Yanni Chen
Yc3337

Big Data Project

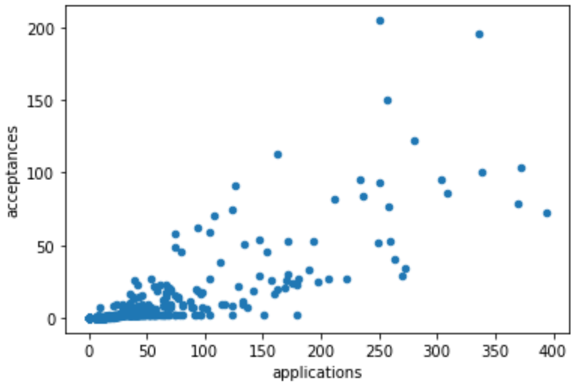
Process before doing questions:

1) Drop the missing value and its correspondent other value in the "application", "acceptance" and "school size" columns when dealing with question 1-3

   With regard that we need to calculate the correlation between students applied and accepted by HSPHS, it is more appropriate to remove both the application and acceptance of that specific school (that row of these two data value) if there is one of them missing, if any. When it comes to school size, since we need to calculate the acceptance rate by the school size, leaving school with unknow school size may make our result in acceptance rate inconsistent with other outcomes. Though we might lose power on deleting the row if one in three missing, it makes sure that all the data we use indicate more valid and consistent correlation instead of imputing the missing data with median/mode that may affect the accuracy of the result.

2) Applying dimension reduction process to columns L-Q and V-X.
   a) Begin with question 4, drop the whole row of data if one value is found missing before PCA, because usually there are not only one value is missing in that school. Dropping the whole row seems like to be the better solution in the situation that 4 of 6 values are missing when we want to get the correlation between them and also other analysis that need the overall performance and information of each school later.
   b) Using z-scoring to prevent skewness before PCA process and apply Kaiser criterion to find valid factors during the process.

Q1: What is the correlation between the number of applications and admissions to HSPHS?
correlation coefficient is 0.8018



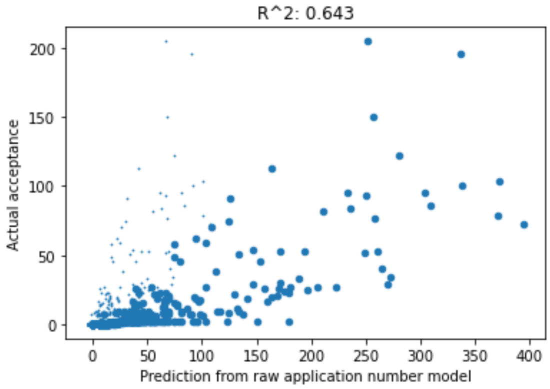
As the number of applications increase, the acceptances increase as well, showing a linear relationship. Therefore, I use Pearson correlation to correlate the number of applications and admissions and get the result of 0.8018.

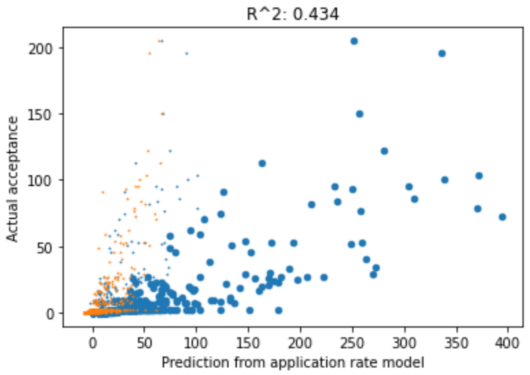
Q2: What is a better predictor of admission to HSPHS? Raw number of applications or application rate?

Raw number of applications is better.

R-square for model of raw number is 0.64

R-square for model of application rate is 0.434





When I do the linear regression model using raw number of application and application rate separately to predict acceptance, I found that R^2 of the raw number model is higher, which means that the proportion of the variance that can be explained by this model is higher. Therefore, the raw application number becomes a better predictor of the actual acceptance.

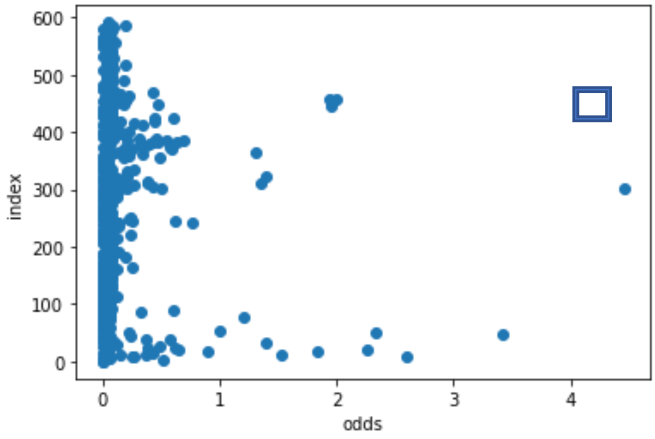
Q3: Which school has the best per student odds of sending someone to HSPHS?

THE CHRISTA MCAULIFFE SCHOOL\I.S. 187

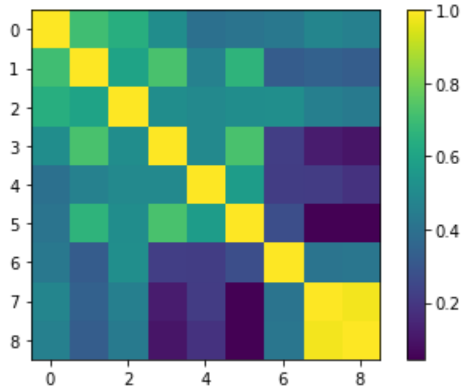
Probability of sending someone to HSPHS= $\frac{Acceptance}{Application}$

Odds of sending someone to HSPHS=$\frac{P(H1)}{P(H0)}$==$\frac{P(sending)}{1-P(sending)}$

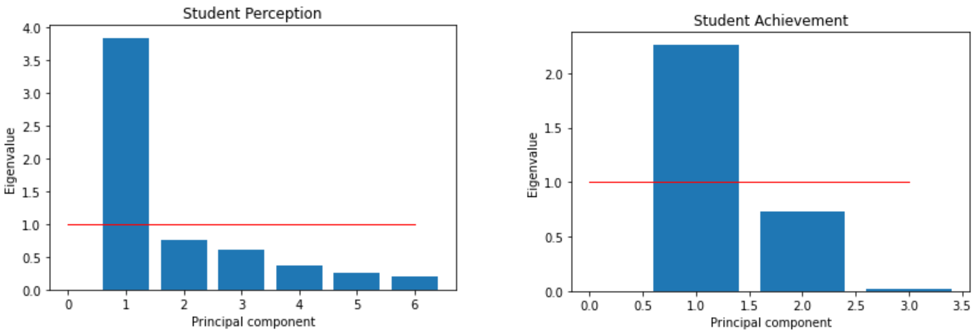
When I applied the above equation to the data, I got THE CHRISTA MCAULIFFE SCHOOL\I.S. 187 with the highest odds which is on index 303 of my cleaned data, with odds value of 4.46.



Q4: Is there a relationship between how students perceive their school and how the school performs on objective measures of achievement?

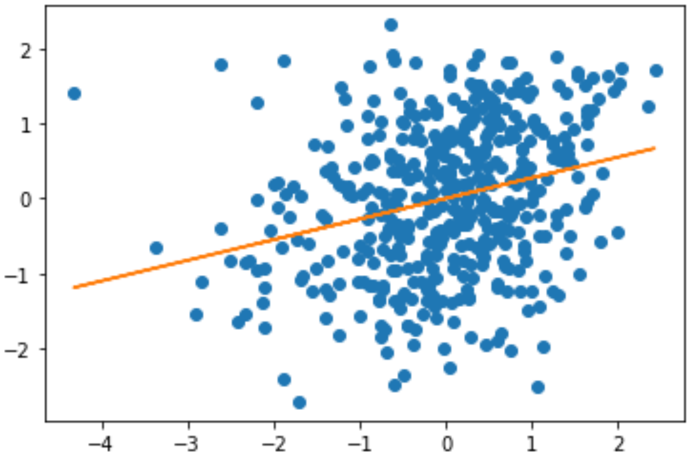


We first plot a correlation matrix and find out that most variables are correlated with each other within two groups: factor 0-5 are students perception; factor 6-8 are students achievement. Therefore we do PCA process in order to do dimension reduction.



As shown in graphs above, according to Kaiser method, both Students Perception and Students Achievement can be reduced to one main component, the one exceeding the red line, so a linear regression between these two main components can be done to check the relationship between two groups.

The linear regression process is done shown below that I conclude there is positive correlation between Students Perception and Students Achievement.

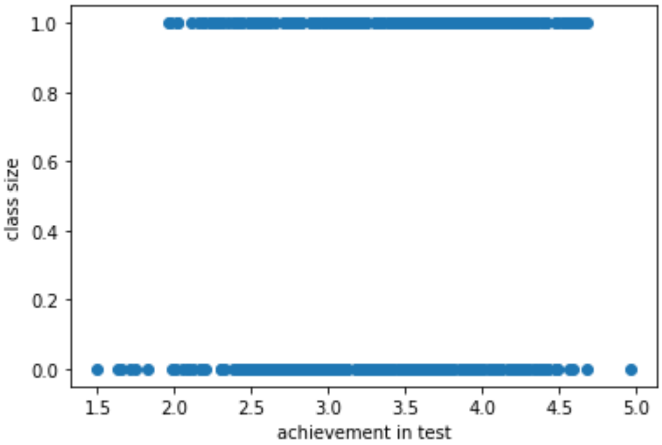


Q5: Test a hypothesis of your choice as to which kind of school performs differently than another kind either on some dependent measure.

H_0: the average class size will not affect the student achievement in standardized test

H_1: the average class size has effect on the student achievement in standardized test

I grouped the school by their average class size first, taking the median of the average class size. The school with average class size above or equal to median is considered as large class size (regarded as value 1)and vice versa (regarded as value 0), which gave me scatter plot shown below.



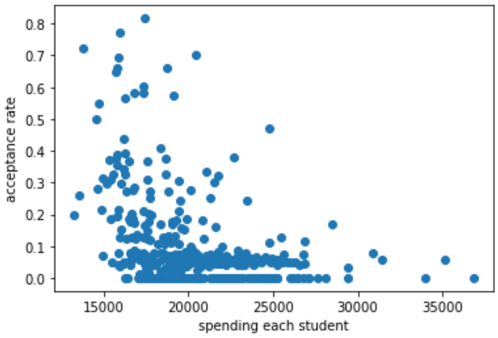
I then applied Mann Witney u-test to compare the test achievement in large class size and small class size, which gives me the result that:

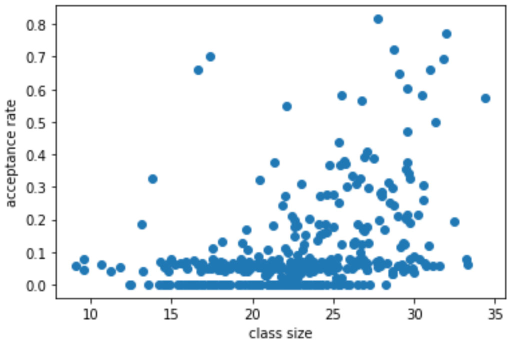
U = 21044.5   p=0.00125399788363813

In that way, p<0.025 and I successfully reject the null hypothesis so the average class size has effect on the student achievement in standardized test.

Q6: Is there any evidence that the availability of material resources impacts objective measures of achievement or admission to HSPHS?

I choose to check whether spending of each student or class size will affect the admission to HSPHS, so I plot scatter plot of spending and admission rate (people applied / people accepted) , and class size and admission rate first.



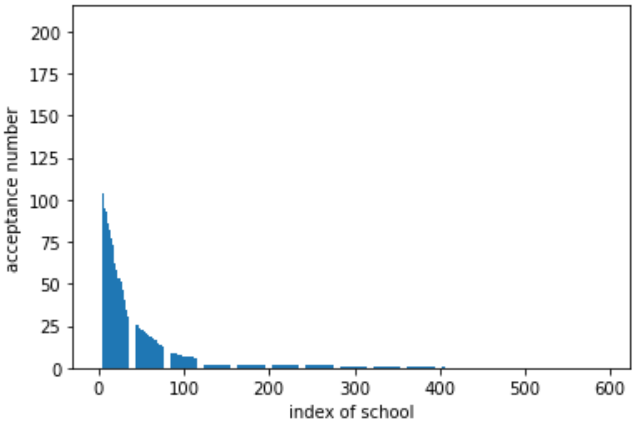


I further split class size and spending as large and small by taking the median of them similar to Q5 and assume that none of them will affect admission rate as H0 and do Kruskal-Wallis H-test tests and get the result that:

The test statistics= 759.1346300013142 p-value= 3.1526808802885166e-164

Since p-value is small enough(smaller than 0.01), we successfully reject H0, so that the availability of material resources, in regard as each student spending and class size have some effect on admission to HSPHS.

