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 which is measured by ACT scores can be predicted by socioeconomic factors.

Here data is obtained from EdGap.org, the National Centre for Education Statistics (CCD, 2016-17) and the U.S Census American Community Survey (ACS, 2017). These datasets are merged by school identifier (NCESSCH) and apart from existing predictors present in dataset that are obtained from EdGap.org additional predictors like student-teacher ratio and percentage of households that have internet access by state were computed to extend the socioeconomic factors.

Descriptive statistics, boxplots, exploratory data analysis 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 couple, household median income, percent of students receiving free lunch, student-teacher ratio in schools and percent of households that has internet access.

Results showed that percentage of students receiving free lunch, unemployment rate shows negative relationship on ACT scores whereas presence of more college educated adults and internet access households positively affected ACT scores.

Findings confirm that socioeconomic factors in some way contribute to ACT score that dictate the performance of schools.

Introduction:

This study examines whether 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 are collecting basic school information, number of students in each school, number of teachers in school and from U.S Census American Community Survey source we are collecting percentage of households with internet access across different states.

The obtained datasets are merged by school ID(NCESSCH) to ensure consistency, one of the new predictors which is student-teacher ratio is added into the merged data, missing ACT value rows are removed, out of range values are replaced by NaN and missing predictor values are imputed to make it ready for exploratory and regression analysis to solve the main agenda of this problem that is how different socioeconomic factors (like percentage college education, unemployment rate, percentage of married couple, household median income, percent of students receiving free lunch, student-teacher ratio in schools and percent of households that has internet access) affects ACT performance in high schools across USA.

Theoretical Background:

Research shows socioeconomic status like income, education level and access to learning resources strongly influence student achievement. Median household income, which is a general economic strength of a community, percentage of college education in a community represents parental support to children's in education, percentage of schools with free lunch which mean schools in poor communities, student-teacher ratio means classroom resources availability, internet access at households means students can learn more through online sources. All these that are discussed either affects or supports students' education. Hence investigating how these different variables effects students' performance in ACT score can provide more insights into US education system.

Methodology:

Data Source:

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

Data Preparation:

- Selected required columns from school information, staff information and internet access information datasets.
- Collected datasets 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).
- New Predictors are created:
    - Student and teacher ratio per school = student count/teacher count per school
    - Internet Access = Percent of households in a state with internet access.
- Rows with missing ACT score are removed, checked variable ranges and replaced incorrect and negatives values by NaN.
- Checked for duplicates, missing values and missing predictor values are imputed.

Exploratory Data Analysis:

- Heatmap to display the correlation between average act and other variable.
- Pair plot to visualize relationship between numerical predictors and average ACT scores by charter and non-charter schools.
- Box plot to compare the spread of socioeconomic proportions across schools.
- Box plot to analyze the spread of median household income across high schools.

Modeling:

