Analysis of New York City High School Data

Yunbin Peng

Introduction

Special High School Admissions Test (SHSAT) is an examination for 8th and 9th grade students in the New York City for admission of the nine specialized high schools such as Stuyvesant High School, High School for Math, Science and Engineering at City College etc. In recent years there is debate of whether those elite high schools have the problem of more homogeneous demographics of admitted students. In this report I used two different datasets to examine high school performance as well as SHSAT registration and admission, and try to identify reasons for underperforming schools. I feel particular interested for this topic since myself is a beneficiary of public education which is a vital tool to boost social mobility.

In particular, I tried to focus on the following four aspects.

1. Out of school quality indicators (rigorous curriculum, school leadership, family involvement etc), which one is the most important for student proficiency in ELA (English Language Art) and Math?
2. Examine schools with very high student absence rate, what are the common features associated with those school such as demographics and economic need?
3. Using both standardized test result, school quality indicators and other variables, is there any groups of schools that desperately need attention for improvement?
4. What are the factors associated with underperforming schools in both SHSAT registration and admission?

Data Source

The main dataset used for this report is the "2016 School Explorer[1]" available on Kaggle provided by PASSNYC, which is a non-profit organization providing outreach service for underrepresented students in New York City. The dataset is in comma-separated values format and contain information in year 2016 for 1272 high schools (including charter schools) such as demographics (percentage of students in each ethnicity), economic need index (a numerical variable which captures temporary housing and eligibility of free lunch program), school quality indicators (numerical variables range between 0 and 1 for factors such as rigorous instruction, supportive environment etc.), as well as number of students achieve more than 4 in proficiency in English Language Arts (ELA) and Math for every grade and every ethnic group. There are 1272 rows, each for one high school and 161 columns in this dataset.

Another supplementary dataset is a comma-separated values file which contains information from each high school for number of students taking SHSAT and number of students receiving offers in year 2017. The information in the dataset is scraped from a New York Times online

[1] The dataset in csv format can be retrieved from https://www.kaggle.com/passnyc/data-science-for-good

article *See Where New York City's Elite High Schools Get Their Students* published in 2018[2] and is available at Kaggle[3]. There are 590 rows (each for one high school) and 8 columns in this dataset and most of variables are numerical.

Question 1

Method

I use school explorer dataset to examine the relationship between students' proficiency and factors of school quality. Since both independent variables and dependent variables (average proficiency in ELA and Math) are continuous numerical variables, linear regression is a suitable method.

There are a total of 6 indicators of school quality: rigorous instruction, collaborative teachers, supportive environment, effective school leadership, strong family and community Ties and trust between students, teachers, parents and administrators. Since each of quality indicator is an index between 0 and 1, there is little need to further convert all variables to similar scale. One major challenge is that each variable is in the format of string and has "%" in it, so I use regular expression to extract number and convert it into integers.

Regarding proficiency in two different subjects (English Language Art and Math), I consider whether it is necessary to run two separate regressions for each subject. The pearson coefficient between average ELA proficiency and average Math proficiency is about 0.94 and they have a linear relationship with slope close to 1, hence I conclude there is no problem to sum them together or to take a simple mean. I add up these two proficiency scores to create a new variable called average proficiency.

Out of 1272 instances in the dataset, there are 55 instances with missing values for either average ELA or Math proficiency. As there are fewer than 5% of instances with missing values, I decide to drop them for the analysis.

Analysis

I run the linear regression where average proficiency as dependent variables and the six quality indicators as independent variables. The model has R-squared of 0.98 which indicates a good fit for a linear model. Coefficients for rigorous instruction, supportive environment and effective leadership are positively and statistically significant from zero. Coefficients of collaborative teachers and strong family ties is close to zero with p-values above 0.85. The most surprising result is that coefficient for trust is negative and significant from zero.

