Specialized High School Scholastic Aptitude Test Admission Performance

Capstone Two: Predictors of test performance for specialized high school admissions offers.

- Jupyter Notebook

Background

Performance on the Specialized High School Admissions Test (SHSAT) determines eligibility to one of the eight specialized high schools in New York City. It is administered by the New York City Department of Education (DOE) to about a third of the city's 8th graders, with **5,000** receiving admissions offers. Of major concern is the racial & ethnic breakdown of admitted students, showing significant underrepresentation from Black and Latinx students.

Data

Two main data sets will be used:

2016 School Explorer (Explorer)

- This dataset consists of 1272 schools in New York city, and 161 variables, provided via Kaggle. Primarily, it's school descriptors, e.g. grades, race & ethnicity student percentages, high/low performing percentages of students. Data is available as a single csv file.
    - Kaggle Dataset (API)

2017-2018 SHSAT Admissions Test Offers By Sending School (Offers)

- This dataset consists of the 2017 SHSAT results, published by NYC in 2018. All test takers are north of 28,000, grade 8 students, Test takers and offers received are grouped by school. Data is available as a single csv file from the NYC Open Data portal.
    - NYC Dataset (CSV)

Approach / Method

The goal of this analysis is to elicit which factors predict performance on the SHSAT. These factors will serve as beacons to direct or draw services, whether education-based or otherwise, towards improving the percentage of Black and Latinx students admitted to the specialized high schools.

This approach aims to quantify which variables lead to admissions offers _beyond_ prior proxies: English Language Learners, Students with Disabilities, Students on Free/Reduced Lunch, and Students with Temporary Housing.

Initially we can assume that those students who perform well on typical standardized tests, throughout the school year, would therefore perform well on the SHSAT. We'll investigate this and extrapolate as to whether this is the case across all schools/students that follow this assumption.

As a bit of forecasting, I'll use linear regression models to determine how many admissions offers schools that fit a certain testing/aptitude standard could be getting based on their current testing scores.

Data Cleaning

To determine what factors are related to receiving admission offers to the specialized high schools, the data feeding into the models need to be not only numeric but free of errors.

We can see that the _2016 School Explorer_ data set has three columns almost entirely of null values. These can be filled with an appropriate value for the data type of those columns.

	Percent
Other Location Code in LCGMS	0.999214
Adjusted Grade	0.998428
New?	0.978774
School Income Estimate	0.311321

Given the test-takers in the *2017-2018 SHSAT Admissions Test Offers By Sending School* are a year away from taking the test in *2016 School Explorer* dataset, I'll focus on the 7th graders.

2016 School Explorer has 20 variables with information 7th graders. This data is broken up into two kinds of information, ELA (English Language Arts) & Math. Scoring on these tests top out at 4, with 1 representing the worst score.

```
Data columns (total 20 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Grade 7 ELA - All Students Tested               1272 non-null   int64
 1   Grade 7 ELA 4s - All Students                   1272 non-null   int64
 2   Grade 7 ELA 4s - American Indian or Alaska Native 1272 non-null   int64
 3   Grade 7 ELA 4s - Black or African American      1272 non-null   int64
 4   Grade 7 ELA 4s - Hispanic or Latino             1272 non-null   int64
 5   Grade 7 ELA 4s - Asian or Pacific Islander      1272 non-null   int64
 6   Grade 7 ELA 4s - White                          1272 non-null   int64
 7   Grade 7 ELA 4s - Multiracial                    1272 non-null   int64
 8   Grade 7 ELA 4s - Limited English Proficient     1272 non-null   int64
 9   Grade 7 ELA 4s - Economically Disadvantaged     1272 non-null   int64
 10  Grade 7 Math - All Students Tested              1272 non-null   int64
 11  Grade 7 Math 4s - All Students                  1272 non-null   int64
 12  Grade 7 Math 4s - American Indian or Alaska Native 1272 non-null   int64
 13  Grade 7 Math 4s - Black or African American     1272 non-null   int64
 14  Grade 7 Math 4s - Hispanic or Latino            1272 non-null   int64
 15  Grade 7 Math 4s - Asian or Pacific Islander     1272 non-null   int64
 16  Grade 7 Math 4s - White                         1272 non-null   int64
 17  Grade 7 Math 4s - Multiracial                   1272 non-null   int64
 18  Grade 7 Math 4s - Limited English Proficient    1272 non-null   int64
 19  Grade 7 Math 4s - Economically Disadvantaged    1272 non-null   int64
```

Therefore, the best students are in the '4s' columns shown above.

Summary of columns:

- All students tested
- All students with 4 scores
- American Indian or Alaska Native with 4 scores
- Black or African American students with 4 scores
- Hispanic or Latino students with 4 scores
- Asian or Pacific Islander students with 4 scores
- White students with 4 scores
- Multiracial students with 4 scores
- Limited English Proficient students with 4 scores
- Economically Disadvantaged with 4 scores

In 2016, the total number of 7th graders in NYC Middle Schools was **69,053**. Of those, **8,320** had ELA scores of 4, and **10,888** had Math scores of 4.

```
Data columns (total 5 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   feeder_school_dbn                     594 non-null    object
 1   feeder_school_name                    594 non-null    object
 2   count_of_students_in_hs_admissions    593 non-null    float64
 3   count_of_testers                      594 non-null    object
 4   count_of_offers                       594 non-null    object
```

