Exploring Key Factors Influencing Student Performance

CSC1143 Visualisation Assignment

Thang-Long Nguyen-Ho

School of Computing

Dublin City University

29 Nov 2024

Abstract

Student performance is a main problem in education, with much debate about its factors. Although time management is often considered a key factor, do personal, family, and social factors play a larger role? This work focuses on analyzing secondary school student data to explore the factors that influence final grades. In particular, the study examines the correlation between personal factors (such as social, study time, and family relationships) and academic performance, in order to find commonalities among excellent students. Visualization of data provides further insight into the factors needed to improve academic performance.

1 Dataset

The UC Irvine Machine Learning Repository is a well-regarded and diverse data repository that I selected for my research. The Student Performance dataset, specifically, focused on the academic performance of high school students in two schools located in Portugal. The dataset encompasses various attributes, including academic grades, demographic information, social and school-related characteristics. The data was collected through school reports and questionnaires, resulting in two distinct datasets for two subjects: Mathematics and Portuguese Cortez 2008.

The choice of this dataset was driven by its specificity. With 649 samples and 30 attributes, the data provides sufficient scope for exploring correlations and gaining insights. The relatively small-scale nature of the analysis is advantageous due to its time-saving and manageable complexity compared to analyzing longitudinal data that necessitates extensive datasets and substantial resources. Additionally, the scarcity of student data due to security concerns and the intricacies of collection further underscores the suitability of this dataset. Given its size and characteristics, it effectively demonstrates relationships and addresses the research questions posed.

2 Data Exploration, Processing, Cleaning and/or Integration

To prepare a dataset for creating charts or graphs, the first step is data cleaning and preprocessing. To understand the data, I prefer to examine the details of the features listed in Table 1. Specifically, the final grade ($G3$) needs to be normalized to classify students into three groups corresponding to increasing levels of academic performance. This process involves calculating the maximum $G3_{max}$ and minimum $G3_{min}$ values of $G3$, then dividing the range of scores into three equal intervals. The classification of groups is defined as follows:

- Excellent Performance: $G3 \geq G3_{min} + \frac{2}{3}(G3_{max} - G3_{min})$
- Good Performance: $G3_{min} + \frac{1}{3}(G3_{max} - G3_{min}) > G3 \geq G3_{min} + \frac{2}{3}(G3_{max} - G3_{min})$
- Poor Performance: $G3 < G3_{min} + \frac{1}{3}(G3_{max} - G3_{min})$

