École Polytechnique Fédérale de Lausanne

Analysis of students performances

Bollero Francesco, Mariantoni Mattia, Mulet-Arabí Pau, Padovano Federica, Rossi Luca

October 26, 2021

Contents

1 Introduction

In this project we aimed to analyse a database containing information about a sample of students attending a Portuguese language course in two possible different schools ('Gabriel Pereira' or 'Mousinho da Silveira'). For each student there are different attributes related both to their academic performances and to their family context. The table in the next page describes what are the variables that we consider, we can notice that some of them are strictly related with the student academic life (for example "G1","G2","G3","absences" and "studytime") and others tend to describe the student social context, as the family (for example "famsize","Mjob" and "Fjob") and the social life.

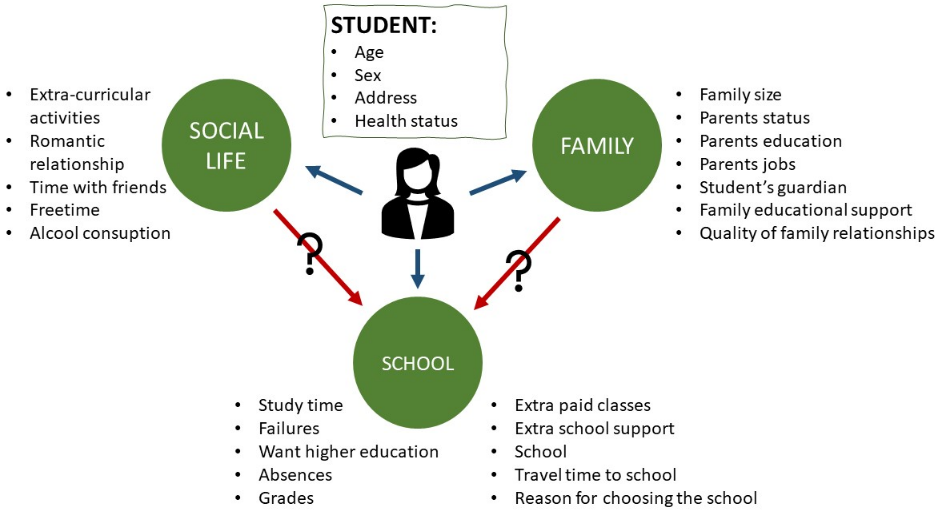


Figure 1: Dataset variables associated with their macroarea.

After a first glance at the dataset these three variable subsets caught our eyes. In our opinion families and social life impact children's academic path in some way. There are many studies that show how parents with higher educational level have most likely children with higher grades or that want higher education. As the same time, we expect that someone who attends many extra-curricular activities and spend lots of time with friends has less time to study and therefore lower grades.
In this project we want to study the data in order to see if we can credit our opinions or not and to try to predict grades knowing all the other variables. For doing that we want to use statistical visualisations and statistical testing in order to give a more accurate and precise point of view of the analysis .

We also decide to split our work in four different points answering different questions:

1. What's the correlation between variables?
2. Do the grades follow a normal distribution ?
3. Which is the capacity of different variables to explain the grades ?
4. Can we train a model to predict grade based on the other variables ?

and for each point we used different statistical methods.

Attributes	Values
school	student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
sex	student's sex (binary: "F" - female or "M" - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: "U" - urban or "R" - rural)
famsize	family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
Pstatus	parent's cohabitation status (binary: "T" - living together or "A" - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at$_h$*ome*”*or*”*other*”)
Fjob	father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at$_h$*ome*”*or*”*other*”)
reason	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
guardian	student's guardian (nominal: "mother", "father" or "other")
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)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: 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)
absences	number of school 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, output target)

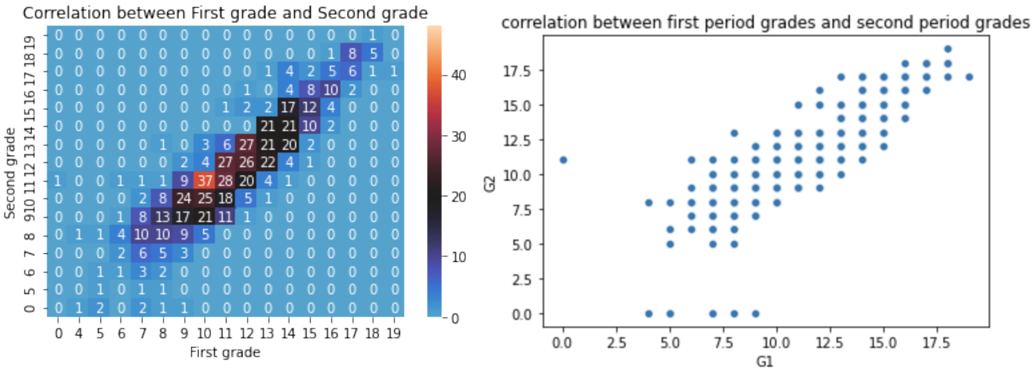
Figure 2: Variables meaning

2 Correlations and Hypothesis testing

In this part we study if there are some correlations between variables, for doing that we calculated Pearson correlation coefficient that measures linear correlation between two sets of data.

- **Correlation between grades across semesters**
  In order to asses which can be the final grade of the students it is very important to know their trajectory.

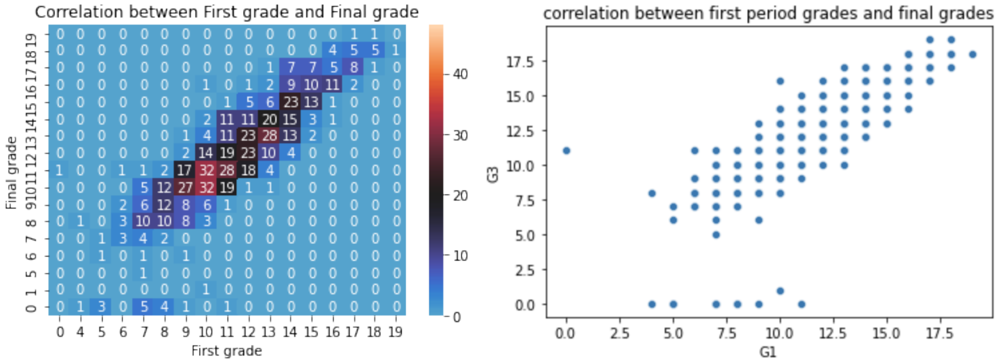
G1 vs G2



(a) Heatmap (b) Scatterplot

r coefficient	0.8649816303085828
Standard error	0.009899545995673489

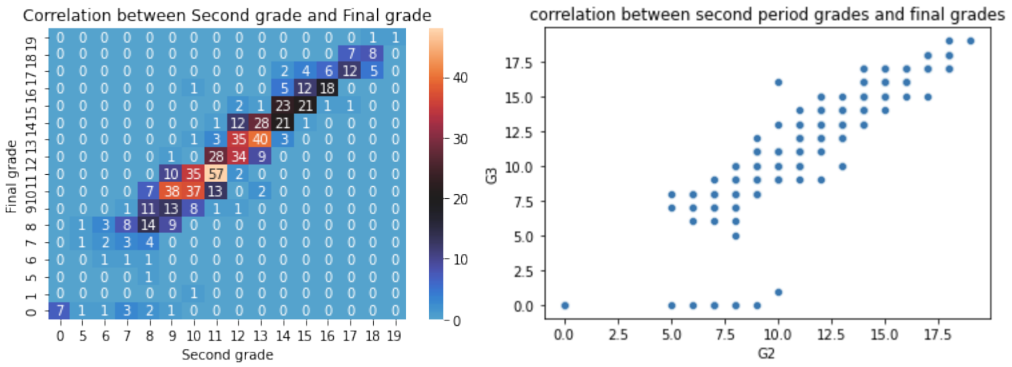
G1 vs G3



(c) Heatmap (d) Scatterplot

r coefficient	0.8263871247890472
Standard error	0.012465870934024606

G2 vs G3



(e) Heatmap

(f) Scatteplot

r coefficient	0.0.9185480035603512
Standard error	0.006143590544771005

From this correlation analysis we expect that the grades of the firs and second semester will have a high impact on the final grade. In addition, the high correlation between G1 and G2 suggest that we may only need one of them to describe the final grade.

- **Correlation between Sex and academic performances**
  For the gender correlation we didn't have precise *a priori* estimations because apparently there's no need to think that belonging to a particular gender can affect grades.

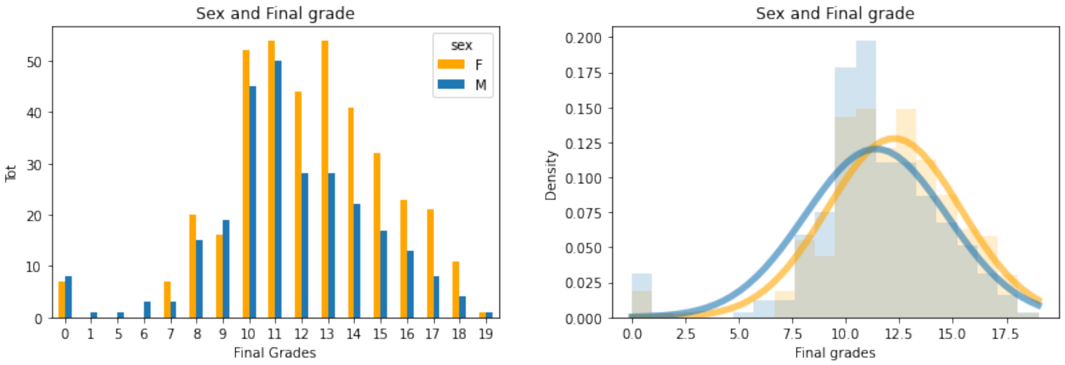


Figure 3: Correlation between Gender and Final grades, considering that the total of female is 383 and the total of males is 266.

The first figure shows that there are more males with lower grades than females and that the grade mean for female is higher than the one for males. At this point we should look if there is any remarkable relationships with the Study time.

So, at this point, we decided to do more quantitative tests to understand whether there is a statistical evidence to believe that there is a real difference between male and female grades using hypothesis tests.

We do not study final grades normality yet, because when we apply hypothesis tests on means, we have that for Central Limit Theorem the distribution of a mean converges to a normal distribution. In our case we can apply this theorem because we have many data and we can believe that final grades distribution is normal.

```
Difference (Male - Female) =    -0.8472
        Degrees of freedom =   549.3077
                         t =    -3.2747
     Two side test p value =     0.0011
    Difference < 0 p value =     0.0006
    Difference > 0 p value =     0.9994
```

Figure 4: Results of the three hypothesis tests between male and female final grades.

As we can observe, the two sides test has a small p value, which makes clear that the two mean are different. Consequently, we apply the one side tests and we see that female has actually better grades than males.

