Student performance dataset report

1. Introduction :

I choose this dataset because it have more than 100 instance and more than 10 variables, also because the theme of this dataset is simple and does not need to have background or do a research in order to understand the variables.

2. Dataset description:

This dataset was collected from two Portuguese schools; it has two files: one mathematics course and the other one for Portuguese course. Math data has 395 row, while Portuguese data has 649. For variables, they have 33 variable are summarized in the table below: (Cortez.P 2014)

school	Character : "GP" for Gabriel Pereira, "MS" for Mousinho da Silveira
sex	Character: "F" female, "M" male
age	numeric: from 15 to 22
address	Character: "U" urban, "R" rural
famsize	Size of family (Character: "LE3" when ≤3, "GT3" – when >3)
Pstatue	cohabitation status of parents (Character: "T" - living together, "A" - apart)
Medu	Education level for mother (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, 4 – higher education)
Fedu	Education level for father (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, 4 – higher education)
Mjob	Job of mother (Character: "teacher", "health","services", "at_home", "other")
Fjob	Job of father (Character: "teacher", "health", "services", "at_home", "other")
reason	reason for choosing this school (Character: close to "home", school "reputation", "course" preference,"other")
guardian	character: "mother", "father", "other"
traveltime	Travelling time between school and home (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, 4 - >1 hour)
studytime	study time per week (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	Failures in past clases (numeric: n if 1<=n<3, else 4)
schoolsup	extra support in education (Character: yes or no)
famsup	family support in education (Character: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (Character: yes or no)
activties	extra-curricular activities (Character: yes or no)
nursery	attended nursery school (Character: yes or no)
higher	wants to take higher education (Character: yes or no)
internet	Internet access at home (Character: yes or no)
romantic	in romantic relationship (Character: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
abcences	number of absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)

G3	final grade (numeric: from 0 to 20)

This dataset have no missing data, and there is 382 student that take Portuguese course and math course.

3. Data cleaning:

I started by finding the students that take both Portuguese and math course. And then calculate the mean of G1 and G2 and G3 to get the mean of each student. At the end I combined the two data sets and added a new column called “subject” that indicates which course is this from: math or Portuguese. Create a new column called “support” which is the combination between the column “famsup” and “schoolsup” that indicate if the student is getting extra educational support (schoolsup), family educational support (famsup), both or none. I also create a new column “grade” to tell if the student has good, fair or poor average.

4. Data exploration:

4.1. Correlation:

the correlation plot in the left of figure 1 shows that there is strong relation between grades G1, G2, G3 and the average (obvious for the average to be correlated with them), and between father education and mother education with 0.65. There is also a weak correlation between number of failures and the grades, and between workday alcohol and weekend alcohol consumption.

