Wiley Wu

Pascal Wallisch

Intro to Data Science

19 May 2021

Final Project

I. Introduction: Handling dimension reduction, data cleaning and data transformation

   a. Dimension Reduction

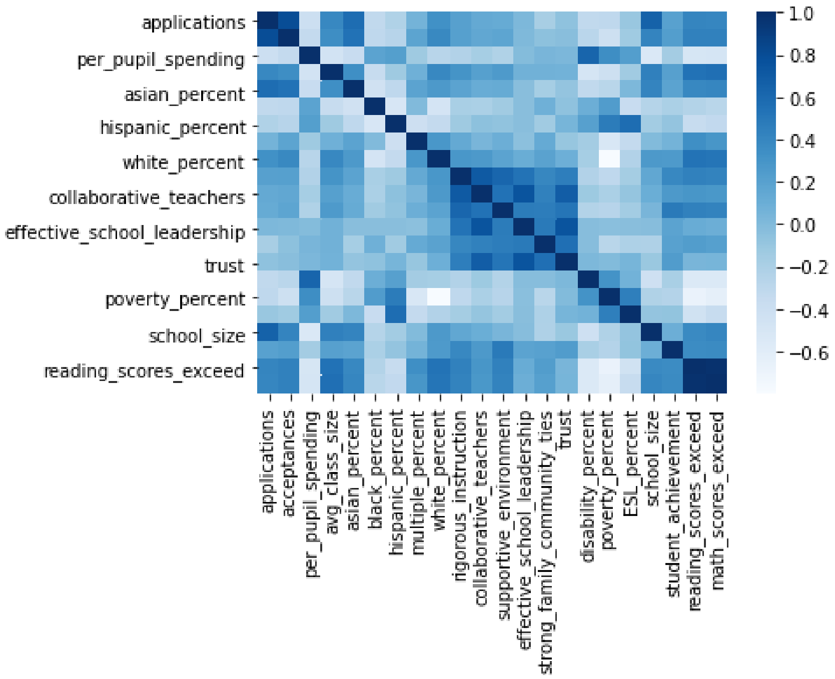
      i. Dimension reduction started with creating a new dataset with the necessary variables. After that, the data was z-scored, and the PCA was then fit with the transformed data. Eigenvalues were then graphed and analyzed, and I chose to use the Kaiser criterion and keep eigenvalues greater than 1.

