EDAV fall 2018 Final Project (hnt2107_nm3086_sdt2134)

Hrishikesh Telang, Naoto Minakawa, Somendra Tripathi

12/10/2018

EDAV fall 2018 Final Project (hnt2107_nm3086_sdt2134)

Hrishikesh Telang, Naoto Minakawa, Somendra Tripathi 12/10/2018

1 Introduction

1.1 Motivation

Grades are the primary indicator of a student's academic performance. While there may not be a correlation between grades and a student's intelligence, there are definitely multiple external factors which contribute to a student's grade. We wanted to analyze the effect of multiple factors like parents' education, alcohol consumption, free time, etc. on students' grades.

1.2 Questions which we are interested in studying

We made several guesses before exploratory data analysis.

As positive factors to grade, we made following guessses.

- Students' grades are higher
  - If study time is longer
  - If they get support from family or school
  - If parents took higher education

As negative factors to grade, we made following guessses.

- Students' grades are lower
  - If students consume more alcohol
  - If students are frequently absent from school

1.3 Team Member and Individual Contribution to Project

Our team consists of Hrishikesh Telang, Naoto Minakawa, and Somendra Tripathi. For exploratory data analysis (EDA) part, we categorize variables into several groups as mentioned in detail later at 3.2. Each individual conducted EDA for a specific group of variables. After EDA, Hrishkesh implemented interactive chart in D3, Somendra conducted further EDA, Naoto and Hrishikesh worked on creating the report. Finally, we reviewed our consolidated work and created executive summary. We regularly discussed issues and progress so that we can be on the same page.

2 Description of data

2.1 The Brief Explanation of Dataset

We chose Student Performance Data Set which is provided on UC Irvine Machine Learning Repository. The data were obtained in a survey of students at math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender-based and study related information about students. Some variables are associated with the questions we are interested in. There are several (382) students that belong to both datasets, i.e they take both Math and Portuguese courses.

(Source)

https://archive.ics.uci.edu/ml/datasets/student+performance

http://www3.dsi.uminho.pt/pcortez/student.pdf

2.2 Data Attributes

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets are as follows; The definition of attributes are cited from original source.

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, output target)

Additional note: there are several (382) students that belong to both datasets. These students can be identified by searching for identical attributes that characterize each student, as shown in the annexed R file.

3 Analysis of data quality

```r
# Install packages
library(tidyverse)
library(ggbeeswarm)
library(ggthemes)
library(ggridges)
library(GGally)
library(cluster)
library(carData)
library(extracat)
library(pgmm)
library(vcd)

# Load datasets
student_por <- read_csv('student-por.csv')
student_mat <- read_csv('student-mat.csv')

# Set theme
theme_set(c(theme_classic(12),plot.title = element_text(face = "bold", size = 12)
            ,plot.subtitle = element_text(face = "bold", color = "grey35", size = 11)
            ,plot.caption = element_text(color = "grey68",size=5)
            ,axis.text = element_text(size=10)))
```

3.1 Observation on dataset

We observed that the number of students were distributed as follows :

- Mathematics : 395
- Portuguese : 677
- Both : 382

It was also obvious from our initial exploration that a lot more students were scoring Zeros in Maths compared to Portuguese. However, the general trend in both the graphs below appears to be the same.

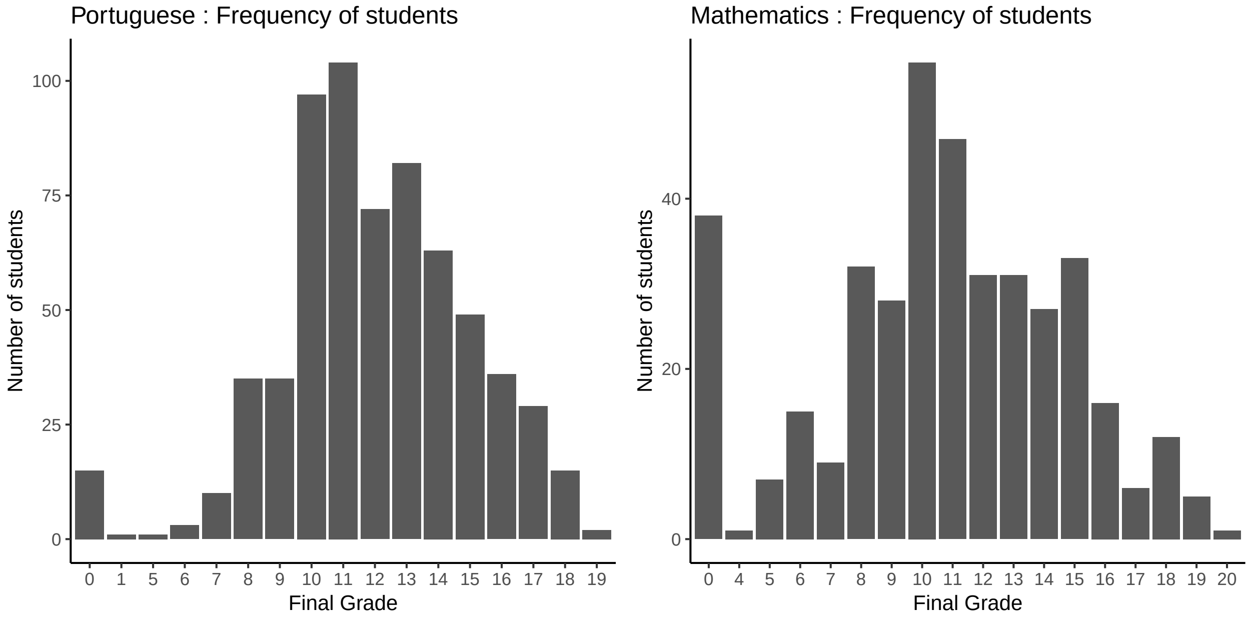
```r
p_por <- student_por %>%
  mutate(grade = as.factor(G3)) %>%
  group_by(grade) %>%
  summarise(freq = n()) %>%
  ggplot() +
  geom_histogram(aes(x = grade ,y = freq), stat = 'identity') +
  ggtitle("Portuguese : Frequency of students")+
  xlab("Final Grade") +
  ylab("Number of students")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
p_mat <- student_mat %>%
  mutate(grade = as.factor(G3))%>%
  group_by(grade) %>%
  summarise(freq = n()) %>%
  ggplot() +
  geom_histogram(aes(x = grade ,y = freq), stat = 'identity') +
  ggtitle("Mathematics : Frequency of students")+
  xlab("Final Grade") +
  ylab("Number of students")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

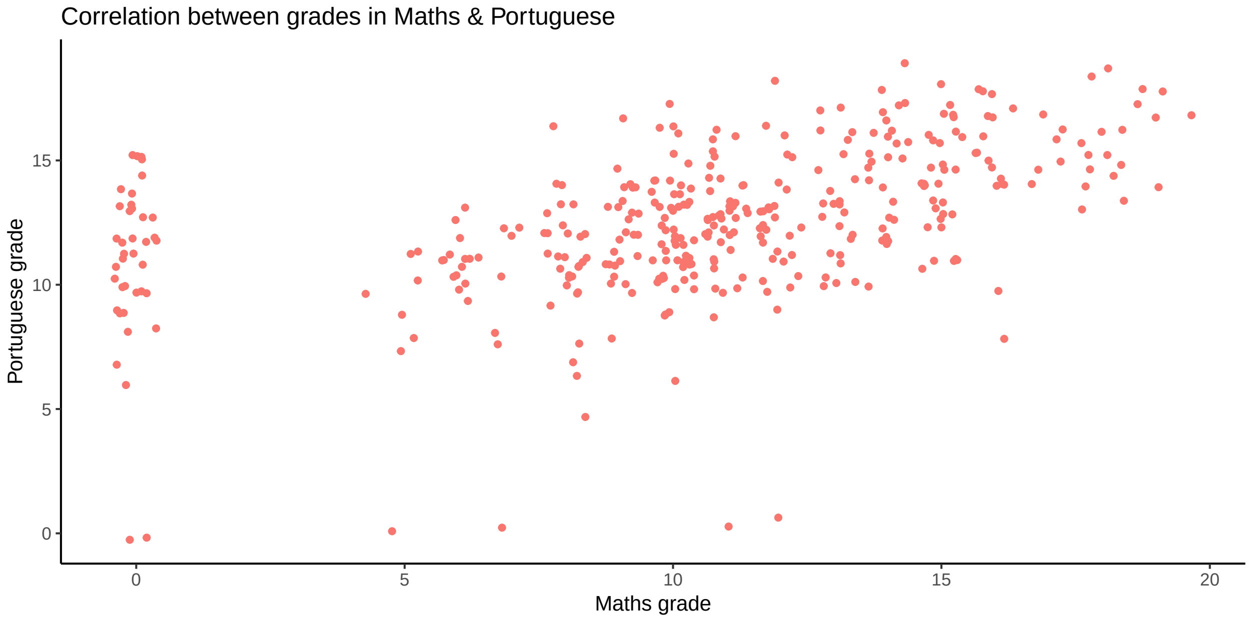
```
gridExtra::grid.arrange(p_por,p_mat,ncol=2)
```



Although some students have performed poorly in Mathematics, we can see below that the two variables are positively correlated. This shows that analysis of final grade of the Mathematics and Portuguese datasets would yield comparable results. Thus, we made a decision to only analyze the Portuguese dataset for our project.

```
# Merge math and Portuguese students data
student_both=merge(student_mat,student_por,by=c("school","sex","age","address","famsize","Pstatus","Medu
                                               "Fedu","Mjob","Fjob","reason","nursery","internet"))
#print(nrow(student_both)) # 382 students

# Plot correlation
ggplot(student_both, aes(G3.x,G3.y,color = "blue"))+
  geom_jitter() +
  theme(legend.position = "none") +
  xlab("Maths grade") +
  ylab("Portuguese grade") +
```

```r
  ggtitle("Correlation between grades in Maths & Portuguese")
```



3.2 Categorization of Data Attributes

We categorized data attributes in 2.2 into several groups so that each individual team member can explore a different group of attributes. There are 30 attirbutes which may be associated with grades. We agreed to categorize variables into broad categories which might be useful for analysis.

Following is our definiton of categorization;

1. Home environment

   We regard the following attritbutes as belonging to home environment. While studytime technically does not have to do a lot with a student's environment, it does not fit as well into any of the other categories. Thus, we thought it best to group it with the following variables:

   - famrel (Family Relationship)
   - internet (Access to internet)
   - studytime (Time spent studying)
   - traveltime (Time taken to travel to school)
   - Pstatus (Whether Parents are separated or together)
   - address (Rural or Urban)
   - famsize (Less than 3 people or more than 3)

2. Social

   We regard that following attritbutes can be categorized as students' social activties. We hypothesize that students consume alcohol at social events such as parties, so we categorized it into this group.

   - freetime (Amount of free time a student has)
   - goout (How often a student goes out with friends)
   - Dalc (Daily alcohol consumption)
   - Walc (Weekend alcohol consumption)
   - romantic (Whether or not the student is in a romantic relationship)

3. Parents

The following attritbutes are related to parents.

```
-   Medu (Mother's education)
-   Fedu (Father's education)
-   Mjob (Mother's job)
-   Fjob (Father's job)
-   guardian (Who is the student's primary guardian)
```

4. Profile

   The variables associated with student profile were:

   - sex
   - age
   - health

5. Academics

   We included following attributes in Academics group. We discussed if we should categorize activities into Social group. However, given that activites is extra-curricular activities, we thought it is more related to academics than to social activites.

   - G1
   - G2
   - G3
   - absences
   - school
   - failures
   - activities

6. Others

   Attributes in this group do not fit well into any of above groups. Therefore, we created one group called Others and categorized following attributes.

   - reason (Reason for choosing the school)
   - higher (whether or not the student is interested in higher education)
   - schoolsup (whether or not the student gets financial support from the school)
   - famsup (whether or not the student gets financial support from family)
   - paid (extra paid classes within the course subject)
   - nursery (Whether or not student attended nursery as a child)

3.3 Issues with the dataset

We found some issues during data quality analysis. According to data source, there are 382 students who study both the math course and Portuguese course. Such overlapping students ideally have to be unique in both the datasets, i.e have the same values for all columns in both the datasets.

However, when we tried to join math course data set and Portuguese some of students data appeared not to be unique. Specifically, when we tried to join students by using school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, nursery, internet as keys, some columns showed different values.

We found that values in such columns were obtained through interview questions to students, and concluded that there may be some human error in tabulating the results or answering questions, resulting in different answers in different surveys by the same students.

It could also be possible that the clashes may be occuring due to the presence of twins going to the same school, as they have the same parents

4 Main analysis (Exploratory Data Analysis)

4.1 Our Approach

As previously mentioned, we categorized 30 attirbutes into 6 groups and explored different groups to gain certain insights. We only analyzed the Portuguese student's dataset for reasons discussed above. We first sought to clean our dataset to better represent our results and then proceeded to analyze each of the 6 groups separately.

4.2 Data Cleaning

We converted GPA - G3, G2, G1 to the Erasmus grading system described in the paper, since it was more clear to understand. The Erasmus methodology suggests that G3, G2, G1 corresponeds to following level in Portugal/France.

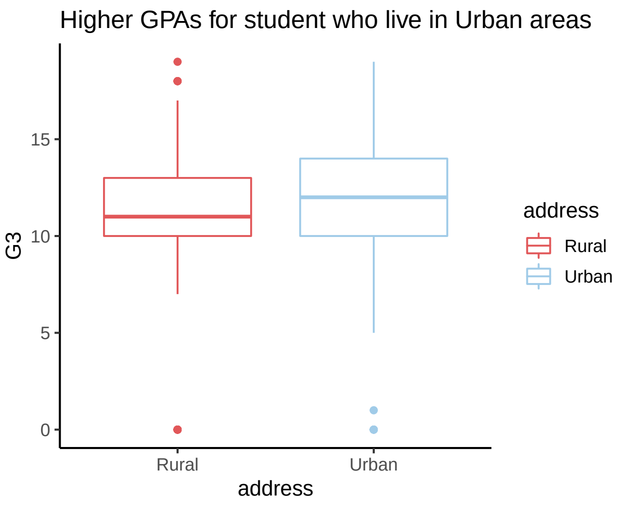
- 16-20: excellent/very good
- 14-15: good
- 12-13: satisfactory
- 10-11: sufficient
- 0-9: fail

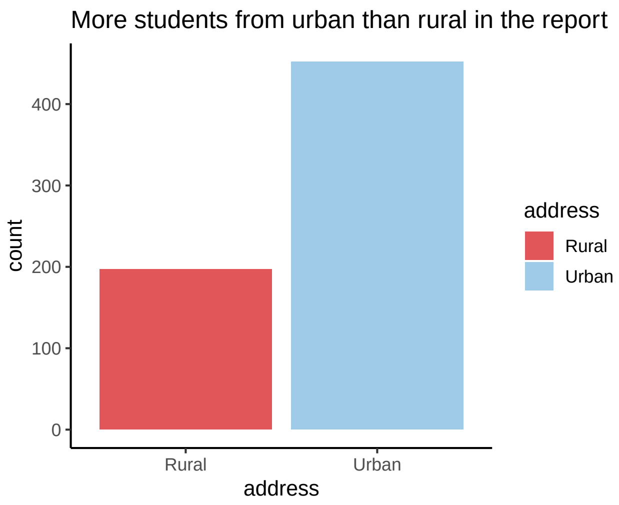
```
student_por <- student_por %>%
  mutate(G3_erasmus = if_else(G3>=16, 'excellent',
                            if_else(G3>=14&G3<=15, 'good',
                                    if_else(G3>=12&G3<=13, 'satisfactory',
                                            if_else(G3>=10&G3<=11, 'sufficient',
                                                    'fail'))))) %>%
  mutate(G3_erasmus = factor(G3_erasmus,
                             c('excellent','good','satisfactory','sufficient','fail')))
```

4.3.1 Analysis for Home Environment Group

Firstly, we found that students who live in urban areas have higher final grades than students in rural areas. It is also crucial to note that there are more students from urban than rural areas in the report.

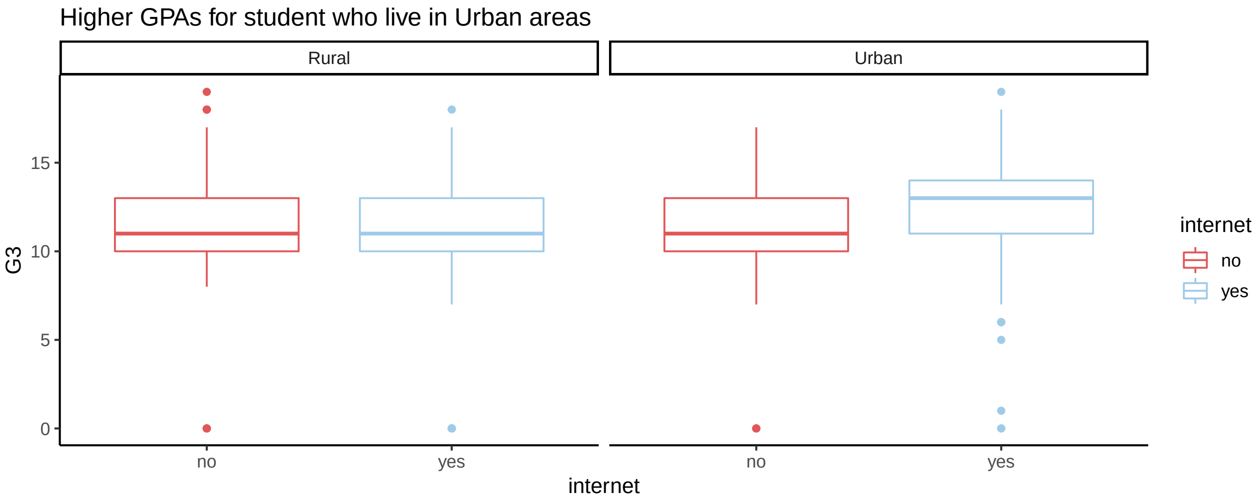
```
student_por$address[student_por$address == 'U'] <- "Urban"
student_por$address[student_por$address == 'R'] <- "Rural"