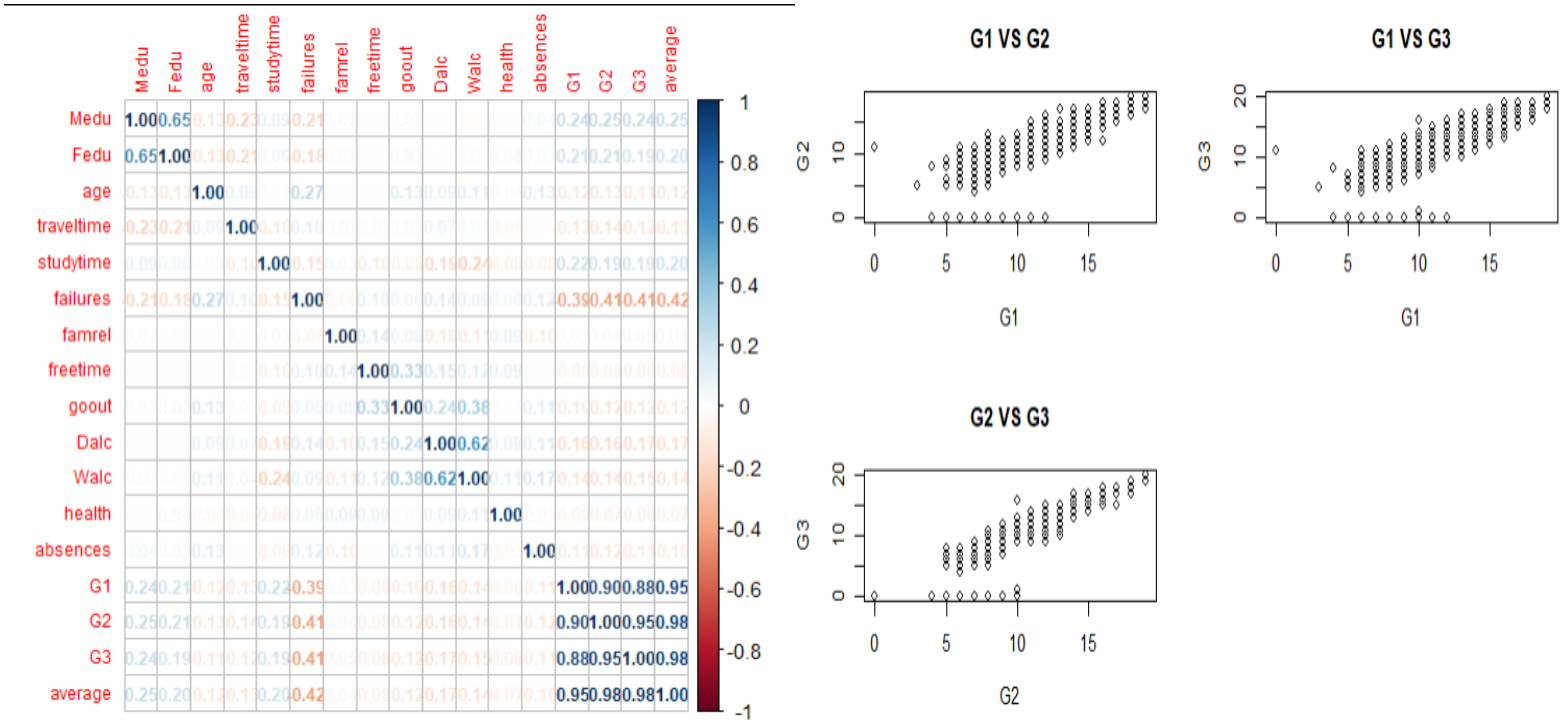

Figure 1: correlation plot and relation between G1, G2 and G3

The plot in the right of figure 1 shows clearly the relation between grades: when students do well in first period, they will do well in period 2 and 3, and the contrast.

Since the correlation plot didn’t show any interesting relations, I will try to find relations between other variables are not in the previous plot.

4.2. Choice of school:

Figure 2 shows that most students of this data set are from Gabriel Pereira school (GP) and the reason why they chose this school, the majority said that because of course preferences and the reason in second place was it's clause to their homes and the reputation of this school. From the plots we can conclude that GP school have better reputation and better course choices for students and it's the closest one to them.

