Prediction of school performance by Socio-economic factors

Raja Sandeep Mukkala
Professor: Dr. Brain Fischer
DATA-5100 Foundation of Data Science
Seattle University
October 22, 2025

Abstract

This study investigates whether school performance, measured by **average ACT scores**, can be predicted by socioeconomic factors.

Data are obtained from **EdGap.org**, the **National Centre for Education Statistics (CCD, 2016-17)**, the **U.S Census American Community Survey (ACS, 2017)** and merged using school identifier (NCESSCH). Apart from existing predictors present in dataset that are obtained from EdGap.org, new predictors **student-teacher ratio** and **percentage of households with internet access** were computed to extend the socioeconomic factors.

Descriptive statistics, exploratory visualizations, and regression analysis were performed to analyze the relationship between average ACT score and socioeconomic factors like percentage college education, unemployment rate, percentage of married couples, household median income, percent of students receiving free lunch, student-teacher ratio in schools and percent of households with internet access.

The regression model was statistically significant ($p < 0.001$) with an R^2 of 0.63 which confirms that socioeconomic factors explained 63% of ACT score variance, with free lunch percentage and unemployment rate being the strongest negative predictors of ACT performance, while more college educated adults and internet access households positively affected ACT scores. These findings highlight that community socioeconomic context significantly shapes the education achievement, suggesting that interventions addressing economic and digital inequality may improve school performance.

Introduction

Educational achievement across the United States continues to reflect persistent socioeconomic disparities. Standardized test outcomes such as the ACT often mirror community differences in income, education and resources access rather than individual ability alone. Understanding these

relationships is critical for designing equitable education policies and resource allocation strategies.

This study explores how school performance which is measured by average ACT score can be predicted by socioeconomic factors. For this analysis data is collected from multiple sources like **EdGap.org** which provides socioeconomic data (like unemployment rate, percentage of adults with college education, percent of children in married couple family, median household income, percentage of schools with free lunch) , from the **National Centre for Education Statistics** source we collected basic school information, number of students in each school, number of teachers in school and from **U.S Census American Community Survey** source we collected percentage of households with internet access across different states.

The aim is to quantify the relationship between community level socioeconomic conditions and school ACT performance.

Hypothesis is, schools in communities with higher income and educational attainment will achieve higher ACT scores, while those in economically disadvantaged or high unemployment areas will perform lower.

The following section reviews the literature that grounds these relationships in prior research.

Theoretical Background

Evidence links socioeconomic status to educational outcomes. Foundational work by Colmen (1966) and later by Reardon (2011) demonstrated that students from wealthier, better educated families consistently outperform peers from lower income backgrounds.

Income and Resources: Median household income, which is a general economic strength of a community, captures community's capacity to provide tutoring, technology and enrichment opportunities.

Parental Education: Higher percentage of college educated adults in a community represents strong parental support to children's in education.

Poverty Indicators: Percentage of schools with free lunch which mean schools in poor communities, which indicates financial hardship in community, which constraints learning environment.

School Resources: student-teacher ratio means classroom resources availability. Lower ratios mean more teachers in a class which enhances more care towards students.

Digital Access: Internet access at households means students can learn more through online sources.

All these variables that are discussed affects students' education and they interact, as economic stability often enhances access to educational resources and internet connectivity.

Hence investigating how these different variables together effects students' performance in ACT score can provide more insights into US education system.

Methodology

Data Source:

- **EdGap.org** - Collected ACT scores and socioeconomic indicators.
- **National Centre for Education Statistics (CCD, 2016-17)** - Collected basic School information, number of students (Student information) and teachers (Staff information) in each school.
- **U.S Census American Community Survey** - Collected percentage of households with internet or broadband subscription (Internet access information).

Data Preparation:

From the data that are collected from the above data sources only required columns are selected from school information, staff information and internet access information. Column names are renamed, and state names are changed to abbreviations to maintain symmetry and uniformity and these datasets are merged on school ID(NCESSCH) to ensure record level consistency across sources.

Apart from the existing socio-economic predictors that are available in datasets, computed two new Predictors:

  - Student and teacher ratio per school = Student count/Teacher count per school
  - Internet Access = Percent of households in a state with internet access.

Records with missing ACT score are removed, replaced incorrect and negatives values by NaN, identified duplicates, missing values and imputed missing predictor values.

