Exploring Factors Related to Mathematics Success in High School Students

Stephanie Chen
Department of Computer Science
San Diego State University
San Diego, USA
schen1700@sdsu.edu

***Abstract*—Several factors listed in this dataset may be responsible for affecting a student's performance. This paper aims to study datasets such as these to discover problems in the national curriculum and help politicians and teachers change or emphasize parts of their teaching. Through the use of statistical and data science techniques, this research hopes to derive meaningful insights from the dataset to develop sound conclusions and predict final math grades using several factors. It also intends to analyze the impact of multiple factors, including participation in additional educational support, overall family circumstances, and time spent in extracurricular activities and studying. This project utilizes the Kaggle dataset "Math Students," which is taken from the UCI Dataset Repository and contains the final math scores of secondary school students, along with potential influencing factors [1]. Examining this data has shown that factors that are considered positive, a greater amount of studying or a lower alcohol consumption, show students having a higher maximum grade score, while factors that are "negative," a lower quality of health or a greater amount of absences, show students having a lower maximum grade score overall.**

***Index Terms*—academic performance, teacher, student, high school, secondary school**

I. Introduction

As the world moves towards increased education levels among the population, it is important to study patterns and draw conclusions between various factors that may help, or hinder, a student's ability to learn. Middle school is a crucial transitional phase in a student's academic journey, and by examining the factors that affect these students, policymakers and educators can create tailored instructional strategies and support systems to better meet the needs of students.

A paper, "Using Data Mining to Predict Secondary School Student Performance" by Paulo Cortez and Alice Silva, takes this dataset and uses regression and classification techniques to predict results based on the factors and on previous grades [2]. Another scientific research paper, "Factors Affecting Students' Academic Achievement According to the Teachers' Opinion," written by Mehmet Ozcan, explores various factors that affect a high school student's academic success [3].

By comparing the factors against G_avg and looking at the maximum and minimum scores for each result of the factor, it is evident that several factors affect G_avg negatively or positively. Box plots were used to compare the data, as many of the factors include nominal and binary values, rather than numeric values, making it easier to draw conclusions from the data.

TABLE I: Revised Dataset, with New Columns, taken from the UCI Dataset Repository

Attribute	Description (Domain)
address	student's home address type (binary: urban or rural)
famsize	family size (binary: $<= 3$ or > 3)
pstatus	parent's cohabitation status (binary: apart or together)
medu	mother's education (numeric: from 0 to 4)
fedu	father's education (numeric: from 0 to 4)
mjob	mother's education (nominal: teacher, health, services, at home, or other)
fjob	father's education (nominal: teacher, health, services, at home, or other)
traveltime	home to school travel time (numeric: 1– $<$ 15 min., 2– 15 to 30 min., 3– 30 min. to 1 hour or 4– $>$ 1 hour)
studytime	weekly study time (numeric: 1– $<$ 2 hours, 2– 2 to 5 hours, 3– 5 to 10 hours or 4– $>$ 10 hours)
freetime	free time after school (numeric: from 1– very low to 5– very high)
failures	number of past class failures (numeric: n if $1 <= n < 3$, else 4)
schoolsup	extra educational school 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	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1– very bad to 5– excellent)
goout	going out with friends (numeric: from 1– very low to 5– very high)
health	current health status (numeric: from 1– very bad to 5– very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)
alc_avg	average workday and weekend alcohol consumption (numeric: from 1– very low to 5– very high)
G_avg	average of first period, second period, and final grades (numeric: from 0 to 20)

II. Approach

A. Cleaning

TABLE I shows the various factors, along with specific descriptions and explanations used in this paper and in Paulo Cortez and Alice Silva's paper [2]. This data set is taken

from the Kaggle website as well [1]. Participants in this study included students from two schools, Gabriel Pereira and Mousinho da Silveira. These two secondary schools are located in Portugal and the dataset pulls data from hundreds of students to study their math scores.

Several additional columns are also populated in the original dataset, but this paper removes them. Columns ”reason,” ”guardian,” ”sex,” and ”age” are not considered external factors as they are used to provide demographics for the students studied in the dataset and may introduce ethical and moral problems. Only data taken from students attending Gabriel Pereira are used to draw conclusions. Students from the school Mousinho da Silveira are not included in this dataset as they make up only 12% of the dataset and could skew the results.

Mother's education and father's education are combined into a new grouping titled, ”Number of Parents that Achieved Higher Education.” ”schoolsup,” ”famsup,” and ”paid” are grouped into ”Total Number of Additional Educational Support.” The original columns, ”dalc” and ”walc” are averaged to create a new column, ”alc_avg.” ”G1,” ”G2”, and ”G3” are averaged as well to create the column ”G_avg.”