address_gpa<- student_por %>%
  ggplot(aes(x = address, y = G3, color = address)) +
  geom_boxplot() +
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Higher GPAs for student who live in Urban areas")

freq <- ggplot(student_por, aes(x = address, fill = address)) +
  geom_bar() +
  scale_fill_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("More students from urban than rural in the report")

gridExtra::grid.arrange(address_gpa,freq,ncol=2)
```





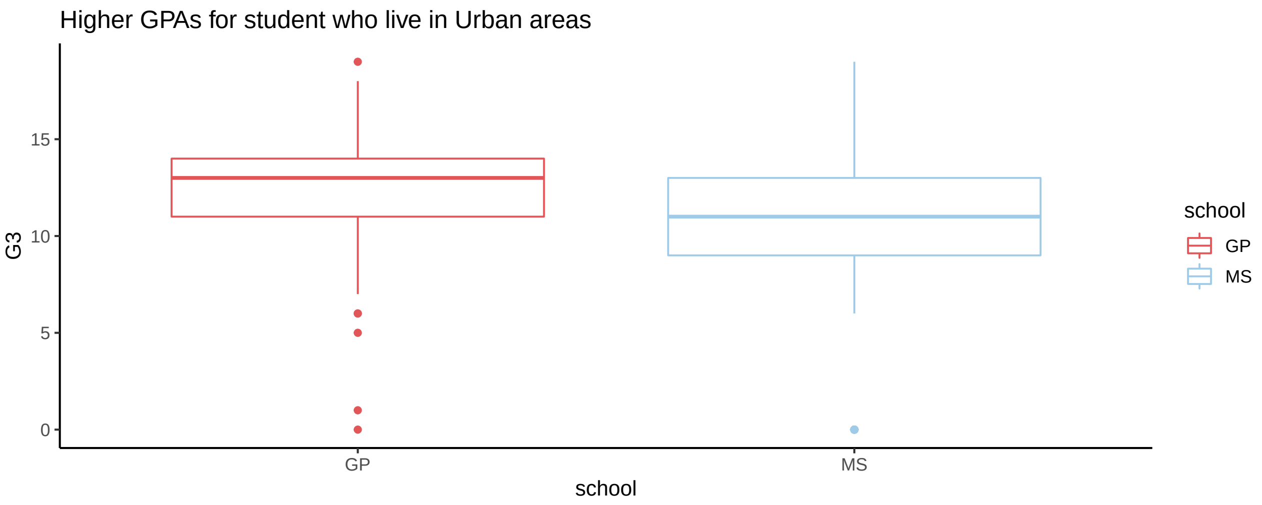
We next checked whether access to internet has any impact on final grades. While access to internet wasn't found to make any change in a student's final grades in Rural areas, Urban students with access to internet outperformed their fellow students without internet.

```
student_por %>%
  ggplot(aes(x = internet, y = G3, color = internet)) +
  geom_boxplot() +
  facet_wrap(~address)+
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Higher GPAs for student who live in Urban areas")
```



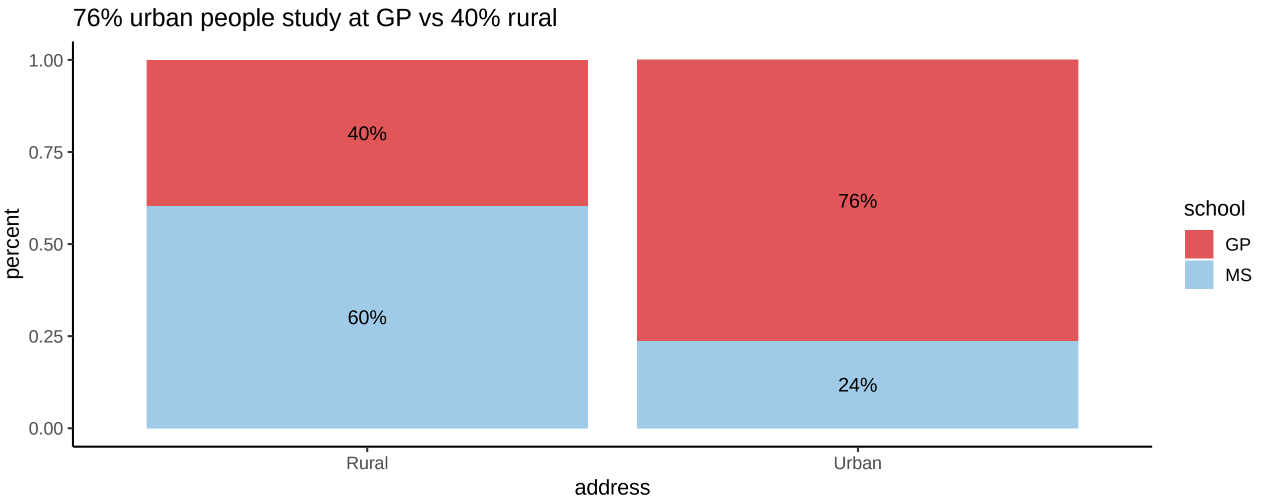
We tried to analyze if going to a particular school has an impact on grades.It can be seen that School GP reported higher grades.

```
student_por %>%
  ggplot(aes(x = school, y = G3, color = school)) +
  geom_boxplot() +
  #facet_wrap(~)+
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Higher GPAs for student who live in Urban areas")
```



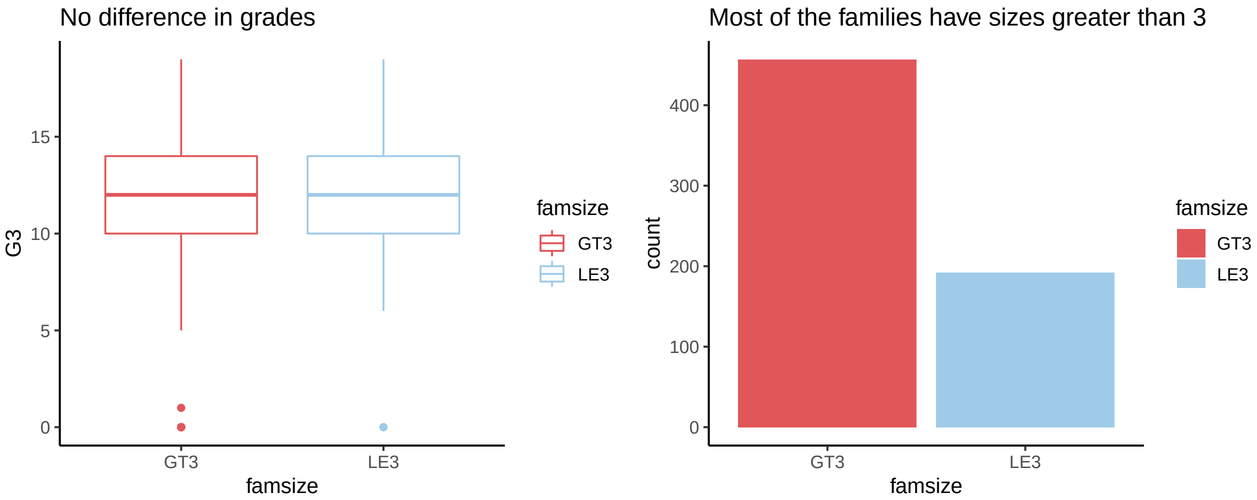
On further analysis, we can see that interestingly, 76% students in urban ares study at GP while 40% of students in rural area do so. This might explain the higher grades found in urban students.

```
student_por %>%
  group_by(school, address) %>% tally() %>%
  group_by(address) %>%
  mutate(percent = n/sum(n))  %>%
  arrange(desc(school)) %>%
  ggplot(aes(x = address, y = percent)) +
  geom_col(aes(fill = school)) +
  geom_text(aes(label = paste(round(percent,2) * 100, "%",sep = "")),
            position = position_stack(vjust = 0.5))+
  scale_colour_manual(values = c("#e15759","#a0cbe8")) +
  scale_fill_manual(values = c("#e15759","#a0cbe8")) +
  labs(fill = "school") +
  ggtitle("76% urban people study at GP vs 40% rural")
```



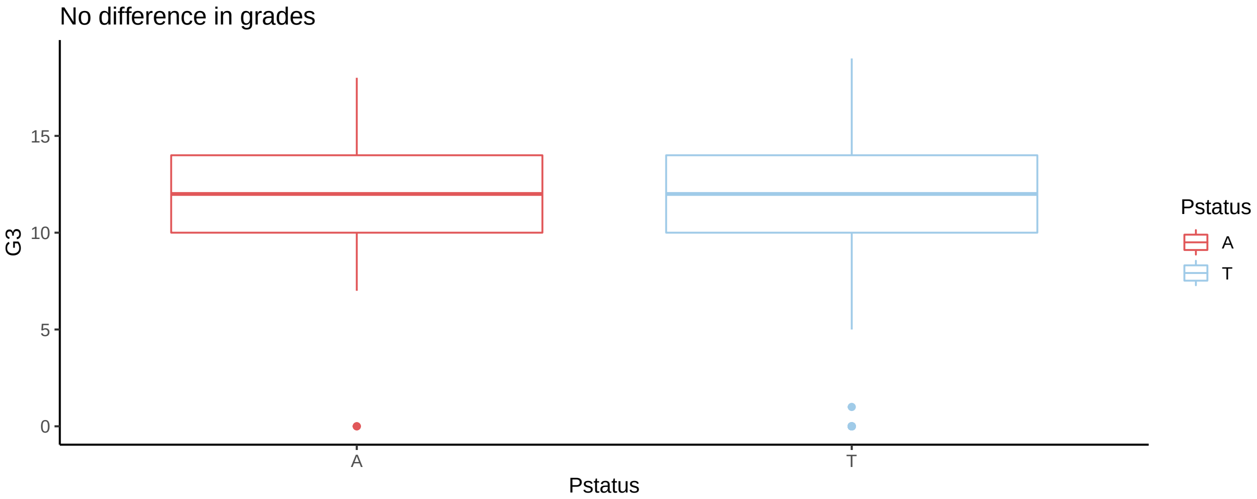
Next, we tried to determine if family size plays any part . We don't see any difference in grades due to family size. Most of the families have sizes greater than 3.

```
famsize_gpa<- student_por %>%
  ggplot(aes(x = famsize, y = G3, color = famsize)) +
  geom_boxplot() +
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("No difference in grades")

freq <- ggplot(student_por, aes(x = famsize, fill = famsize)) +
  geom_bar() +
  scale_fill_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Most of the families have sizes greater than 3")

gridExtra::grid.arrange(famsize_gpa,freq,ncol=2)
```



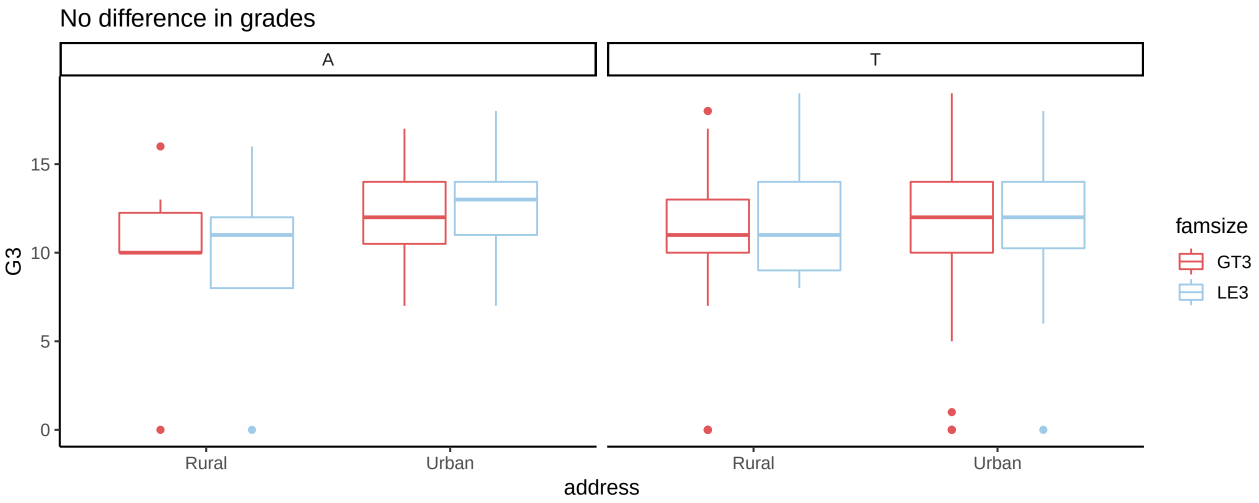
Whether or not parents live separately,it does not have impact on students grades.