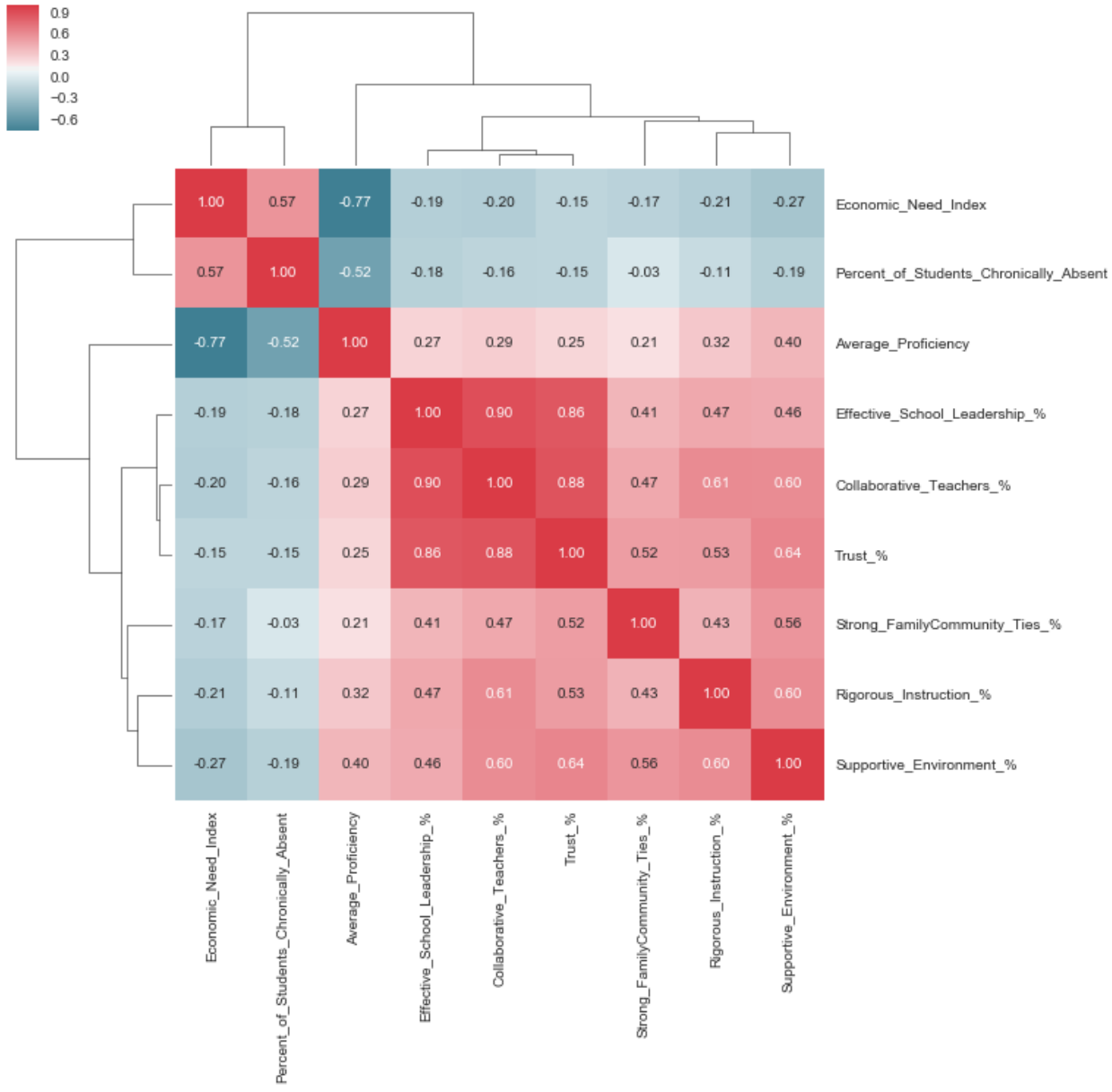
---

[2] https://www.nytimes.com/interactive/2018/06/29/nyregion/nyc-high-schools-middle-schools-shsat-students.html

[3] https://www.kaggle.com/rdisalv2/parsing-nyt-shsat-table/code

I suspect there could be the case of confounding variables and omitted variable bias. If there is a variable that is correlated with both dependent variable and variables included in the regression, omitting that variable from the regression will cause the regression coefficient to fail to capture the marginal effect.

In order to provide more control of regression, I include two additional variables of economic need index and the percentage of students chronically absent (absent for over 10% of school days). I use the following heatmap of correlation coefficient to verify that these two variables are correlated with dependent variables and variables included in the first regression.



I run a different regression of average proficiency against school quality indicators, with control of student family background (economic need index) and motivation (long period of absence). The result from regression is as follows.

