Education data saber 11

Ignacio Almodóvar Cárdenas & Andrés Mejía Rodríguez

10/10/2021

```
saber11=read_delim("~/Downloads/SB11_20211.txt",delim="¬")
```

```
## Rows: 15528 Columns: 78
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "¬"
## chr (58): ESTU_TIPODOCUMENTO, ESTU_NACIONALIDAD, ESTU_GENERO, ESTU_FECHANACI...
## dbl (20): PERIODO, COLE_COD_DANE_ESTABLECIMIENTO, COLE_COD_DANE_SEDE, PUNT_L...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#saber11 %<>% filter(ESTU_MCPIO_PRESENTACION=="PALMIRA")

saber11 %<>% select(ESTU_CONSECUTIVO,FAMI_EDUCACIONMADRE,FAMI_ESTRATOVIVIENDA,starts_with("PUNT"))


saber11$FAMI_EDUCACIONMADRE %<>% factor( levels= c(
"No sabe",
"No Aplica",
"Ninguno",
"Primaria incompleta",
"Primaria completa",
"Secundaria (Bachillerato) incompleta",
"Secundaria (Bachillerato) completa" ,
"Técnica o tecnológica incompleta",
"Educación profesional incompleta",
"Técnica o tecnológica completa",
"Educación profesional completa",
"Postgrado"
  ),
ordered = TRUE  )
```

Introduction

Data comes from the results of Colombian selectivity exam "Saber 11", this exam is taken by students who are about to end their secondary education and their results are used by universites and other terciary education centers to evaluate admision to undergraduate programs.

In addition to the results of the exams in reading, mathematics, science, social science and english a socio-demograpgic survey of the living condition of the student is included. We will only include the stratum of the student's home and the mother's education, leaving the analysis of other variables as future work.

