

Análise das variáveis Saresp Questionário - moda por escola

Série 9EF

Livia Kobayashi

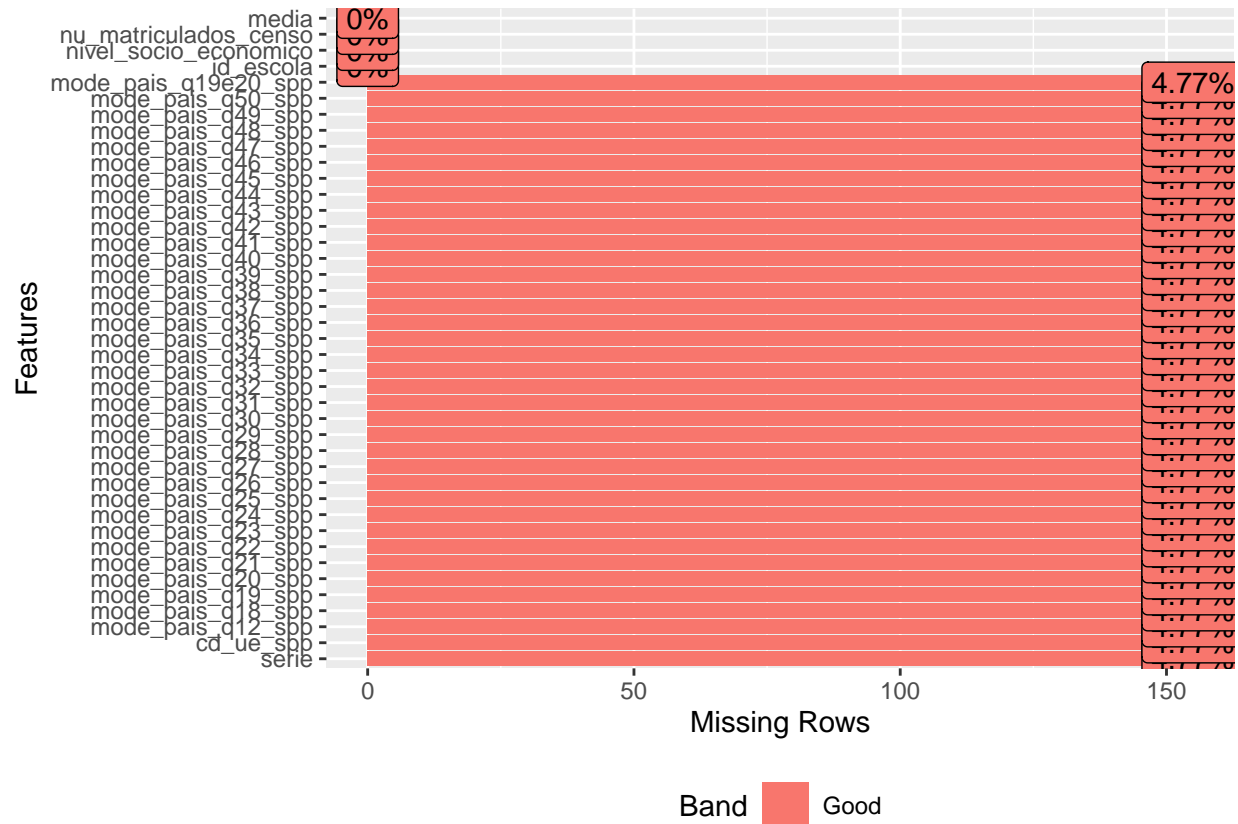
14 junho 2021

```
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      9EF
```

