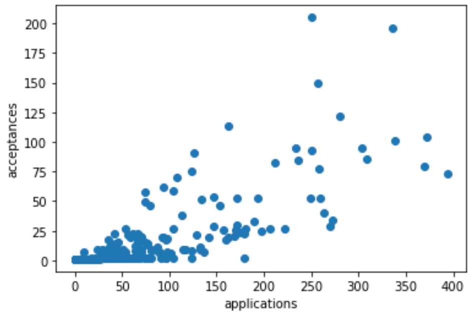
Data Cleaning

I removed the rows with missing data because there did not seem to be too many missing. For charter schools, the columns for ‘per student spending’ and ‘average class size’ are systematically missing, so charter schools were only removed if they had other data missing. After cleaning the data, there were 514 observations left, with 449 public schools and 65 charter schools.

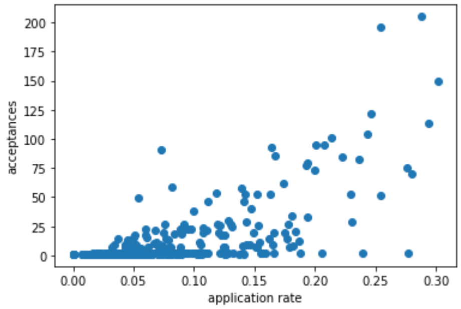
Q1: What is the correlation between the number of applications and admissions to HSPHS?

The Pearson correlation is 0.81. From the plot, it isn’t a very linear relationship, as both factors are extremely skewed. So, I calculated the Spearman correlation which is 0.94, suggesting they are strongly associated with each other.



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

The Pearson correlation between application rate and acceptances is 0.68 and the Spearman correlation is 0.78. It seems that the raw number of applications is a better predictor.



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

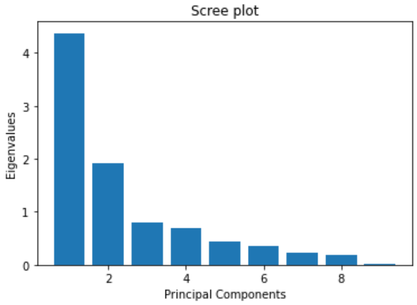
For this question, we might be interested in two things: the proportion of students in the entire school that get into HSPHS and the proportion of students that get in out of the ones that applied. It turns out that for both metrics the Christa Mcauliffe School had the best numbers. The proportion of students out of the entire school that got in was 23.5% and the proportion of students that got accepted out of the ones who applied was 81.7%.

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

I performed a PCA on the nine variables:

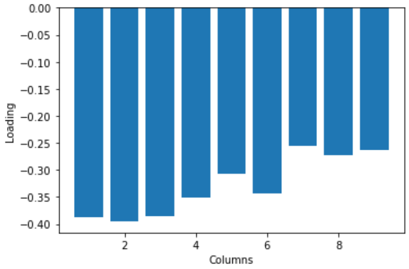
'rigorous_instruction', 'collaborative_teachers', 'supportive_environment', 'effective_school_leadership', 'strong_family_community_ties', 'trust', 'student_achievement', 'reading_scores_exceed', 'math_scores_exceed'

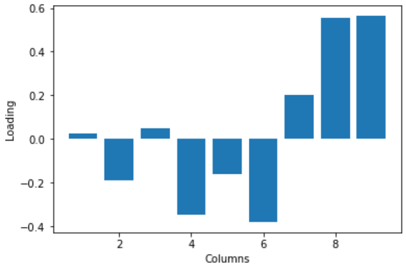
This was the scree plot for the PCA



Using the Kaiser criterion, we end up with 2 principal components.

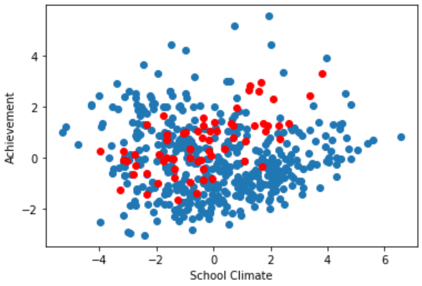
These are the loadings for the first and second component respectively





It seems that the first component can be classified as overall school climate and the second one as objective measures of achievement.

Using the new transformed data, we plot the two components.



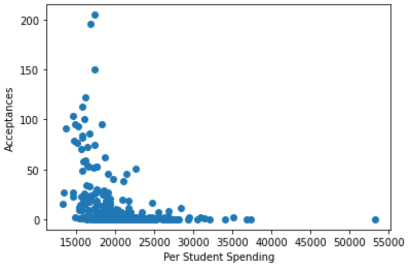
I wanted to see if there was a difference in this relationship between public schools and charter schools, so the charter schools are shown in red dots. The correlation is pretty much zero, so there does not seem to be a relationship between school climate and achievement. There does, however, seem to be a positive relationship for the charter schools.

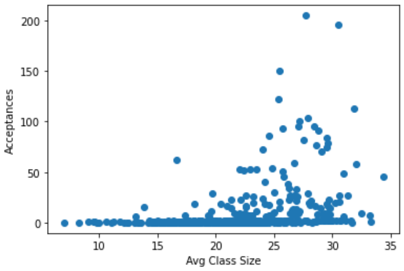
Q5: Test a hypothesis of your choice

I compared the level of achievement from the achievement component from the PCA between charter schools and public schools. I did a t-test with the null hypothesis being that the two kinds of schools have no difference in achievement. The calculated t-statistic was t=3.58 and the p-value was p=0.0004. At a 1% significance level we reject the null hypothesis and conclude that there is a difference in achievement with charter schools performing better than public schools.

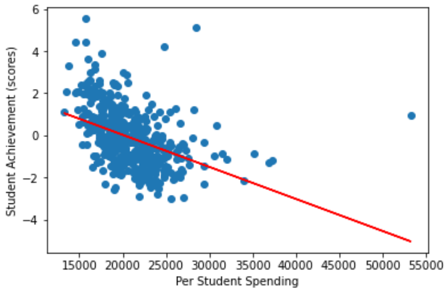
Q6: Is there any evidence that the availability of material resources (e.g. per student spending or class size) impacts objective measures of achievement or admission to HSPHS?

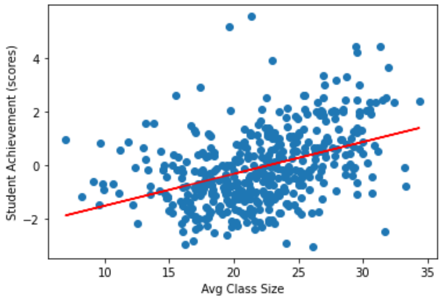
For this analysis, we could only look at public schools because there is no data on per student spending and average class size for charter schools. The first two graphs show the scatterplot of per student spending on acceptances and average class size on acceptances





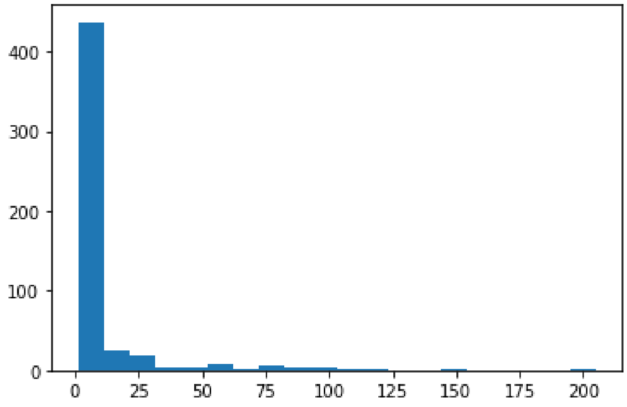
The relationship in these two graphs is counter-intuitive and also highly non-linear. We cannot make a conclusion about causality, but from a policy standpoint, there is certainly something peculiar going on here with these relationships that can be investigated further.
The next two graphs show the scatterplot of per student spending on achievement and average class size on achievement, with achievement being the achievement component we made from the PCA





There is a much weaker association here for both plots, and once again, we cannot conclude causality.

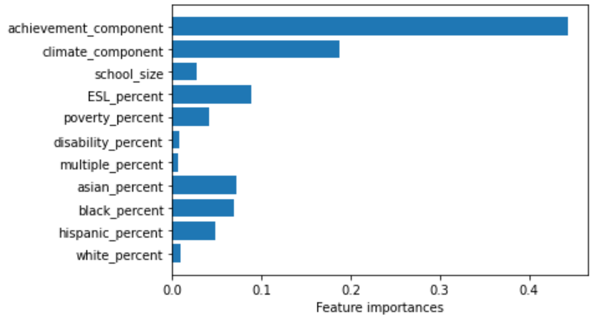
Q7: What proportion of schools accounts for 90% of all students accepted to HSPHS?
From 514 schools there were a total of 4282 acceptances. 103 out of the 514 schools account for 90% of the acceptances to HSPHS.
Here is a histogram of the acceptances for all the schools. It is highly skewed with most schools having very few acceptances.



Q8: Build a model of your choice

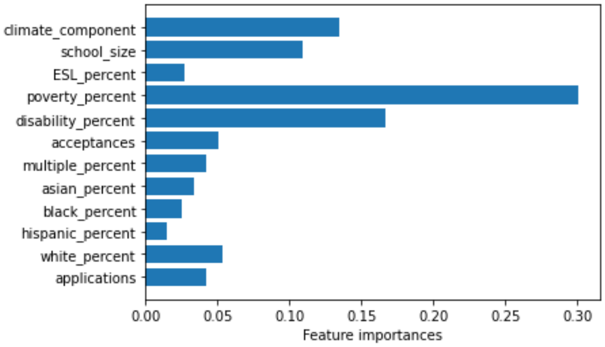
I built a decision tree regressor model to predict the outcomes of student achievement and acceptance rates for schools.

In the first model for predicting acceptance rates, I trained the model using the factors: 'white_percent','hispanic_percent','black_percent','asian_percent','multiple_percent', 'disability_percent','poverty_percent','ESL_percent','school_size', along with the achievement component and school climate component from the PCA. Per student spending and average class size were excluded because the data are missing charter schools and they did not have a strong association with acceptances. The training and test sets were split 80/20 and the resulting RMSE was 0.12. I used the built-in feature importance function to extract the feature importances.



It seems that student achievement and school climate are the most influential in predicting acceptance rates for each school.

In the second model for predicting student achievement, I trained the model using the factors: 'applications','white_percent','hispanic_percent','black_percent','asian_percent','multiple_percent', 'acceptances','disability_percent','poverty_percent','ESL_percent','school_size', along with the achievement component and school climate component from the PCA. Once again, per student spending and average class size were excluded. The training and test sets were split 80/20 and the resulting RMSE was 1.24 in terms of the achievement component.



Poverty percentages and disability percentages were the most influential, followed by school climate and school size.

Q9: Overall summary

Based on the decision tree model, the most relevant school characteristic in determining acceptances to HSPHS are student achievement and school climate, which isn't too surprising. Other relevant factors include, disabilities, language barriers, and ethnic makeup of schools which could imply the need to further evaluate the funding and resource allocation in these areas to remedy the large imbalances in applications and acceptances.

Q10: **What actionable recommendations would you make?**

Based on the predictive models, disability and poverty were the most important features for predicting student achievement, which was, in turn, the most important for determining acceptance rates. Given that, I would suggest further investigating how disability and poverty specifically affect student achievement beyond just the percentage of disabled or impoverished students in schools. I would also suggest gathering more data on the distribution of different types of disabilities and perhaps more levels of socioeconomic status in order to better tackle these problems.