The Solution to America's Education Crisis

A data oriented approach to the troubles of the American education system

THOMAS ALEXANDER, RYAN KINNEAR, AUSTIN TRIPP, and RICHARD WU

Additional Key Words and Phrases: Data Science, America, Education System, Correlation One, Datathon

1 TOPIC QUESTION

Nelson Mandela once said "Education is the most powerful weapon which you can use to change the world." Unfortunately, while the US is a leader in powerful weapons, recent reports show that its education system falls behind other developed nations (DeSilver 2017). The cause of this low performance has been the subject of intense debate, with some people claiming that systemic failures in the US educational system are responsible (Ryan 2013), and others claiming that the root cause is primarily due to class inequality.

In particular, Carnoy (Carnoy and Rothstein 2015) claimed that if the United States' socio-economic distribution were similar to that of other western nations, their standardized test scores would be top-tier. This implies that the solution to solving the United States' educational problem is simply to solve its inequality problem. And, one of the most effective ways to improve people's socio-economic status is through education. So, paradoxically it seems that the best way to improve US education is by reducing inequality, which can be accomplished by improving US education. Clearly, such circular reasoning doesn't get us any closer to a solution.

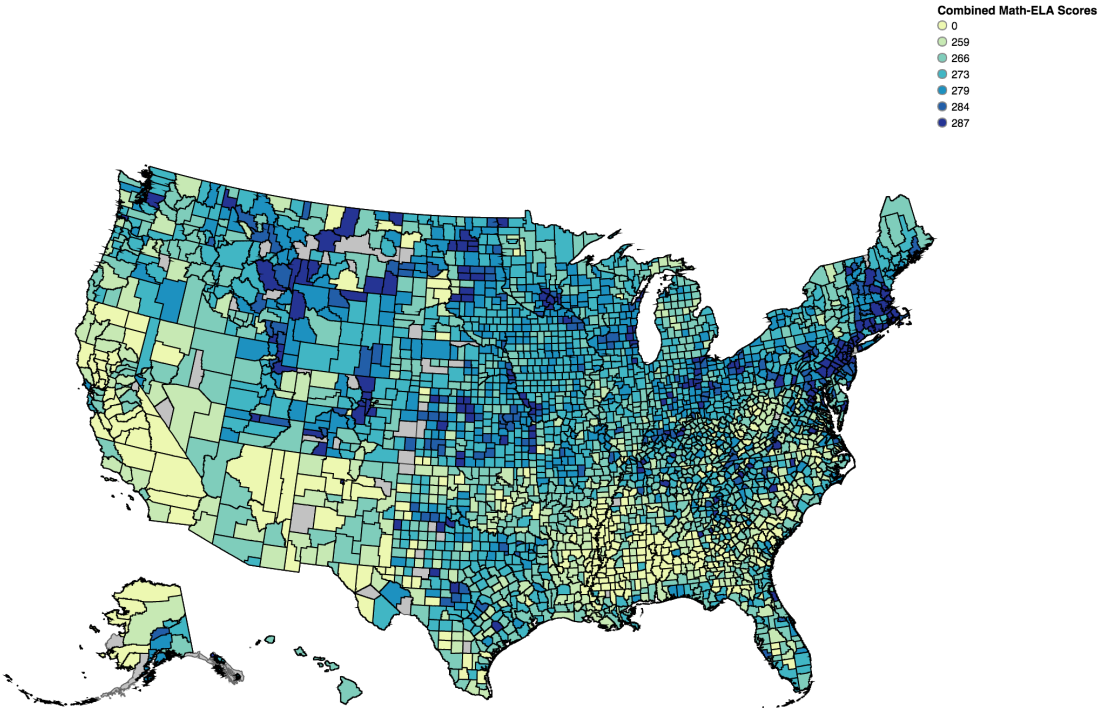
Instead, we focus on the effects of public policy on educational outcomes. Since each state maintains its own educational system, statewide policies naturally have a pronounced effect on the performance of that state. The question we seek to answer is: what are the best educational policies that a state can enact? We use our analysis to make nationwide policy recommendations that will improve the US's performance as a whole.

This is an extremely important question that can be answered with the data provided, because if the goal of the analysis is to help improve the US's lackluster performance, then the most valuable result is to find a method to do so. Using a data-driven approach, we examined the relationships between state education policies and state performance to determine which policies were best. Data on district demographics, NAEP scores, district means grade equivalent, SAT scores, teacher salaries, state crime rates, and state educational budgets were used.

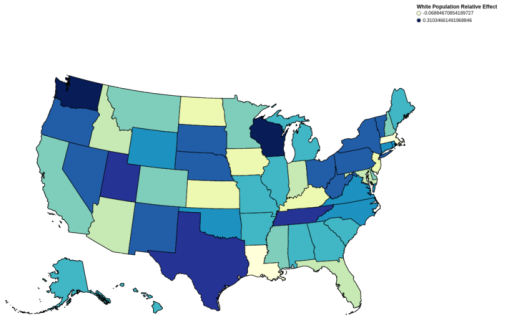
2 EXECUTIVE SUMMARY

From exploratory data analysis with the help of some illuminating geographic heat maps and simple but effective regressions, we briefly looked at how academic achievement, as measured by NAEP test scores, is clustered on a visual map of the United States as shown in figure 1. Race clustering in a geographic context seemingly correlates with the geographic distribution of test scores, whereby conventional minorities are more populous in poorly-performing school districts and White students are more populous in highly-performing districts. Further correlation analysis revealed that there are in fact multiple socioeconomic indicators that also align with test scores: positive indicators such as the

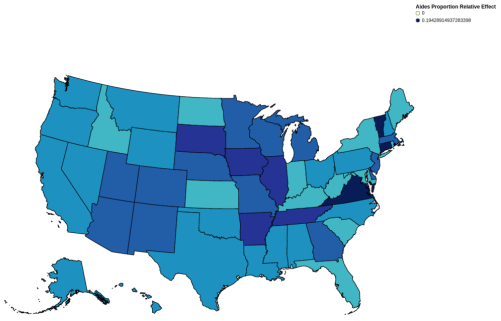
Fig. 1. US Map of Combined NAEP Math-ELA Test Scores (Grade 3 - 2013)



(a) $w^{(s)}$[perwht] Posterior Means (Darker is Bigger)



(b) $w^{(s)}$[aides] Posterior Means (Darker is Bigger)



proportion of adults with a Bachelors degree or higher and negative indicators such as the proportion of households with single moms.

The above correlation analysis draws attention to the controversial affirmative action stance taken by institutions across the United States and its perhaps misguided focus on solely race groups instead of socioeconomic status as a whole. In terms of policy recommendation, negative indicators overwhelmingly describe grim financial and social circumstances underperforming districts and their students experience back at home. We strongly push for direct policies that improve a student that fits this profile's access to extracurricular mentorship. We believe the influence of a guardian is very impactful in the formative years of a child and is essential to a student's academic success.

Using a Bayesian hierarchical model, we have concluded that, at the district level, teachers aides are a robust predictor of good student NAEP scores, even after controlling for some state level covariates like teacher salaries, and crime rates. Our policy recommendation is to significantly increase focus on the limited funds currently allocated to teacher aides across the country. Our posterior weights as shown in figure 2b show that aides have a non-trivial positive influence on students' academic success which intuitively follows from more tailored and undivided attention during classes.

Furthermore, we have shown that, despite many efforts, there are still clear racial advantages. In particular, simply being born white is a strong predictor of student performance as shown in figure 2a. Persistent effort is still necessary to reduce racial segregation and unjust racial advantages.

3 ANALYSIS

To begin the analysis, several basic questions had to be answered:

(1) What kind of public policy is to be considered?
(2) How can educational performance be measured?