```
student_por %>%
  ggplot(aes(x = Pstatus, y = G3, color = Pstatus)) +
  geom_boxplot() +
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
ggtitle("No difference in grades")
```



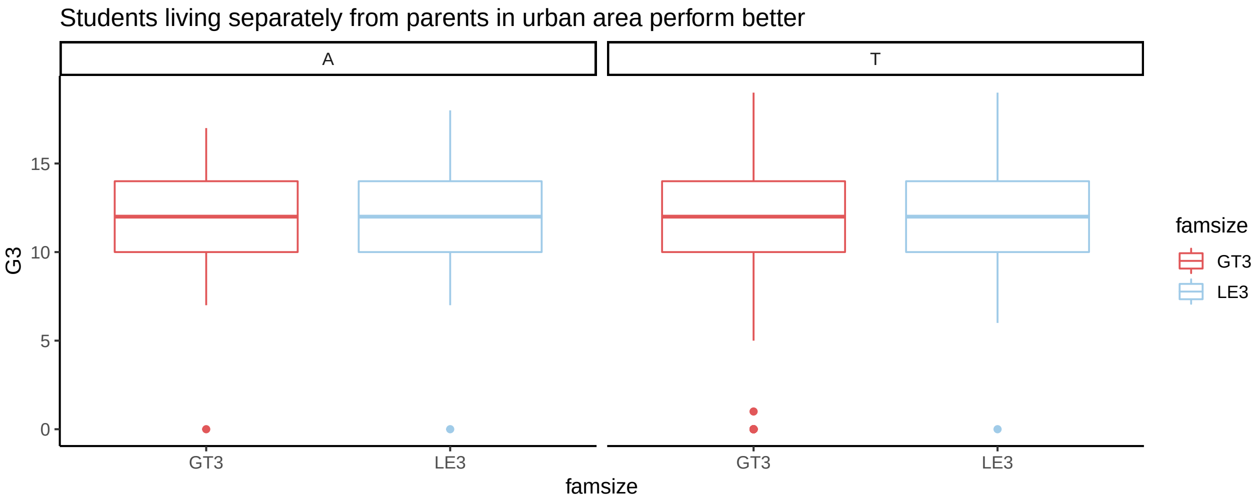
We then check the same trend by urban and rural areas. We see students in urban areas whose parents are separated tend to perform slightly better than their counterparts

```
student_por %>%
  ggplot(aes(x = address, y = G3, color = famsize)) +
  geom_boxplot() +
  facet_wrap(~Pstatus) +
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("No difference in grades")
```



We took a look at the combination of family size and whether parents live separately or not, however, it does not have impact on students grades.

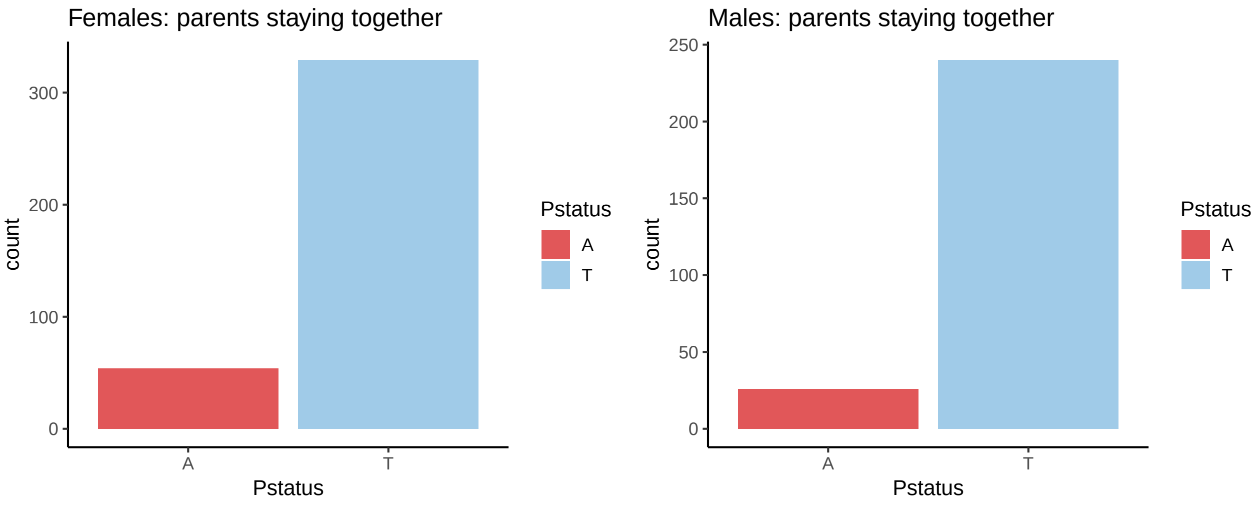
```
student_por %>%
  ggplot(aes(x = famsize, y = G3, color = famsize)) +
  geom_boxplot() +
  facet_wrap(~Pstatus) +
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Students living separately from parents in urban area perform better")
```



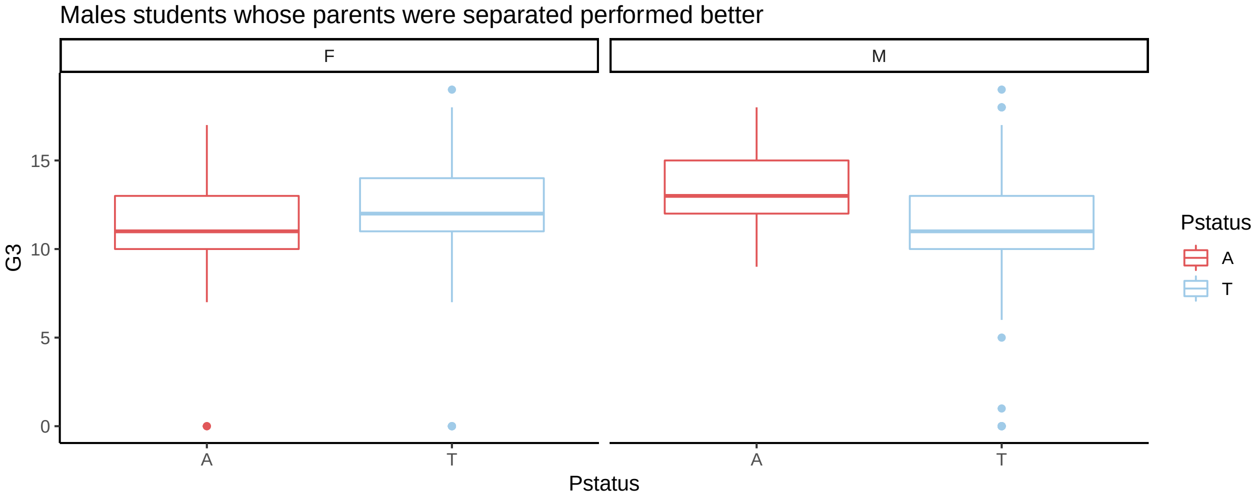
When we look into gender distribution, males students whose parents were separated performed better.

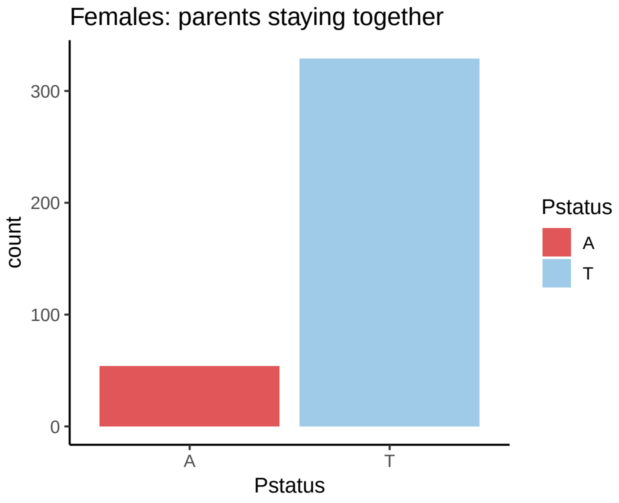
```
x <- student_por %>%
  ggplot(aes(x = Pstatus, y = G3, color = Pstatus)) +
  geom_boxplot() +
  facet_wrap(~sex)+
  scale_color_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Males students whose parents were separated performed better")
```

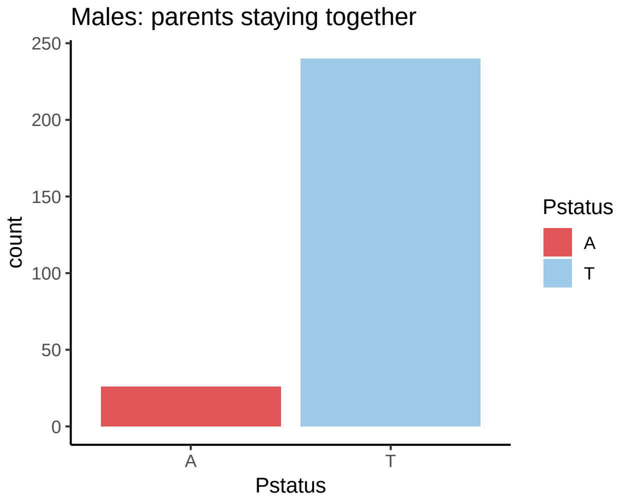
```r
freq_fem <- ggplot(filter(student_por,sex=="F"), aes(x = Pstatus, fill = Pstatus)) +
  geom_bar() +
  scale_fill_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Females: parents staying together")

freq_male <- ggplot(filter(student_por,sex=="M"), aes(x = Pstatus, fill = Pstatus)) +
  geom_bar() +
  scale_fill_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("Males: parents staying together")

y <- gridExtra::grid.arrange(freq_fem,freq_male, ncol = 2)
```



```r
gridExtra::grid.arrange(x,y,ncol=1,nrow=2)
```

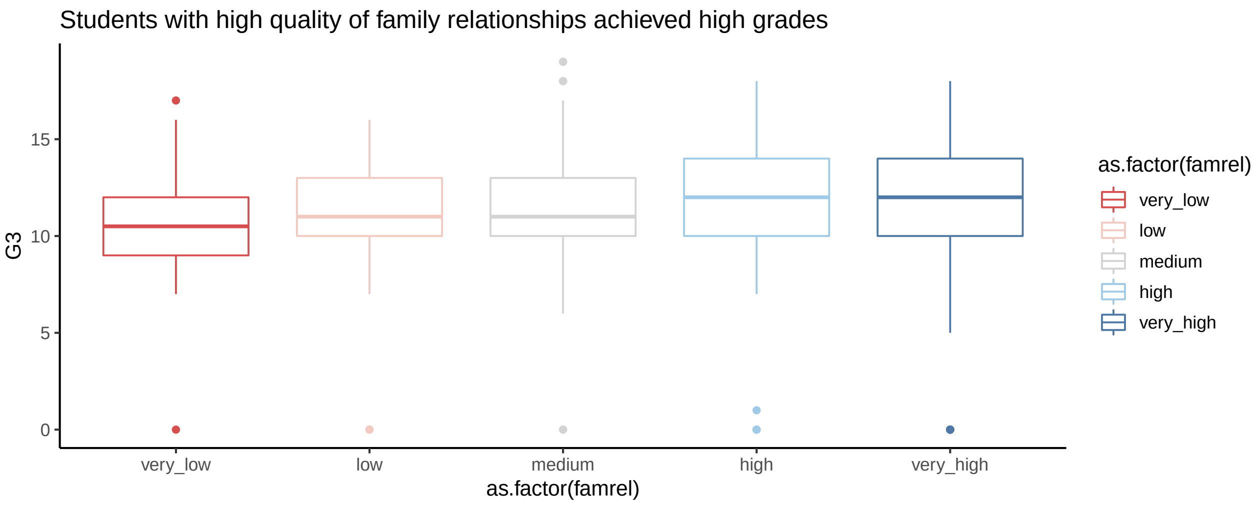






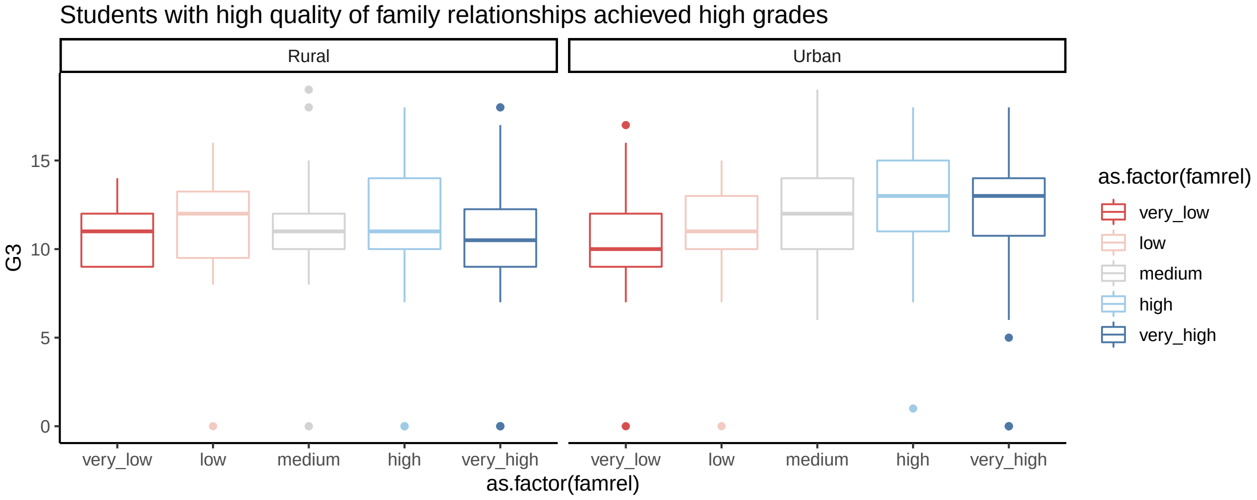
When analyzing family relationship, students who ranked their family relationships high tended to earned better grades.

```
student_por$famrel <- factor(student_por$famrel,
                             labels =  c('very_low', 'low', 'medium', 'high', 'very_high'),
                             ordered = TRUE)

student_por %>%
  ggplot(aes(x = as.factor(famrel), y = G3, color = as.factor(famrel))) +
  geom_boxplot() +
  scale_color_manual(values = c("#D64E4E","#F2CAC1","#d3d3d3","#a0cbe8","#4e79a7"))+
  ggtitle("Students with high quality of family relationships achieved high grades")
```



According to the following chart, such tendency is more obvious for urban families.

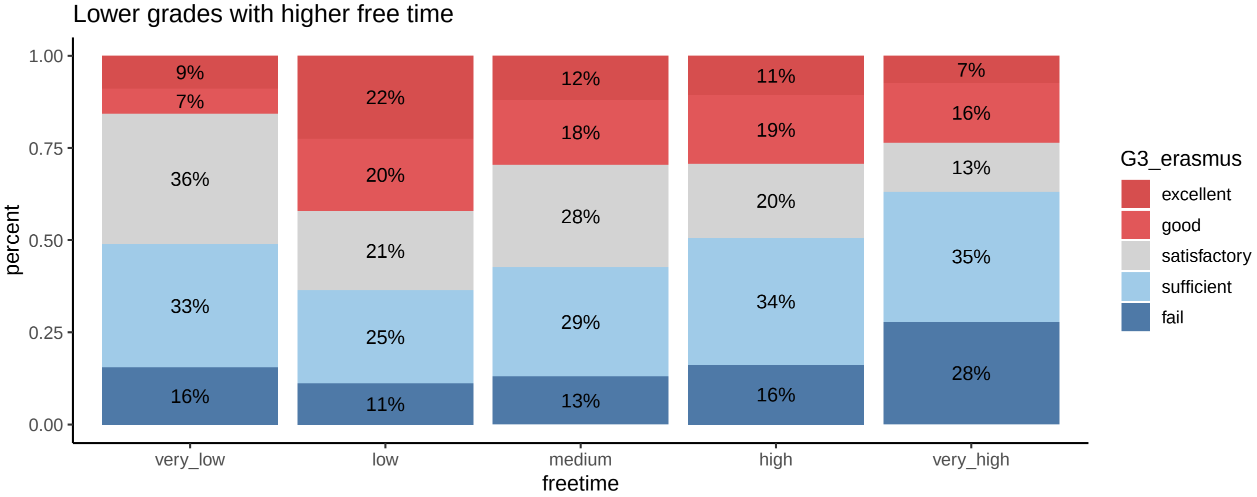
```
student_por %>%
  ggplot(aes(x = as.factor(famrel), y = G3, color = as.factor(famrel))) +
  facet_wrap(~address) +
  geom_boxplot() +
  scale_color_manual(values = c("#D64E4E","#F2CAC1","#d3d3d3","#a0cbe8","#4e79a7"))+
  ggtitle("Students with high quality of family relationships achieved high grades")
```



4.3.2 Analysis for Social Group

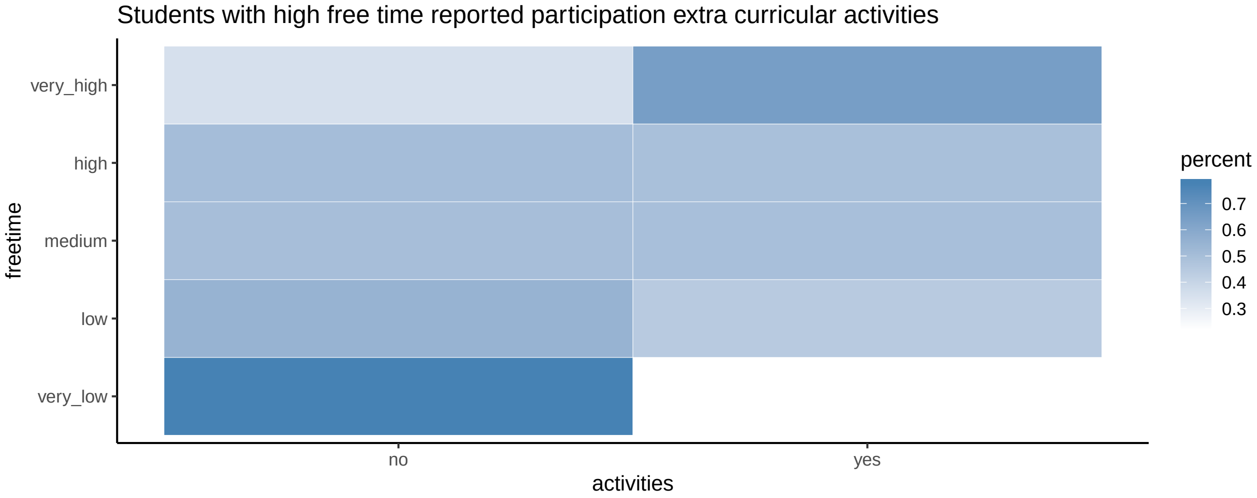
In social group, Freetime was the most interesting variable which affects grade. From the following chart, we can see students with higher free time scoring lower grades.

