Introduction

This project aims to investigate the relationship between the overall academic performance of high school students and their institutional setting. In this study, we analyze data from 372 high schools in New York City. The data consisted of the average SAT scores of 2014-2015 school's cohorts, along with various school and cohort's attributes, such as the school's borough and ethnicity proportion. Our intention is to identify variables that potentially affect the overall academic outcomes, and to quantify the extent that such variables have.

Data

Data	Source	Description
score.csv	https://www.kaggle.com/nycopendata/high-schools	Average SAT scores(Math, Reading, Writing), along with various attributes of 435 schools in NYC. The data pertained to 2014-2015 cohorts
demographics.csv	http://schools.nyc.gov/NR/rdonlyres/46093164-D8AA-40DD-A400-8F80CEBC8DD5/0/DemographicSnapshot201112to201516Public_FINAL.xlsx	Contains information about the gender proportion of each school
survey_2014.csv	http://schools.nyc.gov/documents/misc/2014%20Public%20Data%20File%20SUPPRESSED.xlsx	2014 survey result collected from parents and teachers

Associated Variables

In this analysis, we combine data from the mentioned 3 files. We narrow down the associated features from the original sources to just 26 variables. We use R-script(mungdata.R) to perform the data munging and collect the processed result in processed_score.csv. Noted that the total number of observation that we analyze is reduced to 372 instances because of the missing SAT score in some of the data in score.csv. The explanation of each variable in processed_score.csv is shown in the following table.

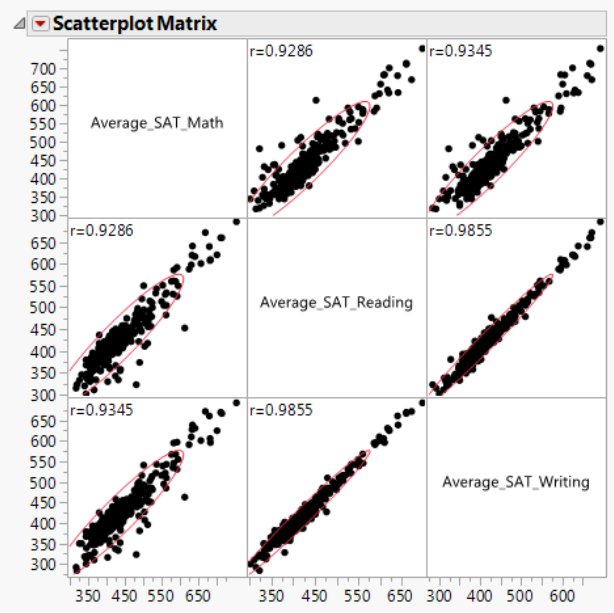
Variable Name	Description	Type
DBN	School's unique identifier	Character
Borough	School's Borough. Comprised of 5 area: Staten Island, Queens, Manhattan, Brooklyn, Bronx	Character
City	City where the school is located	Character
Latitude	School's Latitude	Numeric

