Education Project

Problem Statement

This project addresses inequality of educational opportunity in U.S. high schools. Here we will focus on average student performance on the ACT exams that students take as part of the college application process. Specifically, whether any socio-economic factors that affect the average ACT scores of the schools.

Analysis

Loading the datasets:

To explore the relationship between socio-economic factors and average ACT scores across U.S. high schools, I combined information from three key datasets.

1. EdGap Data:
   This dataset provides important socio-economic indicators at the school or district level, including median household income, percentage of married adults, percentage of students eligible for free or reduced-price lunch, unemployment rate, and percentage of adults with a college degree. These variables serve as potential predictors of educational outcomes.

2. School Information Data:
   This dataset includes detailed information about each school, such as school type (public or private), charter status, state, zip code, school year, and state identification number. These attributes help categorize and contextualize schools in the analysis.

3. Public School Data:
   This dataset provides additional school-level characteristics, including student–teacher ratio, grade levels offered, and other common school attributes relevant to understanding the learning environment.

All three datasets were merged using the common identifier NCESSCH (National Center for Education Statistics School ID) to create a comprehensive dataset for analysis. This integrated dataset enables a deeper investigation into how socio-economic conditions correlate with average ACT performance across schools.

Explore the contents:

To begin the analysis, I performed several initial exploratory checks to understand the structure and quality of the data. This included:

- Inspecting the first five rows of each dataset to get an overview of the variables and their contents.
- Checking the number of rows and columns to understand the dataset dimensions.

- Identifying the number of missing values for each variable to assess data completeness.
- Reviewing the data types of all variables to ensure they were correctly formatted for analysis.

Next, I created a pair plot to visually explore relationships among the key variables in the EdGap dataset. This helped identify potential correlations between socio-economic factors and the average ACT score. The presence of noticeable linear patterns in the plots confirmed that the EdGap dataset is highly relevant for further analysis.

Data Preparation:

For this analysis, I retained all columns from the EdGap dataset, as each feature directly relates to the problem statement and may influence ACT scores. From the other two datasets, I selected only the most relevant variables:

- From the School Information dataset: NCESSCH, school_type, charter, state, zip_code, and school_year.
- From the School Data dataset: student_teacher_ratio and NCESSCH.

Next, I standardized the column names by converting them to lowercase and renaming them to more descriptive and readable labels. The common key NCESSCH was renamed to id for consistency across all datasets.

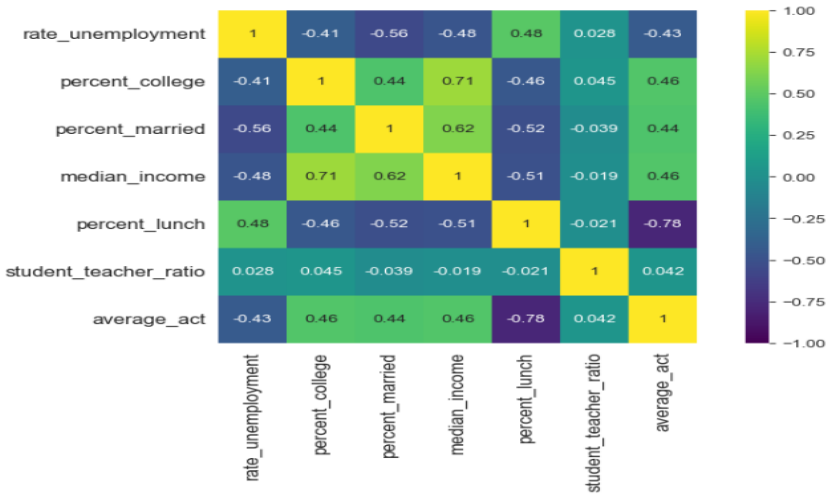
The id column in each dataframe was also converted to the object data type to ensure smooth merging. After that, I merged the School Information and School Data dataframes with the EdGap data using id as the primary key.

During data cleaning, I checked for out-of-range values, and found that some variables—specifically average ACT score, percent lunch, and student–teacher ratio—contained negative values. These were replaced with NaN to avoid skewing the results.

Since the analysis focuses on high schools, I filtered the combined dataset to include only rows corresponding to the high school level. To handle missing data, I used the IterativeImputer from the *scikit-learn* library, which predicts missing values using model-based estimation for higher accuracy. Finally, I created a map visualization to show the distribution and density of schools across different states included in the dataset.

Exploratory data analysis:

I started by plotting a correlation heatmap to understand how the variables relate to each other. The heatmap revealed a strong negative correlation between percent lunch and average ACT scores, and a positive correlation between median income and percent of adults with a college degree. In contrast, student–teacher ratio showed no significant linear relationship with other variables, indicating a weak or random association.



Next, I generated a pair plot to further explore linear relationships among variables and added charter status as the hue to observe how charter and non-charter schools differ. From the plot, I noticed that average ACT scores are approximately normally distributed, and again, there was no visible relationship between student–teacher ratio and other features—its distribution formed a nearly straight line.

To check for outliers, I used boxplots. Because variables like median income and student–teacher ratio have very different ranges, I examined them separately for better visualization. The boxplot for percent lunch displayed a well-defined distribution, while the student–teacher ratio plot appeared mostly flat with several outliers.

Modelling:

Simple linear regression:

To begin modeling, I first created a scatter plot with a regression line between median income and average ACT score to visually assess the relationship. I selected median income as the independent variable since it is a key socio-economic indicator likely to influence academic performance.

Using the StatsModels library, I fitted a simple linear regression model with *median income* as the predictor and *average ACT score* as the response variable.

