December 29, 2023

Advanced Data Analysis with R

Assignment 1: Student Performance Data Set Assignment 15

ANARGYROS TSADIMAS
AM: f3612318
Professor: Ioannis Ntzoufras

Table of Contents

Table of Figures

Table of Tables

Abstract

Despite significant advancements in the educational standards of Portugal over recent decades, the country still trails behind other European nations, largely due to high student failure rates. This problem is particularly pronounced in key subjects like Mathematics and the Portuguese language. To address this issue, we plan to analyze data gathered from school reports and questionnaires. Our approach involves identifying critical factors influencing students' grades and employing regression models for deeper analysis. We aim for our study to provide valuable insights and contribute to resolving this educational challenge.

1.Introduction

The file we analyze includes measurements for 649 observations that each and every one of them describes a student of two Portuguese schools, using 33 variables that are explained in the table below(Table 1).

# of variables	*Variable's name*	*Type*	*Meaning*	*Value*
1	***school***	*binary*	*student's school*	*GP -Gabriel Pereira or MS - Mousinho da Silveira*
2	***sex***	*binary*	*student's sex*	*F- female or M- male*
3	***age***	*numeric*	*student's age*	*from 15 to 22*
4	***address***	*binary*	*student's home address type*	*U -urban or R- rural*
5	***famsize***	*binary*	*family size*	*LE3- less or equal to 3 or GT3- greater than 3*
6	***Pstatus***	*binary*	*parent's cohabitation status*	*T- living together or A- apart*
7	***Medu***	*numeric*	*mother's education*	*0 -none, 1-primary education (4th grade), 2-5th to 9th grade, 3-secondary education, 4-higher education*
8	***Fedu***	*numeric*	*father's education*	*0 -none, 1-primary education (4th grade), 2-5th to 9th grade, 3-secondary education, 4-higher education*
9	***Mjob***	*nominal*	*mother's job*	*'teacher', 'health' care related, civil 'services', 'at home' or 'other'*
10	***Fjob***	*nominal*	*father's job*	*'teacher', 'health' care related, civil 'services', 'at home' or 'other'*
11	***reason***	*nominal*	*reason to choose this school*	*close to 'home', school 'reputation',*

				'course' preference or 'other'
12	***guardian***	*nominal*	*student's guardian*	*'mother' , 'father' , 'other'*
13	***traveltime***	*numeric*	*home to school travel time*	*1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour*
14	***studytime***	*numeric*	*weekly study time*	*1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours*
15	***failures***	*numeric*	*number of past class failures*	*n if 1<=n<3, else 4*
16	***schoolsup***	*binary*	*extra educational support*	*yes or no*
17	***famsup***	*binary*	*family educational support*	*yes or no*
18	***paid***	*binary*	*extra paid classes within the course subject*	*yes or no*
19	***activities***	*binary*	*extracurricular activities*	*yes or no*
20	***nursery***	*binary*	*attended nursery school*	*yes or no*
21	***higher***	*binary*	*wants to take higher education*	*yes or no*
22	***internet***	*binary*	*Internet access at home*	*yes or no*
23	***romantic***	*binary*	*with a romantic relationship*	*yes or no*
24	***famrel***	*numeric*	*quality of family relationships*	*from 1- very bad to 5 - excellent*
25	***freetime***	*numeric*	*free time after school*	*from 1 - very low to 5- very high*
26	***goout***	*numeric*	*going out with friends*	*from 1 - very low to 5- very high*
27	***Dalc***	*numeric*	*work day alcohol consumption*	*from 1 - very low to 5- very high*
28	***Walc***	*numeric*	*weekend alcohol consumption*	*from 1 - very low to 5- very high*
29	***health***	*numeric*	*current health status*	*from 1- very bad to 5 - very good*
30	***absences***	*numeric*	*number of school absences*	*from 0 to 93*
31	***G1***	*numeric*	*first period grade*	*from 0 to 20*
32	***G2***	*numeric*	*second period grade*	*from 0 to 20*
33	***G3***	*numeric*	*final grade*	*from 0 to 20*

Table 1:Variables description

In this assignment we examine their relationship between the grades and all the other variables. We will perform descriptive analysis for the most important variables and pair wise associations between them. Finally ,we will construct a linear model that describes the data and also have predictive power.

2.Descriptive analysis and exploratory data analysis

In this section we will analyze and present our data. After we import them in R studio we will eliminate some of the observations that are considered missing or damaged values which leaves us with the final sample that consists of 632 observations.

First the categorical variables.

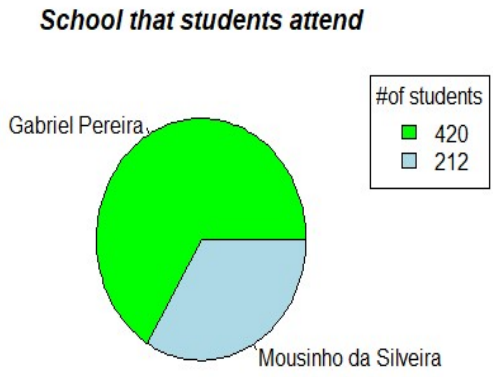


Figure 1:Pie chart of school of origin

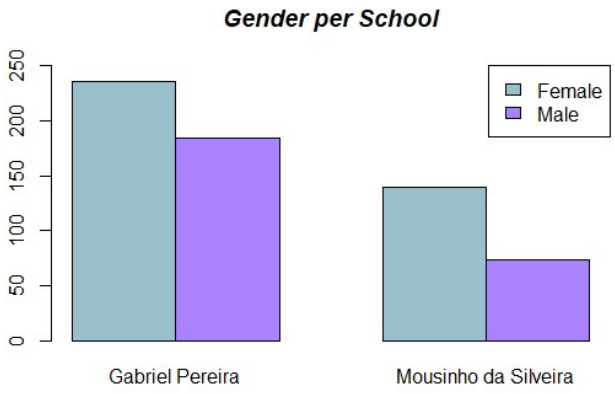


Figure 2:Bar chart of gender per school

Here we see (Figure 1 and Figure 2) how many students from each school we have and also a clearer picture of the gender of our sample.

