Exploratory Analysis of Texas Public High School Rankings and Standardized Exam Scores for 2011

Nivriti Chowdhry, Matt Delhey, Jiandi Mo

December 7, 2012

Contents

1 Introduction

Recent efforts by both national and state governments to increase the quality and accountability of the United States' public high schools has resulted in a data deluge in education. Everything from standardized exam scores to per student budgets are being reported, allowing for more data analysis than has been conducted in the past. Analysis of these data allow for a better understanding of standardized exam metrics and insights on how to improve public education. Specifically, thorough data analysis could potentially identify which aspects of public education play the largest role in the quality of student education.

Today, public education is dominated by quantitative metrics. Many different aspects of a school, from student-teacher ratio to geographic location, can be reduced to a single numerical ranking. Without a deep and clear understanding of each of the metrics that goes into determining a school's ranking, one cannot effectively make use of the data available. For this reason, further evidenced by the findings of the presented analysis, blind dependency on quantitative metrics can lead to inaccurate depictions of a high school's actual educational environment.

Our first finding attempts to explain improvements in school rank from 2010 to 2011 by comparing linear models and found that changes in graduation rates are the most significant factor in changing a school's overall ranking. However, the finding also shows how improvement in rank is dependent on holistic improvement rather than focused efforts to improve the factors most significant in determining school ranking. Our second finding focuses on the cost efficiency of Texas public high schools, in which we found that counties with high

populations tend to be less cost efficient. While counties with larger schools catering to more students are given more money per student to operate, smaller counties tend to achieve similar results and rankings for less expenditure per student. Our third finding takes a closer look at state and national standardized tests and finds evidence that a school's percentile ranking in the social studies segment of the Texas Assessment of Knowledge and Skills is most indicative of that school's SAT and ACT mean scores, more so than the reading, math, or science segments. This part of the analysis also drew similarities between SAT and ACT scores. Our fourth finding is an analysis of race, economic disadvantage, and SAT mean scores and shows that SAT scores are negatively correlated with the percentage of a school's student body that is economically disadvantaged. Furthermore, it can be seen that schools with larger concentrations of White, Asian, and Pacific Islander students tend to be less economically disadvantaged and have higher SAT mean scores. Schools with larger concentrations of African American and Hispanic students tend to be more economically disadvantaged and have lower SAT mean scores.

2 Data Set and Ranking Methodology

The data set was acquired from a state-wide study of Texas schools compiled by the nonprofit research and advocacy group Children at Risk. The study utilized Texas Education Agency (TEA) records and ranked high schools based on a set of 16 academic criteria. The original data set contained $6,386$ public schools and included information on school location, student body demographics, standardized test scores, average expenditure per student, retention rates, and graduation rates. Additionally, the 2010 data set published by Children at Risk was used for the analysis of changes in rank from 2010. The rankings were created using the following metric variable weights:

Metric	Ranking Weight
Graduation Rate	18%
Percent Economically Disadvantaged	17%
Reading Gain/Loss	7.5%
Math Gain/Loss	7.5%
Recommended High School Program	5%
Advanced Courses	5%
AP/IB Test-Takers	5%
AP/IB Students Passing	5%
Attendance Rate	5%
SAT/ACT Test-Takers	5%
Mean SAT Scores	5%
Mean ACT Scores	5%
TAKS Commended Reading	2.5%
TAKS Commended Math	2.5%
TAKS Commended Social Studies	2.5%
TAKS Commended Science	2.5%

Table 1: The ranking is determined half by standardized exams and half by soft factors such as Graduation Rate and Attendance Rate. Notice that several of the metrics only consider participation rather than direct performance.

The ranking methodology[1] used by the Children at Risk organization for Texas public high schools we designed with the goal of capturing the overall quality of education found at each campus. The ranking

[1]The Children at Risk organization also published a detailed explanation of their ranking methodology alongside their data set, which can be found at http://childrenatrisk.org/research/school-rankings/methodology/.

system thus represents a compilation of factors that indicate the degree to which a campus has prepared for post-secondary school success, or in other words the college-readiness of each school's graduates. The ranking methodology is standardized across Texas through the use of the z-score statistic.

The use of percent economically disadvantaged is of particular interest. Children at Risk use this variable as an adjuster for the difference in students who attend each school. Research has consistently shown that poverty is a predictor of student graduation rates and post-secondary academic success, Children at Risk argue, and so it is assumed that more effort is required to successfully educate low-income students. The goal, then, of using the percent economically disadvantaged metric is to give credit to schools who have the increased difficulty of teaching low-income students which is an indirect measure of the quality of the school.

3 Explaining Changes in Rank from 2010

What causes some schools to improve in rank while others fall behind? Do improving schools significantly increase one metric or do they advance slightly across all categories? Is there a pattern to school improvement? In order to explore possible answers to these question and better understand the nature of school improvement, we analyzed the change in school rankings from 2010 to 2011.

Figure 1: The distribution of changes in rank is heavily skewed towards the middle tier schools. At the extremes, we see almost no change in rank. This makes sense because it would be difficult for a school to change from say, rank 2 to rank 22, whereas a change from 500 to 520 would be much easier.

Before detailing our findings from our linear model approach, it is worthwhile to also demonstrate the general behavior of the change in rank. From Figure 1 we can see that the majority of the change in rank is

conducted by the middle-tier schools; Texas's best and worst public high schools stayed about the same in rank from 2010 to 2011. This is an interesting finding in itself because it shows that schools on the extremes are some how more fundamentally consistent, or perhaps extremely different in regards to educational quality to middle-tier schools. This seems to match with our intuitions about the quality of high schools: the best and the worst schools are far and away the best and the worst schools, while the middle-tier schools are more conducive of change in educational environment. This information qualifies the rest of this finding because when we discuss change in rank we are most likely talking about these middle-ranked schools.

Our methodology for quantifying explanatory power was to assume a linear relationship between each metric and the school's rank and then compute the adjusted R-squared value. The adjusted R-squared value is then interpreted as the proportion of rank variation in the model explained by the metric. This gives us a single value describing explanatory power to compare across different metrics. There are, however, some potential issues with our methodology. The first is that the R-squared value does not indicate that the change in the metric is the true cause of the change in rank. Similarly, by leaving out factors likely to be involved in causing the change in rank, we are introducing omitted variable bias into our model. Finally, one could challenge the assumption that the relationship between change in rank and change in each metric is linear. One improvement to our methodology outside of fundamentally abandoning linear models would be to look at changes in rank across several years rather than just one. This was not possible using data sets published by Children at Risk because they did not begin to cover all Texas public high schools until 2009, in which the organization used significantly different ranking methodology compared to 2010 and 2011 effectively rendering cross-year rank comparisons unintelligible. While we recognize the caveats to our analysis, we believe that our results are valid as evidence for our conclusions regarding explaining changes in rank because the general behavior and trends between the change in metrics and the change in rank could reasonably be conceived of as being modeled adequately with a linear approach.

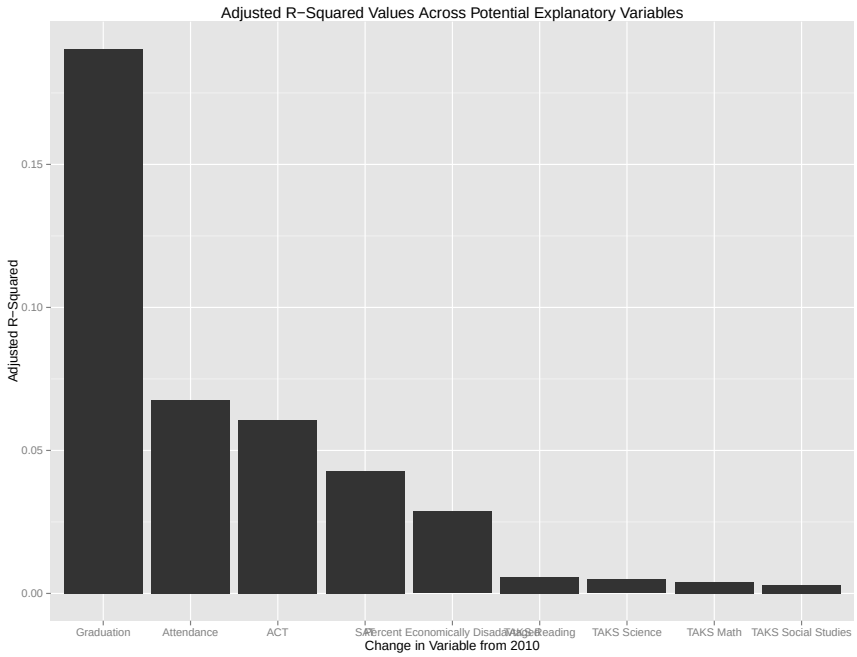

Figure 2: We can notice that several of the metrics plotted here do not match up very well to their position in the ranking methodology outlined in the previous section. Change in soft factors seems particularly explanatory, and change in ACT scores is somewhat more explanatory than change in SAT scores.

