Análise das variáveis Saresp Questionário - moda por escola

Série 3EM

Livia Kobayashi

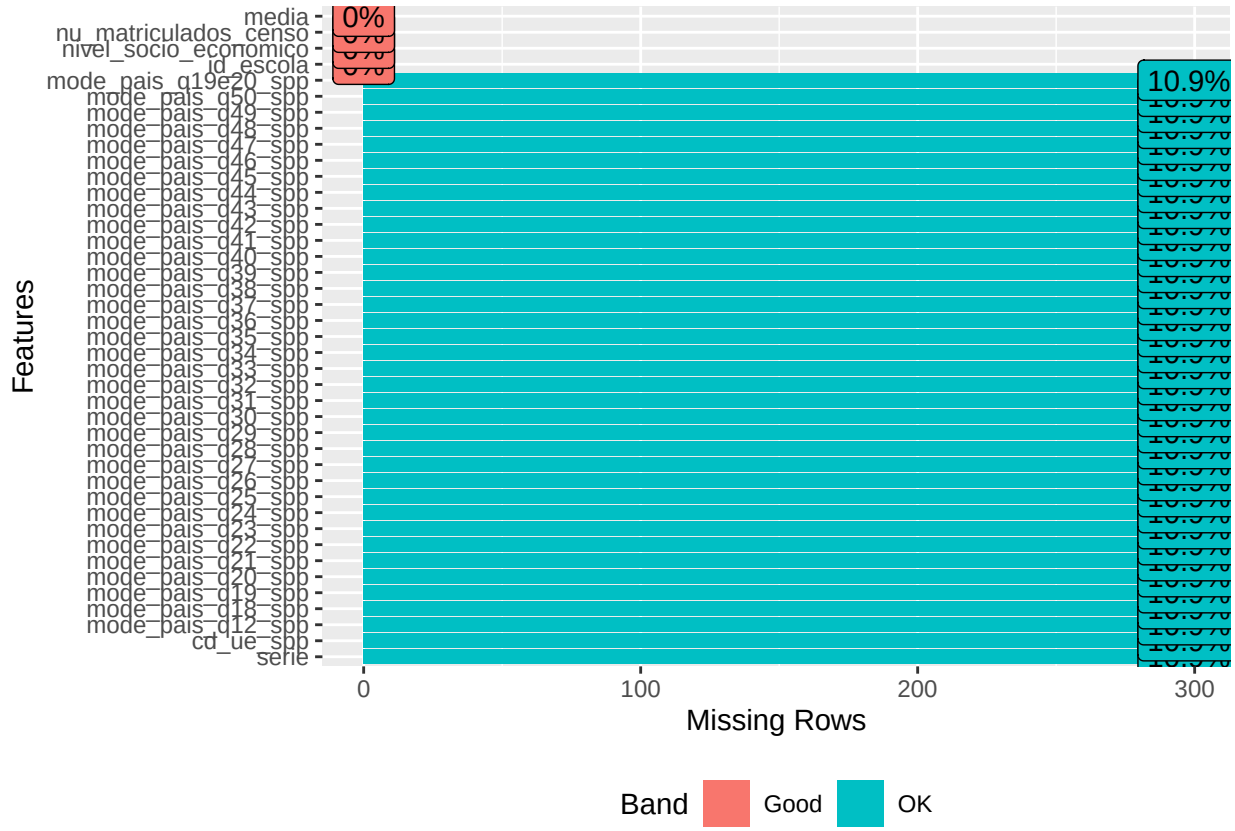
14 junho 2021

```
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      3EM
```

Missing: 10,9% de dados faltantes

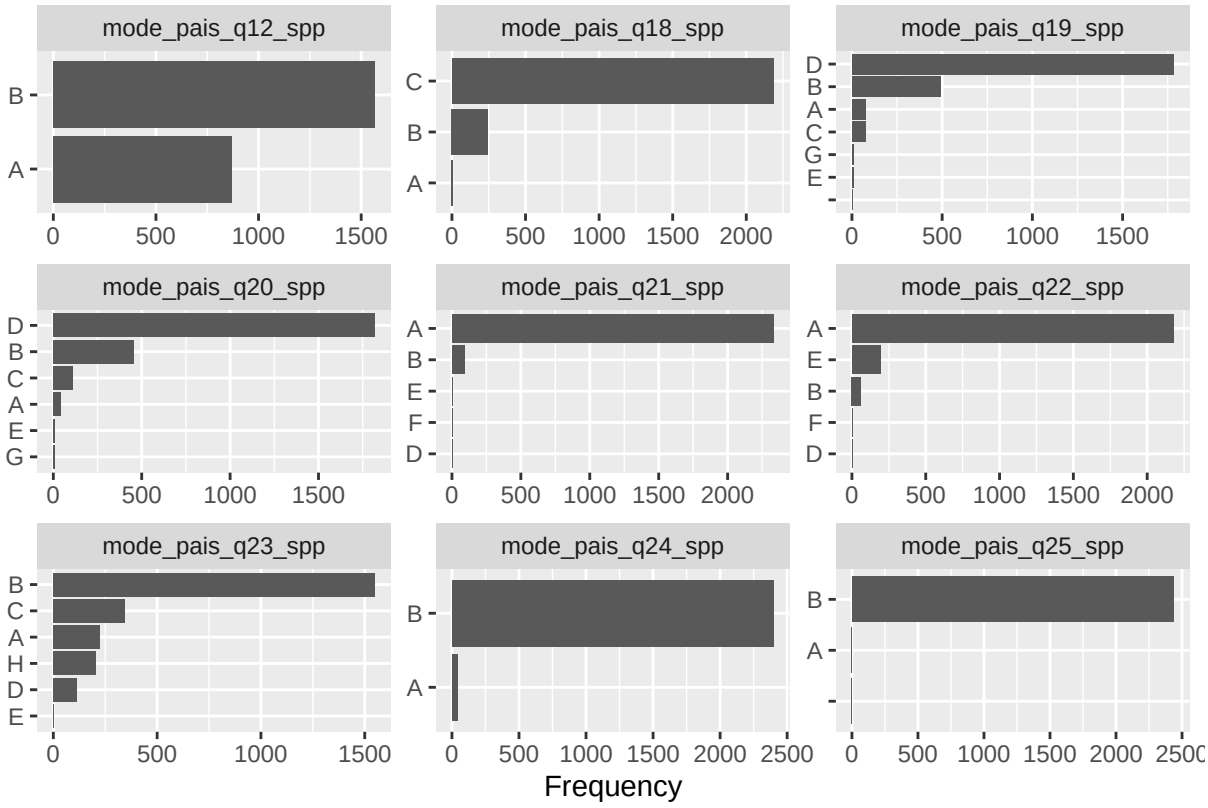
```
plot_missing(df)
```

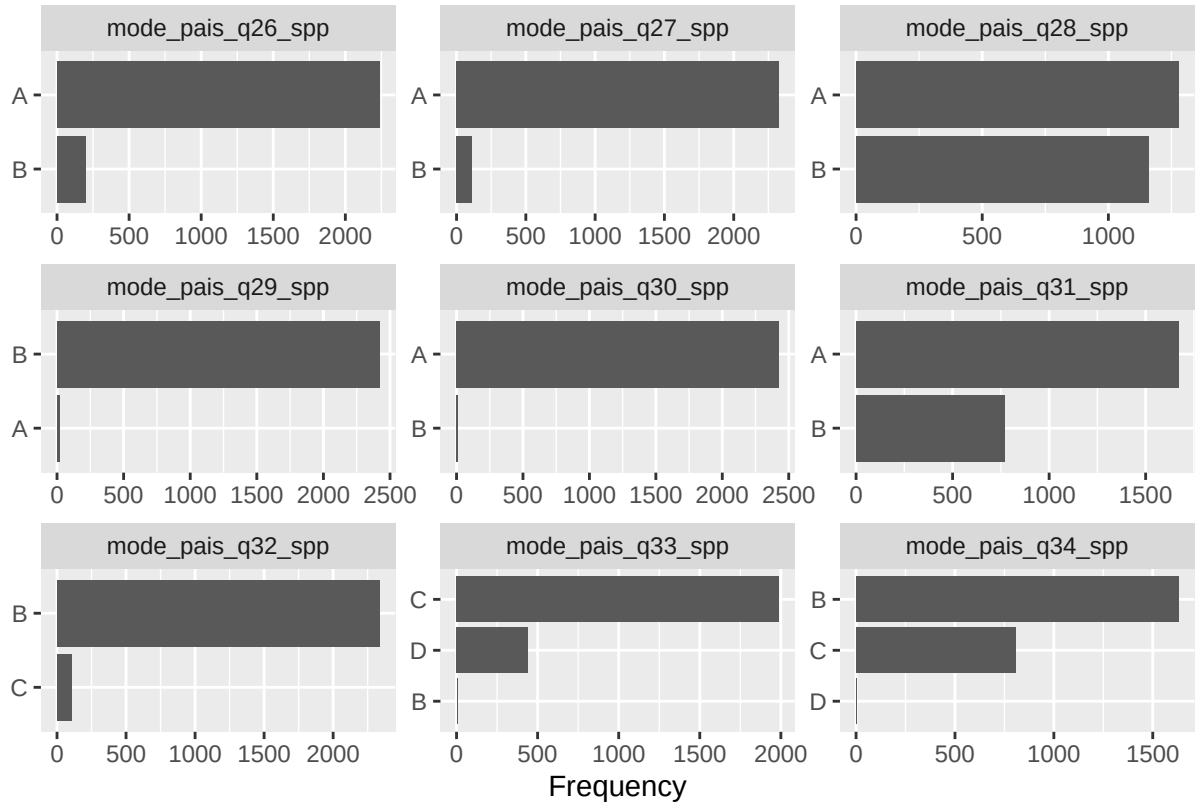


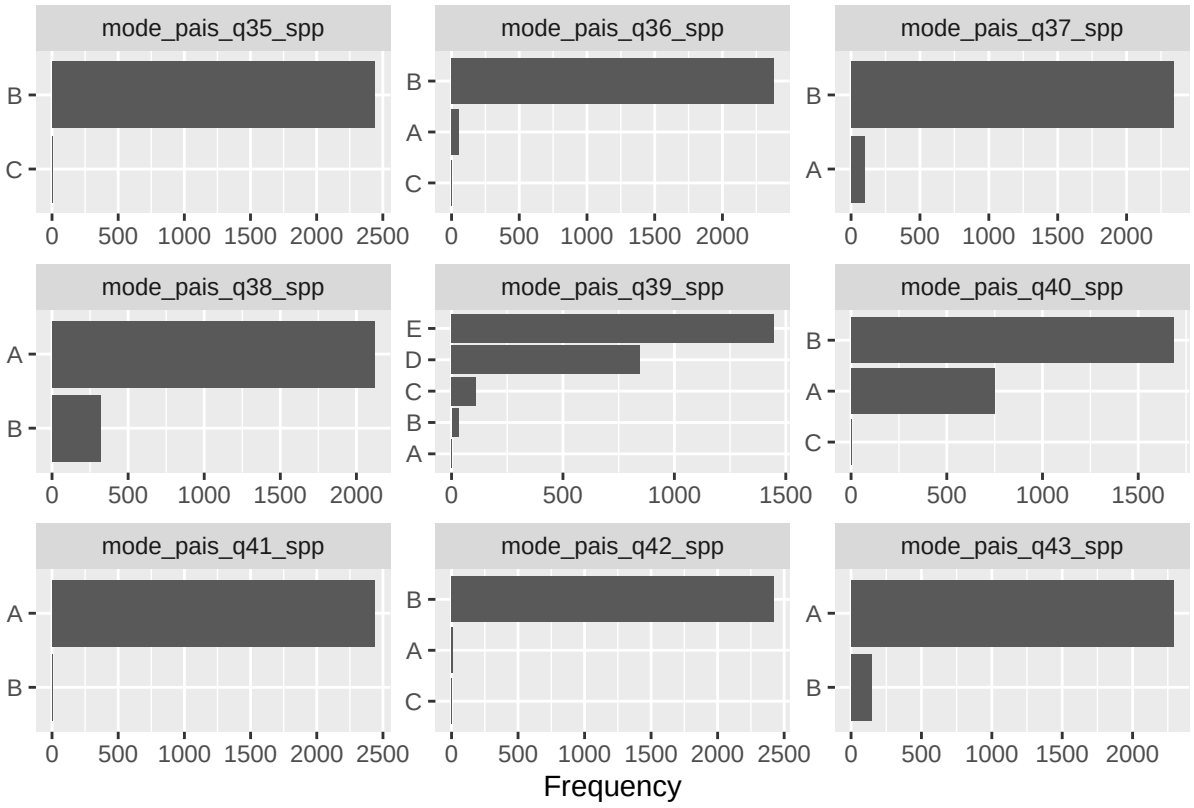
Volume: VAriáveis com bom volume

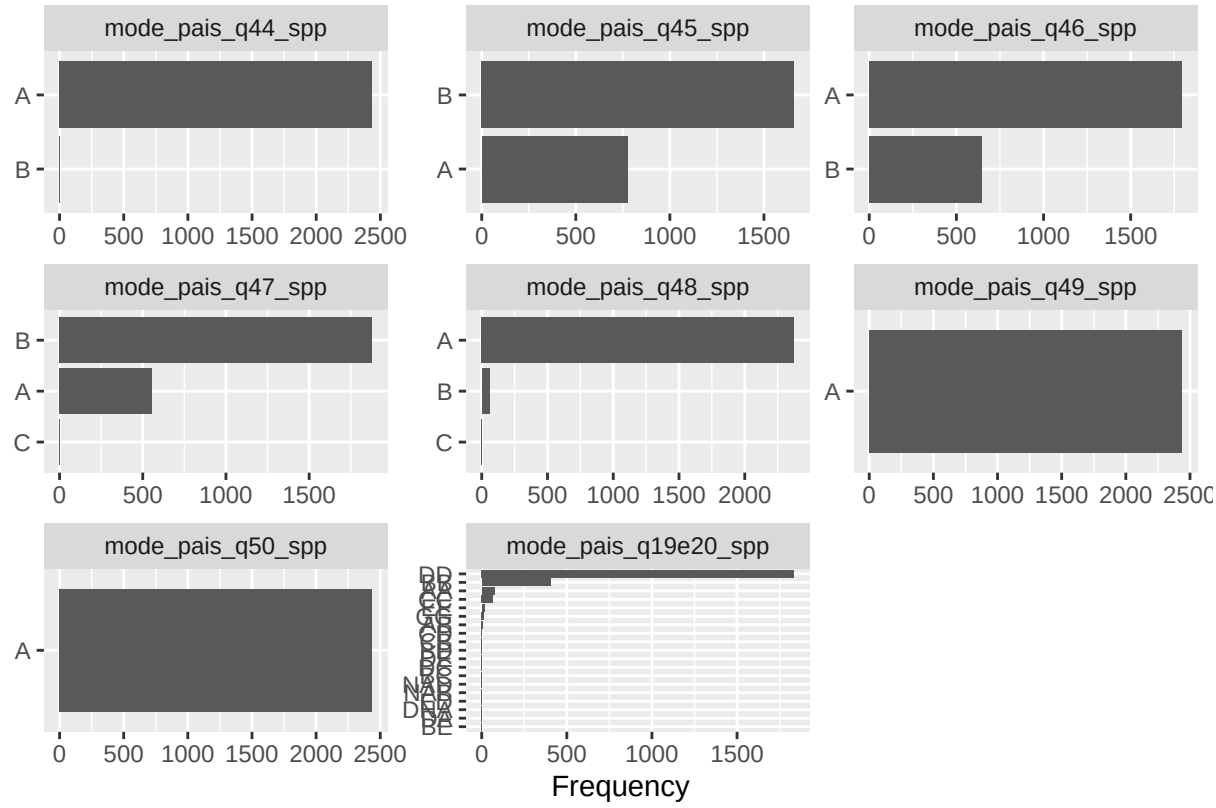
- mode_pais_q12
- mode_pais_q18
- mode_pais_q19
- mode_pais_q20
- mode_pais_q22
- mode_pais_q23
- mode_pais_q26
- mode_pais_q28
- mode_pais_q31
- mode_pais_q33
- mode_pais_q34
- mode_pais_q38
- mode_pais_q39
- mode_pais_q40
- mode_pais_q45
- mode_pais_q46
- mode_pais_q47

```
plot_bar(final_data)
```


mode_pais_q12_spp
B
A
0
500
1000
1500
mode_pais_q18_spp
C
B
A
0
500
1000
1500
2000
mode_pais_q19_spp
D
B
A
C
G
E
0
500
1000
1500
mode_pais_q20_spp
D
B
C
A
E
G
0
500
1000
1500
mode_pais_q21_spp
A
B
E
F
D
0
500
1000
1500
2000
mode_pais_q22_spp
A
E
B
F
D
0
500
1000
1500
2000
mode_pais_q23_spp
B
C
A
H
D
E
0
500
1000
1500
mode_pais_q24_spp
B
A
0
500
1000
1500
2000
2500
mode_pais_q25_spp
B
A
0
500
1000
1500
2000
2500
Frequency


mode_pais_q26_spp
A
B
0 500 1000 1500 2000
mode_pais_q27_spp
A
B
0 500 1000 1500 2000
mode_pais_q28_spp
A
B
0 500 1000
mode_pais_q29_spp
B
A
0 500 1000 1500 2000 2500
mode_pais_q30_spp
A
B
0 500 1000 1500 2000 2500
mode_pais_q31_spp
A
B
0 500 1000 1500
mode_pais_q32_spp
B
C
0 500 1000 1500 2000
mode_pais_q33_spp
C
D
B
0 500 1000 1500 2000
mode_pais_q34_spp
B
C
D
0 500 1000 1500
Frequency


mode_pais_q35_spp
B
C
0 500 1000 1500 2000 2500
mode_pais_q36_spp
B
A
C
0 500 1000 1500 2000
mode_pais_q37_spp
B
A
0 500 1000 1500 2000
mode_pais_q38_spp
A
B
0 500 1000 1500 2000
mode_pais_q39_spp
E
D
C
B
A
0 500 1000 1500
mode_pais_q40_spp
B
A
C
0 500 1000 1500
mode_pais_q41_spp
A
B
0 500 1000 1500 2000 2500
mode_pais_q42_spp
B
A
C
0 500 1000 1500 2000 2500
mode_pais_q43_spp
A
B
0 500 1000 1500 2000
Frequency

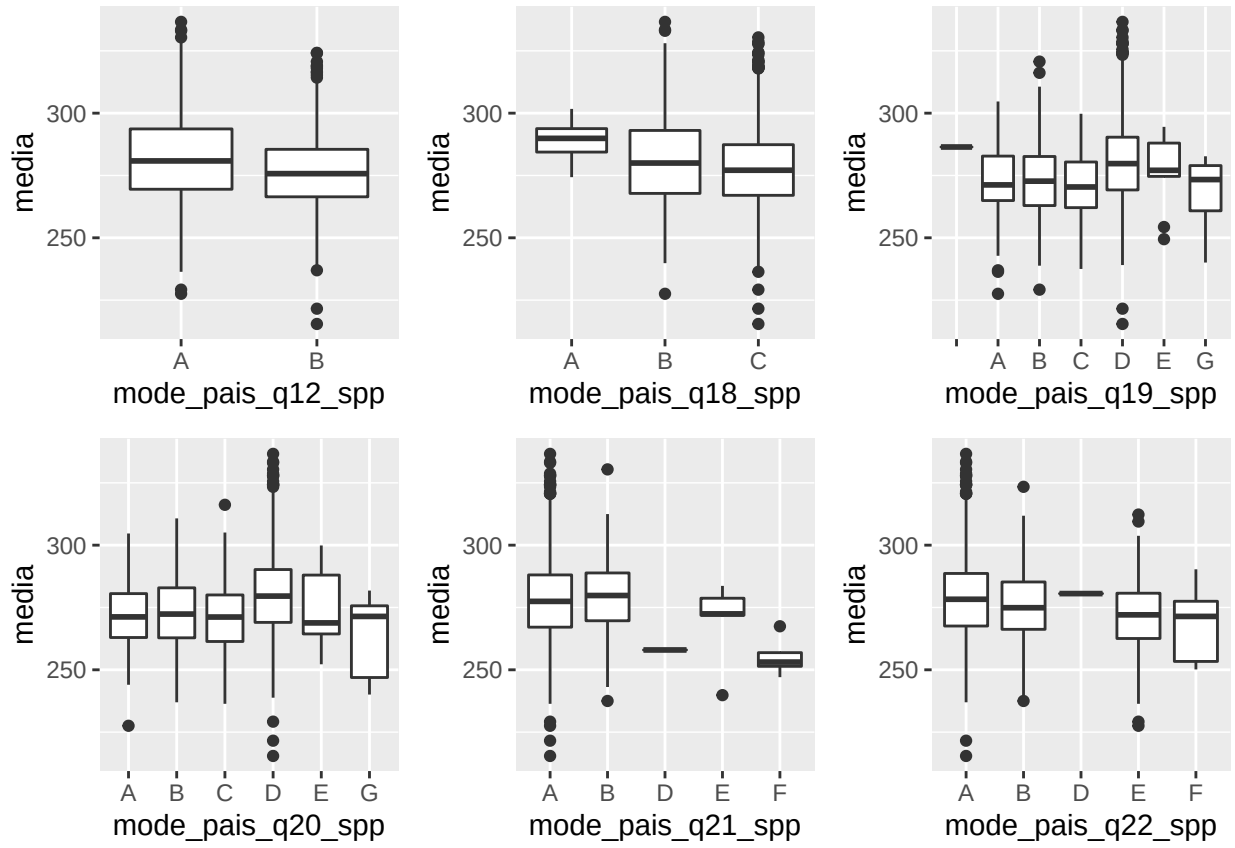


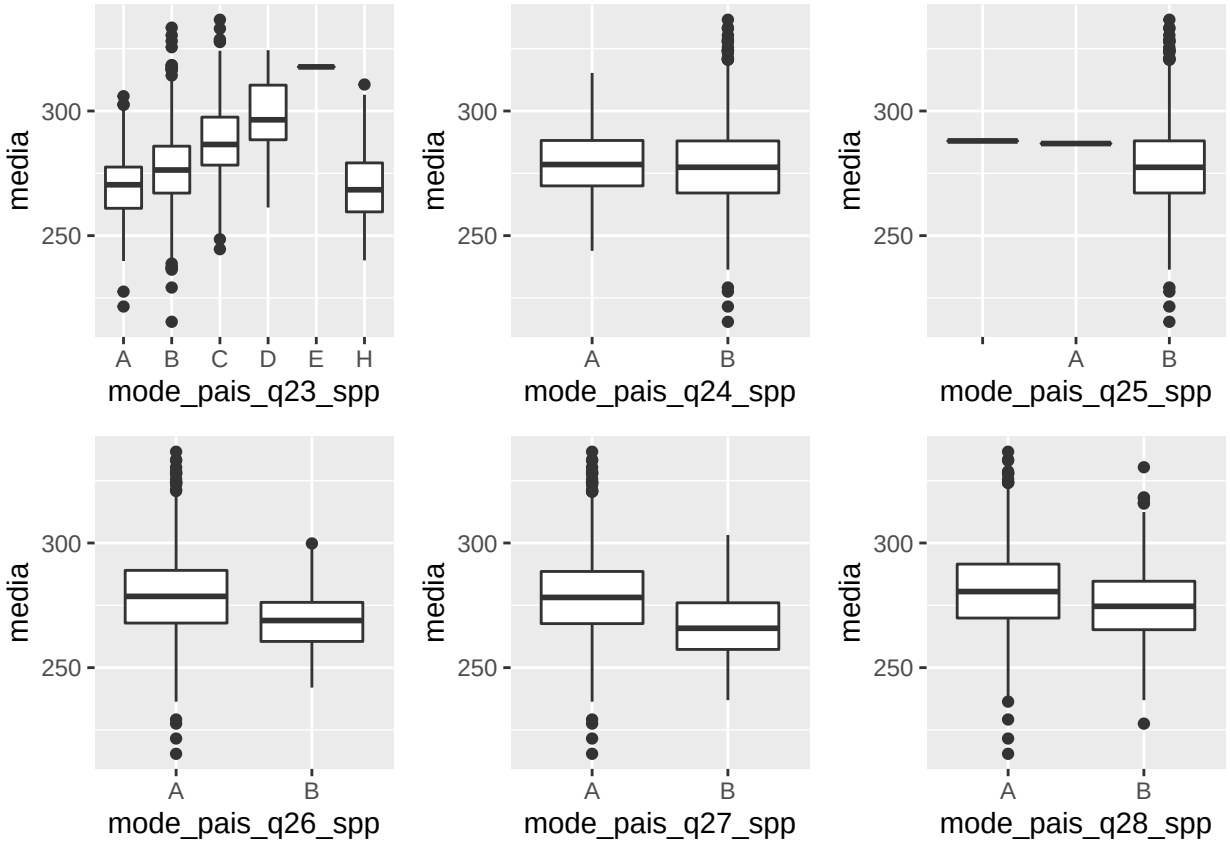
Boxplot: Variáveis com bom volume e variância

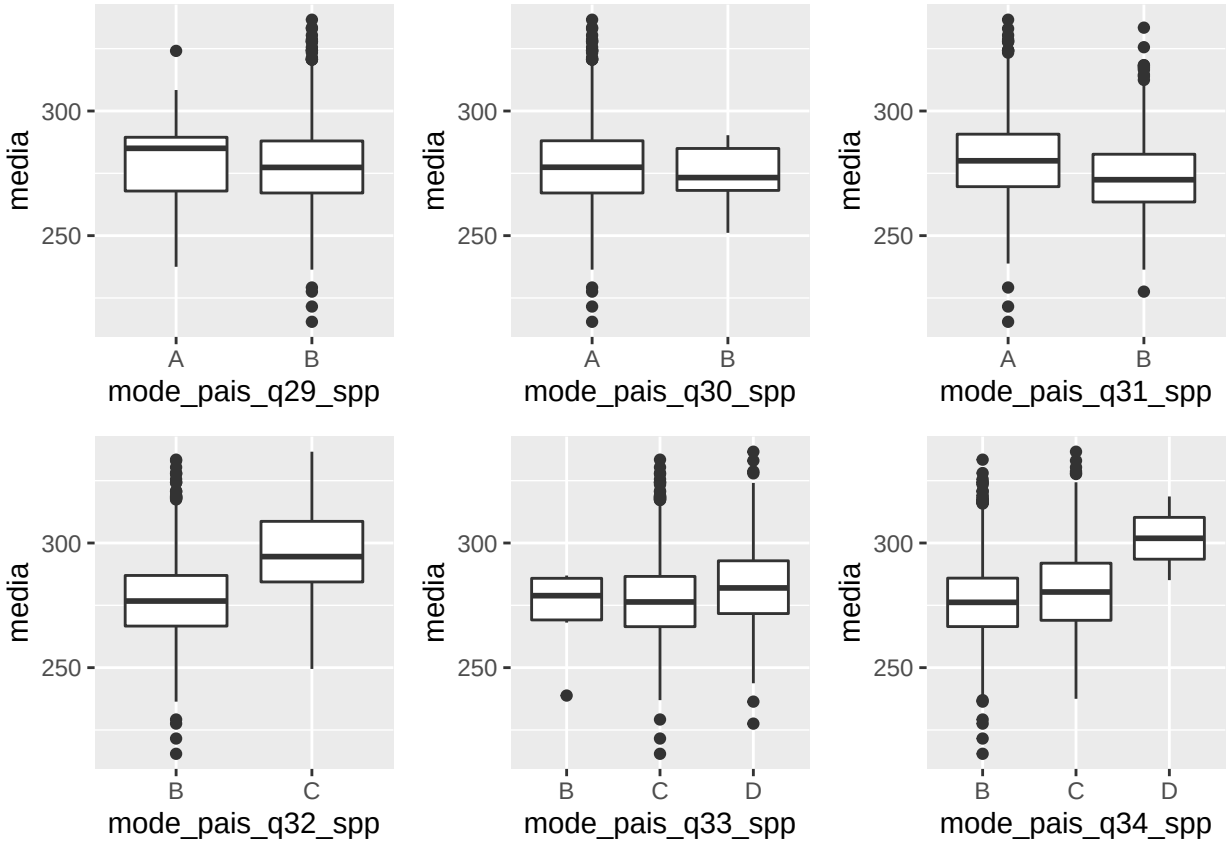
- mode_pais_q19: graduação mãe > ordena (juntar D e E)
- mode_pais_q20: graduação pai > ordena (juntar D e E)
- mode_pais_q23: renda maior=> notas > (juntar A com H)
- mode_pais_q31: TV assinatura => Notas >
- mode_pais_q40: COmputador => Notas > (A x B)
- mode_pais_q46: Aspirador => Notas > (A x B)
- mode_pais_q47: Carro => Notas > (A x B)

