Abstract

The main purpose of this project is to analyze the relationship between certain socioeconomic factors and the academic performance of U.S. high school students. Although many factors can affect a student's grades, this study focuses on variables such as the unemployment rate in their district, the percentage of adults with a college degree, the median income level, the percentage of married adults, the percentage of students on the reduced lunch program, and the number of full-time teacher equivalents in their schools. At the conclusion of our analysis, we want to be able to identify which socioeconomic factors are most strongly associated with academic performance, quantify their influence on ACT scores, and describe how these socioeconomic factors vary across schools and districts and how they are associated with differences in ACT scores.

Introduction

In the news, we often hear the term *"inequality,"* which is most commonly associated with wealth. However, several other forms of inequality receive less attention, one of which is educational inequality. According to *ScienceDirect,* "Educational inequalities refer to the disparities in academic achievement and educational opportunities among children from different socioeconomic backgrounds, influenced by factors such as parental investment and economic conditions." This is a major issue in the United States that often does not receive adequate discussion.

The ACT is a standardized test used in U.S. college admissions to measure students' skills in English, math, reading, and science, with an optional writing section. Scores range from 1 to 36, with the Composite score calculated as the average of the English, math, reading, and science sections. Additional STEM and ELA scores are derived from relevant components. ACT scores are aligned with College and Career Readiness Standards, and score reports include detailed performance by reporting category.

What constitutes a "good" ACT score varies by college selectivity, but national averages and benchmarks provide general guidance. While scores are important for admissions decisions, colleges also consider GPA, extracurricular activities, essays, and recommendations. Some colleges accept SAT scores in place of ACT scores, while others accept both.

Data Description

A review of the most recently published 2024 ACT average scores by state reveals notable differences between states. What could explain these disparities? Examining schools more closely, we observe variations captured in our EdGap dataset, obtained from **EdGap.org**—a platform that provides maps, visualizations, and resources to encourage discussion about the education gap in the United States. The platform offers access to data on various socioeconomic variables of interest, enabling analysis of factors within districts and individual schools.

The EdGap dataset identifies schools only by their National ID. The ID alone does not provide information such as the school's location, name, or type. To obtain more detailed information linked to these IDs, additional data were retrieved from the **National Center for Education Statistics (NCES),** the primary statistical agency of the U.S. Department of Education responsible for collecting, processing, and analyzing educational data. Through this source, we gathered further details about each school corresponding to its ID.

In addition to the socioeconomic variables present in the EdGap dataset, it was also important to understand the influence of other socioeconomic factors. The variable we chose to study, in order to assess its relationship with ACT scores, was the number of **full-time equivalent (FTE)** teachers a school employs. This number was calculated by counting one full-time teacher as 1.0 and one part-time teacher as 0.5, meaning two part-time teachers are equivalent to one full-time teacher. The data was also retrieved from the National Center for Education Statistics(NCES) website. It was important to ensure that the data had the correct date range, school type, and record type, and that it included the attributes we were interested in.

The downloaded datasets were stored locally and subsequently imported into our notebook to begin the data cleaning process. The NCES dataset, which contained school information, included an additional header row that required adjustment to retain only the subheaders relevant to our analysis. To ensure that the datasets were suitable for the intended analyses, we conducted a preliminary analysis to explore the relationship between the target variable (ACT score) and the predictor variables

(socioeconomic factors). This step was important to confirm that the dataset was appropriate for addressing our research questions before investing time in detailed data cleaning and preprocessing.

Initially, it was indicated that our EdGap dataset included several key socioeconomic variables; however, we were also interested in another important variable—the full-time equivalent (FTE) teacher count—which was included in a separate dataset. To create a pairplot comparing ACT scores with FTE teachers, it was necessary to temporarily merge the datasets so that the FTE variable would appear within the same data frame.

Before merging, we ensured that the key which was the school's National ID present in both datasets had a consistent data type. The ID fields were converted to floats to maintain consistency during the matching and joining process. The datasets were then merged using a left-join, with the EdGap dataset as the primary table .

A pairplot was subsequently created to visualize the relationships between the socioeconomic variables and ACT scores. From the pairplot, we observed that some degree of correlation exists between each socioeconomic variable and the ACT score. This finding confirmed that our data were suitable for further analysis.

