Analyzing Educational Inequality Across U.S. Schools

Project Title: Socioeconomic Determinants of Student Performance in U.S. Schools

Anushka Naidu Maddisetty
DATA 5100 – Foundations of Data Science
Date: October 22, 2025

Abstract

This project explores what drives the difference in student performance by studying the relationship between socio-economic factors and average ACT/SAT scores across U.S. high schools. The socio-economic factors initially considered included the unemployment rate, the percentage of adults with a college degree, the free lunch percentage, the percentage of students with married parents, and average family income.
A series of regression models were developed to measure these relationships. Ultimately, a **reduced multiple regression model** was selected, which included only three key predictors: **unemployment rate, adults with a college degree, and free lunch percentage**. This model explained approximately **63%** of the variation in test scores ($R^2 = 0.628$).
The results indicate that schools in areas with a lower unemployment rate, a higher percentage of adults with a college degree, and a lower free lunch percentage achieve better test scores. This finding emphasizes how social and economic conditions shape educational inequality.

Introduction

This report investigates what drives differences in student achievement by studying how socioeconomic factors relate to average ACT and SAT scores across U.S. high schools. Educational inequality remains a significant issue in the United States, where community resources and family backgrounds often shape students' access to learning opportunities and academic success. Understanding these relationships helps explain why achievement gaps persist and provides insight into how targeted interventions can improve outcomes.

The analysis draws from two primary data sources:

1. EdGap.org (2016): Provides school-level ACT/SAT scores along with socioeconomic indicators such as median household income, unemployment rate, adult college education percentage, and family structure.

2. National Center for Education Statistics (NCES, 2016): Includes key school identifiers, such as school name, type, location, and level.

These datasets were merged using school identification numbers to create a unified dataset that combines test performance with community-level socioeconomic context. The data were

cleaned, standardized, and examined for missing or inconsistent values to ensure accuracy before analysis.

This study is important because it quantifies how community conditions such as economic stability, education levels, and family demographics can affect school performance. The scientific aim here is to identify which socioeconomic indicators most strongly predict ACT/SAT scores, and to understand how these relationships contribute to educational inequality across different regions of the United States.

Theoretical Background

Educational outcomes are closely linked to socioeconomic conditions, with many studies showing that students from low-socioeconomic-status (SES) households face greater challenges in academic development. Students from low socioeconomic status (SES) households and areas often develop academic skills more slowly compared to their peers from higher SES groups (American Psychological Association, 2017). These early disadvantages can lead to long term differences in academic success, employment, and health in adulthood.

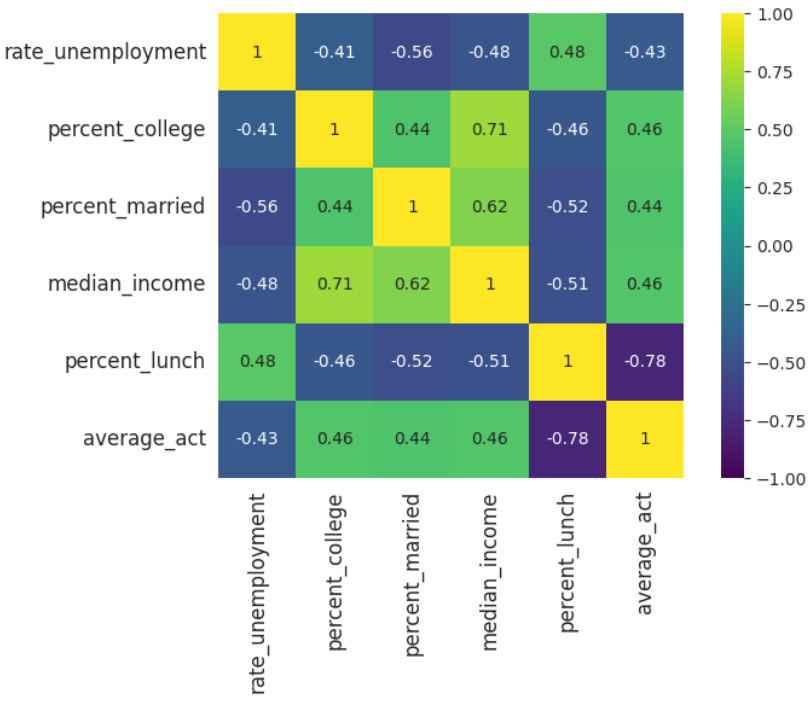
Furthermore, according to the American Psychological Association (2017), schools located in low-SES communities frequently suffer from under-resourcing, which limits access to quality materials, experienced teachers, and learning opportunities. These resource and skill disadvantages help explain the persistent achievement gaps and why schools in communities with stronger economic and educational support systems tend to perform better on standardized tests.

Methodology

To explore the relationship between socioeconomic factors and average ACT/SAT scores, the study followed these main steps :

1. Data Preparation
The datasets were explored, cleaned, and merged to create a single structured dataset. Duplicate or inconsistent entries were removed, and out-of-range values were flagged and replaced with missing indicators. Missing data were then estimated using predictive imputation methods to preserve accuracy and avoid losing valuable information.

2. Exploratory Analysis
Visual tools such as pair plots and correlation heat maps (below) were used to explore the relationships between socioeconomic indicators and school-level test scores. These visuals helped identify which variables were most likely to influence student performance in just one glance.



Few visualization analysis from the above heat map :

- It reveals that higher adult education and household income are linked to higher test scores, while higher unemployment and free lunch percentages are linked to lower scores.
- Among all, the free lunch percentage has the strongest negative correlation with ACT performance, highlighting the strong influence of poverty on academic outcomes.

3. Modeling

Several regression models were developed to measure how strongly socioeconomic factors were associated with average ACT/SAT scores:

• Simple Linear Regression tested the direct relationship between median household income and ACT scores.

• Quadratic Regression examined whether this relationship changed at higher income levels.

• Multiple Linear Regression analyzed how multiple socioeconomic variables jointly influenced school performance.

• Reduced Model included only the most significant predictors to simplify the model while keeping strong explanatory power.

4. Evaluating

Each model was evaluated using R-squared, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) to assess overall fit and accuracy. An ANOVA test compared the models to determine whether adding more predictors significantly improved performance.

Finally, a residual plot was reviewed to confirm that the reduced model met the assumptions of linear regression. The random distribution of residuals around zero indicated that the model

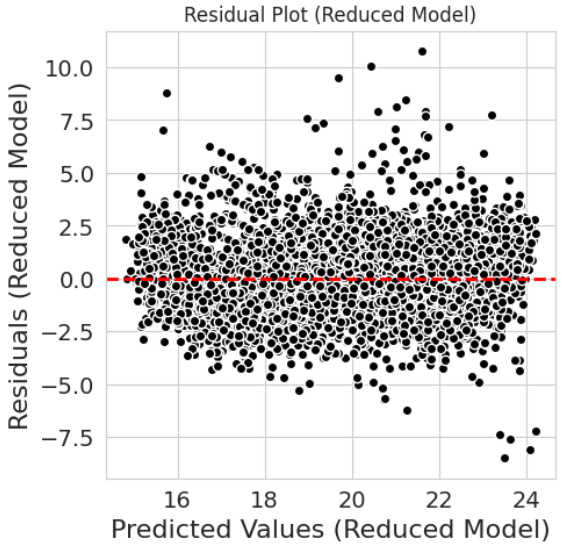
provided a good fit for the data without errors.

Computational Results

The simple linear regression model tested the relationship between median household income and ACT scores. The results showed that income alone explained about 21% of the variation in ACT scores ($R^2 = 0.211$). This means that while income has some effect, it does not fully explain differences in school performance.

Next, a quadratic regression model was used by adding an income-squared term to check if the relationship changed at higher income levels. The R^2 value increased slightly to 0.219 (22%), showing only a small improvement in explaining the data.

A multiple linear regression model was then built, including all socioeconomic factors together such as unemployment rate, adult education, family structure, income, and free lunch percentage. In this model, three predictors stood out as significant: unemployment rate, adult college education, and free lunch percentage.



Finally, a reduced model was created using only the three significant predictors. This simpler reduced performed nearly as well as the full one, explaining about 63% of the variation in ACT scores ($R^2 = 0.628$). The RMSE was 2.26 and the MAE was 1.71, showing a strong and consistent fit. An ANOVA test confirmed that adding more variables did not improve the model, proving the reduced version was efficient and reliable. The residual plot (above graph) showed points scattered randomly around zero, meaning the model fit the data well without major errors.

Discussion

The reduced model found three key factors that had a strong effect on average ACT and SAT scores. First, areas with higher unemployment rates showed to have lower test scores. This shows that families at a economic disadvantage, may have fewer resources and support for their child's learning, which can lead to weaker academic performance.

Second, adult education level in an area was closely linked to student's performance. Schools in areas with more adults who had college degrees generally performed better. This suggests that when adults in a community are more educated, they may place a higher value on learning and create an environment that supports academic growth.

Finally, the percentage of students receiving free lunch had the strongest negative relationship with test performance. Schools with more students from low-income families tended to have lower average ACT and SAT scores. This highlights how poverty can limit opportunities for students and affect their ability to perform well in school.

Together, these three factors explained most of the differences in test scores across schools. This means that economic conditions, family education levels, and access to resources all play a major role in shaping how students perform in a test. When we standardized the data, the results remained almost the same, showing that the findings are consistent and reliable.

These results align with previous research showing that students from low-income households often face more challenges in school due to fewer educational resources and community support (American Psychological Association, 2017). Overall, the findings emphasize that improving economic stability and access to education can help reduce performance gaps and give students a more equal chance to succeed.

Conclusion

This project examined how different socioeconomic factors affect average ACT/SAT scores in schools across the U.S. The results revealed a strong connection between these factors and student performance. Specifically, schools located in areas with lower unemployment rates, higher levels of adult college education, and fewer students receiving free lunch tended to have higher average ACT/SAT scores.

The reduced multiple linear regression model, which included only these three predictors, explained about 62.8% of the variation in test scores ($R^2 = 0.628$). This model performed nearly as well as the full model with all socioeconomic variables, showing that unemployment rate, adult education level, and free lunch percentage are the strongest predictors of student performance.

Overall, this analysis suggests that schools in communities with stronger economic stability and educational support systems achieve better outcomes on standardized tests. These findings

highlight how socioeconomic conditions play a major role in shaping educational opportunities and outcomes across U.S. schools.

References

American Psychological Association. (2017). *Socioeconomic status and education: Research and resources for psychologists and educators*. https://www.apa.org/pi/ses/resources/publications/education

EdGap.org. (2016). *Educational opportunity project: ACT/SAT performance and socioeconomic data*. https://www.edgap.org

National Center for Education Statistics. (2016). *Common Core of Data (CCD): School-level demographics and characteristics*, 2016–2017. https://nces.ed.gov/ccd