Análise das variáveis Saeb - moda por escola

Série 5EF

Livia Kobayashi

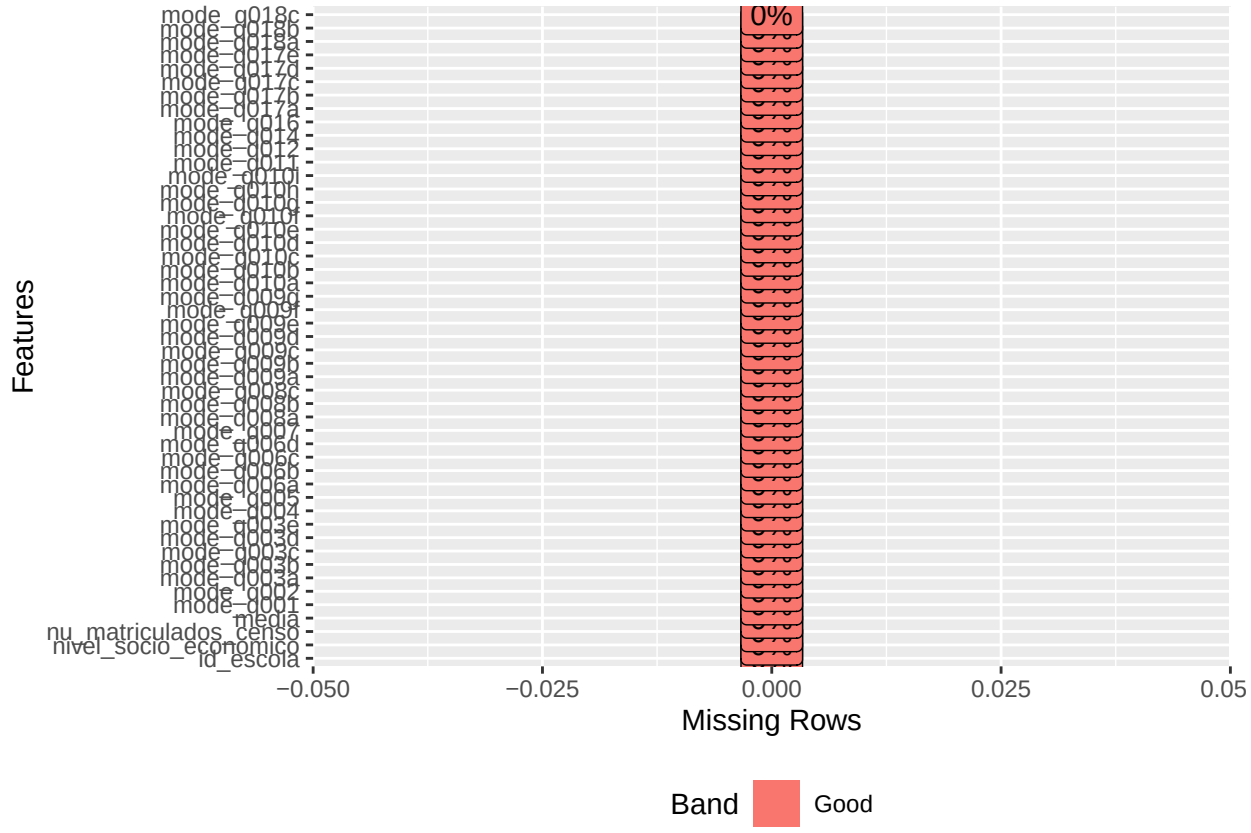
10 junho 2021

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

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

```
##   id_serie
## 1      5EF
```

