# Methods of Visualisation and Exploratory Analysis


## Done by:

## 101068721

## 1. Introduction

Grades are a better measure to predict life than any general intelligence test. Personality plays a huge role in the grades (Lex Borghans et al., 2016). Other major factors influencing are gender, family support (Yousuf Al Husaini and Shukor, 2023), and days of absence (Daily et al., 2020).

### 1.1 Problem statement and relevance

In this report, we will be looking at factors that can impact the student's scores. The factors considered are ones that the student has no control over, but this can help in providing more facilities for students which helps them overcome any struggles which are caused due to the external factors influencing in less marks.

Study suggests that students who fail a class and end up repeating the class generally are reported to have higher chances of dropping out of class. High school dropouts tend to have higher probabilities of being unemployed, incarcerated and have overall poor health (Giano, 2022). The influence of student grades on general life makes it a relevant topic for analysis.

### 1.2 Open themes in the research topic

Some of the open themes are:

I. Influence of demographic factors on student grades
II. Influence of extra educational support on student grades
III. Influence of school on student grades
IV. Influence of health on student grade
V. Influence of parental job and education on student grades

The demographic factors considered from the dataset are gender, age, relationship status. Educational support received from both family and school are explored in this project. Further, the health status of the student is studied to see the influence on grades. Parental job and education along with quality of family relationship data is also used to plot graphs. The entire columns and description for each section can be seen in the codebook section of the reports.

### 1.3 Data cleaning

The dataset has Portuguese subject student grades for two secondary education schools. The initial dataset had 649 rows and 33 columns. Only 16 columns from the existing dataset were chosen for the exploratory data analysis. Three extra columns were added based on the existing data. The first column added was the average of the three scores of G1, G2, G3. The second column added was the grade based on the introductory paper (Cortez and Silva, 2008.).

The columns were not encoded because the categorical columns were used to make specific graphs. Health, Mother's education, family relation, father's education and mother's and father's and mother's job columns were converted to qualitative columns as

they had few ordinal numbers. The dataset did not have any duplicated values or any missing values.

### 1.4 Codebook

The dataset had 33 features out of which 16 were selected for three exploratory data analysis. 3 columns were made from the existing data.

| Survey Item No | Variable name | Variable label | Values and labels | Used in the EDA |
|---|---|---|---|---|
| 1 | School | Name of school | GP: Gabriel Pereira<br>MS: Mousinho da Silveira | Yes |
| 2 | Sex | Gender of students | F: Female<br>M: Male | Yes |
| 3 | Age | Age of students | 15- 22 | Yes |
| 4 | Address | Student's home address | U: Urban<br>R: Rural | No |
| 5 | Famsize | Family size | LE3: <=3<br>GT3: >3 | No |
| 6 | Pstatus | Parent's cohabitation status | T : Together<br>A : Apart | No |
| 7 | Medu | Mother's education | 0: no education<br>1: primary education<br>2: 5th-9th grade<br>3: Secondary education<br>4: Higher education | Yes |
| 8 | Fedu | Father's education | 0: no education<br>1: primary education<br>2: 5th-9th grade<br>3: Secondary education<br>4: Higher education | Yes |
| 9 | Mjob | Mother's job | Teacher,<br>Health, Civil | Yes |

| | | | | |
|---|---|---|---|---|
| | | | services, at_home,other | |
| 10 | Fjob | Father's job | Teacher, Health, Civil services, at_home,other | Yes |
| 11 | reason | Reason to choose this school | Home: close to home<br>Reputation: school reputation<br>Course: course preference<br>Other | No |
| 12 | guardian | Student's guardian | Mother, father or other | No |
| 13 | traveltime | Home to school travel time | 1: <15 mins<br>2: 15-30 mins<br>3: 30 mins -1 hour<br>4: > 1 hour | No |
| 14 | studytime | Weekly study time | 1: < 2 hour<br>2: 2-5 hours<br>3: 5-10 hours<br>4: >10 hours | No |
| 15 | Failures | No of past class failures | 0-3 | No |
| 16 | Schoolsup | Educational support from school | Yes, No | Yes |
| 17 | famsup | Educational support from family | Yes, No | Yes |
| 18 | paid | Extra paid classes | Yes, No | No |
| 19 | Activities | Extracurricular activities | Yes, No | No |
| 20 | nursery | Attended nursery school | Yes, No | No |
| 21 | higher | Wants to take higher education | Yes, No | No |

| | | | | |
|---|---|---|---|---|
| 22 | internet | Internet access at home | Yes, No | No |
| 23 | romantic | In a romantic relationship | Yes, No | Yes |
| 24 | relationship_age | Minor (<18 age) or adult (>18 age) in a relationship or not | my: minor in a relationship<br>Mn: minor not in relationship<br>ay: adult in relationship<br>an: adult not in relationship | Yes |
| 25 | famrel | Quality of family relationship | 5: Excellent<br>4: Very good<br>3: Good<br>2: Bad<br>1: Very bad | Yes |
| 26 | freetime | Free time after school | 5: Very high<br>4: High<br>3: Medium<br>2: Low<br>1: Very low | No |
| 27 | goout | Going out with friends | 5: Very high<br>4: High<br>3: Medium<br>2: Low<br>1: Very low | No |
| 28 | Dalc | Workday alcohol consumption | 5: Very high<br>4: High<br>3: Medium<br>2: Low<br>1: Very low | No |
| 29 | Walc | Weekend alcohol consumption | 5: Very high<br>4: High<br>3: Medium<br>2: Low<br>1: Very low | No |
| 30 | Health | Health status | 5: Excellent<br>4: Very good | Yes |

| | | | | |
|---|---|---|---|---|
| | | | 3: Good<br>2: Bad<br>1: Very bad | |
| 31 | Absences | No of days of absence | 0-32 | Yes |
| 32 | G1 | First period score (max marks = 20) | 0-19 | Yes |
| 33 | G2 | Second period score (max marks = 20) | 0-19 | Yes |
| 34 | G3 | Third period score (max marks = 20) | 0-19 | Yes |
| 35 | avg | Average of G1, G2, G3 | | Yes |
| 36 | grade | Grades based on average scores | a: 16+<br>b : 4-15<br>c: 12-13<br>d: 10-11<br>f: <=9 | Yes |

## 2. Research questions

To understand the factors affecting student grades better the correlation between several columns and the score(G3) and grade columns are studied. Statistical significance of some of these factors are checked to ensure the findings are statistically valid too.

### 2.1 Is the distribution of scores G1, G2, G3 unimodal, bimodal or multimodal?

Density plot is used to see the distribution of all three grades. Density plot was used because it shows a smooth curve while histogram shows the value counts as boxes (Chip, 2023).

### 2.2 Is there a correlation between scores G1, G2, G3?

Since we have three grades which are taken in three periods of time, the linear correlation between each was checked. Pearson correlation was also checked to statistically test the linear correlation.

**2.3 Is there a correlation between schools and G3 score/grades?**
Barplots are used to compare the number of students. A mosaic plot is also used to see the relationship between grade and both the schools. Since both the schools do not have equal numbers of students, the mean scores of both the schools are compared using boxplot and bar charts. Since there are three scores which were taken at period 1, period 2 and period 3 a line plot was plotted to compare how the scores were changed over the period for both the schools.