Addressing the first question, public policy can be divided into two broad categories: budgetary and non-budgetary. Strictly speaking, money does not directly affect students' education, but it has a significant indirect effect because it determines a school's ability to procure new resources and maintain current assets. Therefore it is reasonable to view educational quality as being highly dependent on the budget. So, in broad terms, a state's educational budget should be strongly related to the educational outcomes of that state. Budget data is also widely available and relatively easy to interpret, which makes it possible to compare policy decisions between different states. While non-budgetary policies are also very influential, they tend to be non-numerical and thus only comparable qualitatively. So, for the purposes of this report only budgetary policies were considered. Specifically, state-wide educational budget data from the US Census Bureau from 2013 was used.

Addressing the second question, the most widely-used method for evaluating educational performance is standardized testing. There are several different standardized tests ubiquitous in the US. The National Assessment of Educational Progress (NAEP) is a nation-wide test given to a statistically-balanced group of elementary school students in grades 3-8 every year. It is commonly viewed as the nation's "Report Card". For high school students, the situation is less clear. The NAEP is sometimes administered to twelfth-graders but this is not consistent. Thus, often the best performance evaluations available are standardized college admission exams, namely the SAT and ACT. One of these two tests is taken by almost every single student interested in attending college nationwide.

However, each of these tests only evaluates children of a certain age group, and is therefore not representative of the education system as a whole. While it would be conceivable to combine the results from several different tests into one overall metric, differences in test content and grading methods mean that the method of score combination would have to be carefully considered. To avoid this point of contention, it was decided to only consider a single test in this report.

The only remaining choice was which test to use. To decide this, some initial exploration of the provided SAT test data was done.

3.1 Initial Exploration

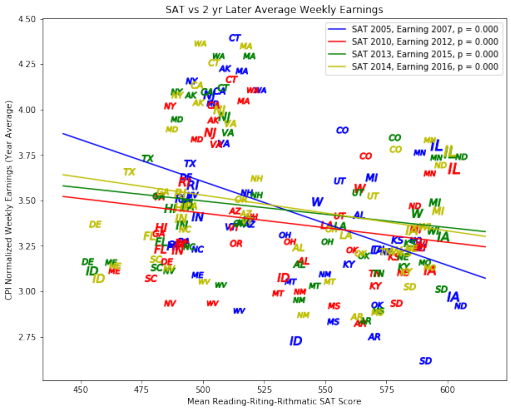
Our analysis has focused on the following datasets:

- districts.csv
- districts_means_naep.csv
- district_means_grade_equivalent.csv
- job_sectors.csv
- state_crime.csv
- state_funding_sources_2013.csv
- state_per_student_funding_2013.csv(Bureau Bureau)

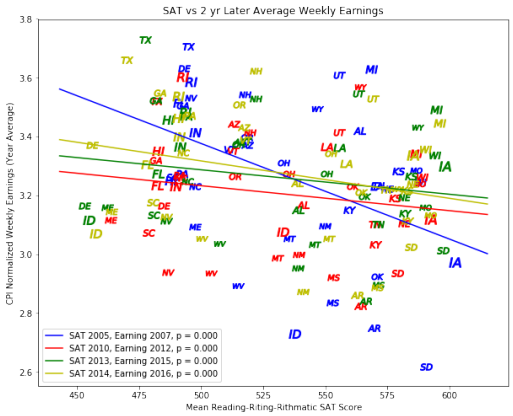
Since our goal is to focus on state and district educational policy, college data and national earnings data were not considered relevant. Further, although it may be within the scope of our research question to analyze job sectors, we have confined our analysis to a more manageable subset.

3.1.1 SAT Scores. SAT scores appear regularly in pop culture, and doing well on SATs are often viewed as a critically important step in the life of a young US student. We have however almost entirely disregarded SAT scores in our analysis, for a myriad of reasons. Firstly, many students do not take the SATs (often being replaced by the ACT), and secondly, some of our early analysis suggests that there are a lot of confounding factors that make it difficult to use SAT scores as a measure of success. In particular, we were surprised that there is an *inverse* relationship between SAT scores and earnings.

(a) SAT Scores vs. Earnings

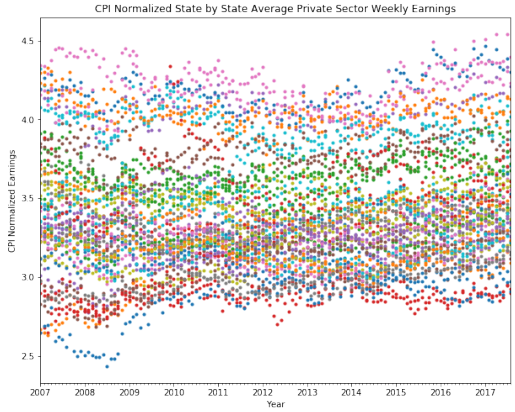


(b) SAT Scores vs. Earnings (Excl. High Earning States)

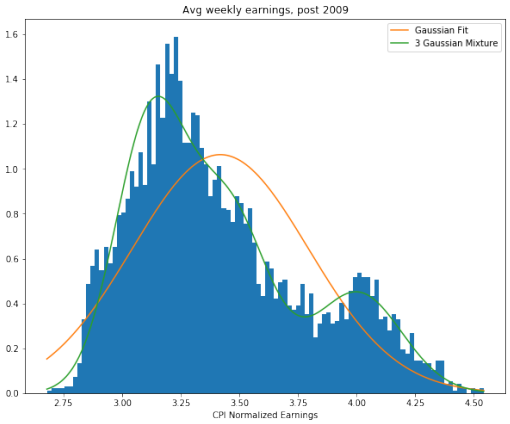


This earnings data comes from the total private weekly earnings given in the file `job_sectors.csv`, and we have been careful to normalize the figures by the "Consumer Price Index - All Urban Consumers" (CPI-U) (?? cpi) in order to account for inflation. We note that simply removing the linear trend is not an acceptable means of accounting for inflation as dollar values alone are not directly interpretable as a measure of prosperity.

(a) CPI Normalized Private Sector Earnings



(b) Private Sector Earnings Post 2009



This inverse relationship passes an F-test of significance, and while this is a crude measure, it is certainly clear that state wide earnings are not positively correlated with SAT scores. What's more surprising, is that the states with the highest incomes have some of the lowest SAT scores.

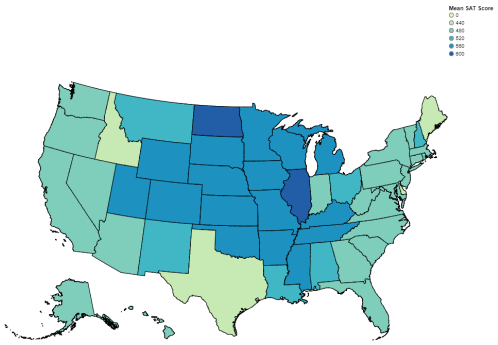
One of the salient features of our dataset is that many observations form time series. Due to inflation, this is particularly important when considering income data. We have normalized every dollar value by the "Consumer Price Index - All Urban Consumers" (CPI-U), which is a statistic published by the Bureau of Labor Statistics (?? cpi) to provide a measure of a dollar's purchasing power. We note that simply removing the linear trend in income data is not an acceptable means for accounting for inflation as dollar values alone are not directly interpretable as a measure of prosperity.

The direct time series plots of median household income is given in figure 4a, along with a histogram of this data in figure 4b. The bi-modality of the data is clear merely from the time series, but the histogram and mixture of 3 Gaussians model corroborates this observation. Further, the histogram includes only observations from 2009 onwards, and hence cannot be explained as a result of the 2008 financial crisis.

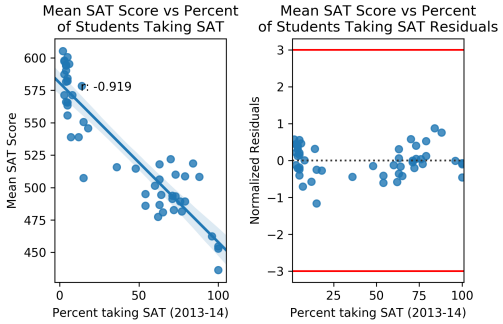
Next, the 2013 mean SAT scores for each state were visualized. Figure 5a shows the distribution of SAT scores across the US. Surprisingly, the states with the highest scores are Illinois and North Dakota, which are not known for being high-achieving states. This mystery is explained however by plotting the mean SAT score of each state versus the percent of students in that state who wrote the SAT, shown in figure 5b. This strong correlation has a very simple explanation: in states where the ACT is the dominant test, only the top students looking to apply to out of state colleges will write the SAT. This means that their average score will be much higher than a state where every student writes it. This raises serious doubts of the efficacy of the SAT as a educational quality metric.