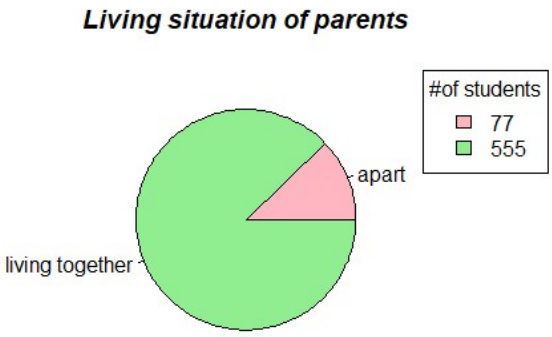


Figure 3:Pie chart of parents' living situation

Only 12% of the parents leave apart (Figure 3).

Then I merged the variables about parents educational level into one with three levels: no parent with higher education, one parent with higher education and two parents with higher education.

Overall Education Level

parents_higher_ed	Freq	Percentage
None	429	67.88
One	110	17.41
Both	93	14.72

Education Level by School

parents_higher_ed	Var2	Freq	Percentage
None	GP	257	40.66
One	GP	93	14.72
Both	GP	70	11.08
None	MS	172	27.22
One	MS	17	2.69
Both	MS	23	3.64

Table 2:Parents' higher education frequency table

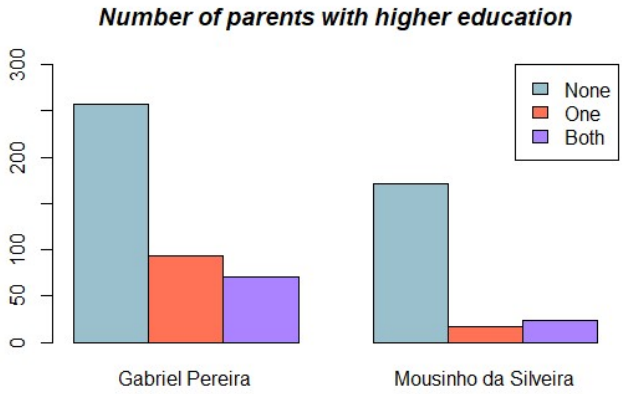


Figure 4:Clustered Bar chart of Parents' higher education

68% of students have none parent with higher education while around 15% has one or two (Table 2 and Figure 4).

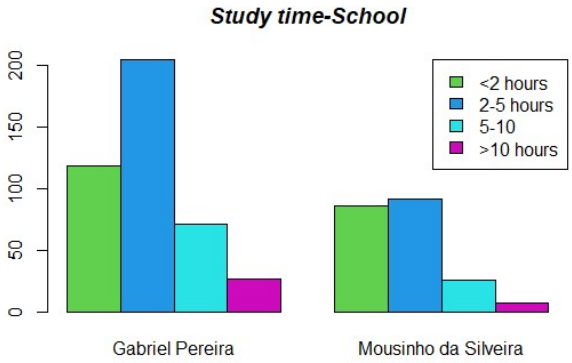


Figure 5:Clustered Bar chart of Student's study time per school

Here is a bar chart of study time with almost half the student studying between two to five hours weekly (Figure 5).

Now the numerical variables.

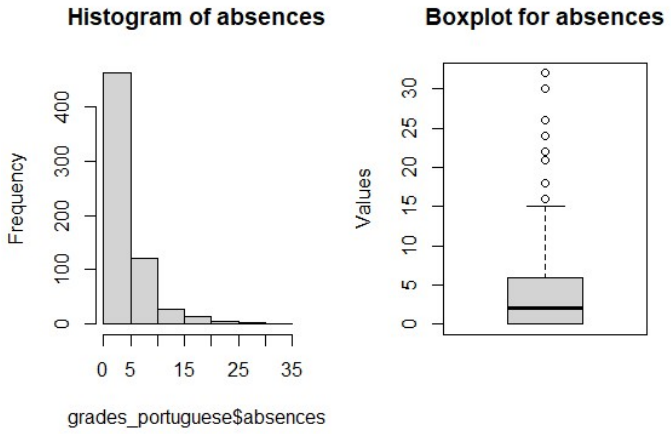


Figure 6:Histogram and boxplot of absences

Here is a histogram and the boxplot for absences (Figure 6). Most students how below 5 with the max being around 30. Though these values appear as outliers they are legitimate values so we will keep them in our sample, and we also see that absences aren't normally distributed.

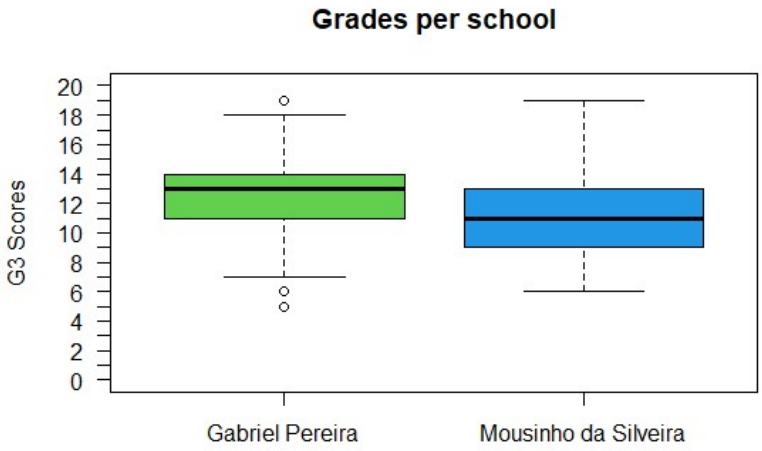


Figure 7:Boxplot of final grades per school

Statistics of G3

	mean	sd	median	min	max
X1	12.21	2.66	12	5	19

Table 3:Description of final grades variable

Here we have the final grades from each school (Figure 7). The mean and the median are around 12 (Table 3) though in GP we have higher mean and median which means that in general the students did better than MS but also more outliers on the low end, and more symmetric distribution which means more consistent grades than MS which is skewed on the right. In both cases no normality appears on the plot.

