FINAL PROJECT

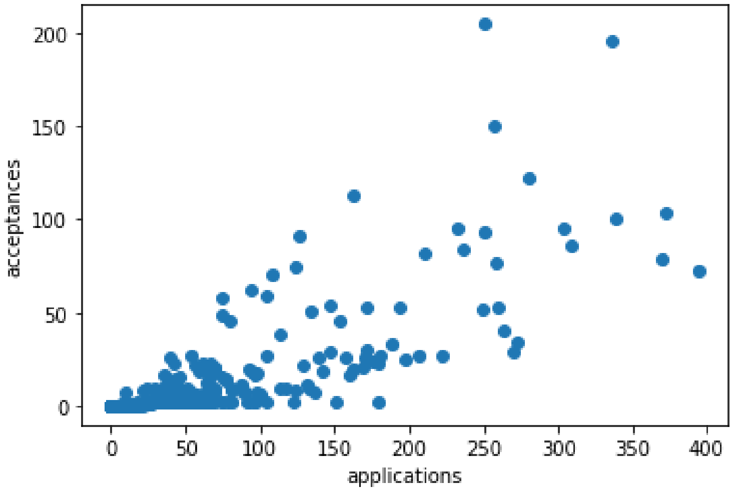
Introduction to Data Science – DS UA 112

OKUBO, Christine Joy C.

This project utilizes a dataset of NYC middle schools provided by the New York City Department of Education. The data includes various information about the academic achievement, population, diversity, and other characteristics of the listed schools which may predict its students' admission to highly selective public high schools (HSPHS). To handle any missing values, the dataset was imputed with the median of the column on which the blank data exists. The presence of outliers and the uncertainty on the underlying distribution of the dataset makes the median a reasonable imputer as it is robust to such conditions. To reduce dimensionality in investigating the relationship between a group of characteristics and outcomes, Principal Component Analysis (PCA) was used. The group of input data was standardized and plotted on a scree plot. This project makes use of the Elbow and Kaiser criteria in determining the number of factors of the PCA, taking the higher number of factors into account.

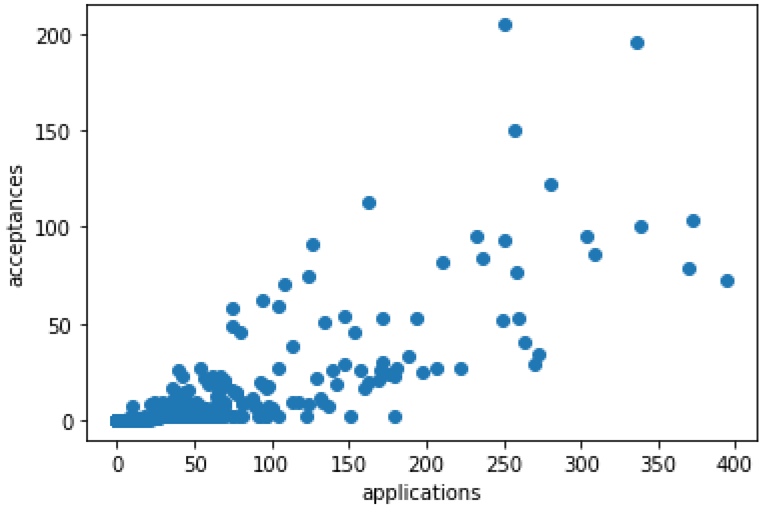
1) What is the correlation between the no. of applications and admissions to HSPHS?

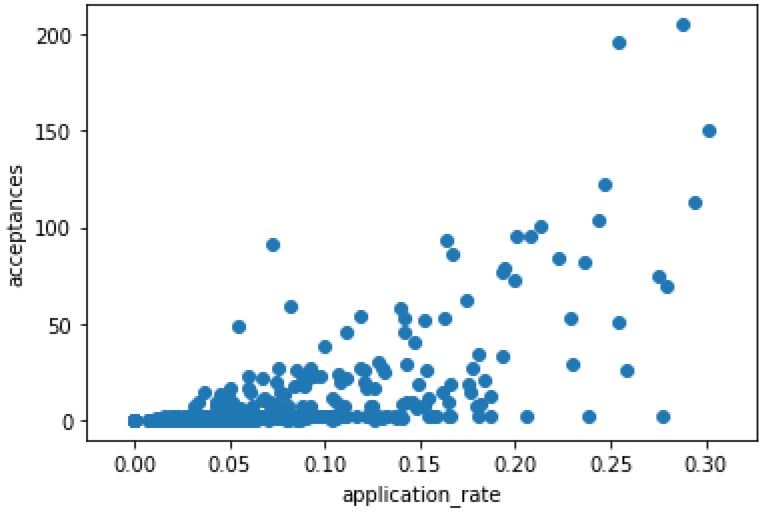
The correlation between the number of applications and admissions to HSPHS is 0.801727. This suggests that we can expect a strong linear and positive relationship between the two characteristics. Below is a scatterplot of the relationship:



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

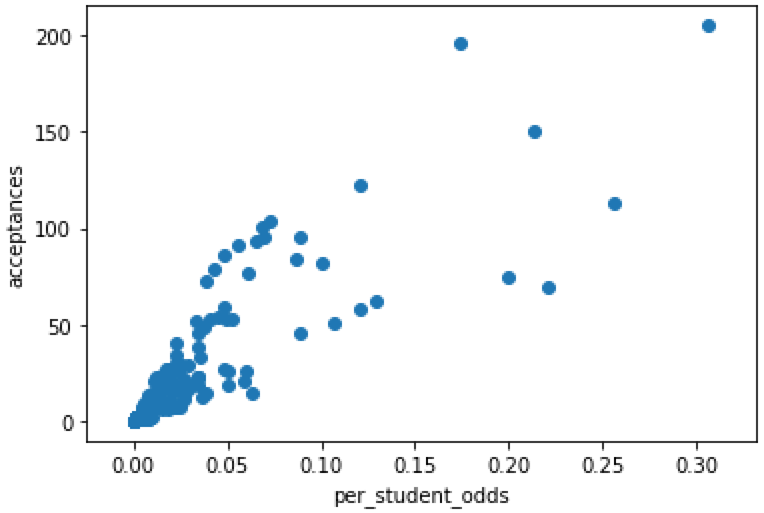
Upon obtaining the correlation of both the raw number of applications and application rate to HSPHS acceptances, I found that the number of applications is a better predictor.





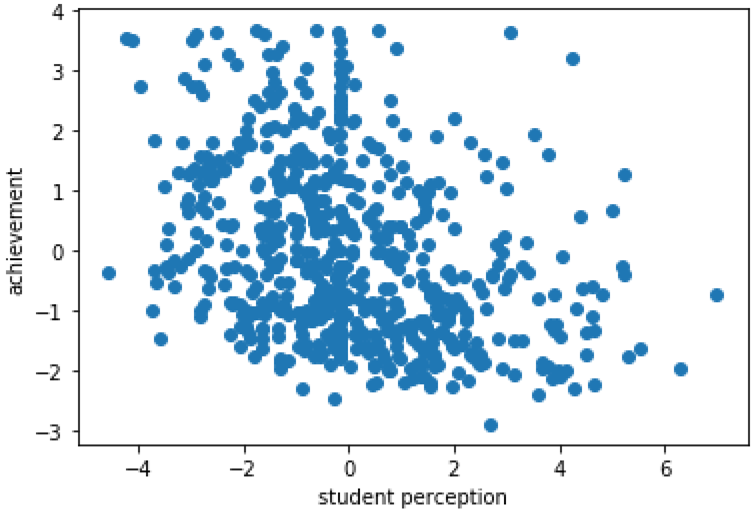
3) Which school has the best per student odds of sending someone to HSPHS?

THE CHRISTA MCAULIFFE SCHOOL\I.S. 187 has the best per student odds of getting accepted into an HSPHS at 0.3069 or 30.69%.



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).

There is a moderate negative linear relationship between student perception and academic achievement of the school. This is based on a -0.35 correlation between the two characteristics. The R-squared between the two is only 0.125 which means that students' perception only accounts for 12.5% of variance in achievement. This suggests that there are other factors that determine achievement or that the relationship is not ideally linear.



5) Test a hypothesis of your choice as to which kind of school performs differently than another kind either on some dependent measure.

Since the dataset provides demographic information on the population of the middle schools, I decided to test if colored schools are more or less likely to get acceptances into less colored schools. I conceptualized colored schools as those whose Asian, Black, Hispanic, and mixed race percentage exceeds the total median percentage of the population. I then decided on a Chi-squared (χ^2) test for the data.