```
student_por$freetime <- factor(student_por$freetime,
                              labels=c('very_low', 'low', 'medium', 'high', 'very_high'),
                              ordered = TRUE)

student_por %>%
  group_by(G3_erasmus, freetime) %>%
  tally() %>%
  group_by(freetime) %>%
  mutate(percent = n/sum(n))  %>%
  arrange(desc(G3_erasmus)) %>%
  ggplot(aes(x = freetime, y = percent)) +
  geom_col(aes(fill = G3_erasmus)) +
  geom_text(aes(label = paste(round(percent,2) * 100,"%",sep = "")),
            position = position_stack(vjust = 0.5)) +
  scale_colour_manual(values = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")) +
  scale_fill_manual(values = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")) +
  labs(fill = "G3_erasmus") +
  ggtitle("Lower grades with higher free time")
```



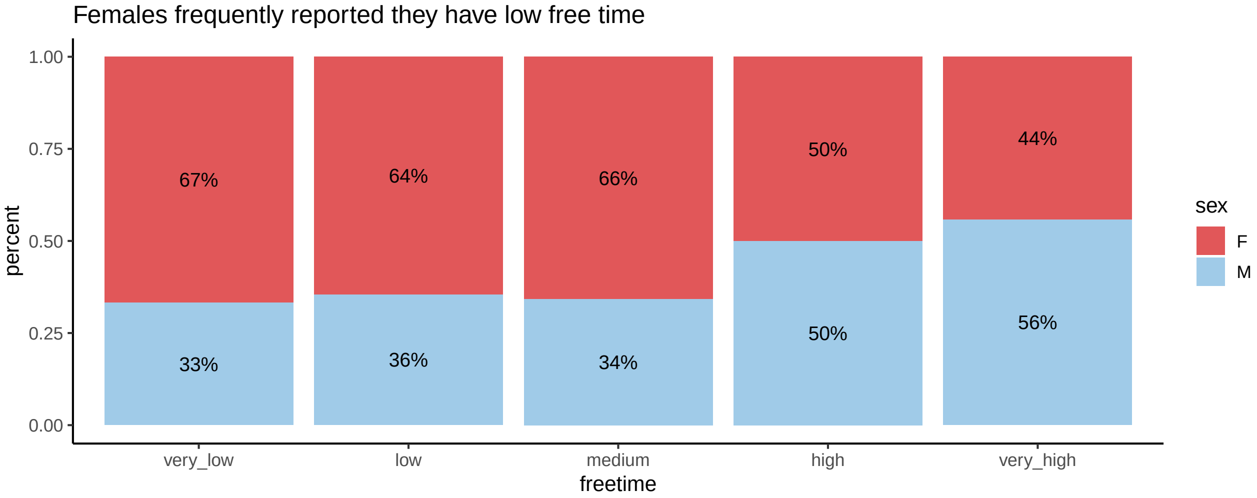
Then, we asked the question - "So what do they do during their free time?". Oddly enough, students with high free time reported the most participation extra curricular activities.

```
student_por %>%
  group_by(activities, freetime) %>%
  tally() %>%
  group_by(freetime) %>%
  mutate(percent = n/sum(n)) %>%
  ggplot(aes(activities, freetime)) +
  geom_tile(aes(fill = percent), colour = "white") +
  scale_fill_gradient(low = "white",high = "steelblue") +
  ggtitle("Students with high free time reported participation extra curricular activities")
```



We also investigated whether a particular gender reports more free time or not. Compared to males, females frequently reported they have lower free time.

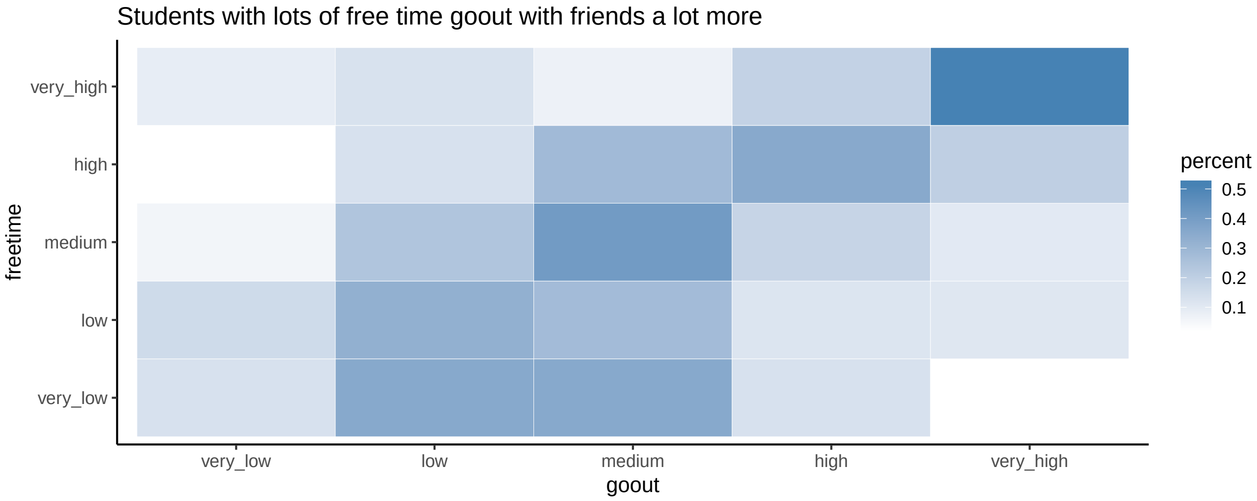
```
student_por %>%
  group_by(sex, freetime) %>%
  tally() %>%
  group_by(freetime) %>%
  mutate(percent = n/sum(n)) %>%
  arrange(desc(sex)) %>%
  ggplot(aes(x = freetime, y = percent)) +
  geom_col(aes(fill = sex)) +
  geom_text(aes(label = paste(round(percent,2) * 100, "%",sep = "")),
            position = position_stack(vjust = 0.5))+
  scale_colour_manual(values = c("#e15759","#a0cbe8")) +
  scale_fill_manual(values = c("#e15759","#a0cbe8")) +
  labs(fill = "sex") +
  ggtitle("Females frequently reported they have low free time")
```



There also seems to be an association between going out with friends and having free time. Students with a lot of free time go out with friends a lot more.

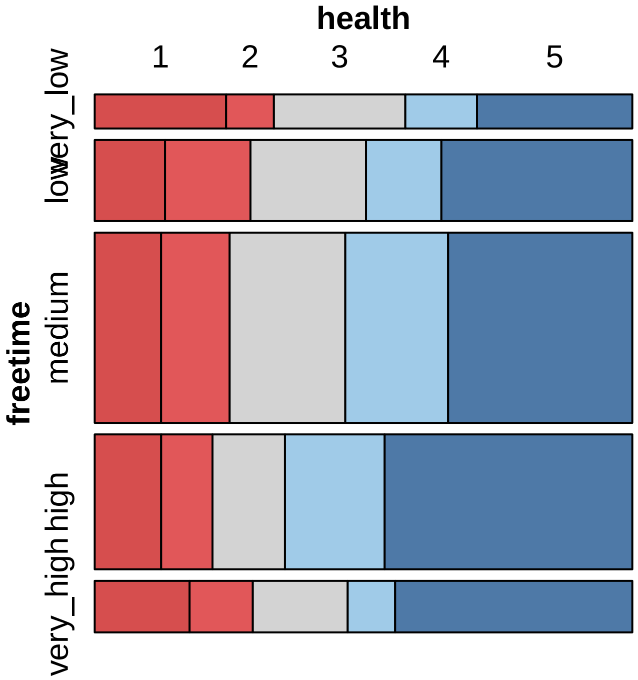
```
student_por$goout <- factor(student_por$goout,
                            labels= c('very_low', 'low', 'medium', 'high', 'very_high'),
```

```r
                            ordered = TRUE)

student_por %>%
  group_by(goout, freetime) %>%
  tally() %>%
  group_by(freetime) %>%
  mutate(percent = n/sum(n)) %>%
  ggplot(aes(goout, freetime)) + geom_tile(aes(fill = percent), colour = "white") +
  scale_fill_gradient(low = "white",high = "steelblue") +
  ggtitle("Students with lots of free time goout with friends a lot more")
```



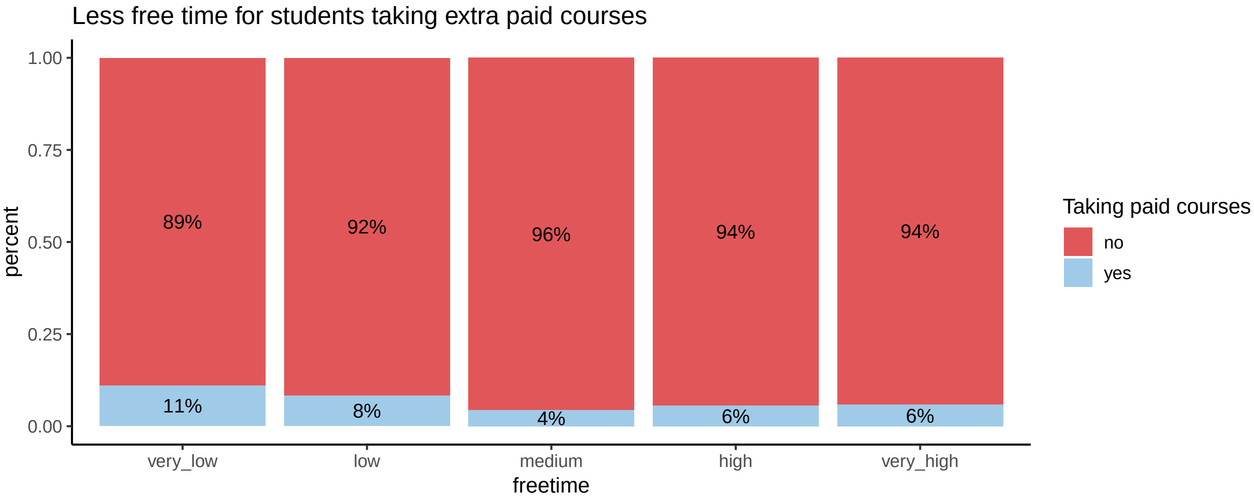
There are some associations between freetime and health. For students with very low free time, health condition is bad compared to their peers.

```r
mosaic(health ~ freetime,student_por,
            gp = gpar(fill = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")))
```



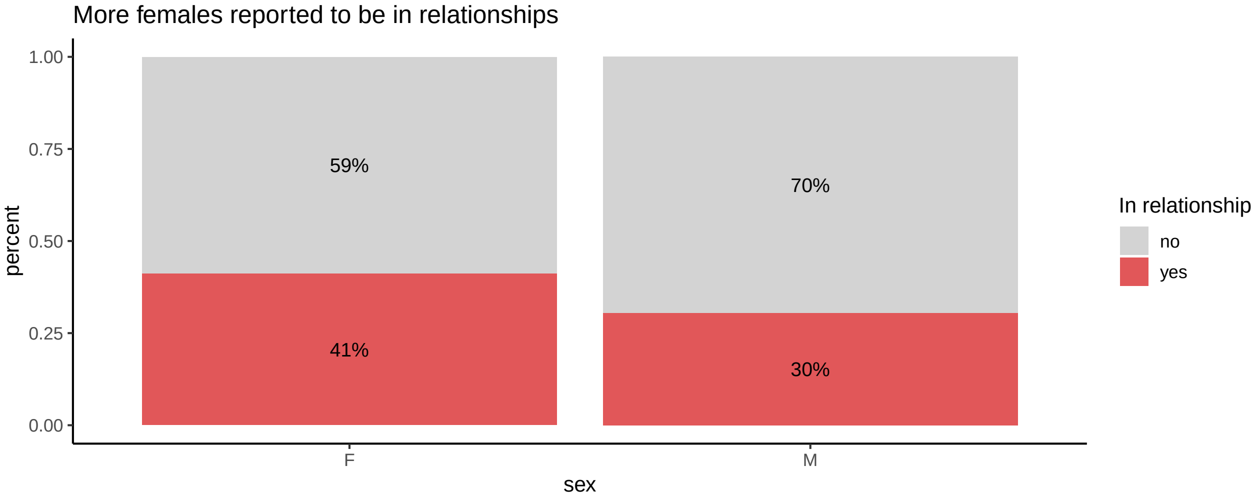
This is a sort of obvious association, but we can observe that students have less free time when they take extra paid courses.

```
student_por %>%
  group_by(paid, freetime) %>%
  tally() %>%
  group_by(freetime) %>%
  mutate(percent = n/sum(n)) %>%
  arrange(desc(paid)) %>%
  ggplot(aes(x = freetime, y = percent)) +
  geom_col(aes(fill = paid)) +
  geom_text(aes(label = paste(round(percent,2) * 100, "%",sep = "")),
            position = position_stack(vjust = 0.5))+
  scale_colour_manual(values = c("#e15759","#a0cbe8")) +
  scale_fill_manual(values = c("#e15759","#a0cbe8")) +
  labs(fill = "Taking paid courses") +
  ggtitle("Less free time for students taking extra paid courses")
```



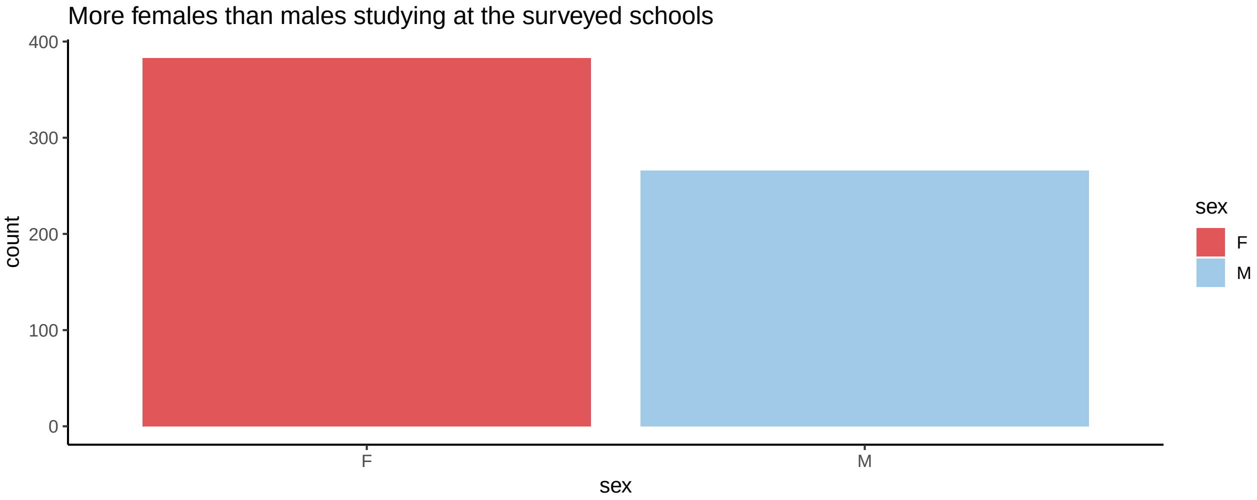
Regarding if the students are in relationship or not, more females reported to be in relationships.