Finally let's see plots between first semester grades, second semester grades and final grades (Figure 8). There appears to be a strong positive relationship as expected and as we will see later on the correlation matrix(Figure 9).

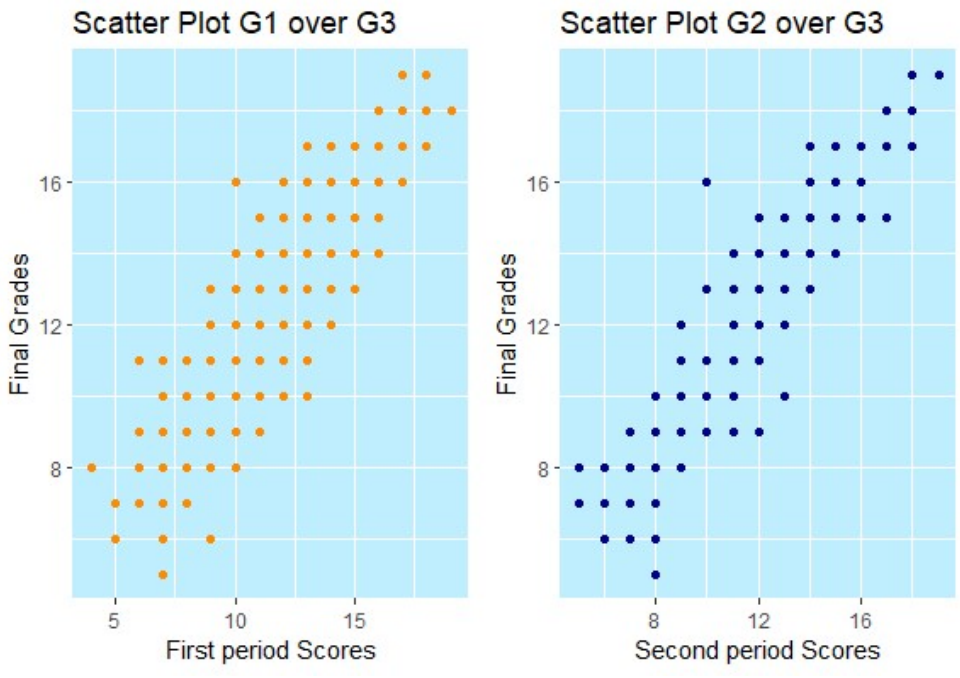


Figure 8:Scatter plot of first and second period against final grades

3.Pairwise comparisons

In this section of the report we will conduct pairwise comparisons between the variables to further analyze our data and draw better conclusions.

First we transformed the three grades variables and merged it into one which is the average and then we made a correlation matrix for all the quantitative variables to see how they relate to each other. As we see no significant relationship exists(Figure 9).

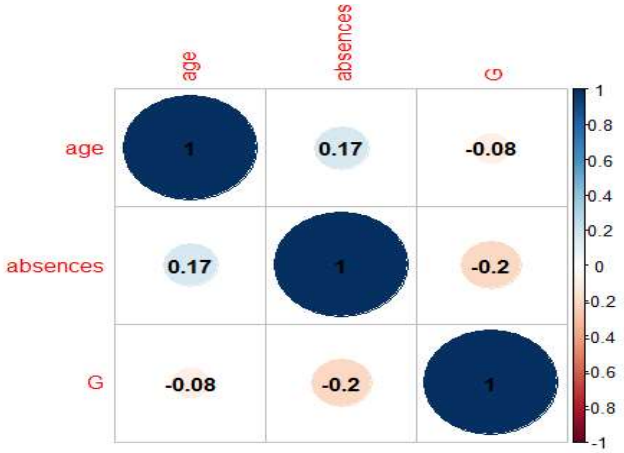


Figure 9:Correlation matrix for numeric variables

Then we will check the variables for normality both visually and using tests to conclude.