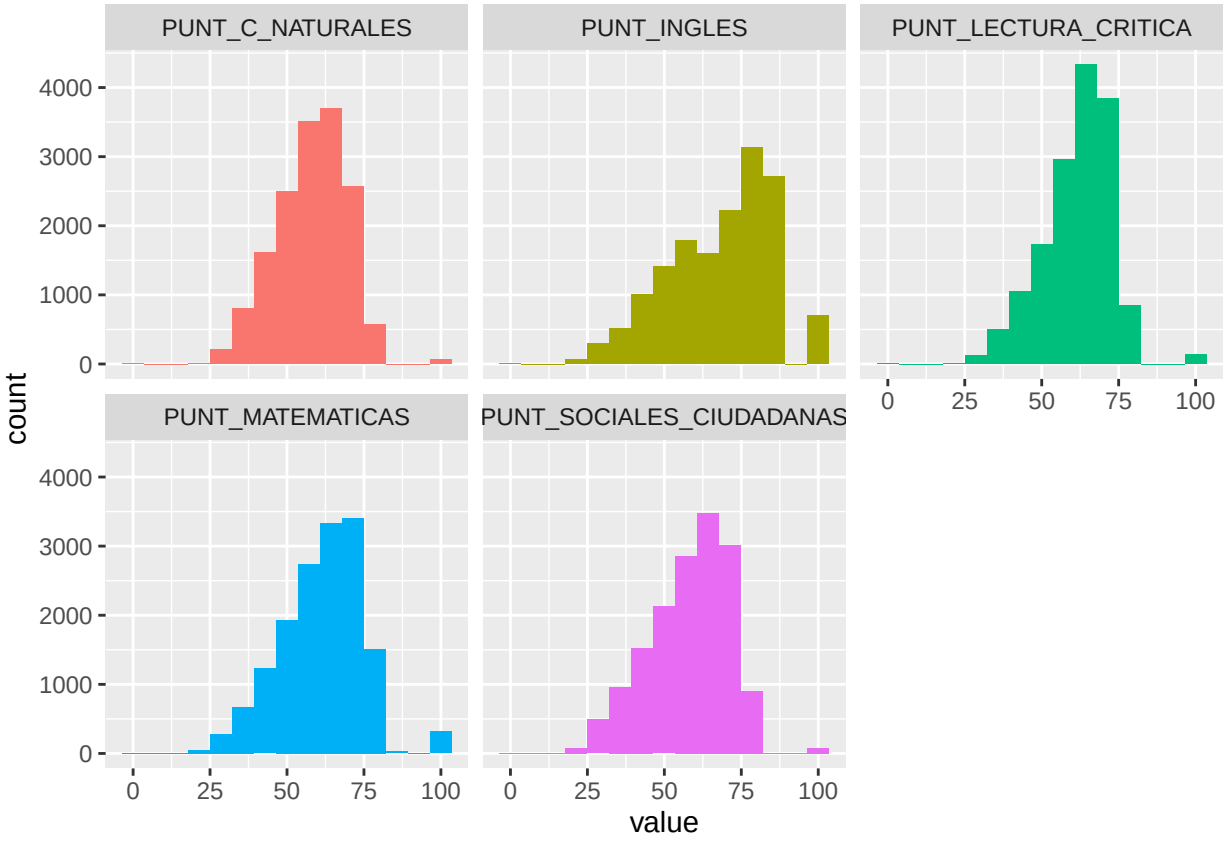
Descriptive study of the variables

Exam scores

The saber 11 exam is divided into five tests: Reading, Mathematics, Social Science, Natural Science and English. Each test is scored in a scale ranging from 0 to 100.

```
puntajes=saber11 %>%  select(starts_with("PUNT")) %>%  select(-PUNT_GLOBAL)
puntajes %<>% pivot_longer(cols = starts_with("PUNT"))



ggplot(puntajes,aes(x=value))+geom_histogram(aes(fill=name),bins=15)+
  facet_wrap(vars(name),nrow=2)+theme(legend.position = "none")
```

```
## Warning: Removed 47 rows containing non-finite values (stat_bin).
```



From the plots we suspect that the scores on each part of the exam follow a normal distribution, this is confirmed as we calculate basic descriptive statistics.

```
puntajes %>%
    group_by(name) %>%
    summarise(mean = mean(value, na.rm = T), sd = sqrt(var(value, na.rm = T)), kurtosis = kurtosis(valu