Moving on to the results of the linear model approach, we can see from Figure 2 that change in graduation rate is the biggest explainer of a school's change in rank. The other predictors (which were the other most significant explainers of change in rank), such as percent economically disadvantaged, are not nearly as significant in terms of explanatory power. We can also see that no particular metric has a significant amount of explanatory power in the linear model. Whereas part of this is due to the non-linear relationship between the change in metric and the change in rank, we also take this as evidence that the schools that did change in rank did so by improving across all metrics rather than just one or two. In other words, we find this evidence in favor of the hypothesis that schools improve and rise in rank holistically, improving across several, if not all, metrics rather than by some sort of magic bullet of increased attendance or standardized exam scores. While it is in the nature of the ranking methodology detailed previously to include multiple metrics, it would also be possible for specific metrics to have a explanatory power that differs significantly from their ranking percentage in the methodology.

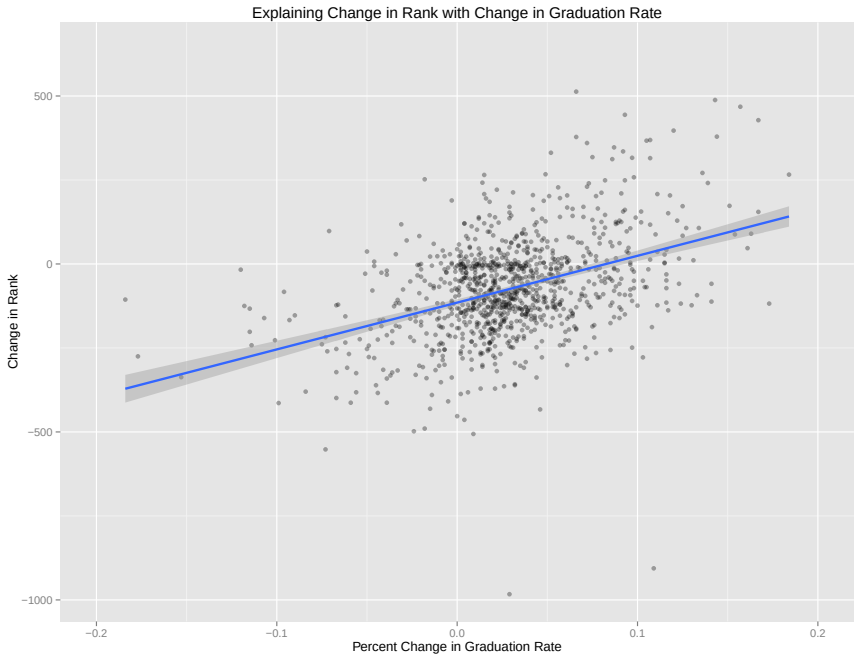

Figure 3: Notice the general linear trend between change in graduation rate and change in rank. Many of the points that do not significantly change in rank also do not significantly change in graduation rate, and this is where we see the non-linear pattern in the center of the graph.

In order to demonstrate more fully the linear relationship being discussed in Figure 2, we have included the change in graduation rate plotted against the change in rank with the linear model included. From the plot we can see that a general positive linear trend between the variables: as schools increased graduation rate, their rank also increased. This plot demonstrates that while a linear trend may exist, the explanatory power of the change in graduation rate alone is not sufficient to model the change in rank. Overall, we have shown that change in rank occurs mostly in middle-tier high schools and that this change in rank is determined holistically rather than by any single metric measured by the data set.

4 Cost Efficiency Across Counties

What high schools and school districts are spending their money most efficiently? This is a question that has become increasingly more important in the face of teacher firings and an overall decrease in public education funding. Its no longer good enough for schools to simply provide high-quality education for their students, instead it can be just as important to do so economically. Additionally, if we can identify which schools are best educating their students most cost effectively, we can learn from these schools and adopt their best practices. Therefore cost efficiency can serve as a more accessible guide for parents, educators, and community members on the performance and efficiency of their local schools while at the same time providing crucial information for educational leaders in the structuring of our public education systems.

For our analysis, we operationally defined cost efficiency as the amount of rank you get for the average expenditure per pupil, where rank is determined by the Children at Risk organization as detailed previously. That is to say, a school would have a high cost efficiency if it had a high rank but a low average expenditure per pupil. It is important to notice here that cost efficiency is partly relative to the performance of other Texas high schools, indicating that it is somewhat robust to state-wide educational changes. More specifically, we calculated a school's cost efficiency by dividing the average expenditure per pupil by it's ranking percentile. For schools with a ranking percentile of zero, we substituted a ranking percentile of 0.05. The substitution is reasonable because the lowest recorded non-zero ranking percentile in the data set was 0.09, and it is reasonable to assume that the zero percentile rankings were rounded down from values most likely less than or equal to 0.05.

While knowledge regarding the cost efficiency of individual schools is useful for local officials, we chose to generalize cost efficiency across counties in order to facilitate a more comprehensive, state-wide view of cost efficiency across Texas. Moreover, budget related concerns are usually decided at the county or district level rather than the individual school level. We performed our generalization by assigning each county a cost efficiency by taking the average cost efficiency of all of its high schools; the top five most efficient and inefficient counties can be seen in Table 2 below. We then created a ranking system for Texas counties based on their average cost efficiency. Finally, we mapped the cost efficiency rank percentile to a map of Texas, which can be see in Figure 3. It's important to note here that while lower cost efficiency values are better, higher cost efficiency percentile is better because it indicates that the county ranks higher in cost efficiency.

	The Best			The Worst	
Rank	County	Cost Efficiency	Rank	County	Cost Efficiency
1	Franklin	86.142	1	Travis	10,159.353
2	Erath	86.471	2	Galveston	1,919.705
3	Fayette	90.353	3	Dallas	1,754.283
4	Mills	91.175	4	Duval	1,544.538
5	Gillespie	91.656	5	Harris	1,475.371

Table 2: Notice that Harris county, where Houston resides, is ranked fifth in worst cost efficiency, falling in line with the trend of large cities correlating with poor cost efficiency.

