Kieren Singh Gill

19 May 2021

Big Data Project

I used both NumPy and pandas to work on this project. To read the .csv file, I used pandas and converted the dataframe to a NumPy array. To reduce the dimensionality of the dataset, I performed a Principal Component Analysis (PCA) when needed. Before conducting calculations or any statistical test, I first cleaned my data of NaNs. I did so by combining the data I need for the calculation/test into a 2D Numpy array, and removing the rows where there were NaNs.

```
temp = np.vstack((applications,schoolSize))
temp = np.transpose(temp)
temp = temp[~np.isnan(temp).any(axis=1)]
```

Figure 1: *Cleaning NaNs from 2D NumPy array "temp".*

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

I spliced the dataset by column to obtain the data for applications and acceptances. Then, I used NumPy to obtain a correlation coefficient of **0.802**, and graphed a scatter plot of the data as shown in Figure 2.

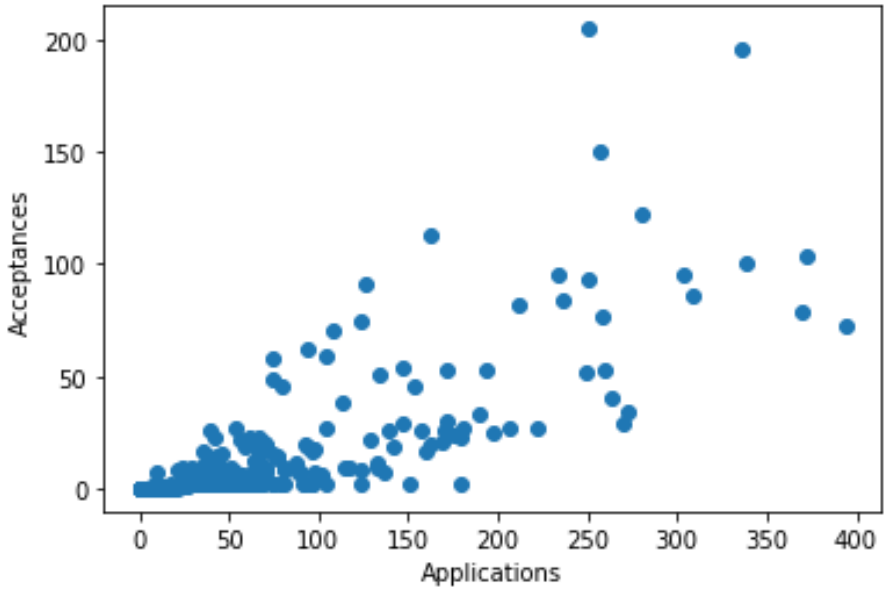

Figure 2: *Scatter plot illustrating the correlation between applications and acceptances to HSPHS*

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

I calculated application rate using the following formula:

$$Application\ Rate = \frac{Applications}{School\ Size}$$

Following this, I used NumPy to calculate the correlation between application rate and acceptances, and obtained a correlation coefficient of **0.659**. This is lower than the value correlation coefficient of **0.802** that was obtained earlier when calculating the coefficient between the raw number of applications and acceptances. Using these values, I would claim that raw number of applications is a better predictor of acceptances, because it has a stronger correlation.

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

I calculated the per student odds by first calculating the acceptance rate per school:

$$Acceptance\ Rate = \frac{Acceptances}{School\ Size}$$

It is important to note why I did not divide acceptances by applications – to determine the best *per student* odds, the question asks for the school from where a student is most likely to be accepted into a HSPHS. This involves the entire student body, not only the number of students who applied to HSPHS. To convert the acceptance rate to odds, I used the following formula:

$$Odds = \frac{Acceptance\ Rate}{1 - Acceptance\ Rate}$$

Then, I used NumPy to sort the schools in order of acceptance rate, and found that **THE CHRISTA MCAULIFFE SCHOOL\I.S. 187** had the best *per student* odds of **0.307**.

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)?

I started off by doing two separate PCAs – one for how students perceive their school and one for how the school performs on objective measures of achievements. Let's first take a look at the eigenvalues of the principal components from both PCAs:

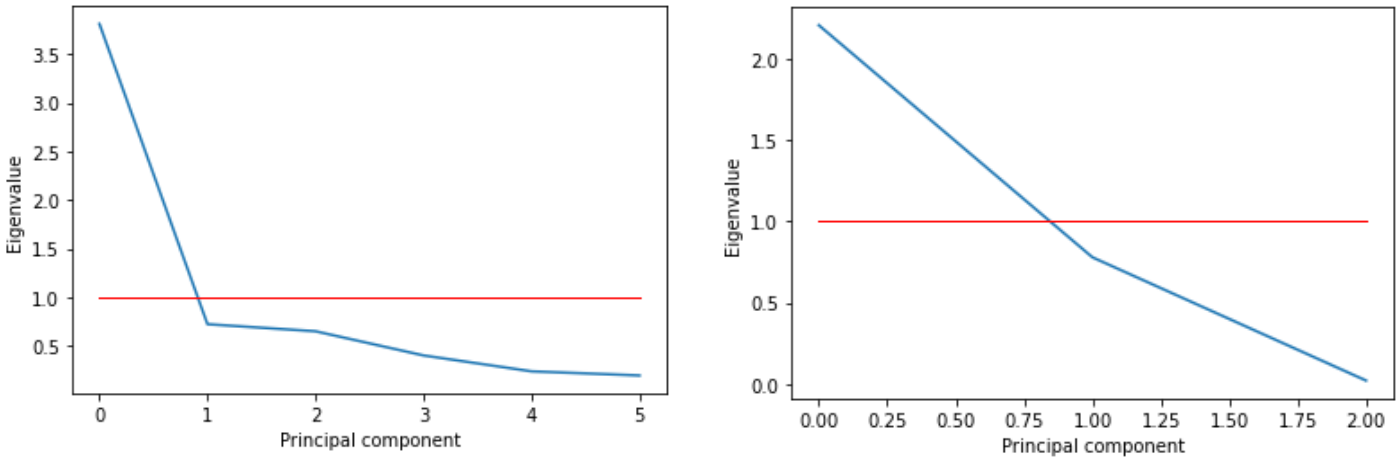

Figure 3: *Eigenvalues of principal components that encompass how students perceive their school (left), eigenvalues of principal components that encompass how the school performs on objective measures of achievements (right).*

To pick the number of factors I interpret meaningfully, I used the Kaiser criterion, which means I will only be looking at the principal components with an eigenvalue greater than 1. This means I have only one component to analyze from each PCA. Here is the breakdown of both components:

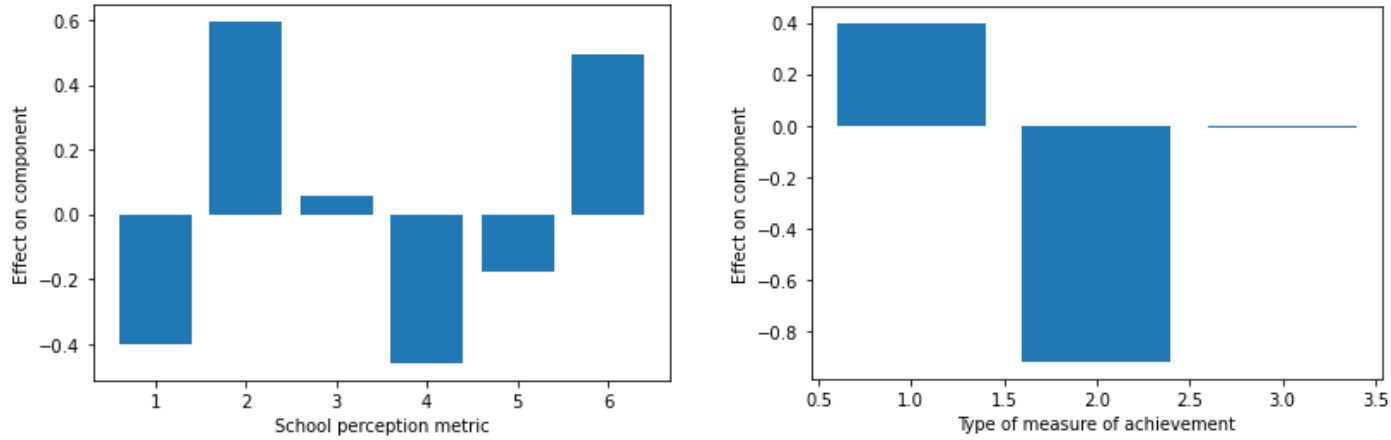

Figure 4: *Component breakdown for the school perception metrics (left), component breakdown for the type of measures of achievement (right).*

While it is easy to notice that the component breakdown for the type of measure of achievement is heavily weighted towards the second measure (which corresponds to “Reading Scores”), it is not as clear cut for the school perception metrics, as many of them seem to carry weight towards this component. Hence, I opted to use a different criterion for the school perception metrics, and decided to take into account the components that account for 90% of the variance. This meant I had to look at the breakdown of 4 components:

	0	1	2	3	4	5
0	−0.399016	−0.447067	−0.385057	−0.432745	−0.349307	−0.428212
1	0.593196	0.162998	0.364934	−0.221362	−0.535691	−0.390394
2	0.0587881	−0.36458	0.50377	−0.426532	0.62176	−0.20329
3	−0.45975	−0.221619	0.640996	0.0735882	−0.436094	0.364754
4	−0.178553	−0.107758	0.178215	0.701554	0.0770108	−0.653175
5	0.492136	−0.761523	−0.148776	0.290186	−0.0912463	0.25143

Figure 5: *Component breakdown for the school perception metrics. Columns represent the components, and rows represent the school perception metrics. I looked at the first 4 columns only because they accounted for 90% of the total variance.*

Based on figure 5, I decided to look at three of the school perception metrics – metric 0, metric 1, and metric 5 (which corresponds to rigorous instruction, collaborative teachers, and trust). Once I determined these three metrics and the one measure of achievement, I ran a multiple regression on NumPy to obtain the variance. Because variance is r^2, I took the square root of that to find r, which was **0.484**. This can be interpreted to mean that there is a positive moderate correlation between how students perceive their school and how the school performs on objective measures of assessment.

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 null hypothesis is that school size has no effect on reading scores. The alternate hypothesis of my choice is that school size has an effect on reading scores. To investigate this hypothesis, I conducted a Mann–Whitney U test. I chose this test because the data doesn't reduce itself to means, the data isn't categorical, there are 2 groups being compared and I am comparing their medians. To run the Mann-Whitney U test, I divided schools into large and small based on their size. I took the median school size (539), and declared that every school that had 539 students and below was small, and every school that had more than 539 students was a big school. I then separated the reading scores for the large and small schools. The large schools had a median of

0.53, whereas the small schools had a reading score of 0.34. Running the Mann-Whitney U test gave me a p-value of 7.934557472038052e-22, which is a lot smaller than p=0.05. This means that I can reject the null hypothesis and accept my alternate hypothesis because my result is statistically significant. Students appear to perform better on reading scores when school size is larger.

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?

To answer this question, I carried out two ANOVA tests. My first ANOVA examined the effects of per student spending, class size, and their interaction effect on an objective measure of achievement. I chose reading score as my measure of achievement because it had the highest weightage of the first component that was determined by the PCA is question 4. I separated per student spending into high spending and low spending, and I separated class size into large classes and small classes. I did this by obtaining the median values of each group and splitting them along the median.

	sum_sq	df	F	PR(>F)
PupilSpending	1.603737	1.0	65.599547	4.815640e-15
ClassSize	1.701512	1.0	69.598939	8.179384e-16
PupilSpending:ClassSize	0.083391	1.0	3.411028	6.539180e-02
Residual	11.465822	469.0	NaN	NaN

Figure 6: *ANOVA that measures the effects of pupil spending, class size, and their interaction effects on reading scores.*

Based on the p-values, I interpret that both per student spending and class size impacts the reading score, as both p values are below p=0.05. My second ANOVA examined the effects of per student spending, class size, and their interaction effect on acceptance rate.

	sum_sq	df	F	PR(>F)
PPS	0.008684	1.0	14.446412	0.000163
ClassSize	0.009589	1.0	15.951651	0.000075
PPS:ClassSize	0.004152	1.0	6.907144	0.008868
Residual	0.280728	467.0	NaN	NaN

Figure 7: *ANOVA that measures the effects of pupil spending, class size, and their interaction effects on acceptance rate.*

Based on the p-values, it appears that both per student spending and class size impacts the acceptance rate, as both p values are below p=0.05. However, the interaction effect is also significant, so this means I needed to examine the interaction plot to see how both factors interact with each other.

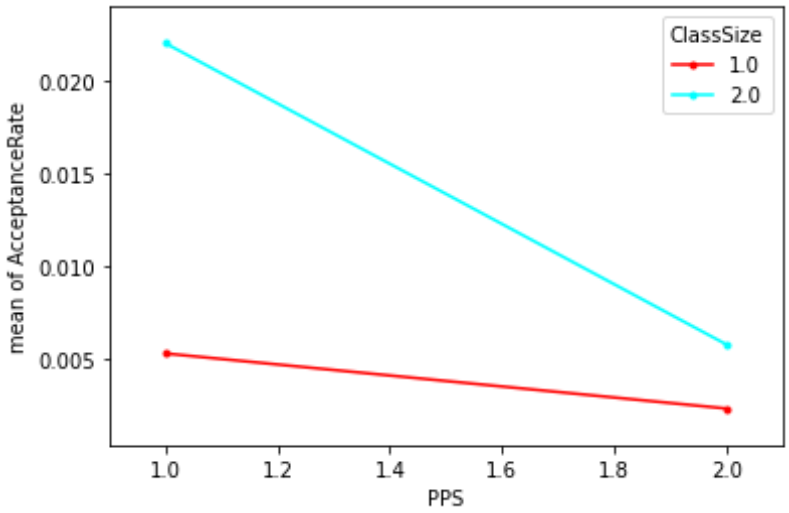

Figure 8: *Interaction plot between per student spending (PPS) and class size, and their effects on acceptance rate.*

Based on the interaction plot, we can see that the relationship between pupil spending and acceptance rate is dependent on class size. If class size is bigger, lower per pupil spending will lead to a more significant increase in acceptance rate as compared to if class size is smaller.

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

The sum of all acceptances is 4461. 90% of that is 4014.9. To calculate the proportion of schools that account for 90% of acceptances, I sorted the schools in descending order of acceptances and used a for loop as so:

```
sortedAccept = np.sort(acceptances)[::-1]
total = 0
counter = 0
for i in sortedAccept:
    total += i
    counter += 1
    if total >= 4014.9:
        break
```

Figure 9: *Code used to count the number of schools that accounts for 90% of acceptances.*

Using this code, I found that 123 out of 594 schools account for 90% of accepted students – this amounts to **20.7%** of schools. The sorted distribution of these schools will be illustrated in the following figure:

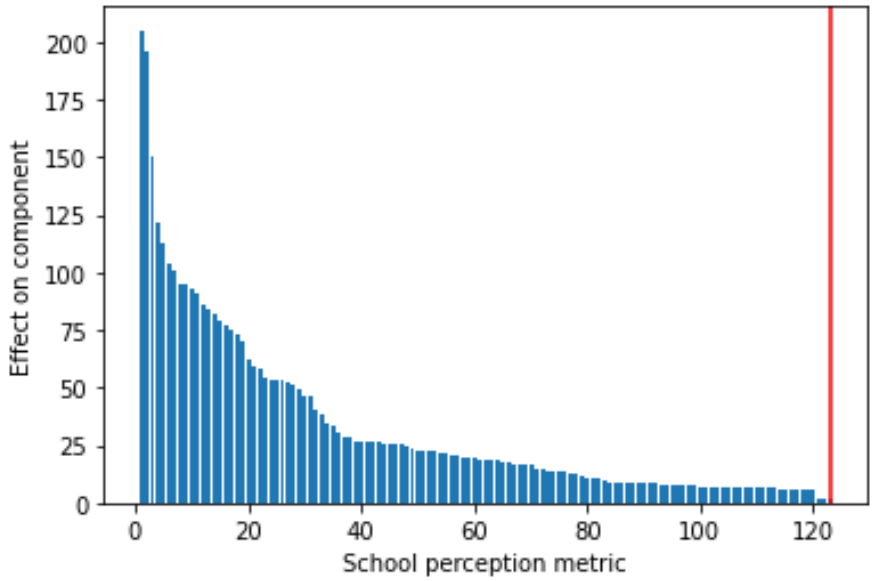

Figure 10: *The 123 schools that account for 90% of the total acceptances.*

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?

For both parts a) and b), I decided to build a prediction model using multiple regression. To do so, I first did a PCA to deal with multi-collinearity.

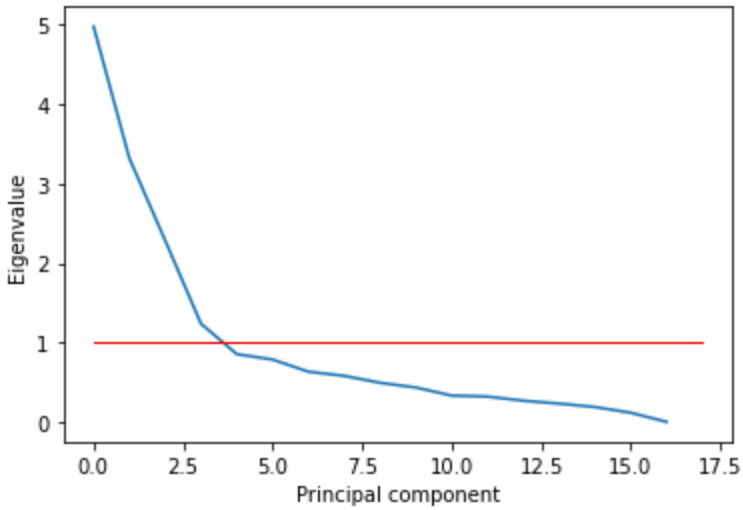

Figure 11: : *Eigenvalues of principal components that encompass all school characteristics.*

I did two PCAs, one for a) and one for b), and both yielded nearly identical graphs and results. To pick the number of factors I interpret meaningfully, I used the Kaiser criterion again, which means I will be looking at 4 components.

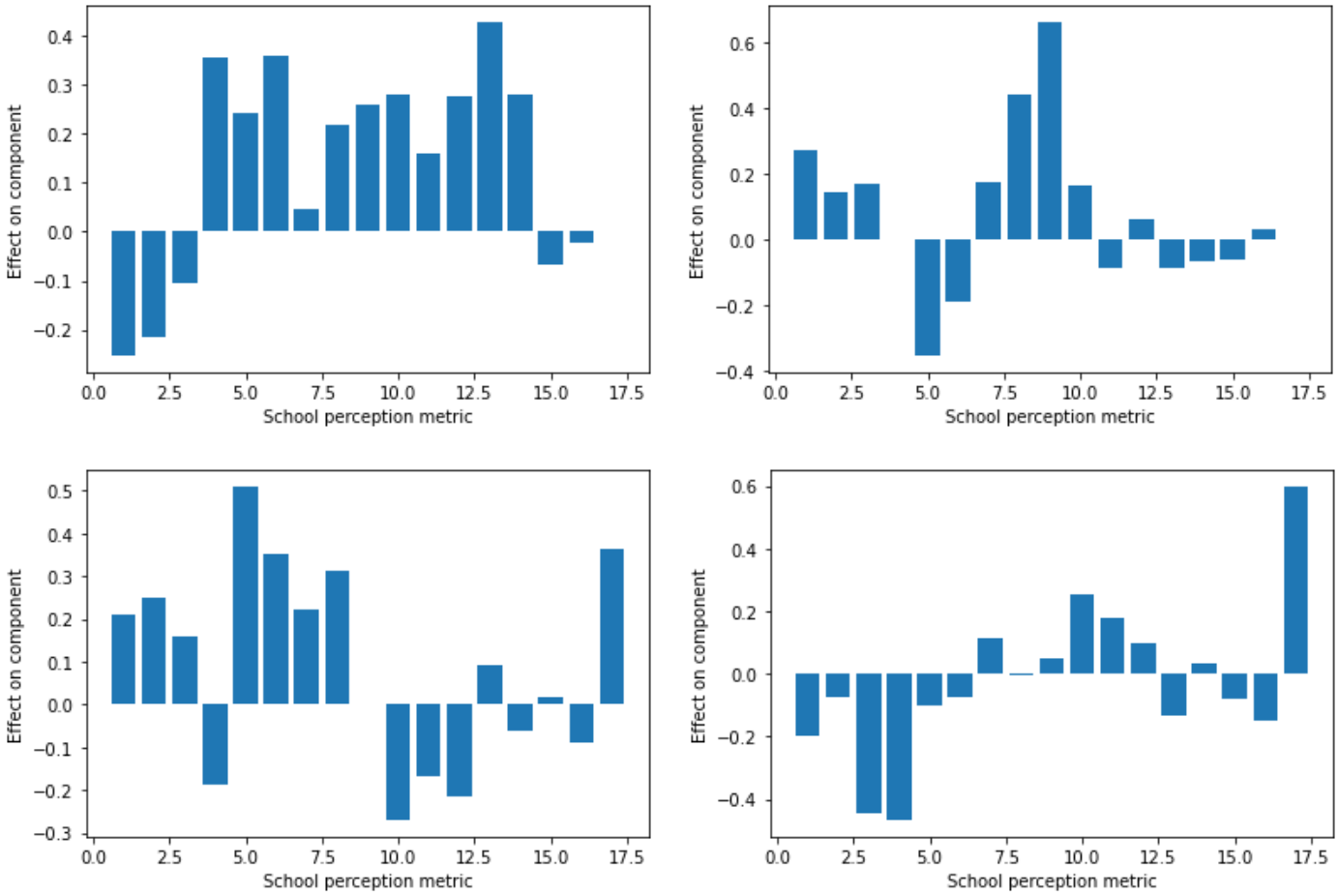

Figure 12: *Component breakdown for all school characteristics. Component 1 (top left), Component 2 (top right), Component 3 (bottom left), Component 4 (bottom right).*

Before running the PCA, I standardized the data by z-scoring the data. After the PCA, I did multiple regressions with four school characteristics (black_percent, multiple_percent, rigorous_instruction, and trust). Because the data was standardized, I could determine which characteristic carried the most weight based on their associated beta value. I did separate multiple regressions for both part a) and b), to investigate the effects of the school characteristics on acceptance rate and on an objective measure of achievement. I chose reading scores as my measure of achievement with the same reason that I explained earlier in question 4.

	0
0	-0.000243719
1	0.00264695
2	0.00711442
3	-0.0021907

	0
0	-0.002288
1	0.0269278
2	0.114014
3	-0.0306697

Figure 13: *Beta values for mutliple regression to investigate the effects of the school characteristics on acceptance rate (left) and on reading scores (right).*

Based on these values, I found that the third feature (which corresponds to **rigorous instruction**) played the largest role in increasing both acceptance rates and reading scores.

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

The correlation between applications and admissions is 0.802, but this information is not necessarily relevant on its own. It seems intuitive that the more applications one sends in, the higher the chance of receiving an acceptance – hence, this isn't all too insightful. Acceptance rate also appears to be dependent on class size (larger class size increases acceptance rate), and per pupil spending (the less spent on a pupil the higher the acceptance rate). However, the interaction between these two factors is also significant, and we can see that the relationship between pupil spending and acceptance rate is dependent on class size. If class size is bigger, lower per pupil spending will lead to a more significant increase in acceptance rate as compared to if class size is smaller. Both class size and per pupil spending are relevant characteristics, with class size being more important. However, the most important characteristic appears to be rigorous instruction, because it played the largest role in increasing both acceptance rates and reading scores (according to the PCA and multiple regression done is Q8) .

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.

When looking to improve schools to send more students to HSPS, we need to look at the bigger picture. Although 90% of acceptances come from 20.7% of schools, there is a high chance that this is because a majority of acceptances come from larger schools. To try and reduce the disparity between acceptances in each school, we need to focus on schools with the lowest *per student* odds, and direct the actionable recommendations we list below to these schools first.

I think there are similar characteristics to look at when looking to improve schools for criteria a) and b). The most important characteristic to look at would be the school's rigor of instruction. To increase the rigor of instruction, schools should look at developing a more challenging curriculum, and possibly re-training teachers. We should also increase class size and decrease per pupil spending, as it appears to have a positive impact on increasing acceptances.