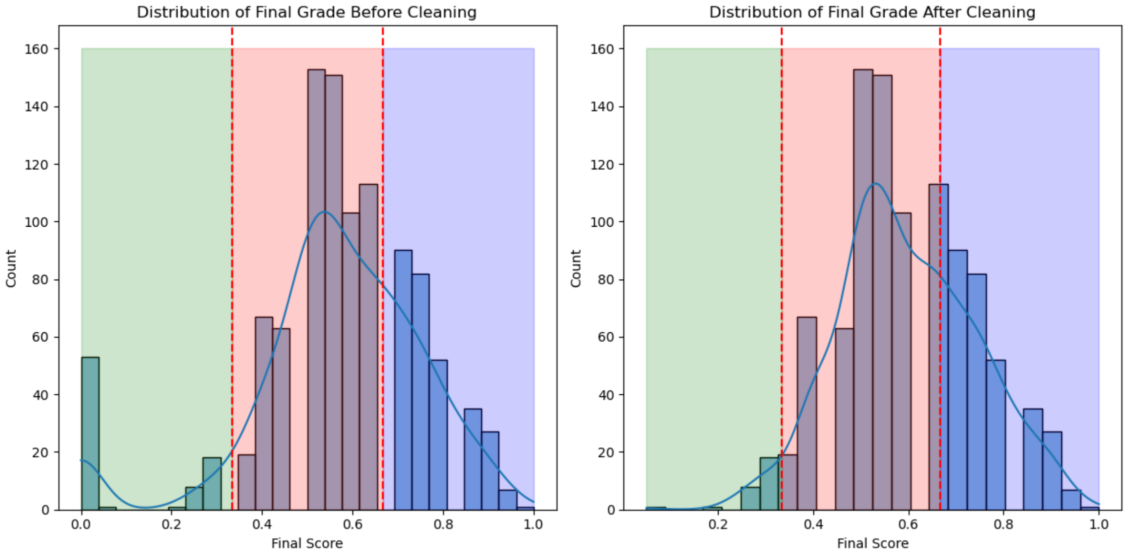

Figure 1: After performing a simple clustering method by dividing the data into three equal groups, I identified the outliers, which are the students who are registered but not studying. These students cannot be classified because they are noise data. Therefore, I removed them before proceeding with the next processing steps.

Standardization facilitate the classification of students based on their final score, establishing a foundation for further analysis. Figure 1 presents my results after dividing the students into three groups. In the subsequent analysis steps, I will not utilize the score but rather categorize the students to draw conclusions. This approach simplifies the reasoning process, enabling the identification of more significant features.

When feature engineering for understanding a dataset, we decide to focus on factors that influence student learning outcomes. Time usage factors, such as free time, study time, or absences, can directly impact learning outcomes. Family factors, including family relationships and parental education, also play a significant role in shaping the learning environment. Additionally, lifestyle factors, such as romantic relationships, time spent socializing with friends, and extracurricular activities, can also influence student learning. Finally, financial conditions also influence learning, including access to additional educational support and paid classes within the course subject. These factors contribute to a comprehensive understanding of the factors that influence learning outcomes and will be utilized for analysis and visualization purposes.

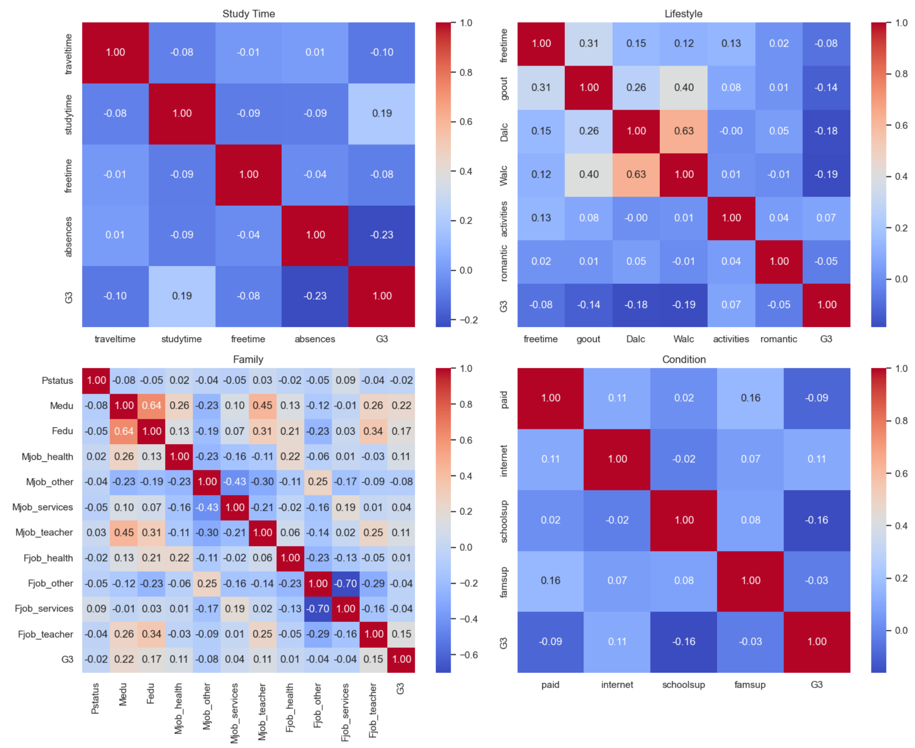

Figure 2: Four correlation matrices are presented to represent the study time, lifestyle, family support, and self-financial condition, respectively.

The details of feature selection are presented in Figure 2 as described above. After dividing the data into four groups, we can easily obtain the correlation matrix between the four perspectives, which reveals the characteristics influencing the final score. Some notable insights are: The learning ability of certain parents is directly proportional to the performance of their students (Figure 5). This can be understood because the family creates conducive learning environments for students (Figure 7). Additionally, alcohol consumption affects the results, as the data shows that the number of excellent students is concentrated in the group that consumes less alcohol (Figure 6). Finally and importantly, self-awareness plays a significant role. Figure 8 demonstrates that excellent students have a higher percentage of time dedicated to studying and fewer absences from class.

3 Visualisation

To address the question ”Which factors are affecting academic performance” we identified four distinct characteristics in each group. Subsequently, we merged these characteristics into the final visualization representation, arranged in the following order: time usage, lifestyle, family background, and personal circumstances. The rationale behind this type of chart is that it not only demonstrates the impact of these factors on academic outcomes but also provides insights into the relationships between them.

A slight different from the correlation matrix, the upper triangular data section of the matrix is redundant, as it merely repeats information. Additionally, we must arrange the groups corresponding to the columns in a sequential manner to facilitate user recognition and comparison of the significance between groups and within groups, ultimately for the final score. Figure 3 represents our preliminary draft before implementing the final version.

The figure 4 is a complete visualization created using libraries seaborn and matplotlib. The masked portion above has been removed to make the image sparse and more readable.

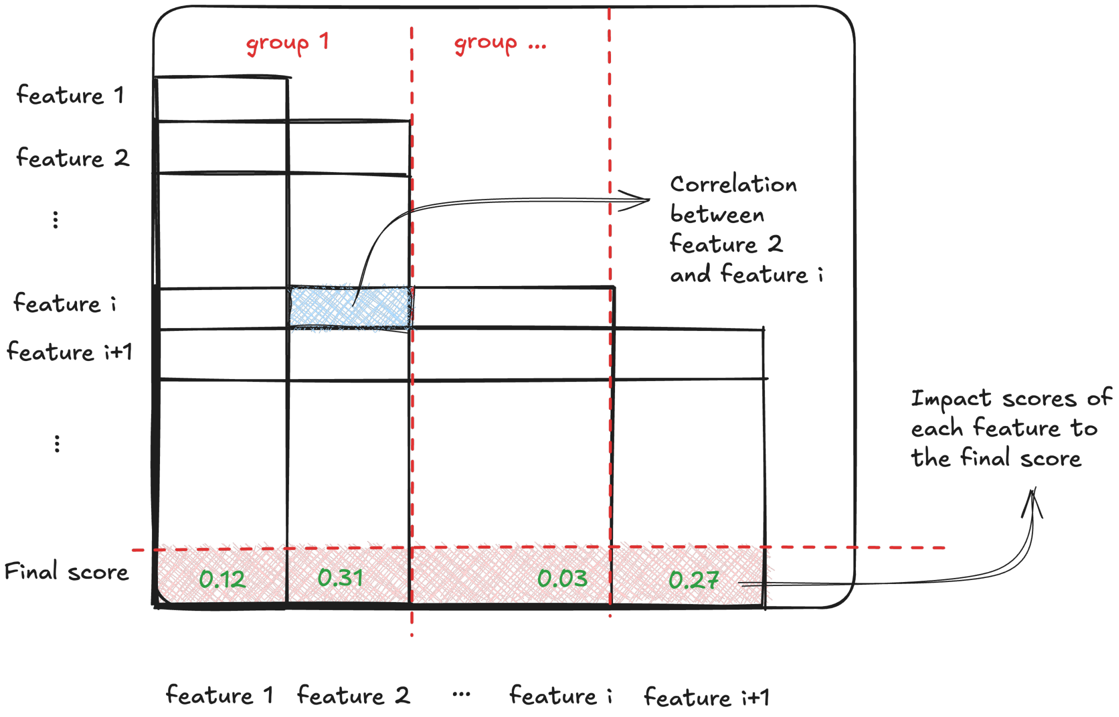

Figure 3: Sketch version of the visualization, which focuses on the correlation between features and the score, between features and other features, and between groups and other groups.

There are some unexpected outcomes we can observe are that individuals with romantic feelings tend to allocate less time to studying. While this may slightly decrease their average emotional factor, it doesn't have a significant impact on their final scores. Extracurricular activities, on the other hand, can help increase a student's score by 0.07. It's important to note that while studying plays a crucial role, exceptional students often manage to strike a balance between studying and participating in other activities. Additionally, families with parents working in the education field tend to provide internet access, extra classes for their children, and these characteristics also contribute to improved learning outcomes.

4 Conclusion

Student academic performance is influenced by various factors beyond just study time. Family dynamics, parental education, and work-life balance all play a crucial role. To enhance the student results, educators should create conducive environ-

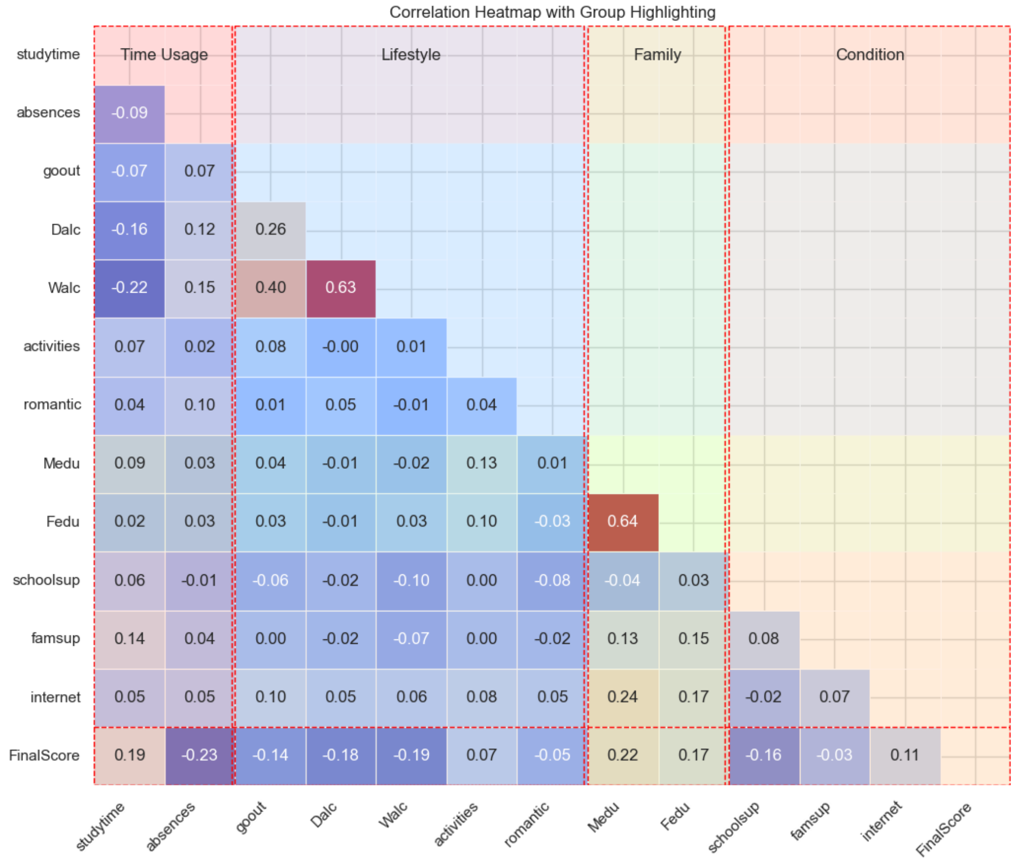

Figure 4: The final visualisation

ments that encourage students to participate in extracurricular activities and foster a supportive family atmosphere conducive to learning.

The analysis dive into a comprehensive exploration of personal, family, and social factors. Practical solutions are proposed to improve academic performance, emphasizing the importance of balancing study and personal life. The presentation employs easy-to-understand visualizations, such as correlation maps, to effectively connect data with significant findings.

The work was supported by references from Stack Overflow, Quora, official Matplotlib documentation, and lectures on visualisation course.

References

Cortez, Paulo (2008). *Student Performance*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TG7T.

Little, Suzanne (2024). *Data Management and Visualisation*. Lecture Notes. Details not specified.

Team, The Matplotlib Development (2024). *Matplotlib: Visualization with Python*. `https://matplotlib.org/stable/index.html`. Accessed: 2024-11-29.

Waskom, Michael and the Seaborn Development Team (2024). *Seaborn Example: Spreadsheet Heatmap*. `https://seaborn.pydata.org/examples/spreadsheet_heatmap.html`. Accessed: 2024-11-29.

Appendix

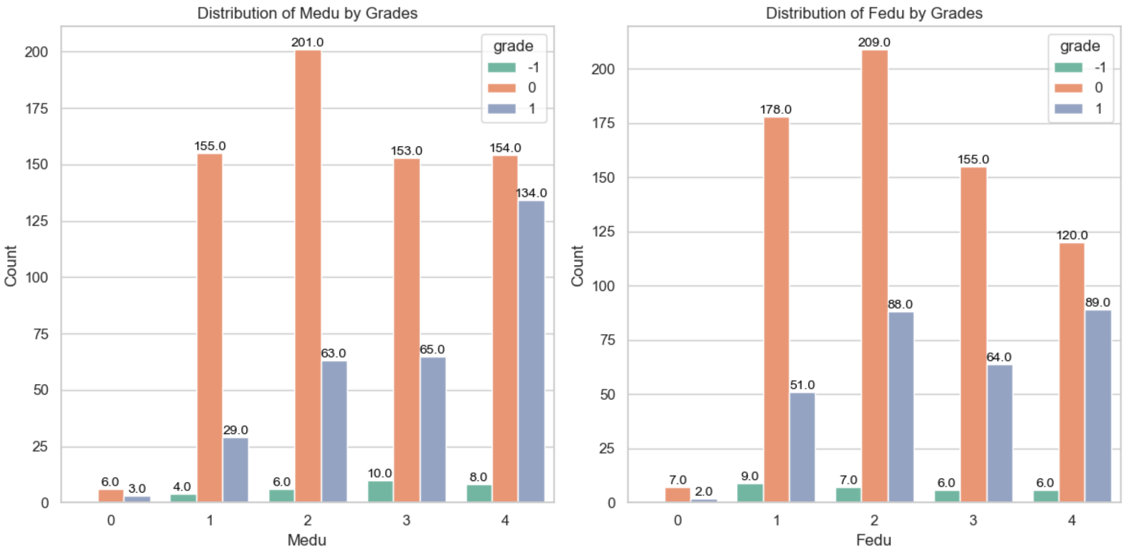

Figure 5: The graph illustrates the varying levels of student performance, categorized as excellent, good, and poor, based on the educational background of both parents. As the graph progresses from left to right, indicating a higher level of parental education, it becomes evident that this correlation leads to improved student performance through various mechanisms.

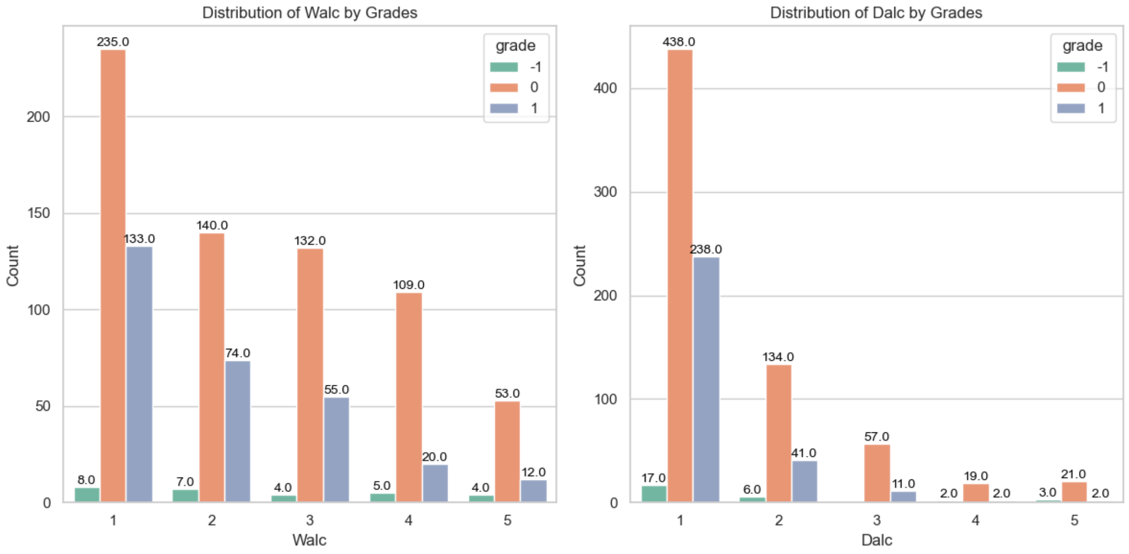

Figure 6: The graph presents varying levels of student performance, categorized as excellent, good, and poor, based on alcohol consumption. Two graphs correspond to weekends and workdays. The results indicate that students with good intentions consume less alcohol.

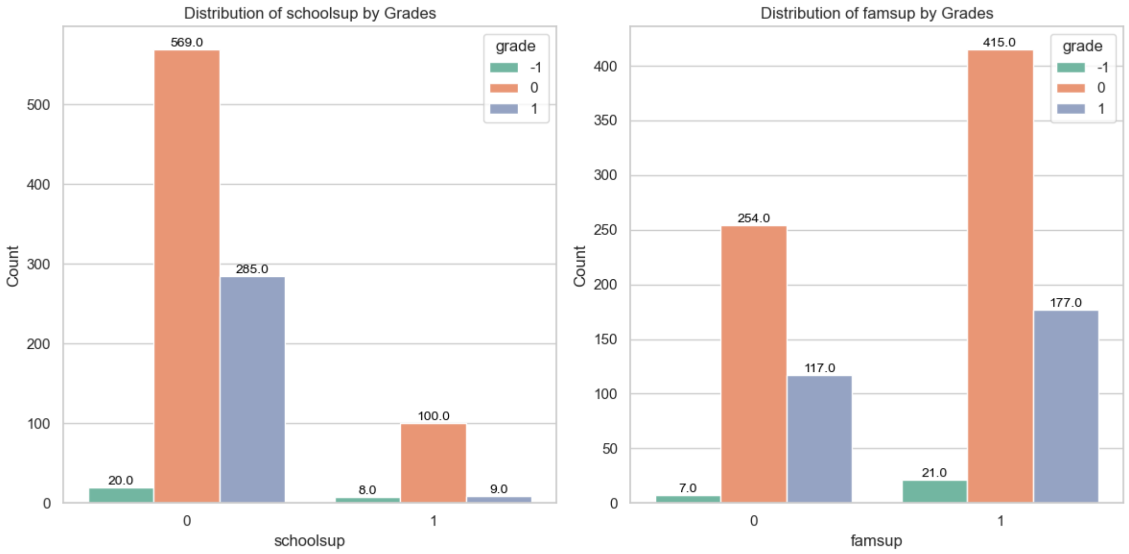

Figure 7: The graph presents varying levels of student performance, categorized as excellent, good, and poor, based on the additional educational support provided by schools and families. The findings indicate that family support positively impacts student outcomes, while school support has a minimal effect on their performance.

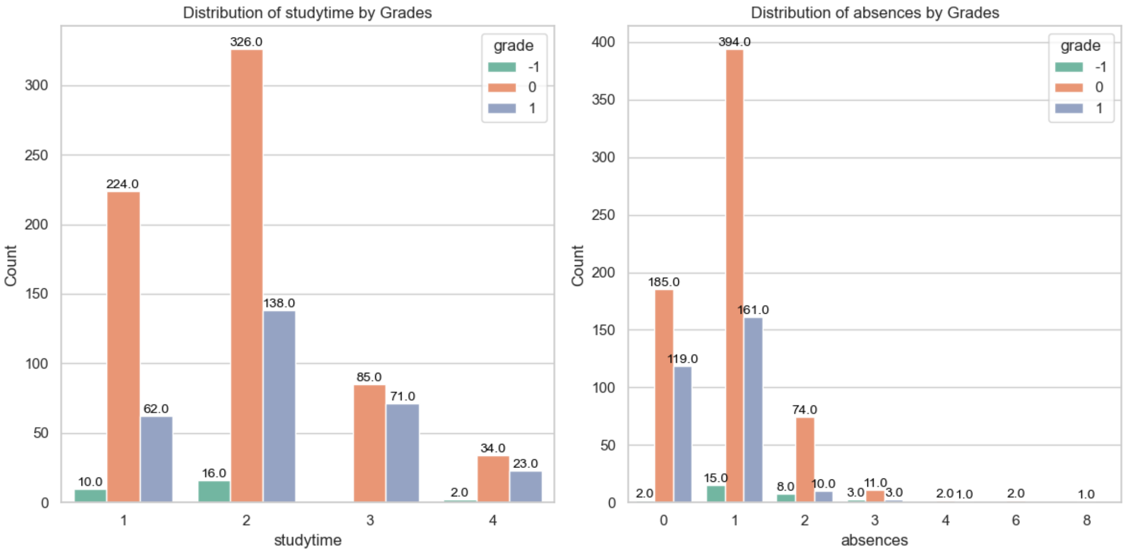

Figure 8: The graph presents varying levels of student performance, categorized as excellent, good, and poor, based on the study time and number of absences. The results indicate that students with fewer absences and who participate more in class tend to have better performance.

Feature	**Meaning**	**Data Type**
school	Student's school	Binary ("GP" or "MS")
sex	Student's sex	Binary ("F" or "M")
age	Student's age	Numeric (15 to 22)
address	Student's home address type	Binary ("U" or "R")
famsize	Family size	Binary ("LE3" or "GT3")
Pstatus	Parent's cohabitation status	Binary ("T" or "A")
Medu	Mother's education	Numeric (0 to 4)
Fedu	Father's education	Numeric (0 to 4)
Mjob	Mother's job	Nominal ("teacher", "health", . . . , "other")
Fjob	Father's job	Nominal ("teacher", "health", . . . , "other")
reason	Reason for choosing this school	Nominal ("home", "course", . . . , "other")
guardian	Student's guardian	Nominal ("mother", "father", "other")
traveltime	Home to school travel time	Numeric (1 to 4)
studytime	Weekly study time	Numeric (1 to 4)
failures	Number of past class failures	Numeric (1 to 4)
schoolsup	Extra educational support	Binary ("yes" or "no")
famsup	Family educational support	Binary ("yes" or "no")
paid	Extra paid classes	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")
internet	Internet access at home	Binary ("yes" or "no")
romantic	In a romantic relationship	Binary ("yes" or "no")
famrel	Quality of family relationships	Numeric (1 to 5)
freetime	Free time after school	Numeric (1 to 5)
goout	Going out with friends	Numeric (1 to 5)
Dalc	Workday alcohol consumption	Numeric (1 to 5)
Walc	Weekend alcohol consumption	Numeric (1 to 5)
health	Current health status	Numeric (1 to 5)
absences	Number of school absences	Numeric (0 to 93)
G1	First period grade	Numeric (0 to 20)
G2	Second period grade	Numeric (0 to 20)
G3	Final grade (target)	Numeric (0 to 20)

Table 1: Features of the Student Datasets (Math and Portuguese Courses)