Middle East Technical University
Department of Statistics

STAT 291
STATISTICAL COMPUTING I
FINAL PROJECT

"Students' Performance in Mathematics Lessons"

AYŞEGÜL BİNBAŞ

DOĞUKAN ÖZTÜRK

FIRAT HACIŞAHİNOĞULLARI

HAZAL KARAKOÇ

Table of Content

Abstract

In this paper, the analysis aims to investigate secondary school students' performance in Mathematics classes. In this analysis, research questions were answered with graphical methods, statistical tests and linear regression model. All works are done on R programming under markdown package.

Keywords

ANOVA, Linear Regression, Two-sided T-test, Correlation, Covariance, Box-Cox

1. INTRODUCTION

There are many things that can affect students' performance such as study time of students, absences and free time activities. These examples do not cover all the effects because there are so many other factors. In fact, "The factors that seem to be related to resiliency can be organized into four categories: individual attributes, positive use of time, family, and school" (Mcmillan & Reed, 1994). According to Mcmillan and Reed, those four factors have an effect on resilient students' performance.

In this study, the main research objective is to find out which factors can affect the students' performance in mathematics classes. For that purpose, student performance data set is taken from the website; https://archive.ics.uci.edu/ml/datasets/Student+Performance. The data named as "student-mat" was used in this project to do research on students' performance in mathematics. Also, "This study will consider data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal." (P. Cortez & A. Silva, 2008)

Data

The data set have 395 observations and 33 variables. There are one continuous, one discrete, three interval, five nominal, ten ordinal and thirteen binary variables in this particular data set. The grades are categorized as first period, second period and final grade, and the range of grades is 0 to 20.

Out of 395 respondents, there are 208 female and 187 male secondary school students in this study. Thus, this shows that there is almost a balanced in gender. Since this study is

conducted on secondary school students, the range of age is between 15 and 22. While 71 % of students have family size greater than three, 29 % of them have family size less than or equal to three. Moreover, 90 % of students' parents are living together; on the contrary, 10 % of students' parents are living apart. Furthermore, 69 % of students' guardian is their mother, 23 % of students' guardian is their father and 8 % of students' guardian is another person.

Significance of the Study

Although this study is conducted on secondary school students, the data set has a variety of variables that can affect students' performance in mathematics classes. Thus, finding out which factors have an effect on students' performance is the main objective in this study. At the end of the study, students can be aware of what can affect their performance in mathematics, and then they may change those factors for their own benefit.

2. AIMS OF RESEARCH

Main Objective

The main research objective is to find out which factors can affect the students' performance in mathematics classes.

Minor Objectives

1. Do final grades in Mathematics of students change according to their study time?
2. Do absences of students differ for being in a romantic relationship or not?
3. Are final grades in Mathematics of students and their absences related?
4. Which factor has more effect than all others on final grades in Mathematics?

3. ANALYSIS

3.1 Descriptive Statistics

Ages of students:

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
15.0	16.0	17.0	16.7	18.0	22.0

Table 1: *Descriptive statistics of ages of students*

The minimum age of students is 15 while the maximum is 22. The average of students' ages is 16.7 which can be rounded as 17. Also, the median of ages of them is 17, too.

Absences of students:

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0.000	0.000	4.000	5.709	8.000	75.000

Table 2: *Descriptive statistics of absences of students*

The minimum number of the absences is 0 while the maximum is 75. The average number of not going to the school is 5.709 which can be rounded as 6 days.

First period grades of students:

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
3.00	8.00	11.00	10.91	13.00	19.00

Table 3: *Descriptive statistics of first period grades of students*

The minimum grade of first period grade is 3 while the maximum is 19. The median of first period grade is 11. Also, the average grade is 10.91 which can be rounded as 11. Those results can be interpreted as; none of the students got zero point out of twenty and the average of the grades equal to the median.

Second period grades of students:

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0.00	9.00	11.00	10.71	13.00	19.00

Table 4: *Descriptive statistics of second period grades of students*

The minimum grade of second period grade is 0 while the maximum is 19. The median is again 11. Also, the average grade is 10.71 which can be rounded to 11. Those results can be interpreted as; some of the students got zero point out of twenty and the average of the grades equal to the median.