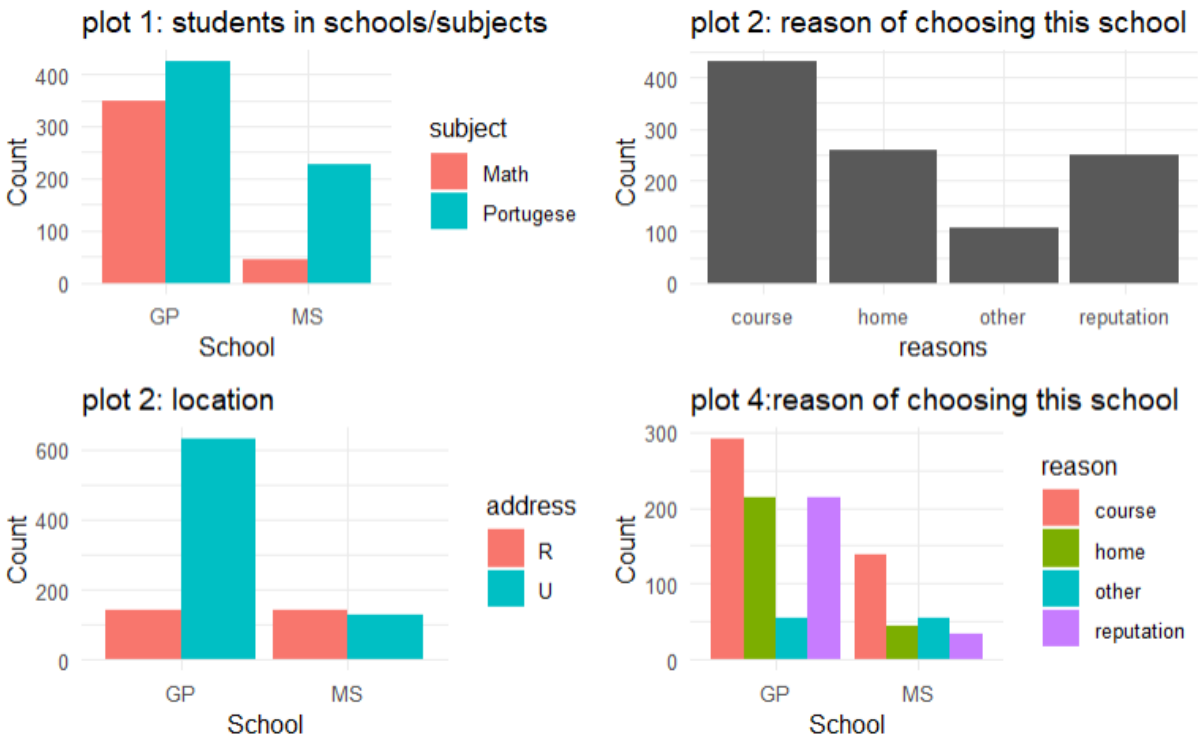

Figure 2: school choice

4.3. Does family affect student performance?

Figure 3 shows that the majority of students have fair grades (between 10 and 15). And students who have good grades (between 15 to 20) are minority. Plot 1 illustrates that the highest number of students who get good grades their mothers are level 4 of education (higher education), and the lower level of education of mother is, the less number of students who got good marks is.

For plots 2, 3, 4, 5 and 6 it seems that there is no trends.

