

Analysis of Portuguese Secondary School

# Executive Summary

**Background**This report provides analysis of the performance of students from two Portuguese secondary schools. The report investigates the impact of influential factors on their grades including such things as: school absences, parent’s education and relationship status, student health, and extracurricular activities to name a few.

**Methods of Analysis**

Descriptive and Inferential statistical analysis has been conducted on a wide range of variables from the data set provided. Descriptive statistic methods used include mean, median, mode, quartiles and standard deviation. Box plots, violin plots and correlation tables have also been included to provide illustrations of important data features.

Inferential statistical analysis is useful for testing hypothesis relating to the effects of one or more variables on another. Inferential statistics methods used include t-tests, ANOVAs, Tukey-tests, linear regressions, Bartletts’s tests and Levene’s tests.

**Areas of Analysis**

Analysis in this report sets out to answer the following questions:

1. *Does the level of education of a student’s parents have any effect on a student’s overall grades?*
2. *Are the grades from G1, G2 and G3 effected by which school the students attend?*
3. *Does the level of study undertaken by a student reflect the number of days taken absent?*
4. *Can the students’ failures be explained by number of absences?*

**Findings**

We find that there is an observable difference in a student’s grade depending on whether the Mother had no education or higher education. We also found this to be true as to whether the student’s Father either had no education or higher education. In both cases, students who had either a mother or father educated to the level of higher education had, on average, better overall grades than student’s whose mother or father had no education.

Also, we’ve found that while the relation between the number of absences and number of failures are statistically significant, it’s complicated to explain failures only by this factor. Thereby, the schools may use these findings to build a model to predict and control students’ failures.

**Conclusions**

As far as the effect of a parent’s level of education on a student’s overall grade it could be concluded that students who have a mother and father

**Limitations**

Whilst this report provides data driven insights into the level of influence outside factors have on students grades no recommendations have been made around development of strategies that might be implemented to sustain or improve students grades in the future.

Contents

[Executive Summary 1](#_Toc528523567)

[Introduction 1](#_Toc528523568)

[Dataset Description 1](#_Toc528523569)

[Descriptive 2](#_Toc528523570)

[Analysis and Results 7](#_Toc528523571)

[Discussion and Conclusion 13](#_Toc528523572)

[Appendix (R-Code) 15](#_Toc528523573)

[References 30](#_Toc528523574)

# Introduction

Brown and Saunders (2008) describe statistics at the many procedures for “gathering, organising, analysing and presenting quantitative data”. Statistics can be used draw conclusions from numerical data by conducting experiments and observations (Dodge, 2006).

There are two main methods used in conducting statistical analysis: Descriptive statistics and inferential statistics (Brown & Saunders, 2008). It is important to first describe that data we are dealing with when conducting analysis. This can be done by identifying mean, median, mode and standard deviation of variables of interest. Descriptive statistics provides useful information about a data source without predicting probability or inferring conclusions about students outside the data collected (Trochim, 2006).

Unlike descriptive statistics, inferential statistics assumes that the data being analysed is a sample obtained from a larger population and can be used to derive the probability of outcomes to a broader population (Kalla & Wilson, 2018). Inferential statistics allows us to draw conclusions based on testing hypotheses to within pre-defined levels of confidence that the results can be generally applied to the larger population (Frost, 2018).

In this study we have obtained the data set for students attending two Portuguese secondary schools (Cortez, 2014). Achievement levels in Portuguese schools are noticeable behind the majority of schools in other European nations (Cortez & Silva, 2008)The primary reason for this analysis is to examine possible links between student grades in Portugal and external influences including parental education levels and relationship status and extracurricular activities.

# Dataset Description

The data set is collected from 2 government schools located in Portugal Alentejo region through sources such as school records and questionaries. The primary reason to build this dataset is to indicate the factors that influence student’s achievement and performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Binary (Yes/No) Type |  | Numeric Variable |  | Nominal Variable (Choose from the list) |
| 1  2.  3.  4.  5.  6  7.  8.  9  10  11.  12  13 | Sex(M/F)  School (GP/MS)  Address (U/R)  Pstatus(A/T)  Schoolsup(yes/no)  Famsup(yes/no)  Activities(yes/no)  Paid(yes/no)  Internet(yes/no)  Nursery(yes/no)  Higher(yes/no)  Romantic(yes/no)  Famsize (LE3/GT3) | 14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29 | Age (15 to 22)  Medu (0 to 4)  Fedu (0 to 4)  Traveltime (1 to 4)  Studytime (1 to 4)  Failures (0 to 4)  Freetime (1 to 5)  gout (1 to 5)  Walc (1 to 5)  Dalc (1 to 5)  Health (1 to 5)  Absences (0 to 93)  G1(0 to 20)  G2(0 to 20)  G3(0 to 20)  Famrel (1 to 5) | 30  31  32  33 | Reason(course/home/reputation/other)  Mjob (0 to 4)  Fjob (0 to 4)  Guardian (mother, father, other) |

Table 1: Data set Student with variable types.

This raw data is applicable in studying through data modelling techniques by which predication can be made to optimize the grades and improve decision making by reacting to the trends and patterns of the data distribution This dataset is based from data, collected from school records (paper based) include 3 grades and the absence records. Also, data collected from questionaries that contains closed questions with 2 predefined options to choose from. All these data were set to build a dataset comprising 649 students with 33 variables. The data set can be categorized into 4 categories or classification. Which are as student grades, demographic, social and school related. The data in this data set is modelled as binary, nominal and numeric. The data set has no missing values.

# Descriptive

**G1, G2 and G3**

When skewness () function of the ‘moments’ package is used to analyse the distribution of variables G1, G2 and G3 respectively the results indicate that all the variables are negatively skewed.

* The variable G1 is very weakly and negatively skewed (skewness(student$G1) = -0.002767222)
* The variable G2 is neutrally and negatively skewed (skewness(student$G2) = -0.3594494)
* The variable G3 is very strongly and negatively skewed (skewness(student$G3) = -0.91079)

Box plots are considered for each variable G1, G2 and G3 respectively which indicate that all these variables are normally distributed and are continuous without any missing values (NA).

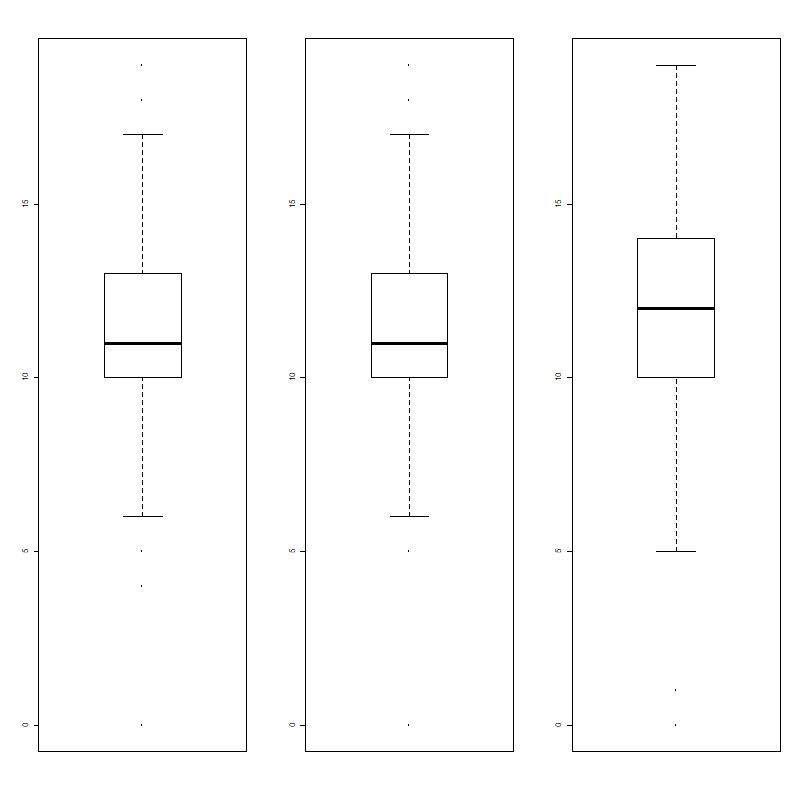


Figure 1. Box plots for g1, g2 and g3

Figure 1 represents the box plots of variables G1, G2 and G3 respectively. From that it is notable that the variances of G1 and G2 are almost similar when compared to a slight change in the variance of G3. Medians of the variables can be compared from (Figure 1) showing that the medians of G1 and G2 are approximately equal when compared to the median of G3 which is greater than both G1 and G2. It can be clearly noted that there exist some outliers in all G1, G2 and G3.