We studied also if there is difference between grades of first and second period.

```
Difference (Male - Female) =    -0.5807 Difference (Male - Female) =    -0.6157
        Degrees of freedom =   591.9478         Degrees of freedom =   592.4999
                         t =    -2.6898                          t =    -2.6879
     Two side test p value =     0.0074      Two side test p value =     0.0074
    Difference < 0 p value =     0.0037     Difference < 0 p value =     0.0037
    Difference > 0 p value =     0.9963     Difference > 0 p value =     0.9963
```

Figure 5: Results of the three hypothesis tests between male and female first and second period grades, on the left and on the right respectively.

This difference is evident also in the first and the second period grades.

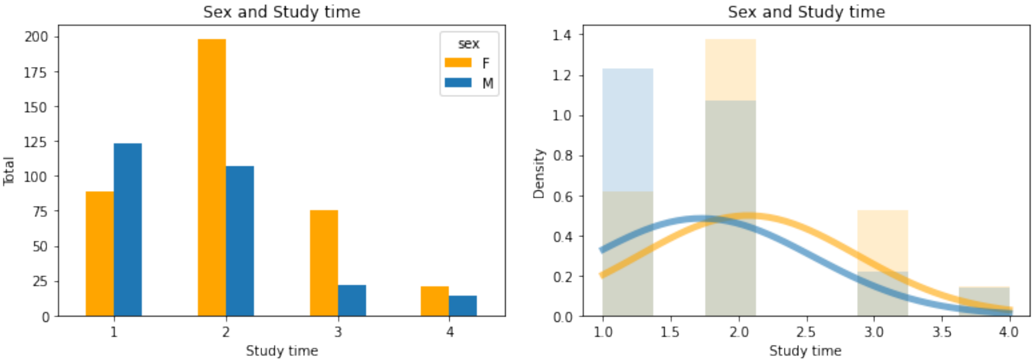
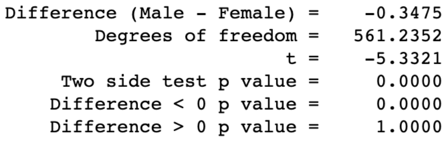


Figure 6: Correlation between Gender and Study time, considering that the total of female is 383 and the total of males is 266.

In the second graphic we can see that the density of men having the value 1 (the lowest one) of study time is way higher than the girls one. However, for higher value of study time the female density is higher than the males one. We can say that on average males study less than females and, as a consequence, on average they also have lower grades.

Again, to see if there is a statistical evidence to say so we apply hypothesis tests to verify the truthfulness of this affirmation.

The test confirms that girls spend more time studying than boys. Now a question comes spontaneously: is it true in general that the study time and the final grades have a positive correlation?



```
Difference (Male - Female) =    -0.3475
        Degrees of freedom =   561.2352
                         t =    -5.3321
        Two side test p value =  0.0000
       Difference < 0 p value =  0.0000
       Difference > 0 p value =  1.0000
```

Figure 7: Hypothesis test between male and female study time.

- **Correlation between Study time and Grades**
  Before running the calculation our *a priori* estimation was that the time of studying influences the grades in a positive way, in theory the more you study the more you succeed in tests. Oppositely, the less you study the more you fail at tests. So, analysing the r coefficient, we got these results:

	Grades	Failures
r coefficient	0.249788689998863	−0.14744054515158145
Standard error	0.036861080126113825	0.03845941964984769

As we expected there's a positive correlation between Study time and Final grades, and a negative correlation between Study time and Failures. However, the value of the coefficient in both cases is pretty low, so it doesn't seem to be a great correlation between the two.

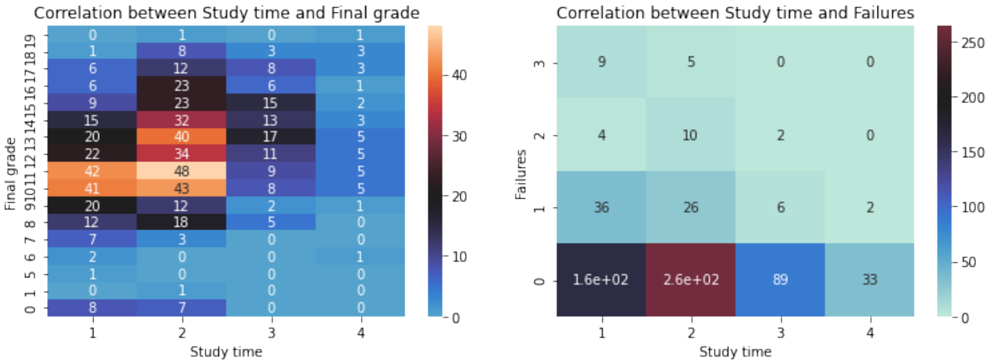
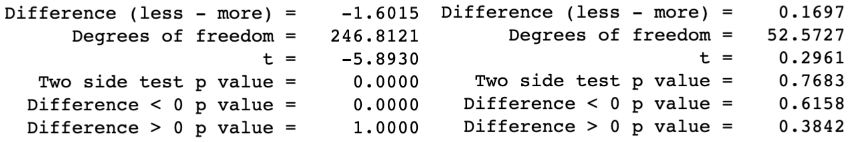


Figure 8: Study times - Final grades and Failures.

In the above figure we can observe two heatmaps describing the correlations we've just studied. The higher is the color in the scale the higher is the amount of students with a certain match 'final grade'-'study time'.

We then applied hypothesis tests to understand whether this visual evidence is statistically true.



```
Difference (less - more) =    -1.6015   Difference (less - more) =     0.1697
      Degrees of freedom =   246.8121         Degrees of freedom =    52.5727
                       t =    -5.8930                          t =     0.2961
   Two side test p value =     0.0000      Two side test p value =     0.7683
  Difference < 0 p value =     0.0000     Difference < 0 p value =     0.6158
  Difference > 0 p value =     1.0000     Difference > 0 p value =     0.3842
```

Figure 9: Test between people who study less vs people who study more than five hours on the left and test between people who study from 5 to 10 hours vs people who study more than ten hours on the right.

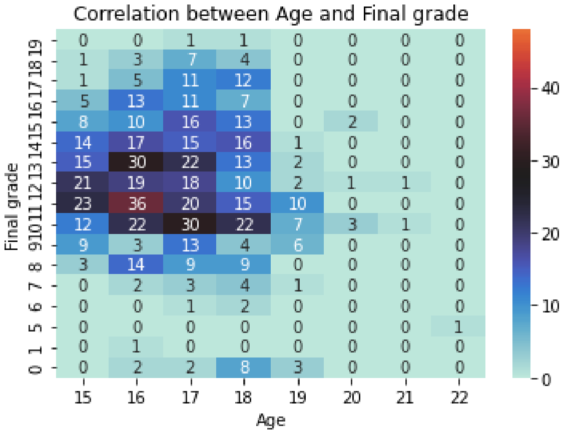
As we observe from the outputs, the first test makes clear that there is a correlation between study time and grades, in fact the p value of the first one side test is 0, which has as alternative hypothesis that the grades mean of people who study more than five hours is bigger than the other clarify that people who study more have better grades. But this correlation can not be linear as the second test demonstrates. This time the p value of the two sides test is 80 percent so we have to accept that the two mean are equals. Additionally, we can notice that people who spend from five to ten hours studying have even a better mean than people who spend more than ten hours studying.

- **Correlation between age and Final grade**

  Another question could be if the students age can have some consequences on the grades they get. Our *a priori* expectation was that there is no correlation between the two because each year students have to study subjects appropriate for they ages . We calculated the Pearson's correlation coefficient and obtained:

r coefficient	-0.008415114730320437
Standard error	0.0393112727067965

  The $r-$coefficient is negative and overall is really low, so it credits our idea of non correlation between age and Final grades. The heatmap below shows the relationships between the two of them.



Final grade	15	16	17	18	19	20	21	22
19	0	0	1	1	0	0	0	0
18	1	3	7	4	0	0	0	0
17	1	5	11	12	0	0	0	0
16	5	13	11	7	0	0	0	0
15	8	10	16	13	0	2	0	0
14	14	17	15	16	1	0	0	0
13	15	30	22	13	2	0	0	0
12	21	19	18	10	2	1	1	0
11	23	36	20	15	10	0	0	0
10	12	22	30	22	7	3	1	0
9	9	3	13	4	6	0	0	0
8	3	14	9	9	0	0	0	0
7	0	2	3	4	1	0	0	0
6	0	0	1	2	0	0	0	0
5	0	0	0	0	0	0	0	1
1	0	1	0	0	0	0	0	0
0	0	2	2	8	3	0	0	0

Figure 10: Correlation between Age and Final grades.

- **Correlation between Student family and grades**

  We now study how the family environment impacts on the students academic performances. First we analyse the family size. In the graphic below we compared student with families larger than three people and the ones equal or smaller than three people.

  We can notice that the mean of grades for students with big family is a bit lower than the one for small family. Furthermore, in higher grades students with small families have higher density than the other ones.

  From the test we notice that this difference is not significant, in fact the two sides test has a p value of more than 20 percent.

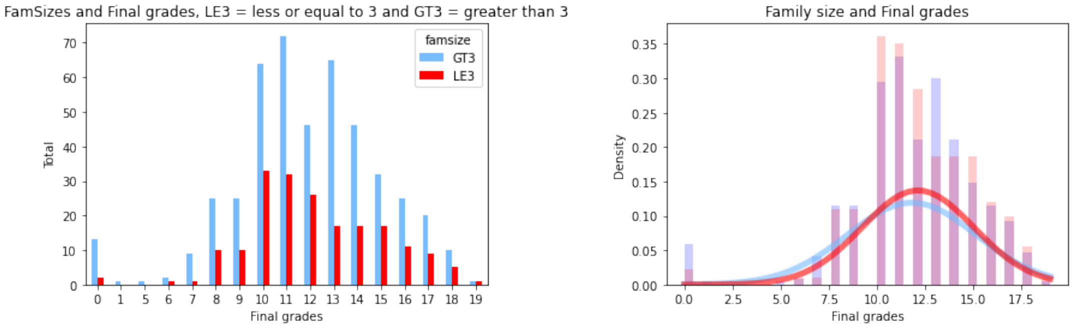


Figure 11: Correlation between Family size and Final grades, considering that the total of family with less than three people is 192 and the total of the other ones is 457.

```
Difference (small - big) =     0.3184
      Degrees of freedom =   410.6020
                       t =     1.2124
   Two side test p value =     0.2261
  Difference < 0 p value =     0.8870
  Difference > 0 p value =     0.1130
```

Figure 12: Test between student with big vs small families.

We believe that parents living together can affect students grades in a positive way, probably because a compacted family may create a less stressful environment for students and therefore can get to higher grades.

