DS-UA 112: Big Data Analysis Project

Svanik Dani
Prof. Wallisch, August 19th, 2021

Abstract.
In this project, I analyzed the dataset ('middleSchoolData.csv'), which contains data from all 594 NYC middle schools, including 485 public schools and 109 charter schools from a randomly picked year in the past 5 years. Since each row of the dataset represents a particular school, so the unit of analysis was "school". The analysis was focused on whether characteristics of NYC middle schools predict admission to one of 8 highly selective public high schools (Stuyvesant, Bronx High School of Science, etc.) in New York (from now on called HSPHS). To deal with missing values, when they were present, those school's were removed from the analysis. This was done as there were always enough remaining schools to still perform an analysis. Furthermore, there seemed no good way to estimate the missing data reliably. In terms of data cleaning, it was done separately for each question, to ensure that we had the maximum amount of samples for each part of the analysis. When needed, data was transformed and scaled using z-scoring. This allowed me to scale data when needed so that it was normalized. Dimensionality reduction was done using principal component analysis (PCA), and the elbow criterion. Scree plots were used to visualize PCA results, and determine how many PCA components ought to be kept.

1 What is the correlation between the number of applications and admissions to HSPHS?

The first step was to take a look at the dataset to ensure no missing values. Both applications and acceptances had 594 observations, so no data cleaning was needed in this step. Then, I found the correlation between number of applications to HSPHS and the correlation was $r = 0.80173$. To visualize the correlation and the data, I created a scatter plot comparing number of applicants to HSPHS and number of students admitted to HSPHS (Fig 1). Each point on the graph represents a school.

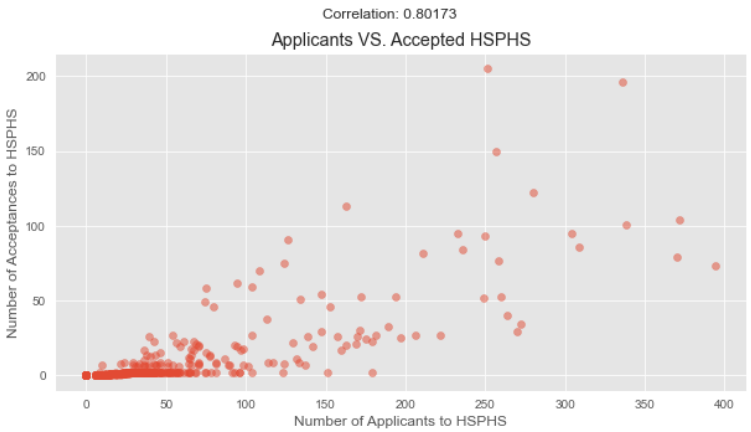

Fig 1 Shown Above

2 What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?

To determine the application rate, which was not given in the data set, I did $\frac{application}{schoolSize}$. This eliminated 2 schools, as there was no school size provided for them, and we cannot divide by zero. On the remaining 592 schools, I then looked at the correlation between number of applicants and number of admits, as well as application rate and the number of admits. The correlation between number of applicants and number of admits was $r = 0.8177$, and the correlation between application rate and number of admits was $r = 0.065875$. The correlation matrix heatmap can be used to visualize the correlation matrix (Fig 2).

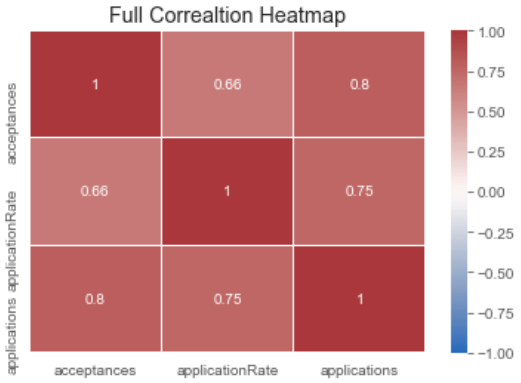

Fig 2 Shown Above

Intuitively it made sense that raw number of applicants was a better predictor. This was backed by the data. The smallest school had 33 students, and the top 10 school in terms of numbers of admits ranged from 205 to 93 admits to HSPHS. Even if this small school with 33 students had a 100% acceptance rate, they would not be in the top 10 schools in terms of number of admits to HSPHS. The scatter plots of number of applicants and application rate vs. number of admits to HSPHS helps to further visualize the correlations (Fig 3).

