

Análise das variáveis Saeb - moda por escola

Série 5EF

Livia Kobayashi

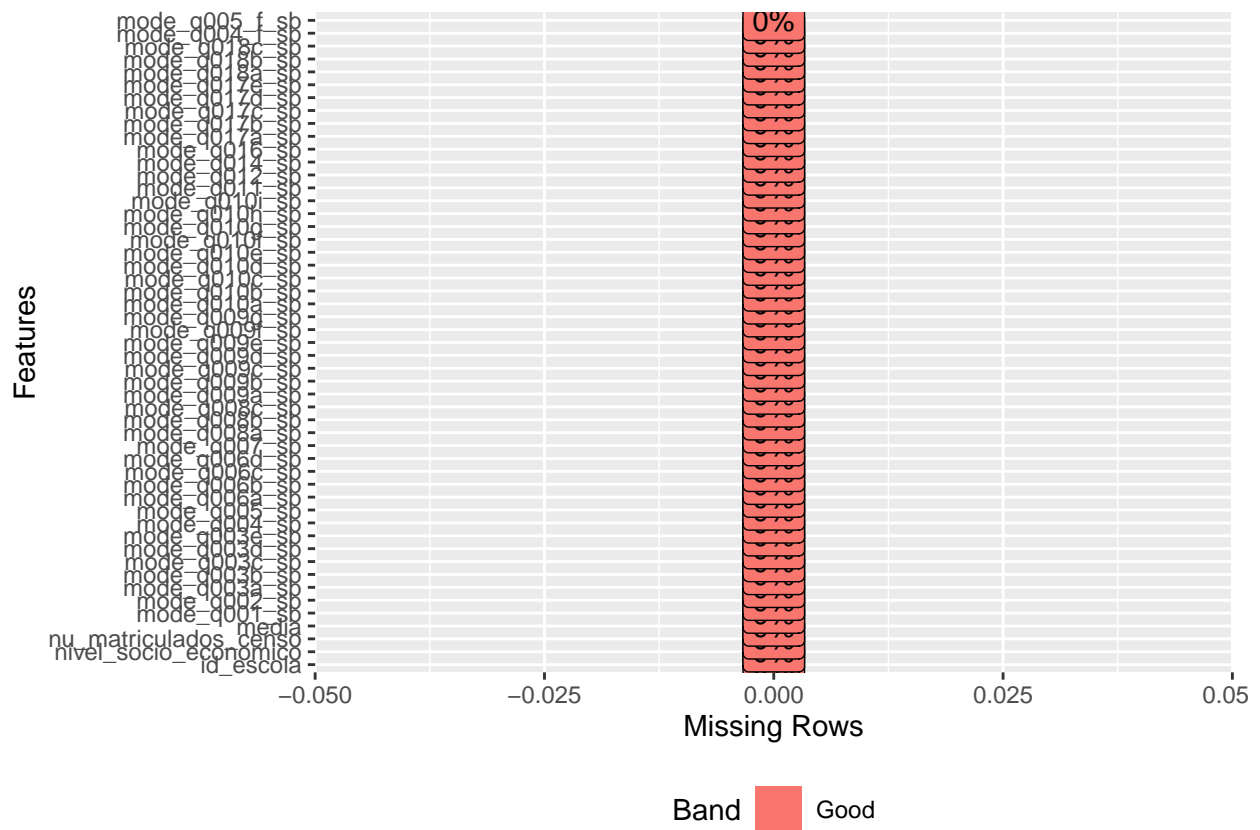
14 junho 2021

```
## # A tibble: 1 x 2
## # Groups:   id_serie [1]
##   id_serie      n
##   <chr>    <int>
## 1 5EF      5764
```