```
student_por %>%
  group_by(romantic, sex) %>%
  tally() %>%
  group_by(sex) %>%
  mutate(percent = n/sum(n))  %>%
  arrange(desc(romantic)) %>%
  ggplot(aes(x = sex, y = percent)) +
  geom_col(aes(fill = romantic)) +
  geom_text(aes(label = paste(round(percent,2) * 100, "%",sep = "")),
            position = position_stack(vjust = 0.5))+
  scale_colour_manual(values = c("#d3d3d3","#e15759")) +
  scale_fill_manual(values = c("#d3d3d3","#e15759")) +
  labs(fill = "In relationship") +
  ggtitle("More females reported to be in relationships")
```



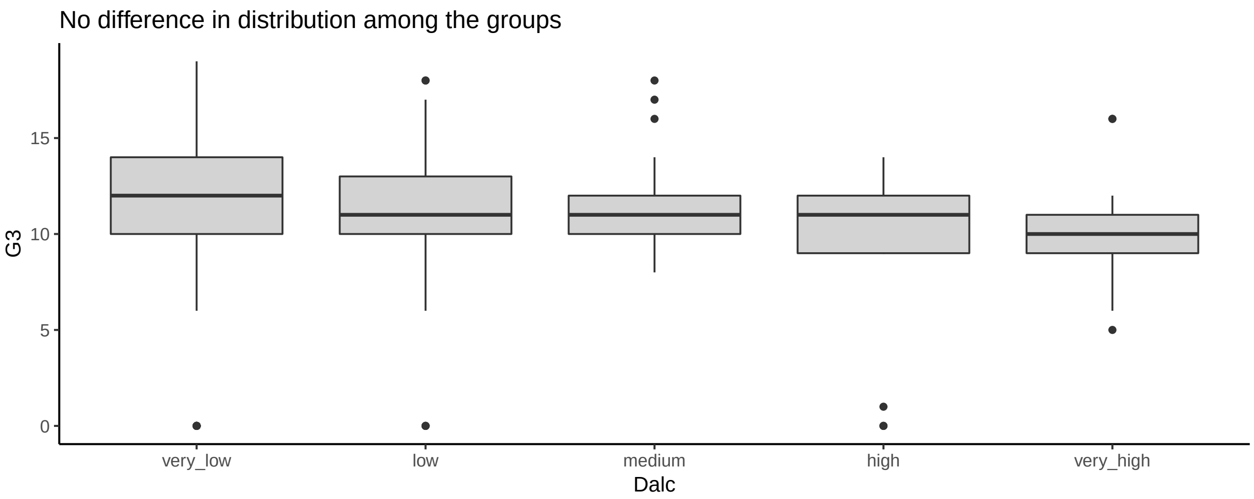
As for distribution of males and females, more females than males studying at the surveyed schools.

```
ggplot(student_por, aes(x = sex, fill = sex)) +
  geom_bar() +
  ggtitle("More females than males in the records") +
  scale_fill_manual(values = c("#e15759","#a0cbe8"))+
  ggtitle("More females than males studying at the surveyed schools")
```

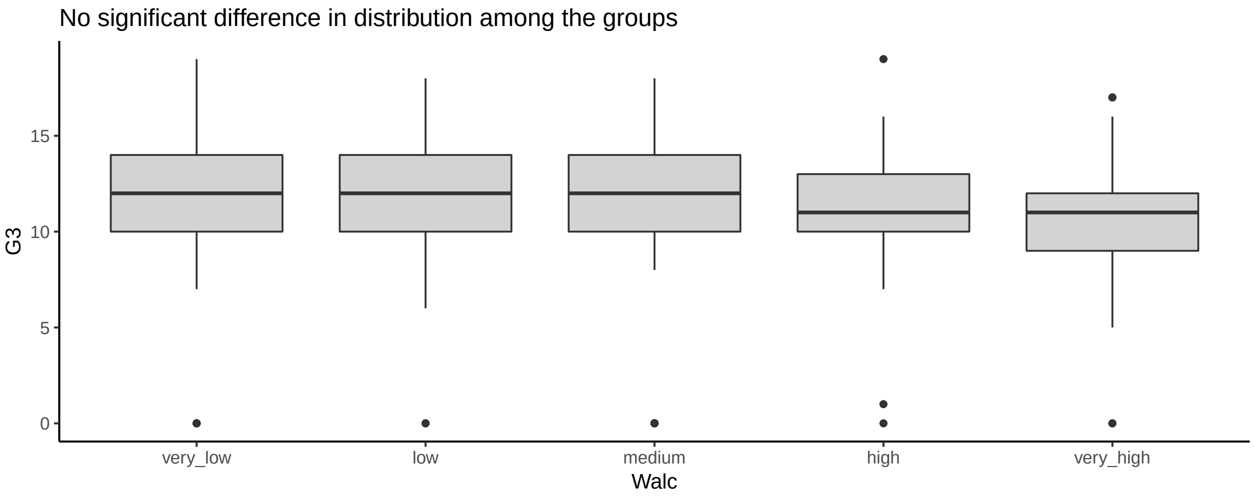


Regarding workday/weekend alcohol consumption, we don't see any significant difference among the groups.

```
student_por$Dalc <- factor(student_por$Dalc,
                           labels =  c('very_low', 'low', 'medium', 'high', 'very_high'),
                           ordered = TRUE)

student_por %>%
  ggplot(aes(x = Dalc, y = G3)) +
  geom_boxplot(fill="#d3d3d3") +
  ggtitle("No difference in distribution among the groups")
```



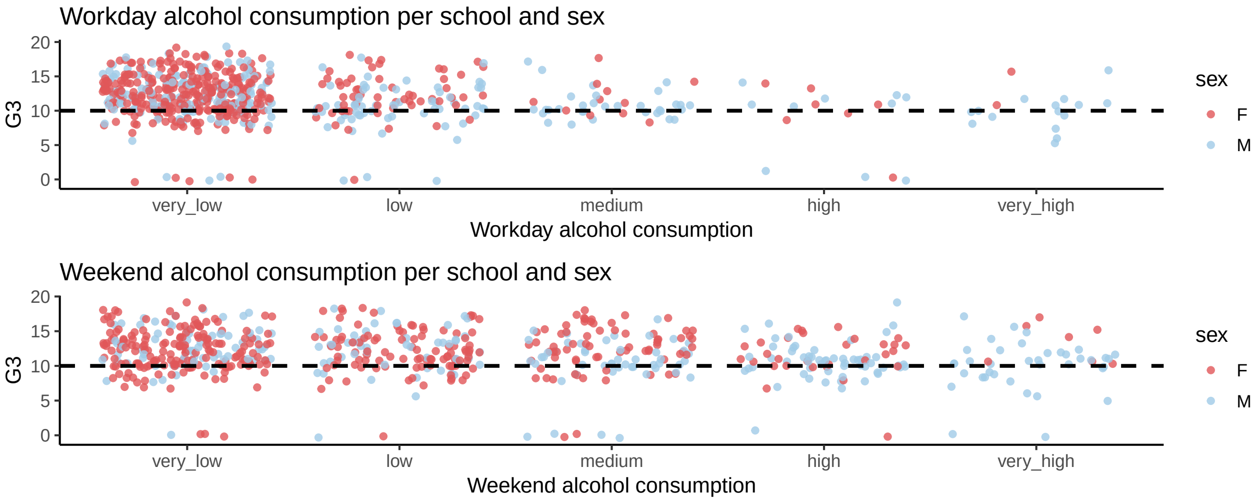
```
student_por$Walc <- factor(student_por$Walc,
                           labels =  c('very_low', 'low', 'medium', 'high', 'very_high'),
                           ordered = TRUE)

student_por %>%
  ggplot(aes(x = Walc, y = G3)) +
  geom_boxplot(fill="#d3d3d3") +
  ggtitle("No significant difference in distribution among the groups")
```



However, we found something interesting when we also observe whether there are any association between workday/weekend alcohol consumption and grades, splitting data by gender. From the chart, we can see males who consume alcohol on weekend are highly likely to get bad grades.

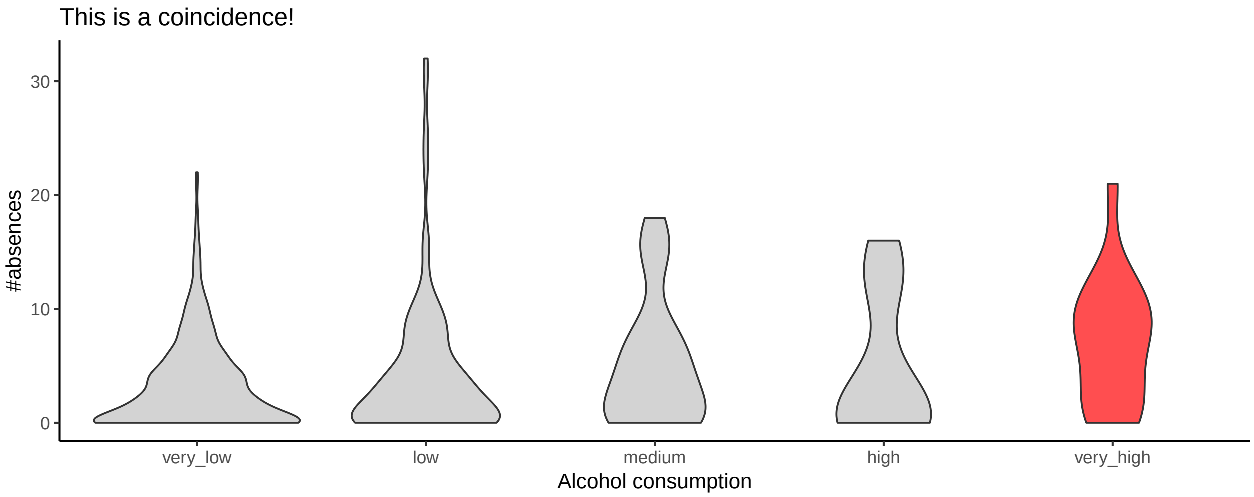
```
c3 <- ggplot(student_por, aes(x=Dalc, y=G3, color=sex))+
      geom_jitter(alpha=0.8)+
       scale_colour_manual(values=c("#e15759", "#a0cbe8"))+
      xlab("Workday alcohol consumption")+
      ylab("G3")+
      ggtitle("Workday alcohol consumption per school and sex")+
      geom_hline(yintercept=10, linetype="dashed",
                color = "black", size=1)
```

```
c4 <- ggplot(student_por, aes(x=Walc, y=G3, color=sex))+
      geom_jitter(alpha=0.8)+
       scale_colour_manual(values=c("#e15759", "#a0cbe8"))+
      xlab("Weekend alcohol consumption")+
      ylab("G3")+
      ggtitle("Weekend alcohol consumption per school and sex")+
      geom_hline(yintercept=10, linetype="dashed",
                color = "black", size=1)


gridExtra::grid.arrange(c3,c4, nrow=2)
```



The following visualization is just being included because coincidentally, the plot for very high alcohol consumption looks like a bottle! Take from that what you may.

```
ggplot(student_por, aes(x=Dalc, y=absences, fill=Dalc))+
      geom_violin()+
      scale_fill_manual(values = c("#d3d3d3","#d3d3d3","#d3d3d3","#d3d3d3","#ff4e50"))+
      theme(legend.position="none")+
      xlab("Alcohol consumption")+
      ylab("#absences") +
      ggtitle("This is a coincidence!")
```



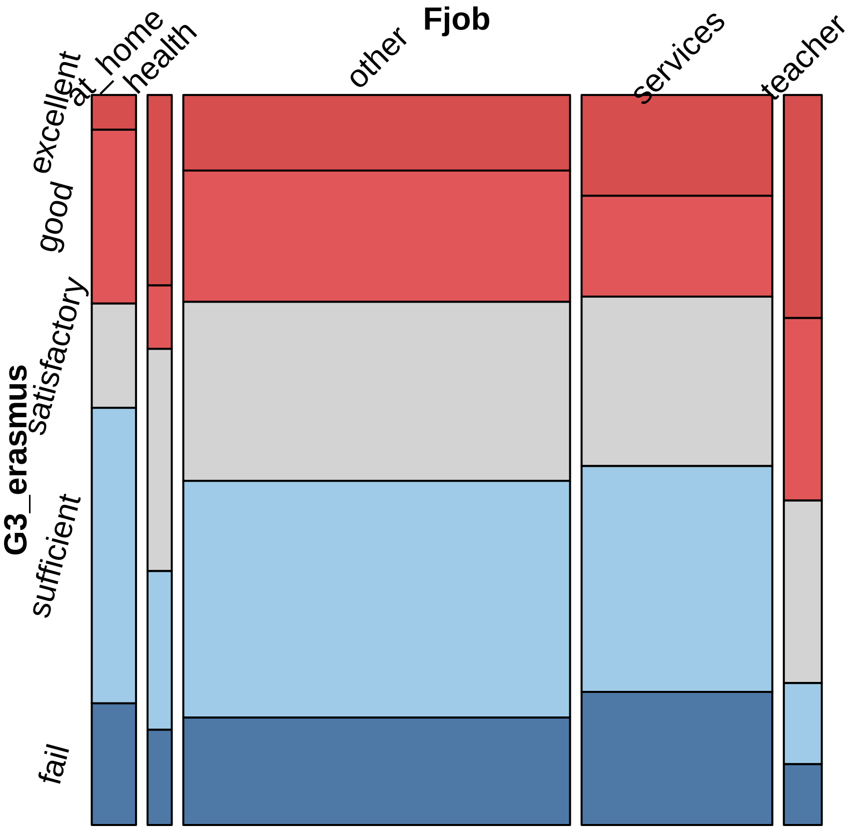
4.3.3 Analysis for Parents Group

We observed parents occupation and education background affects students grade. We will introduce analysis results in the following section.

First of all, students whose parents work as teachers get better grades than other students.

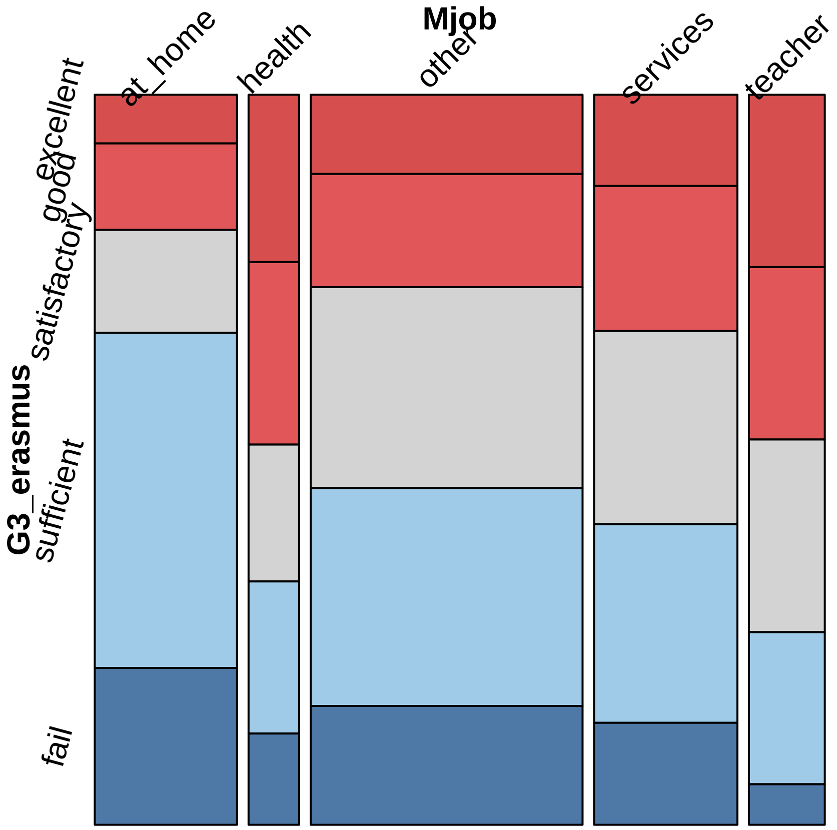
Students whose fathers are teachers perform significantly better than other students. Students whose fathers work in health industry perform relatively well too.