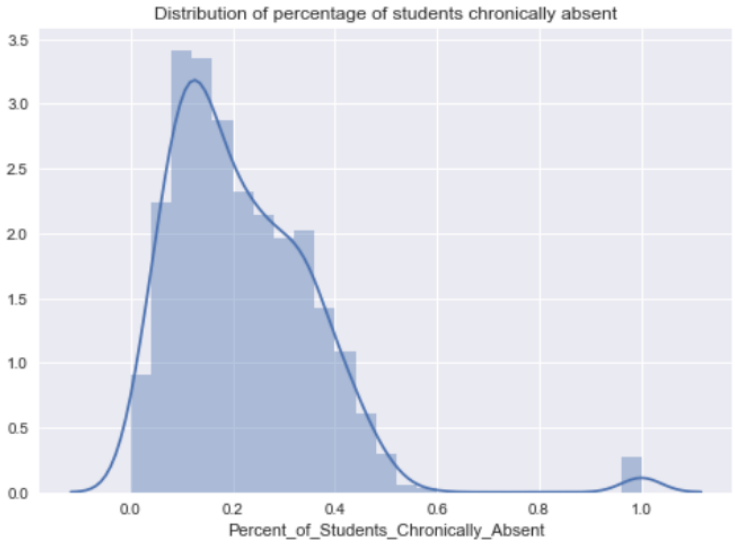
	coef	std err	t	P>|t|	[0.025	0.975]
Rigorous_Instruction_%	1.9998	0.305	6.562	0.000	1.402	2.598
Collaborative_Teachers_%	0.1643	0.617	0.266	0.790	-1.046	1.375
Supportive_Environment_%	2.8388	0.384	7.396	0.000	2.086	3.592
Effective_School_Leadership_%	-1.2926	0.402	-3.211	0.001	-2.082	-0.503
Strong_FamilyCommunity_Ties_%	0.6136	0.310	1.979	0.048	0.005	1.222
Trust_%	3.1556	0.591	5.338	0.000	1.996	4.315
Economic_Need_Index	-2.1049	0.091	-23.198	0.000	-2.283	-1.927
Percent_of_Students_Chronically_Absent	-0.6883	0.139	-4.953	0.000	-0.961	-0.416

The result shows rigorous instruction, a supportive environment and trust between students and teachers are the most important school quality for students' proficiency. Economic hardship and low motivation of students may negatively impact their performance.

Question 2

Woody Allen said 80 percent of success is showing up and previous analysis shows that a school with a higher percentage of student chronically absent tend to have lower performance. I caution from making any causal inference for this aspect, however high percentage can be an indicator of low effort, motivation and morale among students or economic pressure to miss school in order to support families.

Method



Besides school quality indicators, demographics, economics need and income may also affect whether students are absent from schools. I plot a histogram of percentage of chronically absent students. It is right-skewed with a mean and a median around 20%. There is some interesting instance with 100% of absent rate and I will discuss later.

Since I am interested with whether there is significant portion of student chronically absent, rather than the exact level, I use a random forest classification for this question. I create a binary variable of whether a school has over 20% of student absent for more than 10% of school days and use it as dependent variables.

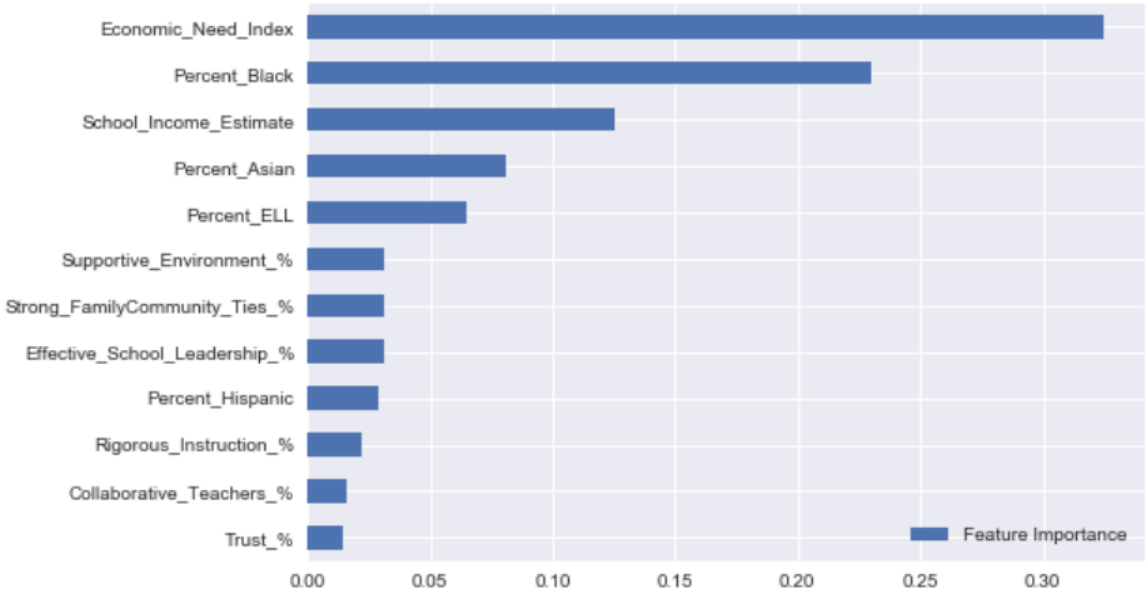
There are 396 schools without estimated average income, I plot the distribution of chronically absent rate for these schools and found it is very similar to the overall distribution. That means dropping them from the analysis will not skew the result in any particular direction. Hence I feel it is safe to drop those instances with missing values.

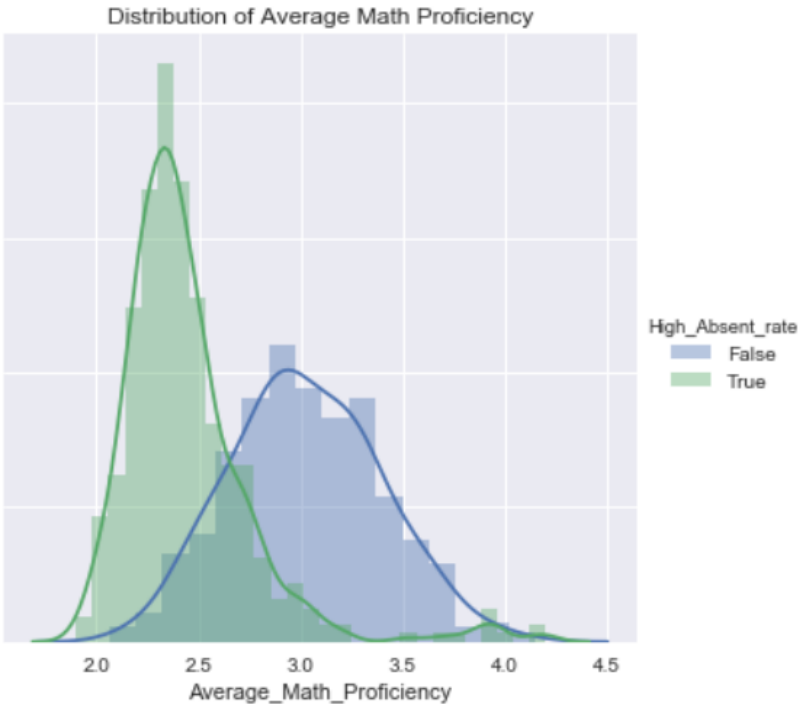
After cleaning the data, I run random forest classification with cross-validation and find the optimal hyperparameter for the classifier. Then I run random forest with these hyperparameter and compute the feature importance.

Analysis

From the visualization of feature importance of random forest classification below, we can see economic background (economics need index and estimated average income) is the most important factor for whether a school has severe problem of student absence. Demographics especially percentage of minority and ELL (English Language Learners) is also important. In contrast, school quality plays a relatively less significant role. To provide help for school troubled

with high student absent rate, it is recommended to address the root problem of economic hardship instead of provide remedy measures such as improving school quality. In fact, a high absence rate basically means students miss the opportunities to benefit from quality education.



