Evaluating the Impact of Socioeconomic Factors on Student Academic Performance

Executive Summary

This study examines how socioeconomic factors influence student academic performance using a dataset of 1,000 students. Our analysis reveals that parental education level, economic status (measured through lunch type), and access to test preparation resources significantly impact student achievement across math, reading, and writing assessments.

Statistical analysis shows that students with standard lunch plans (indicating higher socioeconomic status) scored an average of 11.1 points higher in math, 7.0 points higher in reading, and 7.8 points higher in writing compared to students with free/reduced lunch plans. Similarly, students who completed test preparation courses scored 5.6 points higher in math, 7.4 points higher in reading, and 9.9 points higher in writing than those who did not. Parental education level also demonstrated a significant effect on academic outcomes, with higher education levels generally associated with better student performance.

Cluster analysis identified three distinct student performance groups: high achievers (30.8%), moderate performers (44.3%), and struggling students (24.9%). These clusters showed clear associations with socioeconomic indicators, suggesting targeted interventions could help address educational disparities.

These findings highlight the need for educational policies that address socioeconomic barriers to academic success. Schools should consider implementing comprehensive test preparation programs, providing additional support for students from lower socioeconomic backgrounds, and developing strategies to engage parents across all education levels in their children's academic journey.

Introduction

Education serves as a critical pathway to opportunity in modern society, yet academic achievement continues to be influenced by factors beyond individual effort and ability. Understanding how socioeconomic factors impact student performance is essential for developing effective educational policies and interventions that promote equitable outcomes for all students.

This research examines the relationship between socioeconomic factors and academic performance among a diverse group of students. The study focuses on three key socioeconomic indicators: parental education level, economic status (as indicated by lunch type), and access to educational resources (represented by test preparation course completion). By analyzing how these factors correlate with student achievement in math, reading, and writing, we aim to identify potential intervention points for reducing educational disparities.

The importance of this research extends beyond academic interest. Educational achievement gaps linked to socioeconomic status have profound implications for social mobility, workforce development, and economic prosperity. When students from disadvantaged backgrounds consistently underperform compared to their more affluent peers, society loses valuable human potential and perpetuates cycles of inequality.

Our research addresses three primary questions: 1. How do socioeconomic factors (parental education, economic status, and access to resources) influence student academic performance? 2. Which socioeconomic factors have the strongest predictive power for academic outcomes? 3. How can these insights inform educational interventions and policies to improve outcomes for all students?

While numerous studies have examined educational disparities, this research contributes to the literature by simultaneously analyzing multiple socioeconomic factors and their relative impact on different academic subjects. Additionally, our use of clustering techniques provides a nuanced understanding of how these factors interact to create distinct student performance profiles.

The findings from this study can help educators, administrators, and policymakers develop targeted interventions that address the specific challenges faced by students from different socioeconomic backgrounds. By identifying the most influential factors affecting academic performance, resources can be allocated more effectively to support student success.

Data

Dataset Description

This study utilizes the "Students Performance in Exams" dataset, which contains information on 1,000 students' demographic characteristics, socioeconomic indicators, and academic performance metrics. The dataset was obtained from Kaggle, a platform for data science competitions and datasets.

The dataset includes three numerical variables measuring academic performance: - Math Score (range: 0-100) - Reading Score (range: 0-100) - Writing Score (range: 0-100)

Additionally, the dataset contains five categorical variables representing demographic and socioeconomic factors: - Gender (Male/Female) - Race/Ethnicity (Groups A through E) - Parental Level of Education (six categories ranging from "some high school" to "master's degree") - Lunch Type (Standard or Free/Reduced) - Test Preparation Course (Completed or None)

The lunch type variable serves as a proxy for economic status, with students receiving free or reduced-price lunch typically coming from lower-income households. Parental education level provides insight into the educational environment at home, while test preparation course completion indicates access to additional educational resources.

Data Preprocessing

Initial examination of the dataset revealed no missing values, eliminating the need for imputation techniques. The categorical variables were converted to factors for appropriate statistical analysis. Column names were standardized for consistency and ease of reference (e.g., "parental.level.of.education" was renamed to "ParentalEd").

For regression analysis, categorical variables were encoded numerically. Parental education levels were converted to numeric values based on their ordinal nature. Lunch type was encoded as a binary variable (1 for standard lunch, 0 for free/reduced lunch), as was test preparation course completion (1 for completed, 0 for none).

A total score variable was created by summing the math, reading, and writing scores to provide an overall measure of academic performance. This composite score (range: 0-300) allows for analysis of general academic achievement in addition to subject-specific performance.

Descriptive Statistics

The dataset shows considerable variation in academic performance across all three subjects. Math scores range from 0 to 100, with a mean of 66.09 and a median of 66.00. Reading scores range from 17 to 100, with a mean of 69.17 and a median of 70.00. Writing scores range from 10 to 100, with a mean of 68.05 and a median of 69.00.

The distribution of scores for all three subjects approximates a normal distribution, with slight negative skewness indicating a tendency toward higher scores. The standard deviations for math, reading, and writing scores are 15.16, 14.60, and 15.20, respectively, suggesting similar variability across subjects.