Looking at Table 2 we can see a large discrepancy in spending between the most and least cost efficient counties. This provides evidence that there is a real difference in the way school districts and public high schools are spending their money, or, perhaps more likely, how much money they are able to spend. In particular, Travis county, which is home to the city of Austin, appears significantly less cost efficient than any of the other counties. One possible explanation for this is that the Austin school district, in virtue of being the capital of Texas, is more likely to experiment with alternative methods of education, especially those requiring expensive equipment or new technology. This hypothesis draws evidence from Austin's

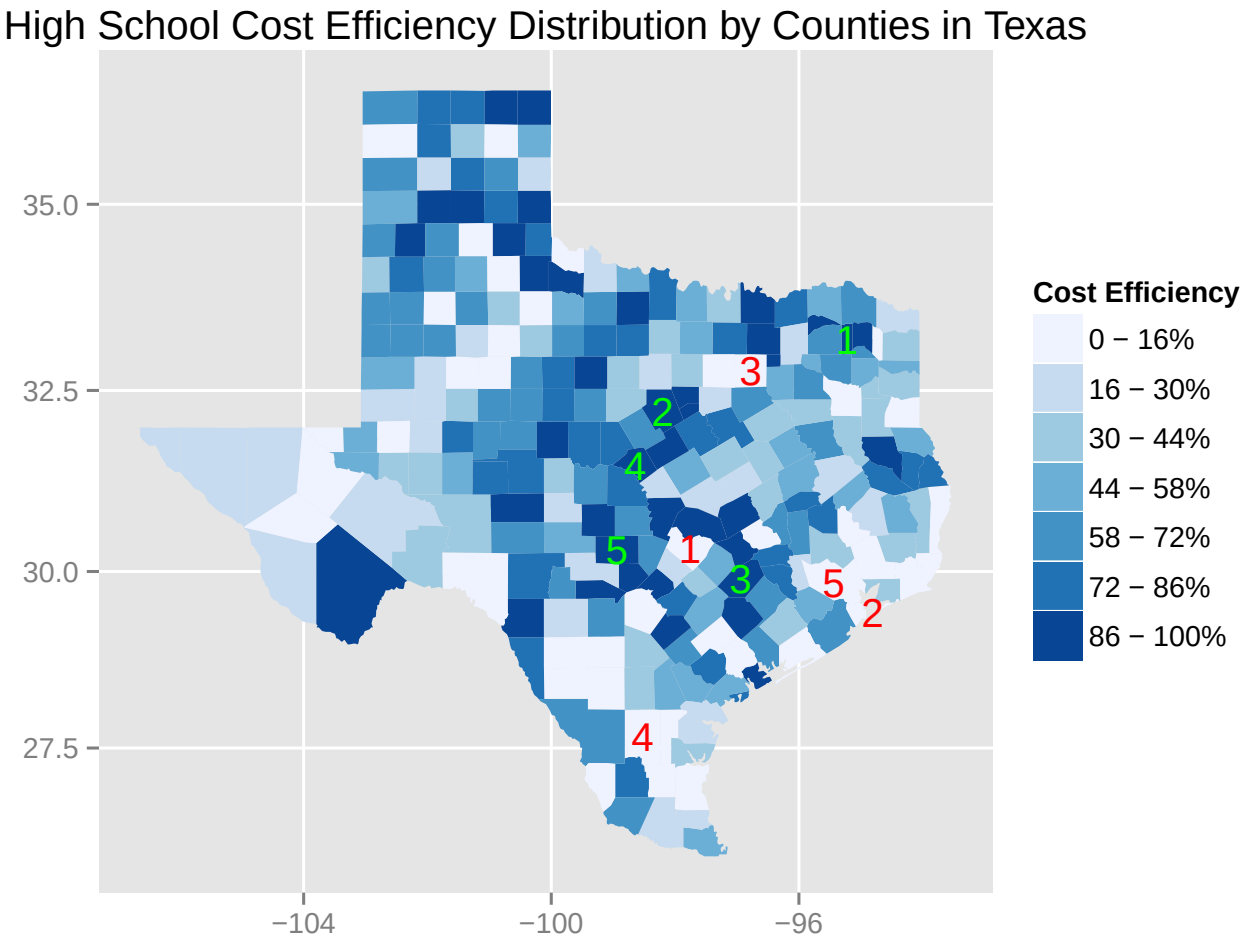

Figure 4: Notice the geographical discrepancies in cost efficiency, with high concentrations of poor cost efficient school districts in East and South Texas. West Texas seems to have more diversity in efficiency.

many magnet schools, most of which involve some kind of technology focus. We believe that it is safe to assume that these kinds of educational experiments could possibly explain the district's very high educational inefficiency. This hypothesis could be taken as evidence that these educational experiments have somehow failed, or at least have not proven useful in increasing the rank of the district relative to the state of Texas.

More generally, there are a two interesting trends we found from the data analysis. The first is a negative relationship between the population of each county, or more specifically, counties with large cities, and the district's cost efficiency. That is, cities with large populations tend to have poor cost efficiency in their public high schools. We can see this trend in Table 2: Travis, Galveston, Dallas, and Harris counties all have some of Texas' largest cities and all appear on the top five worst cost efficient counties. A possible explanation for this trend is that there is an overhead or administration cost required for larger districts that does not equate to an equal improvement in education for students. It could also be the case that these larger cities have a higher percentage of economically disadvantaged students and that the adjustment used in the rankings does not sufficiently take in account this difficulty. It is also important to note here that this is only a trend: Duval county is small relative to the likes of Harris and Travis and is similarly cost inefficient.

The second trend is a pattern regarding cost efficiency based on geographical location. Using the map of Texas in Figure 3, we can see that counties in the South and East Texas tend to be more cost inefficient whereas counties in North and Central Texas tend to be more cost efficient. Explanations for this phenomena seem even more evasive than the size of the district. However, we hypothesize that one possible explanation is the higher cost of education for immigrant and English as a second language students who are more prevalent in these regions than in the rest of Texas. Students who are not native English speakers must struggle with the language barrier before even beginning to learn new material, which could also explain a poorer ranking even if financial concerns were held constant. In fact, it is often the case that immigrant students require more educational resources to teach while still performing poorly in the classroom. While there is not sufficient evidence that this is the case, more analysis on the type of students in these regions could lead to a possible answer as to why we see this regional distribution. Overall, the data presented in this finding indicates that there are important patterns in cost efficiency with regards to size of a school district and its location within Texas.

5 TAKS as an Indicator of SAT/ACT Scores

The TAKS test, or the Texas Assessment of Knowledge and Skills, is a standardized test administered to all public school students in the state of Texas. The test is divided into four subjects - math, reading, sciences, and social studies. The SAT, or Scholastic Aptitude Test, and the ACT, or American College Testing, are similarly broken down into math and verbal sections. Reported SAT scores only cover the math and verbal scores and disregard the writing scores. An analysis of high school students' scores percentile ranking in the social studies TAKS test, as well as in math, reading, and science TAKS test, show that their percentile ranking in the social studies test is the best indicator of how they will perform on the SAT or ACT.

This may be because the social studies tests aims to assess certain skills unique to the exam. These skills could include understanding historical issues and events and the effect that geographic, economic, social, and political influences have on those events. The social studies portion of the Texas standardized test also assesses a student's critical thinking and analysis of questions. Within the test, students can find questions that ask them to analyze graphs and data about measures like population and other questions that require them to read a certain historical document and answer questions based on their understanding.

These data analysis and reading comprehension skills are tested in the SAT and ACT as well. Other subjects tested through the TAKS test, like math, reading, and science, are subject specific. Social studies incorporates the understanding, mathematical analysis and reading comprehension vital for the other subjects and on for the SAT and ACT tests.

A linear model of the relationship between percentile rankings of schools in math, reading, and science

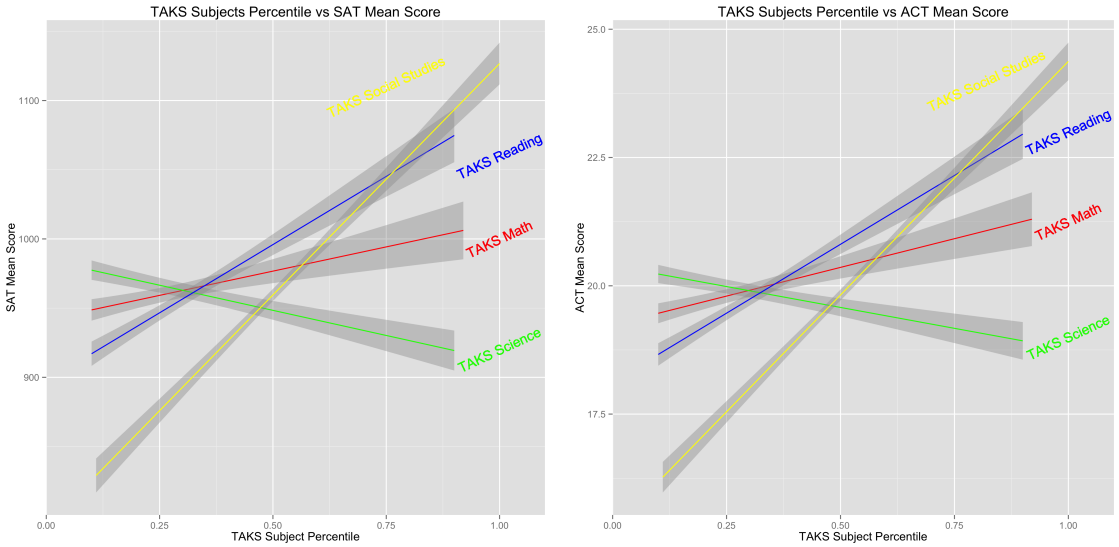

Figure 5: (Left) Layered line graph showing the relationship between TAKS percentiles and SAT mean score. Lines are colored based on subject tested and are accompanied by labels. The yellow line shows the relationship between social science TAKS percentiles and SAT mean scores. Reading, math, and science are represented in blue, red, and green respectively. (Right) Layered line graph showing the relationship between TAKS percentiles and ACT mean score. Lines are colored based on subject tested and are accompanied by labels. The yellow line shows the relationship between social science TAKS percentiles and ACT mean scores. Reading, math, and science are represented in blue, red, and green respectively.