NULL AND ALTERNATIVE HYPOTHESIS

Ho: Colored schools are equally likely to get into HSPHS as less colored schools.

Ha: Colored schools are not equally likely to get into HSPHS as less colored schools.

1. Calculating the test statistic
    a) Sample Information:

Category	n	Observed (O)	Expected (E)	(O-E)^2 / E
Colored	296	421	2222.99	1460.72
Less Colored	298	4040	2238.01	1450.92

$$\boxed{\chi^{2*} = 2911.64}$$

2. Probability Distribution

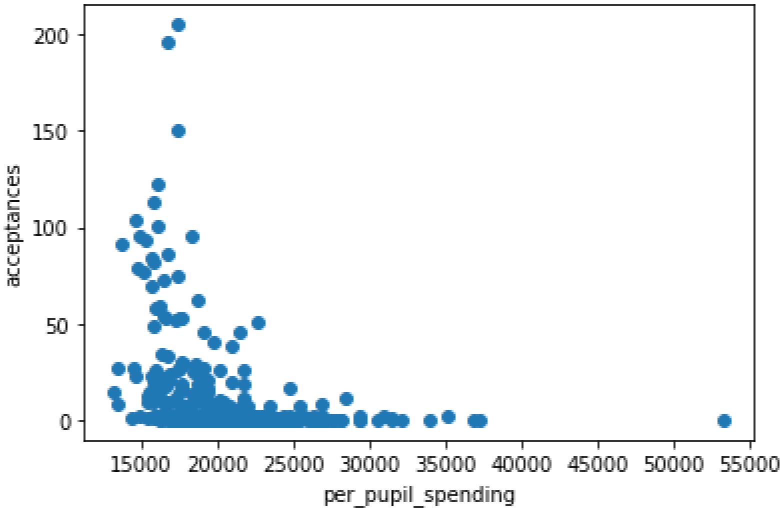
    [p-Value]

    a) $p < 0.0001$; The p-value is extremely significant.

3. Results
    a) Decision: Reject Ho.
    b) **At the 0.01 level of significance, there is sufficient evidence to suggest that colored schools are not equally likely to get into HSPHS than less colored schools.**

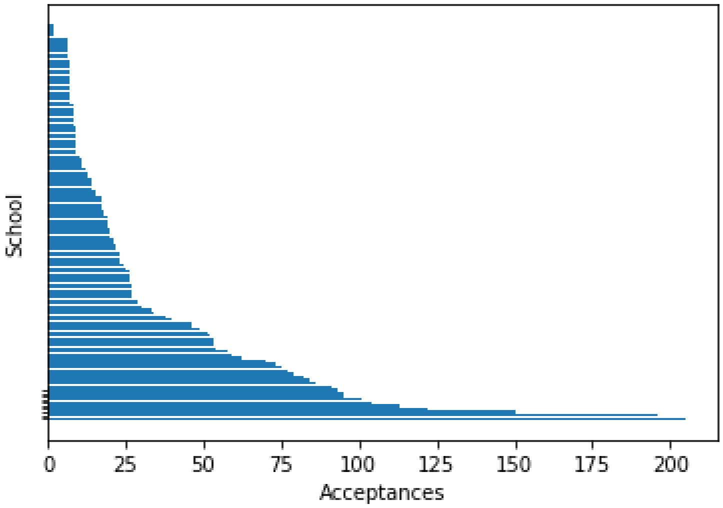
6) Is there any evidence that the availability of material resources impacts objective measures of achievement or admission to HSPHS?

A coefficient of -0.32 suggests a moderate negative and linear relationship between per pupil spending and acceptances to HSPHS.



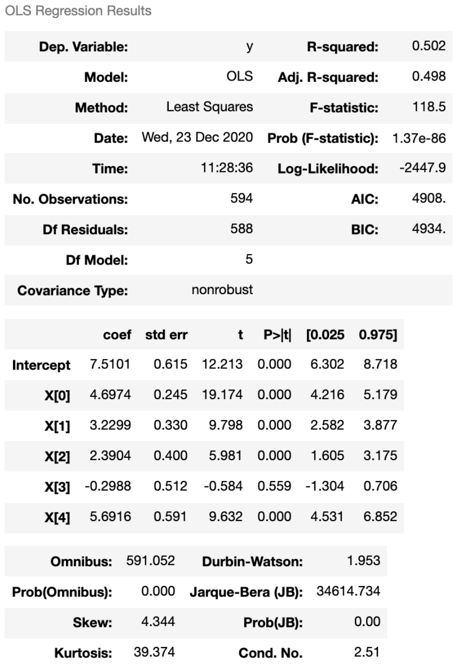
7) What proportion of schools accounts for 90% of all students accepted to HSPHS?

123 or 20.71% of schools account for 90% of all students accepted to HSPHS.



8) Build a model 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.

Regression model with acceptances as dependent variable:



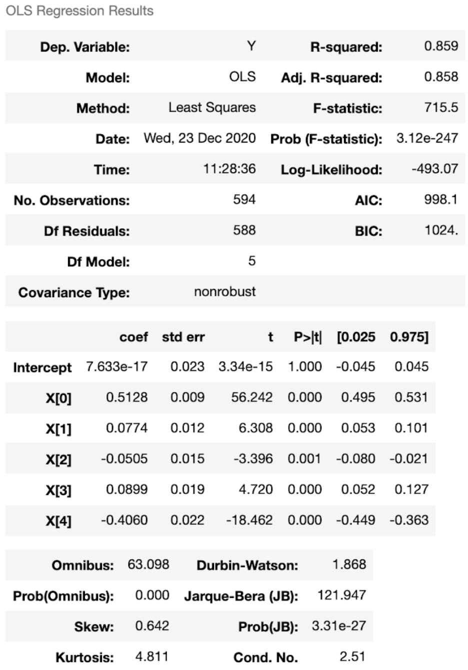
OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.502
Model:	OLS	**Adj. R-squared:**	0.498
Method:	Least Squares	**F-statistic:**	118.5
Date:	Wed, 23 Dec 2020	**Prob (F-statistic):**	1.37e-86
Time:	11:28:36	**Log-Likelihood:**	-2447.9
No. Observations:	594	**AIC:**	4908.
Df Residuals:	588	**BIC:**	4934.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	7.5101	0.615	12.213	0.000	6.302	8.718
X[0]	4.6974	0.245	19.174	0.000	4.216	5.179
X[1]	3.2299	0.330	9.798	0.000	2.582	3.877
X[2]	2.3904	0.400	5.981	0.000	1.605	3.175
X[3]	-0.2988	0.512	-0.584	0.559	-1.304	0.706
X[4]	5.6916	0.591	9.632	0.000	4.531	6.852

Omnibus:	591.052	**Durbin-Watson:**	1.953
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	34614.734
Skew:	4.344	**Prob(JB):**	0.00
Kurtosis:	39.374	**Cond. No.**	2.51

Regression model with achievement as dependent variable:



OLS Regression Results

Dep. Variable:	Y	**R-squared:**	0.859
Model:	OLS	**Adj. R-squared:**	0.858
Method:	Least Squares	**F-statistic:**	715.5
Date:	Wed, 23 Dec 2020	**Prob (F-statistic):**	3.12e-247
Time:	11:28:36	**Log-Likelihood:**	-493.07
No. Observations:	594	**AIC:**	998.1
Df Residuals:	588	**BIC:**	1024.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	7.633e-17	0.023	3.34e-15	1.000	-0.045	0.045
X[0]	0.5128	0.009	56.242	0.000	0.495	0.531
X[1]	0.0774	0.012	6.308	0.000	0.053	0.101
X[2]	-0.0505	0.015	-3.396	0.001	-0.080	-0.021
X[3]	0.0899	0.019	4.720	0.000	0.052	0.127
X[4]	-0.4060	0.022	-18.462	0.000	-0.449	-0.363

Omnibus:	63.098	**Durbin-Watson:**	1.868
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	121.947
Skew:	0.642	**Prob(JB):**	3.31e-27
Kurtosis:	4.811	**Cond. No.**	2.51

9) Write an overall summary of your findings – what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS.

From the scree plot of the characteristics of NYC middle schools (trying out up to 5 principal components), 'collaborative_teachers', and 'supportive_environment' seem to be the most commonly recurring and prominent loadings.

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 NYC middle schools focus on creating a more supportive environment in and out of the classroom. As suggested by the scree plot of the most relevant characteristics, focusing on improving the support for the student is more likely to lead to an increase in important indicators such as admissions to HSPHS and achievement. This can be operationalized by conducting more teacher and counselor trainings, ensuring safety policies against bullying or discrimination, or increasing per pupil spending.

APPENDIX