To check the statistical significance of the difference between mean scores of both the schools, welch's T-test was conducted. Regular T-test was not conducted as there was difference in the number of students in each school The test scores of both schools were checked if they have normal distribution using quantile plot and histogram.

**2.4 Is there a correlation between gender and G3 score/grades?**
There are two genders used in the dataset and they are female and male. A bar chart was used to see the total number of students in each gender. Further, a mosaic plot was made using the genders and grades as they are both categorical variables. A violin plot was plotted to see the distribution of data for each gender and their scores and also see the quantiles in the graph. Welch's t test was conducted to see if there is significant difference between mean scores of both the genders. This was done after checking if both the groups are independent and also have normal distribution.

**2.5 Is there any correlation between health and days of absence on score/grades?**
The number of students with health status from 1-5 were plotted using a bar chart. The boxplot was also done to see the median of the score, with the use of jitters in boxplot, the data distribution could also be seen. Since we are only checking factors that cannot be controlled by students, we are assuming the days of absence are due to health reasons. To check the relationship between days of absence and health, a correlation matrix was also plotted. The distribution of the number of days of absence was checked by grouping it into health status using a boxplot. A scatter plot is also plotted to see if the absence and the grade has a negative linear relationship. To statistically see if grade and health status are independent from each other, we use the chi-square test.

**2.6 Is there any correlation between Age, relationship and scores/grade?**
Relationship is a parameter that can be controlled by the students, but we will be checking the relationship status with respect to the age of the student and its correlation to the scores or grades. The students were divided into minors in relationship and minor without relationship and adults in relationship and adults not in relationship. The histogram with the number of students grouped by age was plotted along with a stacked bar plot of the grade stacked with relationship status. The boxplot of G3 scores grouped by relationship status with respect to age was plotted to see the median and distribution of the scores.

**2.7 Is there any correlation between support received from family and school, parent's education, job, family relation quality and scores/grades?**

Bar plots with the number of students receiving educational support from family/school or not were plotted to compare the family and school educational support. Mosaic plots with education and jobs for both mother and father were compared against the grade to see the impact of the jobs and education of parents on scores.

The family relationship quality was also compared with the grades to see the impact on grades. A new data frame was created for students receiving both family and school educational support and a scatter plot was plotted for just these students to compare their scores. Lastly a chi-square test was done to see if the family support and school support are statistically independent of the grade column.

## 3. Analysis

The insights from the graphs plotted and the statistical significance will be discussed in this section. Before we move into the research questions and the insights from each one of them, we will be looking into the general inspection of the dataset using visualisation.

For general visualisation, scatter plot with G3 score is plotted against the index values. GP school data points were marked as red points and MS school's data points were marked as blue points. As seen in the graph below, we can see that scores are scattered and do not form a pattern for both the schools. A dataset was then shuffled, and a subset of the data was taken and plotted with the age shown in size of the data points. The scores were still scattered around and did not follow any specific pattern based on school or age.

The bar chart for the number of students with each grade was plotted and the majority of the students were graded as "a". Least number of students earned a "b" grade.

Apart from this, a copy of the dataset was made and the non-numerical columns were encoded using a label encoder and a correlation matrix was plotted to see the linear relationship. Mother's education has the highest value for positive linear relationship with the final score.

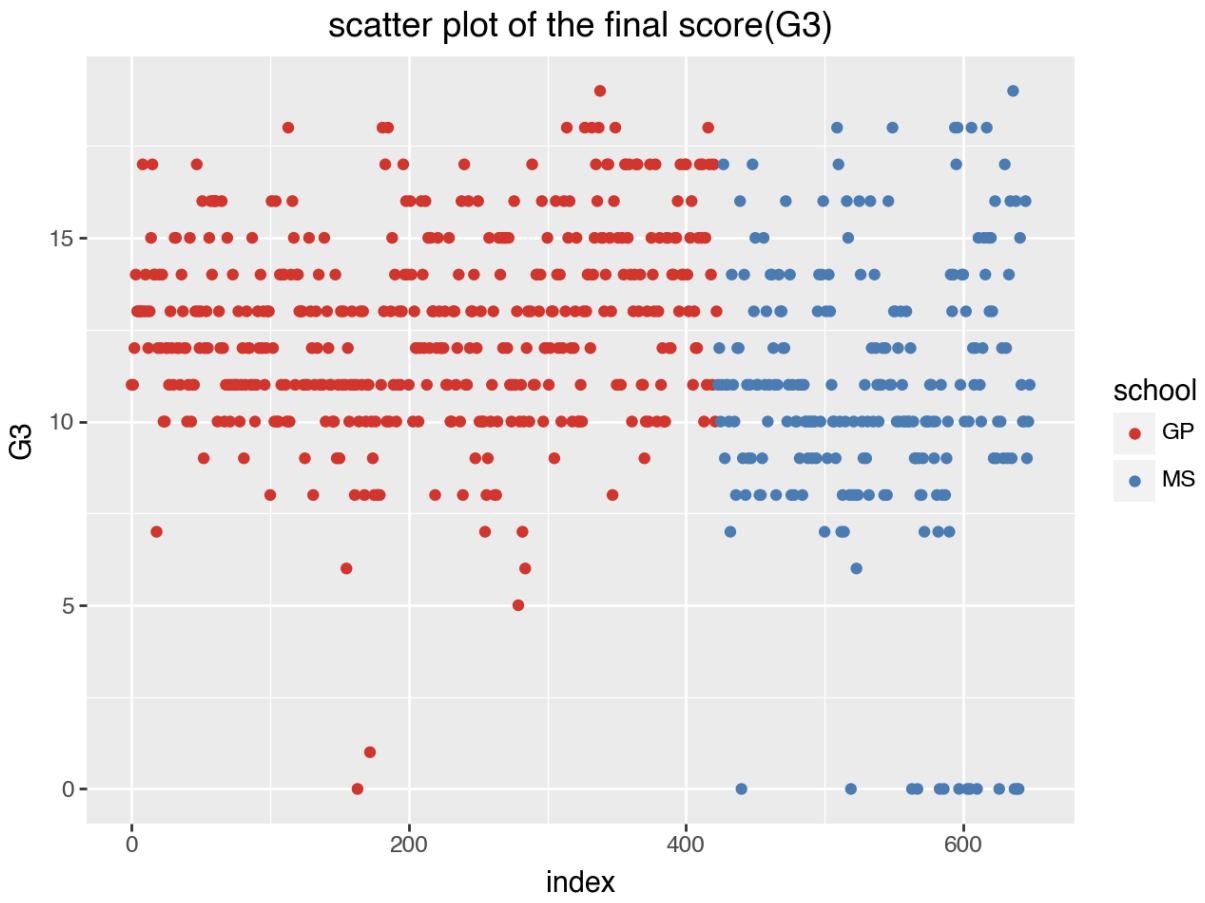


Figure 1: Scatter plot of G1 against the index

**3.1 Is the distribution of scores G1, G2, G3 unimodal, bimodal or multimodal?**
The distribution of the curves for scores G1, G2 and G3 were drawn using the density plot and overall, all the three densities have unimodal distribution. Majority of the students seem to have scored between 10-15 score in this dataset for all three tests.