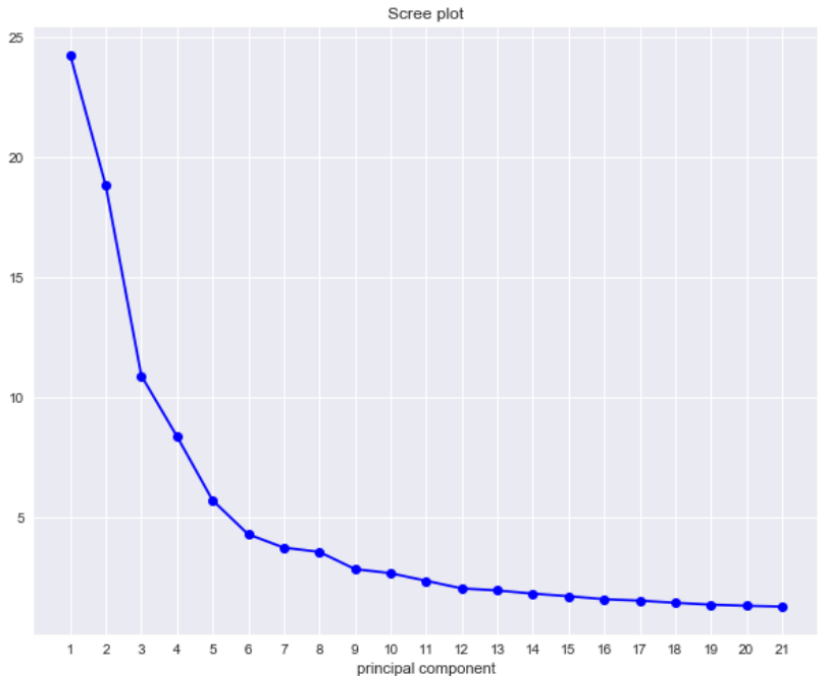


The histogram shows how high percentage of student chronically absent negative impact average proficiency. The distribution of average ELA proficiency is similar. One interesting finding is that the distribution for school with over 20% chronic student absence rate has a little dump on the right tail, representing school with high absent rate but very good test result. There are a total of 10 schools with 100% of students missing over 10% of school days and they are all branches of Success Academy Charter School. Majority of students (over 90%) are from underrepresented groups. These schools have very good school quality score and average ELA proficiency is around 3.5 and average Math proficiency is around 3.9. I am unable to provide a plausible explanation for this phenomenon.

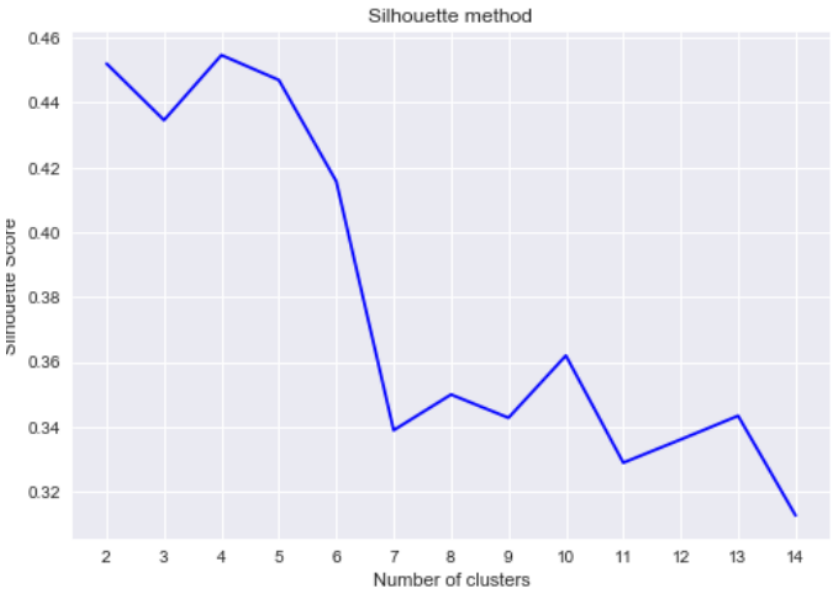
Question 3

Method

For this question I use a dimensionality reduction method to get a low dimensional representation of 137 numerical variables (demographics, school quality and standardized test performance for each grade) via principal component analysis (PCA), then use the PCA transformed data to do a k-means clustering.



