The American University in Cairo

Factors Affecting Secondary School Students’ Performance on a Mathematics Course

Hana Abdelhamid (900225555)

Janna El Kaffas (900231768)

DSCI 1411 Final Project

December 18th, 2023

1. **Introduction**

The dataset being analyzed shows the performance of students from two different schools in Portugal in a math class as well as various other variables which may impact their performance. By analyzing this data, questions surrounding the effect of these variables on students’ performance can be answered, more specifically:

1. Which type of support is most effective, family support, school support or paid support?
2. How do out of school variables, such as absence, study time and going out, impact performance?

Answering the above questions may inform the study habits and types of support students receive and utilize in order to improve their performance as effectively as possible.

1. **Data Description**

The dataset was retrieved from Kaggle from the below link:

(<https://www.kaggle.com/datasets/kukuroo3/student-performance-data-set-competition-form>)

The data was originally gathered by Cortez & Silva (2008) as cited below

Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito, & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12.

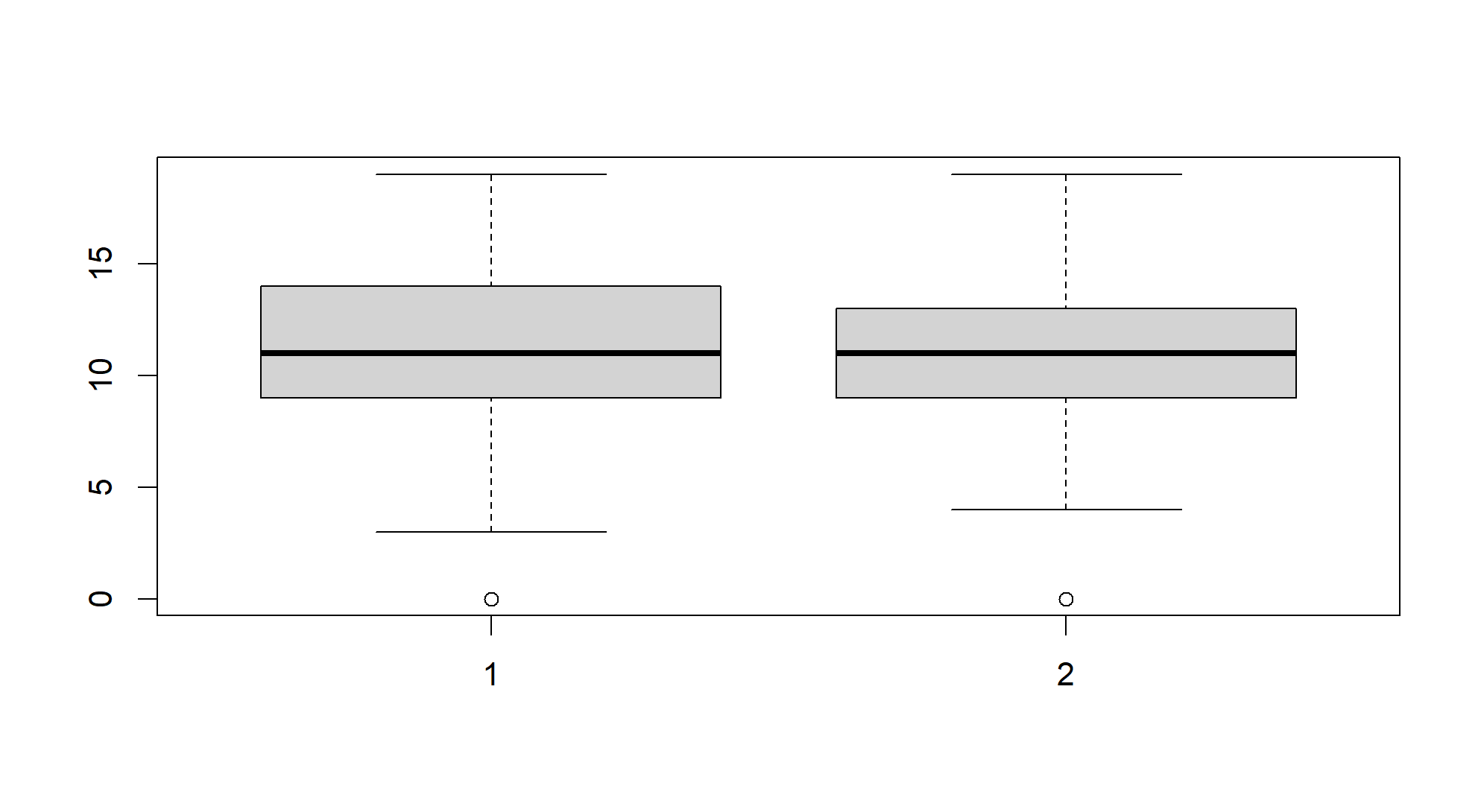
The dataset is composed of 366 rows and 33 columns. It shows secondary school students’ performance in a math course in two schools in Portugal along with several other variables which may impact their performance. Performance is measured through students’ grades in the first and second part of the academic year.

The columns are as follows:

|  |  |
| --- | --- |
| 1.StudentID - The students’ ID number (character) | 2. school - student's school (categorical: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira) |
| 3. sex - student's sex (categorical: "F" - female or "M" - male) | 4.age - student's age (numeric: from 15 to 22) unit of measurement: years old |
| 5. address - student's home address type (categorical: "U" - urban or "R" - rural) | 6. famsize - family size (categorical: "LE3" - less or equal to 3 or "GT3" - greater than 3) |
| 7. Pstatus - parent's cohabitation status (categorical: "T" - living together or "A" - apart) | 8. Medu - mother's education (categorical: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| 9. Fedu - father's education (categorical: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) | 10. Mjob - mother's job (categorical: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other") |
| 11. Fjob - father's job (categorical: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other") | 12. reason - reason to choose this school (categorical: close to "home", school "reputation", "course" preference or "other") |
| 13. guardian - student's guardian (categorical: "mother", "father" or "other") | 14. traveltime - home to school travel time (categorical: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| 15. studytime - weekly study time (categorical: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) | 16. failures - number of past class failures (numeric) |
| 17. schoolsup - extra educational support (logical: yes or no) | 18. famsup - family educational support (logical: yes or no) |
| 19. paid - extra paid classes within the course subject (logical: yes or no) | 20. activities - extra-curricular activities (logical: yes or no) |
| 21. nursery - attended nursery school (logical: yes or no) | 22. higher - wants to take higher education (logical: yes or no) |
| 23. internet - Internet access at home (logical: yes or no) | 24. romantic - with a romantic relationship (logical: yes or no) |
| 25. famrel - quality of family relationships (categorical: from 1 - very bad to 5 - excellent) | 26. freetime - free time after school (categorical: from 1 - very low to 5 - very high) |
| 27. goout - going out with friends (categorical: from 1 - very low to 5 - very high) | 28. Dalc - workday alcohol consumption (categorical: from 1 - very low to 5 - very high) |
| 29. Walc - weekend alcohol consumption (categorical: from 1 - very low to 5 - very high) | 30. health - current health status (categorical: from 1 - very bad to 5 - very good) |
| 31. absences - number of school absences (numeric; from 0 to 93 unit of measurement: number of absences in the academic year ) | 32. G1 - grade for academic period 1 (numeric; from 0-20 unit of measurement: points) |
| 33. G2 - grade for academic period 2 (numeric; from 0-20, unit of measurement : points) |  |

An additional column containing the sum of G1 and G2 was created during analysis. This column represented the total grade and was named GT.

1. **Analysis**

**Fig. 1 Box plot showing the spread of G1 and G2** 

To begin exploring the data set the above box plot was made to compare between performance in G1 and G2. The upper limit scores are identical in both exams. However, the lower limit score is lower in G1 than G2. The spread of the data changes between G1 and G2; G2 has a lower spread as illustrated by the smaller difference between quartile 3 and quartile 1. The median scores are identical. Both exams have outliers at 0. This suggests that scores improve between G1 and G2, as supported by the higher lower limit in G2.

The six number summary of the total grades (GT) was produced in order to act as a point of reference and comparison for the analysis that follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 4.00 | 18.00 | 22.50 | 22.59 | 27.00 | 37.00 |

Measures of center and spread were calculated for the total grade, which is the sum of G1 and G2.

1. Paid Support

|  |  |  |
| --- | --- | --- |
| *Measure of Center* | *Received paid support* | *Did not receive paid support* |
| *Mean* | 22.34 | 22.66 |
| *Median* | 21 | 23 |
| *Quantiles* | 0% 25% 50% 75% 100%  12 18 21 26.5 36 | 0% 25% 50% 75% 100%  4 18.5 23 27 37 |

**Table 1** **Measures of center for students receiving and not receiving paid support**

|  |  |  |
| --- | --- | --- |
| *Measure of Spread* | *Received paid support* | *Did not receive paid support* |
| *Range* | 24 | 33 |
| *Interquartile Range (IQR)* | 8.5 | 8.5 |
| *Mean Absolute Deviation* | 5.93 | 5.93 |
| *Variance* | 30.74 | 39.98 |
| *Standard deviation* | 6.32 | 5.54 |

**Table 2 Measures of spread for students receiving and not receiving paid support**

There is a 0.32 point difference between the mean scores of those who received paid support and those who did not. Moreover, the mean score of those who received paid support is lower than the overall mean of GT (as seen in the six number summary). At first glance this may suggest that paid support does not have a significant impact on the students’ grades. However, there is a difference in the spread of scores as evidenced by the higher variance and range in the scores of those who did not receive paid support. This suggests that while the means are similar, more of those who received paid support scored around the mean , i.e. scored higher, than those who did not receive paid support. This suggests that paid support is effective at preventing the occurrence of low scores, but not at increasing the occurrence of high scores. In addition, the lowest score scored by students who did not receive paid support was 4, while in those who did receive support it was 12.

1. School Support

|  |  |  |
| --- | --- | --- |
| *Measure of Center* | *Received school support* | *Did not receive school support* |
| *Mean* | 20.36 | 22.9 |
| *Median* | 20 | 23 |
| *Quantiles* | 0% 25% 50% 75% 100%  11 17 20 24 35 | 0% 25% 50% 75% 100%  4 19 23 27 37 |

**Table 3 Measures of center for students receiving and not receiving school support**

|  |  |  |
| --- | --- | --- |
| *Measure of Spread* | *Received school support* | *Did not receive school support* |
| *Range* | 24 | 33 |
| *Interquartile Range (IQR)* | 7 | 8 |
| *Mean Absolute Deviation* | 4.45 | 5.93 |
| *Variance* | 26.46 | 38.81 |
| *Standard deviation* | 5.14 | 6.23 |

**Table 4 Measures of spread for students receiving and not receiving school support**

There is a difference between the measures of the center of those who received school support and those who did not. Those who did not receive school support had a higher mean score, 22.9, than the 20.36 mean of those who received it. The median score is also higher in those who did not receive support than those who did, 23 and 20 respectively. Both the mean and median score being higher in those who did not receive school support is an essential indicator as it confirms that the higher mean is not a result of the presence of outliers since the median is not sensitive to outliers. This suggests that school support is ineffective at improving students’ performance. Nevertheless, the variance for the students who received school support is less than those who did not, indicating that more students score around the mean, meaning that not as many low grades are achieved. So the same conclusion drawn about the paid support can be said here.