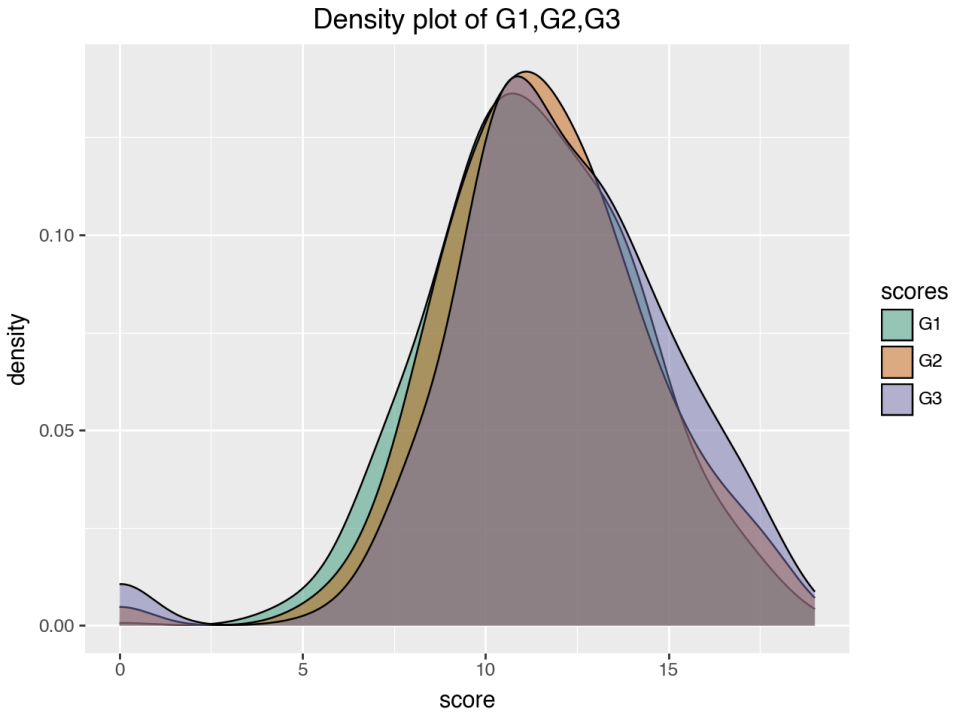


Figure 2: Density plot of G1, G2, G3

**3.2 Is there a correlation between scores G1, G2, G3?**

Using correlation matrix, we see that all the three scores have high positive correlation scores. The Pearson correlation test also shows a high statistical value and extremely low p-value between all the three grades showing that there is positive correlation between all three grades and they are statistically significant.

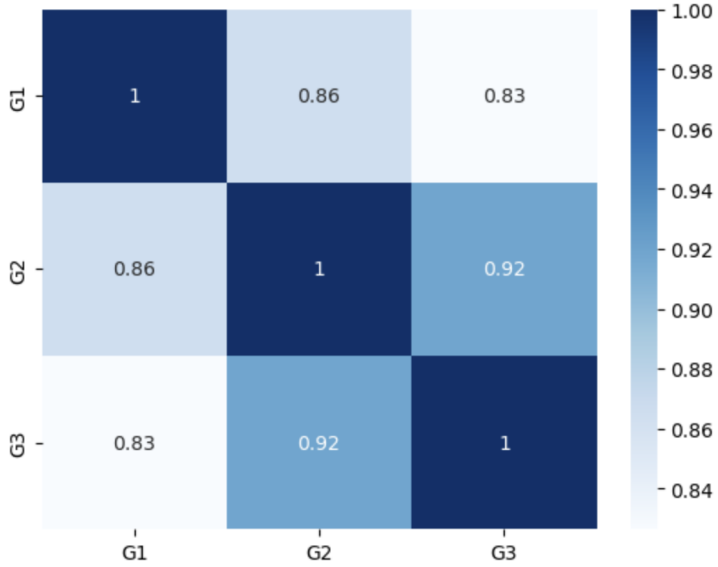


Figure 3: Correlation matrix of G1, G2, G3

As all the three grades have high correlation, G3 will be used for further graphs assuming that all three tests have similar relationships to other variables.

**3.3 Is there a correlation between schools and G3 score /grades?**

Checking the number of students in each school using a bar plot, we can see that GP has more students compared to MS school. The mosaic plot shows that the students in GP have more “a” grade compared to the “a” grades in MS school. The “f” grades seem to be more in MS school.

The boxplot of final scores of both the schools are plotted and we see that GP has a higher median as seen in the figure below.

