43
mode_pais_q45
mode_pais_q46
mode_pais_q47
mode_pais_q19e20