Regarding socioeconomic indicators, approximately 35.5% of students receive free or reduced-price lunch, while 64.5% pay the standard lunch price. Only 35.8% of students completed a test preparation course, with

the remaining 64.2% not participating in such programs. Parental education levels vary widely, with "some college" being the most common category (22.6%), followed by "high school" (19.6%), "associate's degree" (22.2%), "some high school" (17.9%), "bachelor's degree" (11.8%), and "master's degree" (5.9%).

These descriptive statistics provide a foundation for understanding the dataset's characteristics before proceeding to more complex analyses of the relationships between socioeconomic factors and academic performance.

Analysis

Exploratory Data Analysis

Initial exploration of the dataset revealed important patterns in the relationship between socioeconomic factors and student performance. Visualization of score distributions showed that all three academic subjects (math, reading, and writing) followed approximately normal distributions, with math scores showing slightly more variability than reading and writing scores.

Correlation analysis demonstrated strong positive relationships between the three subject scores. The correlation coefficient between reading and writing scores was particularly high ($r = 0.95$), suggesting that these skills are closely related. Math scores showed moderately strong correlations with both reading ($r = 0.82$) and writing ($r = 0.80$) scores, indicating that while mathematical ability is related to verbal skills, it also represents a somewhat distinct cognitive domain.

Examination of academic performance by gender revealed that female students generally outperformed male students across all subjects, with the largest gender gap appearing in writing scores. This gender difference, while interesting, was not the primary focus of our socioeconomic analysis but represents an important demographic factor to consider in educational research.

When analyzing scores by lunch type (our proxy for economic status), a clear pattern emerged. Students with

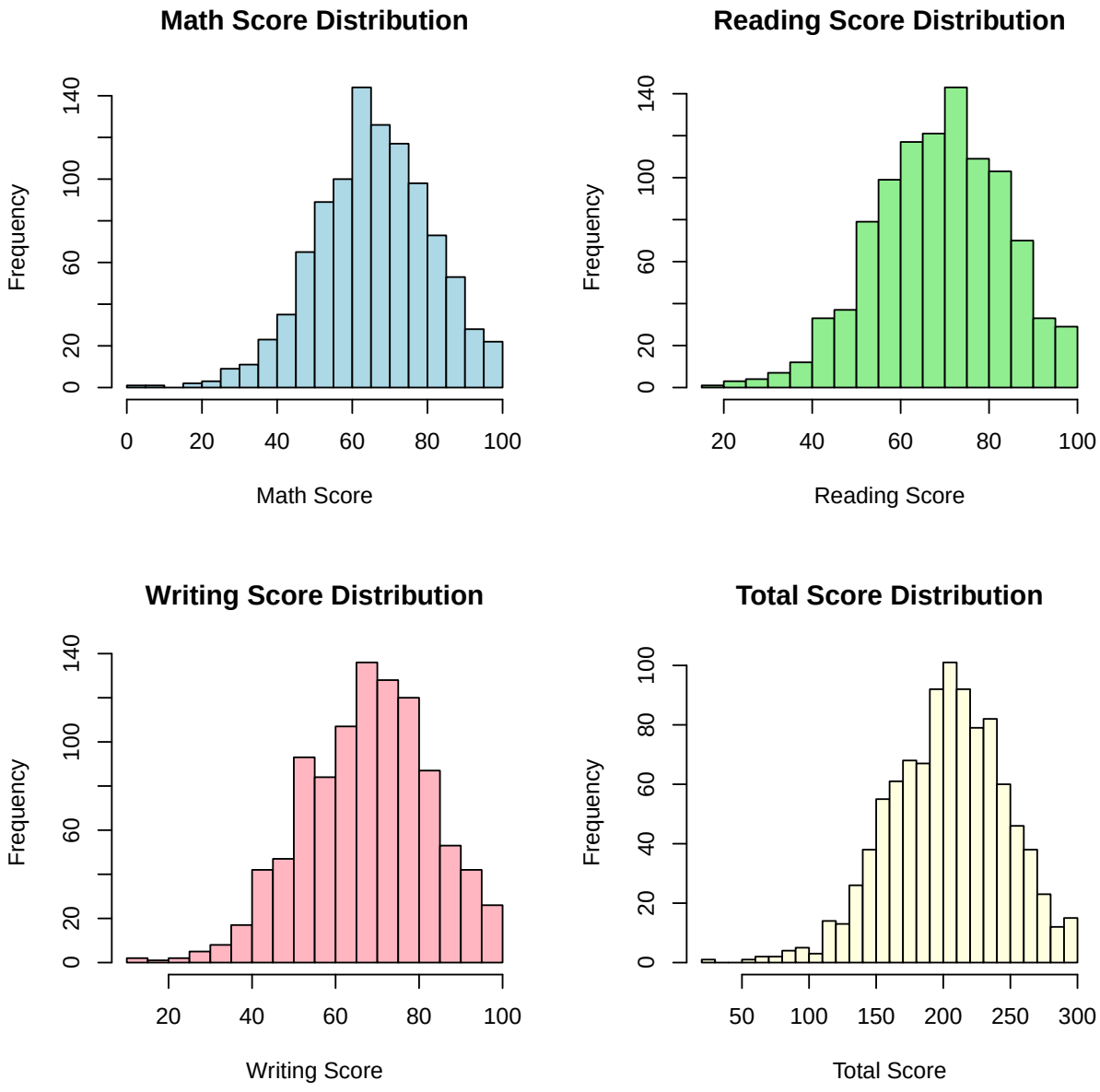

Figure 1: Figure 1: Distribution of Math, Reading, and Writing Scores

standard lunch plans consistently outperformed those with free/reduced lunch plans across all subjects. The performance gap was most pronounced in math (11.1 points) but remained substantial in reading (7.0 points) and writing (7.8 points). These differences suggest that economic factors play a significant role in academic achievement.

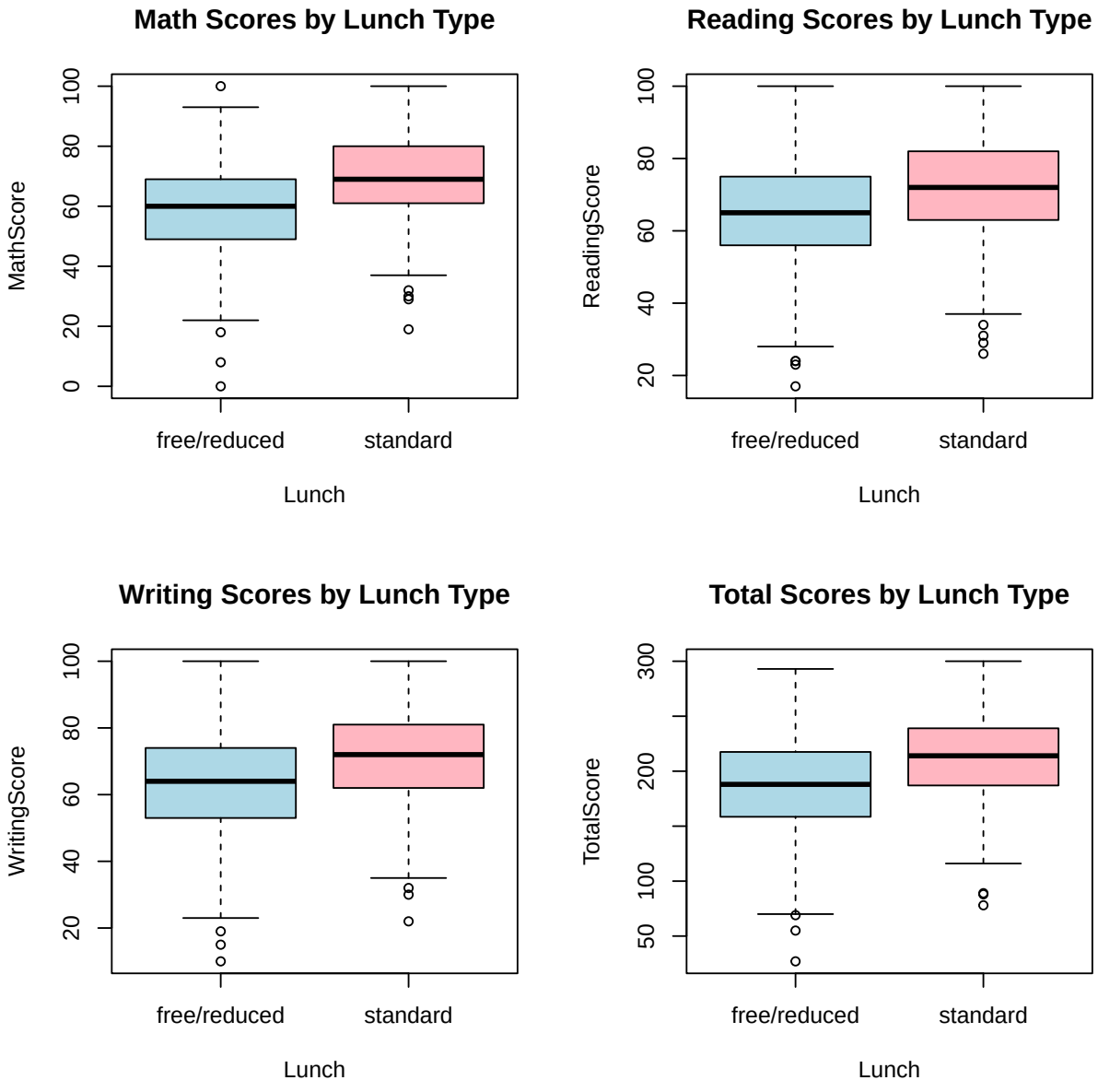

Figure 2: Figure 2: Academic Performance by Lunch Type

Similarly, students who completed test preparation courses showed markedly higher scores than those who

did not. The advantage was most evident in writing (9.9 points) and reading (7.4 points), with a smaller but still significant difference in math (5.6 points). This pattern suggests that access to additional educational resources can substantially impact academic outcomes.

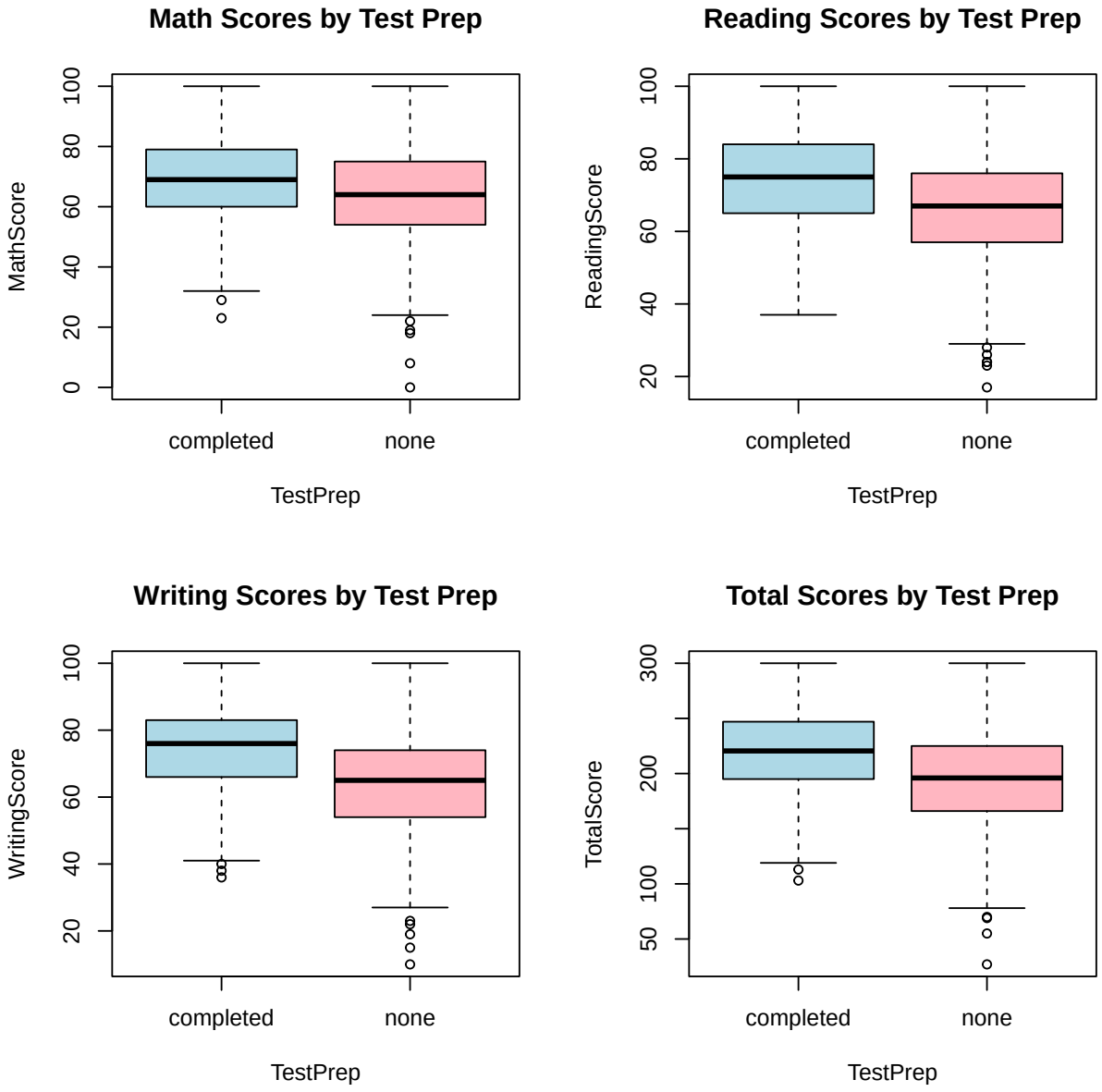

Figure 3: Figure 3: Academic Performance by Test Preparation Course Completion

Analysis of scores by parental education level revealed a general trend of increasing performance with higher levels of parental education, though the relationship was not perfectly linear. Students whose parents

held master's degrees typically achieved the highest scores, while those whose parents had only some high school education tended to score lowest. However, there was considerable overlap between adjacent education categories, indicating that parental education is an influential but not deterministic factor in student achievement.

Statistical Analysis

To quantify the relationships observed in exploratory analysis, we conducted a series of statistical tests to assess the significance of socioeconomic factors on academic performance.

T-tests comparing academic performance based on lunch type confirmed that the observed differences were statistically significant across all subjects ($p < 0.001$). The mean difference in total scores between students with standard lunch and those with free/reduced lunch was 25.9 points (95% CI: 20.5-31.4), representing a substantial achievement gap associated with economic status.

Similarly, t-tests examining the effect of test preparation course completion showed statistically significant differences in all subjects ($p < 0.001$). Students who completed test preparation courses scored an average of 22.9 points higher in total academic performance (95% CI: 17.7-28.1) compared to those who did not complete such courses.

