Análise das variáveis Saeb - moda por escola

Série 3EM

Livia Kobayashi

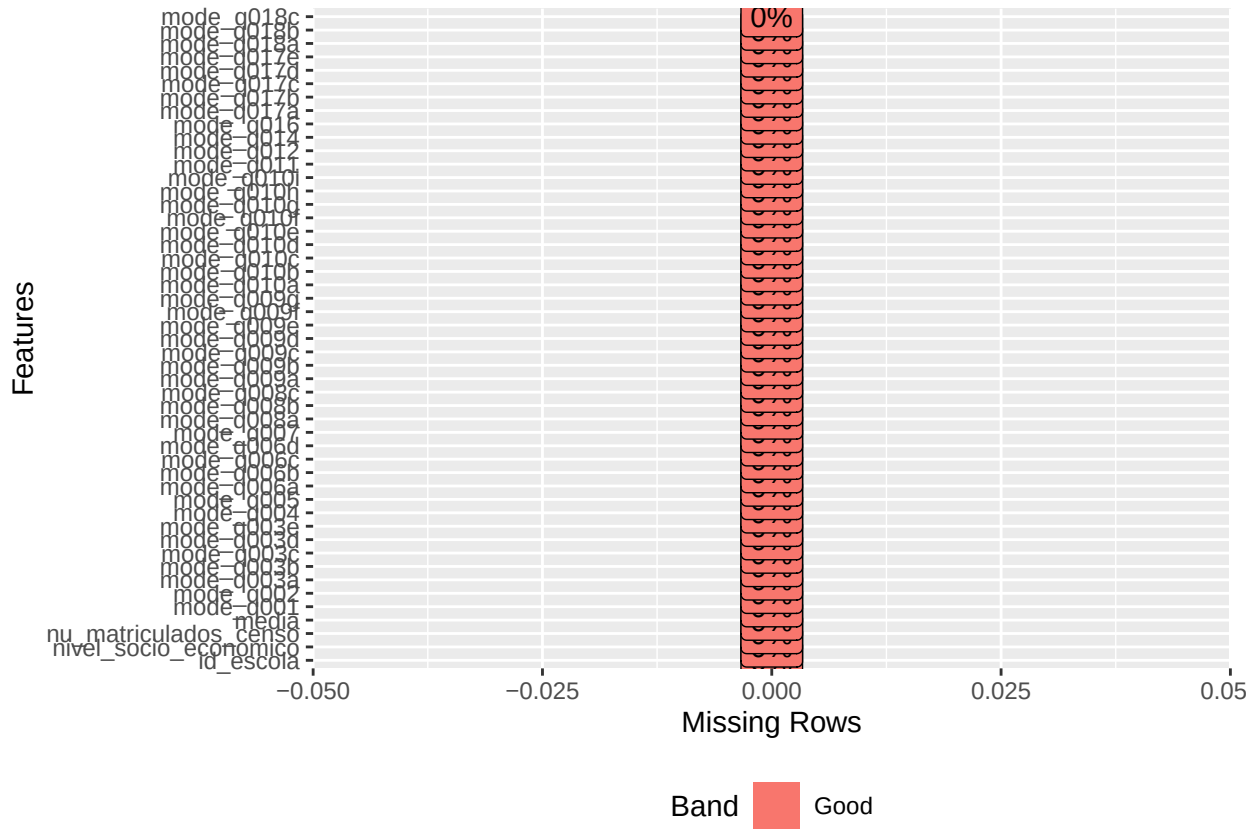
10 junho 2021

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

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

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

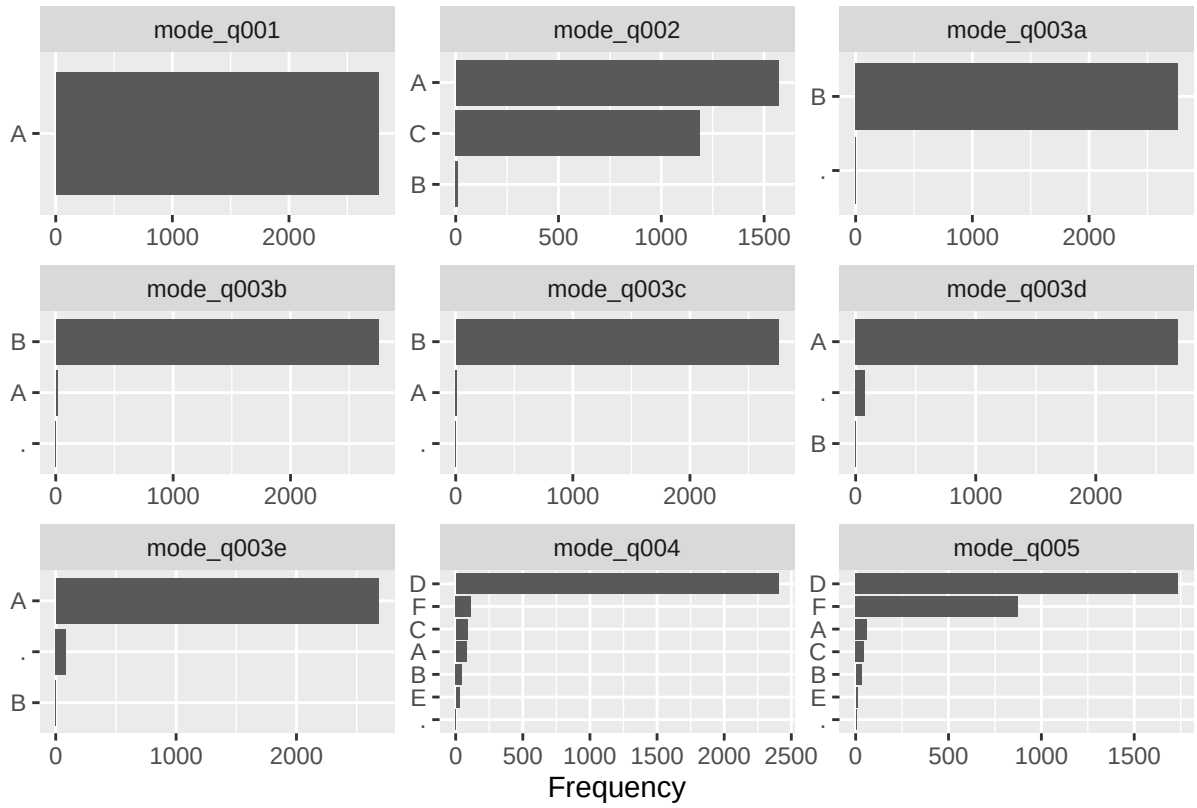
Missing

```
plot_missing(df)
```

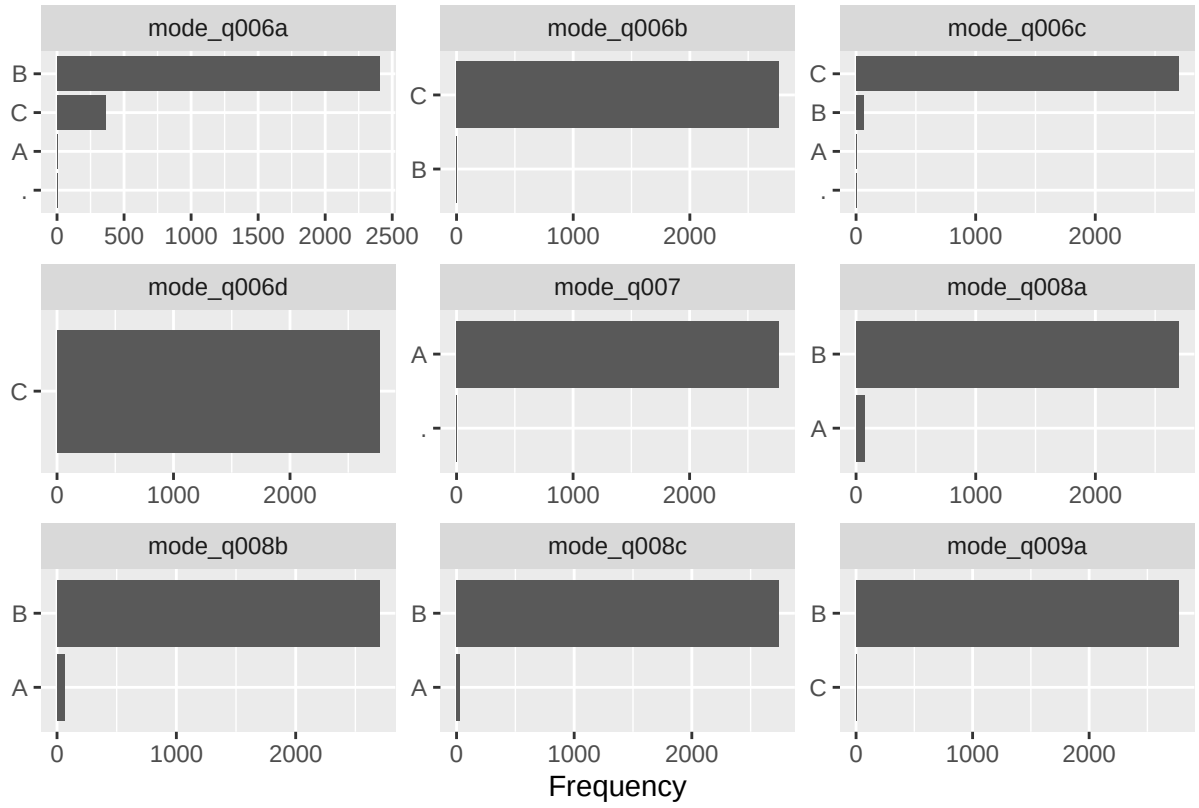


Volume

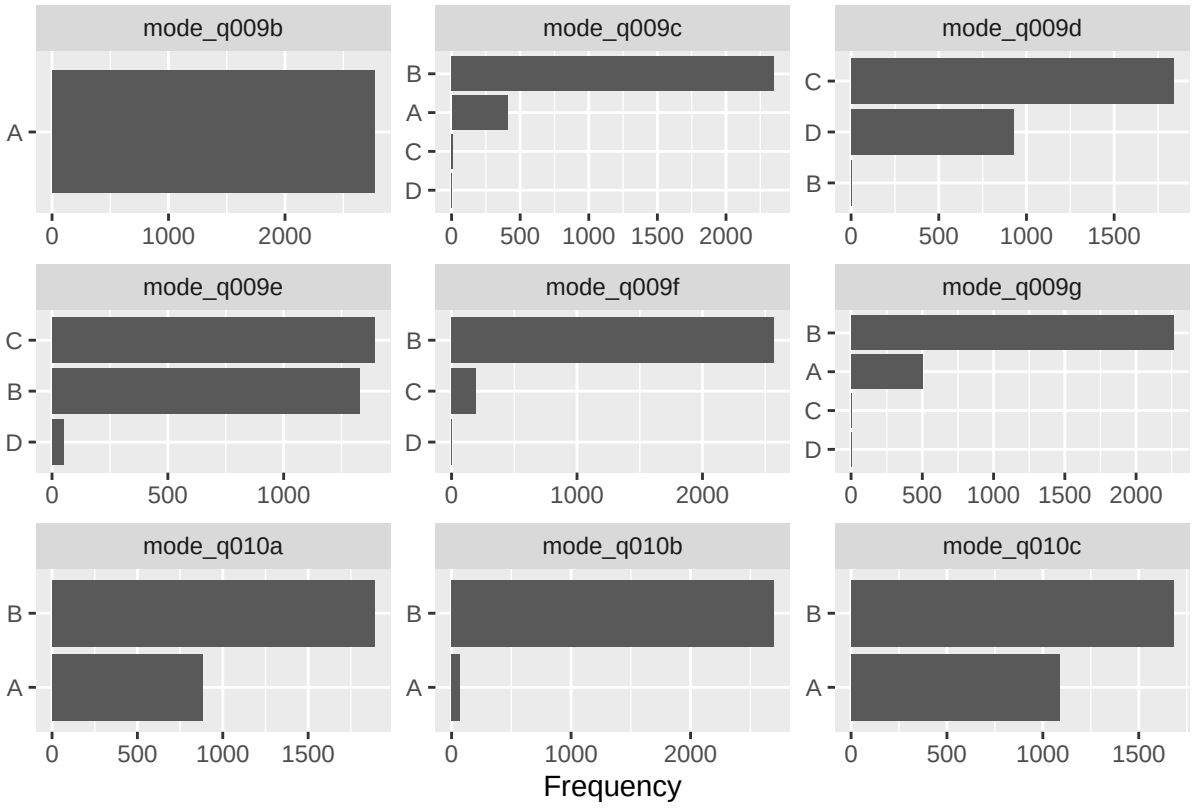
```
plot_bar(final_data)
```

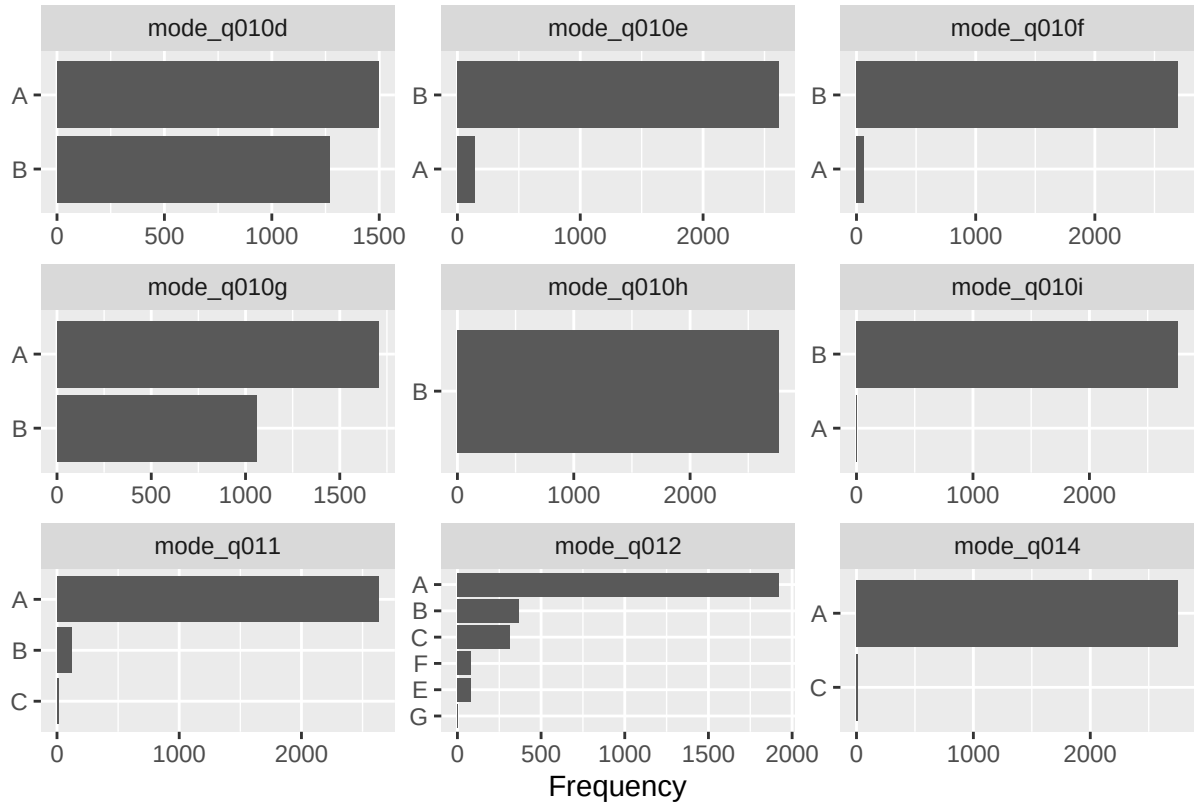


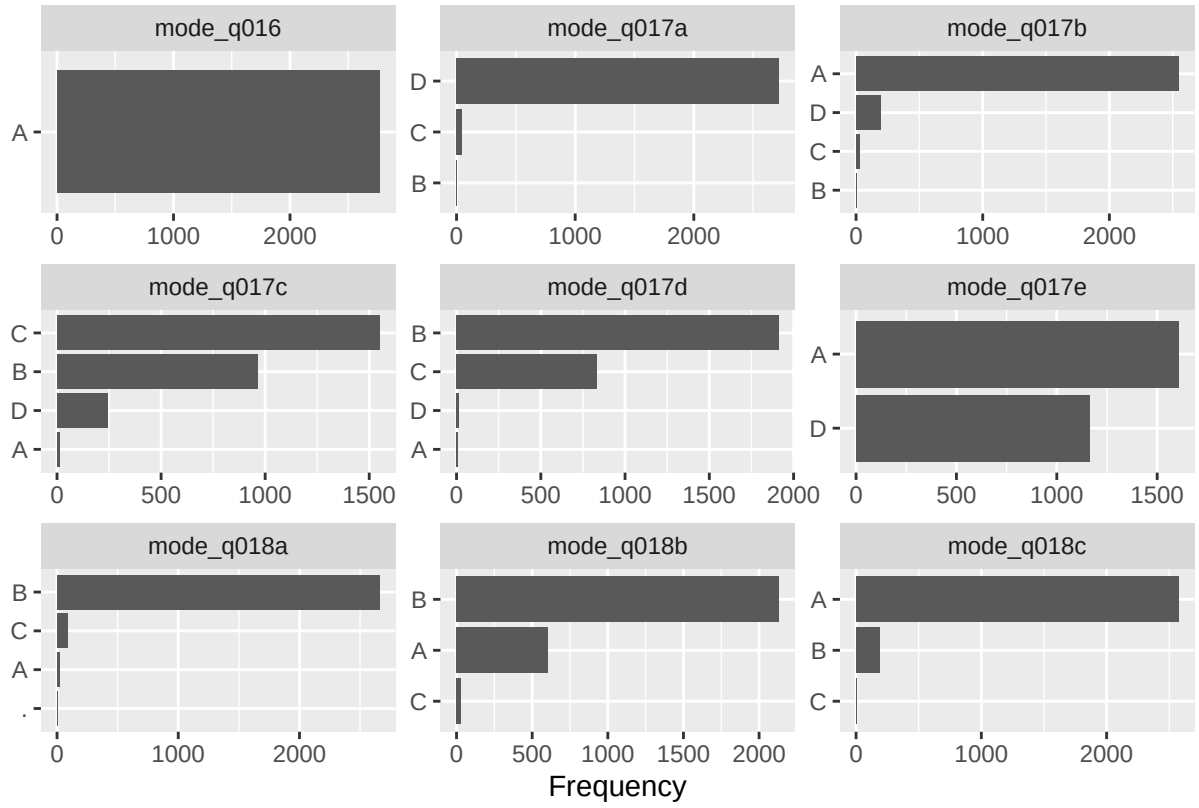
Page 1


mode_q006a
B
C
A
.
0 500 1000 1500 2000 2500
mode_q006b
C
B
0 1000 2000
mode_q006c
C
B
A
.
0 1000 2000
mode_q006d
C
0 1000 2000
mode_q007
A
.
0 1000 2000
mode_q008a
B
A
0 1000 2000
mode_q008b
B
A
0 1000 2000
mode_q008c
B
A
0 1000 2000
mode_q009a
B
C
0 1000 2000
Frequency

Page 2

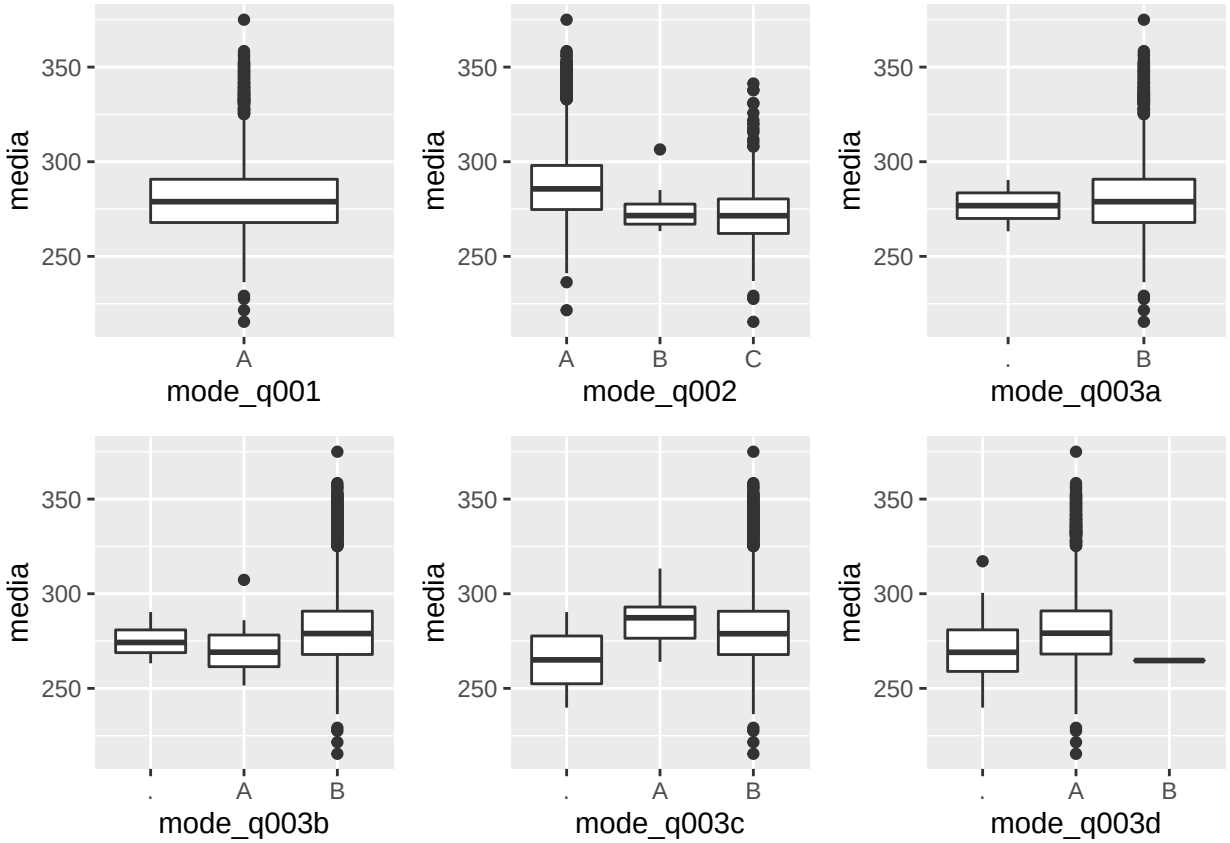

mode_q009b
A
0
1000
2000
mode_q009c
B
A
C
D
0
500
1000
1500
2000
mode_q009d
C
D
B
0
500
1000
1500
mode_q009e
C
B
D
0
500
1000
mode_q009f
B
C
D
0
1000
2000
mode_q009g
B
A
C
D
0
500
1000
1500
2000
mode_q010a
B
A
0
500
1000
1500
mode_q010b
B
A
0
1000
2000
mode_q010c
B
A
0
500
1000
1500
Frequency

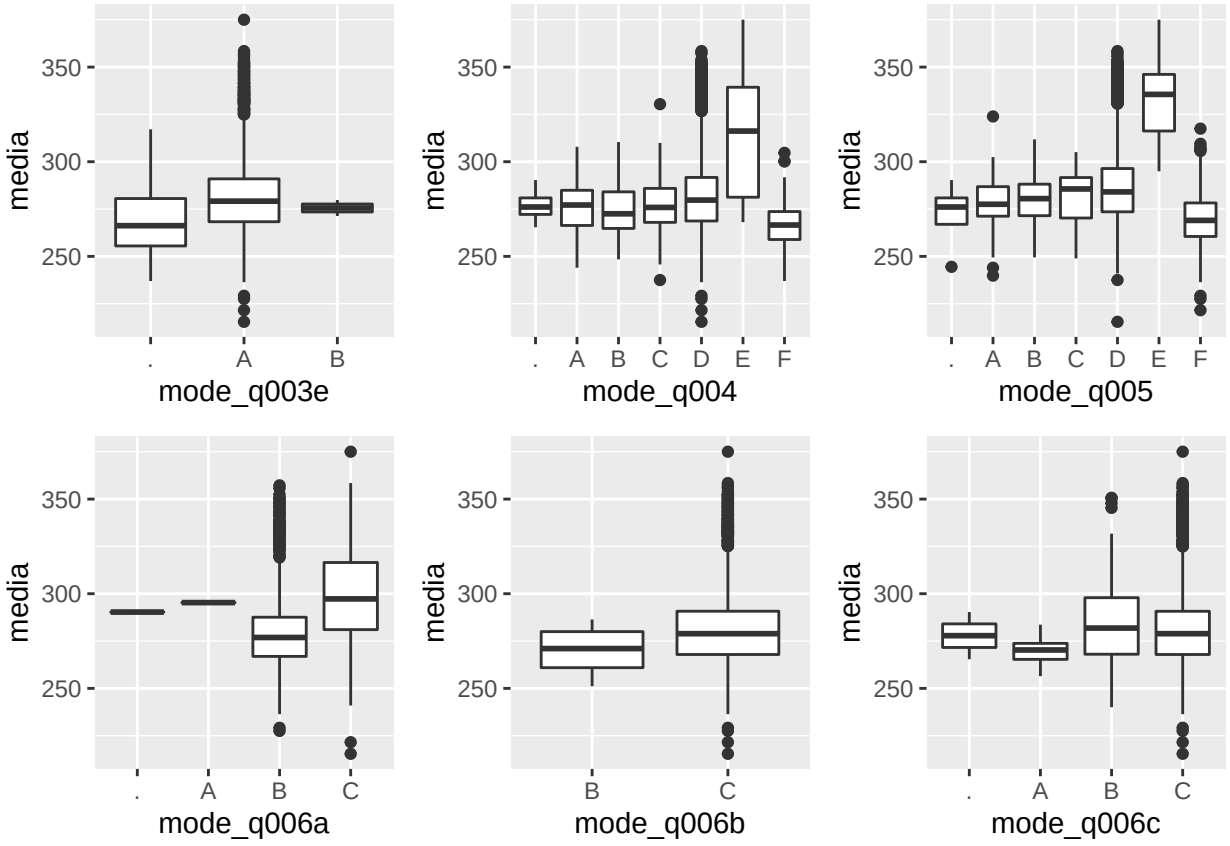

mode_q010d
A
B
0
500
1000
1500
mode_q010e
B
A
0
1000
2000
mode_q010f
B
A
0
1000
2000
mode_q010g
A
B
0
500
1000
1500
mode_q010h
B
0
1000
2000
mode_q010i
B
A
0
1000
2000
mode_q011
A
B
C
0
1000
2000
mode_q012
A
B
C
F
E
G
0
500
1000
1500
2000
mode_q014
A
C
0
1000
2000
Frequency

