Education Project Report

Author: Porhay Rouen
Date: October 2025

Abstract

This report exames how socioeconomic characteristics affect average ACT scores, which are a measure of academic performance across the United States. To find important factors, we used exploratory data analysis and linear regression on the EdGap and CCD datasets. The findings show that the factors that have the biggest effects on ACT scores are local unemployment, adult education level, and student poverty..

Introduction

The goal of this analysis is to comprehend the relationship between school performance in the US and socioeconomic issues at the community level. Higher ACT scores are frequently indicative of communities with greater financial and educational advantages. In order to quantify their relative influence on academic achievement, this initiative looks at factors such median household income, adult education, unemployment, family structure, and student poverty.

Data Sources:

- **EdGap dataset:** School-level ACT scores and socioeconomic indicators link
- **Common Core of Data (CCD):** School characteristics including enrollment, student-teacher ratio, and free/reduced lunch participation link

Theoretical Background

Previous research indicates that while high unemployment and poverty can impair performance, communities with greater adult education and household income levels tend to support students better. By combining regression models to analyze several socioeconomic indices, our project expands on these findings.

Methodology

1. **Data Loading and Cleaning:**
   - Imported datasets in CSV and Excel format.
   - Removed duplicates and out-of-range values
   - Handled missing values by removing incomplete records in key columns.

2. **Data Merging and Preparation:**
   - Merged EdGap and CCD datasets using school IDs.
   - Selected relevant columns and renamed them for clarity.
   - Created derived variables:
     - income_k: median income in thousands
     - act_vs_median: ACT score difference from median
3. **Exploratory Data Analysis (EDA):**

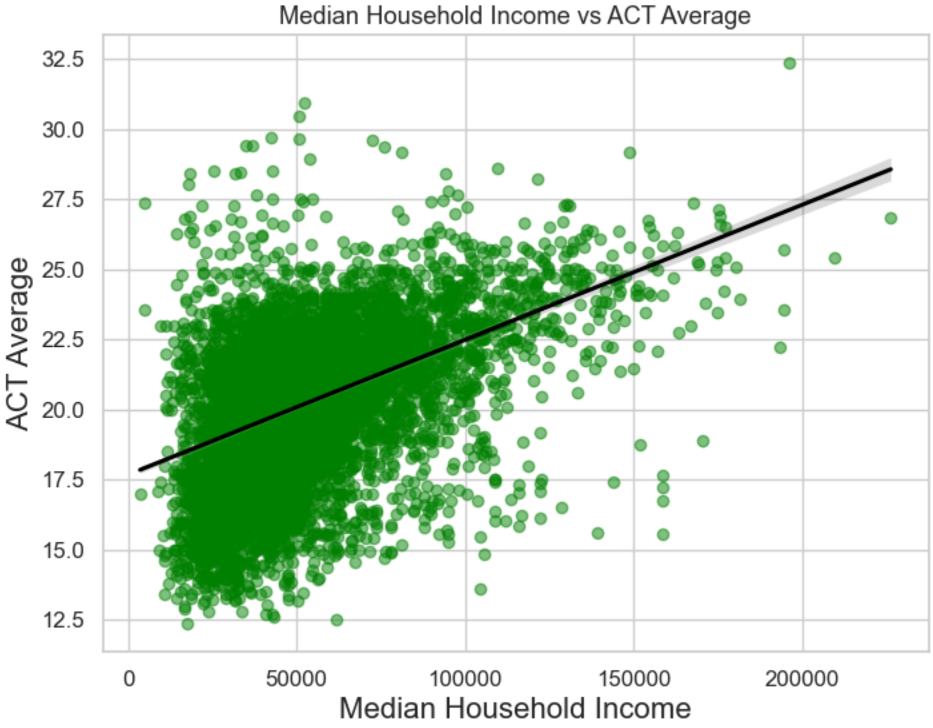


Figure 1: Median Household Income vs. ACT Average, showing a positive correlation between income and ACT scores.

The plot shows a positive correlation between median household income and ACT scores: schools in wealthier areas tend to have higher ACT averages. Most schools fall in the $25,000–$125,000 income range with ACT scores between 17.5 and 27.5. While the trend is clear, scores vary at all income levels, indicating that income is not the only factor influencing performance.