The ddply () function is used to create a summary table indicating the mean, standard deviation and median of G1, G2 and G3 for each school respectively.

Table 1. Grades among the schools

|  | **school** | **N** | **G1\_avg** | **G1\_sd** | **med\_G1** | **G2\_avg** | **G2\_sd** | **med\_G2** | **G3\_avg** | **G3\_sd** | **med\_G3** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | GP | 423 | 11.99 | 2.42 | 12 | 11.99 | 2.42 | 12 | 11.99 | 2.42 | 13 |
| 2 | MS | 226 | 10.3 | 2.98 | 10 | 10.3 | 2.98 | 10 | 10.3 | 2.98 | 11 |

Table 1 clearly indicates that there are 423 observations for which variable school is set as ‘GP’, which means that the 423 observations are collected from ‘Gabriel Pereira’ school and the remaining 226 observations from the school namely ‘Mousinho da Silveira’. (Table 1) indicates the mean, median and standard deviation of the variables for each of the two schools separately.

When a correlation plot is drawn for all the numeric variables in the student dataset, it indicates very strong correlation between the variables G1, G2 and G3.

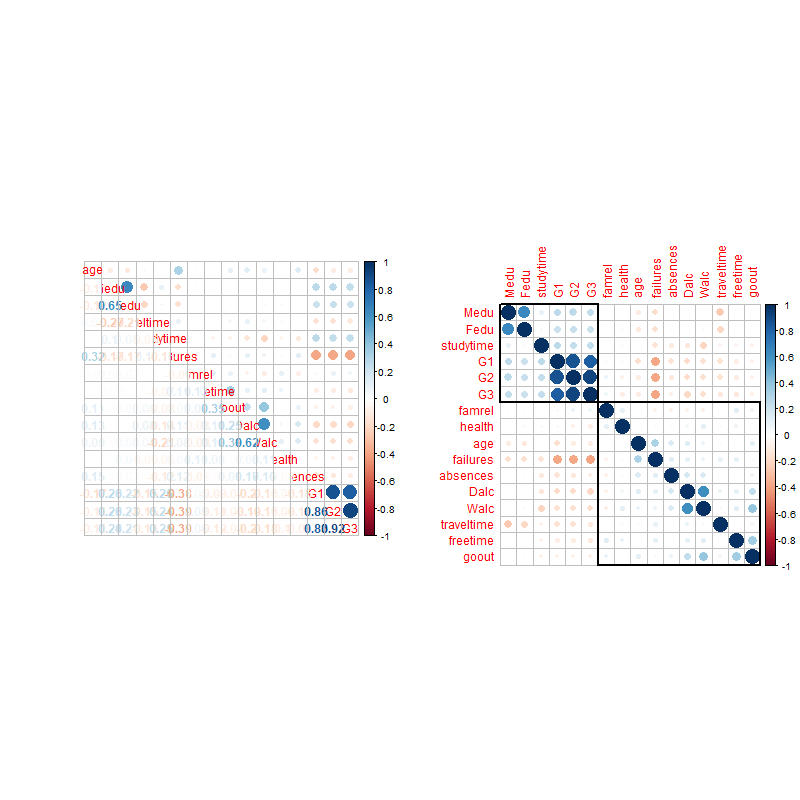


Figure 2. Correlation matrix

Figure 2 clearly indicates that variables G1, G2 and G3 possess strong correlations among them and thus, these variables are taken into consideration for further investigation. Another correlation plot with p-values, combined with significance test is drawn to investigate these grades variables

**Overallgrage**

A new variable, overallgrade, has been created on the assumption that we can average G1, G2 and G3 to a single score. This is to simplify the research as we aim to understand the influence of Parent’s education at a more macro level. As with G1, G2 and G3 the grades ranging from 0 to 19 for students in Period 1. The standard deviation is 2.840744



Figure 3: overallgrade summary statistics

**Medu**

Medu variable describes the level of education obtained by the mother of a student with 0:4 representing education levels of “none”, “primary education (4th grade)”, “5th to 9th grade”, “secondary education”, “higher education” (Figure 4). The standard deviation is 1.134552



Figure 4: Mother Education summary statistics

**Fedu**

Fedu variable describes the level of education obtained by the mother of a student with 0:4 representing education levels of “none”, “primary education (4th grade)”, “5th to 9th grade”, “secondary education”, “higher education” (Figure 5). The standard deviation is 1.099931



Figure 5: Father Education summary statistics

*Box Plots*

Box plots allow us to visualise the distribution of data (Figure 6). We can also see that there are outlying data that can influence the skewness, standard deviation results that might not be visible otherwise.