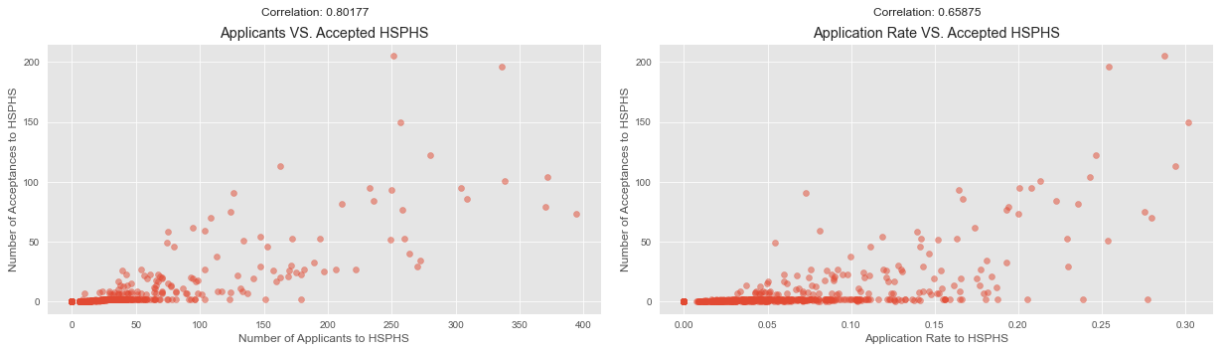

Fig 3 Shown Above

3 Which school has the best *per student* odds of sending someone to HSPHS?

To calculate the odds per student, I did $\frac{acceptance}{schoolSize0}$. As in Q2, since two schools were missing a school size and we cannot divide by 0, we removed these schools and conducted the analysis on the reaming 592 schools. The school with the best odds per student was "THE CHRISTA MCAULIFFE SCHOOL\I.S. 187" (DBN:20K187). The school had 205 students accepted and a school size of 873. This made the odds per student of acceptance to HSPHS at The Christa McAuliffe School 23.482%. A histogram was made to visualize the distribution of odds per student amongst schools (Fig 4). The histogram shows that most schools have low odds, and buckets above 15% odds, contained one school each making them difficult to see on the figure. Furthermore, The Christa McAuliffe School had 90% of students that exceeded both Reading and Math scores. This certainly played a role in them achieving the best odds per student.

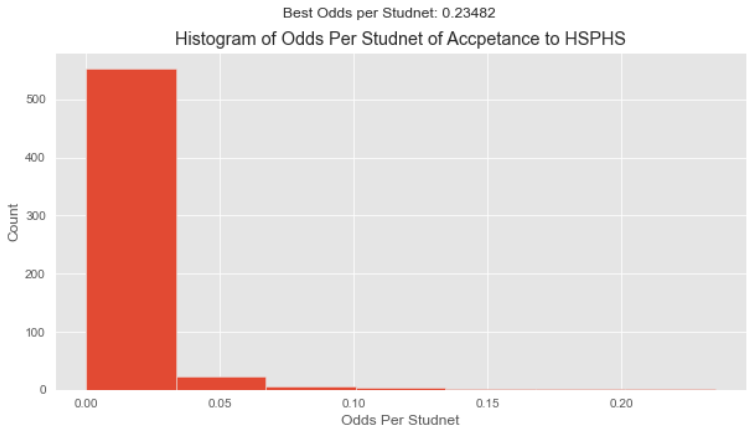

Fig 4 Shown Above

4 Is there a relationship between how students perceive their school (as reported in columns L-Q) and how the school performs on objective measures of achievement (as noted in columns V-X).

In order to analyze the relationship, we began by removing schools that had no data for any of these factors; rigorous instruction, collaborative teachers, supportive environment, effective school leadership, strong family community ties, trust, student achievement, reading scores exceed, math scores exceed. This left 526 schools for analysis. The factors fell into two groups, student perception (rigorous instruction, collaborative teachers, supportive environment, effective school leadership, strong family community ties, trust) and objective measures of achievement (student achievement, reading scores exceed, math scores exceed). I then calculated all the correlation between all the variables. I visualized this using a heatmap (Fig 5). The heatmap showed strong correlations

between factors in the students perceptions and between factors in objective measurements. For example, effective school leadership was strongly correlated ($r = 0.74$) with collaborative teachers. Furthermore, reading scores exceeded and math scores exceeded was a correlation of $r = 0.98$. I then looked at the correlations only between the two groups of variables (students perception and objective measures of achievement) (Fig 5). These correlations were not nearly as strong, but some factors were moderately correlated such as a supportive environment and readings scores with $r = 0.44$.

