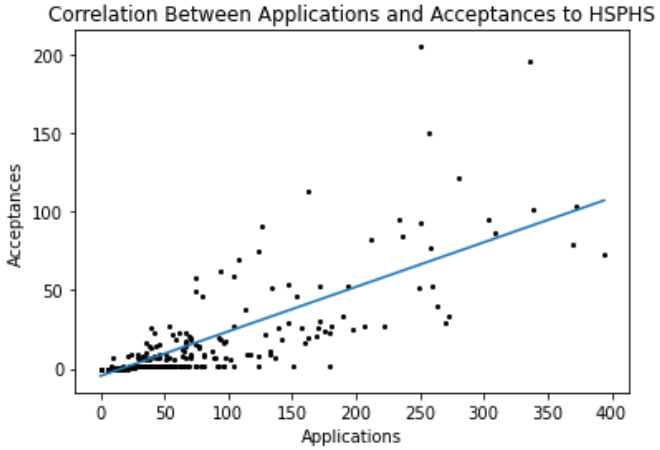
Big Data Project
Richard Zhu

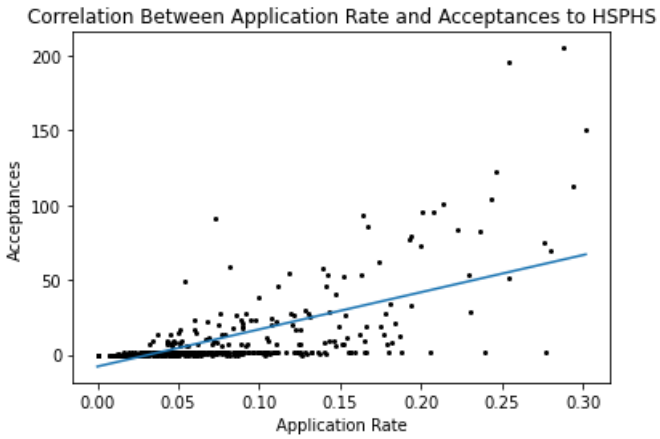
For this project, I used NumPy arrays to store and work with all my data. For each question, I extracted the necessary data and stored it in separate NumPy arrays/matrices. Because I used a dtype of 'unicode' when retrieving the data through np.genfromtxt(), I converted the extracted data to its proper datatype using astype(*datatype*). To deal with missing data and mitigate any inconsistencies, I removed all empty data and their corresponding entries in relevant datasets. I also utilized the Matplotlib, SciPy, and Scikit Learn libraries to plot plots and perform correlation, linear regression, and hypothesis testing functions. To handle dimension reduction, I conducted a PCA on relevant data using sklearn.decomposition's PCA class.

1) The correlation between the number of applications and admissions to HSPHS is 0.801727. I retrieved the applications data which was the third column 'applications' and the admissions data which was the fourth column 'acceptances' from my processed 2d data array and stored them in two separate arrays. I then used Numpy's corrcoef function to find the Pearson product-moment correlation coefficient (r) of these two arrays.



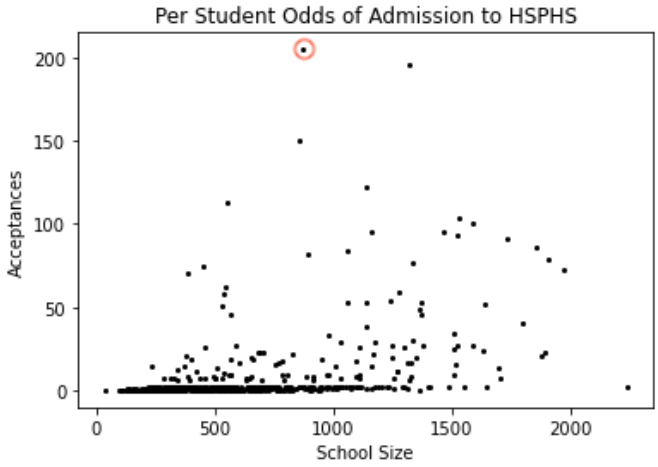
This scatter plot was obtained using Matplotlib and shows the correlation between applications and acceptances to HSPHS as well as a line of best fit.

2) The raw number of applications is a better predictor of admissions to HSPHS than application rate because it has a greater correlation value of 0.801727 compared to 0.658751. To find the application rate, I retrieved the school sizes from the data set as an array and divided it from the applications array. I then used Numpy's corrcoef function to find the Pearson product-moment correlation coefficient of the application rate and admissions rate array.



This scatter plot was obtained using Matplotlib and shows the correlation between application rate and acceptances to HSPHS as well as a line of best fit.

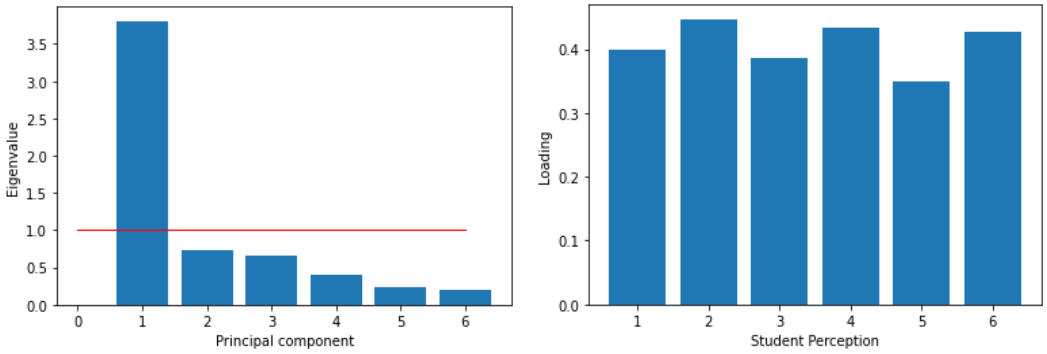
3) The Christa McAuliffe Intermediate School 187 in Brooklyn, NY (20K187) has the best per student odds of sending someone to HSPHS with a rate of 23.48%. To find this, I iterated through a school size array and admissions array obtained from the data set. On each iteration, I checked if the number of admissions divided by the school size of the specific school was greater than the last greatest value I had stored in a variable. If it was, I updated the greatest value variable and a variable that stored the index of the school. At the end of the loop, the greatest value variable stored a value of 0.2348 and I used the index variable to find the corresponding school.



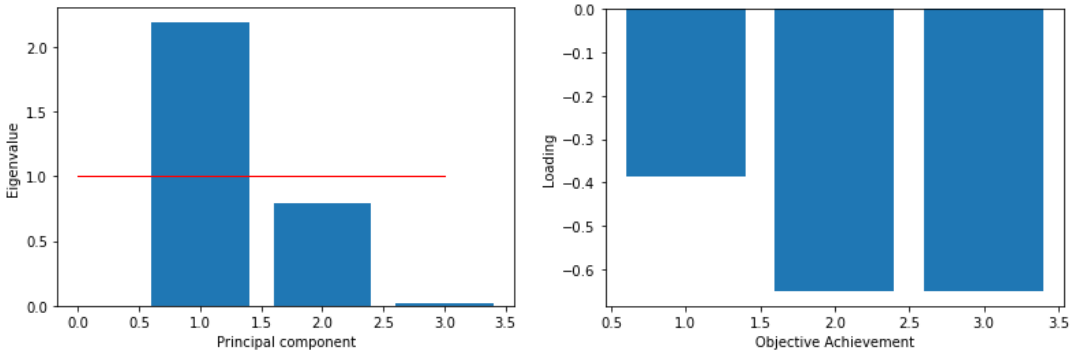
The data point circled red in the above figure represents 20K187 at a per student HSPHS send-off rate of 23.48%.

4) There is a minimal relationship between how students perceive their schools and how the school performs on objective measures of achievement. I tackled this problem by first

performing a PCA on both the student perception data and the objective achievement data. In both PCAs, the Kaiser criterion and "elbow" criterion revealed one principal component, which was consistent with the heat maps that I had analyzed previously. I then used SciPy stat's linregress() function to find the coefficient of determination of the simple linear regression between the principal component from each data set. The low COD of 0.135 reveals that student perception of their schools predicts a low portion of change in objective school performance.



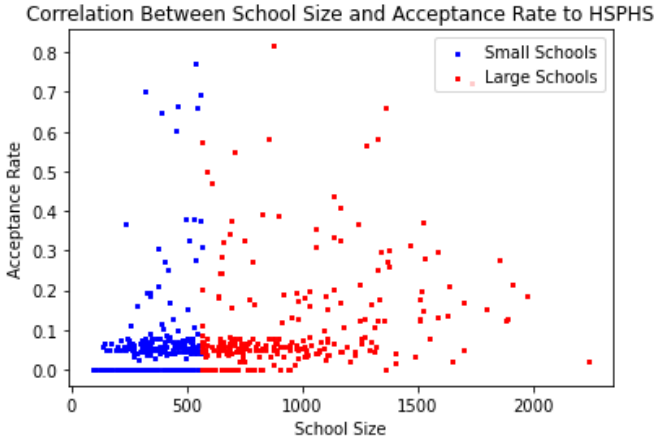
The first plot shows the sorted eigenvalues for student perception with the red line indicating the Kaiser criterion line. The second plot is the loadings, which indicates the weight of each factor. In this case, the second bar has the greatest loading, which corresponds to collaborative teachers.



The two above plots are for objective measures of achievement.

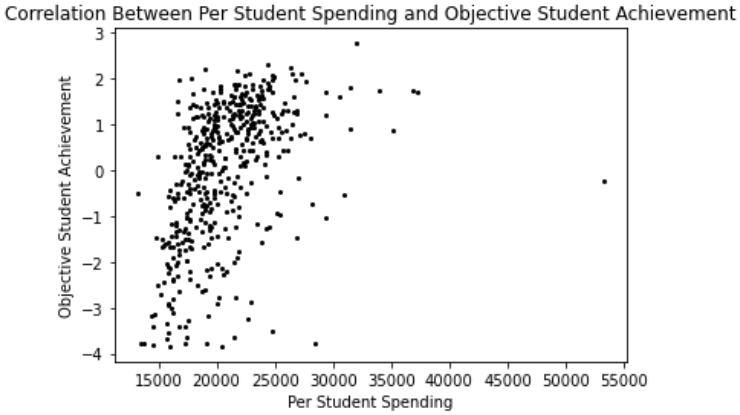
5) The hypothesis I want to test is that large schools have a higher admission rate to HSPHS than small schools. The null hypothesis is that large schools and small schools have identical expected admissions rate values and I assume that the populations have identical variances by default. To test this hypothesis, I first filtered a school size, admissions, and applications array obtained from my data set for only values corresponding to application values greater than 0. I then obtained the admission rates by dividing the new admissions array by the new applications array. The filtering step was done to ensure there was no

dividing by 0. I then computed the median of the new school size array and categorized the schools by size corresponding to the median. If the school size was less than the median, the admission rate of the school was placed in a small school array and if the school size was greater than or equal to the median, the admission rate was placed in a large school array. I then used SciPy stat's ttest_ind() function to run a two sample t-test on the two data sets. Because the computed p-value of 0.00028 is significantly less than the alpha value of 0.05, there is sufficient evidence to reject the null hypothesis and conclude that large schools have higher admission rates to HSPHS than small schools.



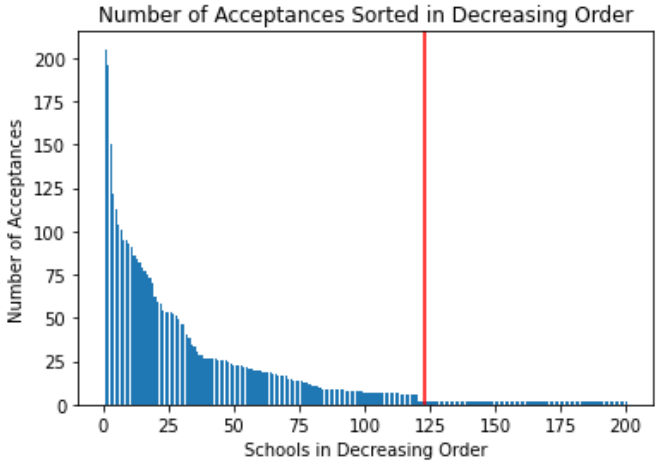
Above is a correlation plot between school size and acceptance rate for small and large schools where 565 is the median school size.

6) There is evidence that lower per student spendings leads to higher objective measures of achievement. The null hypothesis is that high and low per student spendings leads to identical objective measures of achievement. To test this hypothesis, I retrieved the per student spending, student achievement, math scores exceed, and reading scores exceed data. A PCA for the objective measurements of achievements revealed one principal component. I then categorized the schools by per student spending corresponding to the median spending. If the school's per student spending was less than the median, the corresponding student achievement principal component value was placed in a low spending array and if the per student spending was greater than or equal to the median, the corresponding student achievement principal component value was placed in a high spending array. Then, I ran a two sample t-test using the low spending and high spending arrays to obtain a p-value. Because the computed p-value was 2.06e^-23 is significantly less than the alpha value of 0.05, there is sufficient evidence to reject the null hypothesis and conclude that lower per student spendings leads to higher objective measurements of achievement.



Above is the correlation plot for per student spending and objective student achievement. Note that the median per student spending value is $20,099.

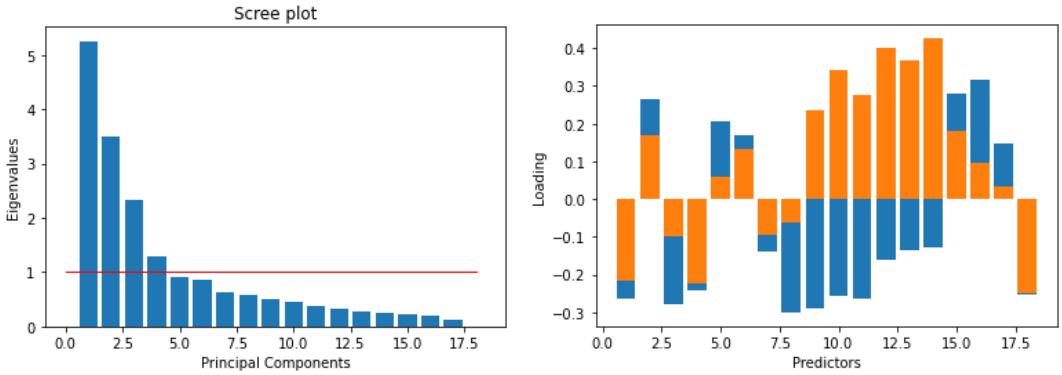
7) 20.7% of the schools account for roughly 90% of all students accepted to HSPHS. To find this, I first sorted my acceptances array in decreasing order of number of admissions. Then, I iterated through the array, checking if the sum of the acceptances of all the schools I iterated through was greater than or equal to 90% of the total number of acceptances to HSPHS. If it was, I stopped the iteration. The total number of acceptances to HSPHS was 4461 and 90% of that is 4014.9. This was met at the 123rd school with a total of 4016 admissions. Thus, 123 out of 594 total schools, which is roughly 20.7%, account for roughly 90% of all students admitted to HSPHS.



Above is a visual representation of the first 200 schools' acceptances to HSPHS. The red line indicates the 123rd school.

8) School district conditions (explained in 9) are a good predictor for acceptance to HSPHS and an even better predictor for objective measures of achievement. I first stored all of the school characteristics in a predictors matrix as well as the three objective measures of

achievement in a separate matrix. I then ran a PCA on both datasets. For the predictors dataset, the "elbow" criterion revealed 2 factors. I then analyzed the loadings to determine how to consolidate the factors into the two principal components. The first component was labeled as "school district conditions" and the second component was labeled as "relational conditions". I then performed linear regressions using these components with the admissions as well as the principal component found from the PCA done with the objective measures of achievement. Relational conditions predict a minimal portion of change in admission to HSPHS and objective measures of achievement, with a COD of 0.096 and 0.011 respectively. However, school district conditions predict a good portion of change in acceptance to HSPHS with a COD of 0.34 and predict a large portion of change in objective measures of achievement with a COD of 0.71.



Above is the scree plot for school characteristics with the red line indicating eigenvalues above 1. The second plot illustrates the loading for each predictor.

9) By analyzing the PCA scree and loading plot and the high COD found above, I found that the most relevant school characteristics in determining acceptance of their students to HSPHS are the school's school district conditions, specifically white percent, rigorous instruction, collaborative teachers, and supportive environment. The set of criteria above are defining qualities that distinguishes different school districts from each other. According to the Fiscal Policy Institute, white families have the highest average income in New York. Those in more affluent neighborhoods tend to be in closer proximity to better school districts that offer more rigorous instruction, collaborative learning styles, and supportive environments. In addition, affluent families have access to tutoring programs and helpful resources, ultimately contributing towards students' success in their academics. Therefore, the characteristics of the school's school district are foremost determinants in determining acceptance of their students to HSPHS.

10) Based on previous evaluations and analysis, schools should not only encourage more applications to HSPHS, but also work on scaling their schools to send more students to HSPHS. There is an extremely high positive correlation between the number of applications and acceptances to HSPHS. The variance in the number of applications to HSPHS can be due to school size, individual students' motivation, and various other factors. However, by encouraging more students to apply to HSPHS, schools not only expose and encourage the possibility of a future at a HSPHS, but also give students a new incentive to succeed (just as college prospects encourage high school students). There is also significant evidence that indicates that larger schools have higher admission rates to HSPHS. By understanding what factors contribute to such success, whether it be an abundance of resources or higher levels of competitiveness, smaller schools can use this knowledge to scale their schools and introduce beneficial change. To improve objective measures of achievement, it is important for schools to analyze and reevaluate their financial spendings on their students. It was shown that lower per student spending leads to higher objective achievement. For schools that can be classified as such, it is likely that they allocate more resources towards improving the quality of teachers and learning resources. On the other hand, schools that are extracurricular-oriented may be spending more on each student, however, the money is used to advance skills such as leadership or athletics. Although these traits can be equally if not more important, by dedicating more financial resources to improving the quality of academics, schools may be able to improve objective measures of achievement.