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.

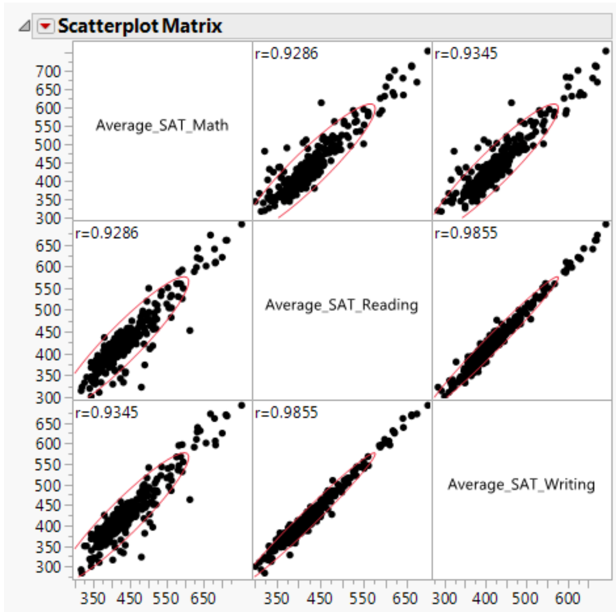
Variable Name	Description	Type
DBN	School's unique identifier	Character
Borough	School's Borough. Comprised of 5 area: Staten Island, Queens, Manhattan, Brooklyn, Bronx	Character
City	City where the school is located	Character
Latitude	School's Latitude	Numeric

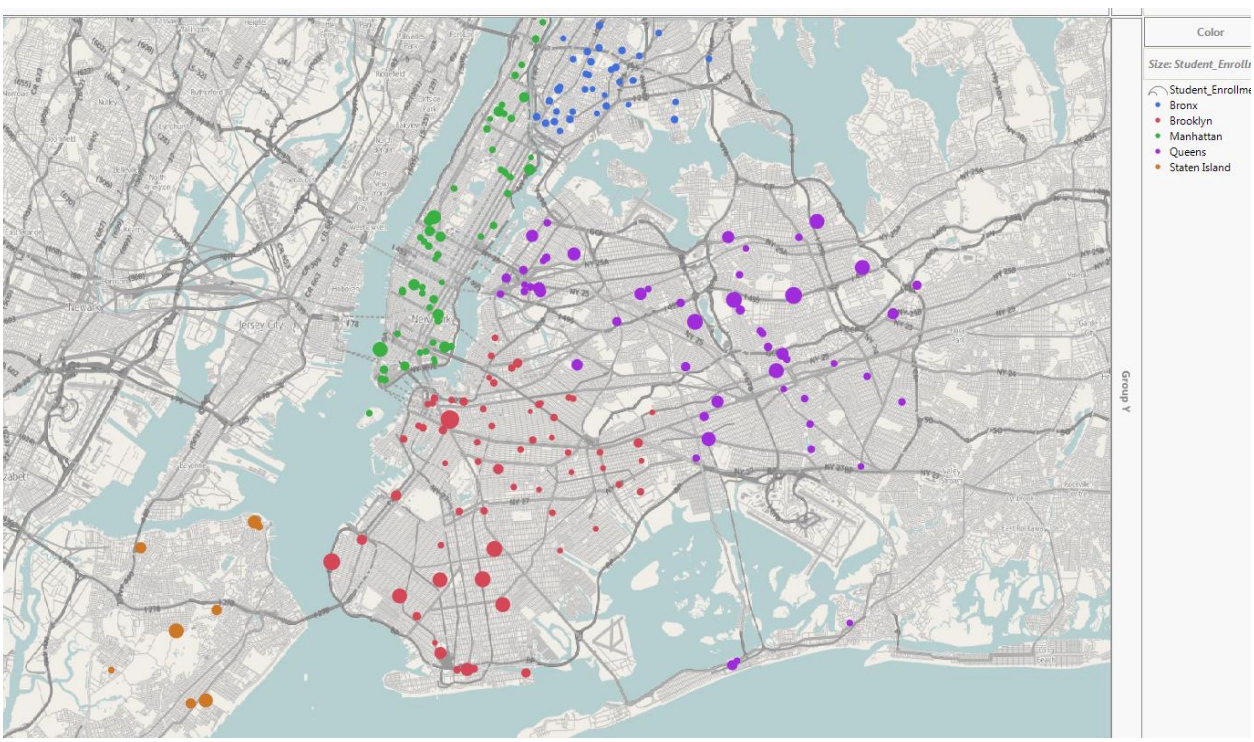
Longitude	School’s Longitude	Numeric
Start_Time	School’s Opening hour	Numeric(e.g.: convert from 8:15 AM to 8.15)
End_Time	School’s Ending hour	Numeric(e.g.: convert from 4:00 PM to 16.00)
Student_Enrollment	Number of school’s enrollment	Numeric
Percent_White	%White students in 2014-2015 cohort	Numeric
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

