Title: Analyzing Key Determinants of Student Academic Performance

Research Question:

How do socioeconomic factors, specifically parental education, lunch type, completion of test preparation courses, and ethnicity influence students' academic performance in traditional exams, and what nuanced connections can be identified between socioeconomic background and educational outcomes?

Project Abstract:

This project aims to explore the impact of socio-economic factors on students' academic performance, focusing on the student performance dataset. Leveraging R programming, the investigation will focus on how variables such as parental education, ethnicity, lunch type, and completion of test preparation courses influence traditional exam scores. The research question delves into understanding the nuanced connections between socioeconomic background and academic outcomes. Through statistical analyses and coding techniques in R, I aim to provide insights into educational disparities. The project adheres to a structured format, including an introduction, data description, methodology, results, and conclusion. Employing a visualization tool aimed to present a comprehensive analysis.

Keywords: Student performance, socio-economic factors, Variables, Visualization.

Introduction

This project delves into the Socio-economic factors influencing student academic achievements. With education being a cornerstone for future success, understanding these determinants is essential for shaping effective educational policies and practices.

Data Used

The analysis is anchored on the “StudentsPerformance.csv” dataset downloaded from Kaggle site Dataset for Student perfomance . (kaggle.com), which comprises categorical variables such as Student Demographics, parental education levels, lunch type, test preparation course, and numeric variables such as exam scores in mathematics, reading, and writing. Each record represents an individual student, offering a comprehensive view of their academic performance about their socio-economic background.

Student Demographics

Gender: Categorical variable indicating the student's gender; Male, Female.

Race/Ethnicity: Categorical variable categorizing students into different race/ethnicity groups; Group A, Group B, Group C

Parental educational Background:

Parental Level of Education: Categorical variable detailing the highest level of education attained by the student's parents e.g. High School, Bachelor's Degree, associate's degree, some college, master's degree. Among others.

Lunch Type: Categorical variable indicating the type of lunch received by the student, for economic background (e.g. Standard, Free/Reduced).

Test Preparation Course: Categorical variable showing whether the student completed a test preparation course (e.g. None, Completed).

Academic Performance:

Math Score: Numerical variable indicating the student's score in mathematics.

Reading Score: Numerical variable showing the student's reading score.
Writing Score: Numerical variable representing the student's writing score.

Methodology

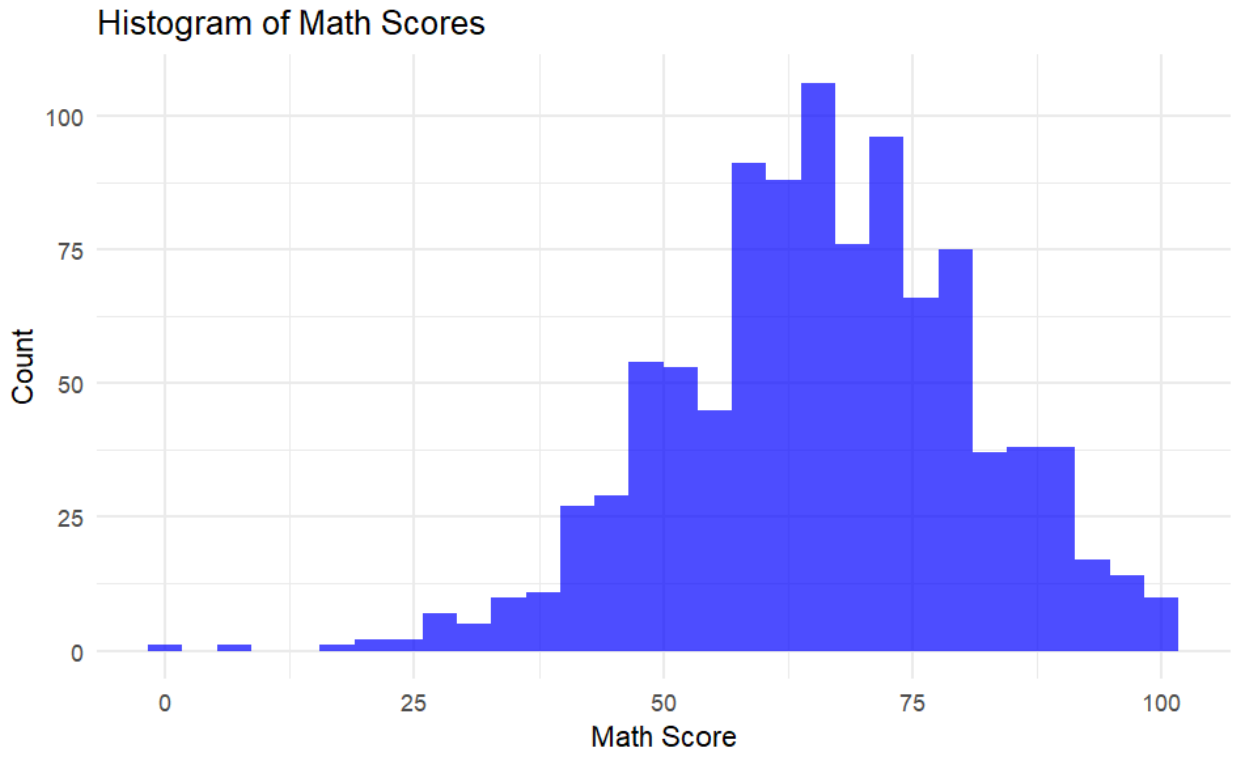
The approach involved thorough data preprocessing followed by exploratory/statistical analysis and regression modeling utilizing R for data manipulation and visualization using libraries such as tidyverse, dplyr, reshape2,tidyr,ggplot2,shiny,readr, rmarkdown,knitr ,among others and employed statistical techniques to explore correlations and causations and. A linear regression model was applied to ascertain the impact of various factors on student scores.

Key findings

Distribution and Skewness.

Histogram of Math Scores

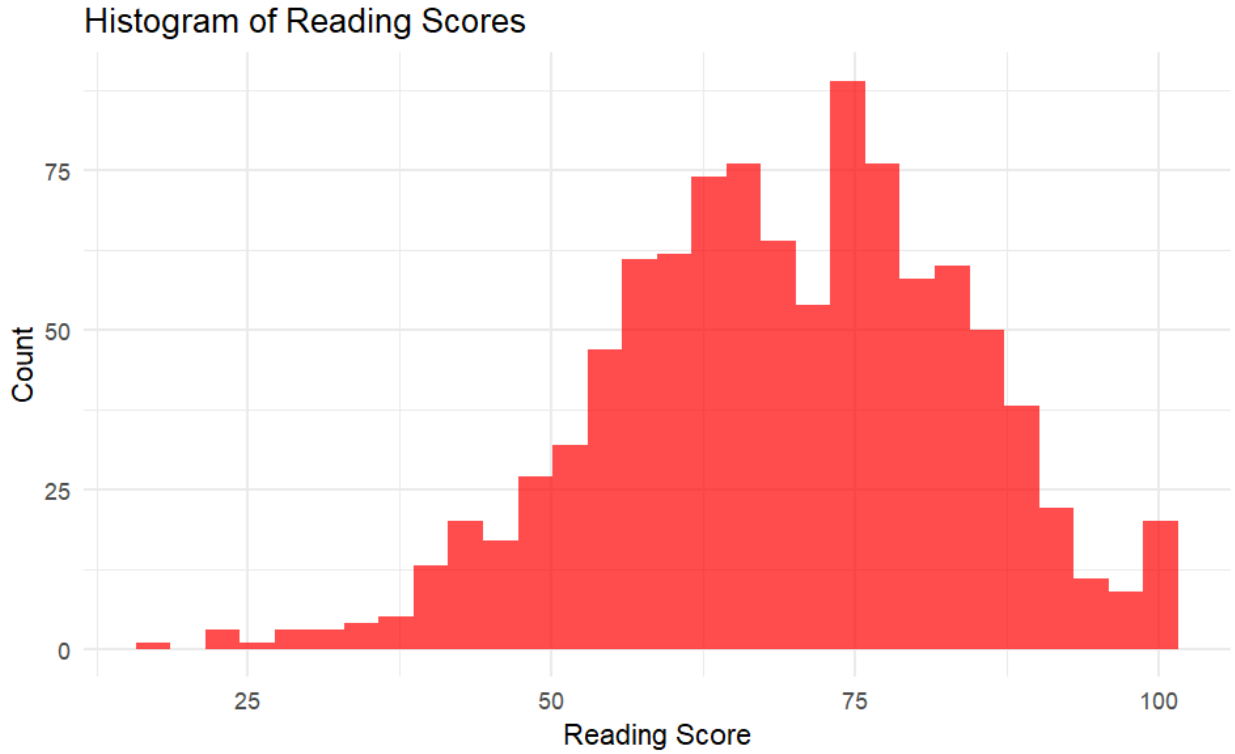
The distribution of math scores showed multiple peaks, indicative of a multimodal distribution. The scores seem to cluster around 60, 70, and 80, which suggests that there may be distinct groups of students achieving similar scores in math. Scores range from near 0 to 100, with a full span of possible outcomes. There is no single central peak; instead, the multiple peaks indicate that there isn't a single most common score but rather several common scores. The distribution shows a slight left skew with a tail extending towards lower scores. There are a few very low scores that could be considered outliers.



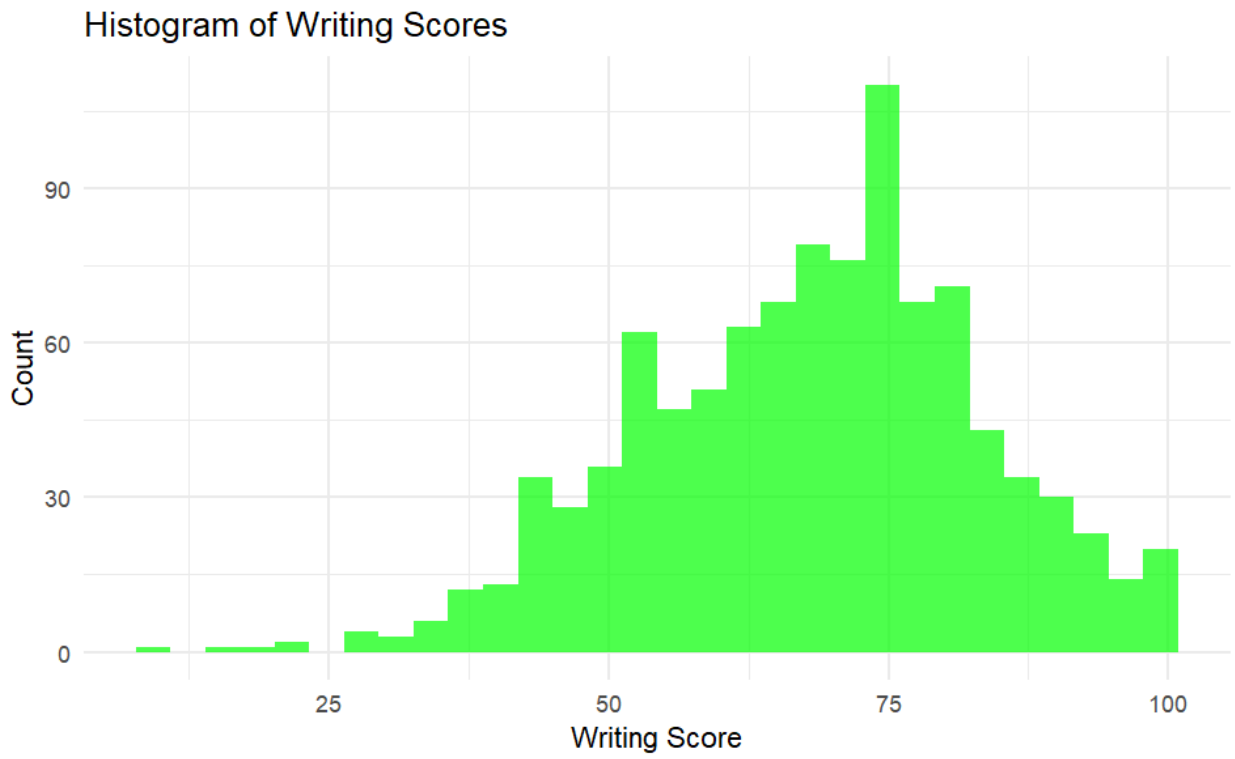
Histogram of Reading Scores (Red)

