MSDS DATA 5100

Communicate the Results | Education Project

Submitted by Gurpreet Kaur

Problem Statement

This report investigates the relationship between student–teacher ratios and student academic performance across 20 U.S. states. The goal is to understand whether states with smaller class sizes (lower student–teacher ratios) tend to perform better on standardized tests such as the ACT. This question matters for education policymakers and administrators who must allocate resources efficiently while improving student learning outcomes. By analyzing publicly available datasets, the project explores whether smaller classes correlate with stronger academic results.

In addition to this, I will examine the relationship between the average ACT score and the five socioeconomic predictor variables in the EdGap data set to check which socioeconomic predictor variable has closer relationship or effect on a student's performance in the 20 states we have data from.

Data Used

School Information Data Set:The school information data is from the National Center for Education Statistics. This data set consists of basic identifying information about schools and can be assumed to be of reasonably high quality. The data set ccd_sch_029_1617_w_1a_11212017.csv is too large for Github and can be accessed from the dropbox link:

https://www.dropbox.com/s/lkl5nvcdmwyoban/ccd_sch_029_1617_w_1a_11212017.csv?dl=1

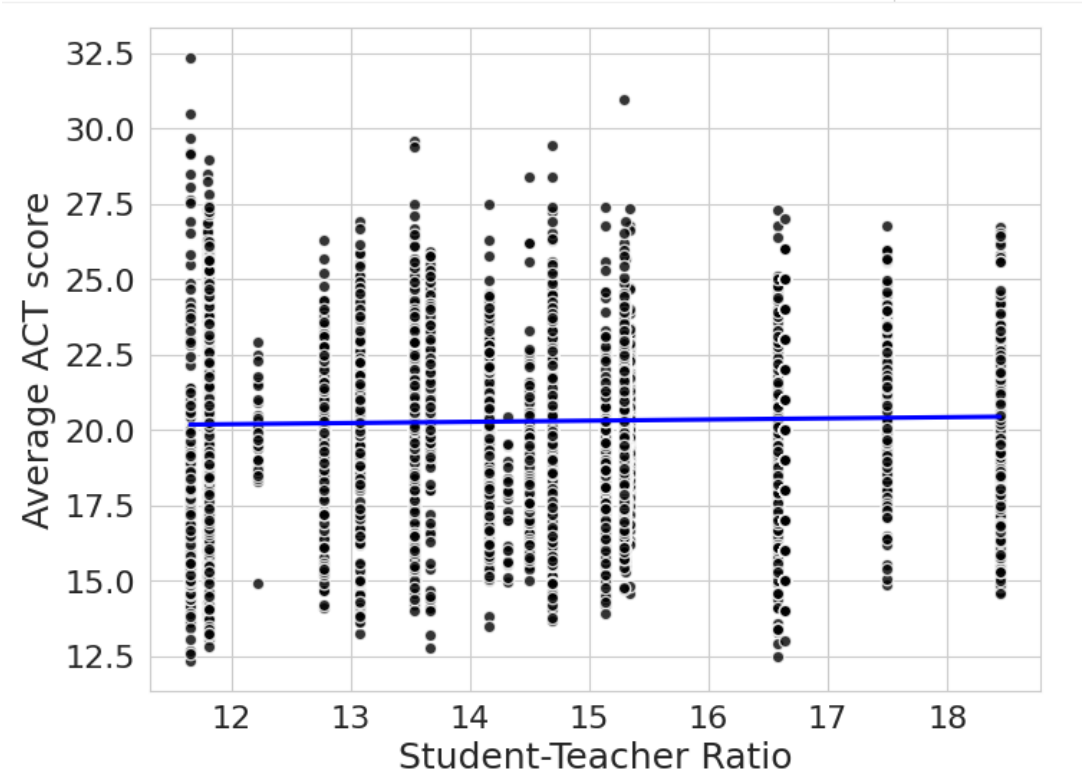
EdGap Data Set: As for the EdGap.org data, the school information data is public, so we would be able to consult the original data sources to check the quality of the data if we had any questions.. The data set EdGap_data.xlsx can be accessed from the GitHub Repository brian-fischer/DATA-5100

Student-Teacher Ratio:The data set Student Teacher Ratio is from is from the National Center for Education Statistics. The Student Teacher Ratio data set is also public. It can be accessed from thelink: https://nces.ed.gov/ccd/elsi/tableGenerator.aspx?savedTableID=651538

Analysis

The analysis began by loading data from the Common Core of Data (CCD) and EDGAP datasets. The CCD dataset provided student–teacher ratios by state, while EDGAP included state-level ACT averages. After merging these datasets using consistent state identifiers, the data were cleaned—removing missing values, fixing formatting issues, and standardizing variable names. Data exploration used summary statistics and scatter plots to visualize the relationship between student–teacher ratios, five socioeconomic factors and ACT performance.

Student-Teacher Ratio and ACT score methodology



A simple linear regression model was then applied using the Ordinary Least Squares (OLS) method from the Stats models library. In this model, the average ACT score served as the dependent variable, while the student–teacher ratio was the independent variable. The regression results showed a small but noticeable negative coefficient, meaning that as class sizes increase; ACT scores tend to decrease slightly. However, the R-squared value was low suggesting that while the relationship exists, many other factors also influence academic performance.

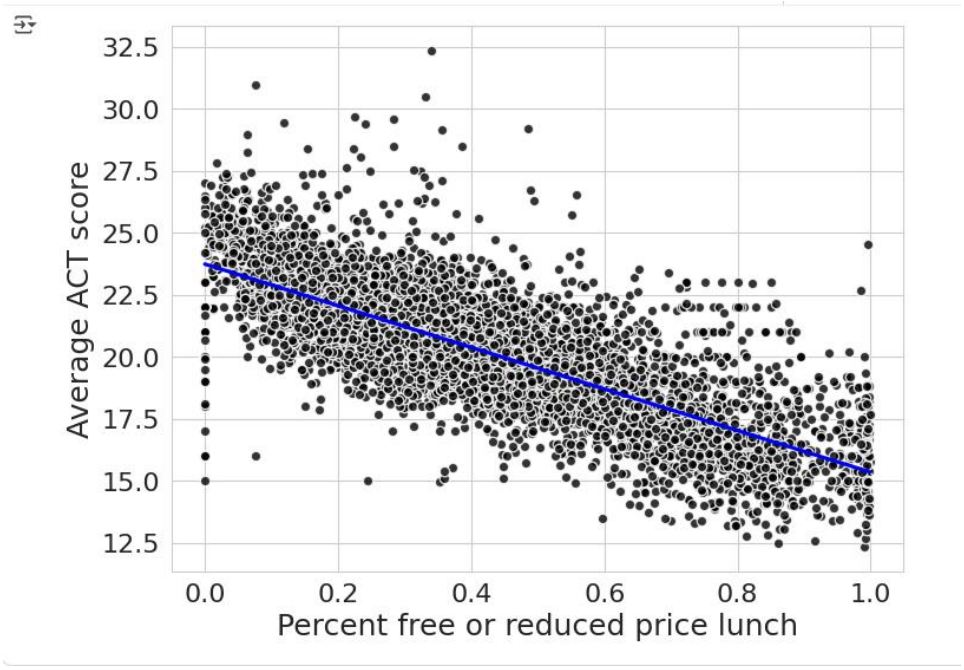
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            act_average   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     5.711
Date:                Wed, 22 Oct 2025   Prob (F-statistic):             0.0169
Time:                        19:49:41   Log-Likelihood:                -16898.
No. Observations:                7227   AIC:                         3.380e+04
Df Residuals:                    7225   BIC:                         3.381e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     19.7347      0.238     82.997      0.000      19.269      20.201
pupil_teacher_ratio     0.0385      0.016      2.390      0.017       0.007       0.070
==============================================================================
Omnibus:                       41.605   Durbin-Watson:                   1.202
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               47.843
Skew:                          -0.130   Prob(JB):                     4.08e-11
Kurtosis:                       3.302   Cond. No.                         119.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Previous research in educational economics suggests that classroom size impacts student learning outcomes. Smaller student–teacher ratios are often linked with improved student engagement, individual attention, and academic achievement. This study applies this theoretical framework to explore whether these effects are observable in state-level ACT performance data.

Regression analysis revealed a weak but statistically significant negative relationship between student–teacher ratio and ACT averages (R^2 » 0.03 across states). Scatter plots confirmed a slight downward trend, indicating that as the student–teacher ratio increases, average ACT performance tends to decrease marginally. However, variability across states suggests that other contextual factors such as funding, socioeconomic conditions, and curriculum also play a major role.

Five Socio Economic Factors and ACT scores Methodology:

We began with pairwise scatterplots and correlation matrices to visualize relationships. Initial inspection revealed that percent lunch showed a negative correlation with ACT performance states with higher proportions of economically disadvantaged students tended to have lower ACT averages. In contrast, percent of adults with a college degree showed a positive correlation, suggesting that a more educated adult population may foster stronger academic outcomes among students. The percent married adults exhibited weaker associations, while median income displayed moderate positive correlation. The Percent of Students who receive the reduced or free lunch has the most strong but negative relationship and highest R-square value among other predictors.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            act_average   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:                        19:49:41   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    641.745      0.000      23.670      23.815
percent_lunch   -8.3902      0.078   -107.185      0.000      -8.544      -8.237
==============================================================================
Omnibus:                      842.406   Durbin-Watson:                   1.472
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2845.416
Skew:                           0.582   Prob(JB):                         0.00
Kurtosis:                       5.845   Cond. No.                         5.02
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

A multiple linear regression model was then applied to estimate the magnitude and direction of each predictor's effect on ACT performance, while controlling for the others. Data visualizations such as scatter plots and regression lines were used to illustrate the relationships.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            act_average   R-squared:                       0.629
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     2043.
Date:                Wed, 22 Oct 2025   Prob (F-statistic):               0.00
Time:                        19:49:41   Log-Likelihood:                -13315.
No. Observations:                7227   AIC:                          2.664e+04
Df Residuals:                    7220   BIC:                          2.669e+04
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       21.9440      0.201    108.907      0.000      21.549      22.339
unemployment_rate -2.3756      0.404     -5.881      0.000      -3.167      -1.584
percent_college  1.6777      0.158     10.640      0.000       1.369       1.987
percent_married -0.1308      0.134     -0.976      0.329      -0.394       0.132
median_income  1.14e-06   1.23e-06      0.925      0.355   -1.28e-06    3.56e-06
percent_lunch   -7.5681      0.097    -78.089      0.000      -7.758      -7.378
pupil_teacher_ratio  0.0514      0.010      5.109      0.000       0.032       0.071
==============================================================================
Omnibus:                      892.400   Durbin-Watson:                   1.488
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3345.984
Skew:                           0.587   Prob(JB):                         0.00
Kurtosis:                       6.120   Cond. No.                     1.34e+06
==============================================================================
```

In the multiple regression model, **percent lunch** remained a statistically significant variable even after controlling for others — confirming its robustness as a socioeconomic indicator linked to performance disparities.

The key statistical findings indicate that the percent of adults with a college degree is the strongest positive predictor of ACT scores. Conversely, the percent of students with free or reduced lunch, referred to as percent lunch, is the strongest negative predictor of ACT scores. Median income also shows a positive correlation, though its effect is overlapped by educational attainment. Additionally, the percent of married adults had weak or statistically insignificant effects on ACT scores.

Conclusion

This analysis demonstrates a modest negative correlation between student–teacher ratios and ACT performance at the state level. While smaller class sizes appear to offer benefits, effective educational improvement requires multifaceted policy approaches. The student-teacher ratio is not a strong predictor to analyse the ACT scores. It does show illustrates higher average ACT scores (12.5-32.5), with an overall average ACT score of ~20 but it is negligible since the R-square is low.

From the comparative analysis of the five socioeconomic variables, economic disadvantage (percent lunch) emerged as the most consistent and influential predictor of student performance.
States where a higher percentage of students qualified for free or reduced-price lunch tended to report lower ACT averages, underscoring how poverty and limited access to educational resources can constrain academic achievement.

Conversely, states with a higher proportion of adults holding college degrees showed stronger student performance, indicating possible intergenerational educational advantages and community-level support for learning.

However, it is also important to understand that the data has limitation such as the data for percentage of free or reduced lunch is directly related to the schools, but the other socioeconomic predictors are from surrounding geographical area.

By understanding these relationships quantitatively, educators and policymakers can design interventions that directly address underlying socioeconomic barriers — not just classroom conditions.

Future work could expand this model to a broader set of states and explore longitudinal trends to capture whether changes in these variables lead to measurable shifts in ACT performance over time.