Intro to Data Science FInal Project

Author: Diego Lopez

7/29/2021

Introduction & Dimension Reduction

This analysis focuses on differences between New York City Middle Schools and admissions of their students to different specialized high schools. We looked at data collected from each Middle school and used this to determine factors that may be impactful to the rate of admissions of students to NYC specialized high schools. In the dataset, there were many instances of missing values across many features. However, in many of the cases where data was missing, there was more data ‘present’ than there was ‘missing’. For this reason, I decided that inputting data would be a better solution than dropping the rows altogether and potentially losing valuable information that was in the other features of the school. I did this by first Z-scoring the dataset, then inputting for missing values the mean value of that z score, which in most cases was a number that was very close to 0. I also removed race and demographics from the dataset that we are working on. Although this information is critical, we are trying to find factors that can be emphasized to schools that can produce improvement across all demographics. In a future study, it will be worthwhile to actually look at which demographics are struggling and find ways to benefit these specific groups as well.

For PCA, I reduced attributes L-Q to ‘perception of school’ and W-Q to ‘objective achievement’. However, ‘perception of school’ was only used in **Q4**, as I thought it was important in my later model to know exactly which was behind the coefficient, as ‘rigorous instruction’ and ‘trust’ have very different implications. Similarly, I only used ‘objective achievement’ as a dependent variable because when this was IV, I found it necessary to know *what* exactly was behind the coefficient once more as in the previous case.

Q1: Correlation between number of applications and admissions to NYCPHS

For finding the correlation between number of acceptances and number of admissions, we first show the scatter plot:

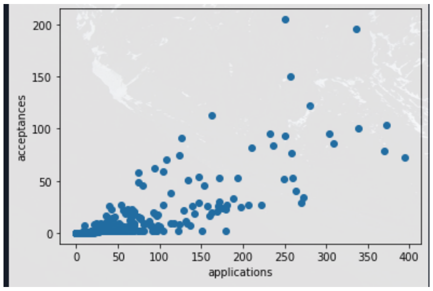


Figure 1: Scatter plot of applications and admissions

We calculate the correlation coefficient using Pearson's method for continuous variables and find **p = 0.801**

Q2: Comparison of predictive models between raw number of applications and application rate for acceptance to NYCSPHS

We first find the application rate by taking applications over school size. We now wish to compare 2 predictive regression models: application rate predicting admissions and applications predicting admissions.

For our model between applications and acceptances, we find **R^2 = 0.643** with a **RMSE of 56.81**. For our model with application rate and acceptances, we find a **R^2 = 0.434** and a **RMSE of 22.3**. We note however that RMSE is scale dependent, and therefore might not be the best measure for evaluating the differences in the models because we are comparing a rate with an aggregate. We therefore look to R^2 and find the raw number of applications to be the better predictor of the two as it accounts for more variance in the model.

Q3: Finding the school with best *per student* odds of acceptance to NYCSPHS

We find the school with the best per student odds of admission. We interpret this as the highest rate of acceptances to school size. We find the school that has the maximal ratio. This turns out to be **I.S. 187 Mcauliffe** with a ratio of **0.23**.

Q4: Analysis of relationship between perception of school and student achievement

We wish to run 2 PCAs, one to reduce the features surrounding ‘student perception of school’ and one to reduce ‘student objective achievement’.

For our PCA on student perception, we have the loadings plot:

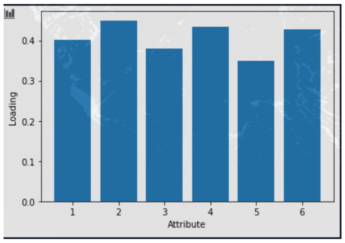


Figure 2: Loadings plot of student perception

We now do another PCA to reduce the objective achievement data:

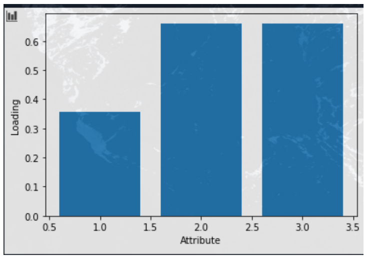


Figure 3: loadings plot of student achievement

We see that in both instances, the data is correlated with each other. We now plot the correlation between these 2 components:

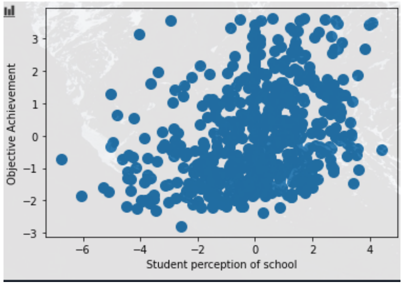


Figure 4: Scatter plot of components

We note that we have a **correlation coefficient of 0.34** and we see that they indeed have a positive relationship.

Q5: Charter school vs non charter school differences in student achievement

For our hypothesis, we test **whether Charter schools have the same distribution for student achievement as non charter schools**. Our null hypothesis is that they have the same distribution, while our alternative hypothesis is that the distributions are not the same.

For this test, we employ the **KS 2 sample test.** This is because it is entirely possible that a few non-charter schools are skewed heavily in terms of high achievement while the majority perform very poorly, which can skew both the mean and distribution while leaving median intact, skewing our perception of median as well. We propose looking at the difference in distributions as a measure of performance.

We note that for the 2 sample KS test, we have **p = 4.2e^-11** and a **K-statistic of 0.3599.**

We plot the CDFs:

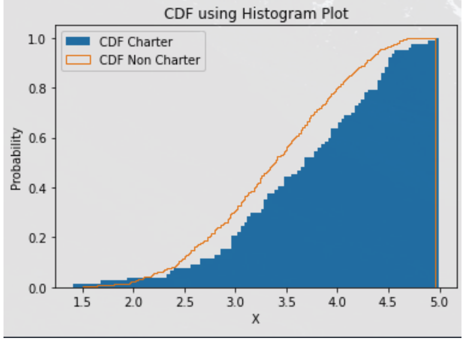


Figure 5: CDF plot

We notice a visual difference as well. This also shows up in the PDFs of the distributions as well:

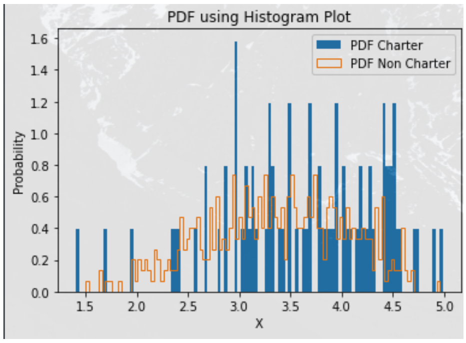
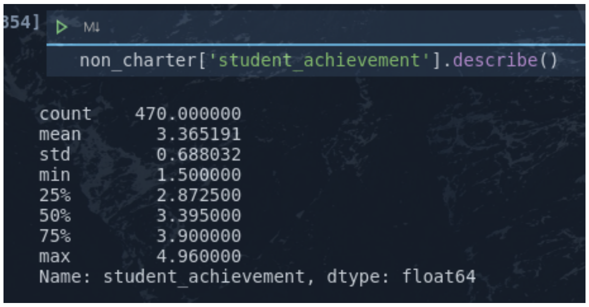


Figure 6: Plot of distributions

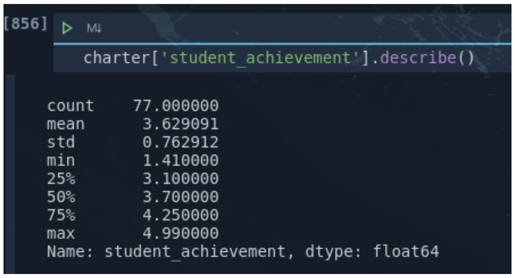
We note that charter schools look visibly skewed higher. We check with the central tendency of the 2 samples:



```
non_charter['student_achievement'].describe()
```

```
count    470.000000
mean       3.365191
std        0.688032
min        1.500000
25%        2.872500
50%        3.395000
75%        3.900000
max        4.960000
Name: student_achievement, dtype: float64
```

Figure 7: Non Charter central tendency



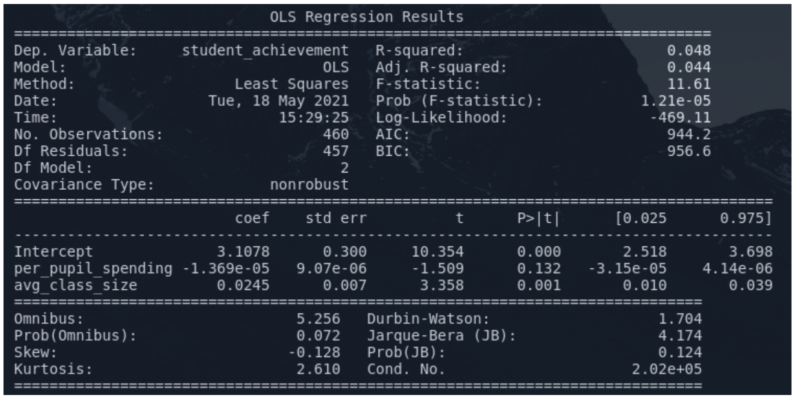
```
[856]
charter['student_achievement'].describe()

count    77.000000
mean      3.629091
std       0.762912
min       1.410000
25%       3.100000
50%       3.700000
75%       4.250000
max       4.990000
Name: student_achievement, dtype: float64
```

Figure 8: Charter central tendency

We see that it is indeed the case that charter schools' distribution as well as averages and median are higher than non charter schools. This can be due to a number of factors, which includes the possibility that charter schools specifically prepare for these types of exams, which could obviously bias their data in favor, and must be examined further at a later time. However we reject the null hypothesis at **p < 0.0001**.

Q6: Analysis of material impact on objective achievement

We want to measure if material resources have an impact on outcomes. We will measure class size and per student spending. We do 2 multiple regressions, one with dependent variable being admissions, and one with dependent variable being student achievement. Regarding student achievement, we see the summary:



```
                             OLS Regression Results
==============================================================================
Dep. Variable:     student_achievement   R-squared:                       0.048
Model:                             OLS   Adj. R-squared:                  0.044
Method:                  Least Squares   F-statistic:                     11.61
Date:                 Tue, 18 May 2021   Prob (F-statistic):           1.21e-05
Time:                         15:29:25   Log-Likelihood:                -469.11
No. Observations:                  460   AIC:                             944.2
Df Residuals:                      457   BIC:                             956.6
Df Model:                            2
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             3.1078      0.300     10.354      0.000       2.518       3.698
per_pupil_spending -1.369e-05   9.07e-06    -1.509      0.132   -3.15e-05    4.14e-06
avg_class_size        0.0245      0.007      3.358      0.001       0.010       0.039
==============================================================================
Omnibus:                        5.256   Durbin-Watson:                   1.704
Prob(Omnibus):                  0.072   Jarque-Bera (JB):                4.174
Skew:                          -0.128   Prob(JB):                        0.124
Kurtosis:                       2.610   Cond. No.                     2.02e+05
==============================================================================
```

Figure 9: Multiple regression of material resources on student achievement

This means holding average class size constant, **per pupil spending is not significant as p=0.13** and the **confidence interval includes 0**. We also see that holding per pupil spending constant, a 1 increase in average class size corresponds with a 0.0245 increase in student achievement and is significant. Looking at each correlation separately:

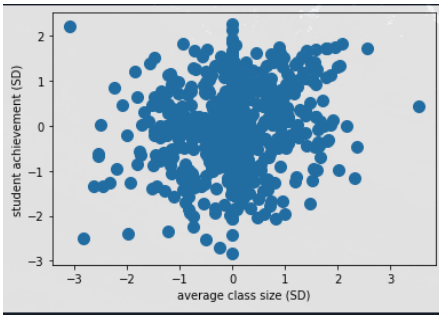


Figure 10: Correlation between average class size & student achievement

We see a correlation between average class size and student achievement of **0.18.**

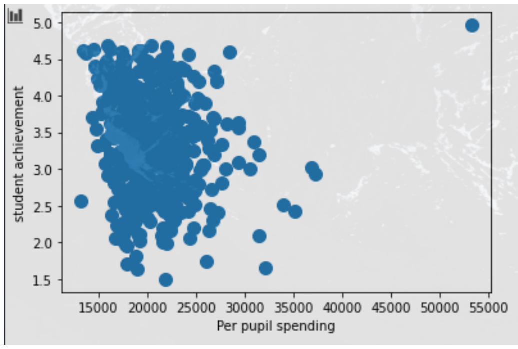
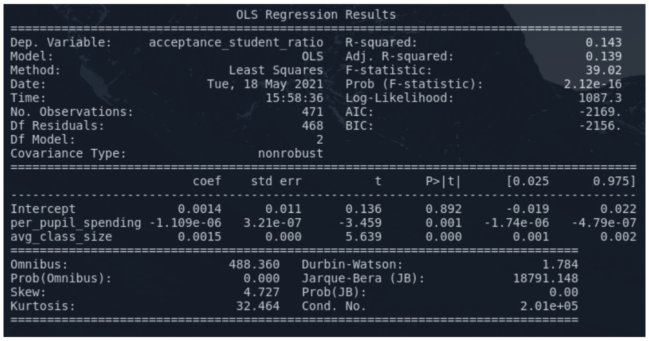


Figure 11: Correlation between per pupil spending & student achievement

We see a **correlation of p = -0.13** for per pupil spending and student achievement.

It appears that per pupil spending is insignificant when holding class size constant, and class size being moderate in impact.

We now observe the effect on acceptances. We will work with the acceptance student ratio to factor in school size. We run multiple regression with a dependent variable being the acceptance student ratio.



```
                               OLS Regression Results
====================================================================================
Dep. Variable:     acceptance_student_ratio   R-squared:                       0.143
Model:                                  OLS   Adj. R-squared:                  0.139
Method:                       Least Squares   F-statistic:                     39.02
Date:                      Tue, 18 May 2021   Prob (F-statistic):           2.12e-16
Time:                              15:58:36   Log-Likelihood:                 1087.3
No. Observations:                       471   AIC:                            -2169.
Df Residuals:                           468   BIC:                            -2156.
Df Model:                                 2
Covariance Type:                  nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              0.0014      0.011      0.136      0.892      -0.019       0.022
per_pupil_spending -1.109e-06   3.21e-07     -3.459      0.001   -1.74e-06   -4.79e-07
avg_class_size         0.0015      0.000      5.639      0.000       0.001       0.002
==============================================================================
Omnibus:                      488.360   Durbin-Watson:                   1.784
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            18791.148
Skew:                           4.727   Prob(JB):                         0.00
Kurtosis:                      32.464   Cond. No.                     2.01e+05
==============================================================================
```

Figure 12: Multiple regression for material resources on acceptances

We see that per pupil spending and average class size are both significant. When holding average class size constant, we see that per pupil spending is negatively associated with the acceptance student ratio. We see that when holding per pupil spending constant, average class size is positively related with the acceptance student ratio.

We see a correlation coefficient between per pupil spending and acceptance student ratio of 0.07

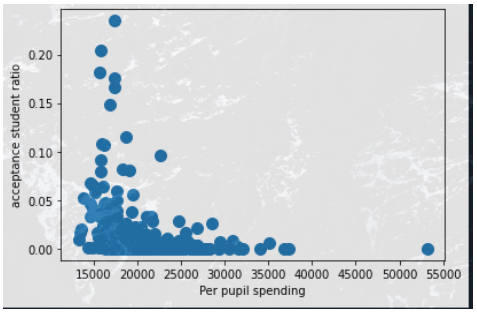


Figure 13: Correlation between per pupil spending and acceptance student ratio

We see a positive correlation between average class size and acceptance student ratio and a **correlation coefficient of 0.34** which does warrant further studies, as this does appear to have an impact.

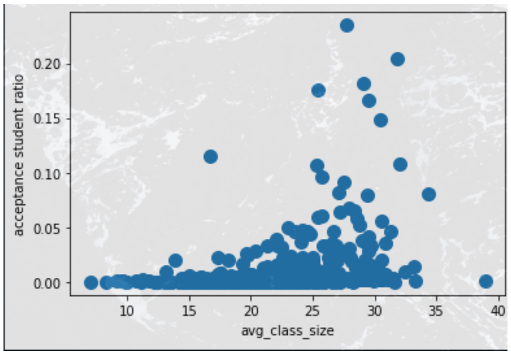


Figure 14: Correlation between average class size and acceptance student ratio

Q7: Proportion of schools that account for 90% of NYCSPHS admission

First we sort the list from max to min. We then iterate over the list and add the acceptances to a partial sum, and every iteration add 1 to a school counter and add the acceptances of the school to partial sum and compare the partial sum / total sum to 0.9. If we are greater than, we stop. We then take that number of schools and divide by the total number of schools to get our proportion. We get **0.207 as our proportion of schools, being 123 schools.**

Q8: Ridge regression model predicting acceptance to NYCSPHS and student achievement

For our model, we will use prediction to determine the most impactful factors for sending students to NYCSPHS and student achievement. We employ **ridge regression** as our model because we have many factors and require a penalty to allow better cross validation. We choose this over LASSO because when tuning the parameter for lasso, we found the alpha that minimizes MSE was actually 0, meaning our data was strange. We tuned the ridge regression parameter and found that the **alpha that minimizes MSE is 1.**

We first use regression to predict acceptances. We note that we use acceptance student ratio to control for school size and do not include school size in our model. We note that this data is standardized and will be providing standardized regression coefficients (SRC) for features.

We find that **applications have the largest SRC of 0.67**, followed by **proportion of math scores exceeded with SRC of 0.31** then followed by a **class size with an SRC of 0.073** then by s**upportive environment with SRC 0.067**. We also take special note of poverty having SRC -0.40. We note this model has a **R^2 of 0.52**.

We now turn to objective student achievement. We see that a **supportive environment has the largest SRC with 0.41**, followed by average **class size with 0.21**, then by **rigorous instruction with 0.12**. We note this model has a **R^2 of 0.45.** We also note that percentage of students in **poverty has a SRC of -0.50**

Q9: Ridge regression analysis

Based on our model, we find that the school characteristics that tend to be the most important for predicting student acceptances to NYCPSHS are acceptances, average class size, supportive environment, and math scores. With regards to objective achievement, our model indicates that a supportive environment is the strongest variable, as well as special importance being placed on class size and rigorous instruction. We see that charter schools have a statistically significant difference in distribution on student standardized test scores. We also note that although we did not test charter schools for acceptance to NYCSPHS, this warrants looking into as it does seem to have an impact on student achievement. We note that in both cases, the poverty percentage of the student population was an important factor for student success.

Q10: Summary

We look for actionable policy to help improve a school’s objective achievement measures and acceptance to student ratio. We note that although applications have the highest SRC, there is little evidence that this could be causal, as we note that the students are more likely to be prepared and high achievers actually submit applications as opposed to their less prepared peers. We see overlap in average class size and having a supportive environment, which increases intimacy between students, their peers, and teachers. Class size however seems to be positively associated with outcomes, and this is unclear as to reasoning and could be erroneous and warrants deeper study. This combination however could assist in both raising scores on objective measures of achievement as well as admissions to NYCSPHS. We also note that math scores were present in both, however were weighted heavily in acceptance to NYCSPHS. A strong math curriculum should be considered by any school which wants to increase its student acceptances to NYCSPHS. We take special note of poverty, which, while not something immediately under the jurisdiction of schools to solve, was a very impactful factor in our model and should be considered as something to solve. Given its incredible weight in our model, we note that reducing poverty in the student population should be of utmost importance to the jurisdiction and is heavily tied to outcomes.