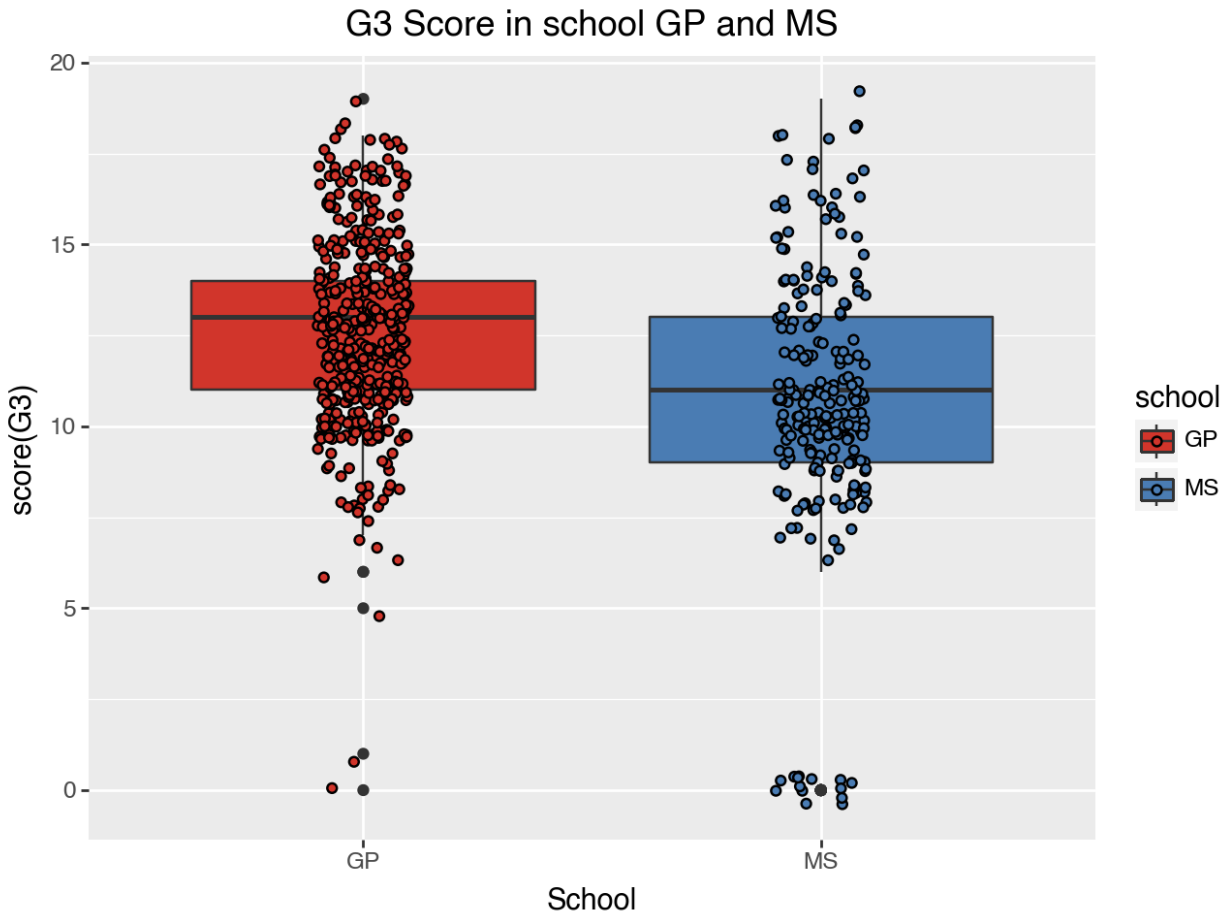


Figure 4: Box plot of G3 score of GP and MS

The mean bar graphs of both schools are plotted, and we see that the mean score of GP is higher than the mean score of MS.

As the three scores are taken in period 1, period 2 and period 3, a line plot of mean scores of both schools at three different time are plotted as shown below.

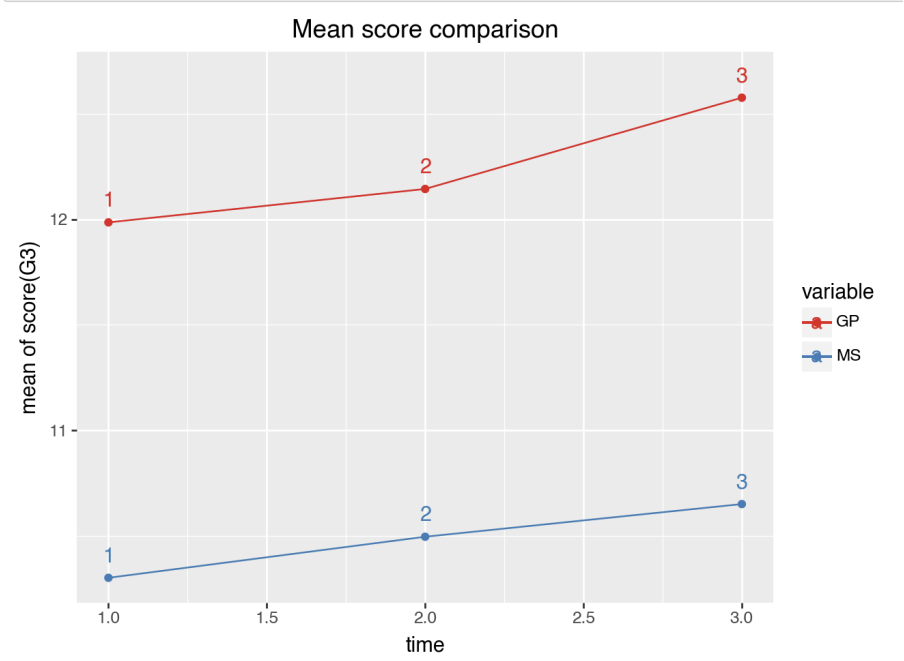


Figure 5: Line plot of mean scores of GP and MS against time/period

To ensure that the mean difference is purely by chance, Welch's T-test was conducted. Before conducting the test, the quantile plot was plotted for G3 scores of both the schools to see if they have normal distribution. They both seem to have a normal distribution with slight deviation as seen below. The histogram also showed a normal distribution.

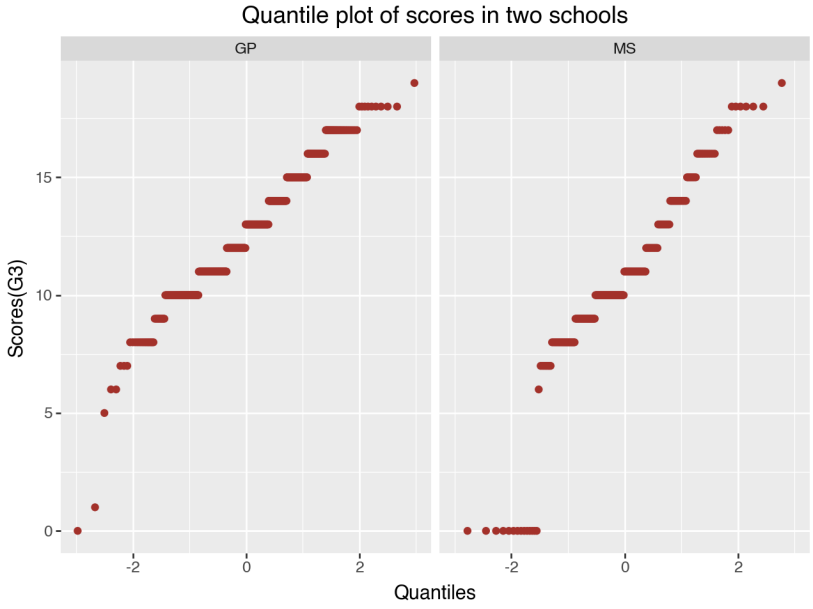


Figure 6: Quantile plots of G3 score of GP and MS

Welch's T-test was conducted after finding they have normal distribution and that they both are independent variables. Welch's was used because the number of students are

not equal in both schools. The null hypothesis was that both the schools have equal mean scores. The p-value was extremely low compared to the significance level of 0.05. This shows that the mean difference of scores between the school is not by chance but is statistically significant. GP has a better mean score than MS.

### 3.4 Is there a correlation between gender and G3 score /grades?

The number of female and male students in both schools is checked using a stacked bar plot. The stacked plot shows the number of female and male students in each school.