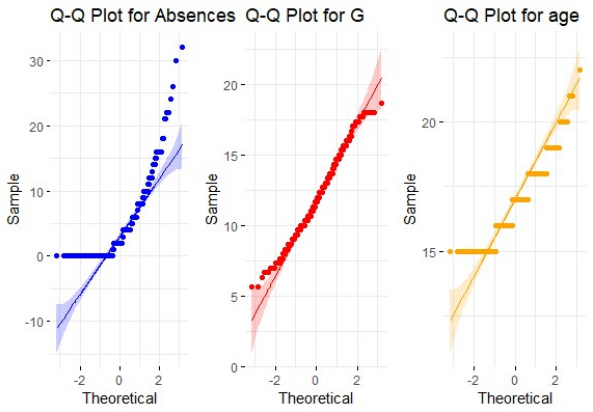


Figure 10:QQ plots of numeric variables

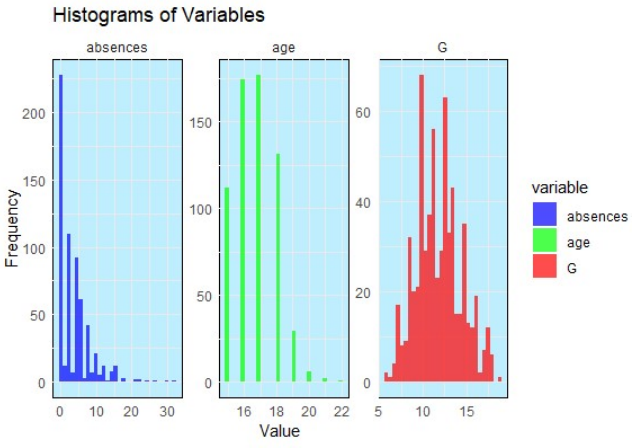


Figure 11:Histogram of numeric variables

Normality is rejected for all the variables both visually (Figure 10 and 11) and by the tests *(for G : Shapiro-Wilks p=0.00004217<0.05, Lilliefors p=0.0000002038<0.05)* , *(for absences : Shapiro-Wilks p=0.00000000000022<0.05, Lilliefors p=0.00000000000022<0.05)*. Then we will take the normality using residuals.

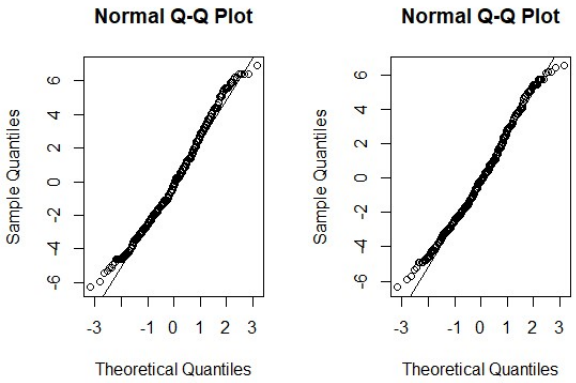


Figure 12:Residual plots of age and absences against grades

Good residual plot (Figure 12) but the tests again reject normality (*for G against age : Shapiro-Wilks p=0.000031<0.05 , for G against absences : Shapiro-Wilks p=0.00014<0.05).*

Now pair wise comparisons for the categorical variables.

Checking school and address variables we see that are relationship exists both in the table below (Table 4) and by the tests *(Pearson's Chi-squared test p=0.00000000000000022 <0.05).*

GP has more urban population while MS is more balanced.

	address		
school	R	U	Row Total
GP	78	342	420
	17.229	7.240	
	0.186	0.814	0.665
	0.417	0.769	
	0.123	0.541	
MS	109	103	212
	34.133	14.344	
	0.514	0.486	0.335
	0.583	0.231	

	0.172	0.163	
Column Total	187 0.296	445 0.704	632

Table 4:Address per school

Other related variables are school and parents higher ed *(Pearson's Chi-squared test p=0.00000099 <0.05)* with GP having more educated parents, school and study time with GP students studying more time *(Pearson's Chi-squared test p=0.0099 <0.05)* and sex and alcohol consumption both during the week and the weekend *(for Dalc: Pearson's Chi-squared test p=0.000000000028 <0.05, for Walc: Pearson's Chi-squared test p=0.0000000000000076 <0.05).*

Finally very wise associations between categorical and numerical variables.

First who will see the relationship of grades and sex (Figure 13).

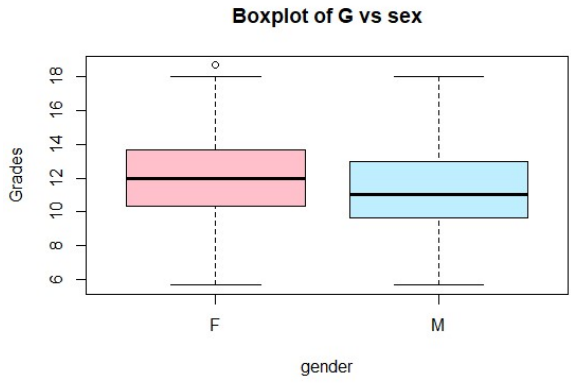


Figure 13:Boxplot of grades per school

Similar variances , normality is rejected for both male and female so we take a non parametric test *(Levene's p=0.42>0.05* , for F : *Shapiro-Wilks p=0.037<0.05. for M: Shapiro-Wilks p=0.00026<0.05, Wilcoxon p=0.001<0.05)* and conclude that medians differ significantly.

Now let's examine the relationship between grades and parents higher education.

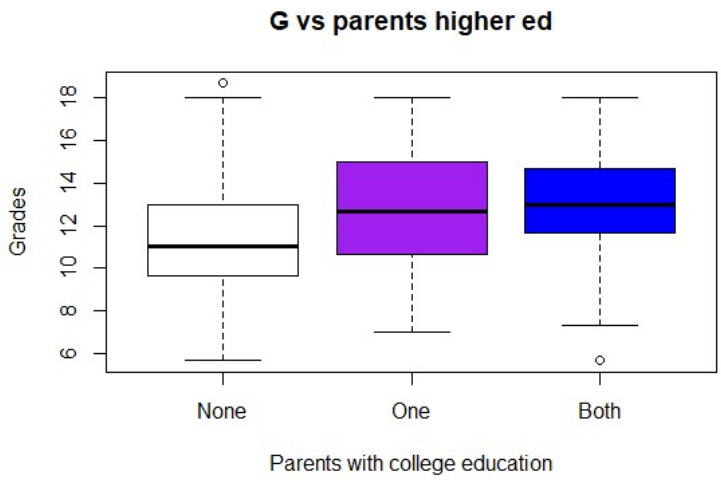


Figure 14:Boxplot of grades per parents education

We see (Figure 14) that going from none two one or both there is difference so not having a parent with college education affects negatively the grades. *(Levene's p=0.30>0.05* similar variances, *anova p=0.000000000017<0.05 parents higher education important, Kruskal-Wallis p=0.0000000024<0.05 parents affect final grades, pairwise t-test p=0.00000024<0.05 from none to one , p=0.000000046<0.05 from none to both , so there is difference).*

4.Predictive and Descriptive models

Now that we have analyzed our variables we can proceed to construct our model.