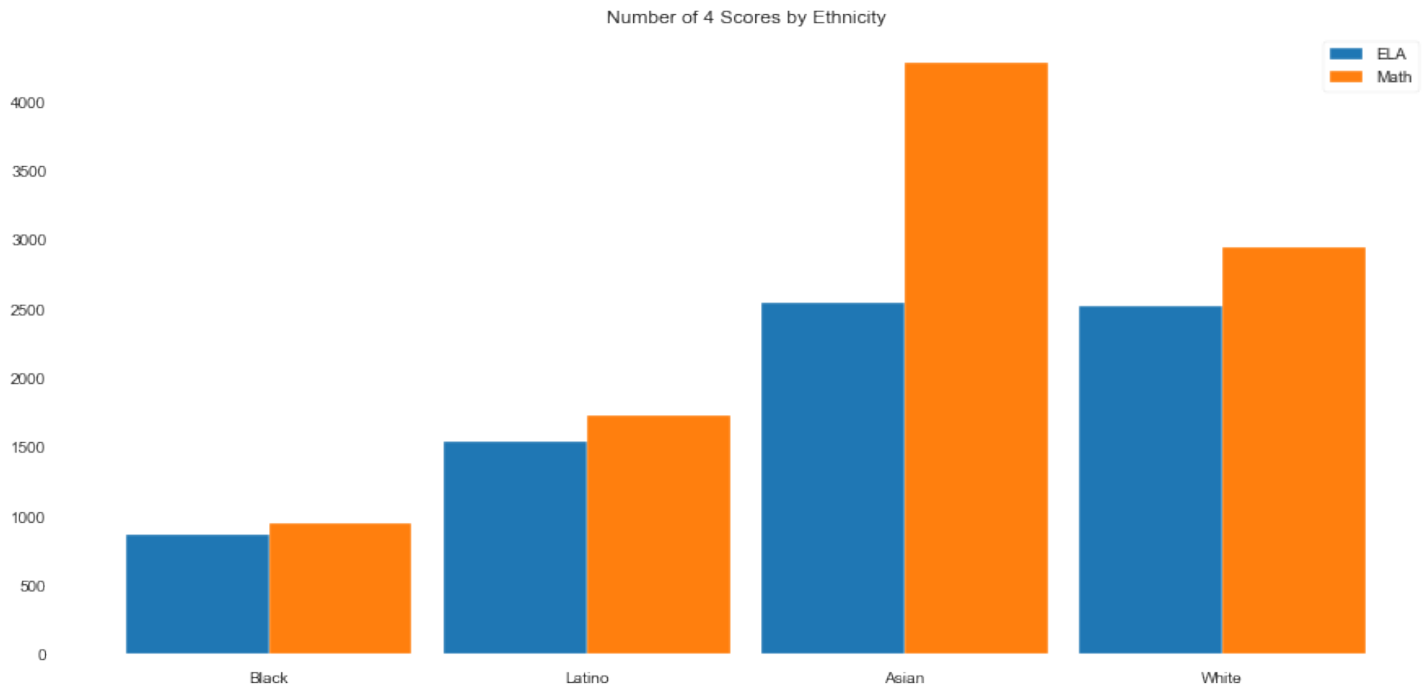
537 NYC Middle Schools sent at least 6 students to SHSAT for a total of **25,349** 8th graders taking the test. **57** schools send 0-5 8th graders to take the test. **121** NYC Middle Schools saw _at least_ 6 of their students receive offers, for a total of **4,018** 8th graders having received an offer. **473** schools saw 0-5 of their 8th graders receive an offer.

Merging Datasets

Using the DBN & Location Code I'll merge *Explorer* data for 7th graders to the *Offers* information for SHSAT testers. In the process it looks like **2** schools in *Explorer* didn't have information in the *Offers* dataset.

Exploratory Data Analysis & Feature Engineering

Looking at the assumption that those students/schools that have the majority of the 4 scores will, in turn, perform well on the aptitude test for the specialized high school, we can see that Black & Latinx students may receive less admittance offers based on this limited criterion.



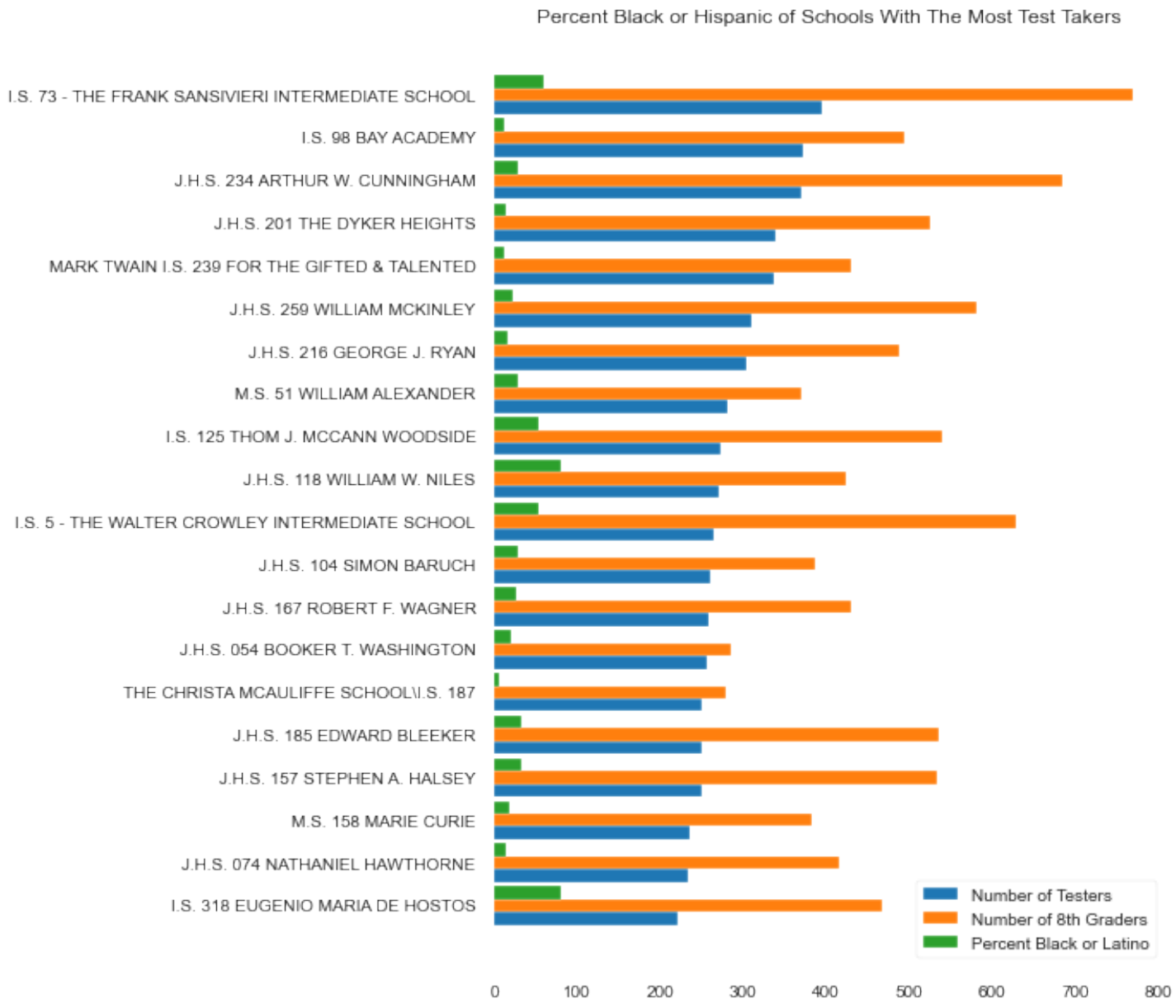
In order to better summarize the schools/students into ranges of test scores, I've added the following features:

- “Percentage of SHSAT takers receiving an offer” (Numbers of SHSAT takers / Number of Offers by school)
- “The total number of Black/Hispanic students in Grade 8” (Number of 8th graders * Percent Black / Hispanic)
- “Percentage of students who did the SHSAT” (Number of SHSAT takers / Number of 8th graders)
- “Average Mark” (the average of Average ELA Proficiency and Average Math Proficiency)
- “Percent of students with Level 4 ELA in Grade 7 (Grade 7 ELA 4s - All Students / Grade 7 ELA - All Students Tested)
- “Percent of students with Level 4 Math in Grade 7 (Grade 7 Math 4s - All Students / Grade 7 Math - All Students Tested)
- “Percent of students with Level 4 in Grade 7” (average of 4 percentages ELA and Math in Grade 7)
- “Average number of Level 4 students” (Grade 7 ELA 4s - All Students + Grade 7 Math 4s - All Students)/2

Schools with the highest number of test takers

What we're seeing in this next plot is that those schools that send the most 8th graders to the SHSAT, have less of their school, percentage-wise, represented by Black or Latinx students.

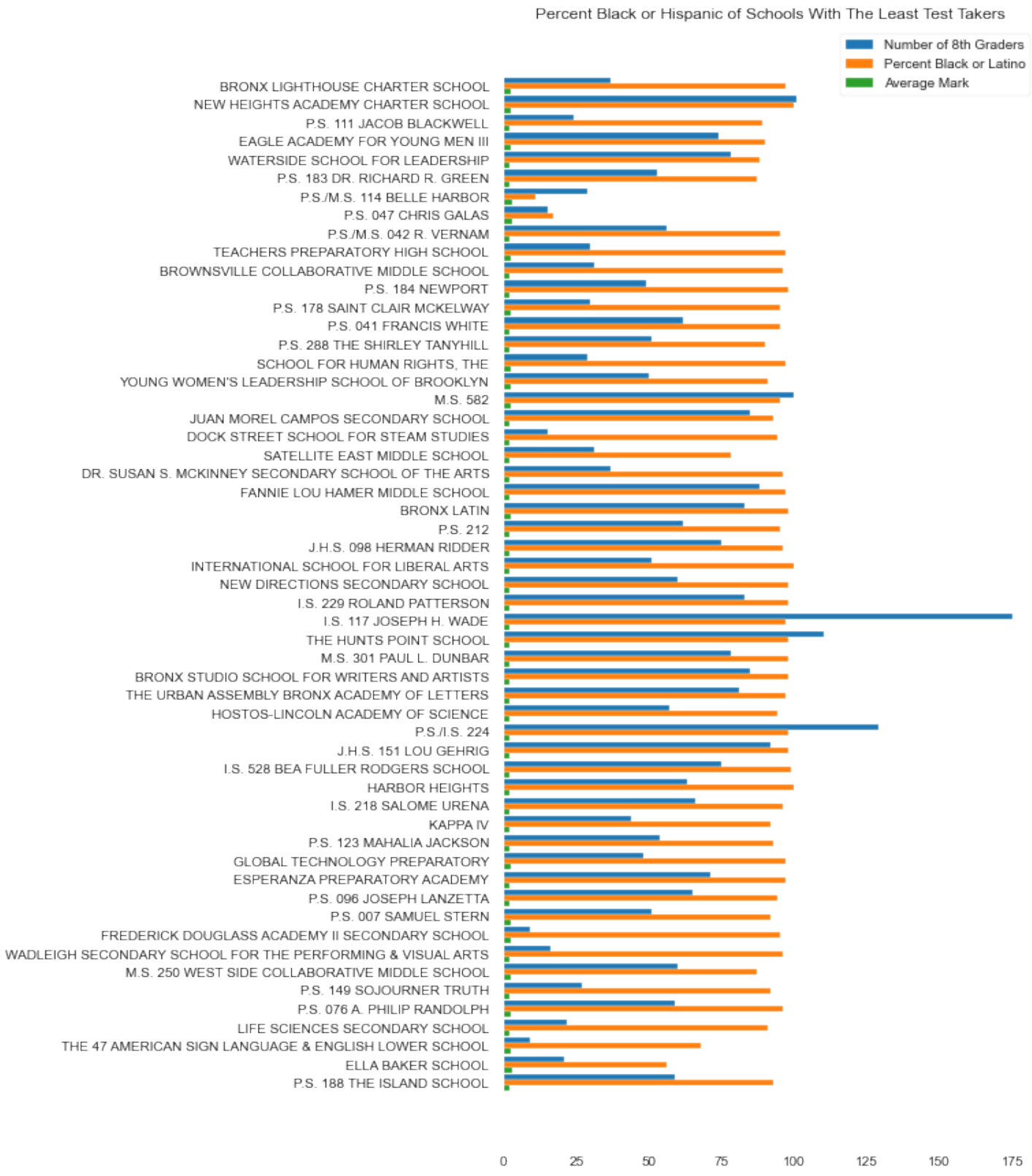
Interestingly, there is a high percentage of Black/Hispanic students (The William W. Niles (**82%**) school and The Eugenio Maria De Hostos (**78%**) school), near the middle of the pack and the lowest, respectively.



