Data Analysis of Average SAT scores of students in NYC Public Schools

Team # 4:
Bhavya Sadrani
Charly Mitchell
Demir McRae
Jayant Bishnoi
Sofia Shur
Banke Choudhry
Tshering Chhoki Sherpa (TC)

Introduction & Background:

The SAT formerly known as a Scholastic Assessment Test is a standardized exam taken by students to apply to graduate schools. The exam structure is divided into 2 sections, Verbal, which constitutes 'Reading', 'Writing' and, Math. The SAT duration is approximately 4 hours. A student can appear for a SAT nearly 5 times in a year.
SAT scores are utilized by colleges across the world as an entrance exam to assess quantitative and qualitative skills of students for future academics.

SAT was introduced in the 1900's and is still considered to be an important aptitude assessment test. One of the biggest goals of SAT is to make sure it's highly relevant to future success of test takers because it focuses on the skills and knowledge at the heart of education. The test taker needs to apply knowledge of what they learned in school, develop an intellect to succeed in college and prove it in the test for critical questions in a stringent timeline. However, The key to a high score in SAT is NOT memorizing words and/or facts one will never use in the real world. Hence, our objective here is to identify patterns, academic facilities, racial influences and external factors that help students ace the SAT

Motivation:

Hypothesis Testing:

A. Higher SAT scores is a collaborative effort of progressive high schools, dedicated teachers and motivated students. Even though the SAT was founded with an intent to test general aptitude and IQ of students, in recent times it is established that there are various factors that can break or make a test taker's SAT scores. *In our analysis, we will prove that students enrolled in AP (Advance Placement) classes have higher test score average than latter.*

What Are AP Classes in High School?

AP (advanced placement) is a program of classes developed by the college board to give high school students an introduction to college-level classes and also gain college credit before

even graduating high school. These courses are more difficult than the usual high school class and also require passing an AP exam at the end of the year to gain college credit.

AP exams are tests on everything you've learned in your AP class that year. They're scored on a scale from 1 to 5 with any score above 3 considered creditworthy.]

B. More than 60000 students took the SAT in New York in 2016; It is said that large racial gaps persist when it comes to students' SAT scores. For instance, among 2015's 11th-graders, white students on average scored 100 points higher than black students and 94 points higher than Hispanics on the math portion of the test, which is scored on a 200 to 800-point scale. Asian students, on average, scored 40 points higher than whites in math. *This motivates us that if (at all) we can identify geographical, psychological, and/or physiological pattern that can be mapped for higher SAT scores*

Some of the other questions that we will answer with the help of our study are as follows:

- What are the average scores of students across New York schools and boroughs?
- What are the influences of good scores, average scores and bad scores? Are there any strong external factors?
- Is there a pattern that defines success in SAT 's for the Test Taker?
- What are the Average SAt scores of students across boroughs in NYC?

Dataset Description:

For detailed analysis, we have opted to merge two datasets that were available online.

Dataset 1:
Link: https://bit.ly/3lZcFUF
This dataset consists of a row for every accredited high school in New York City with its department ID number, school name, borough, building code, street address, latitude/longitude coordinates, phone number, start and end times, student enrollment with race breakdown i.e White, Black, Hispanic or Asian, and average scores on each SAT test section i.e. Reading, Writing and Math, for the 2014-2015 school year.
The high school data of 22 columns and 435 observations was compiled and published by the New York City Department of Education, and the SAT score averages and testing rates were provided by the College Board.

Dataset 2:
Link: https://bit.ly/3GERcIr
Licensed under public domain this dataset has details for schools across the 5 boroughs of New York City with its address, contact details, campus name, website links and program details. In addition to these details, the most important information is on the grade span range, (AP)Advanced Placement courses , online and offline language classes, online AP courses, school sports centers and facilities, partner hospitals, financial and others along with extracurricular activities offered by the school.

The high school data of 64 columns and 435 observations is sourced from NYC open data and merged to our ‘Dataset 1’ to expand the horizon of our study in order to achieve solid results and prove our hypothesis.
Both our datasets are structured and have a combination of quantitative and qualitative variables.

Cleaning process and Final Dataset:
After joining the data by the School ID or DBN, we decided to drop 55 of the 86 resulting variables from merging the two datasets. Some were dropped because this data appeared in both datasets. Some were moved because they were irrelevant to our project. We then took 7 variables which were lists of things offered by the school ("language_classes", "advancedplacement_courses", "online_ap_courses", "online_language_classes", "extracurricular_activites", "psal_sports_boys", "psal_sports_girls", "psal_sports_coed") and created variables counting the length of these lists ("# of language classes", "# of AP courses", "# of online language courses", "# of online ap courses", "# of extracurriculars", "# of boys sports", "# of girls sports, "# of coed sports"). We finally calculated "totalscore", our dependent variable, which is the sum of 3 variables ("Average Score (SAT Math)", "Average Score (SAT Reading)", "Average Score (SAT Writing)").
After dropping these columns and omitting rows that had NA values, we were left with a dataset that had 38 variables and 375 observations. Lastly, we dropped 13 more variables that we thought did not have relevance.

Our primary goal was to weed out irrelevant data and blank columns from the combined sets.
Variable Introduction:

Found the count of this variable	Qualitative	psal_sports_girls	Female PSAL (Public School Athletics League) sports offered at the school, listed under Extracurricular Activities and Clubs in the HS Directory
Found the count of this variable	Qualitative	psal_sports_coed	Co-ed PSAL (Public School Athletics League) sports offered at the school, listed under Extracurricular Activities and Clubs in the HS Directory
Removed for irrelevancy	Qualitative	school_sports	Sports offered by the school outside PSAL (Public School Athletics League)
Removed for irrelevancy	Qualitative	partner_cbo	List of partnerships the school has with community-based organizations
Removed for irrelevancy	Qualitative	partner_hospital	List of partnerships the school has for hospital outreach
Removed for irrelevancy	Qualitative	partner_highered	List of partnerships the school has with higher education institutions
Removed for irrelevancy	Qualitative	partner_cultural	List of partnerships the school has with cultural/arts organizations
Removed for irrelevancy	Qualitative	partner_nonprofit	List of partnerships the school has with not-for-profit institutions
Removed for irrelevancy	Qualitative	partner_corporate	List of partnerships the school has with corporate institutions
Removed for irrelevancy	Qualitative	partner_financial	List of partnerships the school has with financial institutions
Removed for irrelevancy	Qualitative	partner_other	List of other partnerships the school has that do not fall under one of the other partnership categories
Removed for irrelevancy	Qualitative	addtl_info1	Additional information provided by the school (1st of 2 fields), including a dress code or uniform requirement
Removed for irrelevancy	Qualitative	addtl_info2	Additional information provided by the school (2nd of 2 fields), including (but not limited to) extended day program, orientation, internship program, community service requirement.
Removed because this appears in both datasets	Quantitative	start_time	Start time for a typical freshman schedule
Removed because this appears in both datasets	Quantitative	end_time	End time for a typical freshman schedule
Removed for irrelevancy	Qualitative	se_services	Supports for Students with Disabilities
Removed for irrelevancy	Qualitative	ell_programs	ELL Programs offered by the school: ESL, Dual Language, and/or Transitional Bilingual
Removed for irrelevancy	Qualitative	school_accessibility_description	Accessibility of the site where the school is located: Functionally Accessible or Not Functionally Accessible
Removed for irrelevancy	Quantitative	number_programs	Number of distinct programs available at the school (see program table for details). Note that admissions priorities apply for all programs unless otherwise stated in the priority.
Removed for irrelevancy	Qualitative	priority01	Admissions priority #1
Removed for irrelevancy	Qualitative	priority02	Admissions priority #2
Removed for irrelevancy	Qualitative	priority03	Admissions priority #3
Removed for irrelevancy	Qualitative	priority04	Admissions priority #4
Removed for irrelevancy	Qualitative	priority05	Admissions priority #5
Removed for irrelevancy	Qualitative	priority06	Admissions priority #6
Removed for irrelevancy	Qualitative	priority07	Admissions priority #7
Removed for irrelevancy	Qualitative	priority08	Admissions priority #8
Removed for irrelevancy	Qualitative	priority09	Admissions priority #9
Removed for irrelevancy	Qualitative	priority10	Admissions priority #10
Removed for irrelevancy	Qualitative	Location 1	Primary location of school
Removed for irrelevancy	Quantitative	Community Board	Community Board field indicates the New York City Community district where the building is lcocated.
Removed for irrelevancy	Quantitative	Council District	New York City council district whre the building is located.
Removed for irrelevancy	Quantitative	Census Tract	U.S. census tract where the building is located.
Removed for irrelevancy	Quantitative	BIN	Building Identification Number is a unique identifier for each building in the City.
Removed for irrelevancy	Quantitative	BBL	Borough, Block and Lot are a unique identifier for each tax lot in the city.
Removed for irrelevancy	Qualitative	NTA	Neighborhood Tabulation Area field indicates the New York City Neighborhood area where the building is located.

Notes	Type	Name	Description
AVERAGE SAT SCORES FOR NYC PUBLIC SCHOOLS IN 2014-2015			
	Qualitative	School ID	unique ID related to the school
	Qualitative	School Name	name of the school
	Qualitative	Borough	borough the school is located in
	Qualitative	Building Code	code unique to the building the school is located in
	Qualitative	Street Address	street address of the building the school is located in
	Qualitative	City	city the school is located in
	Qualitative	State	state where the school is located in
	Qualitative	Zip Code	zip code of the school
	Quantitative	Latitude	latitude coordinate of school location
	Quantitative	Longitude	longitude coordinate of school location
	Qualitative	Phone Number	phone number of school
	Quantitative	Start Time	typical start time for a freshman schedule
	Quantitative	End Time	typical end time for a freshman schedule
	Quantitative	Student Enrollment	amount of students enrolled in the school
	Quantitative	Percent White	percent of the student population that's White
	Quantitative	Percent Black	percent of the student population that's Black
	Quantitative	Percent Hispanic	percent of the student population that's Hispanic
	Quantitative	Percent Asian	percent of the student population that's Asian
	Quantitative	Average Score (SAT Math)	the average score for the Math section of the SAT amongst students
	Quantitative	Average Score (SAT Reading)	the average score for the Reading section of the SAT amongst students
	Quantitative	Average Score (SAT Writing)	the average score for the Writing section of the SAT amongst students
	Quantitative	Percent Tested	percent of total enrollment that was tested

CALCULATED VARIABLES			
	Quantitative	# of language classes	the number of language classes (other than English) offered by the school
	Quantitative	# of AP courses	the number of the Advanced Placement course offered by the school
	Quantitative	# of online language classes	the number of online language classes (other than English) offered by the school
	Quantitative	# of online ap courses	the number of online Advanced Placement course offered by the school
	Quantitative	# of extracurriculars	the number of extracurriculars offered by the school that are not sports
	Quantitative	# of boys sports	the number of male PSAL sports offered at the school
	Quantitative	# of girls sports	the number of female PSAL sports offered at the school
	Quantitative	# of coed sports	the number of coed PSAL sports offered at the school
	Quantitative	Average Score (SAT Math, Reading and Writing)	totalscore is a sum of Average Scores from SAT Math, Reading and Writing variables

Statistical Analysis:

Before we create a regression model, we will investigate our dataset using Exploratory Data Analysis. This will help us understand better distribution of qualitative and quantitative variables and their relationships. We use visual approaches such as graphs and summary statistics as well as a non-visual approach to explore relationships between dependent and independent variables. We run a multivariate linear regression model on the Total Average SAT Score as our dependent variable. We will use Random Forest and Bagging that randomly selects subsets of features used in each data sample and combines the predictions from all models.

Exploratory Data Analysis (EDA):

We consider Exploratory Data Analysis as an essential step before building our two models in order to select important variables. It provides us with information on distribution of qualitative and quantitative variables and their relationships.

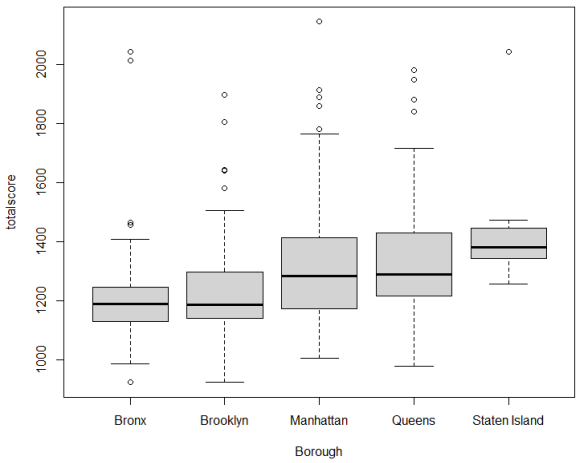
As we can see in the Summary Statistics table below we found that our calculated variable for the total SAT score has a mean of 1,275.9, minimum is 924 and maximum is 2,144 with the Standard Deviation of 194.9.

Table 1: Summary Statistics

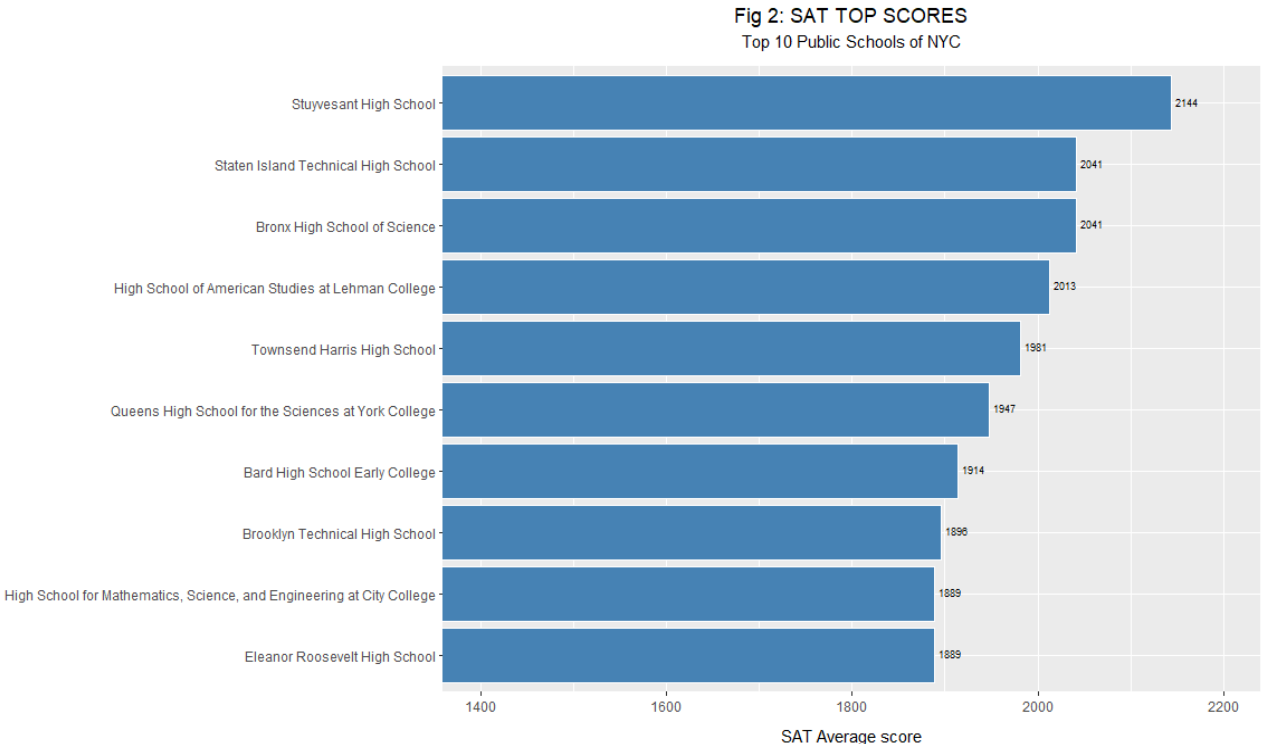
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Zip.Code	375	10,737.8	537.5	10,002	10,306	11,217.5	11,694
Latitude	375	40.7	0.1	40.5	40.7	40.8	40.9
Longitude	375	-73.9	0.1	-74.2	-74.0	-73.9	-73.7
Student.Enrollment	375	764.7	789.5	142	397.5	663.5	5,447
Percent.White	375	0.1	0.1	0.0	0.01	0.1	0.8
Percent.Black	375	0.4	0.3	0.0	0.2	0.5	0.9
Percent.Hispanic	375	0.4	0.2	0.03	0.2	0.6	1.0
Percent.Asian	375	0.1	0.1	0.0	0.02	0.1	0.9
AveMathscore	375	432.9	72.0	317	386	458.5	754
AveReadscore	375	424.5	61.9	302	386	445	697
AveWritescore	375	418.5	64.5	284	382	437.5	693
Percent.Tested	375	0.6	0.2	0.2	0.5	0.8	1.0
X..of.Language.Classes	375	1.7	1.4	0	1	2	11
X..of.online.language.courses	375	0.4	1.5	0	0	0	16
X..of.ap.courses	375	5.2	5.1	0	2	7	31
X..of.online.ap.courses	375	0.6	1.9	0	0	0	19
X..of.extracurriculars	375	19.5	10.8	1	12	24	72
X..of.boys.sports	375	5.8	4.4	0	2	9	17
X..of.girls.sports	375	5.1	3.7	0	2	7	17
X..of.coed.sports	375	0.6	1.0	0	0	1	9
totalscore	375	1,275.9	194.9	924	1,157	1,330.5	2,144

The Boxplot in the Fig 1 displays the distribution of the Average Total SAT Scores grouped by 5 NYC Boroughs. We observe that there is greater variability for Manhattan students' SAT scores as well as larger outliers. Staten Island students have better SAT scores' range, a mean and just one outlier compared to students' scores in other boroughs. To get a better understanding why the SAT scores on average are much lower in the Bronx than in Staten Island can be one of the topics for future research.

Fig 1: The Average Total SAT Score by NYC Borough

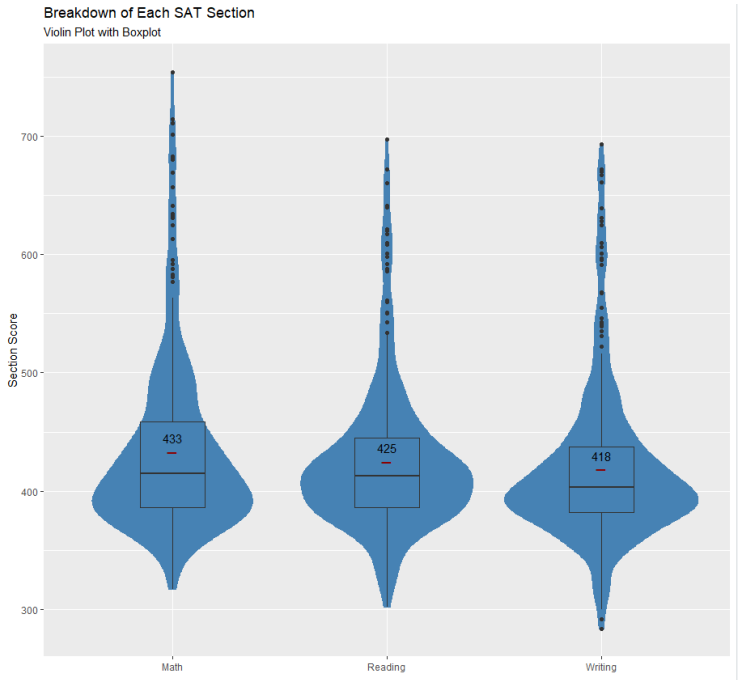


In the Fig 2 we have listed the top 10 schools in NYC. All 5 boroughs have been presented in the top 10 school list according to our dataset.



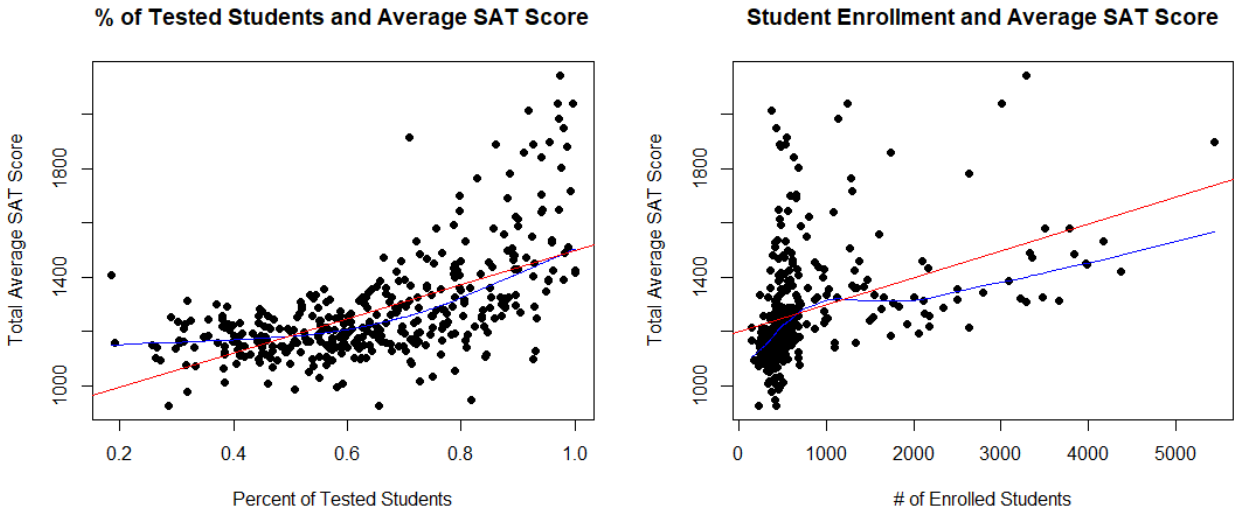
The boxplots in Fig 3 below within each violin plot show that the Math section has the highest average score of all the sections (433), followed by Reading (425), and Writing (418). Math also has several high outliers, indicating that a few schools have a relative strength in Math.

Fig 3:



The first scatter plot on the left in the Fig 4 shows no strong linear relationship between a ratio of students that took the SAT test and the total Average SAT score but it still has some relationship. Similarly, the second scatter plot on the right does not display a linear relationship between a number of enrolled students in school and the total Ave. SAT score.

Fig 4:



We calculated correlation coefficients for all 26 variables to determine the relationship between two variables in the Fig 5. The total score is our dependent variable of interest and we use the rest of variables as independent variables.

Fig 5:

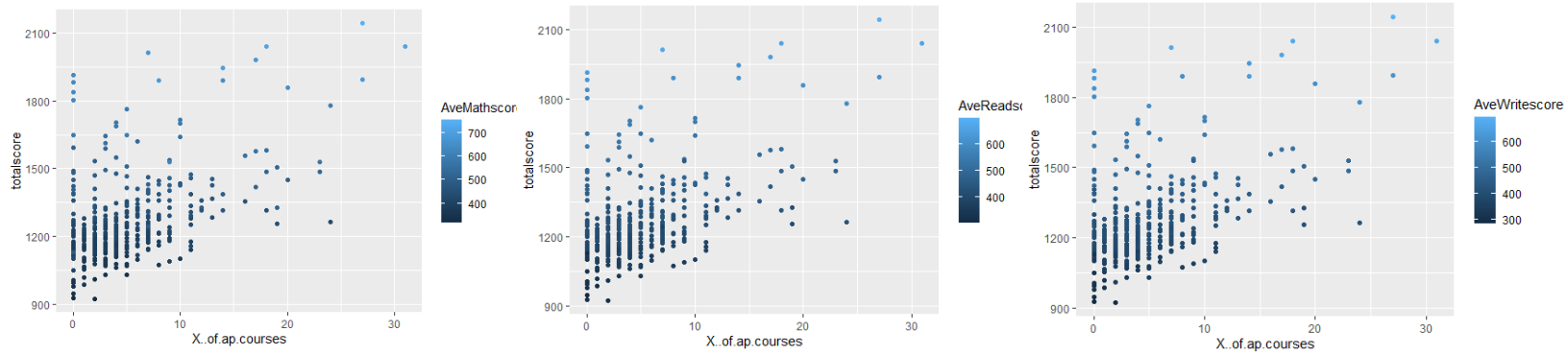
	AveMathscore	AveReadscore	AveWritescore
totalscore	1.0	1.0	1.0
Zip.Code	0.0	-0.1	-0.1
Latitude	-0.1	-0.1	-0.1
Longitude	-0.1	-0.1	-0.1
Student.Enrollment	0.4	0.4	0.4
Percent.White	0.6	0.6	0.6
Percent.Black	-0.4	-0.2	-0.3
Percent.Hispanic	-0.4	-0.4	-0.4
Percent.Asian	0.7	0.5	0.5
AveMathscore	1.0	0.9	0.9
AveReadscore	0.9	1.0	1.0
AveWritescore	0.9	1.0	1.0
Percent.Tested	0.6	0.6	0.6
X..of.Language.Classes	0.4	0.3	0.3
X..of.online.language.courses	0.0	0.0	0.0
X..of.ap.courses	0.5	0.5	0.5
X..of.online.ap.courses	-0.1	-0.1	-0.1
X..of.extracurriculars	0.4	0.4	0.4
X..of.boys.sports	0.1	0.1	0.1
X..of.girls.sports	0.2	0.2	0.2
X..of.coed.sports	0.1	0.0	0.0

As we can see from the correlation table above, we have a positive correlation between the total score and Percent.White/Percent.Asian/Percent.Tested/X of.ap.courses where the

correlation coefficient is greater than 0.4. The results need to be more explored for ethnicity and A.P. courses to explain these relationships. However, the ethnicity ratio is not something that public school can control or change. Therefore, we will focus on A.P. courses and other available resources that schools can provide to improve their students' performance on the SAT exam. This provided us a direction for our further analysis and possible models we would want to utilize for this research.

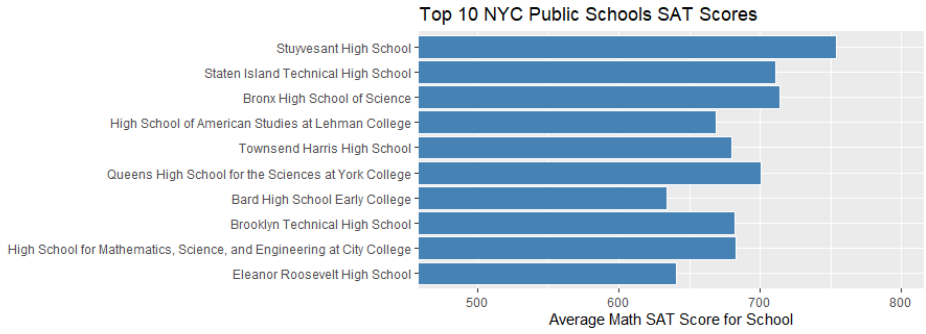
First, let's take a glance at three scatterplots in Fig 6 that represent the relationship between the total score and the number of A.P. courses and provide color coding for the three parts of SAT exam scores: Math, Reading and Writing. The scatter plots represent a positive relationship and it displays higher scores for each SAT part when more than 15 A.P. courses are available.

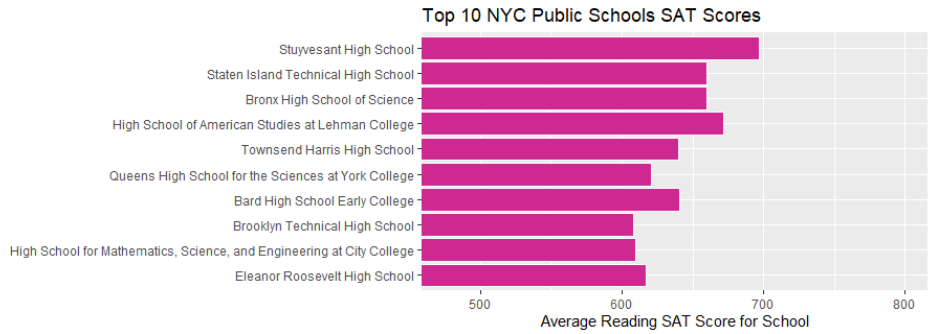
Fig 6:

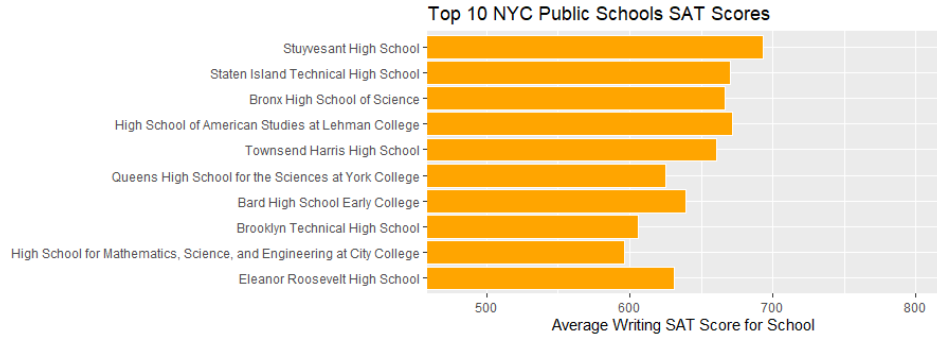


When we compared the top 10 schools with the highest SAT scores, we can see on the Fig 7 with the bar charts that these schools have higher scores on the Math part over Reading and Writing parts in general. It highlights that performance on the Math part could drive the total SAT score to the top tier.
As per below charts, we can conclude that schools in the top 10% tend to do better on the Math section compared to the other two sections: Reading and Writing.

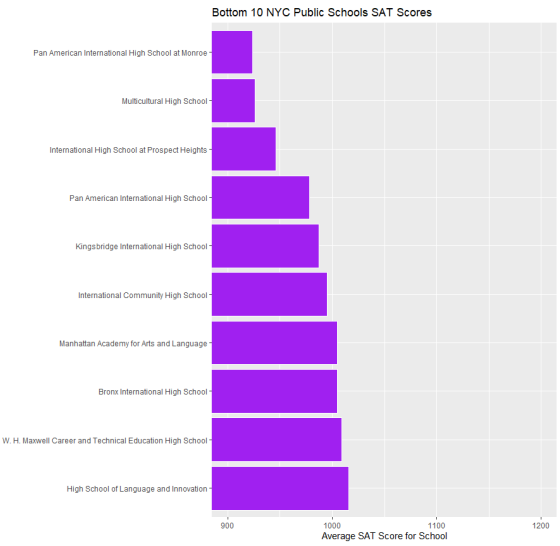
Fig 7:

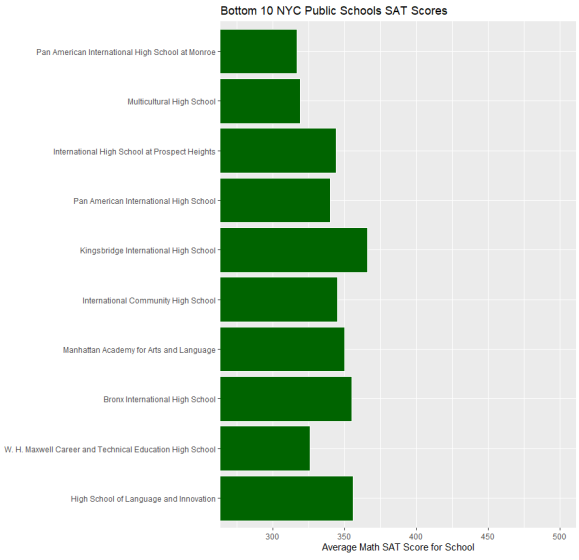


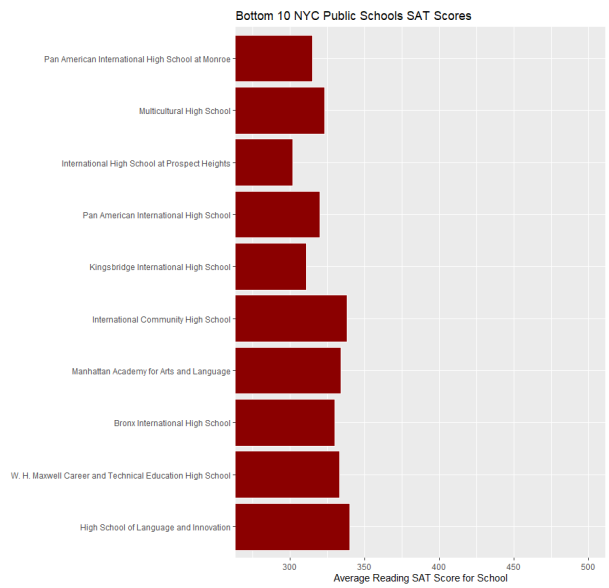


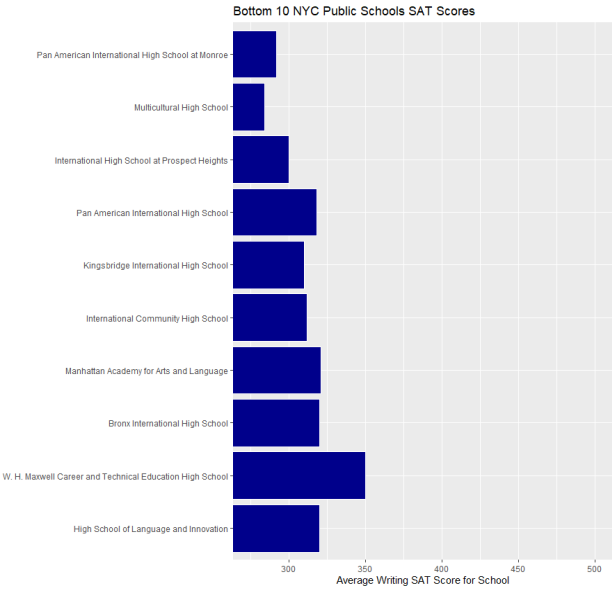


Once we prove our hypothesis by running the models, we can advise public schools that have poor students' performance on the SAT scores to offer various A.P. courses and other additional resources to their students. Our further analysis will investigate variables that have a positive impact on the SAT score.







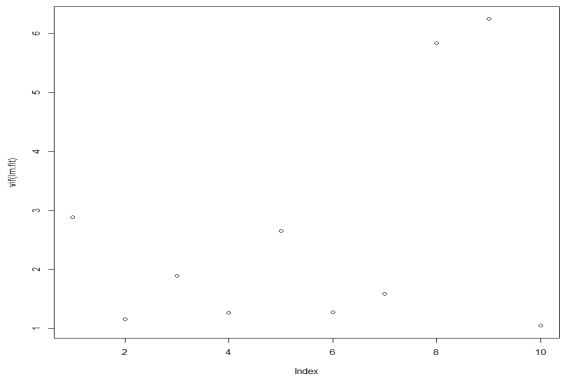


Linear Multiple Regression:

We use the multiple linear regression model to predict an outcome variable (y) on the basis of multiple distinct predictor variables (x)**.**
Here, in our dataset , the column (totalscore) is the dependent variable..
Assumptions of the linear model:

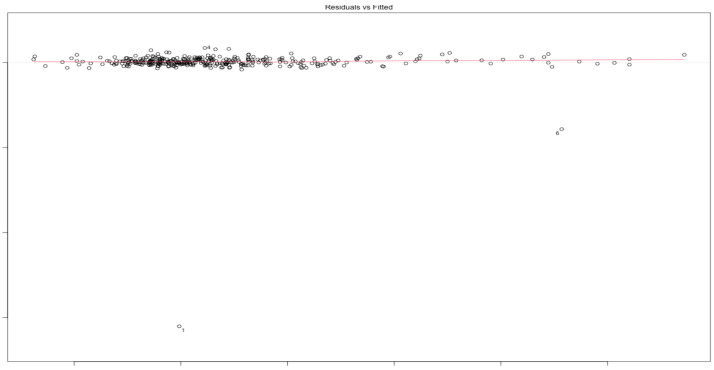
- To have a linear relationship between (totalscore) and the rest of the predictors.
- The independent variables should not be too highly correlated to each other.

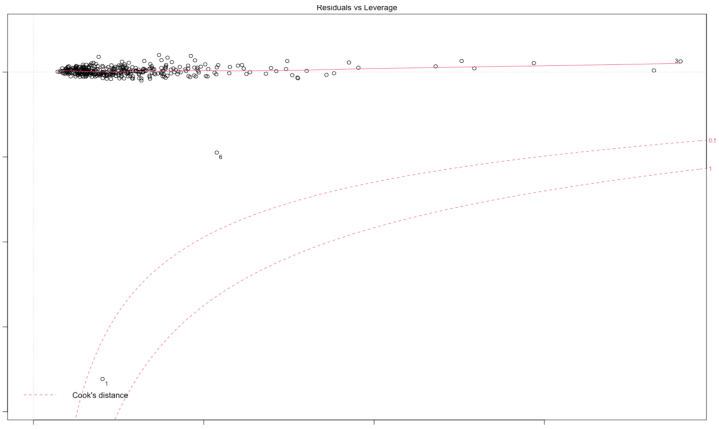
Process:
Firstly, we fit the totalscore against the entire set of predictors. We computed the variance inflation factor(Vif) ,a measure of the amount of multicollinearity among variables used in regression.We observed high multicollinearity among variables which resulted in a flatline plot for regression.
(The plot below , indicates a graphical output of the variance inflation factor as run on the first model which clearly shows high values due to reasons stated above.)



In the residual vs fitted plot, the red line, (that is an indicator of avg. value of residuals at each fitted value), is perfectly **straight ,**there are no discernable non-linear trends to the residuals. The residual points were found to be too close to zero, which meant that the regression had overfit the data. The only significant variables were AveMathscore, AveReadscore, and AveWritescore with very high values ,so it was difficult to predict which of them individually impacts the total score the most; this indicates high multicollinearity.

Zip.Code	1.414e-16	1.590e-16	8.890e-01	0.375	
Latitude	5.097e-13	1.109e-12	4.600e-01	0.646	
Longitude	-1.333e-12	1.065e-12	-1.252e+00	0.211	
Student.Enrollment	7.430e-17	1.022e-16	7.270e-01	0.468	
Percent.White	-4.425e-12	3.100e-12	-1.427e+00	0.154	
Percent.Black	-2.382e-12	3.059e-12	-7.790e-01	0.437	
Percent.Hispanic	-2.651e-12	3.038e-12	-8.730e-01	0.383	
Percent.Asian	-2.883e-12	3.146e-12	-9.160e-01	0.360	
AveMathscore	1.000e+00	2.850e-15	3.508e+14	<2e-16	***
AveReadscore	1.000e+00	4.720e-15	2.119e+14	<2e-16	***
AveWritescore	1.000e+00	4.557e-15	2.194e+14	<2e-16	***
Percent.Tested	2.710e-13	3.210e-13	8.440e-01	0.399	
X..of.Language.Classes	-7.462e-14	4.622e-14	-1.615e+00	0.107	
X..of.online.language.courses	-2.177e-14	3.404e-14	-6.400e-01	0.523	
X..of.ap.courses	2.081e-14	1.541e-14	1.350e+00	0.178	
X..of.online.ap.courses	-1.997e-14	2.744e-14	-7.280e-01	0.467	
X..of.extracurriculars	2.760e-15	5.385e-15	5.120e-01	0.609	
X..of.boys.sports	1.411e-14	2.560e-14	5.510e-01	0.582	
X..of.girls.sports	4.624e-16	3.106e-14	1.500e-02	0.988	
X..of.coed.sports	-1.984e-14	4.479e-14	-4.430e-01	0.658	





We chose to drop Zip.code, Latitude and Longitude, because even though their data is a numeric type,they didn't provide any real-world value to the regression, AveMathscore, AveReadscore, AveWritescore because they were too highly related to each other.

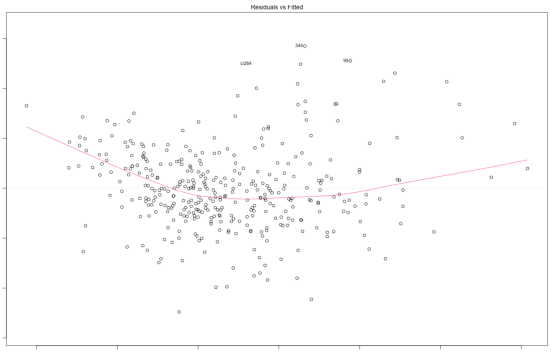
```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.827e+00  2.059e-02 331.563  < 2e-16 ***
Student.Enrollment            1.305e-05  1.063e-05   1.228   0.2203
Percent.Tested                3.720e-01  2.826e-02  13.165  < 2e-16 ***
X..of.Language.Classes       -5.406e-04  4.932e-03  -0.110   0.9128
X..of.online.language.courses -1.852e-03  3.649e-03  -0.508   0.6121
X..of.ap.courses              7.142e-03  1.575e-03   4.534 7.87e-06 ***
X..of.online.ap.courses      -5.121e-03  2.908e-03  -1.761   0.0791 .
X..of.extracurriculars        1.338e-03  5.749e-04   2.327   0.0205 *
X..of.boys.sports            -3.738e-03  2.737e-03  -1.366   0.1728
X..of.girls.sports            5.346e-03  3.308e-03   1.616   0.1069
X..of.coed.sports            -5.010e-04  4.836e-03  -0.104   0.9175
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09556 on 364 degrees of freedom
Multiple R-squared:  0.5408,    Adjusted R-squared:  0.5282
F-statistic: 42.87 on 10 and 364 DF,  p-value: < 2.2e-16
```

Here, Percent.Tested , # of ap.courses , # .of online ap courses , # of extracurriculars all are positive and significant effect in predicting the response variable , with # of ap courses having the most significance.
We ran another model with the remaining predictors. We obtain standardized residuals to identify the outliers in our regression model.
We have observed:

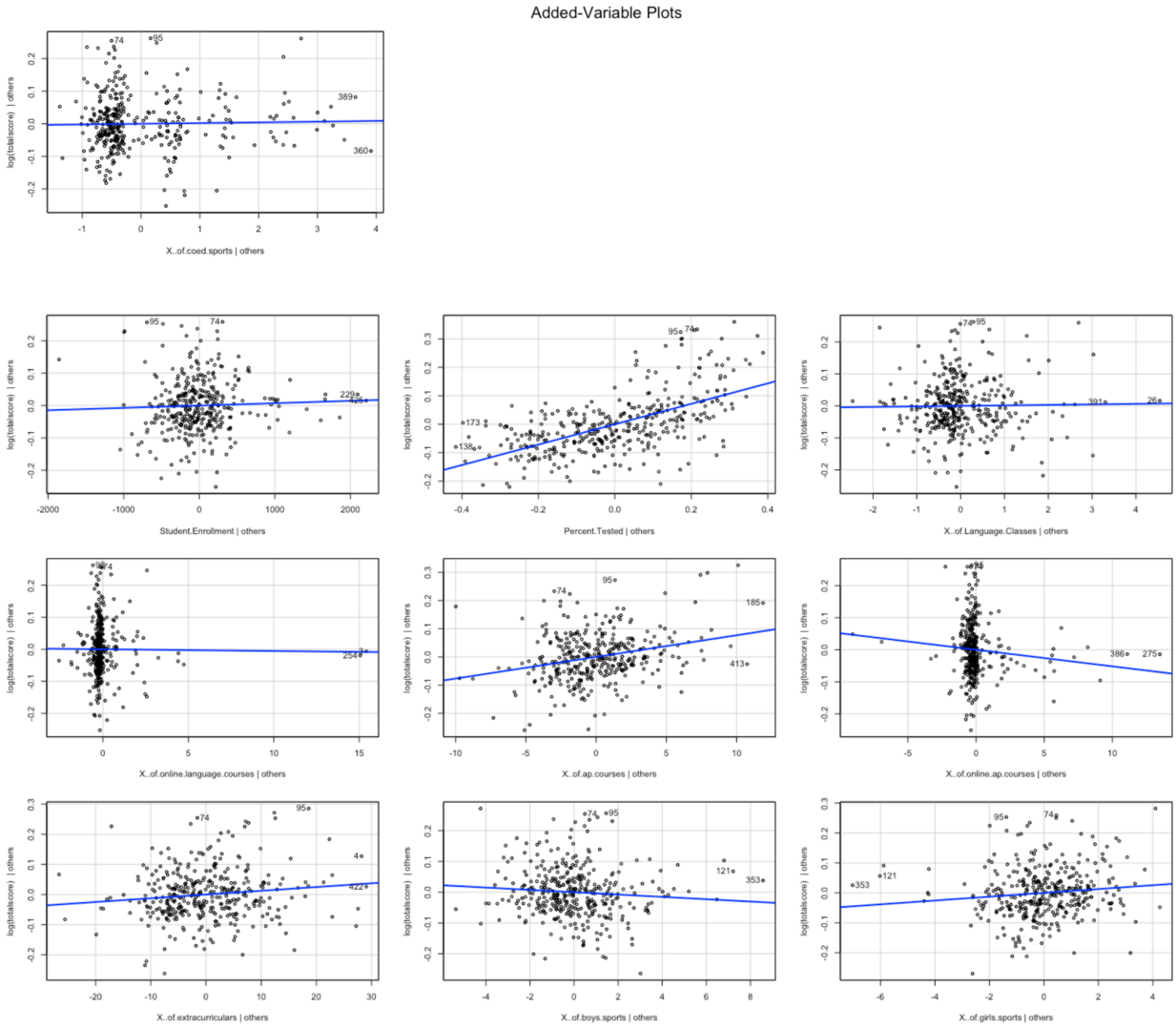
1. many outliers have significant leverage
2. adjusted R square shows us how well the model fits overall (0.5282) for this model.
3. hetrosecardicity

Finally, we ran Cooks' distance function to identify and filter out any values greater than 3 times the mean. then, Based on the information shown in the residuals vs fitted plot and residuals vs leverage we removed observations (6, 39, 154,90, 200, 244, 299, 202, 231, ,264,344,359, 314). We have also dropped all the variables on ethnicity to further reduce collinearity and fit a log function over response and fit the remaining predictors into a del to get rid of any potential herosecadicity.. We repeated this process 2 times and removed a total of 13 observations that we felt we highly impacted the fit of the model. We manually then remove the outliers and run the model to see if the accuracy level improves.



The adjusted R value improves up to 0.57 which provides us with a better model accuracy comparing to its predecessor.

```
Residual standard error: 0.7268 on 351 degrees of freedom
Multiple R-squared:  0.5897,    Adjusted R-squared:  0.5757
F-statistic: 42.05 on 12 and 351 DF,  p-value: < 2.2e-16
```

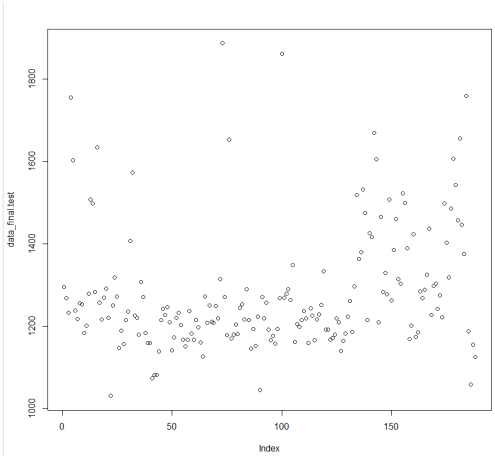


Random Forest:

We would try to use bagging and random forest machine learning techniques to fit our model. It is an ensemble of decision trees, usually trained with the "bagging" method. It is a supervised learning algorithm. Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the bagging variance of a statistical learning method. The general idea of the bagging method is that a combination of learning models increases the overall result and is particularly helpful in the case of decision trees.

"The main difference between bagging and random forests is the choice of predictor subset size m. For instance, if a random forest is built using m = p, then this amounts simply to bagging."

We first apply bagging, which is simply a special case of a random forest with m = p. The argument mtry = 12 indicates that all 12 predictors should be considered for each split of the tree—in other words, that bagging should be done. The test set MSE associated with the bagged regression tree is 15780.4.



We obtain a random forest using the randomForest function, except that we pass a smaller value to the mtry argument. Here we try mtry=12/3. The MSE is 15498.8 , which is lower than the MSE achieved using bagging, so random forest performed better.

Using the importance() function, we can view the importance of each variable.

The results indicate that across all the trees considered in the random forest, the Ethnicity and # of A.P. Courses (X.of.ap.courses) as taken are by far the most important variables.

Practical Implications:

We concluded that there is some relationship between the total score and ethnicity. This could be related to cultural differences in parenting and studying habits. This factor plays some role in students' test performance. However, public schools do not have control over and have limited resources to address it. Therefore, we would focus on other factors that are important for high performance on the SAT exam and also controllable by public schools. The first one is to provide

a variety of in-person A.P. courses to the students. Ideally, public schools should offer over 15 different courses to their students. Proving in person A.P courses over online courses might play a major role in improving students' Math, Reading and Writing skills. As per our data analysis, the math part is important to success in the SAT exam. Schools that are underperforming in SAT exams could shift their focus and priorities to Math oriented subjects and provide in-person Math tutoring. Eventually, it will help students to obtain a higher test score and build their confidence. Our research needs data from private schools to analyze and compare dependent and independent variables. As a result, it will provide better insights on significant variables for the SAT score improvement. Therefore, it will lead to a better translation of findings and recommendations for public schools by us.

The higher scores provide more opportunities for students at disadvantage to get scholarships for college and get into a better college of their choice. High paid salary/jobs. Based on our analysis, retaking SAT exams might help students to obtain better scores.

Limitations:

There are a few limitations that schools may face that can impact their SAT Scores such as lack of funding, some schools lack funding that could provide vouchers for exams so students don't have to come out of pocket or even funding to have schools give the opportunity to host practice exams to give students more help and practice and to take the exam multiple times if needed. Another limitation is the lack of resources the schools have. For example, some schools don't have the proper SAT books to help students prepare or they don't have enough teachers to create a SAT prep class that can be held to guide students through the exam. Background can also be a limitation as a students family background and how involved they are can reflect on their work in school and on the exam. Another factor is how well they are doing in their studies outside of this exam. Sometimes certain students struggle with course work as well as exam taking and that can influence their SAT Scores. Location is also a big factor, depending on your school location and what influences go on around these areas can impact the students and how well they are paying attention or being taught in school due to locational distractions.

References

Data Sources:
Average SAT Scores for NYC Public Schools in 2014-2015:
https://www.kaggle.com/nycopendata/high-schools
2014-2015 DOE High School Directory:
https://data.cityofnewyork.us/Education/2014-2015-DOE-High-School-Directory/n3p6-zve2

R studio package:
Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
R package version 5.2.2. https://CRAN.R-project.org/package=stargazer