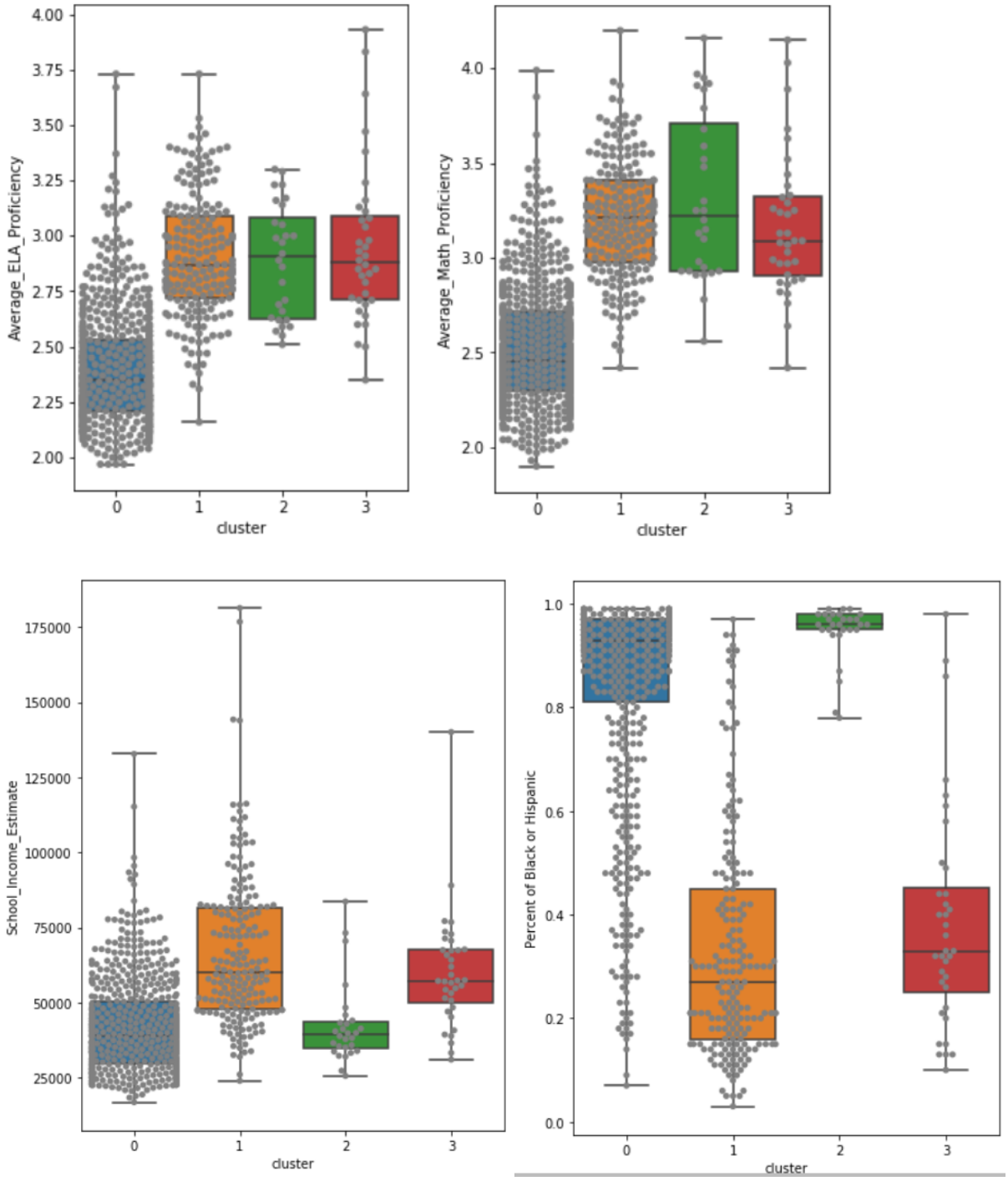
As usually there are 396 instances with missing values being dropped from the analysis (justification of this treatment has been discussed in question 2). For each column, I standardize the values to make the column mean to be 0 and standard deviation to be 1. After the standardization, I use PCA with scree plot to decide to keep the first 5 components as a low dimensional representation of the whole data which is used for k-means clustering.



The main challenge for this part is to determine the optimal number of clusters for k-means clustering. There is no significant “elbow” for the elbow method. With the silhouette method I decide 4 is a reasonable number of clusters. I apply the k-means algorithm with 4 clusters and append the cluster label to the data.

Analysis

K-means algorithm with four clusters gives clusters with size of 571, 193, 28 and 36.



From the boxplot and swarmplot, we can see cluster 0 contains schools with lower average student proficiency and need the most attention. Schools in that cluster are mostly associated with students from low-income family and underrepresented groups. However, it is interesting to

note that cluster 2 also have students coming from similar backgrounds but they achieve significant better proficiency than schools in cluster 0, some examples are the Success Academy Charter School discussed in the previous section, hence it is possible good instruction may offset disadvantage due to family background.