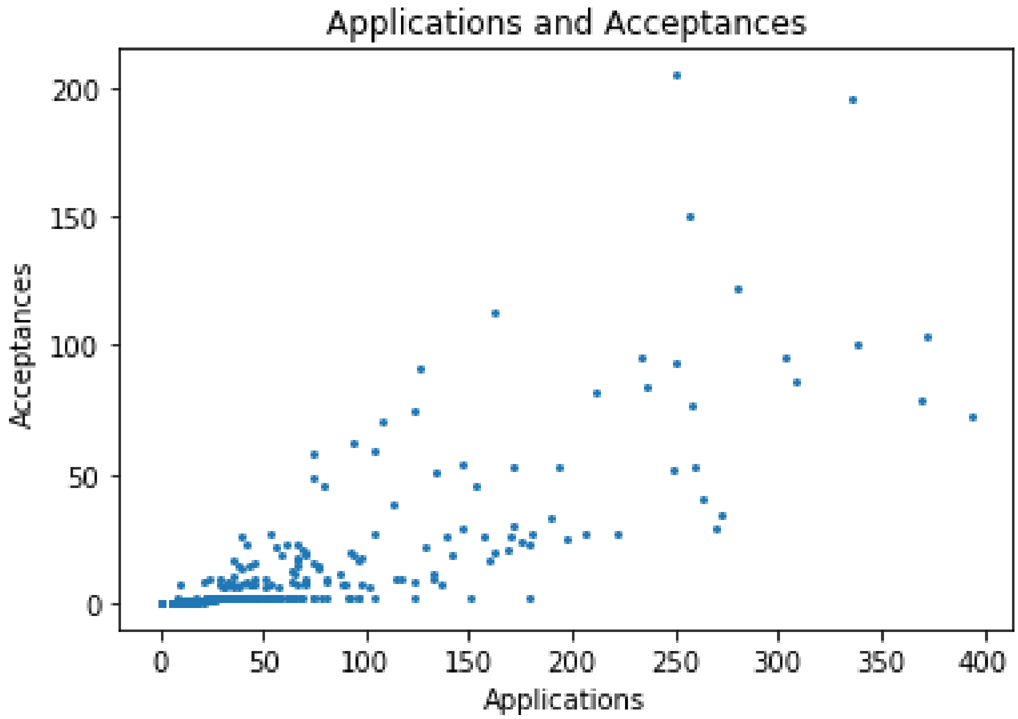
   b. Data Cleaning

      i. The entire dataset was set to a variable at the beginning called *data*. For each question, a new dataset was created from *data*, titled dataQuestion#, where the Question # was swapped out for the question number accordingly. Only the variables necessary would be taken from *data* and would be cleaned at this step to minimize the amount of data lost.

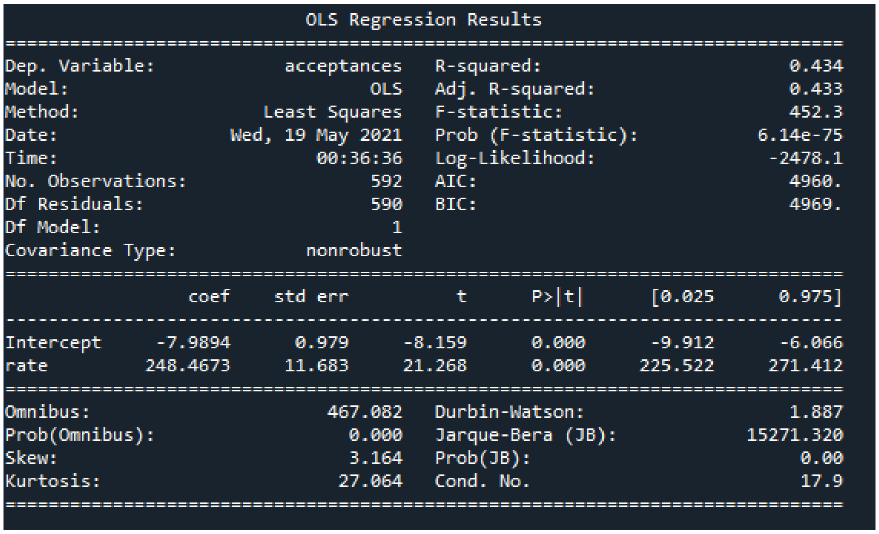
   c. Data transformation

      i. Data transformation was z-scored for use in PCA, and then rotated to graph the old data in the new coordinate field.

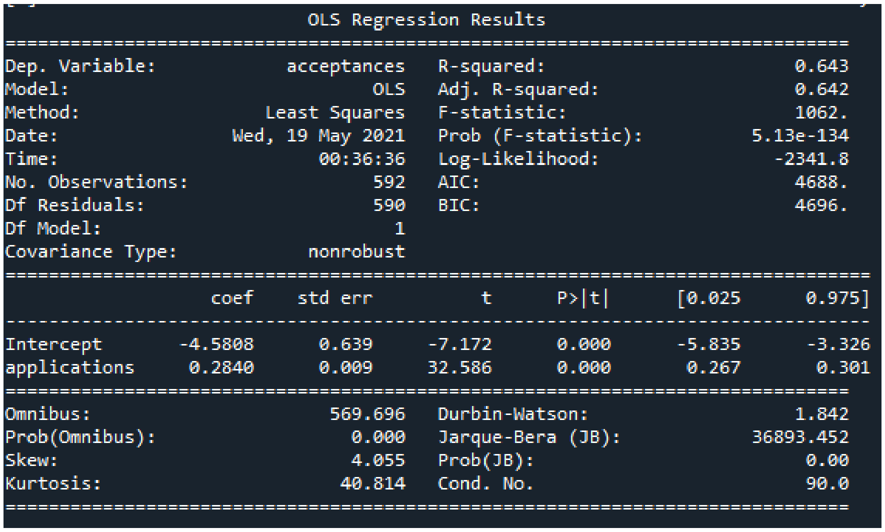




1. The correlation between number of applications and admissions is 0.801727. It was obtained by using the panda’s correlation function to get the correlation table. No cleaning was applied on the dataset, as checking the columns of applications and admissions of null values returned false statements. It makes sense when controlling for all other variables, the more applications a school has for HSPHS, the greater the chance a student is admitted.