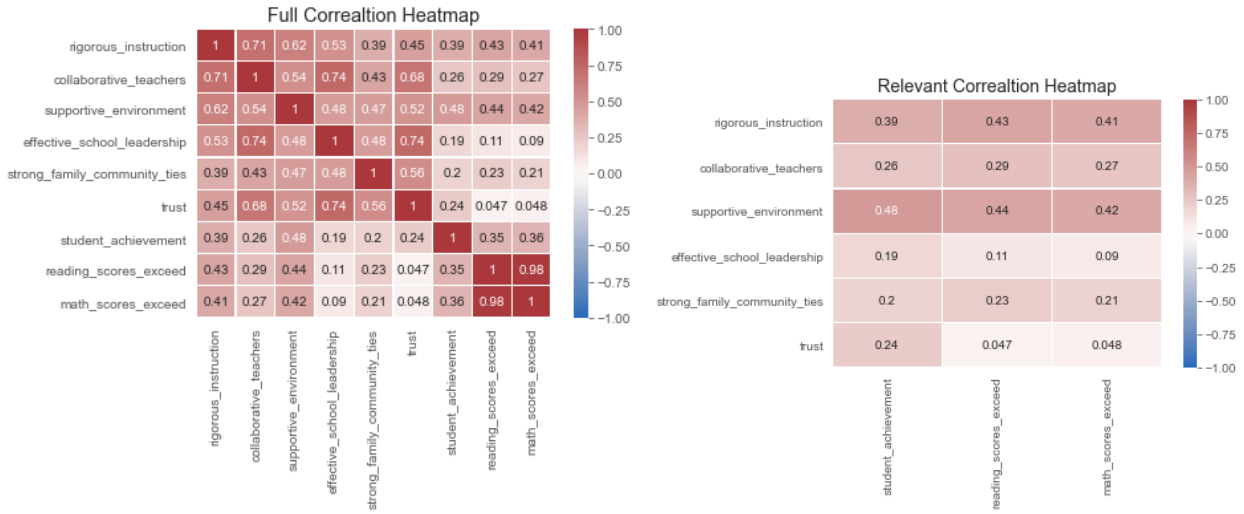

Fig 5 Shown Above

After finding the correlations, I performed multi-regression to test the predictive power of the student perspective factors. The first 3 models took all the student perspective factors as inputs, and the outputs were student achievement, reading scores exceeded, and math scores exceeded. The regression analysis showed the r^2 of the models: student achievement $r^2 = 0.2602$, reading scores exceeded $r^2 = 0.32903$, math scores exceeded $r^2 = 0.29548$. These models did not explain even a third of the variation of the outputs, their performance was rather poor. After looking at the relevant correlation matrix, 3 more multi-regression models were made using only rigorous intuition, collaborative teachers, and supportive environment as inputs. The regression analysis showed the r^2 of the simplified models: student achievement $r^2 = 0.2538$, reading scores exceeded $r^2 = 0.2370$, math scores exceeded $r^2 = 0.2162$. These simpler models performed just as well as the ones that considered all factors, telling us some of the factors are unnecessary.

In order to reduce the dimensionality of the data, PCA was conducted. The PCA on the student perception factors was than visualised (Fig 6). Using the elbow criterion, 3 components were kept. This was because although there was an elbow at 2, both PC2 and PC3 explained approximately the same amount of variance. The 3 components kept explained a total of 86.27% of the variance.

PCA was then also conducted on the objective measures of achievement (Fig 7). This allowed the output space to go from 3 factors to 1. Furthermore, another PCA was done on just reading and math scores (Fig 7). For the PCA on all 3 objective measures of achievement, 1 component was kept, which explained 72.94% of the variance. 1 component was also kept for the PCA on reading and math scores, it explained 98.97% of the variance.

