Data Mining Project

Mohammed Zaki Nassar

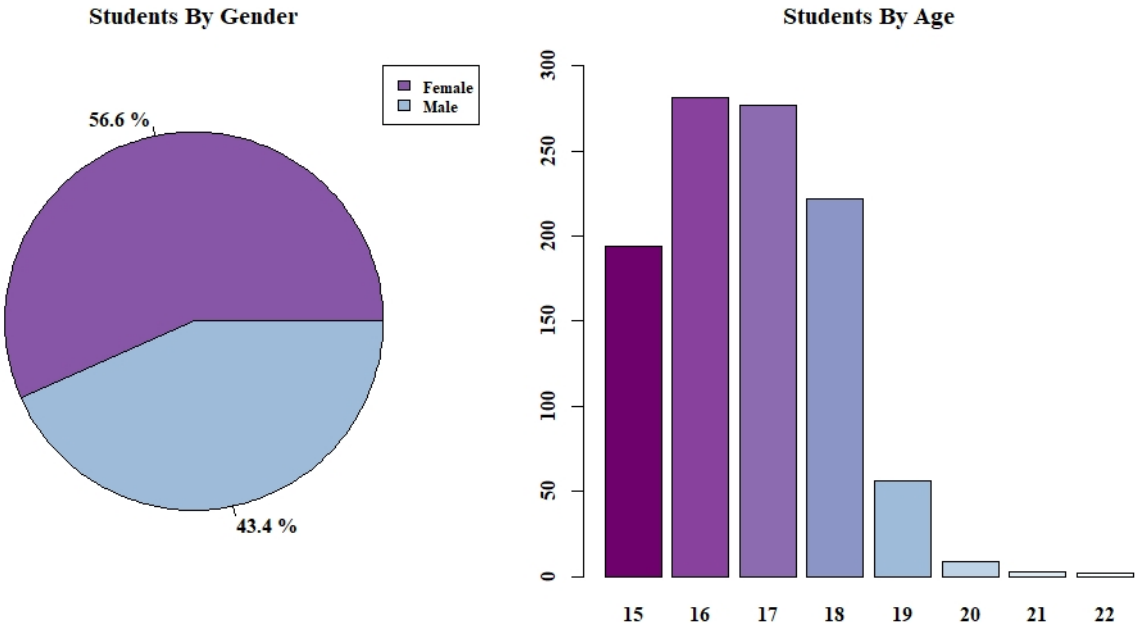
1 Problem understanding

In our dataset we have a lot of details about students' lives, who study math and Portuguese language in a secondary school. The data was collected via a survey of the students in question. I am going to take a deeper look at the composition of the students in the data set. And then take a look at how different aspects of their lives affect their final grades. And, of course, take a look at how alcohol consumption can affect the success of students as well.

The data contains 1044 entries, with 33 observations. This amount of details can provide a chance to take a closer look at what affects the success of students in their education and what other factors are of no consequence when it comes to final grades in secondary school tests.

2 Data understanding

As a first step we take a look at the composition of the students by gender and age:

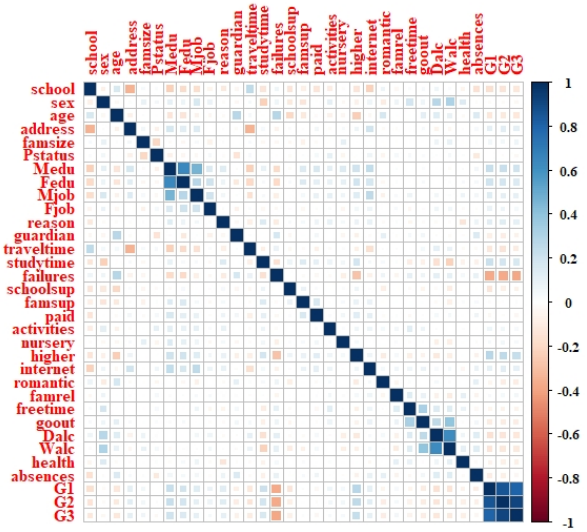


Students by gendre and age.