Missing

Base 100% preenchida

```
plot_missing(df)
```

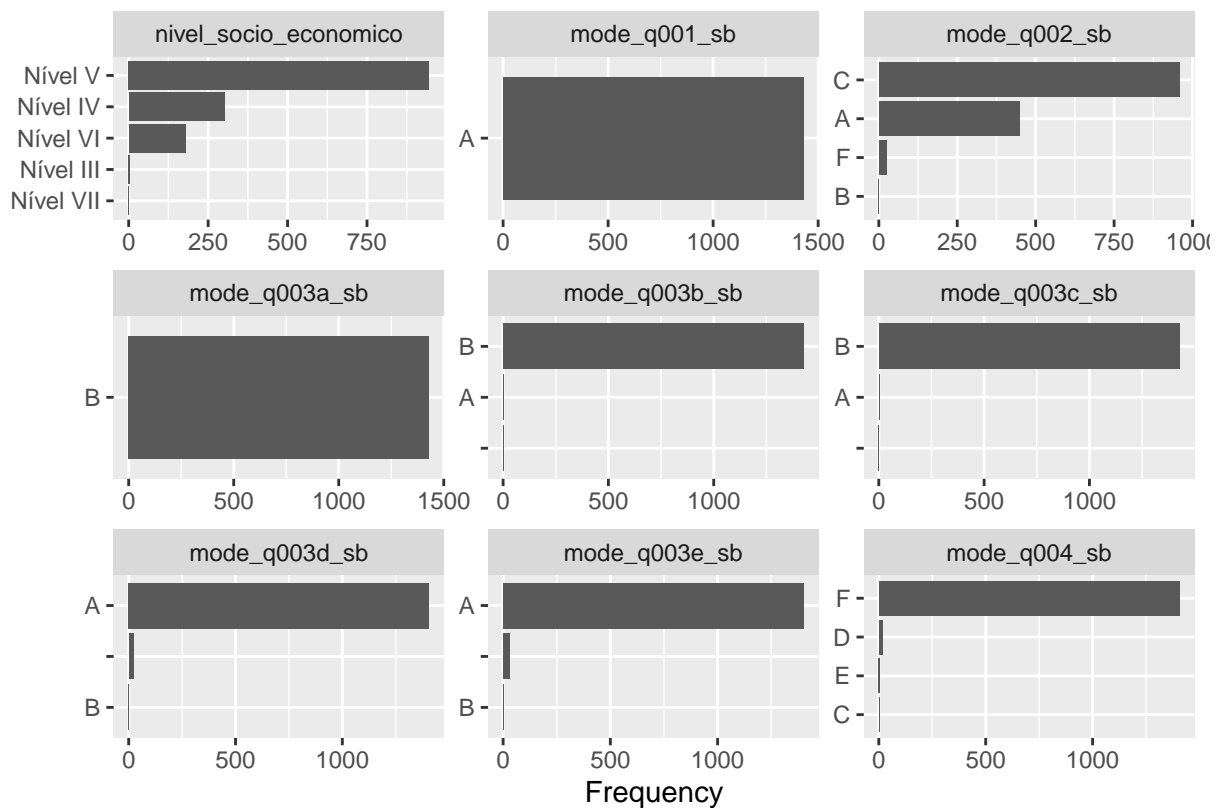


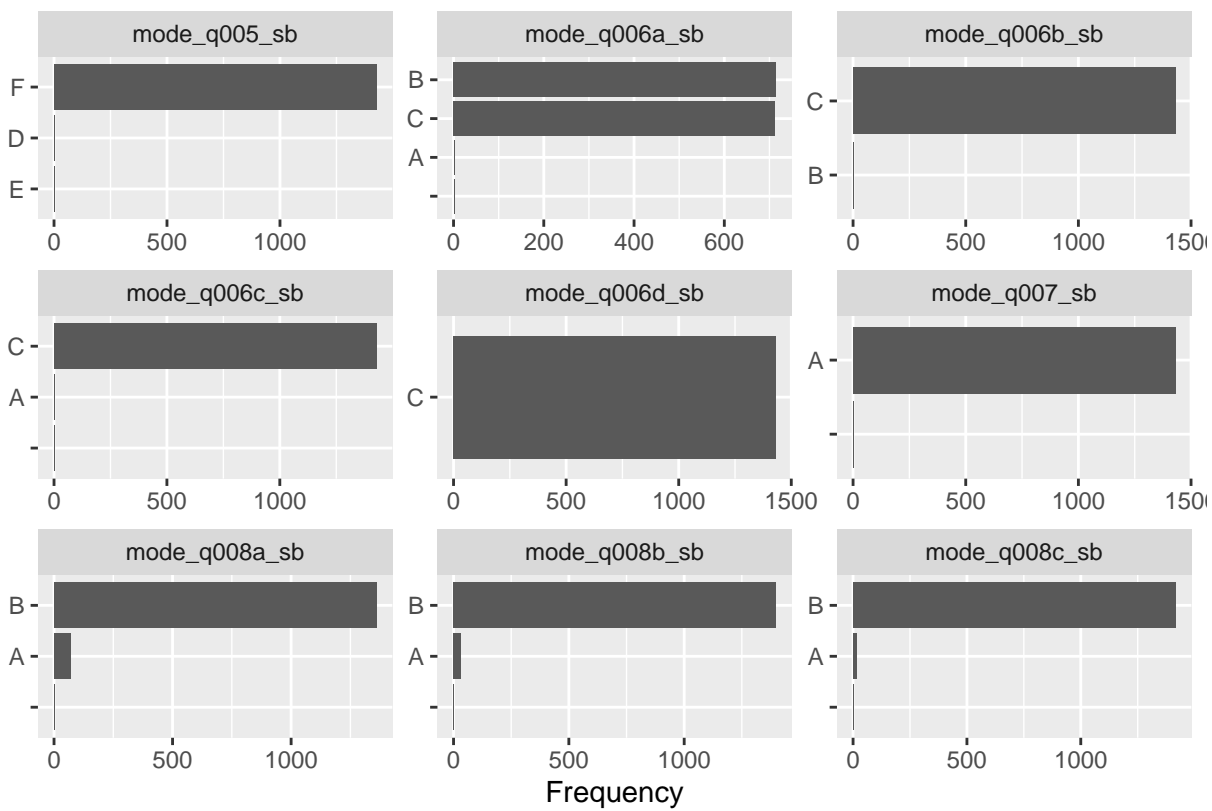
Volume

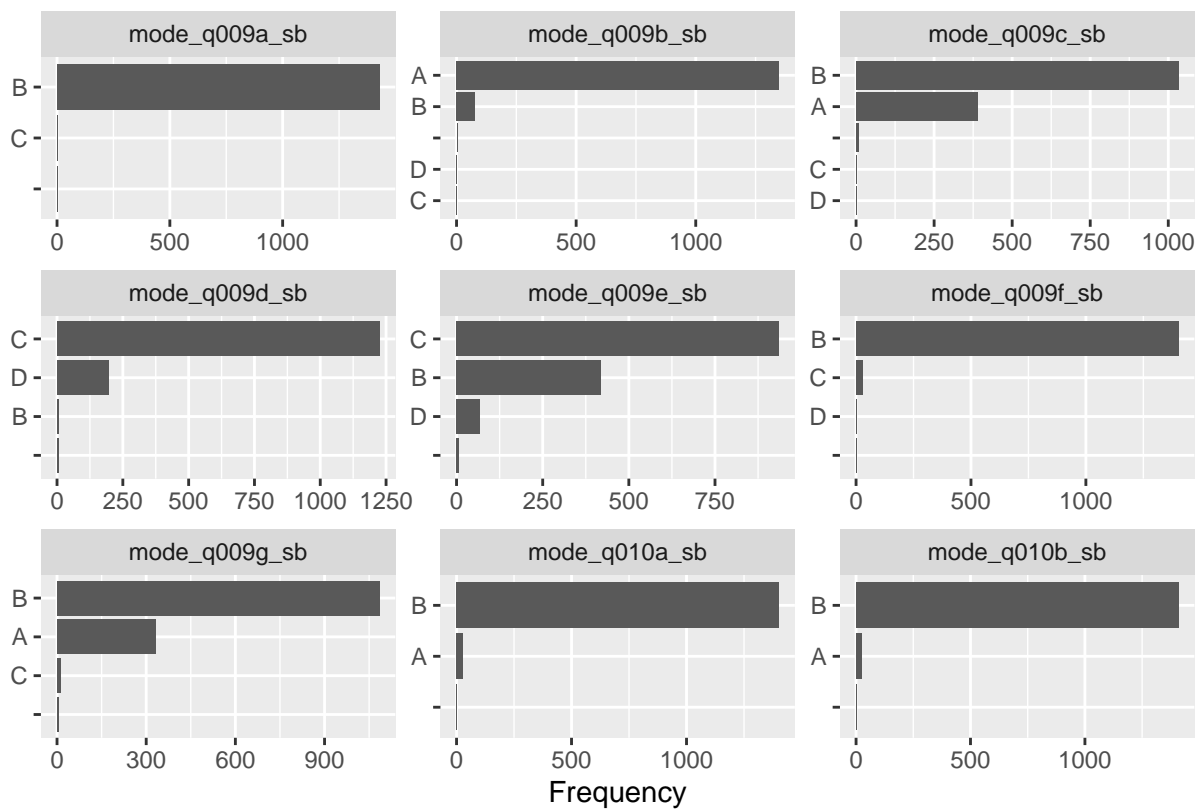
Variáveis com baixa variância:

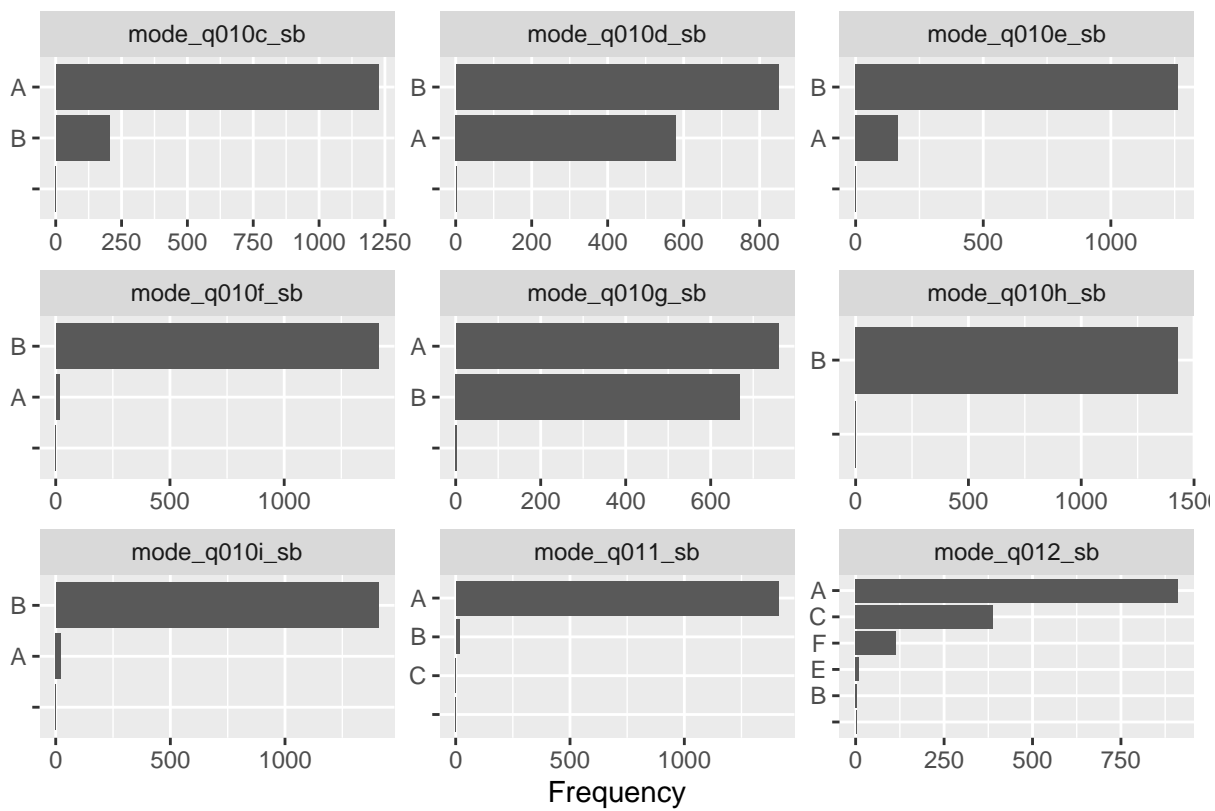
- mode_q001_sb - mode_q003a_sb - mode_q003b_sb - mode_q003c_sb
- mode_q003d_sb - mode_q003e_sb - mode_q006b_sb
- mode_q006c_sb - mode_q006d_sb - mode_q007_sb - mode_q008a_sb
- mode_q008b_sb - mode_q008c_sb - mode_q009a_sb - mode_q009b_sb
- mode_q009f_sb - mode_q010b_sb - mode_q010e_sb - mode_q010f_sb
- mode_q010h_sb - mode_q010i_sb - mode_q011_sb - mode_q014_sb
- mode_q016_sb - mode_q017a_sb - mode_q017b_sb - mode_q017e_sb
- mode_q004_f_sb - mode_q005_f_sb