        na.rm = TRUE), skewness = skewness(value, na.rm = TRUE))
```

```
## # A tibble: 5 x 5
##   name                    mean    sd kurtosis skewness
##   <chr>                  <dbl> <dbl>    <dbl>    <dbl>
## 1 PUNT_C_NATURALES        58.0  11.4   0.156   -0.148
## 2 PUNT_INGLES             68.4  17.0  -0.457   -0.398
```

```
## 3 PUNT_LECTURA_CRITICA      61.5  11.1   0.865   -0.283
## 4 PUNT_MATEMATICAS          61.2  13.5   0.436   -0.0695
## 5 PUNT_SOCIALES_CIUDADANAS  58.1  13.0  -0.0750  -0.324
```

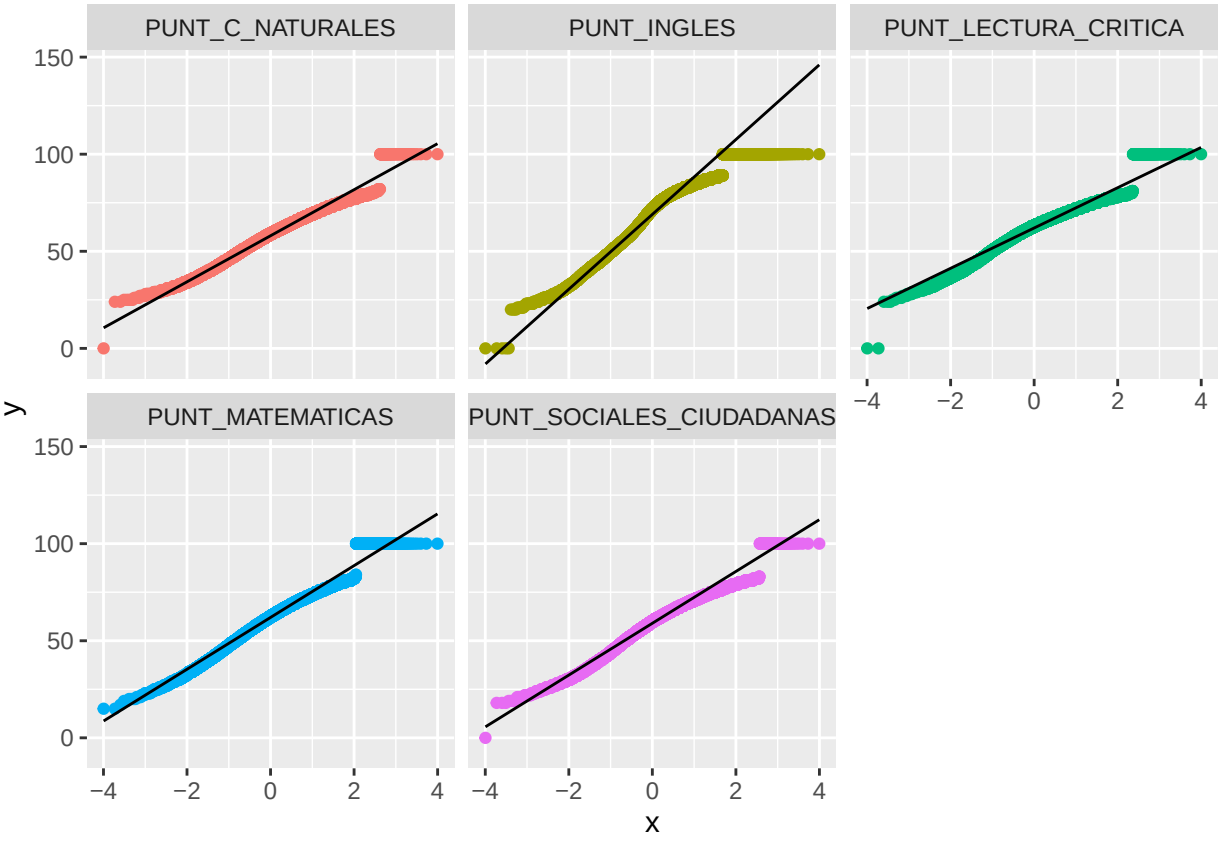
Skewness shows that reading, english and social sciences are sligthly slanted to the right while math and science to the right. Reading, math and science seem to be slightly heavy tailed and social sciences has a light tail. However [1] and other authors suggest a guideline that values of both skewness and kurtosis between -1 and 1 are acceptable when dealing with a normal distribution.