TAKS tests and the average SAT and ACT scores of students at that school show some interesting trends. Both math and reading are positively correlated to SAT and ACT scores, as can be expected. The format of the math and reading portions of the TAKS test are similar to the format of math and verbal sections on the SAT and ACT scores as well. However, TAKS Reading percentiles are more significantly correlated to SAT and ACT mean scores. Higher TAKS Math percentiles do not affect SAT and ACT mean scores nearly as much. This could be because the level of math tested at a state level, and therefore incorporated into the public school curriculum, does not match the level of math tested on national SAT and ACT tests. If this is the case, higher percentile rankings in TAKS Math may give students a slight edge on the SAT or ACT, but not enough to cause a significant positive correlation between state and national tests. TAKS Reading, on the other hand, is significantly correlated to SAT and ACT mean scores. This could indicate that the state curriculum that aims to teach the skills necessary to rank in higher percentiles of the TAKS Reading test also prepare students for SAT and ACT testing.

Surprisingly, TAKS Science percentile rankings are negatively correlated with SAT and ACT scores. This may be because neither the SAT nor ACT tests, as presented in this data set, incorporate scores for a science portion of the exam in the mean score even though the ACT does include a Science Reasoning section. This is similar to how the Writing section is taken out of the SAT score, leaving maximum possible the SAT mean score at 1600 points for two sections worth 800 points each. In this situation, the relationship between TAKS Science scores and SAT and ACT mean scores may not be relevant at all. It seems unlikely that higher scores on any portion of a standardized state exam would place public school students at a disadvantage on standardized exams administered at a national level. The relationship may be entirely of correlation and not at all of causation in either direction. Another point to be noted is that the SAT and ACT tests are similarly correlated to the different TAKS test sections. This could indicate that the SAT and ACT are very similar to each other, and therefore have similar relationships to different TAKS test sections.

A deeper understanding of what topics are covered in each section of the TAKS, SAT, and ACT tests would allow a more thorough comparison and analysis of how the scores are related. It is possible that the topics and skills covered in the TAKS test, which are the same topics incorporated into the Texas public school curriculum, are very similar to the topics covered in the SAT and ACT tests. For example, the Texas public school math curriculum may not cover all the topics tested in the SAT and ACT, while the reading and English curriculum may be more closely matched. Most of all, this analysis would determine possible reasons for why TAKS Social Science percentile rankings are the best indicator of SAT and ACT mean scores. The analysis could show that similar skills are taught and tested in social studies as are needed to perform well on the SAT and ACT tests.

A detailed description of how TAKS percentile rankings and SAT and ACT mean scores are calculated would also be helpful. These descriptions could explain whether certain sections on the SAT and ACT tests are eradicated for the purposes of determining a mean score. As speculated earlier in the report, SAT Writing is not included in the mean score. Sections could similarly have been left out of the calculation for ACT mean scores.

6 SAT Scores by Race

An initial attempt to uncover simple relationships between different variables reported in the data set and SAT mean scores showed that schools in which higher percentages of the student body are economically disadvantaged typically have lower scores. The relationship, as seen in an initial figure, is reasonably linear. However, this simple graph does not convey much about what other variables affect both SAT mean scores and percentage of student bodies that are made up of economically disadvantaged students.

Other regressions that were run showed that the concentration of each race - African American, White, Asian or Pacific Islander, and Hispanic - affects SAT mean scores differently. Some races also tend to be more highly concentrated at different levels of economic disadvantage, which is measured by the percentage of a

school's student population that is economically disadvantaged.

All these relationships were best seen by adding a color gradient to scatter plots. The points are placed according to the relationship of the percentage of each school's student body that is economically disadvantaged and the school's SAT mean score. The gradient within these points shows where higher concentrations of the different races are placed.

The first of the four scatter plots portrays analyzes the distribution of different concentrations of White students in a school's study body over the negatively correlated percentage of disadvantaged students and SAT mean scores. Higher concentrations of White students are represented by lighter points, while lower concentrations of White students are represented by darker, purple points. The graph shows that schools with higher concentrations of White students tend to be less economically disadvantaged and report higher SAT mean scores than schools with low concentrations of White students.

The second graph shows a very similar relationship between the concentration of Asian and Pacific Islander students among the schools. It is important to note that the gradient only goes between 0 and 50 percent as Asian and Pacific Islander students are never reported as making up more than 50 percent of a school's student body. Higher concentrations of Asian and Pacific Islander students are marked in dark purple points while lower concentrations are pink. The dark purple points identifying schools with higher percentages of Asian and Pacific Islander students occur in schools with lower percentages of economically disadvantaged students and higher SAT mean scores. It is also interesting to note that all of the purple points are above the blue line that signifies the average SAT mean score for different values of percentage of students economically disadvantaged. This may mean that Asian and Pacific Islander students perform at levels above the average given their economic situation.

The third graph shows the opposite relationship between African American students across different schools. The color gradient applied to this graph ranges between 0 and 100 percent of the student body being African American. Schools with higher concentrations of African American students are represented by red points, while schools with lower concentrations of African American students are represented with yellow points. The red dots tend to lie towards the bottom right of the graph, indicating higher percentages of the student body that is economically disadvantaged and lower SAT mean scores. Unlike the purple points representing higher concentrations of Asian and Pacific Islander students, the red points representing higher concentrations of African American students tend to lie below the average SAT mean score for each level of economic disadvantage. This could indicate that African American students tend to perform below the mean.

The last graph shows the relationship between Hispanic students across different schools, and the relationship is similar to that of African American students. The color gradient ranges from green points for concentrations of Hispanic students as low as 0 percent to purple points for concentrations of Hispanic students as high as 100 percent. The graph shows that schools with higher concentrations of Hispanic students tend to lie on the higher end of percentage of student body that is economically disadvantaged as well as lower in SAT mean scores. However, the points are spread equally above and below the average SAT mean score. This could indicate that higher percentages of Hispanic students are more closely correlated with higher percentages of economically disadvantaged students, and not as closely to SAT mean scores.

These graphs draw home the point that percentage of a student body that is economic disadvantage is negatively correlated to SAT mean scores. While adding color gradients to the points helps convey the relationship of race to both the levels of economic disadvantage and SAT mean scores, the analysis could have some gaps. The primary issue is that races are reported as percentages of student bodies. The data set provided does not include data on the standardized test scores or economic disadvantage of specific races. Since races are presented as percentages of student bodies, it is unknown how the different percentages mix together. For example, a school could have a student body that is 50 percent White, 25 percent Hispanic, and 25 percent African American or a student body that is 50 percent white and 50 percent Hispanic. The way the points are plotted, the viewer cannot easily differentiate the former and latter of these racial distributions.

