Análise das variáveis Saeb - moda por escola

Série 3EM

Livia Kobayashi

14 junho 2021

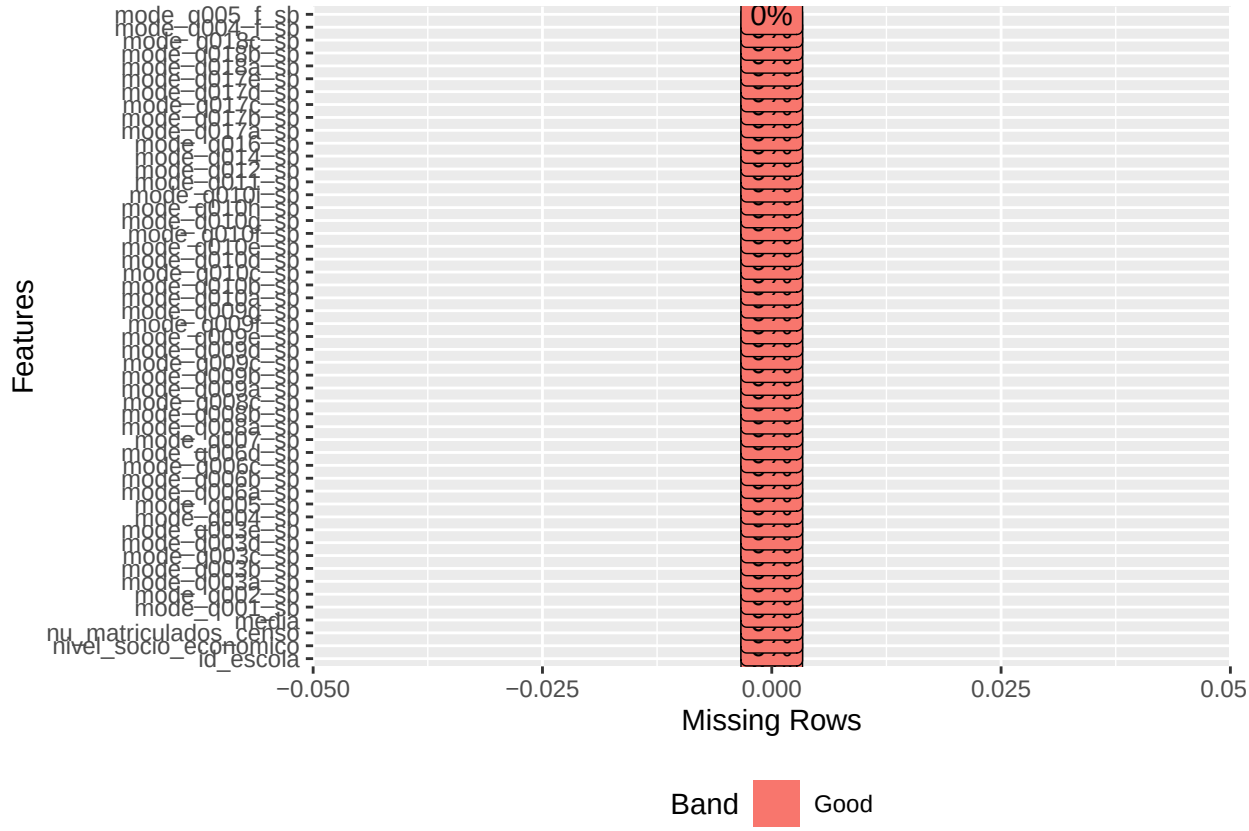
```
df_publico <- read.csv2("../output/books/df_publico.csv")

book <- read.csv2(params$book)
```

```
##   id_serie
## 1      3EM
```

Missing

Base 100% preenchida

```
plot_missing(df)
```

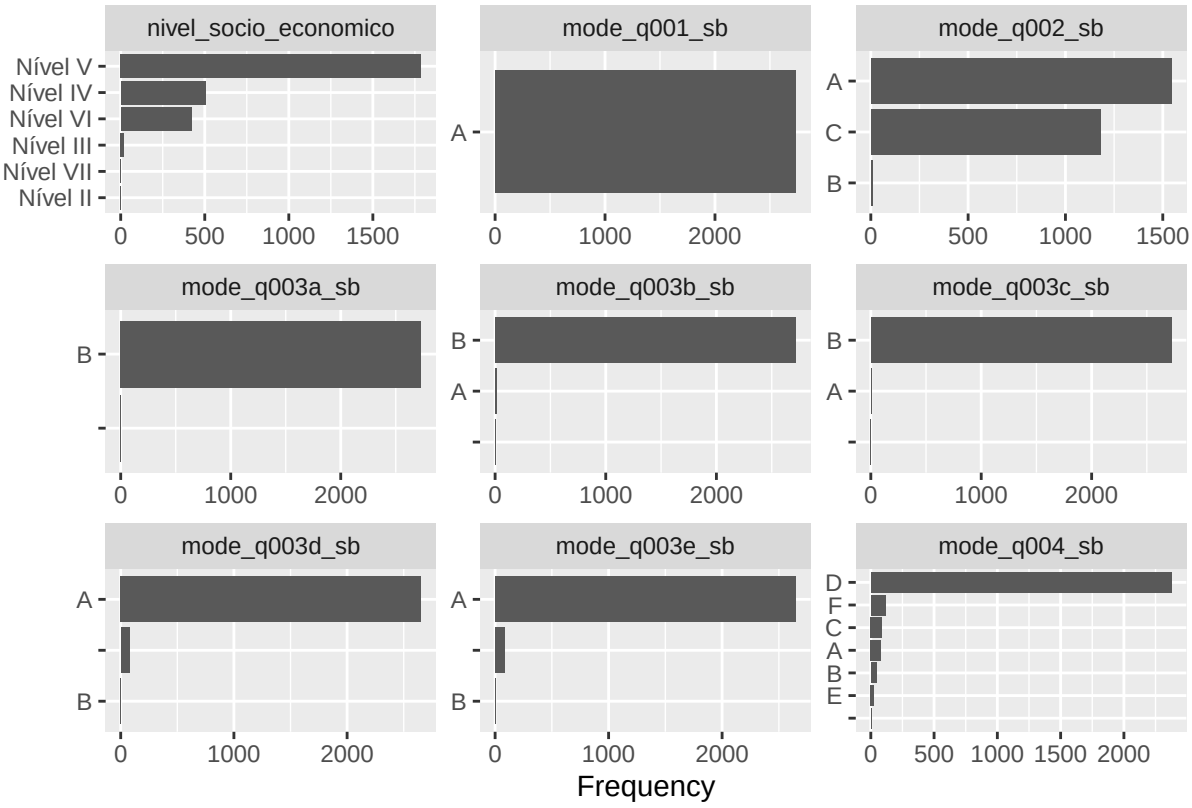


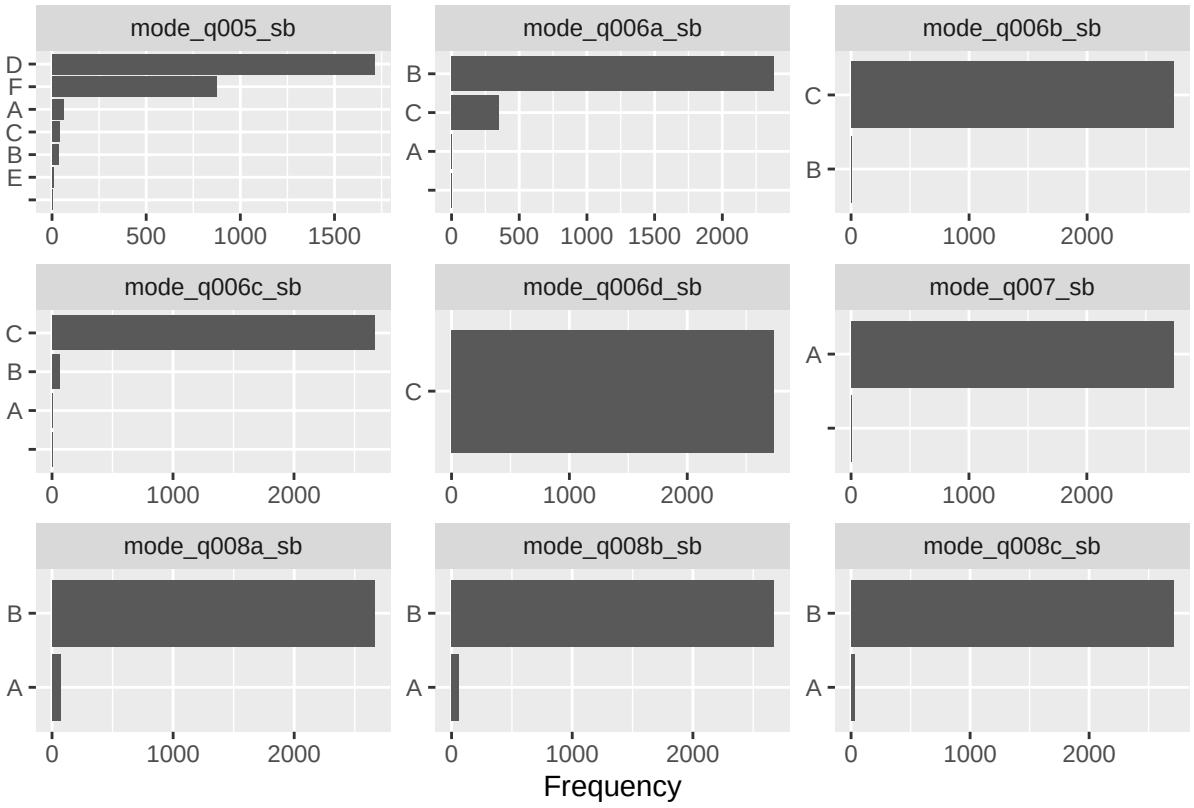
Volume

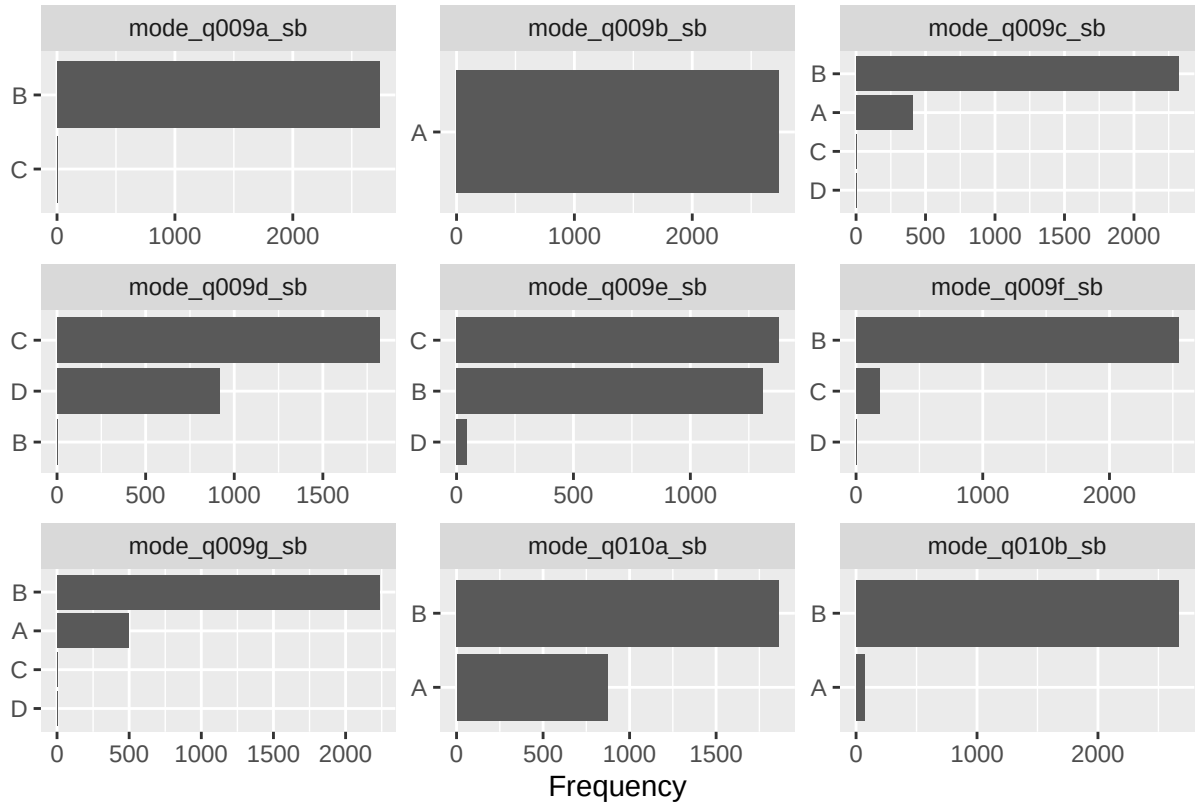
Variáveis com baixa variância:

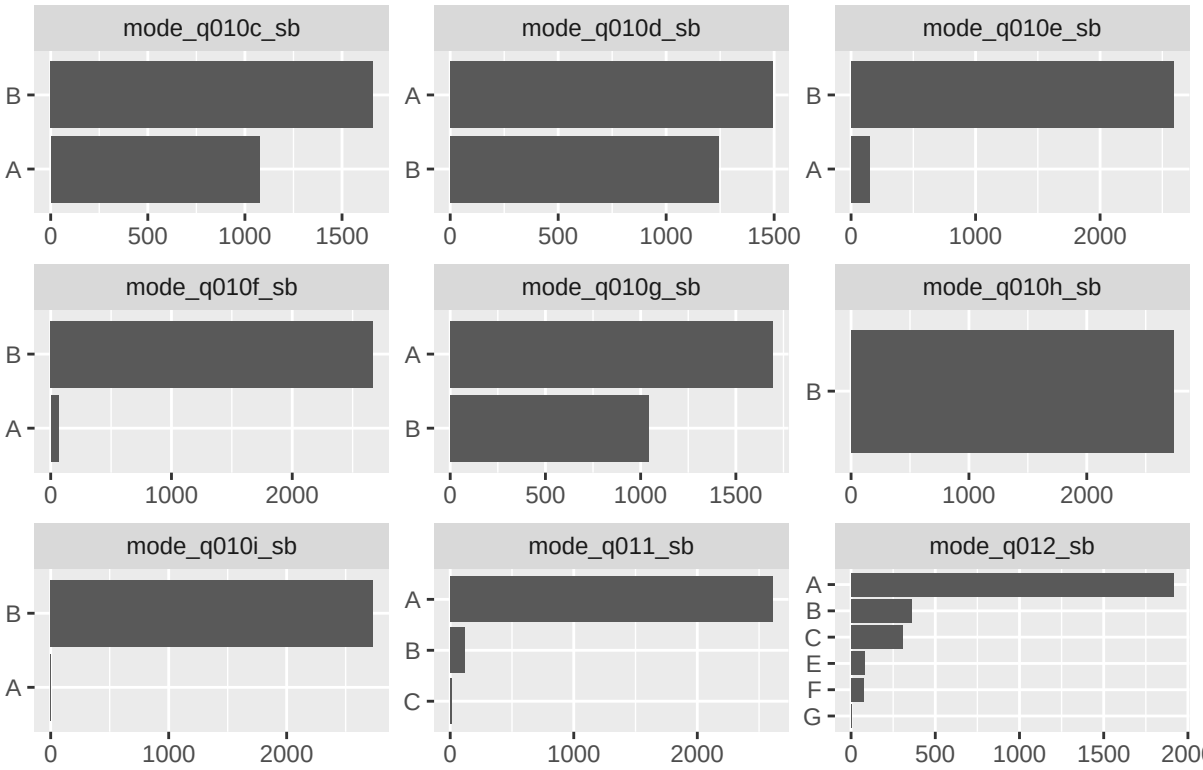
- mode_q001_sb - mode_q003a_sb - mode_q003b_sb - mode_q003c_sb
- mode_q003d_sb - mode_q003e_sb - mode_q004_sb - mode_q006b_sb
- mode_q006c_sb - mode_q006d_sb - mode_q007_sb - mode_q008a_sb
- mode_q008b_sb - mode_q008c_sb - mode_q009a_sb - mode_q009b_sb
- mode_q009f_sb - mode_q010b_sb - mode_q010e_sb - mode_q010f_sb
- mode_q010h_sb - mode_q010i_sb - mode_q011_sb - mode_q014_sb
- mode_q016_sb - mode_q017a_sb - mode_q017b_sb - mode_q018a_sb
- mode_q018c_sb - mode_q004_f_sb - mode_q005_f_sb

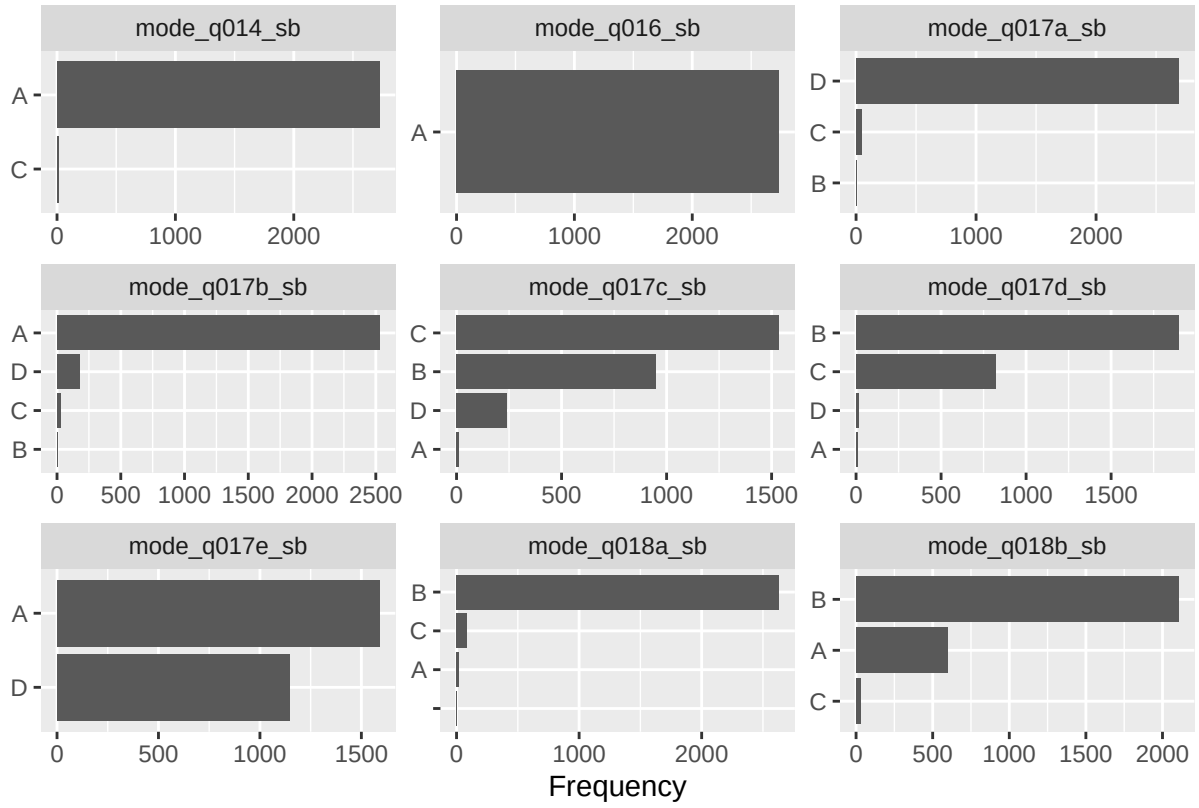
```
plot_bar(final_data)
```


nivel_socio_economico
Nível V
Nível IV
Nível VI
Nível III
Nível VII
Nível II
0
500
1000
1500
mode_q001_sb
A
0
1000
2000
mode_q002_sb
A
C
B
0
500
1000
1500
mode_q003a_sb
B
-
0
1000
2000
mode_q003b_sb
B
A
-
0
1000
2000
mode_q003c_sb
B
A
-
0
1000
2000
mode_q003d_sb
A
-
B
0
1000
2000
mode_q003e_sb
A
-
B
0
1000
2000
mode_q004_sb
D
F
C
A
B
E
-
0
500
1000
1500
2000
Frequency

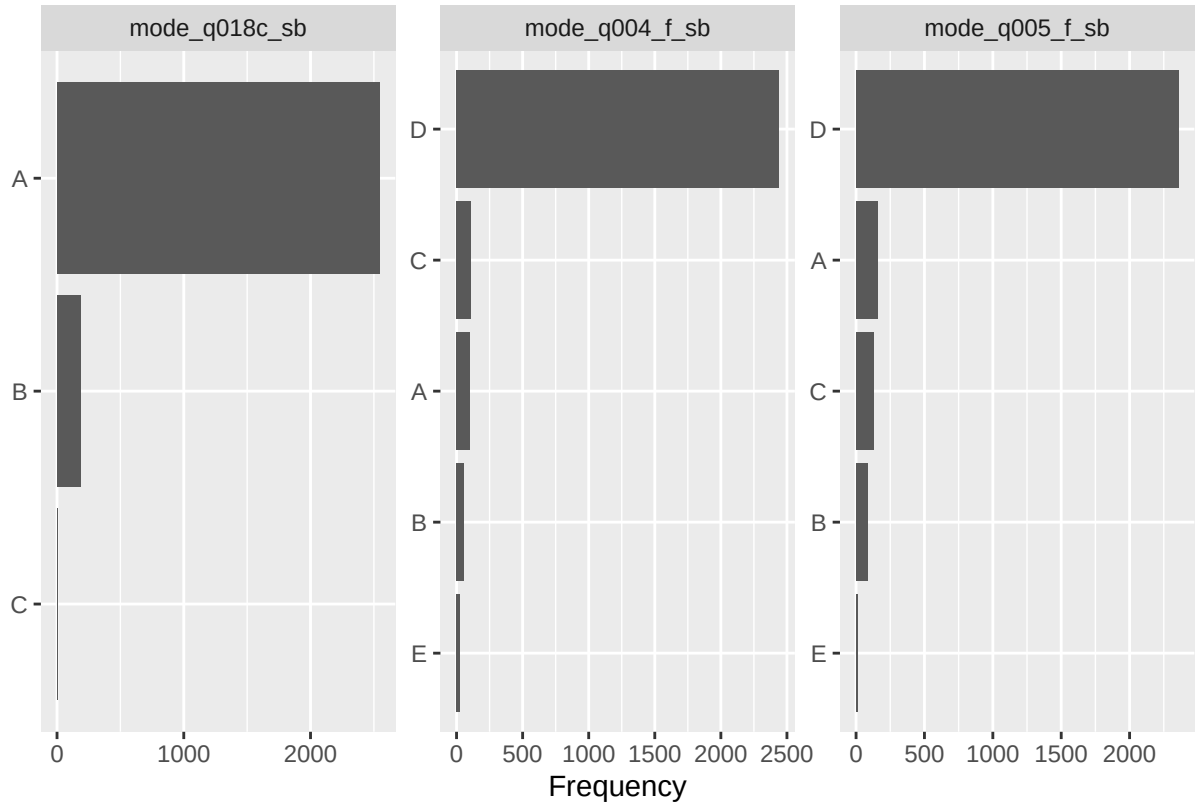

mode_q005_sb
D
F
A
C
B
E
-
0
500
1000
1500
mode_q006a_sb
B
C
A
-
0
500
1000
1500
2000
mode_q006b_sb
C
B
0
1000
2000
mode_q006c_sb
C
B
A
-
0
1000
2000
mode_q006d_sb
C
0
1000
2000
mode_q007_sb
A
-
0
1000
2000
mode_q008a_sb
B
A
0
1000
2000
mode_q008b_sb
B
A
0
1000
2000
mode_q008c_sb
B
A
0
1000
2000
Frequency


mode_q009a_sb
B
C
0
1000
2000
mode_q009b_sb
A
0
1000
2000
mode_q009c_sb
B
A
C
D
0
500
1000
1500
2000
mode_q009d_sb
C
D
B
0
500
1000
1500
mode_q009e_sb
C
B
D
0
500
1000
mode_q009f_sb
B
C
D
0
1000
2000
mode_q009g_sb
B
A
C
D
0
500
1000
1500
2000
mode_q010a_sb
B
A
0
500
1000
1500
mode_q010b_sb
B
A
0
1000
2000
Frequency


mode_q010c_sb
B
A
0
500
1000
1500
mode_q010d_sb
A
B
0
500
1000
1500
mode_q010e_sb
B
A
0
1000
2000
mode_q010f_sb
B
A
0
1000
2000
mode_q010g_sb
A
B
0
500
1000
1500
mode_q010h_sb
B
0
1000
2000
mode_q010i_sb
B
A
0
1000
2000
mode_q011_sb
A
B
C
0
1000
2000
mode_q012_sb
A
B
C
E
F
G
0
500
1000
1500
Frequency


mode_q014_sb
A
C
0
1000
2000
mode_q016_sb
A
0
1000
2000
mode_q017a_sb
D
C
B
0
1000
2000
mode_q017b_sb
A
D
C
B
0
500
1000
1500
2000
2500
mode_q017c_sb
C
B
D
A
0
500
1000
1500
mode_q017d_sb
B
C
D
A
0
500
1000
1500
mode_q017e_sb
A
D
0
500
1000
1500
mode_q018a_sb
B
C
A
0
1000
2000
mode_q018b_sb
B
A
C
0
500
1000
1500
2000
Frequency