```
mosaic(G3_erasmus ~ Fjob,student_por,
            gp = gpar(fill = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")),
            rot_labels=c(45,0,0,75),
            direction=c("v","h"))
```



Similarly, students whose mothera are teachers perform significantly better than other students. Students whose mothers are working in health industry also do well. In fact, having a mother working in the health industry seems to have a bigger impact on student's grades compared to having a father working in the health industry.

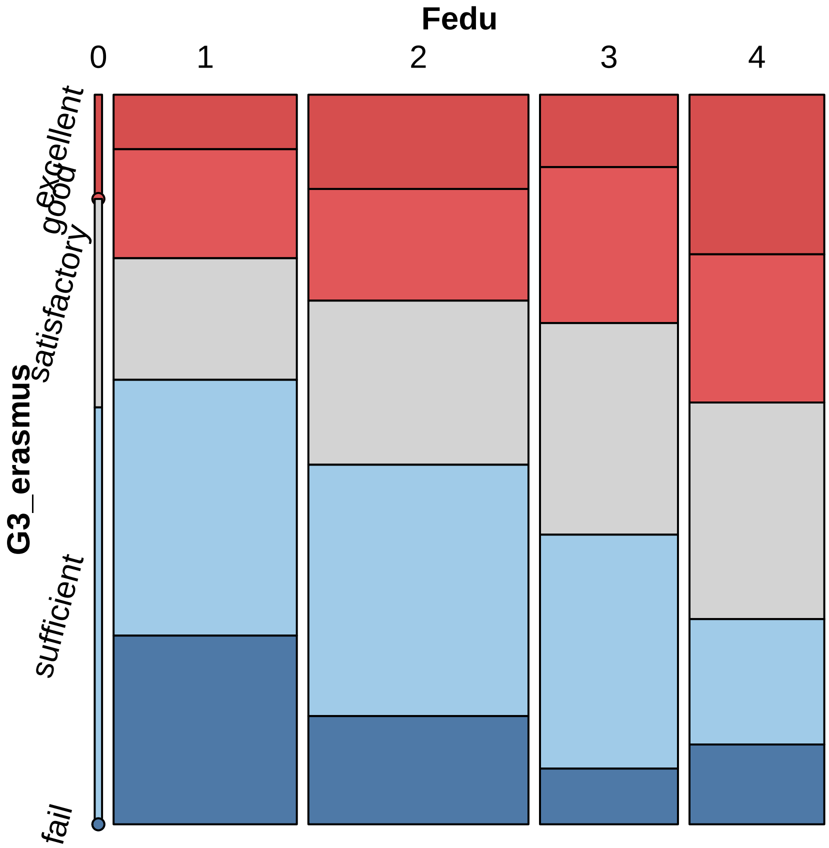
```
mosaic(G3_erasmus ~ Mjob,student_por,
            gp = gpar(fill = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")),
            rot_labels=c(45,0,0,75),
            direction=c("v","h"))
```



Secondly, parent education is also associated with students' grade. As indicated in 2.2, bigger the number in Fedu and Medu means higher the education.

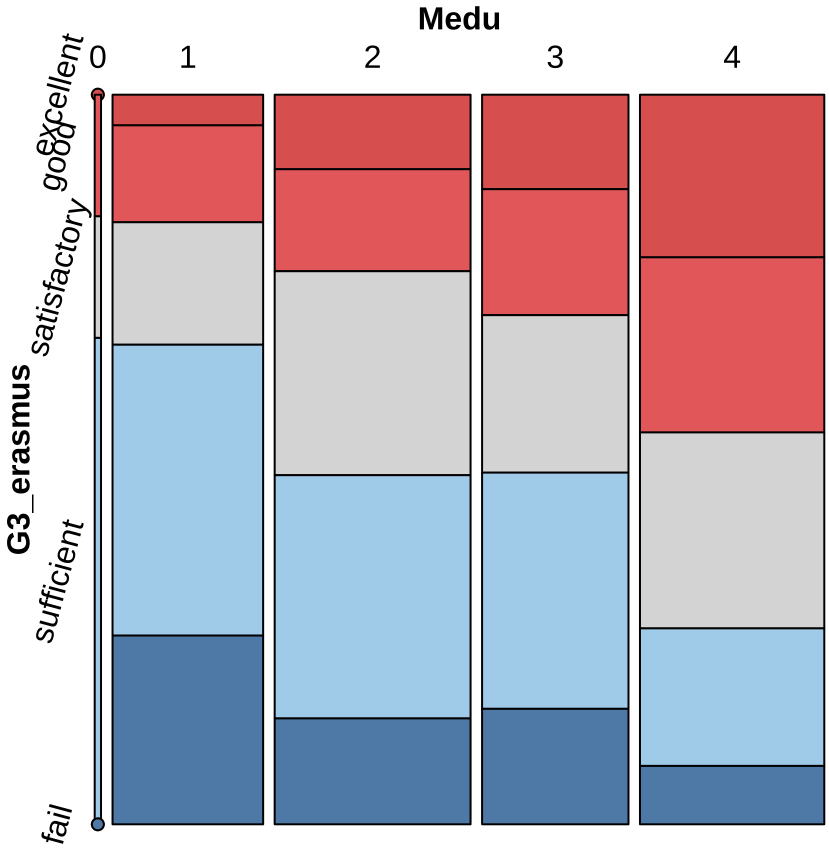
When father took higher education, students perform better in their grades. In the chart below, we can see students grades become better, as father's education goes higher.

```
mosaic(G3_erasmus ~ Fedu,student_por,
            gp = gpar(fill = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")),
            rot_labels=c(0,0,0,75),
            direction=c("v","h"))
```



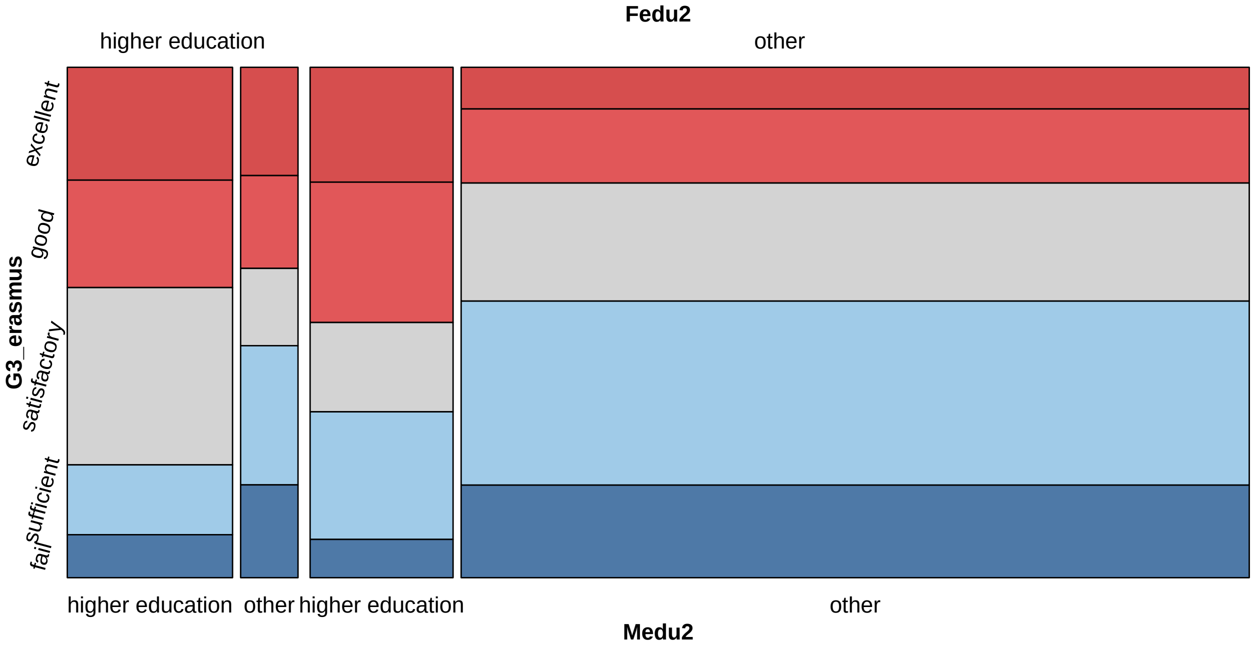
Similarly, when mother took higher education, students perform better in their grades.

```
mosaic(G3_erasmus ~ Medu,student_por,
            gp = gpar(fill = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")),
            rot_labels=c(0,0,0,75),
            direction=c("v","h"))
```



From the chart below, it appears to be important that at least either mother or father took higher education. When both parents did not take higher education, students grades appears to be worse than other case. We can oberve this by looking at the most right combination (Fedu2=other and Medu2=other).

```
student_por <- student_por %>%
  mutate(Medu2=if_else(Medu==4, 'higher education','other')) %>%
  mutate(Fedu2=if_else(Fedu==4, 'higher education','other'))

mosaic(G3_erasmus ~ Fedu2+Medu2,student_por,
            gp = gpar(fill = c("#D64E4E","#e15759","#d3d3d3","#a0cbe8","#4e79a7")),
            rot_labels=c(0,0,0,75),
            direction=c("v","v","h"))
```

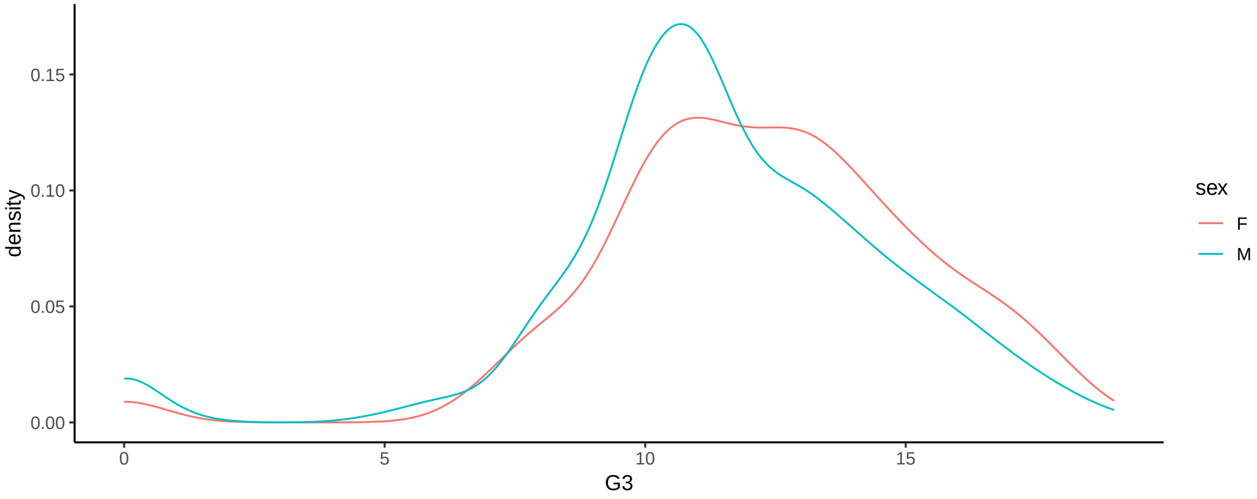


4.3.4 Analysis for Profile Group

Overall, this group appears to be not too related to performance. Having said that, we will describe what we analyzed as follows.

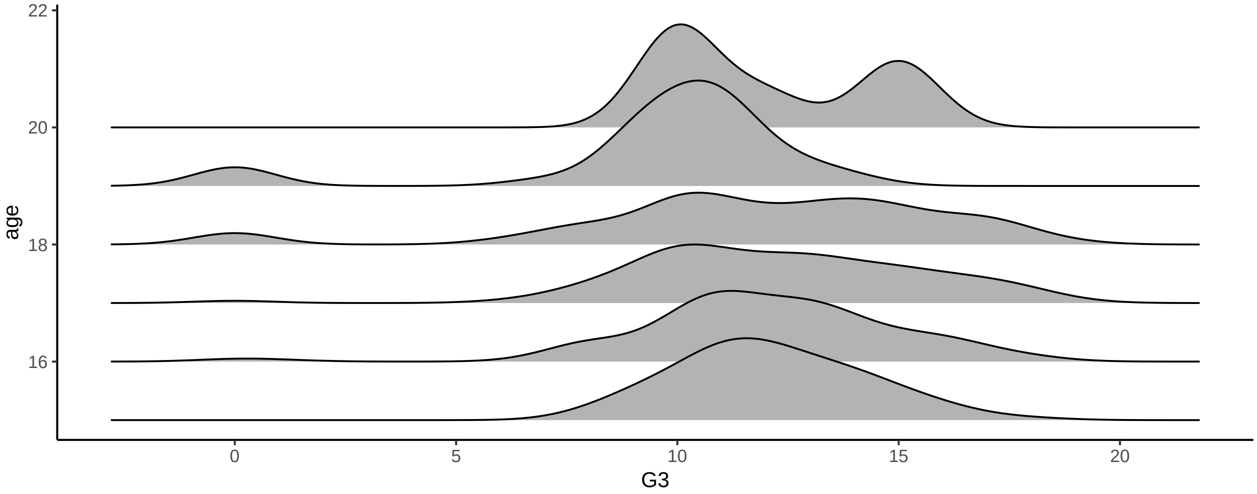
When it came to gender, we don't see any significant difference between male and female grade performance. Grades distribution was almost identical.

```
student_por %>%
  ggplot(aes(x = G3)) +
  stat_density(aes(group = sex, color = sex),position="identity",geom="line")
```



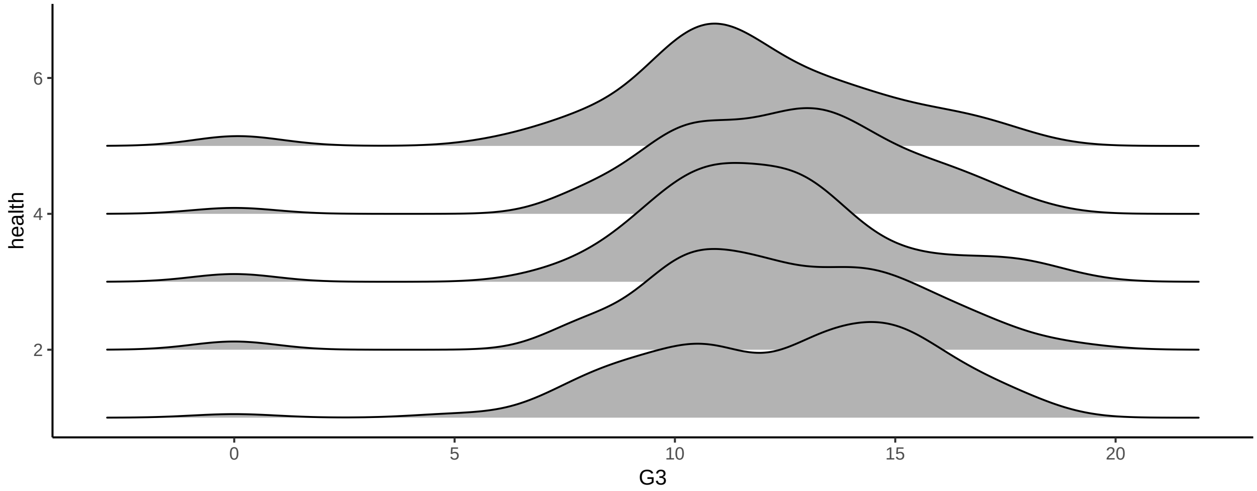
Regarding age, we don't see so much difference regarding grade performance among ages. Overall, their average grades are around 12. However, we can that distribution of students grades have less variance as students are younger. Also, for students whose age are 20, we can see the distribution is bimodal.

