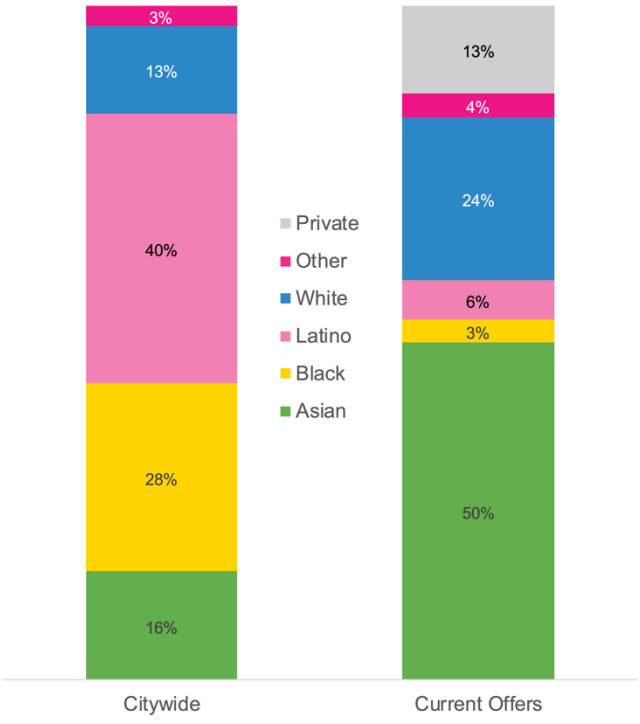
Specialized High School Scholastic Aptitude Test Admission Performance

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 (SHS) 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 the students receiving SHS offers. Black & Latinx students are 68% of the NYC high school population, but only 9% of the SHS offers[1].

Problem Statement

Which factors that predict success on the SHSAT can increase the number of SHS admissions offers received by Black & Latinx students?

[1]Exam Schools: Inside America's Most Selective Public High Schools by Finn and Hockett

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 by school, 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 receiving SHS offers.

An initial assumption is that those students who do well on English Language Arts (ELA) & Math standardized tests, will similarly perform well on the SHSAT. We'll investigate this and extrapolate as to whether this is the case across all schools/students.

Because my response variable is continuous, I'll be using linear regression models to determine how many SHS offers schools that fit a certain testing/aptitude standard could be getting based on their standardized test (ELA + Math) 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. In addition for models to perform best, improvements to data dimensionality were implemented, and several summary features were created.

For example, the *2016 School Explorer* has 20 variables with ELA & Math testing information on just the 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
```

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

These columns were summarized into several other features that I've detailed in the Feature Engineering section below.

We can also 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

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.

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.

```
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
```

In the 2017-2018 SHSAT Admissions Test Offers By Sending School (Offers) we see that **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.

Feature Engineering

In order to better summarize the schools/students into ranges & to allow the models to predict low-dimension data, I've added the following summary features:

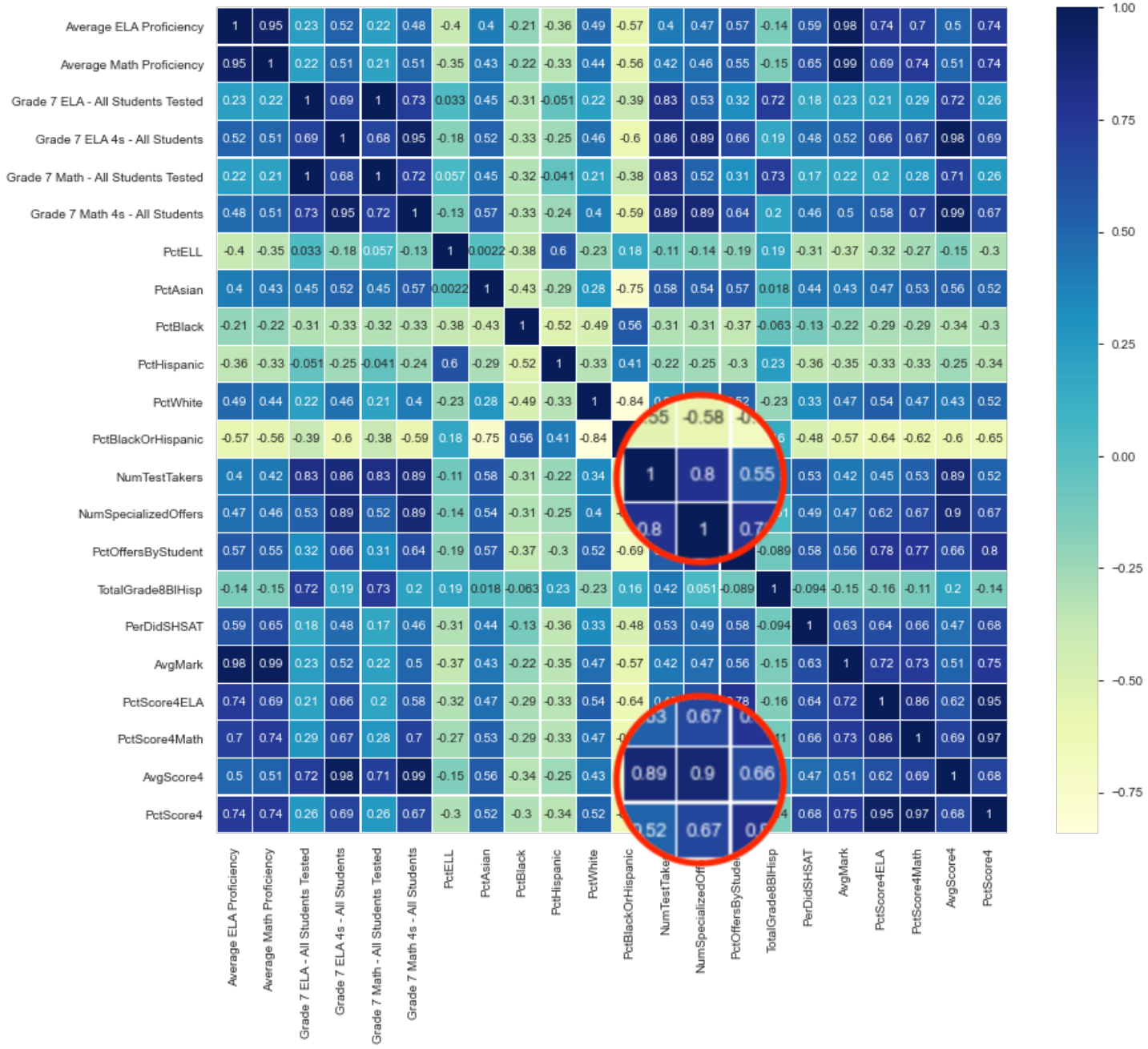
- “Percentage of SHSAT takers receiving an offer” (Numbers of SHSAT takers / Number of Offers by school)
- “The total number of Black/Hispanic students in Grade 8” (Number of 8th graders * Percent Black / Hispanic)
- “Percentage of students who did the SHSAT” (Number of SHSAT takers / Number of 8th graders)
- “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

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.