Page 5



Médias das notas x variáveis

Variáveis que discriminam e tem volume:

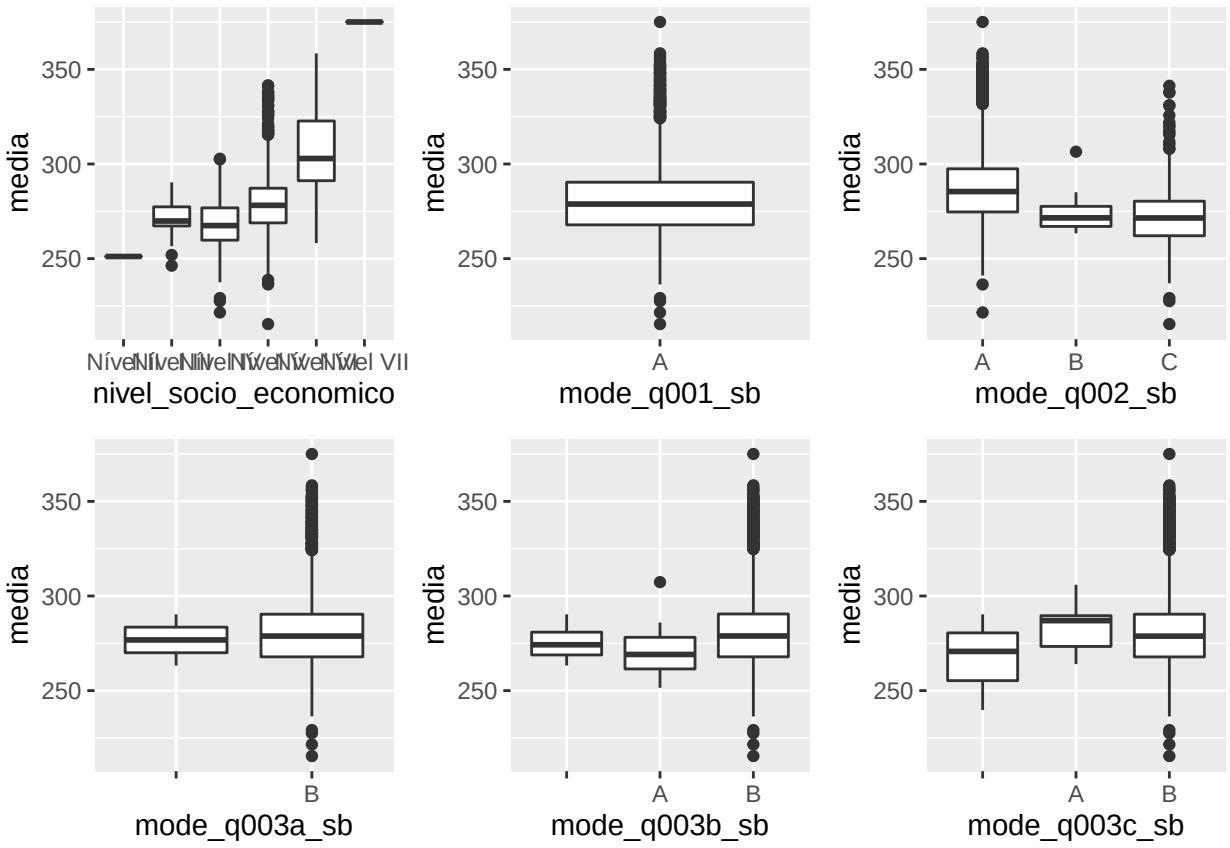
- mode_q002_sb: Maioria branca => notas maiores
- mode_q005_sb: pai com ensino superior => notas maiores e os que não sabem a escolarida do pai => notas menores
- mode_q006a_sb: Os pais conversarem sobre a escola => notas >
- mode_q009c_sb: Ter computador => notas >
- mode_q009g_sb: Ter Carro => Notas >
- mode_q010c_sb: Ter um quarto só para si => notas >
- mode_q010d_sb: Ter escrivaninha => Notas >
- mode_q010g_sb: Ter aspirador => Notas >
- mode_q012_sb: Vai de carro/outros para escola => nots maiores
  - a.À pé.
  - b.De ônibus urbano.
  - c.De transporte escolar.
  - d.De barco.
  - e.De bicicleta.