3.1.2 Scope and Feature Engineering. The NAEP exam is preferable to the SAT or ACT because it administered in a way that is statistically representative of the entire population. The SAT and ACT are exams taken by the academically-inclined who are prospective candidates for college and higher education. Furthermore, as shown above, different regions of the US take the SAT or ACT preferentially, so it is difficult to compare test results between states. The SAT and ACT are also graded differently which makes it difficult to aggregate results between the two. Because of these difficulties, it was decided to use the NAEP in order to have a more representative sample.

(a) Mean 2013 SAT Score for each State



(b) SAT Scores vs Percent of Students Taking It



We choose to use grade 3 NAEP results in 2013 as the educational and academic achievement metric in this report. Demographic data from the year 2013 was also scoped accordingly.

The reason why we scoped down to the year 2013 was to permit a greater time horizon and runway for any insights and policy recommendations we extract from our analysis. The most recent data will allow us to pivot our discussion to the most recent snapshot of the education system in the US while adjusting for any time dependencies, keeping our analysis simple and digestible. Furthermore, it was necessary to scope the NAEP scores down to a single grade since there was a high positive correlation (r = 0.844) between NAEP score (both in math and ELA) and the tested grade. Examining data from grade 3 students means that insights in this report will be relevant as the sample students will still be in school. We expected some districts and transitively counties to be selected out of our dataset. We checked that the original count of counties that had valid data only decreased from 3077 to 3048 after projecting on grade 3 students and the year 2013, which was deemed acceptable to proceed.

Taking a look at the NAEP scores data, two scores are provided per district: one in math and one in ELA (English and languages). Many districts were missing math scores while a few were missing English scores, but no districts were missing both math and ELA scores. A "Combined Math-ELA" score was computed by equation 1 and will be referred to occasionally in this report.

$$
\begin{aligned}
Score(\text{Math-ELA}) &= Score(\text{Math}) && \text{if Score(ELA) is null and Score(Math) is not null} \\
&= Score(\text{ELA}) && \text{if Score(Math) is null and Score(ELA) is not null} \\
&= \frac{Score(\text{Math}) \times Mean(\text{ELA}) + Score(\text{ELA}) \times Mean(\text{Math})}{Mean(\text{Math}) + Mean(\text{ELA})} && \text{otherwise}
\end{aligned}
\tag{1}
$$

Some feature engineering we attempted was to introduce a "diversity" score to indicate the heterogeneity of a district based on its percentage of each race group. This was calculated using information entropy in equation 2.

$$
- \sum_{r \in \{\text{wht, asn, blk, hsp, ind}\}} P(r) \cdot logP(r) \tag{2}
$$

$P(r)$is the percentage of each race in a district

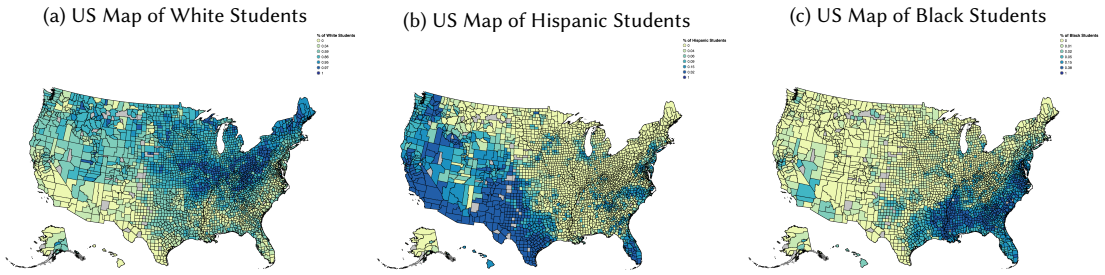


Fig. 6. Proportion of Students by Race Groups in 2013

To further augment our feature set, we noticed that some metrics segmented by race were only provided for Whites, Blacks and Hispanics but the aggregate total was provided in an "all" metric. Since we are provided percentages of each of the three race groups as well as those of Asians and Native Americans, we can "differentially" derive the same segmented metric for Asians and Native Americans combined.

Given an additive metric such as percentage of adults holding a Bachelors degree or higher (as opposed to a metric like the median annual income which cannot be differentially derived), a given metric M we can compute $M(asn + ind)$ in equation 3.

$$M(asn + ind) = \frac{M(all) - \sum_{r \in \{\text{wht, blk, hsp}\}} M(r) * P(r)}{P(asn) + P(ind)} \tag{3}$$

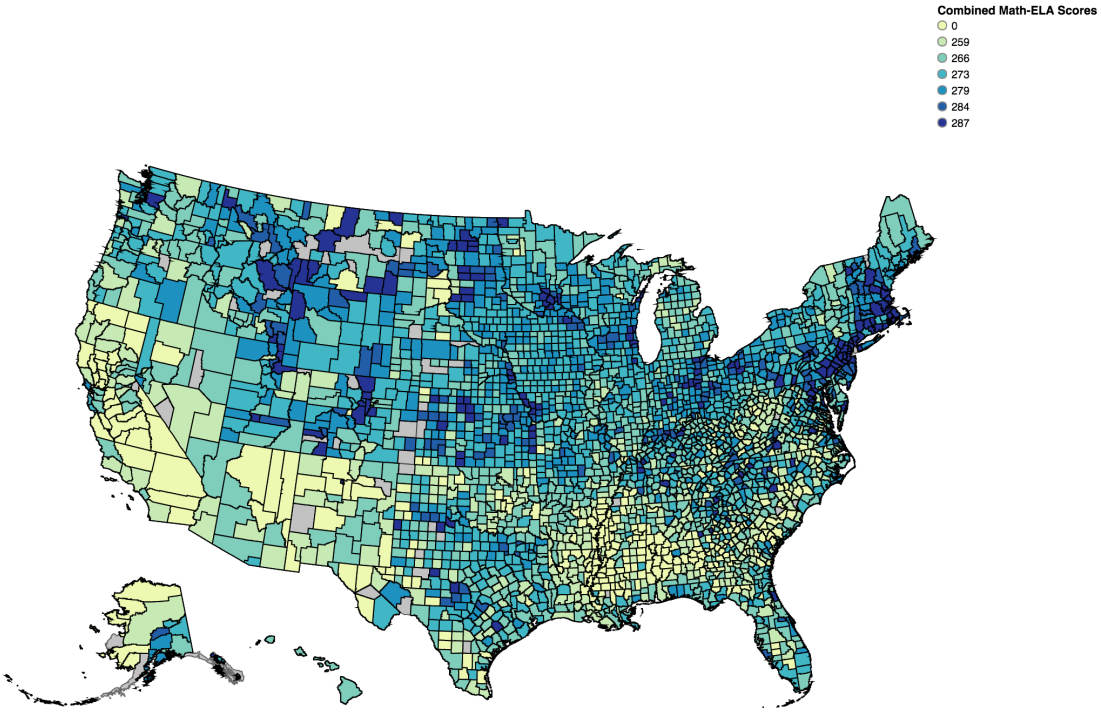
3.1.3 Race and Test Scores. Previous research has shown that race plays an important role in standardized test performance (Camara and Schmidt 1999). Because of this, the next thing we decided to explore was the effect of socio-economic factors on educational performance.