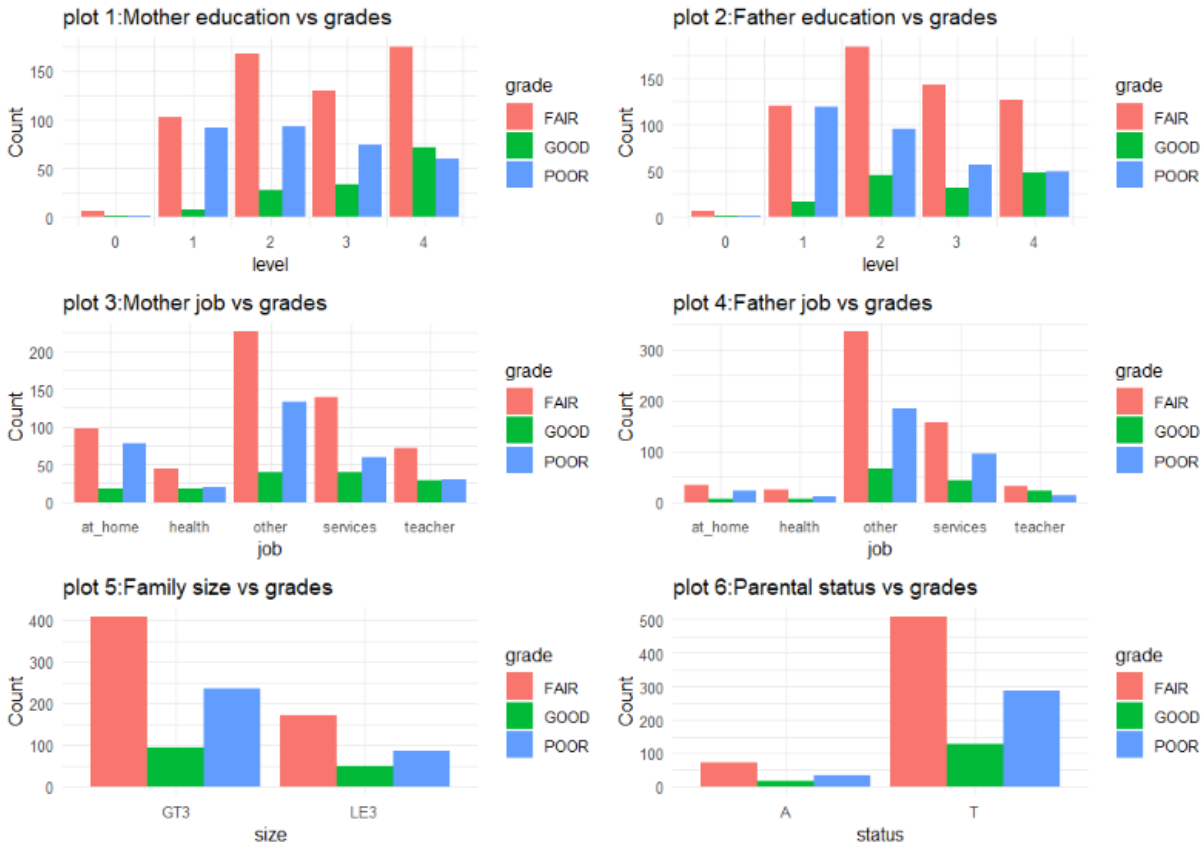

Figure 2: family vs grades

4.4. What actions effect student performance?

Figure 3 contains different activities vs the grades. Plot1 is right skewed plot the only information that it gives, is that there is relation between address and time travelling and most students spend between 1-15 min to arrive to school. Plot 2 and 3 take the shape of normal distributions which means most students are average in how much they have free time, and how many times they go out with friends. For plot 5, it's also a right skewed distribution. The rest of graphs don't give any significant information. In general, there is no clear relation between usual activities of students and their performance.

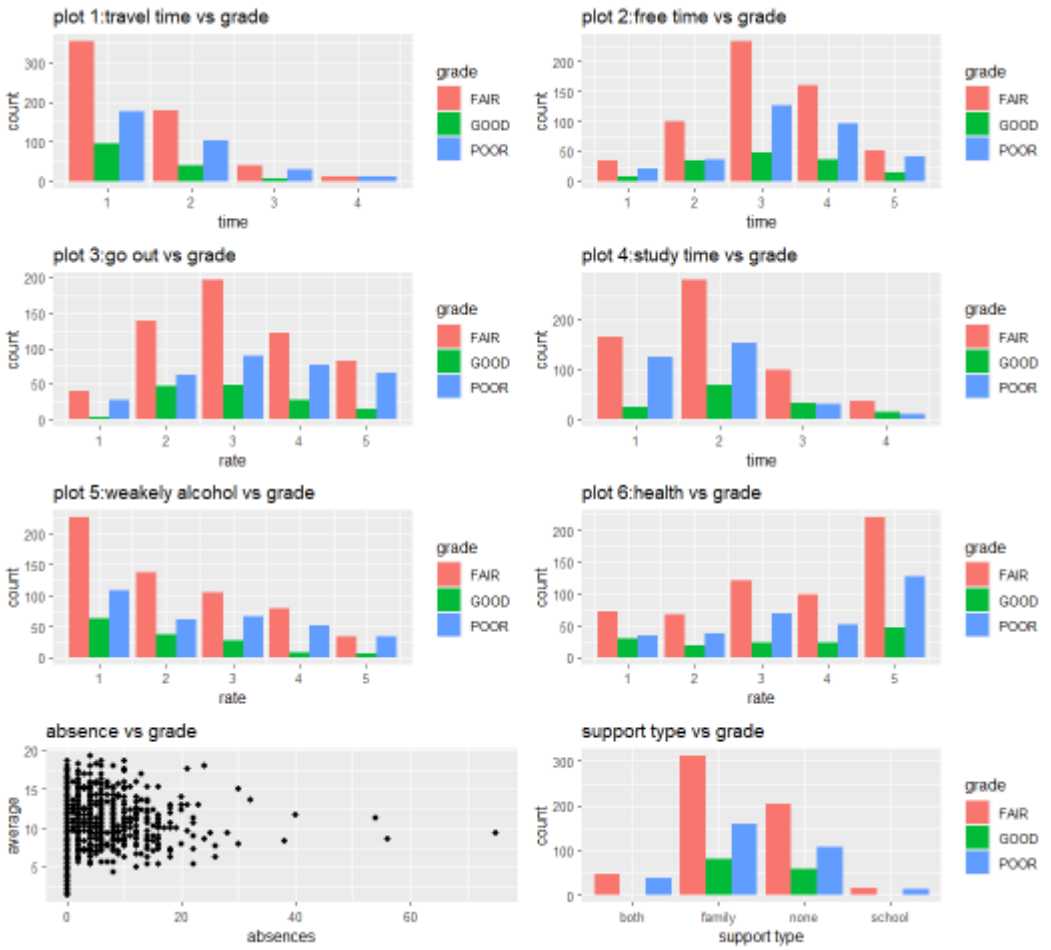

Figure 3: activities vs grades

5. Conclusion:

In conclusion, the students chose their school in first place basing on the course preference, then distance to home and finally reputation of that school.