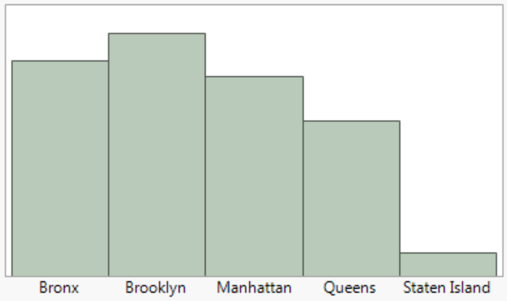
Firstly, we look at the relationship among all SAT scores. They are(unsurprisingly) highly correlated. So, in our analysis, we will put more emphasis on the SAT-Math score and later apply our findings to SAT-Reading and SAT-Writing scores.





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:

Borough

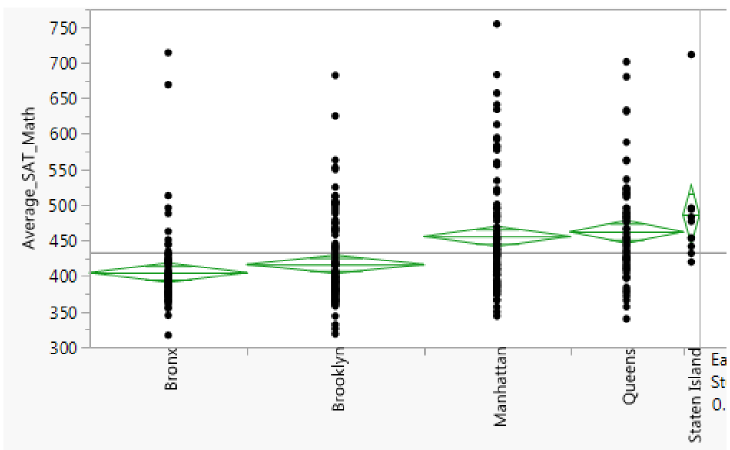


Frequencies

Level	Count	Prob
Bronx	96	0.25806
Brooklyn	108	0.29032
Manhattan	89	0.23925
Queens	69	0.18548
Staten Island	10	0.02688
Total	372	1.00000

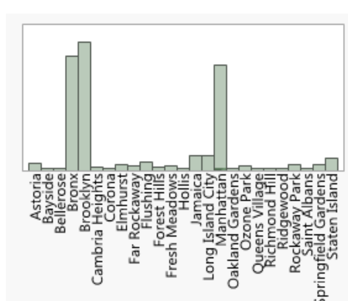
N Missing 0
5 Levels

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.

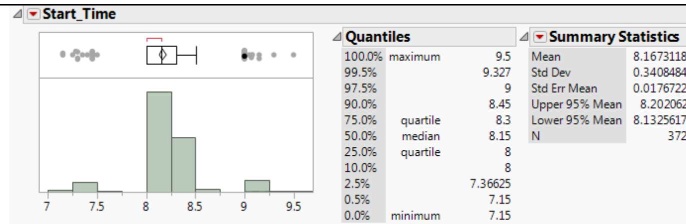
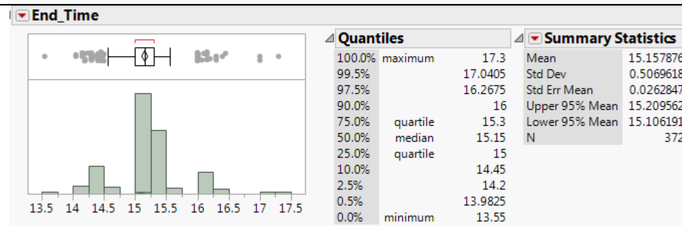
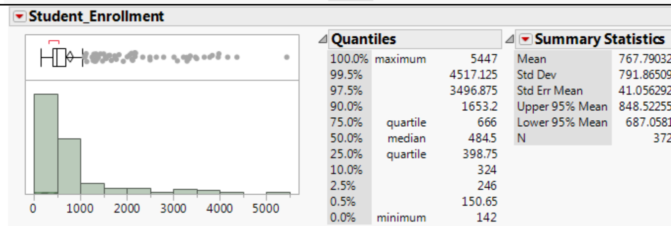
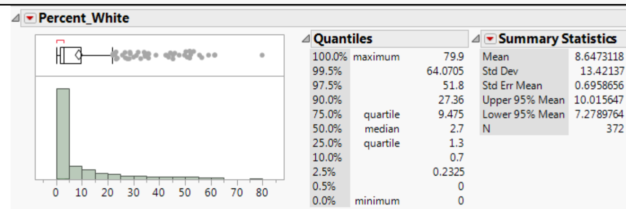
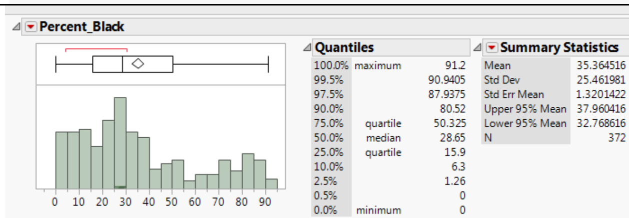
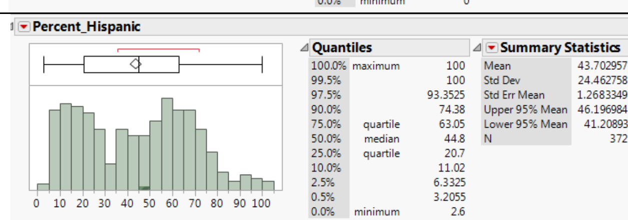
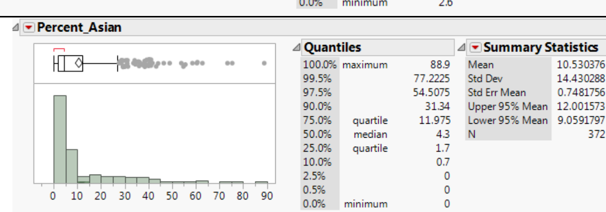


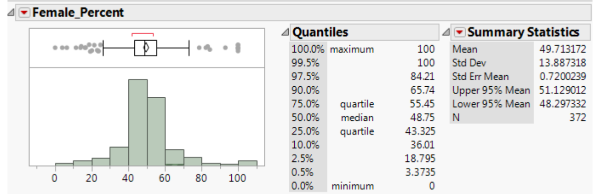
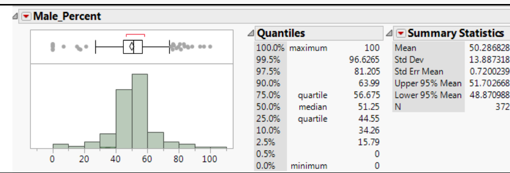
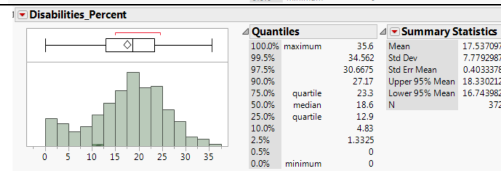
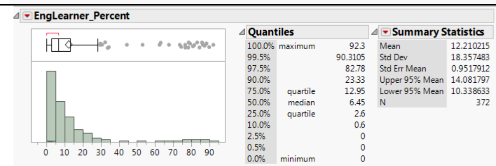
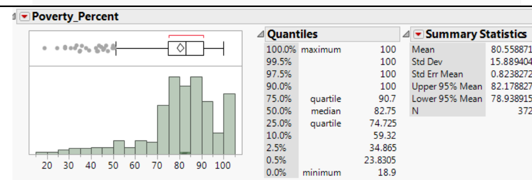
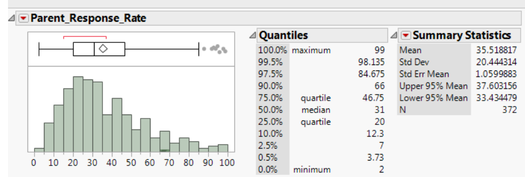
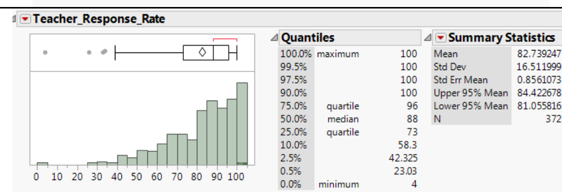
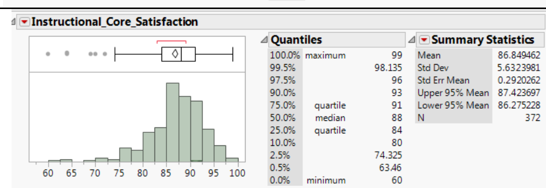
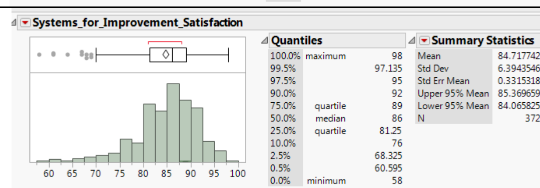
Level			Mean
Staten Island	A		486.20000
Queens	A		462.36232
Manhattan	A		455.88764
Brooklyn		B	416.62963
Bronx		B	404.93750