The basis of our investigation stems from the prevailing institutionalization of "affirmative action". Its origins in the United States come from President John F. Kennedy's Executive Order 10925 whereby government contractors "must take affirmative action to ensure applicants are employed without regard to their race, creed, color or national origin". This well-intentioned, egalitarian movement has since evolved into something far from the initial motives of John F. Kennedy. It is tacitly understood nowadays that "affirmative action" is a policy that partially favors candidates who are from historical marginalized and under-represented cohorts. There have been significant research in identifying private institutions who discreetly exercise this policy with respect to employment, and many criticize that although it's implemented in good faith, affirmative action is illegal and borderline unconstitutional.

Many colleges have publicly declared that they support affirmative action in light of the growing inequalities in education between race and ethnic groups. The basis for their argument is that certain race and ethnic groups have had historically fewer opportunities for academic and professional achievement and the onus is on society to help propel them to the median experience of "well-off" cohorts.

We first set out to explore if this widespread assumption is true: that race is correlated with academic achievement. We first map out the three largest race groups in the US provided in districts.csv: Whites, Hispanics and Blacks in figure 6.

Fig. 7. US Map of Combined NAEP Math-ELA Test Scores (Grade 3 - 2013)



Unsurprisingly, we see that the counties with the highest percentage of White students are in the North and Eastern locales of the United States. The counties with the highest percentage of Hispanic students are in the South, which follows from the proximity to Mexico and other predominantly-Hispanic countries. The highest concentration of Black students are also in the South which follows from the US's history of immigration.

We then map the combined NAEP Math-ELA scores derived from the provided data set of "Estimated District Mean NAEP scores" identify geographically the high-performing counties in figure 7.

The knee jerk response as data explorers is to wonder why academic achievement as measured by NAEP test scores have such a remarkable geographic clustering. Counties in the New England area have some of the highest test scores in the country whereas counties in states such as California, Arizona, and Mississippi have some of the lowest. Even more troubling is how visually similar the heat map of test scores is to that of the proportion of White students (in fact, one of us initially thought we had incorrectly re-plotted the map for White students when in fact it was the heat map of test scores).

Table 1. Top 10 Positively Correlated Covariates to Combined Math-ELA Scores (Grade 3 - 2013)

Covariate Description	Correlation to State-wide Mean NAEP ELA Score
Percentage of Non-Free Lunch Students	0.3696
Median Annual Household Income (All)	0.3302
Percentage with Bachelors Degree or Higher (All)	0.3045
Percentage of Females with Bachelors Degree or Higher	0.2904
Percentage of Males with Bachelors Degree or Higher	0.2890
Percentage of White Students	0.2662
Median Annual Household Income (White)	0.2561
Percentage in the Management Field (White)	0.2519
Percentage with Bachelors Degree or Higher (White)	0.2249
Percentage in the Computer Field (Male)	0.2039

Table 2. Top 10 Negatively Correlated Covariates to Combined Math-ELA Scores (Grade 3 - 2013)

Covariate Description	Correlation to State-wide Mean NAEP ELA Score
Percentage of Free Lunch Students	-0.3696
Percentage in Poverty (Female)	-0.3022
Percentage in Poverty (All)	-0.2995
Percentage Receiving SNAP Benefits (All)	-0.2938
Percentage of Households with Single Moms (All)	-0.2358
Gini Coefficient (All)	-0.2267
Percentage Receiving SNAP Benefits (White)	-0.2127
Percentage in the Service Field (Female)	-0.1940
Percentage of Unemployed (Male)	-0.1924
Percentage of Unemployed (All)	-0.1907

To more quantitatively assess the correlation between the plethora of demographic covariates and test scores, we performed pairwise ordinary least squares (OLS) linear regression between the combined Math-ELA scores and the demographic covariates. The Pearson's correlation coefficient are given in table 1 and table 2.

These covariates are statistically significant as they reject the null hypothesis that the slope of the linear relationship is 0 (their p-values are de facto 0). What's noteworthy about the regression results and the visual geographic results is that the regression results give us a magnifying glass into all covariates and their relationship with test scores. Although the percentage of white students is ranked as the sixth most positively correlated covariate to the test scores, there are far more covariates related to finances and upbringing of the students' families that have a greater correlation coefficient with respect to test scores.

What was surprising was how free and non-free lunch percentages are the most correlated linear features to test scores. After looking into the national free lunch program, it follows from intuition that these are in fact measures of a student's financial circumstances. Districts with many students who require or receive free lunch because of their unfortunate financial situation seem to perform worse on test scores. We do not believe that this is the real confounding variable but rather an indicator of a deeper problem: that financially-disadvantaged students perhaps do not get access to the same resources that are important to academic success.

Fig. 8. Most Positively Correlated Covariates to NAEP Test Scores

(a) US Map of Non-Free Lunch Students (b) US Map of Annual Median Income (c) US Map of Bachelors Degree or Higher

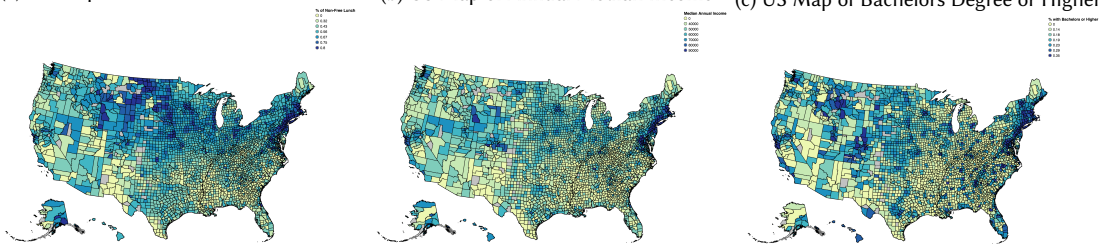
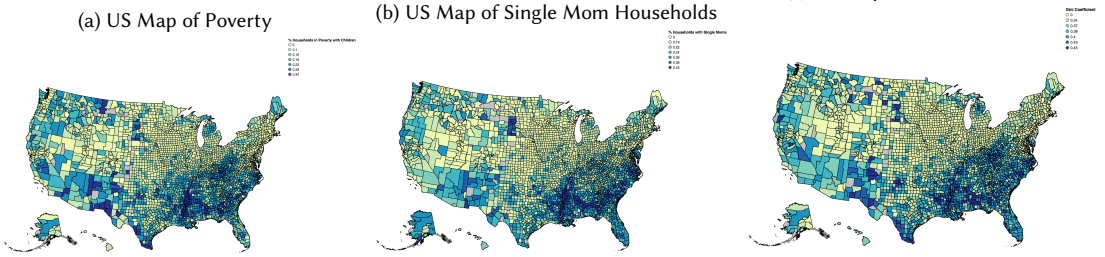


Fig. 10. Most Negatively Correlated Covariates to NAEP Test Scores

(a) US Map of Poverty (b) US Map of Single Mom Households (c) US Map of Gini Coefficient



Let us compare the US maps of the top three "positive" covariates (non-free lunch, median income, and bachelors degree) in figure **??** and the some of the top "negative" covariates (poverty, single moms, and the Gini coefficient) in figure **??** to our initial US maps of race groups and test scores.

These maps reveal to us that many covariates that highlight socioeconomic differences have slight correlations to test scores and that race groups happen to have statistically-significant correlations to these socioeconomic factors and test scores. In fact, one could argue that one should account for one's socioeconomic background more heavily than superficial race categories when it comes to affirmative action. For example the percentage of households in poverty that span all races has a high negative trend with respect to test scores. The Gini Coefficient, a measure of wealth distribution where a higher coefficient denotes greater inequalities, and its negative correlation with test scores reveal a troubling observation that inequality is negatively correlated to academic achievement.