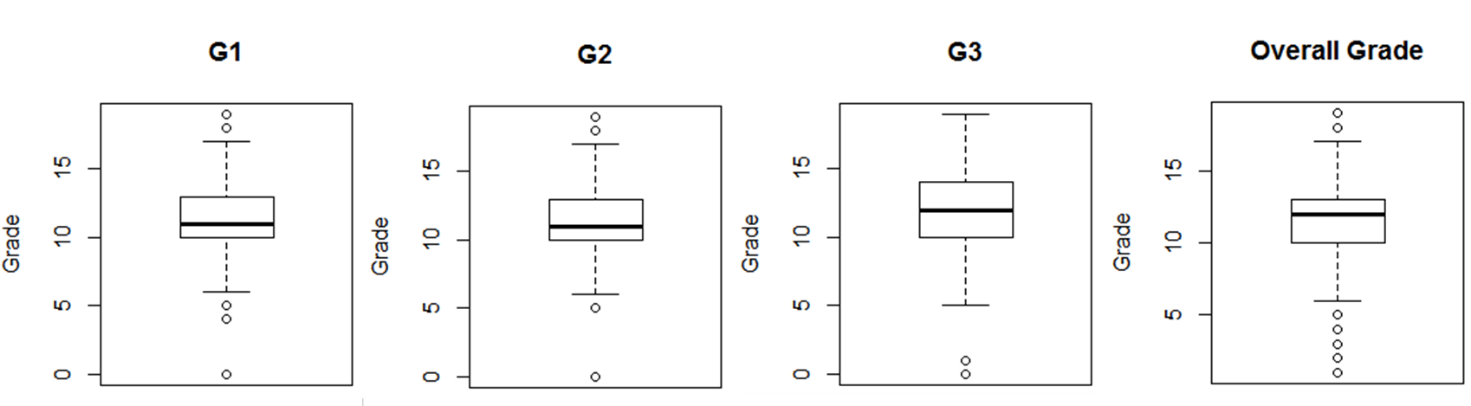


Figure 6: Box plots for G1, G2, G3 and overallgrade

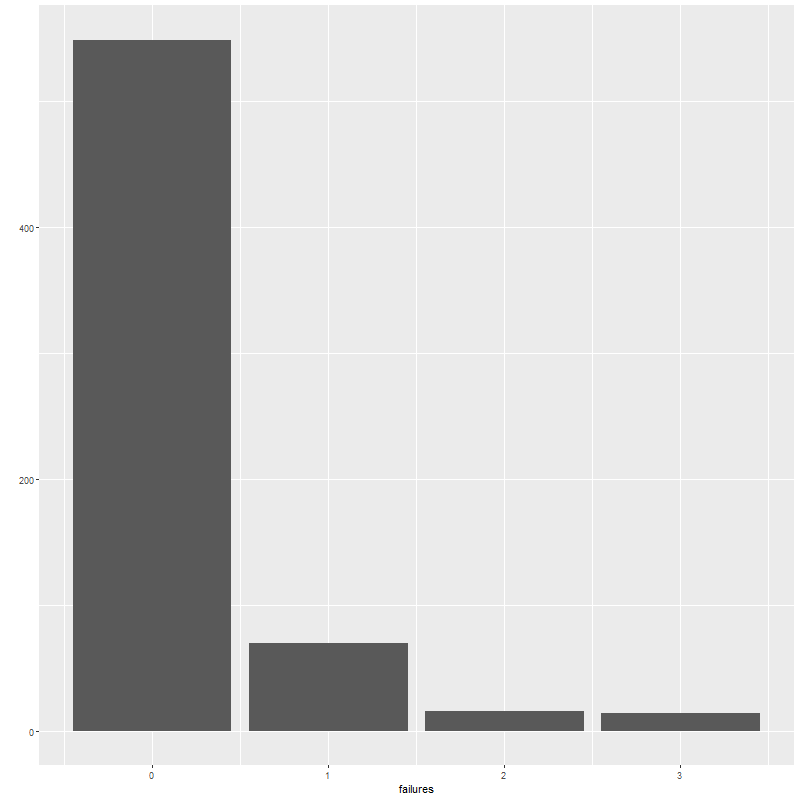
**Failures**

Failures variable (nominal) represents the number past classes failures in a numeric format (n if 1 <=n<3, else 4). The standard deviation is 0.5932351, other descriptive statistics of the variable is presented in Figure 7:



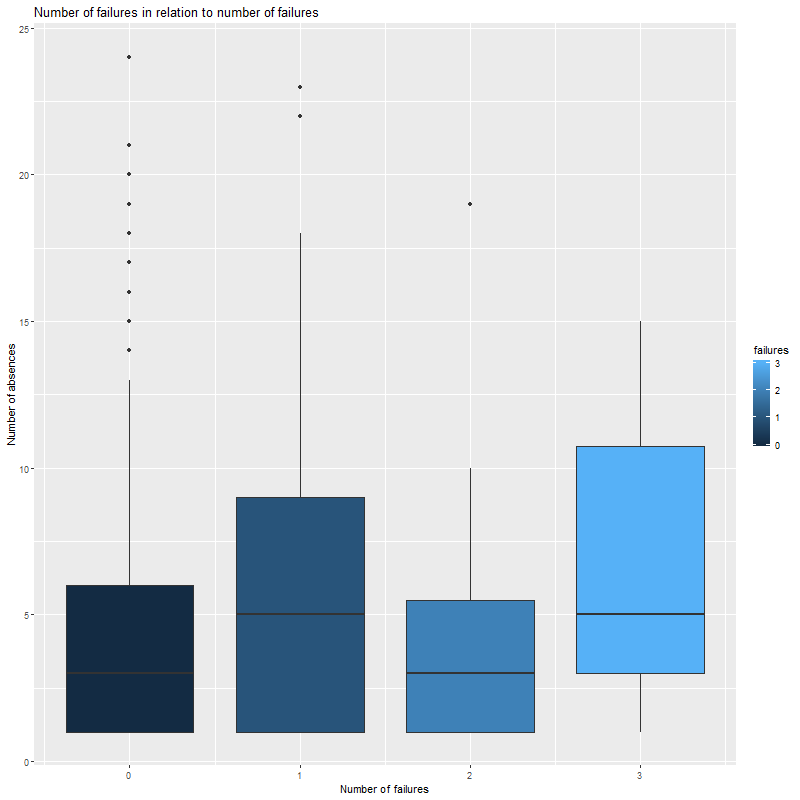
*Figure 7. Descriptive statistics for "Failures" variable*

In order to understand the data distribution, the following boxplot has been created (Figure 8):



*Figure 8. Box plot for Failures distribution*

One of the analysis questions is to analyse the relationship between the number of failures and number of absences. The following boxplot has been created to visualize this possible relation (Figure 9):



*Figure 9. Boxplot of relation between the number of absences and number of failures*

**Address**

Address variable (nominal) represents the students’ home address type: “U” for Urban type and “R” for Rural. “U value” has 452 observations, “R value” – 197.

**Health**

Health variable (ordinal) represents the current health status of a student. Scaled from 1 (very bad) to 5(very good). The standard deviation is equal to 1.446259, other descriptive statistics of the variable is presented in Figure 10:



*Figure 10. Descriptive statistics for "Health" variable*

**Absences**

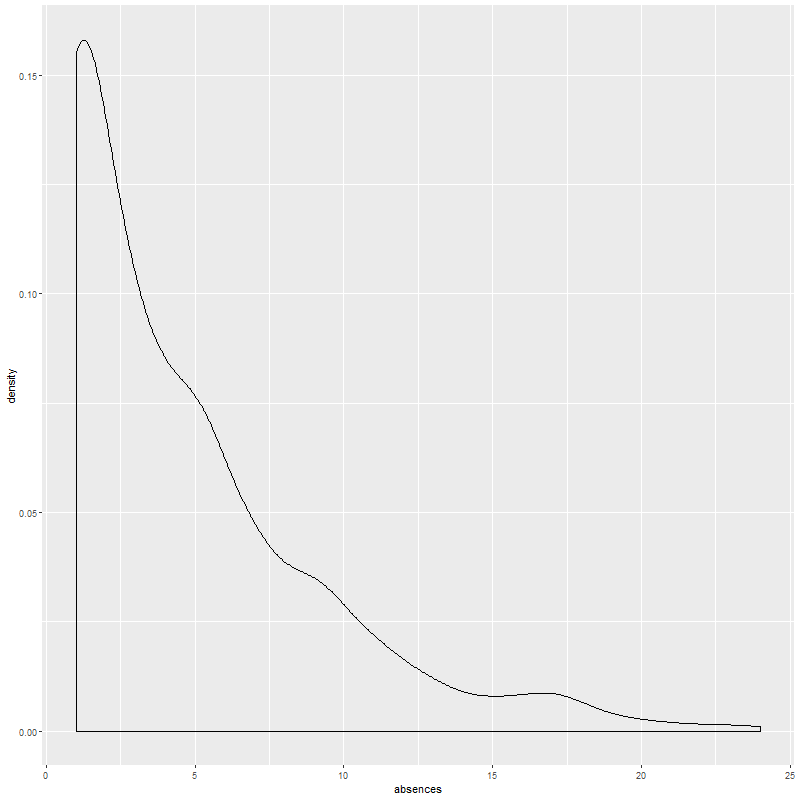
Absences variable contain a number of missing classes by a student (from 0 to 93). The standard deviation of the variable 4.364954 and other descriptive information presented below (Figure 11):



*Figure 11.Descriptive statistics of Absence variable*

According to this, the minimum number of absences is 1, while the maximum one is 24. Students miss 4.596 classes in average.

To understand the data distribution, the density plot has been created (Figure 12).



*Figure 12. Distribution of the absence variable*

To it more useful for the research purposes, the additional variable called “AbsencesCat” has been derived from the original one. Based on the summary statistics and variable distribution, “AbsencesCat” has 5 categories: Very Low (0-5 absences), Low (6-10), Medium (11-15), High (16-20), Very High (21-25).

**Studytime**

Studytime variable represents the time that students spend on study (weekly). Scaled from 1 (<2 hours) to 4 (>10 hours). The following descriptive statistics represents its values among the 2 schools(Figure 13):



Figure 13. Summary statistics

From the summary table we can conclude that Gabriel Pereira school students spent more time on average in studying than Mousinho da Silveira for the year 2005-2006.From standard deviation the study time is distributed closely towards the mean value.