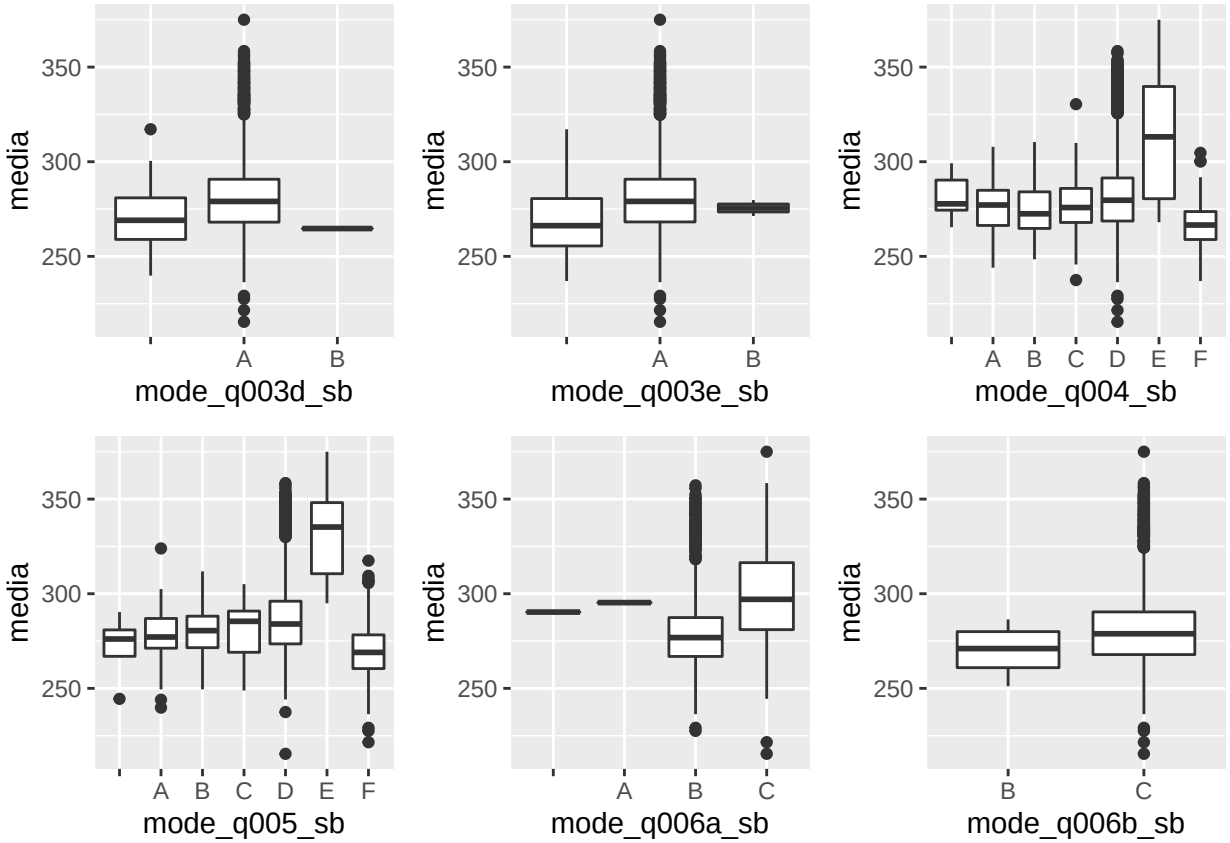
  - f.De carro

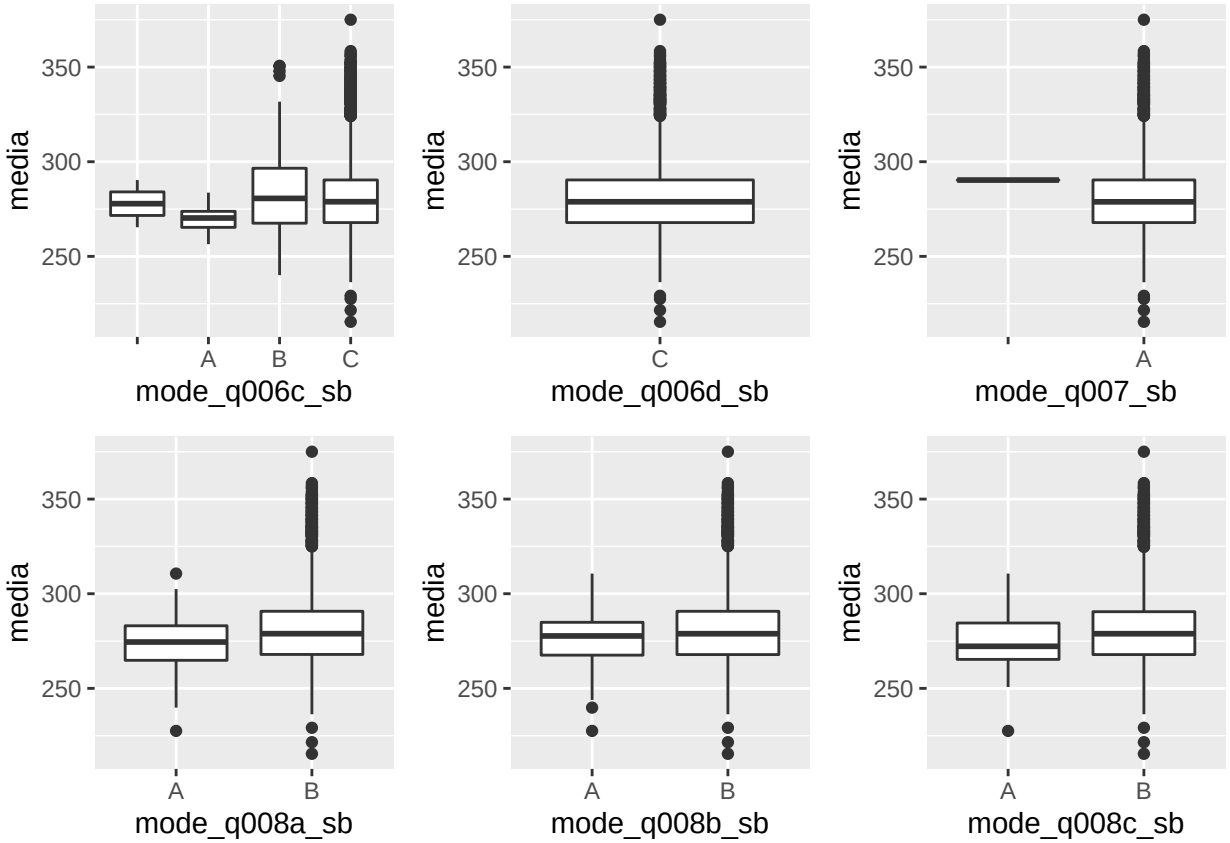
  - g.Outros meios de transporte.

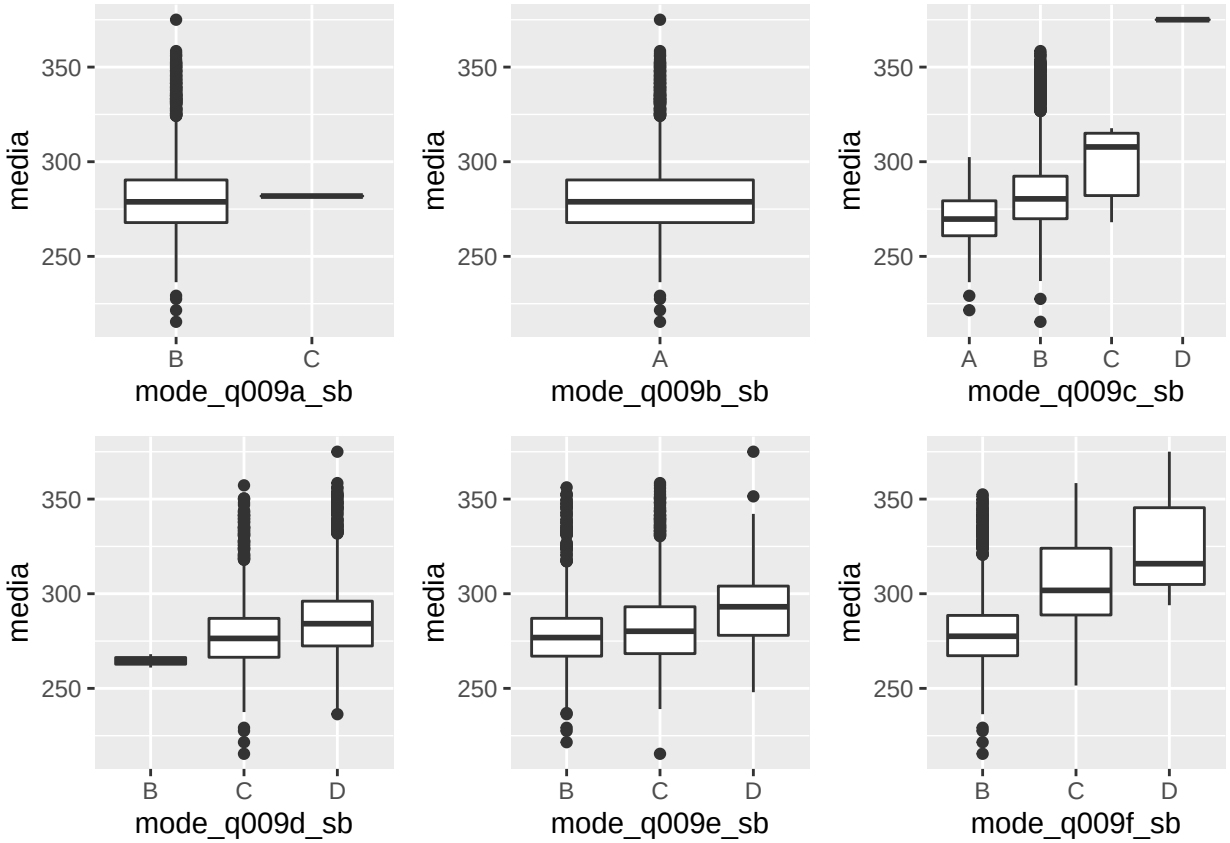
- mode_q017c_sb: menos tempo gasto em tarefas domésticas => notas >
- mode_q017d_sb: Não usar!!!
- mode_q018b_sb: Leitura extraescolar => notas >