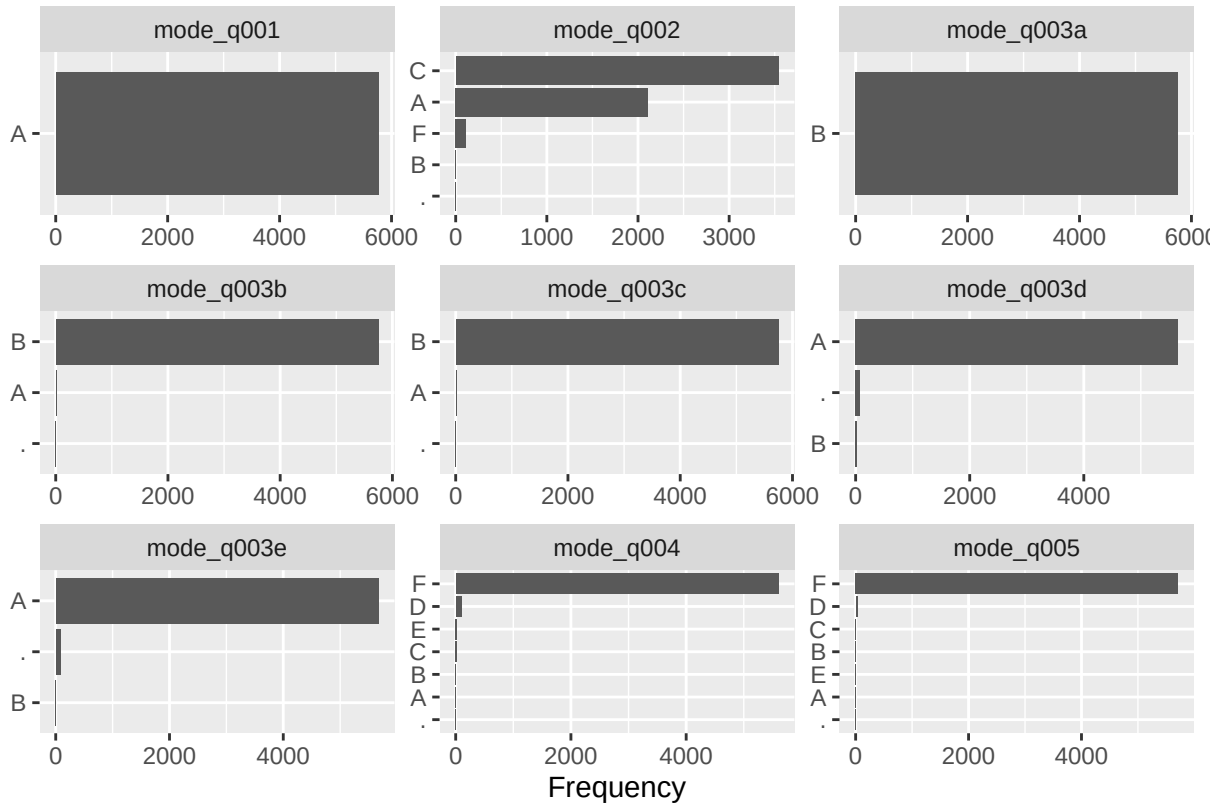
Missing

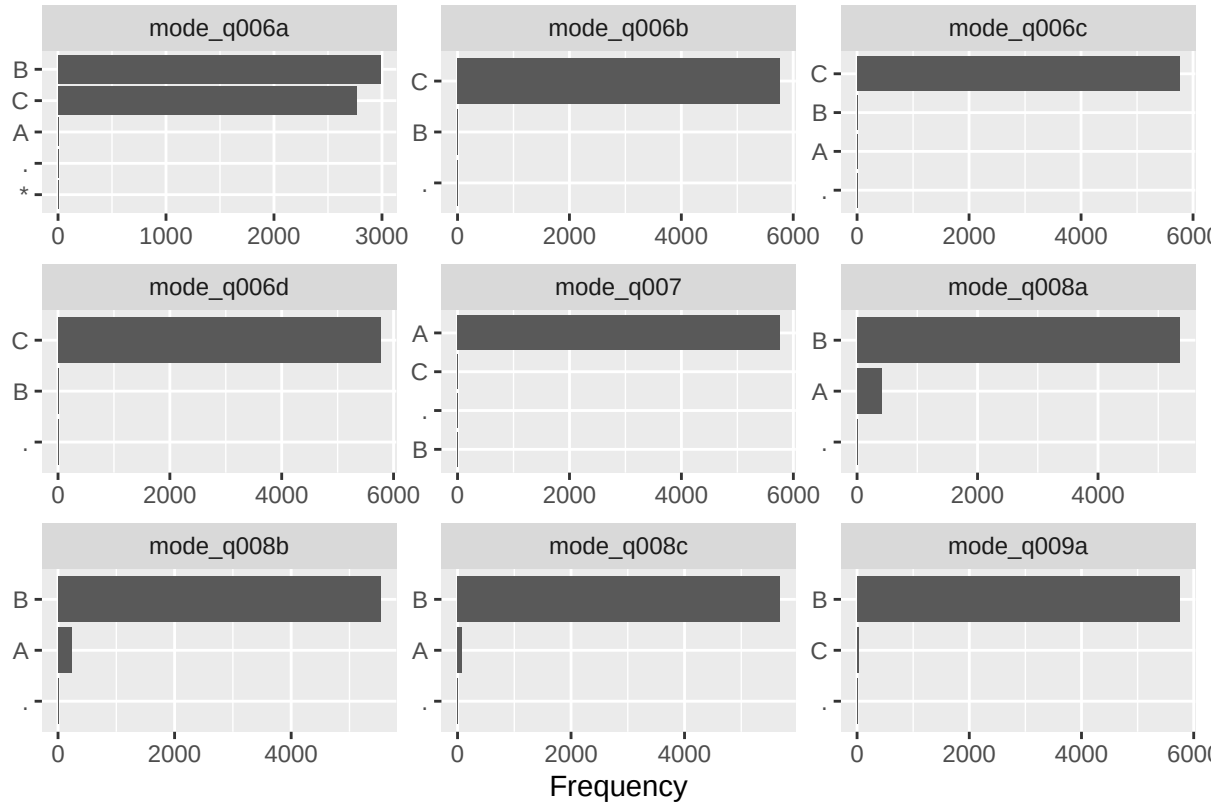
```
plot_missing(df)
```

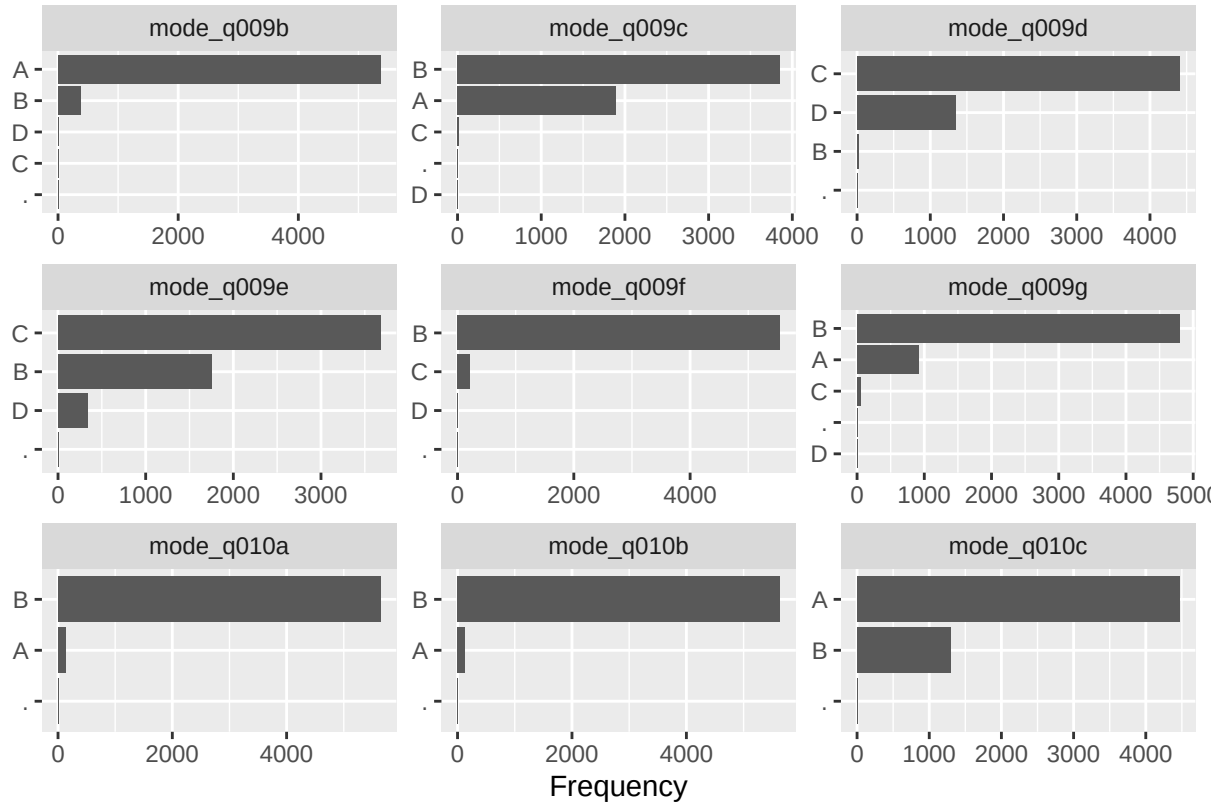


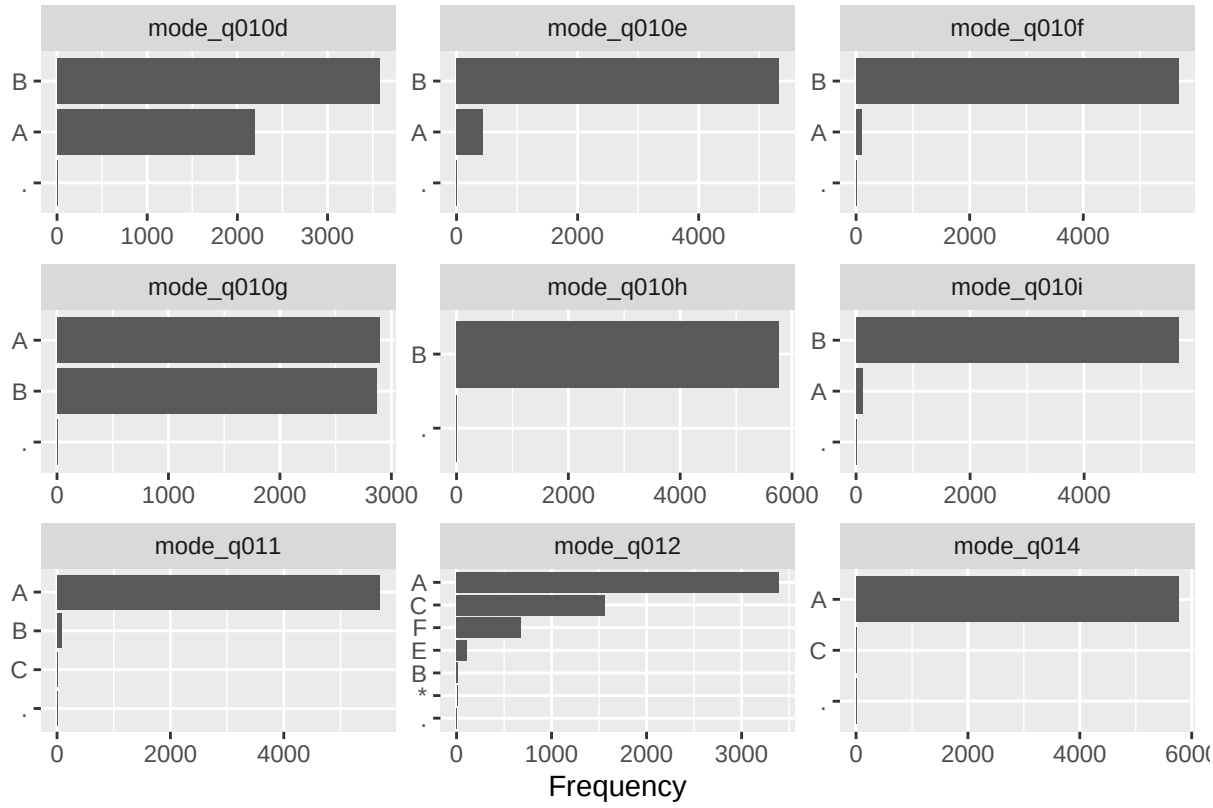
Volume

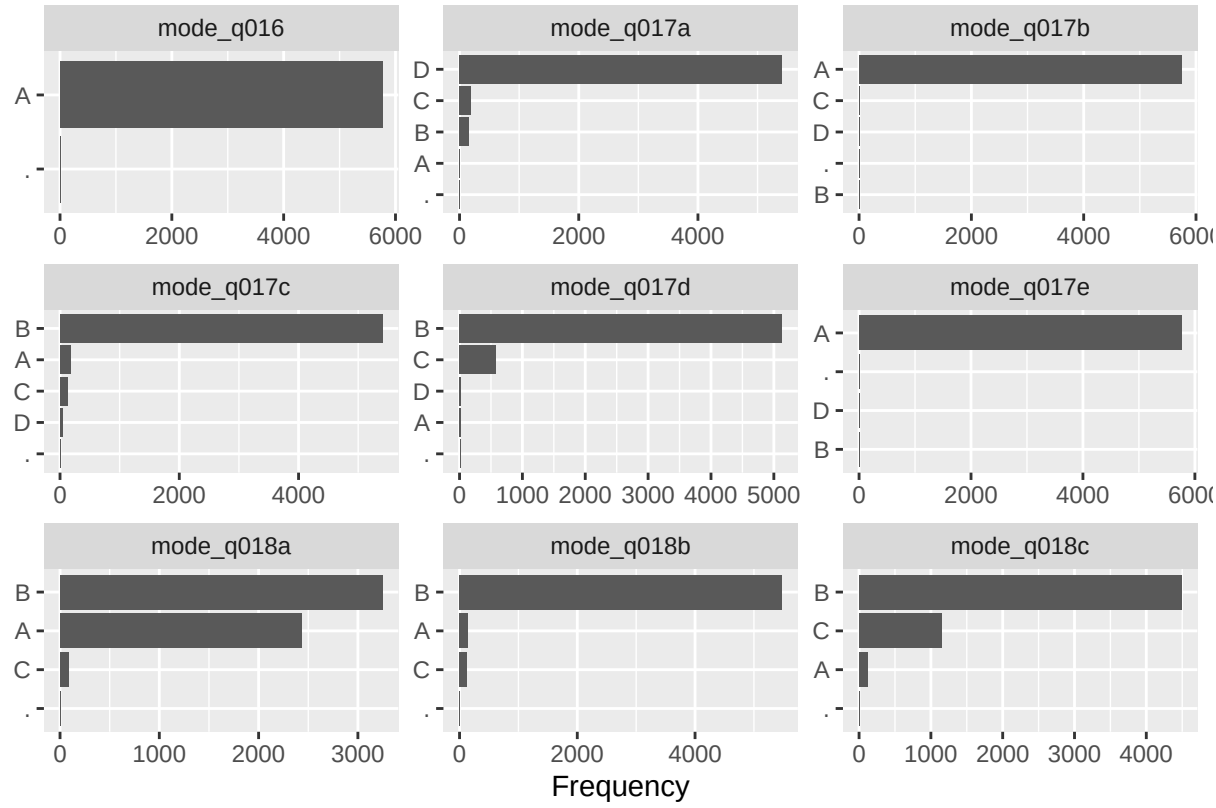
```
plot_bar(final_data)
```




mode_q006a
mode_q006b
mode_q006c
mode_q006d
mode_q007
mode_q008a
mode_q008b
mode_q008c
mode_q009a
Frequency