We start from the full model, the one that takes into account all the variables and examine which of them appear significant (R-squared= 0.43). In the ANOVA table below (Table 5) only 13 of them appear to be statistically significant.

```
Analysis of Variance Table

Response: G
                Df  Sum Sq Mean Sq F value                  Pr(
>F)
school           1  246.72 246.720 60.2806   0.00000000000003
854 ***
sex              1   80.82  80.818 19.7460   0.00001063703412
387 ***
age              1   21.86  21.858  5.3404             0.0211
937 *
address          1   12.97  12.973  3.1697             0.0755
495 .
famsize          1    7.14   7.139  1.7442             0.1871
374
Pstatus          1    3.85   3.849  0.9405             0.3325
617
Mjob             4  168.42  42.104 10.2872   0.00000004769812
943 ***
Fjob             4   71.60  17.899  4.3733             0.0017
210 **
reason           3   98.72  32.907  8.0400   0.00002966046627
806 ***
guardian         2   50.96  25.480  6.2255             0.0021
160 **
traveltime       3   12.56   4.186  1.0227             0.3821
001
studytime        3  152.59  50.862 12.4271   0.00000006993957
498 ***
failures         3  419.59 139.864 34.1727 < 0.00000000000000
022 ***
schoolsup        1   47.65  47.649 11.6419             0.0006
907 ***
famsup           1    6.74   6.736  1.6457             0.2000
712
paid             1    0.07   0.067  0.0164             0.8982
176
activities       1    6.30   6.295  1.5381             0.2154
119
nursery          1    0.28   0.277  0.0678             0.7947
239
higher           1  143.98 143.983 35.1791   0.00000000522754
609 ***
internet         1    0.78   0.784  0.1916             0.6617
459
romantic         1   11.73  11.734  2.8670             0.0909
609 .
famrel           4   29.48   7.371  1.8008             0.1271
422
freetime         4   29.11   7.279  1.7784             0.1316
321
goout            4   52.90  13.224  3.2309             0.0122
885 *
```

```
Dalc                4   12.98   3.245  0.7929                0.5300
341
Walc                4   18.80   4.699  1.1482                0.3328
851
health              4   29.82   7.456  1.8217                0.1231
050
absences            1   45.51  45.507 11.1186                0.0009
104 ***
parents_higher_ed   2   31.11  15.554  3.8002                0.0229
361 *
```

Table 5:ANOVA of full linear model with all the variables

Now we will construct a linear model just with those 13 variables (R-squared= 0.39). This is our first option and we will compare it with other models. Then using the stepwise method where found two more models (R-squared= 0.41, using two way and backwards direction and R-squared= 0.43 using forward direction). So I will choose the first model because the extra variables do not explain the data significant significantly better, and we would prefer a simpler ,faster, less costly model.

Now that we have our model we we take the four assumptions homoscedasticity ,linearity, independence of errors end normality. In the graphs below (Figure 15) homoscedasticity and linearity are violated, so we will need a transformation for our response variable.

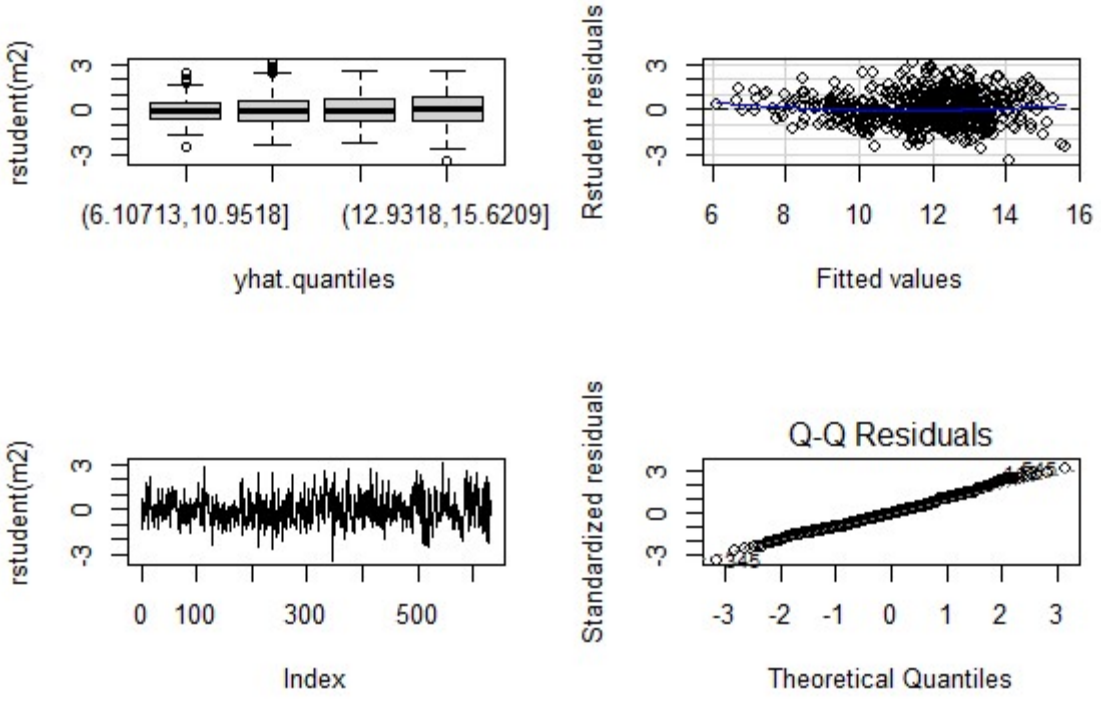


Figure 15:Plots for the four assumptions for model with 13 variables

We have applied a logarithmic transformation to the 'grades' variable. In our revised and final model, all four assumptions are now satisfied. (Figure 16) *(Levene's p=0.397>0.05 homoscedasticity exists,good plot of studentized residuals against the fitted values,independence of errors and normality exists Lilliefors p=0.13>0.05*).We do not have multicollinearity (vif<2 for all variables).

The final model is :

```
m4<-
lm(log(G)~school+sex+age+Mjob+Fjob+reason+guardian+studytime+failures+schoolsup+higher+goout+absences+
parents_higher_ed)
```

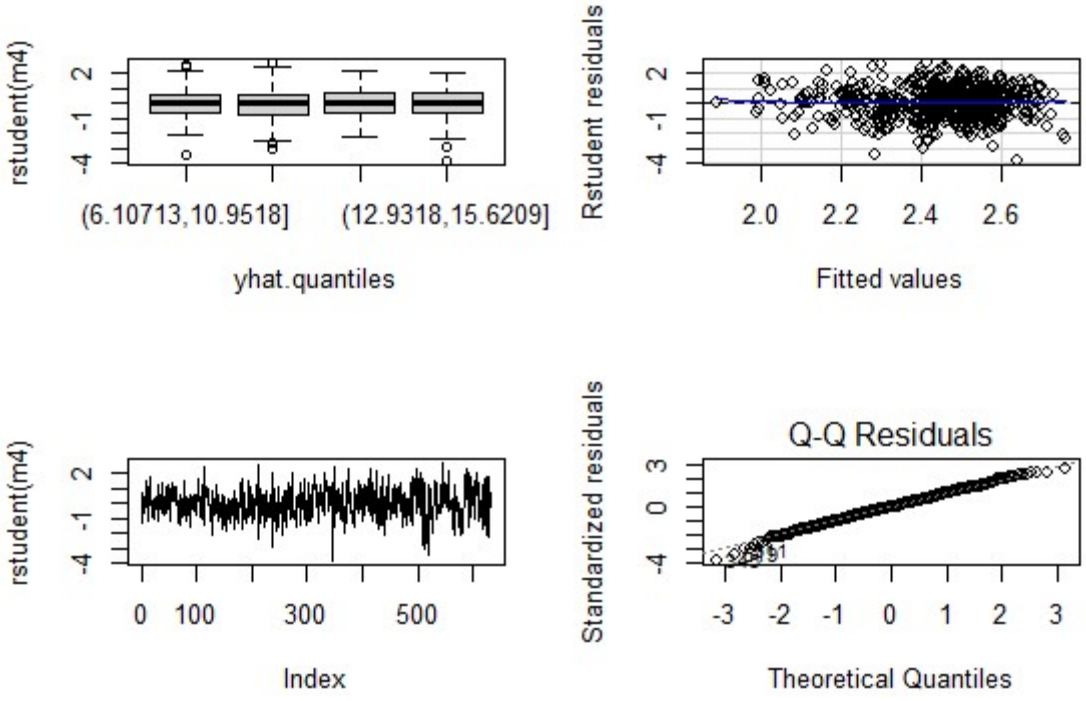


Figure 16:Plots of four assumptions of the final logarithmic model

5.Conclusions

After concluding the analysis, we can make several observations. Factors such as the school, study time, and the jobs and education levels of parents play important roles in determining students' grades. This information can be useful for predicting or influencing students' academic outcomes. Further research and analysis in the future could strengthen and validate our findings.

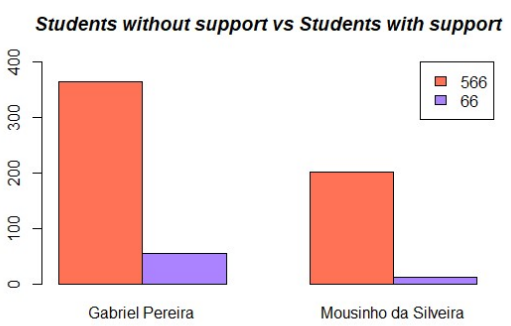
6.References

[1] Ntzoufras I. ,(2023) Advanced data analysis with R, educational notes for MSc program Statistics AUEB

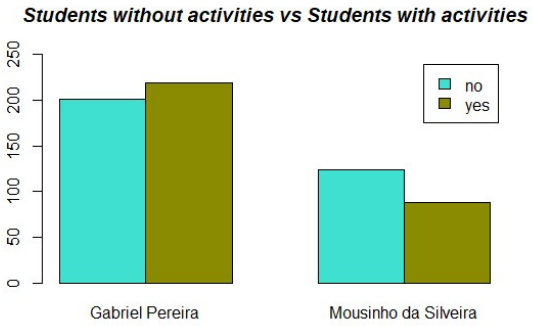
[2] Paulo Cortez and Alice Silva, (2008) , *Using Data Mining to Predict Secondary School Student Performance* , In A. Brito, & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12.

7.Appendix

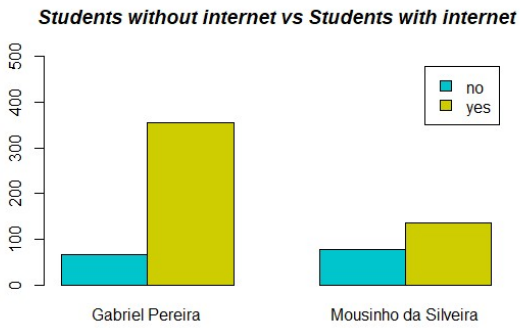
Some extra figures that we checked during the analysis but were considered not that important to make the report.Bar charts and pie charts for categorical variables.



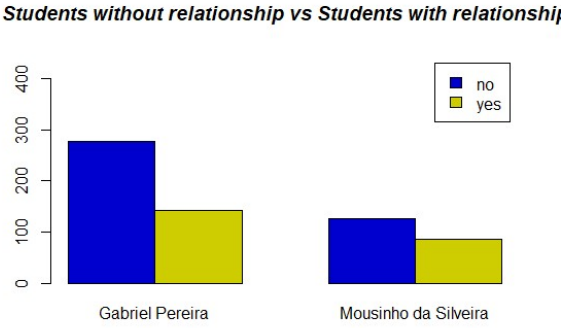
17:Students with support per school



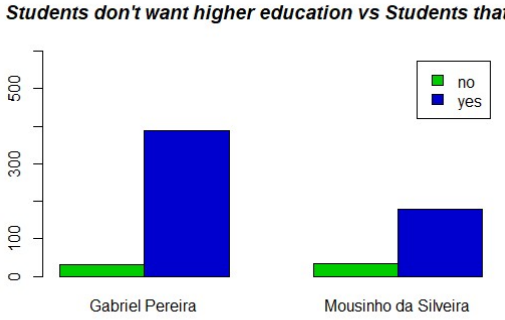
18:students with activities per school

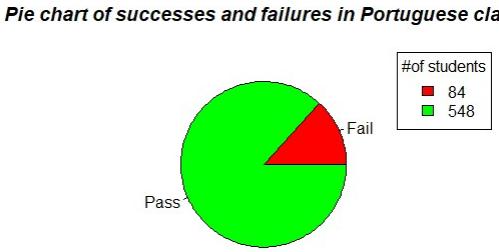


19:Students with internet per school



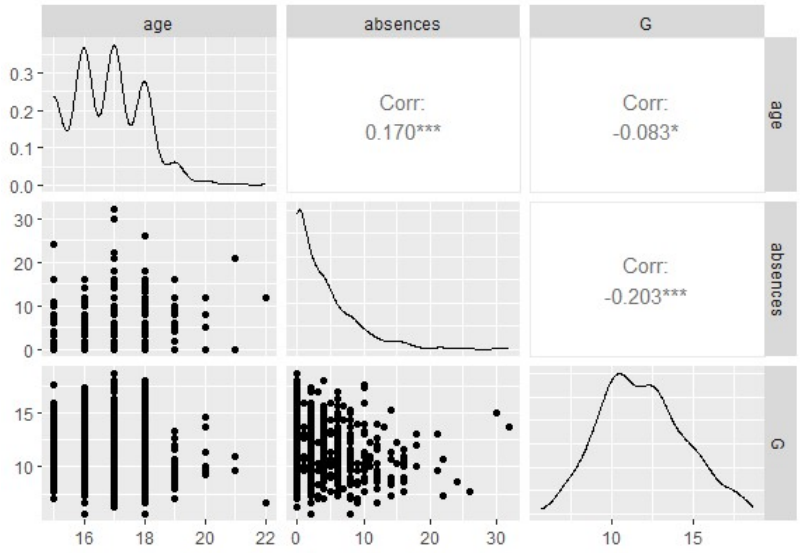
20:Students with relationship per school





21:Students that want higher education per school

22:Successes and failures in Portuguese class



23:Scatterplot matrix for numeric variables

Enhanced scatterplot matrix with GGally package for correlation between numeric variables.

```
G - Mean: 11.82911 Median: 11.66667
age - Mean: 16.71835 Median: 17
absences - Mean: 3.751582 Median: 2
```

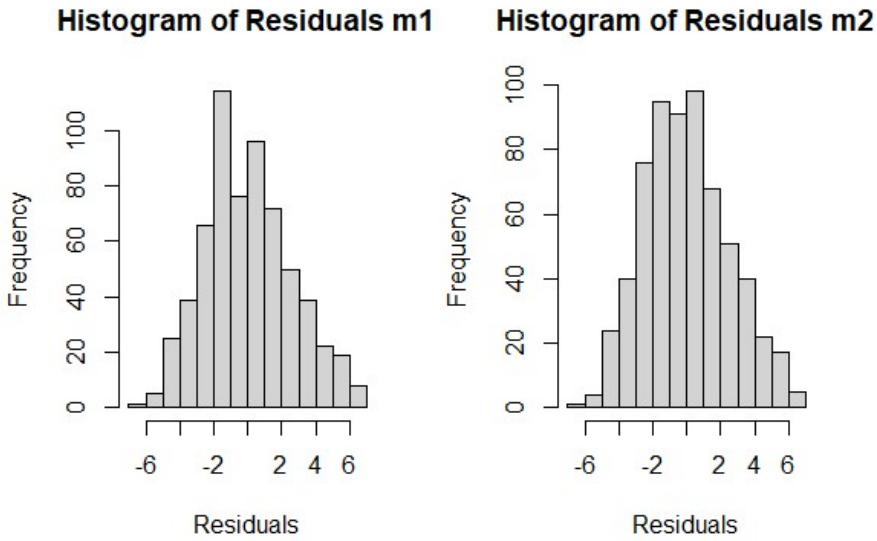
Table 6:mean and median for numeric variables

Checking the mean and the median for symmetry in numeric variables.

```
by(G,school,lillie.test)
school: GP

      Lilliefors (Kolmogorov-Smirnov) normality test

data:  dd[x, ]
D = 0.061166, p-value = 0.0007032

------------------------------------------------------------
----
school: MS

      Lilliefors (Kolmogorov-Smirnov) normality test

data:  dd[x, ]
D = 0.1354, p-value = 0.000000000392
```

Table 7:Lilliefors test for grades per school

Also checking for normality of grades per school using Lilliefors test.Again normality is rejected with small p-values.



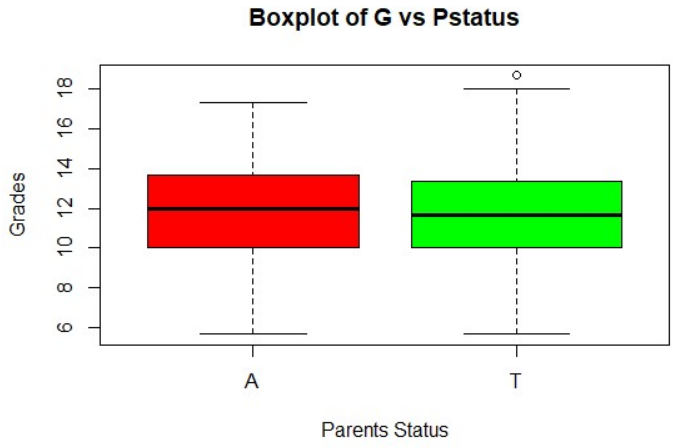
24:Histogram of residuals of age and absences against grades

Checking the relationship between sex and failures we found no association (*Pearson's Chi-squared test p=0.1212>0.05) as well as famrel and Pstatus (Pearson's Chi-squared test p=0.33 >0.05) and famrel and parents higher ed (Pearson's Chi-squared test p=0.919 >0.05) which I*

found to be very interesting . On the other hand the relationship between guardian and Pstatus is statistically significant with fathers being guardians more often when the parents live together (Pearson's Chi-squared test p=0.0002<0.05)

guardian	Pstatus A	Pstatus T	Row Total
father	6	143	149
	8.137	1.129	
	0.040	0.960	0.236
	0.078	0.258	
	0.009	0.226	
mother	61	382	443
	0.915	0.127	
	0.138	0.862	0.701
	0.792	0.688	
	0.097	0.604	
other	10	30	40
	5.393	0.748	
	0.250	0.750	0.063
	0.130	0.054	
	0.016	0.047	
Column Total	77	555	632
	0.122	0.878	

Table 8:Cross table for guardian and Pstatus

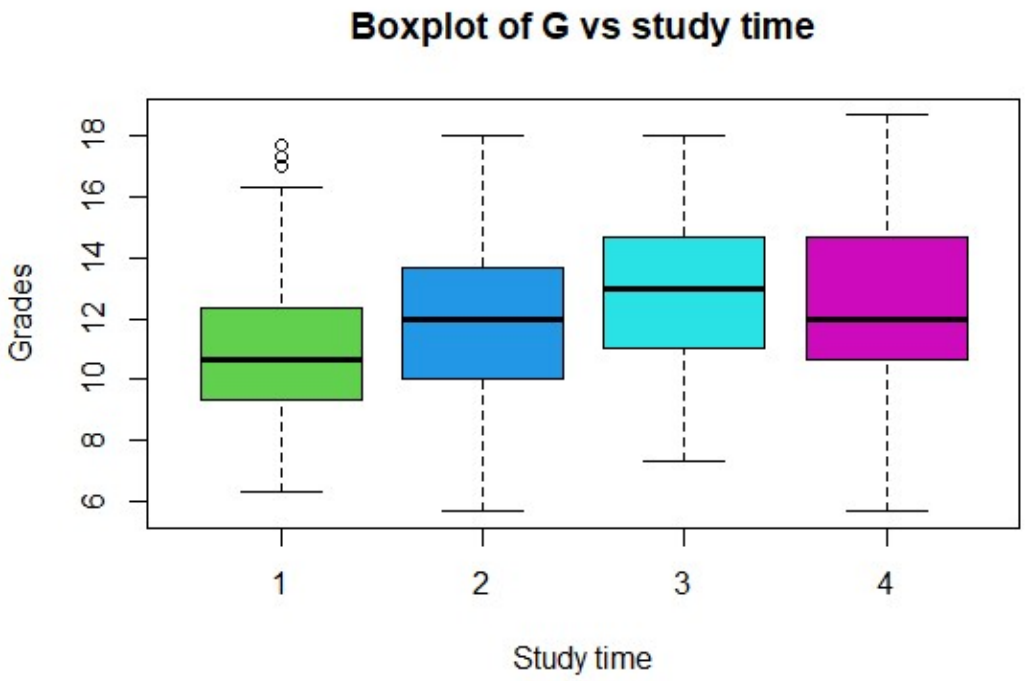


25:Boxplot of grades and Pstatus

Checking the grades and Pstatus per school we see similar variances *(Levene's p=0.99>0.05)*

Similarly and normally distributed (*Kolmogorov-Smirnov test p=0.99>0.05 , Lilliefors test p=0.27>0.05 for parents that live apart*) and finally similar means (*t.test p=0.96>0.05*).

In conclusion for the pairwise association analysis between categorical and numeric variables we have examined grades and study time.



26:Boxplot of grades and studytime

We have similar variances *(Levene's test p=0.136>0.05)* and study time is statistically significant

(ANOVA p=0.00000000055<0.05), and since we don't have normality (Shapiro-Wilks test p=0.002<0.05) and non-parametric test is more suitable (Kruskal-Wallis test p=0.000000077<0.05).

Using the pairwise t-test as well as in the plot , we see that there is difference going from the first level to all the others ,so studing less than two hours per week affects the grades negatively.