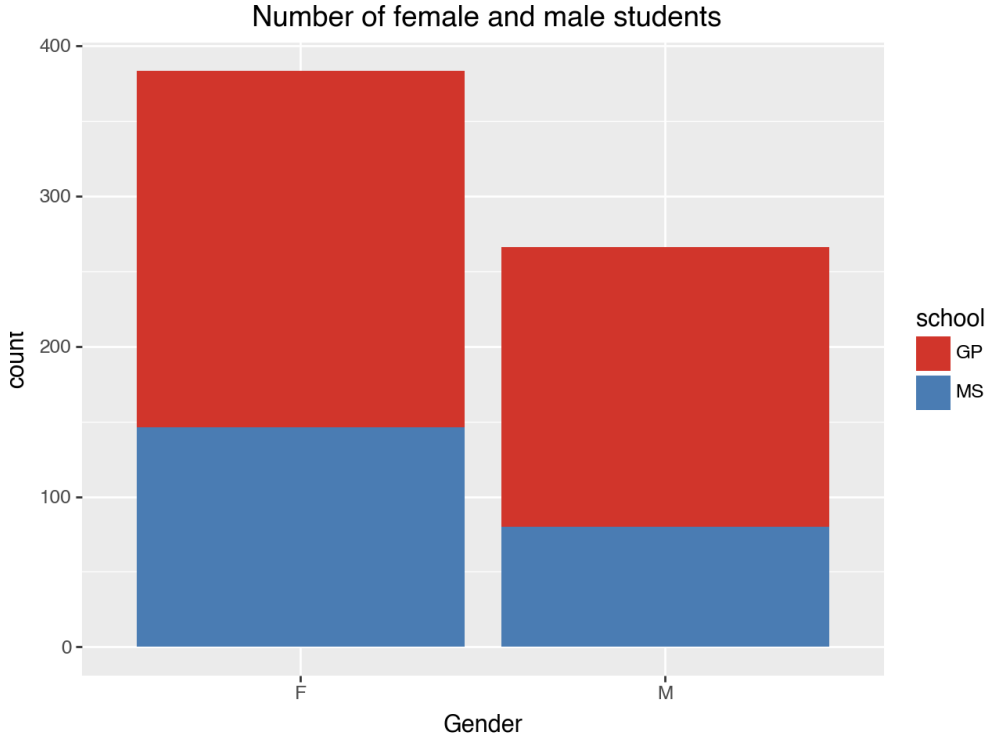


Figure 7: Bar plot of female and male students (stacked by school)

The mosaic plot was plotted after this with grades based on gender, it showed slightly more female students with “a” grade. To check the distribution and evaluate further, violin plot was plotted for G3 scores grouped by gender. As shown in the figure below, we can see that the female group tends to have a slightly higher median compared to the male students.

Figure 8: Violin plot of G3 score distribution grouped by gender

To evaluate the relevance of this in a statistical sense, Welch's T-test was done. To check if the groups have normal distribution histogram was made for both the groups by facet grid. Both the groups have a normal distribution.

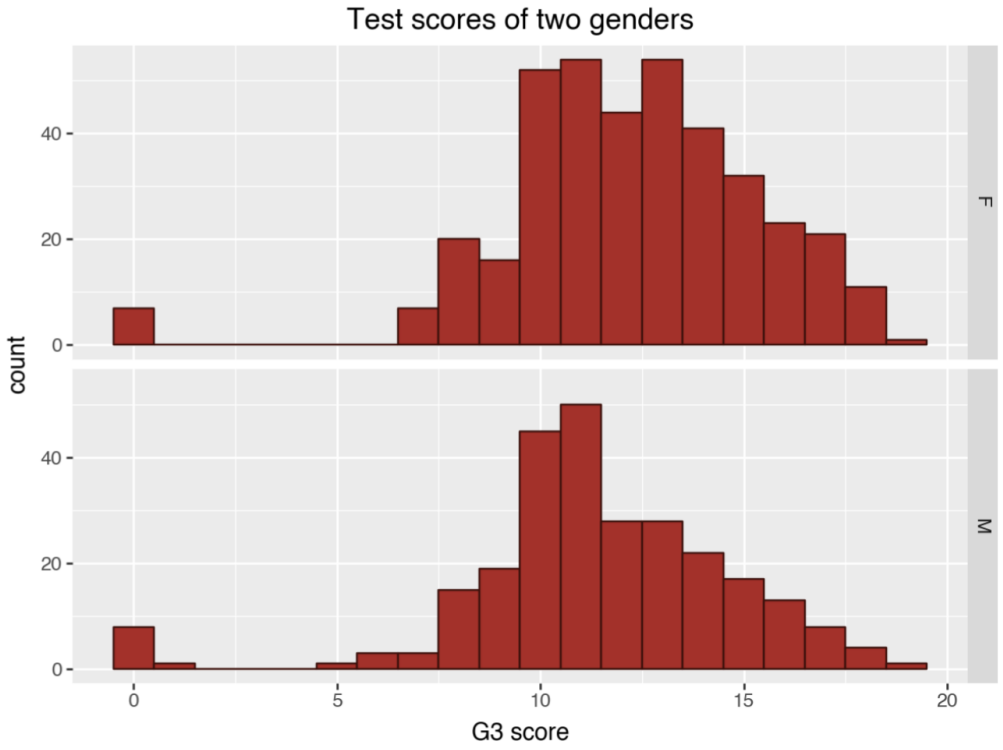

Figure 9: Histogram for G3 score grouped by genders

Welch's T-test was done with the null hypothesis that both the mean scores of female and male students are equal. The p-value was low compared to the significance level. The null hypothesis can be refuted due to this. There is a statistically significant difference between the mean score of female and male students.

### 3.5 Is there any correlation between health and days of absence on G3 score/grades?

A bar plot with sequential colouring was used to plot the number of students in each health status from 1-5. 1 represents very bad health and 5 represents very good health. The bar chart shows that most of the students have very good health.

The boxplot as shown below, the students with very bad health status (1) have a higher median of score compared to students with 5 statuses.