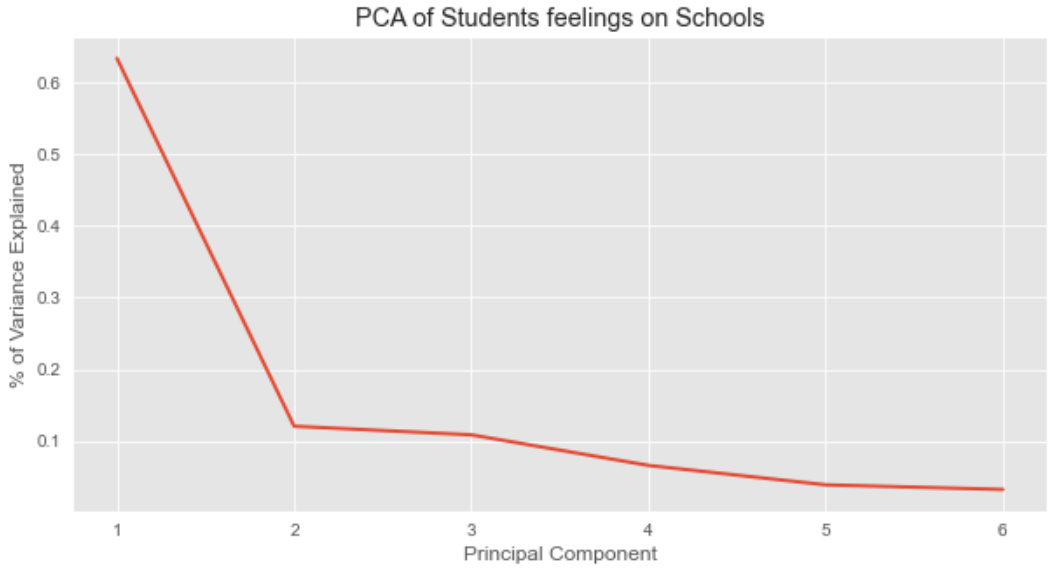

Fig 6 Shown Above

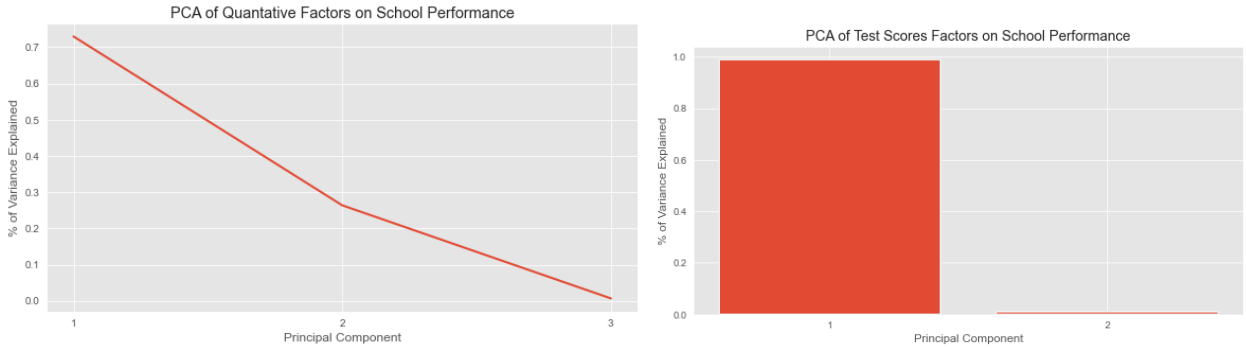

Fig 7 Shown Above

5 multi-regression models were than made. The inputs were the 3 PCA components from the students perception. The 5 outputs were student achievement, reading scores, math scores, the 1 PCA component of the entire objective achievement space, and the 1 PCA component of just math and reading scores. The models r^2 values were as follows: student achievement $r^2 = 0.2364$, reading scores $r^2 = 0.3033$, math scores $r^2 = 0.2781$, entire space $r^2 = 0.35.91$, test scores $r^2 =$

0.2934. These models were just as good was if not better than those that included all the factors. This showed that the PCA retained the information while reducing the dimensions. Furthermore, the model whose output was the PCA component of the objective achievement factors, was the best model. Nonetheless, none of these models even got close to explain 50% of the variance, and were all relatively poor. Hence, we can conclude although there is a relationship between students perception and objective measures of achievement, it is a weak one.

5 Test a hypothesis of your choice as to which kind of school (e.g. small schools vs. large schools or charter schools vs. not (or any other classification, such as rich vs. poor school)) performs differently than another kind either on some dependent measure, e.g. objective measures of achievement or admission to HSPHS (pick one).

The hypothesis I choose to test was the effect of percentage of students living in households below the poverty line and student achievement. The null hypothesis was that living under the poverty line has no effect on a student's achievement. Alternative hypothesis 1 was that living under the poverty line has a negative impact on student's achievement. Alternative hypothesis 2 was that living under the poverty line has a positive impact on student's achievement. I began by removing schools that did not report poverty percent or student achievement, this left 545 schools for analysis. I then converted poverty percent to a categorical variable by splitting school's as those above and below the median of students living in households below the poverty line. Next, I visualized the two distributions and their mean's (Fig 8).