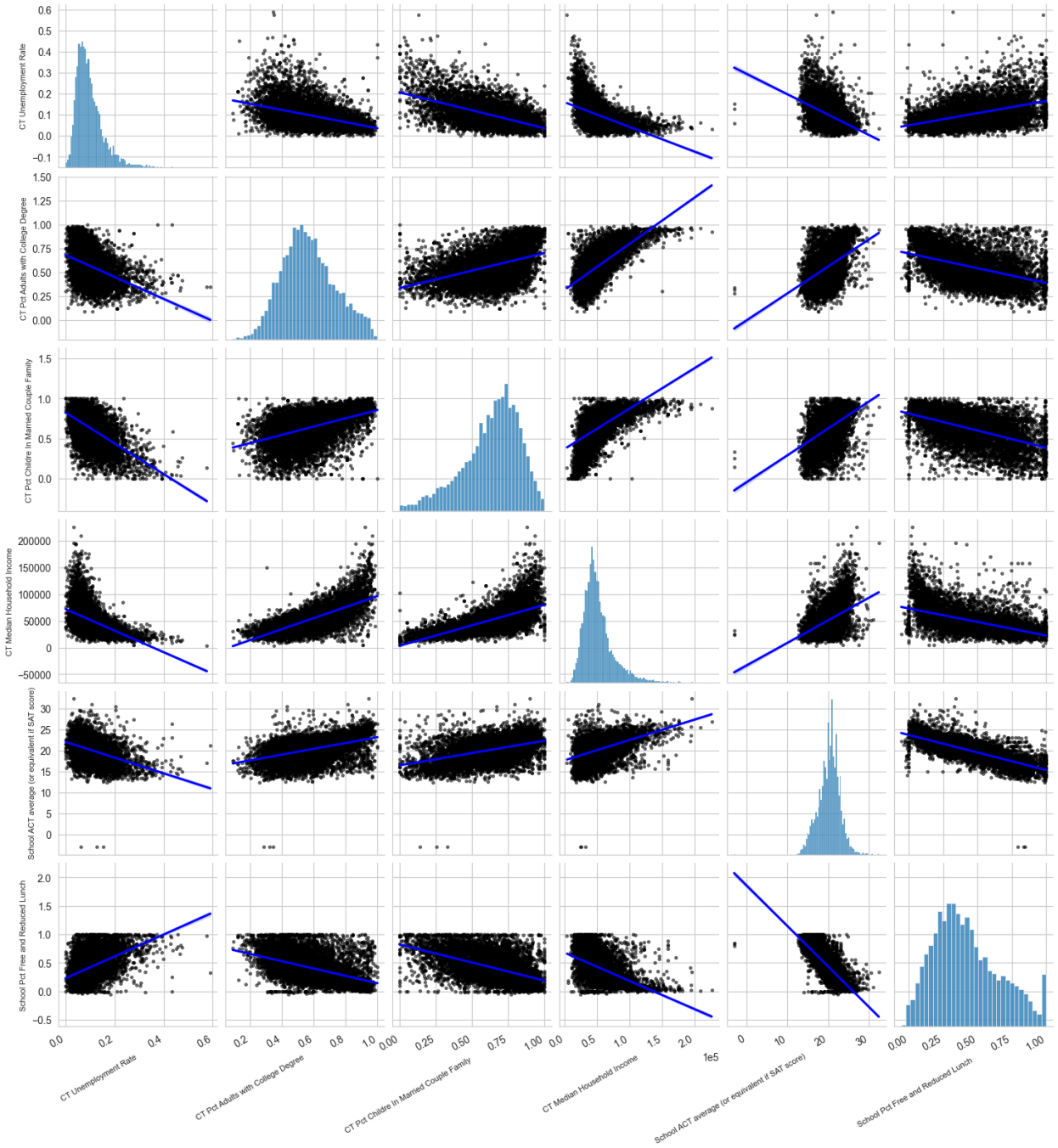

Figure 1 - Pairplot : socioeconomic variables vs ACT scores

As part of the process of further cleaning the data, the columns were given descriptive names that were more meaningful than the default ones. Quality control was also performed to ensure that all values fell within their correct ranges. For example, our dataset included average ACT scores, some of which were outside the accepted range of 1 to 36. These cells were replaced with NaNs. Similarly, some values for the percentage of students on the school's reduced lunch program were outside the normal bounds of 0 to 1.