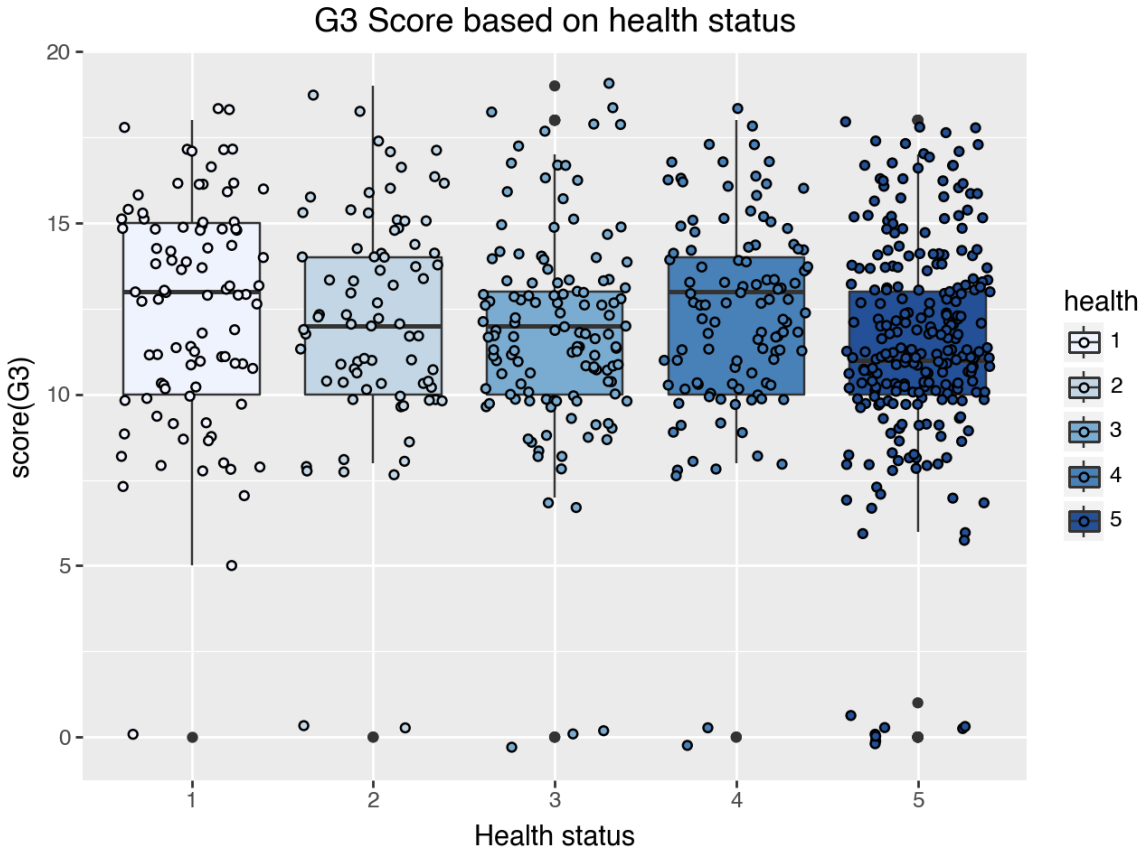


Figure 10: Box plot with G3 score grouped by health status

Unlike the general assumption that students with bad health struggle with studies, this graph does not show any impact of health on G3 score. The bar chart plotted with the number of days of absence and grouped by health status shows that the highest number of leaves were taken by students with very bad health status (1).

To see if the score has any linear relation with the number of days of absence a scatter plot was plotted, and we see that there is no linear correlation. The points tend to scatter around as seen in the graph below. The test score varies from around 0- above 15 for students with 0 leaves. This can be seen from the graph below. The correlation matrix also showed a very low negative relationship.

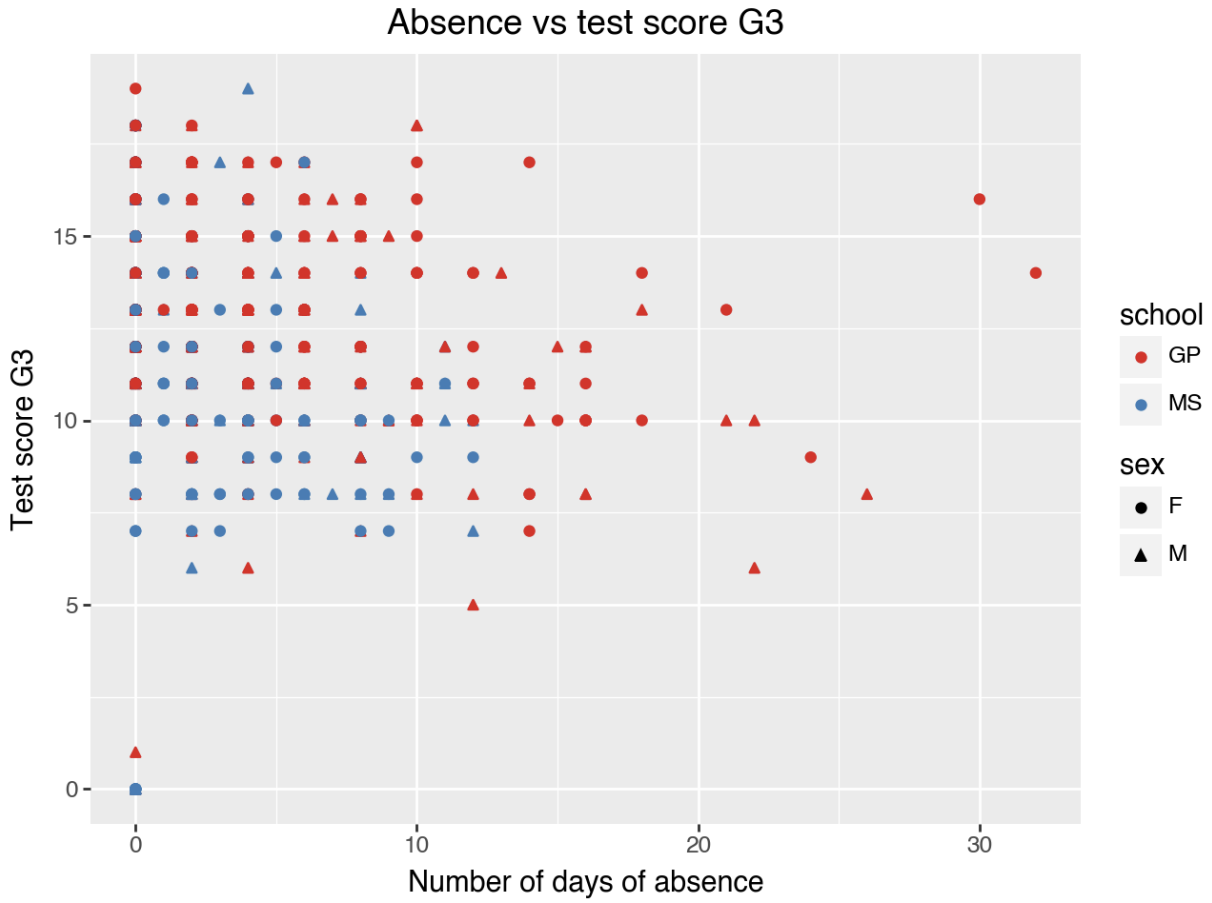


Figure 11: Scatter plot of number of days of absence and test score

A chi-square test was conducted to see if both the variables, grades and health are independent of each other. The null hypothesis is that grades and health are independent variables. The p-value we got is higher than the significance value so the null hypothesis cannot be refuted. Since the statistical score is also low, the data distribution is not very different from the expected data distribution for null hypothesis.

**3.6 Is there any correlation between Age, relationship and scores/grade?**

A histogram of all students in all ages is plotted. We can see that most of the students are minors in the age of 15-18. The rest of the students are in the age 19-22. Rather than comparing if a relationship has any correlation with score, we are comparing if being a minor or adult in a relationship or being single has any correlation to the G3 score.

A stacked graph with all the grades is plotted showing the number of students getting each grade. The relationship status and age are shown with the stacked graph.