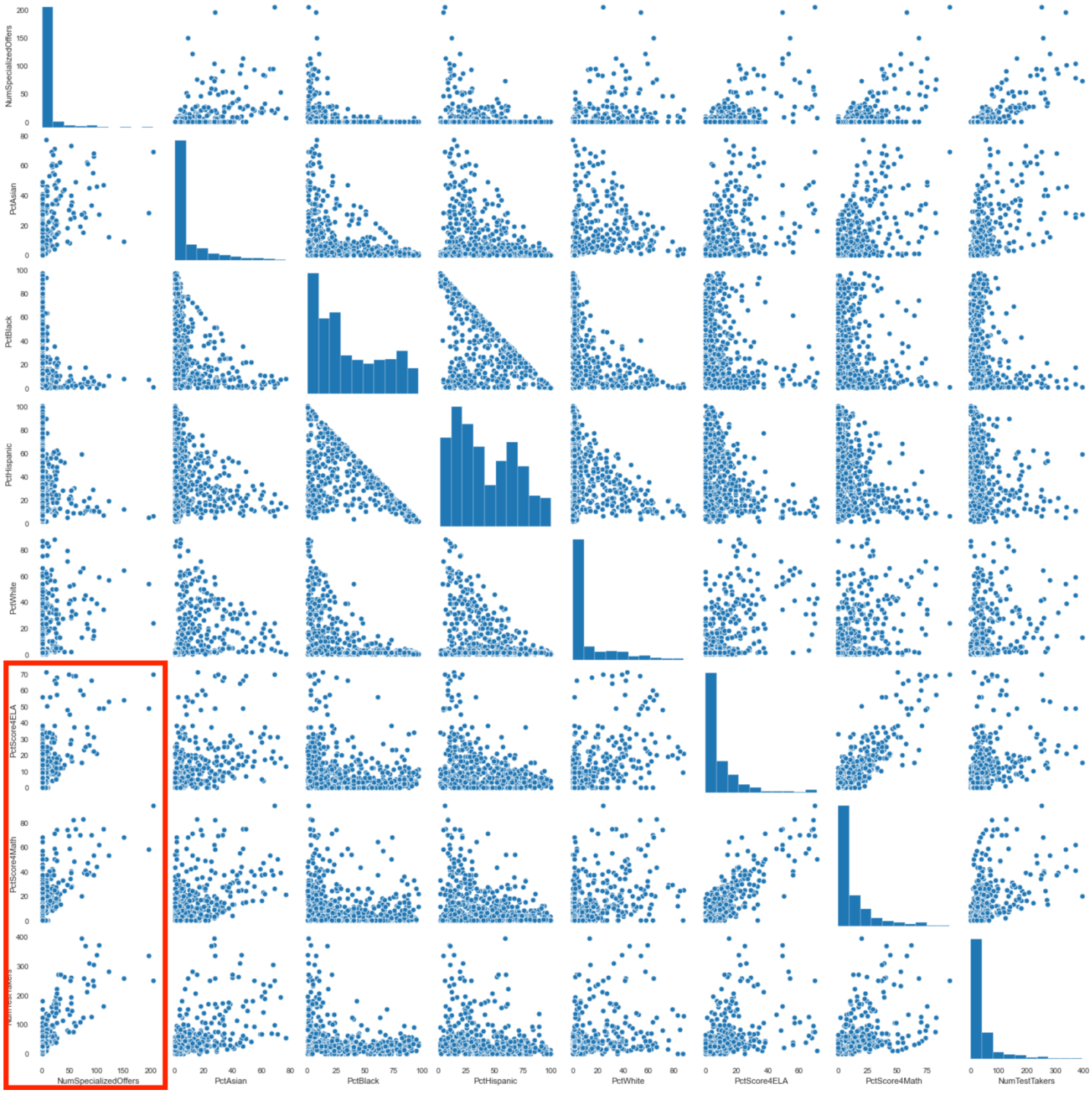
Exploratory Data Analysis

The initial test of our assumption that high performers on the ELA & Math tests translates to SHSAT success used a heatmap:



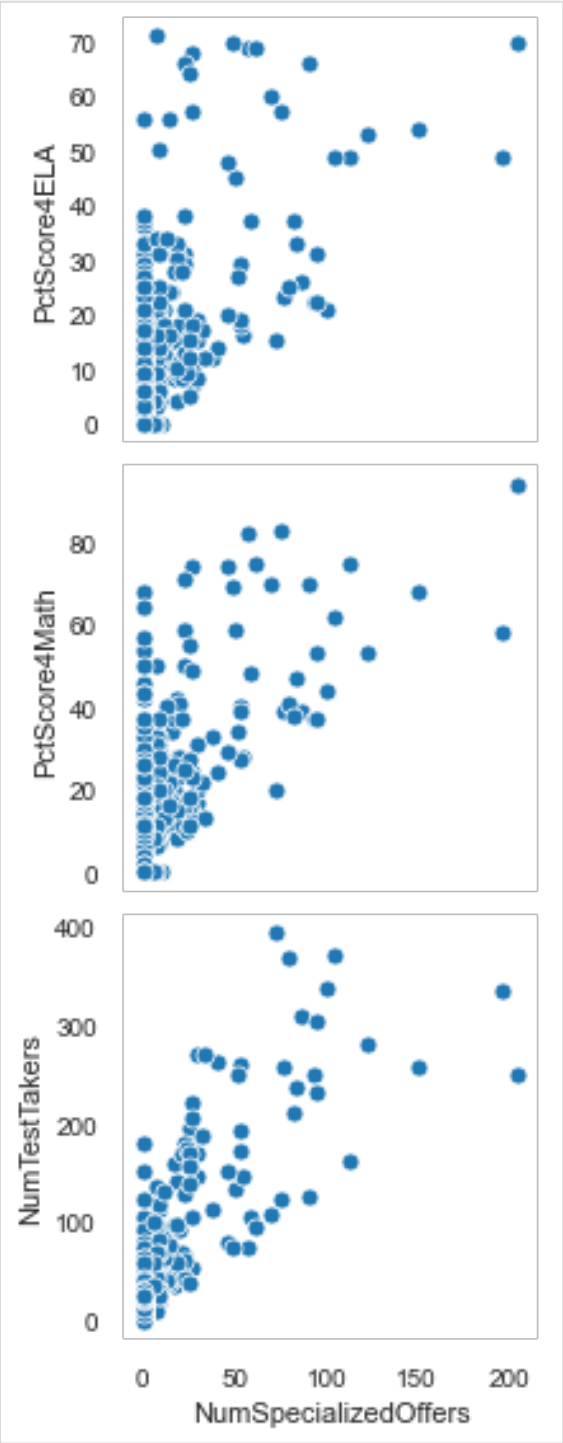
Our assumption is proven true (AVgScore4 vs NumSpecializedOffers), preliminarily, but even more interesting is that the number of SHSAT test takers (NumTestTakers) is also strongly correlated with the SHS offers!

Let's look at another plot to get a better idea of where the values are falling:



We can see again the positive relationship between ELA & Math test scores (PctScore4ELA & PctScore4Math) and the number of SHSAT test takers (NumTestTakers).

