Capstone 1 PASSNYC Data- Milestone Report

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 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.

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.

PASSNYC is conducting a Datascience for Good challenge in collaboration with Kaggle

**Dataset:**

PASSNYC has shared 2016 School Explorer data which has all the demographic and other relevant data about the NYC schools, contained in 1272 Rows and x 161columns

In addition to that, it has also shared District 5‘s SHSAT Registrations and Testers details in

140\*7 R \* C

In addition to the above I am also using publicly available data from City of New York’s publicly available data at NYC data, which has SHSAT offer details for three academic years

**Data Wrangling:**

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.

Datasets that are available for this project are

1) School explorer dataset with demographic and academic details of students in all NYC public schools

2) SHSAT data set that has details of the count of students schools, the number that enrolls for tests, count that take the test and that count that gets offer.

In order to aid PASSNYC, the data sets have to be explored in detail to check whether any trends or patterns can be observed that would help identify the schools that need help

1. **Cleaning and Consolidating the Data**

In order to consolidate PASSNYC data from different years (and different file formats), column headers were standardized. Some unnecessary attributes were dropped to reduce dimensionality. A couple of new columns were introduced, to calculate scores based on values of some feture variables, which would eventually become the labels for the Machine Learning Algorithm. For some columns, the value was categorization with words like ‘ Meeting Target’, Approaching Target, ‘ Exceeded Target’. That was converted to numeric ranking ‘1’,’2’,’3’.

1. **Missing Values**

Some columns in the dataset had missing values and there were a few inconsistencies in notation that were adjusted for ease of future analysis. Dropna and fillna were used to drop if the numbere of rows were insignificant to the data and to replace using mean value as applicable for each of the specific case.

1. **Outliers**

There were not much significant outliers to be worked upon

**Data Visualization and Inferential Statistics**

'School Explorer' dataframe has a huge set of feature variables and doing a thorough EDA is key to finding out , which of this have the maximum impact on the smiddle school's getting their students admitted to the Specialised High Schools(SPHS). Recent studies/ news paper reports show that the 3 highest-status schools—Stuyvesant, Bronx Science, and Brooklyn Techin SPHS —have black and Latino student populations of 4, 9, and 13 percent, respectively, far below the 70 percent in public schools citywide, which corroborates with PASSNYC's observation . The task is in finding out which feature variables are key in helping shift focus to the right set of schools.

Given below are the feature variables:

Economic Need Index float64

School Income Estimate object

Percent ELL object

Percent Asian object

Percent Black object

Percent Hispanic object

Percent Black / Hispanic object

Percent White object

Student Attendance Rate object

Percent of Students Chronically Absent object

Rigorous Instruction % object

Rigorous Instruction Rating object

Collaborative Teachers % object

Collaborative Teachers Rating object

Supportive Environment % object

Supportive Environment Rating object

Effective School Leadership % object

Effective School Leadership Rating object

Strong Family-Community Ties % object

Strong Family-Community Ties Rating object

Trust % object

Trust Rating object

Student Achievement Rating object

Average ELA Proficiency float64

Average Math Proficiency float64

Grade 3 ELA - All Students Tested int64

Grade 3 ELA 4s - All Students int64

Grade 3 ELA 4s - American Indian or Alaska Native int64

Grade 3 ELA 4s - Black or African American int64

Grade 3 ELA 4s - Hispanic or Latino int64

Grade 3 ELA 4s - Asian or Pacific Islander int64

Grade 3 ELA 4s - White int64

Grade 3 ELA 4s - Multiracial int64

Grade 3 ELA 4s - Limited English Proficient int64

Grade 3 ELA 4s - Economically Disadvantaged int64

Grade 3 Math - All Students tested int64

Grade 3 Math 4s - All Students int64

Grade 3 Math 4s - American Indian or Alaska Native int64

Grade 3 Math 4s - Black or African American int64

Grade 3 Math 4s - Hispanic or Latino int64

Grade 3 Math 4s - Asian or Pacific Islander int64

Grade 3 Math 4s - White int64

Grade 3 Math 4s - Multiracial int64

Grade 3 Math 4s - Limited English Proficient int64

Grade 3 Math 4s - Economically Disadvantaged int64

Grade 4 ELA - All Students Tested int64

Grade 4 ELA 4s - All Students int64

Grade 4 ELA 4s - American Indian or Alaska Native int64

Grade 4 ELA 4s - Black or African American int64

Grade 4 ELA 4s - Hispanic or Latino int64

Grade 4 ELA 4s - Asian or Pacific Islander int64

Grade 4 ELA 4s - White int64

Grade 4 ELA 4s - Multiracial int64

Grade 4 ELA 4s - Limited English Proficient int64

Grade 4 ELA 4s - Economically Disadvantaged int64

Grade 4 Math - All Students Tested int64

Grade 4 Math 4s - All Students int64

Grade 4 Math 4s - American Indian or Alaska Native int64

Grade 4 Math 4s - Black or African American int64

Grade 4 Math 4s - Hispanic or Latino int64

Grade 4 Math 4s - Asian or Pacific Islander int64

Grade 4 Math 4s - White int64

Grade 4 Math 4s - Multiracial int64

Grade 4 Math 4s - Limited English Proficient int64

Grade 4 Math 4s - Economically Disadvantaged int64

Grade 5 ELA - All Students Tested int64

Grade 5 ELA 4s - All Students int64

Grade 5 ELA 4s - American Indian or Alaska Native int64

Grade 5 ELA 4s - Black or African American int64

Grade 5 ELA 4s - Hispanic or Latino int64

Grade 5 ELA 4s - Asian or Pacific Islander int64

Grade 5 ELA 4s - White int64

Grade 5 ELA 4s - Multiracial int64

Grade 5 ELA 4s - Limited English Proficient int64

Grade 5 ELA 4s - Economically Disadvantaged int64

Grade 5 Math - All Students Tested int64

Grade 5 Math 4s - All Students int64

Grade 5 Math 4s - American Indian or Alaska Native int64

Grade 5 Math 4s - Black or African American int64

Grade 5 Math 4s - Hispanic or Latino int64

Grade 5 Math 4s - Asian or Pacific Islander int64

Grade 5 Math 4s - White int64

Grade 5 Math 4s - Multiracial int64

Grade 5 Math 4s - Limited English Proficient int64

Grade 5 Math 4s - Economically Disadvantaged int64

Grade 6 ELA - All Students Tested int64

Grade 6 ELA 4s - All Students int64

Grade 6 ELA 4s - American Indian or Alaska Native int64

Grade 6 ELA 4s - Black or African American int64

Grade 6 ELA 4s - Hispanic or Latino int64

Grade 6 ELA 4s - Asian or Pacific Islander int64

Grade 6 ELA 4s - White int64

Grade 6 ELA 4s - Multiracial int64

Grade 6 ELA 4s - Limited English Proficient int64

Grade 6 ELA 4s - Economically Disadvantaged int64

Grade 6 Math - All Students Tested int64

Grade 6 Math 4s - All Students int64

Grade 6 Math 4s - American Indian or Alaska Native int64

Grade 6 Math 4s - Black or African American int64

Grade 6 Math 4s - Hispanic or Latino int64

Grade 6 Math 4s - Asian or Pacific Islander int64

Grade 6 Math 4s - White int64

Grade 6 Math 4s - Multiracial int64

Grade 6 Math 4s - Limited English Proficient int64

Grade 6 Math 4s - Economically Disadvantaged int64

Grade 7 ELA - All Students Tested int64

Grade 7 ELA 4s - All Students int64

Grade 7 ELA 4s - American Indian or Alaska Native int64

Grade 7 ELA 4s - Black or African American int64

Grade 7 ELA 4s - Hispanic or Latino int64

Grade 7 ELA 4s - Asian or Pacific Islander int64

Grade 7 ELA 4s - White int64

Grade 7 ELA 4s - Multiracial int64

Grade 7 ELA 4s - Limited English Proficient int64

Grade 7 ELA 4s - Economically Disadvantaged int64

Grade 7 Math - All Students Tested int64

Grade 7 Math 4s - All Students int64

Grade 7 Math 4s - American Indian or Alaska Native int64

Grade 7 Math 4s - Black or African American int64

Grade 7 Math 4s - Hispanic or Latino int64

Grade 7 Math 4s - Asian or Pacific Islander int64

Grade 7 Math 4s - White int64

Grade 7 Math 4s - Multiracial int64

Grade 7 Math 4s - Limited English Proficient int64

Grade 7 Math 4s - Economically Disadvantaged int64

Grade 8 ELA - All Students Tested int64

Grade 8 ELA 4s - All Students int64

Grade 8 ELA 4s - American Indian or Alaska Native int64

Grade 8 ELA 4s - Black or African American int64

Grade 8 ELA 4s - Hispanic or Latino int64

Grade 8 ELA 4s - Asian or Pacific Islander int64

Grade 8 ELA 4s - White int64

Grade 8 ELA 4s - Multiracial int64

Grade 8 ELA 4s - Limited English Proficient int64

Grade 8 ELA 4s - Economically Disadvantaged int64

Grade 8 Math - All Students Tested int64

Grade 8 Math 4s - All Students int64

Grade 8 Math 4s - American Indian or Alaska Native int64

Grade 8 Math 4s - Black or African American int64

Grade 8 Math 4s - Hispanic or Latino int64

Grade 8 Math 4s - Asian or Pacific Islander int64

Grade 8 Math 4s - White int64

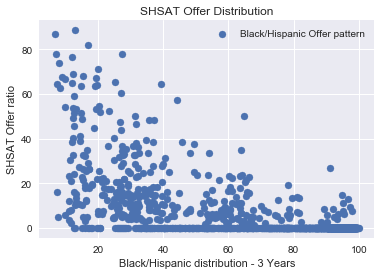
Grade 8 Math 4s - Multiracial int64

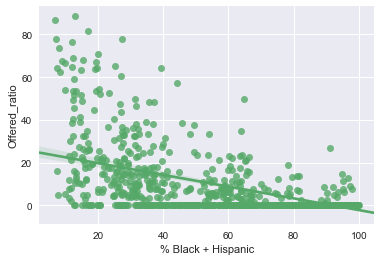
Grade 8 Math 4s - Limited English Proficient int64

Grade 8 Math 4s - Economically Disadvantaged int64

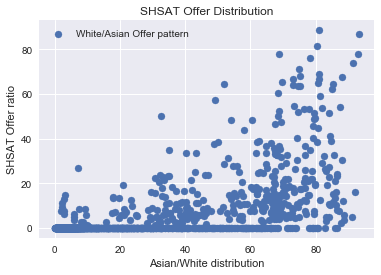
It makes sense to plot a scatter plot between the SHSAT offer ratio against some of the feature variables over the three years to see if there is any co-relation

Given below are some of the scatter plots that I generated. (Please refer the ipynb file for EDA for a comprehensive set)



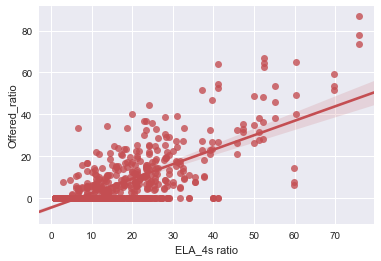


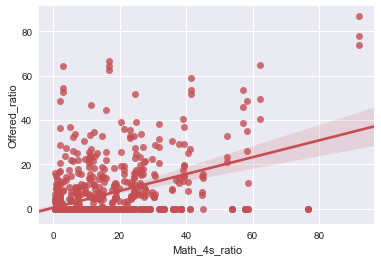
Contrary to my expectation, the correlation co-efficient is only 62%.



Again the Offer ratio plotted against the White Asian population in middle schools also have a correlation coefficient of around 60% only. However it is interesting to note that the offer ratio goes up as the White- Asian population in schools go up, whereas, for the Black Hispaic population, with the increase in population, the offer rate was coming down

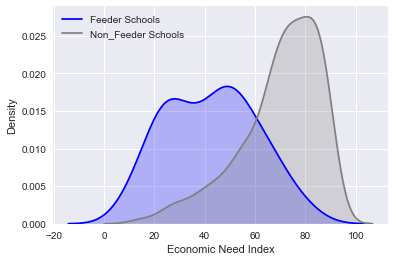
Next I decided to see if there schools with students who scored 4s in their Math and ELA(English Language and Arts) in state exams , had better chances of scoring well in SHSAT exams thereby receiving more offer letters from Specialized High Schools(SPHS).

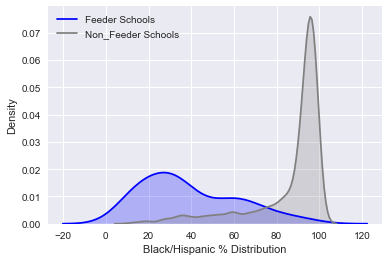


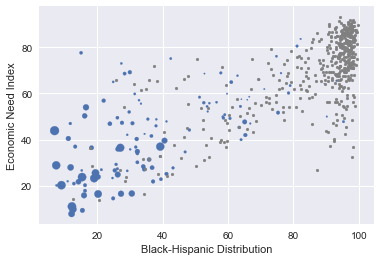


From the scatterplot, though the ELA scores show a somewhat substantial correlation with the Offer Ratio, the Math 4s score does not seem to have a strong correlation with the Offers received from SPHS.

I then categorized the school list into Feeder Schools( schools that have students who qualify the SHSAT exams) and Non Feeder Schools( school where students do not qualify in the test) and plot density graphs with a couple of feature demographics, to see if there is any correlation.







From the above graphs, it can be seen that the feeder schools and non-feeder schools are strikingly different in terms of their racial composition and economic need level. Not only are Asian, Caucasian and high-income students more likely to attend SPHS, but interestingly enough the students attending these middle schools have a higher proportion of Asian and Caucasian students with lower economic need.

The scatterplot above visualizes the positive correlation (r = 0.78) between Economic Need Index and % Black and Hispanic students. Feeder schools with more students admitted to SPHS (larger-sized blue 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 (gray 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.

The above analysis implies that if I assign the schools a performance score based on select features from this data and rank them according to the score, it would help me find the schools that need assistance. And the demographic relations between multiple features and labels from the above EDA indicate, that they would in turn help the low income Black and Hispanic families in the hood who would do well with training programs and benefit from awareness sessions in hood schools.