It was also identified that some records in the dataset did not correspond to high school students. Filtering was therefore performed to include only data for high school students.

Additionally, to handle missing data in the dataframe, an **IterativeImputer** was used to fill cells containing NaN values. This is a data imputation technique that models each feature with missing values as a function of the other features to provide more accurate estimates. Finally, the cleaned dataset was exported for further analysis.

Analysis

To better understand the relationship between the socioeconomic variables and ACT scores, a correlation matrix was generated. For the rate of employment(rate_unemployment), we observed r=−0.43r = -0.43r=−0.43, indicating a negative relationship between employment rate and ACT scores(average_act). This suggests that as unemployment levels increase among adults within a district, the average ACT score tends to decrease. However, it is important to note that this does not establish a causal relationship, as other factors are likely influencing ACT scores.

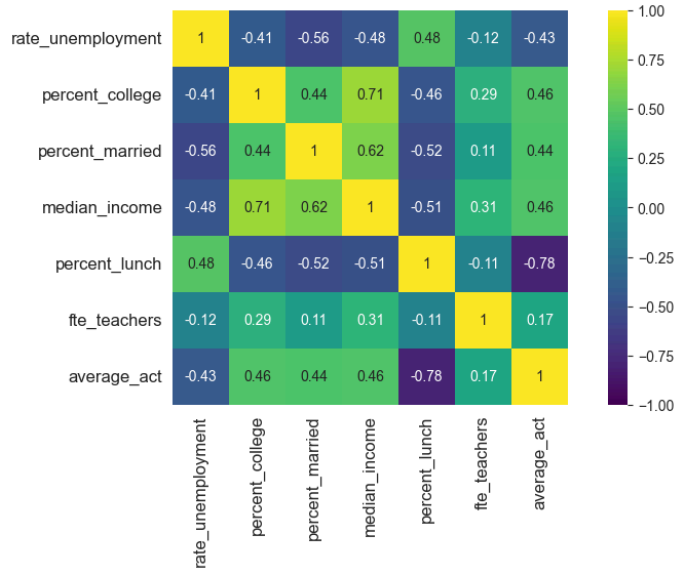

Figure 2 - Correlation Matrix

Looking at the percentage of students on reduced lunch (percent_lunch), the correlation coefficient is r = −0.78. This indicates a strong negative correlation between percent_lunch and ACT scores. Among the socioeconomic variables, percent_lunch has the largest magnitude of correlation, which suggests that it may be one of the strongest predictors of ACT performance in the dataset. Although it is tempting to assume a causal relationship, we cannot do so because correlation does not demonstrate causation. Other confounding variables could be influencing ACT scores, and even though a relationship exists, it is unclear which variable, if any, directly affects the other. The correlation coefficient describes the relationships that exist between our variables, but we need to analyze the variability in the ACT scores as explained by our predictor variables. The coefficient of determination can help us achieve this. The coefficient of determination (R^2) was calculated for each predictor variable to assess how much of the variance in ACT scores could be explained by that variable alone.

To compute the ordinary least squares(OLS) regression models we the smf.ols function in Python statsmodels library. Lets analyze some of our predictor variables:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.210
Model:                            OLS   Adj. R-squared:                  0.210
Method:                 Least Squares   F-statistic:                     1920.
Date:                Wed, 22 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:36:27   Log-Likelihood:                -16049.
No. Observations:                7227   AIC:                         3.210e+04
Df Residuals:                    7225   BIC:                         3.212e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      16.3055      0.095    171.953      0.000      16.120      16.491
percent_college     6.9688      0.159     43.819      0.000       6.657       7.281
==============================================================================
Omnibus:                      353.848   Durbin-Watson:                   1.208
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              526.162
Skew:                          -0.442   Prob(JB):                    5.56e-115
Kurtosis:                       3.982   Cond. No.                         8.10
==============================================================================
```

From the regression results, we can see that about 21% of variance in the response variable - average ACT scores - is explained by the percentage of adults with a college degree.

The positive coefficient of 6.97 would indicate that schools in areas with higher percent_college tend to have higher ACT scores. But this does not mean if we increase the percent_college, it will cause ACT scores to arise. The p-value also being less than 0.05 does indicate that the predictor is statistically significant.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.614
Model:                            OLS   Adj. R-squared:                  0.614
Method:                 Least Squares   F-statistic:                 1.149e+04
Date:                Wed, 22 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:33:10   Log-Likelihood:                -13461.
No. Observations:                7227   AIC:                         2.693e+04
Df Residuals:                    7225   BIC:                         2.694e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     23.7429      0.037    641.759      0.000      23.670      23.815
percent_lunch -8.3902      0.078   -107.187      0.000      -8.544      -8.237
==============================================================================
Omnibus:                      842.255   Durbin-Watson:                   1.472
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2849.644
Skew:                           0.582   Prob(JB):                         0.00
Kurtosis:                       5.848   Cond. No.                         5.02
==============================================================================
```