Final grades of students:

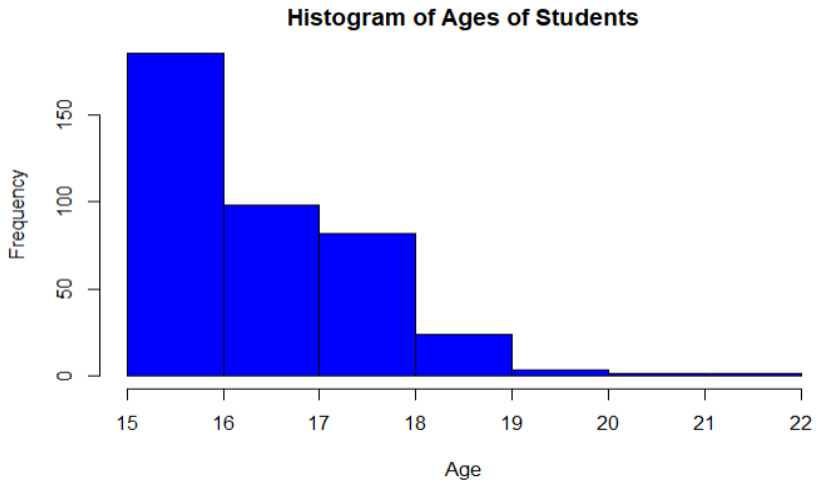
Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0.00	8.00	11.00	10.42	14.00	20.00

Table 5: *Descriptive statistics of final grades of students*

The minimum final grade of students is 0 while the maximum is 20. The median is again 11. Also, the average final grade of students is 10.42 which can be rounded to 10. Those results can be interpreted as; some of the students got zero out of twenty while some of them got twenty out of twenty. Also, the average of grades almost equal to the median.

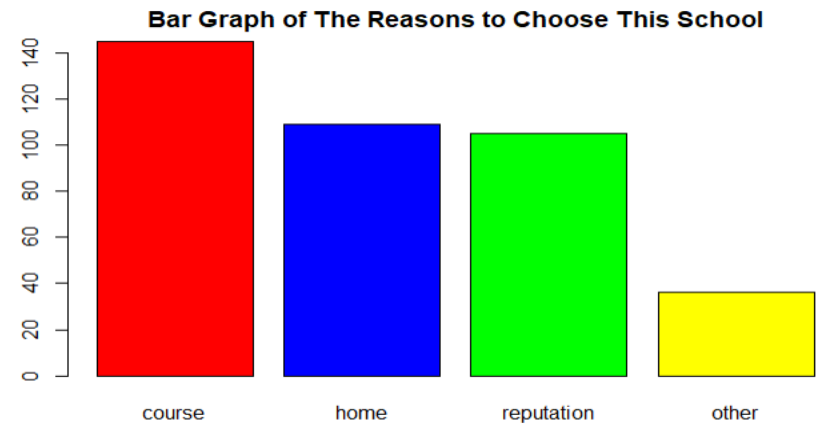
3.2 Background Information

In this part, graphical tools which are histogram, bar graph and pie chart was used to illustrate the background information of respondents.



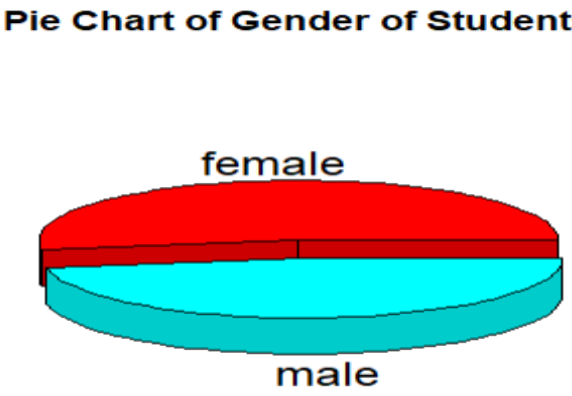
According to Figure 1, age variable has a right skewed shape which means that age variable is rightly skewed. Also, most of the students' ages are between 15 and 16 while the least ones are between 20 and 22.