Schools with the least number of test takers

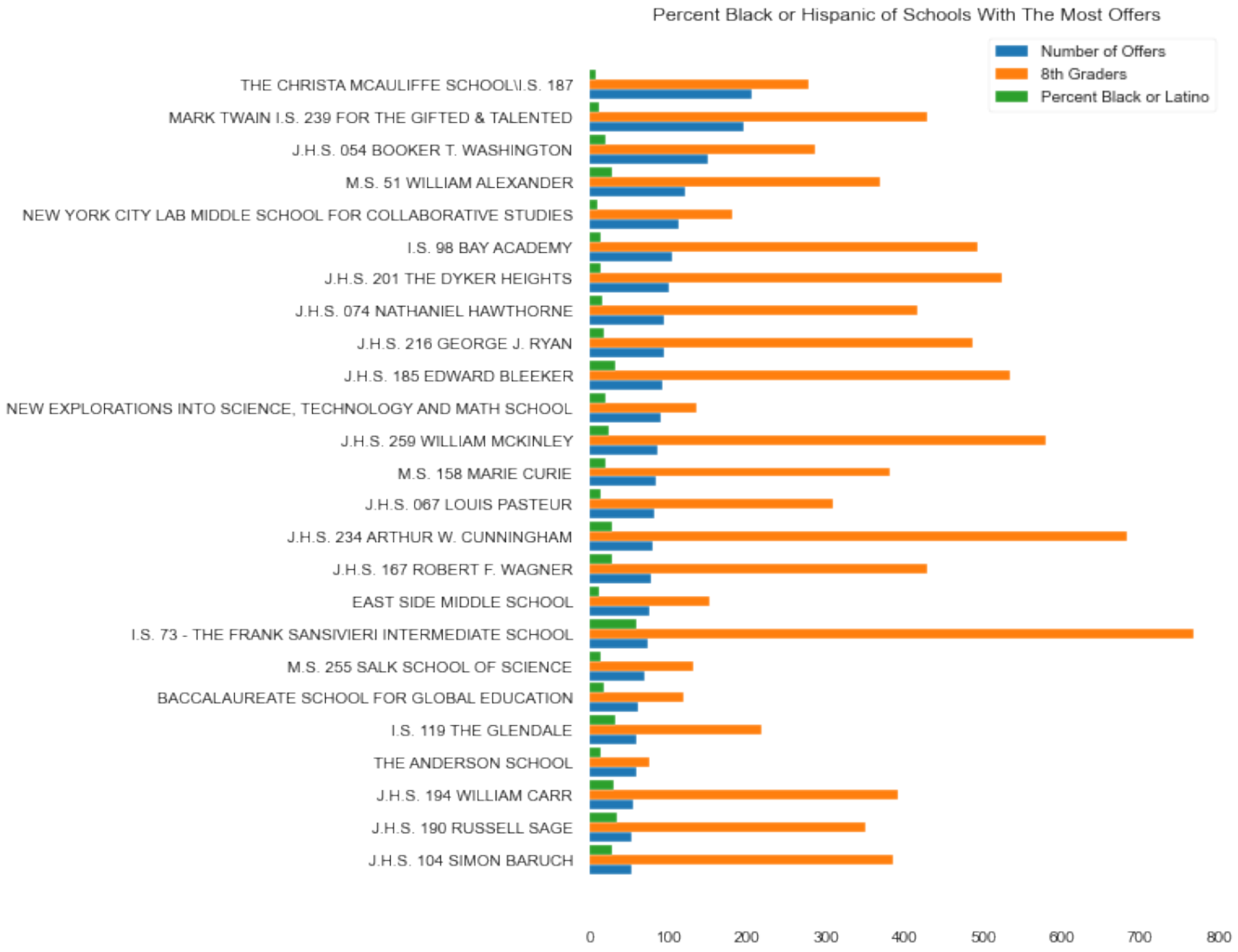
Nearly all of the schools with the least number of test takers in 2017 (55) had low average marks (average of AvgELA and AvgMath).

Also, most schools had a high percentage of Black or Latinx students.



Number of offers by school

The top **25** schools with the most offers received had lower percentages of Black or Hispanic students (highest percentage is at Frank Sansivieri school with **59%** Black or Latinx students).