Missing: 4,7% de dados faltantes

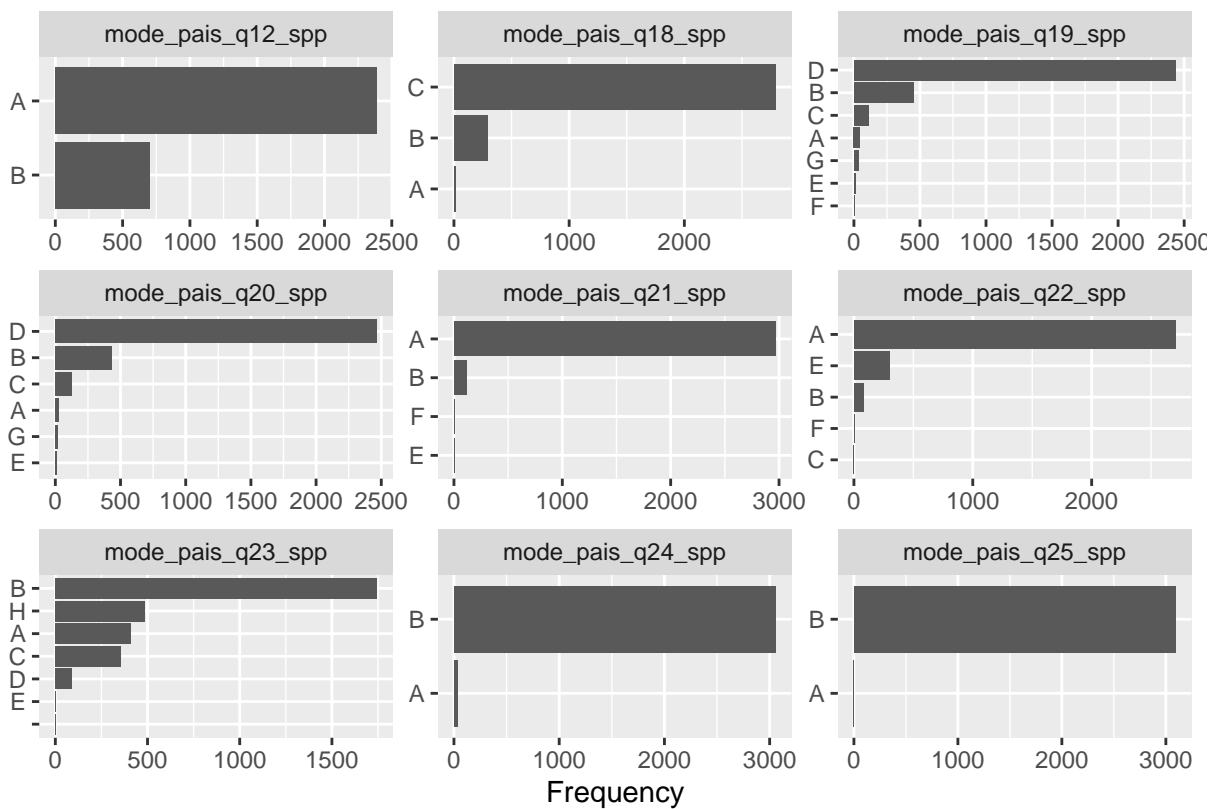
```
plot_missing(df)
```

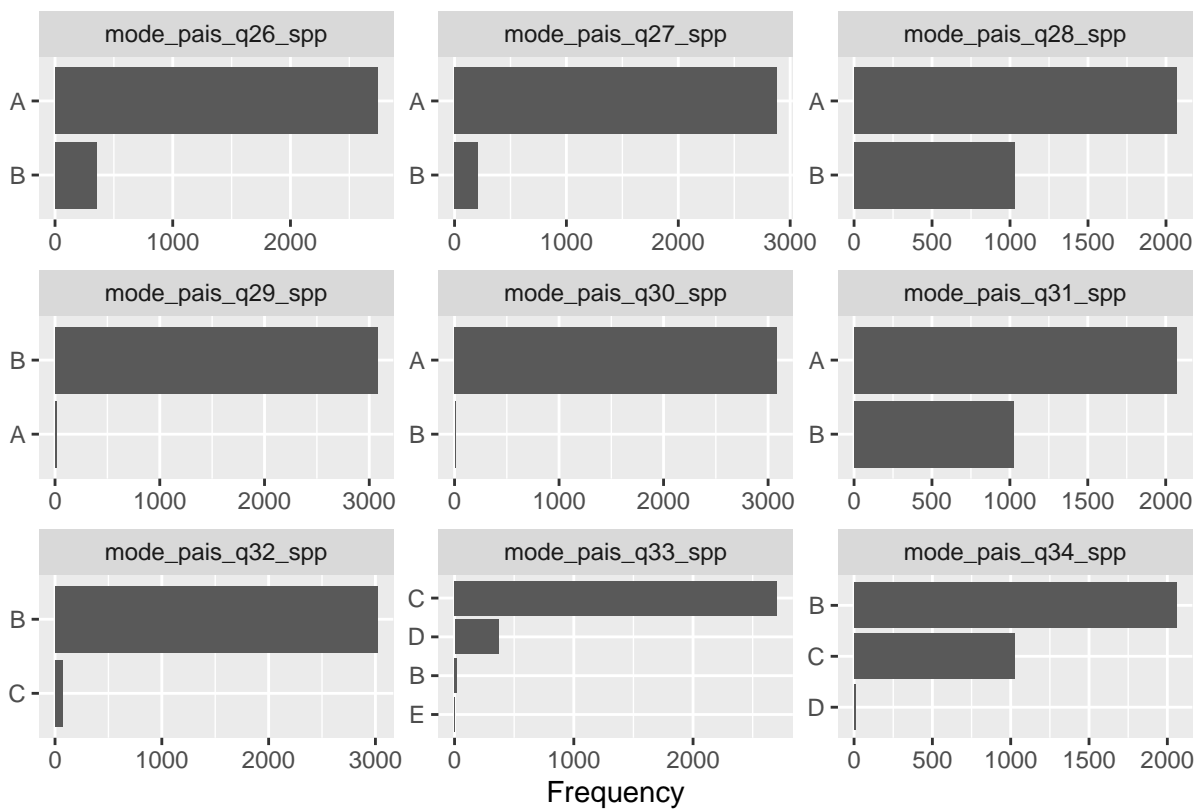


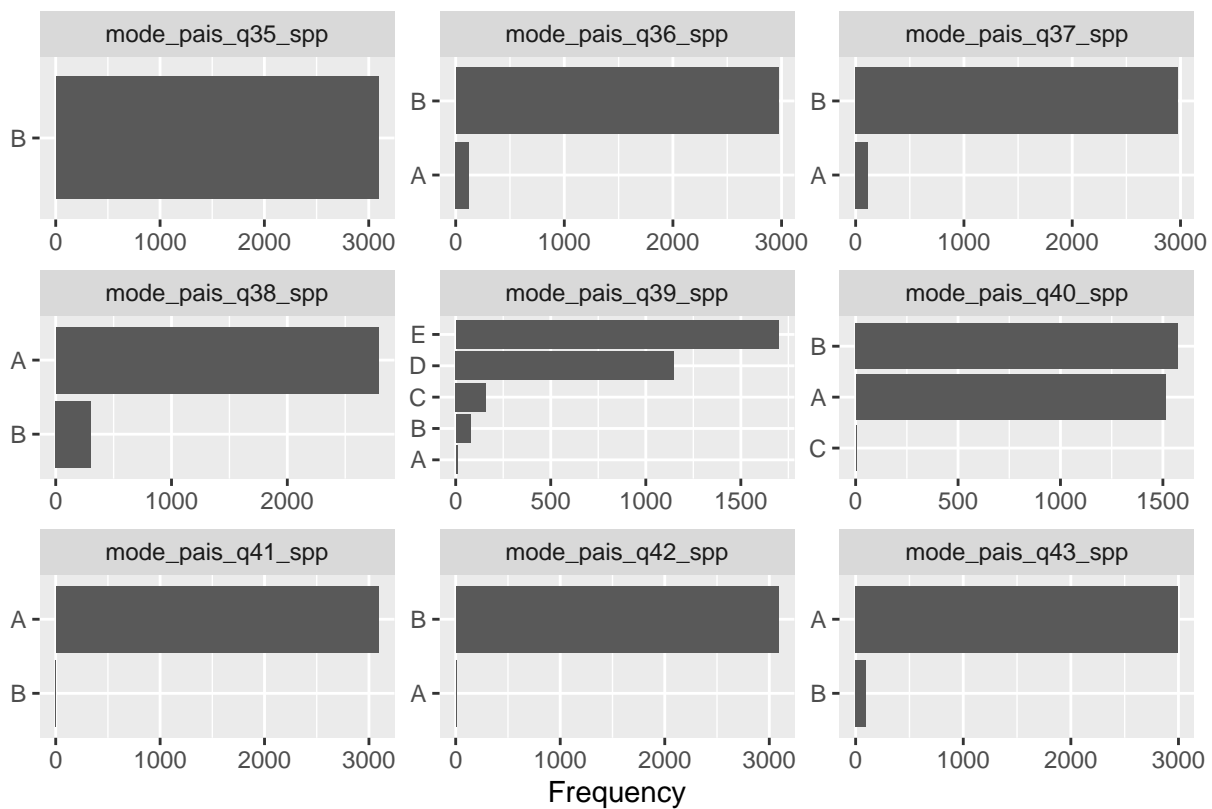
Volume: VARIáveis com bom volume

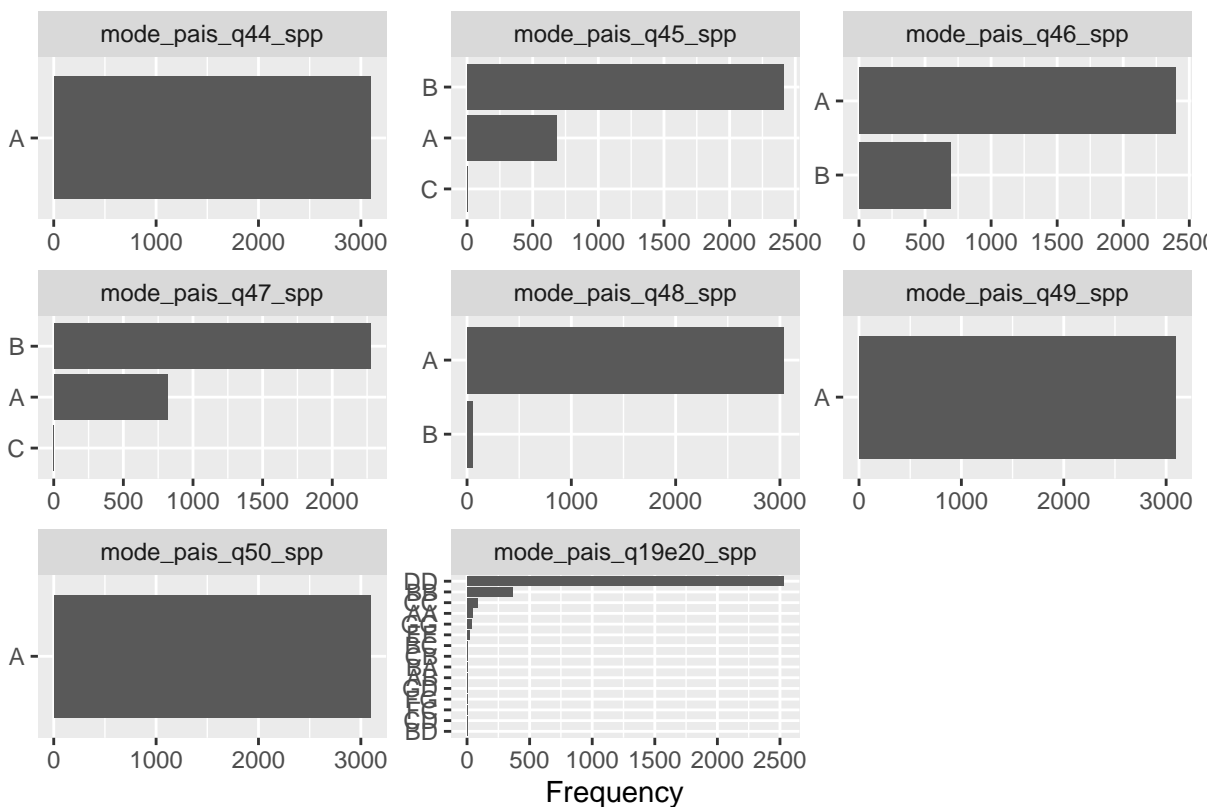
- mode_pais_q12
- mode_pais_q18
- mode_pais_q19
- mode_pais_q20
- mode_pais_q22
- mode_pais_q23
- mode_pais_q26
- mode_pais_q28
- mode_pais_q31
- mode_pais_q33
- mode_pais_q34