Here is a zoomed in version:

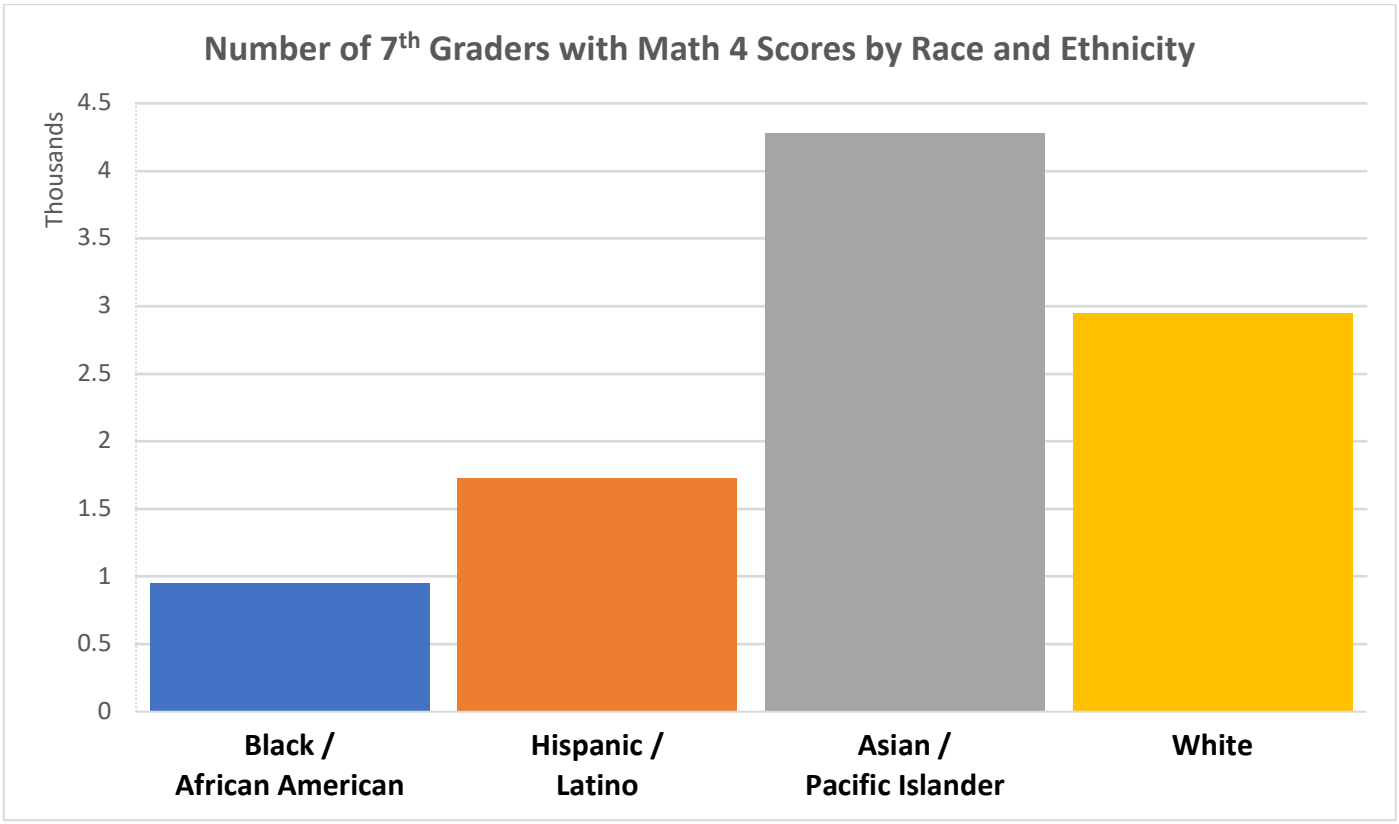


If the assumption holds, that those students/schools that do well on ELA & Math (level 4 scores) will also perform well on the SHSAT, then are Black & Latinx students not in possession of those kinds of scores? Does that hint as to the fewer SHS offers?

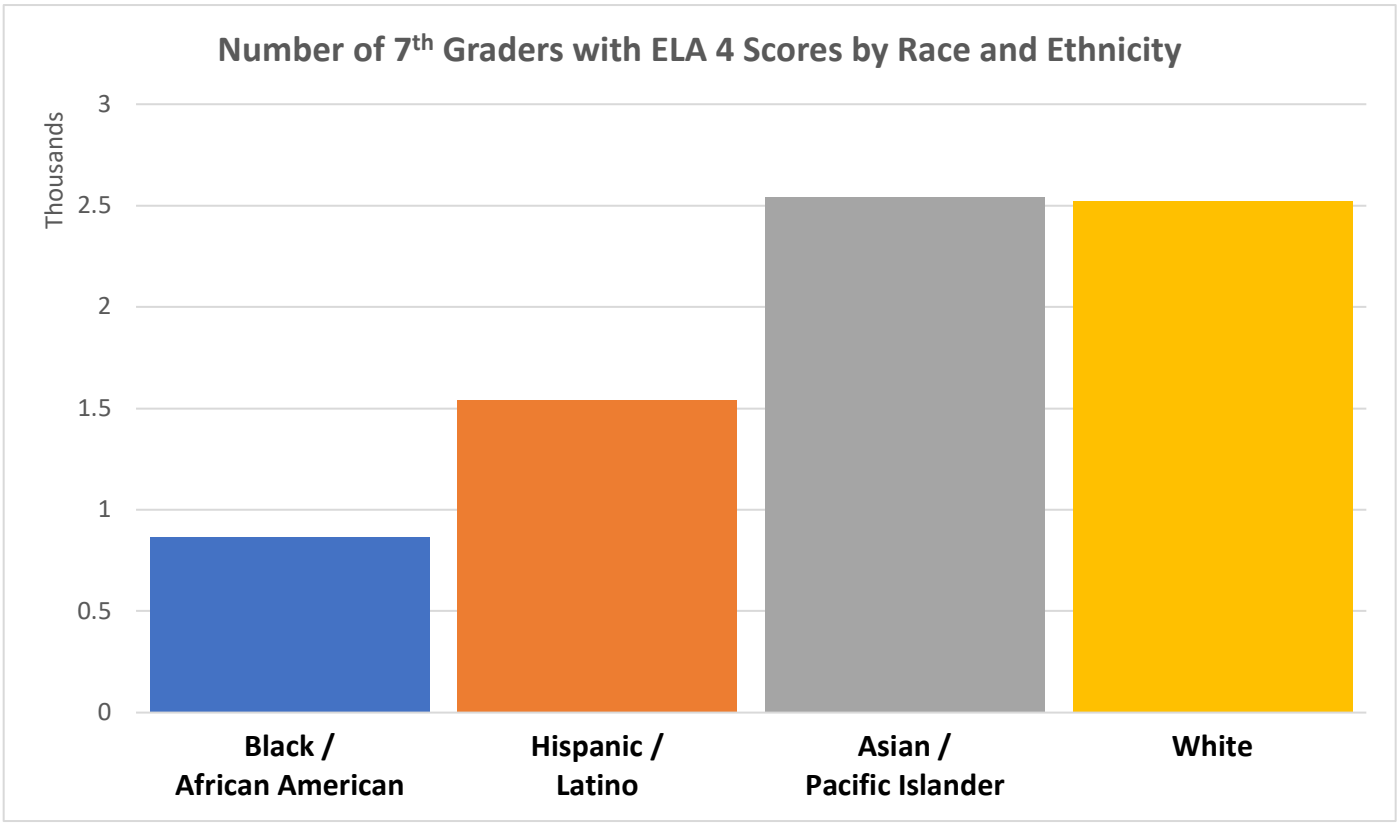
ELA & Math Testing Scores

To investigate this, the following plots dig into the racial and ethnic breakdowns of ELA & Math testing scores, in addition to the number of SHSAT test takers.