The distribution of reading scores appear to be approximately normally distributed, with a single, central peak around the 70 scores. The scores range from near 0 to 100. The central

peak suggests the most common score range is around 70, which may represent the average reading score for this group. There is a slight left skew, with a tail extending towards the lower scores, but less pronounced than in the math scores histogram.
Outliers: There are some low scores below 25, but they seem to be less extreme than those in the math score distribution.

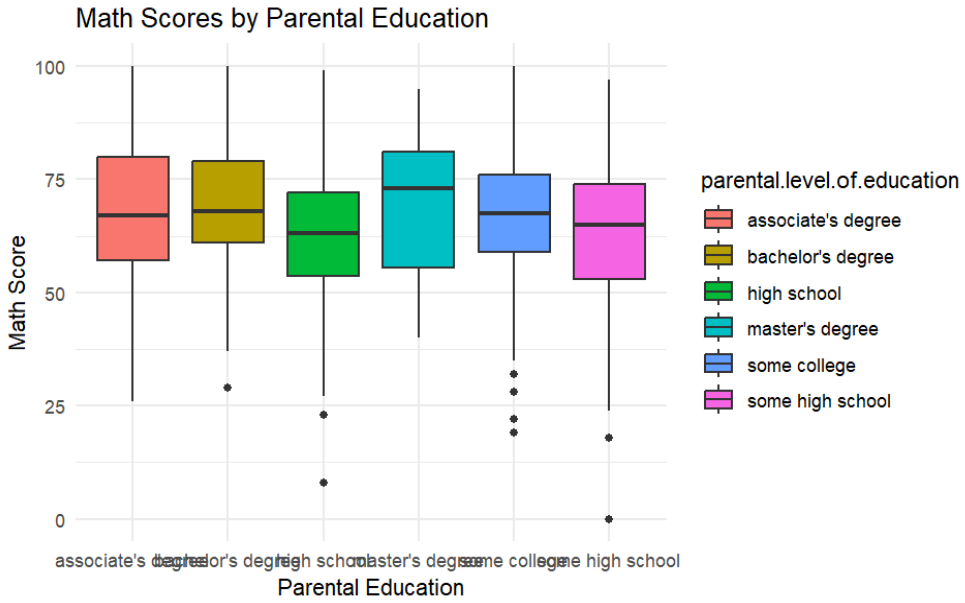


Histogram of Writing Scores (Green)
The distribution of writing scores is skewed to the right, with a peak around the 70-80 score range and a long tail stretching towards higher scores. The scores range from near 0 to 100. The mode of the distribution is around 70-80, suggesting that a significant number of students scored in this range. There is a pronounced right skew, indicating that more students scored higher than the mode compared to those who scored lower. There are a few low scores, but the skewness of the distribution is more towards the higher scores, so the outliers are less prominent.



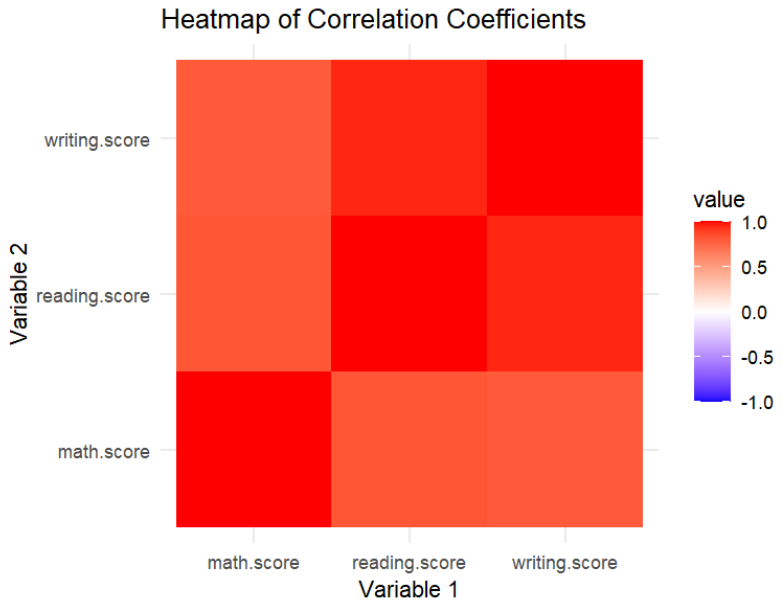
Comparison.
The distributions of scores for reading and writing are more typical of academic performance, with scores clustering around a central value, suggesting a normal distribution. Math scores, however, show a multimodal distribution, which might indicate different levels of proficiency or clusters of students with similar performance levels in math. The slight skews in reading and writing could reflect a trend where a subset of students finds these subjects more challenging or excels in them, respectively. Overall, these distributions offer valuable insights into the academic strengths and weaknesses within the student population, highlighting areas where intervention may be beneficial.

Correlation Between Socio-Economic Factors and Scores: A noticeable correlation was observed between parental education levels and student scores. Students with parents having higher education levels tended to score better across all subjects.



Correlations matrix among numeric variables

Math and Reading Scores: The cell at the intersection of math.score (Variable 1) and reading. score (Variable 2) is red, indicating a positive correlation. This suggests that students who score higher in math also tend to score higher in reading, and vice versa.
Math and Writing Scores: Similarly, the cell at the intersection of math. score and writing score is also red, indicating a positive correlation between math and writing scores.
Reading and Writing Scores: The cell at the intersection of reading score and writing score shows a very bright red color, which implies a very strong positive correlation between reading and writing scores.
Summary.
The heatmap suggests that there are strong positive relationships between all pairs of subjects. The particularly bright red between reading and writing scores may indicate that these two skills are more closely related to each other than either is to math. However, it's important to note that even math scores have a positive correlation with reading and writing scores, which may reflect a general academic proficiency across subjects.