- mode_pais_q38
- mode_pais_q39
- mode_pais_q40
- mode_pais_q45
- mode_pais_q46
- mode_pais_q47

```
plot_bar(final_data)
```









Page 4

Boxplot: Variáveis com bom volume e variância

- mode_pais_q19: graduação mãe > ordena (juntar A,B,C, G e E+F)
- mode_pais_q20: graduação pai > ordena (juntar A,B,C, G)
- mode_pais_q22: notas < para mães desempregadas
- mode_pais_q23: renda maior=> notas > (juntar A com H)
- mode_pais_q31: TV assinatura => Notas >
- mode_pais_q34: #TV => Notas > (B x C)
- mode_pais_q38: Telefone => 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)

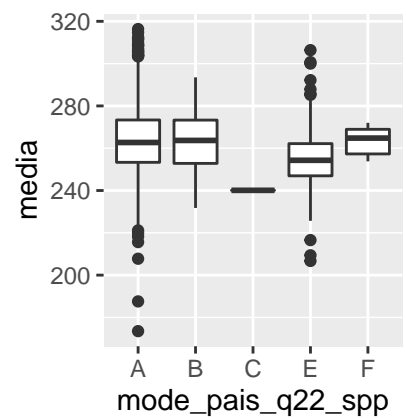
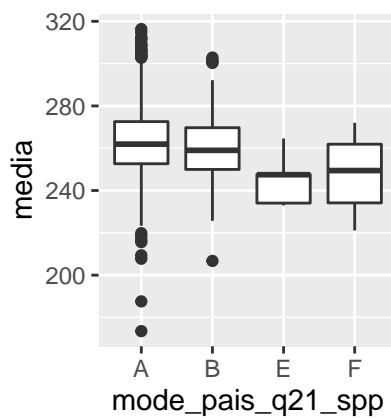
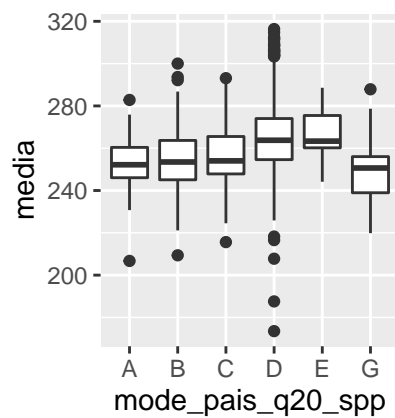
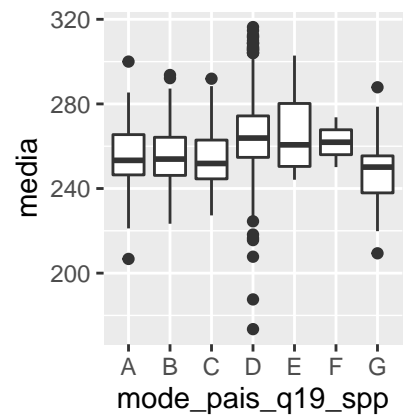
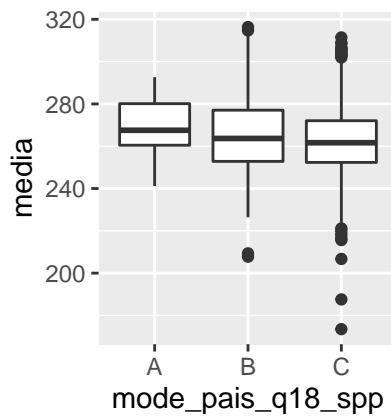
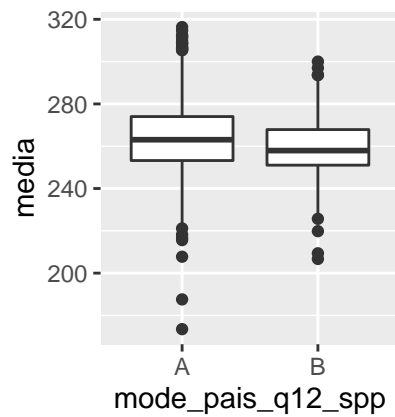
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