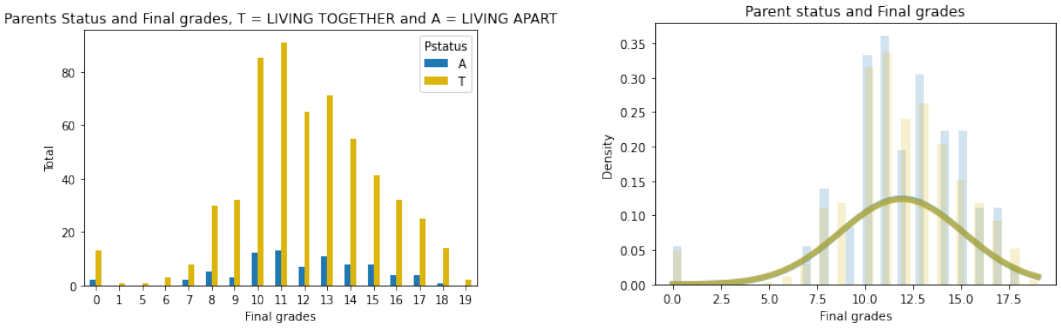
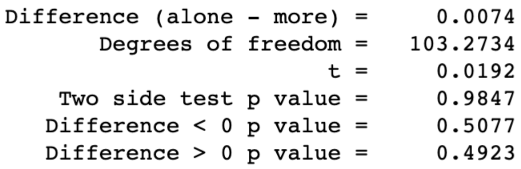


Figure 13: Correlation between Parents status and Final grades considering that that total amount of students with parents living together is 569 and for the other ones is 80.

In the figure above we can notice that the grades mean for students with parents living together and the one with parents living apart is the same. In this case there is no relevant difference between the two density so in this case if parents are separated or not doesn't affect their children grades.

The hypothesis test confirm that there is no evidence to believe that the parents status affect the grades of students.
As we said, we believe that good family relationship can affect students grades in a positive way. Now we verify the Pearson's correlation coefficient between Family relationships and Final grades:

r coefficient	0.06336112772983048
Standard error	0.03915622520852588



```
Difference (alone - more) =      0.0074
      Degrees of freedom =    103.2734
                       t =      0.0192
     Two side test p value =      0.9847
    Difference < 0 p value =      0.5077
    Difference > 0 p value =      0.4923
```

Figure 14: Test between student with big vs small families.

Once again the data analysis destroys our *a priori* expectation, in fact we found a very low correlation coefficient that shows that quality of family relationships doesn't affect students grades.

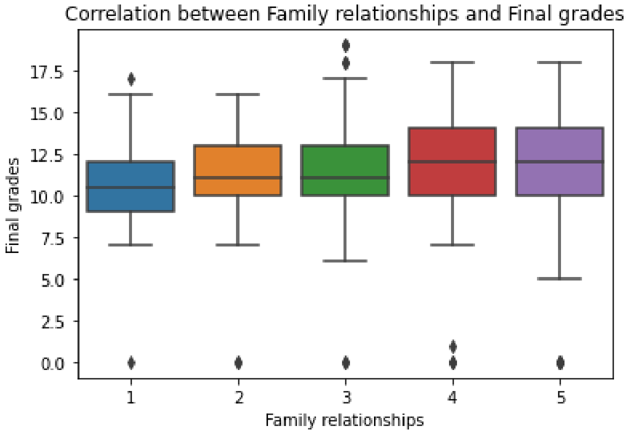
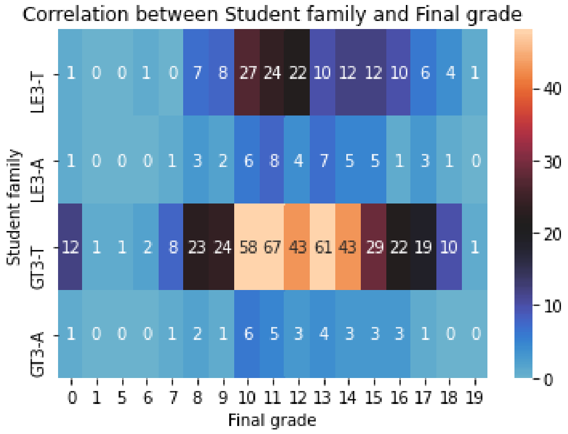


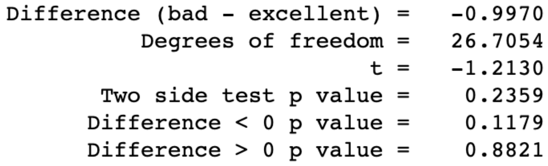
Figure 15: Correlation between Family relationships and Final grades.

In the figure below we show a summary graphic for the correlation between Family and Final grades.



	0	1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
LE3-T	1	0	0	1	0	7	8	27	24	22	10	12	12	10	6	4	1
LE3-A	1	0	0	0	1	3	2	6	8	4	7	5	5	1	3	1	0
GT3-T	12	1	1	2	8	23	24	58	67	43	61	43	29	22	19	10	1
GT3-A	1	0	0	0	1	2	1	6	5	3	4	3	3	3	1	0	0

Figure 16: Correlation between Family environment and Final grades.



```
Difference (bad - excellent) =   -0.9970
           Degrees of freedom =   26.7054
                            t =   -1.2130
        Two side test p value =    0.2359
        Difference < 0 p value =    0.1179
        Difference > 0 p value =    0.8821
```

Figure 17: Hypothesis test based on family relations quality.

Once again the hypothesis test confirm the result obtained by graphic visualization, in fact the two sides test has a p value of 23 percent, which means that there are no statistically significant differences between the two groups mean.

3 Normality of final grades

To understand whether variable G3 has normal distribution we used different techniques. Firstly we tried to plot an histogram with the distribution of the variable. Then over the histogram we plotted a line-plot of a gaussian with same mean and same variance as the variable.

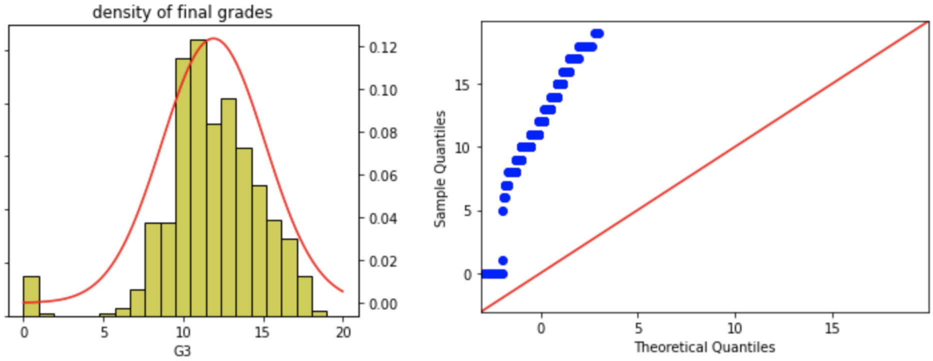


Figure 18: Comparing normal distribution with G3 histogram and qq-plot.

As we can notice from the first plot we can already think that the distributions of the variable is not normal. In fact G3 has too many data in the tails, which underline that it must not be normal.
Just to be sure that our data are actually not normal we then apply more quantitative tests: QQ-plot and Shapiro-Wilks test. The results make clear that the distributions can't be normal, in fact the p value of Shapiro test is smaller than 10^{-17}, which give no evidence to accept the null hypothesis. The QQ-plot then confirms that we can't believe that the distribution of final grades is normal.
We also tried to understand whether it can be normal without the outliers, but again the Shapiro test has a really small p value: 10^{-8}.
Additionally, we conducted a similar analysis on grades G1 and G2 and we arrived at the same result. So, although, we expected that our grades would have a normal distribution, which could be useful for regression analysis the results clarify that they are not normal.

4 Explaining Academic Performance with Linear Regression

The goal of this section is to find which factors determine the grade of a student. This is similar to what we did in section 1 with hypothesis testing, but now we also want to quantify the contribution of each variable and how they interact between them to explain the grades. We take the grade of the third semester as the response and the rest as explanatory variables. We assume that the underlying model is linear with a Gaussian response. We justify this choice in the model diagnosis and testing goodness of fit. For the time being, the use of linear models is motivated by their high interpretability despite not having as much predicting power as more complex models.

We begin by removing the students that have a 0 in one of the grades. This decision is justified by the fact that in most education systems around the world is not possible to get a 0. Therefore it may not be a real grade but simply caused by the student moving to another school for example. Removing these entries will have little impact on our capabilities to explain the data because there are only 17 students in this situation. Now we can have a first look at the data.

3rd Semester Grades					
Min	**1st Quart**	**Median**	**Mean**	**3rd Quart**	**Max**
5	10	12	12.21	14	19

In the following boxplot we can appreciate the symmetry of both the median and maximal values with respect to the IQR.

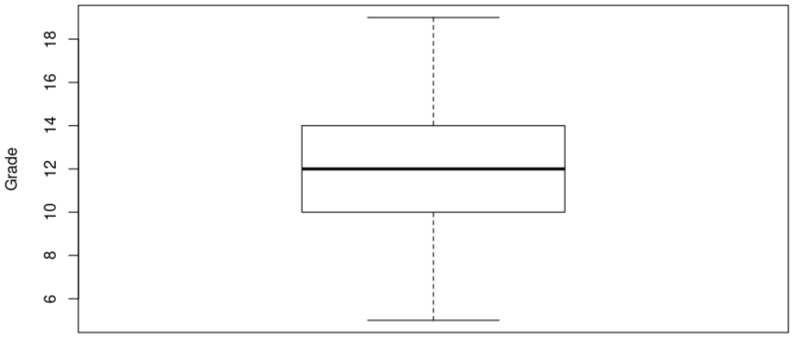


Figure 19: Final grades

To understand in more detail how the grades are distributed we can look again at the histogram and its corresponding smoothed density. If we want to find out how to combine the different input variables to predict the final grade it is crucial to understand the individual effect that each one has in the response variable.

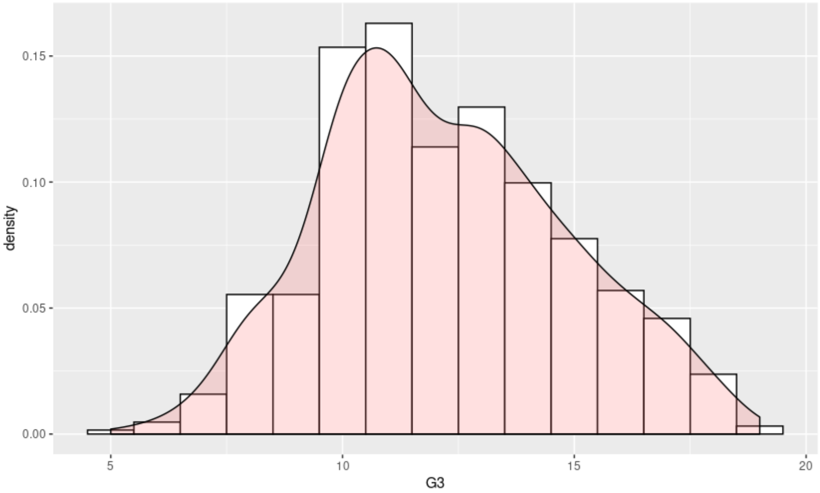


Figure 20: Final grade boxplot and smoothed density

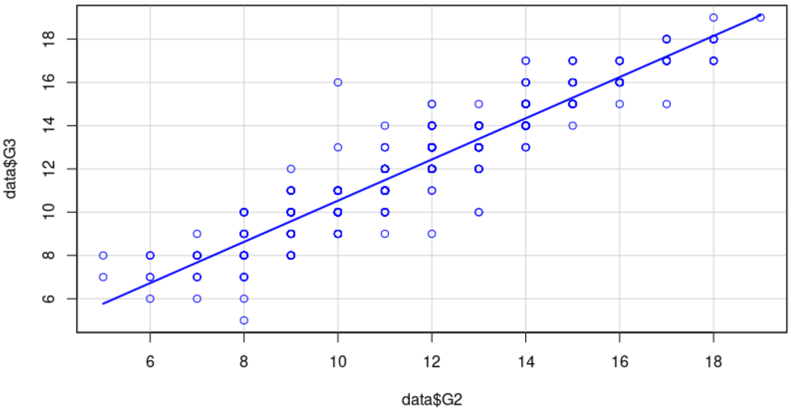


Figure 21: Scatterplot of the grade in the second term against the final grade

We can see that the grade in the second term explains quite well the final grade. Indeed, in most cases the final grade lies in the range of ±3 with respect to the second term grade. However, although the general trend is captured by the second term grade we also want to understand which factors contribute to increase/dicrease the grade with respect to the second term in order to explain the ±3 changes. If we are able to identify categorical variables which have either a positive or negative contribution to the final grade we will be able to better explain this changes. In fact the corresponding intercepts of those different groups would help to fit not only the scale but also the location. Finally it is important to clarify why the points are vertically aligned. This is due to the fact that both variables (G2, G3) only take integer values. However when fitting the data we assume that they take real values because then we can express more accurately how sure is the model about the predicted value. For instance, either if the model outputs 13.9 or 13.5 our predicted grade will be 14, but the uncertainty will be much higher in the second case.

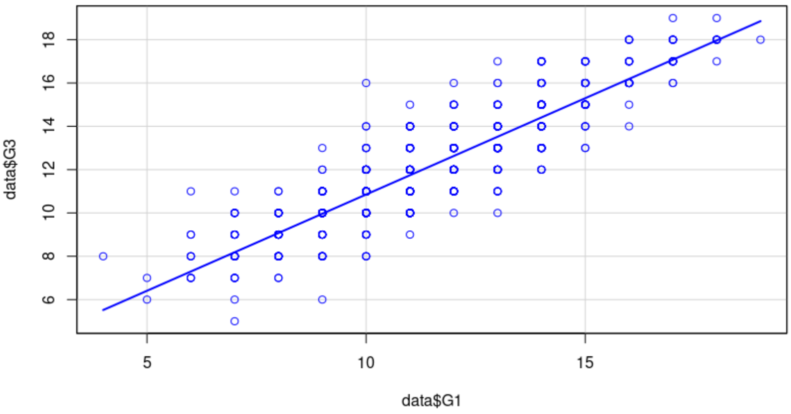
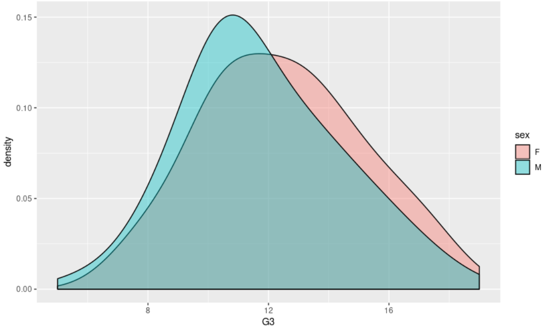


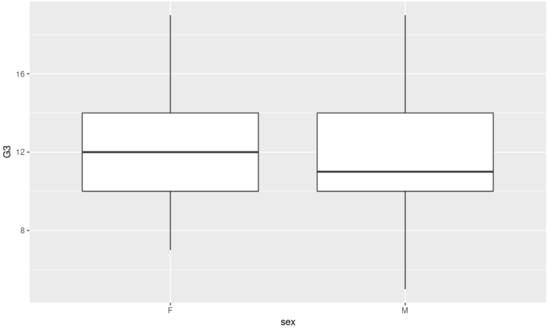
Figure 22: Scatterplot of the grade in the firs term against the final grade

As expected the effect of the first term grade on the final grade is very similar to the one exposed before for the second term grade. Moreover we can appreciate that the location changes are even bigger now. The higher uncertainty can be explained by the longer time interval between the grades. Therefore it is clear that we need to add some categorical variables that can capture this location changes. In a first step we want to visualize how does the final grade distribution changes among different groups defined from a categorical variable. In order to do so we compare the boxplots and smoothed densities of the final grade for different values of each categorical variable.

Sex



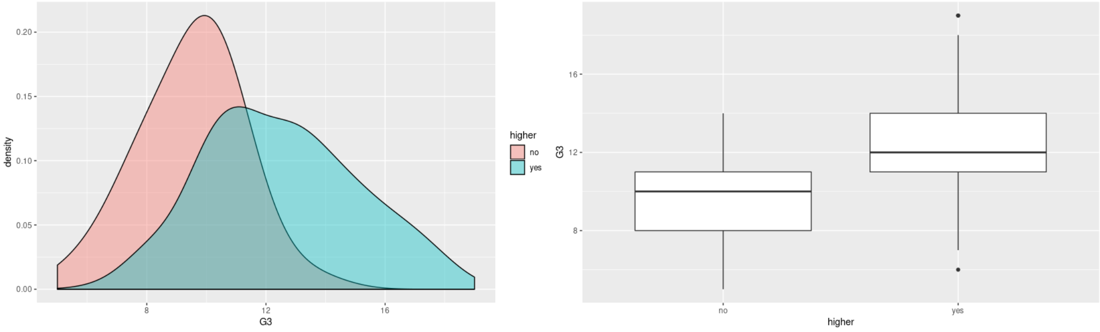
(a) Smoothed densities



(b) Boxplots

We can see that females tend to get better grades than males in our population. However the interquantile range looks exactly the same for both groups. The main difference relies on the median which is centered for females but positively skewed for males.

Higher Education



(c) Smoothed densities

(d) Boxplots

In these plots we can see the difference between the students that want to pursue higher education and the ones that want to enter the job market directly. In this case we can already see at first glance that there are important difference between both groups in our population. Students who want to pursue higher education tend to get higher grades than those that do not want to continue studying. In fact the first quartile of the first group coincides with the third quartile of the second group. However the medians are quite close (only 2 points difference) due to the fact that the first group is positively skewed and the second one is negatively skewed.

Some categorical variable in our dataset take more than 2 values. In this case it is possible that there are only differences between some groups but others are similar in terms of grade distribution.

Mother Job

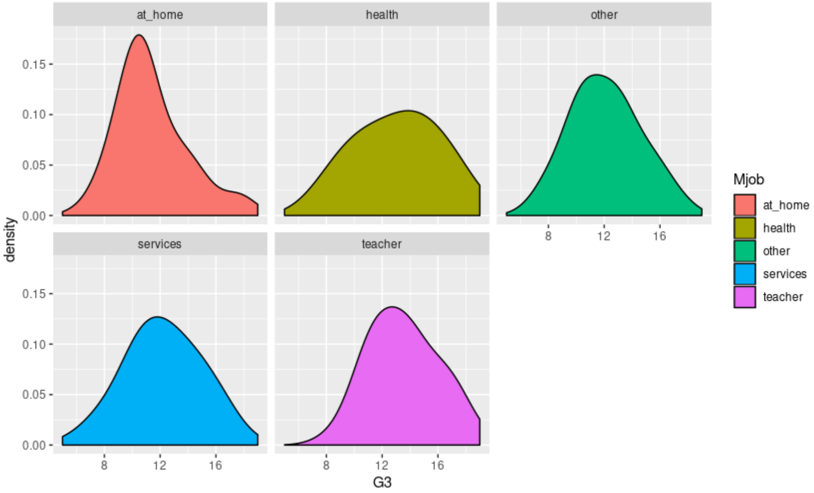


Figure 23: Smoothed densities for different students grouped by their mothers job

In the figure above we can see significant differences between the smoothed densities of some groups. In fact those whose mother works either in health or education seem to be more likely to get higher grades than those in the other groups. Regarding the rest of the groups we can appreciate that the location of the distribution is very similar but they differ in terms of the tail's weight.

Absences

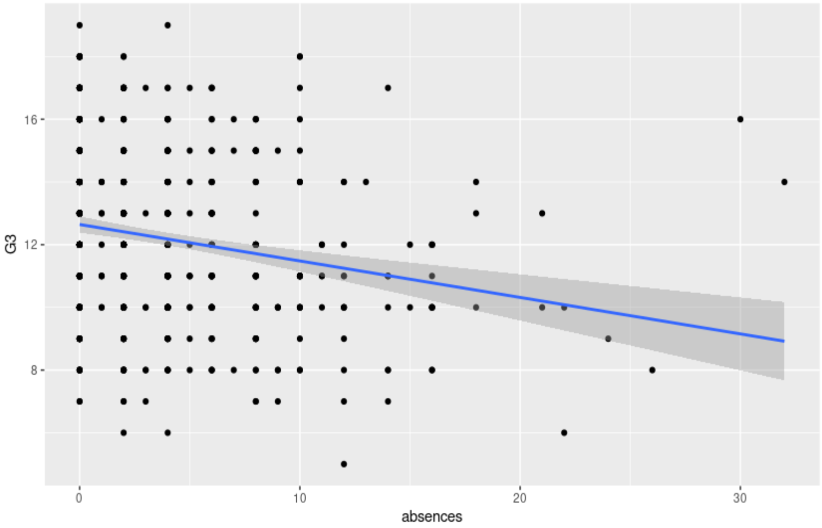


Figure 24: Fitted regression line between absences and final grade

As you can see the number absences tend to have a negative impact on the final grade. However, we have to be cautious because it is not a steep slope. It makes perfect sense that the more absences a student has the lower the grade because the time spent in class is decreased. Nevertheless we also want to look at the relevance of the time that students spend studying in their own.

Study Time

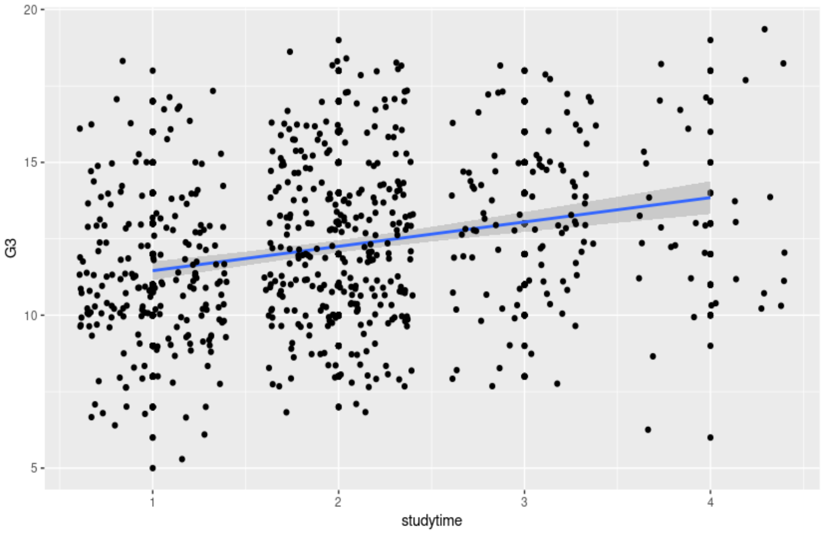
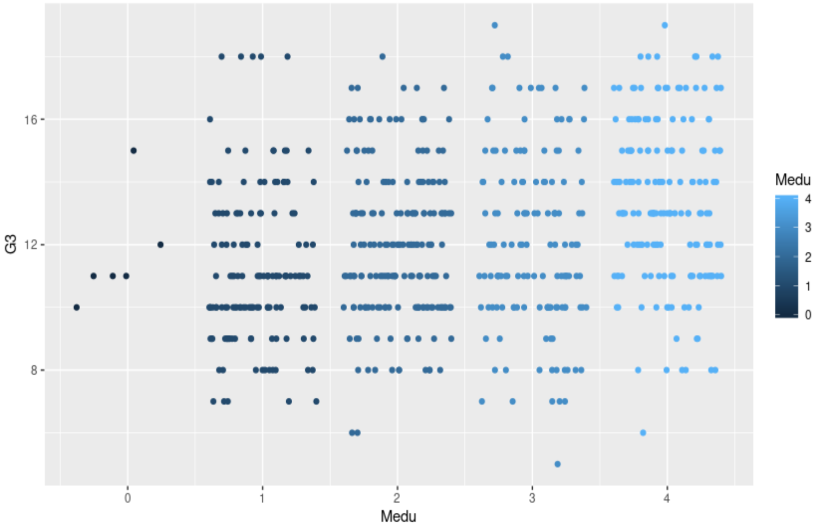


Figure 25: Fitted regression line between studytime and final grade

We are in the same situation as before. Despite having a trend in the data that study time contributes positively to the final grade we must be wary. In fact the slope of the fitted line is not very steep, so it could be possible that there is no statistical significance (seen in sec. 1).

Mother Education



Finally we see that the level of education of the mother also seems to have a positive contribution to the final grade. Indeed, points corresponding to a higher level of education (the lighter ones) tend to have a higher G3 value.

4.1 Data Fitting

The goal in this section is to asses the relationship between the input variables and the final grade. Previously we have got an intuition of which are the individual effects of each explanatory variables but now we want to consider multiple inputs. In addition it is crucial to formalize our procedure beyond descriptive statistics in order to draw rigorous conclusions. We will be trying to fit a linear model of the final grade with respect to the rest of the input variables. In general terms we assume that:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon \quad i = 1, ..., n \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

and want to find $\hat{\beta_0}, ..., \hat{\beta_k}$ that minimize the mean square error $||Y - \hat{Y}||^2$ where

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_{1i} + ... + \hat{\beta_k} X_{ki}$$

Gauss-Markov Theorem states that under our model assumptions this estimator has the lowest sampling variance within the class of linear unbiased estimators. Now that we have exposed the theoretical aspects of our model it is time to move into practice.

First of all we want to check if our explanatory variables are useful for explaining the final grade with this model. As a result we perform an Omnibus test:

$$\begin{cases} H_0 : & \beta_0 = \beta_1 = ... = \beta_k \\ H_1 : & \exists j \quad \beta_j \neq 0 \end{cases} \tag{4.1}$$

$$\boxed{F - statistic := 141.6 \qquad p - value < 2 \cdot 10^{-16}}$$

Therefore we can reject the null hypothesis that our data is useless to predict the final grade through a linear model. Nevertheless, if we want to explain the data in a concise way it is important to remove any redundant information. At this point we are using 32 input variables so it is very likely that some of them are either irrelevant or redundant. In order to select the right variables that explain the model without redundancies we use backward selection with two factor Anova. In other words we begin with a linear model that includes all the parameters and then remove the variable with the highest p-value according to the Anova test. This test is basically a tool to quantify the marginal effect of a variable assuming the presence of the rest. Although this looks similar to what we did when testing hypothesis in section one it is important to point some differences out. In fact, when we remove a variable following this procedure we are not saying that it has no effect on the final grade. In this setting when we remove a variable we only mean that it has no effect on the final grade under the presence of the rest of the inputs. Therefore it would be wrong to say that the removed variables have no contribution to the response. Another aspect of variable selection is the rejection threshold. Although the typical value is 0.05 when doing hypothesis testing it varies more when fitting or predicting. As a result we have decided not to remove some variables with p-values close to 0.05. Finally an important point to bear in mind is that at each iteration we only remove one variable even if the test outputs several variable with p-values larger than the threshold. The reason for this is that the removal of one predictor could increase the significance of another. Indeed this is what happens very often in practice.

Anova Type II

	Sum Sq	Df	F value	Pr(>F)
G2	457.14	1	661.01	0.0000
G1	35.64	1	51.53	0.0000
age	8.86	1	12.81	0.0004
failures	4.67	1	6.76	0.0095
higher	3.98	1	5.76	0.0167
schoolsup	2.59	1	3.74	0.0535
traveltime	2.50	1	3.61	0.0580
absences	2.37	1	3.43	0.0646
school	2.78	1	4.02	0.0454
sex	2.57	1	3.71	0.0545
Residuals	429.48	621		

Table 4.1: Anova test of the selected variables after the backward iterations

We can see in the table above that most of the model variance is explained by the previous grades. As a result we want to test if the other variables are relevant once we already have the previous grades. We call M_1 the current model and M_2 a linear model that only includes the previous grades. Then we perform the test

$$\begin{cases} H_0 : & M_0 \equiv M_1 \\ H_1 : & M_0 \not\equiv M_1 \end{cases} \tag{4.2}$$

$$\boxed{F - statistic := 4.4639 \qquad p - value = 2.8 \cdot 10^{-5}}$$

Therefore we can reject the null hypothesis, which means that the variables that we currently have in the model apart from the grades are not irrelevant at all. From now on we will stick to the model M_1

Goodness of fit

The most widely used measure to test how well a model fits the data is the so-called coefficient of determination. It is basically the percentage of variance explained by the model.

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$$

In order to take into account the model complexity and avoid overfitting there is an extension of the R^2 usually called $R^2 - adjusted$. It has the following expression:

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

In our particular case both are very similar because we have already performed feature selection.

$$\boxed{R^2 := 0.9037 \qquad \tilde{R}^2 = 0.9022}$$

Collinearity

We can express our linear model as $\hat{Y} = X\hat{\beta}$ and find β such that $(X^TX)\beta = X^TY$. This is equivalent to the least square estimate, but it is not unique unless X^TX is a full rank matrix. In the former case it causes many problems in terms of interpretability because the model becomes non identifiable. Then the meaning of the parameters can be misleading and this can lead us to miss the importance of some variables. In order to detect collinearity we can perform a regression of each input variable X_i based on all other inputs. For each explanatory variable we will get an associated coefficient of determination R_i^2. We want these to be as low as possible otherwise we will increase the variance of the estimator. This leads to the concept of **Variance Inflation Factor** $\frac{1}{1-R_i^2}$

Variance Inflation Factors				
G2	**G1**	**age**	**failures**	**higher**
5.13	5.33	1.24	1.31	1.23
schoolsup	**traveltime**	**absences**	**school**	**sex**
1.1	1.08	1.13	1.24	1.06

Most authors say that variables must be removed when its VIF is higher than 5 and some suggest lower threshold values but not lower than 2. So we can proceed to remove grade 1 on our predictors to avoid collinearity. Now we can check that there is not collinearity anymore.

Variance Inflation Factors				
G2	**G1**	**age**	**failures**	**higher**
1.38	X	1.21	1.30	1.21
schoolsup	**traveltime**	**absences**	**school**	**sex**
1.1	1.08	1.13	1.24	1.06

Despite removing a variable that was highly correlated with the response we should not expect a significant deterioration in the goodness of fit.

$$\boxed{R^2 := 0.8957 \qquad \tilde{R}^2 = 0.8942}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0415	0.5709	0.07	0.9420
G2	0.9076	0.0154	59.05	0.0000
age	0.0771	0.0312	2.47	0.0138
failures	-0.2014	0.0678	-2.97	0.0031
higheryes	0.3730	0.1259	2.96	0.0032
schoolsupyes	-0.2614	0.1178	-2.22	0.0268
traveltime	0.0922	0.0478	1.93	0.0541
absences	-0.0152	0.0079	-1.94	0.0532
schoolMS	-0.2495	0.0801	-3.12	0.0019
sexM	-0.1572	0.0720	-2.18	0.0294

In this figure we can see a summary of the fitted model. The first two colums contain the parameter estimates and their standard errors respectvely. The last two columns contain the information from the anova test for the marginal effect of each variable.

The simplicity of the model makes it very easy now to interpret the results and quantify the contribution of each variable. As expected from the visualization the grade of the second semester has a high impact on the final grade. In fact, each point increase in the second term grade will increase by 0.9 the final grade. Conversely the age of the student has very little effect, but older students are more likely to get higher grades. This trend can become significant when we compare 15 year old students against 22 year-old student. The later will have on average 0.5 higher grade when all the other factors are equal. The school attended by the student can also lead to higher or lower grades but there are no big differences. The same goes for the gender that despite having some positive effect for women it is not as determinant as other factors like if the student wants to pursue higher education. Finally we have some variables with negative effect on the grade, the most important being the number of failures. Indeed we could say that a past failure has the same effect as 20 absences. The school support variable also has a negative parameter. However this does not mean that the extra education support causes a lower performance in the final exam. It could be explained by the fact that students that receive support tend to have more difficulties studying and that is the reason why they turn to the extra support.