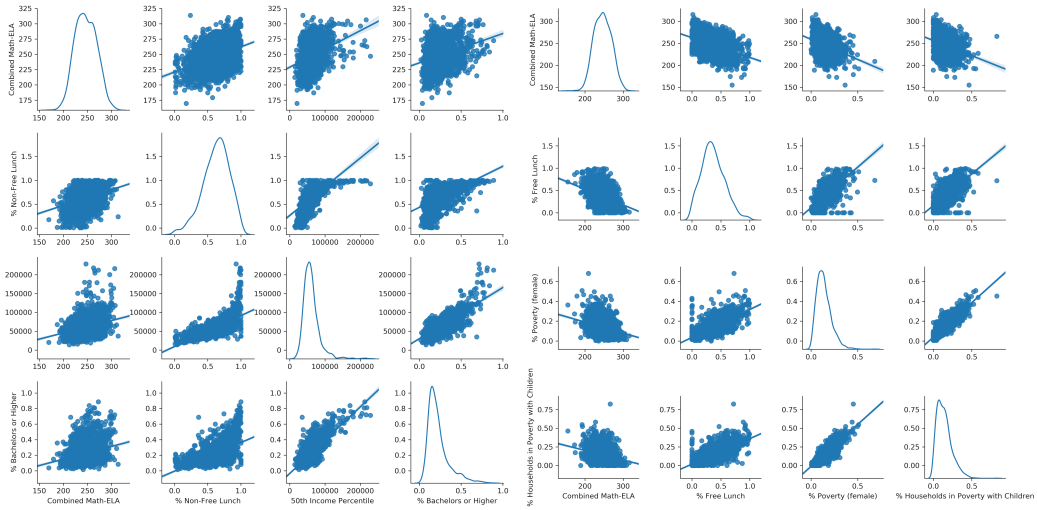
In fact, we can show that all positively and negatively correlated covariates are themselves correlated to each other in figure 12, which supports our assertion that focusing on race by itself is not sufficient to interpret the grander influence of multiple socioeconomic factors such as income and parental support.

Of course these linear correlation coefficients only piece together potential linear relationships between our covariates and test scores (we could further our analysis by attempting to perform higher-order polynomial regression). It does however reveal that perhaps we are misguided in how we approach affirmative action and the movement towards equal opportunities. The fact that the proportion of households that have single moms and the proportion of adults (and

Fig. 12. Pairwise Correlation Matrix for the Most Correlated Covariates to NAEP Test Scores

(a) Top 3 Positively Correlated Covariates

(b) Top 3 Negatively Correlated Covariates



transitively parents) that have Bachelors degree or higher are negatively and positively correlated to our measure of academic achievement highlights an important and often overlooked aspect of a student's education: their situation at home and their upbringing. Intuitively, students with parents who have regular work hours and can help them with homework and extracurricular activities will have a more fulfilling upbringing. Studies have shown that a positive influence from a guardian is extremely valuable in the formative years of a child.

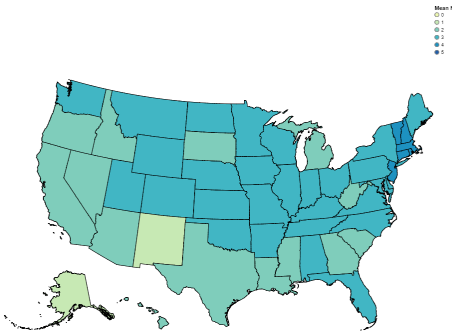
3.1.4 State Policy Variables. After examining socioeconomic factors, the next step in answering this question is to explore the impact of educational policies at the state level on state-wide educational performance.

First, we visualized the average NAEP score in both English and Math for each state. Figure 14a shows the average NAEP scores for English, while figure 14b shows the results for math. As shown in previously at the county level, scores are higher in the Northeast and Midwest, and lower in the South and West Coast. Because some states did not report NAEP math scores in a large number of regions, we concluded that state to state comparison on math scores would be inappropriate, so the remainder of the analysis focuses only on ELA scores.

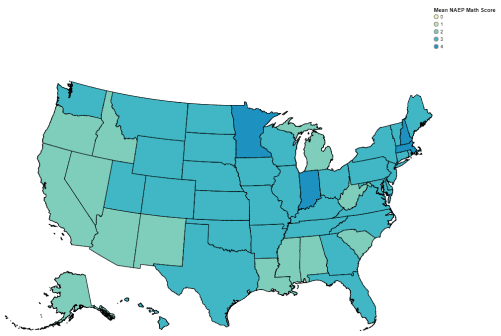
To explore this data, we first visualized the distributions of various policy variables across the 50 states. Figure 15a shows the total spending per student in each state. Notably, the spending is high in Alaska, Wyoming, and the Northeast. Spending is lower in the South and the Midwest. As shown in 15b, spending per student is correlated with improved ELA scores. However there are some notable outliers to this trend, particularly Alaska and DC in the bottom right corner of the graph, who have very high spending but low performance.

However, total spending can be misleading because school funding is spent on many different things, instruction being one of them. Thus it might be more useful to compare the spending per pupil on instructors' salaries. Figure 16 shows the relationship between the spending per pupil on instructors' salaries and NAEP ELA scores for each state. Several things are worth noting in this plot. Firstly, the correlation coefficient is higher than that of total spending, indicating that spending on instructor salaries is more correlated with improved test scores than spending in

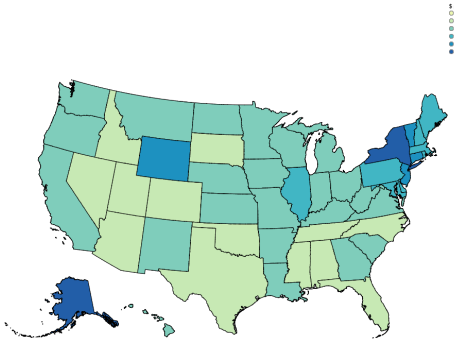
(a) Grade Equivalent English Scores by State in 2013



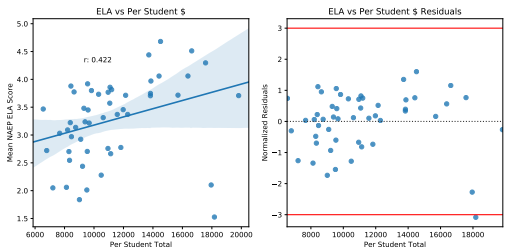
(b) Grade Equivalent Math Scores by State in 2013



(a) Total Annual Spending per Student in 2013 (Per State)



(b) Grade Equivalent ELA Scores vs Spending Per Pupil (Per State)



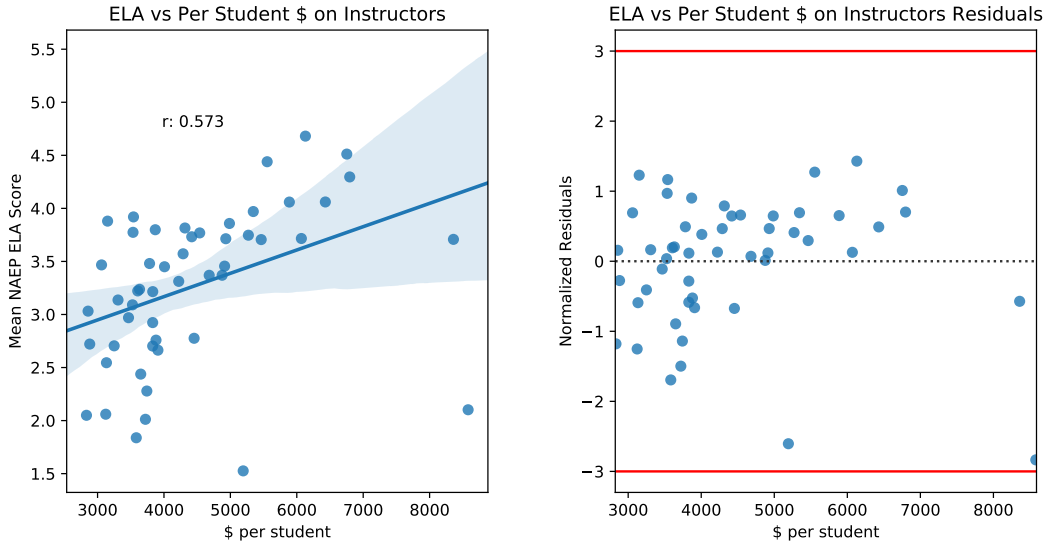
general. Secondly, the values on the x axis are significantly lower than in figure 15b, showing that the amount spent on instructors' salaries is significantly less than the total amount per pupil.