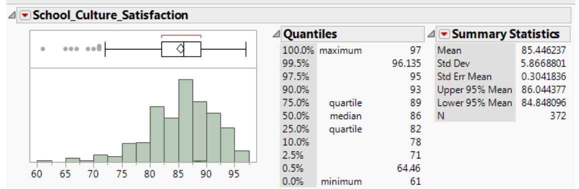
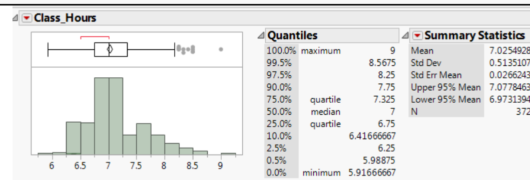
We also investigate the distribution of schools in each City but find that this variable is too fined-grained and decide to drop it as fear of running into overfitting.



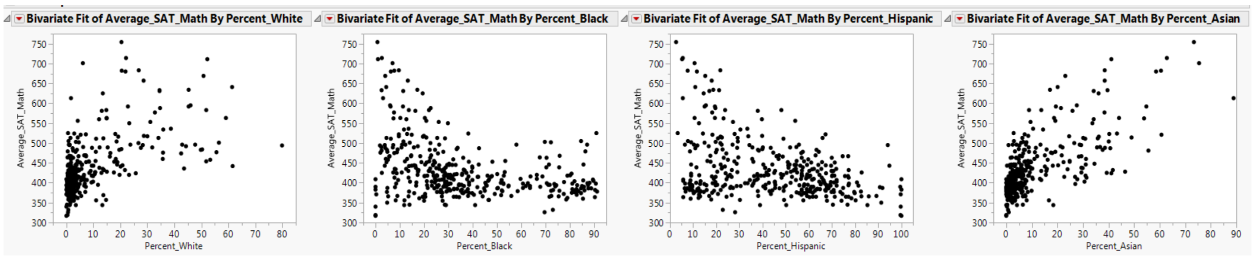
The statistics of other variables are shown in the following table.

<table>
<tr><th>Variable</th><th>Statistics</th></tr>
<tr><td>Start_Time</td><td>
</td></tr>
<tr><td>End_Time</td><td>
</td></tr>
<tr><td>Student_Enrollment</td><td>
</td></tr>
<tr><td>Percent_White</td><td>
</td></tr>
<tr><td>Percent_Black</td><td>
</td></tr>
<tr><td>Percent_Hispanic</td><td>
</td></tr>
<tr><td>Persent_Asian</td><td>
</td></tr>
</table>

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.