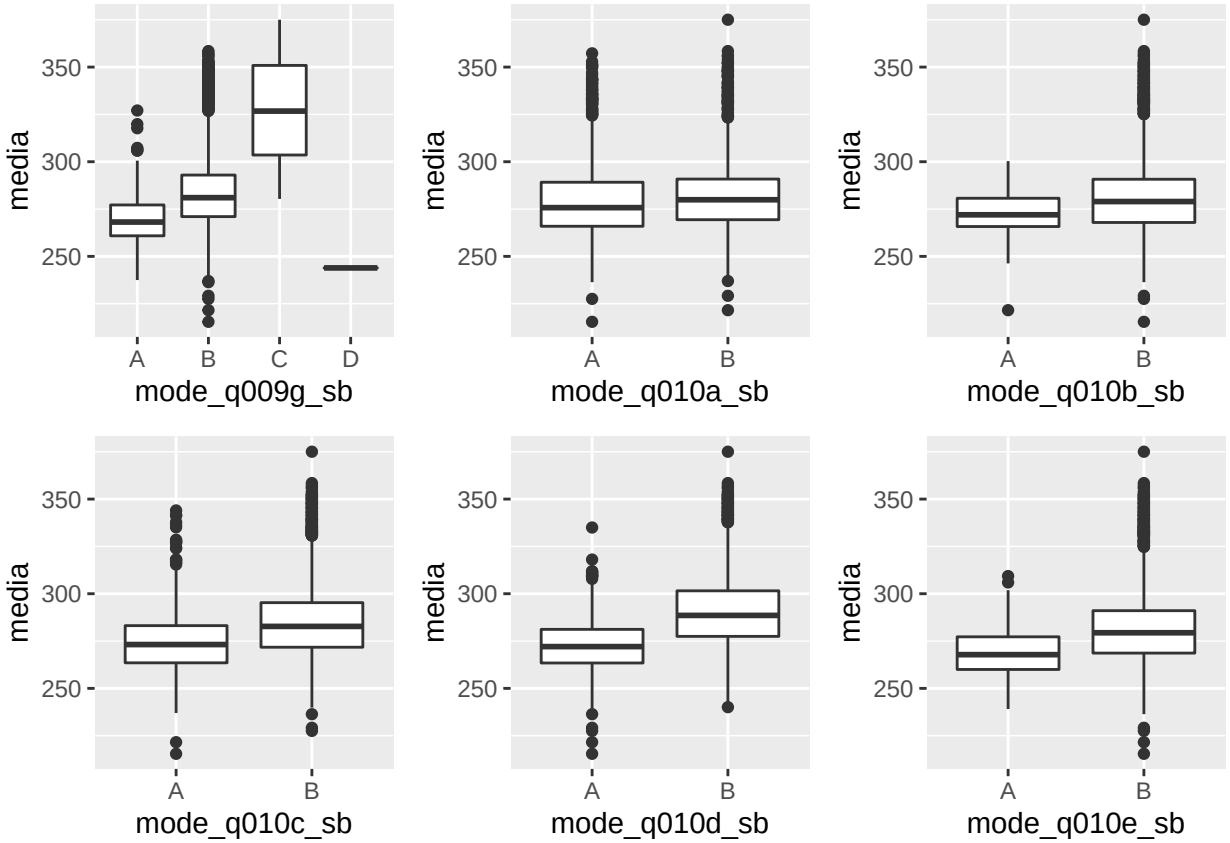
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