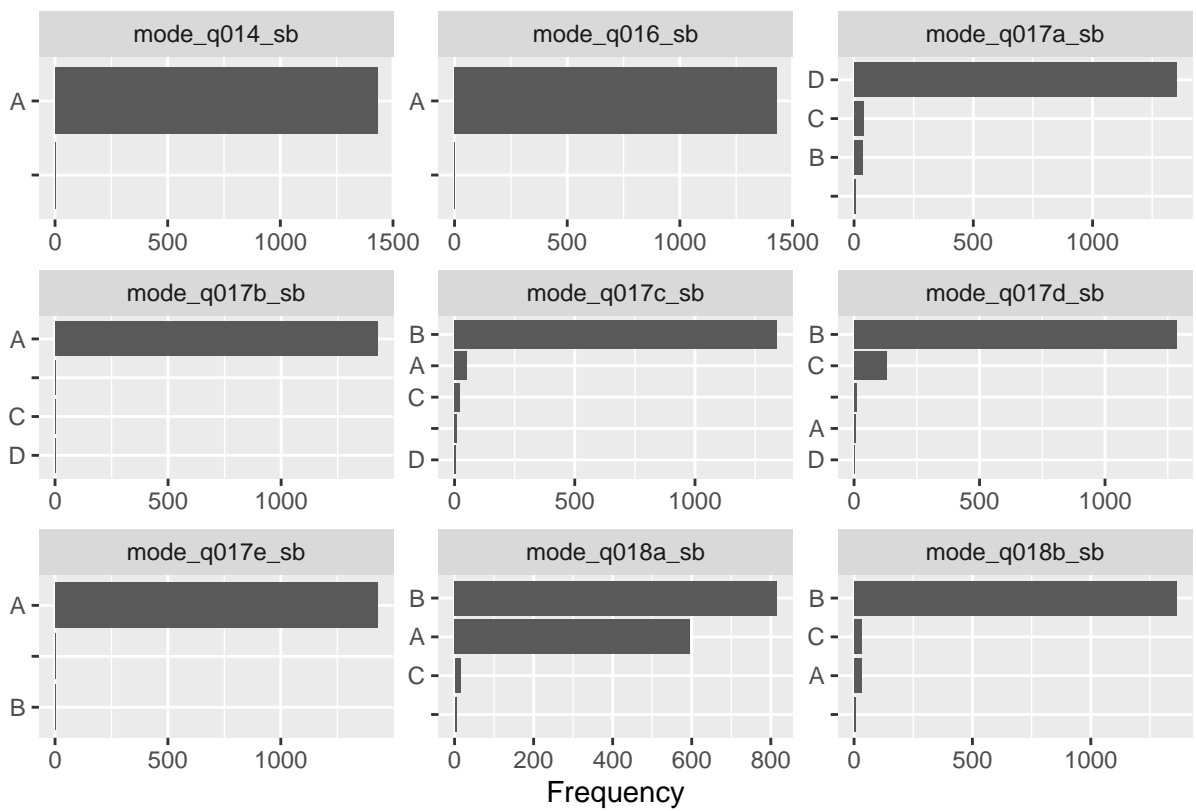
```
plot_bar(final_data)
```

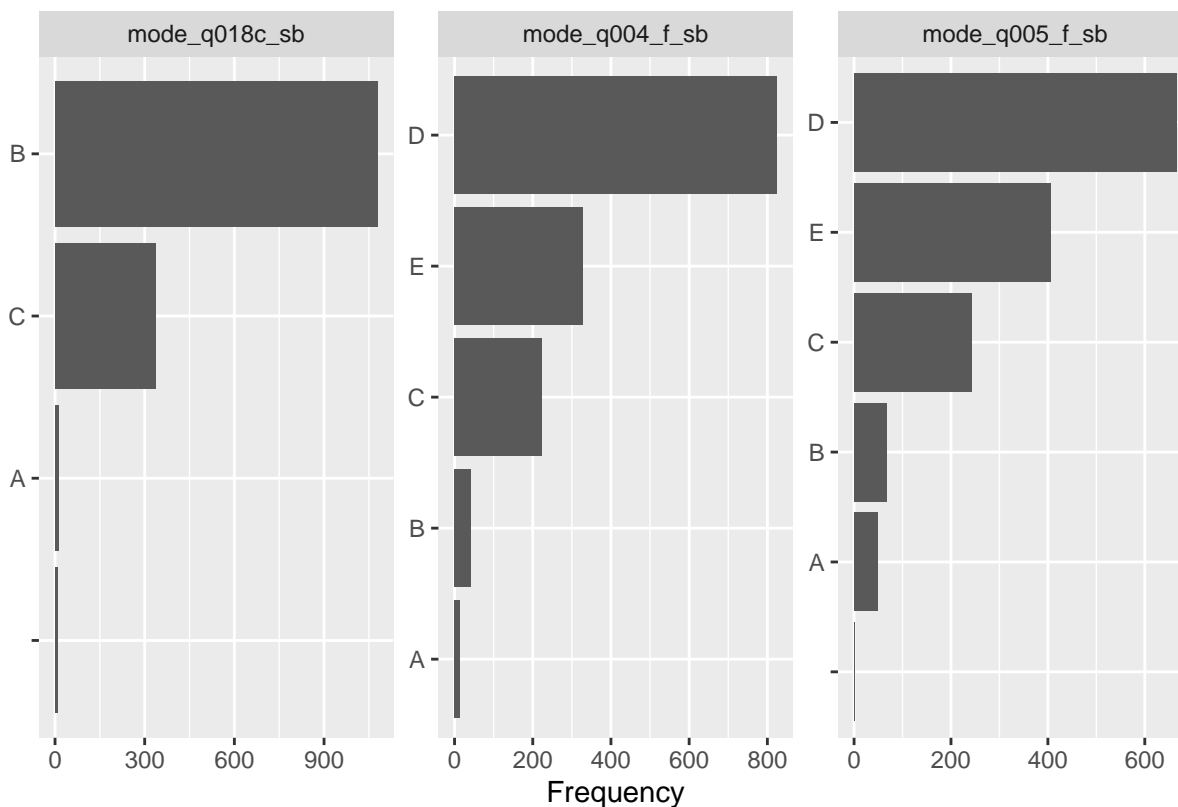












Page 6

Médias das notas x variáveis

Variáveis que discriminam e tem volume nas categorias:

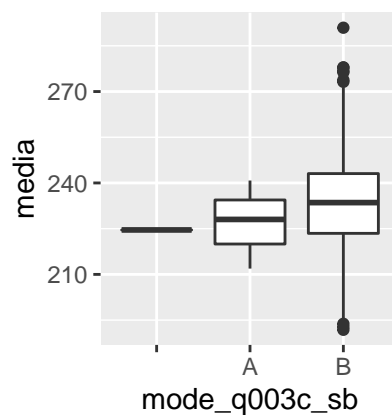
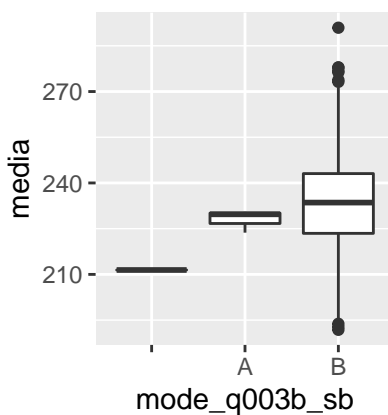
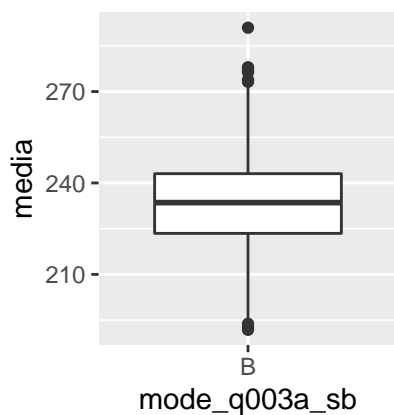
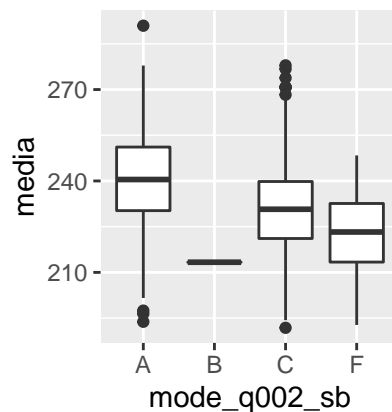
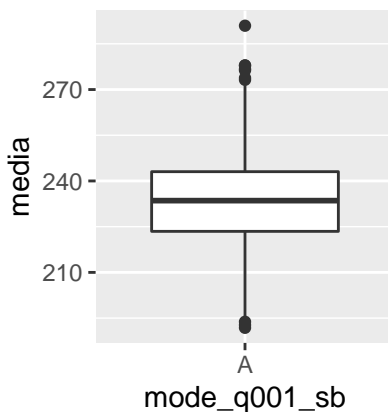
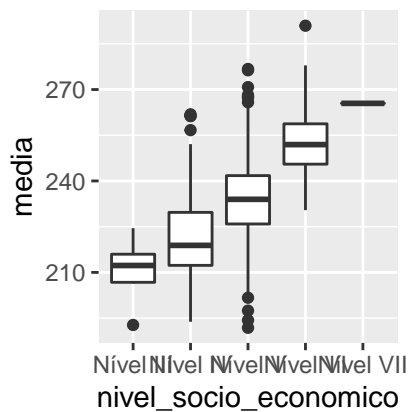
- mode_q002_sb: Maioria branca => notas maiores
- mode_q004_sb: maior graduação => notas maiores (agrupar A, B e C)
- mode_q005_sb: maior graduação => notas maiores (agrupar A, B e C)
- mode_q006a_sb: Os pais conversarem sobre a escola => notas >
- mode_q009c_sb: # computador => notas >
- mode_q009d_sb: # Quartos => notas >
- mode_q009g_sb: # 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 => notas maiores
- a.À pé.