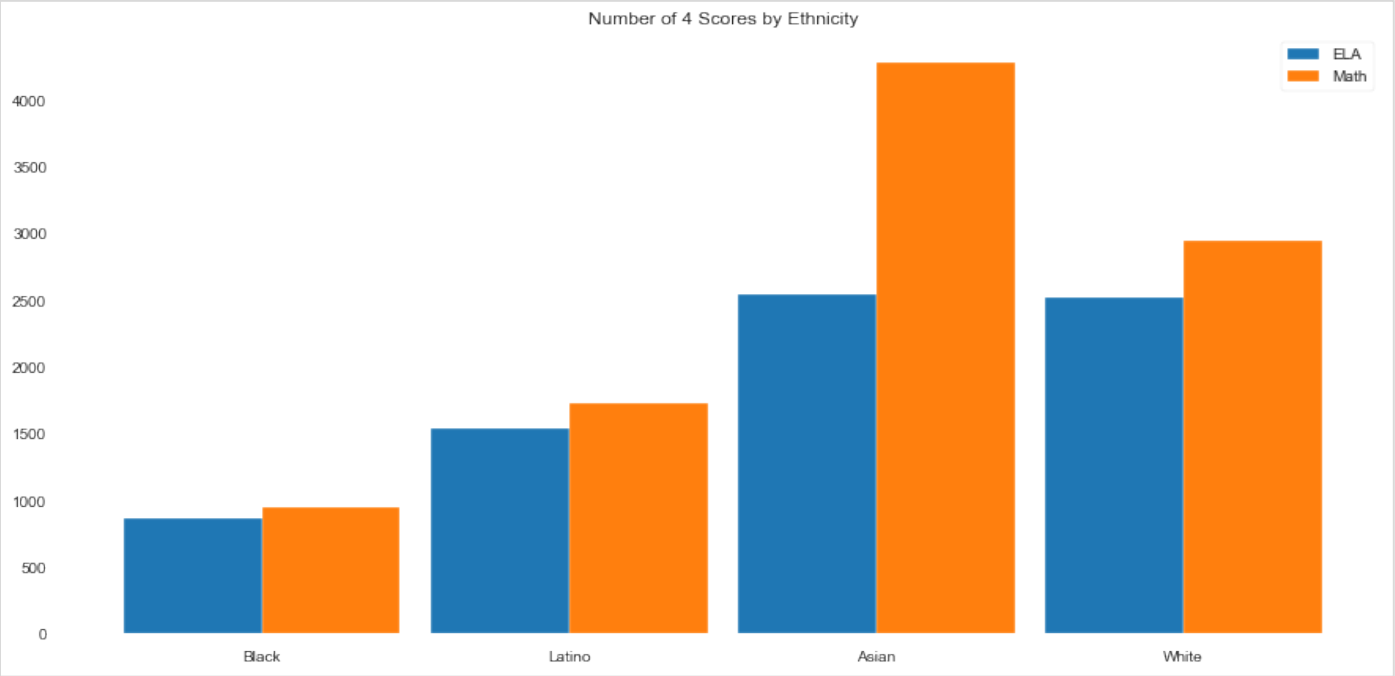
This Math testing score breakdown shows significantly less Black & Latinx students scoring 4s (best) on this standardized test.



This ELA breakdown also shows significantly less Black & Latinx students scoring 4s (best).



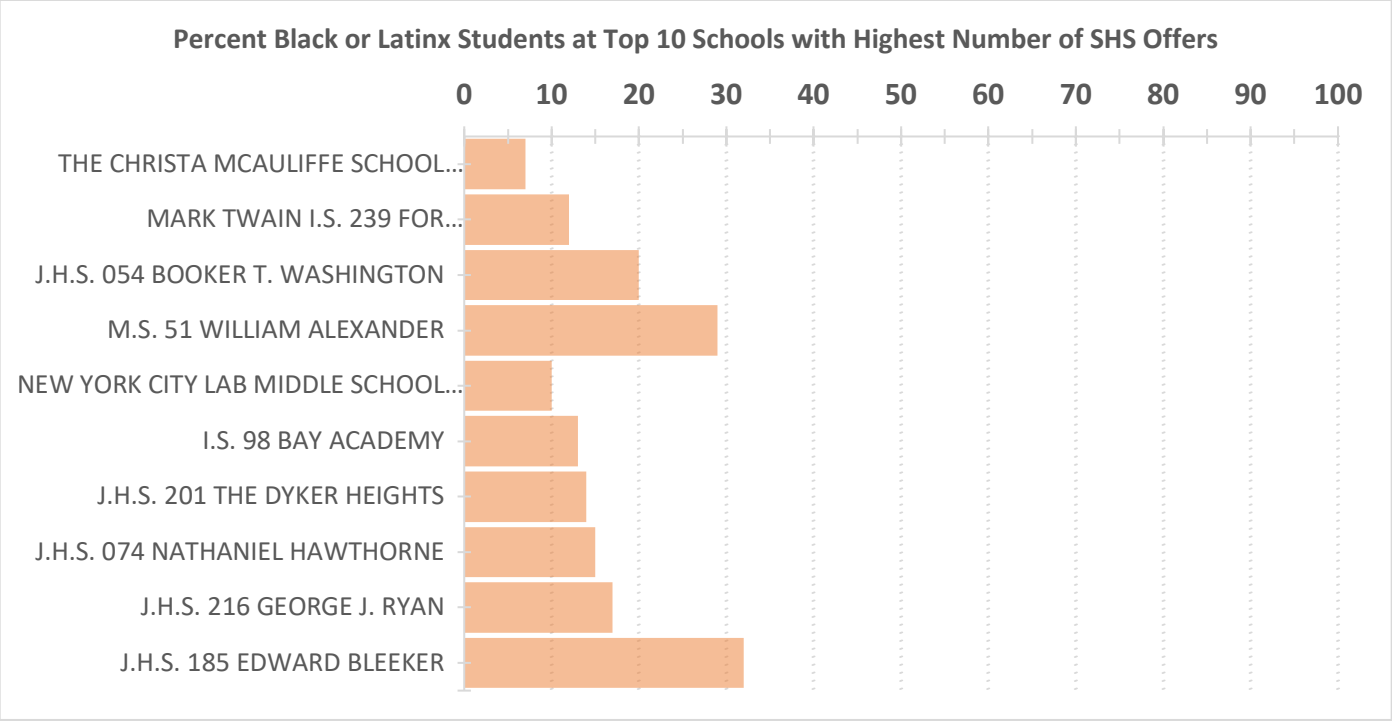
Here are the ELA & Math test scores side by side.