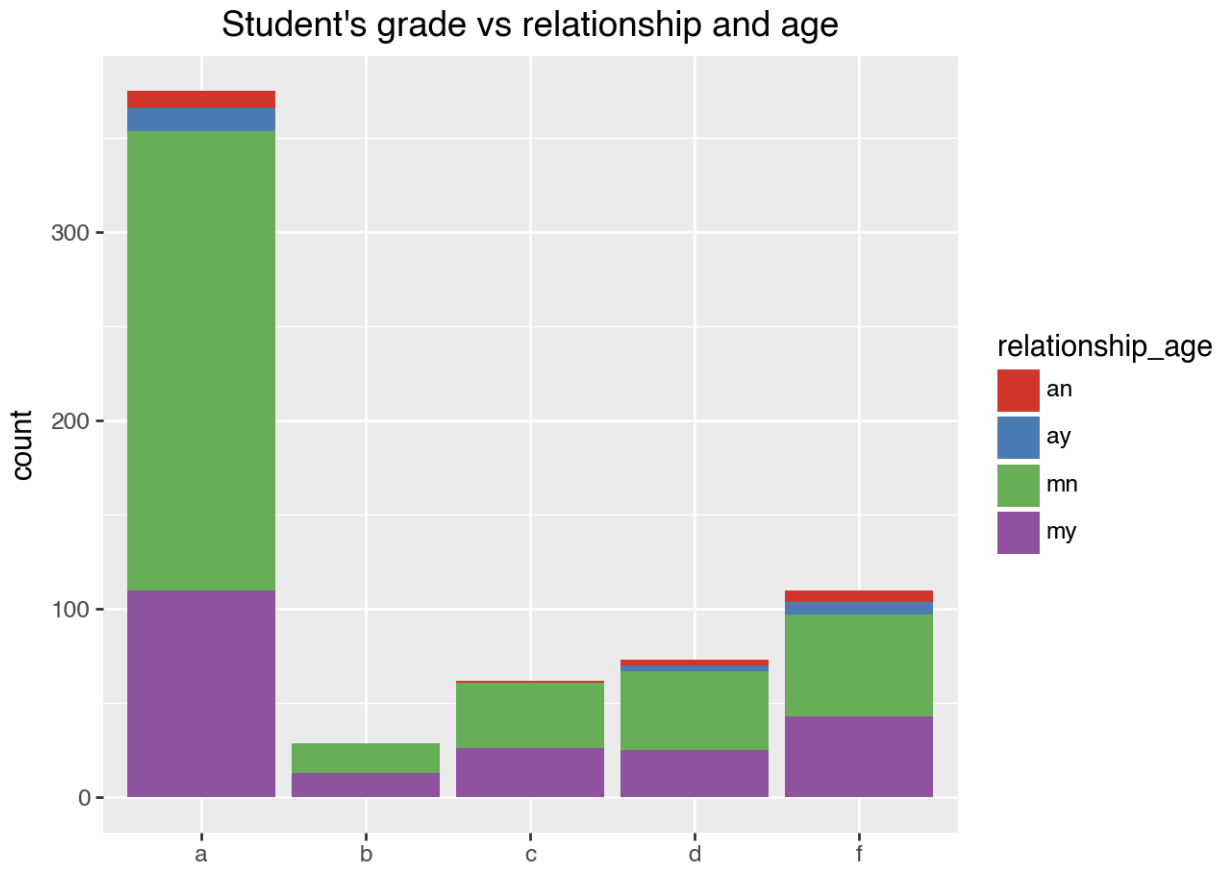


Figure 12: Stacked bar plot with student grade stacked by relationship status and age

From the graph, we can see that the maximum students who got “a” grade are minors who are not in a relationship. We can also see that the maximum number of adult students who have scored “a” grade are in a relationship. We cannot say being single results in better grades as more adults in a relationship have scored “a” grade, also the maximum number of students who score f grade are also minors with no relationship. The number of adult students is also less compared to minor students.

A box plot was also plotted, and it showed that the median of both the minors in a relationship and minors not in a relationship are very similar even though the highest scores seem to be scored by students who are not in a relationship. We cannot conclude on how the relationship and age together is related to the score or grade based on these graphs.

**3.7 Is there any correlation between support received from family and school, parent's education, job, family relation quality and scores/grades?**
A boxplot for both students receiving educational support from family and another graph for students receiving educational support from school is plotted, we see that more students receive support from the family.

Using a mosaic plot, the education of mother and educational support from family is compared and we see that most of the students with mother's who have higher education received educational support from home.

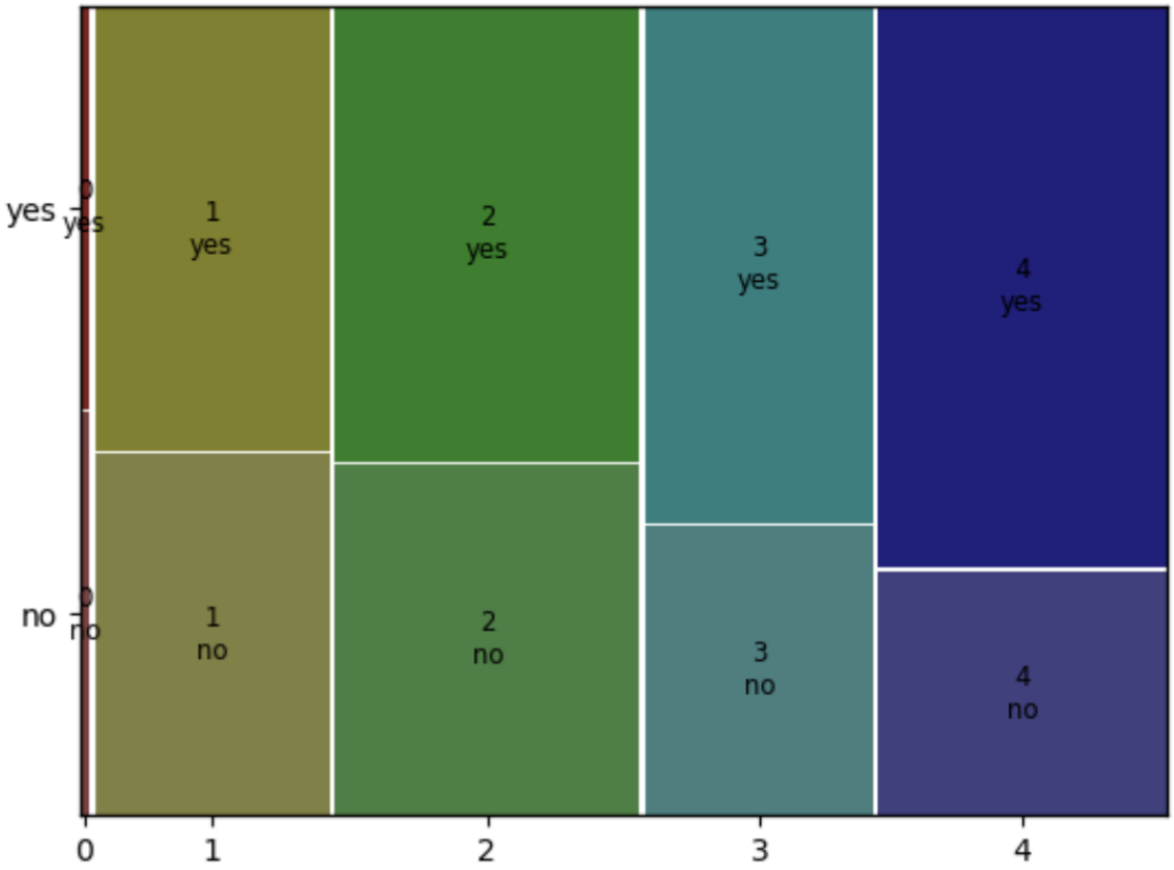

Figure 13: A mosaic plot for education of mother and educational support from family

A mosaic plot for father's education and educational support from family was plotted and it has similar results compared to the above mosaic plot. Many students with fathers who have higher education also received educational support from family.

The relationship between father's and mother's job to grade are plotted using two different mosaic plots and in both the plots, we see that the students with mother or father working as a teacher, a huge percentage of them tends to earn "a" grade but it is to be noted that most student's parents work in other sector or service sector.