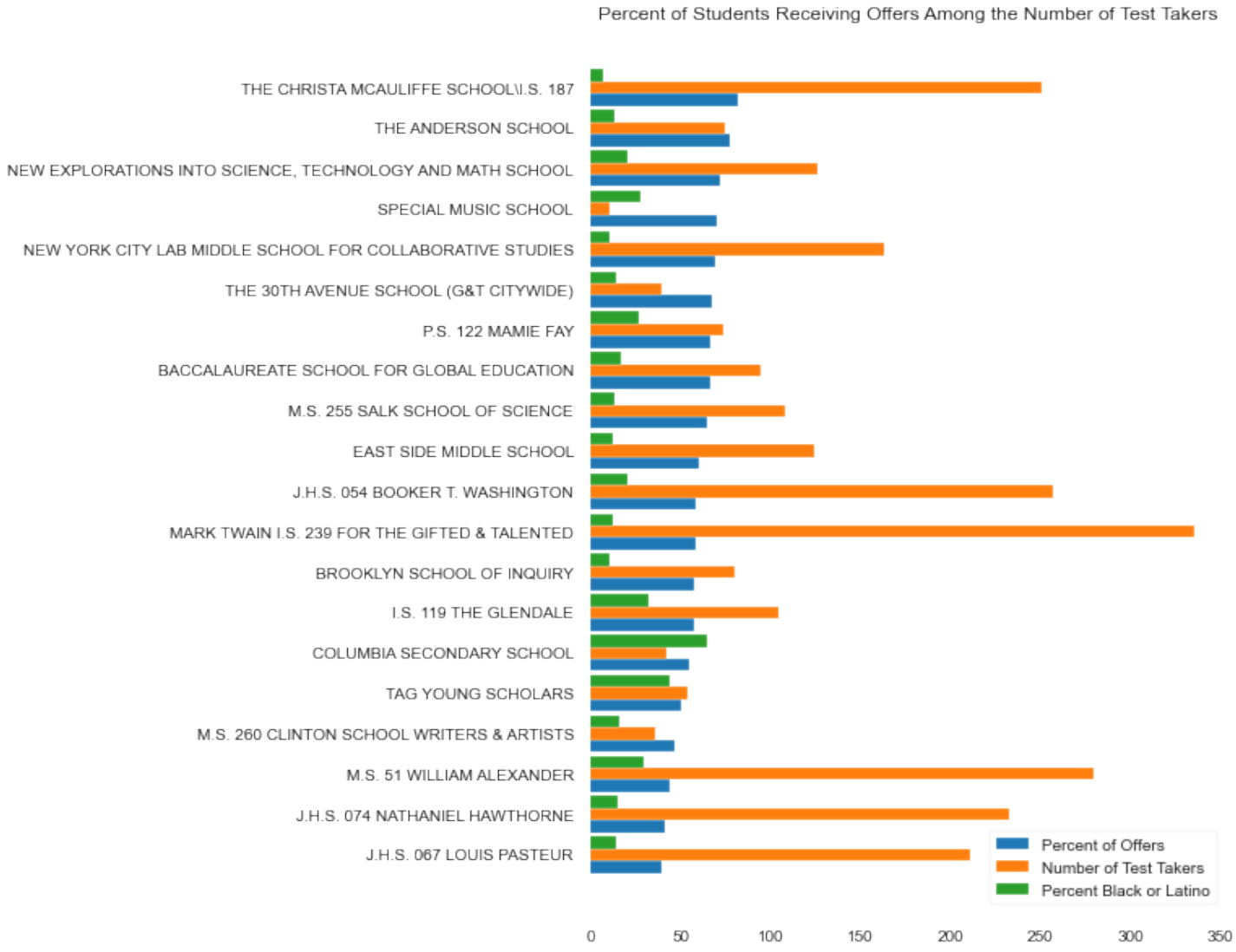


Highest percentage of offers for the number of test takers

Below are the top **20** schools that had the highest percentage of offers for the number of test takers, representing how successful that school was as to the number of students that were admitted to the specialized high school.

The Christa McAuliffe School had the most success with **82%** of 251 students taking the test getting an offer.

The schools scoring best at the percentage of students actually getting an offer are very low in Black or Latinx student percentages (the exception is the small Columbia Secondary School with **64%** Black/Latinx).



Least percentage of offers for the number of test takers

Of those schools which had *at least* 6 offers, the 20 schools with least success are shown below.

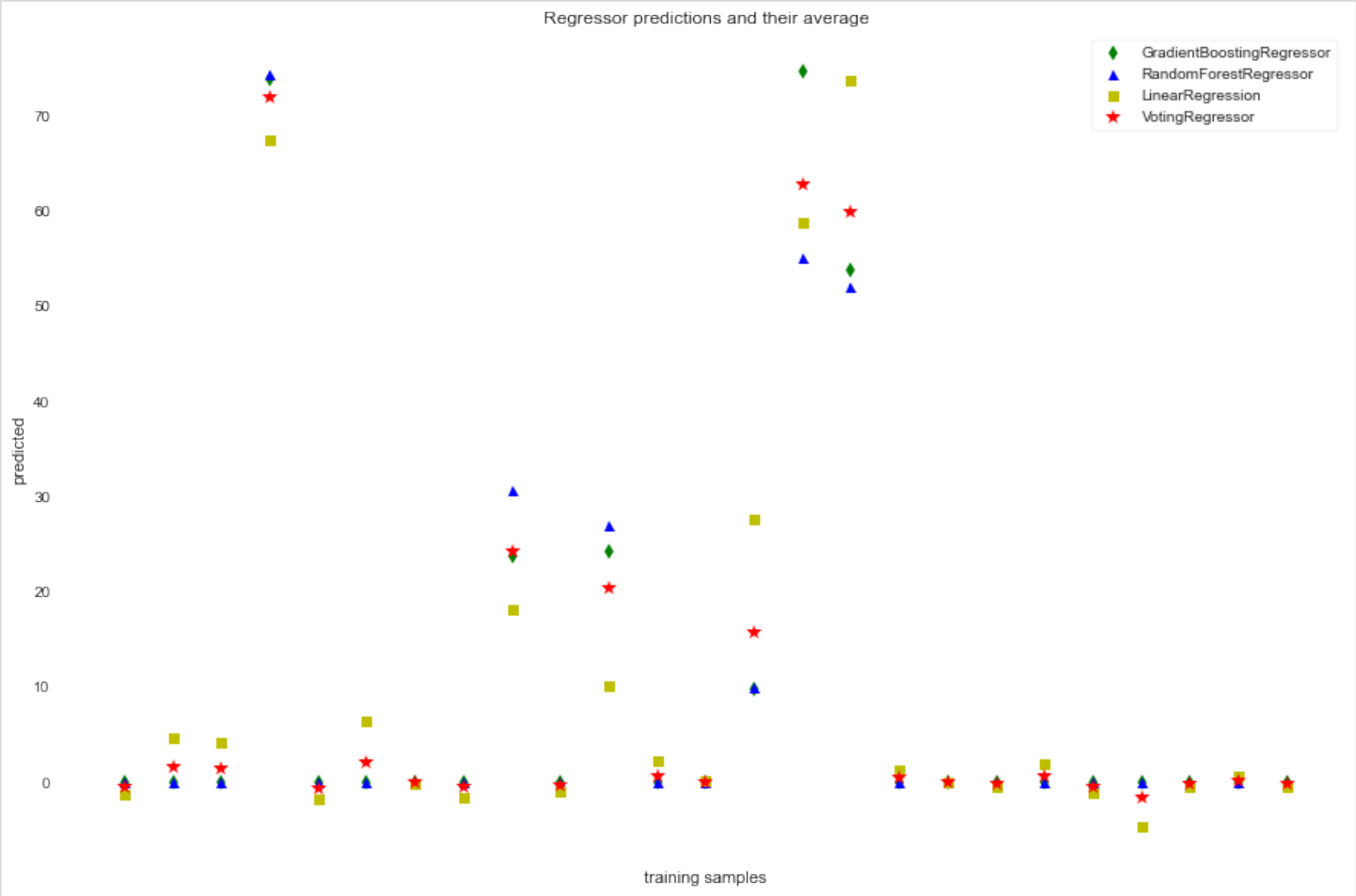
Two of the largest schools that are predominantly Black/Latinx and sent many students to the test are J.H.S 118 William W. Niles school @ **424** 8th graders (**82%** Black/Latinx) and I.S. 318 Eugenio Maria De Hostos @ **467** 8th graders (**78%** Black/Latinx).