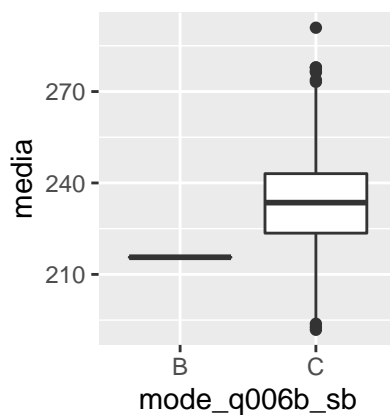
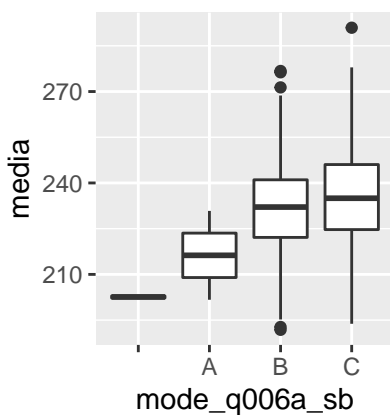
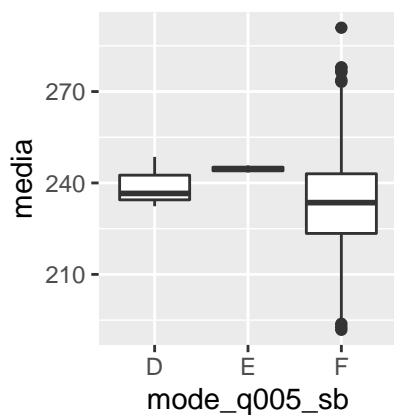
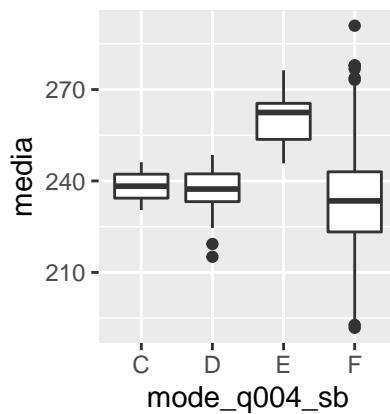
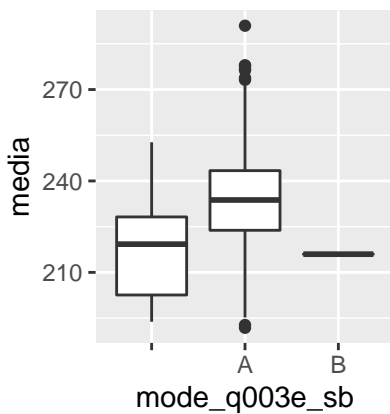
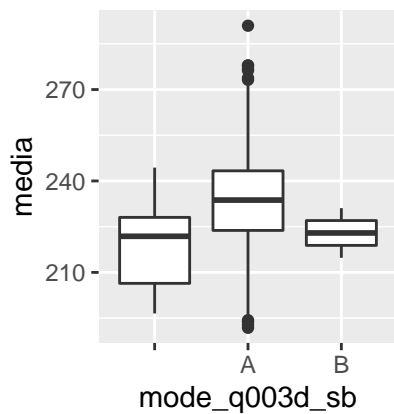
- b.De ônibus urbano.
- c.De transporte escolar.
- d.De barco.
- e.De bicicleta.
- f.De carro
- g.Outros meios de transporte.
- mode_q017c_sb: menos tempo gasto em tarefas domésticas => notas > (Isolar o A e juntar as outras)
- mode_q017d_sb: Não usar!!!
- mode_q018a_sb: Leitura notícias => notas >
- mode_q018b_sb: Leitura extraescolar => notas >

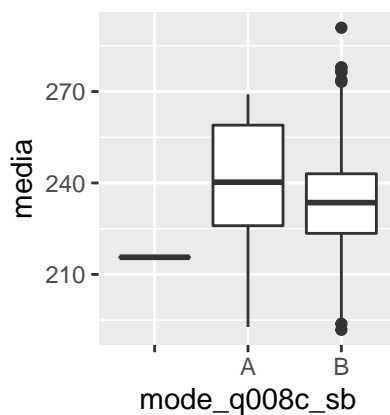
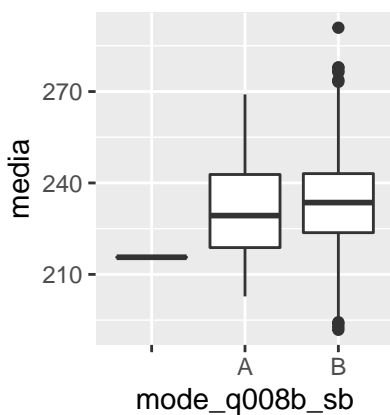
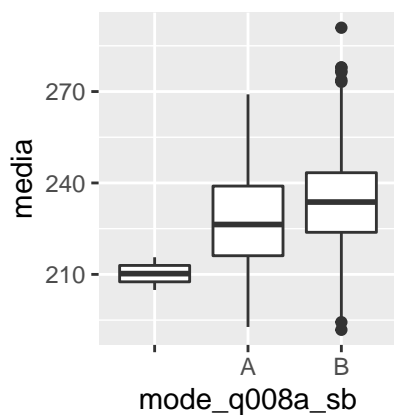
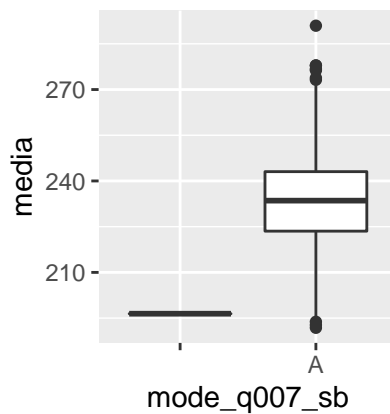
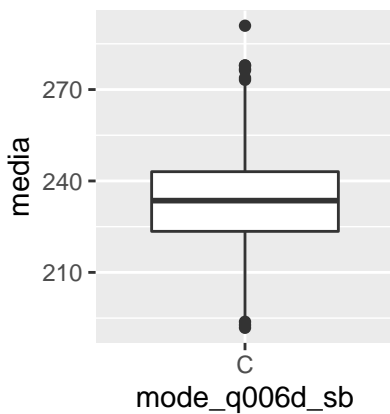
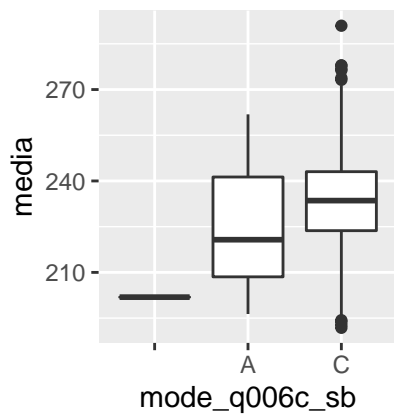
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