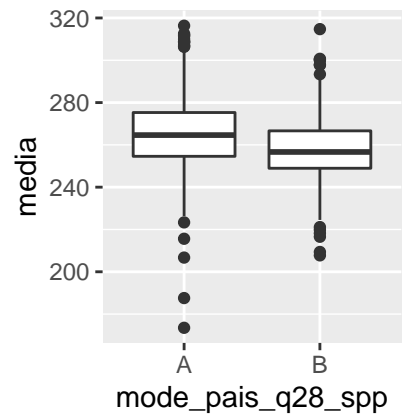
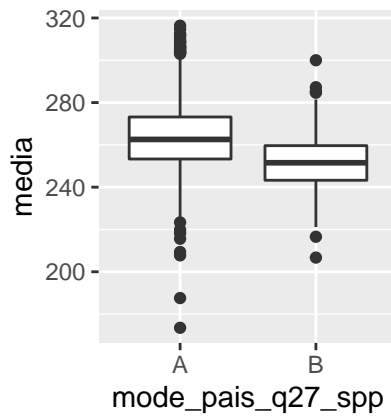
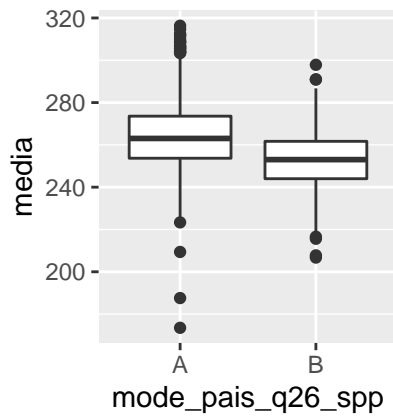
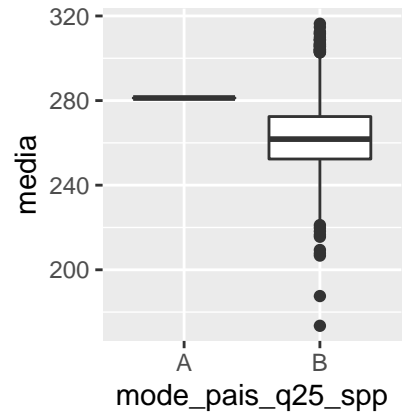
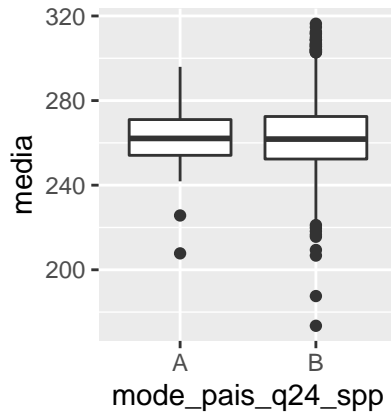
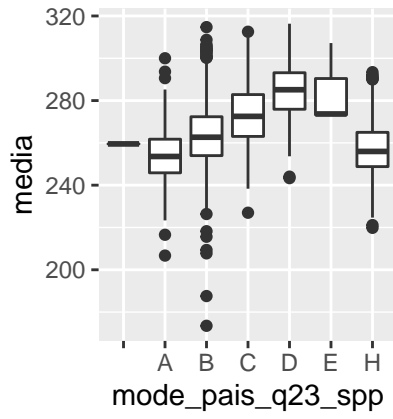
n <- length(plots)
```

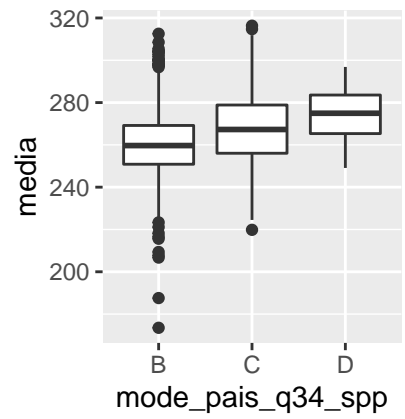
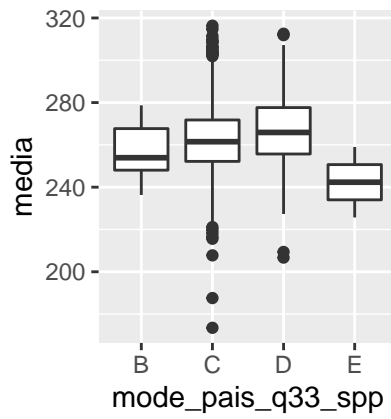
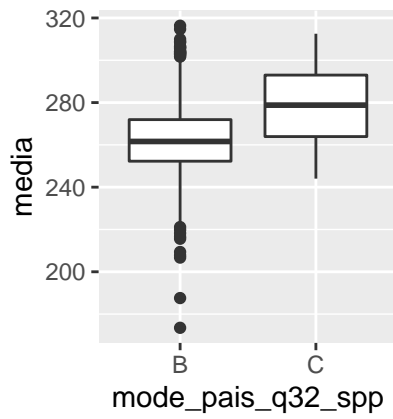
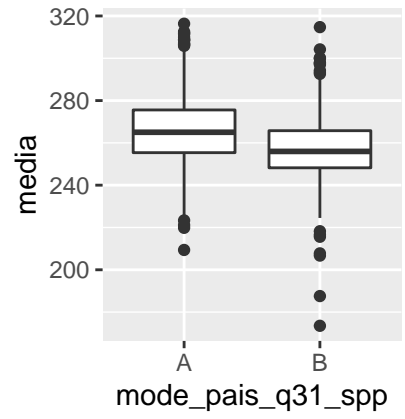
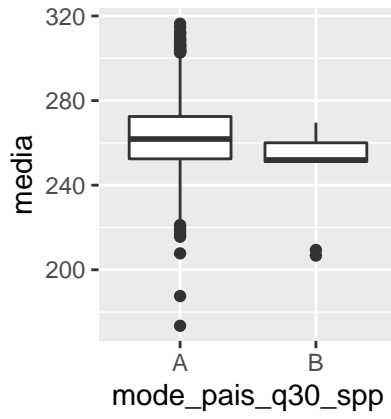
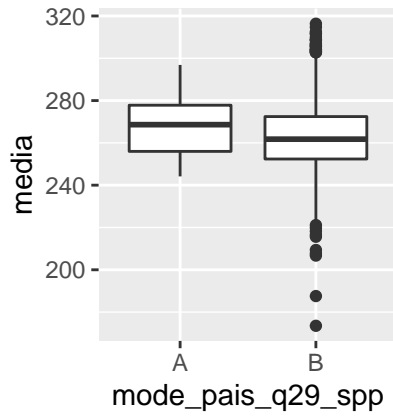
```

