Masaaki Kato

8.17.21

Data Handling:

For the purpose of this analysis, I decided to use row-wise removal to handle all NaN values. This is because many of the missing data included school spending (column E) and class size (column F), which are systematically absent from all 109 charter schools. So, instead of potentially not being able to use these factors, I decided to remove all charter schools and make the analysis only on NYC public middle schools. Fortunately, the row-wise removal resulted in 75% (449 schools) remaining, which I believe is more than enough to conduct sufficient analysis. For the PCA analysis, I looked at the magnitude and number of the loading values pointing in the same direction as the principal component to properly interpret it. In regards to data transformation, I used both the pandas dataframes and numpy arrays depending on the question.

Question 1:

Number of applications was measured using column B and admissions was measured using column D. The correlation of number of applications and admissions to HSPHS is 0.81(p<0.01). This indicates that as the number of applications increases, the admission increases as well. Overall, lower applications (<50) do not increase admissions as much and the increase is readily seen past 100 applications.

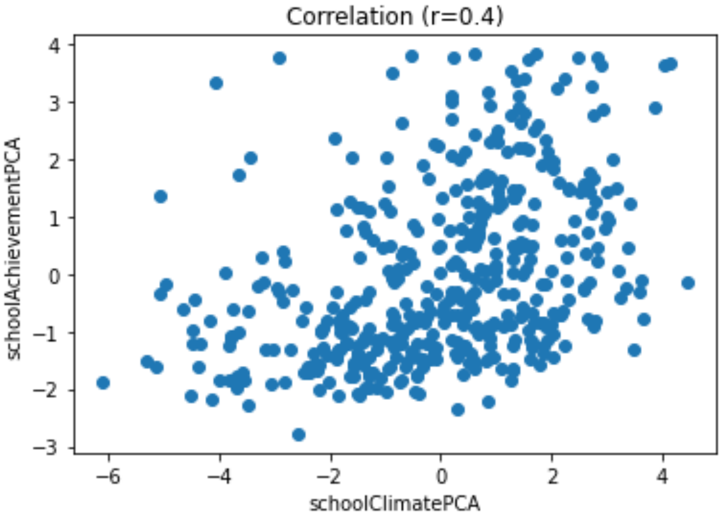

Figure1: Scatterplot between the number of applications and HSPHS admissions

Question 2:

In order to determine to find which factor, raw number of applicants or application rate, is a better predictor of admissions to HSPHS, I conducted two separate OLS regressions. The application rate was calculated by dividing the number of applicants (column C) by the school size (column U). In the first regression, I regressed the number of admissions on the number of applicants and found the beta coefficient is 0.29 (r^2=0.65). In the second regression, I regressed the number of admissions on the number of admissions on the application rate and found the beta coefficient is 272 (r^2=0.47). These findings suggest that the number of applicants is a better predictor of admissions to HSPHS because it has a higher r^2, which means it explains more of the variance. Furthermore, in order to determine whether if these two factors together better predicted the admissions, I conducted a multiple regression and found r^2 is 0.66. Since, the r^2 value did not increase as much compared to first regression, this means that these two factors are likely to be highly correlated and cover the same domain of information.

Question 3:

I interpreted "per student odds" as how likely it for student to be accepted to HPSHS (i.e., acceptance rate). I calculated this acceptance rate by dividing the number of acceptances by the number of applicants. This was applied to every single school, and I added the resulting array as a new column in the dataset. Then, I sorted the dataset in descending order by this acceptance rate and found that the Christa McAuliffe School (I.S. 187) had the highest acceptance rate. To conclude, the Christa McAuliffe School has the best per student odds of sending someone to HSPHS.

Question 4:

In order to analyze the relationship between how students perceive school (6 factors, column L-Q) and their objective achievements (3 factors, column V-X), I, first conducted a PCA for each set of factors (perception and achievements) to ideally explain them in a single factor. In the first PCA on school perception, the Screeplot showed the first principal component has an eigenvalue of 3.8, which is nearly 64% of all variables (out of 6). Based on the elbow and Kaiser criterion, the first principal component was the only interpretable factor and based on the its loading values, where all the factors have high positive values, I interpreted the first principal component to represent the general school climate. In the second PCA on student achievements, the Screeplot showed the first principal component has an eigenvalue of 2.3, which is nearly

77% of all variables (out of 3). Based on the elbow and Kaiser criterion, I chose only the first principal component and based on its loading values, where all the factors have high positive values, I interpreted the first principal component to represent the student's achievement metrics. Then, I correlated the principal components of the two PCA results and found r is 0.4(p<0.01). This suggests that there is a moderate positive correlation between how student's perceive school climate and the student's achievement scores.

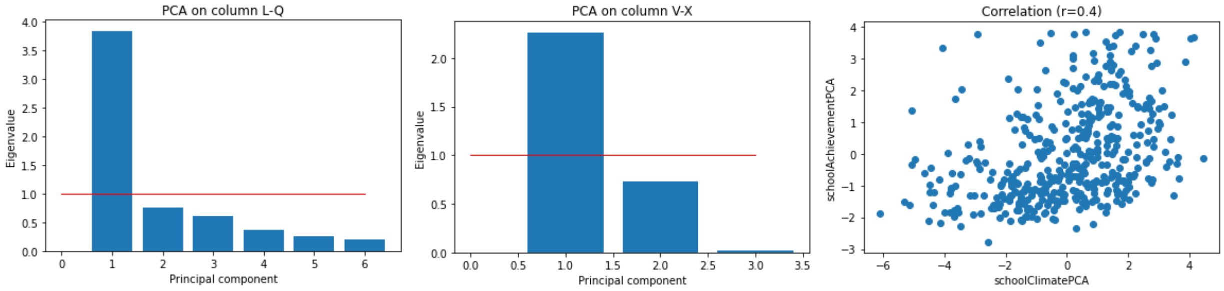

Figure 2: The first two graphs on the left are the Screeplots for PCA on column L-Q and column V-X. The last graph represents the correlation between school climate PCA and achievement PCA.

Question 5:

My hypothesis is that school spending impacts the HSPHS acceptances. School spending is measured by the spending per student (column E) and the HSHPS acceptances is measured by the raw number of acceptances (column D). I tested my hypothesis using an independent t-test to compare the means of the HSPHS acceptances between poor ($\bar{x}$=16.8) and rich schools ($\bar{x}$=1.9). The two groups were split based on the median of the spending per student, where if the school spending was above the median it was grouped as a rich school, while if the school spending was equal or below the median it was grouped as a poor school. The null hypothesis is that there is no differences between the means and even if there was a difference it is due to chance alone, while the alternate hypothesis is that there is a difference between the means that is not due to chance alone. Based on the t-test result, there is a significant difference (t=-6.9,p<0.01) in the HSPHS acceptances between rich and poor schools and thus, I rejected the null hypothesis and concluded school spending negatively impacts HSPHS acceptances. Furthermore, I checked if these two samples violate the homogeneity of variance assumption by using the Levene test for equality of variance and found that sample deviations are not statistically significant (p> 0.05), which means I failed to reject the null hypothesis, that the sample deviations are the same.

Question 6:

I measured the school's availability of material resources using per student spending (column E) and average class size (column F). I measured the school's objective assessment using average achievement scores (column V) and the HSPHS acceptance (column D). I correlated these factors and the results are represented as scatterplots shown in Figure 3. School spending has a negative impact on both achievement scores (r=-0.15,p<0.01) and HSPHS acceptances (r=-0.34,p<0.01), while class size has a positive impact on both achievement scores (r=0.21,p<0.01) and HSPHS acceptances (r=0.36,p<0.01). Furthermore, I inspected the relationship between the two factors, measuring material resources, and found they are negatively correlated (r=-0.46,p<0.01). These findings suggest that schools with larger class sizes perform better in both HSHPS acceptances and achievement scores and these same schools have lower school spending.

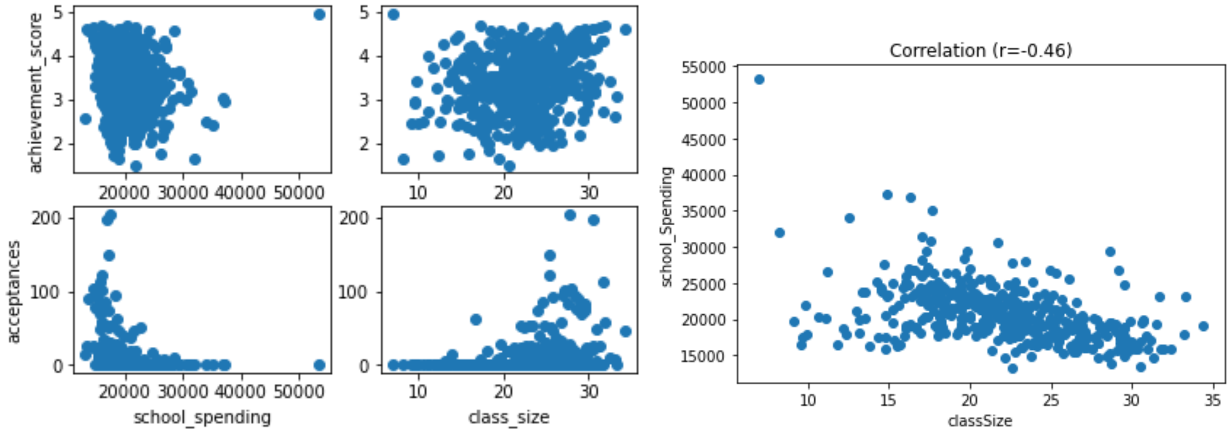

Figure 3: On the subplot on the left, the left side represents the scatterplots between school spending and achievement scores/HSPHS acceptances. The right side of the subplot represents the scatterplot between class size and achievement scores/ HSPHS acceptances. The graph on the right is the correlation between class size and per student spending.

Question 7:

For this analysis, I used the raw number of acceptances (column D). Since, this question is about acceptances, I removed schools with 0 acceptance from my analysis. This resulted in 291 schools and looking at the frequency table, I found that 212 schools have an acceptance of 2 or greater. However, this only makes up 73% of acceptances, so I added schools with 1 acceptance until the proportion reached 90%. I found that 261 schools make up 90% of HSPHS acceptances and found the proportion of schools this corresponds to by dividing by the total number of schools (449). Overall, 58% of schools account for 90% of HSPHS acceptances.

Question 8:

I used a multiple linear regression model to predict HSPHS acceptances and objective achievement metrics. However, before the model, I conducted a PCA to reduce the number of factors. The PCA included all factors except for the two outcomes, which resulted in 18 factors. Based on the Kaiser criterion, I choose four factors and based on the loading values, I labeled them as positive school climate, large population, improvised communities, and financial stability. Then, I used a multiple regression model with these principal components as predictors and ran two models, one with the HSPHS acceptances (column D) as the outcome and another with the average student achievement scores (column V) as the outcome. Both outcomes were standardized using z-scores and since, the principal components are already standardized, the resulting regression coefficients is also standardized. The results of the two regressions are shown in Figure 4, which includes the r^2 and standardized coefficients of each predictor. In terms of model assessment, both models are not reliable prediction models because of their relatively low r^2 values. However, the regression model on achievements captures far less of the variance. In terms of the impact of the predictors, the positive school climate has a moderate impact on both predictions and this principal component. So, it seems that how well the students perceive the school is a strong indication of the school's HSPHS acceptance and their students' achievement scores.

	Regression1(acceptance)	Regression2(achievements)
PCA1 (positive school climate)	0.25	0.20
PCA2 (large population)	0.17	-0.10
PCA3 (improvised communities)	0.04	0.06
PCA4 (financial stability	-0.08	0.09
r^2	0.45	0.25

Figure 4: Multiple Linear regression results. Regression 1 is where the HSPHS acceptances are the outcome, while regression 2 is where the achievement scores are the outcome.

Question 9:

The largest factor that impacts HSPHS acceptance is how well the students perceive the school. This factor is represented by columns L-Q, which include metrics such as how much students believe that school can be trusted, supportive, and collaborative. This indicates that a safe study environment may impact the student's performances. The second largest factor that impacts HSPHS acceptance is the size of the school, both in terms of class size and applicants. This is supported in the correlation results in question 1 and question 6. In addition, these two metrics are also correlated because they both represent PCA2 in question 8. These results suggest that the larger the class size, the more students apply, and this will also lead to higher HPSHS acceptances without necessarily changing the acceptance rates.

One surprising finding is how school spending negatively impacts HSPHS acceptance and this is supported by both the t-test and correlation results. This may seem confusing, both when you consider the negative correlation between class size and school spending (r=-0.46), it illustrates a clearer picture. Larger schools have a disproportional per student spending when compared to that of smaller schools, and because these large schools have higher acceptances, it makes it seem as if more spending leads to lower acceptances. So, if the per student spending was proportional to the class size across all schools, the school spending may positively impact HSPHS acceptances. However, to note, this is only speculation, and further analysis and experiments need to be conducted to investigate this.

Question 10:

My recommendation would be the same for both HSPHS acceptances and achievement scores because how students perceive the school is the most important factor. I would first recommend to design surveys that meaningfully represents how students perceive the school in fewer questions. This is to make it easily accessible to students and to for the schools to clearly envision future initiatives. Then, I would recommend to post these initiatives to school boards and emails to show the students that the schools takes their feedback seriously. This may lead students to trust the school and feel that it is more supportive. Furthermore, I would recommend to post quarterly reports on how the school fulfilled these initiatives and ask for more feedback from the students. Hopefully, by accomplishing these actions, it will lead students to perceive the school better and feel more comfortable to ask more questions in class and explore more opportunities. This may lead more students to apply to HSPHS, which may lead to more acceptances. The key to this is that it does not artificially increasing the applicants, but rather it cultivates a positive learning environment that encourages students, who may not have even tried to apply otherwise and improve their achievement scores. This also helps establish a positive feedback loop that helps both the schools and students.