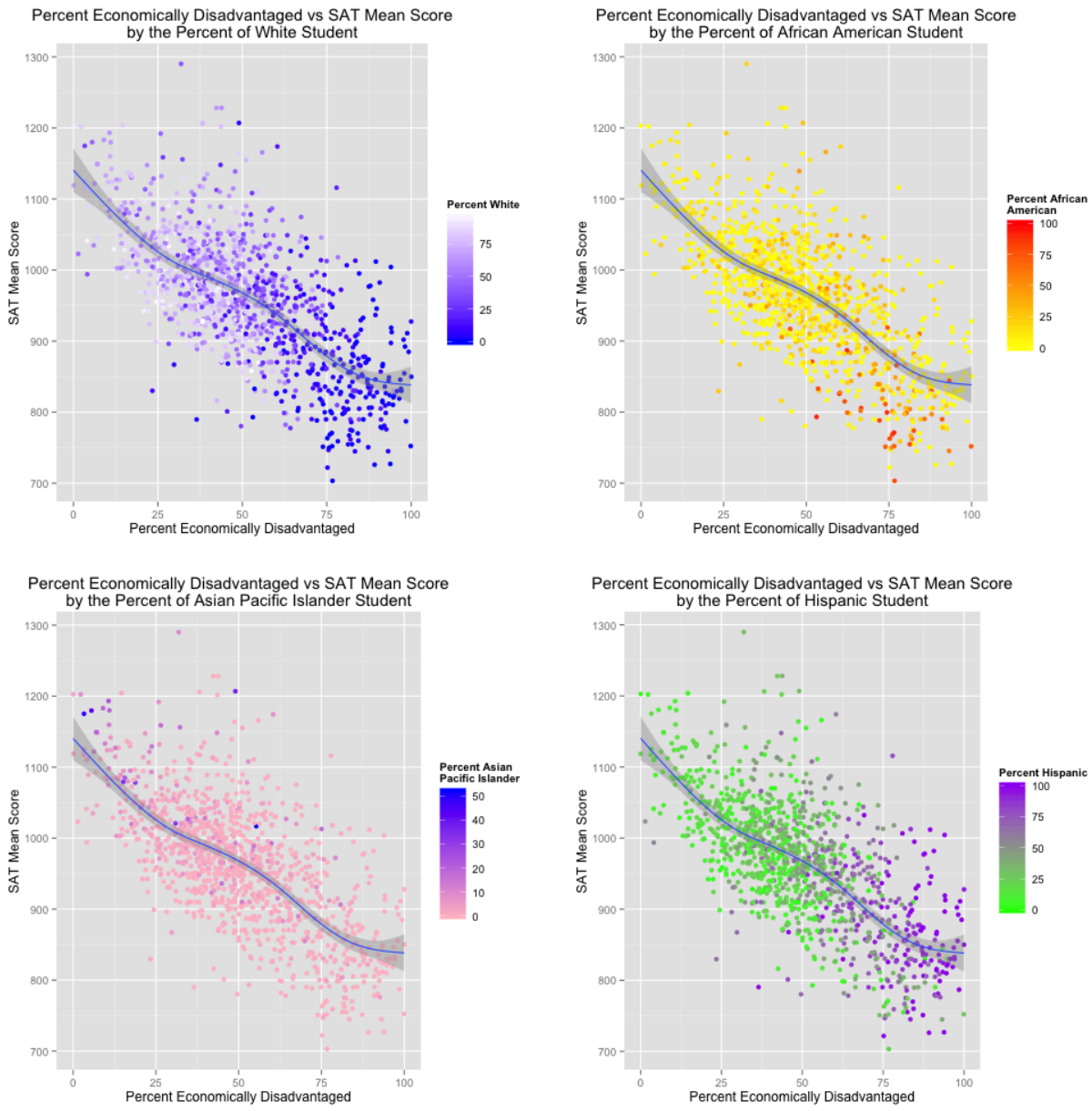

Figure 6: (Top left) Scatterplot, with a smoothed line to show average SAT mean score, showing the relationship between percentage of a school's student body that is economically disadvantaged and the school's SAT mean score along with the spread of different concentrations of White students in schools. Lighter points represent higher concentrations of White students. (Bottom left) Scatterplot, with a smoothed line to show average SAT mean score, showing the relationship between percentage of a school's student body that is economically disadvantaged and the school's SAT mean score along with the spread of different concentrations of Asian and Pacific Islander students in schools. Darker, purple points represent higher concentrations of Asian and Pacific Islander students, while pink points represent lower concentrations. Note that the gradient only goes up to 50 percent and that the purple points lie primarily above the smoothed curve. (Top right) Scatterplot, with a smoothed line to show average SAT mean score, showing the relationship between percentage of a school's student body that is economically disadvantaged and the school's SAT mean score along with the spread of different concentrations of African American students in schools. Red points represent higher concentrations of African American students and yellow points represent lower concentrations. Notice how the red points lie primarily below the smoothed curve. (Bottom right) Scatterplot, with a smoothed line to show average SAT mean score, showing the relationship between percentage of a school's student body that is economically disadvantaged and the school's SAT mean score along with the spread of different concentrations of Hispanic students in schools. Purple points represent higher concentrations of Hispanic students and green points represent lower concentrations.

The effect that different distributions of races have on both economic disadvantage and SAT mean scores is difficult to determine unless someone closely examines the location of each point, the concentration of the different races its colors refer to, and then develop an understanding of the mixture.

Another possible gap and area to further investigate is whether the concentration of Hispanic students is actually indicative of SAT mean scores. Since the points colored for similar levels are spread evenly above and below the average SAT mean scores, it seems that race is more closely correlated to economic disadvantage than SAT mean scores. There could be an entirely different factor that correlates Hispanic students and their SAT mean scores. As mentioned earlier, without race-specific data, these relationships are difficult to determine with only proportions of student bodies that are of a certain race.

Another problem that arises from using proportional data is that it disallows a realistic and holistic understanding of how many students represent certain races and how they are related to economic disadvantage and SAT mean scores. The schools more concentrated with African American or Hispanic students could be small schools located in neighborhoods concentrated with those races. These schools could also be very large inner city schools. Incorporating the size of the student population, especially specific numbers of students of different races, into this analysis could allow a better understanding of how the entire Texas public school student body is broken down. These sections can then be analyzed further to extract trends related to race and other measurements.

However, this analysis provides an overview of the demographics of Texas public education. The analysis shows that Hispanic and African American students tend to be more concentrated in more economically disadvantaged schools and have lower reported SAT mean scores. On the other hand, White and Asian or Pacific Islander students are more concentrated in less economically disadvantaged schools and tend to have higher SAT mean scores. These relationships provide interesting insight into how demographics of a school's student body can potentially affect its rankings.

7 Conclusion

The increase in available data in the education sector is only as useful as the insights we can further derive from this data. In order to create the types of data-driven educational reform desired by our politicians and educational leaders, a deep understanding of the implications, abilities, and limitations of this data is necessary. Our findings aimed to explore this data with regards to Texas public high schools in an effort to better this understanding. We believe that we have shown throughout our findings that a dependency on one metric, such as SAT scores or attendance, does paint the entire educational picture at a given school. For example, we have seen how factors such as geographical location and the size of one's school district might effect the educational quality of a high school, which is not depicted solely through an analysis of SAT scores. The implication, then, for educators is to continue to collect data and attempt to improve our schools with a holistic approach rather than specific focus on a single metric.

Our first finding reiterates this point; we found that the change in any single metric does not fully explain the change in rank. Changes in rank were found to occur primarily in middle-tier schools, and were almost entirely non-existent at Texas's best and worst high schools. Change in graduation rate was found to be the best explanatory metric for the change in rank but was still far away from providing a satisfying, full account of school improvement. A more longitudinal study of school improvement could be used here to get a better sense of changes in school quality over time. An interesting question for further research could delve into the relationship between the popularity of certain standardized exams and the average scores that schools received on these exams.

We then analyzed the educational cost efficiency of each school district in Texas. Bigger counties, especially those containing Texas's largest cities, were found to be less cost efficient than smaller counties, with a massive dependency between Travis county and the rest of Texas. We also recognized a geographical trend in the data, indicating that counties in East and South Texas tended to be less cost efficient. It could

also be interesting to look at cost efficiency on a national scale, comparing the cost efficiency of counties in Texas with similar counties across the country in regard to standardized exam scores. Within Texas, further exploration of school demographics in each county could lead to possible explanations for the geographical discrepancies.