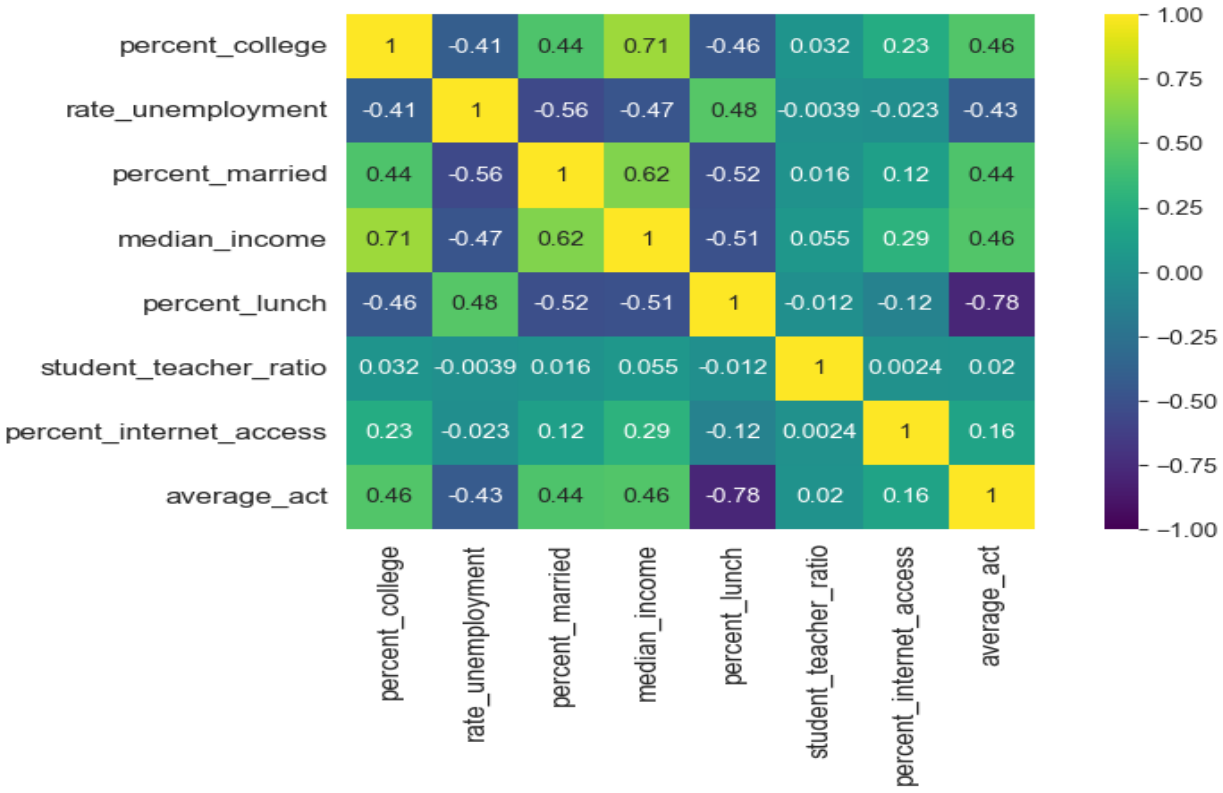
- Performed regression analysis:
    - A simple linear regression was performed between average ACT score and median income.
    - To capture nonlinearity a quadratic regression model was fitted between median income and average ACT score.
    - To investigate how different socio-economic factors affect average ACT score, we implemented:
        - Multiple linear regression model to investigate how different socio-economic factors like rate of unemployment, percent college education, percent married, median household income, percent free lunch, student-teacher ratio and percentage of internet access against average ACT score.
        - After assessing different predictors significance, implemented:
            - Reduced multiple linear regression with significant predictors: **average_act ~ rate_unemployment + percent_college + percent_lunch + percent_internet_access**

Software used:

- Python libraries: Pandas, scikit-learn, stats models, Matplotlib, NumPy and Seaborn

Computational Results:

Heatmap Correlation Results:



- **percent_lunch** shows strong negative correlation with **average_act**
- Schools with more low-income students have low ACT scores.
- **median_income, percent_college** have positive correlation with ACT scores.
- Schools that are around high income and better educated population has good ACT scores.
- **student_teacher_ratio**, **percent_internet_access** have some positive correlations.
- **rate_unemployment** is negatively correlated, which means higher unemployment rate led to lower ACT scores.

Multiple linear 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 ACT performance.

Schools in wealthier areas, with higher median income areas and college educated population are generally performing better.

And schools with higher percentage of free lunch which is a sign of poverty are performing lower in ACT scores which indicates that socio-economic disadvantage negatively impacting academic outcome of students.

The weaker correlation between student-teacher ratio and average ACT scores suggest that less teachers in school or large classroom size might affect student learning capabilities, but income and educational factors have more influence than this, in the same way internet access even though important in today's digital world to learn much information from online yet has only very little effect on average ACT scores.

Conclusion:

Finally, this analysis finds that school performance is significantly associated with socio-economic factors within the community. Higher income, good education and stable households within the community affects positively ACT performance. While poverty, instability(unemployment) within the community negatively affects ACT performance. Which indirectly suggests that improving access to resources in economically disadvantage communities might improve school performance.

References

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