First, we visualized the average grade equivalent score in both English and Math for each state. Figure 14a shows the average equivalent scores for English, while figure 14b shows the results for math. As shown in previous sections, scores are higher in the Northeast and Midwest, and lower in the South and West Coast. Because some states did not report grade equivalent math scores in a large number of regions, we concluded that state to state comparison on math scores would be inappropriate, so the remainder of the analysis focuses only on ELA scores.

To explore this data, we first visualized the distributions of various policy variables across the 50 states. Figure 15a shows the total spending per student in each state. Notably, the spending is high in Alaska, Wyoming, and the Northeast. Spending is lower in the south and the Midwest. As shown in 15b, spending per student is correlated with improved ELA scores. However there are some notable outliers to this trend, particularly Alaska and DC in the bottom right corner of the graph, who have very high spending but low performance.

However, total spending can be misleading because school funding is spent on many different things, instruction being one of them. Thus it might be more useful to compare the spending per pupil on instructors' salaries. Figure 16

Fig. 16. NAEP ELA Scores vs Spending Per Pupil on Instructor Salaries (Per State)

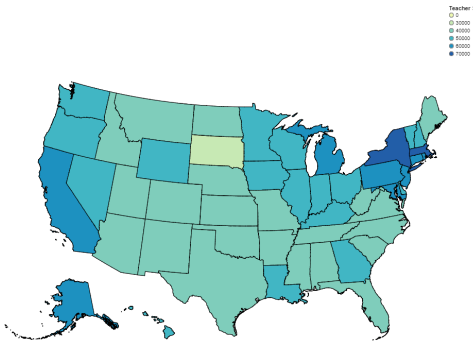


shows the relationship between the spending per pupil on instructors' salaries and grade equivalent ELA scores for each state. Several things are worth noting in this plot. Firstly, the correlation coefficient is higher than that of total spending, indicating that spending on instructor salaries is more correlated with improved test scores than spending in general. Secondly, the values on the x axis are significantly lower than in figure 15b, showing that the amount spent on instructors' salaries is significantly less than the total amount per pupil. This means that most of the money spent per pupil does not go to instructor salaries. The outliers in this data group are different than before. The rightmost points are New York and DC, which have high pupil expenditure but lower average scores. The bottom-most point is Alaska, which has mid-level expenditure but an extremely low score. It is worth noting that while Alaska had among the highest total spending per pupil, its spending on instructors' salaries per pupil is mid-range, showing that comparatively little money goes to instructor salaries. This is likely due to the low population density of Alaska, which necessitates a lot of small schools in remote areas, increasing the overhead cost.

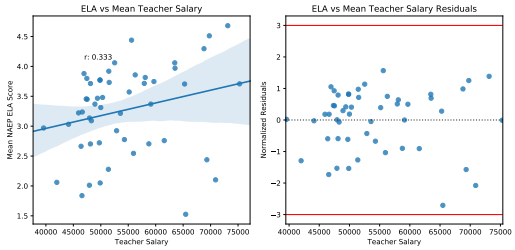
Another variable of interest is teacher salary, which presumably is related to teacher quality. Figure 17a shows the mean teacher salary in each state. Salaries are higher on the West Coast and in the Northeast, and lower in the central states. Figure 17b shows the effect of teacher salary on grade equivalent ELA score. It shows a weak positive correlation between the two, implying that higher teacher salaries are beneficial to students but not largely responsible for their scores.

Another major educational policy variable is funding source. Schools in the US are funded with a mixture of federal, state, and local funding. Figure 18a shows the percent of educational funding from the federal government for each state. Federal funding tends to be higher in the southern states, and lower in the Northeast. Figure 18b shows the effect of federal funding on grade equivalent scores. Surprisingly, increased federal funding is extremely well correlated with lower test scores. This is likely not a causal relationship, instead being because the federal government gives more money to low-achieving districts.

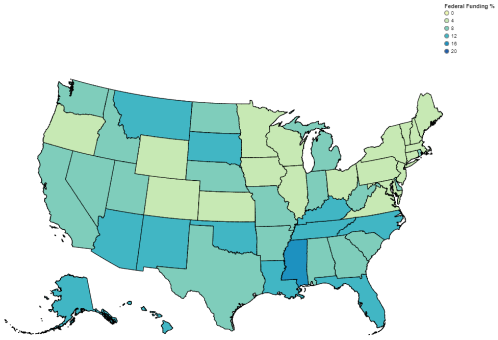
(a) Mean Teacher Salaries Per State



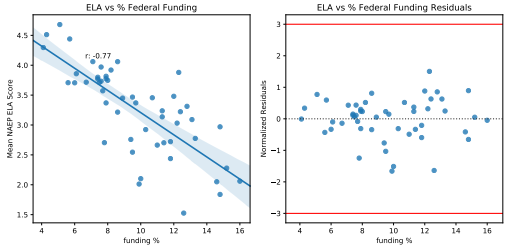
(b) Grade Equivalent ELA Scores vs Teacher Salaries (per state)



(a) Federal Educational Funding Per State



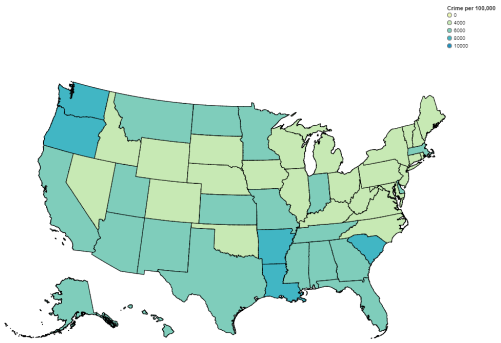
(b) Grade Equivalent ELA Scores vs Federal Educational Funding



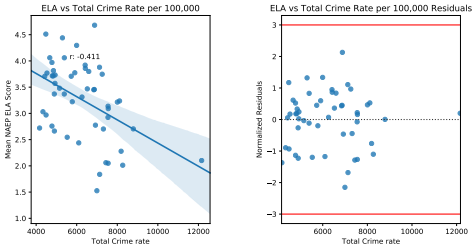
Finally, the effect of crime rate on grade equivalent was examined. Figure 19a shows the crime rate (number of crimes per 100,000 people) for each state. Crime rates are generally higher in the South and West Coast, and lower in the Northeast. Figure 19b shows the effect of crime rates on grade equivalent scores. Unsurprisingly, increased crime is correlated with lower test scores.

3.1.5 Conclusions of Initial Exploration. The initial exploration of the data revealed a large number of interesting trends, some of which were expected, and others unexpected. However, such a cursory analysis has several significant shortcomings. Most importantly, the only thing that has been examined is correlations. Since many different variables are correlated together (for example race and income), it is difficult to pinpoint the root causes of low test scores. Ultimately a good policy recommendation must be made based on causal data, so to do this a more sophisticated model is needed.

(a) Crime Rate Per State



(b) Grade Equivalent ELA Scores vs Crime Rate



3.2 Feature Selection

This section pertains primarily to the feature selection process for our hierarchical Bayesian model detailed in section 3.3.

In order to quantify educational outcomes, we have focused on mean NAEP scores in ELA and MATH for grade 3 students. The restriction to grade 3 students is due firstly to data availability as well as to ensure that we have a manageable scope, even though it would be possible to blend together various performance outcomes.

One of the primary difficulties with Bayesian analysis is the computational burden. Under time constraints, we have resorted to some rather crude feature selection schemes, in large part to the correlation between district level covariates, and the district mean NAEP scores. An excerpt of these results are given in the following table.

Table 3. District Level Covariate Correlations to District Mean ELA Scores

Covariate	ELA Correlation
flunch_all	-0.698712
perfrl	-0.698712
pernonfrl	0.698712
flunch_hsp	-0.692111
inc50all	0.615278
snap_all	-0.550066
baplus_mal	0.549585

From the entire set of district level features, we have selected the following:

3.3 Bayesian Hierarchical Modeling

In order to better understand the reasons for educational performance differences, a Bayesian hierarchical model was used.