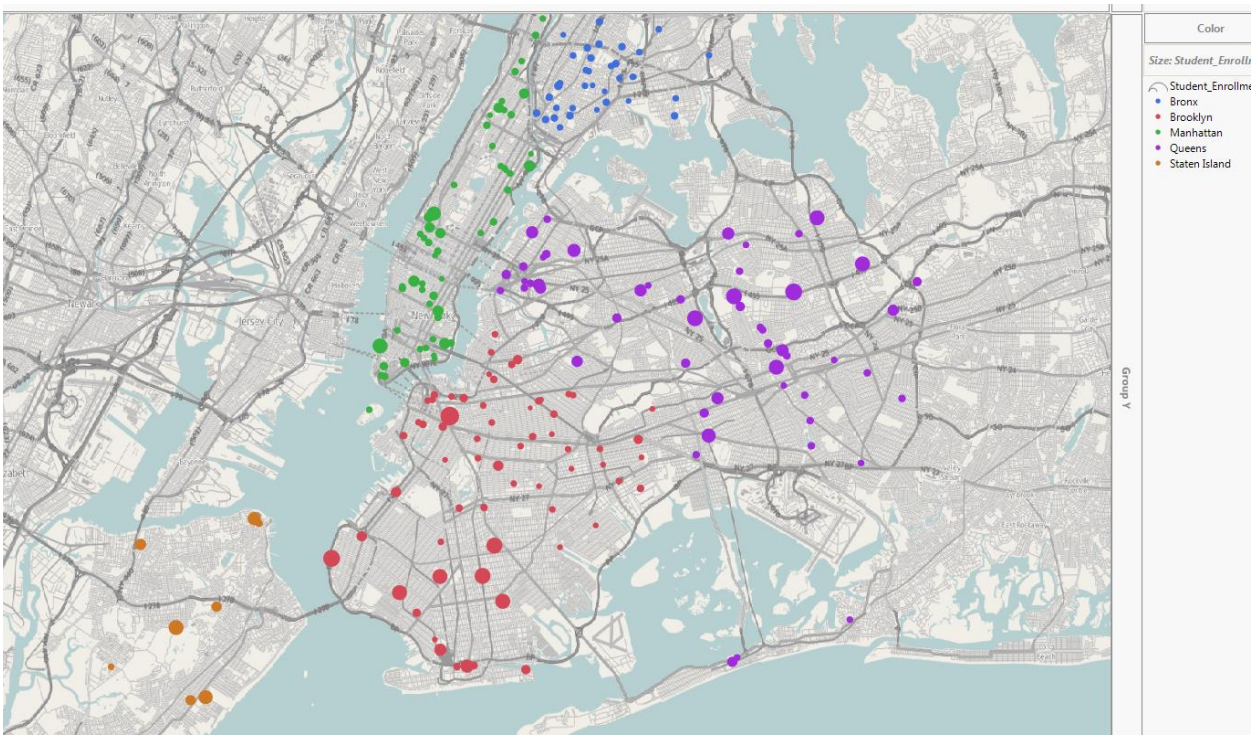
Longitude	School's Longitude	Numeric
Start_Time	School's Opening hour	Numeric(e.g.: convert from 8:15 AM to 8.15)
End_Time	School's Ending hour	Numeric(e.g.: convert from 4:00 PM to 16.00)
Student_Enrollment	Number of school's enrollment	Numeric
Percent_White	%White students in 2014-2015 cohort	Numeric
Percent_Black	%Black students in 2014-2015 cohort	Numeric
Percent_Hispanic	%Hispanic students in 2014-2015 cohort	Numeric
Persent_Asian	%Asian students in 2014-2015 cohort	Numeric
*Average_SAT_Math	Average SAT Math score of 2014-2015 cohort	Numeric
*Average_SAT_Reading	Average SAT Reading score of 2014-2015 cohort	Numeric
*Averate_SAT_Writing	Average SAT Writing score of 2014-2015 cohort	Numeric
Female_Percent	%Female students in 2014-2015 cohort	Numeric
Male_Percent	%Male students in 2014-2015 cohort	Numeric
Disabilities_Percent	%Disability students in 2014-2015 cohort	Numeric
EngLearner_Percent	%English learner students in 2014-2015 cohort	Numeric
Poverty_Percent	%Poverty students in 2014-2015 cohort	Numeric
Parent_Response_Rate	Parent response rate on 2014 school's survey	Numeric
Teacher_Response_Rate	Teacher response rate on 2014 school's survey	Numeric
Instructional_Core_Satisfaction	%Response regarding instructional satisfaction	Numeric
Systems_for_Improvement_Satisfaction	%Response regarding system satisfaction	Numeric
School_Culture_Satisfaction	%Response regarding culture satisfaction	Numeric
Class_Hours	School's operating duration	Numeric(difference in hour: end_time – open_time)

* - Dependent variable

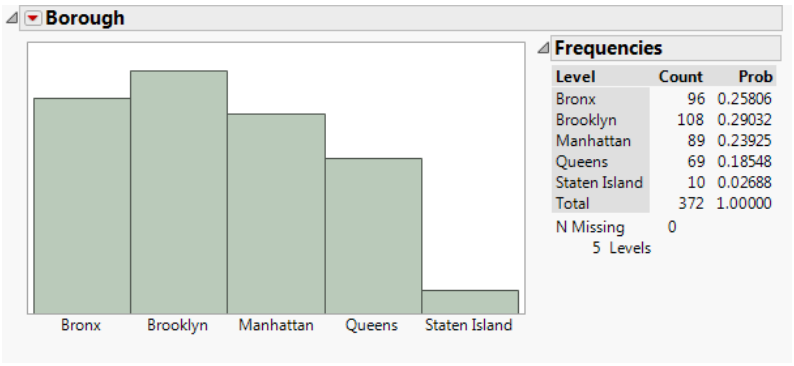
Exploratory Analysis

Firstly, we look at the relationship among all SAT scores. They are(unsurprisingly) highly correlated. So, in our analysis, we will put more emphasis on the SAT-Math score and later apply our findings to SAT-Reading and SAT-Writing scores.

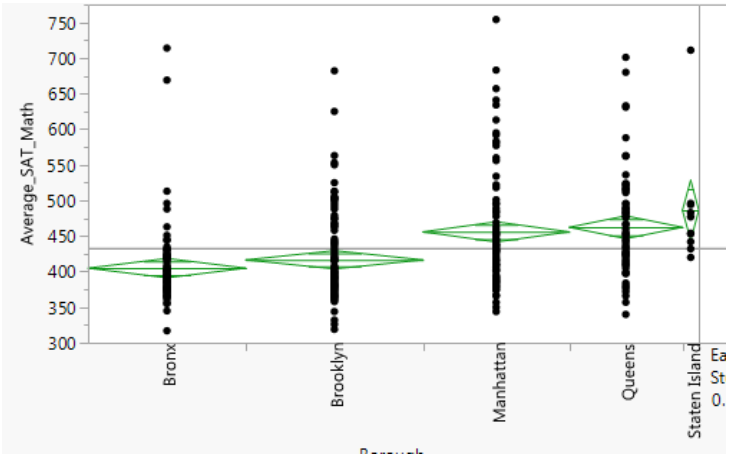




We indicate the location of each school in the above graph. The circle's size corresponds to the size of enrollment. The school's distribution in each borough can be summarized as follows:

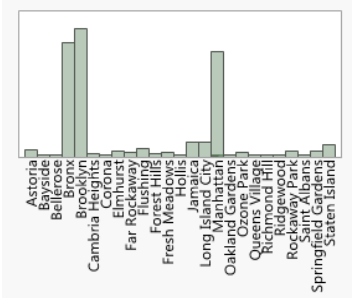


We investigate the effect of spatial information (Borough) on SAT Math score by performing ANOVA. We find that this variable maybe helpful in predicting the SAT score.

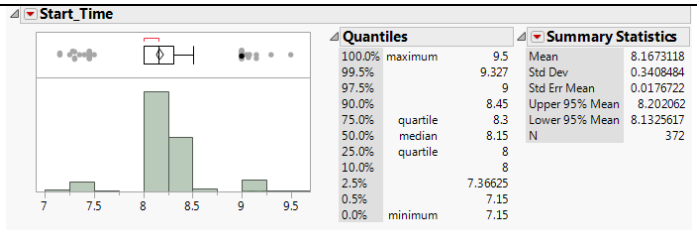
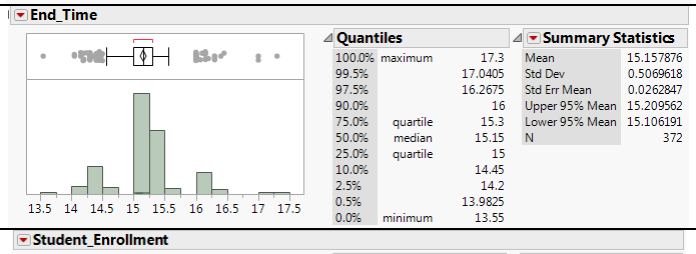
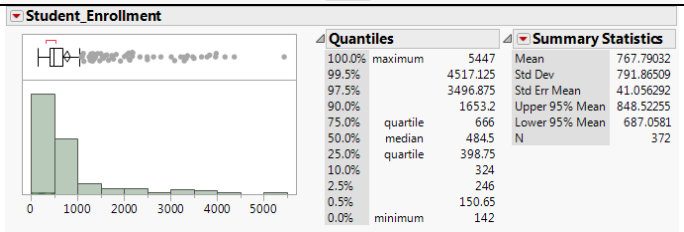
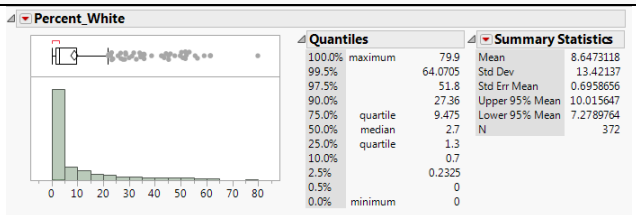
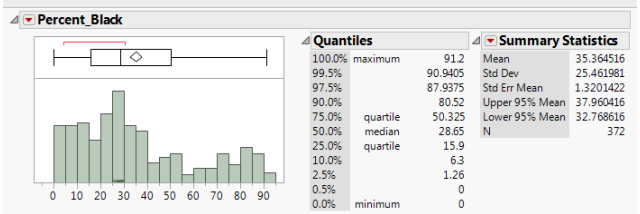
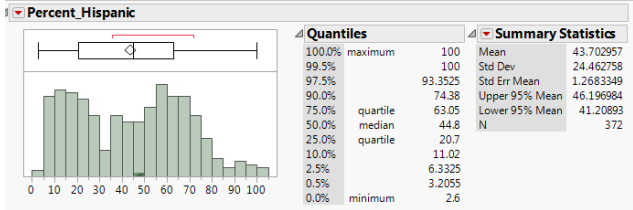
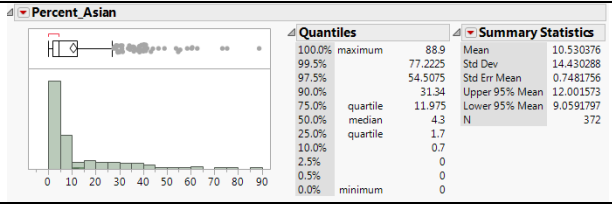


Level		Mean
Staten Island	A	486.20000
Queens	A	462.36232
Manhattan	A	455.88764
Brooklyn	B	416.62963
Bronx	B	404.93750

We also investigate the distribution of schools in each City but find that this variable is too fined-grained and decide to drop it as fear of running into overfitting.



The statistics of other variables are shown in the following table.

Variable	Statistics
Start_Time	
End_Time	
Student_Enrollment	
Percent_White	
Percent_Black	
Percent_Hispanic	
Persent_Asian	

Start_Time

Quantiles		
100.0%	maximum	9.5
99.5%		9.327
97.5%		9
90.0%		8.45
75.0%	quartile	8.3
50.0%	median	8.15
25.0%	quartile	8
10.0%		8
2.5%		7.36625
0.5%		7.15
0.0%	minimum	7.15

Summary Statistics	
Mean	8.1673118
Std Dev	0.3408484
Std Err Mean	0.0176722
Upper 95% Mean	8.202062
Lower 95% Mean	8.1325617
N	372

End_Time

Quantiles		
100.0%	maximum	17.3
99.5%		17.0405
97.5%		16.2675
90.0%		16
75.0%	quartile	15.3
50.0%	median	15.15
25.0%	quartile	15
10.0%		14.45
2.5%		14.2
0.5%		13.9825
0.0%	minimum	13.55

Summary Statistics	
Mean	15.157876
Std Dev	0.5069618
Std Err Mean	0.0262847
Upper 95% Mean	15.209562
Lower 95% Mean	15.106191
N	372

Student_Enrollment

Quantiles		
100.0%	maximum	5447
99.5%		4517.125
97.5%		3496.875
90.0%		1653.2
75.0%	quartile	666
50.0%	median	484.5
25.0%	quartile	398.75
10.0%		324
2.5%		246
0.5%		150.65
0.0%	minimum	142

Summary Statistics	
Mean	767.79032
Std Dev	791.86509
Std Err Mean	41.056292
Upper 95% Mean	848.52255
Lower 95% Mean	687.0581
N	372

Percent_White

Quantiles		
100.0%	maximum	79.9
99.5%		64.0705
97.5%		51.8
90.0%		27.36
75.0%	quartile	9.475
50.0%	median	2.7
25.0%	quartile	1.3
10.0%		0.7
2.5%		0.2325
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	8.6473118
Std Dev	13.42137
Std Err Mean	0.6958656
Upper 95% Mean	10.015647
Lower 95% Mean	7.2789764
N	372

Percent_Black

Quantiles		
100.0%	maximum	91.2
99.5%		90.9405
97.5%		87.9375
90.0%		80.52
75.0%	quartile	50.325
50.0%	median	28.65
25.0%	quartile	15.9
10.0%		6.3
2.5%		1.26
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	35.364516
Std Dev	25.461981
Std Err Mean	1.3201422
Upper 95% Mean	37.960416
Lower 95% Mean	32.768616
N	372

Percent_Hispanic

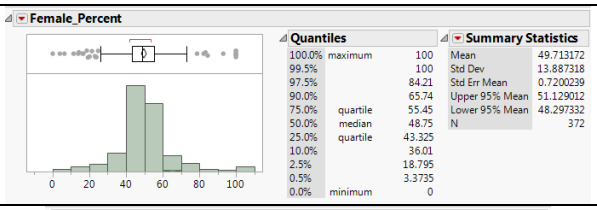
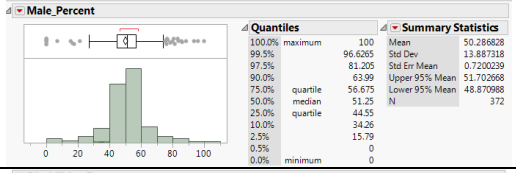
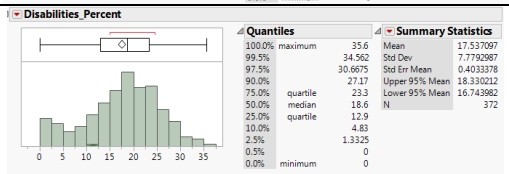
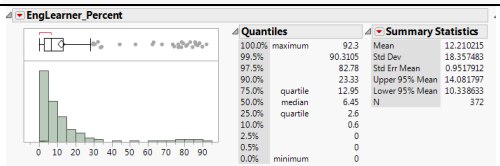
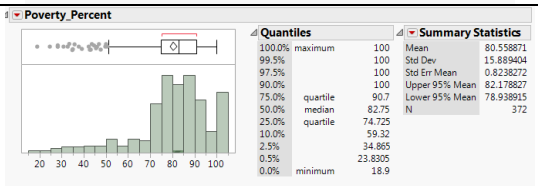
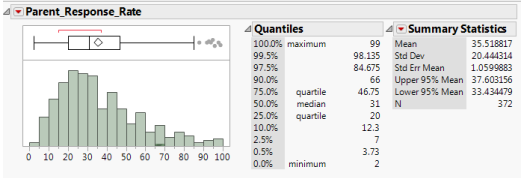
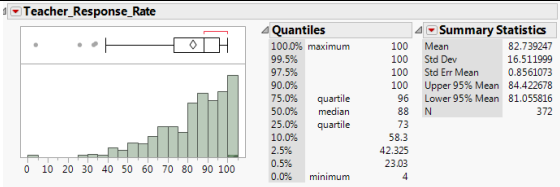
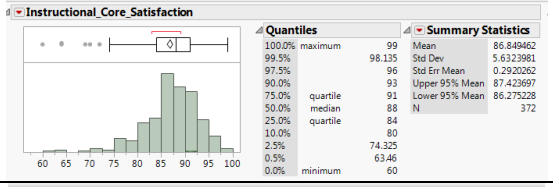
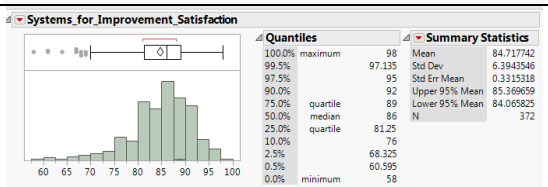
Quantiles		
100.0%	maximum	100
99.5%		100
97.5%		93.3525
90.0%		74.38
75.0%	quartile	63.05
50.0%	median	44.8
25.0%	quartile	20.7
10.0%		11.02
2.5%		6.3325
0.5%		3.2055
0.0%	minimum	2.6

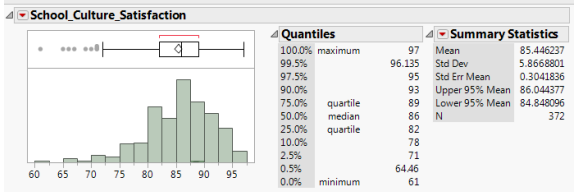
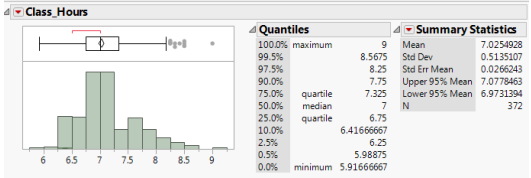
Summary Statistics	
Mean	43.702957
Std Dev	24.462758
Std Err Mean	1.2683349
Upper 95% Mean	46.196984
Lower 95% Mean	41.20893
N	372

Percent_Asian

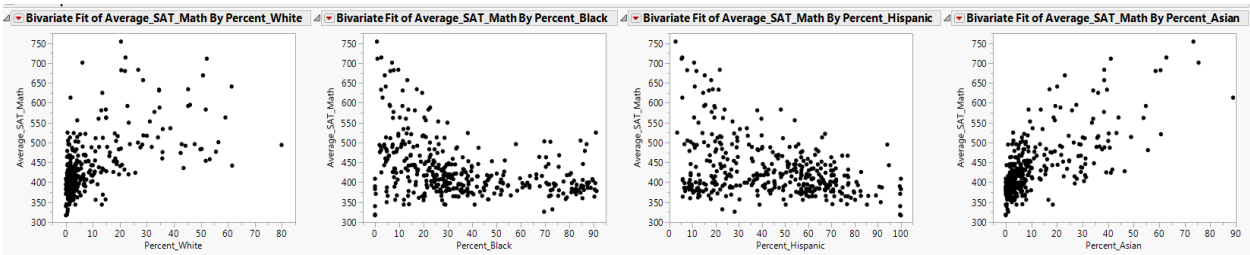
Quantiles		
100.0%	maximum	88.9
99.5%		77.2225
97.5%		54.5075
90.0%		31.34
75.0%	quartile	11.975
50.0%	median	4.3
25.0%	quartile	1.7
10.0%		0.7
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	10.530376
Std Dev	14.430288
Std Err Mean	0.7481756
Upper 95% Mean	12.001573
Lower 95% Mean	9.0591797
N	372

Female_Percent	
Male_Percent	
Disabilities_Percent	
EngLearner_Percent	
Poverty_Percent	
Parent_Response_Rate	
Teacher_Response_Rate	
Instructional_Core_Satisfaction	
Systems_for_Improvement_Satisfaction	

School_Culture_Satisfaction	
Class_Hours	

For preliminary analysis, the scatter plot between ethnicity proportion and SAT-Math score shows some predictive power and indicates that these variables should be included in the model.

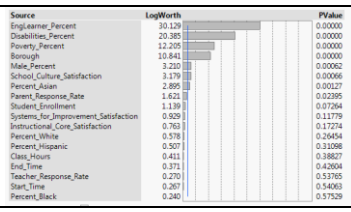
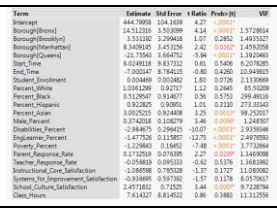
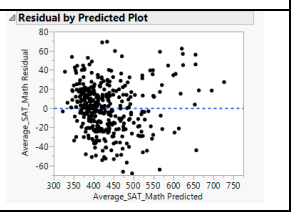


Project Objective

Uncover the relationship between the academic setting and cohort's academic outcomes(as measured by the average SAT scores). We also quantify the variable's effect on 3 different sections of SAT exams(Math, Reading, Writing) to determine their predictive power on each section. Linear Regression is chosen as our base model to fit the data on because of its simplicity and interpretability.

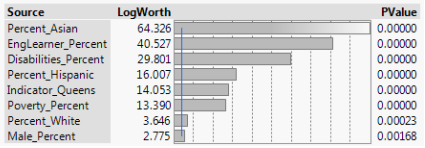
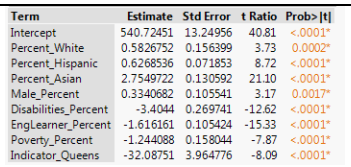
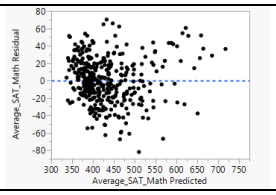
Model Development

As stated earlier, our model will be developed based on using Average SAT-Math score as a dependent variable. The initial model building consisted of all independent variables (exclude DBN, City, Latitudes, Longitudes, and Female_Percent (as this variable is reflected in Male_Percent)). We obtain the following model:

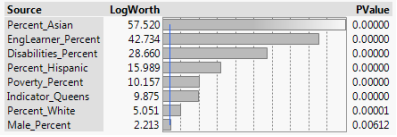
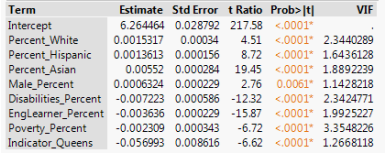
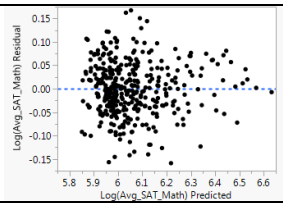
Effect Summary	Parameter Estimates	Residual Plot
		

Although the base model shows a strong predictive power (RSquare = 0.87 and RMSE = 26), it has many undesirable properties; the model contains many variables that are not statistically significant, some independent variables are highly correlated (as shown by VIF), and the Residual Plot shows Heteroscedasticity problem (verified by Park-Test). To attenuate theses effects, we perform a series of model development, which can be summarized as follows:

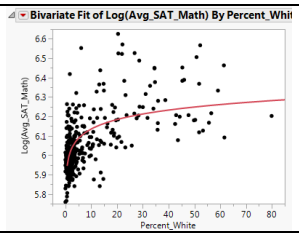
1. Manually create a dummy variable based on Borough. As opposed to the one generated by JMP, this will allow us to remove an individual borough that we found not significant. Bronx is treated at the base level since it has the lowest Average SAT-Math score means.
2. Re-fit the model. Iteratively remove variables with high VIF and P-Value that exceeds 0.01 significant threshold.

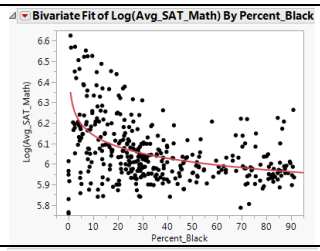
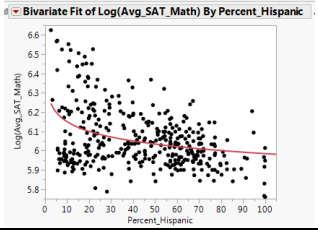
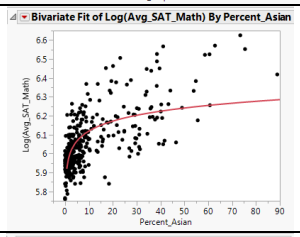
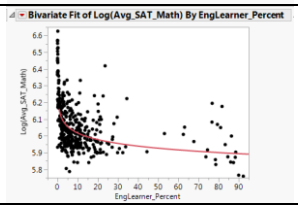
Effect Summary	Parameter Estimates	Residual Plot
		

3. As Heteroscedasticity is still presented, we apply Log transformation to Average SAT-Math score and re-fit the model. Drop any unnecessary variable as stated in 2.)

Effect Summary	Parameter Estimates	Residual Plot
		

4. We try to eliminate Heteroscedasticity by plotting every independent variable against Log(Avg_SAT_Math) and transform them appropriately if we think that leads to a more linear relationship.

Variable	Plot of variable vs Log(Avg_SAT_Math) and transformation fit	Transformation Taken
*Percent_White		Log(Percent_White + 0.1)

*Percent_Black		Log(Percent_Black + 0.1)
*Percent_Hispanic		Log(Percent_Hispanic + 0.1)
*Percent_Asian		Log(Percent_Asian + 0.1)
*EngLearner_Percent		Log(EngLearner_Percent)

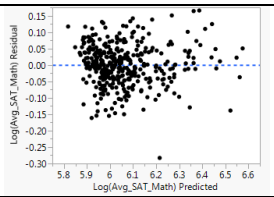
* Add 0.1 to the original value before applying the Log transformation because some instances have 0 value

5. Re-fit the model using the transformed variables. Drop any unnecessary variable as stated in 2.)

Effect Summary	Parameter Estimates	Residual Plot

Source	LogWorth	PValue
Log(Percent_Asian_adj)	26.680	0.00000
Log(EngLearner_Percent_adj)	21.873	0.00000
Student_Enrollment	12.281	0.00000
Log(Percent_Black_adj)	7.769	0.00000
School_Culture_Satisfaction	7.332	0.00000
Systems_for_Improvement_Satisfaction	6.633	0.00000
Poverty_Percent	5.344	0.00000
Indicator_Queens	4.504	0.00003
Indicator_Staten Island	2.065	0.00862
Disabilities_Percent	2.054	0.00884

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.1194874	0.06625	92.37	<.0001*	.
Indicator_Queens	-0.040766	0.009668	-4.22	<.0001*	1.3138748
Indicator_Staten Island	-0.058806	0.022263	-2.64	0.0086*	1.206477
Student_Enrollment	4.0076e-5	5.348e-6	7.49	<.0001*	1.6645134
Disabilities_Percent	-0.001553	0.00059	-2.63	0.0088*	1.9542378
Poverty_Percent	-0.001524	0.000327	-4.66	<.0001*	2.5098501
Systems_for_Improvement_Satisfaction	-0.005927	0.001124	-5.27	<.0001*	4.7956064
School_Culture_Satisfaction	0.0075669	0.001355	5.58	<.0001*	5.8679113
Log(Percent_Black_adj)	-0.022268	0.003859	-5.77	<.0001*	1.8124027
Log(EngLearner_Percent_adj)	-0.037004	0.003532	-10.48	<.0001*	2.5515644
Log(Percent_Asian_adj)	0.0377099	0.003195	11.80	<.0001*	1.9398177



6. We have lessened the effect of Heteroscedasticity but the issue is still presented.