Exploratory Data Analysis:

Descriptive statistics, correlation heatmaps and pair plots were used to visualize associations between ACT scores and predictors variables. Boxplots compared socioeconomic distributions across schools

Modeling Approach:

Three regression models were tested:

- **Simple Linear Regression** between average ACT score and median income.

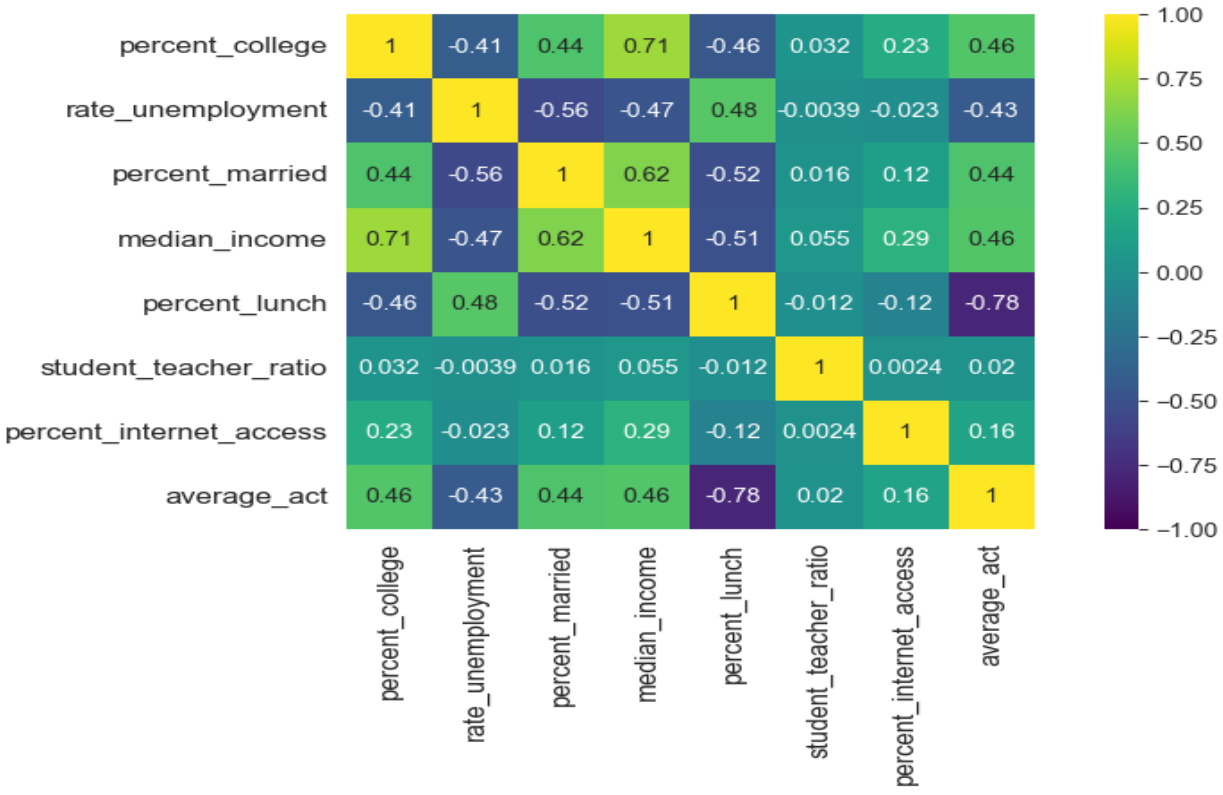
- **Quadratic Regression** to capture possible nonlinearity between median income and average ACT score.
- **Multiple Linear Regression** to investigate how different socio-economic factors (rate of unemployment, percent college education, percent married, median household income, percent free lunch, student-teacher ratio and percentage of internet access) affect average ACT score. Predictors with statistically significant coefficients ($p < 0.05$) were retained in the reduced MLR model.

ACT_{avg} = $beta_0$ + $beta_1$(Unemployment Rate) + $beta_2$(Percent College Education) + $beta_3$(Percent Free Lunch) + $beta_4$(Percent Internet Access household)

Complete analysis is performed in Python using pandas, scikit-learn, stats models, matplotlib, numpy and seaborn

Computational Results

Heatmap Correlation Results:



- **Free lunch percent** shows strong negative correlation with **average ACT.**
- Schools around low-income population have low ACT scores.
- **Median income and college education** have positive correlation with ACT scores which means schools that are around high income and better educated population has good ACT scores.