```
student_por %>%
  ggplot(aes(x = G3, y = age, group = age)) +
  geom_density_ridges()
```



As for health, we don't see any significant difference regarding grades performance among students' health condition. Grades distribution was almost identical.

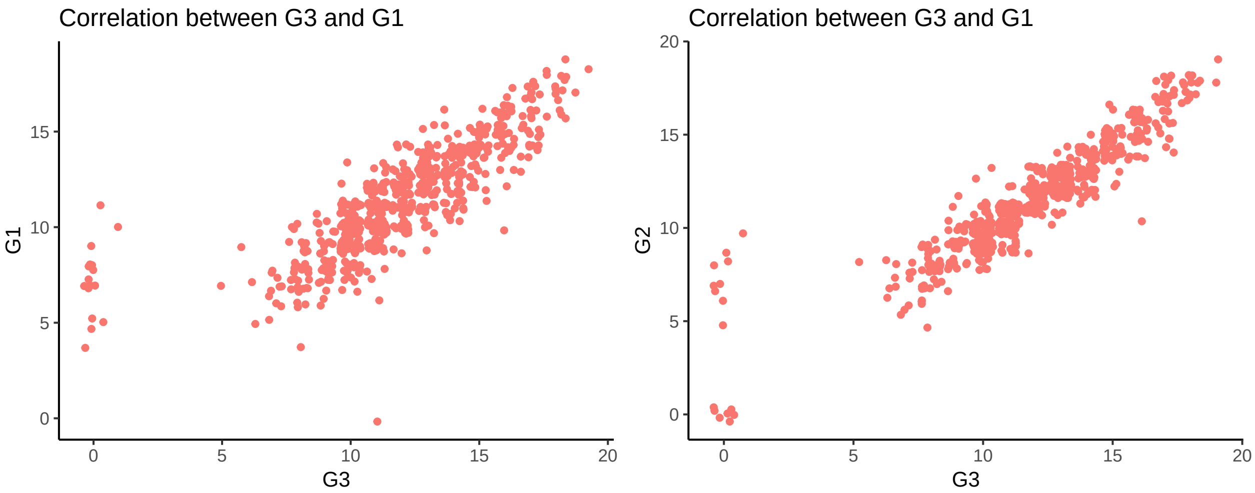
```
student_por %>%
  ggplot(aes(x = G3, y = health, group = health)) +
  geom_density_ridges()
```



4.3.5 Analysis for Academics Group

We observed that final grades are strongly correlated with 1st period grade and 2st period grade.

```
g1 <- ggplot(student_por, aes(G3,G1,color = "blue"))+
  geom_jitter() +
  theme(legend.position = "none") +
  xlab("G3") +
  ylab("G1") +
  ggtitle("Correlation between G3 and G1")

g2 <- ggplot(student_por, aes(G3,G2,color = "blue"))+
  geom_jitter() +
  theme(legend.position = "none") +
  xlab("G3") +
  ylab("G2") +
  ggtitle("Correlation between G3 and G1")

gridExtra::grid.arrange(g1,g2, nrow=1)
```

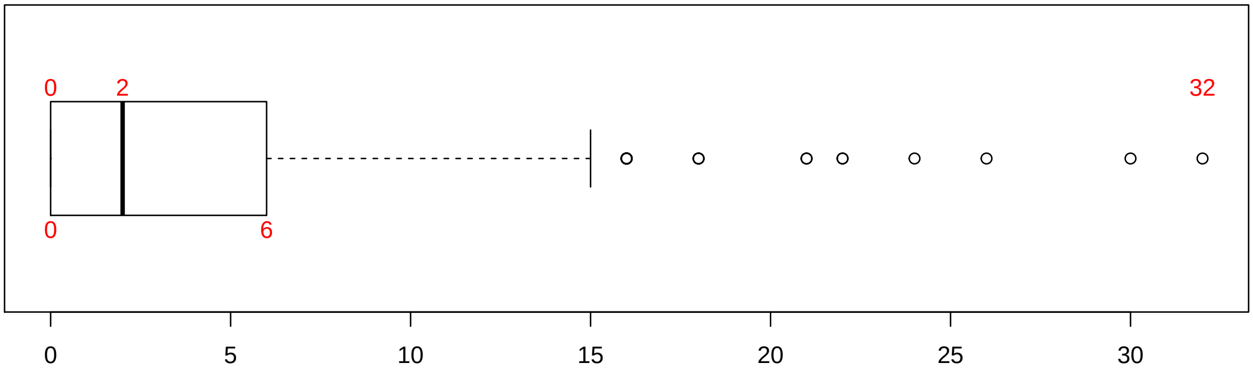


To figure out association between absences frequency and grades, we obtained quantiles for absence frequency. We categorize absences>6 as most frequent, absences = 3 to 6 as second frequent, absences 1 or 2 to average, and no absence.

```
# absences quantiles
boxplot(student_por$absences, horizontal = TRUE, las = 1)
fivenumnames <- c("min", "lower-hinge", "median", "upper-hinge", "max")
D <- student_por$absences
fivenum(D) %>% set_names(fivenumnames)
```

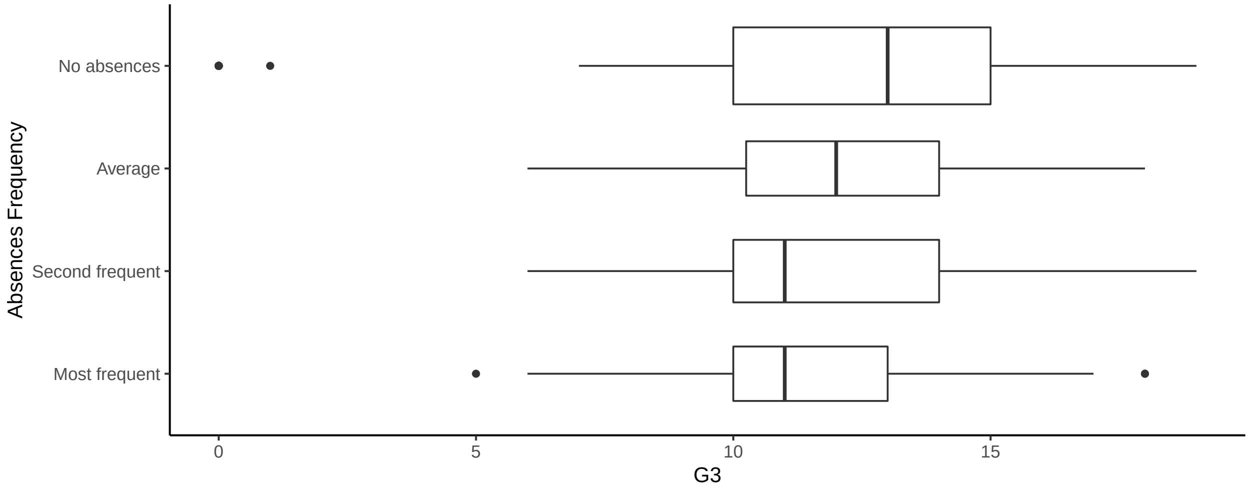
```
##         min lower-hinge      median upper-hinge         max
##           0           0           2           6          32
```

```
text(fivenum(D)[c(1,3,5)], 1.25, round(fivenum(D)[c(1,3,5)],1), col = "red")
text(fivenum(D)[c(2,4)], .75, round(fivenum(D),1)[c(2,4)], col = "red")
```



We can see students most frequently absent from school have poorer grades than students with lesser absences.

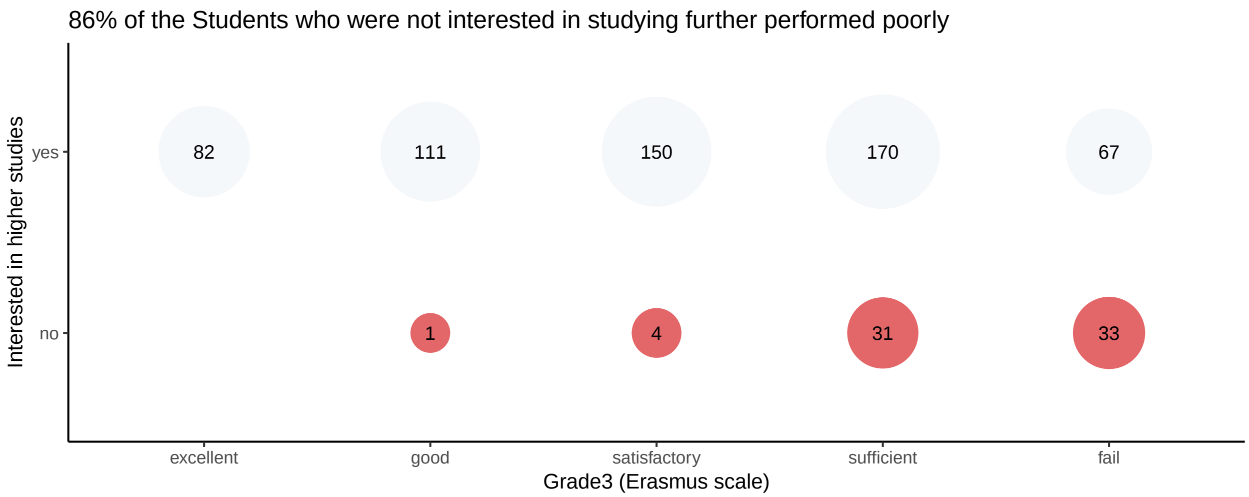
```
student_por %>%
  mutate(AbsenceFrequency = if_else(absences>6, 'Most frequent',
                                   if_else(absences>2&absences<=6, 'Second frequent',
                                                  if_else(absences>0&absences<=2, 'Average',
                                                          'No absences')))) %>%
  mutate(AbsenceFrequency = factor(AbsenceFrequency,
                                   c('Most frequent','Second frequent','Average','No absences'))) %>%
  ggplot(aes(x = reorder(`AbsenceFrequency`, G3, median),
                  y = G3)) +
  geom_boxplot(varwidth = TRUE) +
  coord_flip() +
  xlab("Absences Frequency")
```



4.3.6 Analysis for Others Group

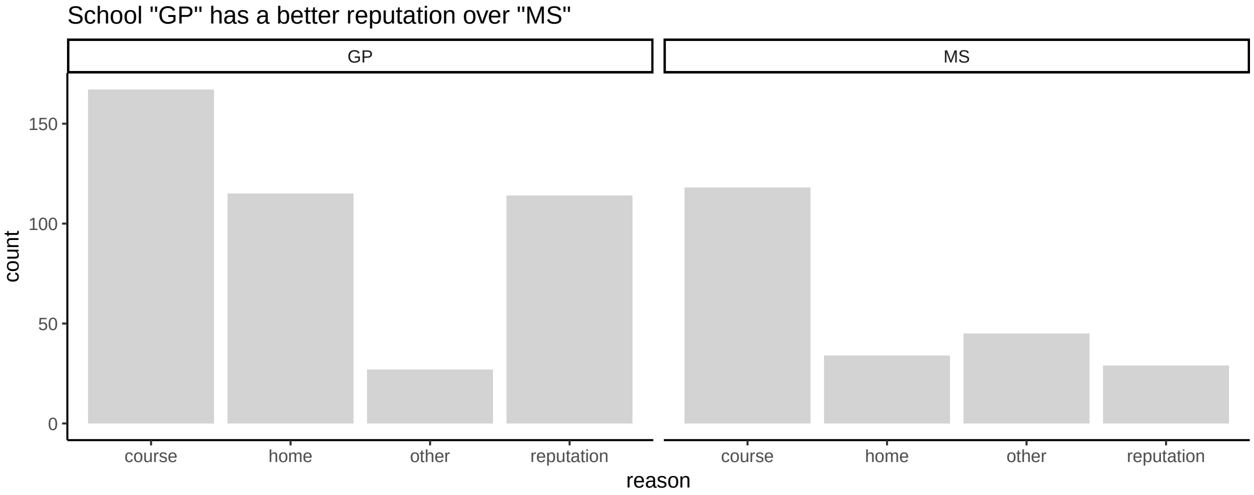
Regarding Higher variable, which represents if students want to take higher education, 86% of the students who were not interested in studying further performed poorly. Therefore, motivation for higher education is significantly associated with grades performance.

```
student_por %>%
  group_by(G3_erasmus, higher) %>%
  tally() %>%
  ggplot(aes(G3_erasmus, higher)) +
  geom_point(aes(size = n,fill = higher, color = higher)) +
  xlab("Grade3 (Erasmus scale)") +
  ylab("Interested in higher studies") +
  scale_size_continuous(range=c(10,30)) +
  geom_text(aes(label = n)) +
  theme(legend.position = "none") +
  scale_color_manual(values = alpha(c("#e15759","#4e79a7"),c(0.9,0.05)))+
  ggtitle("86% of the Students who were not interested in studying further performed poorly")
```



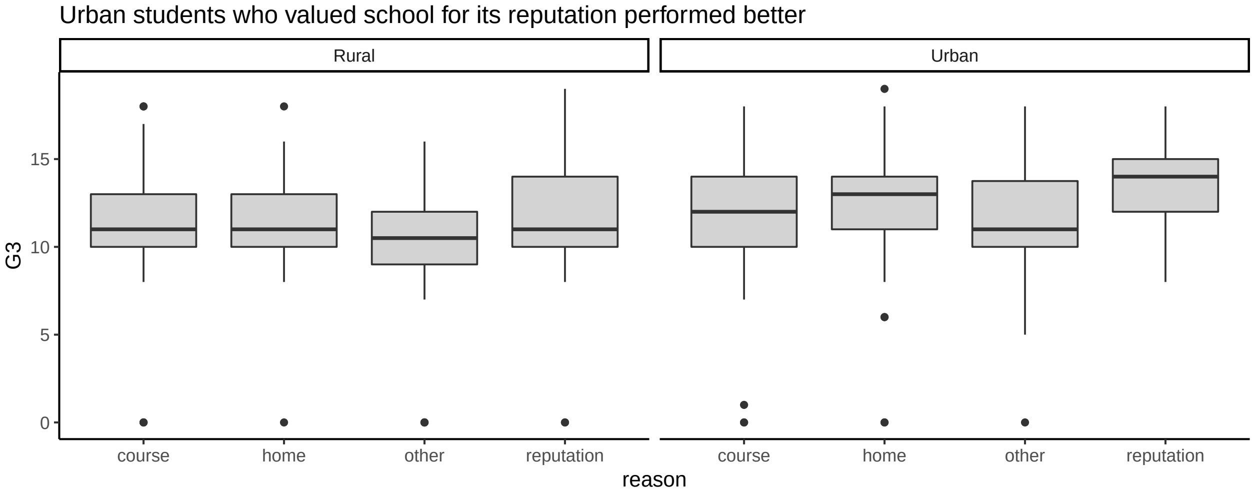
Regarding Reason, which is reason to choose this school, School GP has a better reputation over MS.

