Vedant Gupta, Noah Chen, Tara Standard

What Influences Student Test Scores?

Introduction

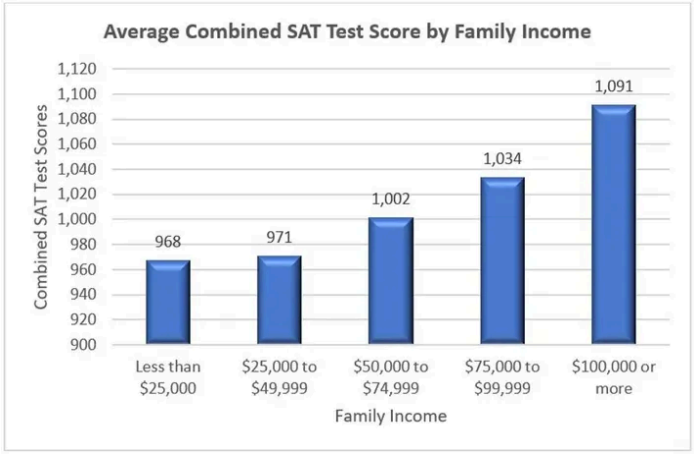
Student performance is a critical factor in a student's future, but the factors that influence performance can be complex and multidimensional, spanning across academic and social habits, among other circumstances. By studying these relationships, we hope that educators, policymakers, and others involved in students' lives can walk away with valuable insights on how to maximize each student's academic potential, no matter their current level. We believe that data-driven analysis can provide practical solutions and possible alternatives to teaching methods and greater education system functions.

Reference Papers

https://www.sciencedirect.com/science/article/pii/S1877050922022529 - This paper tests the correlation of many variables that are similar to our dataset using machine learning techniques. This is relevant as it could help us identify what features to focus on the most.

https://link.springer.com/article/10.1057/eej.2011.33 - This paper specifically talks about the relationship between sleep and student performance. We foresee sleep being a huge factor, so this paper is extremely relevant to what we are doing.

Static Images



Data Source

Our data surrounds factors that may influence student performance such as hours studied, attendance, parental involvement, extracurricular activities, hours of sleep, etc. in relation to final exam scores. It measures 20 variables including exam scores and includes 6,607 records. Our data is in the format of a CSV file which we found on Kaggle, an open-source website for data science materials and resources.

Analysis Plan

First, we plan on performing exploratory data analysis, mainly checking for null/invalid values. Then, we will consider if we should continue to use these observations, and if we should perform imputation or not. To start with, we would first visualize the overall distribution of exam scores in the dataset by using a histogram. As we explore further, one of the main types of plots we think will be insightful are side-by-side boxplots with exam scores on the y-axis. For

example, we are interested in seeing how the level of parental involvement influences exam scores, and we could use that categorical feature on the x-axis with exam score on the y-axis in a side-by-side boxplot. Thirdly, we can utilize scatter plots to see the relationship between exam scores and the continuous features of the dataset. In our scatter plots, we allow users to view other relevant information using a tool tip. Another aspect is color encoding, which would make scatter plots more informative.

Group Member Duties

We will be splitting all the work evenly to ensure that no one is left out and we all contribute to the final product of the project. We will each do at least one visualization, and will evenly split the work done for the remaining visualizations and the website. Depending on the complexity and difficulty of our tasks, some of us will spend more time working on code, while others will start on reports and final deliverables. We have not designated anybody to work on one specific thing, since we plan to make those decisions when we begin working on specific plots and website features. At the same time, we are willing to help each other out if there are problems or issues.

Visualizations

1. Side by side scatterplots with exam score vs hours studied, attendance, sleep hours, previous scores
    - Interactivity: select one point and have them correspond to each other when point is highlighted, rest greyed out

2. Bar chart with motivation level and average final exam score

3. Bar chart with categories being parental education level
    a. Interactivity: drop down where you can pick y axis variable being compared (exam score, hours studied, attendance, sleep hours, tutoring sessions, family income)

4. Correlation heatmap with all variables to find the ones that correlate the most (static)

5. Parallel coordinates plot of different variables vs test score - D3
    a. Interactivity: Brushing to isolate high vs low performer trends

Scatterplots (“scatterplots.html”) :