The grades of periods 1, 2 and 3 are highly related: when G1 goes higher G2 and G3 also go higher, and if G2 goes high G3 also will go high. Another relation is the higher level of education the mother is, the higher grades there is. The last trend is that the majority of students have an average free time.

References:

Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Example code:

```
library(dplyr);library(reshape);library(corrplot);library(gridExtra)
library(tidyr);library(ggplot2) # libraries
math=read.csv("student-mat.csv", stringsAsFactors=T,sep=";",header=TRUE)
portuguese=read.csv("student-por.csv",stringsAsFactors=T,sep=";",header=TRUE)
#students that study both modules
math_port=merge(math,portuguese,by=c("school","sex","age","address","famsize","Pstatus","Me
du","Fedu","Mjob","Fjob","reason","nursery","internet"))
print(nrow(math_port)) # 382 students
# calculate the average of each student
math$average=round(apply(select(math, G1, G2, G3), 1, mean), 2)
portuguese$average=round(apply(select(portuguese, G1, G2, G3), 1, mean), 2)
# create subject row
math$subject=rep("Math", nrow(math))
portuguese$subject=rep("Portugese", nrow(portuguese))
data=rbind(math, portuguese) # merge data and Portuguese
data$support= "0" # create support column and delete famsup and schoolsup
data[data$schoolsup=="yes"& data$famsup=="yes",]$support= "both"
data[data$schoolsup=="yes"& data$famsup!="yes",]$support= "school"
data[data$schoolsup!="yes"& data$famsup=="yes",]$support= "family"
data[data$schoolsup!="yes"& data$famsup!="yes",]$support= "none"
data$support = as.factor(data$support)
data = select(data, -c("schoolsup","famsup"))
grades= data # grades column
good=data[((data$average>=15) & (data$average<= 20)),]
fair=data[((data$average>=10) & (data$average< 15)),]
poor=data[((data$average>=0) & (data$average< 10)),]
good$grade=c("GOOD"); fair$grade=c("FAIR"); poor$grade=c("POOR")
grades = list(good,poor, fair);grades = (merge_recurse(grades))
str(data);summary(data)
# figure 2
d=ggplot(data, aes(x = school, fill = reason))+geom_bar(position = 'dodge') +labs(title = '
plot 4:reason of choosing this school', x = 'School', y = 'Count') + theme_minimal()
b=ggplot(data, aes(x = reason)) +geom_bar(position = 'dodge') +labs(title = 'plot 2: reason
 of choosing this school', x = 'reasons', y = 'Count') +
    theme_minimal()
c=ggplot(data, aes(x = school, fill = address)) +
    geom_bar(position = 'dodge') +labs(title = 'plot 2: location', x = 'School', y = 'Count
') +theme_minimal()
a=ggplot(data, aes(x = school, fill = subject)) +geom_bar(position = 'dodge') +labs(title =
 'plot 1: students in schools/subjects', x = 'School', y = 'Count') +theme_minimal()
grid.arrange(a,b,c,d,ncol=2)
par(mfrow=c(2, 2)) # scatter plots in figure 1
plot(data$G1,data$G2,xlab="G1",ylab="G2",main="G1 VS G2")
plot(data$G1,data$G3,xlab="G1",ylab="G3",main="G1 VS G3")
plot(data$G2,data$G3,xlab="G2",ylab="G3",main="G2 VS G3")
# correlation plot
num_data=data[,numeric_var];correlation2=cor(num_data, method ='spearman')
corrplot(correlation2, method="number",tl.cex=0.7,number.cex=0.7)
plot_list=list()# plots in figure 3
variables=c("traveltime","freetime","goout","studytime","Walc","health","absences","support
")
for (var in variables) {
  plot=ggplot(grades, aes_string(x = var, fill = "grade")) +geom_bar(position = 'dodge') +l
abs(title = paste("Plot:", var, "vs grade"), x = var)
  plot_list[[length(plot_list) + 1]]= plot}
scatter_plot=ggplot(grades, aes(x = absences, y = average)) +geom_point() +labs(title = 'Ab
sence vs Grade', x = 'Absences')
plot_list[[length(plot_list) + 1]]=scatter_plot
grid.arrange(grobs = plot_list, ncol = 2)
```