SCHOOL DATA DEEP DIVE

By Annie Kroo, Dieter Brehm, and Case Zito - Fall 2019

The state of education in the U.S. right now is one that is divided. Test scores and achievement varies greatly from state to state, and much within the schools and students of those states. In this project, we wanted to explore what factors correlate with that very different achievement and to find possible explanations for the gaps.

Our Data

We explored two different databases for this project, both of which explore U.S. public middle and high school data, but with different focuses. One of the databases has 1992 to 2017 state-wide education data with National Assessment of Educational Progress (NAEP) math and english scores, National Center for Education Statistics (NCES) enrollment, gender breakdown, and race demographic infordmation, U.S. census finance information. We specifically use math NAEP scores as the results for data analysis. The other database from the Civil Rights Data Collectionis a wide reaching national government survey of public schools with data about funding, enrollment, and classes. It contains data from every federally funded highschool in the US.

Algorithm

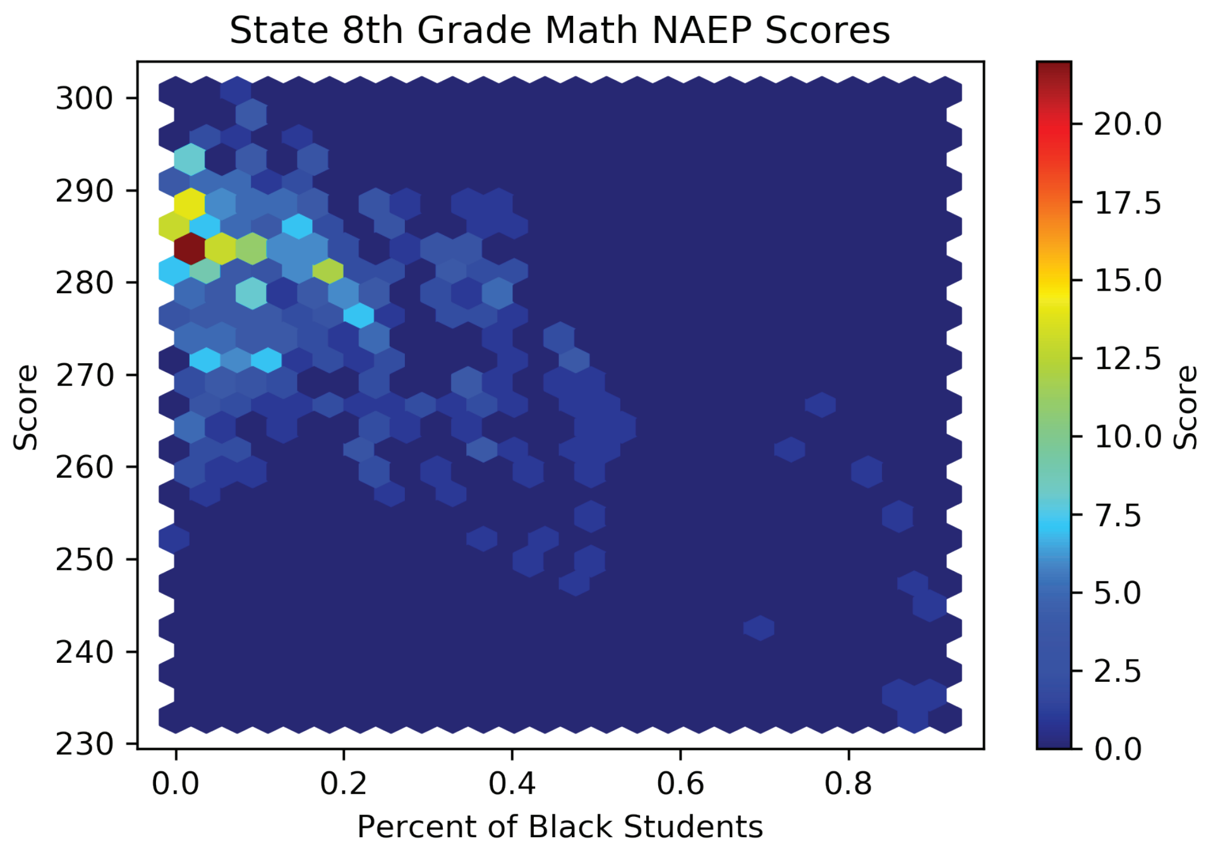
Starting with statistical analysis and plotting, we explored data trends and shapes.