Another mosaic plot for relationship between quality of family relation and grades was plotted. The best quality of relation is rated as 5 and very bad relation is marked as 1. we can see that more students had a good quality relation with their family with a 4 and 5 score. Very few students seem to fail from these two categories.

A new data frame was created for students with educational support from both school and family. A scatter plot was plotted for these students only. We can see from the below graph that there are only a few students who received educational support from both school and family. The red line indicates the passing score. There are four students who failed and there are only two students who scored 16 or above.

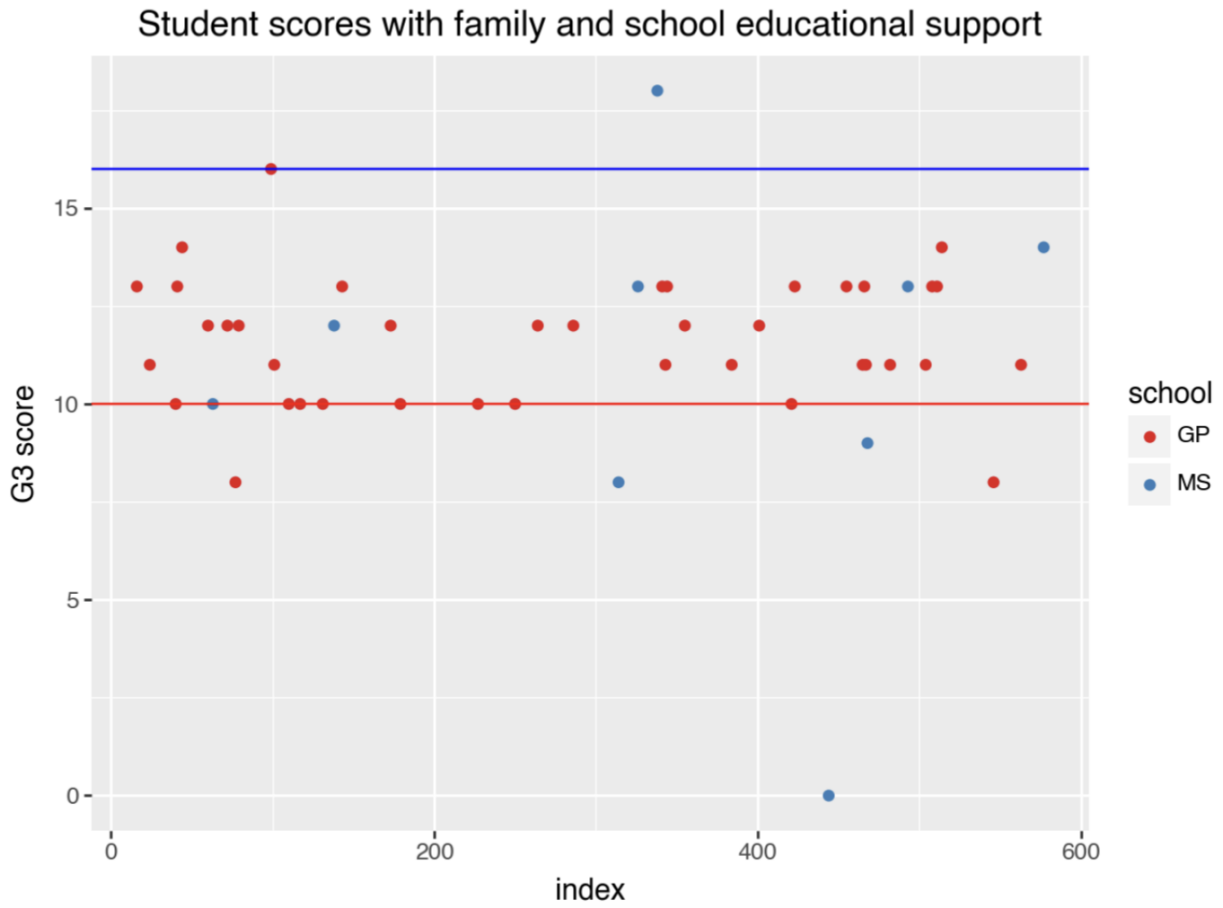


Figure 14: The G3 score scatter plot of students getting educational support from family and school

To check if the family/school support variable and grades are independent or not, a chi-square test was conducted. The null hypothesis is that the family educational support and school educational support are independent variables from grade. The p-value is greater than the significance level value, this means that we cannot reject the null hypothesis based on the data we have. The statistical value is also not high which shows that the data distribution is not very different from the expected distribution for the null hypothesis

distribution. This suggests that there is some association between family and school educational support.

## 4. Conclusions and Further developments

Even though we have found presence of correlations between several variables and grades/score in this dataset, we still cannot say that these are the causal factors of student grade/score because linear or non-linear correlation only lets us know there is a relationship but it is not necessary that it causes the target variable to change (Madhavan, 2019). We also need to acknowledge the presence of confounder variables in the dataset. To conclude on which factors can impact the student grades/scores, we need to make sure the confounders are removed.

There is a dataset of maths student grades of the same schools with lesser number of students, this can be used to compare the conclusions made from this dataset. It was not merged with the dataset and used for this project because the number of students in that dataset was less compared to the number of students in this one.

Since the dataset is small, it cannot be generalised easily to a larger population. A larger dataset needs to be used to come to a profound conclusion about relationships in data.

**References**


Chip (2023). *Density Plots vs Histograms: How Do They Compare?* [online] Quanthub.com. Available at: https://www.quanthub.com/how-do-histograms-compare-to-density-plots-in-terms-of-what-they-show/ [Accessed 21 Jul. 2024].

Cortez, P. and Silva, A. (n.d.). *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE*. [online] Available at: http://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf.

Daily, S.M., Smith, M.L., Lilly, C.L., Davidov, D.M., Mann, M.J. and Kristjansson, A.L. (2020). Using School Climate to Improve Attendance and Grades: Understanding the Importance of School Satisfaction Among Middle and High School Students. *Journal of school health*, [online] 90(9), pp.683–693. doi:https://doi.org/10.1111/josh.12929.

Giano, Z. (2022). *Grade Retention and School Dropout: Comparing Specific Grade Levels Across Childhood and Early Adolescence - Zachary Giano, Amanda L. Williams, Jennifer N. Becnel, 2022*. [online] The Journal of Early Adolescence. Available at: https://journals.sagepub.com/doi/full/10.1177/02724316211010332 [Accessed 20 Jul. 2024].

Lex Borghans, Bart, Heckman, J.J. and John Eric Humphries (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences of the United States of America*, [online] 113(47), pp.13354–13359. doi:https://doi.org/10.1073/pnas.1601135113.

Madhavan, A. (2019). *Correlation vs Causation: Understand the Difference for Your Product*. [online] Amplitude.com. Available at: https://amplitude.com/blog/causation-correlation# [Accessed 21 Jul. 2024].

Yousuf Al Husaini and Shukor, A. (2023). *Factors Affecting Students' Academic Performance: A review*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/367360842_Factors_Affecting_Students'_Academic_Performance_A_review [Accessed 20 Jul. 2024].