Hypothesis and Regression Model Analysis

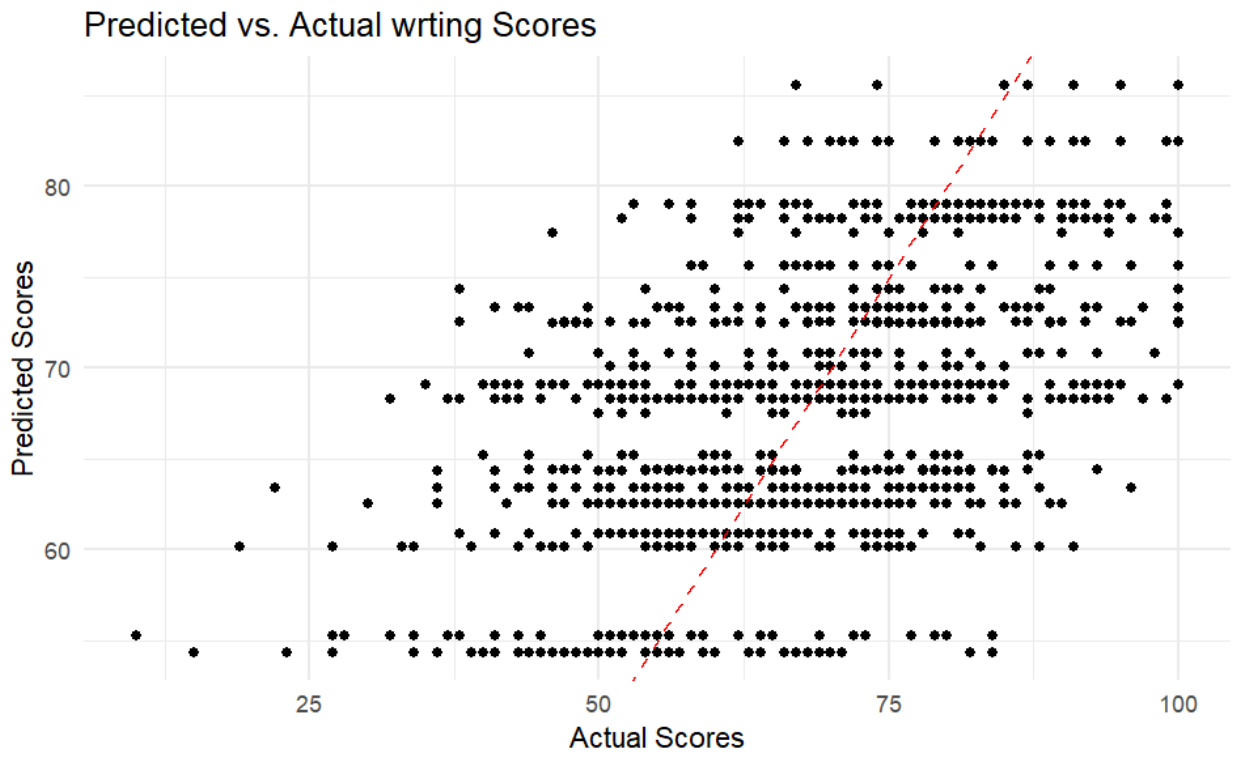
Hypothesis: The study hypothesized that parental education, lunch type, and test preparation significantly affect student academic scores.

Regression Model Findings:

The regression model indicated that higher parental education levels were positively associated with better student performance in all subjects.

Participation in test preparation courses strongly predicted higher scores, especially in mathematics and writing.

Students receiving standard lunch had consistently higher scores, suggesting an underlying economic influence.



There is a positive correlation between actual and predicted writing scores, as indicated by the upward slope of the trend line. This suggests that, generally, as the actual scores increase, so do the predicted scores.
The spread of predictions around the trend line appears to increase as the actual scores increase. This suggests that the model's accuracy might vary more at higher score levels.
There seems to be a concentration of data points above the trend line in the lower half of the actual scores (below 50), indicating that the model may overestimate lower scores. If the dashed red line represented perfect predictions, the closer the dots are to this line, the better the model's predictions. Since many dots are close to the line, it suggests the model has a degree of predictive accuracy. The vertical spread of points at any given value on the x-axis represents the variability of the predictions. Where there's more spread, the predictions are less consistent. In conclusion, the model seems to predict writing scores with some accuracy, as indicated by the concentration of points along the trend line. However, there is variability in its accuracy, and there may be a tendency to overestimate lower actual scores.

Hypothesis Testing:

The positive coefficients for parental education and test preparation in the regression model supported our hypothesis, indicating a significant impact on student scores.
The lunch type variable also significantly correlated, affirming its role in academic performance.

Parental Education: A strong positive correlation was observed, with higher education levels in parents leading to better student performance.
Test Preparation: Engagement in test preparation courses was a significant determinant of improved scores.

Economic Factors: Economic background, as inferred from lunch type, emerged as a crucial factor.

Conclusion

The project confirmed that socio-economic factors, particularly parental education, test preparation, ethnicity, gender, and economic background, significantly influence student academic performance. The use of advanced statistical tools and interactive dashboards facilitated a deeper understanding of these complex relationships, providing valuable insights for educational stakeholders.

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical 1. 1. Learning with Applications in R, 2013

2.WhatIsaRegressionModel?:https://www.imsl.com/blog/what-is-regression-model#:~:text=A%20regression%20model%20provides%20a,by%20a%20linear%20regression%20model,2021

3.Shiny app documentation
shinyapps.io user guide (posit.co)