Figure 1: Histogram of Ages of Students



According to Figure 2, most of the students chose their school because of the course preference. The other reasons for choosing those schools are being close to home and school reputation, respectively. Other reasons are not priority to choose the school.

Figure 2: Bar Graph of the Reasons to Choose This School



According to Figure 3, there seems to be a balance in the gender of students. That is, the percentages of female and male students are very close.

Figure 3: Pie Chart of Gender of Students

3.3 The Relationship between Final Grades and Study Time

In order to investigate the relationship between final grades of students and their study time, some statistical tests and graphical tool were used.

H_o: $\mu_1=\mu_2=\mu_3=\mu_4$

H_1: $\mu_1\neq\mu_2\neq\mu_3\neq\mu_4$

The null hypothesis of this research question is the average of final grades equal to in each group of study time. Since the study time is an ordinal variable, ANOVA Table can be performed.

In order to do diagnostic check before conducting the ANOVA Table, Shapiro-Wilk Normality test was performed on the each group. Only the group which is students' weekly study time is more than 10 hours was found as normally distributed. After, Barttlet Test of Homogeneity of Variances was performed in order to check the homogeneity between groups. The p-value of the test which is 0.1683 is significantly higher than the significance level 0.05, so there is not enough evidence to reject the null hypothesis. That is, variances of groups are all equal to each other. Thus, there is homogeneity between groups' variances.

ANOVA	Degrees of Freedom	Sum of Squares	Mean Squares	F value	Pr(>F)
Study Time	3	108	36.07	1.728	0.161
Residuals	391	8162	20.87		

Table 6: *Analysis of Variances Table*

According to ANOVA Table, the p-value is 0.161 and it is significantly greater than the significance level 0.05. Thus, there is not enough evidence to reject the null hypothesis that the mean values of final grades in terms of study time are all equal to each other.

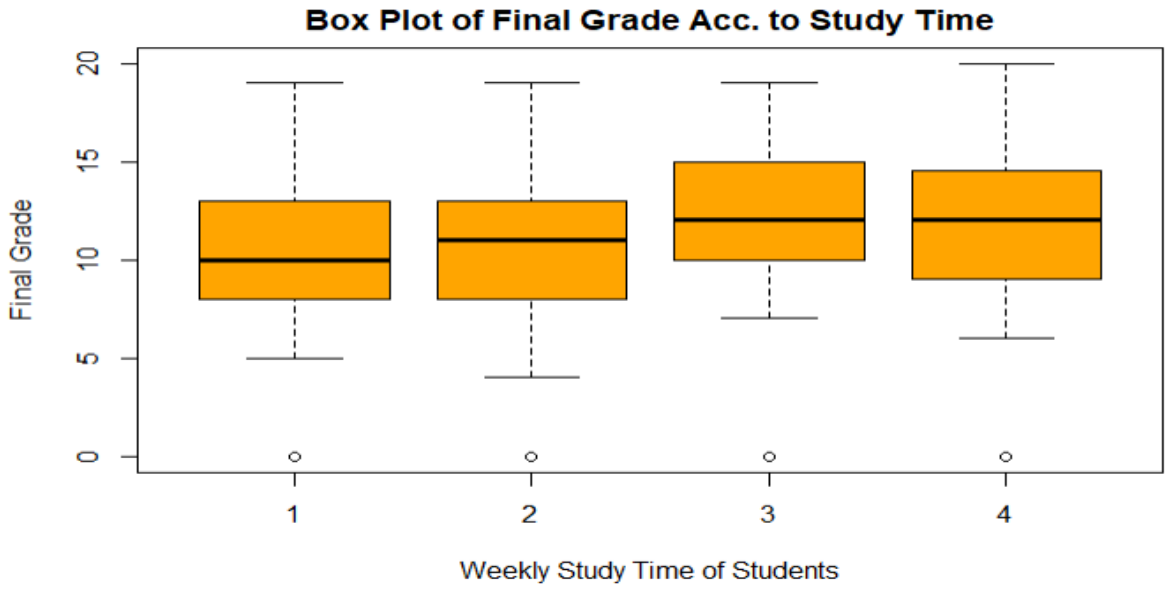


Figure 4: Box Plot of Final Grade According to Study Time

In order to illustrate the relationship between final grades of students and weekly study time of them, the Box Plot was used. In the Figure 4, the horizontal axis represents weekly study times. "1" represents less than 2 hours, "2" represents 2 to 5 hours, "3" represents 5 to 10 hours and "4" represents more than 10 hours in terms of weekly study times. On the other hand, the vertical axis represents final grades of students and the range is between 0 and 20.