Our next finding then compared the results of standardized exams across Texas public high schools. Looking at both the SAT and ACT, we found evidence that a school's ranking in the social studies segment of the Texas Assessment of Knowledge and Skills is the best predictor of that school's SAT and ACT mean scores. We also found that average scores on all TAKS exams relate nearly identically with SAT and ACT scores, providing evidence that the SAT and ACT are indeed similar.

In our last finding, we examined the well-documented phenomena of correlation between a school's concentration of different races, percentage of economically disadvantaged students, and SAT scores. Our findings supported the research status-quo; we showed that SAT mean scores are negatively correlated with the percentage of a school's student body that is economically disadvantaged. With regards to race, it was found that schools with high proportions of African Americans and Hispanics correlated with lower SAT scores and higher percentages of economically disadvantaged students. Similarly, schools with high proportions of White, Asian, and Pacific Islander students were have higher overall SAT scores and less economically disadvantaged students. This finding, along with our other explorations, show that Texas is not immune to the socio-economic issues documented in America's public high schools.

A Code

```
library(ggplot2)
library(plyr)
library(stringr)
library(sp)
library(maps)
library(gmm)
library(reshape2)
library(xtable)

###
### Data cleaning
###

# Load the origin school rankings data set
schools <- read.csv("school_rankings_2011.csv", stringsAsFactors = F)

# Create subset for high school
schools <- subset(schools, School_Level == "High")

# Clean Average_Expenditure_Per_Pupil  [-999 is NA]
schools$Average_Expenditure_Per_Pupil[schools$Average_Expenditure_Per_Pupil
  == "-$999"] <- NA

# Convert character column to numeric without $
a <- str_extract(schools$Average_Expenditure_Per_Pupil, "[0-9]+,[0-9]+")
b <- str_replace(a, ",", "")
schools$Average_Expenditure_Per_Pupil <- as.numeric(b)

# Assign NA to ce when its Texas_Rank_Percentile equals to 0
for(i in seq_along(schools$School_Level)) {
  if(schools[i, ]$Texas_Rank_Percentile == 0){schools[i, ]$ce <- NA}
}

# Write the clean data set to a new csv file
write.csv(schools, "schools_clean.csv")

###
### Finding 1: Explaining Changes in Rank from 2010
###

# Load data from 2010 report
schools_2010 <- read.csv("school_rankings_2010.csv", stringsAsFactors = F)

# Create subset for high school
schools_2010 <- subset(schools_2010, School.Level == "High School")
```

```
# Rename Texas Rank for 2010 & Campus Number for joining the two data sets
schools_2010$Texas_Rank_2010 <- schools_2010$Texas.Rank
schools_2010$Campus_Number <- schools_2010$Campus.Number
schools_2010$Texas.Rank <- NULL
schools_2010$Campus.Number <- NULL

# Label our 2010 variables as such for distinction when joining
schools_2010$Campus_Name_2010 <-
  schools_2010$Campus.Name # Ensure correct matchup
schools_2010$Campus.Name <- NULL
schools_2010$School_Type_2010 <-
  schools_2010$School.Type
schools_2010$School.Type <- NULL
schools_2010$Grade_Span_2010 <-
  schools_2010$Grade.Span
schools_2010$Grade.Span <- NULL
schools_2010$Charter_2010 <-
  schools_2010$Charter
schools_2010$Charter <- NULL
schools_2010$Student_Teacher_Ratio_2010 <-
  schools_2010$Student.Teacher.Ratio
schools_2010$Student.Teacher.Ratio <- NULL
schools_2010$Average_Expenditure_Per_Pupil_2010 <-
  schools_2010$Average.Expenditure.Per.Pupil
schools_2010$Average.Expenditure.Per.Pupil <- NULL
schools_2010$TEA_Accountability_Rating_2010 <-
  schools_2010$TEA.Accountability.Rating
schools_2010$TEA.Accountability.Rating <- NULL
schools_2010$Total_Student_Count_2010 <-
  schools_2010$Total.Student.Count
schools_2010$Total.Student.Count <- NULL
schools_2010$Percent_Economically_Disadvantaged_2010 <-
  schools_2010$Percent.Economically.Disadvantaged
schools_2010$Percent.Economically.Disadvantaged <- NULL
schools_2010$Percent_African_American_2010 <-
  schools_2010$Percent.African.American
schools_2010$Percent.African.American <- NULL
schools_2010$Percent_Asian_Pacific_Islander_2010 <-
  schools_2010$Percent.Asian.Pacific.Islander
schools_2010$Percent.Asian.Pacific.Islander <- NULL
schools_2010$Percent_Hispanic_2010 <-
  schools_2010$Percent.Hispanic
schools_2010$Percent.Hispanic <- NULL
schools_2010$Percent_White_2010 <-
  schools_2010$Percent.White
schools_2010$Percent.White <- NULL
schools_2010$Percent_Limited_English_Proficient_2010 <-
  schools_2010$Percent.Limited.English.Proficient
```

```
schools_2010$Percent.Limited.English.Proficient <- NULL
schools_2010$TAKS_English_Language_Arts_2010 <-
  schools_2010$TAKS.English.Language.Arts
schools_2010$TAKS.English.Language.Arts <- NULL
schools_2010$TAKS_Mathematics_2010 <-
  schools_2010$TAKS.Mathematics
schools_2010$TAKS.Mathematics <- NULL
schools_2010$TAKS_Writing_2010 <-
  schools_2010$TAKS.Writing
schools_2010$TAKS.Writing <- NULL
schools_2010$TAKS_Reading_2010 <-
  schools_2010$TAKS.Reading
schools_2010$TAKS.Reading <- NULL
schools_2010$TAKS_Science_2010 <-
  schools_2010$TAKS.Science
schools_2010$TAKS.Science <- NULL
schools_2010$TAKS_All_Tests_2010 <-
  schools_2010$TAKS.All.Tests
schools_2010$TAKS.All.Tests <- NULL
schools_2010$TAKS_Social_Studies_2010 <-
  schools_2010$TAKS.Social.Studies
schools_2010$TAKS.Social.Studies <- NULL
schools_2010$Attendance_2010 <-
  schools_2010$Attendance
schools_2010$Attendance <- NULL
schools_2010$Advanced_Courses_2010 <-
  schools_2010$Advanced.Courses
schools_2010$Advanced.Courses <- NULL
schools_2010$AP_IB_Test_Takers_2010 <-
  schools_2010$AP.IB.Test.Takers
schools_2010$AP.IB.Test.Takers <- NULL
schools_2010$AP_IB_Passing_Rate_2010 <-
  schools_2010$AP.IB.Passing.Rate
schools_2010$AP.IB.Passing.Rate <- NULL
schools_2010$SAT_ACT_Test_Takers_2010 <-
  schools_2010$SAT.ACT.Test.Takers
schools_2010$SAT.ACT.Test.Takers <- NULL
schools_2010$SAT_Mean_Score_2010 <-
  schools_2010$SAT.Mean.Score
schools_2010$SAT.Mean.Score <- NULL
schools_2010$ACT_Mean_Score_2010 <-
  schools_2010$ACT.Mean.Score
schools_2010$ACT.Mean.Score <- NULL
schools_2010$Graduation_Rate_2010 <-
  schools_2010$Graduation.Rate
schools_2010$Graduation.Rate <- NULL
schools_2010$Recommended_High_School_Program_2010 <-
  schools_2010$Recommended.High.School.Program
```

```
schools_2010$Recommended.High.School.Program <- NULL
schools_2010$Retention_Grade_7_2010 <-
  schools_2010$Retention.Grade.7
schools_2010$Retention.Grade.7 <- NULL
schools_2010$Retention_Grade_8_2010 <-
  schools_2010$Retention.Grade.8
schools_2010$Retention.Grade.8 <- NULL
schools_2010$District_Rank_2010 <-
  schools_2010$District.Rank
schools_2010$District.Rank <- NULL
schools_2010$State_Tier_2010 <-