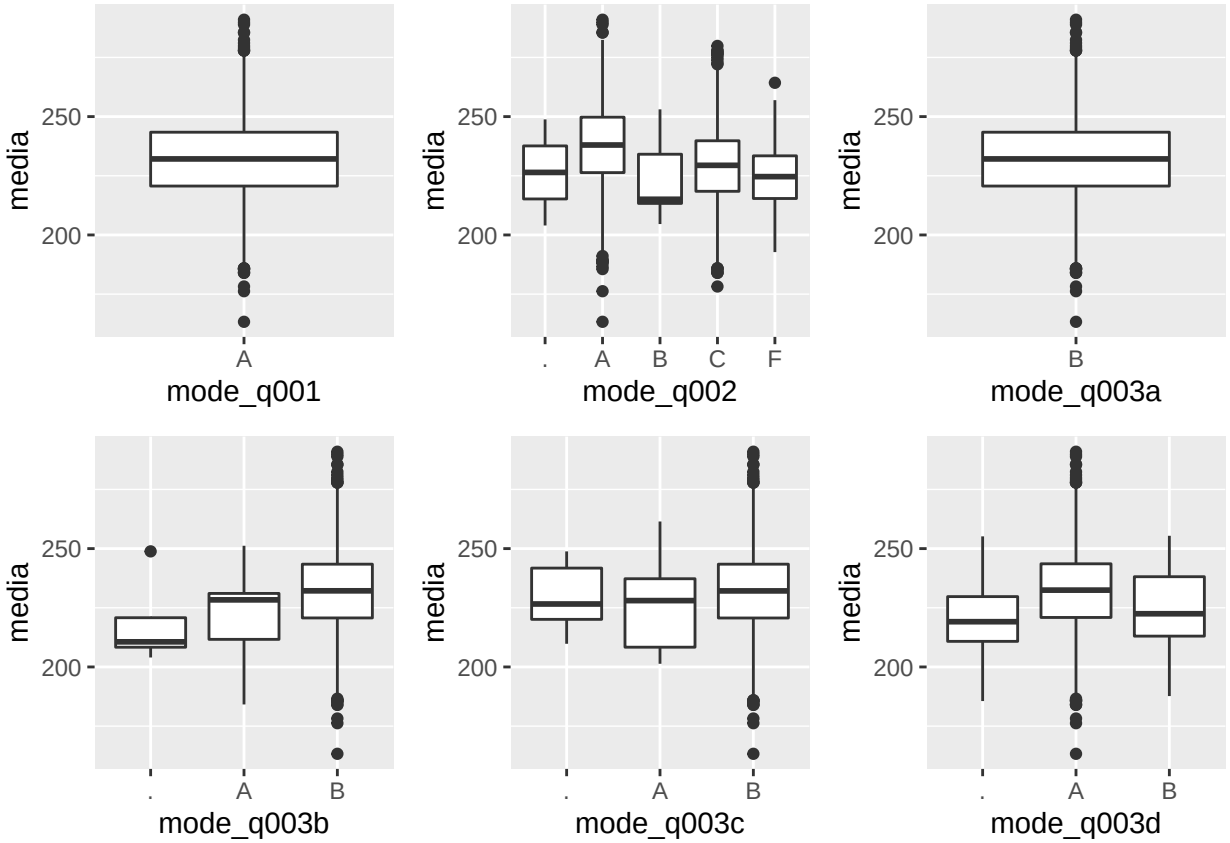

mode_q009b
A
B
D
C
.
0
2000
4000
mode_q009c
B
A
C
.
D
0
1000
2000
3000
4000
mode_q009d
C
D
B
.
0
1000
2000
3000
4000
mode_q009e
C
B
D
.
0
1000
2000
3000
mode_q009f
B
C
D
.
0
2000
4000
mode_q009g
B
A
C
.
D
0
1000
2000
3000
4000
5000
mode_q010a
B
A
.
0
2000
4000
mode_q010b
B
A
.
0
2000
4000
mode_q010c
A
B
.
0
1000
2000
3000
4000
Frequency

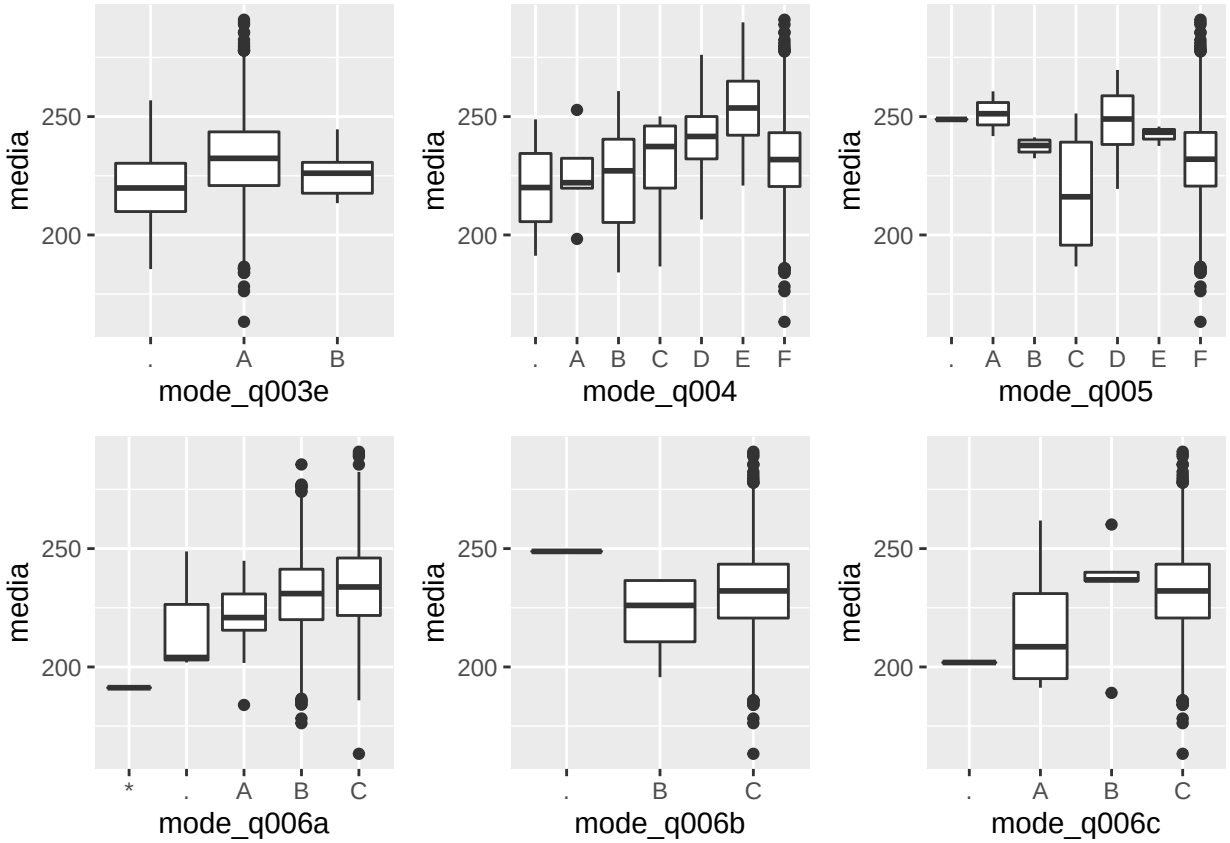

mode_q010d
B
A
.
0
1000
2000
3000
mode_q010e
B
A
.
0
2000
4000
mode_q010f
B
A
.
0
2000
4000
mode_q010g
A
B
.
0
1000
2000
3000
mode_q010h
B
.
0
2000
4000
6000
mode_q010i
B
A
.
0
2000
4000
mode_q011
A
B
C
.
0
2000
4000
mode_q012
A
C
F
E
B
*
.
0
1000
2000
3000
mode_q014
A
C
.
0
2000
4000
6000
Frequency

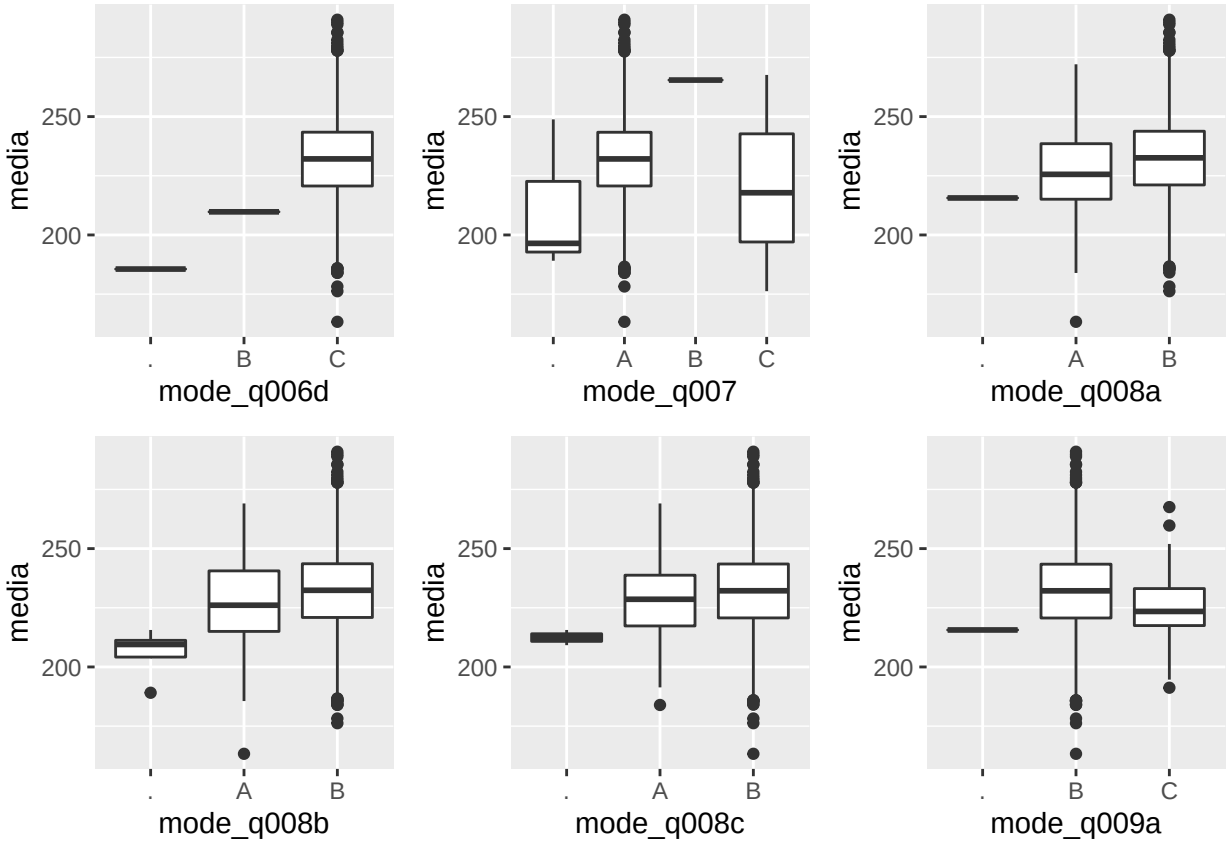


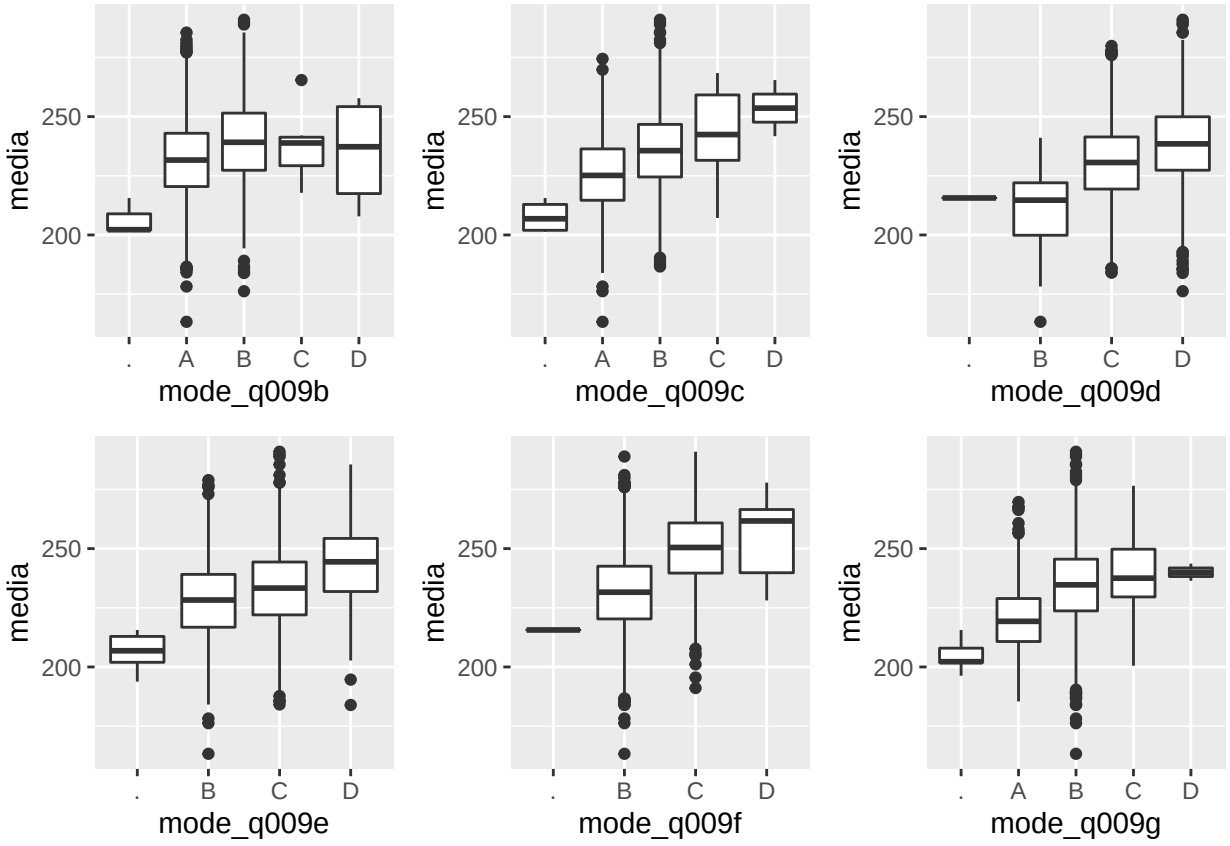
Boxplot

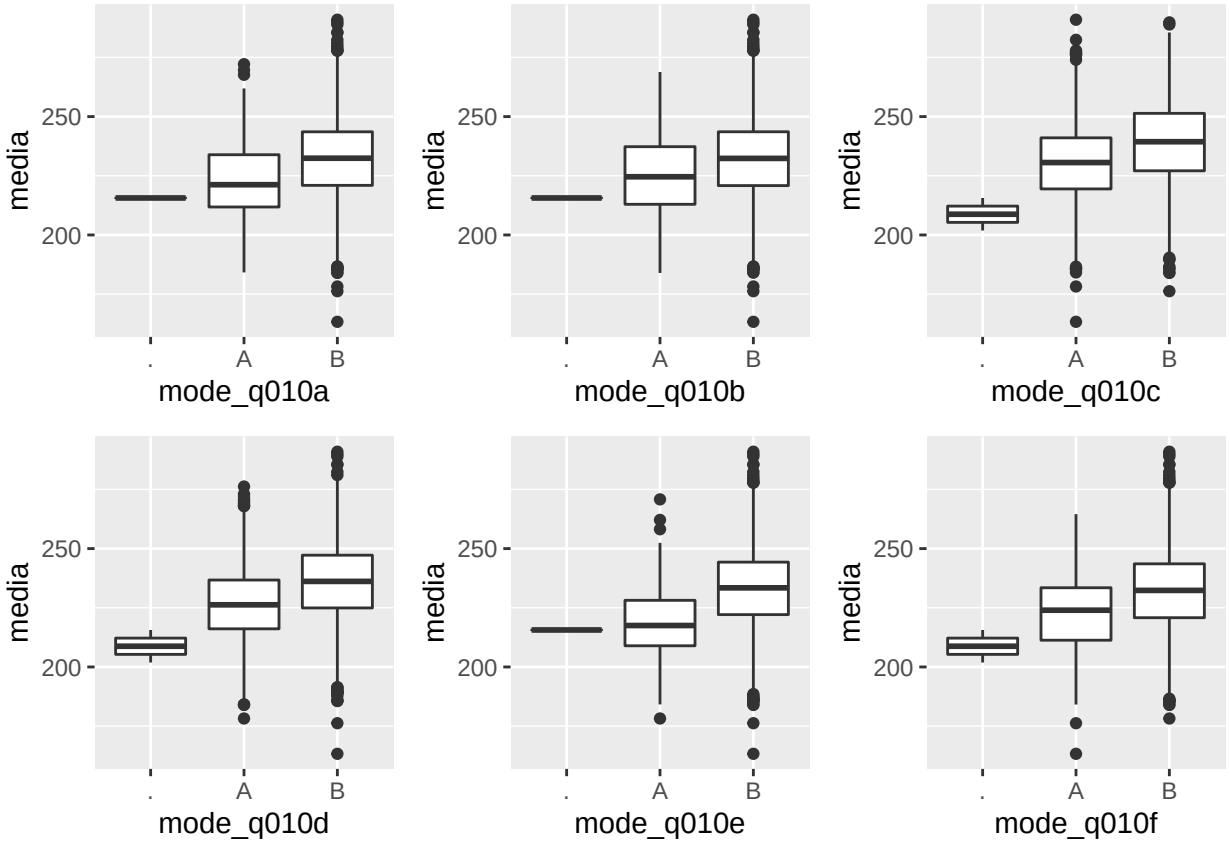
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

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
}
```

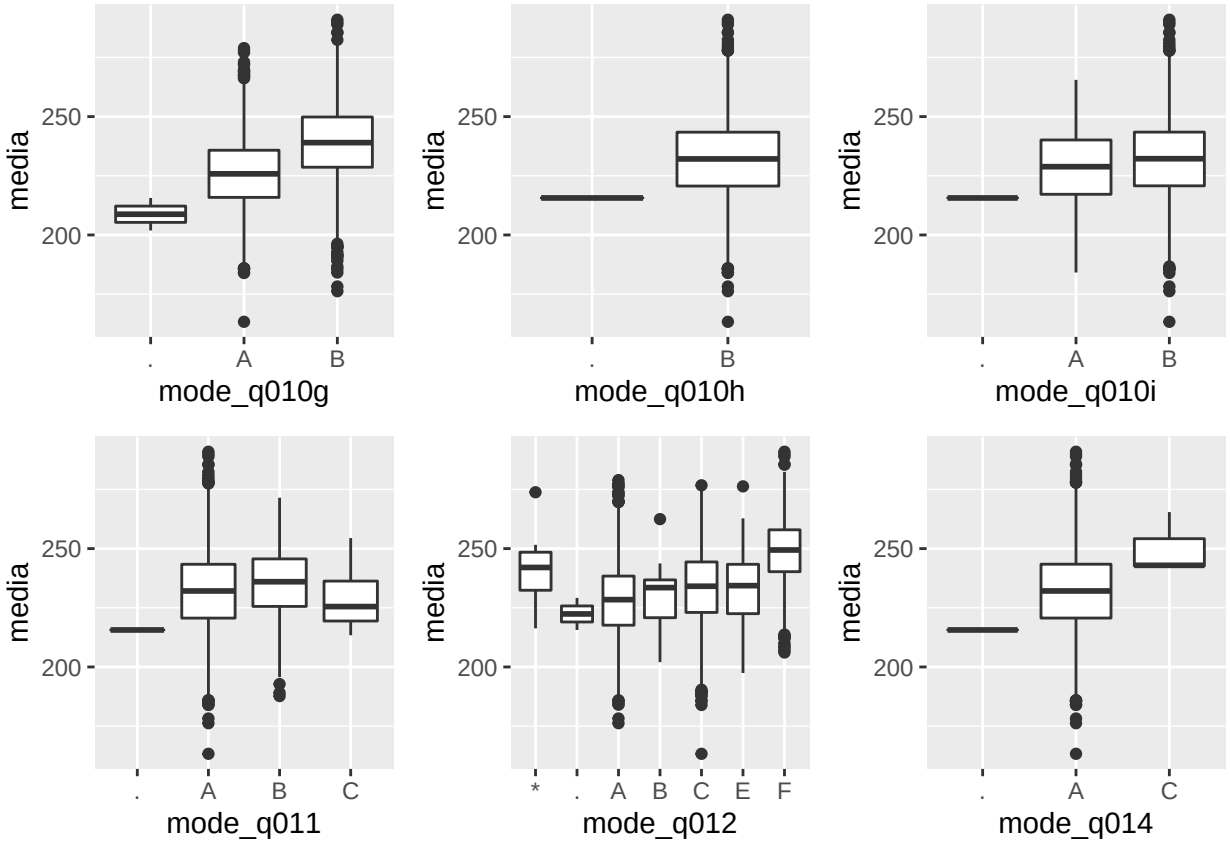

media
250
200
A
mode_q001
media
250
200
.
A
B
C
F
mode_q002
media
250
200
B
mode_q003a
media
250
200
.
A
B
mode_q003b
media
250
200
.
A
B
mode_q003c
media
250
200
.
A
B
mode_q003d

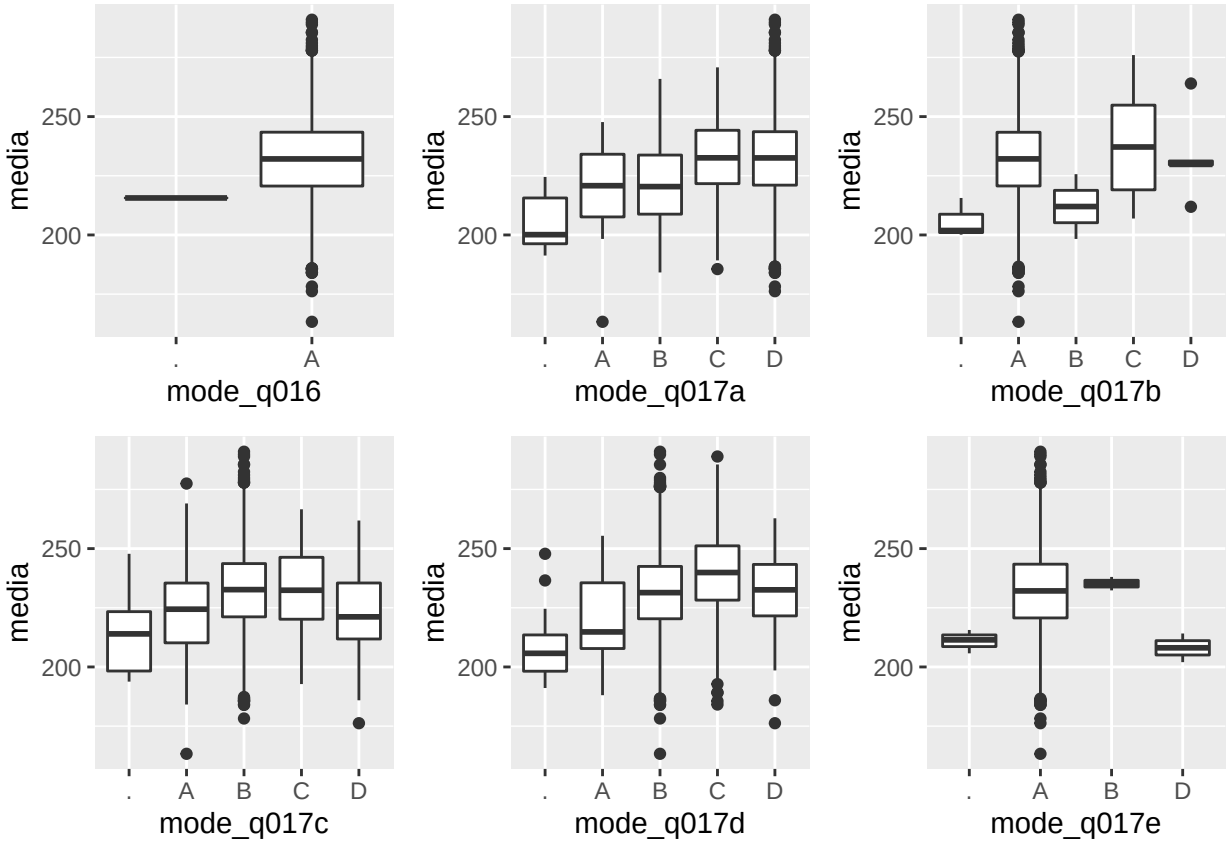

media
250
200
. A B
mode_q003e
media
250
200
. A B C D E F
mode_q004
media
250
200
. A B C D E F
mode_q005
media
250
200
* . A B C
mode_q006a
media
250
200
. B C
mode_q006b
media
250
200
. A B C
mode_q006c

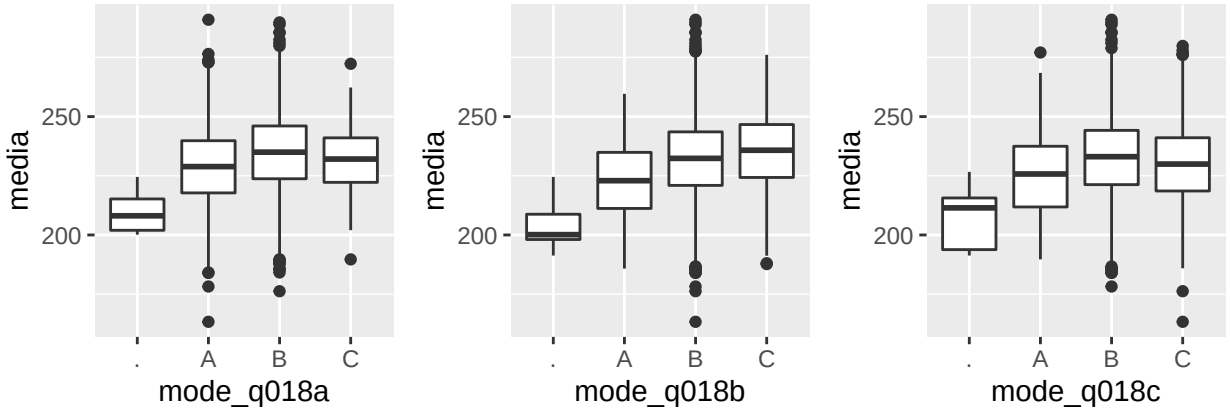

media
250
200
. B C
mode_q006d
media
250
200
. A B C
mode_q007
media
250
200
. A B
mode_q008a
media
250
200
. A B
mode_q008b
media
250
200
. A B
mode_q008c
media
250
200
. B C
mode_q009a


media
250
200
. A B C D
mode_q009b
media
250
200
. A B C D
mode_q009c
media
250
200
. B C D
mode_q009d
media
250
200
. B C D
mode_q009e
media
250
200
. B C D
mode_q009f
media
250
200
. A B C D
mode_q009g


media
250
200
.
A
B
mode_q010a
mode_q010b
mode_q010c
mode_q010d
mode_q010e
mode_q010f


media
250
200
. A B
mode_q010g
media
250
200
. B
mode_q010h
media
250
200
. A B
mode_q010i
media
250
200
. A B C
mode_q011
media
250
200
* . A B C E F
mode_q012
media
250
200
. A C
mode_q014


media
250
200
. A
mode_q016
media
250
200
. A B C D
mode_q017a
media
250
200
. A B C D
mode_q017b
media
250
200
. A B C D
mode_q017c
media
250
200
. A B C D
mode_q017d
media
250
200
. A B D
mode_q017e



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
```

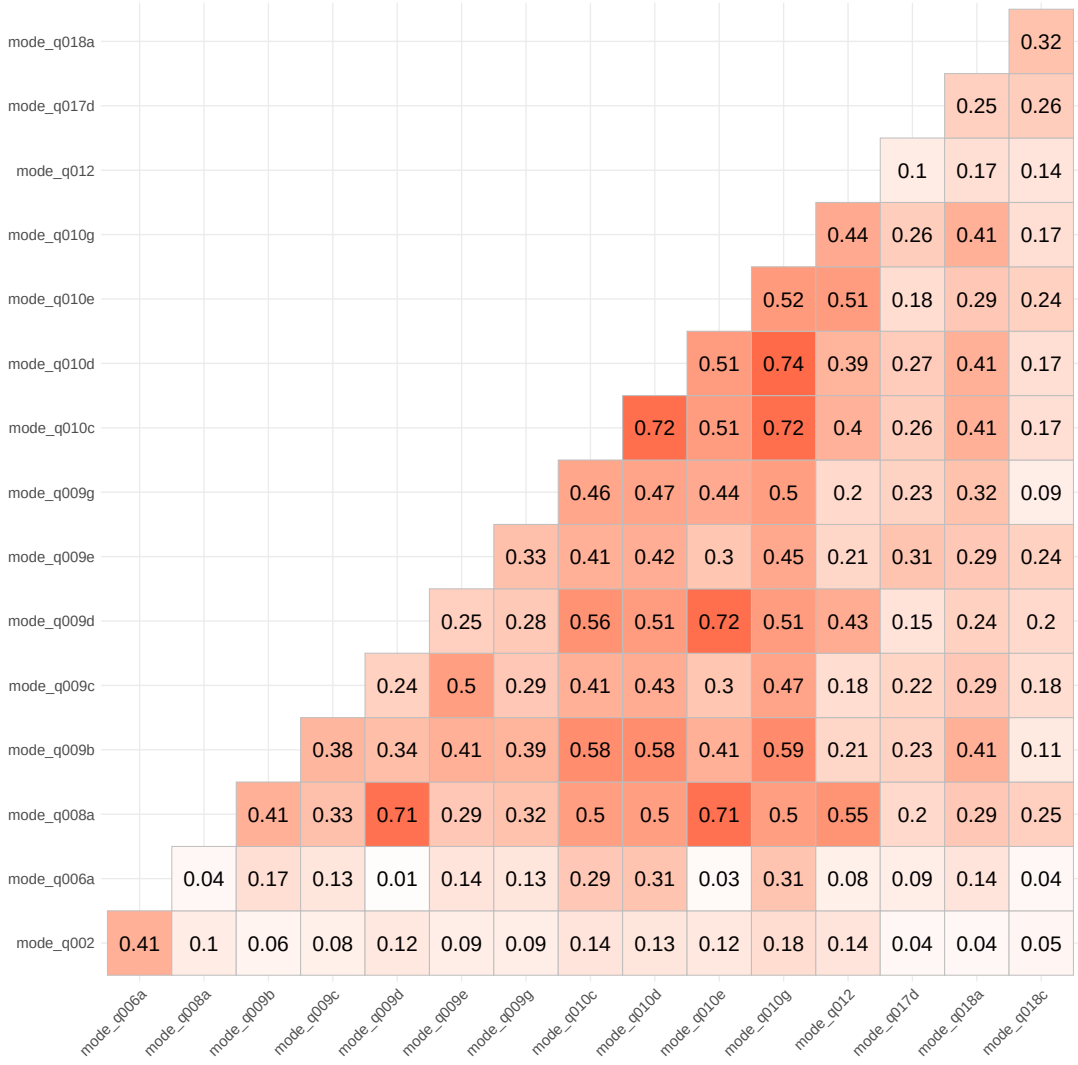
```
##  [1] "mode_q002"  "mode_q006a" "mode_q008a" "mode_q009b" "mode_q009c"
##  [6] "mode_q009d" "mode_q009e" "mode_q009g" "mode_q010c" "mode_q010d"
## [11] "mode_q010e" "mode_q010g" "mode_q012"  "mode_q017d" "mode_q018a"
## [16] "mode_q018c"
```

```
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_q002 0.062522069
## 2  mode_q006a 0.010957795
## 3  mode_q008a 0.008745463
## 4  mode_q009b 0.010104012
## 5  mode_q009c 0.088235288
## 6  mode_q009d 0.041654080
## 7  mode_q009e 0.051546087
## 8  mode_q009g 0.107866155
## 9  mode_q010c 0.044307188
## 10 mode_q010d 0.076031445
## 11 mode_q010e 0.056197622
## 12 mode_q010g 0.151463517
## 13  mode_q012 0.152800928
## 14 mode_q017d 0.027381297
## 15 mode_q018a 0.032481521
## 16 mode_q018c 0.011337705
```

```
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_q018a
mode_q017d
mode_q012
mode_q010g
mode_q010e
mode_q010d
mode_q010c
mode_q009g
mode_q009e
mode_q009d
mode_q009c
mode_q009b
mode_q008a
mode_q006a
mode_q002
0.32
0.25 0.26
0.1 0.17 0.14
0.44 0.26 0.41 0.17
0.52 0.51 0.18 0.29 0.24
0.51 0.74 0.39 0.27 0.41 0.17
0.72 0.51 0.72 0.4 0.26 0.41 0.17
0.46 0.47 0.44 0.5 0.2 0.23 0.32 0.09
0.33 0.41 0.42 0.3 0.45 0.21 0.31 0.29 0.24
0.25 0.28 0.56 0.51 0.72 0.51 0.43 0.15 0.24 0.2
0.24 0.5 0.29 0.41 0.43 0.3 0.47 0.18 0.22 0.29 0.18
0.38 0.34 0.41 0.39 0.58 0.58 0.41 0.59 0.21 0.23 0.41 0.11
0.41 0.33 0.71 0.29 0.32 0.5 0.5 0.71 0.5 0.55 0.2 0.29 0.25
0.04 0.17 0.13 0.01 0.14 0.13 0.29 0.31 0.03 0.31 0.08 0.09 0.14 0.04
0.41 0.1 0.06 0.08 0.12 0.09 0.09 0.14 0.13 0.12 0.18 0.14 0.04 0.04 0.05
mode_q006a
mode_q008a
mode_q009b
mode_q009c
mode_q009d
mode_q009e
mode_q009g
mode_q010c
mode_q010d
mode_q010e
mode_q010g
mode_q012
mode_q017d
mode_q018a
mode_q018c