# Analysis and Results

*Question: Are overall grades of students influenced by the level of education of their parents?*

To answer this question, the following hypothesis has been developed:

Hₒ : The statistical relationship between the Parent’s level of education and student’s grades is not significant

Hₐ : The statistical relationship between the Parent’s level of education and student’s grades is significant

First we break down the grades by level of Mother’s and Father’s levels of education so we can examine key statistics including skewness, mean, standard deviation, variability, median and quartile ranges.

All of the skewness scores except for FatherEdu None are close to zero (Figure 13,Figure 14). Therefore, we could be confident in using these data in prediction models. FatherEdu None has a is positively, or left, skewed by 1.47. That means that these data may be less reliable in predicting the grades of students based on Father’s level of education when it is none.

We can see from the data that the variability for student’s grades for all levels of parent’s education a fairly and relatively similarly variable (Figure 14,Figure 15).

|  | **MotherEdu** | **N** | **Overall**  **grades**  **\_skewness** | **Overall**  **grades**  **\_avg** | **Overall**  **grades**  **\_sd** | **Overall**  **grades**  **\_var** | **Overall**  **grade**  **s\_med** | **First**  **\_quartile** | **Third**  **\_Quartile** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | none | 6 | 0.5 | 11.17 | 1.72 | 11.17 | 11 | 10.25 | 11.75 |
| 2 | primary edu (4th grade) | 143 | -0.14 | 10.58 | 2.71 | 10.58 | 10 | 9 | 12 |
| 3 | 5th to 9th grade | 186 | -0.29 | 11.39 | 2.59 | 11.39 | 11 | 10 | 13 |
| 4 | secondary ed | 139 | -0.05 | 11.58 | 2.85 | 11.58 | 11 | 10 | 13 |
| 5 | higher ed | 175 | -0.62 | 12.81 | 2.84 | 12.81 | 13 | 11 | 15 |

Figure 14: overallgrades by Mother education summary statistics

|  | **FatherEdu** | **N** | **Overall**  **grades**  **\_skewness** | **Overall**  **grades**  **\_avg** | **Overall**  **grades**  **\_sd** | **Overall**  **grades**  **\_var** | **Overall**  **grades**  **\_med** | **First**  **\_quartile** | **Third**  **\_Quartile** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | none | 7 | 1.47 | 11.57 | 2.57 | 11.57 | 11 | 10.5 | 11.5 |
| 2 | primary edu (4th grade) | 174 | -0.08 | 10.74 | 2.82 | 10.74 | 10.5 | 9 | 12 |
| 3 | 5th to 9th grade | 209 | -0.31 | 11.53 | 3.03 | 11.53 | 11 | 10 | 13 |
| 4 | secondary ed | 131 | 0.24 | 12.02 | 2.24 | 12.02 | 12 | 10 | 13 |
| 5 | higher ed | 128 | -0.36 | 12.62 | 2.76 | 12.62 | 13 | 11 | 15 |

Figure 15: overallgrades by Father education summary statistics

When we observe the Box plots we can see that in almost all cases there are a number of outlying data points that have a significant effect on the distribution of the data (Figure 16). When we look at the Violin plots we can see that the distributions are fairly normal except for the Fathers Education-none and the Mother’s Education-secondary (Figure 17). In both sets of plots, we can observe an increased median and average score for students whose mother or father have higher education compared with students whose mother or father have no education. We can also see that the distribution for parents with no education is much tighter. This is likely due to the much smaller number of observations for mothers and fathers with no education compared to the other categories. Noticeably, in all categories there are a number of significant outliers doing much better and much worse than the majority of students whose parents have some level of education.