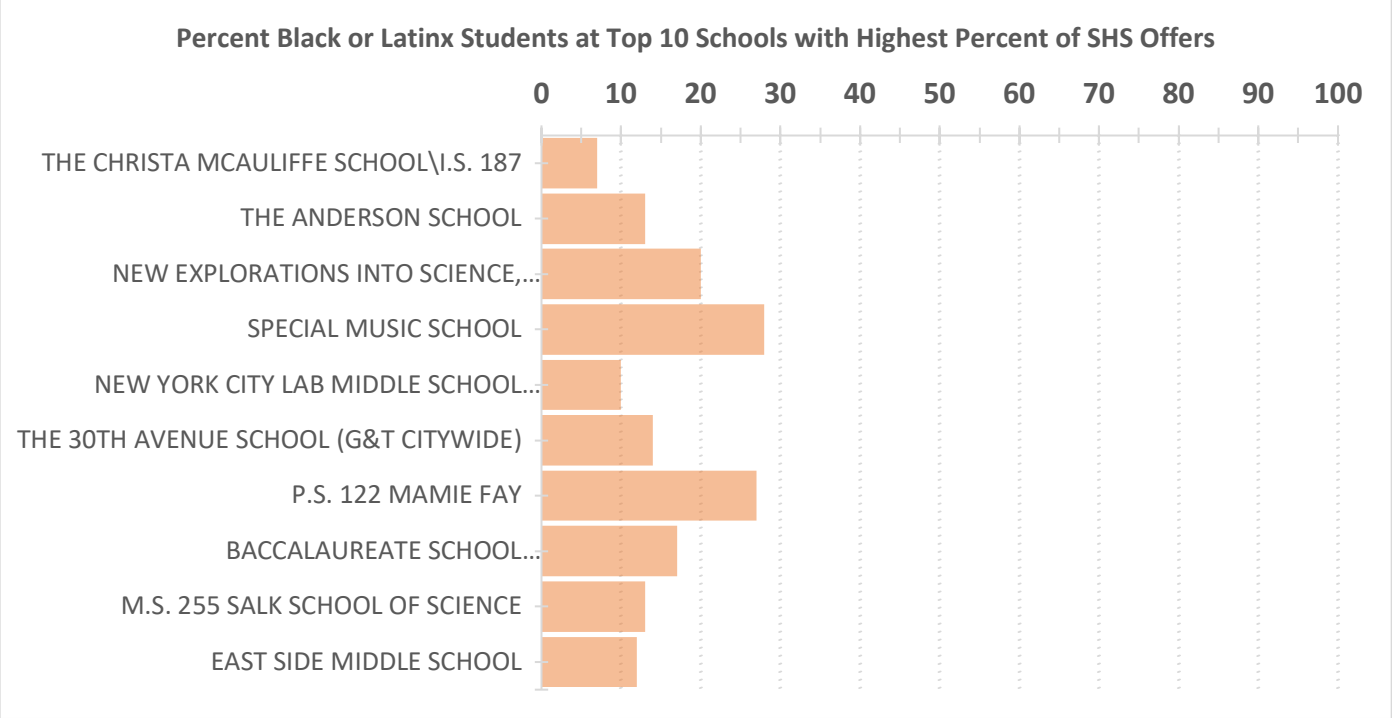
With the strength of the relationship between ELA & Math 4 scores and SHS offers, we can potentially see why Black or Latinx students are generally receiving less SHS offers.

SHS Offers

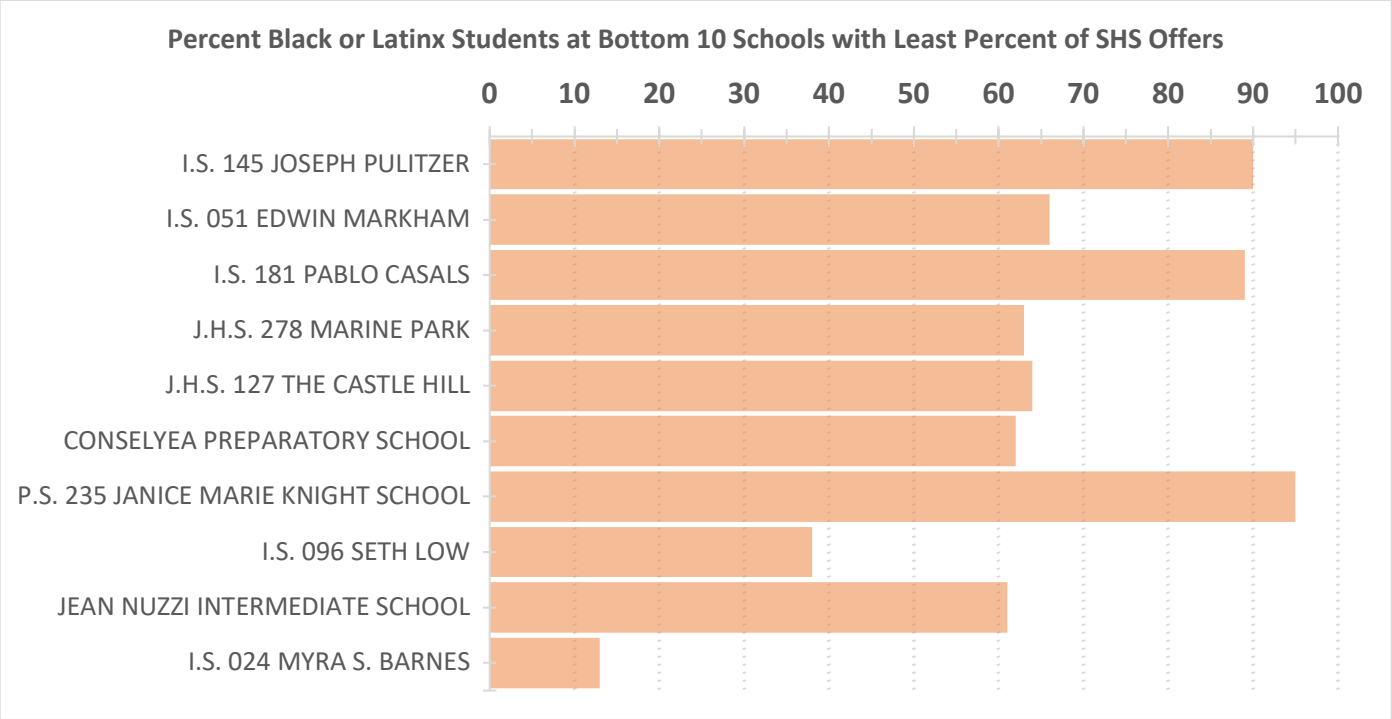
Looking at the number of SHS offers we can further depict the disparity of Black & Latinx students



Among the schools that received the **highest** number of SHS offers, nearly all are composed of fewer than 30% Black or Latinx students.



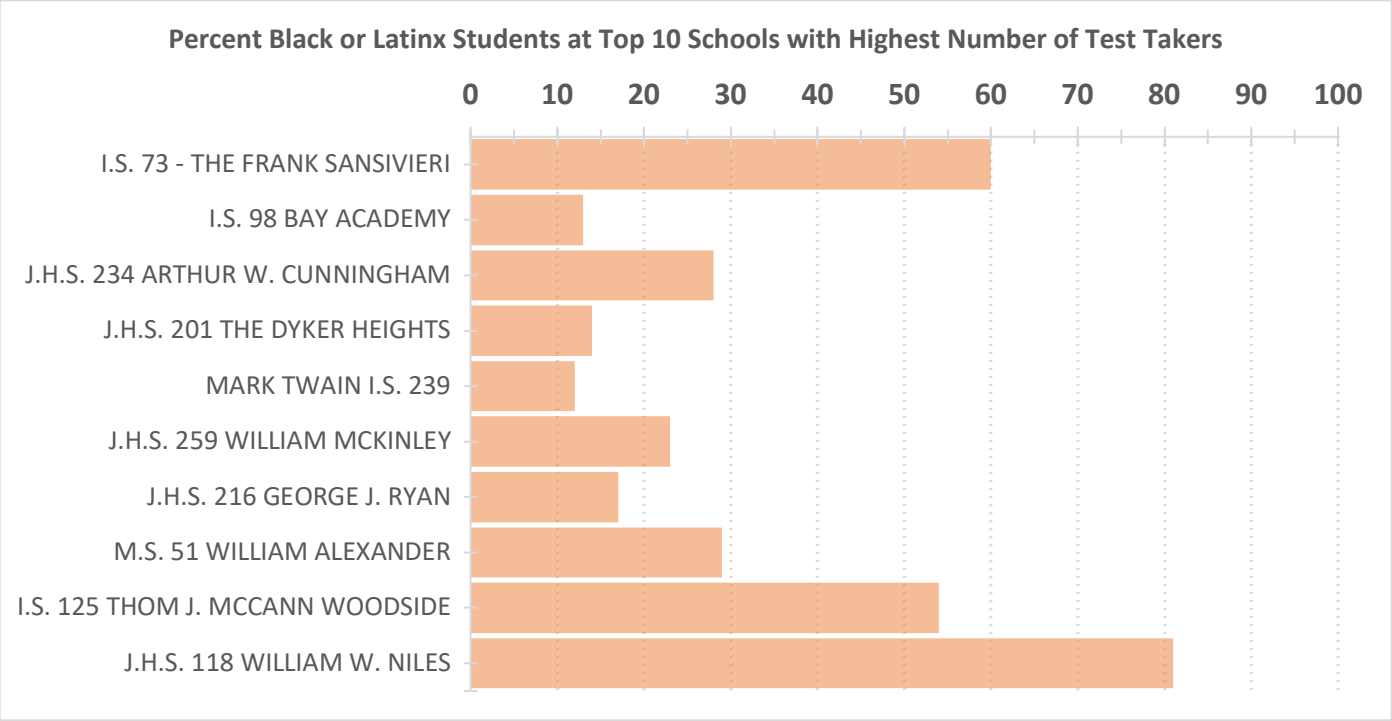
Among schools with the **highest** percentage of their test takers/students receiving SHS offers, **all** are composed of less than 30% Black or Latinx students.



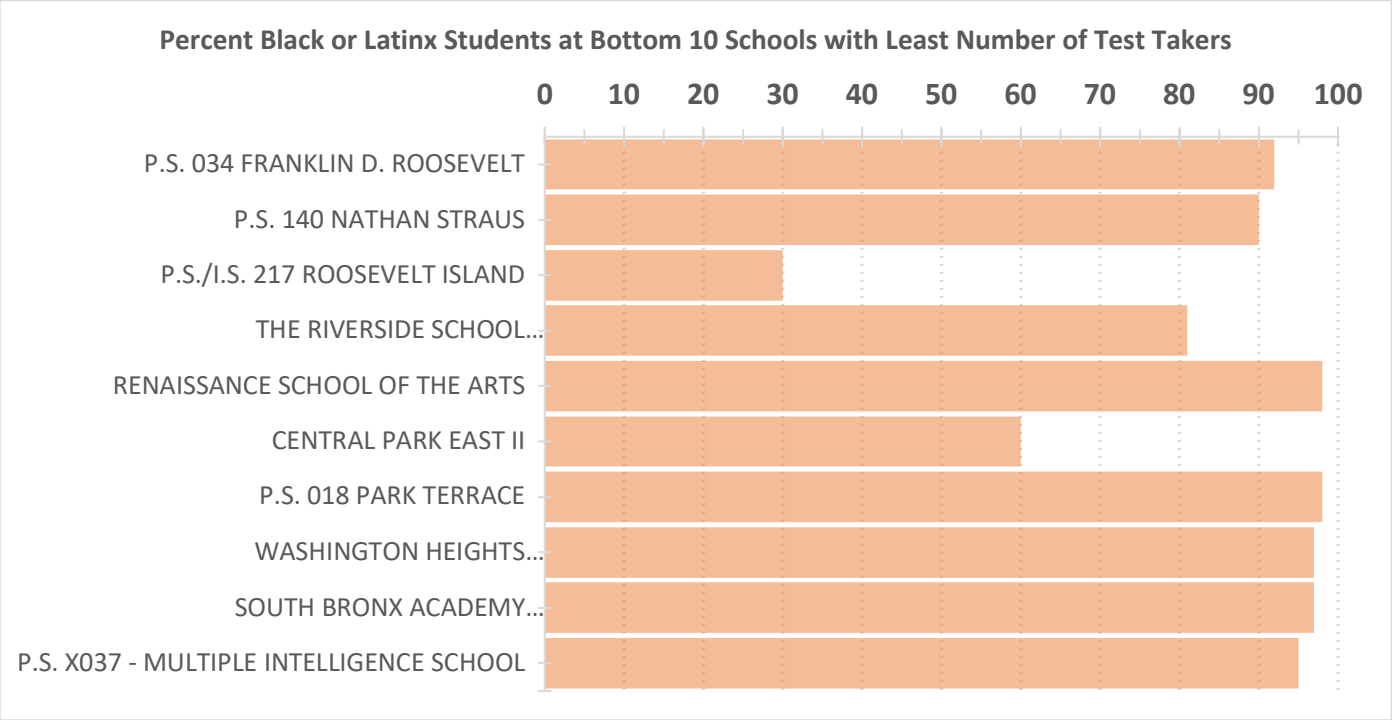
Among the schools where the **least** percentage of their students received SHS offers, **nearly all** are composed of 60% or more Black or Latinx students.

SHSAT Test Takers

Looking at relationship between SHS offers and the number of SHSAT test takers, we can explore the relationship between the Black & Latinx students and their representation at the SHSAT test.



Most of the schools with the **highest** number of SHSAT test takers are composed of less than 30% Black or Latinx students.



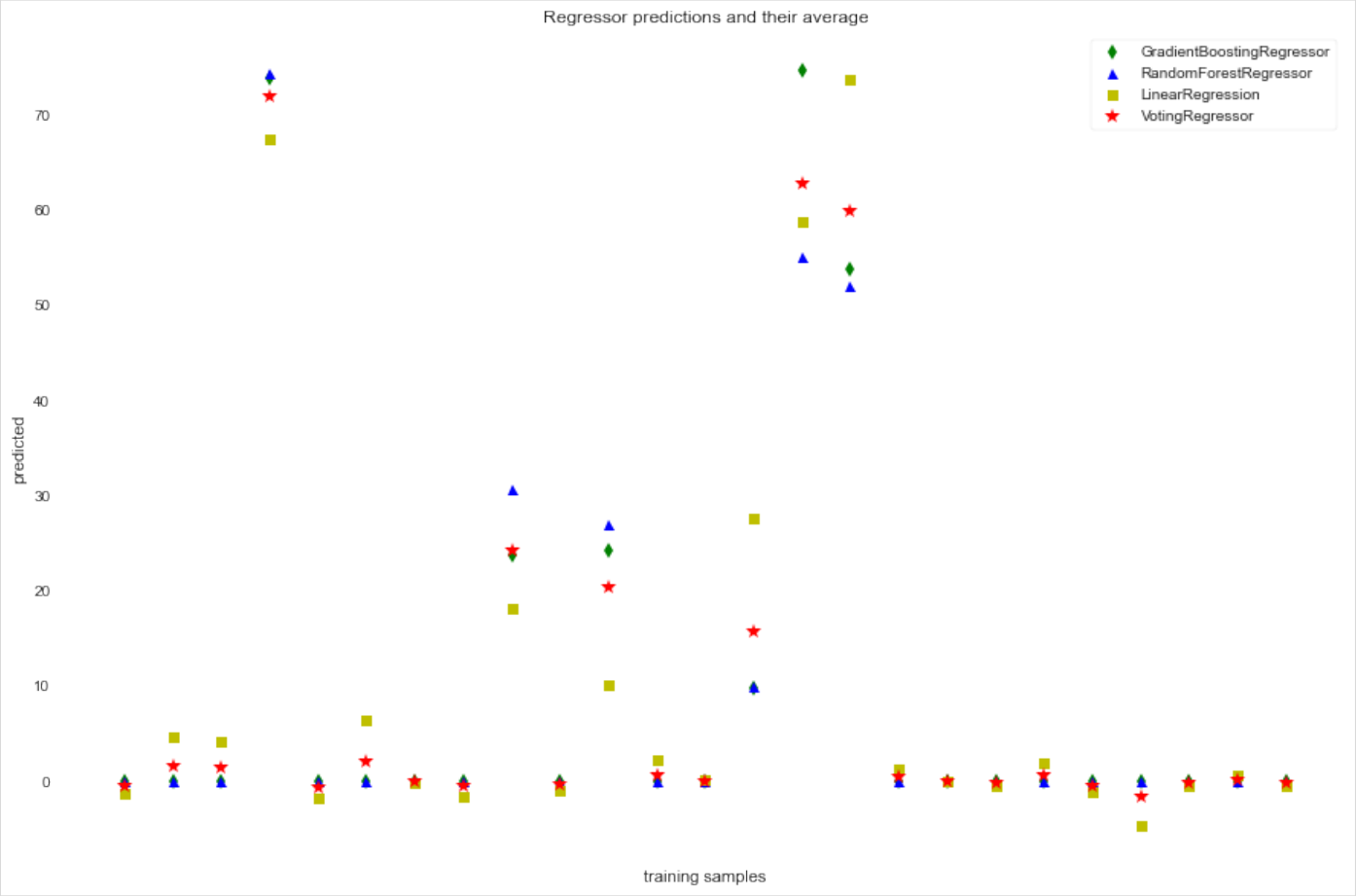
Nearly all schools with the **least** number of SHSAT test takers are composed of 80% or more Black or Latinx students.

Models & Evaluation

My intent is to use regression-based models because my response variables are continuous 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 averaged several base estimators.

Each of the base estimators is using the first 25 schools, as my training & testing sample, to make predictions on the number of SHS offers a school received.



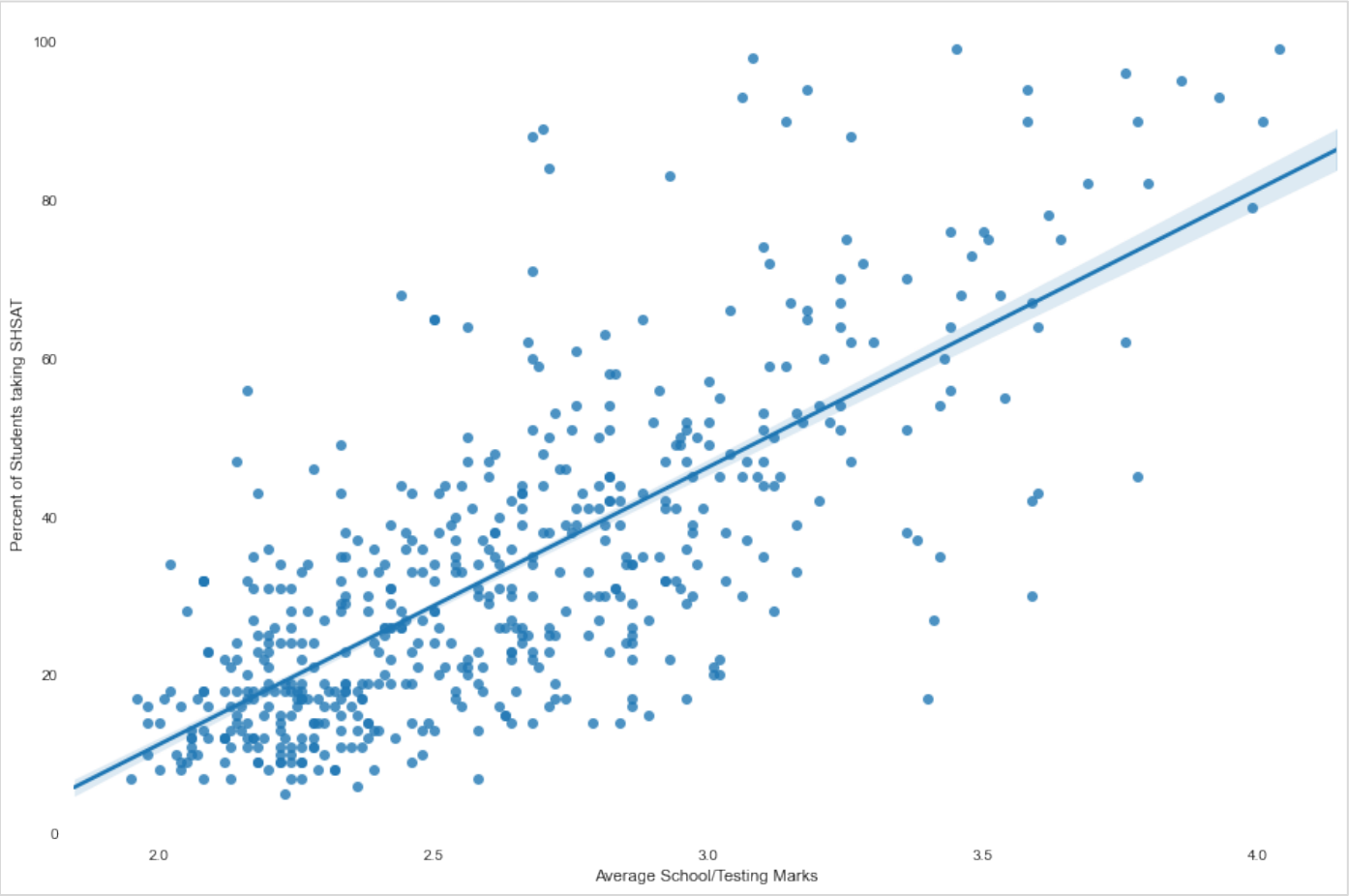
- Gradient Boosting Regressor R^2: 0.930
- Random Forest Regressor R^2: 0.946
- Linear Regression R^2: 0.930
- Voting Regressor R^2: 0.966

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

Predictions & Recommendations

As we saw earlier, there was a strong correlation between the number of SHSAT test takers and the number of SHS offers. By that logic, a school should just send *all* of their eligible students (8th graders), right? That probably wouldn't be a good idea.

The best idea is to send more of your students that're level 4 scorers on the ELA & Math tests. I explored the relationship between the level 4 students and the percentage of them that took the SHSAT.



A higher testing score means a higher percentage of students (in a school) took the SHSAT.

Based on the ELA & Math test scores, which schools could've sent more students to take the SHSAT?

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.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            PerDidSHSAT   R-squared (uncentered):                   0.820
Model:                            OLS   Adj. R-squared (uncentered):              0.820
Method:                 Least Squares   F-statistic:                              2412.
Date:                Tue, 27 Oct 2020   Prob (F-statistic):                   3.24e-199
Time:                        15:11:50   Log-Likelihood:                         -2233.7
No. Observations:                 530   AIC:                                      4469.
Df Residuals:                     529   BIC:                                      4474.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
AvgMark       13.1628      0.268     49.114      0.000      12.636      13.689
==============================================================================
Omnibus:                       94.880   Durbin-Watson:                   1.647
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              147.944
Skew:                           1.148   Prob(JB):                     7.49e-33
Kurtosis:                       4.196   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Here we see a relatively strong relationship (R2 = 0.820) between the percent of the school that took the SHSAT & the ELA/Math scores.

Recommendation #1: Top 25 schools that can send more students (level 4) to the take the test

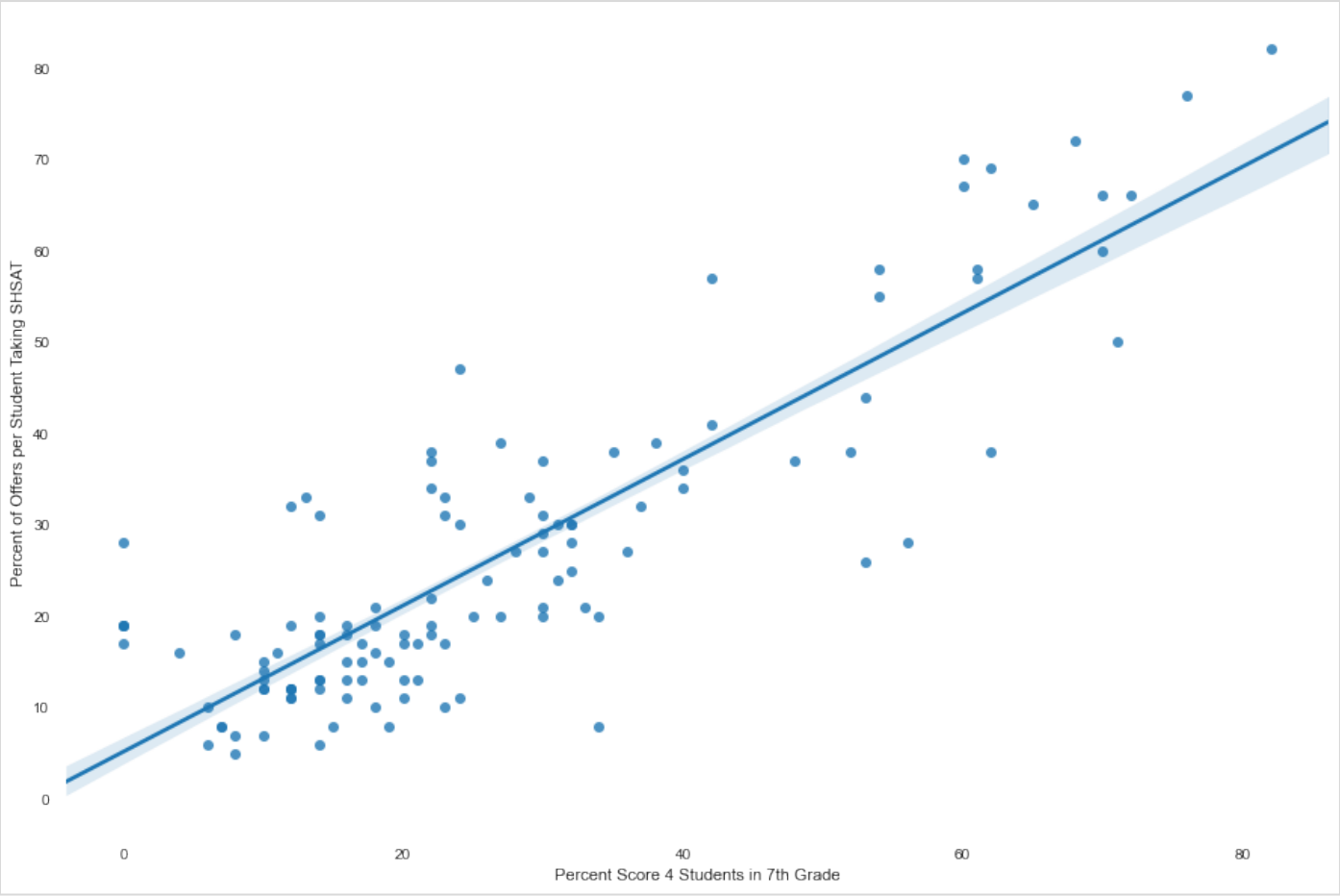
Below, are the predictions by school regarding the percent of students that could have taken the SHSAT (PerModelDidSHSAT). The PotentialTakers takes the difference between the PerModelDidSHSAT and PerDidSHSAT (ModAgainstDidSHSAT) multiplied by the total number of 8th grade students in each school (count_of_students_hs_admissions).

	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 to improving the number Black or Latinx students receiving SHS offers is to look at the likelihood of SHS offers given a high performing student (ELA & Math 4 scorers).

I'm predicting (successful) performance on the SHSAT by looking at the percent of offers among test takers (PctOffersByStudent) against the percent of level-4 students in 7th grade (PctScore4).



Here we see that the higher the percentage of ELA/Math 4 scorers you send to take SHSAT, the more of them receive SHS offers.

Based on the ELA & Math testing scores, which schools could've seen more SHS offers?

```
                            OLS Regression Results
==============================================================================================
Dep. Variable:     PctOffersByStudent   R-squared (uncentered):                   0.914
Model:                            OLS   Adj. R-squared (uncentered):              0.913
Method:                 Least Squares   F-statistic:                              1268.
Date:                Tue, 27 Oct 2020   Prob (F-statistic):                    1.25e-65
Time:                        20:26:42   Log-Likelihood:                         -443.30
No. Observations:                 121   AIC:                                      888.6
Df Residuals:                     120   BIC:                                      891.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
PctScore4      0.9278      0.026     35.602      0.000       0.876       0.979
==============================================================================
Omnibus:                        3.571   Durbin-Watson:                   1.734
Prob(Omnibus):                  0.168   Jarque-Bera (JB):                3.953
Skew:                           0.093   Prob(JB):                        0.139
Kurtosis:                       3.866   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Here we can see a very strong relationship (R2 = 0.914) between those high performing students and SHS offers.

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 **NOT** 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 by sending their level 4 students, 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 SHS. Their high percentage of Black or Latinx students would've improved the rate at which those ethnicities receive offers.

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]