**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**

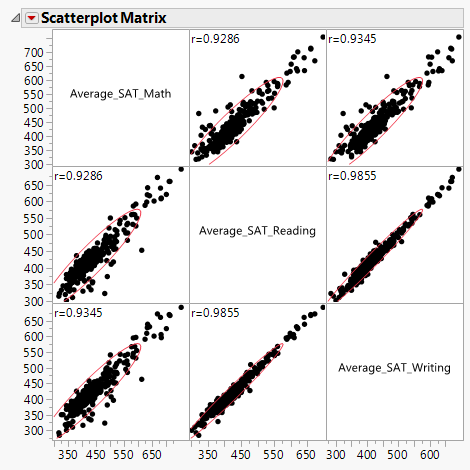
In this analysis, we combine data from the mentioned 3 files. We narrow down the associated features from the original sources to just 26 variables. We use R-script(mungdata.R) to perform the data munging and collect the processed result in processed\_score.csv. Noted that the total number of observation 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.

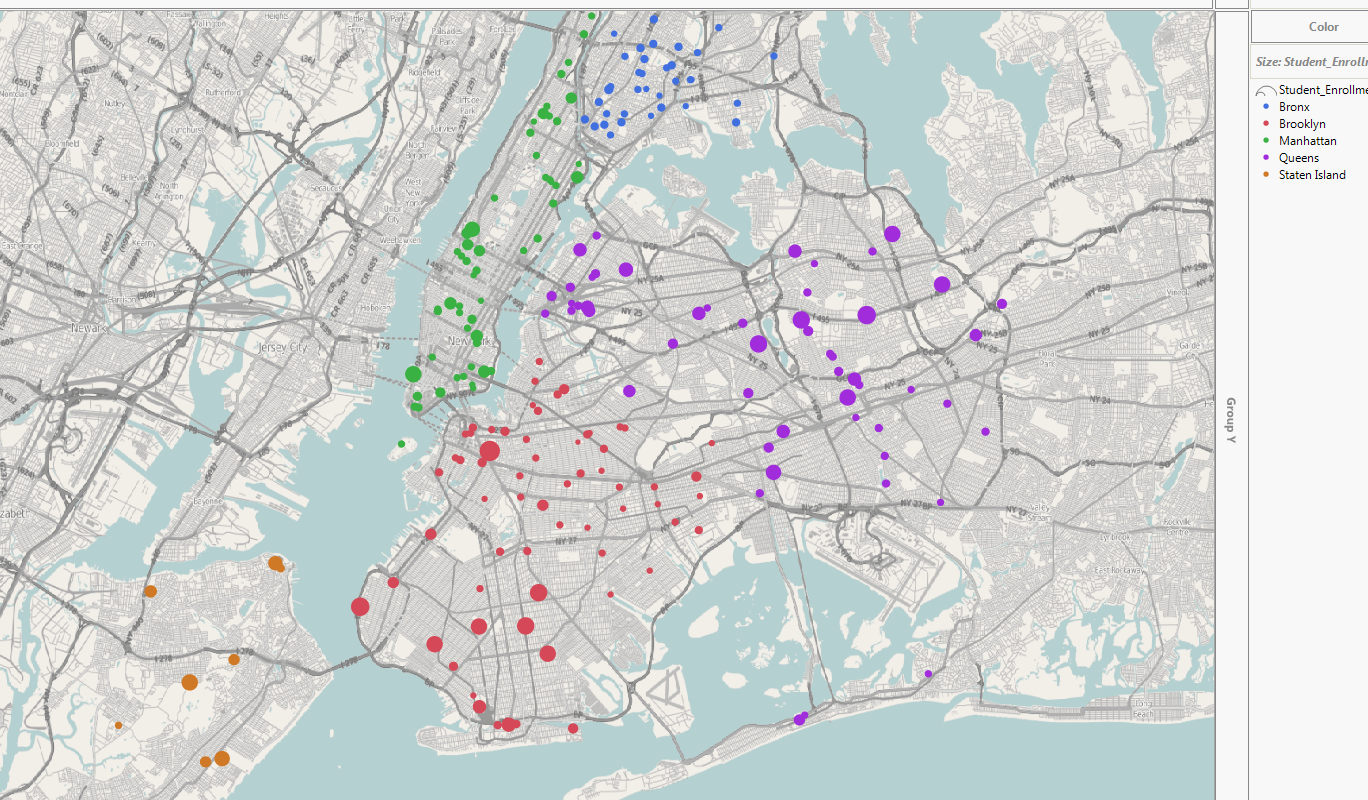
|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Type |
| DBN | School’s unique identifier | Character |
| Borough | School’s Borough. Comprised of 5 area: Staten Island, Queens, Manhattan, Brooklyn, Bronx | Character |
| City | City where the school is located | Character |
| Latitude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

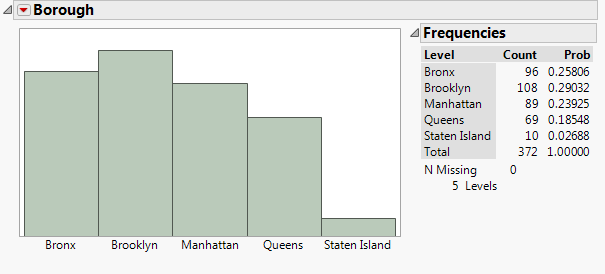
**Exploratory Analysis**

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.

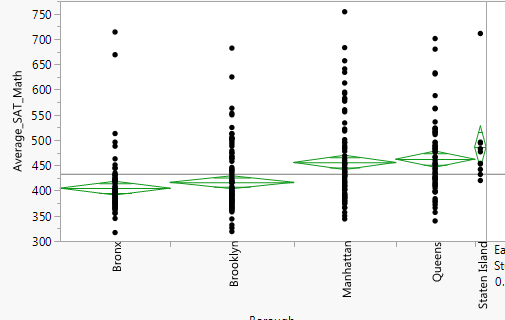


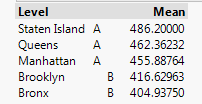


We indicate the location of each school in the above graph. The circle’s size corresponds to the size of enrollment. The school’s distribution in each borough can be summarized as follows:

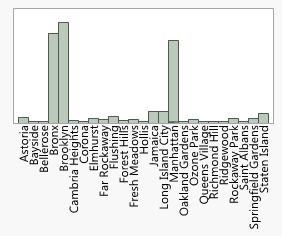


We investigate the effect of spatial information (Borough) on SAT Math score by performing ANOVA. We find that this variable maybe helpful in predicting the SAT score.





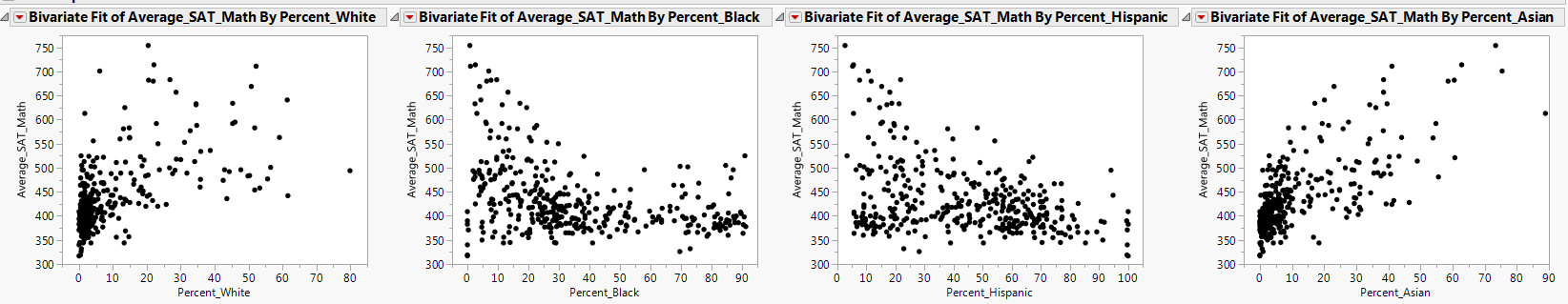
We also investigate the distribution of schools in each City but find that this variable is too fined-grained and decide to drop it as fear of running into overfitting.



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 |  |
| 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 |
|  |  |  |

1. 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 |
|  |  |  |

1. 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.

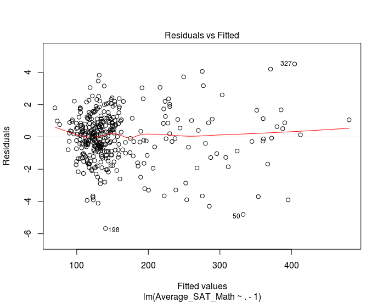
|  |  |  |
| --- | --- | --- |
| Variable | Plot of variable vs Log(Avg\_SAT\_Math) and transformation fit | Transformation Taken |
| \*Percent\_White |  | Log(Percent\_White + 0.1) |
| \*Percent\_Black |  | Log(Percent\_Black + 0.1) |
| \*Percent\_Hispanic |  | Log(Percent\_Hispanic + 0.1) |
| \*Percent\_Asian |  | 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

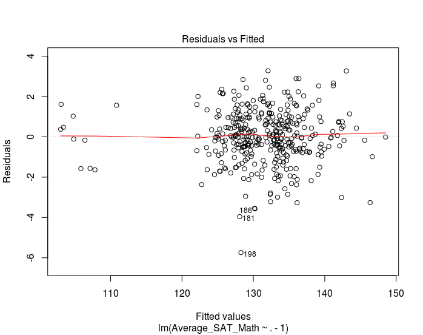
1. Re-fit the model using the transformed variables. Drop any unnecessary variable as stated in 2)

|  |  |  |
| --- | --- | --- |
| Effect Summary | Parameter Estimates | Residual Plot |
|  |  |  |

1. 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.



1. 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.



1. 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 | Parameter Estimates | 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

|  |
| --- |
|  |

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

|  |
| --- |
|  |