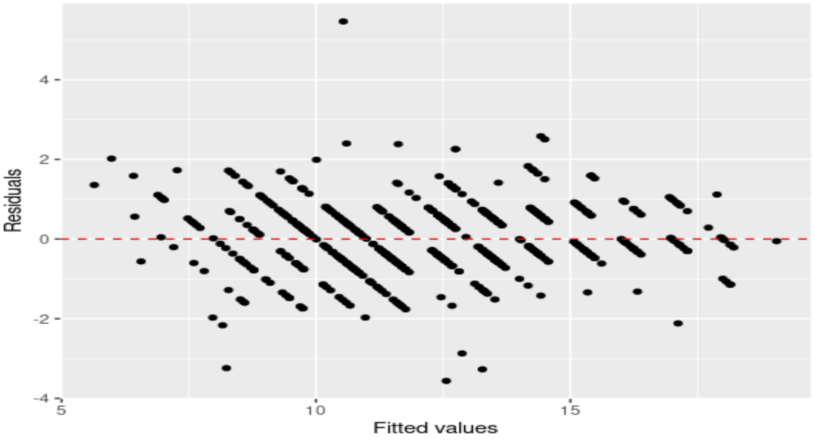
Before going back to visualization we show some other measures to asses the model fit to the data.

$$\boxed{\textbf{MSE} = 0.7447322 \qquad \textbf{MAE} = 0.6739461}$$

4.2 Diagnostics

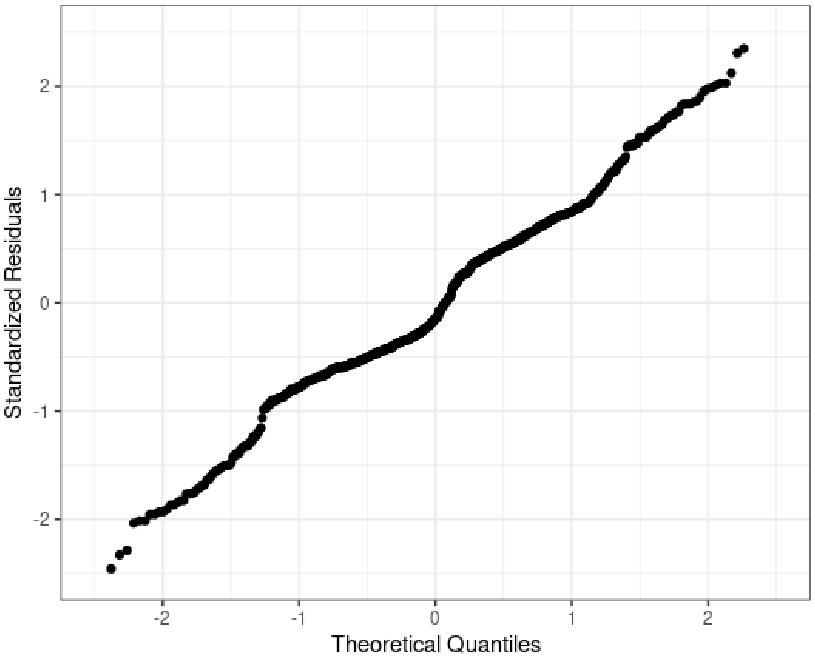
In order to fit the data we have made some previous assumptions that are encoded in our model. Now we need to check that these assumptions are valid. In addition we want to detect possible observations that do not fit the model.

Standarized residuals



The previous plot shows the predicted values against their standarized residuals. These allow us to check that the variance is homogeneous across the whole range of grades as we assumed in our model.

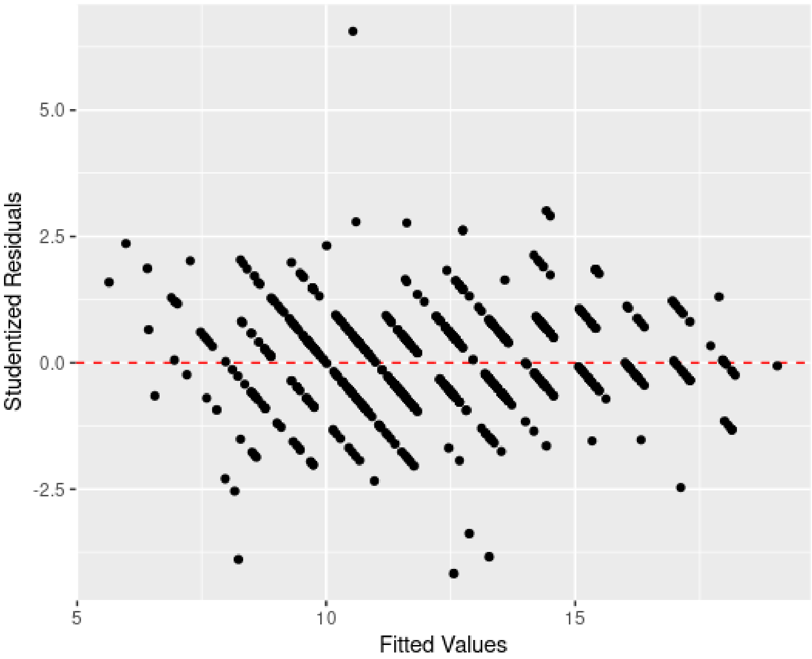
QQ-Plot for residuals



The previous plot shows that the residual variance follows a normal distribution as assumed in our model.

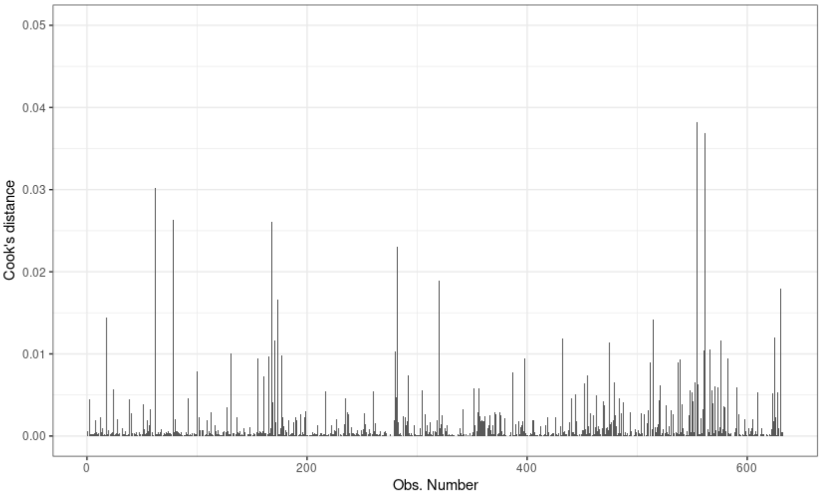
Now we want to detect possible outliers that are not explained well by the model. Identifying them is very important as we could propose that those students are individually analyzed by a human expert.

Studentized residuals



In the previous plot we can see the fitted values against standarized residuals which are a good measure to detect outliers. Most authors suggest that entries with stud. residuals higher than 3 in absolute value should be considered outliers.

Cook's distance



In the last figure we can see the cook's distance plot, that measures the influence of the different data points. It summarizes how much the model would change if we remove a data point.

5 Student classification

In this section, we want to try to predict if a student will pass the final exam or not. This can be crucial in real life because the fact that a student fails can have many undesirable consequences. On the one hand it can have a huge economic cost in the case that the students have to retake the whole school year. On the other hand, it could demotivate the students and affect their self-esteem. Since now the focus will be more on prediction rather than interpretability we will choose a slightly more complex model. The linear model can be generalized to:

$$y|X \sim F(\theta(x))$$

for a general distribution F and a function $\theta(x)$ that has an associated link function satisfying

$$g(\theta(x)) = \beta^T x$$

.
We will use the logistic regression that is defined as:

$$y|X \sim Bern(p) \qquad p = \frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)}$$

According to the general formulation we would have that

$$F \sim Bern(p) \qquad g(t) = \log\left(\frac{t}{1-t}\right)$$

Fitting a Logistic Regression Model

We split the data in two groups: Train (70%) & $Test(30\%)$

Note that we do not create any validation set because we will not be tunning hyperparameters. Hence we simply fit the regression using the training set and measure the performance in the test set. Before presenting the results there are some aspects of the training procedure that we want to point out. First of all we have not performed feature selection. The reason for this is that we have realized that including all the variables not only resulted in a higher fit (as expected) but also in higher generalization. In second place we want to remark that the train/test sets have been selected completely at random without any modification to avoid imbalances. After those clarification we can proceed to present the results

$$\boxed{\textbf{Test Accuracy:}\ \ 92\%}$$

Confusion Matrix

		True values		
		Fail	**Pass**	Total
Predicted Values	**Fail**	38	5	43
	Pass	10	137	147
	Total	48	142	190

- **Sensitivity:** 0.7917
- **Specificity:** 0.9648
- **Pos Pred Values:** 0.8837
- **Neg Pred Values:** 0.9320

Mcnemar's Test

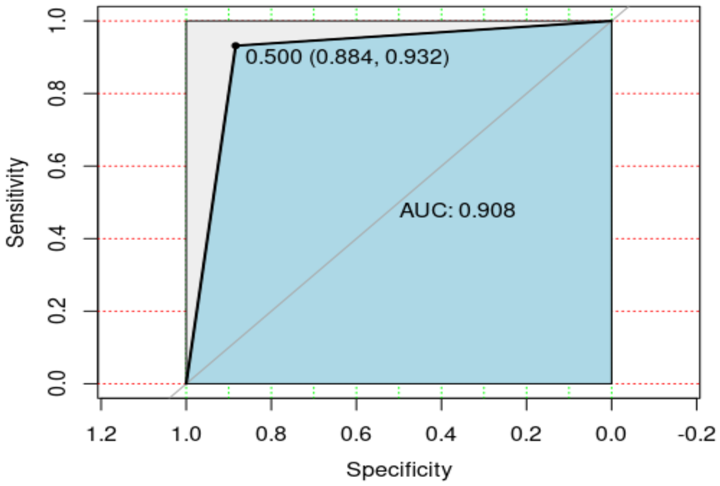
It is a test for marginal homogeneity between the parameter $\hat{p}$ induced by the model and the true probability p.

$$\begin{cases} H_0: & p = \hat{p} \\ H_1: & p \neq \hat{p} \end{cases}$$

$$\boxed{p - value = 0.3017}$$

Therefore we cannot reject the null hypothesis, which means that there are no statistically significant differences between our estimator and the true parameter.

ROC Curve



So far our model is making good predictions but we can adapt it to make it even better for our purposes. In particular, we are more interested to predict well those students who fail that the ones that pass. As explained before we would like to avoid that students have to retake the school year because it has several unsatisfactory consequences. However our data is not particularly good for this purpose. In fact we have much more data about students who pass that about student who fail. As a result, if we treat all the data points equally our model will be more prone to predict well students who pass rather than students who fail because it will lead to a lower loss. To avoid this behaviour on our predictor we can add some weights to the errors incurred when classifying students that failed. The general procedure is as follows: Let $\mathcal{E}(\mathcal{D}, \beta)$ be the error in the sample $\mathcal{D}$ when using the parameter β with our logistic regression model. Usually:

$$\mathcal{E}(\mathcal{D}, \beta) = \sum_{i=1}^{n} \mathcal{E}(x_i, \beta)$$

which as mentioned leads to skewed predictors when there are imbalances in the dataset. In order to solve this we can redifine the error as:

$$\mathcal{E}(\mathcal{D}, \beta) = \sum_{i=1}^{n} w_i \mathcal{E}(x_i, \beta) \qquad w_i = \begin{cases} 0 & x_i \in C_0 \\ w & x_i \in C_1 \end{cases} \tag{5.1}$$

where C_0, C_1 indicate the two classes. Hence w can be chosen according to the frequencies so that both classes have the same contribution into the total error. That is the approach we followed to select w, but we could also find the optimal w in a validation set via grid search for example. These modifications led to the following results:

Confusion Matrix

		True values		
		Fail	**Pass**	Total
Predicted Values	**Fail**	44	22	66
	Pass	4	120	124
	Total	48	142	190

- **Accuracy:** 0.8632
- **Sensitivity:** 0.9167
- **Specificity:** 0.8451
- **Pos Pred Values:** 0.0.667
- **Neg Pred Values:** 0.967

We can see that the number of missclassified students has decreased from 10 to 4. However this has caused the model to decrease its accuracy. This is not a surprise because we have forced the model to pay less attention to the passed students which are the majority.

6 Machine Learning

The aim of the Machine Learning section is to apply some of the most popular Machine Learning techniques to our dataset in order to generalise relevant information contained in the data. Before diving into the questions there are some aspects of our methodology that we must clarify.

All the models have been trained with a 80 to 20 percentage split for training to testing. We have implemented our models using Sklearn and Keras libraries. Finally to determine if a model is non trivial we compare it against a baseline, which is a constant prediction model which outputs either the mean or the median of the training set.

Once we have established our procedures we can already answer some relevant questions about our data. The easiest and first question which we want to address is the following:

- *Can we build a model which can predict the final grade G3 based on the other features ?*

To answer this question we tried different models and we compared their performances. First of all, we preprocessed the data. Indeed, as we can see from the table above (which contains the information of the dataset) our original dataset is made of discrete and categorical features. For instance, the attribute 'address' is categorical because it can be 'U' for urban and 'R' for rural so it can belong to two distinct classes and there is no ordering relationship between the classes. However, we have also some variables which assume discrete range of values like 'age' which is a number from 15 to 22 or 'G1' which is an integer from 0 to 20. The first thing that we need to do is saving the dataset in a pandas dataframe and then perform one hot encoding on categorical variables. Indeed, for each categorical variable v assuming values in a set S with $S = \{s_1, s_2, \dots, s_k\}$, we create k new columns and for each row we set the n-th created column equal 1 if and only if in the original column v assumed the value s_n, we set it to zero otherwise, we finally delete the original column. In this way we are mapping each value s_i to a vector with k components whose i-th component is equal to one and the other components are equal to zero. This is a common technique in Machine Learning because it allows to encode categorical values in numerical values on the top of which we can train our machine learning model. Then we plot the correlation matrix to discover which variables are the most correlated to the output we would like to predict (G3). The result is shown below. As we can easily observe, there is a high correlation between G3 and G1, G2, we have a low but significant positive correlation between G3 and the study time, the mother and father education and the desire of pursuing a higher education. On the other hand we have a negative correlation between G3 and the number of failures, the number of workday/weekend alchool consumption and the desire not to pursue a higher education. All this information are exactly what we would have expected. Then we split the data in X, a matrix containing all the features a part from G3 and Y a column vector containing just the G3 feature. We then split each of them in X_{train}, X_{test}, Y_{train}, Y_{test}. We will use X_{train} and Y_{train} to train our model and then we will evaluate the performance using X_{test} and Y_{test}. To answer this question we have considered the problem as a regressive problem and we have used the following models/approaches.

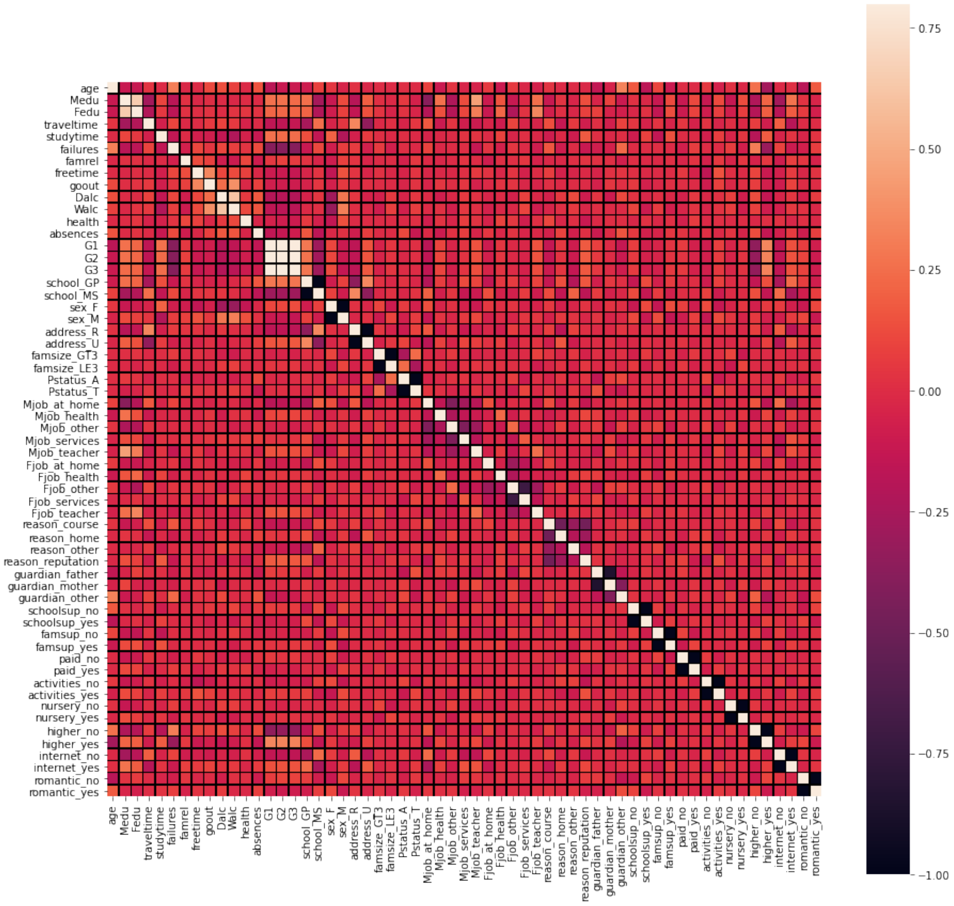
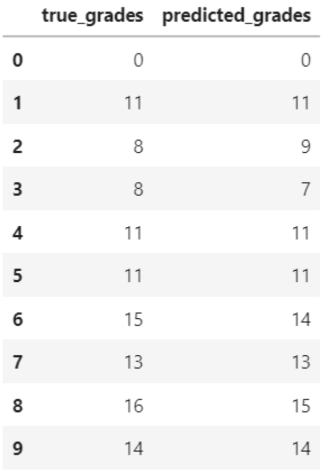


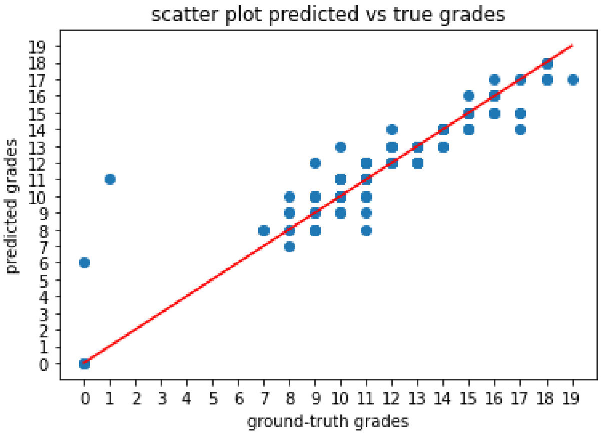
Figure 26: Correlation between the variables of the preprocessed dataset

1. We used a dense Neural Network with 7 layers with a large number of neurons for each layer. We used 'relu' as the activation function and we chose Adam as the optimizer. In addition to that, in order to prevent overfitting we applied the following early stopping technique. We have taken a subset of the training data as validation data and at the end of each epoch we computed the loss that the current model would have had on the validation data and we have saved the model corresponding to the smallest validation loss. After we have trained our model, we applied the model to X_{test} in order to obtain the predictions corresponding to Y_{test} we then have rounded the predictions because G3 has to be an integer value between 0 and 20. We computed the Mean Squared Error between the predicted and the ground-truth grades and we have obtained a value of 1.94. This information is not enough for us to understand if the model made a good fit or not for G3. Therefore, we tried to visualise the results obtained. This is how our results look like compared to the real grades.
   As we can see from the first 10 students the predictions are quite accurate, however we cannot really understand how good the predictions are just comparing the true grades and the predicted grades with a table. We will now use different plots to have a better perspective of the fit.



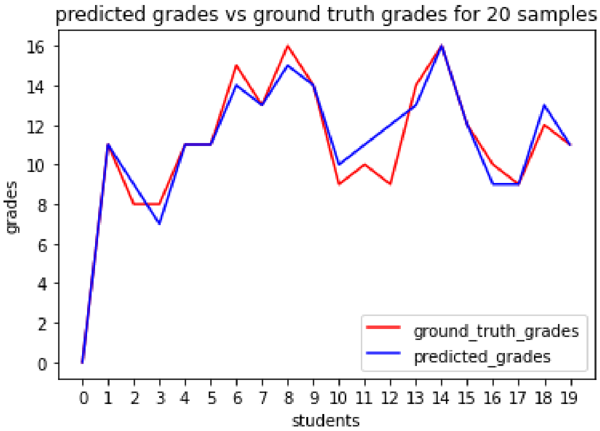
	true_grades	predicted_grades
0	0	0
1	11	11
2	8	9
3	8	7
4	11	11
5	11	11
6	15	14
7	13	13
8	16	15
9	14	14

The first plot is a scatter plot of the true and predicted grades.



scatter plot predicted vs true grades

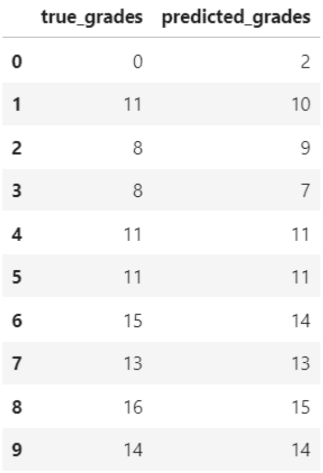
We would expect the points to be on the line y=x since our predicted grades should be equal to the true grades. As we can see from the plot this is indeed true but of course not all the points lie on the line and some variance around the line is observed. We then used the following line plot to compare the predictions with the true grades for the first 20 students.



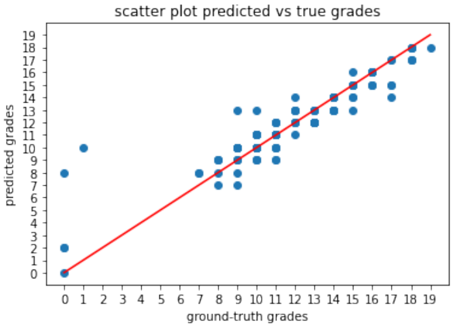
predicted grades vs ground truth grades for 20 samples

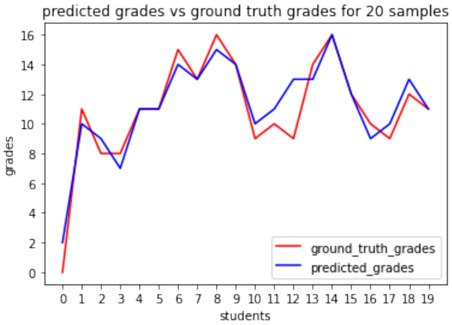
Again we observe that the predictions are quite accurate. All this plots suggested a good fit but in order to confirm our hypothesis we have used the Kolmogorov-Smirnov test for goodness of fit. Having as a null hypothesis that the distributions of the true grades and the one of the predicted grades were the same distribution we obtained a very large p-value suggesting us that we cannot reject the null hypothesis and consequently that our fit is a good one.

2. As a second model we have used the support vector machine for regression. We have tuned the hyperparameters of the model using grid-search trying out different kernels, degree and coefficients. This model showed a worse performance compared to neural nets, indeed the Mean Squared Error in this case turned out to be 2.11. The plots reflect the drop in performance.



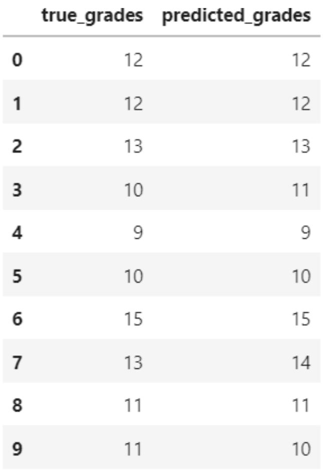
	true_grades	predicted_grades
0	0	2
1	11	10
2	8	9
3	8	7
4	11	11
5	11	11
6	15	14
7	13	13
8	16	15
9	14	14



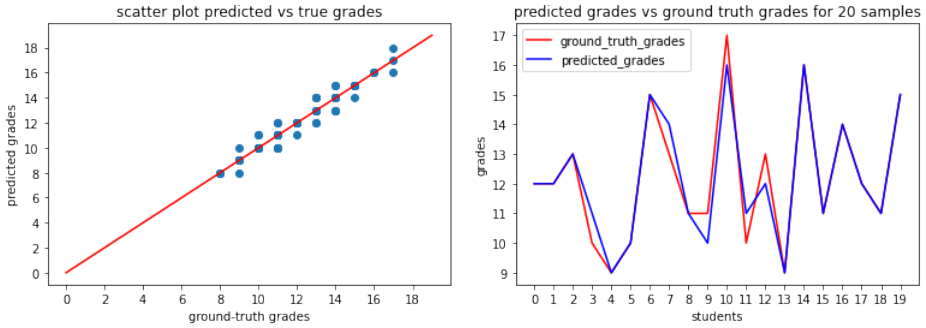


However, in this case too, we cannot reject the null hypothesis that the predicted grades and the true grades come from the same distribution using Kolmogorov-Smirnov test for goodness of fit.

3. To further increase the performance we can now train the neural net model on the same dataset but preprocessed such that no outliers are now present in the dataset. In this case we reach a mean squared error of 0.22 and as we can see the predicted grades are almost always correct.



	true_grades	predicted_grades
0	12	12
1	12	12
2	13	13
3	10	11
4	9	9
5	10	10
6	15	15
7	13	14
8	11	11
9	11	10

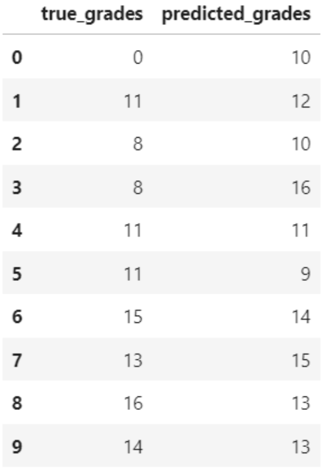


Clearly also in this case we cannot reject the null hypothesis of the Kolmogorov-Smirnov test for goodness of fit having a p-value of 0.9999999.

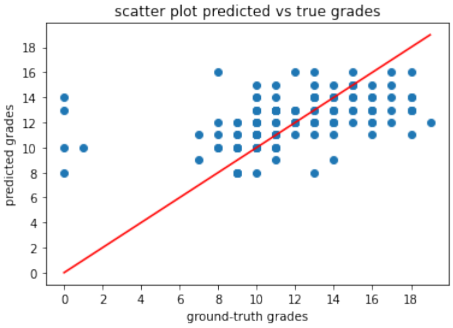
As we have seen our models work quite well and we can be satisfied of the reached fit. The next questions that might arise spontaneously is the following

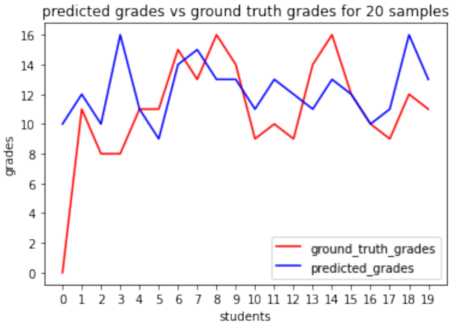
- *Can we train a model which can predict the final grade G3 based on all the other features except the previous grades G1 and G2 ?*

We have tackled this question training the neural net model on a new X'_{train} which is exactly the same as the X_{train} described before a part from the column related to G1 and G2 which are not present in X'_{train}. The same applies to X'_{test}. We followed the exact same procedure described in the answer to the previous question but in this case we have reached a mean squared error of just 11.02. The following plots testify this poor results.



	true_grades	predicted_grades
0	0	10
1	11	12
2	8	10
3	8	16
4	11	11
5	11	9
6	15	14
7	13	15
8	16	13
9	14	13

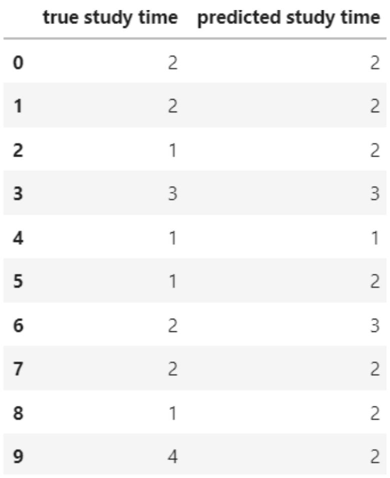




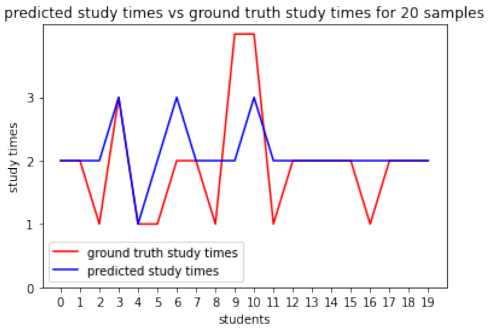
As we can see from the scatter plot there is a high variety in the difference between the predictions and the true grades and how can we see in the line plot this drives to a poor fit. Also the Kolmogorov-Smirnov test notice the poor fit and now the p-vaue for the null hypothesis is only 0.067. This does not indicate that we should reject the null hypothesis (with a significance level $\alpha = 0.05$) but it is an indicator of the poorness of the fit (since the p-value is quite small). We could justify the results using the correlation matrix plotted before. Indeed, how we can see, G3 are highly positive correlated with G1 and G2 so, removing this two information, the model cannot predict G3 as well as before. The next question that we are going to address is the following:

- *Can we predict the studytime based on the other features of the dataset ?*

We are going to answer to this question using a neural network based model since the best results obtained in the previous tasks were reached with this model. Obviously in this case after having split the dataset in training and test, Y_{train} will be just the column related to the studytime and X_{train} the columns for all the other features. The same applies to X_{test} and Y_{test}. In this case when we apply the model to X_{test} creating the prediction and when we compare them with Y_{test} we have the following results.



	true study time	predicted study time
0	2	2
1	2	2
2	1	2
3	3	3
4	1	1
5	1	2
6	2	3
7	2	2
8	1	2
9	4	2



We have obtained another poor fit testified by a really small p-value in Kolmogorov-Smirnov test (0.0042). The reason why we have such a bad fit lies probably on the fact that we do not have enough data correlated with the feature 'studytime'. Indeed, the dataset was collected to explain the grades of the students and not their studytime. As a direct consequence, despite the model we use to predict the studytime, we will never have a good fit due to the lack of crucial missing variables such as the time that students spend at home or in different other activities. The last question that we would like to give an answer to is the following:

- *Can we predict whether a student is in a romantic relationship or not based on the other features of the dataset?*

We have used three different models to answer to this question.

1. A first method based on dense neural networks which scored 0.62 in accuracy

2. A second method based on Support Vector Machine tuned with grid-search which scored 0.65 in accuracy

3. A third model based on Logistic Regression tuned with grid-search which scored 0.65 in accuracy.

Recalling the definition of accuracy:

$$\text{accuracy} = \frac{\text{number of correctly classified samples}}{\text{number samples}}$$

we notice that an accuracy around 0.65 is not high at all. Since we tried different models and we have always tuned the hyperparameters, we can claim that the problem lies in the lack of crucial variables as happened in the studytime analysis. Indeed, as we can observe from the correlation matrix, there are not highly positive or highly negative correlated variables with the romantic relationship variable.

7 Conclusions

First and foremost we must point out the importance of education in society. The school stage tends to have a high impact in the future of people, specially in their welfare. As a result it is crucial to set the education process in such a way that it leads to the highest possible students success. In order to do so we must analyze which factors contribute the most to the students performance. In this project we tried to better understand how the different aspects of the students life affect their performance at school. In particular we have studied which variables are either positively or negatively correlated with the grades. Moreover we tested if we can draw rigorous conclusions from these correlations. We have placed a special focus on the role of gender, study time and the family status on the final grade of the student. Apart from the grades we have also tested other relationships like gender and study time.

After our analysis, we have concluded with all the methodologies used that the previous grades are by far the most relevant feature to explain the final grade. For example in machine learning when we excluded the previous grades from the training set the performance of the learnt model dropped significantly. This happened independently of the particular model used. At this point we thought that maybe the rest of the data was just redundant to predict the final grade. However we tested in section 4 that this is not the case, so that the other data is also useful. Therefore we decided to study more in depth the effect of the other variables on the final grade. We conclude that females tend to get higher grades than males. This could be caused by the fact that girls also study more as we have discovered from our hypothesis tests. Nevertheless, although we would expect that the study time is an important factor for the final grade we have found that this is only the case once the student studies at least 5 hours. Therefore the study time variable may not be as relevant as we thought at first. In fact when we built the model to explain the final grades of the students this variable was not necessary to achieve a good fit. Another factor that could seem to be relevant at first sight is the family relations but from our analysis we conclude that it is not the case. It was not necessary to include it in the linear model neither. Other aspects for which we have not tested its statistical significance to predict the grade but that were useful in the linear model are the variables: school, higher, absences and failures. Among these ones the one that turned out to be the most important is if the student wants to pursue higher education or not. Finally among the factors that could negatively affect the grade we found through the linear model that the number of absences and failures are the most relevant.

Beyond understanding which factors are important for the students success we wanted to provide a powerful methodology to predict the final grade in advance. This could have many positive implications because then we would be able to asses the possible needs of the students. We mainly built two type of models. On the one hand we have a model to classify students who pass or fail. It achieves 92% general accuracy and we have provided some modifications to achieve the highest possible rate of failed students prediction. On the other hand we have the most complex models among our methodologies. These achieve very good performance to predict the exact grade of the students with the MSE being lower than 2 on unseen data.