Unfortunately, the clustering method did not do a good job identifying which schools need the most urgent attention as cluster 0 contains over 500 schools, and with experiment of using more clusters show little improvement in the result. One possible reason is that although the data is high dimensional with over 100 variables, most variables are highly correlated. Nevertheless, clustering still captures some common features associated with schools needing help and can be exploited as initial screening when directing public assistance to underperforming schools.

Question 4

Method

The main variables of interest from the 2017 SHSAT data is number of students taking the test and number of students receiving admission from specialized high schools after taking SHSAT. One major challenge encountered for this section is missing values. If there are fewer than 6 students in the two variables, no exact number is available. I suspect the missing values are intentional in order to protect student privacy. This problem is less significant for the number of students taking the test (52 out of 589 schools) but much more problematic for the number of students receiving admission (469 out of 589).

Due to this complication, I decide not to use thepercentage of students receiving admission out of students taking SHSAT as a performance measure. For example, if there are only a few students taking the test, having 0 to 5 students getting admission can sway the percentage by a large extent.

Compare to percentage, the absolute value is relatively less prone to missing values since it is bounded between 0 and 5. I decide to replace all the missing values with 0 and avoid methods that estimate marginal effect for one additional student (i.e regression).

After cleaning the SHSAT data, I perform an inner join with the 2016 School Explorer data to add school information. There are 580 schools in the combined data. I perform exploratory data analysis with mainly scatter plots to examine the interaction between different variables.

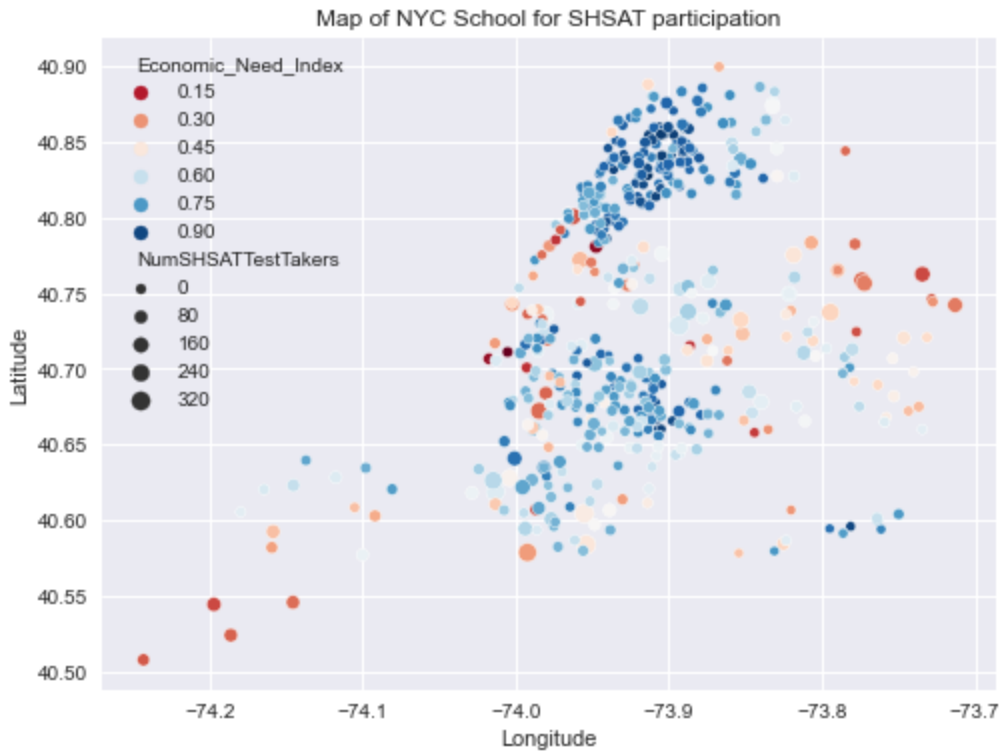
Analysis

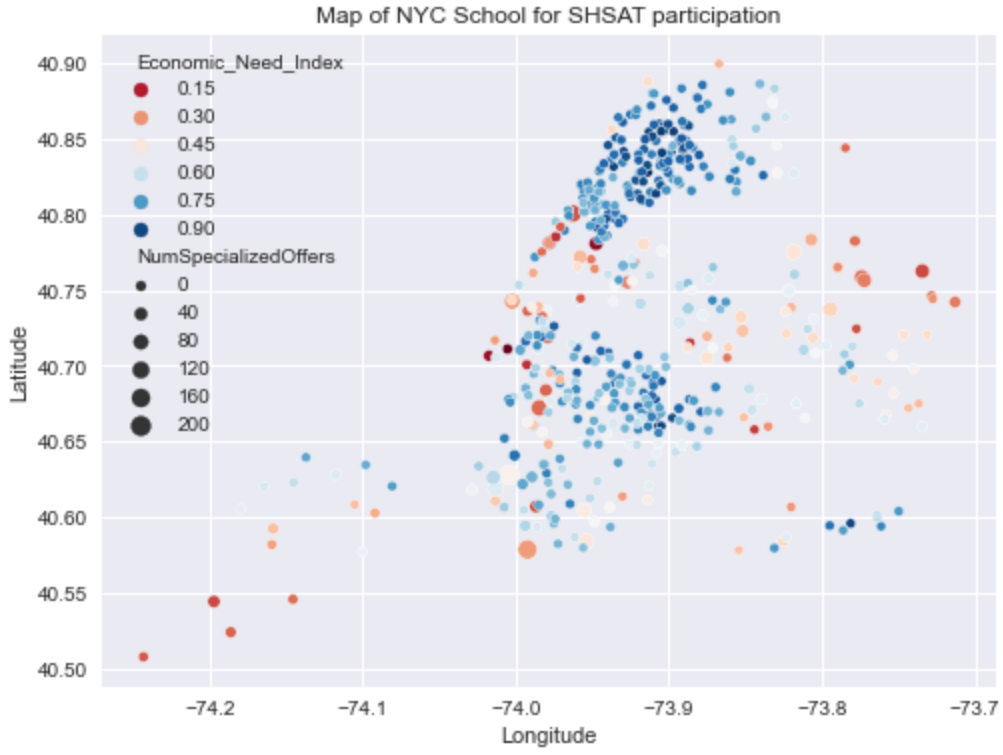
For the first visualization, I plot the geographical distribution of high schools with the number of students receiving admission after taking SHSAT in 2017. What stands out is that admission is concentrated within a small fraction of high schools, mainly in south Manhattan, Queens, and southeast of Brooklyn.

Naturally, a school with better student proficiency in grade 8 should have more students taking the SHSAT. I use a scatter plot by putting proficiency on the horizontal axis and the number of students taking SHSAT on the vertical axis. There is a positive correlation between the two variables, and that confirms the claim. The schools on the right bottom corner should be considered as the ‘diamond’ where there are many students with good proficiency but do not register for SHSAT, I believe PASSNYC should have more outreach programs and encourage registration for those schools.

However, having a high number of students taking SHSAT does not automatically translate into a high number of students receiving admission to specialized high schools. There are schools with over 100 students taking exams but with fewer than 6 students receiving admission. We can see this issue is more significant for schools with high economic needs or a majority of minority groups.


Map of NYC School for SHSAT participation
Latitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude
City
BROOKLYN
NEW YORK
BAYSIDE
FLUSHING
LITTLE NECK
MASPETH
LONG ISLAND CITY
GLENDALE
WHITESTONE
FOREST HILLS
REGO PARK
ASTORIA
STATEN ISLAND
ELMHURST
WOODSIDE
FLORAL PARK
BRONX
ROCKAWAY PARK
EAST ELMHURST
JACKSON HEIGHTS
JAMAICA
OZONE PARK
MIDDLE VILLAGE
QUEENS VILLAGE
BELLEROSE
ARVERNE
BROAD CHANNEL
ROOSEVELT ISLAND
SAINT ALBANS
FAR ROCKAWAY
CAMBRIA HEIGHTS
SPRINGFIELD GARDENS
ROSEDALE
HOWARD BEACH
SOUTH OZONE PARK
RIDGEWOOD
CORONA
HOLLIS
NumSHSATTestTakers
0
80
160
240
320


Map of NYC School for SHSAT participation
Economic_Need_Index
0.15
0.30
0.45
0.60
0.75
0.90
NumSHSATTestTakers
0
80
160
240
320
Latitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude


Map of NYC School for SHSAT participation
Economic_Need_Index
0.15
0.30
0.45
0.60
0.75
0.90
NumSpecializedOffers
0
40
80
120
160
200
Latitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude