Abstract: The purpose of this research report is to explore the characteristics of middle school data provided by the New York City Department of Education and to determine whether any characteristics can yield meaningful insights about acceptance rate to one of New York's specialized high schools (HSPHS). The dataset was composed of multivariate data of 595 middle schools, including zero and missing values. For the purposes of calculations and models, the schools with missing relevant components were excluded for all purposes. Dimension reduction techniques were applied and a PCA test was done to decrease the number of variables used for modeling. Using the correlation matrix and variances of the eigenvalues, the test determined that for qualitative student responses, a factor of three can explain most of the variances in student response scores (above 0.8) as opposed to six. For the three objective measures, the dependent variable must be only one measure and due to the significant correlation between reading and math scores, reading/math scores were used arbitrarily. Student achievement did not correlate as strongly as the latter. The correlation of reading and math scores is 0.98, so using either score is acceptable. The PCA test is shown in **Figure A**. For the purposes of the PCA test, the data was transformed by z-scoring. The qualitative components also share similar correlations with each other as shown in **Figure B**, reinforcing that even if you choose to use just one

category, you'd still be covering a chunk of the explained variance, so it wouldn't be completely unreasonable to use any one of the variables.

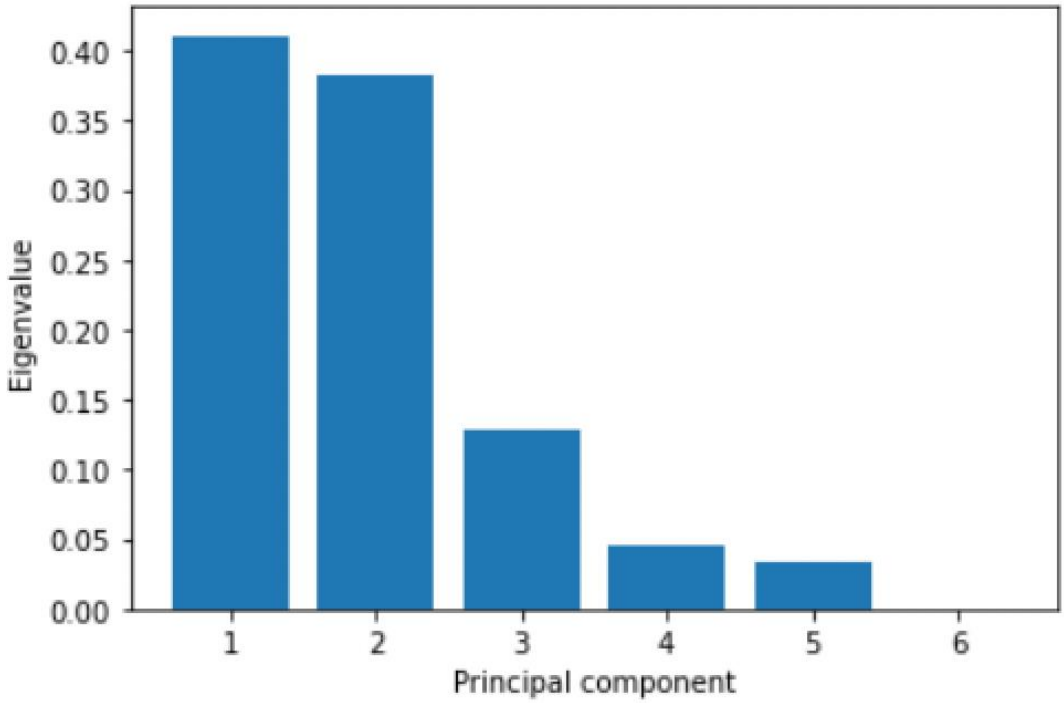

Figure A: PCA Test for Student Responses

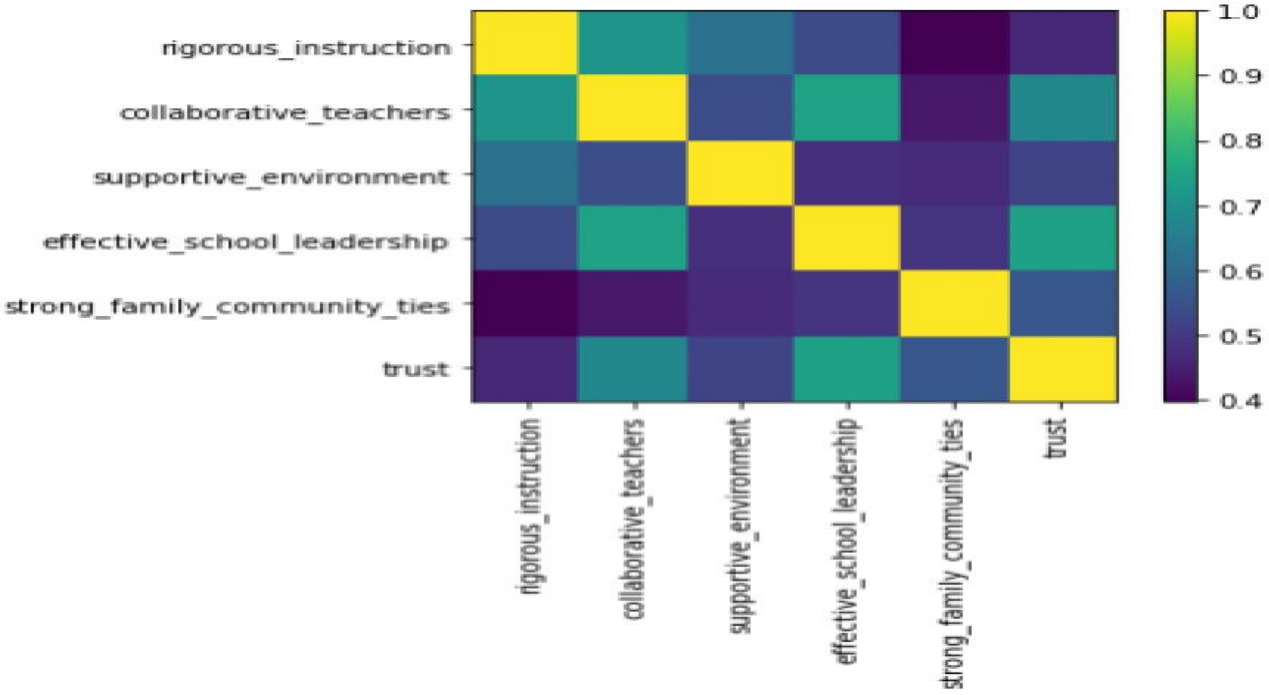

Figure B: Correlation Heat Matrix

1) The correlation between the number of applicants and admissions to HSPHS appears to fit a linear model, that is, as the number of applications from a given middle school increases, the number of admissions also increases. This was determined because the dataset has a Pearson correlation coefficient of around 0.80, representing a strong positive trend. The Pearson correlation coefficient is a standard when assessing the efficacy of using a linear model to fit a dataset. **Figure 1** depicts the scatter plot for the data set of applications vs. admissions.

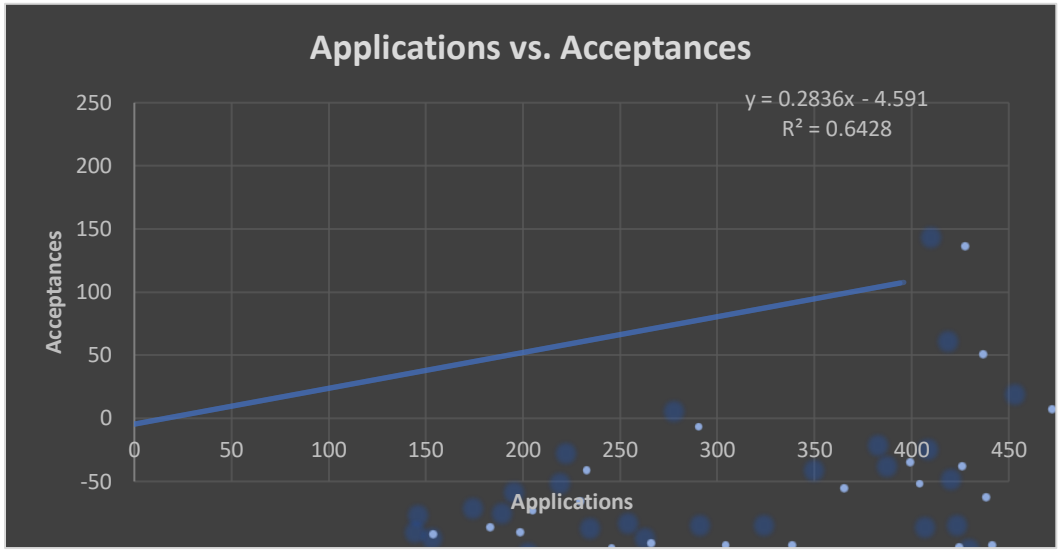

Figure 1: Scatter Plot for Acceptances based on Applications

2) Raw number of applications is a better predictor for acceptances than the application. Application rate was calculated by dividing the total number of applications by the school size. The application rate and acceptances were graphed on a scatter plot. Based on the general look of the data in **Figure 2**, it seemed to follow a linear model, so the Pearson coefficient was calculated to be nearly 0.65, indicating a moderately strong positive trend, meaning, as the application rate increases, so does the acceptances. This correlation is lower than the correlation calculated from **Figure 1**, indicating a worse

predictor for acceptances. Due to the Pearson coefficient being lower when using application rate, raw number of applications fits a better linear model. Moreover, there seems to be a moderately strong positive trend (R = 0.65) in school size and applications, meaning as the school size increases, so does the number of applications from that school. This is significant because it demonstrates that there may be a correlation between school sizes and acceptances to HSPHS. Thus, the admission rate was calculated to incorporate school sizes as a variable for the model. For example, if a school of 1000 students had 100 applications, 10% of the student body applied to the HSPHS. If a school of 100 had 20 applications, 20% of the student body applied to the HSPHS. Even though the raw number of applications is fewer in the latter school, it is double the student body, having a profound effect on the data while not having nearly as many acceptances as the former school. Due to this degree of variability when incorporating school sizes, raw number of applications is a better predictor.

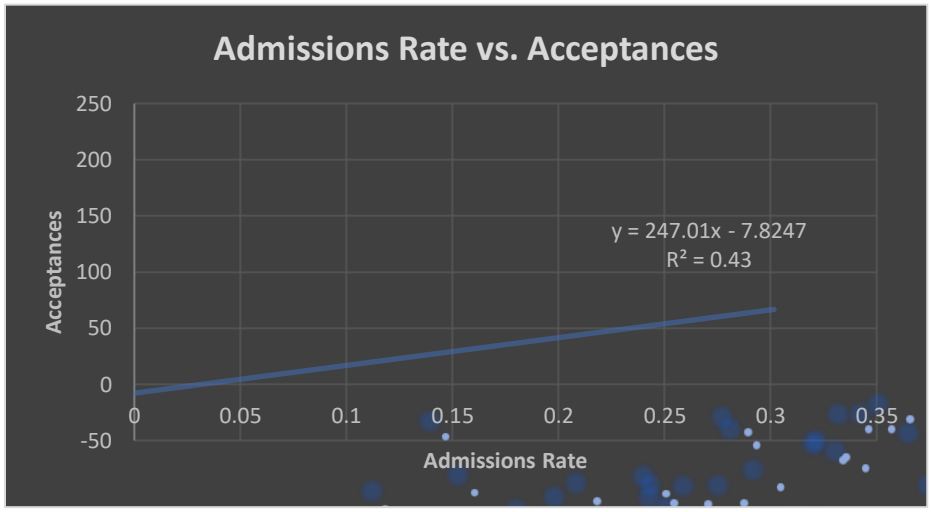

Figure 2: Scatter Plot for Admissions Rate vs. Acceptances

3) The school with the best person student odds of getting into a HSPHS is most likely The Christa McAuliffe School I.S. 187. From a quantitative point of view, this school has the highest acceptance rate among all other middle schools within the dataset analyzed, with slightly over 23% of its students getting accepted into a HSPHS and 80% of those students that applied received admissions offer. It is extremely rare for a school to have over 20% of its student body attend a specialized HS, less than ten schools have reached this mark. Deducing this middle school to be the best, however, required more evidence to make such a bold claim. The next parameters to be tested are the student achievement, math, and reading scores. This parameter was chosen because an assessment of the student's relative intelligence was needed. I.S. 187 ranked the highest among math and reading scores at 0.9 and had a student achievement score of 4.36. **Figures 3-4** show the frequencies of these scores.

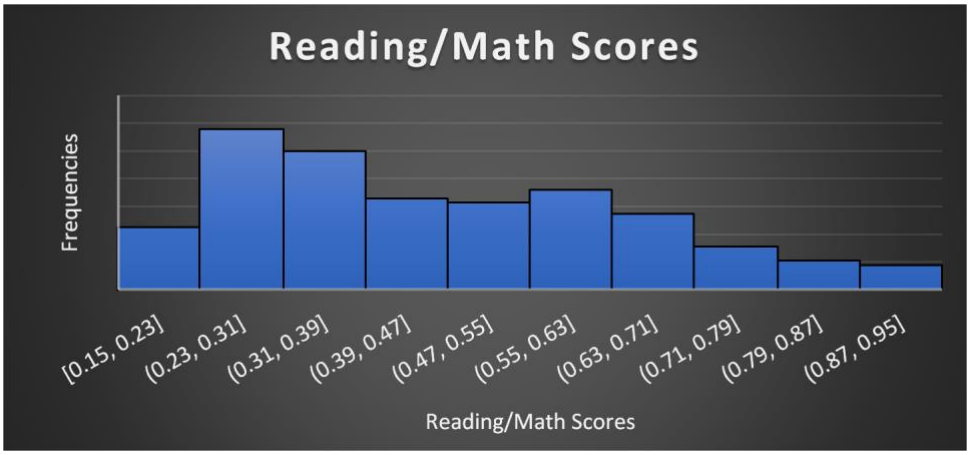

Figure 3: Reading/Math Score Frequencies

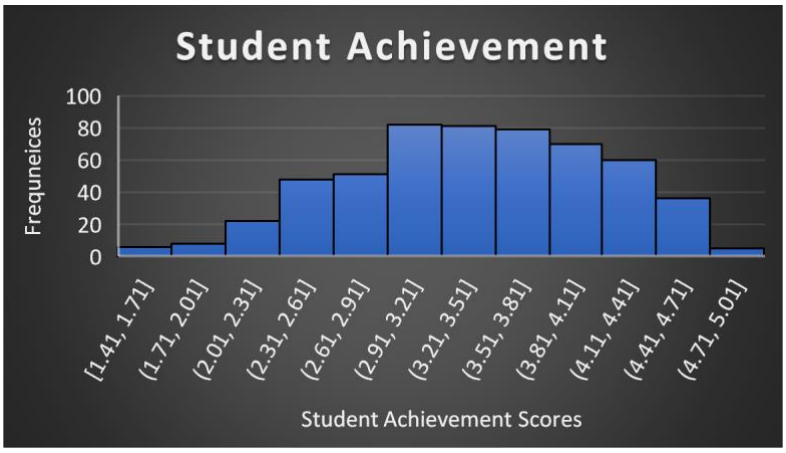

Figure 4: Frequency of Student Achievement Scores

There were other schools with above average acceptance rates that were compared against this middle school. Those schools also had relatively high student achievement scores and math/reading scores. There were schools that had superior scores to I.S. 187, but had no applicants to HSPHS, therefore cannot be compared against I.S. 187, that's where acceptance rate came into play to distinguish the school.

4) There does not appear to be a strong correlation between how students perceive their school and how the school performs on objective measures of achievements. The Pearson coefficients calculated for each independent student response were low for all categories except rigorous instruction, collaborative teachers, and supportive environment, each of which had a correlation between 0.3-0.5. These three categories have weak positive correlations to math/reading performance. The other three categories had correlations in the range of 0-0.2 indicating an extremely weak or no correlation whatsoever. The data was highly clustered in a specific range of values (3.0 - 4.5), the average of each qualitative component was around 3.5. Anything below this value was considered relatively poor student perception and anything above was

considered strong student perception. The average score was 0.44, skewed to the lower end of the average. The aggregated frequencies of the qualitative components are shown in **Figure 5**. The aggregated frequencies of reading/math scores are shown in **Figure 6**. Schools that had poor student perception were categorized into a subset of the data. The math scores of the schools within this subset were analyzed and for most categories, half the students had good scores and the other half had poor scores. This indicates that student perception did not have a significant impact on objective achievements and varies immensely. In addition, the correlation

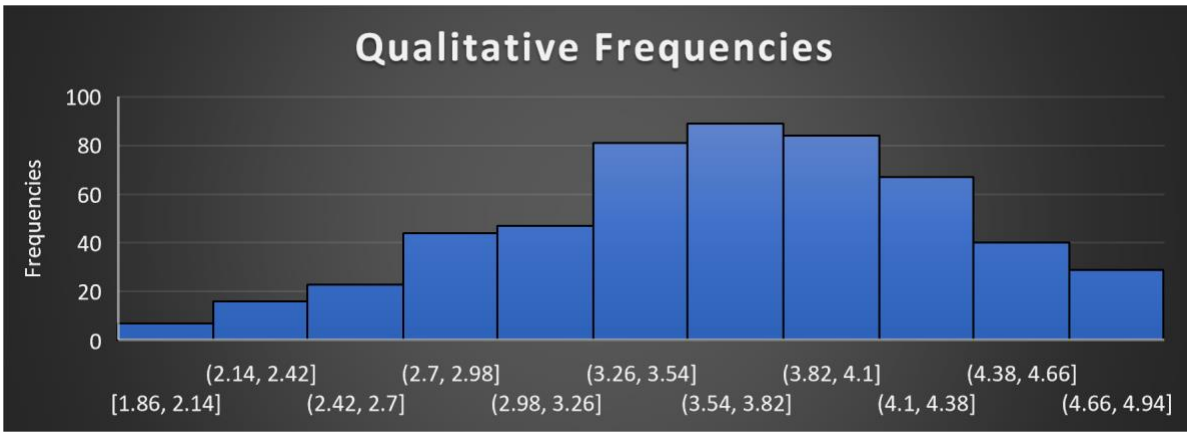

Figure 5: Aggregated Frequencies of Student Perception Responses

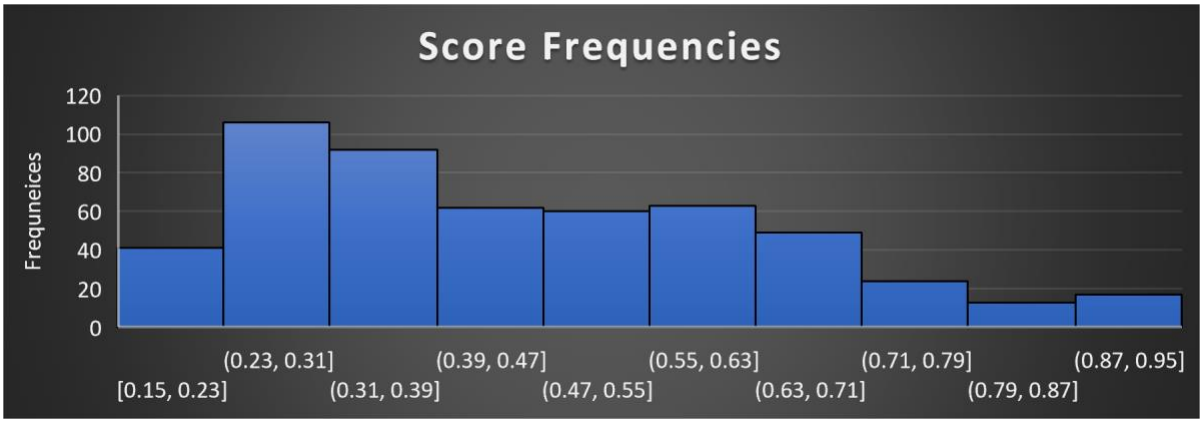

Figure 6: Aggregated Frequencies of Math/Reading Scores

5) The hypothesis to be tested is whether the percent population of Asians within a given student body is correlated with the acceptance rate to HSPHS. The acceptance rates were calculated by dividing the number of acceptances per school-by-school size, the average acceptance rate is approximately 8%. If the school had an acceptance rate above 8%, they were above average and were categorized in a separate dataset. Only 125 schools out of the 587 schools examined were above average. Within this above average subset, the Asian percentages of each school were iterated over. If a school had an Asian percentage of above 25%, the ethnicity was a majority in the school. This benchmark was derived by dividing the four ethnicities examined (Asian, Black, Hispanic, White), which would account for 100% of the student population and divided that by 4 to arrive to 25%. This benchmark may not necessarily be the majority ethnicity of the school however, as this was calculated assuming schools had a normally distributed ethnicity count (which may not always be the case as some schools have disproportionate number of a specific ethnicity). Out of the 125 above average schools, 59 of those schools had an Asian population that is greater than or equal to 25%. This represents 47% of the above average schools, indicating that some of the higher performing schools (Using admission to HSPHS as a benchmark of high performance) have a majority or at the very least a sizeable Asian student body.
6) There appears to be a moderately positive correlation between the availability of material resources and the performance of students on math/reading. The Pearson correlation for class sizes and math scores was calculated to be 0.56. This is shown in **Figure 7**. The Pearson correlation for spending per pupil was calculated to be -0.49. This

is shown in **Figure 8**. Based on these findings, class sizes seem to impact student performance more than spending per pupil.

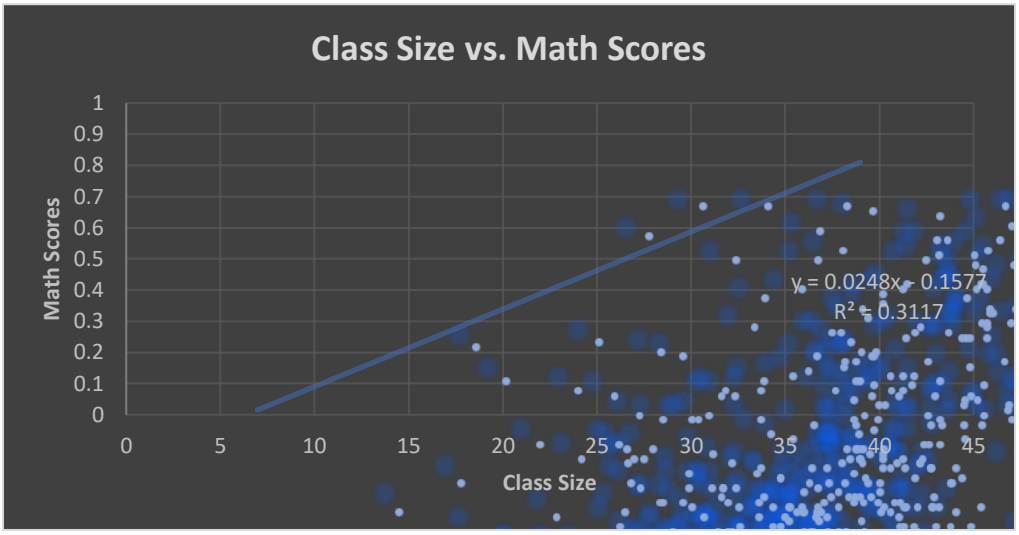

Figure 7: Class Size vs. Math Scores

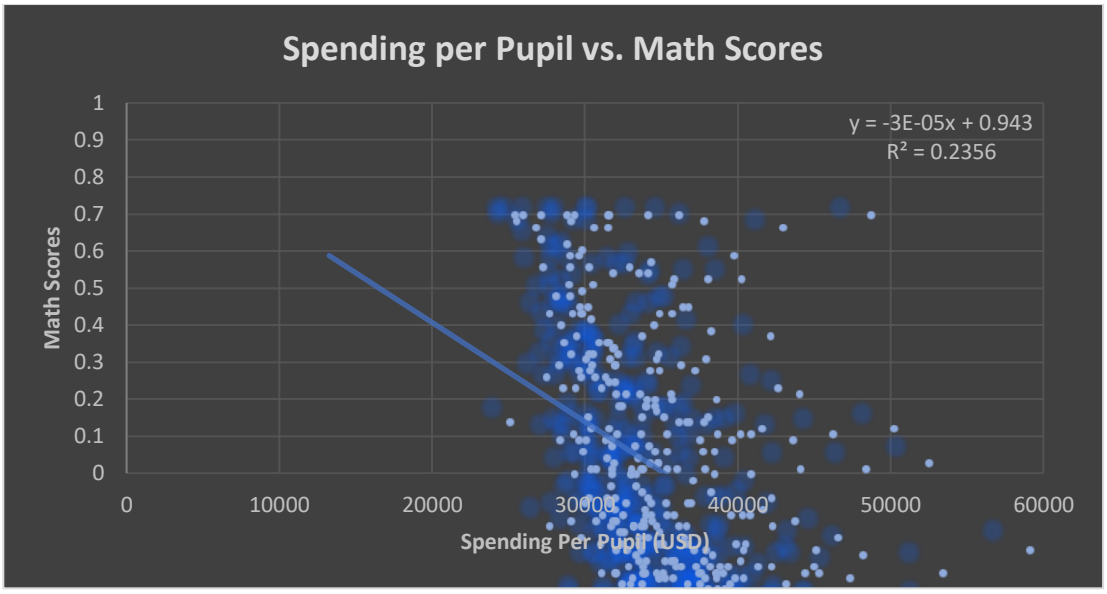

Figure 8: Spending Per Pupil vs. Math Scores

7) The proportion of schools that account for 90% of all HSPHS acceptances is 122 over 595 schools, which equates to approximately 21% of the schools. The total number of acceptances within the dataset was calculated to be 4461, and 90% of that is 4014. Upon sorting an array of all acceptance values, starting from the greatest integer, a

resting count of the acceptances was taken until the program hit the 4014 marks. It took approximately 122 iterations to get to this mark, weighing schools with greater number of acceptances more significantly. A good number of schools had only a few acceptances, arbitrarily affecting the count. **Figure 9** shows a bar graph of acceptances.

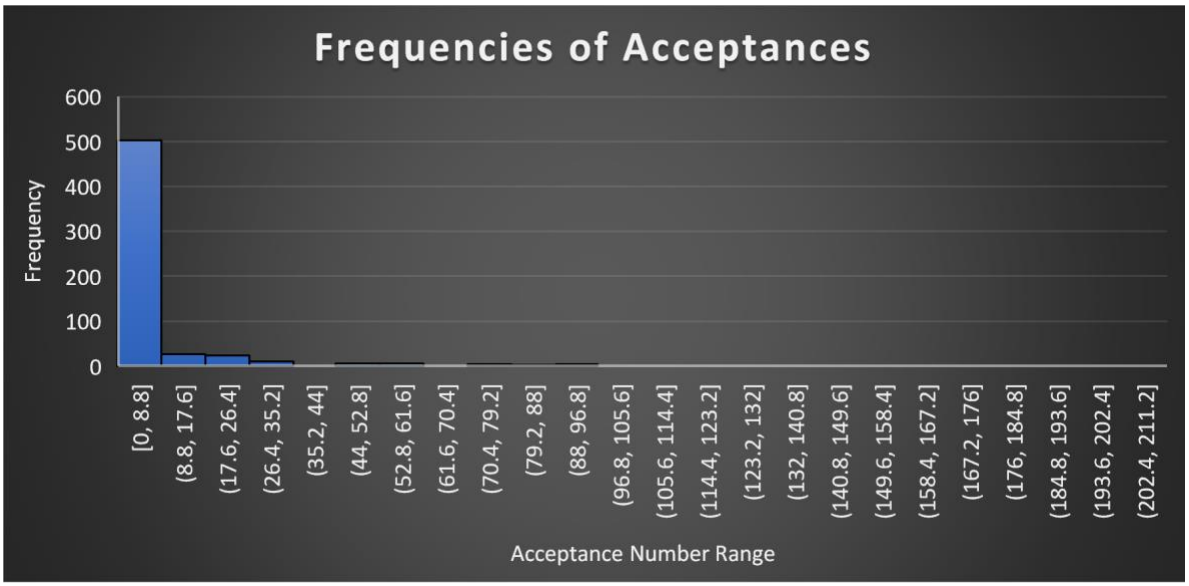

Figure 9: Frequencies of Acceptances

8) A multiple linear regression model was used to predict the math scores for each middle school. This regression models allows multiple predictor variables to be incorporated in the modeling of a single dependent variable, in this case, it would be math scores. **Figure 10** shows the summary of the major test components of a multiple linear regression model with the following predictors: collaborative teachers, rigorous instruction, disability/poverty/ESL percentage and White/Black/Asian/Hispanic percentage. The first step in figuring out which predictor variables to use was a PCA test and dimension reduction occurred. Multiple variables that correlate strongly with each other allowed the reduction of some variables. Moreover, some variables had

particularly stronger affects to math scores when using simple linear regression models, so those variables were also included. The correlation coefficient of this model was over 0.80, demonstrating that the model follows a strong positive correlation. The R squared value of 0.65 indicates that 65% of the data's variance can be explained using these nine predictor variables.

Regression Statistics								
Multiple R	0.80720803							
R Square	0.65158481							
Adjusted R Sc	0.64571264							
Standard Erro	0.12673407							
Observations	544							
ANOVA								
	df	*SS*	*MS*	*F*				
Regression	9	16.039908	1.782212	110.961575	3.4021E-116			
Residual	534	8.57685378	0.01606152					
Total	543	24.6167618						
	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.49802754	0.26459425	1.88223116	0.06034815	-0.02174573	1.01780082	-0.0217457	1.01780082
rigorous_inst	0.05217059	0.01257099	4.15007884	3.8673E-05	0.027475935	0.07686524	0.02747594	0.07686524
collaborative	-0.0024034	0.01186044	-0.2026362	0.83949654	-0.02570221	0.0208955	-0.0257022	0.0208955
disability_per	-0.0082915	0.00083857	-9.8877074	2.7777E-21	-0.00993885	-0.0066442	-0.0099389	-0.0066442
poverty_perc	-0.0035891	0.00063001	-5.6969929	2.0168E-08	-0.00482674	-0.0023515	-0.0048267	-0.0023515
ESL_percent	-0.0042563	0.00067214	-6.3324488	5.1234E-10	-0.00557669	-0.0029359	-0.0055767	-0.0029359
asian_percer	0.00478027	0.0029117	1.64174705	0.10123119	-0.00093952	0.01050006	-0.0009395	0.01050006
white_percer	0.00293472	0.00278827	1.05252625	0.2930342	-0.00254259	0.00841204	-0.0025426	0.00841204
black_percer	0.00129819	0.00286898	0.45249189	0.65109843	-0.00433768	0.00693406	-0.0043377	0.00693406
hispanic_per	0.00256387	0.0028367	0.90382015	0.36649832	-0.0030086	0.00813633	-0.0030086	0.00813633

Figure 10: Multiple Linear Regression Summary

9) The school characteristic that seems most relevant in determining admissions to HSPHS by far must be number of applications. If many students apply to HSPHS, it is only natural that the possible chances of acceptances from that middle school increases. Word of mouth is the most effective way to market taking the exam for kids, if all your friends in your class are taking this test that will supposedly benefit you in the future, you, as a middle student will most likely be inclined to take the exam yourself. Other key characteristics that are moderate to strongly correlated with acceptances are: Asian

percentage, school size, and math/reading scores. School size is strongly correlated with acceptances because the number of potential applicants only increases as students increase. Admission into HSPHS is dependent on how well you score on a standardized exam called the SHSAT, which simply tests your reading/math skills, so having exceptional math/reading scores in the school will naturally boost your chances of admission. The Asian percentage correlation is a more controversial subject to touch upon. There are often political debates about the number of Asians who are attending high class schools and are labeled as “the model minority”. Political agendas often influence data in unforeseen ways, but the purpose of this paper is not about furthering political agenda, it’s to draw insights. As a Bengali American raised in an Asian household in Queens, I have experienced the NYC education system up until college. I have both attended middle school here (My middle school is in the dataset) and received admission to a HSPHS, even going as far as tutoring kids for the SHSAT. Personally, I believe that reason for the high amounts of Asians in HSPHS is a cultural matter. Studying for the SHSAT was something non-negotiable in the childhood of my siblings, cousins, and me. Everybody in my generation took that exam and studied for it because of the strong emphasize of education in Asian families, and pursuing the top schools is a demonstration of that. My upbringing is arbitrary in the grand scheme of the dataset, and I chose to include it in the hopes of bringing more meaning to the elaborate narrative painted by the data. Moreover, when I attended high school, most of my classmates were in fact, Asian American.

10) Based on the findings from the data analysis of NYC middle schools, to improve admissions rate to HSPHS, application numbers need to rise. Many middle schools simply had zero applications to HSPHS, and an even larger number of schools had less than 50 applications. Increasing awareness both inside and outside of school environments are necessary to boost application numbers from schools. Furthermore, implementing support programs during school to help aid the students in the application process would be tremendously useful. Incorporating after-school study sessions or additional tutoring for students who wish to practice exam problems can also help. To combat the disproportionate number of ethnicities within HSPHS, opportunity programs that help low-income and minority groups can be created. These additional programs can help boost objective-measures of achievement as well, because ethnicities seem to correlate strongly with performance in school. Socioeconomic status plays a huge role in the quality of education that you receive and addressing these concerns would be the optimal course of action when trying to increase student performance.