How does Location and Socioeconomic Variables Impact Average ACT scores?

By Sydney Golden

**Abstract**

This project assesses how 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. Exploratory analysis included creating a pair plot of the numerical data. 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 in general did not have a statistically significant impact 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 therefore 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.

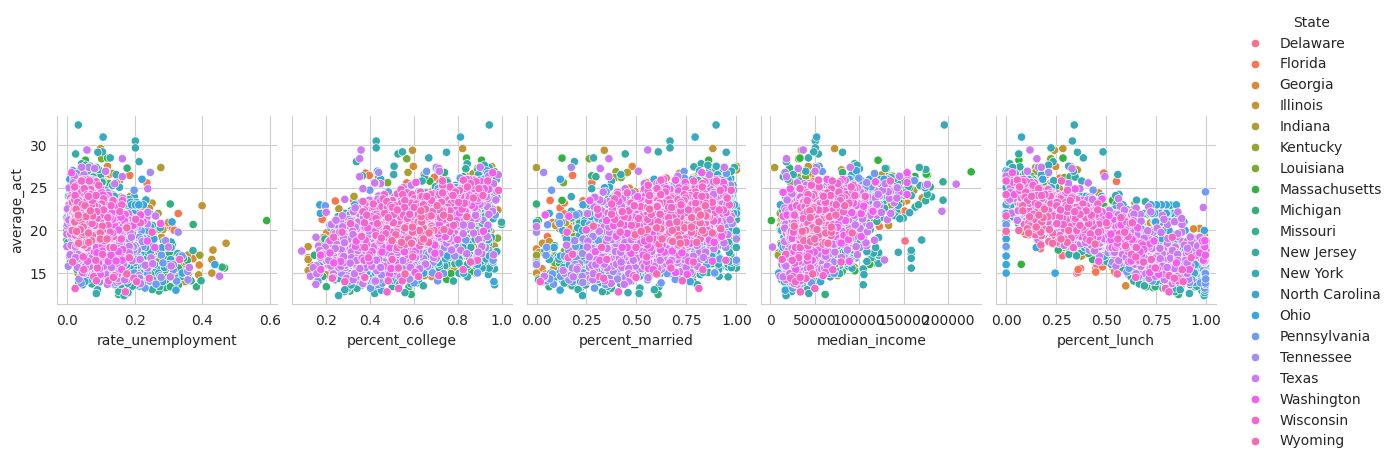
**Theoretical background**

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. The EdGap data set comes from a sample of specific schools, while the ACT score information is representative of the entire U.S high school population. Data from the ACT state scores only is from the years 2017 and 2021, and only the states listed in the EdGap data set will be analyzed. Only the composite scores will be taken into account in this study.

**Methodology**

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 (ACT composite scores) comes from the [National Center for Education Statistics](https://nces.ed.gov/ccd/pubschuniv.asp). The data set was cleaned by dropping unnecessary rows and columns that were not the state, year or composite score. The EdGap data set was revised to have state names labelled fully instead of abbreviations, and an additional “years” column was created to keep similar formatting. These two data sets were merged by an inner join on state and year into “merged\_df” to keep only state data that is included in both datasets. 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 for numerical values. Single input models of linear regressions were conducted for median income. A multiple linear regression was created including state data. 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. Pair plot 
2. Multiple Regression Results (States)

OLS Regression Results

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

Dep. Variable: average\_act R-squared: 0.708

Model: OLS Adj. R-squared: 0.705

Method: Least Squares F-statistic: 219.6

Date: Thu, 30 Oct 2025 Prob (F-statistic): 0.00

Time: 05:27:13 Log-Likelihood: -12449.

No. Observations: 7227 AIC: 2.506e+04

Df Residuals: 7147 BIC: 2.561e+04

Df Model: 79

Covariance Type: nonrobust

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

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

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

Intercept 20.0647 2.276 8.818 0.000 15.604 24.525

C(State)[T.Florida] 3.7296 2.329 1.601 0.109 -0.836 8.295

C(State)[T.Georgia] 2.2133 2.310 0.958 0.338 -2.316 6.742

C(State)[T.Illinois] 0.0393 2.300 0.017 0.986 -4.470 4.549

C(State)[T.Indiana] 1.3371 2.323 0.576 0.565 -3.216 5.890

C(State)[T.Kentucky] 1.9482 2.374 0.821 0.412 -2.706 6.602

C(State)[T.Louisiana] 1.7219 2.345 0.734 0.463 -2.874 6.318

C(State)[T.Massachusetts] 1.1590 2.372 0.489 0.625 -3.490 5.808

C(State)[T.Michigan] 1.3194 2.316 0.570 0.569 -3.220 5.859

C(State)[T.Missouri] 4.0068 2.315 1.731 0.084 -0.531 8.545

C(State)[T.New Jersey] 2.4266 2.337 1.038 0.299 -2.154 7.007

C(State)[T.New York] 6.3507 2.351 2.701 0.007 1.742 10.960

C(State)[T.North Carolina] 4.8123 2.335 2.061 0.039 0.235 9.390

C(State)[T.Ohio] 2.6417 2.294 1.152 0.250 -1.855 7.139

C(State)[T.Pennsylvania] 2.3833 2.296 1.038 0.299 -2.117 6.884

C(State)[T.Tennessee] 3.5183 2.338 1.505 0.132 -1.064 8.101

C(State)[T.Texas] 2.7880 2.293 1.216 0.224 -1.706 7.282

C(State)[T.Washington] 3.1489 2.397 1.314 0.189 -1.549 7.847

C(State)[T.Wisconsin] 3.5717 2.326 1.535 0.125 -0.989 8.132

1. C(State)[T.Wyoming] -2.8277 3.281 -0.862 0.389 -9.260 3.605

\*Please note that only information from states is displayed.

1. Regression model with additional predictor variables

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, 30 Oct 2025 Prob (F-statistic): 0.00

Time: 05:26:57 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C(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

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

Omnibus: 1192.719 Durbin-Watson: 1.710

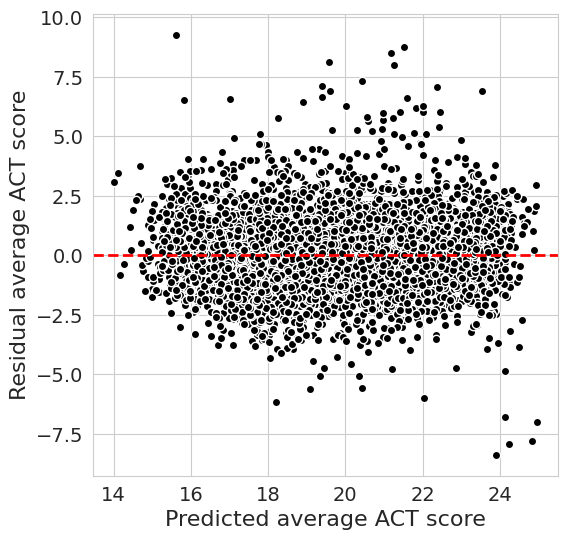
Prob(Omnibus): 0.000 Jarque-Bera (JB): 7107.782

Skew: 0.655 Prob(JB): 0.00

Kurtosis: 7.678 Cond. No. 97.8

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

1. Residual model

****

**Discussion**

As shown by the variation in different colored points, the first pair plot shows a clear linear relationship between the state an ACT score was taken in and different socioeconomic variables. This clear linear relationship encourages a deeper dive into the impacts of different predictor variables.

It is important to note that in the multiple regression analysis, the state of reference is Delaware. In this model, only New York has a p-value less than 0.05, showing a statistical significance between the comparison of ACT test scores taken in New York vs. Delaware. The high R squared value (0.708) shows that all the predictor variables together created a statistical significance.

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 value (0.679) shows that all the reduced predictor variables together created a statistical significance. The p values for states all display as 0.00, hinting at possible error with integrated categorical variables into the analysis.

The residual model is derived from the model that was used in the second computational results. It displays most points somewhat along the dotted red line, with a few outliers showing a weak negative relationship. Since the relationship appears weak, it shows that there is most likely not another quadratic or additional relationship between the variables that is not being accounted for.

**Conclusions**

Whether a student received aid for lunch was the greatest factor in impacting their ACT score as opposed to their location. Only the state of New York displayed statistical significance in ACT score differences in comparison to the state of Delaware. Future analysis may include having the average ACT score in the entire U.S be the reference state for comparison instead of Delaware. While location and environment are factors that do shape children, the reality is that if 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.