PASSNYC: Data Science for Good Challenge

Help PASSNYC determine which schools need their services the most

I. Definition

Project Overview

PASSNYC is a not-for-profit organization that facilitates a collective impact that is dedicated to broadening educational opportunities for New York City's talented and underserved students. New York City is home to some of the most impressive educational institutions in the world, yet in recent years, the City's specialized high schools - institutions with historically transformative impact on student outcomes - have seen a shift toward more homogeneous student body demographics.

PASSNYC uses public data to identify students within New York City's under-performing school districts and, through consulting and collaboration with partners, aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT). By focusing efforts in under-performing areas that are historically underrepresented in SHSAT registration, we will help pave the path to specialized high schools for a more diverse group of students.

What is the Specialized High School Admissions Test (SHSAT) exactly?

The Specialized High School Admissions Test (SHSAT) is the only criterion for admissions to eight of the nine New York City Specialized High Schools. The only exception is the Fiorello H. LaGuardia High School of Music & Art and Performing Arts, which requires an audition or portfolio for admission.

The SHSAT is administered by the New York City Department of Education and is only available to New York City residents in the 8th grade. 9th grade students may also choose to take the 9th grade version of the SHSAT for a very limited number of seats that may become available at the Specialized High Schools.

The maximum score is 800, and Mathematics and English Language Arts (Verbal) are weighted equally. The Specialized High Schools that require the SHSAT are:

Bronx High School of Science

Brooklyn Latin School

Brooklyn Technical High School

High School for Math, Science and Engineering at City College

High School for American Studies at Lehman College

Queens High School for Sciences at York College

Staten Island Technical High School

Stuyvesant High School

In 2016, approximately 28,000+ students took the SHSAT; less than 20% of those students were accepted to a New York City Specialized High School

English Language Learners (ELLs) taking the SHSAT are granted extended testing time (2.0x standard testing time). Bilingual mathematics glossaries will also be provided by the NYCDOE on the day of the SHSAT at each test administration site in the NYCDOE’s nine major languages: Arabic, Bengali, Chinese (Traditional and Simplified), French, Haitian-Creole, Korean, Russian, Spanish, and Urdu. Note: Students are not permitted to bring their own bilingual mathematics glossaries.

Problem Statement

PASSNYC and its partners provide outreach services that improve the chances of students taking the SHSAT and receiving placements in these specialized high schools. The current process of identifying schools is effective, but PASSNYC could have an even greater impact with a more informed, granular approach to quantifying the potential for outreach at a given school. Proxies that have been good indicators of these types of schools include data on English Language Learners, Students with Disabilities, Students on Free/Reduced Lunch, and Students with Temporary Housing.

Part of this challenge is to assess the needs of students by using publicly available data to quantify the challenges they face in taking the SHSAT. The best solutions will enable PASSNYC to identify the schools where minority and underserved students stand to gain the most from services like after school programs, test preparation, mentoring, or resources for parents.

Metrics

To better evaluate the need of students at different schools and help PASSNYC prioritize its outreach effort, I will create a simple index score, the Underrepresentation Score, will be assigned to each middle school to quantify how likely students at a given school are underperforming during SPHS application process as the origin in the plot represents middle school without Hispanic, Black and low-income student.

Underrepresentation Score = Scaled Euclidean distance between a given point and the origin

Underrepresentation Score closer to 1 indicates high level of underrepresentation at SPHS. Please see Table 1 in the appendix for the full list of middle schools with their corresponding Underrepresentation Scores.

II. Analysis

Data Exploration

- **School explorer file** dataset has 1272 samples with 161 features each.
- **The District 5 SHSAT file** dataset has 140 samples with 7 features each.
- Below you can find a profiling report of our main data set – school explorer

myoutputfile.html

School explorer file - As the last 120 variables all show numbers by grade for specific groups, I am only showing a glimpse of the first 50 variables below. Variables 50-161 are all variables similar to the ones at the bottom (such as Grade3ELA_Black4s), and basically specify the number of high performing students for each Grade by group (the number of "Level 4" students, which is the highest level in New York). These groups are discussed in detail in section 6.3 (Analysis of the Grade 7 numbers in the school explorer).

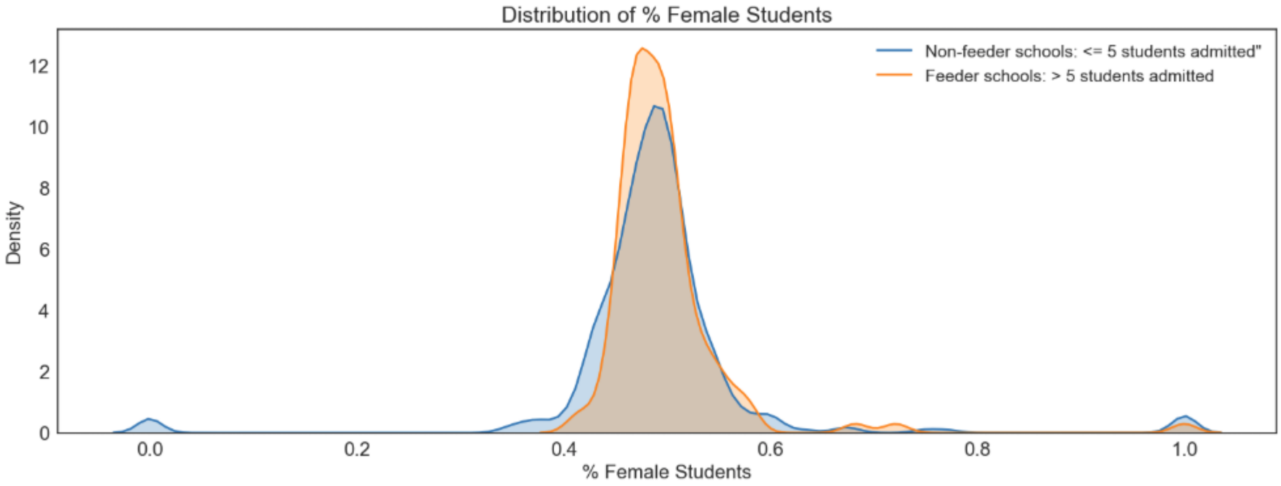
The District 5 SHSAT file - Altogether, we have school data from 1270 schools and unfortunately only SHSAT (time series) data for 28 schools in District 5 (Central Harlem).

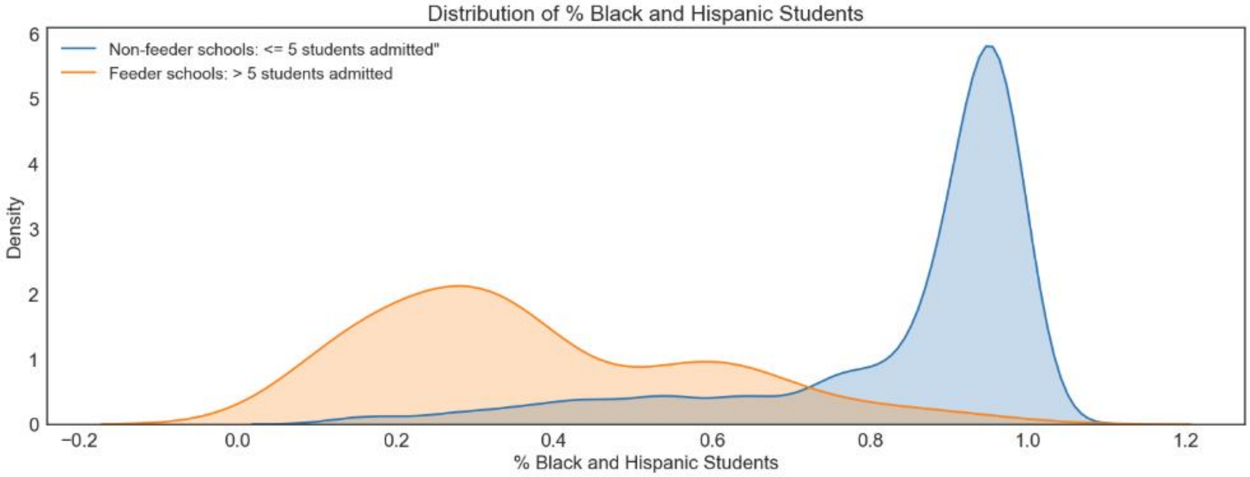
In addition, I've added few other data sets (CSVs) also published by open data NYC.