QQ plots show also the nomality of most of the variables, note that the outliers seem to impact the kurtosis of the variables.

```
ggplot(puntajes,aes(sample=value))+geom_qq(aes(color=name))+
  stat_qq_line()+facet_wrap(vars(name),nrow=2)+theme(legend.position = "none")
```

```
## Warning: Removed 47 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 47 rows containing non-finite values (stat_qq_line).
```

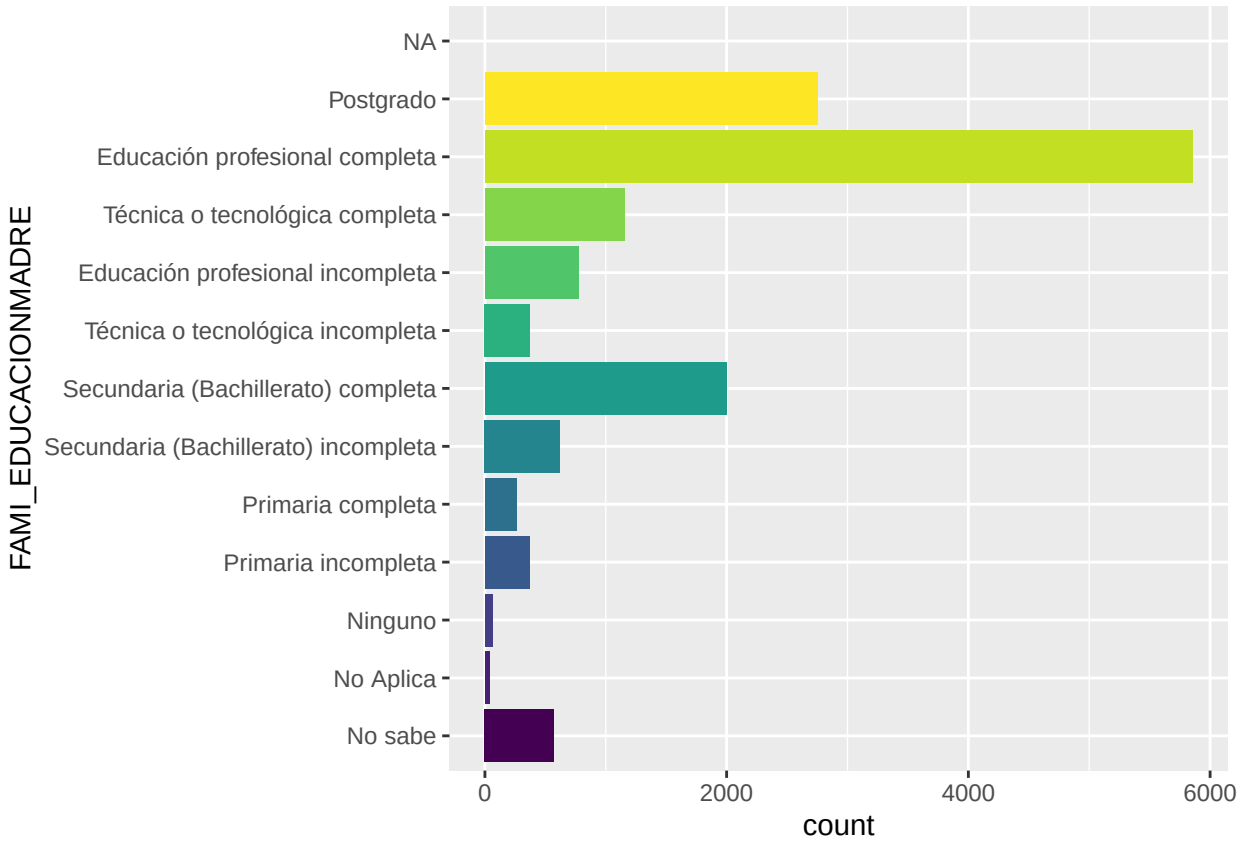


Socio demographic variables

We include two variables that reflect the socioacademic background of the student, education of the mother, and stratum of the student's home.

```
ggplot(saber11,aes(y=FAMI_EDUCACIONMADRE,fill=FAMI_EDUCACIONMADRE))+geom_bar()+
  theme(legend.position = "none")
```

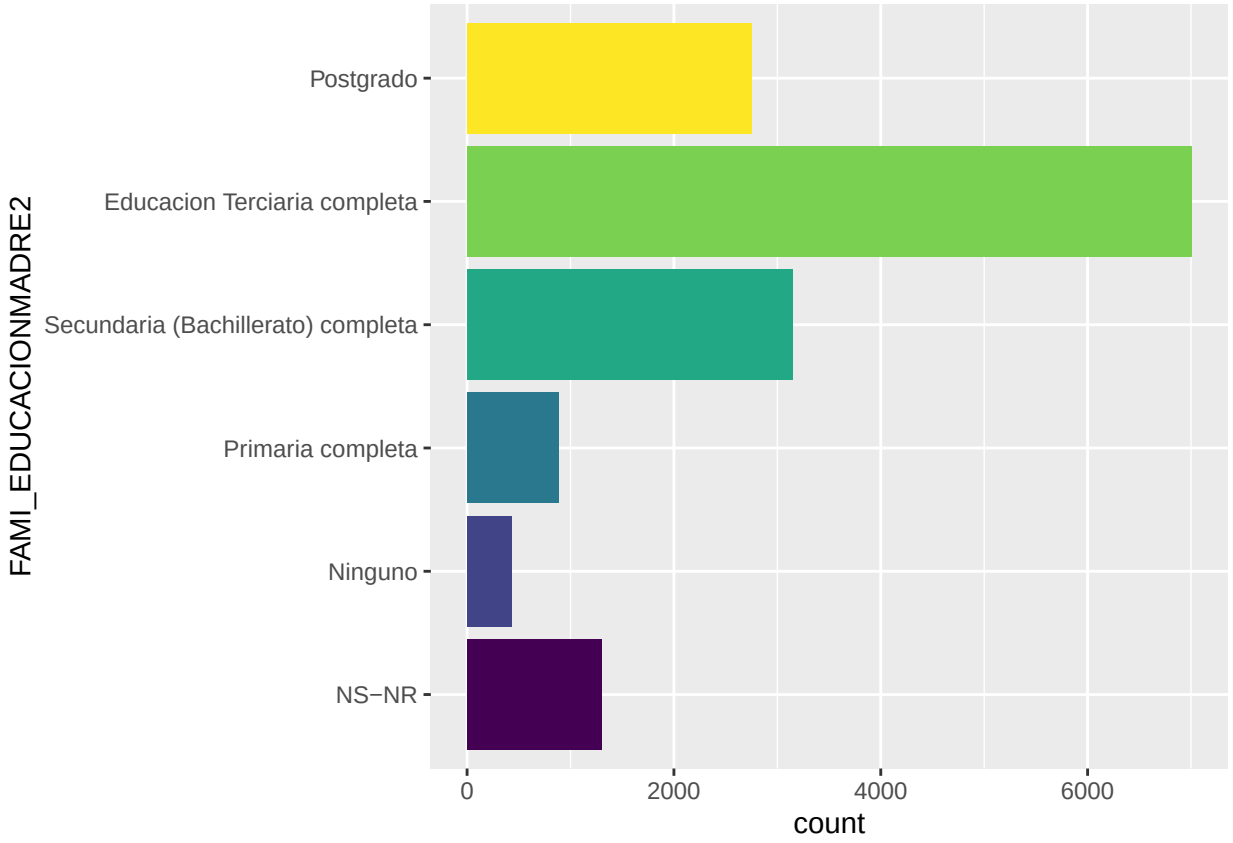
Mother's education level



This data uses the highest level of education achieved by the student's mother. As there are a lot of categories and some seem to fit well together, for example joining complete primary education and incomplete secondary education.

```
saber11 %<>%  mutate(FAMI_EDUCACIONMADRE2=case_when(
  FAMI_EDUCACIONMADRE=="No sabe"~"NS-NR",
  FAMI_EDUCACIONMADRE=="No Aplica"~"NS-NR",
  FAMI_EDUCACIONMADRE=="Ninguno"~"Ninguno",
  FAMI_EDUCACIONMADRE=="Primaria incompleta"~"Ninguno",
  FAMI_EDUCACIONMADRE=="Primaria completa"~"Primaria completa",
  FAMI_EDUCACIONMADRE=="Secundaria (Bachillerato) incompleta"~"Primaria completa",
  FAMI_EDUCACIONMADRE=="Secundaria (Bachillerato) completa"~"Secundaria (Bachillerato) completa" ,
  FAMI_EDUCACIONMADRE=="Técnica o tecnológica incompleta"~"Secundaria (Bachillerato) completa" ,
  FAMI_EDUCACIONMADRE=="Educación profesional incompleta"~"Secundaria (Bachillerato) completa" ,
  FAMI_EDUCACIONMADRE=="Técnica o tecnológica completa"~"Educacion Terciaria completa",
  FAMI_EDUCACIONMADRE=="Educación profesional completa"~"Educacion Terciaria completa",
  FAMI_EDUCACIONMADRE=="Postgrado"~"Postgrado"  ,
  is.na(FAMI_EDUCACIONMADRE)~"NS-NR"
))


saber11$FAMI_EDUCACIONMADRE2 %<>% factor( levels= c(
"NS-NR",
"Ninguno",
"Primaria completa",
"Secundaria (Bachillerato) completa" ,
"Educacion Terciaria completa",
"Postgrado"
```

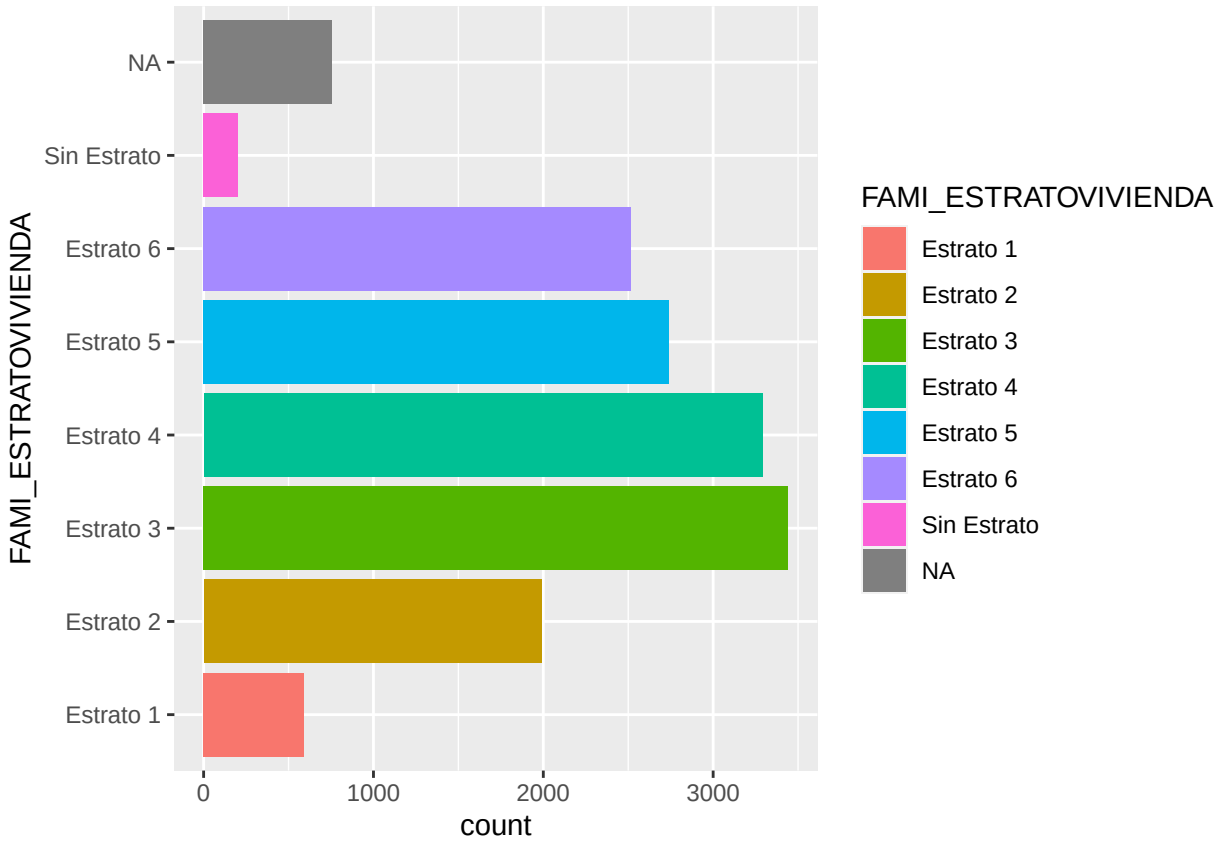
```
  ),
ordered = TRUE  )


ggplot(saber11,aes(y=FAMI_EDUCACIONMADRE2,fill=FAMI_EDUCACIONMADRE2))+geom_bar()+
  theme(legend.position = "none")
```



Socio-economic stratum

The stratum is a number ranging from 1 to 6 and based on physical characteristics of the structure (such as material used, repairs needed, access to public utilities)[2], that number is also often asociated with to the people living in that space. Colombians who are in a similar socioeconomic condition, are assigned by the same number. The higher this number is, the better their socio economical situation is. Categorical variable

```
ggplot(saber11,aes(y=FAMI_ESTRATOVIVIENDA,fill=FAMI_ESTRATOVIVIENDA))+geom_bar()
```

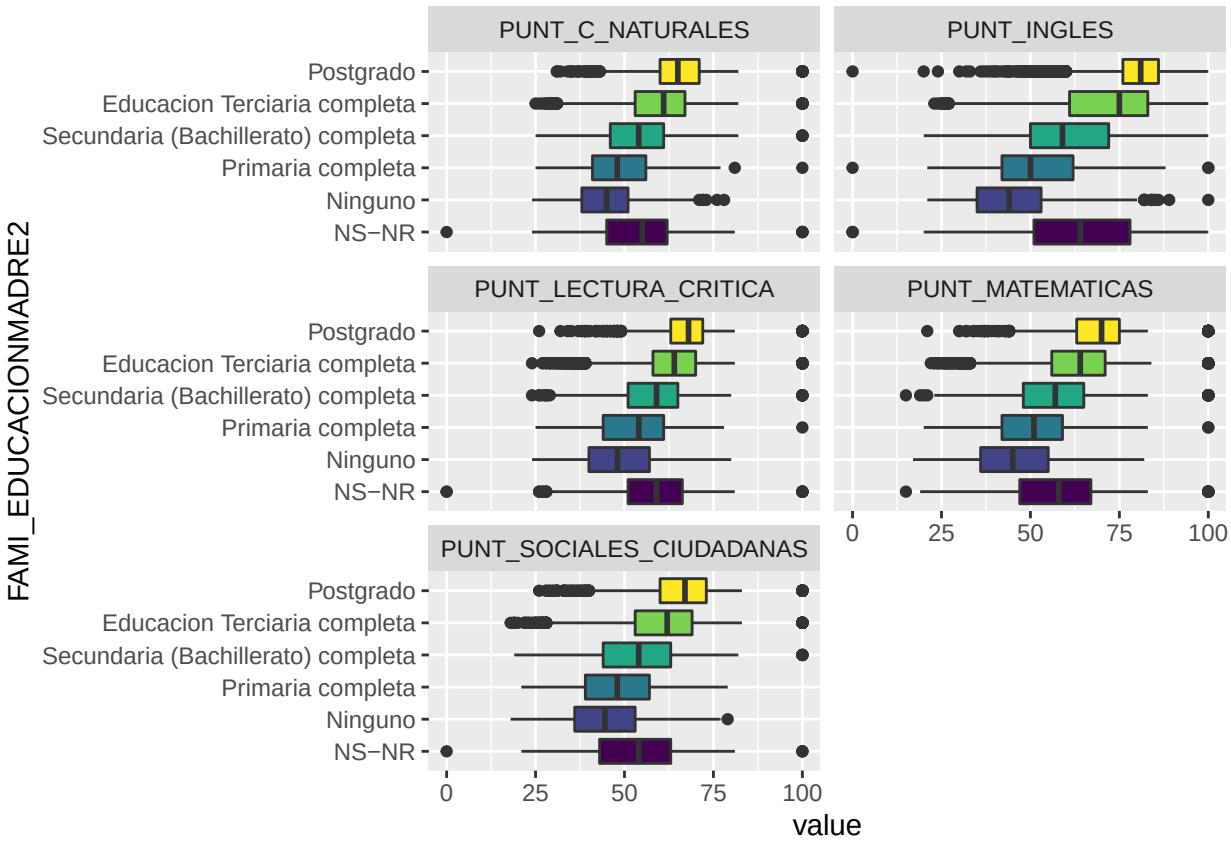


Impact of socio-economic factors in student's performance

We suspect the mother's education is a good preductor of the students results as shown by the following boxplot.

```
puntajes2=saber11 %>%  select(starts_with("PUNT"),FAMI_EDUCACIONMADRE2,FAMI_ESTRATOVIVIENDA) %>%  selec
puntajes2 %<>% pivot_longer(cols = starts_with("PUNT"))


ggplot(puntajes2,aes(x=value,y=FAMI_EDUCACIONMADRE2,fill=FAMI_EDUCACIONMADRE2))+
  geom_boxplot()+
  facet_wrap(vars(name),ncol =  2)+theme(legend.position = "none")
```

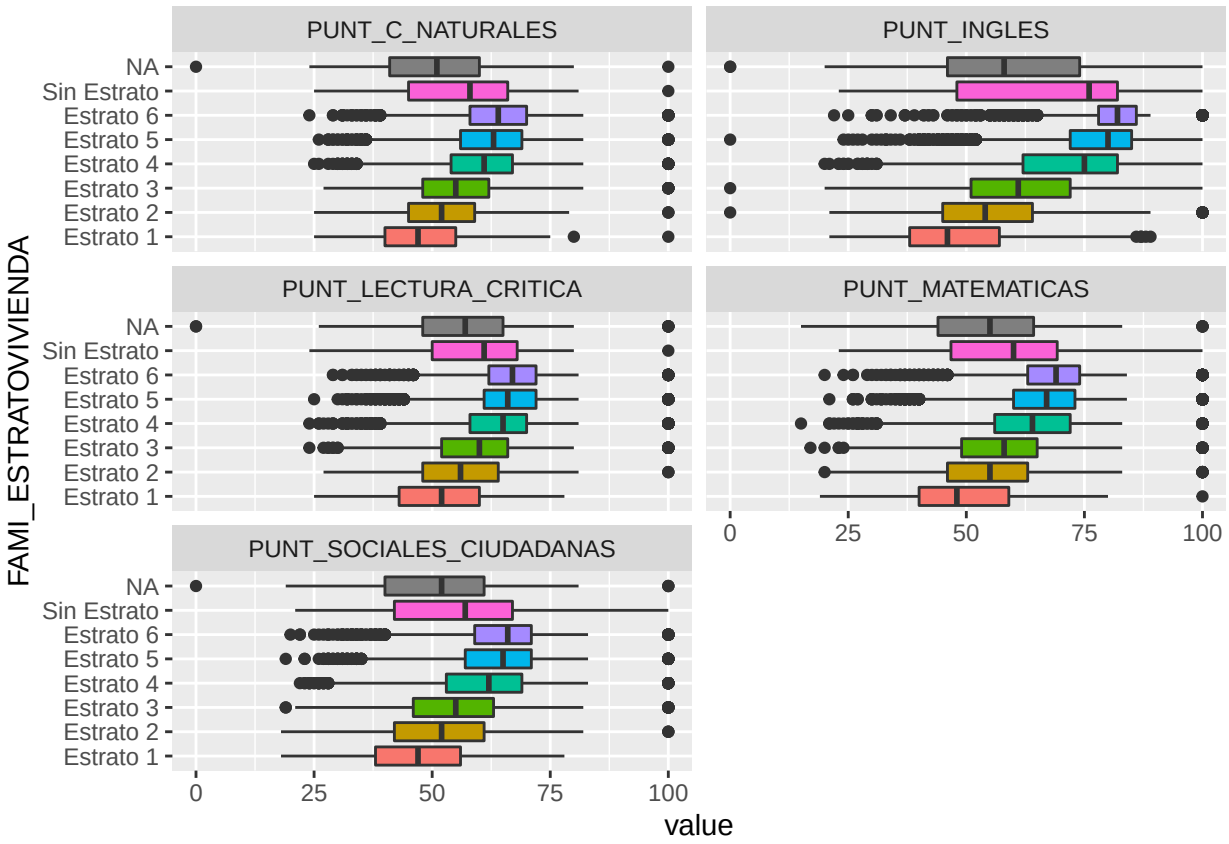
```
## Warning: Removed 47 rows containing non-finite values (stat_boxplot).
```



A similar relationship is expected if we use strata as a factor to explain student's performance in each test.