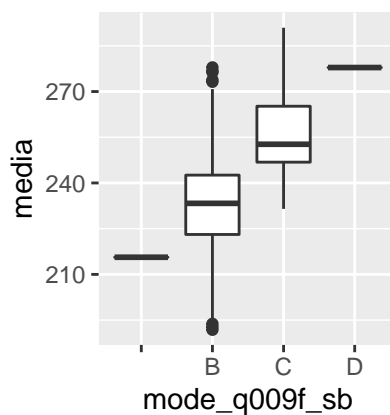
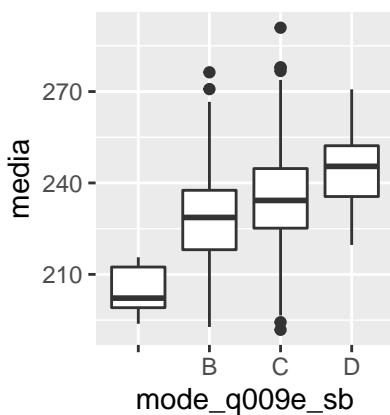
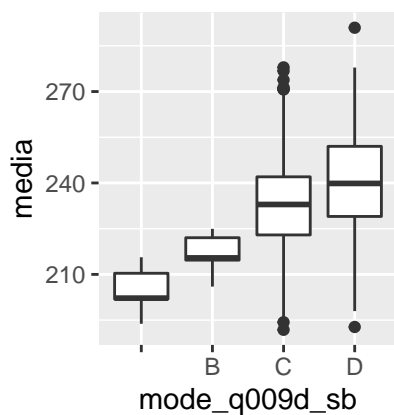
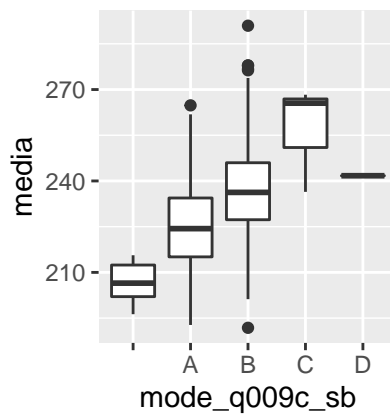
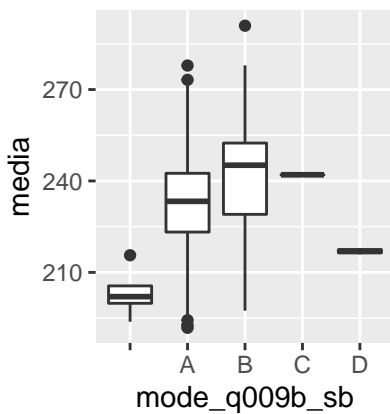
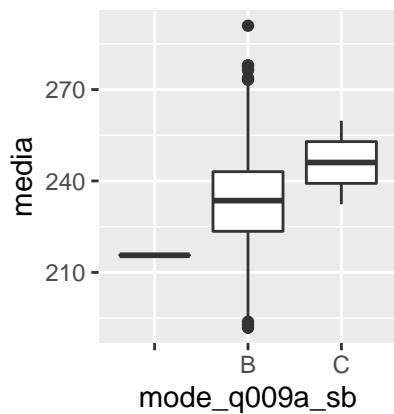
n <- length(plots)