The model included 7,227 observations and produced an R^2 of 0.614, indicating that approximately 61.4% of the variability in average ACT scores can be explained by the proportion of students on reduced lunch. The p-value indicates that the predictor is statistically significant. We can conclude that percent_lunch is a strong predictor of ACT score in the data set.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.211
Model:                            OLS   Adj. R-squared:                  0.211
Method:                 Least Squares   F-statistic:                     1932.
Date:                Wed, 22 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:33:11   Log-Likelihood:                -16044.
No. Observations:                7227   AIC:                         3.209e+04
Df Residuals:                    7225   BIC:                         3.211e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     17.8035      0.063    284.773      0.000      17.681      17.926
median_income  4.73e-05   1.08e-06     43.959      0.000    4.52e-05    4.94e-05
==============================================================================
Omnibus:                      191.606   Durbin-Watson:                   1.274
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              362.301
Skew:                          -0.196   Prob(JB):                     2.13e-79
Kurtosis:                       4.024   Cond. No.                     1.39e+05
==============================================================================
```

This regression result indicates that the coefficient of determination (R^2) for median_income is 0.211, meaning that approximately 21.1% of the variation in ACT scores is explained by median income.

We also create a multiple linear regression to …….. that takes into account all the socioeconomic predictors:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.631
Model:                            OLS   Adj. R-squared:                  0.631
Method:                 Least Squares   F-statistic:                     2062.
Date:                Wed, 22 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:33:12   Log-Likelihood:                -13293.
No. Observations:                7227   AIC:                         2.660e+04
Df Residuals:                    7220   BIC:                         2.665e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept           22.6068      0.137    164.550      0.000      22.337      22.876
rate_unemployment   -2.2591      0.402     -5.616      0.000      -3.048      -1.471
percent_college      1.5947      0.158     10.114      0.000       1.286       1.904
percent_married      0.0228      0.134      0.170      0.865      -0.239       0.285
median_income    -2.103e-06   1.23e-06     -1.714      0.087   -4.51e-06    3.02e-07
percent_lunch       -7.6402      0.096    -79.206      0.000      -7.829      -7.451
fte_teachers         0.0037      0.000      8.322      0.000       0.003       0.005
==============================================================================
Omnibus:                      960.120   Durbin-Watson:                   1.488
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3481.073
Skew:                           0.645   Prob(JB):                         0.00
...
```

From our regression result, we see that percent_lunch has the largest correlation coefficient r = -7.64 compared to the standard error. The other predictor variables do not have a coeff

Conclusion

We can conclude that strongest predictor of ACT scores given that it has the largest magnitude of coefficient relative to standard error, and is highly significant given p-value of 0. It also explains about 61.4% of the variance individually. In simple terms, in contributes more to the variability in our ACT scores.

The other predictor variables such as unemployment rate(rate_unemployment), percent of adults with a college degree(percent_college), and full-time teachers equivalents also contribute meaningfully to the model.

Note that given that there are cofounding factors, we cannot establish a causal relationship between variables.

References

ScienceDirect. "Educational Inequality - an Overview | ScienceDirect Topics." *Sciencedirect.com*, 2015, www.sciencedirect.com/topics/social-sciences/educational-inequality.

"The National Center for Education Statistics: Who We Are | IES." *Ed.gov*, 2023, nces.ed.gov/national-center-education-statistics-nces/about. Accessed 22 Oct. 2025.