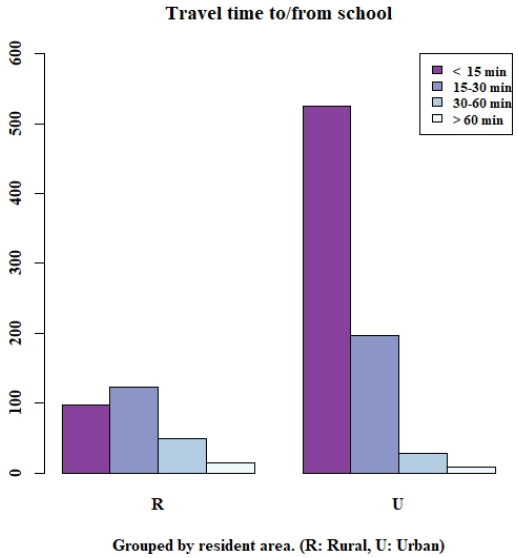
So, we are dealing a an age range of mostly teenagers between 15-18 years old, where the number of girls a little more, but not by a huge margin.

3 Data Preparation

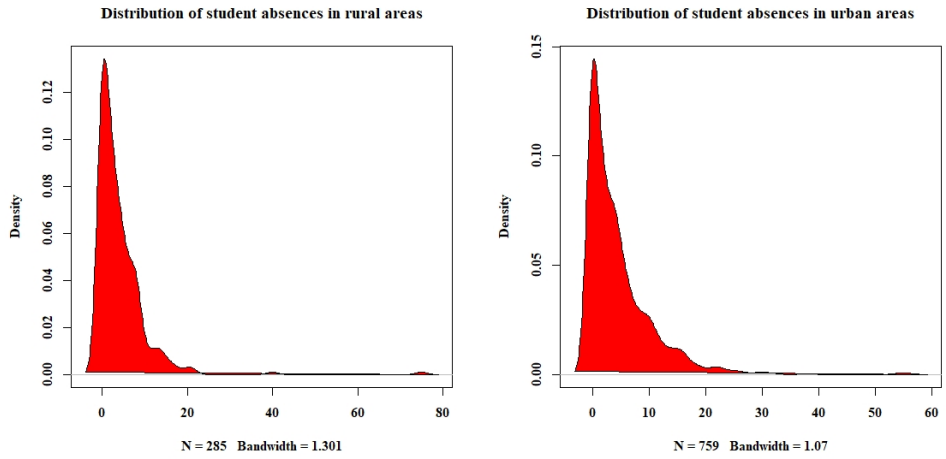
For the data preparation we try to find the factors that affect the students grades the most. So, we start with a correlation graph:



Now, we can take a closer look at some of the variables that affect the grades the most and some others that I suspect can be of some significance. First, let's take a look at how long does it take most students to get to school and back home:



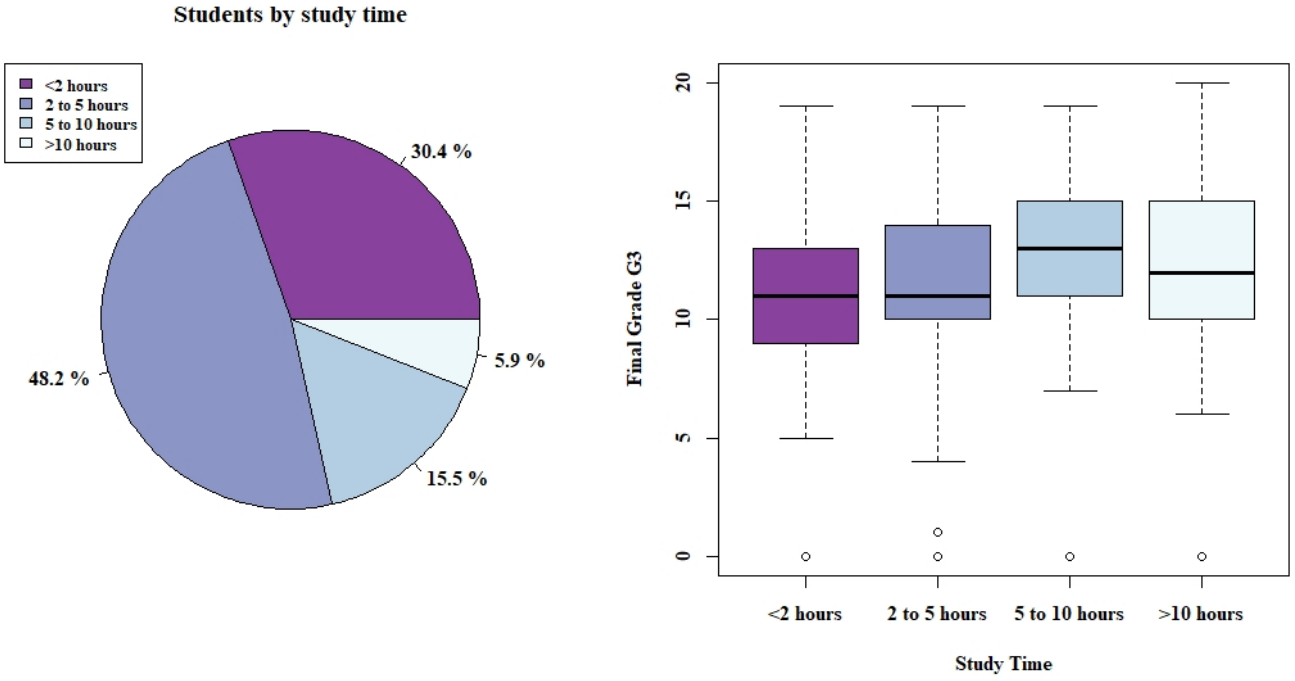
Let's take a look at how this travel time affects the number of absences from school:



Distribution of absences for students by urban and rural residence.

We can notice that students in rural areas have a slightly higher percentage of absences compared to city students. However, this can be attributed to factors other than their distance from school.

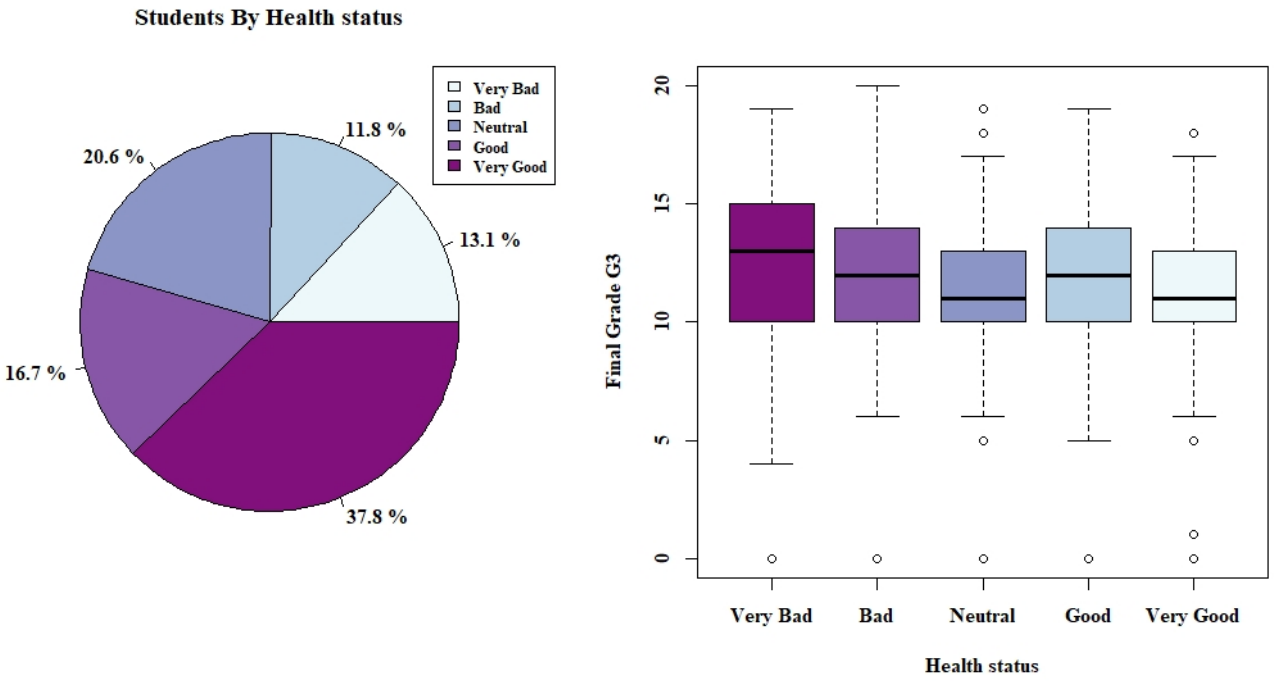
One of the most important factors, of course, is the amount of time the student spends each week on his studies. In the following graph we see the students by their reported time spent studying each week and a box plot showing the correlation between this time and the final grade:



Study time and final grades.