We then used sklearn's linear regression in Python on different school information vs. test scores. We found coefficients for the correlations between these factors and scores and explored the meaning of them. For validation, we performed train-test split to validate the prediction quality of the models we explored.

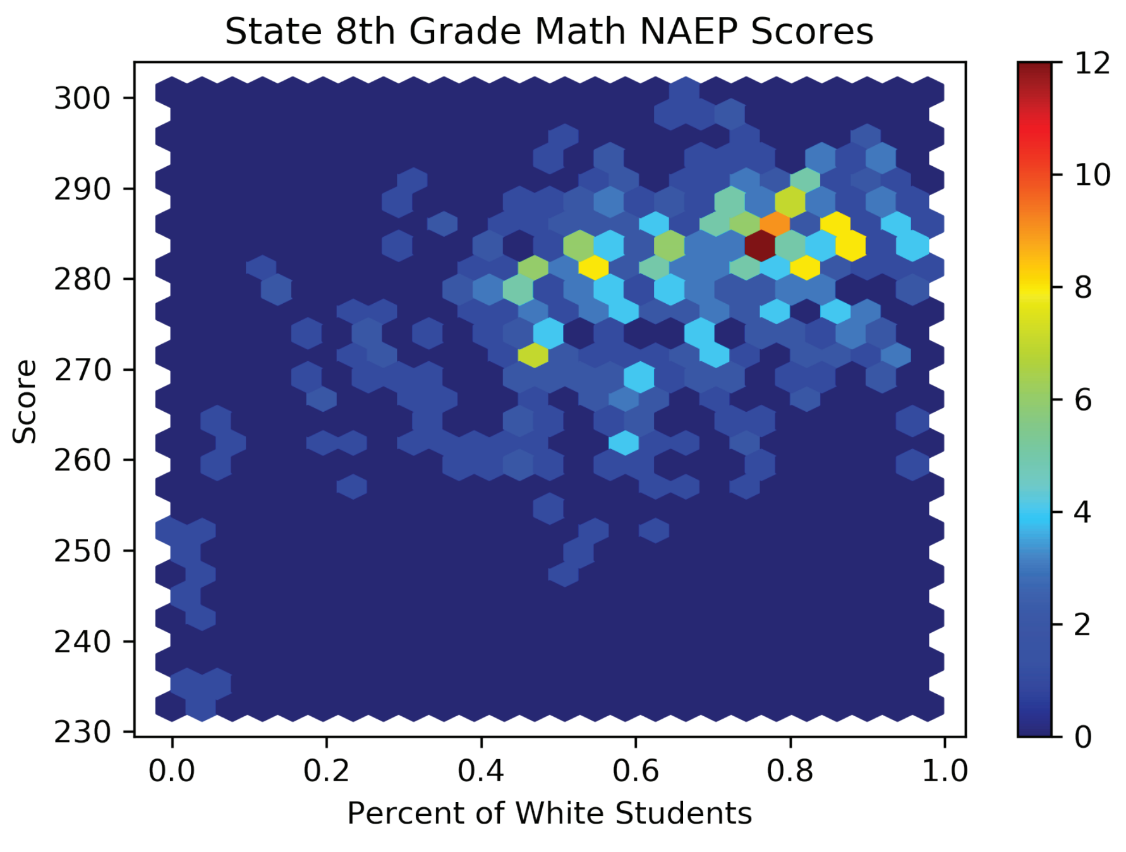
Data Visualization

For the first dataset, native american and asian percentages show a higher correlation coefficient than we expected. This is due to their low populations. There is a centering of points at around the average score for 0% of those populations and then a few lower score points for higher % of the populations. This means that there is a negative correlation but it may appear higher than it actually is because there's a small amount of schools with high populations.