The first and the third groups' box plots have a right skewed shaped. The second group's box plot has a left skewed shape, and the fourth group's box plot has almost a symmetric shape. Also, some outliers were detected in the box plot.

To conclude, it can be said that there is no difference in final grades of students in terms of study time. This is proven by ANOVA Table and Box Plot.

3.4 The Relationship between Absences and Romantic Relationship

In order to investigate the relationship between absences of students and being in a romantic relationship, some statistical tests and graphical tool were used.

H_o: $\mu_{yes}=\mu_{no}$

H_1: $\mu_{yes}\neq\mu_{no}$

The null hypothesis for this research question is the mean values of students who have a romantic relationship and who do not have in terms of their absences are equal to each other. In order to test this hypothesis two sample t-test was used. As a first step, 30 observations were sampled randomly from the data set. Secondly, F-test was performed in order to compare variances of two groups which are having a romantic relationship and not having it. The F-test's p-value which is 0.07923 is clearly higher than the significance level taken 0.05, so there is not enough evidence to reject the null hypothesis. That is, variances of those groups are equal. Thirdly, the Shapiro-Wilk Normality Test was performed. Since the p-values of each group are less than the significance level, there is enough evidence to reject the null hypothesis that sample follow the normal distribution. This can be due to having a small sample size. If the normality is assumed, two sample t-test can be conducted.

T	Degrees of Freedom	p-value	Mean of x	Mean of y
1.2112	28	0.2359	6.13	3.86

Table 7: *Two-Sample T-Test Results*

According to Table 7, since the p-value which is 0.2359 is significantly higher than the significance level 0.05, there is not enough evidence to reject the null hypothesis. That is, we are 95 % confident that there is not a significant difference between average value of absences of being in a romantic relationship and average value of absences of not being in a romantic relationship.

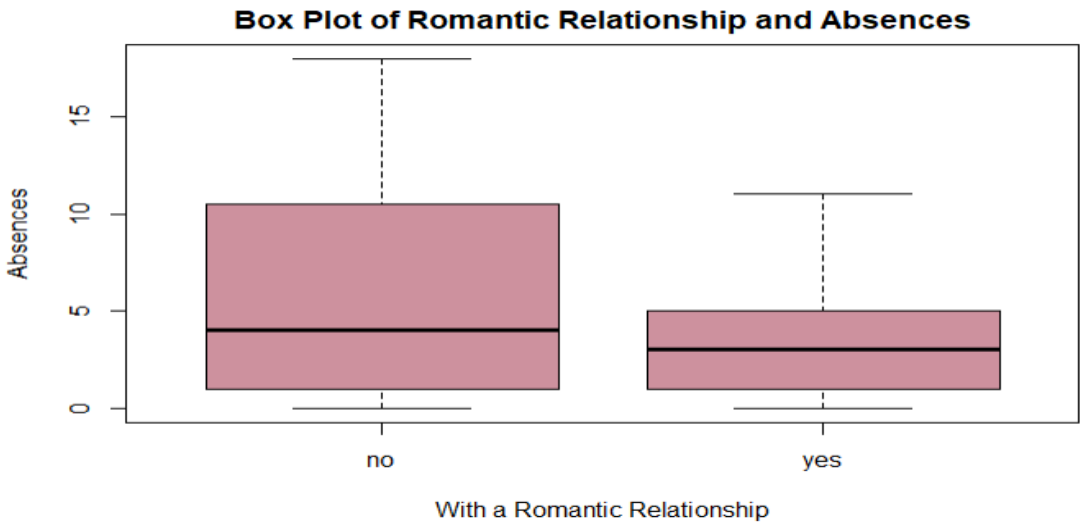


Figure 5: Box Plot of Romantic Relationship and Absences

In Figure 5, the horizontal axis represents the variable which is with a romantic relationship. “No” stands for students who do not have a romantic relationship while “Yes” stands for students who have a romantic relationship. In the vertical axis, absences of students were shown.

The box plot of not having a romantic relationship follows right skewed distribution whereas the box plot of having a romantic relationship follows almost symmetric distribution. Moreover, there is not any outlier in this particular sample. Furthermore, the median values of absences seem to very close which means there is not significant difference between the mean values of absences in terms of being in a romantic relationship or not.

3.5 The Relationship between Final Grades and Absences

In order to investigate the relationship between final grades of students and absences, statistic and graphical tool were used.

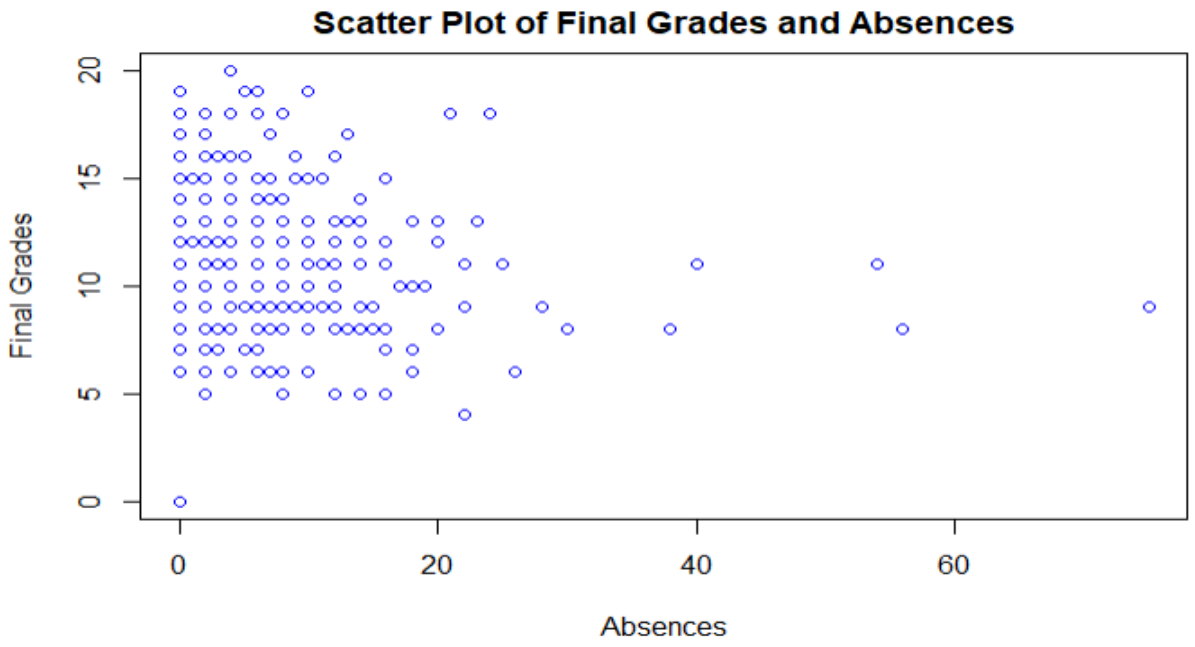


Figure 6: Scatter Plot of Final Grades and Absences

According to Figure 6, there is not significant linear relationship between the final grades and absences of students. However, there might be a weak and positive relationship between them. In order to be sure about that, the correlation between them should be investigated.

Correlation Matrix	Absences	Final Grade
Absences	1.00000	0.03424
Final Grades	0.03424	1.00000

Table 8: *Correlation Matrix between Absences and Final Grades*

According to Table 8, there is a very weak positive relationship between absences and final grades of students. Also, by looking at the covariance value between them, since the covariance was found as 1.256, it can be said that there is a positive relationship between them.

3.6 The Relationship between Final Grades and All Other Variables

In order to investigate this relationship, linear regression model was used.

As a first step, linear regression modelling was performed on the whole data set with response variable as final grades and all other independent variables. The model was statistically significant since its p-value is less than the significance level 0.05. Also, the Adjusted R-squared

value was found as 0.83 which is good for analysis. However, there were lots of insignificant coefficients and the residuals of the model did not follow the normal distribution. Consequently, the Box-Cox transformation was done on the response variable, but problems were not solved.

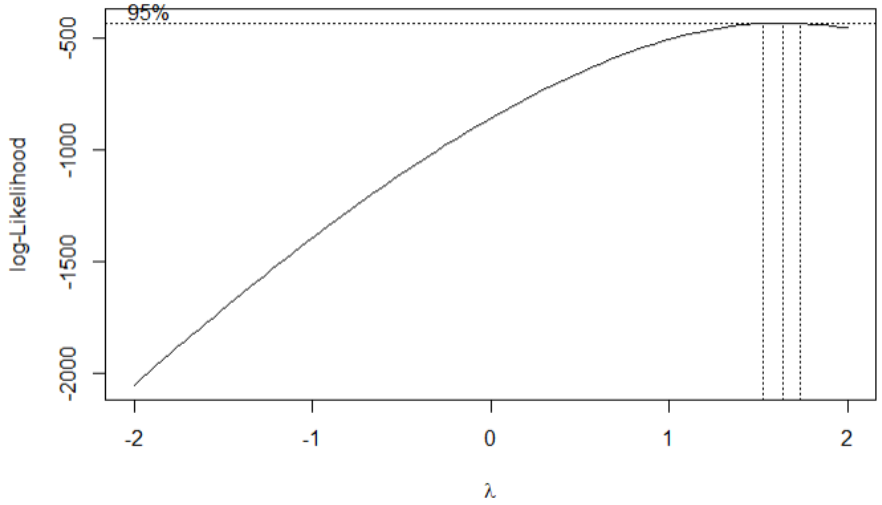


Figure 7: Box Cox Lambda of Final Grades Variable

As it can be seen from Figure 7, the Box-Cox lambda is between 1 and 2. Thus, the Box-Cox transformation was performed.

After the transformation, in order to decide which variables should be eliminated from the model, the Stepwise Elimination was done. Both Backward and Forward elimination technique was used. The last model according to Stepwise elimination is;

```
Call:
lm(formula = ((G3^lambda - 1)/lambda) ~ age + Fedu + guardian +
    famsup + paid + activities + romantic + famrel + G1 + G2,
    data = student)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5689  -2.8952   0.5631   3.5847  14.9997

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.3454     6.5207  -1.433  0.15263
age             -0.7293     0.2672  -2.729  0.00664 **
Fedu1           -0.1590     4.2796  -0.037  0.97037
Fedu2           -2.6245     4.2535  -0.617  0.53759
Fedu3           -0.5813     4.2577  -0.137  0.89147
Fedu4           -1.4218     4.2595  -0.334  0.73873
guardianmother   1.4967     0.7329   2.042  0.04184 *
guardianother    0.3385     1.3329   0.254  0.79965
famsupyes        0.9731     0.6624   1.469  0.14266
paidyes         -1.0496     0.6429  -1.633  0.10339
activitiesyes   -1.3333     0.6058  -2.201  0.02834 *
romanticyes     -0.9425     0.6558  -1.437  0.15147
famrel2         -2.6523     2.5406  -1.044  0.29717
famrel3          0.5995     2.2363   0.268  0.78877
famrel4          0.6437     2.1552   0.299  0.76537
famrel5          2.7991     2.1996   1.273  0.20397
G1               1.6090     0.1774   9.071  < 2e-16 ***
G2               3.2479     0.1577  20.595  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.904 on 377 degrees of freedom
Multiple R-squared:  0.8969,    Adjusted R-squared:  0.8923
F-statistic:   193 on 17 and 377 DF,  p-value: < 2.2e-16
```

Although the model seems significant and adjusted R-squared value is very high, there are still some coefficients that are not significant. Thus, they should be eliminated from the model, too. Thus, the last model should be;

```
Call:
lm(formula = ((G3^lambda - 1)/lambda) ~ age + guardian + activities +
    G1 + G2, data = student)

Residuals:
     Min       1Q   Median       3Q      Max
-25.2715  -2.8184   0.1386   3.6670  14.9312

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.4212     4.6369  -2.032  0.04286 *
age             -0.7157     0.2663  -2.687  0.00752 **
guardianmother   1.2855     0.7410   1.735  0.08356 .
guardianother    0.1045     1.3551   0.077  0.93857
activitiesyes   -1.1048     0.6160  -1.794  0.07365 .
G1               1.5824     0.1776   8.908  < 2e-16 ***
G2               3.2244     0.1581  20.394  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.079 on 388 degrees of freedom
Multiple R-squared:  0.8875,    Adjusted R-squared:  0.8858
F-statistic: 510.4 on 6 and 388 DF,  p-value: < 2.2e-16
```