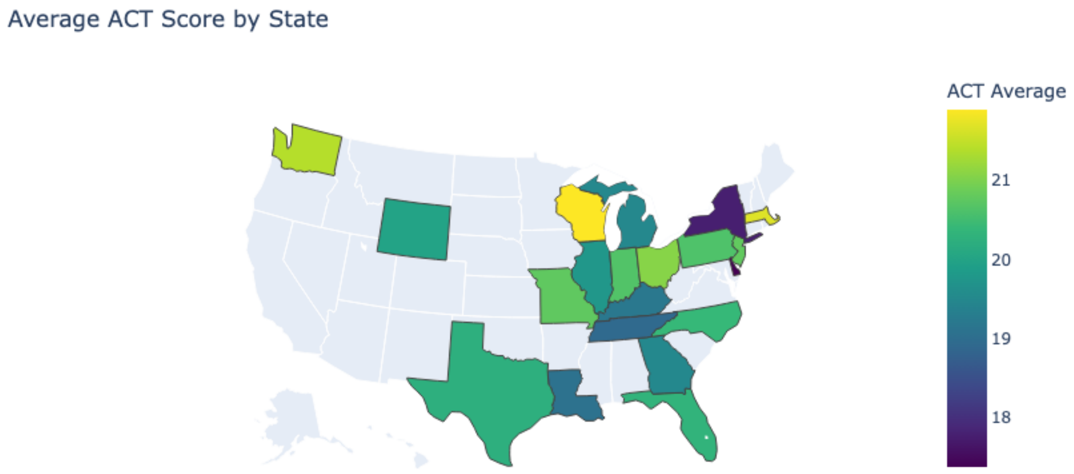


Figure 2: Average ACT Score by State, highlighting regional differences

Average ACT Score by State

- The map shows state-level average ACT scores using a Viridis color scale: darker colors indicate lower scores (~18), and lighter colors indicate higher scores (~21+).

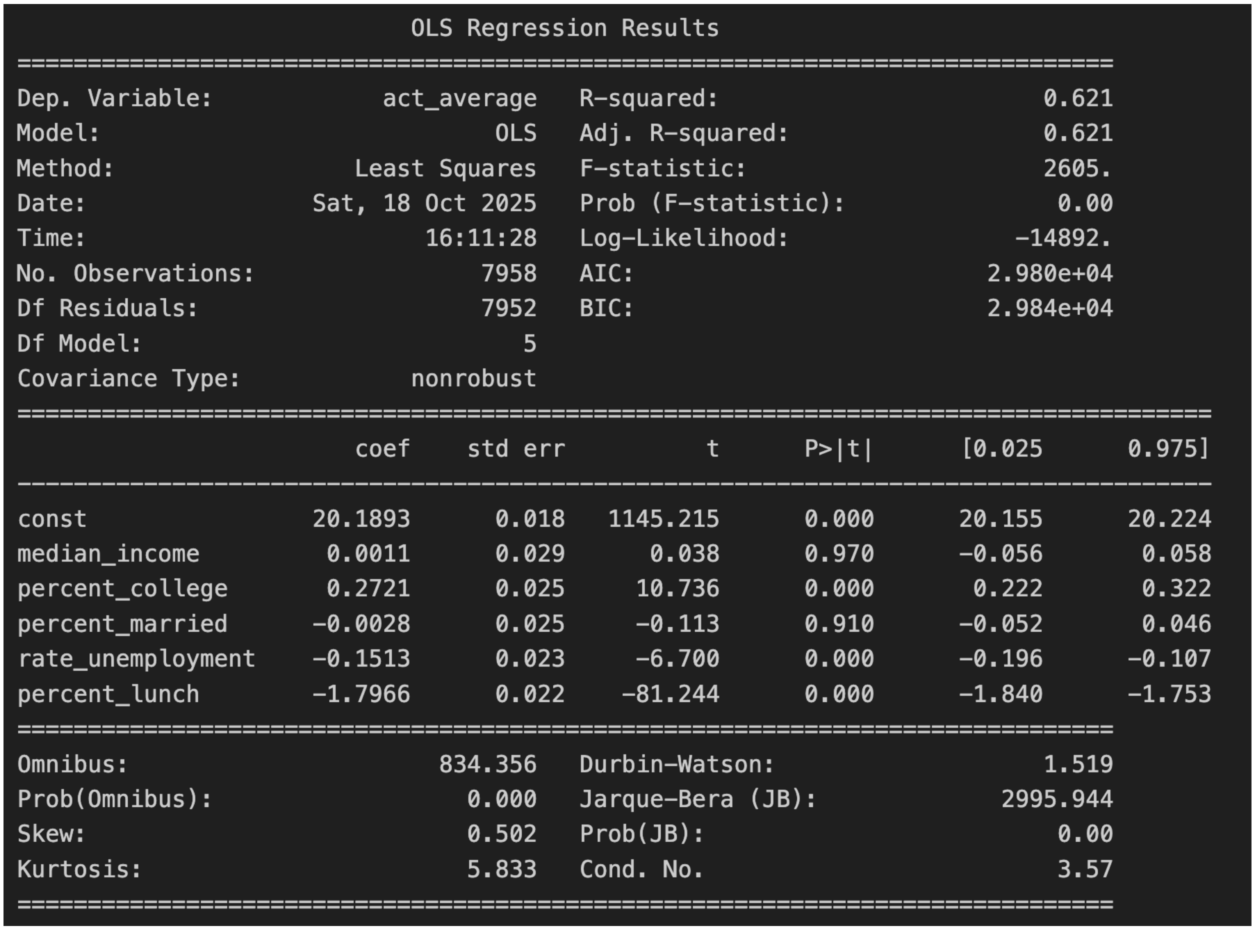
Key Observations:

- Lowest scores: Concentrated in the Southeastern states (e.g., Louisiana, Mississippi, Alabama).
- Highest scores: Seen in parts of the Midwest (e.g., Michigan, Wisconsin), the Western U.S. (e.g., Washington), and some Northeastern states (e.g., Massachusetts, Connecticut).
- Mid-range scores: Spread across the Midwest and Southwest.
- Missing data: Some states appear uncolored, likely due to reporting gaps or preference for SAT over ACT.

4. **Linear Regression Analysis:**
   - Ordinary Least Squares (OLS) regression to model ACT scores using predictors:
     - Median household income
     - Percent of adults with college degrees
     - Percent of children in married-couple families
     - Local unemployment rate
     - Percent of students on free/reduced lunch
   - Scaled predictors using StandardScaler to compare effect sizes.

Computational Results

Figure 3 shows the OLS **Scaled Regression Output:**



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            act_average   R-squared:                       0.621
Model:                            OLS   Adj. R-squared:                  0.621
Method:                 Least Squares   F-statistic:                     2605.
Date:                Sat, 18 Oct 2025   Prob (F-statistic):               0.00
Time:                        16:11:28   Log-Likelihood:                -14892.
No. Observations:                7958   AIC:                         2.980e+04
Df Residuals:                    7952   BIC:                         2.984e+04
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                20.1893      0.018   1145.215      0.000      20.155      20.224
median_income         0.0011      0.029      0.038      0.970      -0.056       0.058
percent_college       0.2721      0.025     10.736      0.000       0.222       0.322
percent_married      -0.0028      0.025     -0.113      0.910      -0.052       0.046
rate_unemployment    -0.1513      0.023     -6.700      0.000      -0.196      -0.107
percent_lunch        -1.7966      0.022    -81.244      0.000      -1.840      -1.753
==============================================================================
Omnibus:                      834.356   Durbin-Watson:                   1.519
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2995.944
Skew:                           0.502   Prob(JB):                         0.00
Kurtosis:                       5.833   Cond. No.                         3.57
==============================================================================
```

In the text, explain which predictors have the strongest impact. For example:

- Student poverty (percent_lunch) has the largest standardized coefficient → strongest impact.
- Percent of adults with college degrees (+) and unemployment (-) are also significant.
- Median income and percent of children in married-couple families are not significant after controlling for other factors.

Discussion

According to the data, socioeconomic circumstances have a significant impact on academic achievement. Local unemployment and community education level are the next most important predictors, after student poverty. After scaling, income was not important on its own, although it probably interacts with other factors. The findings imply that programs aimed at reducing poverty and providing educational assistance in low-income areas could enhance academic performance..

Conclusions

1. School performance is influenced by multiple socioeconomic factors.
2. Percent of adults with college degrees positively predicts ACT scores.
3. Student poverty and local unemployment negatively predict ACT scores.
4. Approximately 62% of ACT score variation is explained by the model.
5. Addressing economic disadvantage and supporting education in low-income communities could improve school outcomes.

References

- EdGap dataset, GitHub
- Common Core of Data (CCD), Dropbox
- Python libraries: pandas, numpy, matplotlib, seaborn, plotly, statsmodels, scikit-learn