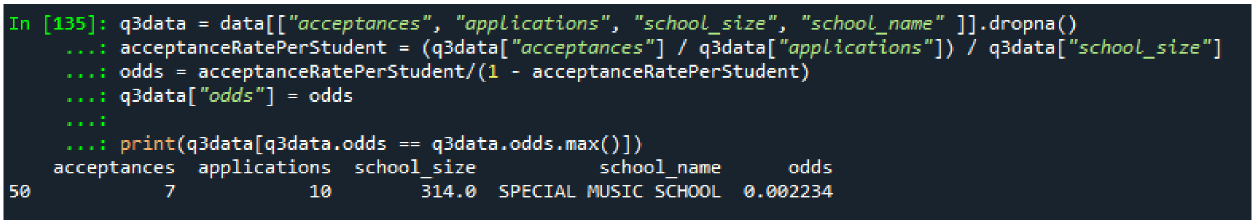


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            acceptances   R-squared:                       0.434
Model:                            OLS   Adj. R-squared:                  0.433
Method:                 Least Squares   F-statistic:                     452.3
Date:                Wed, 19 May 2021   Prob (F-statistic):           6.14e-75
Time:                        00:36:36   Log-Likelihood:                -2478.1
No. Observations:                 592   AIC:                             4960.
Df Residuals:                     590   BIC:                             4969.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -7.9894      0.979     -8.159      0.000      -9.912      -6.066
rate         248.4673     11.683     21.268      0.000     225.522     271.412
==============================================================================
Omnibus:                      467.082   Durbin-Watson:                   1.887
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15271.320
Skew:                           3.164   Prob(JB):                         0.00
Kurtosis:                      27.064   Cond. No.                         17.9
==============================================================================
```



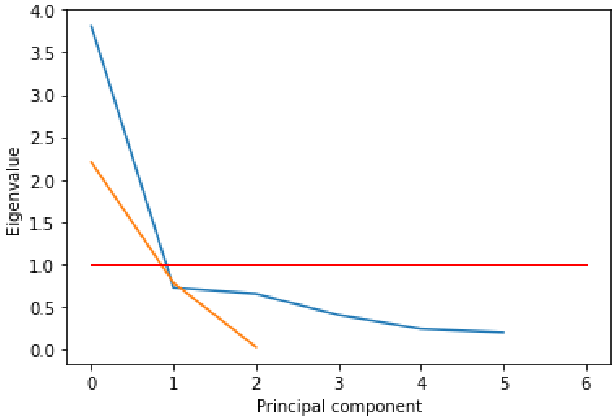
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            acceptances   R-squared:                       0.643
Model:                            OLS   Adj. R-squared:                  0.642
Method:                 Least Squares   F-statistic:                     1062.
Date:                Wed, 19 May 2021   Prob (F-statistic):          5.13e-134
Time:                        00:36:36   Log-Likelihood:                -2341.8
No. Observations:                 592   AIC:                             4688.
Df Residuals:                     590   BIC:                             4696.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       -4.5808      0.639     -7.172      0.000      -5.835      -3.326
applications     0.2840      0.009     32.586      0.000       0.267       0.301
==============================================================================
Omnibus:                      569.696   Durbin-Watson:                   1.842
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            36893.452
Skew:                           4.055   Prob(JB):                         0.00
Kurtosis:                      40.814   Cond. No.                         90.0
==============================================================================
```

