Sydney Golden

10/22/2025

Data 5100 Education Project

How does State along with Socioeconomic variables impact Average ACT scores?

**Abstract**

This project intends to assess how does location and different socioeconomic variables impact average ACT scores across the U.S. Different socioeconomic variables are assumed to impact a student’s performance on their assessments. Data from the EdGap and the National Center for Education Statistics were joined with state ACT data also from the National Center of Education Statistics to analyze the impact of different predictor variables. Modeling methods including linear regressions and residual plots were used to assess the statistical significance of the impact of different variables. Results showed that percent lunch was the predictor variable that had the greatest impact on ACT score, while a student’s location had some significance on their ACT score.

**Introduction**

Children are heavily impacted by their environment, whether it is their home or physical location. In an ideal world, all students would have equal access to education and opportunities. It is important to determine which socioeconomic variables may be influencing a student’s road to success to work towards possible solutions on small and large scales. In this study, we will determine if the state a student is located in along with different socioeconomic factors will have an impact on their ACT performance using data from the National Center for Education Statistics and EdGap.

**Theoretical background**

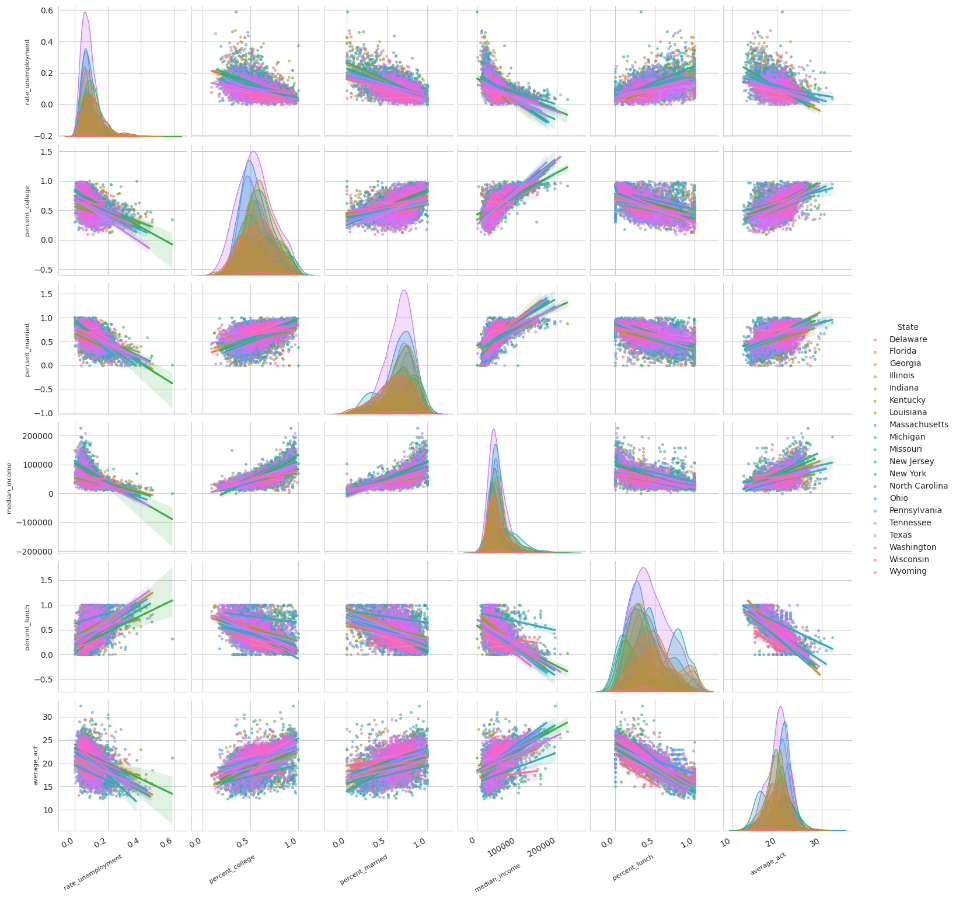
In regards to the quality of our data, the data from the census and department of education is assumed to be reasonably accurate. Please note that the EdGap data is not processed and can be prone to human error. Data from the ACT state scores only is from the years 2017 and 2021, and only the composite scores are taken into account in this study.

**Methodology**

Cleaned education data was utilized. The primary data set is the EdGap data set from [EdGap.org](https://www.edgap.org/#5/37.875/-96.987) and the secondary data set comes from the [National Center for Education Statistics](https://nces.ed.gov/ccd/pubschuniv.asp). These tow data sets have already been cleaned and merged into “df”. The ACT score data by state was also from the National Center for Education Statistics. The ACT score by state date was imported, formatted and tidyed. The ACT by state scores and the “df” date frame were merged by an inner join on state to keep data from matching states. Composite ACT data from year 2017 and 2021 were carried over to states in the education dataset. A correlation matrix of the predictor variables was created. Single input models of linear regressions were conducted for median income and state. From analysis of the correlation matrix, a reduced model was created to reflect the variables that are the strongest predictors. A residual plot was created to reflect the new reduced model. Finally, numerical variables were scaled and assessed.

**Computational results**

1. Pairplot



1. Reduced model analysis

OLS Regression Results

==============================================================================

Dep. Variable: average\_act R-squared: 0.679

Model: OLS Adj. R-squared: 0.678

Method: Least Squares F-statistic: 693.1

Date: Thu, 23 Oct 2025 Prob (F-statistic): 0.00

Time: 06:00:17 Log-Likelihood: -12793.

No. Observations: 7227 AIC: 2.563e+04

Df Residuals: 7204 BIC: 2.579e+04

Df Model: 22

Covariance Type: nonrobust

===========================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------

Intercept 18.7382 0.307 61.128 0.000 18.137 19.339

State[T.Florida] 3.9933 0.299 13.344 0.000 3.407 4.580

State[T.Georgia] 3.9667 0.300 13.206 0.000 3.378 4.555

State[T.Illinois] 3.2127 0.297 10.824 0.000 2.631 3.795

State[T.Indiana] 3.5758 0.301 11.870 0.000 2.985 4.166

State[T.Kentucky] 3.0654 0.308 9.960 0.000 2.462 3.669

State[T.Louisiana] 3.5836 0.309 11.615 0.000 2.979 4.188

State[T.Massachusetts] 3.9517 0.304 12.991 0.000 3.355 4.548

State[T.Michigan] 2.6234 0.298 8.813 0.000 2.040 3.207

State[T.Missouri] 4.0350 0.301 13.418 0.000 3.446 4.625

State[T.New Jersey] 3.2180 0.301 10.699 0.000 2.628 3.808

State[T.New York] 3.6176 0.304 11.883 0.000 3.021 4.214

State[T.North Carolina] 3.9385 0.299 13.163 0.000 3.352 4.525

State[T.Ohio] 4.4299 0.296 14.975 0.000 3.850 5.010

State[T.Pennsylvania] 3.6886 0.297 12.424 0.000 3.107 4.271

State[T.Tennessee] 2.9838 0.304 9.827 0.000 2.389 3.579

State[T.Texas] 3.9198 0.294 13.311 0.000 3.343 4.497

State[T.Washington] 4.4044 0.304 14.508 0.000 3.809 5.000

State[T.Wisconsin] 4.4206 0.300 14.723 0.000 3.832 5.009

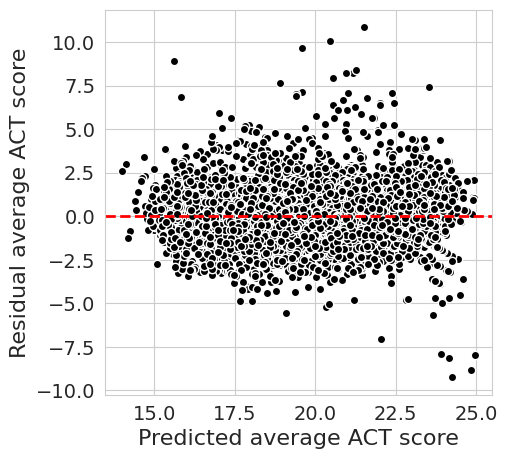
State[T.Wyoming] 1.5896 0.371 4.283 0.000 0.862 2.317

rate\_unemployment -0.9356 0.364 -2.567 0.010 -1.650 -0.221

percent\_college 1.9133 0.125 15.254 0.000 1.667 2.159

percent\_lunch -7.6812 0.095 -80.772 0.000 -7.868 -7.495

1. Residual model



**Discussion**

The first pairplot shows a clear linear relationship between the state an ACT score was taken and the average score. Some states will impact the ACT more than others.

In the reduced model, all states have a p value that is less than or equal to 0.01, showing a great statistical significance between the state a student took their exam and their ACT score. Percent lunch is shown to have the greatest impact on a student’s ACT score, with the strongest coefficient of the all (-7). The high R squared values shows that all the reduced predictor variables together created a statistical significance.

The residual model displays points all somewhat along the dotted red line, showing that there is most likely not another quadratic or additional relationship between the variables that is not being accounted for.

**Conclusions**

Although the location a student took their exam is shown to have some statistical significance in the result of their test scores, whether they received aid for lunch was the greatest factor in impacting their ACT score. A low p value and a high R squared all showed the statistical significance of the reduced model. While location and environment are factors that do shape children, the reality is if a student is not well-nourished, they will not have the energy to thoughtfully take assessments.

**References**

“Digest of Education Statistics, 2021.” *National Center for Education Statistics (NCES) Home Page, a Part of the U.S. Department of Education*, nces.ed.gov/programs/digest/d21/tables/dt21\_226.60.asp. Accessed 22 Oct. 2025.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |