Socioeconomic Factors as Predictors for ACT Performance in Highschool Students

Introduction

In the United States, the American College Test (ACT) is well known. For some students, it represents a looming opportunity to prove their college aspirations; for others, a required junior-year exam they cannot avoid; and for still others, a test to which they pay no mind. Around the country, different factors can drastically alter the demographics of who take this test. Some states, like Colorado once required full ACT testing amongst all students, yet now, they solely require the Scholastic Aptitude Test (SAT). Some states, require simply a subset of the ACT be taken. Others require their own standardized tests to be taken, or no test at all. With such a wide range of combination by state, and by student, one can quickly imagine how complicated it can be in trying to understand patterns in students test scores. To complicate things further, it is often thought that socioeconomic factors can also play significant roles in students' performance in schools, and even standardized testing. These factors vary widely amongst individuals and even through time. Moreover, many of these factors, such as income and free and reduced lunch qualifications, can be highly associated or have compounding effects when considered in tandem.

In this report, we will consider data from the following 20 states:

1. Washington
2. Wyoming
3. Texas
4. Louisiana
5. Missouri
6. Illinois
7. Wisconsin
8. Michigan
9. Indiana
10. Kentucky
11. Tennessee
12. Georgia
13. Florida
14. North Carolina
15. Ohio
16. Pennsylvania
17. New York
18. New Jersey
19. Delaware
20. Massachusetts

From these states, we will consider the following socioeconomic variables and their potential to predict students ACT Score (or SAT Equivalent).

- Percent Adults with College Degree | percent_college (by census tract)
- Unemployment Rate | rate_unemployment (by census tract)
- Percent Children in Married-Couple Families | percent_married (by census tract)
- Median Household Income | median_income (by census tract)
- Percent Free and Reduced Lunch | percent_lunch (by school)
- Average Teacher Salary per Pupil | salary_pupil (by school district)

From these variables we will run iterative imputation methodology to fill knowledge gaps eventually running predictive Multiple Linear Regression (MLRs)Model and Single Linear Regression Models (LRs). We will then use these model's Mean Absolute Error to assess their accuracy and determine if these predictors in tandem, or one alone, is best for modeling student's ACT scores from predictors related to their home life.

Limitations

This report has limitations that must be mentioned up front. To start, our cleaned data set of 7227 school ACT averages is from a sample of 20 of the 50 states in the union, representing less than half the full demographics of the country. Furthermore, a large minority of those schools come from Texas alone, 913 to be exact, with the next largest contributor, Ohio, at 654 schools. Potentially more challenging, is the fact that for one of our predictors, Average Teacher Salary per Pupil, were not reported for New York or Washington State. Therefore, all salary data from those states represents imputed data from an iterative imputer fit to the combination of the other predictor variables from other states. These data represent a further 558 data points combined from both states. In total, Average Teacher Salary, was by far the largest imputed variables, representing 766 imputations. Additionally, 5 of our 6 predictor values are measured at the census tract or district level, representing, themselves, summary statistics of multiple schools. Therefore, we are in some fashions, comparing apples (district level averages) to oranges (school level ACT averages)

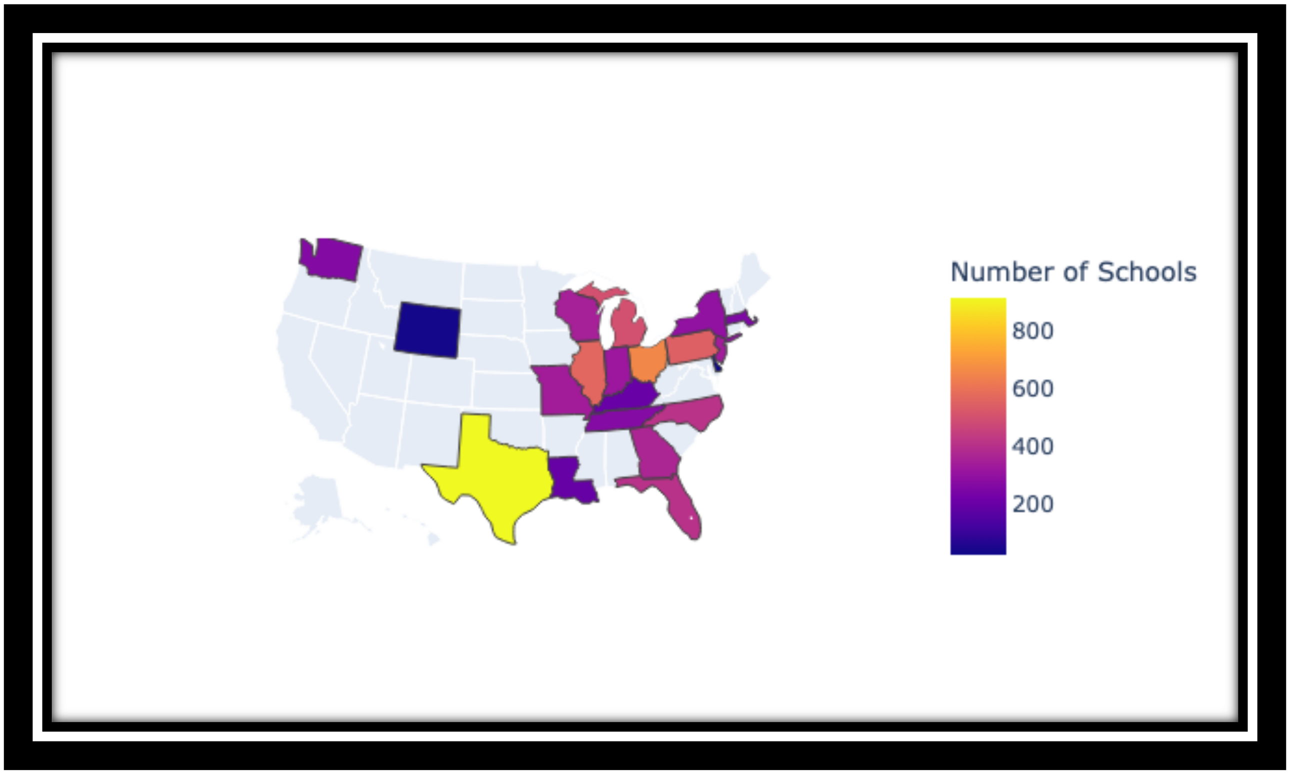

Fig 1. *Depicts the spatial distribution of our data set's observations in the United States as well as their relative density by stat*