As expected, less than two hours per week for studying is not enough time to get high grades, and higher grades are correlated with more studying. We can also notice that the number of students spending more than 10 hours per week studying is low. But they achieve the highest average of grades.

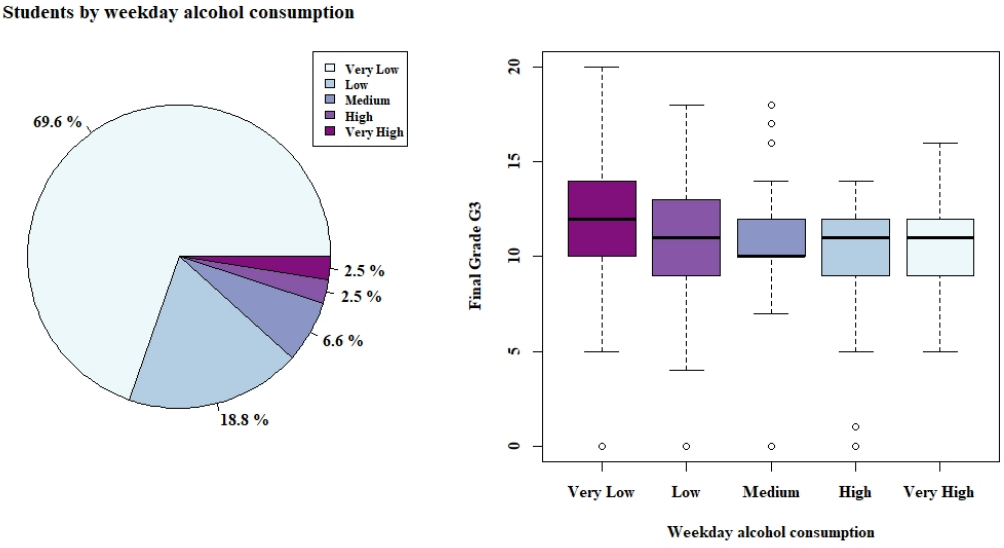
Another factor to look into is the students' health:



Health status and final grades.

An interesting observation is that students who report bad health status have high average grades. Which could suggest that their efforts to excel at school is at the expense of their health.

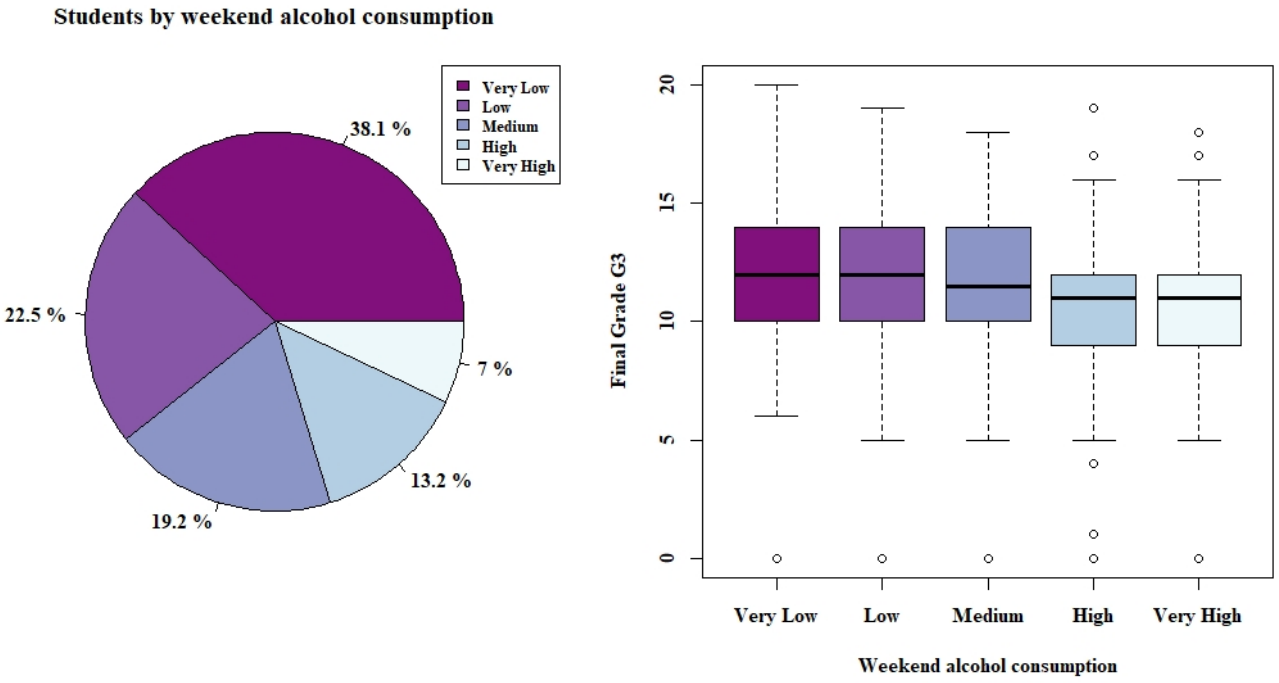
Now we analyse how alcohol consumption affects students' performance. First, we look at their weekday consumption:



Weekday alcohol consumption and final grades.

So clearly, we can see how weekday alcohol consumption negatively affects the students grades. And thankfully, most students prefer little to no alcohol during the week.

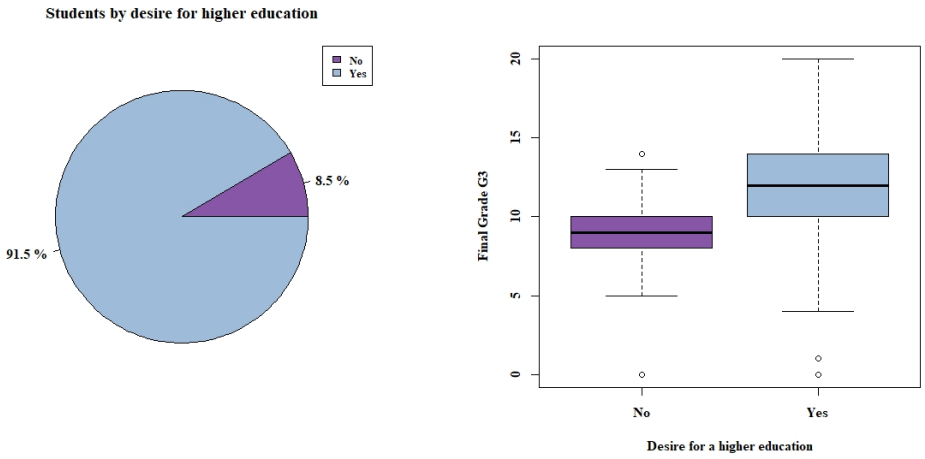
We also take a look at the weekend alcohol consumption and its effects:



Weekend alcohol consumption and final grades.

For the weekend, alcohol consumption increases among students. However, it doesn't show that it affects the students' grades as long as they don't abuse it.

One final factor to consider is the student's desire for a higher education and how would that desire motivate his performance:



Higher education desire and final grades.

As it's quite apparent, the desire for a higher education motivates the students to increase their performance in secondary school.

4 Modeling

For modeling, a linear regression model was created based on the most significant factors found from the analysis above. And these factors are:

- Study time
- Health status
- Weekday alcohol consumption.

With the following model:

```
Call:
lm(formula = G3 ~ studytime + health + Dalc, data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max
-13.0396  -1.6770   0.2301   2.3705   8.0794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.35152    0.55107  20.599  < 2e-16 ***
studytime    0.60586    0.16102   3.763  0.00018 ***
health      -0.15072    0.09356  -1.611  0.10759
Dalc        -0.43390    0.14466  -3.000  0.00278 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.804 on 831 degrees of freedom
Multiple R-squared:  0.03712,   Adjusted R-squared:  0.03364
F-statistic: 10.68 on 3 and 831 DF,  p-value: 6.829e-07
```

5 Evaluation

For the evaluation of the model, the test subset was use for prediction with the following results:

	Actual Values	Predicts
7	13	11.67718
8	13	11.97862
10	13	11.37574
14	13	11.67718
20	12	10.76988