B. Plotting

This paper uses comparative box plots in order to compare grade averages against the various factors included in this dataset and shown in TABLE I. These box plots aim to show the distribution of the grade averages. It would be difficult to use linear regression or gradient descent as many values are nominal or binary. These box plots directly compare and contrast the distribution of grade averages among different groups. An example would be the comparison of median grades between students whose parents live together or apart.

III. Data-Analysis

A. Evaluation Goals

The goal of this project was to find correlations between grades and circumstances a student may encounter or even be born into and to draw conclusions from interpreting such datasets. Theoretically, more positive factors should positively affect a student's grades, and more negative responses should negatively affect a student's grades.

B. Metrics Used for Evaluation

Minimum and maximum grades are compared among the different groupings in the factors that may affect a student's grades. The medians are also looked at, along with the interquartile ranges, or the length of the boxes in the box plots, the spread, signs of skewness, and potential outliers.

C. Results

Fig. 1 shows the comparative box plot for comparing students' grades based on their address locations: urban or rural. The median for students living in urban locations is higher by around one grade score compared to students living in rural locations. The lengths of the boxes are around the same, however, the distribution for urban locations has no skew, while the box plot for rural locations shows some slight right-skew with an outlier that differs by around one grade score, which should not affect the values too much. The overall conclusion is that there is not much difference in grade scores for students living in rural and urban locations, but students living in rural addresses have an average grade score that is lower than students living in urban addresses.

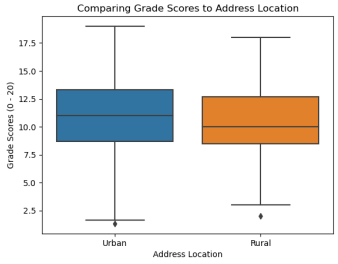


Fig. 1: Comparative Box Plot for Address Locations

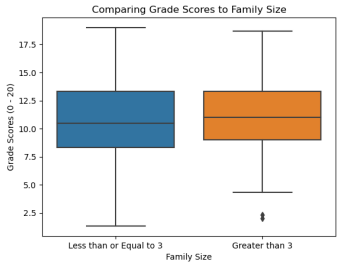


Fig. 2: Comparative Box Plot for Family Sizes

Fig. 2 compares the grade scores for students with less than or equal to three family members and students with greater than three family members. The box plots for both groups show slight right-skew. However, there are multiple outliers present in the box plot for family members greater than three, so calculations of the skewness should be treated with some skepticism. On average, students with larger families had a median grade score greater than students with smaller families.

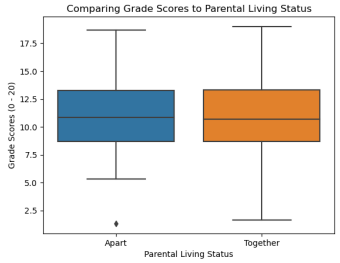
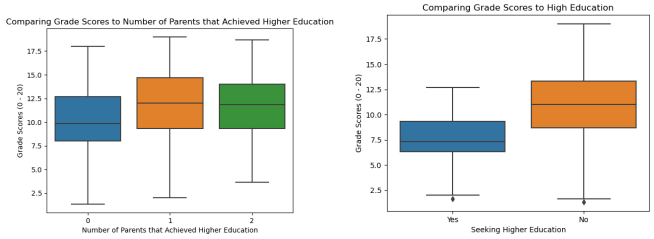


Fig. 3: Comparative Box Plot for Parental Living Status

Comparing students with parents who live together and students with parents who live apart (Fig. 3), show an almost equal median grade score. The box plot for students whose parents live apart is right-skewed with a very extreme outlier, while students whose parents live together are more symmetric. The median grade score is almost exactly equal between

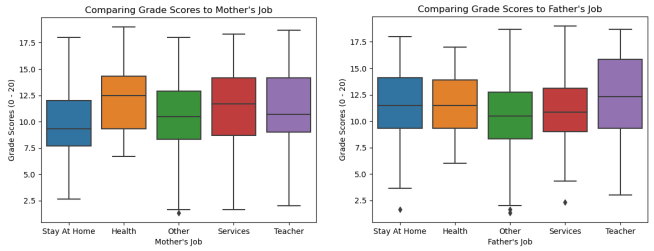
the two groups. To conclude, more students whose parents live apart have a score greater than the median and the median grade score is around the same for both groups. However, the outlier may affect the data for students whose parents live apart.