Analysis of Variance (ANOVA) tests were conducted to assess the impact of parental education on academic performance. The results showed significant differences in scores across parental education levels for math ($F(5, 994) = 6.52$, $p < 0.001$), reading ($F(5, 994) = 9.29$, $p < 0.001$), writing ($F(5, 994) = 14.44$, $p < 0.001$), and total scores ($F(5, 994) = 10.75$, $p < 0.001$). These findings confirm that parental education level significantly influences student academic achievement.

Regression Analysis

To understand the combined effect of multiple socioeconomic factors on academic performance, we conducted regression analyses. Simple linear regression examining the relationship between parental education level and math scores showed a statistically significant but relatively weak relationship ($p = 0.031$, $R^2 = 0.005$). This suggests that while parental education has an influence, it alone explains only a small portion of the variance in math performance.

Multiple linear regression models incorporating all three socioeconomic factors (parental education, lunch type, and test preparation) provided a more comprehensive understanding of their combined impact. For math scores, the model was statistically significant ($F(3, 996) = 64.41$, $p < 0.001$) and explained 16.0% of the variance in performance (adjusted $R^2 = 0.16$). Lunch type emerged as the strongest predictor ($b = 11.23$, $p < 0.001$), followed by test preparation ($b = 5.87$, $p < 0.001$) and parental education ($b = -0.62$, $p = 0.010$).

Similar patterns were observed in the multiple regression models for reading and writing scores, with all three socioeconomic factors showing significant effects. The models explained 11.7% of the variance in reading scores and 16.7% of the variance in writing scores. Across all subjects, lunch type and test preparation consistently emerged as stronger predictors than parental education.

For total academic performance, the multiple regression model was statistically significant ($F(3, 996) = 63.33$, $p < 0.001$) and explained 15.8% of the variance (adjusted $R^2 = 0.158$). The model indicated that students with standard lunch scored an average of 26.37 points higher than those with free/reduced lunch ($p < 0.001$), while students who completed test preparation courses scored 23.53 points higher than those who did not ($p < 0.001$). Parental education level, while statistically significant ($p = 0.003$), had a smaller effect size.

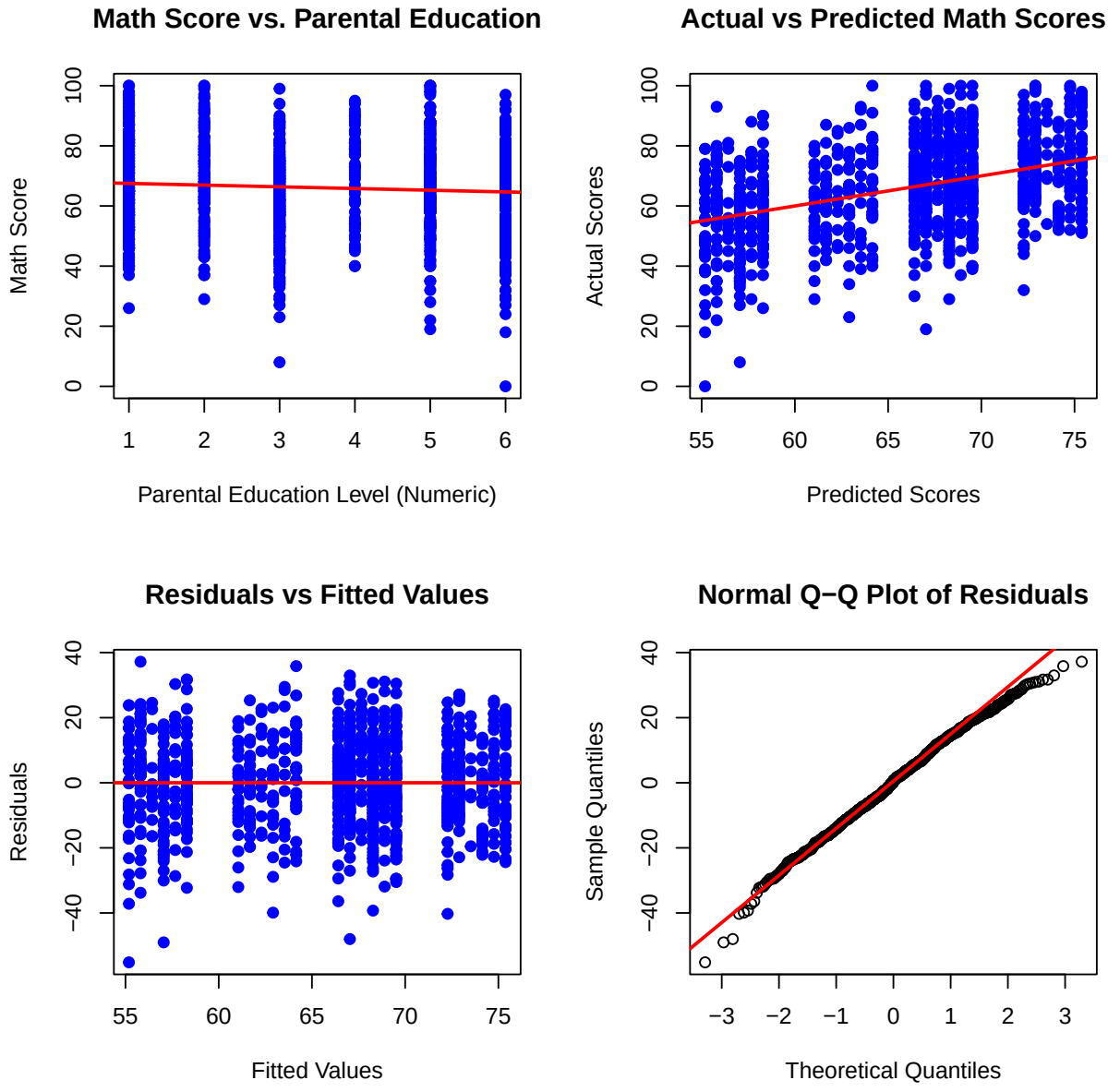

Figure 4: Figure 4: Regression Analysis Results - Actual vs. Predicted Scores and Residual Plots

Clustering Analysis

To identify natural groupings of students based on their academic performance, we conducted k-means clustering analysis. After examining the within-cluster sum of squares for different numbers of clusters, we determined that three clusters provided an optimal balance between simplicity and explanatory power.

The clustering algorithm identified three distinct student performance groups: 1. High Achievers (30.8% of students): Characterized by high scores across all subjects (mean math: 81.7, reading: 85.1, writing: 84.2) 2. Moderate Performers (44.3% of students): Characterized by average scores across all subjects (mean math: 65.3, reading: 68.5, writing: 67.8) 3. Struggling Students (24.9% of students): Characterized by below-average scores across all subjects (mean math: 48.1, reading: 50.7, writing: 48.5)

Cross-tabulation of these clusters with socioeconomic factors revealed significant associations. Among high achievers, 77.0% had standard lunch plans, compared to 44.2% of struggling students. Similarly, 49.7% of high achievers completed test preparation courses, versus only 20.5% of struggling students. Parental education also showed clear patterns, with higher education levels more prevalent among high achievers than struggling students.

These clustering results provide a nuanced understanding of how academic performance profiles relate to socioeconomic factors. The clear association between cluster membership and socioeconomic indicators suggests that interventions targeting specific student groups could be an effective approach to addressing educational disparities.

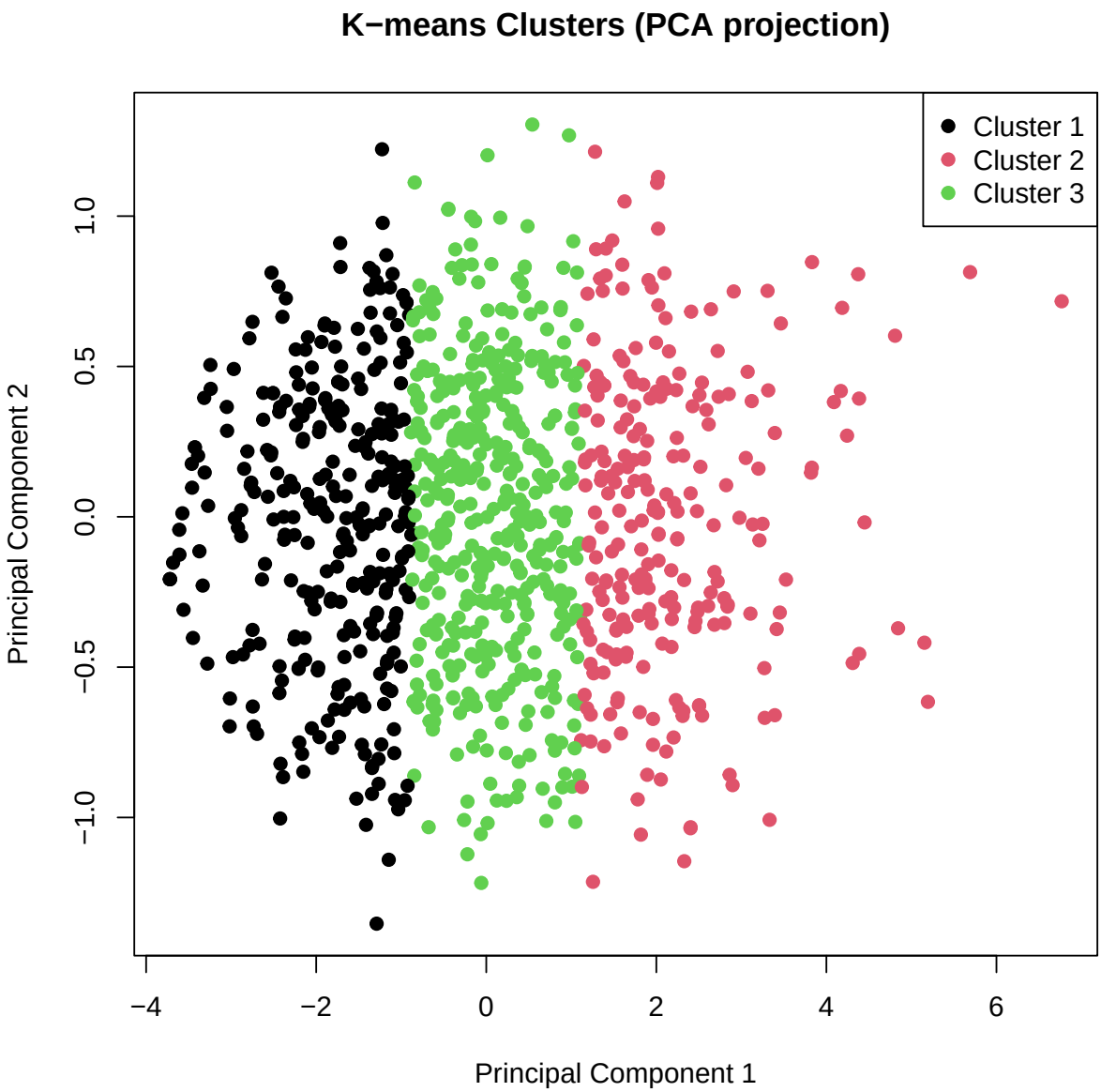

Figure 5: Figure 5: K-means Clustering Results Showing Three Student Performance Groups

Discussion and Conclusion

Summary of Key Findings

This research examined the impact of socioeconomic factors on student academic performance using a dataset of 1,000 students. Our analysis revealed several important findings that contribute to understanding educational disparities and potential intervention points.

First, economic status (as indicated by lunch type) emerged as a powerful predictor of academic achievement across all subjects. Students with standard lunch plans consistently outperformed those with free/reduced lunch plans, with an average difference of 25.9 points in total academic scores. This finding aligns with extensive research documenting the relationship between economic resources and educational outcomes.

Second, access to educational resources (represented by test preparation course completion) showed a substantial impact on performance. Students who completed test preparation courses scored significantly higher across all subjects, with an average advantage of 22.9 points in total scores. This highlights the importance of supplemental educational opportunities in supporting academic success.

Third, parental education level demonstrated a significant but more complex relationship with student performance. While higher parental education was generally associated with better academic outcomes, its effect was smaller than that of economic status and test preparation when all factors were considered simultaneously. This suggests that while the educational environment at home matters, other socioeconomic factors may have more immediate impacts on student achievement.

Finally, cluster analysis identified three distinct student performance profiles that showed clear associations with socioeconomic indicators. High achievers were more likely to have standard lunch plans, complete test preparation courses, and have parents with higher education levels compared to struggling students. This pattern underscores how socioeconomic advantages tend to accumulate, creating substantial performance

gaps between different student groups.

Implications for Education Policy and Practice

These findings have several important implications for education policy and practice. First, they highlight the need for targeted interventions to address economic disparities in education. Schools serving students from lower socioeconomic backgrounds may need additional resources to provide the support services necessary for academic success. Free or subsidized meal programs should be maintained and potentially expanded, as they address a basic need that can affect learning readiness.

Second, the strong impact of test preparation suggests that access to supplemental educational resources is a key factor in academic achievement. Schools should consider implementing universal test preparation programs that ensure all students, regardless of economic background, have access to these valuable resources. After-school tutoring, summer enrichment programs, and online learning platforms could help level the playing field for students who cannot afford private test preparation services.

Third, while parental education showed a smaller effect than other factors in our regression models, it remains an important consideration. Schools can develop programs that engage parents across all education levels in their children's academic journey. Parent education initiatives, family literacy programs, and clear communication about educational expectations and opportunities could help mitigate the impact of varying parental education levels.

Fourth, the identification of distinct student performance clusters suggests that a one-size-fits-all approach to education may be ineffective. Personalized learning strategies that address the specific needs of different student groups could help struggling students improve while challenging high achievers to reach their full potential. Early identification of students at risk of falling into the struggling cluster could enable timely interventions before performance gaps widen.

Limitations and Future Research

This study has several limitations that should be acknowledged. First, the dataset provides a snapshot of student performance at a single point in time, limiting our ability to assess causal relationships or track how socioeconomic factors influence academic trajectories over time. Longitudinal studies following students throughout their educational careers would provide more robust insights into these dynamics.

Second, while lunch type serves as a proxy for economic status, it is an imperfect measure that does not capture the full complexity of socioeconomic status. Future research could incorporate more comprehensive measures of family income, wealth, neighborhood characteristics, and other socioeconomic indicators to provide a more nuanced understanding of economic influences on education.

Third, our analysis explained approximately 16% of the variance in academic performance, indicating that many other factors beyond the socioeconomic variables in our dataset influence student achievement. Future studies should consider additional variables such as school quality, teacher effectiveness, peer influences, student motivation, and learning disabilities to develop more comprehensive models of academic performance.

Finally, our dataset did not include information on interventions or support programs that might mitigate the effects of socioeconomic disadvantages. Research evaluating the effectiveness of specific interventions for students from different socioeconomic backgrounds would provide valuable guidance for educational practice.

Conclusion

This study demonstrates that socioeconomic factors—particularly economic status and access to educational resources—significantly impact student academic performance. The substantial achievement gaps associated with these factors highlight the need for educational policies and practices that address socioeconomic disparities to promote more equitable outcomes.

By implementing targeted interventions that provide additional support for economically disadvantaged students, ensuring universal access to test preparation resources, and engaging parents across all education levels, schools can work toward reducing the influence of socioeconomic factors on academic achievement. Such efforts are essential not only for improving individual student outcomes but also for building a more equitable education system that allows all students to reach their full potential regardless of their socioeconomic background.

The clear relationship between socioeconomic factors and academic performance underscores a fundamental reality: educational success is influenced by circumstances beyond students' control. Acknowledging this reality is the first step toward creating an education system that truly provides equal opportunity for all students. By addressing the socioeconomic barriers to academic achievement, we can move closer to this ideal and help ensure that every student has the chance to succeed based on their abilities and efforts rather than their socioeconomic circumstances.

References

1. Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of educational opportunity. Washington, DC: US Government Printing Office.
2. Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417-453.
3. Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), Whither opportunity? Rising inequality, schools, and children's life chances (pp. 91-116). Russell Sage Foundation.
4. Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. Journal of Family Psychology,

19(2), 294-304.

5. Byun, S. Y., & Park, H. (2012). The academic success of East Asian American youth: The role of shadow education. Sociology of Education, 85(1), 40-60.

6. Kaggle. (n.d.). Students Performance in Exams dataset. Retrieved from https://www.kaggle.com/datasets

7. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

8. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686.

Appendix

Additional Statistical Outputs

Correlation Matrix for Academic Scores

```
             MathScore ReadingScore WritingScore TotalScore
MathScore    1.0000000    0.8175797    0.8026420  0.9187458
ReadingScore 0.8175797    1.0000000    0.9545981  0.9703307
WritingScore 0.8026420    0.9545981    1.0000000  0.9656672
TotalScore   0.9187458    0.9703307    0.9656672  1.0000000
```

Multiple Regression Model for Total Academic Score

```
Call:
lm(formula = TotalScore ~ ParentalEd_num + Lunch_num + TestPrep_num,
    data = students_reg)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-145.766  -24.310    1.091   28.164   95.833

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    184.9802     3.2711  56.550  < 2e-16 ***
ParentalEd_num  -2.0356     0.6791  -2.998  0.00279 **
Lunch_num       26.3653     2.5947  10.161  < 2e-16 ***
TestPrep_num    23.5262     2.5905   9.082  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.26 on 996 degrees of freedom
Multiple R-squared:  0.1602,    Adjusted R-squared:  0.1577
F-statistic: 63.33 on 3 and 996 DF,  p-value: < 2.2e-16
```

Cluster Characteristics

	Cluster	MathScore	ReadingScore	WritingScore	TotalScore
1	1	81.71429	85.14610	84.20779	251.0682
2	2	48.09639	50.67068	48.48594	147.2530
3	3	65.33860	68.45824	67.82167	201.6185

Selected R Code Snippets

Data Preprocessing

```r
# Convert categorical variables to factors
students$gender <- as.factor(students$gender)
students$race.ethnicity <- as.factor(students$race.ethnicity)
students$parental.level.of.education <- as.factor(students$parental.level.of.educat
students$lunch <- as.factor(students$lunch)
students$test.preparation.course <- as.factor(students$test.preparation.course)

# Rename columns for easier access
colnames(students) <- c("Gender", "Race", "ParentalEd", "Lunch", "TestPrep",
                        "MathScore", "ReadingScore", "WritingScore")

# Create a total score variable
students$TotalScore <- students$MathScore + students$ReadingScore + students$Writin
```

Multiple Regression Analysis

```r
# Convert categorical variables to numeric for regression
students_reg <- students
students_reg$ParentalEd_num <- as.numeric(students_reg$ParentalEd)
students_reg$Lunch_num <- ifelse(students_reg$Lunch == "standard", 1, 0)
students_reg$TestPrep_num <- ifelse(students_reg$TestPrep == "completed", 1, 0)

# Multiple linear regression
```

```r
multi_lm_total <- lm(TotalScore ~ ParentalEd_num + Lunch_num + TestPrep_num,
                     data=students_reg)
summary(multi_lm_total)
```

K-means Clustering

```r
# Prepare data for clustering
students_cluster <- students[, c("MathScore", "ReadingScore", "WritingScore")]
students_cluster_scaled <- scale(students_cluster)

# Apply k-means with 3 clusters
set.seed(123)
kmeans_result <- kmeans(students_cluster_scaled, centers=3, nstart=10)

# Add cluster assignments to original data
students$Cluster <- as.factor(kmeans_result$cluster)

# Analyze clusters
cluster_summary <- aggregate(students[, c("MathScore", "ReadingScore", "WritingScor
                             by=list(Cluster=students$Cluster), mean)
```