Our analysis is inspired by the hypothesis that state-level policies have an impact on county-level or local education agency levels. A natural methodology for exploring this hypothesis is a Bayesian hierarchical model.

Table 4. Selected District Level Covariates

Covariate	Reasoning and Comments
elmgui	(Number of Elementary School Guidance Counselors) Quantifying the effect of guidance counselors is valuable because governments can take action to increase or decrease the number of guidance counselors, depending on their effect on educational outcomes.
aides	(Number of Instructional Aides) Similarly as above
tottch	(Total number of Teachers) Similarly as above
perhsp	(Percent Hispanics in Grade) We have included various racial statistics in order to account for effects of racial discrimination or disadvantage
perwht	(Percent White in Grade)
perblk	(Percent Black in Grade)
perasn	(Percent Asian in Grade)
baplus_all	(% of Adults With Bachelor's+) It seems natural that children whose parents are educated would perform well in school.
inc50all	(Median Income) One of our key interests is on the effect of inequality. We expect wealthier districts to perform better.
unemp_all	(Unemployment) A key measure of regional prosperity
perell	(% of English Language Learners) Finding significant effects of learning english as a second language could lead to recommending increased funding for language education
poverty517_all	(Child Poverty) Another key indicator of inequality
perfrl	(% Receiving Free Lunch) Checking for effect of free lunch program

Suppose $y_t^{(s)} \in \mathbb{R}^m$ is a set of measurable educational outcomes in each of m small geographic locales at time t and in state s. These measurable outcomes are, for instance, the estimated mean math score for grade 3 students. Next, suppose that $X_t^{(s)} \in \mathbb{R}^{m\times p}$ is a set of p measurable covariates (e.g. percentage of households in poverty) in each of the m locales, at time t and in state s. We may then fit the following Bayesian linear model:

$$\begin{aligned} y_t^{(s)} &\sim \mathcal{N}(X_t^{(s)} w^{(s)} + \Delta_t^{(s)}, \Sigma^{(s)}), \\ w^{(s)} &\sim \mathcal{N}(\mu_s, \Sigma_s) \\ \Delta_t^{(s)} &\sim \mathcal{N}(\mu_\Delta^{(s)}, \Sigma_\Delta^{(s)}), \\ \Sigma^{(s)} &\sim \mathcal{W}^{-1}(\Psi, \nu). \end{aligned} \tag{4}$$

The vector $w^{(s)}$ provides a measurement of the "effects" of the covariates in $X_t^{(s)}$ on the outcome $y_t^{(s)}$ (in state s) and the time dependent noise vector $\Delta_t^{(s)}$ serves to measure how stationary these effects are over time. If $\Delta_t^{(s)} \approx \Delta_{t+1}^{(s)}$ then the interpretation is that the distribution relating $y_t^{(s)}$ and $X_t^{(s)}$ does not vary significantly over time. A simpler approach may be to flat out ignore the yearly labels in the data, but the Bayesian approach gives us a natural way to keep account of this potential yearly variation.

The result of fitting this model is a posterior distribution $p(w^{(s)}, \Delta^{(s)} \mid X^{(s)}, y^{(s)})$, where $\Delta_t^{(s)} = (\Delta_1^{(s)}, \Delta_2^{(s)}, \ldots)$, and similarly for $X^{(s)}$ and $y^{(s)}$. Our analysis then depends on the interpretations of the posterior of the coefficients $w^{(s)}$. If this posterior of $w_1^{(s)}$ is centered at 0 with a very small spread, then our model is "confident" that the associated covariate has little effect on the outcome. On the other hand, if $w_1^{(s)}$ is centered entirely away from 0 with a small spread, then the model is confident of the importance of the associated covariate. The size of the spread in the posterior is the Bayesian alternative to frequentist confidence intervals, so if the posterior is centered away from zero, but with a large spread, we can still conclude that the covariate is relevant, but are less sure about it's precise value.

Table 5. State Level Covariate Correlations

Covariate Description	Correlation to State-wide Mean NAEP ELA Score
Mean NAEP Math Score	0.901474
Federal Total (funding %)	-0.753588
Federal Title I (funding %)	-0.629793
Total Crime rate	-0.458988
Instruction: Total Salaries and wages ($ per student)	0.397472
Total Salaries and wages ($ per student)	0.396004
Instruction: Total ($ per student)	0.375053
Per Student Total	0.323200
Instruction: Total Employee Benefits ($ per student)	0.319817
Total Employee Benefits ($ per student)	0.288471
State Total (funding %)	-0.253791
Local Total (funding %)	0.248456
Teacher Salary	0.248302
Support Services: Total ($ per student)	0.231382
Support Services: Pupil Support ($ per student)	0.226809
Local Taxes and Parent Govt Contributions (funding %)	0.220160
State General Formula Assistance (funding %)	-0.207616
Other Government (funding %)	0.116060
Support Services: School Administration ($ per student)	0.072524
Charges (funding %)	0.066854
Support Services: General administration ($ per student)	-0.029237
Support Services: Instructional Staff Support ($ per student)	0.001390

Table 6. Selected State Level Covariates

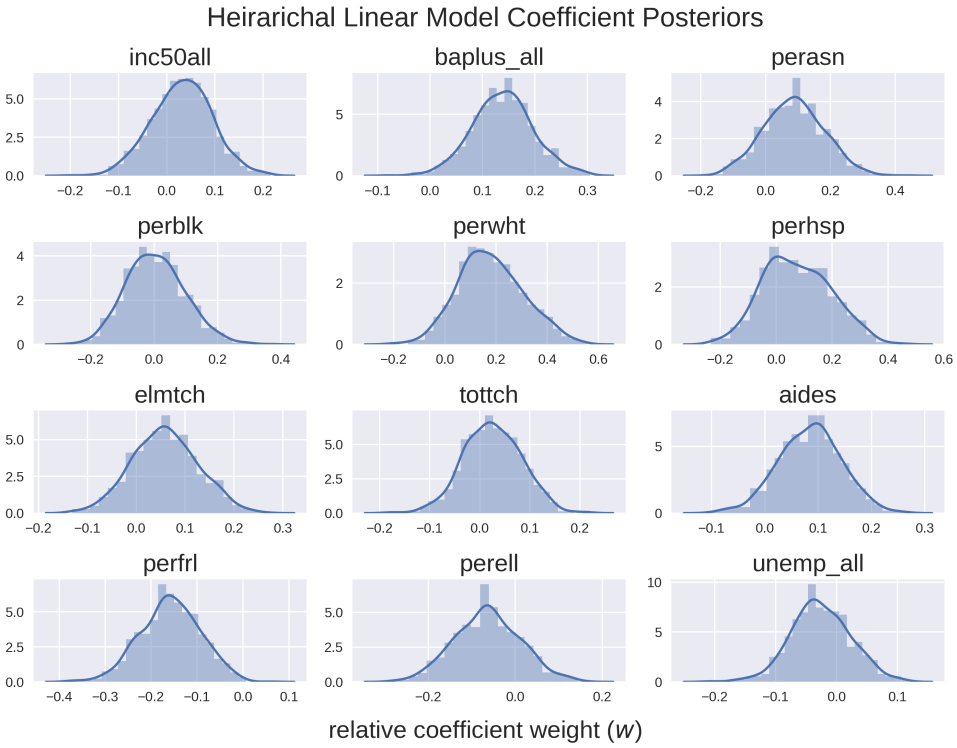
Covariate	Reasoning and Comments
`ts_2012`	(Teacher Salaries in 2012) We wish to control for the effects of well paid teachers, looking for less obvious means of improving educational attainment.
`per_student_total`	(Per Student Total Funding) Similarly, we want to find ways to improve student performance aside from simply increasing funding
`crime`	(Aggregated Crime Rates) We want to avoid any confounding effects of varying crime rates.

Now, fitting 50 state by state models, aside from being a rather arduous task, does not provide us very much information about the effect of state policies. Our dataset provides a number of state-level covariates, for instance crime levels or teacher salaries, and we want to find relationships, or "effects", of these covariates on the county level. In order to do this, we build a hierarchy into the model (4) by replacing the generic prior $w^{(s)} \sim \mathcal{N}(\mu_s, \Sigma_s)$ with

$$\begin{aligned} w^{(s)} &\sim \mathcal{N}(Z^{(s)}w + \mu_s, \Omega), \\ w &\sim \mathcal{N}(\mu_0, \Omega_0), \end{aligned} \tag{5}$$

where $Z^{(s)} \in \mathbb{R}^{50\times q}$ are state level covariates.

Fig. 20. Posterior μ_s



This second level w vector is to be interpreted as quantifying the effect of state level policies (e.g. teacher salary) on the coefficients in $w^{(s)}$. That is, if $w_i^{(s)}$ is considered desirable, then our state level policy recommendation would be to modify whatever state level covariate is predicted by the model to improve $w_i^{(s)}$.

Moreover, the vector μ_s is now an indicator essentially of the mean of $w^{(s)}$, controlling for the effects of the state level covariates. Through this model we can quantify the effects of the district level controlling for the state level's policies. Trace plots for this variable are shown in 21.

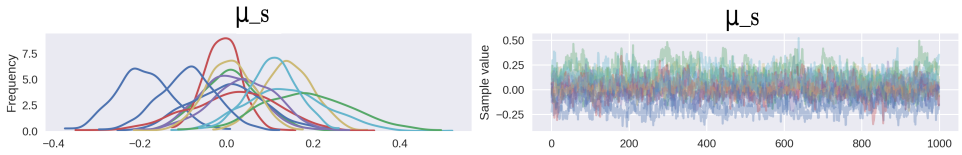
The figure 20 shows the MCMC posterior approximations of μ_s, from which we can draw some conclusions.

The posterior of μ_s[aides] is quite clearly supported away from 0, and hence the number of teacher's aides is positively associated with higher NAEP scores, controlled for the state level covariates (e.g. crime rates and teacher salaries) summarized in table 6. So, teacher aides improving performance can't be a result of confounding with regions having higher teacher salaries.

We can also see that μ_s[perwht] is a consistent predictor of good NAEP performance – racial inequality is persistent, after controlling for other state level covariates.

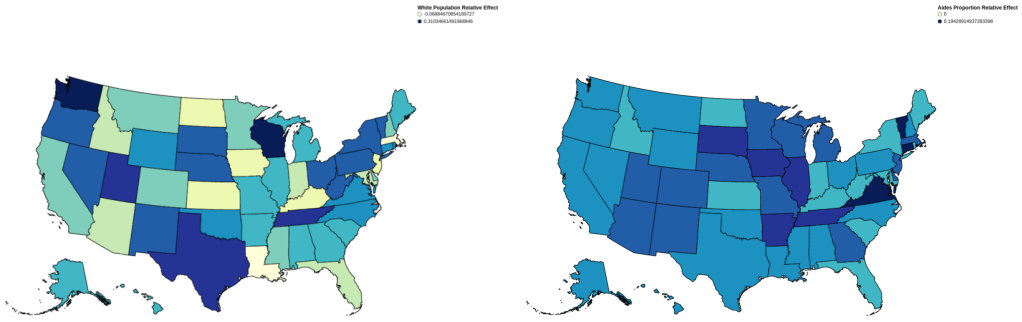
Finally, μ_s[perfrl] is a consistent predictor of poor NAEP performance. We suspect that this is due to uncontrolled confounding in the data – students who need free school lunches tend to already be disadvantaged. We have not been able to fully develop our modeling methodology within the time constraints.

Fig. 21. Trace Plots for μ_s



(a) $w^{(s)}$[perwht] Posterior Means (Darker is Bigger)

(b) $w^{(s)}$[aides] Posterior Means (Darker is Bigger)



At the district level, we consider the vectors $w^{(s)}$. We have plotted the `perwht` and the `aides` posterior means in figures 22a and 22b. From these figures, we see that some states give less racial advantage to white students than others. However, there are still many states, Texas stands out, in which simply being born white confers significant advantages. Teachers aides on the other hand, are seen to uniformly improve student performance, and should be a focus of policymakers.

4 CONCLUSIONS AND POLICY RECOMMENDATIONS

Many districts and policy makers have attempted to help students from lower socioeconomic backgrounds by providing additional guidance counselors and after-school programs. From the scope of our data sets, we see that although the median test scores across the most troubling states like California has increased over the years (from a median combined score of 232.7 to 235.6 from 2009 to 2013), they still pale in comparison to a high achieving state like New Jersey (whose median combined score increased from 253.9 to 257.2).

In order for achievement scores of students from all parts of the US to catch up to those of the high performing districts, **we argue that policies addressing a child's access to mentorship both inside and outside of school need to be prioritized.** This could potentially be a huge stepping stone towards remedying the inequalities in the socioeconomic class that has shown to be non-trivially correlated to test scores and academic achievement.

While the No Child Left Behind act of 2001 and the Race to the Top Fund of 2009 has increased coverage on the inequalities of education around the country, there needs to be more direct policies that help address the root issue: the socioeconomic differences between the high-performing and low-performing districts. Instead of incrementally

improving the conventional guidance counselor programs, **a more effective solution would be to set up a mentorship program where older students with positive life goals can help mentor the younger students who demonstrate a need for a guardian figure.**

Our exploration of the effect of state education budgets revealed that education performance increases with teacher salary, spending per student, and spending per student on instructor salaries. It decreases with increasing crime and increasing federal funding, although we think it is unlikely that the latter is a causal relationship.

The results of our Bayesian hierarchical modeling suggests that teachers aides provide robust improvements in NAEP scores for young students, controlling for teacher salaries, per student funding, and state wide aggregate crime. Hence, we conclude that an increased focus on teacher's aides can improve US educational attainment, without simply increasing funding across the board.

Furthermore, our Bayesian analysis reveals that higher district level percentages of white students are robust predictors of good NAEP performance, which implies that racial inequality is a persistent problem.

REFERENCES

Consumer Price Index for All Urban Consumers (CPI-U), United States. (????). https://www.statbureau.org/en/united-states/cpi-u

US Census Bureau. Public Education Finances: 2013. (????). https://www.census.gov/library/publications/2015/econ/g13-aspef.html

Wayne J. Camara and Amy Elizabeth Schmidt. 1999. *Group Differences in Standardized Testing and Social Stratification. Report No. 99-5.* College Entrance Examination Board. https://eric.ed.gov/?id=ED562656

Martin Carnoy and Richard Rothstein. 2015. What International Test Scores Tell Us. *Society* 52, 2 (April 2015), 122–128. https://doi.org/10.1007/s12115-015-9869-3

Drew DeSilver. 2017. U.S. studentsfi academic achievement still lags that of their peers in many other countries. (2017). http://www.pewresearch.org/fact-tank/2017/02/15/u-s-students-internationally-math-science/

Julia Ryan. 2013. American Schools vs. the World: Expensive, Unequal, Bad at Math. *The Atlantic* (Dec. 2013). https://www.theatlantic.com/education/archive/2013/12/american-schools-vs-the-world-expensive-unequal-bad-at-math/281983/

5 APPENDIX

5.1 Predictive Modelling

Over the course of our analysis, we also fitted a few predictive models, namely a linear regression model (L1, L2, and Elastic Net penalized), a random forest model, and an XGBoost tree model. The feature vector was composed of all our district demographic covariates and our target was the combined NAEP Math-ELA score. We obtained r^2 values of $0.96, 0.95$ and 0.92, respectively on our test set which indicates that there is very good predictive potential for test scores by demographics data.

We noted that these almost perfect predictive models were in a sense overfitting the tested grade as the tested grade had a very high linear relationship to score. After adjusting to a single grade, we had r^2 score values of $0.85, 0.82$ and 0.82, respectively.

PCA decomposition also revealed that 80% of the variance in the dataset could be explained with only 12 principal components out of 140+ initial features, which denotes that much of the feature set is correlated as shown in our analysis.