The model is:

Final grades = -9.4212 + (-0.7157)*age + (1.2855)*guardianmother + (0.1045)*guardianother + (-1.1048)*activitiesyes + (1.5824)*G1 + (3.2244)* G2

Interpretations of the last model;

- Since the p-value of the model is significantly less than the significance level, there is enough evidence to state that the model is significant.
- 89 % of variability in final grades of students can be explained by age, guardian, extra-curricular activities, first period grade and second period grade.
- The highest effect on final grades belongs to second period grade because its estimate is the highest while the least effect belongs to guardian another.
- If the all variables equal to zero which means they do not have any effect on final grades, the expected value for final grades is -9.4212.
- One unit increment in age results in 0.72 percentage decline in final grades when all other variables remain constant.
- One unit increment in guardian mother results in 1.29 percentage increment in final grades when all other variables remain constant.
- One unit increment in guardian another results in 0.11 percentage increment in final grades when all other variables remain constant.
- One unit increment in activities yes results in 1.11 percentage decline in final grades when all other variables remain constant.
- One unit increment in first period grade results in 1.58 percentage increment in final grades when all other variables remain constant.
- One unit increment in second period grade results in 3.22 percentage increment in final grades when all other variables remain constant.

All in all, final grades of students in Mathematics lessons are affected by age, guardian, extra-curricular activities, first period grade and second period grade. The second period grade has the highest effect while guardian another has the least effect on the final grades of the students.

4. CONCLUSION

There are lots of variable that can have an effect on students' performance in Mathematics lessons as shown in this study. In this project, the main purpose was to investigate the students' performance in Mathematics lessons. For that purpose, the data set conducted on secondary school students in the Alentejo region of Portugal was used. In order to answer the research questions, ANOVA Table, two-sided t-test, correlation, covariance and linear regression model were used. Also, Bar Graph, Box Plot, Pie Chart, Scatter Plot and Histogram were used to have better understanding from the data set.

According to the findings of this study, it can be stated that final grades in Mathematics of students do not change according to their study time. Besides, absences of students do not differ for being in a romantic relationship or not. Also, final grades in Mathematics of students and their absences are related, but their relation is very weak and positive. Furthermore, the final grades of students in Mathematics lessons are affected by age, guardian, extra-curricular activities, first period grade and second period grade. Actually, age and extra-curricular activities have negative effect on the final grades.

All in all, this study shows that students' performance in Mathematics is affected by their age, guardian, extra-curricular activities, first period grade and second period grade. Thus, the students should consider those factors while studying. For instance, if they reduce their extra-curricular activities, they might have better performance. This is because, the extra-curricular activities has a negative effect on the final grades in Mathematics lessons.

5. REFERENCES

(n.d.). Retrieved January 14, 2021, from https://archive.ics.uci.edu/ml/machine-learning-databases/00320/

Mcmillan, J. H., & Reed, D. F. (1994). At-Risk Students and Resiliency: Factors Contributing to Academic Success. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 67*(3), 137-140. Doi:10.1080/00098655.1994.9956043

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

6. APPENDIX

```{r}
student<-read.csv("student-mat.csv",sep=";",header=T)
head(student)
student$Medu<-as.factor(student$Medu)
student$Fedu<-as.factor(student$Fedu)
student$traveltime<-as.factor(student$traveltime)
student$studytime<-as.factor(student$studytime)
student$famrel<-as.factor(student$famrel)
student$freetime<-as.factor(student$freetime)
student$goout<-as.factor(student$goout)
student$Dalc<-as.factor(student$Dalc)
student$Walc<-as.factor(student$Walc)
student$health<-as.factor(student$health)
```