- **Student Teacher Ratio**, **Percent Internet Access** have some positive correlations.
- **Unemployment Rate** is negatively correlated, which means higher unemployment rate led to lower ACT scores.

Regression Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.630
Model:                            OLS   Adj. R-squared:                  0.630
Method:                 Least Squares   F-statistic:                     3075.
Date:                Thu, 23 Oct 2025   Prob (F-statistic):               0.00
Time:                        09:22:29   Log-Likelihood:                -13307.
No. Observations:                7227   AIC:                         2.662e+04
Df Residuals:                    7222   BIC:                         2.666e+04
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
Intercept                         20.2986      0.018   1130.711      0.000      20.263      20.334
rate_unemployment_normalized      -0.1355      0.021     -6.397      0.000      -0.177      -0.094
percent_college_normalized         0.2532      0.021     11.843      0.000       0.211       0.295
percent_lunch_normalized          -1.7700      0.022    -81.819      0.000      -1.812      -1.728
percent_internet_access_normalized     0.1184      0.019      6.385      0.000       0.082       0.155
==============================================================================
Omnibus:                      857.639   Durbin-Watson:                   1.493
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3217.567
Skew:                           0.562   Prob(JB):                         0.00
Kurtosis:                       6.070   Cond. No.                         1.95
==============================================================================
```

Multiple linear regression:

average ACT ~ normalized rate of unemployment + normalized percent of college education + normalized percent of lunch + normalized percent of internet access

- The R-squared value 0.63 indicates that 63% variance in ACT scores can be explained by predictors in this multiple regression model.
- Percent of students receiving free or reduce lunch is a strong negative predictor. Which means higher poverty levels are strongly associated with lower ACT performance.
- Unemployment rate shows negative relationship with ACT scores.
- Percent of internet access has some positive relationship with ACT scores.
- College educated household positively correlated to ACT scores, that is more college educated adults in a community leads to higher ACT performance of schools nearby.

Discussion

The analysis suggests that economic prosperity and educational standards in the community strongly affecting the academic performance.

Schools in wealthier areas like around higher median income households and college educated population (coef = 0.2532, $p < 0.001$) are generally performing better.

And schools with higher percentage of free lunch (coef = -1.7700, $p < 0.001$) which is a sign of poverty are performing lower in ACT scores which indicates that socio-economic disadvantage like financial hardship and limited access to learning resources negatively impacting academic outcome of students.

Though weaker correlation between student-teacher ratio and average ACT scores exists but it suggests that less teachers in school or large classroom size affect student learning capabilities. However, income and educational factors have more influence than this.

While internet access provides important opportunities in today's digital world for online learning, its relatively small effect (coef = 0.1184, $p < 0.001$) on average ACT scores indicates that connectivity alone may not enhance outcomes unless coupled with digital literacy and structure academic use.

Limitations and Implications:

This analysis uses cross sectional data (2016 - 2017) hence causal inference is limited and regional variations may also affect generalizability. However, the results emphasize that addressing **economic inequality**, **education among adult population** and **digital access** are essential strategies to improve equitable academic outcomes.

Conclusion

This analysis demonstrates that school performance is significantly associated with socio-economic factors within the community. Higher income, good education and stable households within the community positively affects ACT performance. While poverty, instability(unemployment) within the community negatively affects ACT performance. These findings confirm that improving economic stability, education among adults and improving access to learning resources will improve students' achievement. Addressing these socio-economic disparities could ultimately reduce achievement gaps and promote long-term educational equity.

References

Coleman, J.S (1966). Equality of Education Opportunity Report. U.S. Department of Health, Education and Welfare.

Reardon, S.F. (2011). The Widening Academic Achievement Gap Between the Rich and the Poor. Community Investments, 23(2), 19-39.

EdGap.org: EdGap. (2016). Education Opportunity Project Data Porta*l.* Retrieved from https://edgap.org

National Center for Education Statistics (NCES): U.S. Department of Education, National Center for Education Statistics. (2017). Common Core of Data (CCD): Public Elementary/Secondary School Universe Survey Data, 2016–17 [Data set]. Retrieved from https://nces.ed.gov/ccd/pubschuniv.asp

U.S. Census Bureau – American Community Survey (ACS): U.S. Census Bureau. (2017). American Community Survey (ACS) 5-Year Estimates [Data set]. Retrieved from https://www.census.gov/programs-surveys/acs