1. Family Support

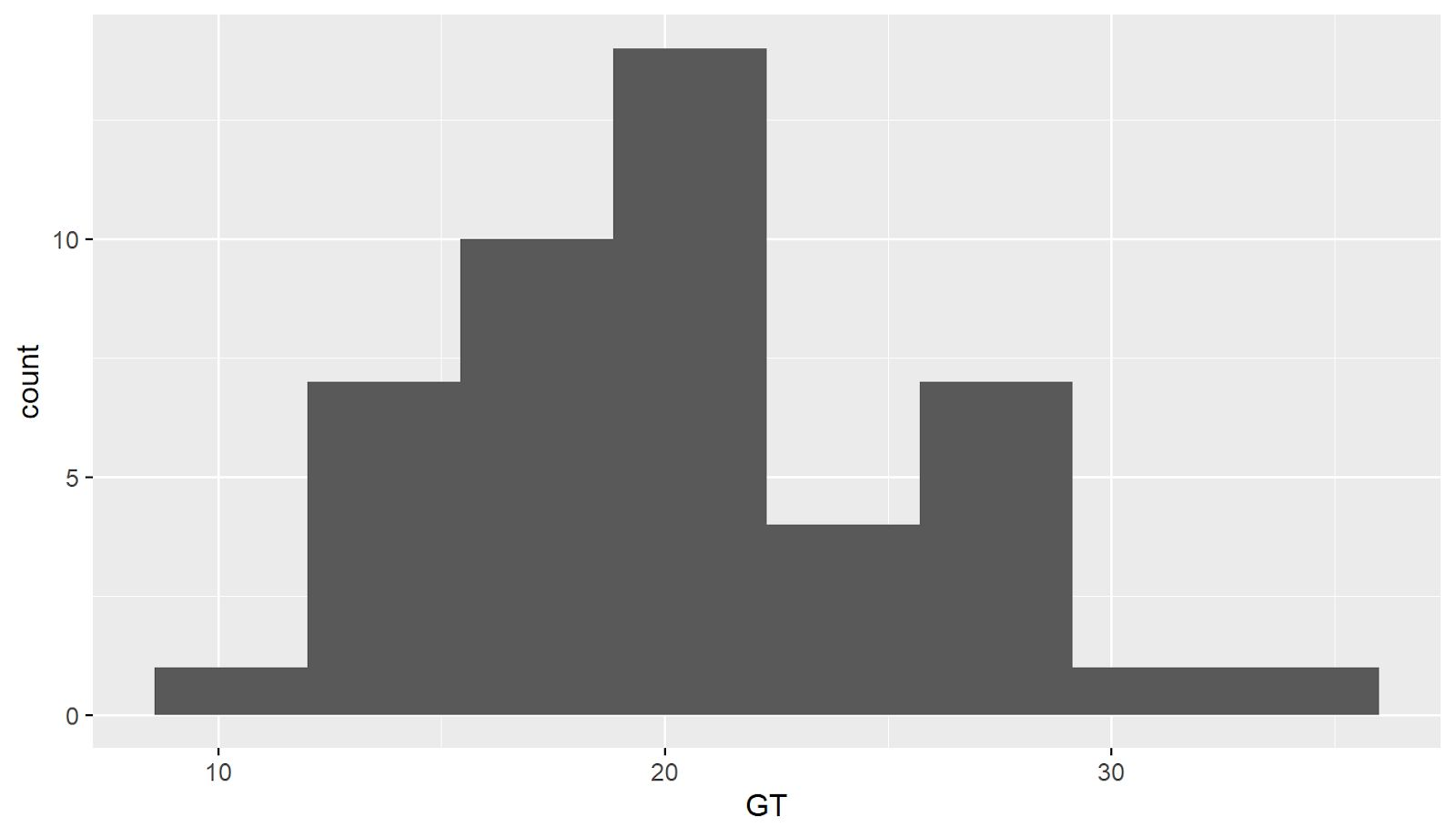
|  |  |  |
| --- | --- | --- |
| *Measure of Center* | *Received family support* | *Did not receive family support* |
| *Mean* | 22.45 | 22.8 |
| *Median* | 22 | 23 |
| *Quantiles* | 0% 25% 50% 75% 100%  4 19 22 27 35 | 0% 25% 50% 75% 100%  4 18 23 27 37 |

**Table 5 Measures of center for students receiving and not receiving family support**

|  |  |  |
| --- | --- | --- |
| *Measure of Spread* | *Received family support* | *Did not receive family support* |
| *Range* | 31 | 33 |
| *Interquartile Range (IQR)* | 8 | 9 |
| *Mean Absolute Deviation* | 5.93 | 5.93 |
| *Variance* | 36.29 | 40.58 |
| *Standard deviation* | 6.02 | 6.37 |

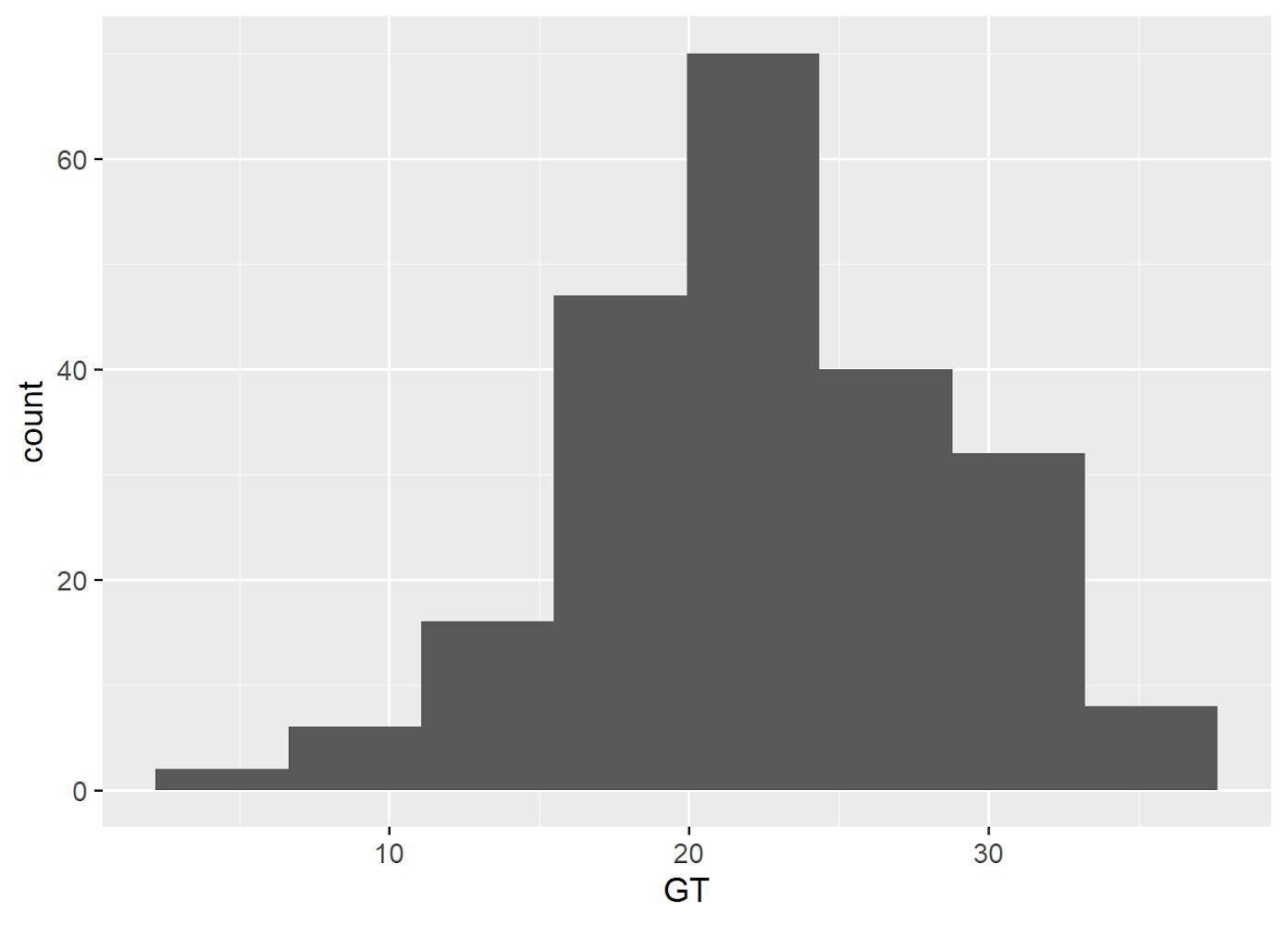
**Table 6 Measures of spread for students receiving and not receiving family support**

The difference in the measures of center of those who received family support and those who did not is small, with the mean and median of those who did not receive family support being slightly higher than those who received family support as well as slightly higher than the overall mean and median of GT. All measures of spread, except for variance are also similar. Compared to the measures of spread of the previous 2 types, the difference between the spread of the data of students who received family support and those who did not is the slightest, highlighting the fact that this third type of support does not affect the students’ grade as much as the other types of support do. For instance, the measure of spread with the highest difference is the variance at a difference of only 4.29 between those who received family support and those who did not. Meanwhile, the difference in the variance between those who received school support and those who did not is 12.35, more than double the aforementioned value. Overall, family support appears to be the least effective type of support and paid support and school support are effective at preventing low scores and not at increasing scores. To get further insight into the differences in effectiveness between paid support and school support histograms of students’ total grades when receiving each type of support were constructed.



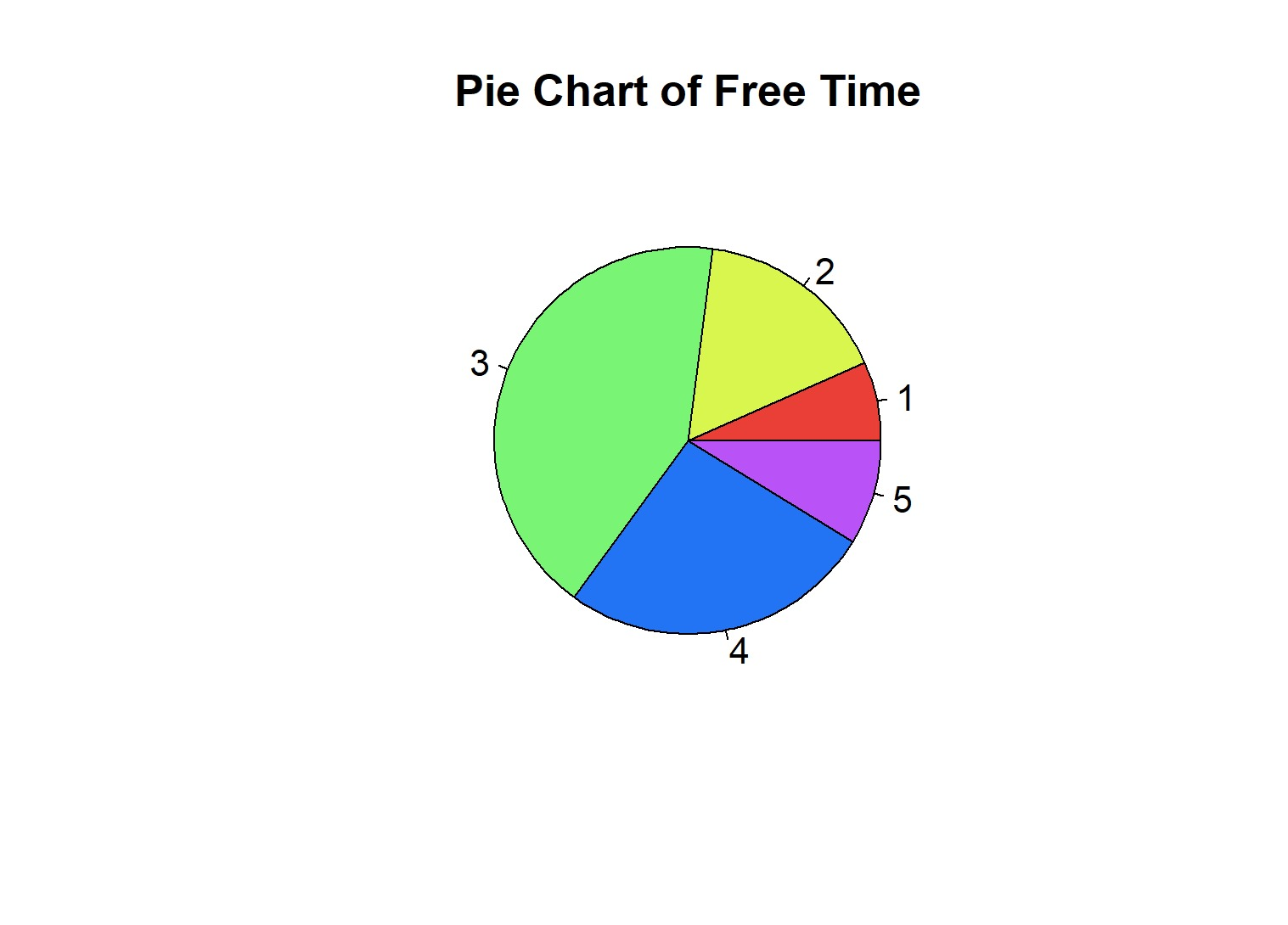
**Fig.2 Histogram of total grades of students receiving school support**

In Figure 1, the x-axis displays the student's total grade divided into classes of class width 5 . On the y-axis is the frequency, or the number of students who obtained a score within each class. It is clear that the histogram is unimodal as there are no bars of equal height. The modal class is (20-25] at a frequency of 14. Additionally, the classes which the least number of students belong to are [0,10], the lowest class, and (35-40], the highest class.

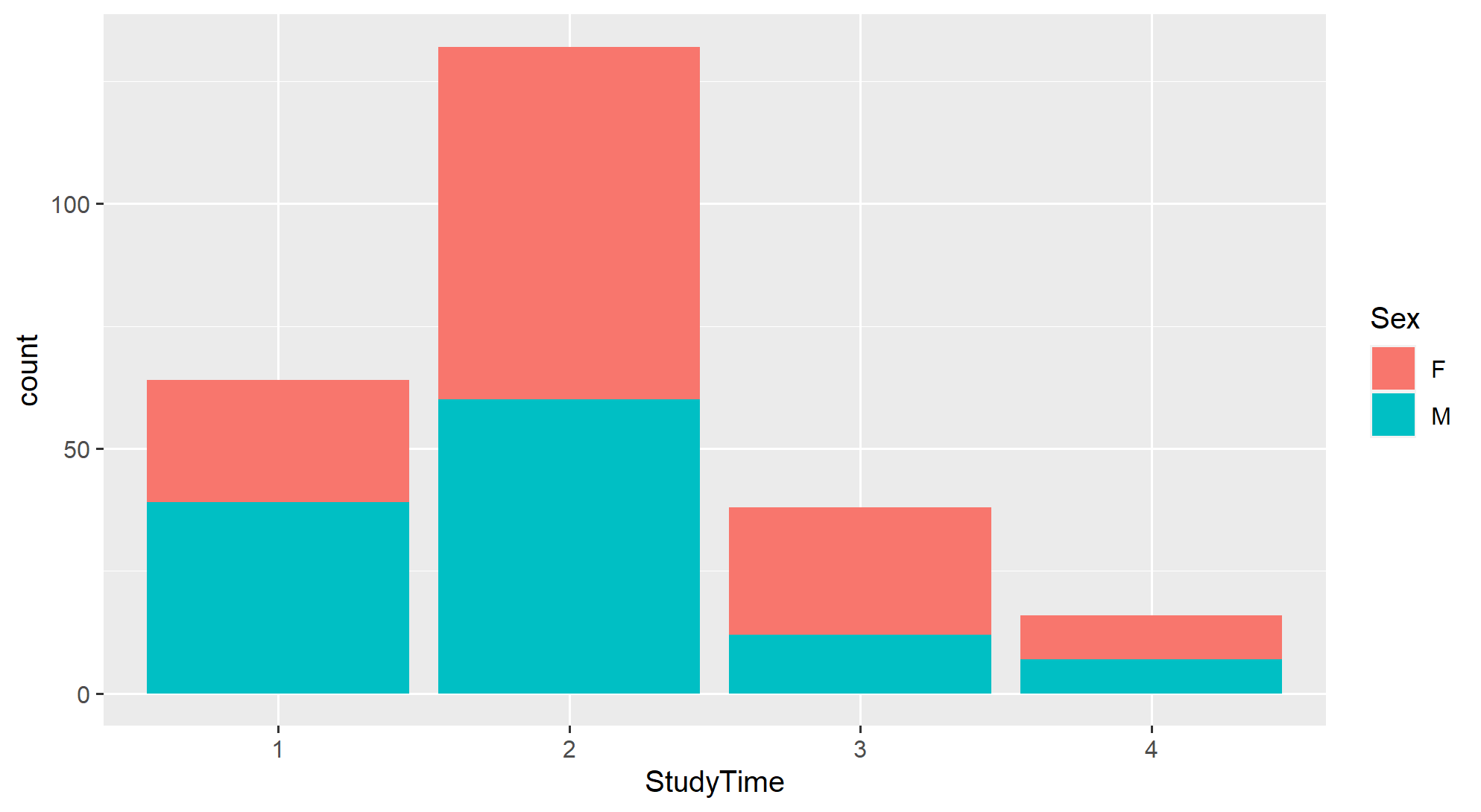
**Fig. 3 Histogram of total grades of students receiving paid support** 

This histogram demonstrates the spread of the total grade of the students who received paid support. On the x-axis, the student’s total grade is split into intervals of 10. On the y-axis, the count or frequency for each of these intervals is shown. Similarly to the previous figure, this histogram is unimodal with the highest frequency being 70 at the modal class of (20-35]. When figure 1 and 2 are compared, a conclusion can be drawn about the effectiveness of both support types. Since most data in figure 2 is concentrated around the right side of the histogram (classes where score is higher than 20), while in figure 1, most data is concentrated around the left side (classes where score is lower than 20), it can be argued that paid support is a more effective method in comparison to family support.

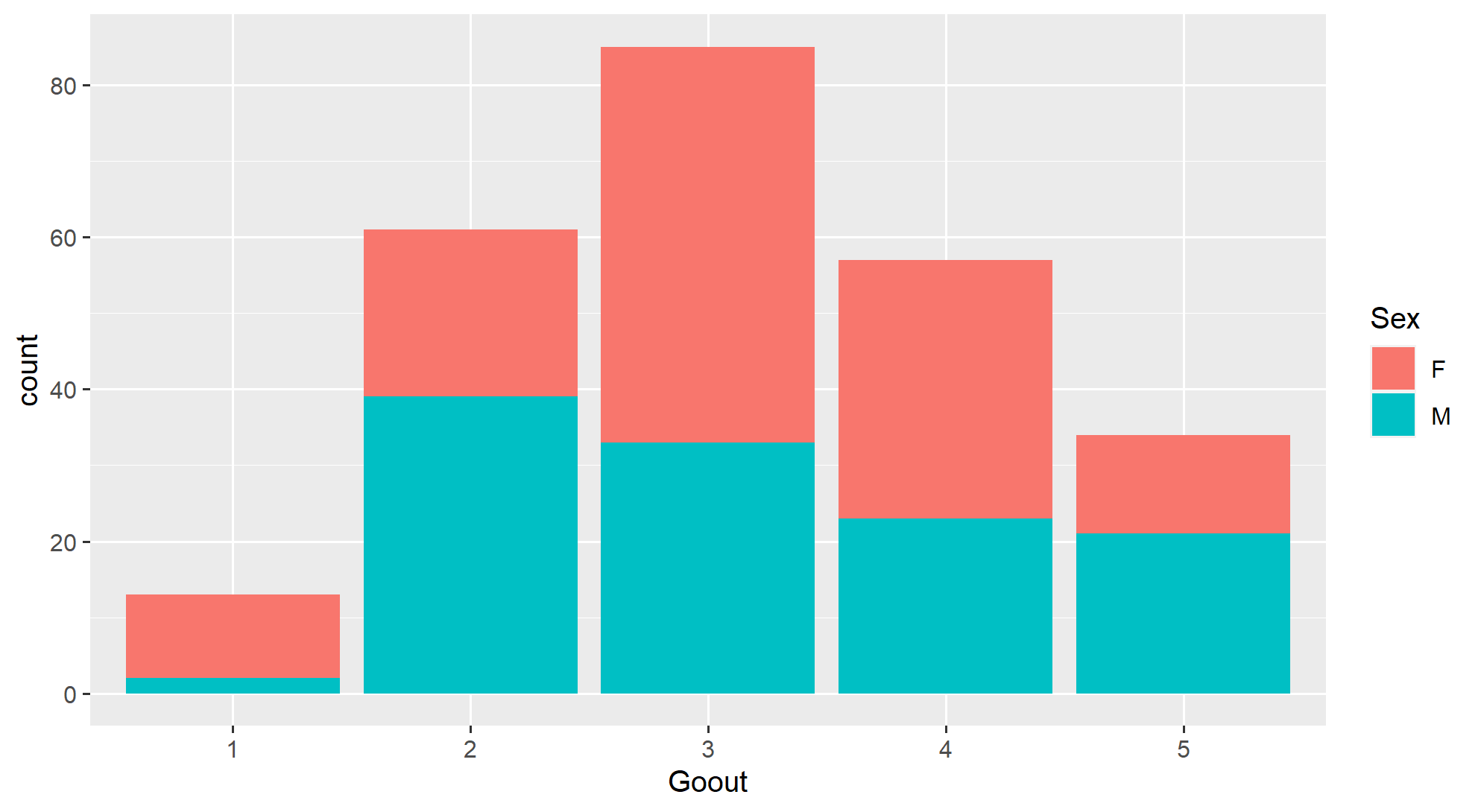
After establishing the impact of the different types of support on students’ performance, the impact of how they spend their time was investigated, this included investigating the amount of free time they have, as well as the relationship between the time spent studying and going out with their total grades.



**Fig. 4 Pie chart showing how much free time students have**

Figure 4 was constructed using the frequency table attached in the appendix. Its purpose is to explore the time available to students. It shows that most students belong to category 3, i.e. have moderate amounts of free time. While a minority of students belong to category 1 and category 5, i.e. have very low amount of free time or very high amount respectively. 

**Fig. 5 Bar chart showing study time of students who passed, filled by sex**

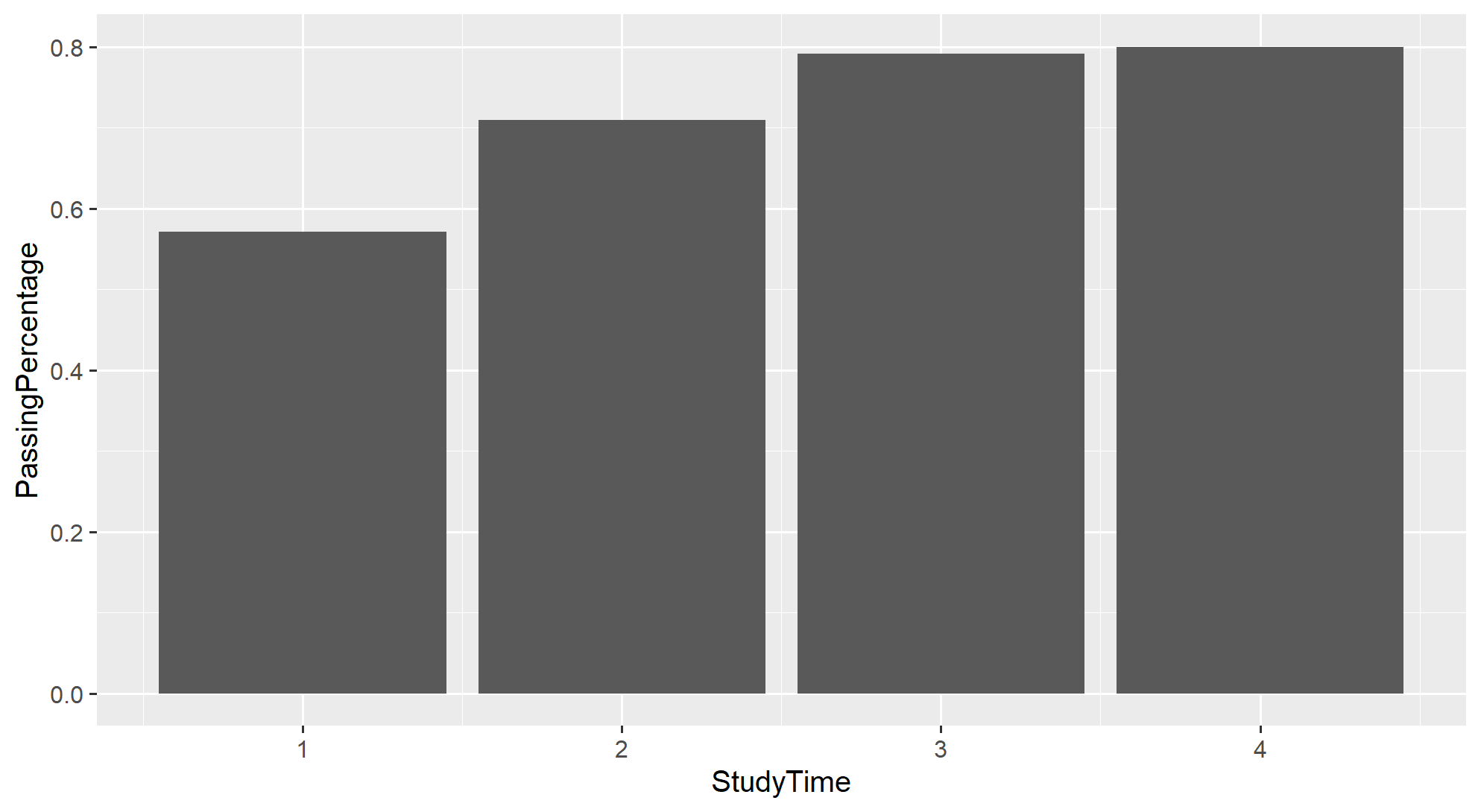


**Fig. 6 Bar chart showing going out time of students who passed, filled by sex**

Figure 5 and 6 are bar charts comparing time spent on studying and frequency of passing, and going out with friends and frequency of passing respectively. Upon constructing the charts it was noted that they fail at giving insights into the relationship time spent studying and going out have with the likelihood of passing. This is because the graphs reflect the number of people in each category. For example, category 4 for study time will always show the lowest frequency of passing because there are less students in this category overall. This does not mean that being in category 4 decreases likelihood of passing.

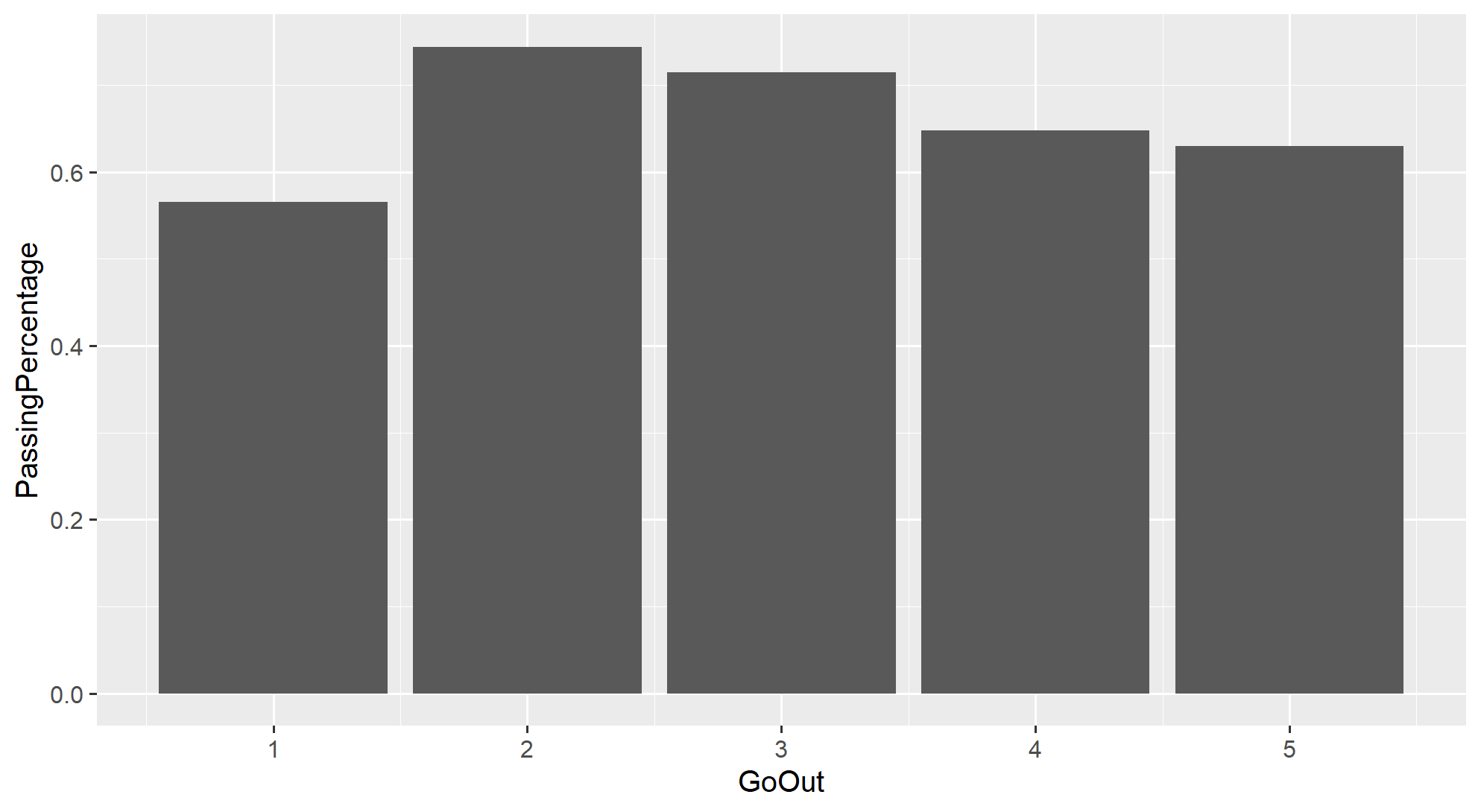
However, these charts give insight into gender demographics. Figure 5 shows that those in the lowest study time category (1) and pass are more likely to be male than female. As the time spent studying increases, students belonging to each category and achieving a passing grade are more likely to be female than male. In category 4, the highest category they are equally likely to be male or female. These differences reflect the difference in the number of male and female students in each category as supported by table 1 and table 2, rather than reflecting each gender’s ability to study for a certain number of hours, or go out for a certain number of hours as well as pass. This suggests that there is a difference between the habits of males and females which should be taken into account when attempting to improve performance.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | F | M | | 1 | 40 | 72 | | 2 | 109 | 77 | | 3 | 36 | 12 | | 4 | 12 | 8 |   **Table 1 Contingency table of study time and sex** | |  |  |  | | --- | --- | --- | |  | F | M | | 1 | 20 | 3 | | 2 | 38 | 44 | | 3 | 74 | 45 | | 4 | 49 | 39 | | 5 | 16 | 38 |   **Table 2 Contingency table of going out time and sex** |

In order to compensate for figure 5 and 6 not allowing for comparisons between the likelihood of belonging to each category and passing being made, bar charts were made of study time and percentage of students who passed (Fig.7), as well as going out and percentage of people who passed (Fig.7). 

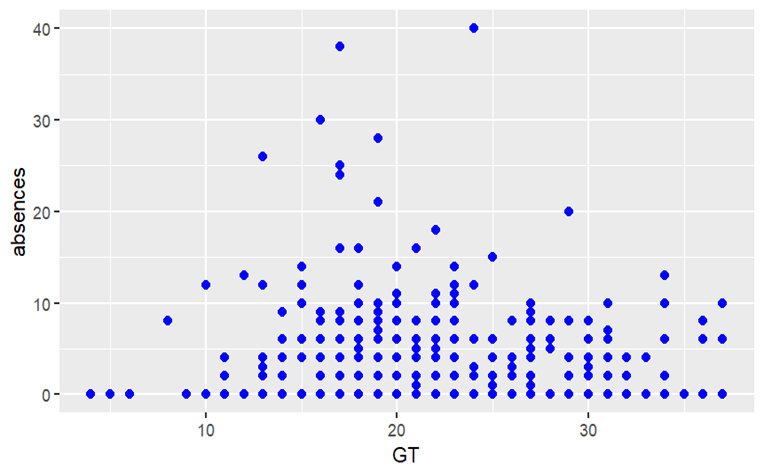
**Fig. 7 Bar chart showing the percentage of students who passed from each category of study time**

In figure 7, the highest percentage of passing is in the maximum category of study time while the lowest percentage of students passing is in the minimum category of study time. A trend is evident where the higher the number of hours spent studying the higher the percentage of students who passed. The two variables are positively correlated.

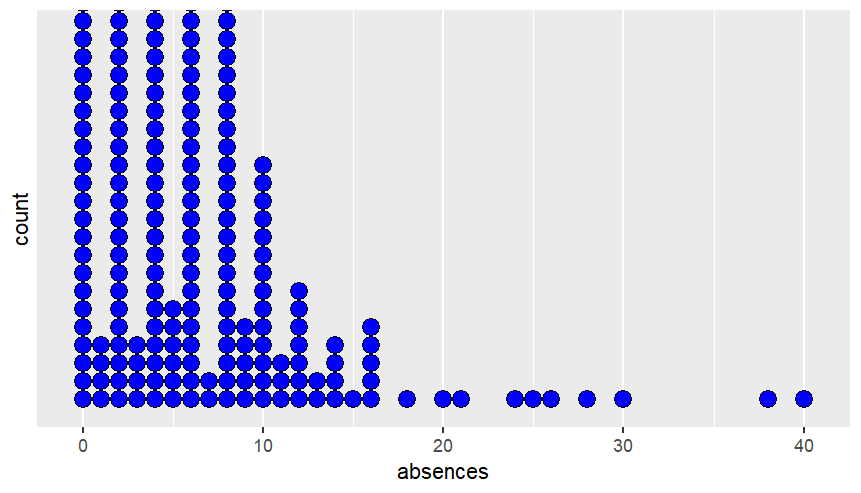


**Fig. 8 Bar chart showing the percentage of students who passed from each category of going out**

Figure 8 displays an inverse relationship between going out and the percentage of people who pass, from category 2 to category 6. Category 1 is an exception as these are the students who go out with friends the least, however they have the lowest percentage of students who passed.

Another factor which may impact grades is attendance. To gain insight into the relationship between attendance and performance a scatter plot of absences and grade total was constructed. 

**Fig. 9 Scatter plot showing the relationship between number of absences and total score**

This scatter plot shows the student’s total grade, the independent variable, on the x-axis. The dependent variable is the absences on the y-axis. The y-axis shows the number of absences, the dependent variable. The scatter plot shows no linear relationship between these two variables, indicating that absences do not have an impact on the student’s academic performance. 

**Fig. 10 Dot plot showing the spread of the frequency of absences**

A possible explanation for absences not affecting grades is that absences were few. Figure 10 demonstrates that the majority of students actually do not exceed 10 absences, which one can consider a relatively small number. This may explain the lack of the correlation between absences and the students’ total grades in the scatter plot.

1. **Conclusion**

This investigation set out to answer the following questions:

1. Which type of support is most effective, family support, school support or paid support?
2. How do out of school variables, such as absence, study time and going out, impact performance?

With regard to the first question, the conclusion reached is that family support is the least effective, while school support and paid support are more effective. More specifically, school support and paid support are effective at preventing the occurrence of low scores, not at increasing scores. Through the histograms, it was determined that paid support is more effective at reducing low scores than school support. As for the second question, absence was found to have no correlation with the student’s academic performance. Meanwhile, time spent on studying was found to have a positive correlation with student performance and going out was negatively correlated with student performance.

**Appendix:**

Frequency Table of Free Time

level freq perc cumfreq cumperc

1 1 24 6.6% 24 6.6%

2 2 60 16.4% 84 23.0%

3 3 154 42.1% 238 65.0%

4 4 96 26.2% 334 91.3%

5 5 32 8.7% 366 100.0%