DATA5100 – Education Project

Hannah Belville

Socioeconomic Determinants of School Performance

Abstract

This study investigates whether school performance, measured by the average ACT composite score, can be predicted from socioeconomic factors. Using a data science workflow, the analysis combines several publicly available datasets from EdGap.org, the National Center for Education Statistics (NCES), and U.S. Census EDGE IPR estimates.

1. Introduction

The central question is:

Can socioeconomic indicators such as income, education level, and family structure meaningfully explain variations in high school ACT performance across the United States?

Understanding this relationship provides valuable insight into educational inequality and the broader social context shaping academic outcomes.

2. Theoretical Background

The analysis integrates multiple datasets, including socioeconomic indicators (unemployment rate, adult education, family structure, income, and free/reduced lunch rates), NCES school-level data, and EDGE program IPR data.

3. Data and Methods

Datasets were merged on the NCESSCH school ID. Data cleaning steps included type standardization, high school filtering, invalid value replacement, duplicate removal, and Iterative Imputer-based missing value imputation. Outliers were detected using the interquartile range (IQR) method, and continuous variables were scaled for consistency.

Exploratory plots revealed strong relationships between income, education, and ACT performance. Schools with higher community education levels and income showed higher ACT averages. Correlation analysis found ACT score positively correlated with median income (r ≈ 0.65) and adult college attainment (r ≈ 0.70), while negatively correlated with free/reduced lunch (r ≈ -0.60).

4. Computational Results

A multiple linear regression model predicted ACT score using socioeconomic and school-level predictors. Predictors included median income, adult education, family structure, free/reduced lunch, unemployment, IPR, and charter status.

The model achieved R² ≈ 0.72, indicating 72% of variance explained. Median income and adult education were the strongest **positive** predictors (p < 0.001). However, free/reduced lunch had a significant negative effect (p < 0.01). IPR added context to the explanation of the correlation for predictors.

5. Discussion and Conclusion

The analysis confirms a robust link between socioeconomic advantage and academic performance. Economic context and parental education significantly predict higher ACT outcomes, while free/reduced lunch remains a strong ACT scoring indicator.

Limitations include uncertainty on data completeness, state-level reporting, and imputation. Future work could explore hierarchical models, nonlinear algorithms, and inclusion of school funding variables.

In conclusion, socioeconomic conditions explain over two-thirds of ACT score variance, emphasizing that educational equity requires addressing broader social and economic disparities alongside school-level reforms.

6. References

Belville, H. (2025). *Education Analysis Notebook (Education.ipynb).* GitHub repository: <https://github.com/belvilleh/education>

EdGap.org. (2023). *Education and Socioeconomic Data by School District.* Retrieved from https://edgap.org

National Center for Education Statistics (NCES). (2023). *Common Core of Data (CCD) and ACT Score Reports.* U.S. Department of Education. Retrieved from https://nces.ed.gov/ccd/

U.S. Census Bureau. (2023). *Education Demographic and Geographic Estimates (EDGE) Program – Income to Poverty Ratio (IPR) Data.* Retrieved from https://www.census.gov/programs-surveys/edge.html