**STAT 413 (Statistical Learning)**

Second Project Report

**Analyzing Student Performance Dataset**



A picture containing text, sign

Description automatically generated

**For dr. Ali Duman**

**By Group 6:**

**Omer Alabs / 201769570**

**Ahmed Alturfi / 201743590**

**Omar Al Bumadrah / 201636140**

**Abdullah Mohmmed / 201740190**

Table of Contents

[Table of figures 3](#_Toc100406059)

[Introduction 4](#_Toc100406060)

[Results 5](#_Toc100406061)

[Data preprocessing 5](#_Toc100406062)

[Variable’s type 5](#_Toc100406063)

[missing values 7](#_Toc100406064)

[Outliers 8](#_Toc100406065)

[Feature creation 9](#_Toc100406066)

[Data Visualization 9](#_Toc100406067)

[Correlation matrix 9](#_Toc100406068)

[alcohol consumption vs grades 11](#_Toc100406069)

[Internet access at home 11](#_Toc100406070)

[higher education 12](#_Toc100406071)

[Parents education 13](#_Toc100406072)

[Guardian effects on student's grades 13](#_Toc100406073)

[Age effect on Grades 14](#_Toc100406074)

[Grades vs Travel Time 15](#_Toc100406075)

[Grades vs Study Time 15](#_Toc100406076)

[Relationship and Patterns 16](#_Toc100406077)

[grades distribution 16](#_Toc100406078)

[Male vs Female comparison 17](#_Toc100406079)

[Alcoholic vs non-alcoholic students 17](#_Toc100406080)

[Students in relationship vs students with no relationship 18](#_Toc100406081)

[Parents together or apart 18](#_Toc100406082)

[High absences students vs low absences students 18](#_Toc100406083)

[Tentative Schedule 19](#_Toc100406084)

[Conclusion 20](#_Toc100406085)

# Table of figures

[Figure 1- missing values 8](#_Toc100406086)

[Figure 2- Outliers 9](#_Toc100406087)

[Figure 3- Correlation matrix 10](#_Toc100406088)

[Figure 4- alcohol workday consumption 11](https://kfupmedusa-my.sharepoint.com/personal/s201636140_kfupm_edu_sa/Documents/Second%20Report.docx#_Toc100406089)

[Figure 5- alcohol weekend consumption 11](https://kfupmedusa-my.sharepoint.com/personal/s201636140_kfupm_edu_sa/Documents/Second%20Report.docx#_Toc100406090)

[Figure 6- Internet access at home 11](#_Toc100406091)

[Figure 7- higher education 12](#_Toc100406092)

[Figure 8- higher education count 12](#_Toc100406093)

[Figure 9- Parents education 13](#_Toc100406094)

[Figure 10- Guardian of the student 13](#_Toc100406095)

[Figure 11- Age effect on Grades 14](#_Toc100406096)

[Figure 12- Grades vs Travel Time 15](#_Toc100406097)

[Figure 13- Grades vs Study Time 15](#_Toc100406098)

[Figure 14- grades distribution 16](#_Toc100406099)

# Introduction

Data preprocessing and data visualization play a crucial role in any data science project. To begin preprocessing the student grades dataset, we first search for the missing data and attempt to deal with these missing data.

We then investigate which columns are irrelevant to our objective and which columns can be combined in order to streamline the process. After preprocessing which is the most essential part to move to the second part, which is visualizing the preprocessed dataset.

We used various approaches for various kinds of data such as count plot and box plot for categorical data and scatter plot for numerical data. Data visualization empowers us to look deeper into the data, understand more about the relations between different columns, and interpret the relations with a clearer vision which leads us to complete the third part of this phase of the project.

To explore the data and describe it by the graphs and the data statistics, we discover a lot of interesting facts from the relationships between the grades and various attributes of the dataset. We also observed a lot of patterns that would help us in building the models in the following phase.

# Results

## Data Preprocessing

### Variable’s type

|  |  |  |
| --- | --- | --- |
| Variables | | |
| Variable name | Description | Variable type |
| school | binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira | Categorical |
| Sex | binary: 'F' - female or 'M' - male |
| address | tudent's home address type (binary: 'U' - urban or 'R' – rural) |
| famsize | family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) |
| Pstatus | parent's cohabitation status (binary: 'T' - living together or 'A' - apart) |
| Medu | mother's education (numeric: 0 - none, 1 – primary education (4th grade), 2 â€ “5th to 9th grade, 3 â€“secondary education or 4 â€“higher education) |
| Fedu |
| Mjob | job ('teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at\_home' or 'other') |
| Fjob |
| reason | reason to choose this school (close to 'home', school 'reputation', 'course' preference or 'other') |
| guardian | student's guardian ('mother', 'father' or 'other') |
| traveltime | home to school travel time (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| studytime | weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| failures | number of past class failures (n if 1<=n<3, else 4) |
| schoolsup | extra educational support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| paid | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no |
| activities | extra-curricular activities (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no |
| Variables | | |
| Variable name | Description | Variable type |
| internet | Internet access at home (binary: yes or no) Internet access at home (binary: yes or no) | Categorical |
| romantic | with a romantic relationship (binary: yes or no) |
| famrel | quality of family relationships (from 1 - very bad to 5 - excellent) |
| freetime | free time after school (from 1 - very low to 5 - very high) |
| goout | going out with friends (from 1 - very low to 5 - very high) |
| Dalc | workday alcohol consumption (from 1 - very low to 5 - very high) |
| Walc | weekend alcohol consumption (from 1 - very low to 5 - very high) |
| health | current health status (from 1 - very bad to 5 - very good) |
| age | student's age (from 15 to 22) | Numeric |
| absences | number of school absences (from 0 to 93) |
| G1 | first period grade (from 0 to 20) |
| G2 | second period grade (from 0 to 20) |
| G3 | final grade (from 0 to 20,) |
| avg | Average of the three periods grades |

### Missing Values

There are no missing values as shown below:

Table

Description automatically generated

Figure 1- missing values

### Outliers

Graphical user interface, text, application

Description automatically generated

Figure 2- Outliers

### Feature Creation

We created two more features, both are the average of other variables

1. **averageG:** which is the outcome of (G1+ G2+G3)/3
2. **Pedu:** which is the outcome of (Fedu + Medu )/2 .Merging Fedu and Medu is very useful to the model since their effects are almost identical as proved in the Data Visualization section.

## 2. Exploratory Data Visualization

### Correlation Matrix

Timeline

Description automatically generated

Figure 3- Correlation matrix

We can observe that **failures**, **Pedu**, **Dalc** and **studytime** are the variables that have a significant correlation with **averageG.**

### Alcohol Consumption

### Chart, box and whisker chart Description automatically generated Chart, box and whisker chart Description automatically generated

Figure 4- alcohol workday consumption

Figure 5- alcohol weekend consumption

As alcohol consumption on weekends increases, the average grade decreases.

### Internet Access at Home

Chart, box and whisker chart

Description automatically generated

Figure 6- Internet access at home

We can see clearly that having internet access at home would increase your chance to get a higher grade.

### Higher Education

Chart, box and whisker chart

Description automatically generated

Figure 7- higher education

Chart, bar chart

Description automatically generated

Figure 8- higher education count

Regardless of the small number of students who are not willing to take a higher education. It is obvious that students who want to complete their education and enter college have a higher chance to score a higher grade.

### Parents Education

Chart, box and whisker chart

Description automatically generated

Figure 9- Parents education

As the level of parents education increases, it reflects positively on the students' scores as it is noticeable in the boxplot graph in figure 9.

### Guardian Effects on Student's Grades

Chart, box and whisker chart

Description automatically generated

Figure 10- Guardian of the student

We can notice that there is almost no difference between the mother guardian and father guardian while clearly there is a significant difference when the guardian is other than the parents. It reflects negatively on the student performance on the average grade.

### Age Effect on Grades

Chart, line chart

Description automatically generated

Figure 11- Age effect on Grades

We can see that higher age corresponds to lower grades, the reason behind this could be that older students are students who repeated school, therefore will have a worse performance compared to younger students.

### Grades vs Travel Time

Chart, box and whisker chart

Description automatically generated

Figure 12- Grades vs Travel Time

The time from home to school has a strong effect on students’ performance!

### Grades vs Study Time

Chart, box and whisker chart

Description automatically generated

Figure 13- Grades vs Study Time

Interestingly, higher study time does not mean higher grades.

## 3. Relationships and Patterns

## Grades Distribution

Chart, histogram

Description automatically generated

Figure 14- grades distribution

We can see clearly that the distribution is almost little right skewed normal since mode < median < mean.

mode = 0 10.0

median = 11.666666666666666

mean = 11.677382966723075

### Male vs Female Comparison

Referring to the detailed analysis section in Collab Notebook, we can see clearly that female student on average 11.905714 has a much higher average grade than male student 11.345781.

Moreover, we can notice that female student study on average more time than male student also they are less likely to have a failure then male student.

|  |  |
| --- | --- |
| **Study time** | **Failures** |
| 2.060000 | 0.165714 |
| 1.734440 | 0.232365 |

### Alcoholic vs Non-Alcoholic Students

Referring to the detailed analysis section in Collab Notebook, it is clear that students who consume more alcoholic drinks on the weekend “ Walc” score 10.972222 on average, while students who consume less alcoholic drinks score a much higher grade of 11.857042 on average.

The same situation is on the weekday alcoholic consumption “ Dalc”. But since the count of students who consume alcoholic drinks during the weekdays is very small (27), we cannot conclude a statement.

### Students in Relationship vs Students with no Relationship

Referring to the detailed analysis section in Collab Notebook, we can see that there is almost no relationship between students in relationship and average grade 11.635945, as well as students with no relationship and average grade 11.633690. Almost no effect is noted.

### Parents Together or Apart

Referring to the detailed analysis section in Collab Notebook, we can see there is a little difference between students' average grade with the living situation ***together*** or ***apart***. Parents together with average grade 11.653153. Parents apart with average grade 11.849315. which means parents living apart have little positive effect on student performance.

However, the mean of the absences is significant between them, we can notice that it is 5.419355 for students whose parents are apart. While it is 3.577681 for students whose parents live together.

### High Absences Students vs Low Absences Students

Referring to the detailed analysis section in the Collab Notebook, we can see clearly that students with absences more than five score on average 11.226337 while the student who has absences less than five 11.847708 which is something significant to be looked at.

We can see clearly that students with absences more than five are more likely to have failures 0.265432, while the student who has absences less than five is less likely to have a failure 0.165501.

*\*Note: refer to the Collab Notebook for more analysis:*

<https://drive.google.com/file/d/1ErR7z4SOM30WD1yyiY3UmfWBf4IknhhV/view?usp=sharing>

# Tentative Schedule

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Week(s)** | **Assigned to** | **Phase** | **state** |
| **Problem Definition** | **1 - 4** | **All members** | **1** | **completed** |
| **Data Gathering** | **2 - 6** | **All members** | **completed** |
| **Submit First Report** | **6** | **Omer Alabas** | **Completed** |
| **Data Preprocessing and Cleaning** | **7** | **Omer Alabas**  **Omar Albumadrah** | **2** | **Completed** |
| **Exploratory Data Analysis (EDA)** | **7** | **Abdullah Mohammed**  **Ahmad Alturfi** | **Completed** |
| **Statistical Modelling** | **8 - 14** | **All members** | **Pending** |
| **Submit Progress Report** | **10** | **Omar Albumadrah** | **Completed** |
| **Model Evaluation and Improvement** | **10 - 15** | **All members** | **3** | **Pending** |
| **Submit Final Report** | **16** | **Abdullah Mohammed** | **Pending** |

*\*Note: some weeks are overlapping due to the flexible nature of some tasks.*

# Conclusion

Through the analysis that has been discussed earlier and throughout the statistical examination of the data, and after the data preprocessing, it has been proven that some features play a significant role on the average grade. However, we found that some of these features should be combined, like Father education ‘Fedu’, Mother education ‘Medu’ to a new feature called Parents education ‘Pedu’.

Moreover, it has been noticed that as alcohol consumption increases, the average grade decreases. Also having an internet access at home would increase the chance to get a higher grade. Furthermore, the desire of student to Complete the education in the future affects positively in getting higher grade.

Also, it has been observed that the student is more likely to get higher grades if the education level of their parents was higher. However, when the student age surpasses 18 year, it indicates that the student is more likely to have a failure before, as a result he would probably score low grades.

Additionally, the gender is a crucial factor, because the female students tend to study more than the male students, which results in getting higher grade for female students. Yet, high study time does not necessarily reflect high grades.