# ABSTRACT

We examine how socioeconomic factors predict high school ACT scores using the 2016-2017 EdGap.org dataset, NCES Common Core Data, and state-level American Community Survey indicators. Through data analysis and linear regression modeling, we find that socioeconomic indicators—especially free/reduced-price lunch participation—strongly predict ACT outcomes. These relationships persist after controlling for other variables. We discuss the study’s limitations and implications for education policy.

# INTRODUCTION

We address the question of whether school performance, measured by the average ACT score, is predicted by socioeconomic factors. We leverage multiple public datasets for the 2016-2017 school year, including EdGap.org and the NCES Common Core of Data (CCD), and we augment the analysis with additional socioeconomic indicators (e.g., from the American Community Survey).

Our analysis follows these steps: load and clean datasets, merge datasets, handle missing values, create derived variables, build regression models, and quantify ACT score relationships.

Data sources:

* CCD data: github.com/benjaminjonjessop/education/data/ccd\_sch\_029\_1617\_w\_1a\_11212017.csv
* EdGap data: github.com/benjaminjonjessop/education/data/EdGap\_data.xlsx
* ACS data: api.census.gov/data/2017/acs/acs5/subject

# THEORETICAL BACKGROUND

Research consistently shows strong links between socioeconomic conditions and academic achievement. Key factors include family income, local poverty rates, parents’ education levels, and free/reduced-price lunch participation. While these correlations don't prove causation, they reveal systematic disparities in educational opportunity. Our study examines how these factors predict school-level ACT performance, acknowledging two key limitations:

1. Omitted-variable bias: Important unmeasured factors may influence both socioeconomic status and test scores
2. Selection effects: Which students take the ACT varies by school and region

In addition, while we have not identified prior research directly linking household internet access to ACT performance, we believe this factor may be an interesting one to explore. Internet connectivity reflects broader access to educational resources, communication tools, and learning opportunities outside the classroom, and therefore may serve as a modern socioeconomic indicator worth testing empirically in this analysis.

# METHODOLOGY

**Software and libraries.** We conducted analysis in Python using the following packages: matplotlib, numpy, pandas, plotly, requests, seaborn, sklearn, statsmodels.

**Data preparation.** We inspected each dataset for schema, types, and duplicates; converted data types where necessary; removed out-of-range values; and addressed missing values through deletion or imputation as appropriate. We joined the EdGap data with school-level CCD records and added internet access data at state level from ACS. We standardized variables and normalized where required for interpretability.

We created transformations relevant to interpretation (e.g., percent variables as 0-100, z-scores for standardized effects), and harmonized naming to lowercase, snake\_case conventions. The final “clean” dataset includes one observation per school and one column per variable.

**Exploratory analysis.** We produced scatterplots, regression trend lines, histograms, and correlation summaries to identify bivariate relationships and potential confounders.

**Statistical modeling.** We estimated linear regression models with average ACT score as the dependent variable and socioeconomic predictors as independent variables (e.g., free/reduced-price lunch participation, income, connectivity). We assessed model fit (e.g., R², adjusted R²) and examined coefficients for magnitude, direction, and statistical uncertainty.

**Reproducibility.** We documented all process steps alongside the code, exported the cleaned dataset to CSV, and included code for regenerating all figures and tables.

# COMPUTATIONAL RESULTS

**Figure 1. ACT score vs. free/reduced-price lunch share**  
A scatterplot with fitted regression line visually confirms the linear relationship between socioeconomic disadvantage and performance outcomes. The line represents the fitted OLS prediction.

Figures (pairplots and regression fits) consistently indicate negative relationships between socioeconomic disadvantage and average ACT scores. For example, the chart clearly demonstrates the negative relationship between % free/reduced lunch and ACT scores. Descriptive statistics and histograms confirm reasonable distributions without extreme outliers after cleaning. Trends are statistically distinguishable from zero at conventional levels for the primary predictors.

**Table 1. Descriptive statistics of key variables**  
*This table summarizes the distribution of variables used in the regression model. Values reflect school-level means and variation after data cleaning.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **N** | **Mean** | **Std. Dev.** | **Min** | **Max** |
| **Average ACT score** | 7227 | 20.3 | 2.51 | 12.36 | 32.36 |
| **Unemployment rate (%)** | 7227 | 0.1 | 0.06 | 0 | 0.59 |
| **College graduates (%)** | 7227 | 0.57 | 0.16 | 0.09 | 1 |
| **Free/reduced lunch (%)** | 7227 | 0.41 | 0.23 | 0 | 1 |

**Table 2. Correlation matrix**  
Pairwise Pearson correlations among variables illustrate expected associations in sign and strength. Correlations were calculated across all available school observations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ACT** | **Unemployment** | **College** | **Lunch** |
| **ACT** | 1 | -0.43 | 0.46 | -0.78 |
| **Unemployment** | -0.43 | 1 | -0.41 | 0.48 |
| **College** | 0.46 | -0.41 | 1 | -0.46 |
| **Lunch** | -0.78 | 0.48 | -0.46 | 1 |

**Table 3. Regression Results: Reduced model**  
Ordinary Least Squares (OLS) regression of average ACT score on selected socioeconomic predictors. Coefficients are expressed in ACT scale units.

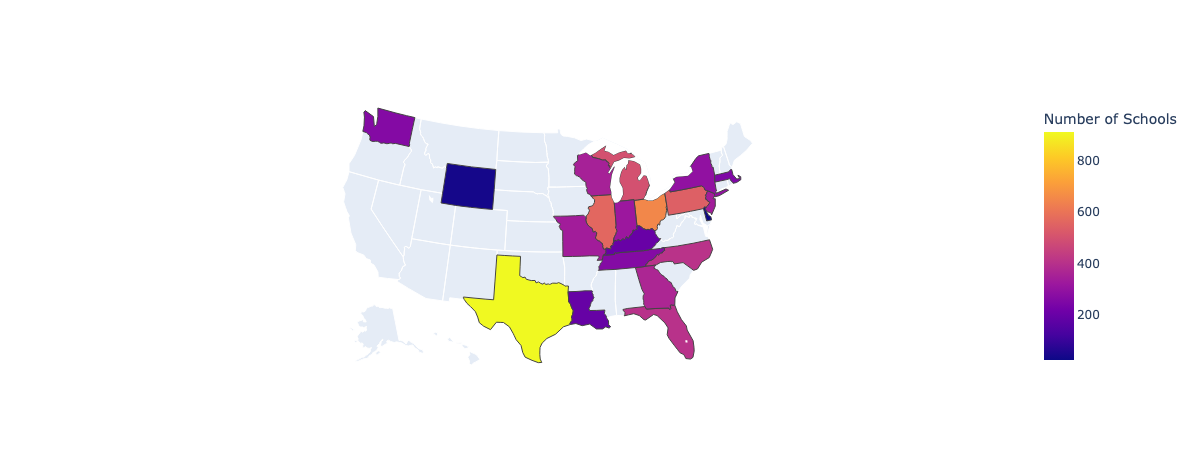
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Coefficient** | **Std. Error** | **t-Statistic** | **p-Value** |
| **Intercept** | 22.639 | 0.102 | 220.93 | 0 |
| **rate\_unemployment** | -2.174 | 0.374 | -5.817 | 0 |
| **percent\_college** | 1.717 | 0.127 | 13.53 | 0 |
| **percent\_lunch** | -7.586 | 0.093 | -82.01 | 0 |

# DISCUSSION

**Key findings.** Our analysis suggests that schools with higher levels of socioeconomic disadvantage tend to report lower average ACT scores. Each percentage point increase in disadvantage correlates with meaningful ACT score decreases.

**Robustness and sensitivity.** The results remain consistent when using slightly different model setups, such as expressing percentages as fractions or adjusting the level of aggregation (state vs. school). However, results should be interpreted with caution: the analysis is observational and subject to confounding. Unobserved factors (e.g., instructional quality, peer effects, local policy) may bias estimates.

**Important limitations.** Data integration required harmonizing identifiers and imputing or dropping missing values; one socioeconomic measure (internet access) was available only at aggregated geographies (i.e., state), limiting within-state variation and attenuating explanatory power. The school dataset included records from only **20 U.S. states (illustrated below)**, which constrains the generalizability of findings and may underrepresent regional diversity in socioeconomic and educational contexts. Further, the ACT score is itself an imperfect proxy of school performance and can reflect selection effects in test-taking.



# CONCLUSIONS

Our analysis reveals strong connections between socioeconomic conditions and school ACT scores in the 2016-2017 academic year. Drawing from EdGap and CCD data, supplemented with American Community Survey indicators, we developed an OLS model to examine these relationships.

Socioeconomic factors explain substantial ACT score variation across schools. Higher unemployment rates and increased free lunch program participation correlate with lower ACT scores, while areas with more college graduates show higher scores. Though correlation does not prove causation, these patterns demonstrate clear links between community economic health and student achievement.

Several constraints affect our findings. Our dataset covers only 20 states, limiting national insights. We relied on state-level socioeconomic measures rather than school-specific data. While we explored internet access as a predictor, it added little explanatory power to our model.

The evidence confirms that socioeconomic conditions and school performance intertwine closely. This connection demands policies that address community-level disadvantage alongside educational reforms. Future studies should examine these relationships through longitudinal data and causal analysis to uncover the specific mechanisms connecting socioeconomic status to academic achievement.