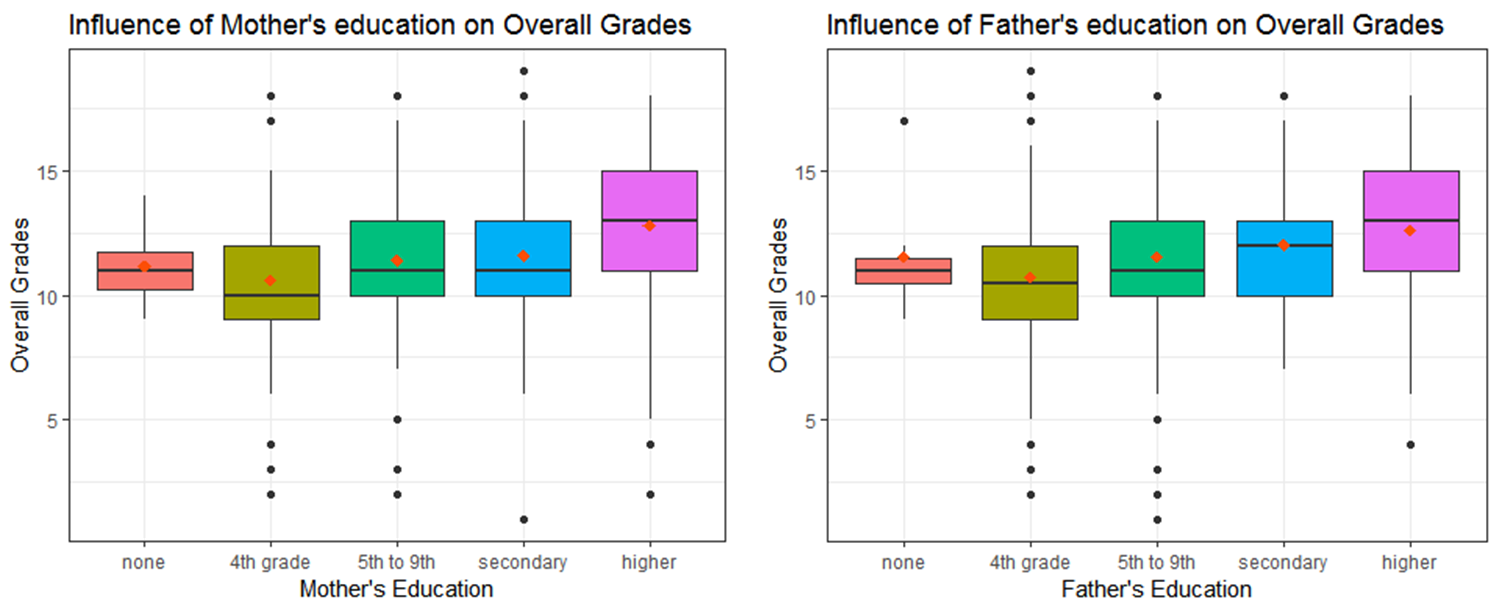


Figure 16: Box plots of overallgrades by Mother and Father education

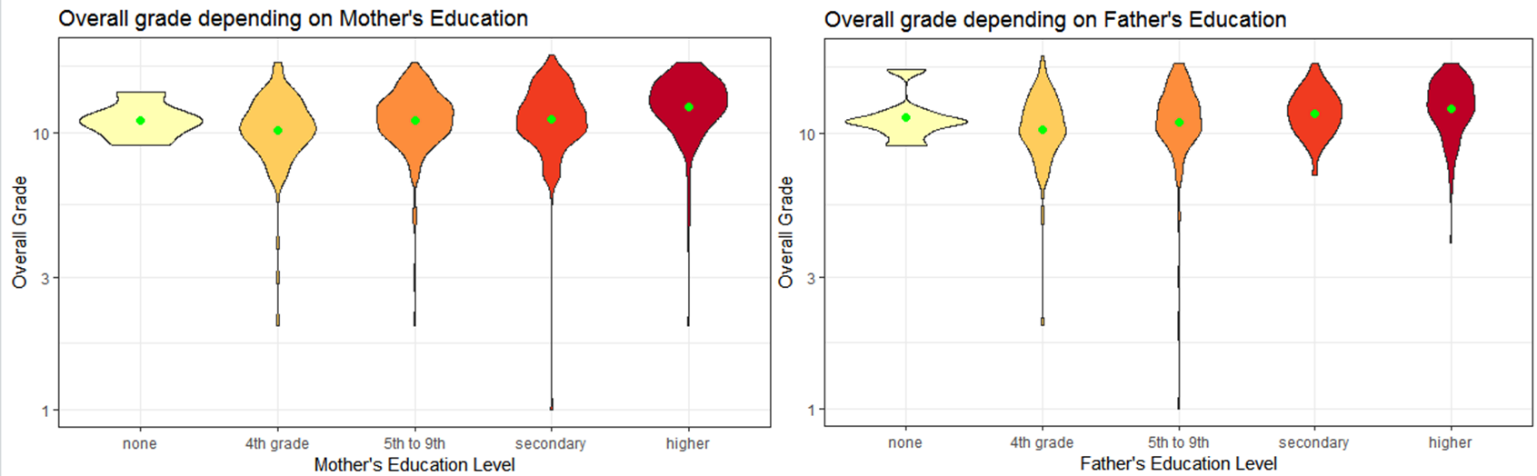


Figure 17: Violin plots of overallgrades by Mother and Father education

*ANOVA*

One-way ANOVA Null Hypothesis

From the analysis we can see that the relationships between the level of the Mother’s education is highly significant [F (4,644)=13.77, p.001] and Father’s education is highly significant [F(4,644)=9.371,p<.001] at 99% confidence we accept Hₐ : The statistical relationship between the Parent’s level of education and student’s grades is significant (Figure 18).



Figure 18: ANOVA statistics of overallgrades by Mother and Father education

*Regression*

Multiple regression analysis was used to test if the education level of student’s Mothers and Fathers significantly predicted student’s overall grades. The results of the regression indicated the Mother’s education level explained 7.3% of the variance (R2=0.07,F(1,647)=50.03,p<.001)(Figure 19), whilst the Father’s education level explained slightly less at 5.2% of the variance (R2=0.052,F(1,647)=35.47,p<.001)( Figure 20).

That means we can reject the null hypothesis that there is no relationship. Further our modelling estimates that for every level of education increase for the mother a student’s grades will increase, on average, by 0.671 (rounded) and for the father’s education the increase will be 0.589 (rounded).



Figure 19: Regression statistics of overall grades by Mother education



Figure 20. Regression statistics of overall grades by Father education

*Question 2: Are the grades from G1, G2 and G3 effected by which school the students attend?*

The following tests are performed to answer a question, Does the school which students attend have a significant effect on their grades G1, G2 and G3?

Test for variance homogeneity

* The variances of G1 across two different schools (GP & MS) are not homogeneous (df=1, p=0.0002754)
* The variances of G2 across two different schools (GP & MS) are not homogeneous (df=1, p<0.001)
* The variances of G3 across two different schools (GP & MS) are not homogeneous (df=1, p<0.001)

Thus, we run a t-test assuming non-homogeneous variances for grades G1, G2 and G3 across different schools.

Welch test (or) t-test assuming non-homogeneous variances

* The difference in G1 between GP (M = 11.98582) and MS (M = 10.30088) schools is significant (t (385.77) = 7.3114, p < .001).
* The difference in G2 between GP (M = 12.14421) and MS (M = 10.49558) schools is significant (t (354.63) = 6.4844, p < .001).
* The difference in G3 between GP (M = 12.57683) and MS (M = 10.65044) schools is significant (t (340.49) = 6.7545, p < .001).

One-Way ANOVA (F-test)

* There is a significant relationship between G1 and its corresponding school (F (1, 647) = 60.59, p<.001).
* There is a significant relationship between G2 and its corresponding school (F (1, 647) = 50.78, p<.001).
* There is a significant relationship between G3 and its corresponding school (F (1, 647) = 56.89, p<.001).

PostHoc Analysis

* The different PostHoc analysis procedures come to different results with respect to the G1 difference between GP and MS schools. This is most likely due to the non-homogeneous variances.
* The different PostHoc analysis procedures come to different results with respect to the G2 difference between GP and MS schools. This is most likely due to the non-homogeneous variances.
* The different PostHoc analysis procedures come to different results with respect to the G3 difference between GP and MS schools. This is most likely due to the non-homogeneous variances.

*Question 3: Does the level of study undertaken by a student reflect the number of days taken absent?*

Conduct a One-Way Analysis of Variance (ANOVA):to study time variable to that of absences.

Box plot showing relationship study time to that of absences (Figure 21).

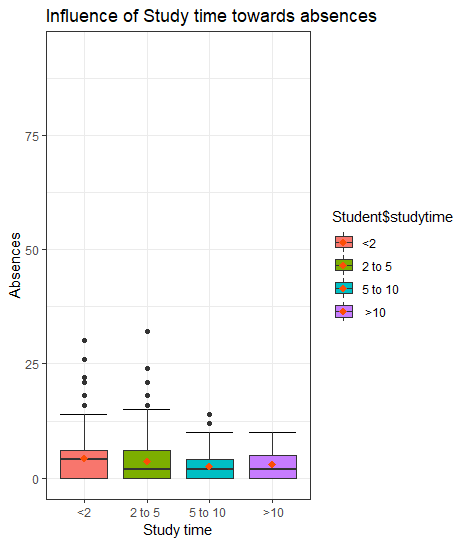


Figure 21. Box plot for Q3

Result: Analysis of variance showed a statistically significant difference at the p < .05 level in absence scores for the four studytime groups: F (3, 649) = 3.479, p = .0157

*Question 4: “Can the students’ failures be explained by number absences?”*

Another analysis was conducted to answer the following question: “Can the students’ failures be explained by number absences?”

To answer this question, the following hypothesizes has been developed:

Hₒ : The statistical relationship between the variables is not significant

Hₐ : The statistical relationship between the variables is significant

Because the “Failures” and “AbsencesCat” are categorical variables with 3 or more values, we can apply ANOVA to see if there is a significant difference in means. A one-way between subjects ANOVA was conducted to compare the effects of absences on failures (Figure 22). The analysis of variance showed that the effect of absences on failures is significant [F(4,644) = 2.614, p<0.05]. Thereby, with 95% confidence, we accept Hₐ: The statistical relationship between absences and failures is significant.



*Figure 22. ANOVA analysis for overall relationship*

However, the ANOVA determines if there is at least 1 statistical difference. To see the detailed picture, the POSTHOC analysis (Tukey HSD test) has been conducted (Figure 23).



*Figure 23. POSTHOC analysis (Tukey HSD test)*

# Discussion and Conclusion

*Question 1:* *Are overall grades of students influenced by the level of education of their parents?*

From the analysis we can see that a student’s grades are influenced by the level of education obtained by their parents. Of the schools surveyed we can see that a very small number of students have parents that have no education. So even though these results may have some influence on the overall findings the data still suggests that the higher the parent’s level of education the better the student’s grades will get.

If we were to advise on strategies for schools to improve grades those strategies might be to provide better support programs or additional education for those students whose parents have lower levels of education.

*Question 2: Are the grades from G1, G2 and G3 effected by which school the students attend?*

The test for variance homogeneity draws a conclusion that the variances of all the grades G1, G2 and G3 across the two different schools GP and MS are not homogeneous. Hence, we test the variables using a welch test. The welch test performed leads to an understanding that the difference in all the grades G1, G2 and G3 across the two different schools is significant. The conclusion drawn from the welch test is cross checked using a One-Way ANOVA which indicates that all the grades G1, G2 and G3 have a significant relationship with the schools GP and MS. A PostHoc analysis is conducted using both Pairwise t-Tests with Bonferroni Adjustment and Tukey HSD Test for all the grades G1, G2 and G3 across the different schools but they come to different results indicating that it wold be most likely due to non-homogeneous variances of the grades across the schools.

*Does the level of study undertaken by a student reflect the number of days taken absent?*

One-way ANOVA is used see whether a categorical variable (with K ≥ 2 levels) has a significant influence on the variance of a variable. In this data set categorical variable studytime has K more than 2,ie <2, 2to 5,5 to 10 and >10 categories which is used to see its significant influence on absence. As per the result from the test, there is a significant relationship between studytime and absence which is known from the F statistic as it is greater than 1. Therefore, we reject null hypothesis. The null hypothesis is that the variances among study period groups are the same.