Models & Evaluation

My intent is to use regression-based models because my label/y is numeric in nature and the interest I have is in how many offers a school ought to expect given the features/independent variables one could supply to the model.

Initially I will determine which regressor algorithm performs best then I will use an ensemble meta-estimator that fits several base regressors, each on the whole dataset then averages the individual predictions to form a final prediction.

Each of the Regressors is used to used to make the first **25** predictions.



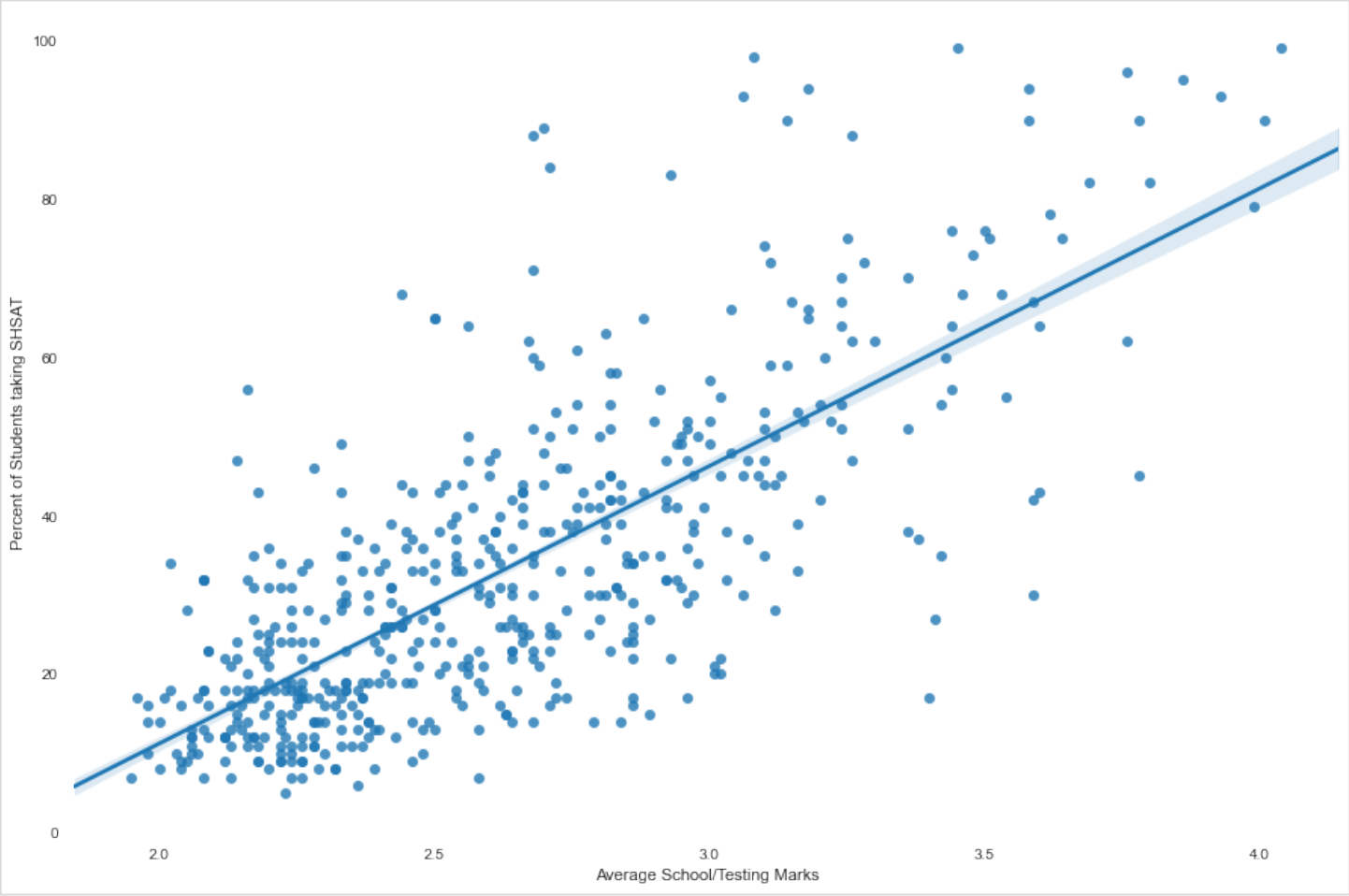
- Gradient Boosting Regressor R^2: 0.930
- Random Forest Regressor R^2: 0.946
- Linear Regression R^2: 0.930
- Voting Regressor R^2: 0.966

Given the R^2 scores are so close, I'll lean towards simplicity rather than running several base regressors and an ensemble to gain just a **3%** gain in explained behavior/R² by choosing to do a **linear regression** going forward.

Predictions & Recommendations

There's a strong correlation between the percent of test takers in a school and the number of offers that that high school received.

One approach to achieving more admissions offers is to send more students to take the SHSAT, based on the average number of level 4 scorers in the school.



The model is based on 530 schools (536 schools with at least 6 SHSAT takers, as SHSAT is unknown for category 0-5 takers. For 6 out of those 536 schools the AvgMark is 0 as a result.

Recommendation #1: Top 25 schools that can send more students to the take the test

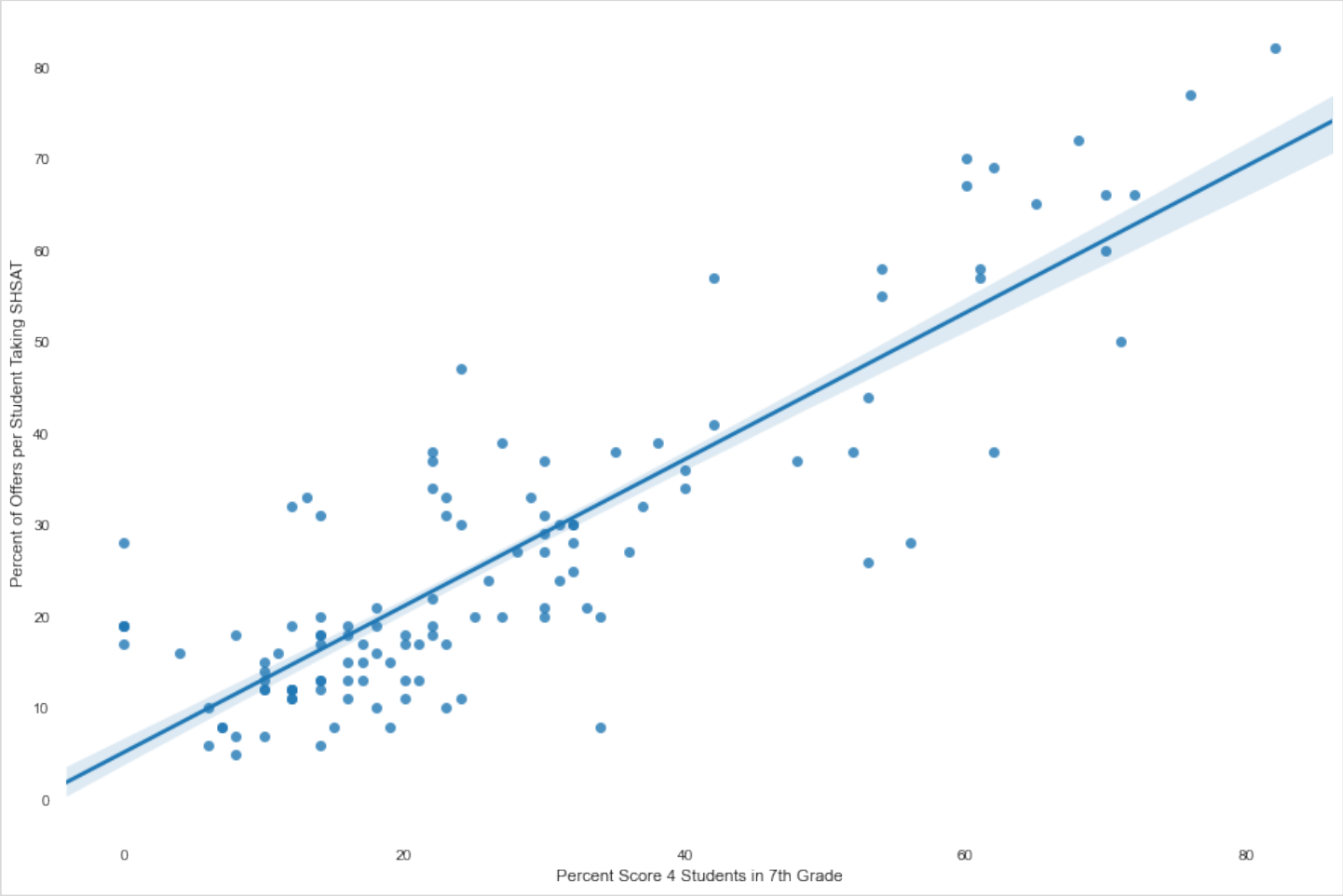
Below, are the predictions by school regarding the percent of students that could have taken the SHSAT (PerModelDidSHSAT). The PotentialTakers takes the difference between the PerModelDidSHSAT and PerDidSHSAT (ModAgainstDidSHSAT) multiplied by the total number of 8th grade students in each school (count_of_students_hs_admissions).

	feeder_school_name	count_of_students_in_hs_admissions	NumTestTakers	PerDidSHSAT	PerModelDidSHSAT	PotentialTakers	PctBlackOrHispanic
357	I.S. 061 LEONARDO DA VINCI	716.0	93.0	13.0	31.590739	133.0	92
464	I.S. 034 TOTTENVILLE	378.0	81.0	21.0	39.620051	70.0	11
468	I.S. 061 WILLIAM A MORRIS	325.0	40.0	12.0	31.327482	63.0	78
402	J.H.S. 210 ELIZABETH BLACKWELL	640.0	169.0	26.0	34.486556	54.0	67
450	I.S. 145 JOSEPH PULITZER	581.0	137.0	24.0	32.907019	52.0	90
397	M.S. 137 AMERICA'S SCHOOL OF HEROES	642.0	179.0	28.0	36.066093	52.0	37
361	I.S. 093 RIDGEWOOD	378.0	78.0	21.0	33.696788	48.0	72
359	I.S. 077	345.0	61.0	18.0	30.800970	44.0	79
403	J.H.S. 226 VIRGIL I. GRISSOM	331.0	58.0	18.0	31.064226	43.0	55
433	I.S. 238 - SUSAN B. ANTHONY ACADEMY	529.0	123.0	23.0	30.800970	41.0	67
475	J.H.S. 162 THE WILLOUGHBY	161.0	8.0	5.0	29.353061	39.0	95
467	I.S. 051 EDWIN MARKHAM	436.0	101.0	23.0	31.590739	37.0	66
462	I.S. 027 ANNING S. PRALL	328.0	77.0	23.0	33.960044	36.0	63
119	M.S. 302 LUISA DESSUS CRUZ	166.0	13.0	8.0	28.958177	35.0	98
178	THE ANGELO PATRI MIDDLE SCHOOL	190.0	17.0	9.0	26.852128	34.0	97
294	LIBERTY AVENUE MIDDLE SCHOOL	152.0	12.0	8.0	30.537714	34.0	92
460	I.S. 007 ELIAS BERNSTEIN	394.0	129.0	33.0	41.594473	34.0	7
284	I.S. 171 ABRAHAM LINCOLN	151.0	11.0	7.0	29.484689	34.0	91
443	I.S. 010 HORACE GREELEY	250.0	56.0	22.0	35.276325	33.0	58
470	I.S. 072 ROCCO LAURIE	463.0	139.0	30.0	36.987490	32.0	29
400	J.H.S. 202 ROBERT H. GODDARD	374.0	98.0	26.0	34.618184	32.0	55
431	I.S. 192 THE LINDEN	162.0	18.0	11.0	29.221433	30.0	95
466	I.S. 49 BERTA A. DREYFUS	241.0	42.0	17.0	29.616317	30.0	76
93	M.S. 324 - PATRIA MIRABAL	149.0	20.0	13.0	32.643763	29.0	95
168	I.S. 254	143.0	13.0	9.0	29.221433	29.0	95

- The above-referenced schools ought to send more students to take the SHSAT as their average marks may translate into more of their students receiving offers to attend the specialized high schools.
- Increasing the number of test-takers from schools with higher percentages of Black &/or Latinx test-takers will help address the deep disparity of offers being received by White and Asian students.

Another approach is to only send test takers that're high performing students, or those who are level 4.

All possible performance related predictors (mark and level4) are very strongly correlated with each other (multicolinear). Another way to look at what can predict (successful) performance on the SHSAT is to look at percent of offers by student (PctOffersByStudent) that took the SHSAT to the percent of level-4 students in 7th grade (PctScore4).



Recommendation #2: Top 25 schools that can increase the percent of their students receiving offers

Here I'm making predictions by school regarding the percent of offers per student according to the model (mod_offers). Real offers (RealOffers) looks at the modeled percent offers (PctModelOffers) multiplied by the number of SHSAT takers. PotentialOffers takes the difference of RealOffers from actual offers (NumSpecializedOffers).

	feeder_school_name	count_of_students_in_hs_admissions	NumSpecializedOffers	PctModelOffers	RealOffers	PotentialOffers	PctBlackOrHispanic
334	J.H.S. 234 ARTHUR W. CUNNINGHAM	684.0	79.0	30.619022	113.0	34.0	28
411	SCHOLARS' ACADEMY	231.0	27.0	49.176006	51.0	24.0	37
276	P.S. 235 JANICE MARIE KNIGHT SCHOOL	91.0	7.0	31.546872	28.0	21.0	95
415	J.H.S. 157 STEPHEN A. HALSEY	533.0	52.0	27.835475	69.0	17.0	33
461	I.S. 024 MYRA S. BARNES	384.0	11.0	17.629134	23.0	12.0	13
424	QUEENS GATEWAY TO HEALTH SCIENCES SECONDARY SC...	83.0	23.0	57.526648	35.0	12.0	46
161	J.H.S. 118 WILLIAM W. NILES	424.0	29.0	14.845587	40.0	11.0	81
452	I.S. 227 LOUIS ARMSTRONG	460.0	27.0	18.556983	38.0	11.0	59
464	I.S. 034 TOTTENVILLE	378.0	8.0	21.340531	17.0	9.0	11
54	TAG YOUNG SCHOLARS	56.0	27.0	65.877291	36.0	9.0	44
188	I.S. 181 PABLO CASALS	290.0	8.0	12.989888	16.0	8.0	89
355	I.S. 5 - THE WALTER CROWLEY INTERMEDIATE SCHOOL	629.0	40.0	17.629134	47.0	7.0	54
236	CONSELYEA PREPARATORY SCHOOL	169.0	7.0	13.917737	13.0	6.0	62
460	I.S. 007 ELIAS BERNSTEIN	394.0	22.0	21.340531	28.0	6.0	7
469	MARSH AVENUE SCHOOL FOR EXPEDITIONARY LEARNING	156.0	7.0	16.701285	12.0	5.0	27
465	P.S. 048 WILLIAM G. WILCOX	118.0	9.0	19.484832	14.0	5.0	22
113	M.S. X101 EDWARD R. BYRNE	147.0	7.0	18.556983	12.0	5.0	66
453	I.S. 230	437.0	26.0	19.484832	31.0	5.0	57
366	P.S. 128 THE LORRAINE TUZZO, JUNIPER VALLEY EL...	117.0	9.0	27.835475	13.0	4.0	21
362	P.S. 102 BAYVIEW	142.0	17.0	29.691173	20.0	3.0	36
449	I.S. 141 THE STEINWAY	348.0	26.0	15.773436	27.0	1.0	41
458	BACCALAUREATE SCHOOL FOR GLOBAL EDUCATION	119.0	62.0	66.805140	63.0	1.0	17
397	M.S. 137 AMERICA'S SCHOOL OF HEROES	642.0	23.0	12.989888	23.0	0.0	37

The above table is filtered to exclude schools in eight overperforming districts, detailed below. "Overperforming" means total offers (NumSpecializedOffers) divided by total 8th graders (count_of_students_in_hs_admissions).

The table displays schools that should have received at least 10 Extra Offers according to the model in under and average performing districts (listed below).

	District	count_of_students_in_hs_admissions	NumSpecializedOffers	Performance
1	2	2551.0	493.0	19.0
25	26	2182.0	397.0	18.0
2	3	1708.0	266.0	16.0
19	20	4011.0	591.0	15.0
0	1	968.0	128.0	13.0
20	21	3088.0	363.0	12.0
14	15	2272.0	247.0	11.0
24	25	2764.0	313.0	11.0
29	30	3374.0	253.0	7.0
23	24	4569.0	252.0	6.0

In particular, **P.S. 235 Janice Marie Knight School** & **J.H.S. 118 William W. Niles** are great candidates that should've seen their students receive more **21** and **11** more offers, respectively, for admission to the specialized high schools. Their high percentage of Black &/or Latinx students would've improved the rate at which those ethnicities receive offers across NYC.

Future Analyses

- An aspect that I wasn't able to explore was using GIS to determine if there are any differences in admissions offers to schools/students based on how close the feeder school is to the specialized high school.
- I was only able to use one year of admissions and test data. It would have been interesting to determine if there're any trends in the data across more than one year of data.
- The data only contained the performance on tests that are administered to students during a typical school year.
    - It would be interesting to see how preparatory tests for the SHSAT relate to the number of offers received by schools/students.
    - Also, looking at any after-school prep programs' impact on the number of offers received, would be interesting.

Credits

- Project structure based on [cookiecutter data science project template]