The model results showed an R-squared value of 0.211, indicating that median income alone explains about 21% of the variation in average ACT scores. The p-value was less than 0.001, suggesting that the relationship is statistically significant and not due to random chance.

Although median income is a meaningful predictor, it is not a dominant factor — implying that other elements such as school resources, teacher quality, demographics, and community factors also play important roles in shaping student performance.

Next, I introduced a quadratic term to capture potential non-linear patterns in the relationship. However, the improvement in model performance was minimal, with the Mean Absolute Error (MAE) decreasing only slightly from 1.7129 to 1.697, and the Root Mean

Squared Error (RMSE) remaining unchanged.This suggests that while income has a statistically significant effect on ACT scores, its predictive power is limited when considered in isolation.

Multi-linear regression model:

After fitting a multiple linear regression model using average ACT score as the dependent variable and unemployment rate, percent college, percent married, median income, percent lunch, and student–teacher ratio as predictors, the model's performance improved significantly. The R-squared value increased to 0.628, indicating that these socio-economic and school-related factors together explain about 63% of the variation in average ACT scores.

Based on the p-values, the strongest predictors of ACT performance are percent lunch, percent college, and unemployment rate. In contrast, median income and percent married were not statistically significant once the effects of other factors were considered.

The model's accuracy also improved, with a Mean Absolute Error (MAE) of 1.145 and a Root Mean Squared Error (RMSE) of 1.529, showing better predictive performance compared to the simple linear model.

Finally, the residual plot revealed that the residuals were randomly distributed around zero, suggesting that the model fits the data reasonably well and that the assumptions of linear regression were not violated.

Fit a reduced model with significant predictors:

After fitting a reduced model using only the three significant socio-economic predictors, the model continues to perform strongly. It remains highly predictive, numerically stable, and easier to interpret compared to the full model.

- Percent of students on free/reduced lunch (a proxy for poverty) shows the largest negative effect on average ACT scores.
- Percent of adults with a college education has a positive effect, indicating higher education levels in the community are linked to better student performance.
- Local unemployment rate has a negative effect, suggesting higher unemployment is associated with lower ACT scores.

The dropped variables — *median income*, *percent married*, and *student–teacher ratio* — do not significantly influence ACT scores once the effects of these three main predictors are considered.

The model fit remains strong ($R^2 = 0.628$), meaning these three variables alone explain most of the variation in average ACT scores. The residual plot also shows that residuals are evenly clustered around the center, confirming a good model fit and consistent performance.

Scaling:

To bring all variables onto a comparable scale, I applied scaling to transform their ranges between 0 and 1.

After scaling, the overall model results remained consistent; however, normalizing the predictors helped clarify which variables have the strongest standardized impact on average ACT scores.

OLS Regression Results

Dep. Variable:	average_act	R-squared:	0.629
Model:	OLS	Adj. R-squared:	0.628
Method:	Least Squares	F-statistic:	2036.
Date:	Wed, 22 Oct 2025	Prob (F-statistic):	0.00
Time:	16:27:50	Log-Likelihood:	-13322.
No. Observations:	7227	AIC:	2.666e+04
Df Residuals:	7220	BIC:	2.671e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	22.6013	0.141	160.663	0.000	22.326	22.877
rate_unemployment	-2.3050	0.404	-5.706	0.000	-3.097	-1.513
percent_college	1.6987	0.158	10.760	0.000	1.389	2.008
percent_married	-0.0611	0.134	-0.456	0.648	-0.323	0.201
median_income	6.864e-08	1.21e-06	0.057	0.955	-2.3e-06	2.44e-06
percent_lunch	-7.5927	0.097	-78.414	0.000	-7.782	-7.403
student_teacher_ratio	0.0061	0.002	3.226	0.001	0.002	0.010

Omnibus:	865.659	Durbin-Watson:	1.483
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3087.559
Skew:	0.583	Prob(JB):	0.00
Kurtosis:	5.982	Cond. No.	1.34e+06

Multi-linear

OLS Regression Results

Dep. Variable:	average_act	R-squared:	0.628
Model:	OLS	Adj. R-squared:	0.628
Method:	Least Squares	F-statistic:	4064.
Date:	Wed, 22 Oct 2025	Prob (F-statistic):	0.00
Time:	16:47:02	Log-Likelihood:	-13328.
No. Observations:	7227	AIC:	2.666e+04
Df Residuals:	7223	BIC:	2.669e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.2986	0.018	1127.631	0.000	20.263	20.334
rate_unemployment_normalized	-0.1227	0.021	-5.801	0.000	-0.164	-0.081
percent_college_normalized	0.2826	0.021	13.504	0.000	0.242	0.324
percent_lunch_normalized	-1.7771	0.022	-82.000	0.000	-1.820	-1.735

Omnibus:	873.220	Durbin-Watson:	1.483
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3128.447
Skew:	0.587	Prob(JB):	0.00
Kurtosis:	6.002	Cond. No.	1.93

Scalling

Among all predictors, percent lunch showed the highest standardized coefficient (-1.77), confirming that it is the most influential factor affecting ACT performance.

Conclusion:

Areas with more educated adults, higher household incomes, and fewer economically disadvantaged students tend to have higher average ACT scores. Among all factors, the percent of students on free or reduced lunch stands out as the strongest predictor, serving as a clear indicator of economic disadvantage.

The analysis also shows that simpler models containing only the key socio-economic predictors perform almost as well as more complex models, suggesting that a few core factors explain most of the variation in ACT performance.

With an R^2 value of 0.628, the model captures a substantial portion of the variation in ACT scores, though not all. This indicates that while socio-economic conditions strongly influence academic performance, other factors—such as school quality, teacher effectiveness, and student motivation—also play meaningful roles.