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
}
```

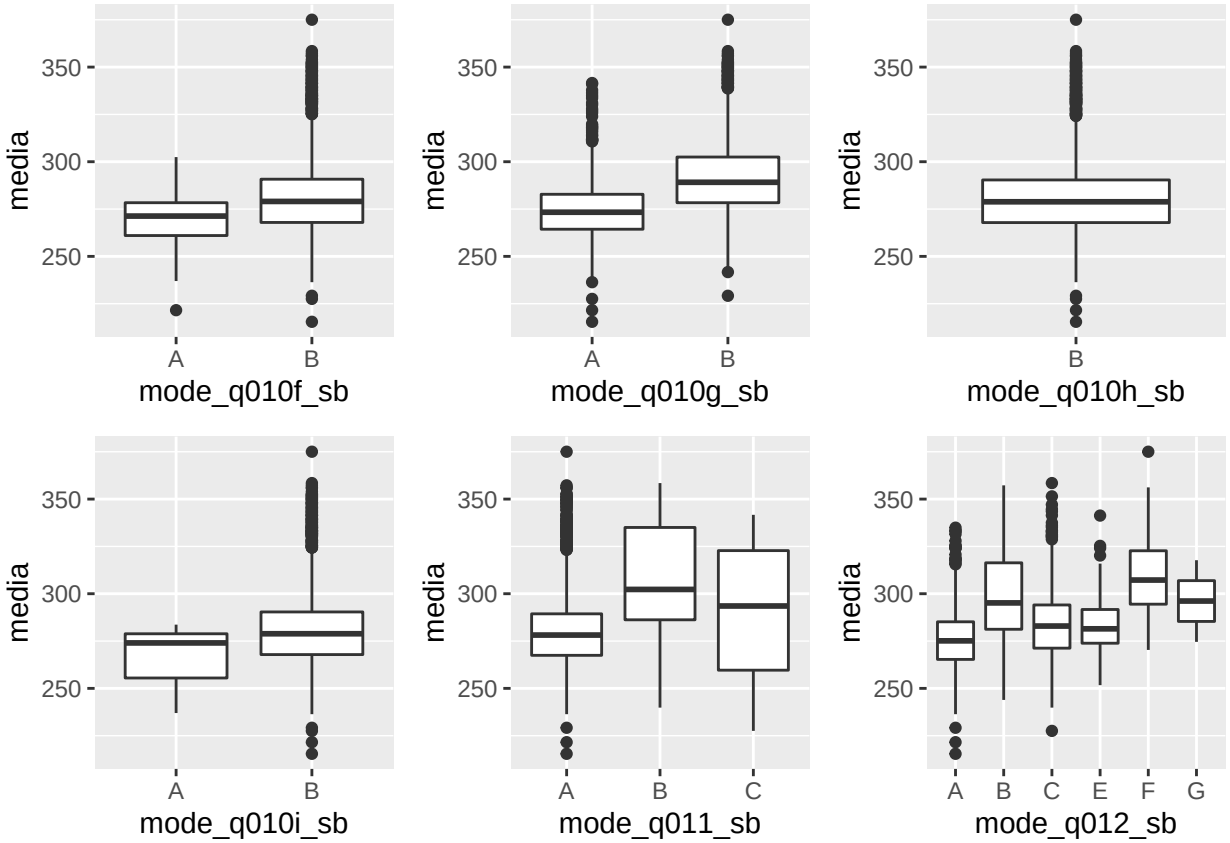


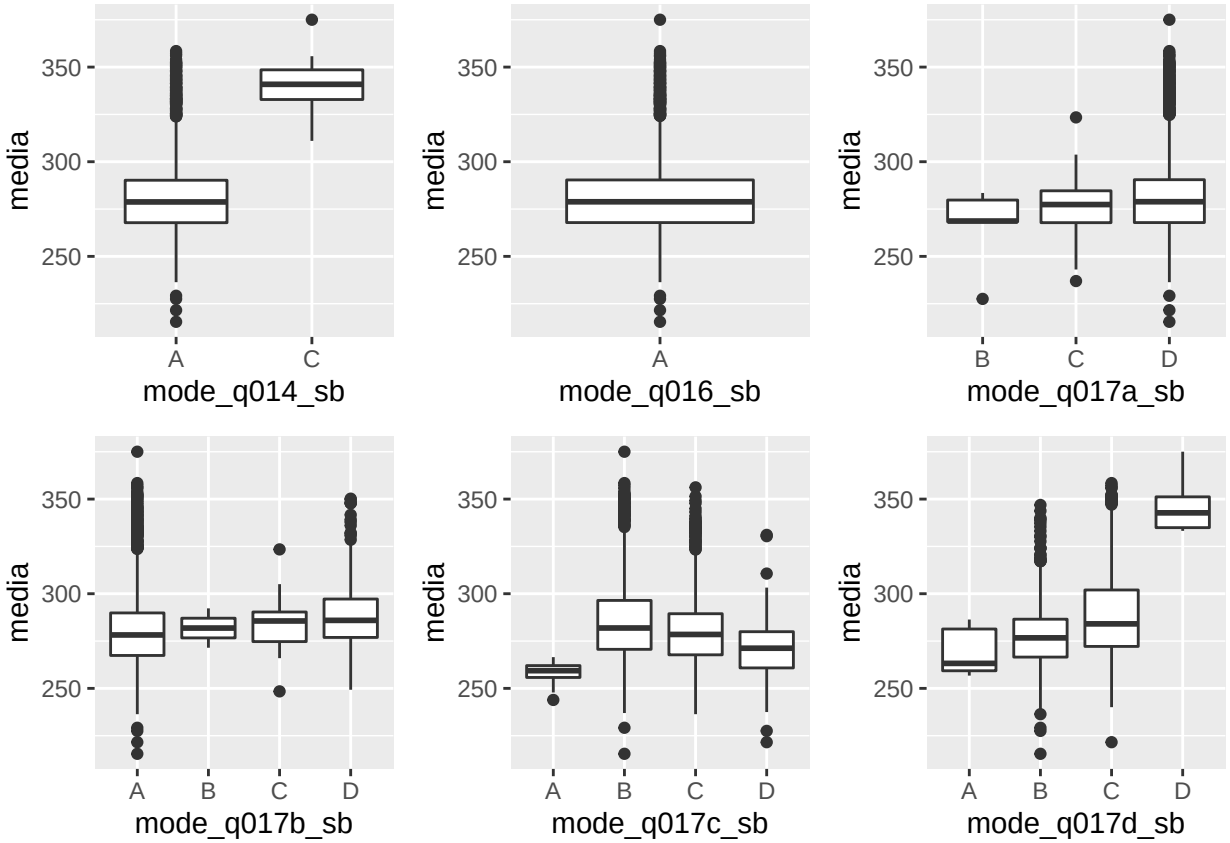

media
350
300
250
A
B
mode_q003d_sb
media
350
300
250
A
B
mode_q003e_sb
media
350
300
250
A
B
C
D
E
F
mode_q004_sb
media
350
300
250
A
B
C
D
E
F
mode_q005_sb
media
350
300
250
A
B
C
mode_q006a_sb
media
350
300
250
B
C
mode_q006b_sb

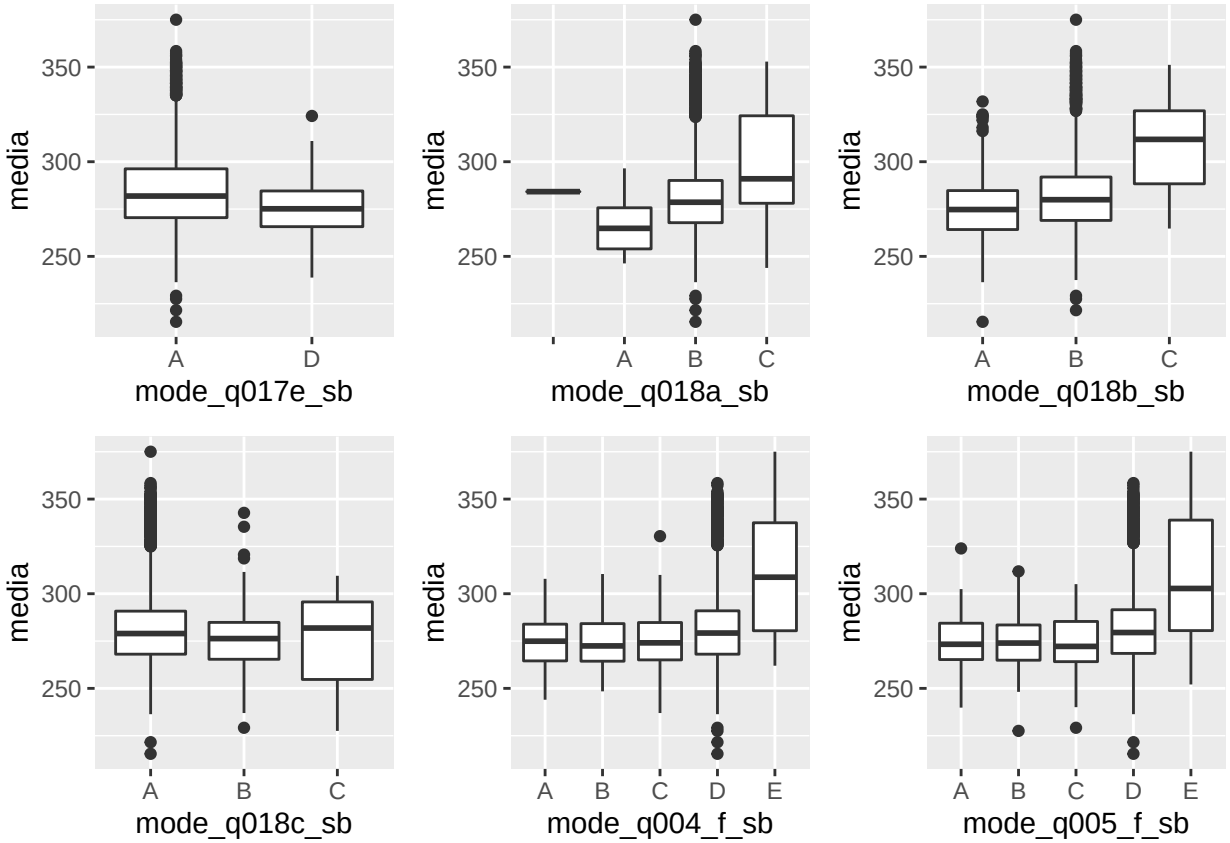

media
350
300
250
A
B
C
mode_q006c_sb
media
350
300
250
C
mode_q006d_sb
media
350
300
250
A
mode_q007_sb
media
350
300
250
A
B
mode_q008a_sb
media
350
300
250
A
B
mode_q008b_sb
media
350
300
250
A
B
mode_q008c_sb


media
350
300
250
B
C
mode_q009a_sb
media
350
300
250
A
mode_q009b_sb
media
350
300
250
A
B
C
D
mode_q009c_sb
media
350
300
250
B
C
D
mode_q009d_sb
media
350
300
250
B
C
D
mode_q009e_sb
media
350
300
250
B
C
D
mode_q009f_sb


media
350
300
250
A
B
C
D
mode_q009g_sb
media
350
300
250
A
B
mode_q010a_sb
media
350
300
250
A
B
mode_q010b_sb
media
350
300
250
A
B
mode_q010c_sb
media
350
300
250
A
B
mode_q010d_sb
media
350
300
250
A
B
mode_q010e_sb


media
350
300
250
A
B
mode_q010f_sb
media
350
300
250
A
B
mode_q010g_sb
media
350
300
250
B
mode_q010h_sb
media
350
300
250
A
B
mode_q010i_sb
media
350
300
250
A
B
C
mode_q011_sb
media
350
300
250
A
B
C
E
F
G
mode_q012_sb


media
350
300
250
A
C
mode_q014_sb
media
350
300
250
A
mode_q016_sb
media
350
300
250
B
C
D
mode_q017a_sb
media
350
300
250
A
B
C
D
mode_q017b_sb
media
350
300
250
A
B
C
D
mode_q017c_sb
media
350
300
250
A
B
C
D
mode_q017d_sb



Análise Univariada

Variáves mais significativas

- **mode_q012_sb:** Considerando a maior distância percorrida, normalmente de que forma você chega à sua escola?
- **mode_q010d_sb:** Na sua casa tem: - Mesa para estudar (ou escrivaninha).
- **mode_q010g_sb:** Na sua casa tem: - Aspirador de pó.
- **mode_q005_sb:** Qual é a maior escolaridade de seu pai (ou homem responsável por você)?
- **mode_q002_sb:** Qual é a sua cor ou raça?