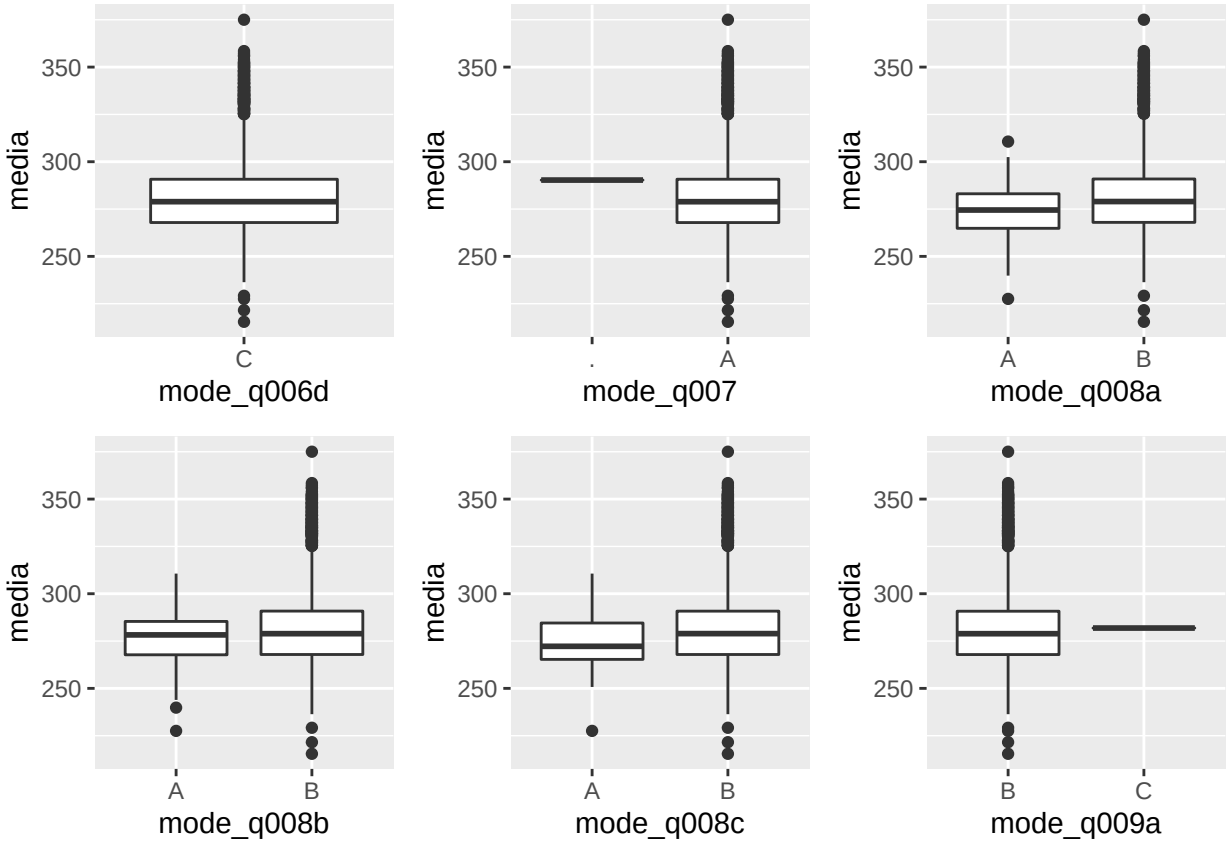


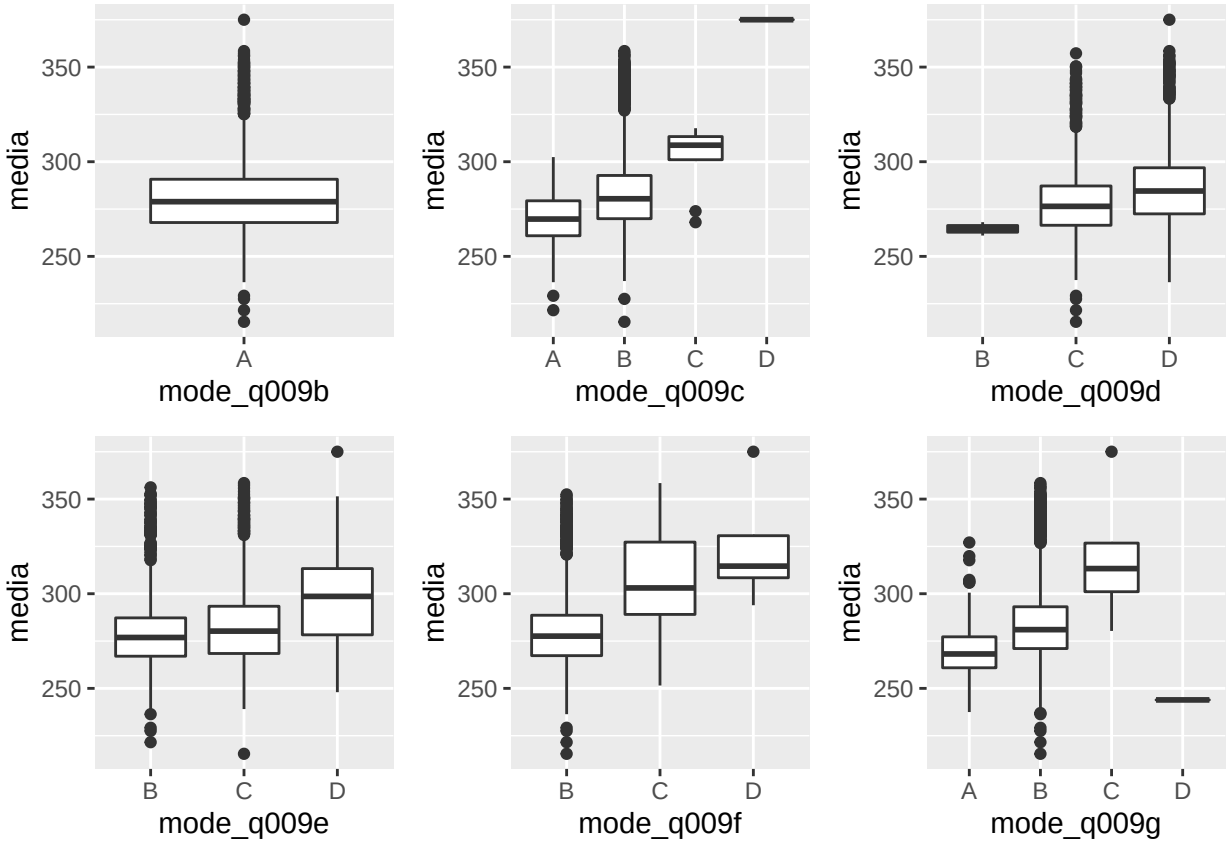
Boxplot

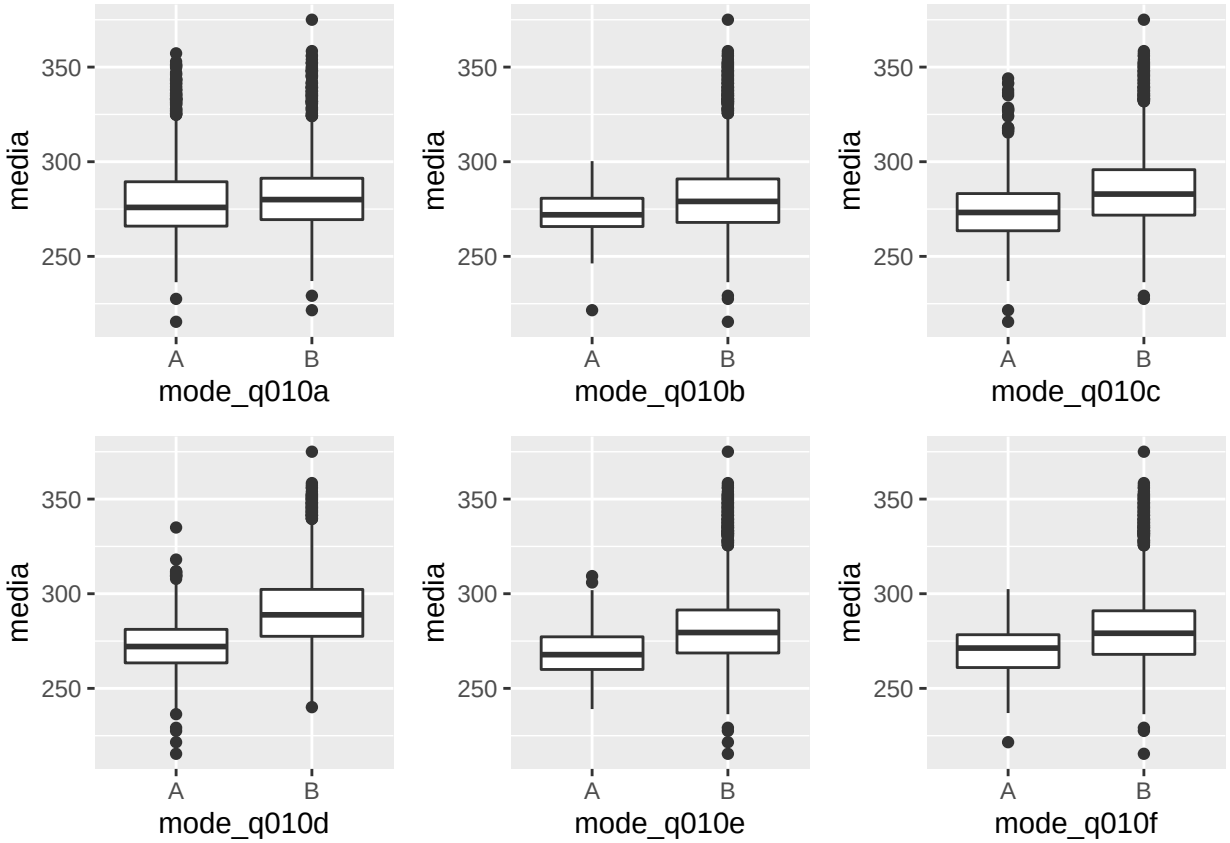
```
vars <- colnames(final_data)
vars <- vars[-c(1,2)]
plots <- list()
i <- 1
for (variable in vars) {
  #plots[[i]] <- plot_boxplot(final_data, by = variable)
  plots[[i]] <- ggplot(final_data, aes_string(variable, "media")) + geom_boxplot()