(a) Parents that Achieved Higher Education (b) Students Seeking Higher Education

Fig. 4: Comparative Box Plots for Higher Education

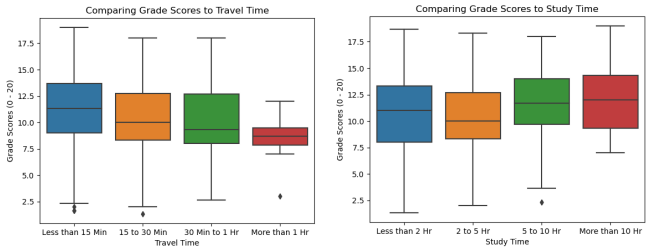
Fig. 4 (a) compares students with no parents, one parent, and both parents who completed higher education. For at least one parent who achieved higher education, the median grade score was around the same and was greater than students with no parents who achieved higher education. There are no outliers for any of the three groups, but the box plot for zero parents is right-skewed so more students are scoring higher than the median, the box plot for one parent is left-skewed with more students scoring less than the median, and the box plot for two parents is about symmetrical. However, the minimum scores increase slightly with more parents who achieved higher education. In conclusion, having at least one parent who completed higher education yields better grade scores when comparing grade scores to students with no parents who achieved higher education. Students who hope to pursue higher education and those who do not want to pursue higher education are compared in Fig. 4 (b). Surprisingly, students seeking higher education had lower maximum and median grade scores than those who were not seeking higher education. This could also be due to a smaller amount of students seeking higher education, 17 compared to 332. On average, students not seeking higher education had no skew with higher grade averages compared to students who were seeking higher education.



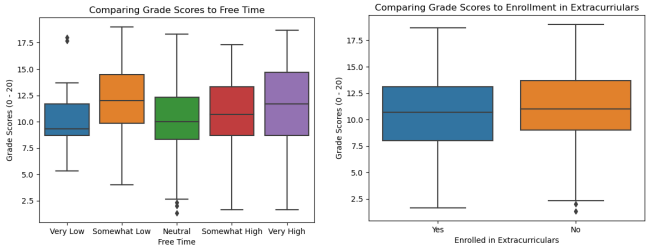
(a) Mother's Job (b) Father's Job

Fig. 5: Comparative Box Plots for Parent's Job

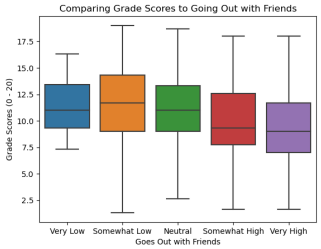
Fig. 5 shows two comparative box plots for the mother's job (a) and the father's job (b). Surprisingly, for both the father's job and the mother's job, those with parents in the health industry had a higher minimum grade score when compared to parents who stayed at home, worked in services, were teachers, or had other jobs. For both mothers and fathers who worked as teachers, their students had consistently high third-quartile grade scores and maximum scores. For mothers, the box plot was right-skewed, while students with fathers working as teachers had left-skewed box plots. Students whose parents had "other" jobs had a lower minimum score, but the outliers could have also skewed the data, so it should be evaluated with caution. A smaller amount of responses accounted for box plots with smaller ranges. Overall, median grade scores for all job groups are around the same, with differences of less than three grade scores, but those with parents in the health sector had a higher minimum score.



(a) Travel Time (b) Study Time



(c) Free Time (d) Extracurricular Activities



(e) Going Out

Fig. 6: Comparative Box Plots for Student Time

Fig. 6 (a), (b), and (c) illustrate several box plots on student travel time, study time, and free time. Those that traveled for more than 1 hour represented a small portion of students, compared to the other groupings, so the data should be evaluated with skepticism. On average, students who traveled less had higher median grade scores. All of the box plots for travel time (a) were right-skewed, with most students falling at the higher end of average grade scores. For study time, minimum grade scores increased as study time increased. There are no clear increases in median grade score as study

time (b) increases, but the box plots become gradually more right-skewed with increased study time. The box plots shown in the free time graph (c) show the exact opposite progression as students increase their free time. Those with very high free time had a lower minimum score than those with very little free time. However, having little free time cannot be equated to having increased study time. Students with very little free time had, on average, lower median grade scores than other students, with few outliers scoring higher than 17.5.