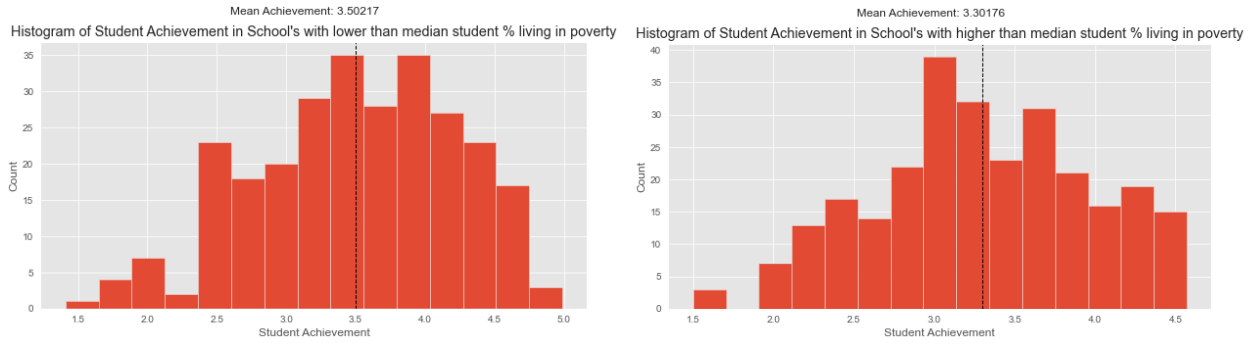

Fig 8 Shown Above

Since these distributions do not look normal, I used a Mann-Whitney U test instead of a T-Test (a T-Test assumes a normal distribution). The Mann-Whitney U test gave me a p-value of $p < 0.001$ and was therefore highly significant. This allowed me to reject the null hypothesis. Since schools with less student living under the poverty line had a higher average achievement, I was able to reject alternative hypothesis 2. Therefore, I accepted alternative hypothesis 1: living under the poverty line has a negative impact on student's achievement

6 Is there any evidence that the availability of material resources (e.g. per student spending or class size) impacts objective measures of achievement or admission to HSPHS?

For evidence of material resource I used: per pupil spending, average class size, and poverty percent. The objective measures of achievement were student achievement, reading scores exceed, and math scores exceed. Schools without these factors were not included in the analysis, this left 458 schools. I then visualized the correlation matrix (Fig 9). Poverty percent was strongly correlated with test scores, and average class size as well as per pupil spending were modernly correlated with test scores. There were no definitive correlations with student achievement.

