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

Introduction

This report explores what drives differences in student achievement by studying how socioeconomic factors relate to average ACT and SAT scores across U.S high schools.

The analysis uses two main data sources:

1. EdGap.org provides school-level ACT/SAT scores along with several socioeconomic indicators, including median household income, unemployment rate, adult college education percentage, and family structure.
2. National Center for Education Statistics includes school identification details such as school name, type, location, and level.

Method

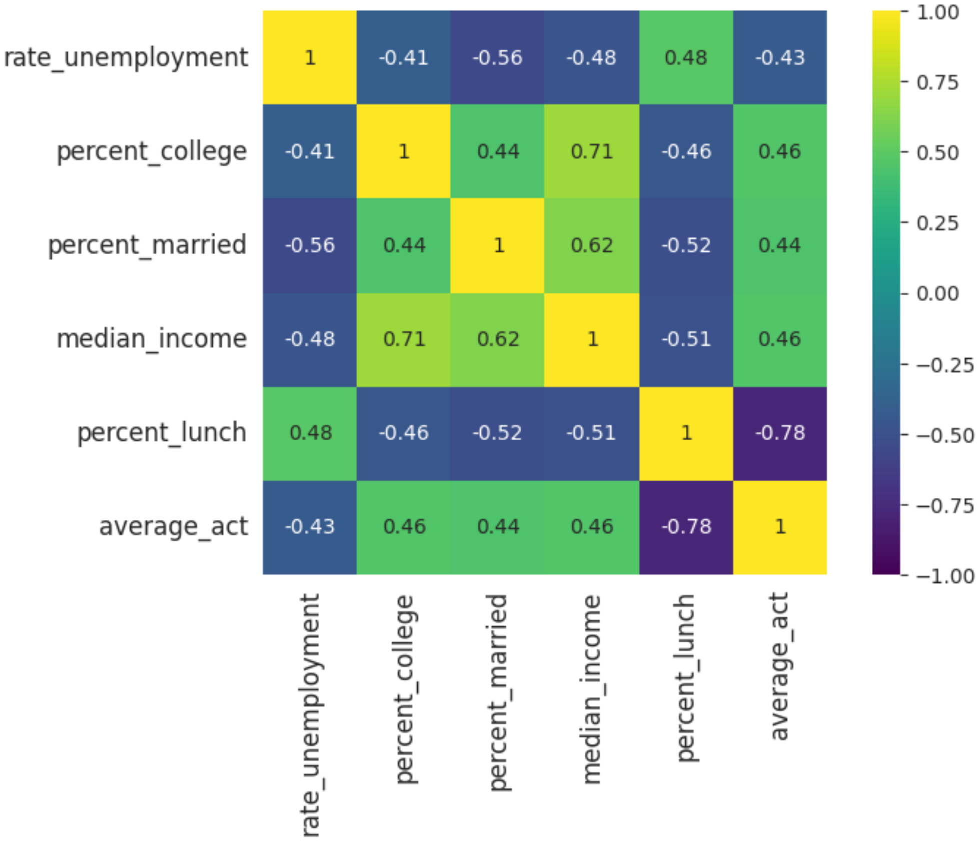
To explore the relationship between socioeconomic factors and average ACT/SAT scores, the study followed these main steps :

Data Preparation

The datasets were explored, cleaned and later merged together. We also aimed to remove duplicate or inconsistent entries. Out-of-range values were identified and replaced with missing indicators. Missing data were then estimated using predictive methods to maintain accuracy and avoid losing valuable information.

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.



Analysis

Few early visualization analysis were drawn such as :

- Schools in communities with higher unemployment rates and higher percentages of students receiving free or reduced-price lunch tended to have lower ACT/SAT scores.

- Schools in areas with more adults holding college degrees had higher average scores.
- Household income was positively correlated with performance, though less strongly than education and poverty-related measures.

These findings suggested that economic stability and educational opportunity in a community strongly affect student outcomes.

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.

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.

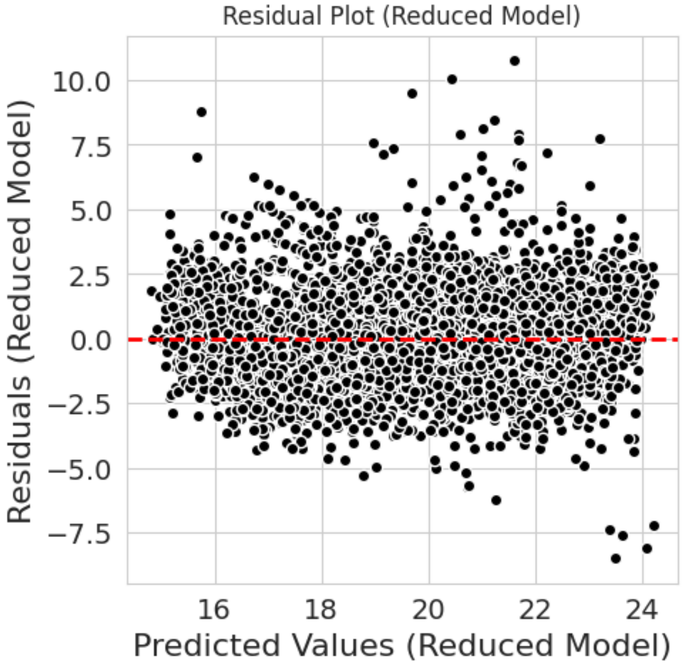
Finally, a residual plot was reviewed to confirm that the reduced model met the assumptions of linear regression. The random distribution of residuals around zero indicated that the model provided a good fit for the data without errors.

Regression 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 or reduced lunch percentage. In this model, three predictors stood out as significant: unemployment rate, adult college education, and free/reduced lunch percentage.



Finally, a reduced model was created using only these three key predictors. This simpler model 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.

Results

The reduced model revealed three main factors that strongly influenced average ACT and SAT scores. First, higher unemployment rates in a area were clearly linked with lower test performance, suggesting that economic hardship negatively affects academic outcomes. Second, adult education level showed a positive relationship with student achievement such as schools located in areas with more college-educated adults tended to have higher average scores. Finally, the percentage of students receiving free lunch was the strongest negative predictor. Schools with a larger share of economically disadvantaged students generally scored lower on standardized tests.

Together, these three factors explained most of the variation in school-level test performance across the dataset. Standardizing the predictors did not change the accuracy of the model, confirming that the findings were consistent and reliable.

Conclusion

In this project, we looked at how different socioeconomic factors affect average ACT/SAT scores in schools across the U.S.

Our results showed a strong connection between these factors and student performance. In particular, schools in areas with lower unemployment rates, more adults with college degrees, and fewer students receiving free or reduced-price lunch tended to have higher average ACT/SAT scores.

The reduced multiple linear regression model, which used only these three factors, explained about 62.8% of the variation in test scores (R-squared = 0.628). This model performed just as well as the full model with all the variables, showing that these three predictors are the most important ones in explaining student achievement.

Overall, this analysis suggests that students in communities with stronger economic and educational support systems perform better on standardized tests. It highlights how much socioeconomic conditions can shape educational opportunities and outcomes.