```
student_por %>%
  ggplot(aes(x = reason)) +
  geom_bar(fill="#d3d3d3") +
  facet_wrap(~school) +
  ggtitle("School \"GP\" has a better reputation over \"MS\"")
```



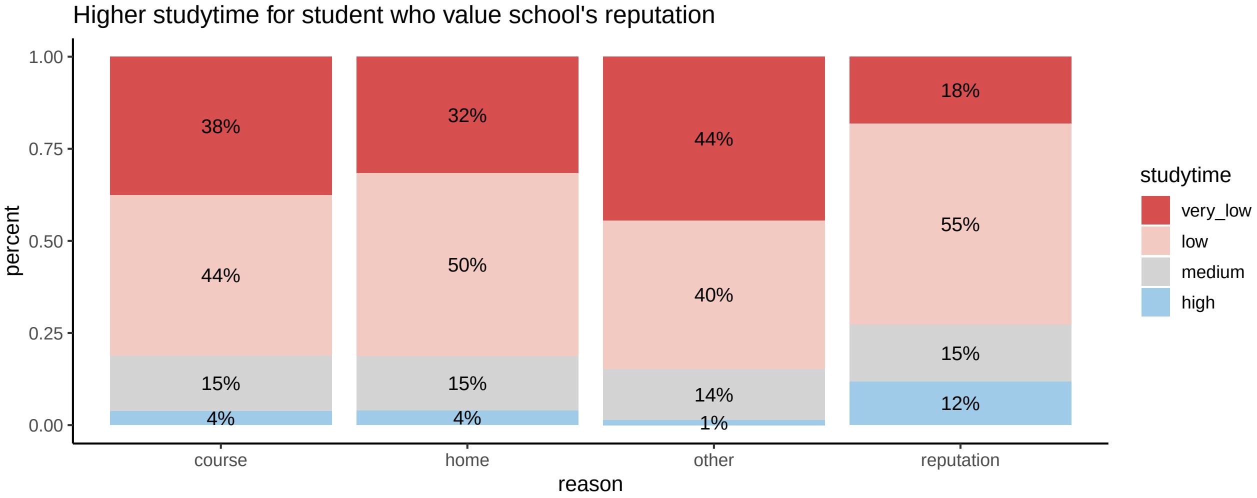
We analyzed what rural and urban kids value while choosing a school. In the chart below, we can see urban students who valued school for its reputation performed better.

```
student_por$address[student_por$address == 'U'] <- "Urban"
student_por$address[student_por$address == 'R'] <- "Rural"

student_por %>%
  ggplot(aes(x = reason, y = G3)) +
  facet_wrap(~address) +
  geom_boxplot(fill="#d3d3d3") +
  ggtitle("Urban students who valued school for its reputation performed better")
```



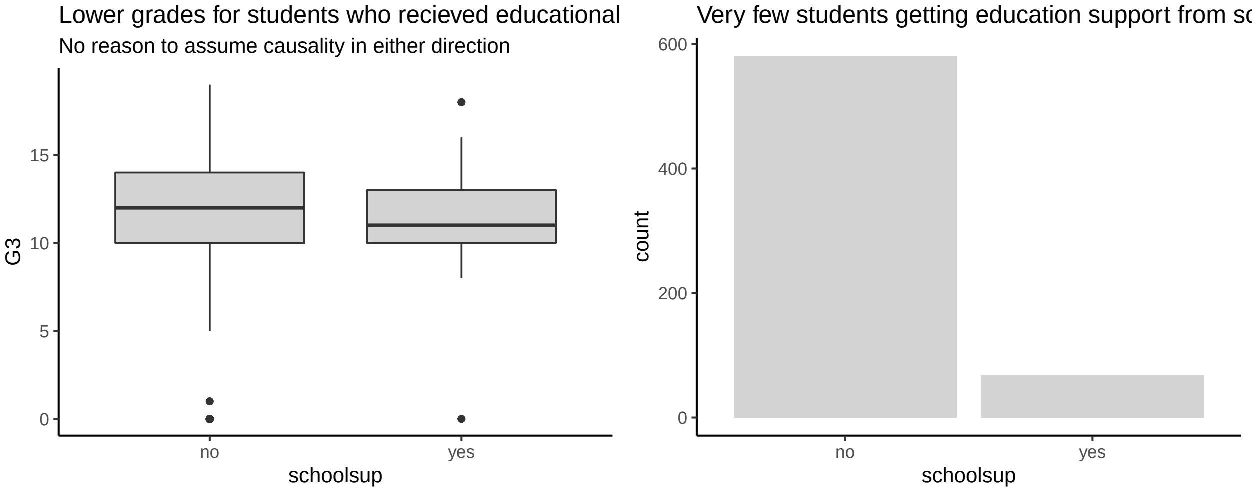
The following chart shows if school reputation is linked to higher study time. Students who values school reputation tend to have higher studytime.

```
student_por$studytime <- factor(student_por$studytime,
                                labels =  c('very_low', 'low', 'medium', 'high'),
                                ordered = TRUE)

student_por %>%
  group_by(studytime, reason) %>%
  tally() %>%
  group_by(reason) %>%
  mutate(percent = n/sum(n))  %>%
  arrange(desc(studytime)) %>%
  ggplot(aes(x = reason, y = percent)) +
  geom_col(aes(fill = studytime)) +
  geom_text(aes(label = paste(round(percent,2) * 100, "%",sep = "")),
            position = position_stack(vjust = 0.5))+
  scale_colour_manual(values = c("#D64E4E","#F2CAC1","#d3d3d3","#a0cbe8","#4e79a7")) +
  scale_fill_manual(values = c("#D64E4E","#F2CAC1","#d3d3d3","#a0cbe8","#4e79a7")) +
  labs(fill = "studytime") +
  ggtitle("Higher studytime for student who value school's reputation")
```



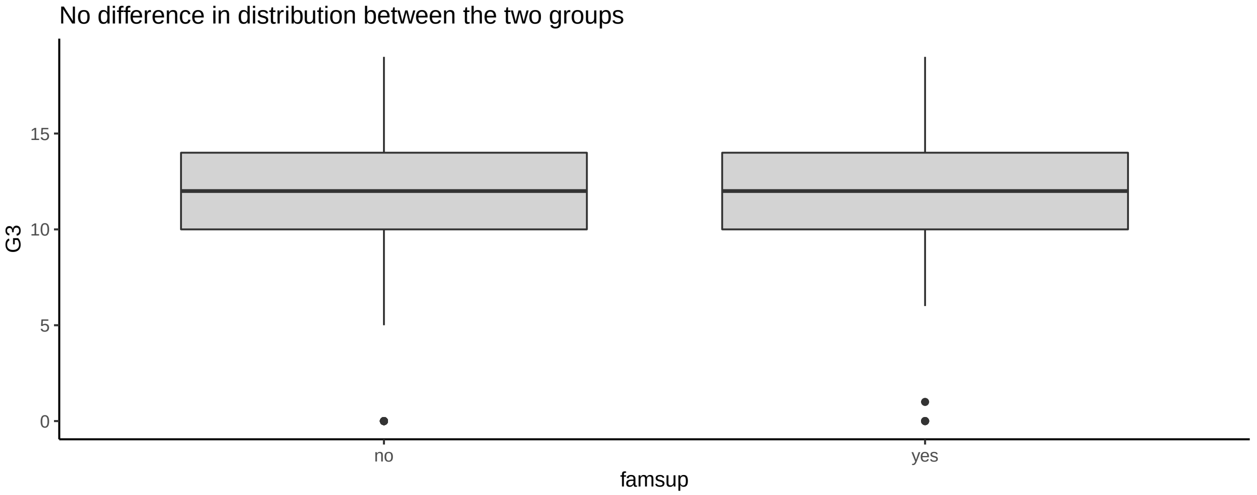
Regarding schoolup (school suppport) variable, students who recieved educational support from school showed lower grades. We don't have any reason to assume causality in either direction. Also, very few students get education support from school.

```
box <- student_por %>%
  ggplot(aes(x = schoolsup, y = G3)) +
  geom_boxplot(fill="#d3d3d3") +
  ggtitle("Lower grades for students who recieved educational support from school",
          subtitle = "No reason to assume causality in either direction")

bar <- student_por %>%
  ggplot(aes(x = schoolsup)) +
  geom_bar(fill="#d3d3d3") +
  ggtitle("Very few students getting education support from school")

gridExtra::grid.arrange(box,bar, nrow=1,ncol=2)
```

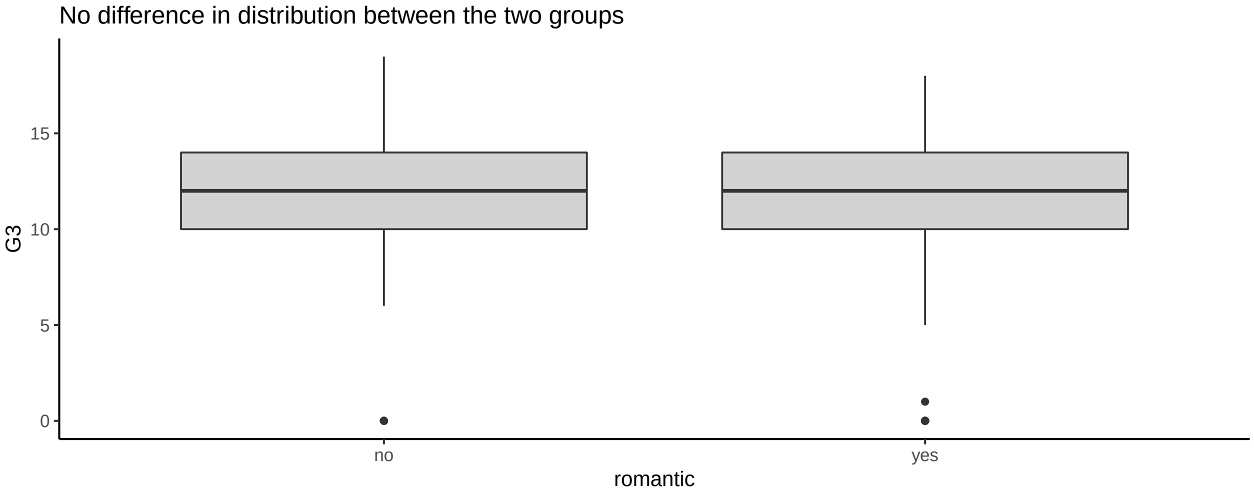


As for famsup (family support) and romantic (in a romantic relationship), we don't see any difference in distribution of grades between two groups.

```
student_por %>%
  ggplot(aes(x = famsup, y = G3)) +
  geom_boxplot(fill="#d3d3d3") +
  ggtitle("No difference in distribution between the two groups")
```



```
student_por %>%
  ggplot(aes(x = romantic, y = G3)) +
  geom_boxplot(fill="#d3d3d3") +
  ggtitle("No difference in distribution between the two groups")
```



4.4 Challenges

Actual data was not so easy to work with. As Prof. Joyce Robbins mentioned in the class, EDA cycle is not clearly sequential. We needed to go back and forth through datasets. As mentioned in 3.3, we also identified some data issues. We realized it is important to conduct data quality anaysis to see if there are any anomalies and human error we need to remove.

Also, we sometimes observed no associations for single variables, however, when we facet by another variable, we could find some patterns. There was no definite method to find such patterns, we needed to formulate our own initial hypothesis and conduct verification of such hypothesis though trial and error. It took a lot of time to do so, and it was challenging to do so while facing deadlines.

5 Executive summary (Presentation-style)

Link: https://github.com/Somendratripathi/edav__hnt__nm__sdt/blob/master/FP__executive__summary__v4.pdf

6 Interactive component

The variables which were found to be most correlated with final grade (G3) were:

1. Grade in first half of semester (G1)
2. Grade in the second half of semester (G2)
3. Relationship with family (famrel)
4. Number of times student has failed in the past (fail)
5. Time devoted to studying (studytime)

A student is said to fail if his/her final grade (G3) is below 9. We wanted to build a decision tree based on the above variables to help us understand trends in the data better. Eg. Will a student pass if he/she has failed twice before but their study time is high?

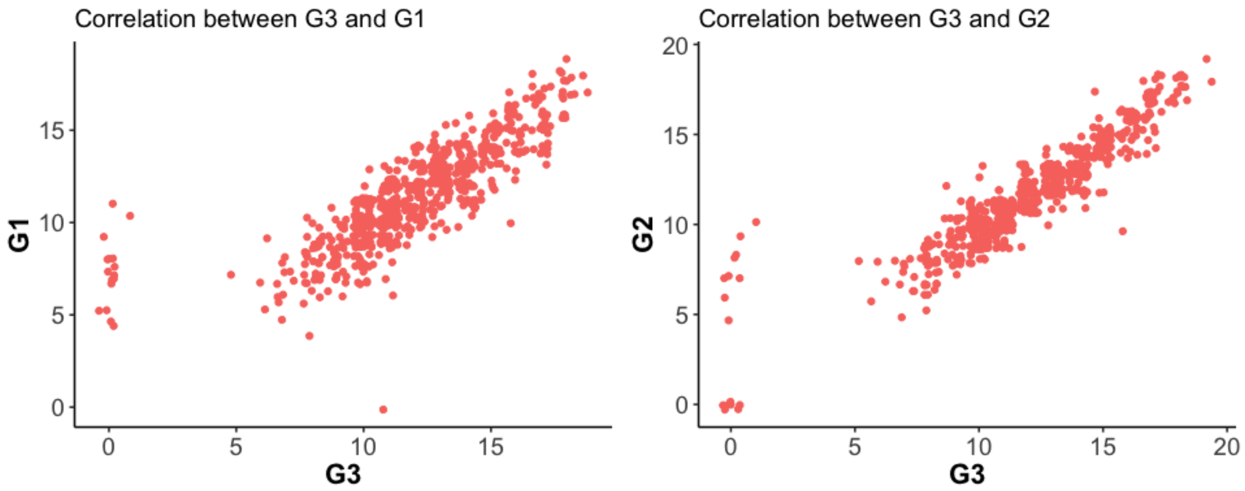


Figure 1:

However, as seen from the graphs below, we can see that G1 and G2 are highly correlated with G3. It implies that G3 is highly sensitive to these two variables and thus, constructing a decision tree which involves these two variables would lead to a tree with minimal depth and the splitting variables would only be G1 and G2. Thus, we decided to exclude these two variables from the task of constructing a decision tree.

Feel free to construct your own hypothetical student using variables famrel, failures and studytime and see if they will Pass or Fail.

Demo: https://imgur.com/N7jS4A7

Link: https://blockbuilder.org/hrishifishy/c0289a2bbaa8b20d52b7c483c5816f77

Technical Execution:

We used the 'rpart' package in R to construct a classification tree for our data. The resultant tree structure was stored in a .json file (structure.json) that was then used to construct our interactive decision tree in D3.

While we chose to focus on the variables that were most correlated with final grade, one could also construct a decision tree using other numeric variables to see what conclusion they lead to.

We would have liked to play around with such a feature in our interactive component, where a user could select variables they wanted to construct a decision tree with and then display a custom tree for the selected variables, but doing so would have become very cumbersome in d3 and we would have had to run the decision tree construction algorithm in the backend after the variables were selected

7 Conclusion

We started with a few broad questions we wanted to answer, and along the way discovered some very interesting insights from the data, which we have detailed in the Executive Summary section of the report.

What we learned through the exploratory phase is that it is highly effective to focus on a specific variable and formulate certain educated questions around it. After a preliminary visualization or after uncovering some trend in the data we found that drilling down by asking the question "why?" proved to be very effective in analyzing trends. We could thus construct a narrative around a variable. For example, in analysis of the Social Group and the variable 'freetime', we found that students with lower grades tend to have higher free time. After asking 'Why?', we discovered that students with higher free times not only go out with friends more often but are also more likely to engage in extra curricular activities. After asking does free time impact any other variable, we found that students with lesser free time reported low health. In this way we learnt that though there may be a definite answer regarding how to conduct EDA, this approach was effective in analysing the data.

One of the major limitations of the dataset was that the data was collected through surveys, which means that most of the answers (apart from Grades) were self-reported by students. This brings in a lot of subjectivity into the data, for example students may have very different definitions of what constitutes high drinking or low family relationship.

Secondly, the data is from high school students studying in Portugal in 2006, which was 12 years ago. Societal and cultural norms may differ in 2018, which are not reflected in the conclusions drawn from the data. Lastly, the data is only derived from two high schools in Portugal and thus cannot be said to represent general trends in students' grades in Portugal or the world at large.

In terms of the future scope of this analysis, we would definitely want to work with a larger and more diverse dataset, ideally having standardized definitions for variables and taken from multiple high schools. We hope our results and methodology can be used to analyse similar datasets and measure the impact of external factors on a student's grades.