#Descriptive Statistics of variables

```{r}
summary(student)
```

```{r}
summary(student$age)
summary(student$absences)
summary(student$G1)
summary(student$G2)
summary(student$G3)
```

```{r}
```

```
tab1(student$sex, sort.group = "decreasing", cum.percent = TRUE)
```

```{r}
tab1(student$famsize, sort.group = "decreasing", cum.percent = TRUE)
```

```{r}
tab1(student$Pstatus, sort.group = "decreasing", cum.percent = TRUE)
```

```{r}
tab1(student$guardian, sort.group = "decreasing", cum.percent = TRUE)
```

##ANALYSIS

#Backround analysis

```{r}
hist(student$age,breaks =9,col="blue",main="Histogram of Ages of Students",xlab="Age")
```

```{r}
library(plotrix)
table(student$sex)
tbl <- c(208,187)
pielbl <- c("female","male")
pie3D(tbl,labels=pielbl,explode=0.1, main="Pie Chart of Gender of Students")
```

```{r}
freq<-table(student$reason)
barplot(sort(freq,decreasing = T),main="Bar Graph of The Reasons to Choose This School",col=c("red","blue","green","yellow"))
```

#Research Question 1

Do Final Grades of Students change according to their Study Time?

```{r}
boxplot(G3~as.factor(studytime),data=student,main="Box Plot of Final Grade Acc. to Study Time",xlab="Weekly Study Time of Students",ylab="Final Grade",col="orange")
```

```{r}
shapiro.test(student$G3[student$studytime=="1"])
shapiro.test(student$G3[student$studytime=="2"])
shapiro.test(student$G3[student$studytime=="3"])
shapiro.test(student$G3[student$studytime=="4"])
```

```{r}
bartlett.test(G3~as.factor(studytime),data=student)
```

```{r}
anova<-aov(G3~as.factor(studytime),data=student) #fitting ANOVA model
summary(anova)
```

#Research Question 2

Do absences of students differ for being in a romantic relationship or not?

```{r}
set.seed(147)
abse<-sample(student$absences,size=30)
rom<-sample(student$romantic,size=30)
df<-data.frame(abse,rom)
```

```{r}
```

boxplot(abse~rom,data=df,main="Box Plot of Romantic Relationship and Absences",xlab="With a Romantic Relationship",ylab="Absences",col="pink3")

```

```{r}

abs<-split(df$abse,df$rom)

```

```{r}

var.test(abs$no,abs$yes)

```

```{r}

shapiro.test(abs$no)

shapiro.test(abs$yes)

```

```{r}

t.test(abs$no,abs$yes,alternative = c("two.sided"),var.equal = TRUE,conf.level = 0.95)

```

#Research Question 3

Are final grades and absences related?

```{r}

plot(student$absences,student$G3,main = "Scatter Plot of Final Grades and Absences",xlab="Absences",ylab="Final Grades",col="blue")

```

```{r}

corr<-cor(student$absences,student$G3)

corr

student1<-data.frame(student$absences,student$G3)

corrplot<-cor(student1)

corrplot

cov(student$absences,student$G3)

```

#Research question 4

Which variable has more effect on final grades of students?

```{r}
genrlmdl<-lm((G3+1)~.,data=student)
summary(genrlmdl)
```

```{r}
shapiro.test(residuals(genrlmdl))
```

```{r}
library(MASS)
boxcox(genrlmdl)
a<-boxcox(genrlmdl)
lambda<-a$x[which.max(a$y)]
transformed<-lm(((G3^lambda-1)/lambda) ~ .,data=student)
summary(transformed)
```

```{r}
stepAIC(transformed,direction = "both")
```

```{r}
lastmodel<-lm(formula = ((G3^lambda - 1)/lambda) ~ age + Fedu + guardian +
    famsup + paid + activities + romantic + famrel + G1 + G2,
    data = student)
summary(lastmodel)
```
```

```
```{r}
shapiro.test(residuals(lastmodel))
```
```{r}
mdl<-lm(formula = ((G3^lambda - 1)/lambda) ~ age  + guardian +
     activities  + G1 + G2,
    data = student)
summary(mdl)
shapiro.test(residuals(mdl))
```
```