```
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] "nivel_socio_economico" "mode_q002_sb"          "mode_q005_sb"
##  [4] "mode_q006a_sb"         "mode_q009c_sb"         "mode_q009d_sb"
##  [7] "mode_q009e_sb"         "mode_q009f_sb"         "mode_q009g_sb"
## [10] "mode_q010a_sb"         "mode_q010c_sb"         "mode_q010d_sb"
## [13] "mode_q010e_sb"         "mode_q010g_sb"         "mode_q012_sb"
```

```
## [16] "mode_q017b_sb"         "mode_q017c_sb"         "mode_q017d_sb"
## [19] "mode_q017e_sb"         "mode_q018b_sb"         "mode_q018c_sb"
## [22] "mode_q005_f_sb"
```

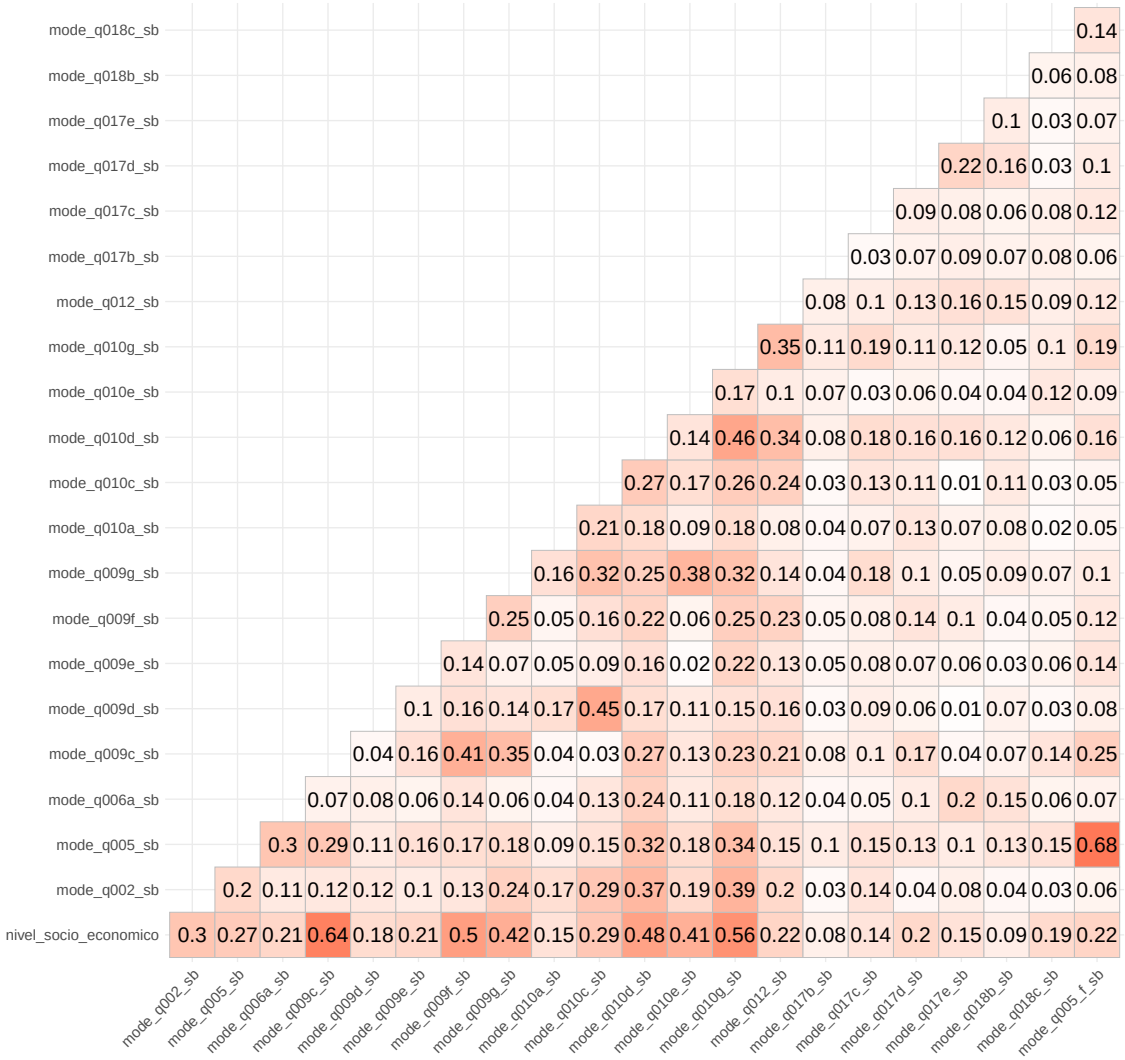
```
tb_r2 <- data.frame(var = final_cols)

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  nivel_socio_economico 0.361536143
## 2           mode_q002_sb 0.167977605
## 3           mode_q005_sb 0.187664789
## 4          mode_q006a_sb 0.119331996
## 5          mode_q009c_sb 0.065726193
## 6          mode_q009d_sb 0.039633040
## 7          mode_q009e_sb 0.017049430
## 8          mode_q009f_sb 0.127835792
## 9          mode_q009g_sb 0.080912696
## 10         mode_q010a_sb 0.001317518
## 11         mode_q010c_sb 0.082522023
## 12         mode_q010d_sb 0.234957292
## 13         mode_q010e_sb 0.021941578
## 14         mode_q010g_sb 0.198409218
## 15          mode_q012_sb 0.238266731
## 16         mode_q017b_sb 0.012940587
## 17         mode_q017c_sb 0.052411301
## 18         mode_q017d_sb 0.125845922
## 19         mode_q017e_sb 0.065242898
## 20         mode_q018b_sb 0.045742328
## 21         mode_q018c_sb 0.005495327
## 22        mode_q005_f_sb 0.030805750
```

MAtriz de correlação

Sem variáveis correlacionadas (> 50%)
Com exceção da variável de formação do pai transformada

```
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_q018c_sb 0.14
mode_q018b_sb 0.06 0.08
mode_q017e_sb 0.1 0.03 0.07
mode_q017d_sb 0.22 0.16 0.03 0.1
mode_q017c_sb 0.09 0.08 0.06 0.08 0.12
mode_q017b_sb 0.03 0.07 0.09 0.07 0.08 0.06
mode_q012_sb 0.08 0.1 0.13 0.16 0.15 0.09 0.12
mode_q010g_sb 0.35 0.11 0.19 0.11 0.12 0.05 0.1 0.19
mode_q010e_sb 0.17 0.1 0.07 0.03 0.06 0.04 0.04 0.12 0.09
mode_q010d_sb 0.14 0.46 0.34 0.08 0.18 0.16 0.16 0.12 0.06 0.16
mode_q010c_sb 0.27 0.17 0.26 0.24 0.03 0.13 0.11 0.01 0.11 0.03 0.05
mode_q010a_sb 0.21 0.18 0.09 0.18 0.08 0.04 0.07 0.13 0.07 0.08 0.02 0.05
mode_q009g_sb 0.16 0.32 0.25 0.38 0.32 0.14 0.04 0.18 0.1 0.05 0.09 0.07 0.1
mode_q009f_sb 0.25 0.05 0.16 0.22 0.06 0.25 0.23 0.05 0.08 0.14 0.1 0.04 0.05 0.12
mode_q009e_sb 0.14 0.07 0.05 0.09 0.16 0.02 0.22 0.13 0.05 0.08 0.07 0.06 0.03 0.06 0.14
mode_q009d_sb 0.1 0.16 0.14 0.17 0.45 0.17 0.11 0.15 0.16 0.03 0.09 0.06 0.01 0.07 0.03 0.08
mode_q009c_sb 0.04 0.16 0.41 0.35 0.04 0.03 0.27 0.13 0.23 0.21 0.08 0.1 0.17 0.04 0.07 0.14 0.25
mode_q006a_sb 0.07 0.08 0.06 0.14 0.06 0.04 0.13 0.24 0.11 0.18 0.12 0.04 0.05 0.1 0.2 0.15 0.06 0.07
mode_q005_sb 0.3 0.29 0.11 0.16 0.17 0.18 0.09 0.15 0.32 0.18 0.34 0.15 0.1 0.15 0.13 0.1 0.13 0.15 0.68
mode_q002_sb 0.2 0.11 0.12 0.12 0.1 0.13 0.24 0.17 0.29 0.37 0.19 0.39 0.2 0.03 0.14 0.04 0.08 0.04 0.03 0.06
nivel_socio_economico 0.3 0.27 0.21 0.64 0.18 0.21 0.5 0.42 0.15 0.29 0.48 0.41 0.56 0.22 0.08 0.14 0.2 0.15 0.09 0.19 0.22
mode_q002_sb
mode_q005_sb
mode_q006a_sb
mode_q009c_sb
mode_q009d_sb
mode_q009e_sb
mode_q009f_sb
mode_q009g_sb
mode_q010a_sb
mode_q010c_sb
mode_q010d_sb
mode_q010e_sb
mode_q010g_sb
mode_q012_sb
mode_q017b_sb
mode_q017c_sb
mode_q017d_sb
mode_q017e_sb
mode_q018b_sb
mode_q018c_sb
mode_q005_f_sb