5/19/2021
Justin Chang

Introduction

Data Cleaning: I cleaned the data on a per question basis. First, I would assess what variables I would be needing to answer the question. Then have all the columns (variables) I needed in one composite array.

If I needed to calculate any new variables (i.e. application rate) I would do so before the data cleaning.

Afterwards, using the method shown in class, I iterated row by row seeing if any of the elements in each row had any Nans. If yes, then the row was deleted. This way I could keep as much data as possible when conducting tests per question.

Dimension Reduction: When there were simply too many variables, I used PCA and Kaiser's justification to reduce each group of variables down into one.

Data Transformation: I used pandas to extract the CSV file as I was running into some problems using the numpy method. Then I copied the pandas dataframe into a numpy array. If there was any skewed data or data being used for a PCA, the data was z-scored and normalized.

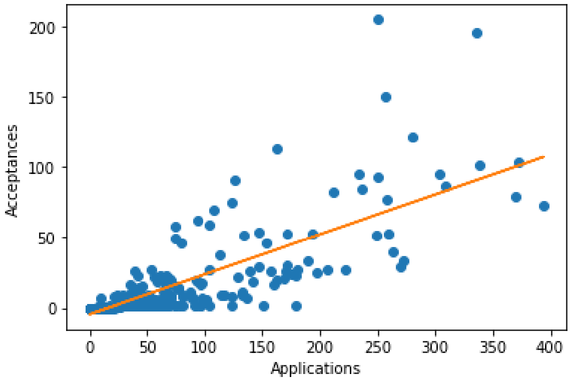
Questions

Question 1:

To test the correlation between number of applications and admission to HSPHS, I first extracted the columns from the entire data set. Then used numpy's corrcoef function with these two columns as inputs to calcuate an r value of 0.8017265370719316. There appears to be a strong positive linear relationship between number of applicants and number of accepted students to HSPHS.

There was no missing data so I felt there was no need to clean up the data. There were only 2 variables as well so I felt no need to do any dimension reduction here.

In order to get the plot, the stats package (stats.linregress) from scipy was used in order to get all the values needed for the best fit line.



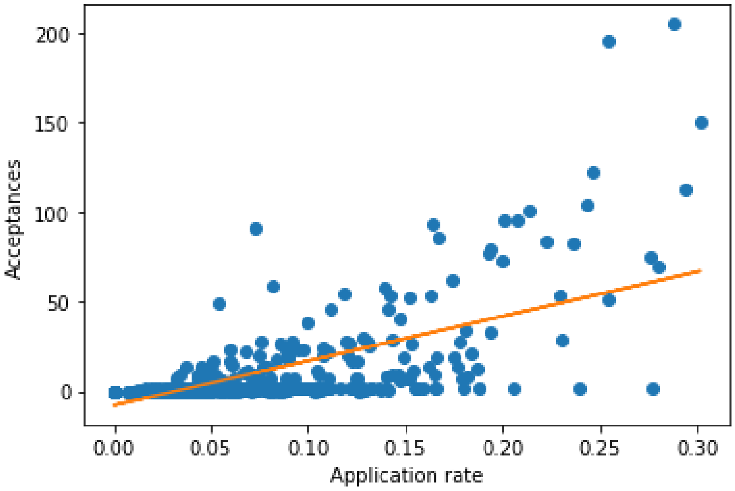
5/19/2021
Justin Chang

Question 2:

The columns that indicated applications, acceptances, and school size were extracted the array containing all data.

To get the application rate, the number of applications in each row was divided by the school size. This was done using a for loop and iterating over the indices for the array of data and each result was put into a new array of application rates.

The application rates and acceptances were then put into the np.corrcoef function. The result was an r value of 0.6587507529002681. So there appears to be a moderate positive linear relationship between the two.



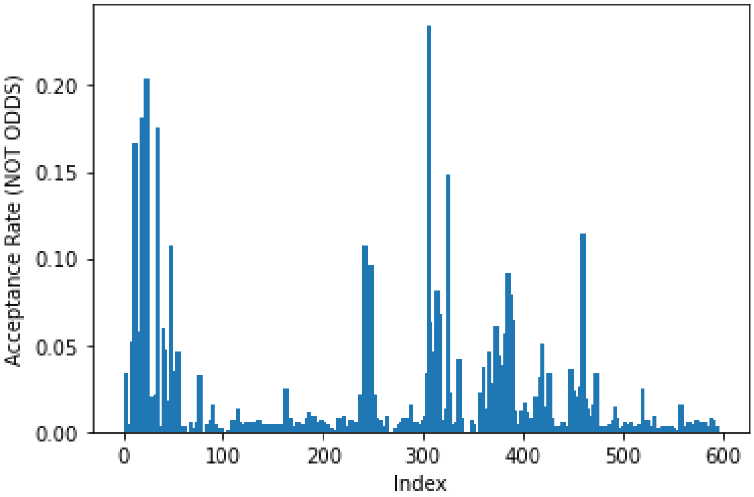
The best fit line was calculated with the same function in Question 1, stats.linregress.

Since the correlation between number of applications and the acceptances was already tested. We just compare the two correlations that we calculated. Applications appears to be a better predictor that application rate as the correlation is comparably larger. Additionally, the p value for applications as a predictor is 1.929258245307361e-134, this is much smaller compared to the p value of rate of applications which is 6.141061848423162e-75.

Question 3:

To get the acceptance rate per student, the number of acceptances in each row was divided by the school size. This was done using a for loop and iterating over the indices for the array of data and each result was put into a new array of acceptance rates.

The maximum acceptance rate was found in the new array using np.nanmax and its index was found using np.where. This index was then traced back to the original array of data to find the name. The odds were then calculated using the accepted students and school size of that school.



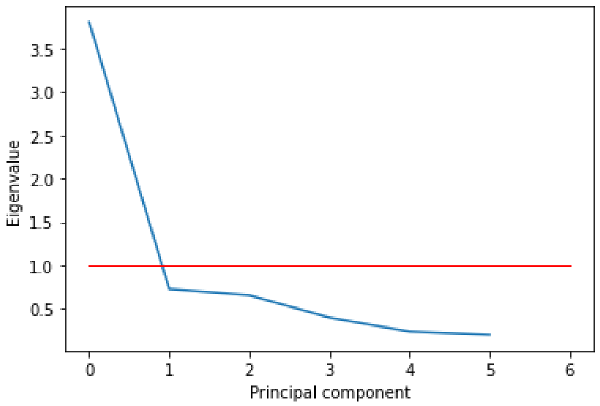
The Christa Mcauliffe School has the best per student odds of sending someone to HSPHS. The odds are 0.3068862275 or (205/668).

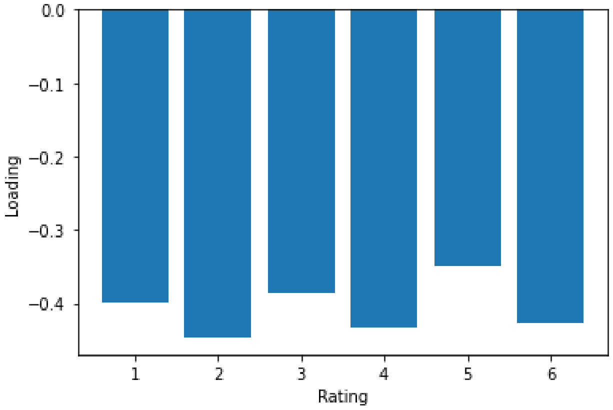
Question 4:

Because we need to fine if there exists a relationship between the student perception of a school and their performances, Columns L-Q and V-X were extracted from the data array. Then all of these selected columns are cleaned of Nans.

Columns L-Q are all representative of student opinions of the school, so we group those 6 variables into one PCA. Columns V-X are representative of student performance so these 3 variables into another PCA. We do PCA here because there are too many variables are many of these variables within their groups are seemingly redundant.

PCA 1 (Rigorous Instruction, Collaborative teachers, Supportive Environment, Effective School Leadership, Strong Family Community Ties, and Trust):

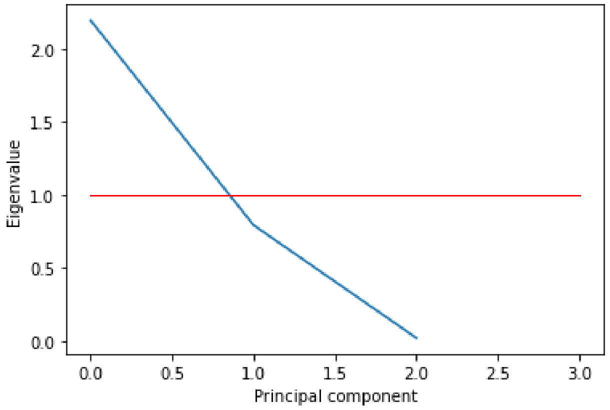


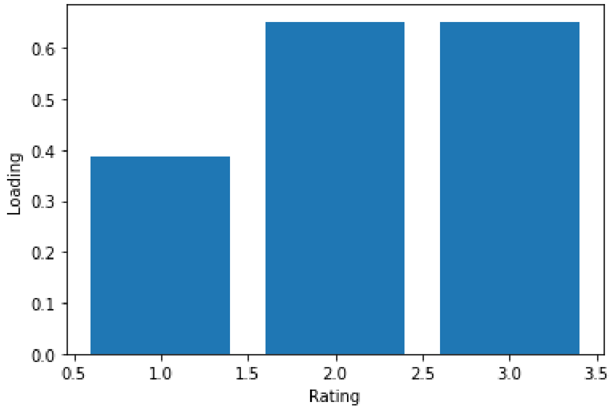


We see that factor 1 (marked as 0 in the eigenvalue graph) is accounting for a lot of information. On the loadings it shows that it has negative commonalities with all the originally variables, generally accounting for how poor the student perception of the school is. It can be interpreted as How poor the student's overall perception of the school is.

None of the other factors are used because of Kaiser's criteria, only Factor 1 is used.

PCA 2 (Student Achievement, Reading Scores, and Math Scores):



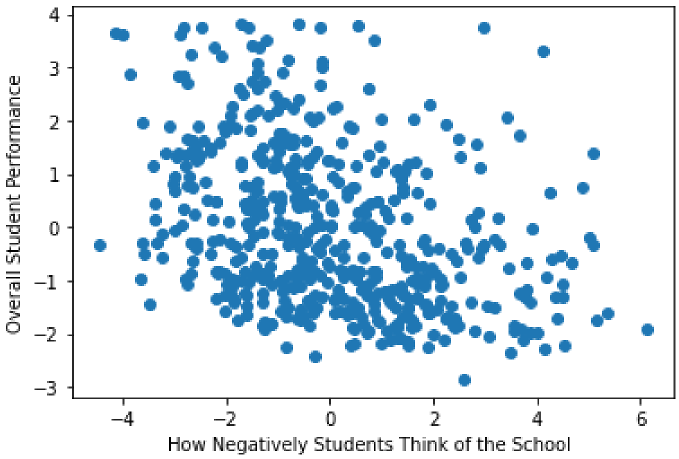


With PCA2 factor 1 (0 in the eigenvalue graph) also seems to account for most of the information in the original 3 variables. The loadings show that is has positive commonalities with all 3 of the original variables. This can be interpreted as overall performances of the students within that school.

Using Kaiser's criteria once again, we see that factor 1 is the only factor with an eigenvalue greater than 1. As a result, this is the only factor being used.

Relation:

There appears to be weak relation between the student perception of the school and their performances at said school. When correlating the 2 factors from PCA, the correlation value was only -0.36744642711810915. This suggests a weak negative correlation between how badly the students think of the school and their performance, or in other words, a weak positive correlation between the Student perception of their school and their performance. The p value for this is 2.935859466721121e-18.



Question 5:

I hypothesized that there was a difference between number of acceptances in High Poverty Schools (Poverty rate is 75 or above) and Not High Poverty Schools (Poverty rate is below 75) on the dependent variable of acceptance rate.

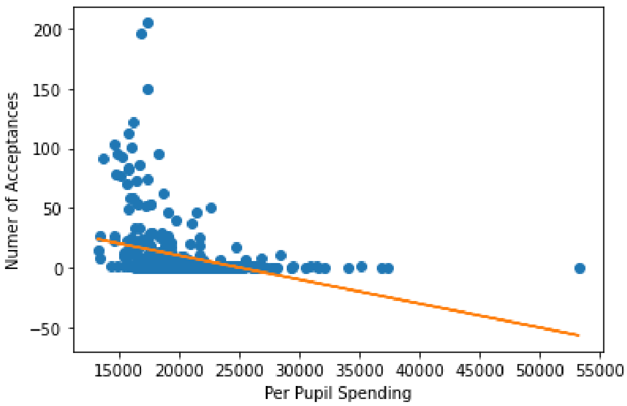
Null hypothesis being that there is no differences between the mean acceptances between these 2 samples.

Using an independent samples t test, I tested to see if the mean number of acceptances was different between the two samples' means. The p value was 0.47107246308651074, the t score was 0.8814016624013438.

This is not evidence as the p value is far too high for the result to be considered significant, so we fail to reject null hypothesis.

Question 6:

*NOTE: Because all charter schools lack per pupil spending, and missing values were removed, this result should not be interpreted to include charter schools.



To test whether there is any evidence material resources impacting student success. I correlated the Per pupil spending with the number of acceptances. The results were an correlation coefficient of -0.335125031707362 and a p value of 7.058248165264534e-14. So there appears to be a weak negative correlation between per pupil spending and number of acceptances. While this does appear to affect acceptances negatively, do the reason is not clear.

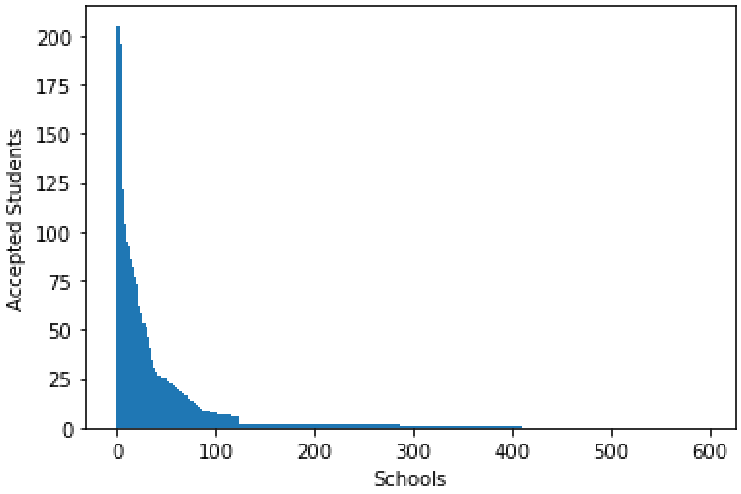
The y axis has a lower bound of 0, so no values (acceptances) go below 0.

Question 7:

To find what proportion of schools make up 90% of acceptances to HSPHS, I first extracted the number of acceptances column from the data set. This array was then sorted from greatest to least values using np.sort(arrayname)[: :-1]. The sum of all acceptances was calculated suing np.sum on the acceptances column. Total acceptances were 4016.

This array was then put through a for loop that adds accepted students from each row into a count 1 by 1 until the cumulative number is greater than or equal to 90% of the total sum (4016*0.9). The loop also counts each iteration to keep tack the number of schools needed to reach 90% of acceptances.

The result was that 123 out of 594 or 20.7071% of all schools make up 90% of acceptances.



5/19/2021
Justin Chang

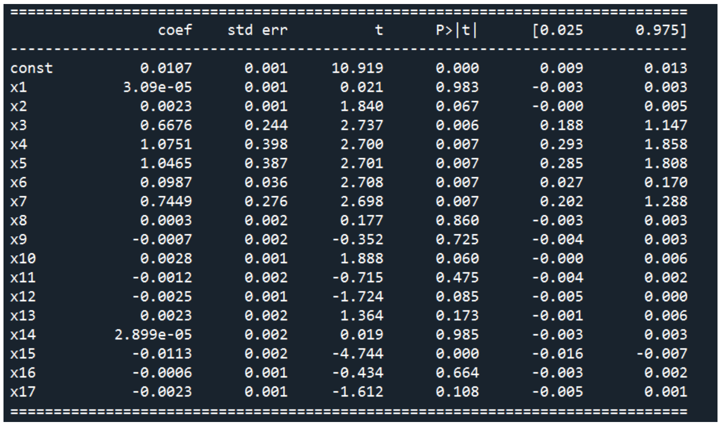
Question 8:

*NOTE: Because all charter schools lack per pupil spending, and missing values were removed, this result should not be interpreted to include charter schools.

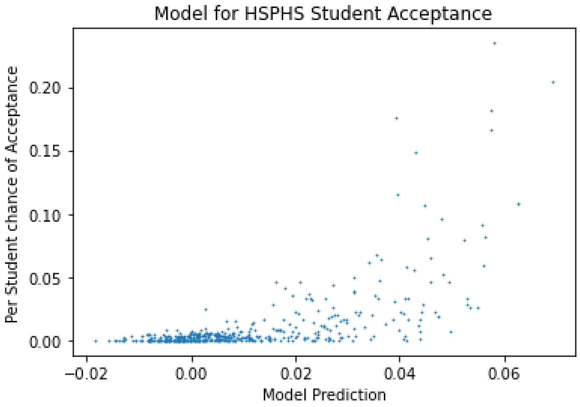
I chose multiple regression as my model for both parts A and B. As it seemed to be the simplest model to use for both of the dependent variables.

A:

The dependent variable (acceptance rate) was calculated the same way as question 5. While all the independent variables were z scored just in case of any skewed data. The independent variables used were all columns except objective performance variables, applications, and acceptances. I chose not to use object performance variables because we are testing for it in another model. While number of applications seems to be a good characteristic to base our predictions on.



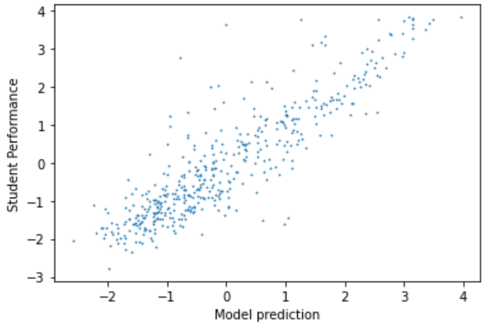
```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0107      0.001     10.919      0.000       0.009       0.013
x1           3.09e-05      0.001      0.021      0.983      -0.003       0.003
x2             0.0023      0.001      1.840      0.067      -0.000       0.005
x3             0.6676      0.244      2.737      0.006       0.188       1.147
x4             1.0751      0.398      2.700      0.007       0.293       1.858
x5             1.0465      0.387      2.701      0.007       0.285       1.808
x6             0.0987      0.036      2.708      0.007       0.027       0.170
x7             0.7449      0.276      2.698      0.007       0.202       1.288
x8             0.0003      0.002      0.177      0.860      -0.003       0.003
x9            -0.0007      0.002     -0.352      0.725      -0.004       0.003
x10            0.0028      0.001      1.888      0.060      -0.000       0.006
x11           -0.0012      0.002     -0.715      0.475      -0.004       0.002
x12           -0.0025      0.001     -1.724      0.085      -0.005       0.000
x13            0.0023      0.002      1.364      0.173      -0.001       0.006
x14         2.899e-05      0.002      0.019      0.985      -0.003       0.003
x15           -0.0113      0.002     -4.744      0.000      -0.016      -0.007
x16           -0.0006      0.001     -0.434      0.664      -0.003       0.002
x17           -0.0023      0.001     -1.612      0.108      -0.005       0.001
==============================================================================
```



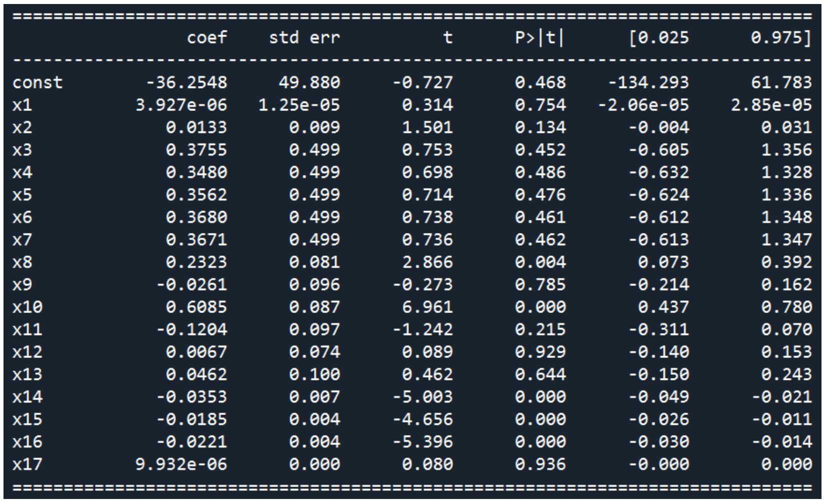
X1 to X17 corresponds with columns L-U in the same order (x1 = Per pupil spending, x2 = average class size, etc.)

- Several race related factors (x3 to x7) seem to be positively associated while also being the most impactful.
- Poverty percent (x15) seems to be negatively associated with acceptance rate.
- All other factors seem too unreliable to count on as the p values are too high to be used.

B:



The dependent variable this time was the objective performance variables put into a PCA and using only factor 1 (same PCA2 from Question 4). I chose this to narrow student achievement into one variable and it seems to be a good indicator of general student performance. The independent variables are the same as part A. (Student performance is measured in PCA factor 1).



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -36.2548     49.880     -0.727      0.468    -134.293      61.783
x1          3.927e-06   1.25e-05      0.314      0.754   -2.06e-05    2.85e-05
x2             0.0133      0.009      1.501      0.134      -0.004       0.031
x3             0.3755      0.499      0.753      0.452      -0.605       1.356
x4             0.3480      0.499      0.698      0.486      -0.632       1.328
x5             0.3562      0.499      0.714      0.476      -0.624       1.336
x6             0.3680      0.499      0.738      0.461      -0.612       1.348
x7             0.3671      0.499      0.736      0.462      -0.613       1.347
x8             0.2323      0.081      2.866      0.004       0.073       0.392
x9            -0.0261      0.096     -0.273      0.785      -0.214       0.162
x10            0.6085      0.087      6.961      0.000       0.437       0.780
x11           -0.1204      0.097     -1.242      0.215      -0.311       0.070
x12            0.0067      0.074      0.089      0.929      -0.140       0.153
x13            0.0462      0.100      0.462      0.644      -0.150       0.243
x14           -0.0353      0.007     -5.003      0.000      -0.049      -0.021
x15           -0.0185      0.004     -4.656      0.000      -0.026      -0.011
x16           -0.0221      0.004     -5.396      0.000      -0.030      -0.014
x17         9.932e-06      0.000      0.080      0.936      -0.000       0.000
==============================================================================
```

X1 to X17 corresponds with columns L-U in the same order (x1 = Per pupil spending, x2 = average class size, etc.)

- Once again race related factors (x3 to x7) seem to be positively associated while also being impactful.
- While small, class size (x2) does seem to have a positive association with student achievement.
- Rigorous instruction (x8) seems to be moderately associated with student success.
- The variable with by far the most association with student success appears to be a supportive environment (x10).
- Collaborative-ness of teachers oddly looks to be negatively correlated with student success (x11).
- Disability, Poverty, and ESL percent (x14 to x16) all seem to have a small negative impact on how successful students are.
- All other factors are too unreliable (p-value too high) to notice.

Question 9:

It appears that there appear several unknown race related factors of impact student acceptance and success. It would be unwise to say that race determines success and acceptance to HSPHS, as correlation does not equal causality. Deeper socioeconomics analysis to discover the reasons behind this observed association is needed. A supportive environment has a big positive impact on student success. Disability, Poverty, and ESL percent all have a negative correlation with student success. This makes sense as these 3 characteristics are all disadvantages when it comes to an English-speaking educational experience.

Question 10:

I would highly recommend increasing how supportive a school's environment is to students. This was a factor that had a clear impact on student success. This could possibly be due a variety of reasons such as mental health of students or other factors. As for increasing the number of HSPHS, this seems to be heavily race related. But race is an entirely different issue and requires deeper socio-economic analysis. All other factors either had too small of an impact or were seemingly too unreliable to say it had an impact on either acceptances or student success.