As per the results from ANOVA test studytime has influence over absence, POSTHOC analysis can be conducted further to see , the difference in mean among the studytime group variables .

*Question 4: “Can the students’ failures be explained by number of absences?”*

Based on the results obtained from ANOVA analysis (Figure 22) and further POSTHOC (Figure 23), we can conclude that there is a relationship between these to variables. An F value obtained from the analysis [F(4,644) = 2.614 is slightly higher that F critical for these particular degrees of freedom [F-crit(4, 644) = 2.38576613] that means that there is a small relationship between the number of absences and number of failures. Further POSTHOC analysis showed that there is a significant difference in means between “Very Low” and “Medium” groups of values [p<0.05].

Overall, that means that according to this particular piece of data, number of student’s missed classes slightly influence his or her chances to fail the course (the more missed classes lead to higher failure chances). However, the relationship is not so strong that means that there are other factors that can explain student’s failures better.

# Appendix (R-Code)

#######################################

############ Item 5 ###############

#######################################

# Generate summary data for key variables for testing whether Parent's education level influences student grades

summary(student$G1)

capture.output(summary(student$G1), file = "Descriptive\_G1.doc")

summary(student$G2)

capture.output(summary(student$G2), file = "Descriptive\_G2.doc")

summary(student$G3)

capture.output(summary(student$G3), file = "Descriptive\_G3.doc")

# Create Overall Grade for the 3 periods

student$overallgrade <- round((student$G1 + student$G2 + student$G3)/3)

summary(student$overallgrade)

capture.output(summary(student$overallgrade), file = "Descriptive\_overallgrade.doc")

summary(student$Medu)

capture.output(summary(student$Medu), file = "Descriptive\_Medu.doc")

summary(student$Fedu)

capture.output(summary(student$Fedu), file = "Descriptive\_Fedu.doc")

# Standard deviations for key variables

sd(student$G1)

sd(student$G2)

sd(student$G3)

sd(student$overallgrade)

sd(student$Medu)

sd(student$Fedu)

#######################################

############ Item 6 ###############

#######################################

# Create categorical variable for level of Mother's education

student$MotherEdu <- factor(student$Medu, levels = c(0, 1,2,3,4), labels = c("none","4th grade","5th to 9th","secondary","higher")) #Transform the data to factor

# Summary tables for Mother's education levels

MotherEduLevel\_data<-ddply(student, c("MotherEdu"), summarise, N = length(MotherEdu),

overallgrades\_skewness=round(skewness(overallgrade),digits = 2),

overallgrades\_avg=round(mean(overallgrade),digits=2),

overallgrades\_sd=round(sd(overallgrade),digits=2),

overallgrades\_var=round(overallgrades\_avg,overallgrades\_sd),

overallgrades\_med=round(median(overallgrade),digits=2),

first\_quartile=round(summary(overallgrade)[2],digits = 2),

third\_Quartile=round(summary(overallgrade)[5],digits=2))

##RTF file is used to export the above created summary table.

MotherEdu.rtf<-RTF("MotherEdu.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(MotherEdu.rtf ,MotherEduLevel\_data , font.size=8, row.names=TRUE, NA.string="-")

done(MotherEdu.rtf)

# Create categorical variable for level of Fathers's education

student$FatherEdu <- factor(student$Fedu, levels = c(0, 1,2,3,4), labels = c("none","4th grade","5th to 9th","secondary","higher")) #Transform the data to factor

# Summary tables for Father's education levels

FatherEduLevel\_data<-ddply(student, c("FatherEdu"), summarise, N = length(FatherEdu),

overallgrades\_skewness=round(skewness(overallgrade),digits = 2),

overallgrades\_avg=round(mean(overallgrade),digits=2),

overallgrades\_sd=round(sd(overallgrade),digits=2),

overallgrades\_var=round(overallgrades\_avg,overallgrades\_sd),

overallgrades\_med=round(median(overallgrade),digits=2),

first\_quartile=round(summary(overallgrade)[2],digits = 2),

third\_Quartile=round(summary(overallgrade)[5],digits=2))

##RTF file is used to export the above created summary table.

FatherEdu.rtf<-RTF("FatherEdu.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(FatherEdu.rtf ,FatherEduLevel\_data , font.size=8, row.names=TRUE, NA.string="-")

done(FatherEdu.rtf)

# Create Box Plot for overall grades depending on Mother's education

MotherEduBoxPlot <- ggplot(student, aes(x=MotherEdu, y=overallgrade,fill=student$MotherEdu)) +

geom\_boxplot() +

theme\_bw()+

labs(title = "Influence of Mother's education on Overall Grades")+

xlab("Mother's Education")+

ylab("Overall Grades")+

stat\_summary(fun.y = mean, geom = "point",

shape = 18, size = 2.5, color = "#FC4E07") # Add a Box Plot on Top with red color on the mean.

png(file ="MotherEduBoxPlot.png", width = 900, height = 900) #Create and edit PNG file

print(MotherEduBoxPlot) #Print the plot to the file

dev.off() #Save file and close PNG editor

# Create Violin Plot for overall grades depending on Mother's education

MotherEduViolin <- ggplot(student, aes(x=MotherEdu, y=overallgrade, fill=student$MotherEdu)) +

scale\_y\_log10() +

geom\_violin() + # Make it a Violin Plot

theme\_bw() + # Change Background Color

labs(title = "Overall grade depending on Mother's Education") + # Add a Title

xlab("Mother's Education Level") + # Label for X-Axis

ylab("Overall Grade") + # Label for Y-Axis

scale\_fill\_brewer(palette = "YlOrRd") +

guides(fill=FALSE) + # Remove the legend

stat\_summary(fun.y=mean, geom="point", size=2, color="green") # violin plot with mean points

png(file ="MotherEduViolin.png", width = 900, height = 900) #Create and edit PNG file

print(MotherEduViolin) #Print the plot to the file

dev.off() #Save file and close PNG editor

# Create Box Plot for overll grades depending on Father's education

FatherEduBoxPlot <- ggplot(student, aes(x=FatherEdu, y=overallgrade, fill=student$FatherEdu)) +

geom\_boxplot() +

theme\_bw()+

labs(title = "Influence of Father's education on Overall Grades")+

xlab("Father's Education")+

ylab("Overall Grades")+

stat\_summary(fun.y = mean, geom = "point",

shape = 18, size = 2.5, color = "#FC4E07") # Add a Box Plot on Top with red color on the mean.

png(file ="FatherEduBoxPlot.png", width = 900, height = 900) #Create and edit PNG file

print(FatherEduBoxPlot) #Print the plot to the file

dev.off() #Save file and close PNG editor

# Create Violin Plot for overall grades depending on Father's education

FatherEduViolin <- ggplot(student, aes(x=FatherEdu, y=overallgrade, fill=FatherEdu)) +

scale\_y\_log10() +

geom\_violin() + # Make it a Violin Plot

theme\_bw() + # Change Background Color

labs(title = "Overall grade depending on Father's Education") + # Add a Title

xlab("Father's Education Level") + # Label for X-Axis

ylab("Overall Grade") + # Label for Y-Axis

scale\_fill\_brewer(palette = "YlOrRd") +

guides(fill=FALSE) + # Remove the legend

stat\_summary(fun.y=mean, geom="point", size=2, color="green") # violin plot with mean points

png(file ="FatherEduViolin.png", width = 900, height = 900) #Create and edit PNG file

print(FatherEduViolin) #Print the plot to the file

dev.off() #Save file and close PNG editor

# Conduct a One-Way Analysis of Variance for overall grades for Mother's Education (ANOVA)

aovResult = aov(overallgrade ~ MotherEdu, data=student)

summary(aovResult)

# Conduct a One-Way Analysis of Variance for overall grades for Father's Education (ANOVA)

aovResult = aov(overallgrade ~ FatherEdu, data=student)

summary(aovResult)

# Create a simple linear regression for Medu on overallgrade

lm\_result <- lm(student$overallgrade~student$Medu)

lm\_result

summary(lm\_result)

# Create a simple linear regression for Fedu on overallgrade

lm\_result <- lm(student$overallgrade~student$Fedu)

lm\_result

summary(lm\_result)

#Getting a summary information for the descriptive part

summary(failures)

capture.output(summary(failures), file = "Descriptive.doc")

summary(address)

capture.output(summary(address), file = "Descriptive.doc")

summary(health)

capture.output(summary(health), file = "Descriptive.doc")

summary(absences)

capture.output(summary(absences), file = "Descriptive.doc")

#Creating a density plot for absence variable and exporting

plot1 <- ggplot(student, aes(x=absences, fill = absences)) + geom\_density(alpha=.75)

png(file = "density\_absences.png", width=800, height = 800)

plot1

dev.off()

#Creating a bar chart for the failure variable and exporting

plot2 <- qplot(failures, data = student, geom = "bar")

png(file = "bar\_failures.png", width = 800, heigh = 800)

plot2

dev.off()

#Creating a boxplot for the relation between failures and absences

Plot3 <- ggplot(student, aes(x=failures, y = absences, fill = failures, group = failures)) + geom\_boxplot() + labs(title = "Number of failures in relation to number of failures") + ylab("Number of absences") + xlab("Number of failures")

png("boxplot.png", width = 800, height = 800)

plot3

dev.off()

#Creating a new categorical variable from absences variable

student$absencesCat[student$absences>=0 & student$absences<=5] <- "Very Low"

student$absencesCat[student$absences>=6 & student$absences<=10] <- "Low"

student$absencesCat[student$absences>=11 & student$absences<=15] <- "Medium"

student$absencesCat[student$absences>=16 & student$absences<=20] <- "High"

student$absencesCat[student$absences>=21 & student$absences<=25] <- "Very High"

#Conducting ANOVA for failures and absences + exporting the data

aov = aov (failures~absencesCat, data = student)

capture.output(summary(aov), file = "ANOVA.doc")

#Conduct a POSTHOC analysis

TukeyHSD(aov, conf.level = 0.95)

capture.output(TukeyHSD(aov, conf.level = 0.95), file = "POSTHOC.doc")

###study the distribution of the key variable absences.

##skewness function is used to see the distribution pattern of the absence variable.ie.normal distribution.

skewness(Student$absences)

# [1] 2.01602

##A pdf is created to show the box plot for the variable absence.

pdf(file="student\_absences.pdf", width=9, heigh=9)

par(mfrow =c(2,2))

boxplot(Student$absences)

dev.off()

##Below summary commands are used to find the mean ,median,standard deviation and 3rd quartile for the absence variable.

absence\_data<-ddply(Student, c("school"), summarise, N = length(school), absences\_avg=round(mean(absences),digits=2),

absences\_sd=round(sd(absences),digits=2),med\_absences=round(median(absences),digits=2),third\_Quartile=round(summary(absences)[5],digits=2))

##RTF file is used to export the above created summary table.

output.rtf<-RTF("output.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(output.rtf ,absence\_data , font.size=8, row.names=TRUE, NA.string="-")

done(output.rtf)

##The below corr and corrmatrix plot finds the relationship among the variables like Medu,Fedu,traveltime,studytime,failures,famrel with that of absences.

corr1 = tapply(rownames(Student),Student$school, function(x) cor(Student[x,c(7,8,13,14,15,24,30)],

method="pearson", use="pairwise.complete.obs"))

pdf(file="corrplots.pdf", width=9, heigh=9)

par(mfrow =c(2,2))

corrplot.mixed(corr1$GP,lower="number",upper="circle", order="hclust",title="GP",mar=c(0,0,0,0))corrplot.mixed(corr1$MS,lower="number",upper="circle", order="hclust",title="MS",mar=c(0,0,0,0))

dev.off()

#####################################################################################################################################################

##Finding the data distribution of the Study time data in the Student CSV file

##create factor for school GP and MS with their school names respectively.

Student$school <- factor(Student$school, levels = c("GP", "MS"),

labels = c(" Gabriel Pereira", "Mousinho da Silveira"))

##Density plot showing the study time spent by students of each school

png(file="study\_schools.png", width=900, height=900)ggplot(Student, aes(x=Student$studytime, fill=Student$school)) + geom\_density(alpha = .75) + # Make it a Density Plot

scale\_fill\_brewer(palette="Set2", name="Schools") + # Use a filling color from RColorBrewer

labs(title = "Weekly Study time spent for both the schools ") + # Add a Title

theme\_bw() + # Change Background Color

ylab("Both Schools ") + # Label for Y-Axis

xlab("Time spent per week on study") # Label for X-Axis

dev.off()

##create a summary table to show the mean,median,sd and 3 rd quartile of the studytime varaibale data distribution.

studytime\_data<-ddply(Student, c("school"), summarise, N = length(school), study\_avg=round(mean(studytime),digits=2), study\_sd=round(sd(studytime),digits=2),med\_study=round(median(studytime),digits=2),third\_Quartile=round(summary(studytime)[5],digits=2))

##RTF file is craeted and exported that contains the summary table.

study.rtf<-RTF("output.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(study.rtf ,studytime\_data , font.size=8, row.names=TRUE, NA.string="-")

done(output.rtf)

##################################################################################

##failures data distribution is studied.

##Violin plot showing the number of failures for both schools and exported as a png file

png(file="Failures\_gender.png", width=900, height=900)

ggplot(Student, aes(x = Student$sex, y = Student$failures, fill = Student$se)) +

geom\_violin() +# Make it a Violin Plot

theme\_bw() + # Change Background Color

labs(title = "Comparing failure for both schools") + # Add a Title

xlab("Male and female students")+ #label for X=axis

ylab("Failure") + # Label for Y-Axis

ylim(1,4) +

guides(fill=FALSE) +# Remove the legend

scale\_fill\_brewer(palette="YlOrRd")+ # Change Fill Color to Yellow orange and red from RColorBrewer library

geom\_boxplot(width=0.2)+

stat\_summary(fun.y = mean, geom = "point", shape = 18, size = 2.5, color = "#FC4E07") # Add a Box Plot on Top with red color on the mean.

dev.off()

## create a rtf document called "output.rtf" and as a data frame "summarystat" as a table in this document.

summarystat=data.frame(ddply(forestfires, c("quarter"), summarise, N = length(quarter), FFMC\_avg=round(mean(FFMC),digits=2),DMC\_avg=round(mean(DMC),digits=2),DMC\_avg=round(mean(DMC),digits=2), DC\_avg=round(mean(DC),digits=2),ISI\_avg=round(mean(ISI),digits=2),temp\_avg=round(mean(temp),digits=2),RH\_avg=round(mean(RH),digits=2),wind\_avg=round(mean(wind),digits=2),rain\_avg=round(mean(rain),digits=2), area\_avg=round(mean(area),digits=2)))

output.rtf<-RTF("output.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(output.rtf ,summarystat , font.size=8, row.names=TRUE, NA.string="-")

done(output.rtf)

##create summary table showing mean and sd for both schools and export the table as a rtf

failures\_data=data.frame(ddply(Student, c("school","sex"), summarise, N = length(school), failures\_avg=round(mean(failures),digits=2), failures\_sd=round(sd(failures),digits=2)))

output\_failures.rtf<-RTF("failures.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(output\_failures.rtf ,failures\_data , font.size=8, row.names=TRUE, NA.string="-")

done(output.rtf)

##################################################################################

# Conduct a One-Way Analysis of Variance (ANOVA) for study time variable to that of absences.

##create a factor for various levels in the study time.