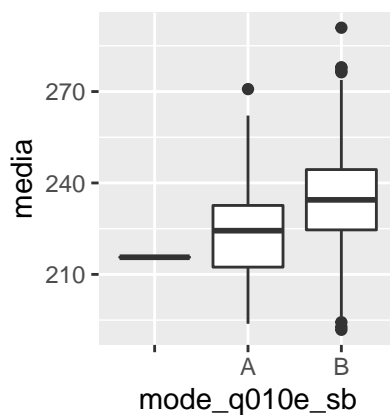
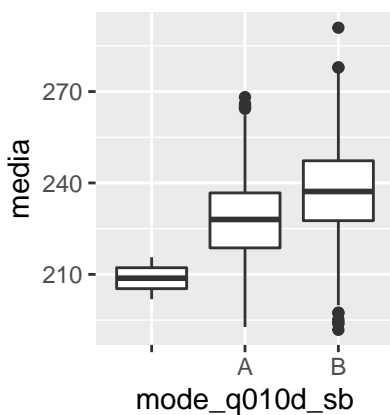
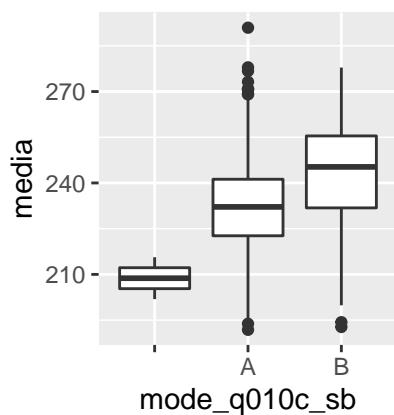
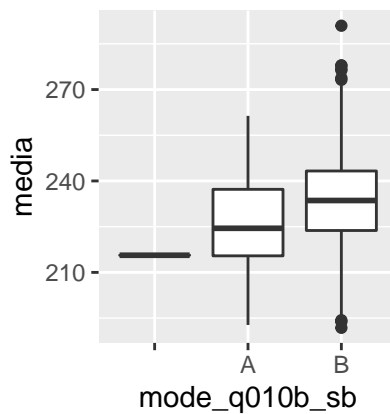
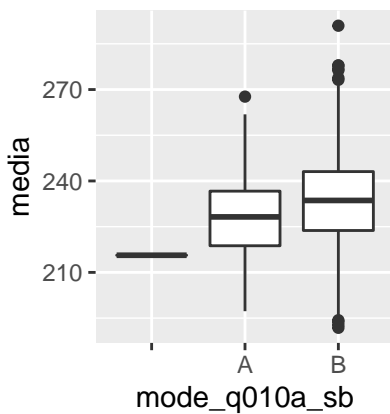
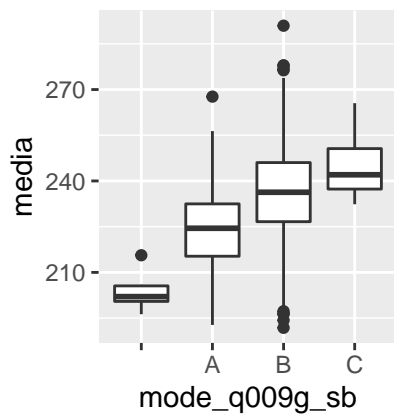
i <- 1
while (i <= n) {
  do.call("grid.arrange", c(plots[i:(min(i+5, n))], ncol=3, nrow = 2))
  i <- i + 6
}
```

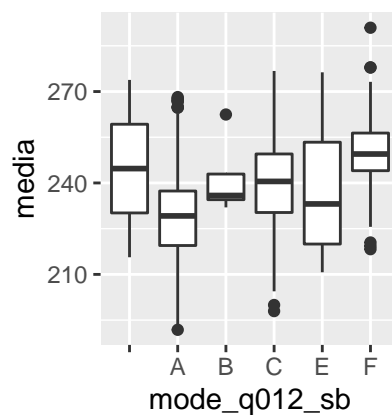
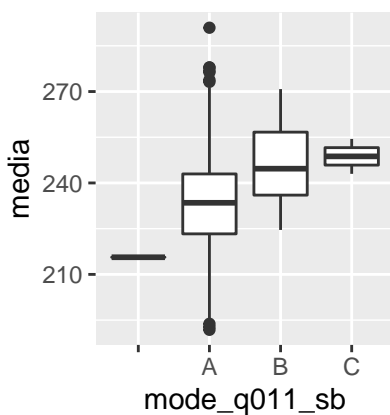
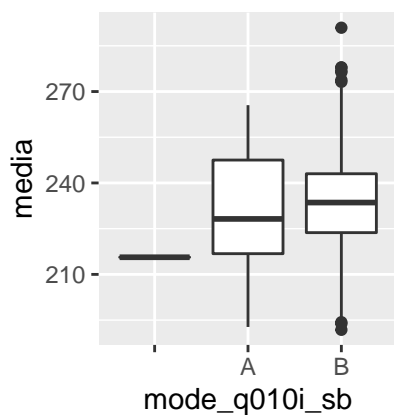
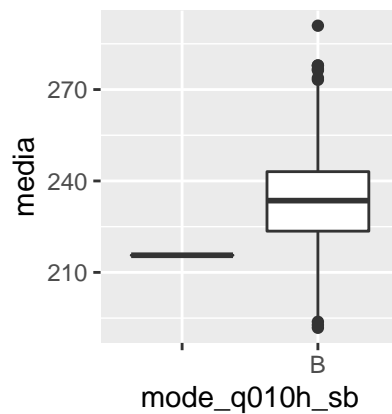
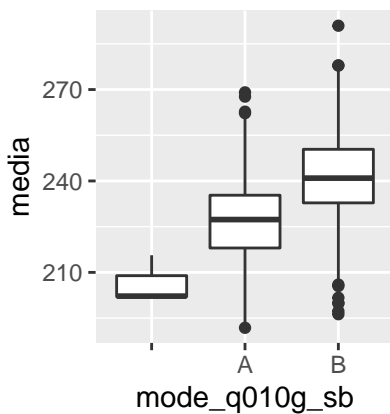
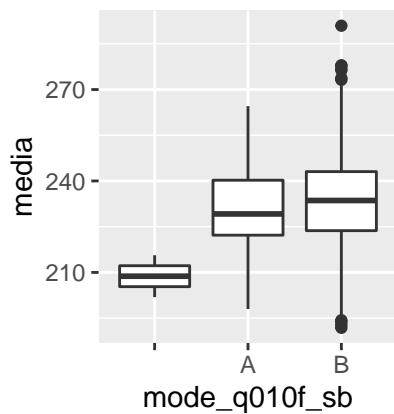



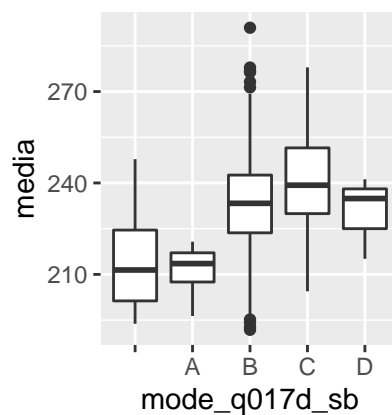
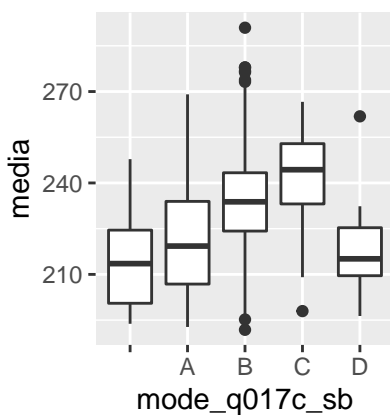
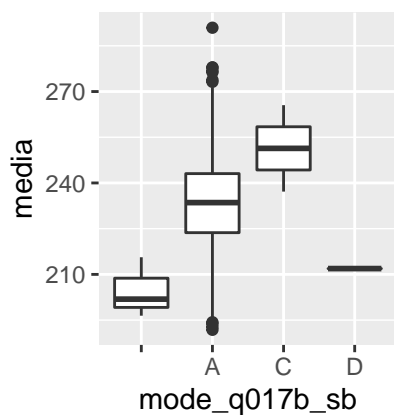
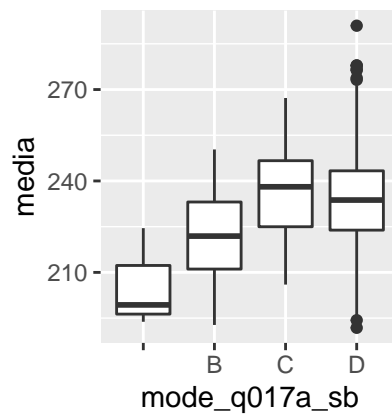
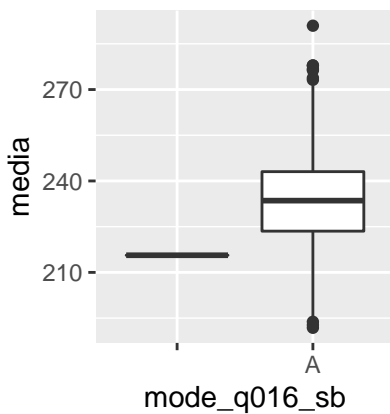
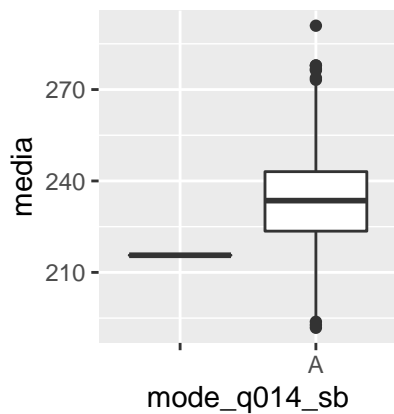


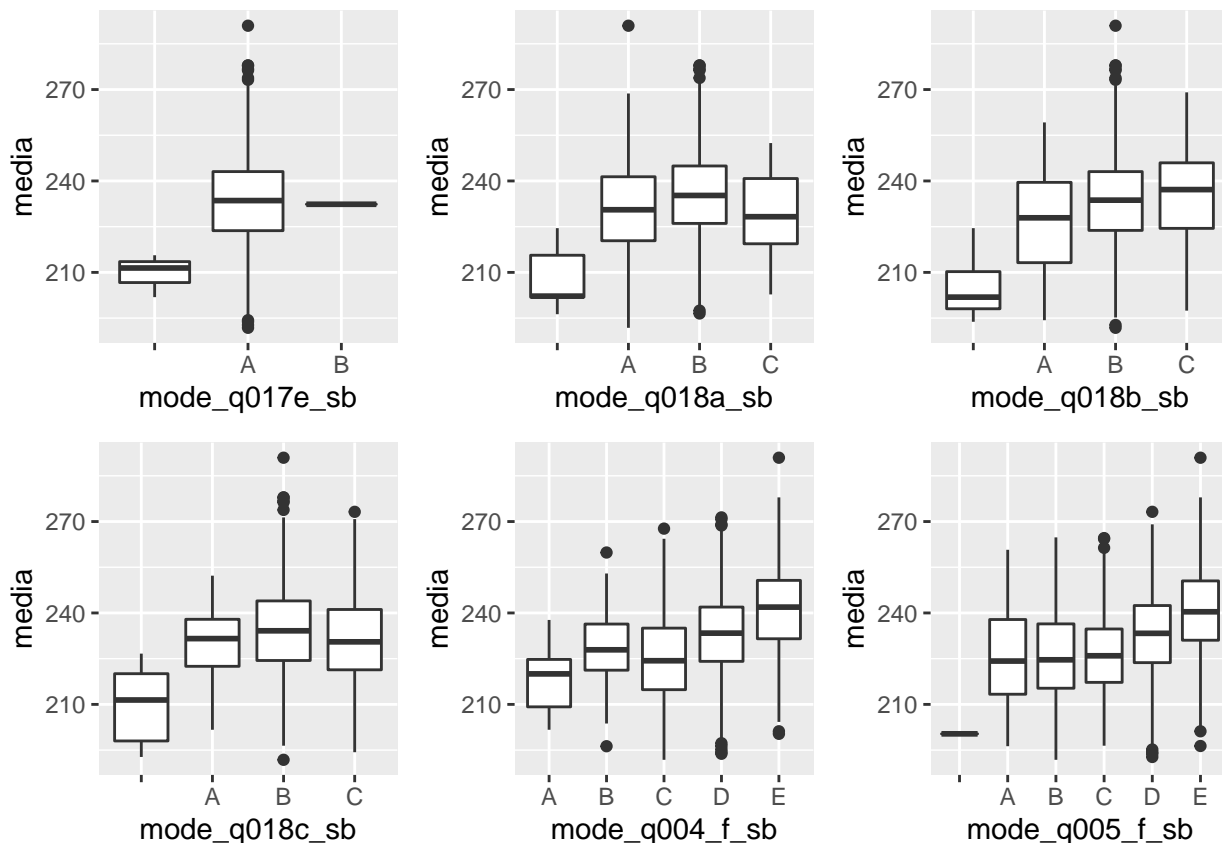












Análise Univariada

Variáveis mais significativas

- **mode_q010d_sb:** Na sua casa tem: - Mesa para estudar (ou escrivaninha).
- **mode_q010g_sb:** Na sua casa tem: - Aspirador de pó.
- **mode_q004_sb:** Qual é a maior escolaridade da sua mãe (ou homem responsável por você)?
- **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_q006a_sb"