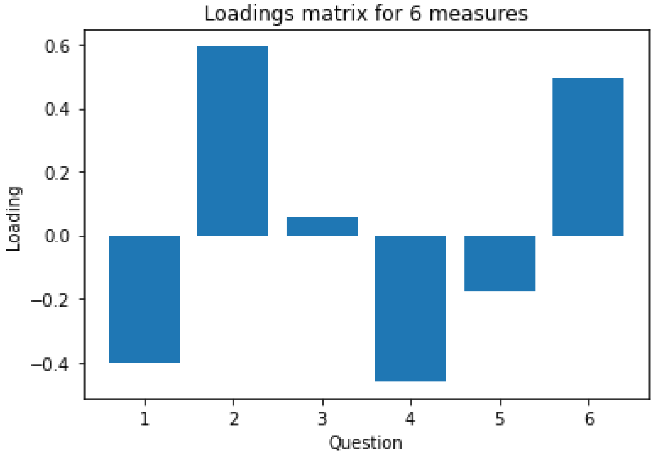
2. The raw number of applications is a better predictor of admission to HSPHS over application rate. Running a linear regression on the raw number of applications and acceptances yielded a R^2 value of 0.643. A linear regression applied to the application rate and acceptances yielded a R^2 value of 0.433. The difference could possibly be attributed to the calculation for the odds, as for schools without large numbers of applicants, the fraction of the odds could jump significantly, leading the data to be more volatile and therefore worse at explaining the model.

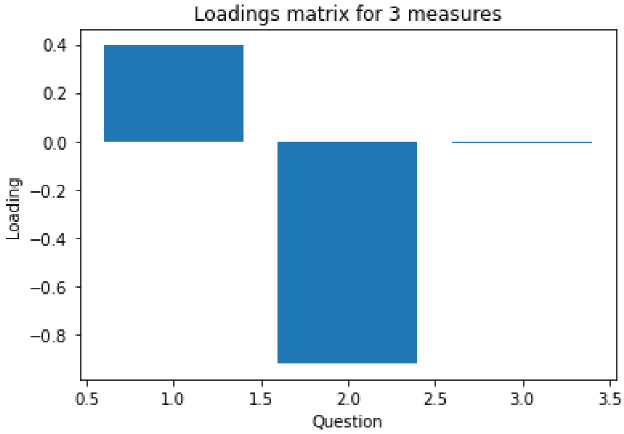


```
In [135]: q3data = data[["acceptances", "applications", "school_size", "school_name" ]].dropna()
     ...: acceptanceRatePerStudent = (q3data["acceptances"] / q3data["applications"]) / q3data["school_size"]
     ...: odds = acceptanceRatePerStudent/(1 - acceptanceRatePerStudent)
     ...: q3data["odds"] = odds
     ...:
     ...: print(q3data[q3data.odds == q3data.odds.max()])
    acceptances  applications  school_size           school_name      odds
50            7            10        314.0  SPECIAL MUSIC SCHOOL  0.002234
```

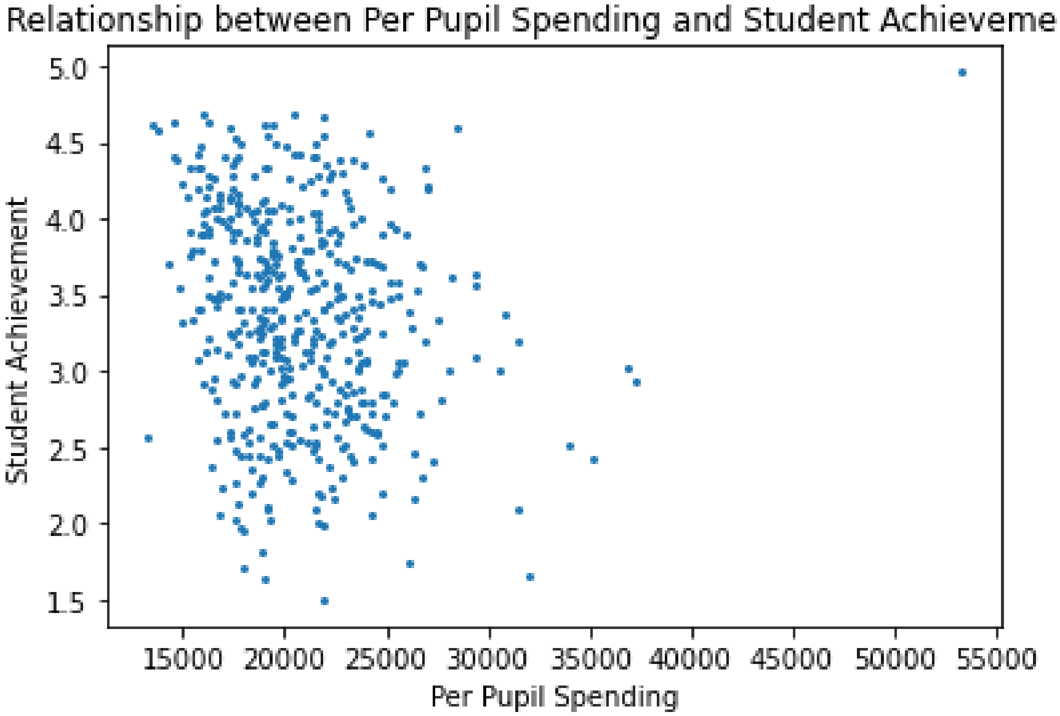
3. The school with the best per student odds of getting into HSPHS was Special Music School with a calculated odds value of 0.002234. Odds was calculated by dividing acceptances by applications, and then dividing by school size to get the per student odds.





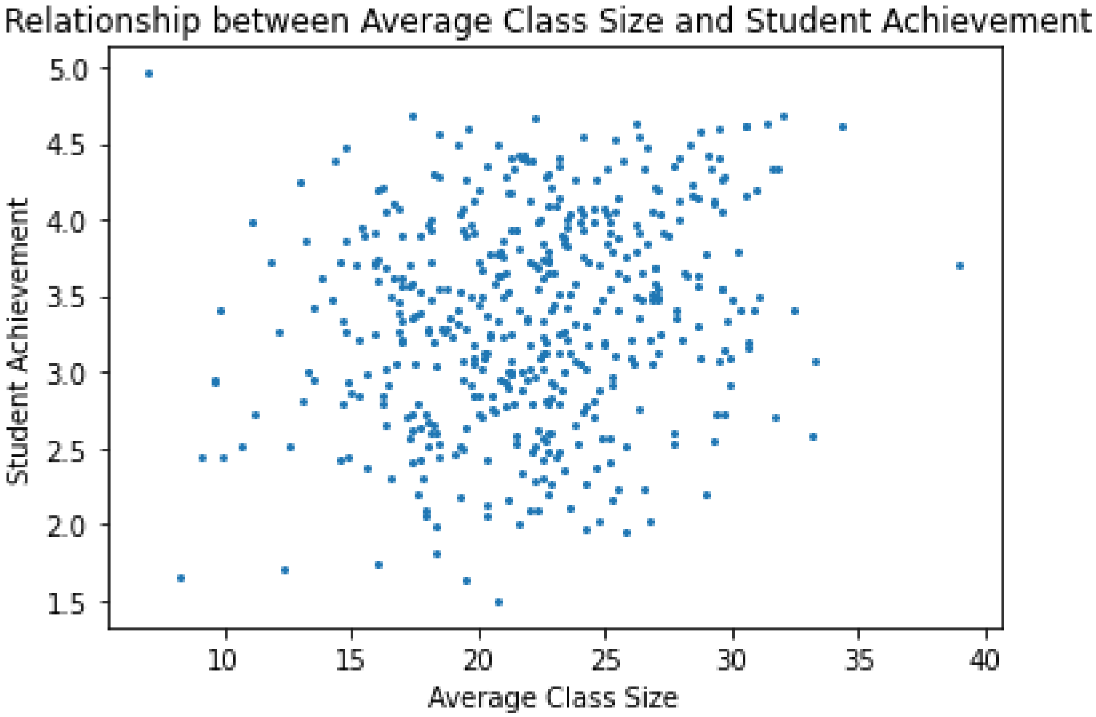


4. There is no relationship between how students perceive their school and how the school performs on measures of achievement. Doing a PCA on the first 6 elements(blue) revealed one eigenvalue that explained more variance than what it brought in. An analysis of the loadings matrix for the first eigenvalue pointed that collaborative teachers and trust could be grouped together as the first principal component. Doing a PCA for the for the second performance PCA, there was also one eigenvalue that should be looked at. Applying a correlation on the transformed data of both parts yielded a correlation of -.05, of which was interpreted as no correlation.

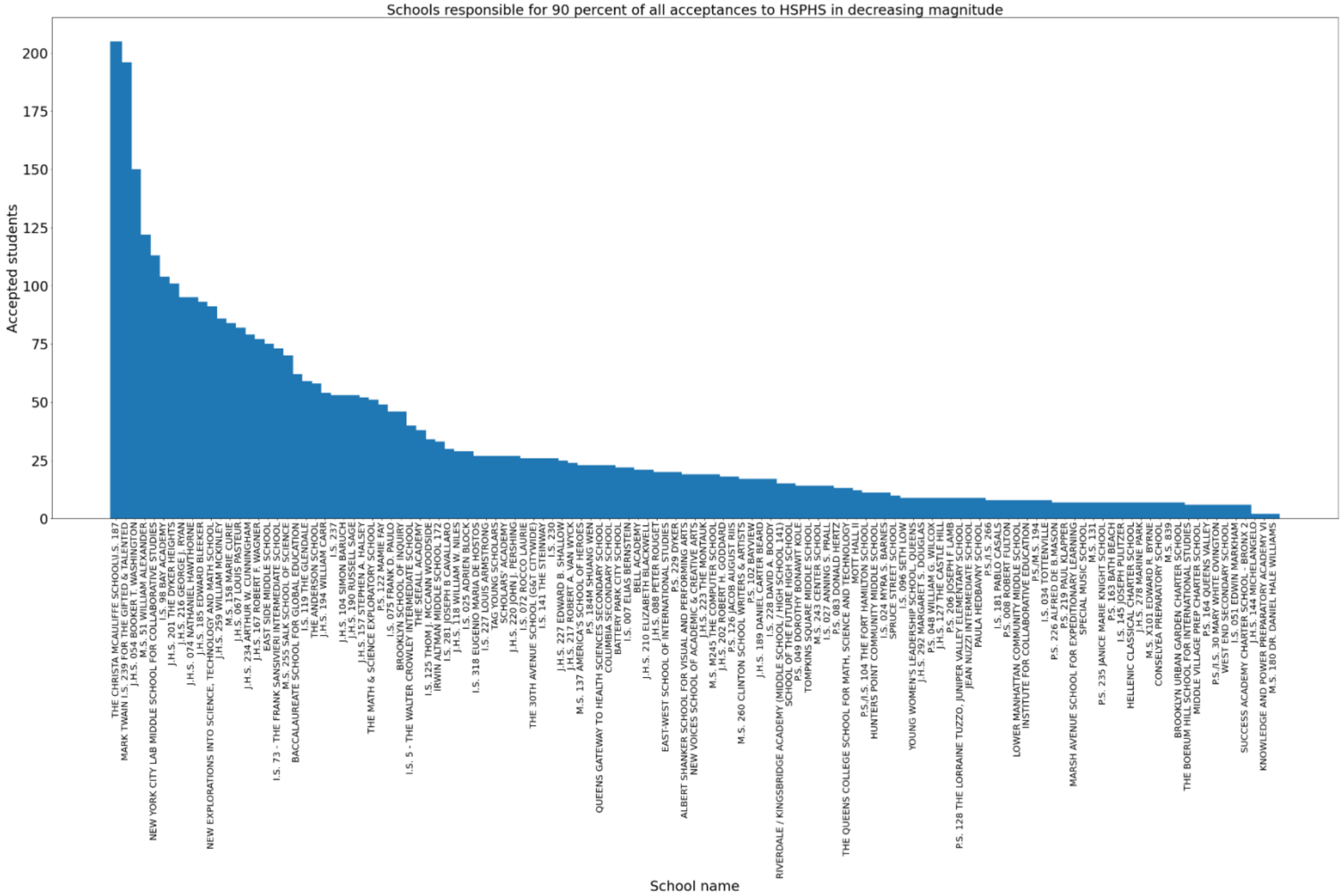


5. The hypothesis of the effect of per pupil spending on student achievement was tested. If schools were rich, then they could afford to spend more per pupil on resources which should translate into a higher academic achievement score. To do this, the two columns of student achievement and average class size were cleaned of nan variables. After, the median of the median per pupil spending column was taken, which would categorize whether a school was rich or not. The dataset was then split into 2 parts, the first part containing values greater than and the second part containing values less than the median which was accomplished by indexing into the average class size column and comparing each row's average class size and splitting accordingly. Then a related groups t-test was taken, yielding a p value of 0.006. As it is less than our alpha level of 5%, the null hypothesis that a difference in per pupil spending has no effect on the student achievement level is rejected. A related groups t-test was chosen over an independent groups t-test because of high levels of school variability, which a related groups t-test is much better at dealing with. While per pupil spending does have an effect, the median of the data is clustered at around $22,000, with a wide variance in student achievement. Spending $22,000 per student when spent wisely is enough to produce high achievement

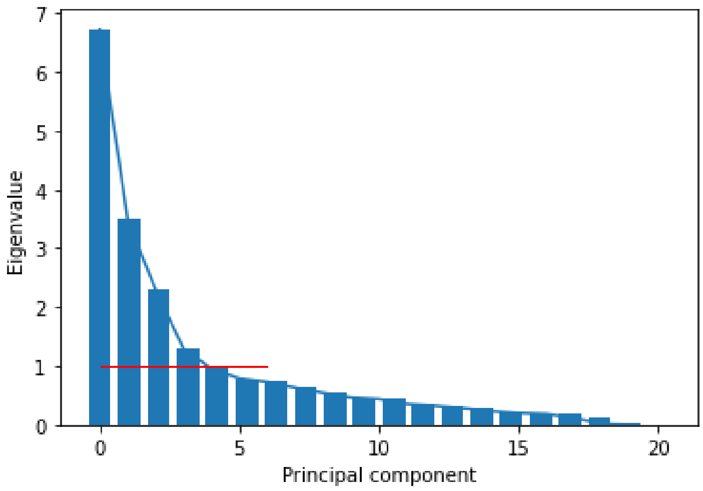
scores, but it doesn't leave very much leeway if the funds are spent poorly, as there are a lot of schools that don't perform well.

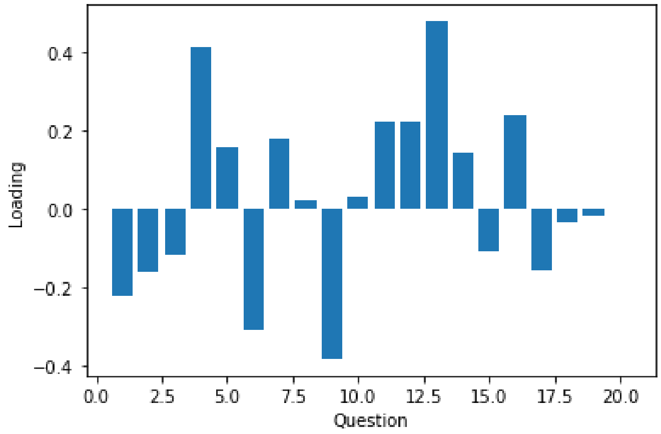


6. The impact of class size was tested on student achievement levels. To do this, the two columns of student achievement and average class size were indexed into and cleaned of nan variables. After, the median of the average class size column was taken. Using the median, the dataset was split into 2 parts, one greater than and one less than the median by indexing into the average class size column and comparing each row's average class size and splitting accordingly. Then a related groups t-test was taken, yielding a p value of .0018. As it is less than our alpha level of 5%, we reject the null hypothesis that the class size effects have no effect on objective measures of achievement. Student achievement seems to be positively correlated with average class size, but this may not consider other factors, such as how well the teacher manages the class or how well the students get along with each other.

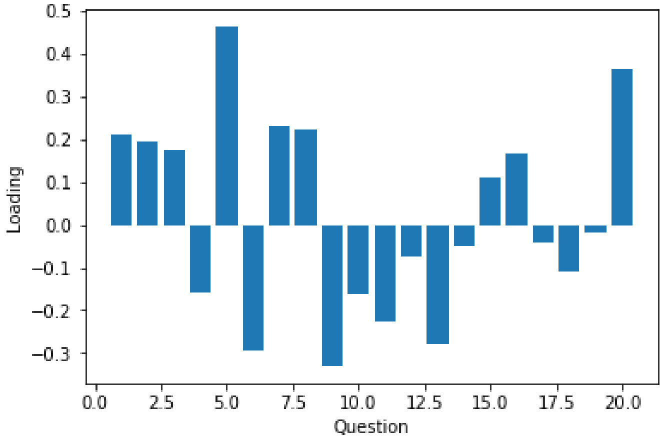


7. The proportion of schools was interpreted in descending order. Therefore, three columns of the dataset were selected – school name, acceptances, and applications. It was cleaned of nan variables and sorted along the acceptance’s column in descending order. A sum was taken to get the total number of students accepted into HSPHS. The dataset was then iterated over by row, with two variables. One variable was used to track the row currently on, which would represent the school. The other variable tracked the total students accepted as each school was iterated over. Once this number met or exceeded the 90% threshold, the loop would break. Out of the 594 schools that were present in the data, 123 schools were responsible for 90% of the students that were admitted to HSPHS, or 20.7% of schools that had submitted data.





Loadings matrix of principle component 1



Loadings matrix of principle component 2

8. Part A) A PCA was run on all the variables excluding the database number, the school’s name, and applications and acceptances as running an OLS on the variables could possibly create issues with multicollinearity. It returned four eigenvalues of significance. When modeling acceptance rate using OLS with the four principal components, the first

two principal components had p-values less than the 5% alpha level. The first principal component was a combination of effective school leadership, strong family community ties, and trust, a possible representation of the environment that encouraged a student to learn or not. The second principal component was a combination of rigorous instruction, collaborative teachers, a supportive environment, and trust again.
Part B) A similar approach to part A was taken, but instead of acceptance rate, student achievement was set as the dependent variable instead. An OLS regression was then performed on the principal components, and the result was the principal components containing per pupil spending, rigorous instruction, a supportive environment, and trust were significant, the structure of the school.

9. Admittance to HSPHS seems to be contingent on the principal components discussed in 8, student environment and trusted guidance. Having both allowed the student to establish a road map through consulting the community that they trust, and therefore set achievable goals to motivate them to their destination. Having a strong environment that encouraged students to learn tends to lead to higher student achievement scores, a significant factor in dictating admissions to HSPHS. But in order to have high student achievement scores, the school structure would have needed to be established, as that is a leading factor in causing high student achievement scores.

10. Part A) To improve chances of getting into HSPHS, schools should try to create an environment where students to be highly motivated to learn and achieve goals, as well trust the school enough to seek guidance. Since student achievement is a key factor in admission to HSPHS, the school should try to set up an environment that encourages students to learn rather than a chore or something they are forced to do. This could be done through retraining teachers and being more transparent about why students are required to do what they're doing, as well as making counselors available to help them decide on jobs or roles they'd like to be in the future.
Part B) To improve objective measures of achievement, there would be a need to establish a strong foundation and structure of the school. From the analysis done in 8, increasing per pupil spending, rigorous instruction, supportive environments and trust made up a principal component, so the school could focus on those. Additionally, from the analysis done on per pupil spending and class size we know that per pupil spending should be increased and class size should be decreased if they are too low and too large respectively. However, I do realize that increasing per pupil spending would require increased funding, which doesn't seem extremely likely as a tax increase would be frowned upon by the public, and redistribution of city funds from other areas would likely be a lengthy process. In the short run, schools should try to increase rigorous instruction levels, while at the same time fostering a trustful supportive environment that students would feel comfortable and willing to learn at a more rigorous pace.