Student$studytime <- factor(Student$studytime, levels = c(1, 2,3,4), labels = c("<2","2 to 5 ","5 to 10 ", " >10"))

##Box plot showing relationship study time to that of absences.

png(file="Study\_absences.png", width=900, height=900)

ggplot(Student[!is.na(Student$studytime),],aes(x=Student$studytime, y=Student$absences, fill=Student$studytime))+

stat\_boxplot(geom= 'errorbar')+

geom\_boxplot()+

theme\_bw()+

labs(title = "Influence of Study time towards absences")+

xlab("Study time")+

ylab("Absences")+ ylim(0,93)+

stat\_summary(fun.y = mean, geom = "point", shape = 18, size = 2.5, color = "#FC4E07") # Add a Box Plot on Top with red color on the mean.

dev.off()

# Conduct a One-Way Analysis of Variance (ANOVA)

aovResult = aov(absences ~ Student$studytime, data=Student)

summary(aovResult)

########################################################################

skewness(student$G1)

skewness(student$G2) ##skweness() function from moments package

skewness(student$G3)

par(mfrow=c(1,3))

boxplot(student$G1, title="G1")

boxplot(student$G2, title="G2") ##boxplots for grades

boxplot(student$G3, title="G3")

dev.copy(png, 'boxplots.png', width=800, height=800) ###Exporting the plot as a png file using dev.copy() function

dev.off()

summarytable<-ddply(student, .(school), summarize,

N=length(school),

G1\_avg = round(mean(G1), 2),

G1\_sd = round(sd(G1), 2),

med\_G1=round(median(G1),digits=2),

G2\_avg = round(mean(G1), 2), ###ddply() to create summary table

G2\_sd = round(sd(G1), 2),

med\_G2=round(median(G1),digits=2),

G3\_avg = round(mean(G1), 2),

G3\_sd = round(sd(G1), 2),

med\_G3=round(median(G3),digits=2))

summarytable.rtf<-RTF("summarytable.rtf", width=10, height=12, font.size=9, omi=c(1,1,1,1))

addTable(summarytable.rtf , summarytable , font.size=6, row.names=TRUE, NA.string="-") ###creating rtf file and adding data as a table to it.

done(summarytable.rtf)

winsor <- function(x, multiplier)

{

if(length(multiplier) != 1 || multiplier <= 0)

{

stop("bad value for 'multiplier'")

}

quartile1 = summary(x)[2] # Calculate lower quartile

quartile3 = summary(x)[5] # Calculate upper quartile

iqrange = IQR(x) # Calculate interquartile range

y <- x ###Creating a function called winsor to winsorize data

boundary1 = quartile1 - iqrange \* multiplier

boundary2 = quartile3 + iqrange \* multiplier

y[ y < boundary1 ] <- boundary1

y[ y > boundary2 ] <- boundary2

y

}

par(mfrow=c(1,2)) ###dividing the plot

corr1 <- cor(student[c(3, 7, 8, 13, 14, 15, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33)], method="pearson",

use="pairwise.complete.obs") ###creating correlation matrix

corrplot.mixed(corr1, lower="number", upper="circle") ###plotting correlation graphs

corrplot(corr1, order="hclust", addrect = 2)

dev.copy(png, 'correlationplot.png', width=800, height=800) ###Exporting the plot as a png file using dev.copy() function

dev.off()

cor.mtest <- function(mat, method="pearson", conf.level = 0.95) {

mat <- as.matrix(mat)

n <- ncol(mat)

p.mat <- lowCI.mat <- uppCI.mat <- matrix(NA, n, n)

diag(p.mat) <- 0

diag(lowCI.mat) <- diag(uppCI.mat) <- 1 ### creating function cor.mtest() for significance test

for (i in 1:(n - 1)) {

for (j in (i + 1):n) {

tmp <- cor.test(mat[, i], mat[, j],

method=method, conf.level = conf.level)

p.mat[i, j] <- p.mat[j, i] <- tmp$p.value

lowCI.mat[i, j] <- lowCI.mat[j, i] <- tmp$conf.int[1]

uppCI.mat[i, j] <- uppCI.mat[j, i] <- tmp$conf.int[2]

}

}

return(list(p.mat, lowCI.mat, uppCI.mat))

}

corr2 <- cor(student[c(31, 32, 33)], method="pearson",

use="pairwise.complete.obs") ##creating correlation matrix

corr3 <- cor.mtest(student[c(31, 32, 33)], method="pearson", 0.95) ###creating significance matrix

par(mfrow=c(1,1)) ###dividing the plot

corrplot.mixed(corr2, p.mat=corr3[[1]], lower="number", upper="circle", sig.level = 0.10, title="correlation among grades with significance test", order="hclust", addrect=2, mar=c(0,0,1,0))

dev.copy(png, 'strongcorrelationssignificancetest.png', width=800, height=800) ###Exporting the plot as a png file using dev.copy() function

dev.off()

corr4 = tapply(rownames(student), student$school, ##creating correlation matrix for different schools

function(x) cor(student[x,c(31, 32, 33)],

method="pearson", use="pairwise.complete.obs"))

par(mfrow=c(1,2)) ###dividing the plot

corrplot.mixed(corr4$GP, lower="number", upper="circle",

title="GP", mar=c(0,0,1,0)) ###plotting the graphs

corrplot.mixed(corr4$MS, lower="number", upper="circle",

title="MS", mar=c(0,0,1,0))

dev.copy(png, 'strongcorrelationsfordifferentschools.png', width=800, height=800) ###Exporting the plot as a png file using dev.copy() function

dev.off()

var(student$G1[student$school=="GP"]) ###calculating variances with var()

var(student$G1[student$school=="MS"])

bartlett.test(G1 ~ school, data=student) ###bartlett's test for G1 in different schools

var(student$G2[student$school=="GP"])

var(student$G2[student$school=="MS"]) ###calculating variances with var()

bartlett.test(G2 ~ school, data=student) ###bartlett's test for G2 in different schools

var(student$G3[student$school=="GP"])

var(student$G3[student$school=="MS"]) ###calculating variances with var()

bartlett.test(G3 ~ school, data=student) ###bartlett's test for G3 in different schools

with(student, t.test(G1[school=="GP"], G1[school=="MS"], var.equal=FALSE))

with(student, t.test(G2[school=="GP"], G2[school=="MS"], var.equal=FALSE)) ###t-test for grades in different schools when variances are non homogeneous.

with(student, t.test(G3[school=="GP"], G3[school=="MS"], var.equal=FALSE))

aovResult = aov(G1 ~ school, data=student)

summary(aovResult)

aovResult1 = aov(G2 ~ school, data=student) ###one way anova test for grades in different schools.

summary(aovResult1)

aovResult2 = aov(G3 ~ school, data=student)

summary(aovResult2)

with(student, pairwise.t.test(G1, school, p.adjust="bonferroni", pool.sd=FALSE))

TukeyHSD(aovResult, conf.level = 0.95)

with(student, pairwise.t.test(G2, school, p.adjust="bonferroni", pool.sd=FALSE)) ###PostHoc analysis using bonferroni adjustment and TurkeyHSD

TukeyHSD(aovResult1, conf.level = 0.95)

with(student, pairwise.t.test(G3, school, p.adjust="bonferroni", pool.sd=FALSE))

TukeyHSD(aovResult2, conf.level = 0.95)

# References

Brown, R. B., & Saunders, M. (2008). *Dealing with Statistics: What you need to know*. New York: Open University Press.

Cortez, P. (2014). *Student Data Performance Set*. Retrieved from: <https://archive.ics.uci.edu/ml/datasets/Student+Performance>

Cortez, P., & Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance*. Paper presented at the 5th FUture Business TEChnology Conference, Porto, Portugal. <https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf>

Dodge, Y. (Ed.) (2006) The Oxford Dictionary of Statistical Terms. Oxford: Oxford University Press.

Frost, J. (2018). Difference between Descriptive and Inferential Statistics. Retrieved from <http://statisticsbyjim.com/basics/descriptive-inferential-statistics/>

Kalla, S., & Wilson, L. T. (2018). Inferential Statistics. Retrieved from <https://explorable.com/inferential-statistics>

Trochim, W. M. K. (2006). Descriptive Statistics. Retrieved from <http://www.socialresearchmethods.net/kb/statdesc.php>