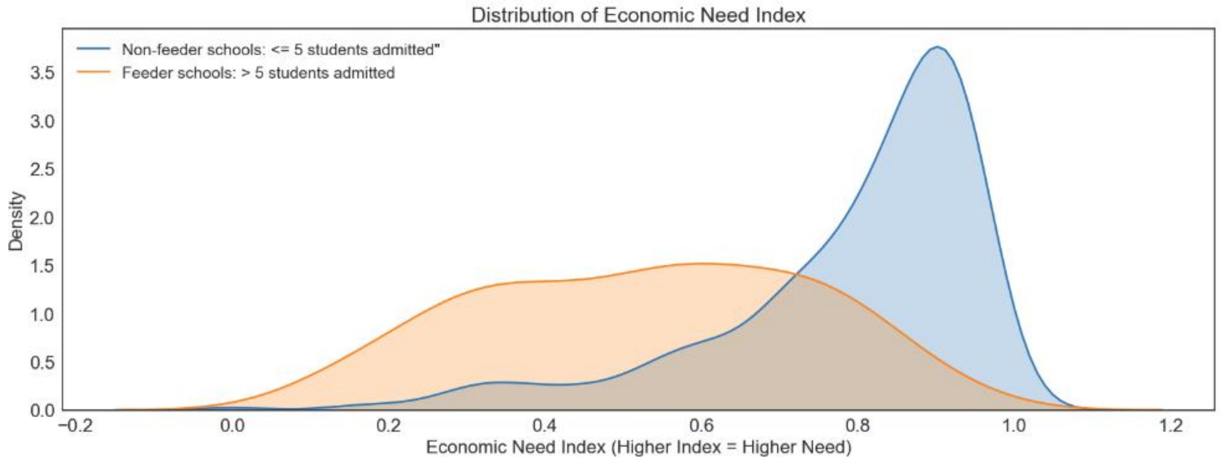
1. (NYC Open Data) Middle school directory
2. (NYC Open Data) Offers received by students in different middle schools
3. (NYC Open Data) Student composition
4. (NYC Open Data) State test result
5. (NYC Open Data) Average class size
6. (NYC Open Data) Pupil to teacher ratio
7. (NYC Open Data) School district demo breakdown
8. (NYC Open Data) High school directory

Exploratory Visualization

To examine whether these feeder schools where offers concentrate are compositionally different from other schools, the following density plots illustrate the distribution of three metrics: % female students, % Black and Hispanic students, and Economic Need Index (calculated as % temp housing + % HRA eligible * 0.5 + % free lunch eligible * 0.5).







- Figure 1 shows that the gender ratio of feeder schools and non-feeder schools display similar distribution centered at ~50%. Similarity in gender ratio across schools make it difficult to study the middle school impact using school-level data alone.

- Figure 2 and Figure 3 show that feeder schools and non-feeder schools are strikingly different in terms of their racial composition and economic need level.

- The following section, the last two metrics will be use to calculate a scaled score to measure how likely students from a given middle school will be underrepresented at SPHS.