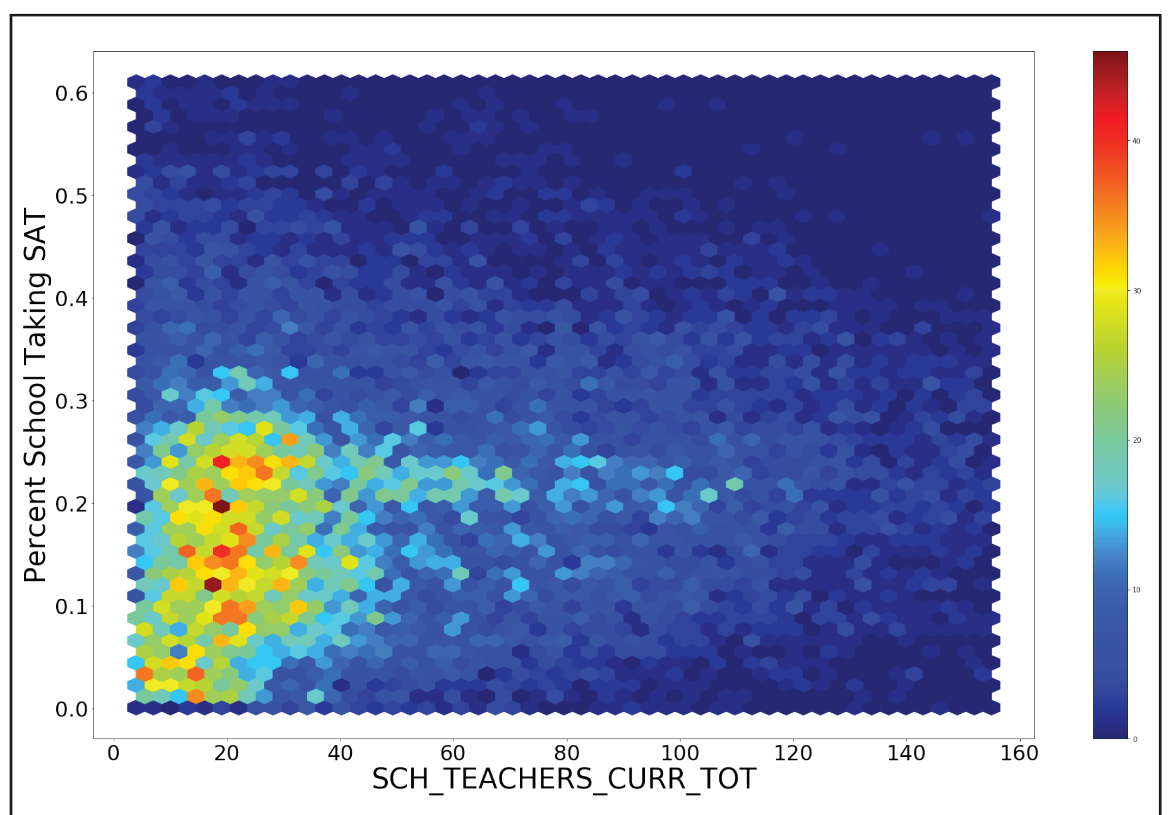
White percentage and revenue per student show a lower coefficient than I would expect, possibly due to scattering with some of the points that may make it seem like a weaker correlation.
Black and hispanic percentage coefficients are fairly reasonable, but perhaps stronger than I would expect when compared to my predictions for white percentage and revenue per student.



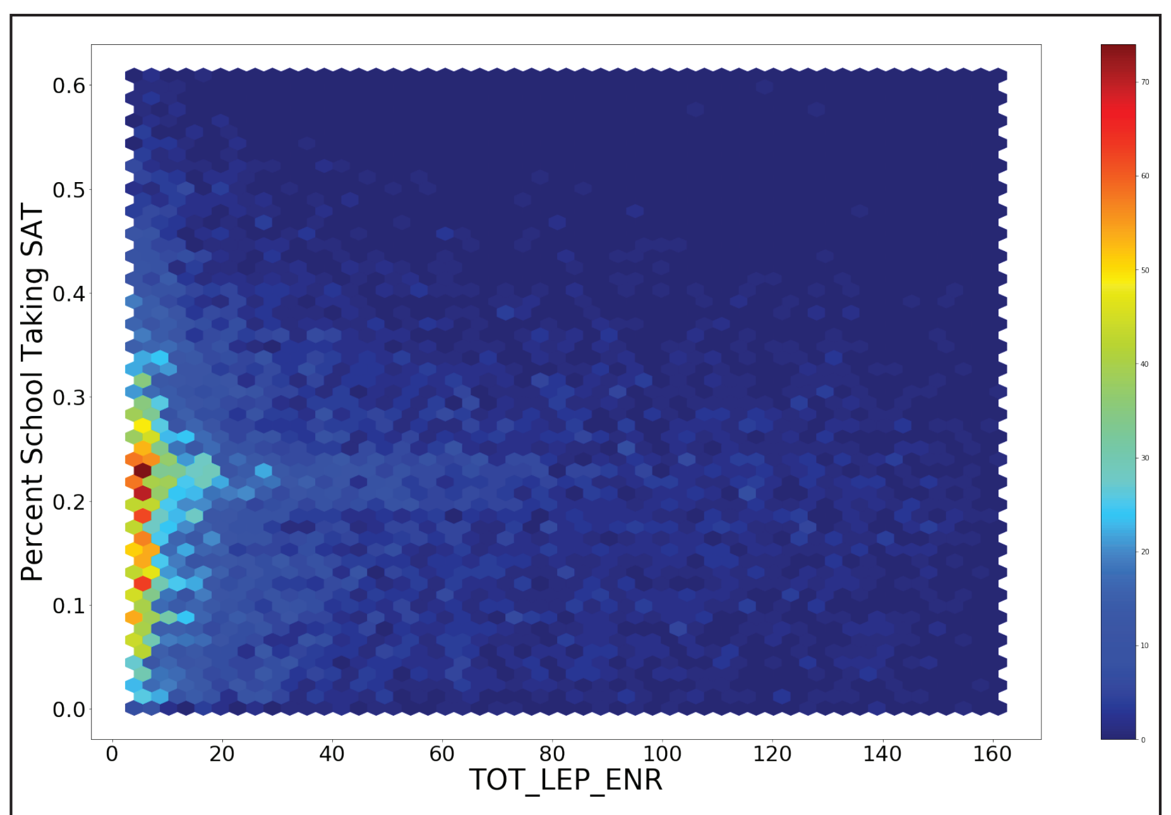
Representation of relationship between percentage of black student population and aggregate test scores.



Representation of relationship between percentage of white student population and aggregate test scores.



Current number of teachers vs. percent of the school participating in the SAT/ACT, a proxy for students preparing to attend college.



Number of students enrolled in limited english proficiency programs vs. percent of school participating in the SAT/ACT.

For the second dataset, which surveyed a wide range of fields in public schools ranging from out of school suspensions in the 2015-16 school year broken down by race, English language proficiency, and disability status, all the way to school resources and SAT/ACT participation. We used our model to explore the predictors of a common metric for success: percent of the enrolled student's participation in the SAT/ACT.

In this exploration, we found several expected predictors of success such as the number of full time teachers at the school, but also identified key failings of our schooling system. Most evident among these was our public school systems ineptitude in its assistance of immigrants. We found a relatively strong negative correlation with participation in SAT/ACT and enrollment in Limited English Proficiency Programs (LEPs).

Results

From a train/test split regression model, we found that:

Revenue/student with white pop% is the best predictor for student test scores of the 2 variable linear regressions we ran. It is also the most consistent from train to test scores.

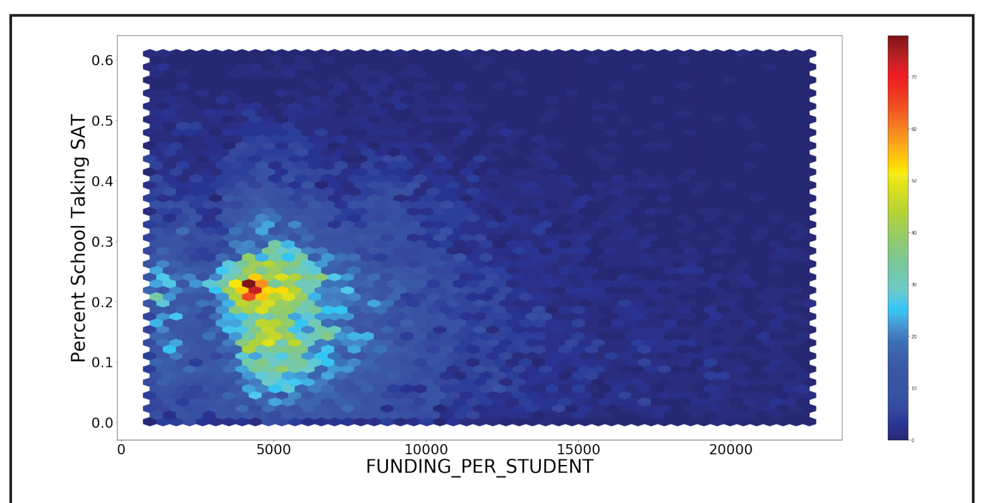
Revenue/student with non-white race pop % predicted test scores with nearly 90% accuracy, showing the significance of these factors.

NAEP dataset

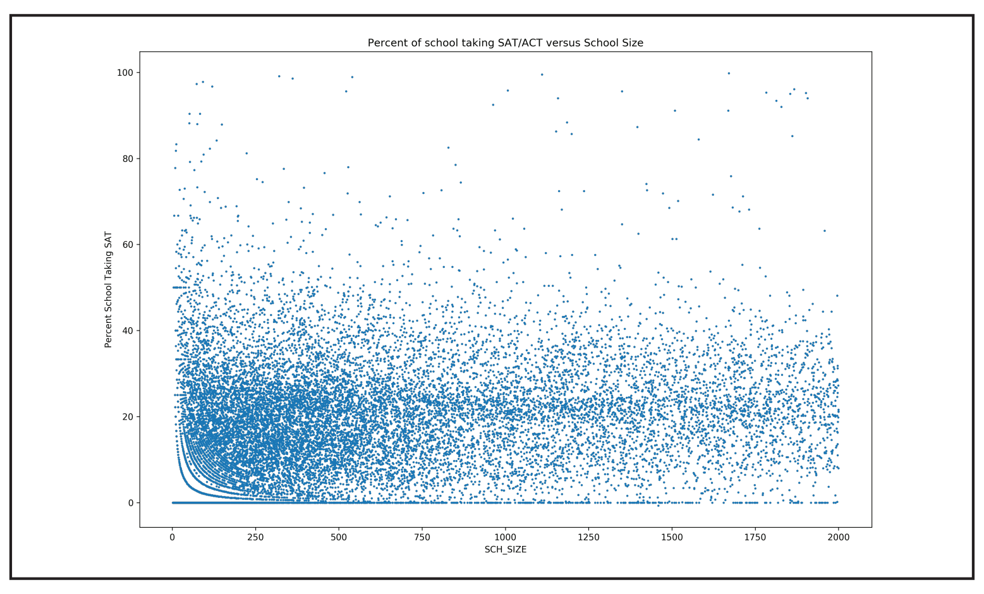
Inputs	*Output (math scores)*
revenue / student + white pop%	train: 84% test: 86%
revenue / student + black pop%	train: 84% test: 85%
revenue / student + [other listed races] pop %	train: 78% test: 76%
revenue / student + non-white pop %	train: 87% test: 87%

In manipulating the CDRC dataset, we identified several redundant variables, isolated and included confounding variables when possible, and identified inaccessible confounding variables. The primary lacking information, was the average household income for the students' families. We were hoping to use some funding data provided, but found that it was inaccurately reported and surprisingly sparse. Our model at the end had a 55% accuracy of estimating the SAT participation within 10% given the information displayed to the right.

CDRC dataset

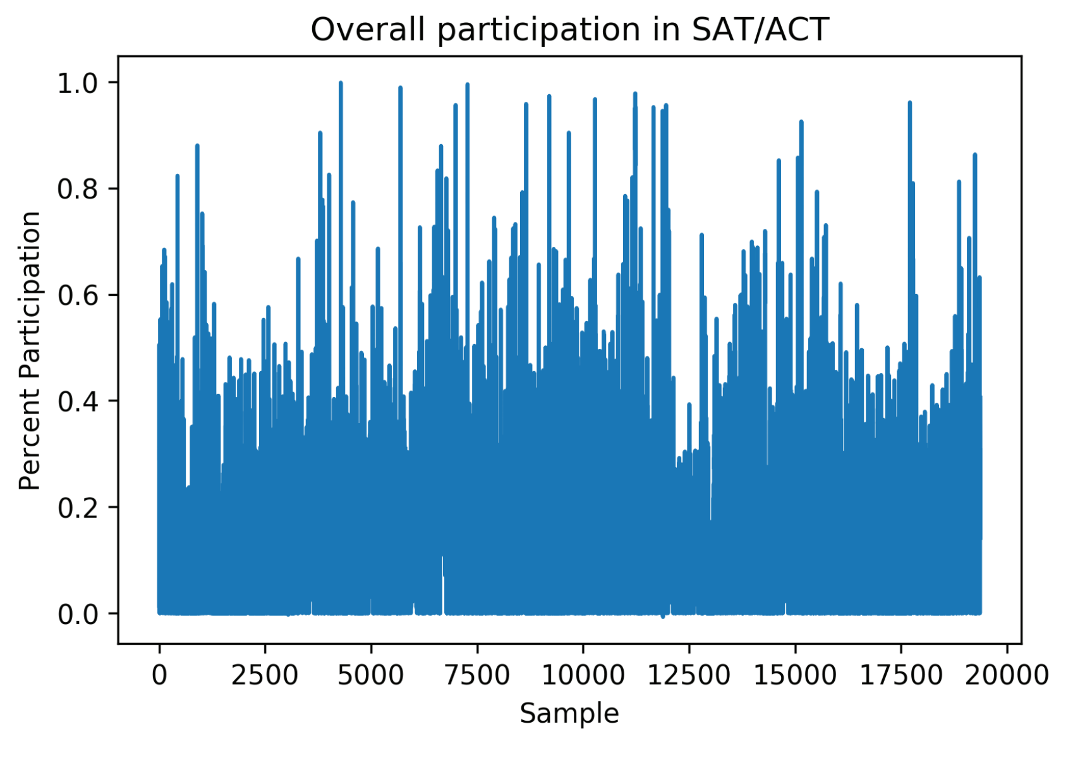


SAT/ACT Participation over funding/student



Conclusions

We believe that the percentage of white (or non-white) students and the total revenue of the state correlate highly with NAEP math scores because of institutional racism in creating the test and because poorer states/students are not able to afford the same quality of teachers/teaching materials/etc. Schools should be given more funding based on need and there needs to be more programs seeking to help underprivileged students of minority backgrounds.



Sources

National Center for Education Statistics, State Education Survey Data, 1992-2017. https://nces.ed.gov/ccd/stnfis.asp
United States Census, School System Finances, 1992-2017.
https://www.census.gov/programs-surveys/school-finances/data/tables.html
The Nation's Report Card, NAEP, 1992-2017.
https://www.nationsreportcard.gov/ndecore/xplore/NDE
Civils Right Data Collection, School Data, 2015-2016.