  i <- i + 1
}

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
}
```

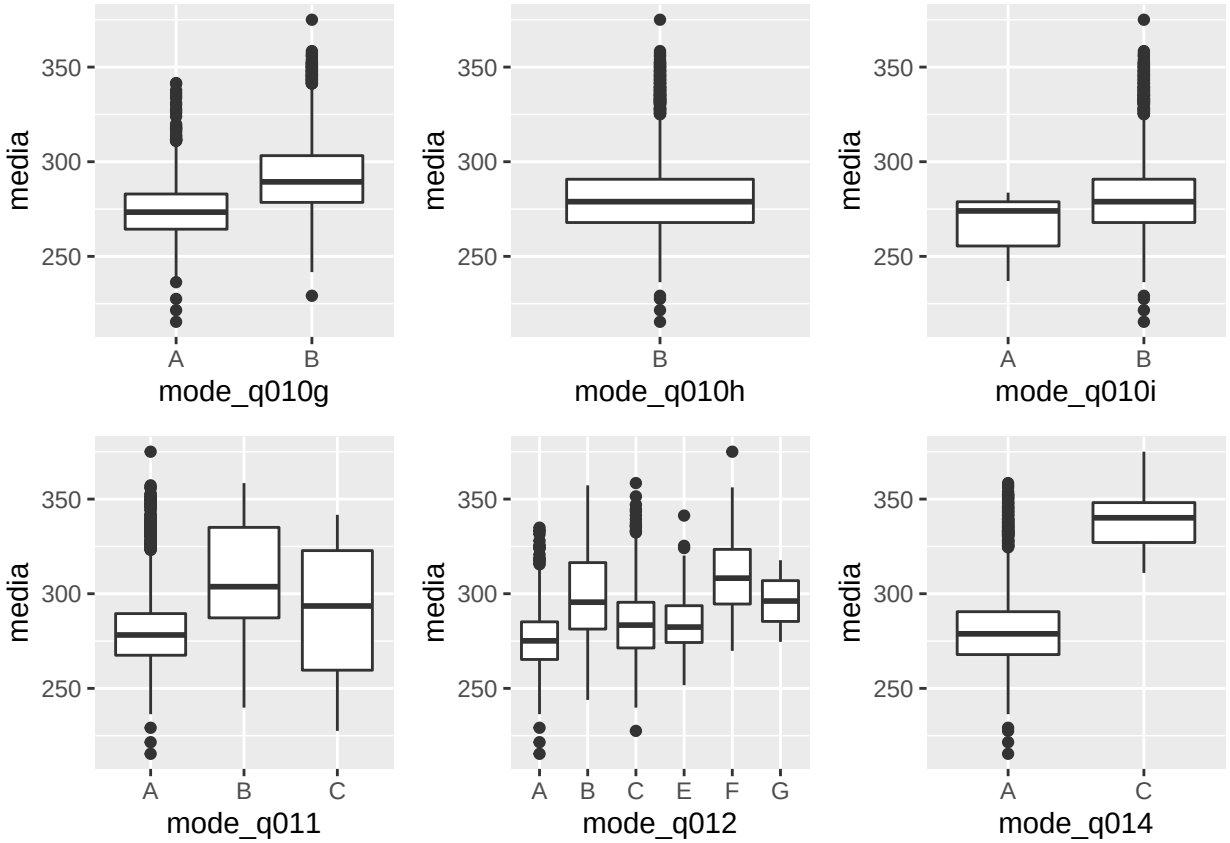

media
350
300
250
A
mode_q001
media
350
300
250
A
B
C
mode_q002
media
350
300
250
.
B
mode_q003a
media
350
300
250
.
A
B
mode_q003b
media
350
300
250
.
A
B
mode_q003c
media
350
300
250
.
A
B
mode_q003d

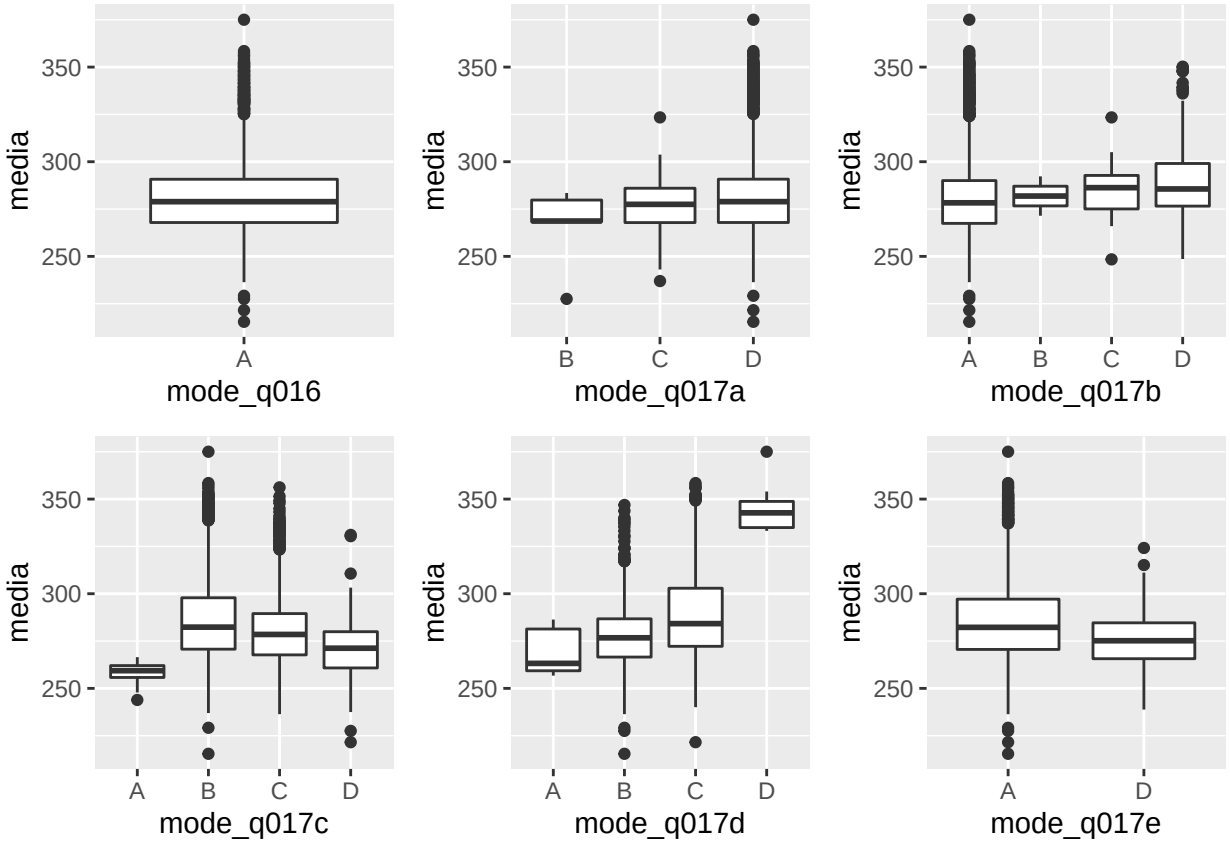

media
350
300
250
. A B
mode_q003e
media
350
300
250
. A B C D E F
mode_q004
media
350
300
250
. A B C D E F
mode_q005
media
350
300
250
. A B C
mode_q006a
media
350
300
250
B C
mode_q006b
media
350
300
250
. A B C
mode_q006c

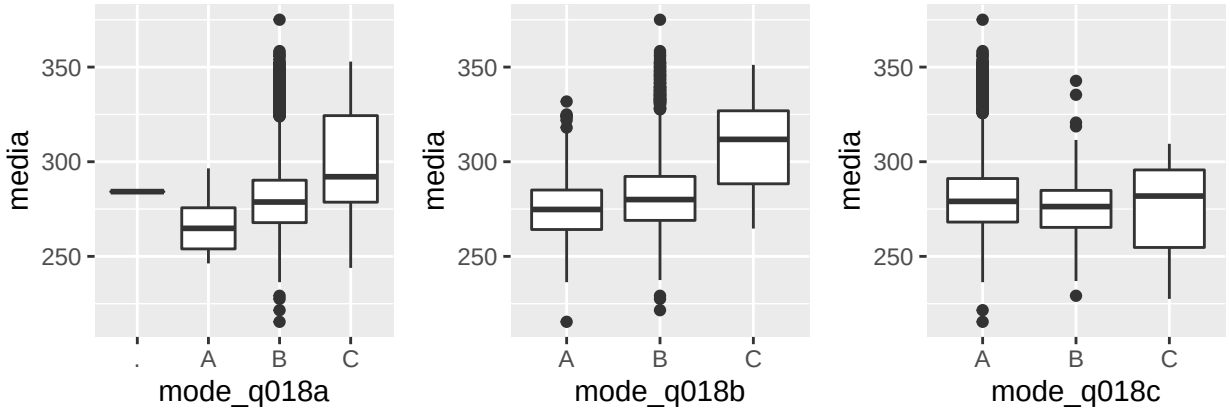

media
350
300
250
C
mode_q006d
.
A
mode_q007
A
B
mode_q008a
A
B
mode_q008b
A
B
mode_q008c
B
C
mode_q009a


media
350
300
250
A
mode_q009b
media
350
300
250
A
B
C
D
mode_q009c
media
350
300
250
B
C
D
mode_q009d
media
350
300
250
B
C
D
mode_q009e
media
350
300
250
B
C
D
mode_q009f
media
350
300
250
A
B
C
D
mode_q009g


media
350
300
250
A
B
mode_q010a
mode_q010b
mode_q010c
mode_q010d
mode_q010e
mode_q010f


media
350
300
250
A
B
mode_q010g
media
350
300
250
B
mode_q010h
media
350
300
250
A
B
mode_q010i
media
350
300
250
A
B
C
mode_q011
media
350
300
250
A
B
C
E
F
G
mode_q012
media
350
300
250
A
C
mode_q014


media
350
300
250
A
mode_q016
B
C
D
mode_q017a
A
B
C
D
mode_q017b
mode_q017c
mode_q017d
A
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
```

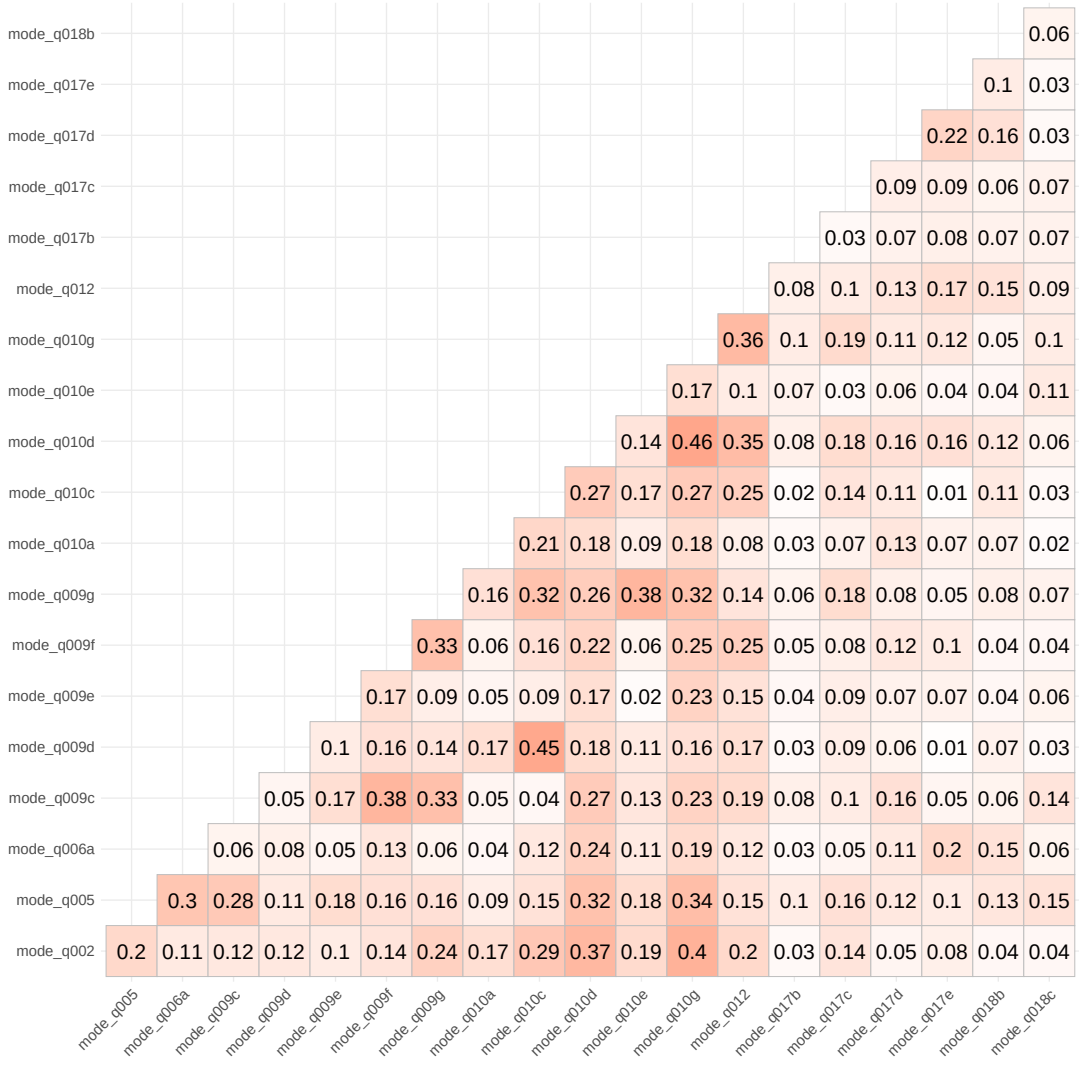
```
##  [1] "mode_q002"  "mode_q005"  "mode_q006a" "mode_q009c" "mode_q009d"
##  [6] "mode_q009e" "mode_q009f" "mode_q009g" "mode_q010a" "mode_q010c"
## [11] "mode_q010d" "mode_q010e" "mode_q010g" "mode_q012"  "mode_q017b"
## [16] "mode_q017c" "mode_q017d" "mode_q017e" "mode_q018b" "mode_q018c"
```

```
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 0.172024763
## 2   mode_q005 0.189264059
## 3  mode_q006a 0.118279384
## 4  mode_q009c 0.067153183
## 5  mode_q009d 0.041264240
## 6  mode_q009e 0.021960999
## 7  mode_q009f 0.133583996
## 8  mode_q009g 0.081161650
## 9  mode_q010a 0.001456487
## 10 mode_q010c 0.085155696
## 11 mode_q010d 0.238139638
## 12 mode_q010e 0.021976558
## 13 mode_q010g 0.200047050
## 14  mode_q012 0.244322771
## 15 mode_q017b 0.013469196
## 16 mode_q017c 0.055194980
## 17 mode_q017d 0.131416633
## 18 mode_q017e 0.067819502
## 19 mode_q018b 0.044652873
## 20 mode_q018c 0.005859695
```

```
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_q018b 0.06
mode_q017e 0.1 0.03
mode_q017d 0.22 0.16 0.03
mode_q017c 0.09 0.09 0.06 0.07
mode_q017b 0.03 0.07 0.08 0.07 0.07
mode_q012 0.08 0.1 0.13 0.17 0.15 0.09
mode_q010g 0.36 0.1 0.19 0.11 0.12 0.05 0.1
mode_q010e 0.17 0.1 0.07 0.03 0.06 0.04 0.04 0.11
mode_q010d 0.14 0.46 0.35 0.08 0.18 0.16 0.16 0.12 0.06
mode_q010c 0.27 0.17 0.27 0.25 0.02 0.14 0.11 0.01 0.11 0.03
mode_q010a 0.21 0.18 0.09 0.18 0.08 0.03 0.07 0.13 0.07 0.07 0.02
mode_q009g 0.16 0.32 0.26 0.38 0.32 0.14 0.06 0.18 0.08 0.05 0.08 0.07
mode_q009f 0.33 0.06 0.16 0.22 0.06 0.25 0.25 0.05 0.08 0.12 0.1 0.04 0.04
mode_q009e 0.17 0.09 0.05 0.09 0.17 0.02 0.23 0.15 0.04 0.09 0.07 0.07 0.04 0.06
mode_q009d 0.1 0.16 0.14 0.17 0.45 0.18 0.11 0.16 0.17 0.03 0.09 0.06 0.01 0.07 0.03
mode_q009c 0.05 0.17 0.38 0.33 0.05 0.04 0.27 0.13 0.23 0.19 0.08 0.1 0.16 0.05 0.06 0.14
mode_q006a 0.06 0.08 0.05 0.13 0.06 0.04 0.12 0.24 0.11 0.19 0.12 0.03 0.05 0.11 0.2 0.15 0.06
mode_q005 0.3 0.28 0.11 0.18 0.16 0.16 0.09 0.15 0.32 0.18 0.34 0.15 0.1 0.16 0.12 0.1 0.13 0.15
mode_q002 0.2 0.11 0.12 0.12 0.1 0.14 0.24 0.17 0.29 0.37 0.19 0.4 0.2 0.03 0.14 0.05 0.08 0.04 0.04
mode_q005 mode_q006a mode_q009c mode_q009d mode_q009e mode_q009f mode_q009g mode_q010a mode_q010c mode_q010d mode_q010e mode_q010g mode_q012 mode_q017b mode_q017c mode_q017d mode_q017e mode_q018b mode_q018c