## [4] "mode_q009b_sb" "mode_q009c_sb" "mode_q009d_sb"
## [7] "mode_q009e_sb" "mode_q009g_sb" "mode_q010c_sb"
## [10] "mode_q010d_sb" "mode_q010e_sb" "mode_q010g_sb"
## [13] "mode_q012_sb" "mode_q017d_sb" "mode_q018a_sb"
## [16] "mode_q018c_sb" "mode_q004_f_sb" "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 0.34817539
## 2           mode_q002_sb 0.08573031
## 3           mode_q006a_sb 0.02152228
## 4           mode_q009b_sb 0.03311918
## 5           mode_q009c_sb 0.14083616
## 6           mode_q009d_sb 0.04435681
## 7           mode_q009e_sb 0.08283319
## 8           mode_q009g_sb 0.13808660
## 9           mode_q010c_sb 0.06442949
## 10          mode_q010d_sb 0.08728693
## 11          mode_q010e_sb 0.06280649
## 12          mode_q010g_sb 0.22197681
## 13          mode_q012_sb 0.20341427
## 14          mode_q017d_sb 0.03392041
## 15          mode_q018a_sb 0.03746296
## 16          mode_q018c_sb 0.01826739
## 17          mode_q004_f_sb 0.12168962
## 18          mode_q005_f_sb 0.12809613
```

MAtriz de correlação

- mode_q002_sb (raça) e mode_q006a_sb (pais conversam com os filhos) => manter mode_q002_sb ou combinar ambas
- mode_q004_sb (escolaridade mãe) e mode_q006a_sb (pais conversam com os filhos) => manter mode_q004_sb ou combinar ambas
- mode_q018b_sb (leitura livros) e mode_q018c_sb (leitura em quadrinhos) => baixo r2

```
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)
```

[illegible]