Fig. 6 (d) and (e) feature two graphs on how students spend their free time. In regards to extracurricular activities (d), students who participated and did not participate in extracurricular activities had around the same median grade scores and spread. Both box plots are roughly symmetrical, so it can be concluded that extracurricular activities do not have a large impact on student grades. Students who reported going out often (very high) had similar median grade scores. Those that reported "very low" had a smaller amount of 20 students compared to other groups, all of which had 49 students or more, which may explain the smaller spread in scores. Overall, the grade scores are very similar to each other, but have a decreasing trend. The amount of time students spend "going out" has a slight effect on student performance.

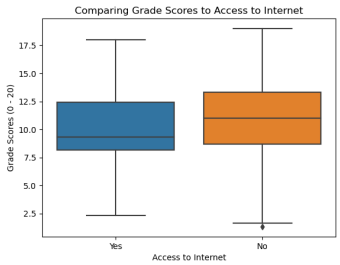


Fig. 7: Comparative Box Plot for Internet Access

Fig. 7 compares students that have internet access and those that do not. The median for those who do not have internet access is higher than those who do, although the box plot is left-skewed. For students with internet access, most score higher than the median, as the box plot is right-skewed. Overall, it can be concluded that those without internet access score higher than those with internet access.

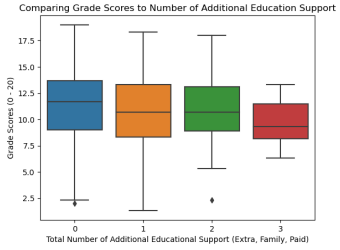
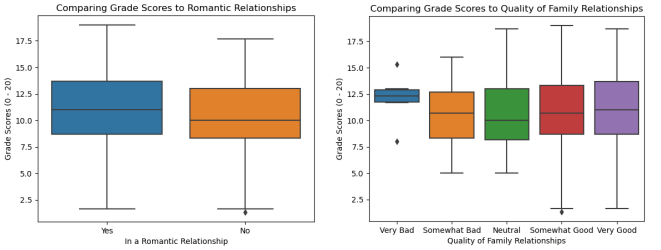


Fig. 8: Comparative Box Plot for Number of Additional Educational Support

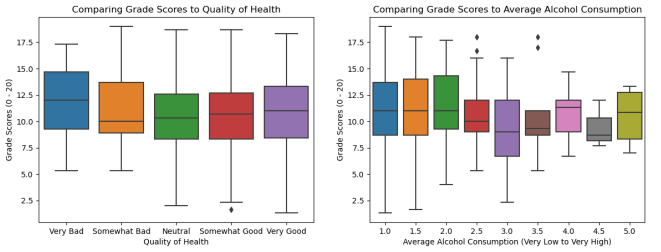
Students with a total of zero, one, two, or three additional educational support (familial support, extra support, paid support) are compared in Fig. 8. Those with three additional supports are few, and may explain the small spread in scores. There is a very noticeable outlier for students with two types of additional educational support, which may affect the median and quartiles shown in the box plot. On average, the medians of all numbers of educational support are around the same, with the box plots for students with zero and one type of educational support being symmetrical and the box plots for students with two or three supports being right-skewed. Surprisingly, students with three types of support have a lower maximum score than other groups, even though they have a higher minimum score.



(a) Romantic Relationships (b) Quality of Family Relationships

Fig. 9: Comparative Box Plots for Relationships

Fig. 9 compares grade scores to relationships. For romantic relationships (a), students in a relationship and not in a relationship have no skew. However, those in a romantic relationship have higher median and maximum grade scores. Average grade scores based on the quality of familial relationships are reflected by the next comparative box plot (b). There were very few students who chose "very bad" as a response, leading to a small spread in the data. In conclusion, most of the medians are around the same value, except for "very bad" responses due to their small spread, and all except "neutral" responses have no skew. The quality of family relationships does not affect grade averages significantly.

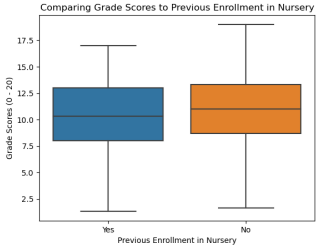


(a) Quality of Health (b) Average Alcohol Consumption

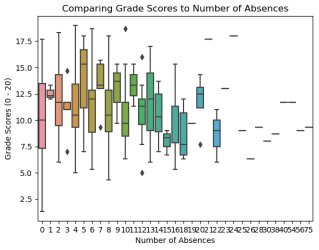
Fig. 10: Comparative Box Plots for Health and Habits

Fig. 10 illustrates the effect of health (a) and habits (b) on average grade scores. Surprisingly, students with "very bad" health had a higher or similar median grade score compared to other responses. However, this could also be due to the smaller number of responses given for those with "very bad"