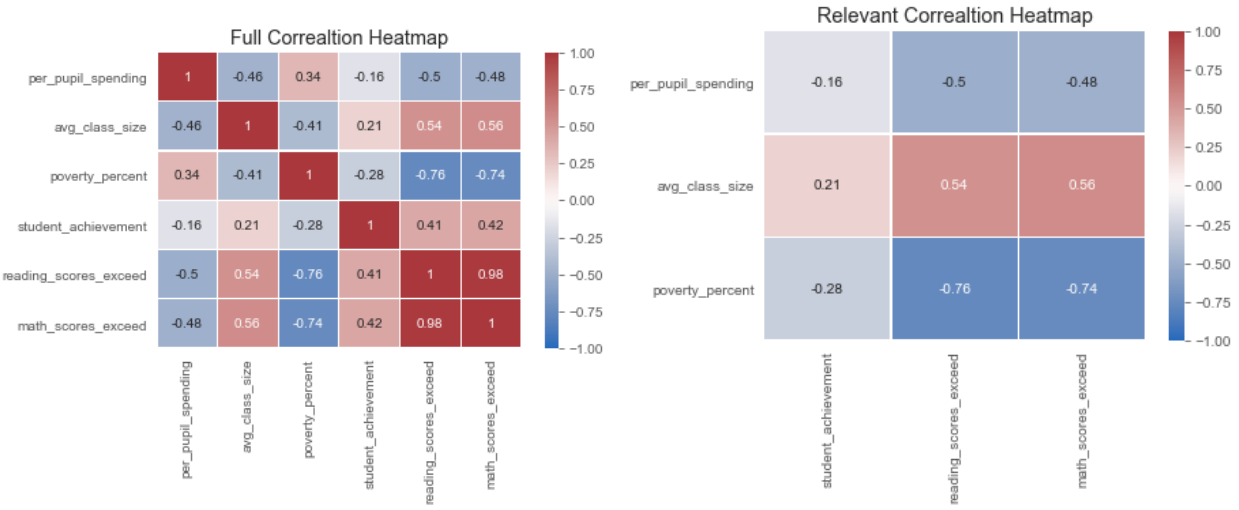

Fig 9 Shown Above

I then made three multi-regression models. They took all 3 measures of material resources as inputs. The 3 outputs were student achievement, reading scores, and math scores. The models r^2 were as follows: student achievement $r^2 = 0.089$, reading scores $r^2 = 0.6714$, math scores $r^2 = 0.6511$. As expected these factors were not great at predicting student achievement, however they were quite good at predicting test scores as compared to our previous models. Hence, we can say there is a strong relationship between material resources and test scores. This backs up previous studies that showed the SAT score is only correlated with household income, as that allows access to tutors and prep resources.

7 What proportion of schools accounts for 90% of all students accepted to HSPHS?

To find what proportion of schools account for 90% of all student accepted to HSPHS, I began by finding the total number of students admitted to HSPHS across all schools. The total number of admits was 4,461. 90% of this was 4,014.9 which I rounded to 4,015 as you cannot have a partial admit. 123 schools were required to cross this threshold and they sent a total of 4016 students to

HSPHS. This means that 20.707% of schools accounted for 90% of the admits. I then visualized the number of students accepted to HSPHS from each school as a bar graph (Fig 10). X ticks were removed for ease of viewing, as were schools with 0 admits. The graph seemed to follow a logistic decay.

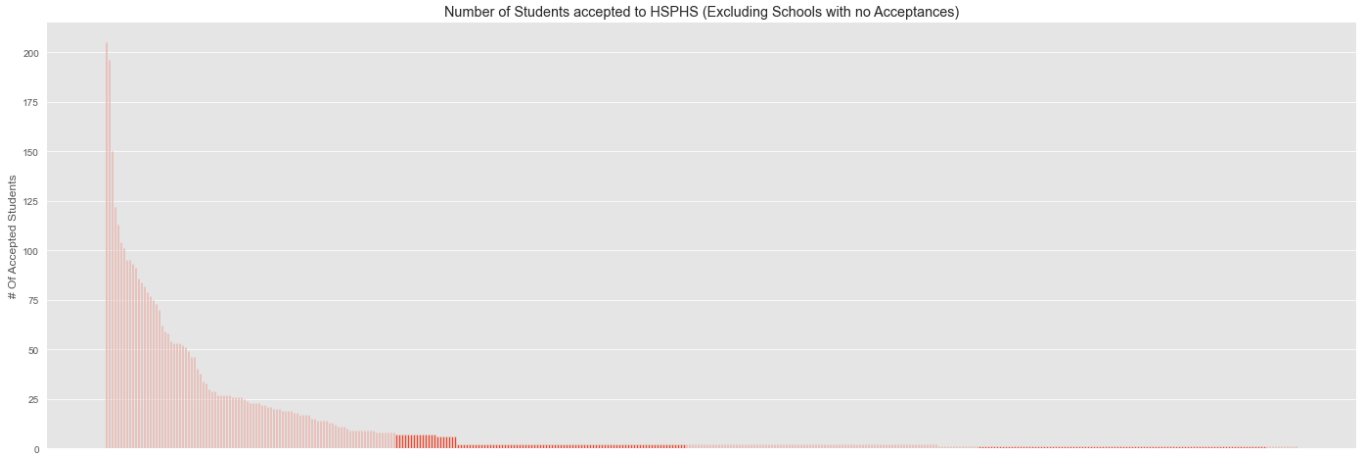

Fig 10 Shown Above

8 Build a model of your choice – clustering, classification or prediction – that includes all factors – as to what school characteristics are most important in terms of a) sending students to HSPHS, b) achieving high scores on objective measures of achievement?

I decided to build a prediction model using multi-regression. I wanted to predict test scores exceeded, as these were objective measures of achievement. I filtered out all schools with any missing data, this left 449 schools for analysis. I began with a full correlation heat map (Fig 11) , and a relevant correlation heat map (Fig 12). These correlation helped me pick out two extremely important variables for predicting test scores: poverty percent and disability percent. I made 2 multi-regression models using these two factors as inputs, and the outputs were reading scores and math scores. The models performance was as follows: reading scores $r^2 = 0.6855$ and math scores $r^2 = 0.6674$. These models were fairly good but I felt using more variables would likely yield better results.

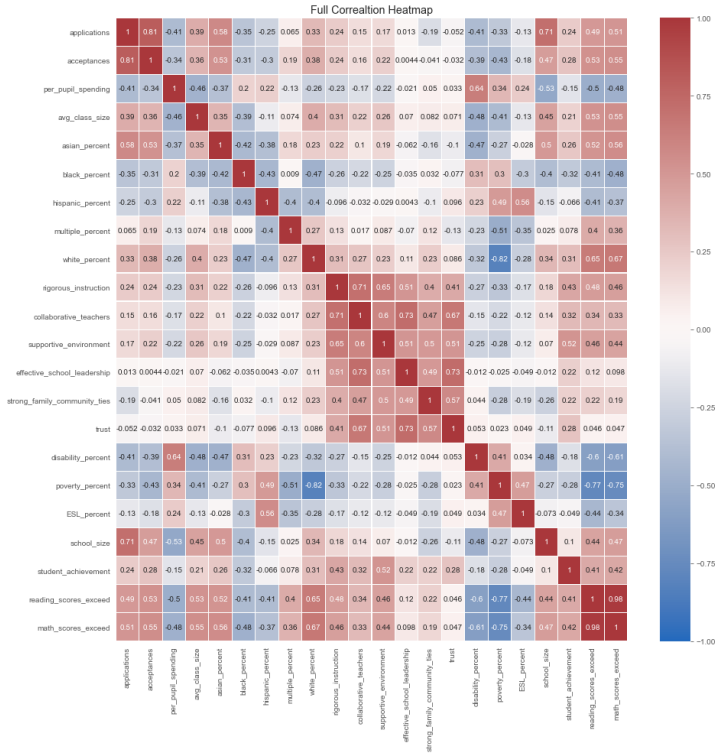

Fig 11 Shown Above

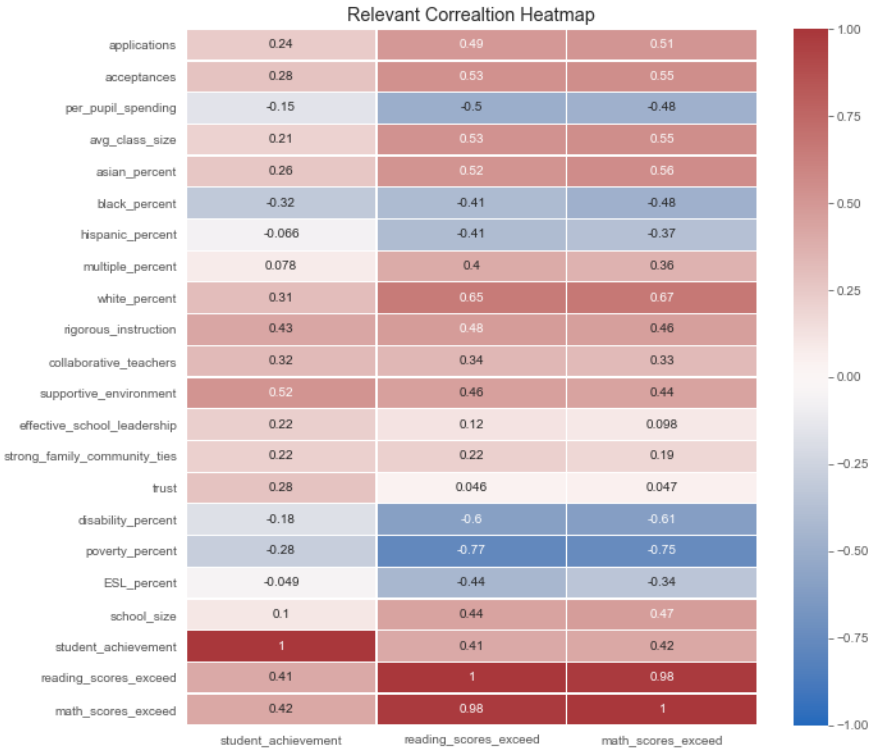

Fig 12 Shown Above

Hence, I performed a PCA on these variables : per pupil spending, avg class size, asian percent, black percent, hispanic percent, multiple percent, white percent, rigorous instruction, collaborative teachers, supportive environment, effective school leadership, strong family community ties, trust, disability percent, poverty percent, ESL percent, school size. I began by z-scoring the data and this was done before every PCA as scale matters for PCA. I then visualized the results (Fig 12). Looking at this Scree Plot, and using the elbow criterion 5 components were kept. I also performed a PCA on math and reading scores, as they were highly correlated ($r^2 = 0.98$) and this allowed me to create a shared output space between both tests (Fig 12).

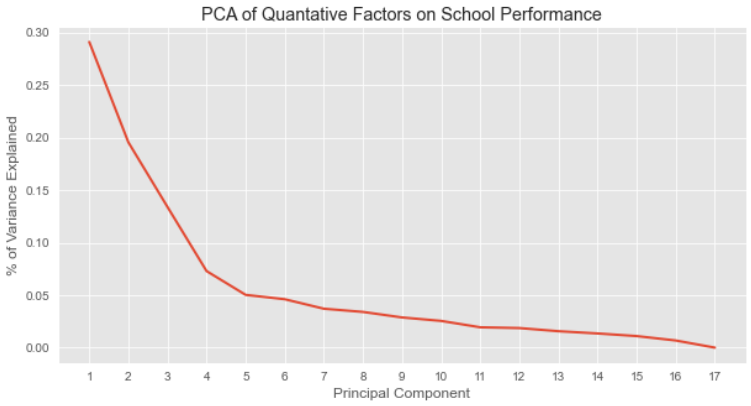

Fig 13 Shown Above

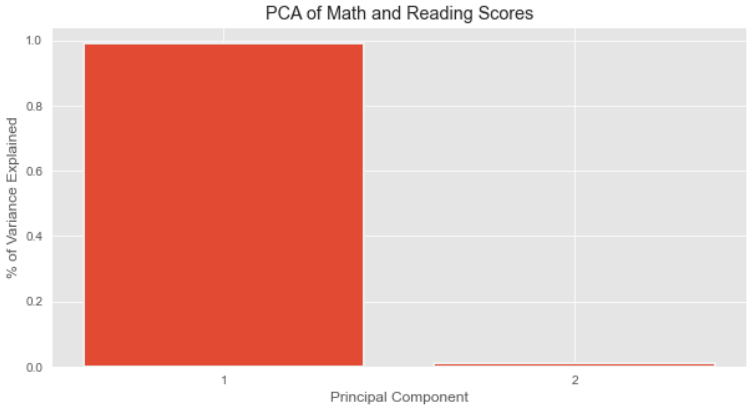

Fig 14 Shown Above

I then created 3 multi-regression models that took the 5 principal components as inputs. The output space for the models were math scores, reading scores, and the shared PCA space. The models performance was as follows: math scores $r^2 = 0.7809$, reading scores $r^2 = 0.7814$, and the shared PCA space $r^2 = 0.7858$. These models performed better than our original simple

models. I then retrained these models using 80% of the data for training and reserved 20% for testing. I wanted to ensure that an out-sample test would produce similar results - that I had not overfit the models. The models performance was as follows: math scores $r^2 = 0.74529$, reading scores $r^2 = 0.75569$, and the shared PCA space $r^2 = 0.75572$. Although these are lower than the in sample results, these models still performed better than our simple models, and maintained good performance. Using just 5 factors I was able to explain about 75% of the variance in test scores.

9 Build a model of your choice – clustering, classification or prediction – that includes all factors – as to what school characteristics are most important in terms of a) sending students to HSPHS, b) achieving high scores on objective measures of achievement?

Perhaps obviously, applications are correlated to admissions. After all, if you do not apply you cannot get in. Overall, poverty percentage and disability percentage were very important to test scores. Furthermore, white percent and poverty percent had a strong negative correlation $r = -0.82$. This indicates how wealth is distributed unevenly between the races. Poverty percentage also was a strong predictor of reading and math scores ($r = -0.77$ and $r = -0.75$). This means that students in poverty were at a rather large disadvantage, and so were students who had some form of disabilities.

10 Imagine that you are working for the New York City Department of Education as a data scientist (like one of my former students). What actionable recommendations would you make on how to improve schools so that they a) send more students to HSPHS and b) improve objective measures or achievement.

I would recommend that the state reallocate resources. The city should focus on and spend money to promote students to excel and want to get into these schools. Perhaps, the city could add prep classes after school and tutoring by teachers for students who want to focus on these schools. Students in low-income neighborhoods schools should have more spent on their education. Wealthy students have access to prep, and do not need smaller class sizes. Furthermore, working on fixing income inequality would also help, as wealthier student perform better in testing as well as in getting into HSPHS. Finally, I would recommend additional spending on students with disabilities. These students require special focus and attention from teachers. By reallocating resources and providing low-income and disabled students with the resources they need to succeed the average test scores would rise and total admits to HSPHS would also rise.