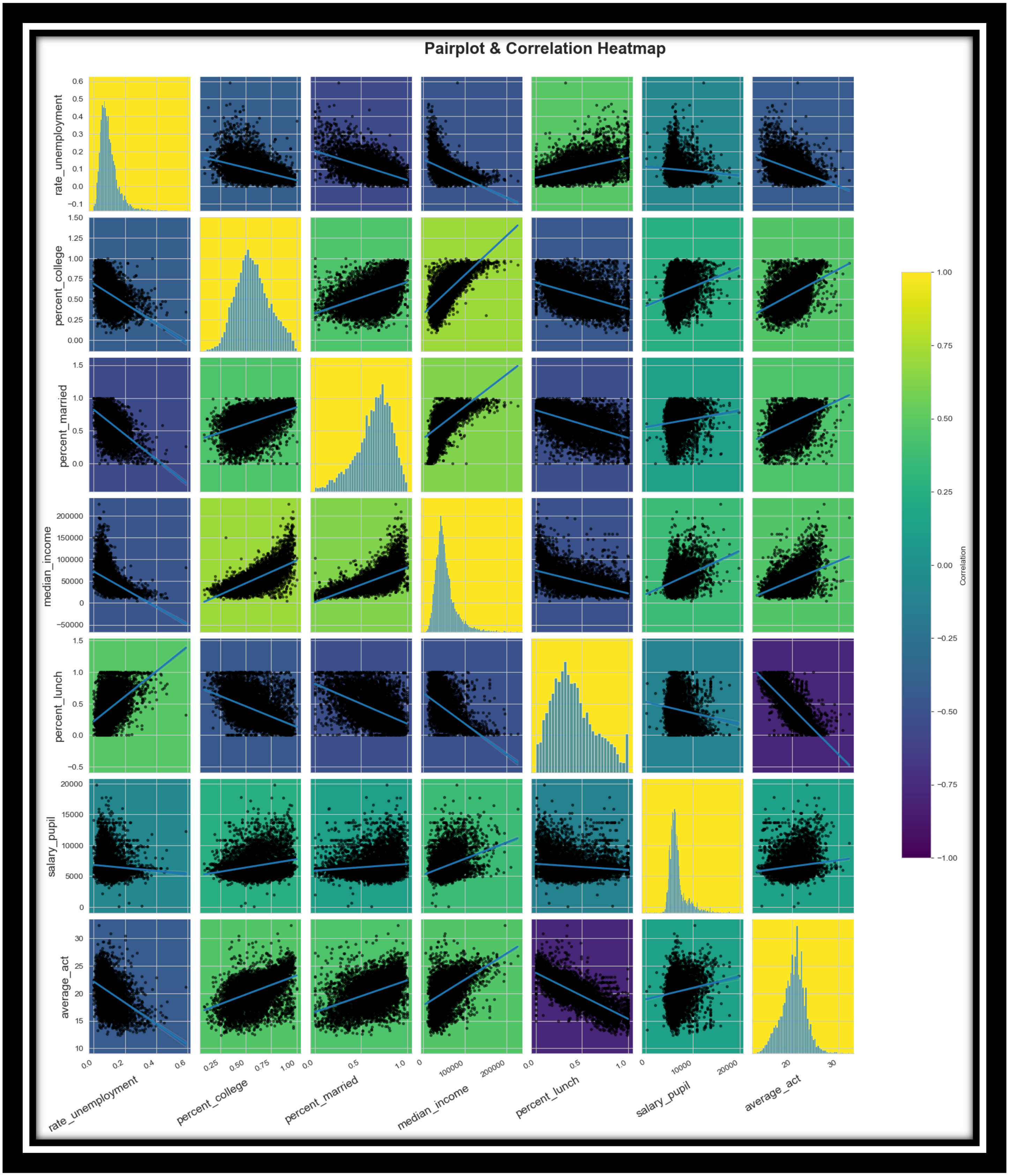

Fig 2. *Depicts the Pair Plot and Correlation Matrix with Each Socioeconomic Predictor & Average ACT Score Plotted on Both the x & y Axis. Across the Diagonal One Will See Each Variables Distribution.*

Methods & Analysis

We began with exploratory data visualization using a pair plot combined with a correlation heatmap (Fig 2). From here we determined a preliminary assessment of relationships between predictor variables and average ACT score by school. Using these tools, we identified overall correlation shapes and distributions of the data; revealing which variables had potential linear relationships and providing a sense of which socioeconomic variables might be strong predictors of ACT performance. Next, a full MLR model was fitted using all available predictor variables, in an effort to get an idea for the statistical significance of each variable within prediction model. If p values in the MLR corresponding to a variable were equal to 0, they were assumed to be important enough to be used in the reduced model. Greater attention was paid to median_income and salary_pupil, as both visually seemed they could need second order LR models in the pair plot. However, after testing, these higher-order models did not improve fit or interpretability and were therefore not retained (Fiq 3 & 4).

From here we could begin to refine our model and simplify it further. Since salary_pupil did have a significant p-value (0.002) in the initial full MLR, it was kept as a key predictor for the simplified MLR along with rate_unemployment, percent_college, and percent_lunch. As another layer of analysis, the reduced set of variables were normalized and re-run through another, normalized, MLR. This normalized MLR, shed light on the relative importance of each variable in making ACT predictions could be determined by the magnitude of its coefficient. This revealed percent_lunch to be the largest contributor and was then used for a Single Linear Regression (LR) for completeness and simplicity's sake.