Underrepresentation Score

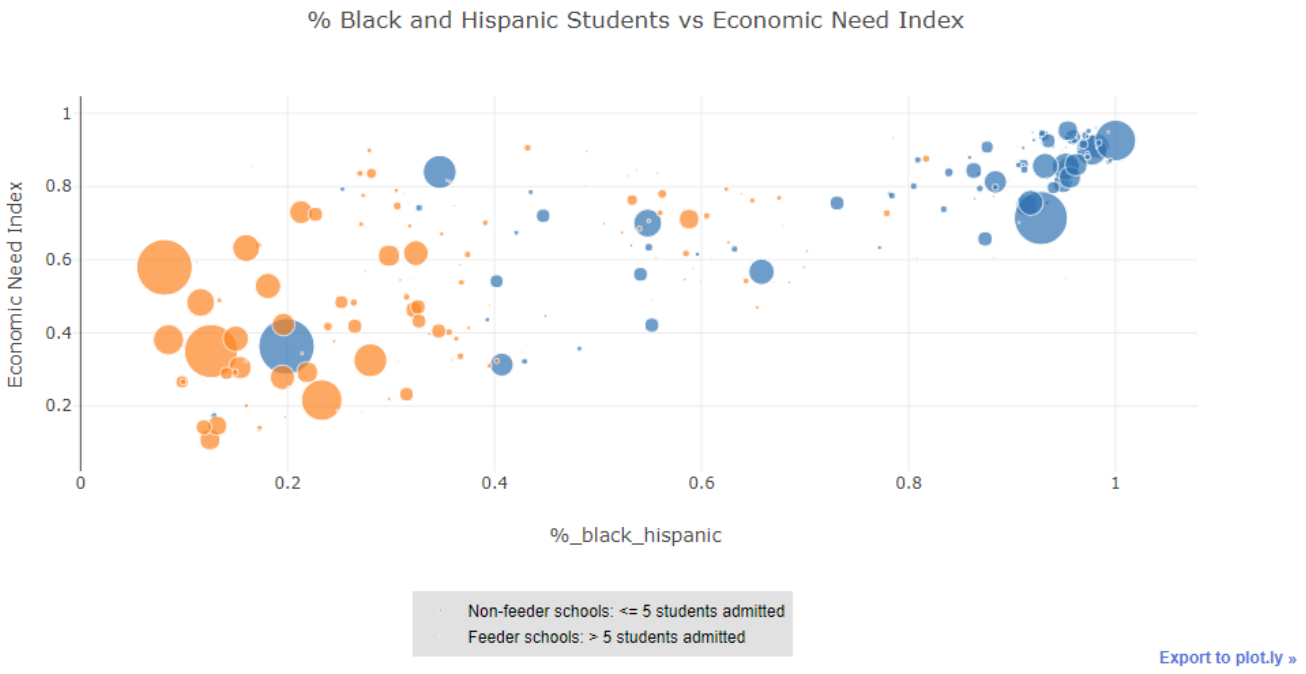
- The scatterplot above visualizes the positive correlation (r = 0.77, p < 0.01) between Economic Need Index and % Black and Hispanic students.

- Feeder schools with more students admitted to SPHS (larger-sized orange points) tend to have low-to-medium economic need and lower proportion of Black or Hispanic students, while a noticeable number of non-feeder schools (blue points) cluster around the upper right corner of the plot. In other words, students from these non-feeder schools are mostly low-income Hispanic or Black student that are underrepresented at SPHS.

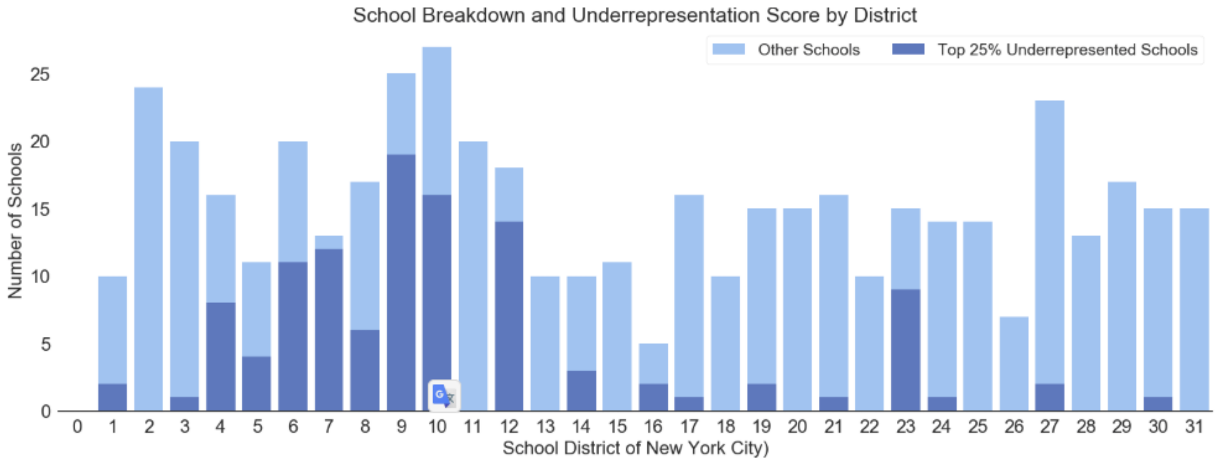
- To better evaluate the need of students at different schools and help PASSNYC prioritize its outreach effort, I will create a simple index score, the Underrepresentation Score, will be assigned to each middle school to quantify how likely students at a given school are underperforming during SPHS application process As the origin in the plot represents middle school without Hispanic, Black and low-income student.

- Underrepresentation Score = Scaled Euclidean distance between a given point and the origin

- Underrepresentation Score closer to 1 indicates high level of underrepresentation at SPHS. Please see Table 1 in the appendix for the full list of middle schools with their corresponding Underrepresentation Scores.



- Top 25% schools with the highest Underrepresentation Scores represent 20 out of 31 school districts in New York City, and districts with higher index scores tend to have higher proportion of underrepresented middle schools.



- The SPHS of New York City have a long history of supporting the educational needs of students with strong academic or artistic performance. Eight out of nine SPHS admit students based on a single entrance exam, SHSAT. As the sore means of admission to the city's most prestigious high schools, SHSAT tests for students' abilities in English and Math.

Assuming that students from middle schools with stronger academic performance are more likely to be motivated to apply and be admitted to SPHS, this section will focus on academic performance, and aims to answer one question through k-means clustering: which non-feeder schools have education quality and academic performance similar to those of feeder schools?

Algorithms and Techniques

- The k-means problem is solved using Lloyd's algorithm. The average complexity is given by O(k n T), were n is the number of samples and T is the number of iteration. The worst case complexity is given by O(n^(k+2/p)) with n = n_samples, p = n_features. (D. Arthur and S. Vassilvitskii, 'How slow is the k-means method?' SoCG2006) In practice, the k-means algorithm is very fast (one of the fastest clustering algorithms available), but it falls in local minima. That's why it can be useful to restart it several times.

- Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.
  Each observation starts in its own cluster, and clusters are successively merged together. The linkage criteria determine the metric used for the merge strategy:

  Ward minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

Benchmark

To create an initial benchmark for the clustering, I used the winner solution in Kaggle competition (Erik Bruin).

- My goal was to find at least 5 schools that similar to Erik soluation and need the PASSNYC help.

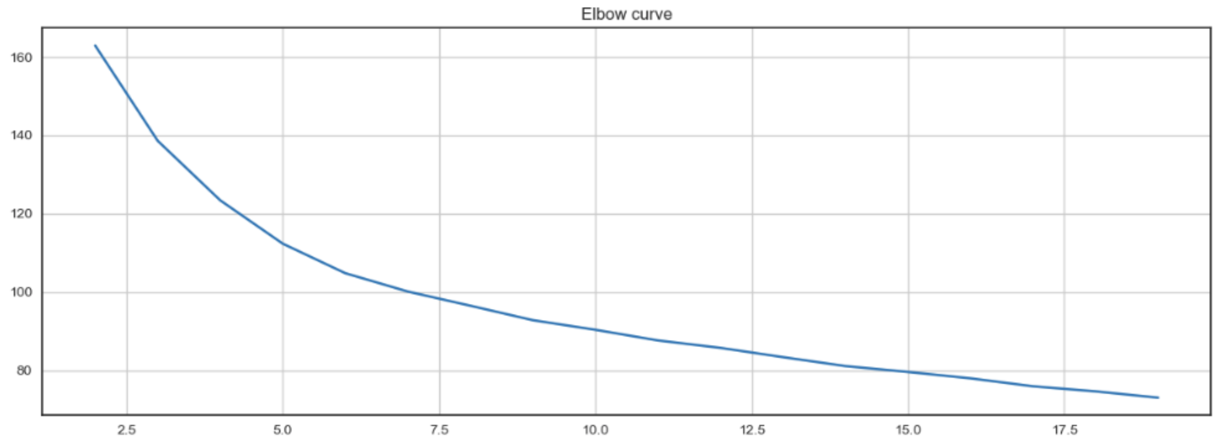
III. Methodology

Data Preprocessing

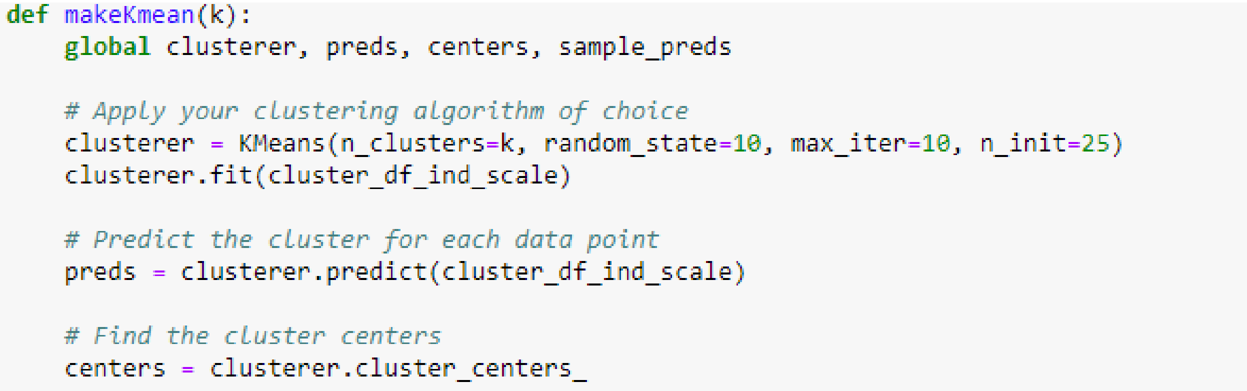
- Clean unnecessary characters, format string columns to integer and floats in the school explore file.
- Remove correlated column according to the profiling report in the school explore file.
- Create diversity table from all the data sets that will use us during the analysis.

Implementation

- Reaching final data set for processing (cluster_df)
- Scaling the data with Min-Max scalar
- Use the elbow method in order to determine the best number of clusters for the analysis



- Build a function that will run the clustering and can be reused.



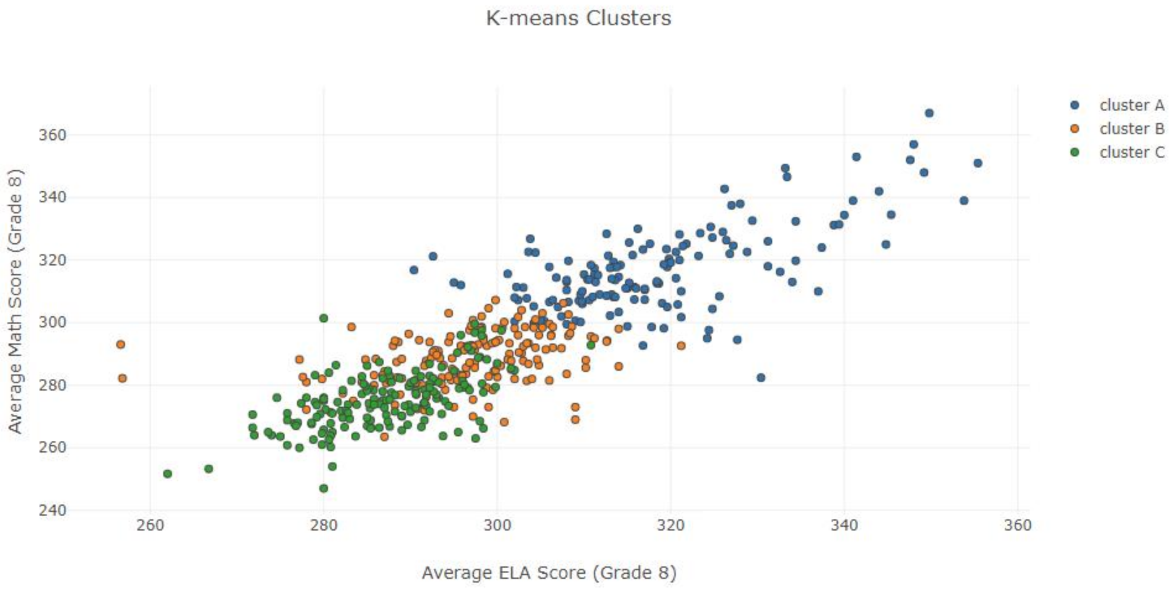
```
def makeKmean(k):
    global clusterer, preds, centers, sample_preds

    # Apply your clustering algorithm of choice
    clusterer = KMeans(n_clusters=k, random_state=10, max_iter=10, n_init=25)
    clusterer.fit(cluster_df_ind_scale)

    # Predict the cluster for each data point
    preds = clusterer.predict(cluster_df_ind_scale)

    # Find the cluster centers
    centers = clusterer.cluster_centers_
```

Refinement

- After trying few number of clusters and few method of clustering I found that K-means and hierarchical clustering perform best.

IV. Results

Model Evaluation and Validation



Cluster name	Number of Schools	Avg Math Score	Avg Ela Score	Avg Student Attendance Rate	Average Class Size	Diversity Score	Feeder Schools (# of Students)
A	145	318	317	95.80 %	17.70	46.49 %	3541
B	174	296	287	93.05 %	14.14	80.29 %	268
C	153	287	275	90.72 %	12.14	88.51 %	64

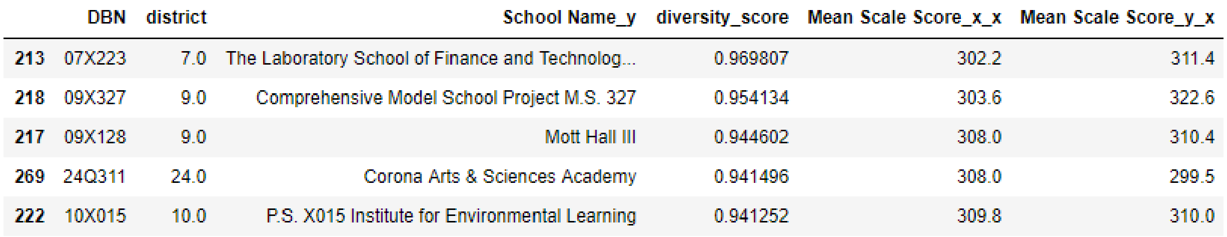
***Cluster A*: Most likely to have students qualified for SPHS**

***Cluster B*: Somewhat likely to have students qualified for SPHS**

***Cluster C*: Least likely to have students qualified for SPHS**

Recommendations to PASSNYC

- Cluster A and B, indicating high level of education quality and academic potential. Students from these middle schools are more like to be qualified for and therefore be motivated to apply for SPHS, and should take higher priority in PASSNYC's outreach effort.



	DBN	district	School Name_y	diversity_score	Mean Scale Score_x_x	Mean Scale Score_y_x
213	07X223	7.0	The Laboratory School of Finance and Technolog...	0.969807	302.2	311.4
218	09X327	9.0	Comprehensive Model School Project M.S. 327	0.954134	303.6	322.6
217	09X128	9.0	Mott Hall III	0.944602	308.0	310.4
269	24Q311	24.0	Corona Arts & Sciences Academy	0.941496	308.0	299.5
222	10X015	10.0	P.S. X015 Institute for Environmental Learning	0.941252	309.8	310.0

- Cluster A represents middle schools that are most likely to have students qualified for SPHS.

- Of all middle schools in the top quantile of the Underrepresentation Score, 5 schools are listed above are assigned to Cluster A, but none of them have more than 5 students accepted by SPHS in 2018. Students from these schools are likely to benefit the most from services provided by PASSNYC and its partners, and these schools should take the highest priority in PASSNYC's plan.

- Same idea for Cluster B below – **showing the full table in the python notebook**.

Justification

- The winning solution has picked 5 schools that similar to my results (Erik Bruin).

V. Conclusion

Reflection

The process used for this project can be summarized using the following steps:

1. An initial problem and relevant, public datasets were found

2. The data was downloaded and preprocessed (clean and formats)

3. A benchmark was chosen

4. The clustering was preformed using the data

5. Conclusions and recommendations was drafted from the results

I found steps 2 and 3 the most difficult, as I had to familiarize myself with the files.

Improvement

- Run More clustering methods
- Plot geolocation information that can help PASSNYC to come up with conclusions