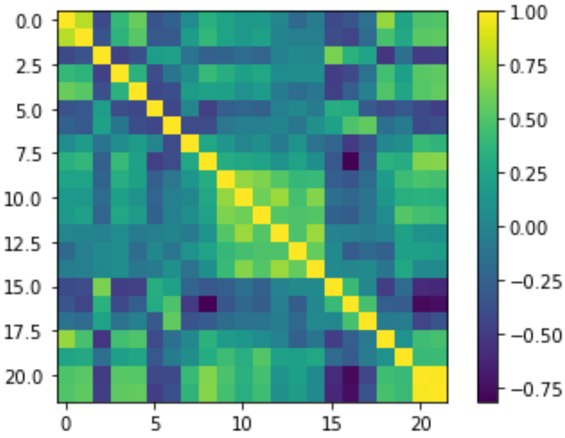
Q7: What proportion of schools accounts for 90% of all students accepted to HSPHS?



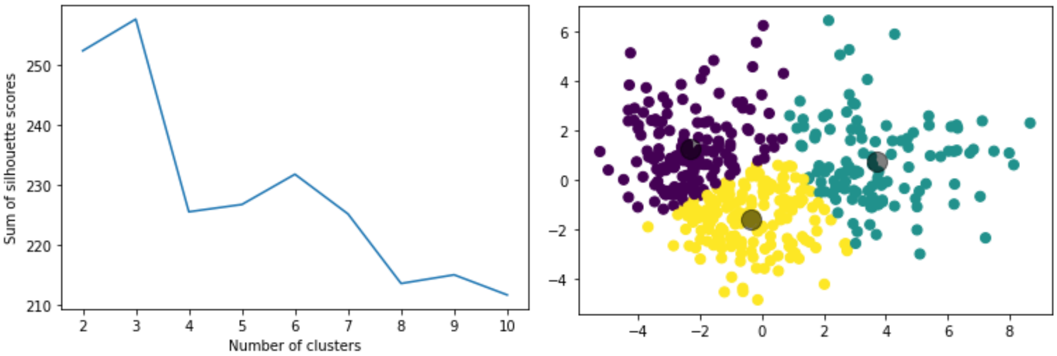
The scatterplot in descending order of acceptance number of each school is shown above, and I further get the result that 123 school accounts for 90% of all students accepted to HSPHS by calculating the 90% of the sum of the acceptance number and see until which index the number is reached.

Q8: 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?

First, plot a correlation matrix and apply PCA as Q4 mentioned.



As I applied silhouette scores, I found that 3 clusters are optional as suggested. Therefore, I build a cluster model with three centers. I regard these three centers as availability of materials (spending, effective class size, etc.), racial diversity and overall atmosphere of the school (collaborative teachers and community ties, etc.).



Q9: Write an overall summary of your findings – what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS?

the relative more abundant availability of materials resources plays a role in making their students have higher objective test score therefore more easily accepted by HSPHS. Also, though it seems obvious, when there are more applications to HSPHS, the school get more admission to HSPHS in turn.

Q10: Imagine that you are working for the New York City Department of Education as a data Scientist.. 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.

As mentioned, the overall atmosphere of the school that student perceive – including teachers, environment, leadership, community ties and trust- and diversity of races in school have both positive effect on improving the objective measures and acceptance rate of HSPHS. Therefore, improving these factors in whole will be highly possible to improve the overall level of the school. Also, it has some evident that spending is an influential factor as well. In that way, certain degree of economic support may help the school to improve as well. Though we know from question above that class size of the school can be determinant to objective achievement, the direction of the class and exact appropriate size need be further investigated, so maybe keep the class size in median is the current better option for school.