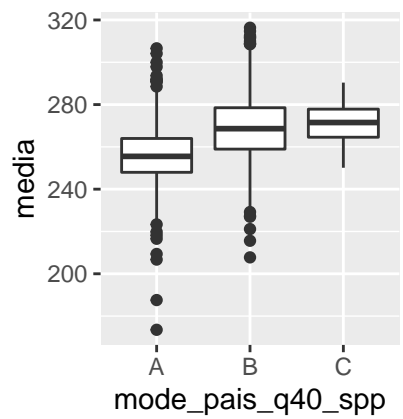
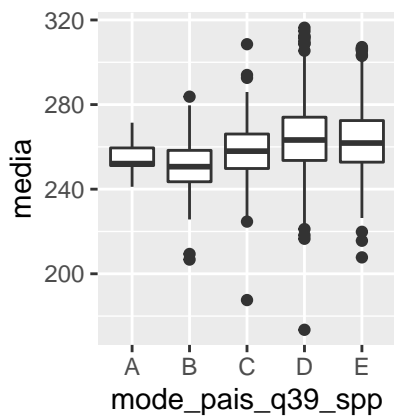
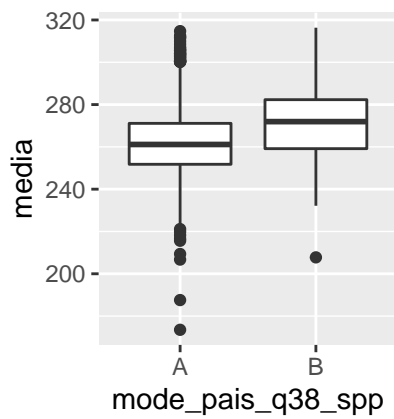
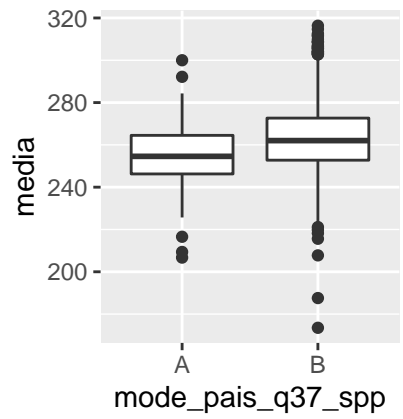
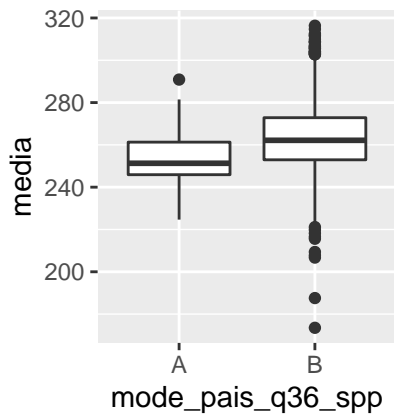
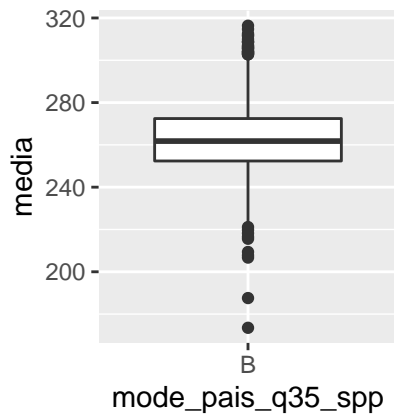
i <- 1
while (i <= n) {
  do.call("grid.arrange", c(plots[i:(min(i+5, n))], ncol=3, nrow = 2))
  i <- i + 6
}

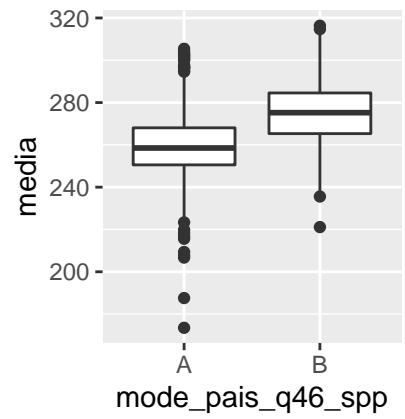
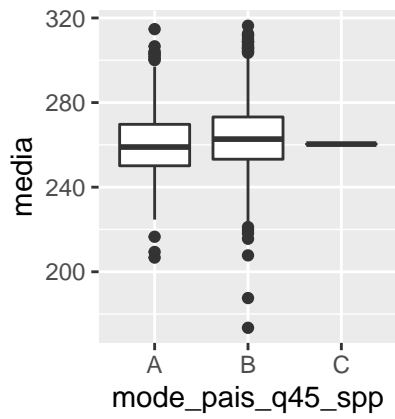
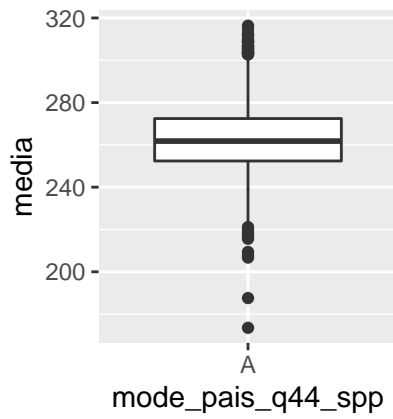
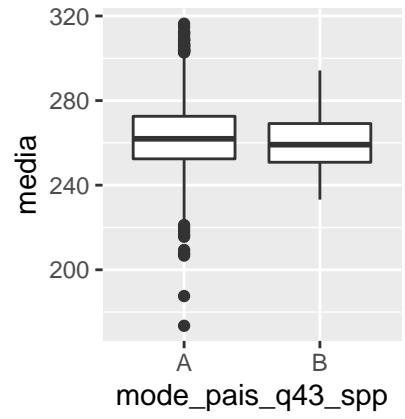
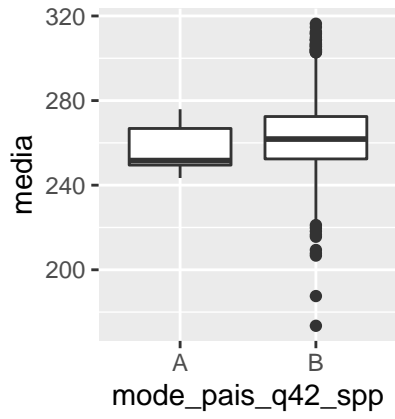
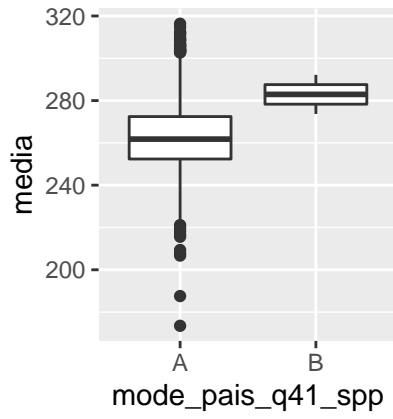
```

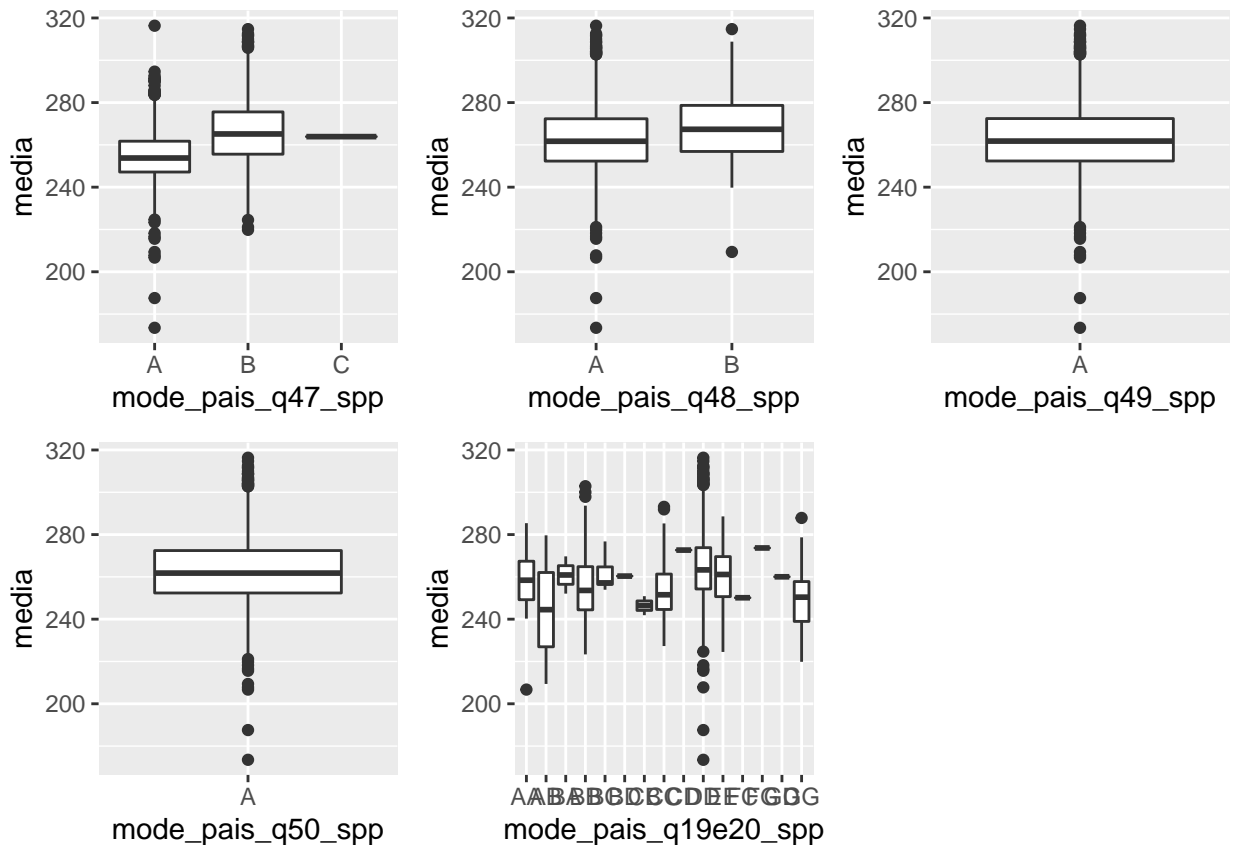












Análise Univariada

- mode_pais_q40: CComputador => 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_q12_spp" "mode_pais_q18_spp" "mode_pais_q19_spp"
## [4] "mode_pais_q20_spp" "mode_pais_q22_spp" "mode_pais_q23_spp"
## [7] "mode_pais_q26_spp" "mode_pais_q27_spp" "mode_pais_q28_spp"
## [10] "mode_pais_q31_spp" "mode_pais_q33_spp" "mode_pais_q34_spp"
## [13] "mode_pais_q38_spp" "mode_pais_q39_spp" "mode_pais_q40_spp"
## [16] "mode_pais_q45_spp" "mode_pais_q46_spp" "mode_pais_q47_spp"
## [19] "mode_pais_q19e20_spp"
```

```
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_spp 0.018614633
## 2  mode_pais_q18_spp 0.004551118
## 3  mode_pais_q19_spp 0.078644949
## 4  mode_pais_q20_spp 0.071754293
## 5  mode_pais_q22_spp 0.031474338
## 6  mode_pais_q23_spp 0.176849647
## 7  mode_pais_q26_spp 0.049320062
## 8  mode_pais_q27_spp 0.037100386
## 9  mode_pais_q28_spp 0.054967895
## 10 mode_pais_q31_spp 0.074907827
## 11 mode_pais_q33_spp 0.010990986
## 12 mode_pais_q34_spp 0.051084242
## 13 mode_pais_q38_spp 0.031791165
## 14 mode_pais_q39_spp 0.026458933
## 15 mode_pais_q40_spp 0.191434824
## 16 mode_pais_q45_spp 0.008219789
## 17 mode_pais_q46_spp 0.186038043
## 18 mode_pais_q47_spp 0.108420230
## 19 mode_pais_q19e20_spp 0.060433159
```

Matriz de correlação (60%)

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

```
catcorrmm <- function(vars, dat) sapply(vars, function(y) sapply(vars, function(x) assocstats(table(dat[,
matriz <- catcorrmm(final_cols, data_corr)

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