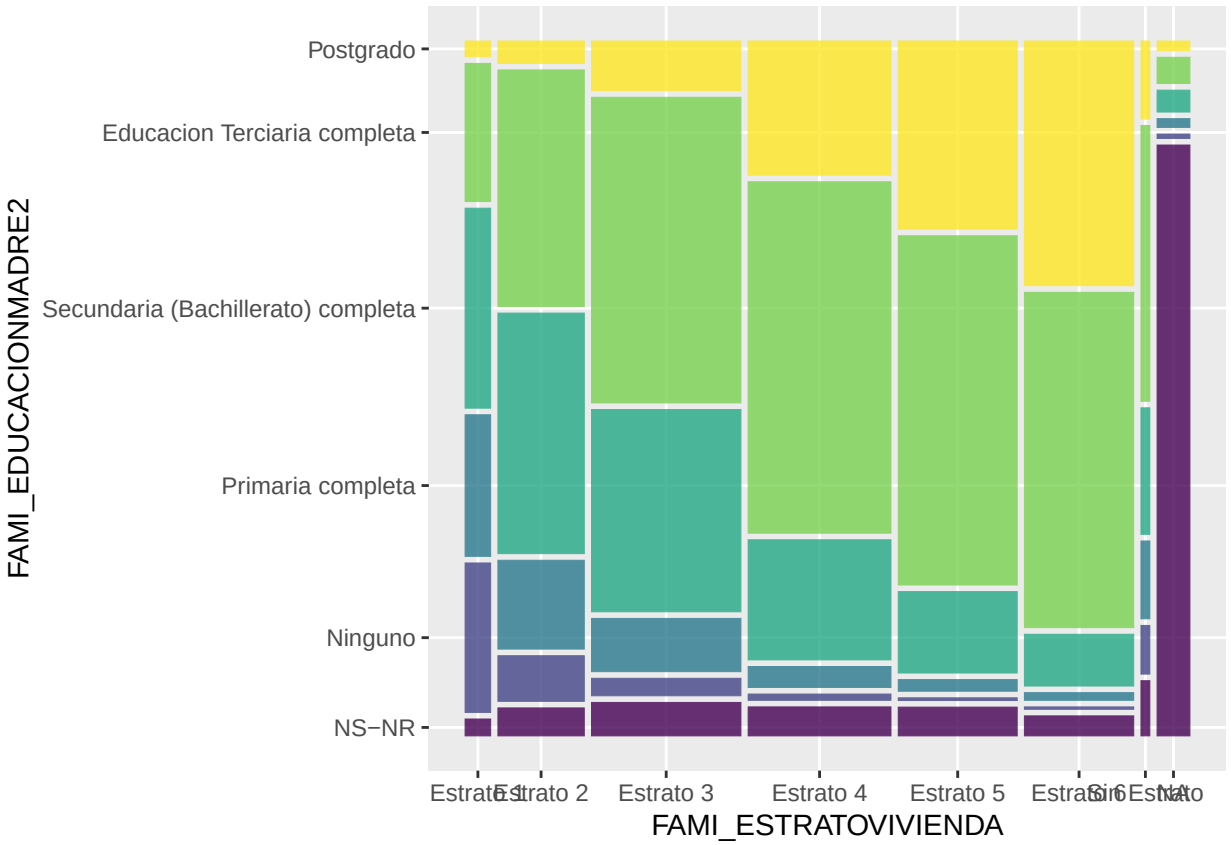
```
ggplot(puntajes2,aes(x=value,y=FAMI_ESTRATOVIVIENDA,fill=FAMI_ESTRATOVIVIENDA))+
  geom_boxplot()+
  facet_wrap(vars(name),ncol =  2)+theme(legend.position = "none")
```

```
## Warning: Removed 47 rows containing non-finite values (stat_boxplot).
```



This is expected as the higher stratum the higher the mother's education is expected to be, this is shown in the following graph.

```
ggplot(saber11)+
  geom_mosaic(aes(x=product(FAMI_EDUCACIONMADRE2,FAMI_ESTRATOVIVIENDA),
                  fill=FAMI_EDUCACIONMADRE2))+
  theme(legend.position = "none")
```



References

[1] George, D. and Mallery, P. (2010) SPSS for Windows Step by Step: A Simple Guide and Reference 17.0 Update. 10th Edition, Pearson, Boston.

[2] Secretaría distrital de planeación. Estratificación económica-Generalidades. Retrieved from http://www.sdp.gov.co/gestion-estudios-estrategicos/estratificacion/generalidades