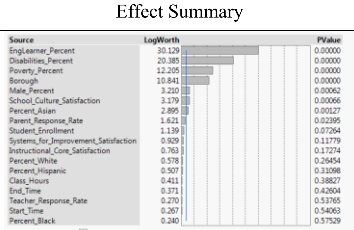
Project Objective

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

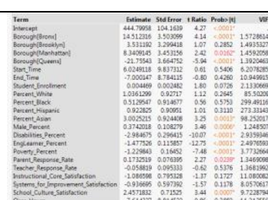
As stated earlier, our model will be developed based on using Average SAT-Math score as a dependent variable. The initial model building consisted of all independent variables (exclude DBN, City, Latitudes, Longitudes, and Female_Percent (as this variable is reflected in Male_Percent)). We obtain the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

Effect Summary



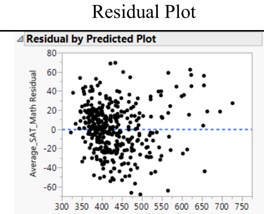
Source	LogWorth	PValue
EngLearner_Percent	30.129	0.00000
Disabilities_Percent	20.385	0.00000
Poverty_Percent	12.205	0.00000
Borough	10.841	0.00000
Male_Percent	3.210	0.00062
School_Culture_Satisfaction	3.179	0.00066
Percent_Asian	2.895	0.00127
Parent_Response_Rate	1.621	0.02395
Student_Enrollment	1.139	0.07264
Systems_for_Improvement_Satisfaction	0.929	0.11779
Instructional_Core_Satisfaction	0.763	0.17274
Percent_White	0.578	0.26454
Percent_Hispanic	0.507	0.31098
Class_Hours	0.411	0.38827
End_Time	0.371	0.42604
Teacher_Response_Rate	0.270	0.53765
Start_Time	0.267	0.54063
Percent_Black	0.240	0.57529

Parameter Estimates



Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	444.79958	104.1639	4.27	<.0001*	.
Borough[Bronx]	14.512316	3.503099	4.14	<.0001*	1.5728614
Borough[Brooklyn]	3.531192	3.299418	1.07	0.2852	1.4935327
Borough[Manhattan]	8.3409145	3.453156	2.42	0.0162*	1.4592058
Borough[Queens]	-21.75543	3.664752	-5.94	<.0001*	1.3920463
Start_Time	6.0249118	9.837312	0.61	0.5406	6.2078285
End_Time	-7.000147	8.784115	-0.80	0.4260	10.949915
Student_Enrollment	0.004469	0.002482	1.80	0.0726	2.1330669
Percent_White	1.0361299	0.92717	1.12	0.2645	85.50209
Percent_Black	0.5129547	0.914677	0.56	0.5753	299.49116
Percent_Hispanic	0.922825	0.90951	1.01	0.3110	273.33143
Percent_Asian	3.0025215	0.924408	3.25	0.0013*	98.252017
Male_Percent	0.3742018	0.108279	3.46	0.0006*	1.248507
Disabilities_Percent	-2.984675	0.296415	-10.07	<.0001*	2.9359346
EngLearner_Percent	-1.477526	0.115857	-12.75	<.0001*	2.4976593
Poverty_Percent	-1.229843	0.16452	-7.48	<.0001*	3.7732664
Parent_Response_Rate	0.1732519	0.076395	2.27	0.0239*	1.3469098
Teacher_Response_Rate	-0.058819	0.095333	-0.62	0.5376	1.3681992
Instructional_Core_Satisfaction	-1.086598	0.795328	-1.37	0.1727	11.080082
Systems_for_Improvement_Satisfaction	-0.936695	0.597392	-1.57	0.1178	8.0570617
School_Culture_Satisfaction	2.4571832	0.71525	3.44	0.0007*	9.7228794
Class_Hours	7.614327	8.814522	0.86	0.3883	11.312556

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.

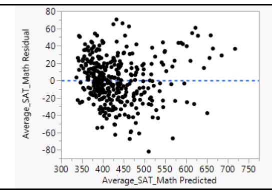
Effect Summary

Source	LogWorth	PValue
Percent_Asian	64.326	0.00000
EngLearner_Percent	40.527	0.00000
Disabilities_Percent	29.801	0.00000
Percent_Hispanic	16.007	0.00000
Indicator_Queens	14.053	0.00000
Poverty_Percent	13.390	0.00000
Percent_White	3.646	0.00023
Male_Percent	2.775	0.00168

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	540.72451	13.24956	40.81	<.0001*
Percent_White	0.5826752	0.156399	3.73	0.0002*
Percent_Hispanic	0.6268536	0.071853	8.72	<.0001*
Percent_Asian	2.7549722	0.130592	21.10	<.0001*
Male_Percent	0.3340682	0.105541	3.17	0.0017*
Disabilities_Percent	-3.4044	0.269741	-12.62	<.0001*
EngLearner_Percent	-1.616161	0.105424	-15.33	<.0001*
Poverty_Percent	-1.244088	0.158044	-7.87	<.0001*
Indicator_Queens	-32.08751	3.964776	-8.09	<.0001*

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)

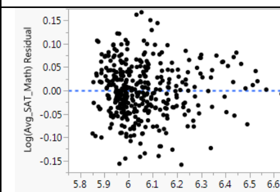
Effect Summary

Source	LogWorth	PValue
Percent_Asian	57.520	0.00000
EngLearner_Percent	42.734	0.00000
Disabilities_Percent	28.660	0.00000
Percent_Hispanic	15.989	0.00000
Poverty_Percent	10.157	0.00000
Indicator_Queens	9.875	0.00000
Percent_White	5.051	0.00001
Male_Percent	2.213	0.00612

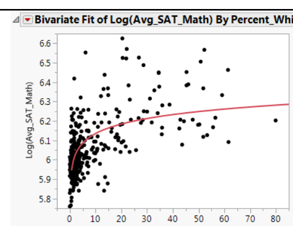
Parameter Estimates

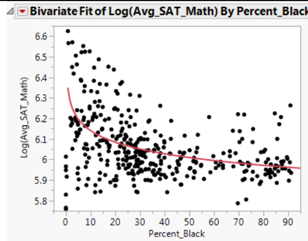
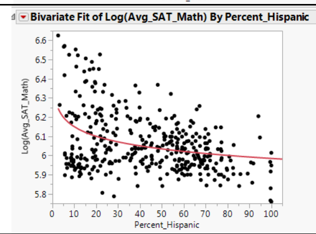
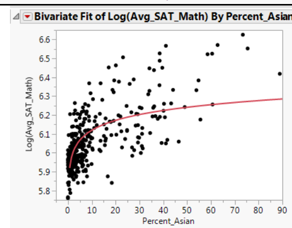
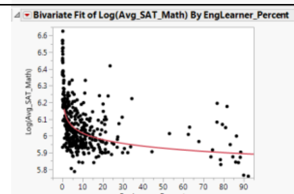
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.264464	0.028792	217.58	<.0001*	.
Percent_White	0.0015317	0.00034	4.51	<.0001*	2.3440289
Percent_Hispanic	0.0013613	0.000156	8.72	<.0001*	1.6436128
Percent_Asian	0.00552	0.000284	19.45	<.0001*	1.8892239
Male_Percent	0.0006324	0.000229	2.76	0.0061*	1.1428218
Disabilities_Percent	-0.007223	0.000586	-12.32	<.0001*	2.3424771
EngLearner_Percent	-0.003636	0.000229	-15.87	<.0001*	1.9925227
Poverty_Percent	-0.002309	0.000343	-6.72	<.0001*	3.3548226
Indicator_Queens	-0.056993	0.008616	-6.62	<.0001*	1.2668118

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

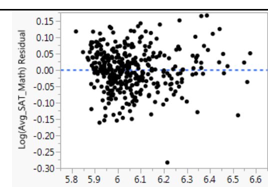
Effect Summary

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

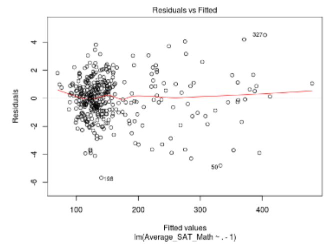
Parameter Estimates

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

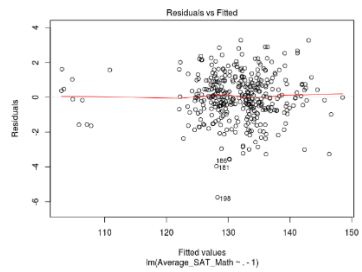
Residual Plot



6. We have lessened the effect of Heteroscedasticity but the issue is still presented. Now, we turn to Weighted Least Squares Regression approach. We switch to using R to conduct the analysis at this point (as performing the analysis in JMP can be quite tedious). The analysis code can be found in analysis.R . We use the transformed data collected from step 4). As the observations come from aggregated result, we firstly try to weight the data by the enrollment size (that is, multiply every variable by sqrt(Student_Enrollment)) and fit the regression model with no intercept. The residual plot indicates that Heteroscedasticity is still presented.

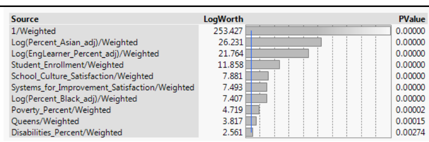


7. Now, we try Weighted Least Squares Regression with two-stage approach; firstly, fit the regression model using transformed variables in step 4) and then use the mean square residual of each borough as a weight for WLS. The residual plot indicate that Heteroscedasticity issue is now fixed.



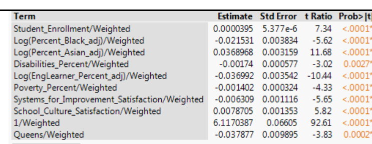
8. We import the transformed data back to JMP(weighted_score.csv) and drop any unnecessary variable as stated in 2). We obtain the following model:

Effect Summary



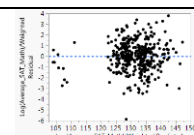
Source	LogWorth	PValue
1/Weighted	253.427	0.00000
Log(Percent_Asian_adj)/Weighted	26.231	0.00000
Log(EngLearner_Percent_adj)/Weighted	21.764	0.00000
Student_Enrollment/Weighted	11.858	0.00000
School_Culture_Satisfaction/Weighted	7.881	0.00000
Systems_for_Improvement_Satisfaction/Weighted	7.493	0.00000
Log(Percent_Black_adj)/Weighted	7.407	0.00000
Poverty_Percent/Weighted	4.719	0.00002
Queens/Weighted	3.817	0.00015
Disabilities_Percent/Weighted	2.561	0.00274

Parameter Estimates



Term	Estimate	Std Error	t Ratio	Prob>|t|
Student_Enrollment/Weighted	0.0000395	5.377e-6	7.34	<.0001*
Log(Percent_Black_adj)/Weighted	-0.021531	0.003834	-5.62	<.0001*
Log(Percent_Asian_adj)/Weighted	0.0368968	0.003159	11.68	<.0001*
Disabilities_Percent/Weighted	-0.00174	0.000577	-3.02	0.0027*
Log(EngLearner_Percent_adj)/Weighted	-0.036992	0.003542	-10.44	<.0001*
Poverty_Percent/Weighted	-0.001402	0.000324	-4.33	<.0001*
Systems_for_Improvement_Satisfaction/Weighted	-0.006309	0.001116	-5.65	<.0001*
School_Culture_Satisfaction/Weighted	0.0078705	0.001353	5.82	<.0001*
1/Weighted	6.1170387	0.06605	92.61	<.0001*
Queens/Weighted	-0.037877	0.009895	-3.83	0.0002*

Residual Plot



We then perform Park-Test to verify for Heteroscedasticity. The P-Value for Y_hat when regressed on r^2 is 0.75 and when regressed on log(r^2) is 0.68. This indicates that the Heteroscedasticity is now fixed.

The model in step 8) is our selected model. The model equation is

$$\frac{Log(AvgSatMath)}{Weight} = 6.1*\frac{1}{Weighted} + 0.000039*\frac{StudentEnrollment}{Weighted} - 0.02*\frac{Log(PercentBlack)}{Weighted} + 0.03*\frac{Log(PercentAsian)}{Weighted} - 0.001*\frac{DisabilityPercent}{Weighted} - 0.03 * \frac{Log(EngLearnerPercent)}{Weighted} - 0.001*\frac{PovertyPercent}{Weighted} - 0.006*\frac{SystemForImprovementSatisfaction}{Weighted} + 0.007*\frac{SchoolCultureSatisfaction}{Weighted} - 0.03*\frac{Queens}{Weighted}$$

Where Weighted is the means of Residual Square per group and has the following values:

Group	Weighted
Bronx	0.0463926118015564

Brooklyn	0.0447919246176774
Manhattan	0.0446127098917998
Queens	0.0484199559569425
Staten Island	0.0584236025734739

The equation can be simplified to

$$Log(AvgSatMath) = 6.1 * \ 0.000039 * StudentEmrollment - 0.02 * Log(PercentBlack) + 0.03 * Log(PercentAsian) - 0.01 * DisabilityPercent - 0.03 * Log(EngLearnerPercent) - 0.001 * PovertyPercent - 0.006 * SystemForImprovementSatisfaction + 0.007 * SchoolCultureSatisfaction - 0.03 * Queens$$