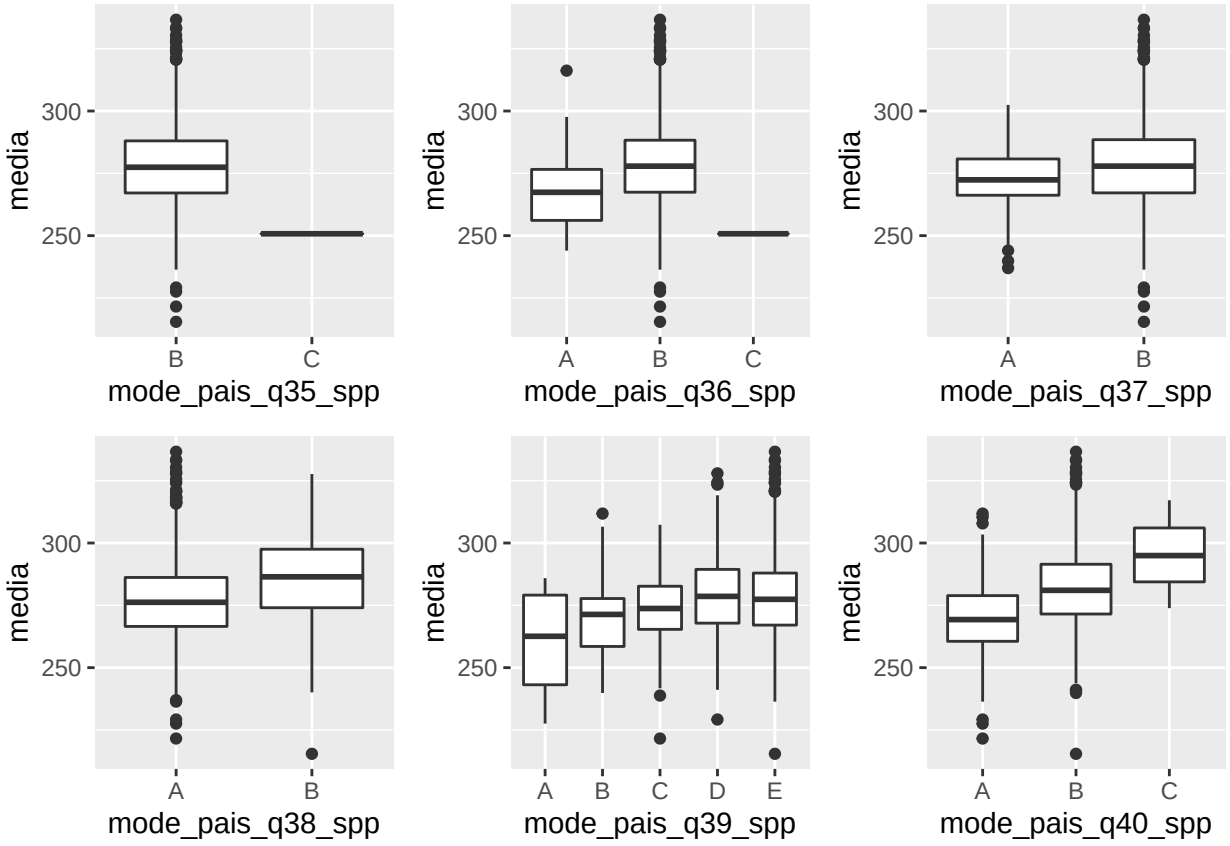
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

n <- length(plots)

i <- 1
while (i <= n) {
```

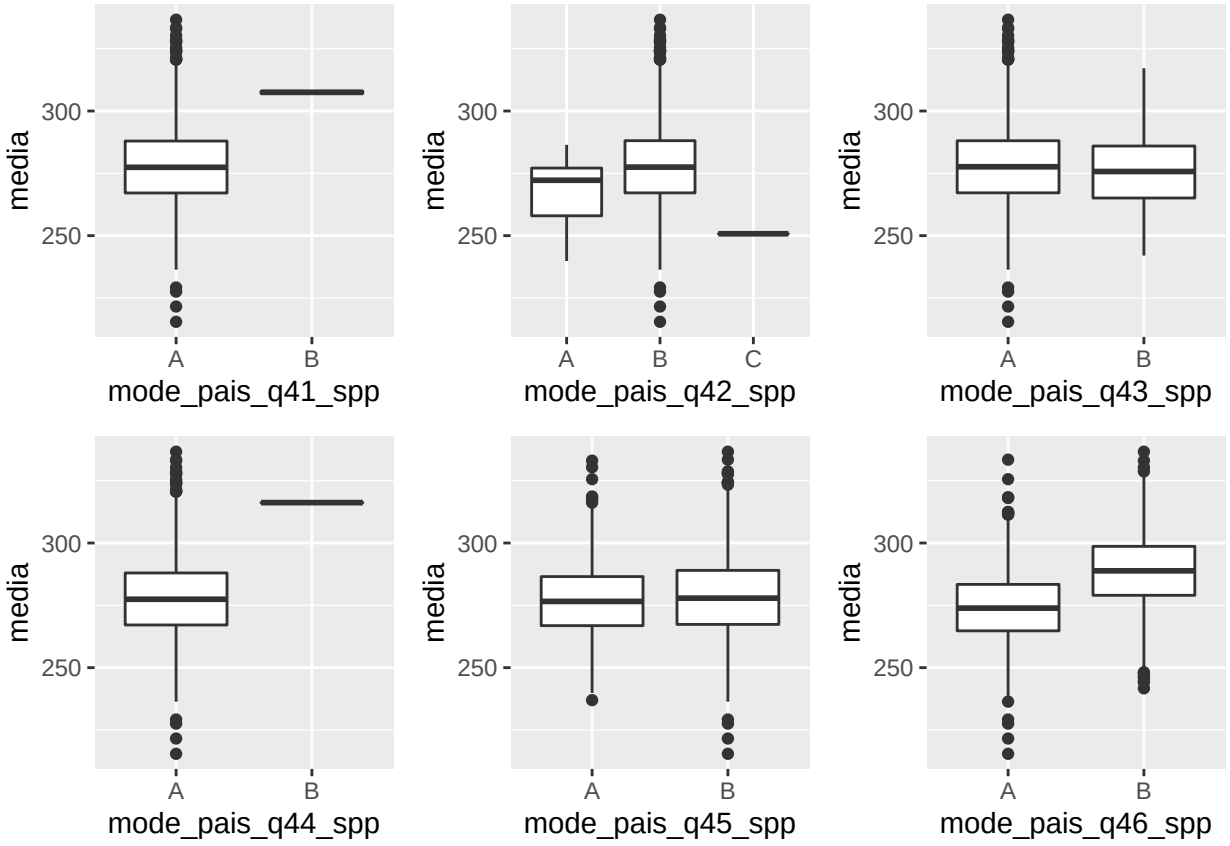
```
  do.call("grid.arrange", c(plots[i:(min(i+5, n))], ncol=3, nrow = 2))
  i <- i + 6
}
```

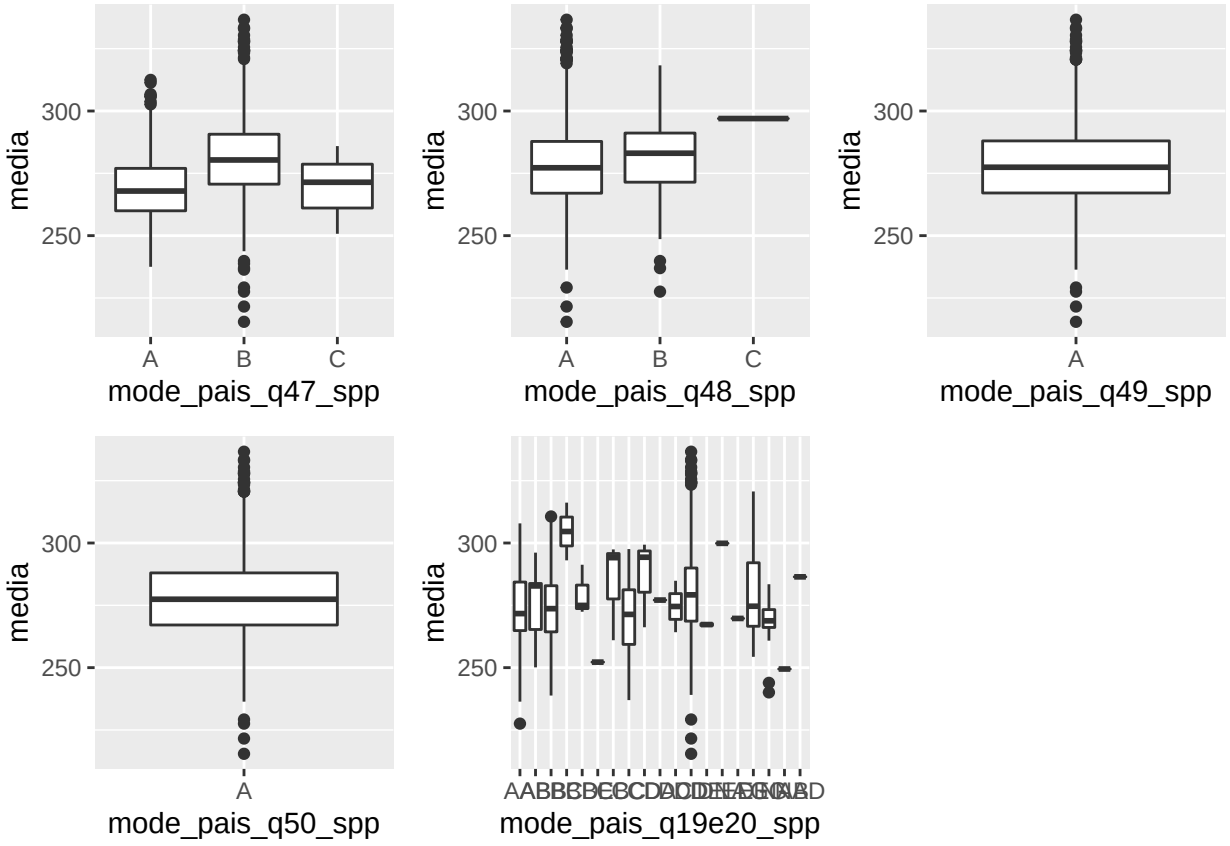



media
300
250
A B C D E H
mode_pais_q23_spp
A B
mode_pais_q24_spp
A B
mode_pais_q25_spp
A B
mode_pais_q26_spp
A B
mode_pais_q27_spp
A B
mode_pais_q28_spp


media
300
250
A
B
mode_pais_q29_spp
media
300
250
A
B
mode_pais_q30_spp
media
300
250
A
B
mode_pais_q31_spp
media
300
250
B
C
mode_pais_q32_spp
media
300
250
B
C
D
mode_pais_q33_spp
media
300
250
B
C
D
mode_pais_q34_spp


media
300
250
B
C
mode_pais_q35_spp
media
300
250
A
B
C
mode_pais_q36_spp
media
300
250
A
B
mode_pais_q37_spp
media
300
250
A
B
mode_pais_q38_spp
media
300
250
A
B
C
D
E
mode_pais_q39_spp
media
300
250
A
B
C
mode_pais_q40_spp


media
300
250
A
B
mode_pais_q41_spp
media
300
250
A
B
C
mode_pais_q42_spp
media
300
250
A
B
mode_pais_q43_spp
media
300
250
A
B
mode_pais_q44_spp
media
300
250
A
B
mode_pais_q45_spp
media
300
250
A
B
mode_pais_q46_spp



Análise Univariada

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

```
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
```

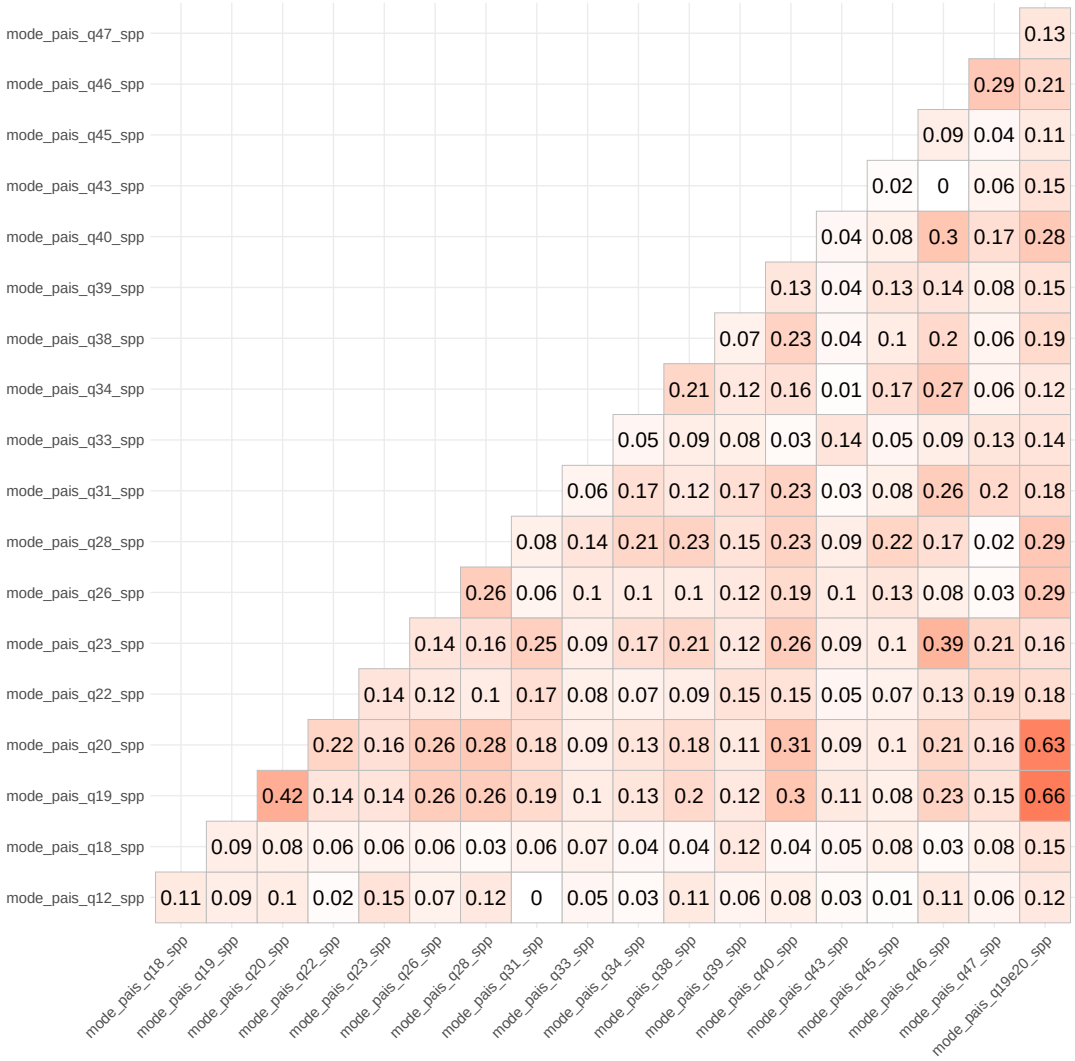
```
##  [1] "mode_pais_q12_spp"    "mode_pais_q18_spp"    "mode_pais_q19_spp"
##  [4] "mode_pais_q20_spp"    "mode_pais_q22_spp"    "mode_pais_q23_spp"
##  [7] "mode_pais_q26_spp"    "mode_pais_q28_spp"    "mode_pais_q31_spp"
## [10] "mode_pais_q33_spp"    "mode_pais_q34_spp"    "mode_pais_q38_spp"
## [13] "mode_pais_q39_spp"    "mode_pais_q40_spp"    "mode_pais_q43_spp"
## [16] "mode_pais_q45_spp"    "mode_pais_q46_spp"    "mode_pais_q47_spp"
## [19] "mode_pais_q19e20_spp"
```

```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_q12_spp 0.0305360025
## 2     mode_pais_q18_spp 0.0063650511
## 3     mode_pais_q19_spp 0.0511221223
## 4     mode_pais_q20_spp 0.0489465496
## 5     mode_pais_q22_spp 0.0139599360
## 6     mode_pais_q23_spp 0.1747604750
## 7     mode_pais_q26_spp 0.0300737222
## 8     mode_pais_q28_spp 0.0366676140
## 9     mode_pais_q31_spp 0.0417335441
## 10    mode_pais_q33_spp 0.0181893585
## 11    mode_pais_q34_spp 0.0190818420
## 12    mode_pais_q38_spp 0.0354549924
## 13    mode_pais_q39_spp 0.0102577151
## 14    mode_pais_q40_spp 0.1282199753
## 15    mode_pais_q43_spp 0.0008581503
## 16    mode_pais_q45_spp 0.0023292533
## 17    mode_pais_q46_spp 0.1682761862
## 18    mode_pais_q47_spp 0.0956628970
## 19 mode_pais_q19e20_spp 0.0439934525
```

Matriz de correlação (60%)

```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_spp
mode_pais_q46_spp
mode_pais_q45_spp
mode_pais_q43_spp
mode_pais_q40_spp
mode_pais_q39_spp
mode_pais_q38_spp
mode_pais_q34_spp
mode_pais_q33_spp
mode_pais_q31_spp
mode_pais_q28_spp
mode_pais_q26_spp
mode_pais_q23_spp
mode_pais_q22_spp
mode_pais_q20_spp
mode_pais_q19_spp
mode_pais_q18_spp
mode_pais_q12_spp
0.13
0.29 0.21
0.09 0.04 0.11
0.02 0 0.06 0.15
0.04 0.08 0.3 0.17 0.28
0.13 0.04 0.13 0.14 0.08 0.15
0.07 0.23 0.04 0.1 0.2 0.06 0.19
0.21 0.12 0.16 0.01 0.17 0.27 0.06 0.12
0.05 0.09 0.08 0.03 0.14 0.05 0.09 0.13 0.14
0.06 0.17 0.12 0.17 0.23 0.03 0.08 0.26 0.2 0.18
0.08 0.14 0.21 0.23 0.15 0.23 0.09 0.22 0.17 0.02 0.29
0.26 0.06 0.1 0.1 0.1 0.12 0.19 0.1 0.13 0.08 0.03 0.29
0.14 0.16 0.25 0.09 0.17 0.21 0.12 0.26 0.09 0.1 0.39 0.21 0.16
0.14 0.12 0.1 0.17 0.08 0.07 0.09 0.15 0.15 0.05 0.07 0.13 0.19 0.18
0.22 0.16 0.26 0.28 0.18 0.09 0.13 0.18 0.11 0.31 0.09 0.1 0.21 0.16 0.63
0.42 0.14 0.14 0.26 0.26 0.19 0.1 0.13 0.2 0.12 0.3 0.11 0.08 0.23 0.15 0.66
0.09 0.08 0.06 0.06 0.06 0.03 0.06 0.07 0.04 0.04 0.12 0.04 0.05 0.08 0.03 0.08 0.15
0.11 0.09 0.1 0.02 0.15 0.07 0.12 0 0.05 0.03 0.11 0.06 0.08 0.03 0.01 0.11 0.06 0.12
mode_pais_q18_spp
mode_pais_q19_spp
mode_pais_q20_spp
mode_pais_q22_spp
mode_pais_q23_spp
mode_pais_q26_spp
mode_pais_q28_spp
mode_pais_q31_spp
mode_pais_q33_spp
mode_pais_q34_spp
mode_pais_q38_spp
mode_pais_q39_spp
mode_pais_q40_spp
mode_pais_q43_spp
mode_pais_q45_spp
mode_pais_q46_spp
mode_pais_q47_spp
mode_pais_q19e20_spp