This interactive scatter plot visualization, created using Altair, explores the relationships between "Exam_Score" and various numerical variables—Hours_Studied, Attendance, Previous_Scores, Tutoring_Sessions, and Sleep_Hours—within the "StudentPerformanceFactors" dataset. By incorporating a selection interval along the x-axis, the visualization not only highlights correlations between exam scores and these metrics but also allows for dynamic comparison of how these variables interact with one another across specific ranges. Key takeaways include a clear positive correlation between exam scores and both hours studied and attendance, indicating that increased study time and class presence tend to boost performance. In contrast, exam scores and previous scores show only a slight positive correlation, suggesting little consistent influence, while sleep hours exhibit no strong correlation with exam scores, implying minimal impact on performance. Interestingly, students with the highest exam scores tend to have fewer tutoring sessions, possibly indicating self-sufficiency or diminishing returns from additional support.

Bar charts (“supportDistributions.html”) :

This visualization comprises three side-by-side bar charts, each depicting the distribution of "Exam_Score" for subsets of students filtered by a categorical variable—Teacher_Quality, Access_to_Resources, and Parental_Involvement—representing levels of support and resources available to students. Each chart is paired with an interactive dropdown menu, allowing users to select a specific category (Low, Medium, High) and view the corresponding exam score distribution. Key takeaways reveal distinct patterns: low parental involvement skews left with lower scores, medium involvement shows slightly better scores, and high involvement skews

right with even higher scores, suggesting a positive link between parental support and academic performance. For access to resources, medium access yields more students with higher scores than low access, while high access boasts the highest mean exam score, indicating that resource availability enhances outcomes. Teacher quality shows less variation—low and medium quality distributions are similar in average, but high quality skews right, suggesting a modest benefit in score distribution despite a comparable mean to medium quality. Together, these interactive charts highlight how varying levels of support influence exam score distributions, with parental involvement and resource access showing clearer impacts than teacher quality.

Bound Bar and Scatter (“ScoreAndMotivation.html”) :

This data visualization features a scatterplot and a linked bar chart to explore the relationship between hours studied, exam scores, and student motivation levels within the "StudentPerformanceFactors" dataset. The scatterplot plots "Hours_Studied" on the x-axis against "Exam_Score" on the y-axis, with each point colored by motivation level (e.g., Low, Medium, High), revealing a positive correlation between study time and exam performance at a glance. However, the intermixed colors suggest no clear relationship between motivation level and exam score, a pattern confirmed by the even distribution of motivation levels among the highest exam scores. The bar chart below, dynamically bound to a selection interval on the scatterplot, displays the count of students by motivation level for the selected data points. For the lowest exam scores, low motivation dominates, though high and medium levels are nearly equal, indicating some influence of motivation at the lower end. Among students who studied the most hours, motivation levels are evenly split, while those who studied the least show a majority with medium motivation, with high and low levels equally represented. This visualization highlights

that while hours studied strongly predict exam scores, motivation's impact is less consistent, varying subtly across different study and performance ranges.

Side by Side Bar Plot (exam_score_barplot.html)

This bar chart compares the average exam scores of students with and without learning disabilities, separated by public vs. private school. The chart shows a slight performance gap between students with and without learning disabilities, regardless of school type. In both school types, students without learning disabilities perform slightly better on average than those with learning disabilities. Because the difference in average scores is so small, it indicates that while learning disabilities may have some impact on test performance, it is not significantly larger for public and private schools. This suggests that public and private schools are providing fairly equitable academic support, but there may still be some room for improvement when it comes to students with learning disabilities to help close the gap fully.

Histogram (exam_histogram.js)

These histograms display the distribution of exam scores across all students, with additional filters for male and female students. The overall distribution is bell-shaped, and is akin to a normal curve with the center being around the mid-to-high 60s. This indicates that most students score in that range. When broken down by gender, both male and female student distributions have the same shape and center, which suggests that there is no significant performance gap between genders. However, the male distribution has a slightly higher spread, having more students at both ends of the score range. In turn, the female distribution is more concentrated in the center and has fewer outliers on either side. This could suggest that female students might

score more consistently than male students, but further analysis would be required to test the significance of this observation. Overall, the histograms shows an expected distribution of test scores with slight differences in variability between male and female students.