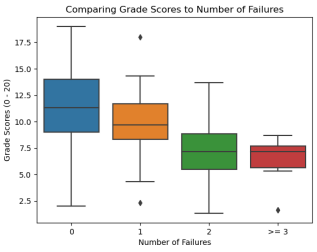
health. This is evidenced in the larger spread for those with "neutral," "somewhat good," and "very good" health. Most of the box plots, except for "somewhat bad" which has a right-skew, have no skew at all. There is no significant effect on average grade scores due to quality of health, but students with "very bad" health scored a higher median grade score. Fig. 10 (b) features several box plots to compare average grade scores between different groupings of students who drink and do not drink alcohol. Most students do not drink, which explains the smaller spread for students who would rate themselves higher on alcohol consumption. The box plots are roughly symmetrical, and the medians are similar, but third-quartile scores and maximum scores decrease as alcohol consumption increases.



(a) Nursery Attendance



(b) Number of Absences



(c) Number of Failures

Fig. 11: Comparative Box Plots for General Schooling

In Fig. 11 there is no significant effect on grade scores for students regardless of nursery attendance (a). The medians are roughly the same, and although the box plot for students who did attend nursery is slightly left-skewed, the spread for most of the students is symmetrical. The effect of absences on grade scores is compared in Fig. 11 (b). Spread decreases with more absences due to a smaller amount of responses. Most students have a smaller amount of absences, and as absences increase, the maximum grade score decreases. However, there is a stark difference when the data for the number of failures is analyzed in Fig. 11 (c). As the number of previous failures increases, the median grade score and maximum score decrease.

IV. Evaluation and Discussion

In summary, it can be seen that internet access, students seeking higher education, and parental occupation and education level have significant effects on a student's performance. Travel time, study time, free time, and time spent going out can also determine a student's grades. Furthermore, educational support, quality of familial relationships, quality of health, average alcohol consumption, number of absences, and number of failures also significantly affect a student's grades. While, Address location, family sizes, parental living status, extracurricular activities, romantic relationships, and previous nursery attendance have little to no effect on a student's grades.

V. Related Work

"Using Data Mining to Predict Secondary School Student Performance," by Paulo Cortex and Alice Silva uses the same data set to predict student grades. Cortex and Silva use RMiner, an open-source library that runs in the R environment to display classification and regression results [2]. They found that there were certain rules that "show the influence of the mother's job (rules 1–2), going out with friends (rules 3–4) and the number of absences (rules 5–6)" as "some of the leading variables that affected a student's grades" [2]. Although their original goal was to predict grades, they were also able to find several features that impacted student performance. However, there is a flaw in their methods. Their prediction is highly dependent on previous grades instead of on student factors [2]. Without knowing "G2", second period, grades, their analysis has a lower predictive accuracy [2]. The study does not base its predictions of student performance solely on external factors; rather, it primarily relies on their past grade results.

A separate paper published in the Education Reform Journal, titled "Factors Affecting Students' Academic Achievement according to Teachers' Opinion" by Mehmet Ozcan, lists several factors, given by real teachers, that they believe can affect a student's grades and performance [3]. Interviews were conducted to gather their data, and their methods "have been specified by researchers ... to ensure validity and reliability" of their responses [3]. Teachers were asked several questions about the family's education level, school physical conditions, school management, school environment, and the effect of teachers to see if these problem questions had a significant effect on students' academic performance [3]. Ozcan concluded that these problem questions are "factors that influence students' academic achievement," and that they parallel "previous and similar studies" [3]. A concern of this study may arise from a lack of concrete results from students. If the report included actual grades or student results outside of these interviews, it would have created a stronger conclusion.

VI. Conclusion

This study has gone in-depth on the various factors that may affect math scores in students. By analyzing and comparing average grade results in different categories and groupings, several factors have been found to influence a student's performance in mathematics. This is integral to developing teaching methodologies, increasing student engagement, and improving grades overall.

In this paper, several factors are shown to have a significant impact on student grades: higher education, health, student time and how it is spent, parental occupation and education level, and school absences and failures. With a larger data set, more concrete conclusions can be drawn. It is important to

continue studying these factors, and other indicators including a student's work, career goals, and mental health to further understand how external factors can affect a student's grades.

REFERENCES

[1] Cortez, P., & Silva, A. M. G. (2017). Math Students, Version 1. Retrieved April 1, 2024 from https://www.kaggle.com/datasets/janiobachmann/math-students.

[2] 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.

[3] Özcan, M. (2021). Factors Affecting Students' Academic Achievement according to the Teachers' Opinion. Education Reform Journal, 6(1), 1-18.