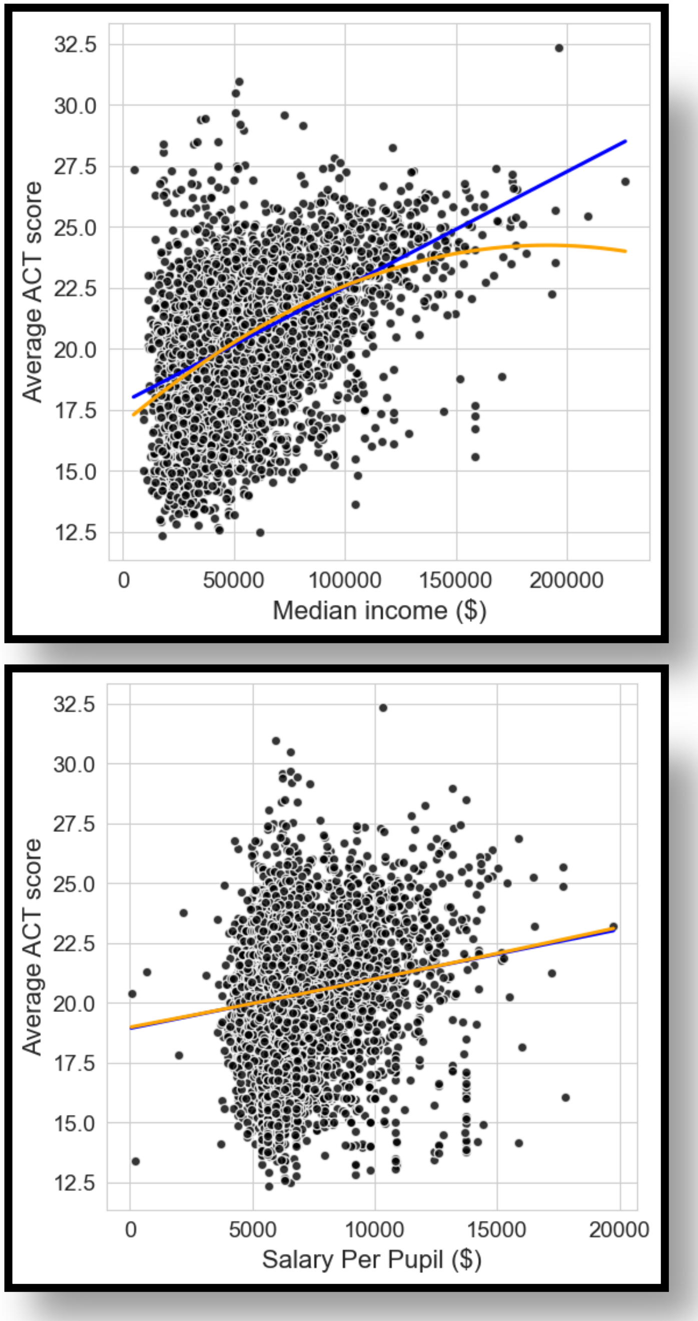

Fig 2 & 3. *Depicts Scatter Plots and First (in blue) and Second (in gold) Order Regression Lines Fitted to their Respective Data*

Results

Model	Variables	Mean Absolute Error: ACT Prediction	R^2
Full MLR	Percent Adults with College Degree	1.1429	0.6285
	Unemployment Rate		
	Percent Children in Married-Couple Families		
	Median Household Income		
	Percent Free and Reduced Lunch		
	Average Teacher Salary per Pupil		
Reduced MLR	Unemployment Rate	1.1434	0.6284
	Percent Free and Reduced Lunch		
	Average Teacher Salary per Pupil		
Normalized MLR	Unemployment Rate	1.1434	0.6285
	Percent Free and Reduced Lunch		
	Average Teacher Salary per Pupil		
LR	Percent Free and Reduced Lunch	1.1690	0.6139

Conclusion

Overall, we have found that these socioeconomic variables are very good predictors for average ACT score, being able to predict an ACT score within ~1.0 – 1.2 points, with all our models having strong r-squared values in the range of 0.63 – 0.61. We can have strong confidence in these findings. In fact, while our full MLR was more accurate and statistically significant, practically, the simplest model using just Percent Free and Reduced Lunch as a predictor would be best to use in real world applications.

There are some likely explanations for this. Percent Free and Reduced Lunch, opposed to all other predictor variables, is on the scale of individual schools, while the other describes districts. Therefore, these rougher scale variables could be obscuring the true relationships in these values; especially, as ACT averages, themselves, are on the scale of the school. With that said, any of these models seems adequate for the scope of our analysis.