  schools_2010$State.Tier
schools_2010$State.Tier <- NULL
schools_2010$Grade_Type_2010 <-
  schools_2010$Grade.Type
schools_2010$Grade.Type <- NULL
schools_2010$Percent_Native_American_2010 <-
  schools_2010$Percent.Native.American
schools_2010$Percent.Native.American <- NULL
schools_2010$Texas_Rating_Description_2010 <-
  schools_2010$TEA_Rating_Description
schools_2010$TEA_Rating_Description <- NULL

# Some of these variables we know won't change, so we can remove them
schools_2010$District <- NULL
schools_2010$County <- NULL
schools_2010$Street.Address <- NULL
schools_2010$City <- NULL
schools_2010$State <- NULL
schools_2010$Zip <- NULL
schools_2010$Phone.Number <- NULL
schools_2010$Website <- NULL
schools_2010$School.Type <- NULL
schools_2010$School.Level <- NULL
schools_2010$Retention.Grade.4 <- NULL
schools_2010$Retention.Grade.5 <- NULL
schools_2010$Class.Size.Grade.1 <- NULL
schools_2010$Class.Size.Grade.2 <- NULL
schools_2010$Class.Size.Grade.3 <- NULL

# Different number of schools in 2010 vs. 2011!
length(schools$Campus_Number) - length(schools_2010$Campus_Number)
# Difference = 173

# Join the data sets based on Campus Number
schools_join <- join(schools_2010, schools, type = "inner")

# Compare rankings from 2010 to 2011 (Texas_Rank vs. Texas_Rank_2010)
```

```
ggplot(aes(Texas_Rank_2010, Texas_Rank), data = schools_join ,
  xlab = "Texas Rank in 2010", ylab = "Texas Rank in 2011") +
  geom_point() +
  ggtitle("Change in Texas School Rank from 2010 to 2011")
ggsave("Texas_Rank_2010_2011.pdf")
# Most of the changes in rank happen in the middle, RARELY AT THE EXTREMES

# Look at the difference in Rank (positive means higher in 2011)
schools_join$rank_difference <-
  schools_join$Texas_Rank_2010 - schools_join$Texas_Rank

# Re-arrange so that it is ordered with highest rank difference first
schools_join <- arrange(schools_join, -rank_difference)

# Look at the difference in most signficant metrics
schools_join$graduation_difference <-
  schools_join$Graduation_Rate - schools_join$Graduation_Rate_2010/100
schools_join$TAKS_Science_difference <-
  schools_join$TAKS_Science - schools_join$TAKS_Science_2010
schools_join$TAKS_Math_difference <-
  schools_join$TAKS_Math - schools_join$TAKS_Mathematics_2010
schools_join$TAKS_Social_Studies_difference <-
  schools_join$TAKS_Social_Studies - schools_join$TAKS_Social_Studies_2010
schools_join$TAKS_Reading_difference <-
  schools_join$TAKS_Reading - schools_join$TAKS_English_Language_Arts_2010
schools_join$Teacher_Ratio_difference <-
  schools_join$Student_Teacher_Ratio_2010 - schools_join$Student_Teacher_Ratio
schools_join$expenditure_difference <-
  schools_join$Average_Expenditure_Per_Pupil - schools_join$Average_Expenditure_Per_Pupil_2010
schools_join$advanced_courses_difference <-
  (schools_join$Advanced_Courses*100) - schools_join$Advanced_Courses_2010
schools_join$attendance_difference <-
  schools_join$Attendance - schools_join$Attendance_2010

# Find adjusted R-squared
ar2.1 <- summary(lm(rank_difference ~
  graduation_difference, data = schools_join))$adj.r.squared
ar2.2 <- summary(lm(rank_difference ~
  SAT_Mean_difference, data = schools_join))$adj.r.squared
ar2.3 <- summary(lm(rank_difference ~
  ACT_Mean_difference, data = schools_join))$adj.r.squared
ar2.4 <- summary(lm(rank_difference ~
  TAKS_Science_difference, data = schools_join))$adj.r.squared
ar2.5 <- summary(lm(rank_difference ~
  TAKS_Math_difference, data = schools_join))$adj.r.squared
ar2.6 <- summary(lm(rank_difference ~
  TAKS_Social_Studies_difference, data = schools_join))$adj.r.squared
ar2.7 <- summary(lm(rank_difference ~
```

```
  TAKS_Reading_difference, data = schools_join))$adj.r.squared
ar2.8 <- summary(lm(rank_difference ~
  econ_difference, data = schools_join))$adj.r.squared
ar2.9 <- summary(lm(rank_difference ~
  expenditure_difference, data = schools_join))$adj.r.squared
ar2.10 <- summary(lm(rank_difference ~
  advanced_courses_difference, data = schools_join))$adj.r.squared
ar2.11 <- summary(lm(rank_difference ~
  attendance_difference, data = schools_join))$adj.r.squared

# Combined model
summary(lm(rank_difference ~ ., data = schools_join))

# Create dataframe holding adjusted R-squared values
r <-
  data.frame(variable = c("Graduation", "SAT",
  "ACT", "TAKS Science", "TAKS Math",
  "TAKS Social Studies", "TAKS Reading",
  "Percent Economically Disadavtaged",
  "Attendance"),
  r2 = c(0.1904, 0.04279, 0.06067, 0.004887, 0.003969,
  0.002835, 0.005815, 0.02864, 0.06749))

# Plot the adjusted R-squared values in a bar chart
ggplot(data = r, aes(x = reorder(variable, -r2), y = r2)) + geom_bar() +
  ggtitle("Adjusted R-Squared Values Across Potential Explanatory Variables") +
  xlab("Change in Variable from 2010") + ylab("Adjusted R-Squared")
ggsave("r-sqaured_chart.pdf")

# Plot the Graduation Difference versus the Rank Difference
ggplot(data = schools_join, aes(x = graduation_difference, y = rank_difference)) +
  geom_point(alpha = I(1/3)) + geom_smooth(method = "lm", size = 1) +
  ggtitle("Explaining Change in Rank with Change in Graduation Rate") +
  xlab("Percent Change in Graduation Rate") + ylab("Change in Rank") +
  xlim(-.2,.2)
ggsave("graduation_vs_rank.pdf")

###
### Finding 2: Cost Efficiency Across Counties
###

# Calculate the cost efficiency of each school by using formule:
# Average Expenditure Per Pupil / Texas Rank Percentile
schools$ce <-
  schools$Average_Expenditure_Per_Pupil/schools$Texas_Rank_Percentile

# Calculate the average cost efficiency of each county
ce_ave <- ddply(schools[!is.na(schools$ce), ], "County", summarise,
```

```
                ce_ave = mean(ce))

# Lowercase the county value in ce_ave
ce_ave$County <- tolower(ce_ave$County)

# Load Texas borders data set
borders <- read.csv("tx-borders.csv", stringsAsFactors = FALSE)

# Change the variable name of data borders "county" to "County"
names(borders)[6] <- "County"

# Join the two data sets
tx_ce <- join(borders, ce_ave)

# Convert continuous values of cost efficiency to discrete values
tx_ce$ce_ai <- NA
x <- data.frame(quantile(tx_ce$ce_ave, probs= seq(0,1, 0.14), na.rm=T))
for(i in seq_along(tx_ce$ce_ave)) {
  if(tx_ce[i, ]$ce_ave < x[2, ]){tx_ce[i, ]$ce_ai <- "< g139.59412"}
  else{
    if(tx_ce[i, ]$ce_ave < x[3, ]){tx_ce[i, ]$ce_ai <- "< f185.79975"}
    else{
      if(tx_ce[i, ]$ce_ave < x[4, ]){tx_ce[i, ]$ce_ai <- "< e244.62007"}
      else{
        if(tx_ce[i, ]$ce_ave < x[5, ]){tx_ce[i, ]$ce_ai <- "< d360.80025"}
        else{
          if(tx_ce[i, ]$ce_ave < x[6, ]){tx_ce[i, ]$ce_ai <- "< c559.66221"}
          else{
            if(tx_ce[i, ]$ce_ave < x[7, ]){tx_ce[i, ]$ce_ai <- "< b936.24064"}
            else{
              if(is.na(tx_ce[i, ]$ce_ave) == T){tx_ce[i, ]$ce_ai <- "< a97205.00"}
              else{
                tx_ce[i, ]$ce_ai <- "< a97205.00"
              }
            }
          }
        }
      }
    }
  }
}

# Save the csv file
write.csv(tx_ce,"tx_ce.csv")

# Create funtion to have the middle value of the map
mid_range <- function(x) mean(range(x, na.rm = TRUE))
centers <- ddply(tx_ce, "County", summarise,
```

```
                long = mid_range(long),
                lat = mid_range(lat))

# State the best five counties in cost efficiency rank
centers$best_rank <- NA
for(i in seq_along(centers$County)) {
  if(centers[i, ]$County == "franklin"){centers[i, ]$best_rank <- 1}
  else{
    if(centers[i, ]$County == "erath"){centers[i, ]$best_rank <- 2}
    else{
      if(centers[i, ]$County == "fayette"){centers[i, ]$best_rank <- 3}
      else{
        if(centers[i, ]$County == "mills"){centers[i, ]$best_rank <- 4}
        else{
          if(centers[i, ]$County == "gillespie"){centers[i, ]$best_rank <- 5}
          else{
            centers[i, ]$best_rank <- NA
          }
        }
      }
    }
  }
}

# State the worst five counties in cost efficiency rank
centers$worst_rank <- NA
for(i in seq_along(centers$County)) {
  if(centers[i, ]$County == "travis"){centers[i, ]$worst_rank <- 1}
  else{
    if(centers[i, ]$County == "galveston"){centers[i, ]$worst_rank <- 2}
    else{
      if(centers[i, ]$County == "dallas"){centers[i, ]$worst_rank <- 3}
      else{
        if(centers[i, ]$County == "duval"){centers[i, ]$worst_rank <- 4}
        else{
          if(centers[i, ]$County == "harris"){centers[i, ]$worst_rank <- 5}
          else{
            centers[i, ]$worst_rank <- NA
          }
        }
      }
    }
  }
}

# Draw school cost efficiency map over Texas
# Also point out five best and worst counties
ggplot(data = tx_ce, aes(long, lat)) +
```

```
  geom_polygon(aes(group = group, fill = ce_ai)) +
  scale_fill_brewer("Cost Efficiency Percentile",
  labels = c("< g139.59412" = "86 - 100%" , "< f185.79975" = "72 - 86%",
  "< e244.62007" = "58 - 72%", "< d360.80025" = "44 - 58%",
  "< c559.66221" = "30 - 44%", "< b936.24064" = "16 - 30%",
  "< a97205.00" = "0 - 16%"), palette = "Blues") +
  scale_x_continuous("") +
  scale_y_continuous("") +
  labs(xlab = "Longitude", ylab = "Latitude",
  title = "High School Cost Efficiency Distribution acrossed Counties in Texas") +
  coord_map() +
  geom_text(aes(label = best_rank), data = centers, col = I("green")) +
  geom_text(aes(label = worst_rank), data = centers, col = I("red"))
# Save the plot
ggsave("cost_efficiency.pdf")

###
### Finding 3: TAKS as an Indicator of SAT/ACT Scores
###

# Relationship between TAKS Subjects Percentile and SAT Mean Score
ggplot(aes(TAKS_Math, SAT_Mean_Score), data = schools) +
  geom_smooth(aes(TAKS_Math, SAT_Mean_Score), col = "red", method = "lm") +
  geom_smooth(aes(TAKS_Science, SAT_Mean_Score), col = "green", method = "lm") +
  geom_smooth(aes(TAKS_Reading, SAT_Mean_Score), col = "blue", method = "lm") +
  geom_smooth(aes(TAKS_Social_Studies, SAT_Mean_Score),
  col = "yellow", method = "lm") +
  labs(x = "TAKS Subject Percentile", y = "SAT Mean Score",
  title = "TAKS Subjects Percentile vs SAT Mean Score") +
  geom_text(label = "TAKS Math", x = 1.0, y = 1000,
  data= NULL, col = "red", angle = 25) +
  geom_text(label = "TAKS Science", x = 1.0, y = 920,
  data= NULL, col = "green", angle = 25) +
  geom_text(label = "TAKS Reading", x = 1.0, y = 1060,
  data= NULL, col = "blue", angle = 25) +
  geom_text(label = "TAKS Social Studies", x = 0.75, y = 1110,
  data= NULL, col = "yellow", angle = 25) +
  coord_cartesian(xlim = c(0,1.1))
ggsave("taks_sat.pdf")

# Relationship between TAKS Subjects Percentile and ACT Mean Score
ggplot(aes(TAKS_Math, ACT_Mean_Score), data = schools) +
  geom_smooth(aes(TAKS_Math, ACT_Mean_Score), col = "red", method = "lm") +
  geom_smooth(aes(TAKS_Science, ACT_Mean_Score), col = "green", method = "lm") +
  geom_smooth(aes(TAKS_Reading, ACT_Mean_Score), col = "blue", method = "lm") +
  geom_smooth(aes(TAKS_Social_Studies, ACT_Mean_Score), col = "yellow",
  method = "lm") + labs(x = "TAKS Subject Percentile", y = "ACT Mean Score",
  title = "TAKS Subjects Percentile vs ACT Mean Score") +
```

```
  geom_text(label = "TAKS Math", x = 1.0, y = 21.25,
  data= NULL, col = "red", angle = 25) +
  geom_text(label = "TAKS Science", x = 1.0, y = 19,
  data= NULL, col = "green", angle = 25) +
  geom_text(label = "TAKS Reading", x = 1.0, y = 23,
  data= NULL, col = "blue", angle = 25) +
  geom_text(label = "TAKS Social Studies", x = 0.82, y = 24, data= NULL,
  col = "yellow", angle = 25) + coord_cartesian(xlim = c(0,1.1))
ggsave("taks_sat.pdf")

###
### Finding 4: SAT Scores by Race
###

# Create the function to put multiple plots in one graph
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {

    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {

      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
```

```
                                    layout.pos.col = matchidx$col))
    }
  }
}

# Colored race graphs
p1 <-
  qplot(Percent_Economically_Disadvantaged, SAT_Mean_Score, data = schools,
  xlab = "Percent Economically Disadvantaged", ylab = "SAT Mean Score", main =
  "Percent Economically Disadvantaged vs SAT Mean Score
    by the Percent of White Student",
  geom = c("jitter", "smooth"), color = Percent_White) +
  scale_colour_gradient(low = "blue", high = "white") +
  labs(color = "Percent White")
p2 <-
  qplot(Percent_Economically_Disadvantaged, SAT_Mean_Score, data = schools,
  xlab = "Percent Economically Disadvantaged", ylab = "SAT Mean Score", main =
  "Percent Economically Disadvantaged vs SAT Mean Score
     by the Percent of Asian Pacific Islander Student",
  geom = c("jitter", "smooth"), color = Percent_Asian.Pacific_Islander) +
  scale_colour_gradient(low = "pink", high = "blue") +
  labs(color = "Percent Asian\nPacific Islander")
p3 <-
  qplot(Percent_Economically_Disadvantaged, SAT_Mean_Score, data = schools,
  xlab = "Percent Economically Disadvantaged", ylab = "SAT Mean Score", main =
  "Percent Economically Disadvantaged vs SAT Mean Score
    by the Percent of African American Student",
  geom = c("jitter", "smooth"),
  color = Percent_African_American) +
  scale_colour_gradient(low = "yellow", high = "red") + labs(color = "Percent African\nAmerican")
p4 <-
  qplot(Percent_Economically_Disadvantaged, SAT_Mean_Score, data = schools,
  xlab = "Percent Economically Disadvantaged", ylab = "SAT Mean Score", main =
  "Percent Economically Disadvantaged vs SAT Mean Score
    by the Percent of Hispanic Student",
  geom = c("jitter", "smooth"),
  color = Percent_Hispanic) +
  scale_colour_gradient(low = "green", high = "purple") +
  labs(color = "Percent Hispanic")
multiplot(p1, p2, p3, p4, cols = 2)
ggsave("race_econ.pdf")
```