Regression on SAT Scores of 374 High Schools and K-means on Clustering Schools

Abstract

In this project, we study 374 public high schools in New York City. The project seeks to use regression techniques generating an optimal model to predict average SAT score given a school, and to apply clustering algorithms to group public high schools in New York City.

Firstly, we visualize the data. Then, checked the assumptions of linear regression model, we fit random forest and apply variable-selection methods to linear models. With a given model, we use 5-folds cross validation to generate its test error. Lastly, after dimension reduction by PCA, we apply K-means to cluster schools and find the most important covariates.

The study suggests the optimal model is the 2-degree polynomial regression with Lasso, and the schools can be clustered into three clusters with emphasis on location and race. This project helps students in New York City to select schools, as well as schools to improve education.

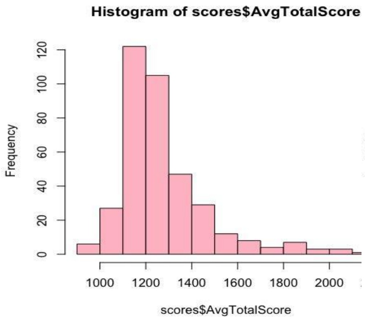
Data

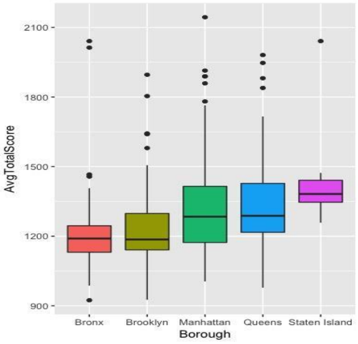
This dataset (Kaggle, 2016) consists of a row for every high school in New York City with its department ID number, school name, borough, building code, street address, latitude/longitude coordinates, phone number, start and end times, student enrollment, race breakdown, and average scores on each SAT test section for the 2014-2015 school year.

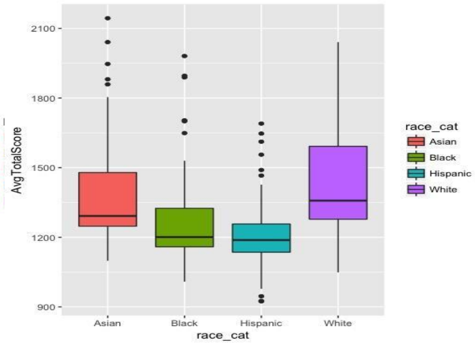
After we select our dataset, we preprocess the data set in such five ways and we have 25 variables in total. Among these data, there are 10 categorical variables and 15 quantitative variables.

- Delete missing values (374 in total after deleting missing values)
- Add a variable AvgTotalScore (sum of scores on each section)
- Add a variable(race_cat) notifying the dominant race of each school
- Add a variable(School_time) notifying end-time minus start-end
- Change percentiles to numeric variable data type

1. Visualizations







We obtain the frequency counts for AvgTotalScore and construct a histogram. The median is around 1200. The histogram is skewed to the right, implying there are outliers in the high score range. The third quantile is further from the median than the first quantile is.

We use median and inter-quantile range to describe the center and spread of the observations in 5 different boroughs. Outliers are in the high score range, which is consistent with the conclusion from the histogram. The scores in Staten Island are apparently higher than other boroughs since it has a higher center and smaller spread. Manhattan and Queens have similar centers, while Bronx and Brooklyn have similar centers. Five spreads are different. It is plausible that scores in different boroughs can differ from each other significantly. Thus, Borough could be a potential predictor for AvgTotalScore.

We plot the side-by-side boxplots to compare set of observations of four races, Asian, Black, Hispanic, and White. White does not have outliers. Though the centers of four races are all close to 1200, the spreads are obviously different from each other. It is plausible that scores in different races can differ from each other significantly. Thus, race could be a potential predictor for AvgTotalScore.

2. Regression Analysis

The goal of this section is to predict the average SAT scores given eight predictors - Borough, Student.Enrollment, Percent.White, Percent.Black, Percent.Hispanic, Percent.Asian, Percent.Tested and School_time. Since we have shown Borough and Race has influence on AvgTotalScore in the Visualization part, we include Borough and percentages of each race into the model. Other quantitative predictors are also included.

2.1 Assumption checking

Before we do linear regression analysis, we check five assumptions of the model- linearity, normally distributed errors, constant variance, no outliers and leverage points and small collinearity (See Appendix A Linear Assumptions for graphs and interpretation). The assumptions are not met well, so we also apply other algorithms to the data.

2.2 Applying 19 models

19 TECHNIQUES	TEST MSE
OLS	13525.735
OLS with AIC backward	13713.629
OLS with AIC forward	13525.735
OLS with AIC stepwise	13713.629
OLS with BIC backward	14366.751
OLS with BIC forward	13525.735
OLS with BIC stepwise	14366.751
OLS with Ridge regression	13100.068
OLS with Lasso regression	12472.333
Polynomial	10245.024
Polynomial with AIC backward	10110.36
Polynomial with AIC forward	10245.024
Polynomial with AIC stepwise	10107.665
Polynomial with BIC backward	10575.604
Polynomial with BIC forward	10245.024
Polynomial with BIC stepwise	10821.946
Polynomial with Ridge regression	8900.983
Polynomial with Lasso regression	7642.741
Random Forest	10121.406

We will compare results from 19 techniques (See Appendix B Models for specific technique names):

We use five-folds cross validation to estimate the test errors. We present the table of test errors of each model below. According to the output table, the 2-degree polynomial model with Lasso gets a significantly lower test MSE, which is 7642.741. Consider that Random Forest usually has much better prediction power, its test MSE is much higher than 7642.741. It implies that the true model should be quite similar to the polynomial model with degree 2 given by Lasso (See Appendix B Optimal for the equation).

2.3 Best fitted model

Thus, we choose 2-degree polynomial model with Lasso as the optimal model. If we were to predict a value with the optimal model, the deviation from the true SAT score is estimated to be 87.

3. Clustering

We employ the K-means technique to cluster schools based on Latitude, Longitude, Student.Enrollment, Percent.White, Percent.Black, Percent.Hispanic, Percent.Asian, Average.Score.SAT.Math, Average.Score.SAT.Reading, Average.Score..SAT.Writing, Percent.Tested, School_time. Since some variables among these 12 variables are correlated, as the matrix plot shows, we should reduce dimension before we apply clustering methods.

3.1 Dimension Reduction

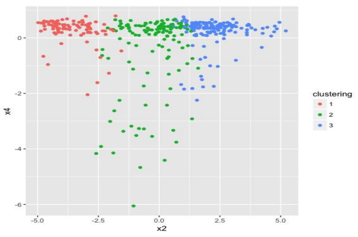
Components	PC1	PC2	PC3	PC4	PC5	PC6
Cumulative Proportion	0.4074	0.5735	0.6830	0.77042	0.84973	0.90120

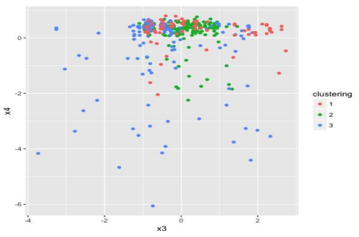
After scaling the variables, we apply PCA on these scaled variables. The cumulative Proportion is around 0.9 which is a regular threshold, when we have 6 components. Then, we use the first six components when applying K-means.

3.2 Important Components

We use K-means for K =1 to 10. Then we compare the WSS of each number of clusters. From plot (See Appendix C WSS plot) that the difference is not significantly large after a certain cut-off point. The cut-off point is K=3 or 4 or 5, but we can not tell exactly which one is the cut-off point by simply looking at the plot. So we apply K-means on K = 3,4,5 respectively. This result is actually consistent with our prior information since we have four races and five boroughs.

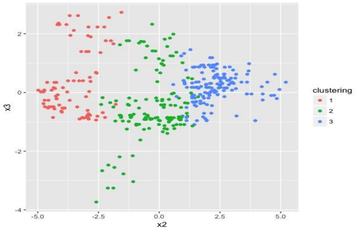
Fixing the number of cluster, we compare the cluster plots generated by all possible pairwise predictors. We find out plots with component2 generally have a neat boundary between clusters, while this is not the case for other components. For example, the two plots below are from 3 clusters. The first is component2 vs component4, while the second is component3 vs component4 (See Appendix C Important Components for clearer graphs).

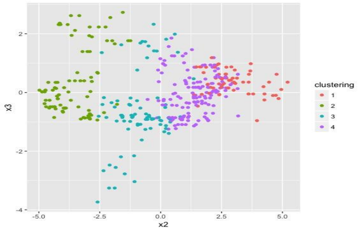


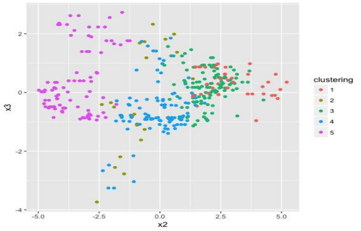


3.3 Compare different K

Then we fix component2 and compare the performances of different number of clusters. We find out K = 3 gives the best clustering. Here, we are just going to show component2 vs component3 for all three Ks, but all other pairs including component2 show the similar pattern (See Appendix C Compare K for clearer graphs).







We can see that the boundary of three clusters are clearer than other two methods. When K = 3, we can see the boundaries are about linear. Inside each cluster, the observations are compact. Different clusters spread out from one another. For K = 4 and 5, clusters are nested together. The sizes of clusters are different.

Thus, we conclude there might be three clusters for schools in New York City. The component2 is the most important factor in clustering.

3.4 Overall interpretation

	Latitude	Longitude	Student.Enrollment	Percent.White	Percent.Black	Percent.Hispanic	Percent.Asian	Average.Score..SAT.Math	Average..Score..SAT.Reading	Average.Score..SAT.Writing	Percent.Tested	School_time
PC2	0.5337	0.0399	0.0392	-0.0021	-0.5863	0.5726	0.0697	0.0878	0.0247	0.0405	0.0429	0.1537

Looking at the component 2 from PCA, we can see Latitude, Percent.Black and Percent.Hispanic are important variables because the absolute values of coefficients are much larger than other variables. Since we have three clusters, from the plot, we can see the observations in the blue cluster has larger latitude, larger Hispanic percentage and smaller Black percentage. In the green cluster, three variables are all close to 0. In the red cluster, we have a small latitude, large Black percentage, and small Hispanic percentage. Thus, locations and race are important when clustering.

4. Limitations

We only apply polynomial of degree = 2 in regression. Models of higher degree could perform better. However, with more than ten variables, higher degreesAppe are very likely to overfit. Also, the true model of clustering is unknown, so we are not sure whether the size, density and the shape of the clusters do not satisfy the conditions to apply K-means.

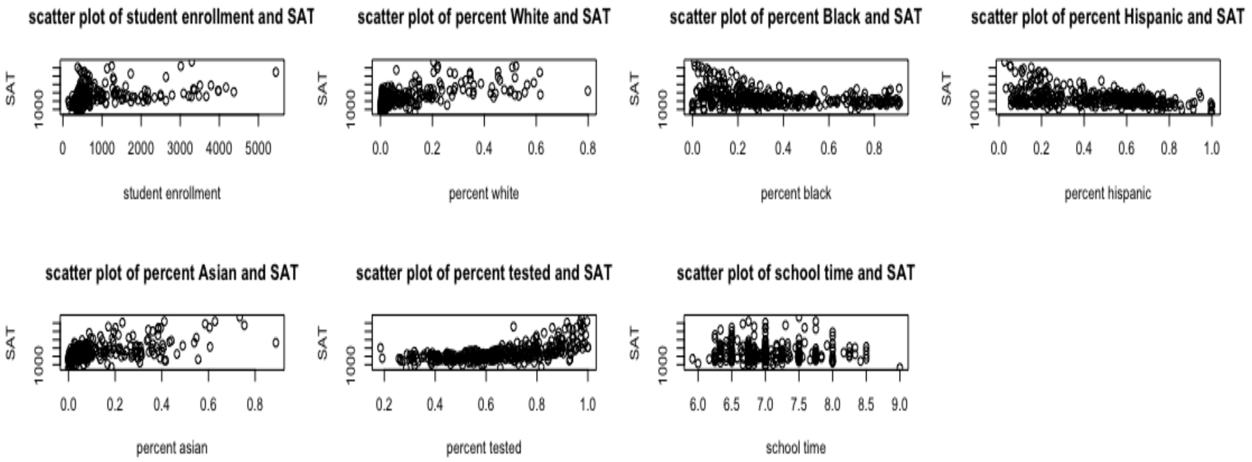
Reference

Average SAT Scores for NYC Public Schools | Kaggle. N.p., n.d. Web. Mar. 2016. <https://www.kaggle.com/nycopendata/high-schools>.

Appendix A : Linear Regression Model Assumptions

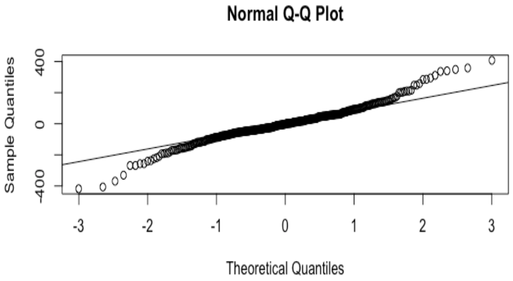
• Linearity Assumptions

According to scatterplots below, the model seems not linear. Therefore, it may be helpful to run a polynomial model later.



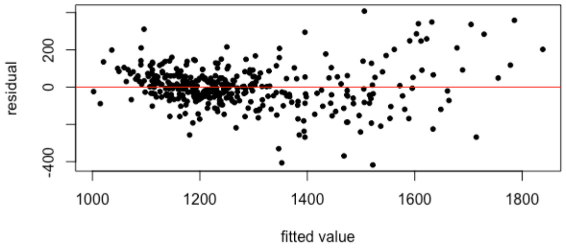
• Normally distributed errors

According to the QQ plot below, the errors are roughly normally distributed.

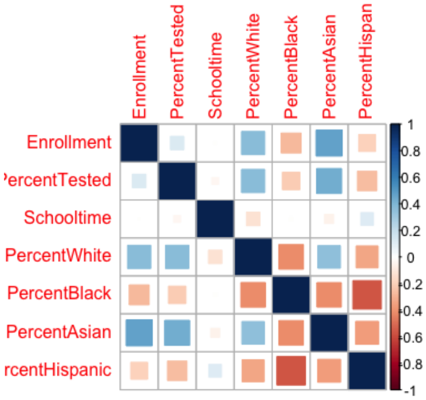


• Constant Variance

According to the residual plot, there seems to be no apparent pattern. Thus, the errors have constant variance.



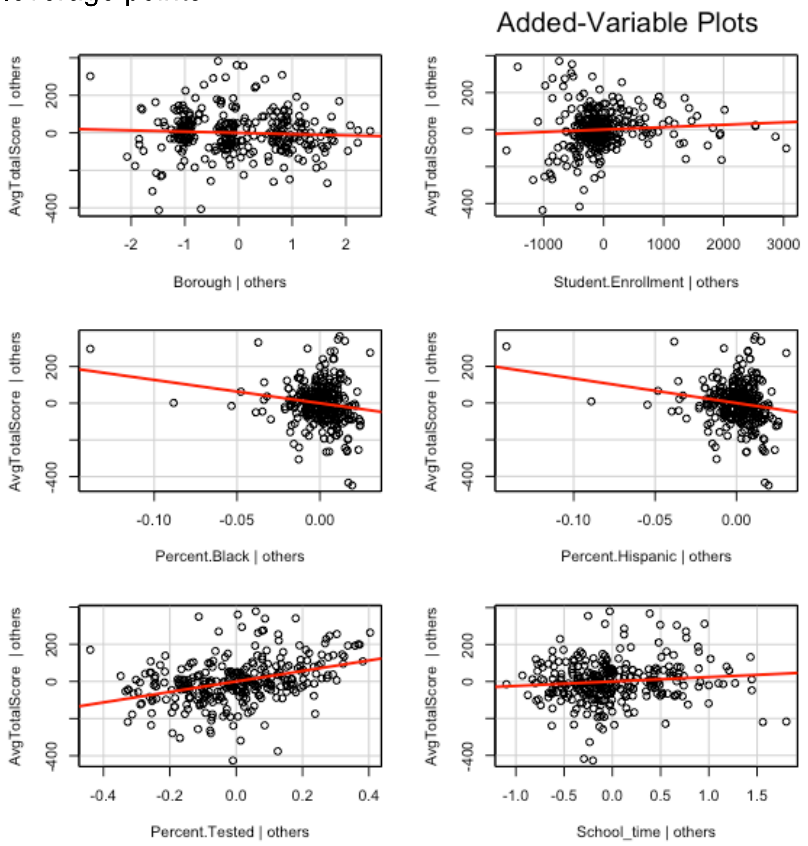
• Small collinearity



According to the correlation matrix, some variables are correlated. Therefore, variable selection, such as AIC and BIC should be performed later.

• No outliers and leverage points

According to the Added-Variable plots attached in Appendix, the model has several outliers and leverage points.



Appendix B : Linear Regression Model Assumptions

• Models

linear model(OLS): OLS / AIC with backward / AIC with forward / AIC with stepwise / BIC with backward / BIC with forward / BIC with stepwise / Ridge regression / Lasso regression

Polynomial: Polynomial / AIC with backward / AIC with forward / AIC with stepwise / BIC with backward / BIC with forward / BIC with stepwise / Ridge regression / Lasso regression

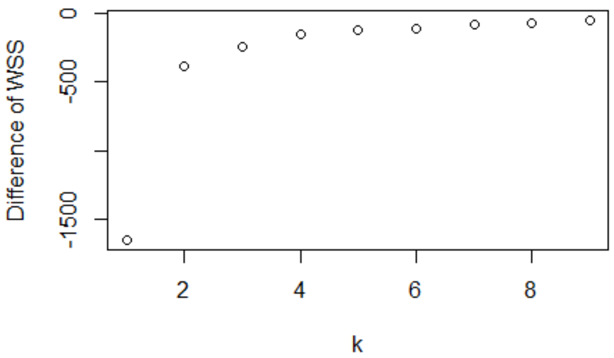
Random Forest

• Optimal

AvgTotalScore = 1236.13 + 18.73 * School_time - 0.000019 * Student_Enrollment^2 - 115.93 *Percent.Black^2 - 192.62 * Percent.Hispanic^2 - 579.56 * Percent.Asian^2 + 663.92 *Percent.Tested^2 -48.96 *Borough * Percent.White - 7.16 * Borough * Percent.Black + 14.06 * Borough* Percent.Hispanic + 76.20 * Borough * Percent.Asian - 0.097*Student.Enrollment * Percent.White - 0.104 * Student.Enrollment * Percent.Hispanic + 0.0028 * Student.Enrollment * Percent.Asian + 0.12 * Student.Enrollment * Percent.Tested + 0.00157 * Student.Enrollment * School_time - 66.75 * Percent.White * Percent.Black -210.42 * Percent.White * Percent.Hispanic + 312.80 * Percent.White * Percent.Asian + 10.17 * Percent.White * School_time + 0.89 * Percent.Black * Percent.Hispanic - 708.25 * Percent.Black * Percent.Asian - 614.55 * Percent.Black * Percent.Tested - 1514.41 * Percent.Hispanic * Percent.Asian - 737.37 * Percent.Hispanic * Percent.Tested

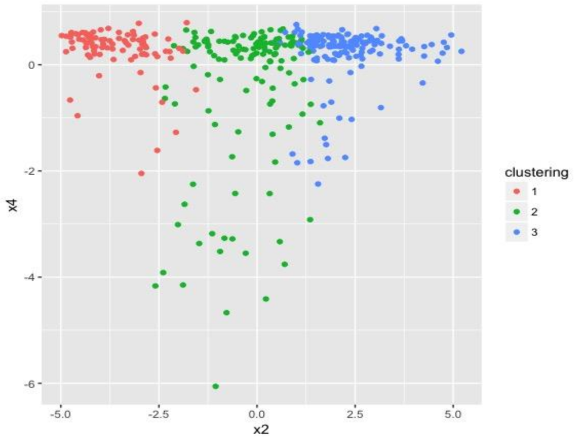
Appendix C : Clustering

• WSS plot

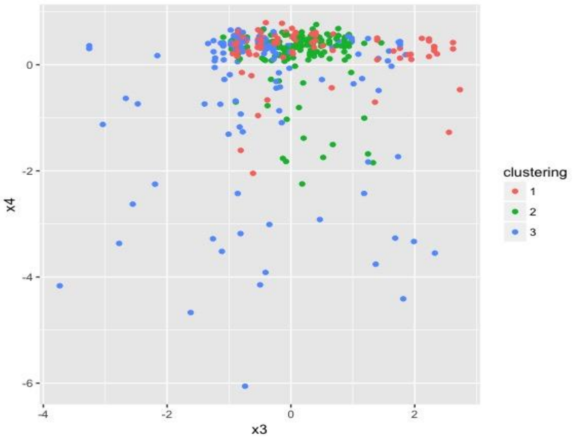


WSS is within cluster sum of squares (SSE). WSS = $\sum_k\sum_{x\in Ck} ||x - m_k||^2$ The larger k is, the smaller WSS is. For each k, we run K-means for 10 times to pick the smallest WSS. Difference of WSS is obtained by subtracting WSS of smaller k from WSS of larger k, e.g. WSS2 – WSS1. The larger the difference is, the more significant the increase of k is. Here the difference levels off from k=3 or 4 or 5.

• Important Component

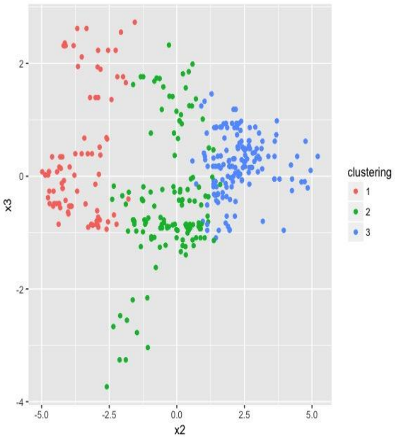


component2 vs component4

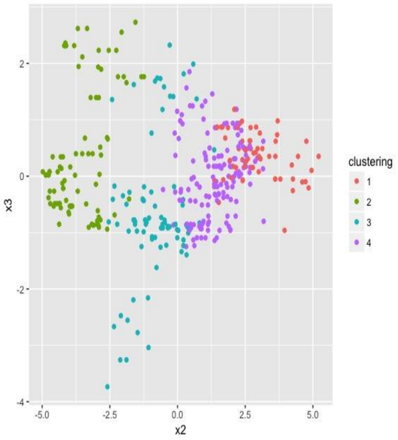


component3 vs component4

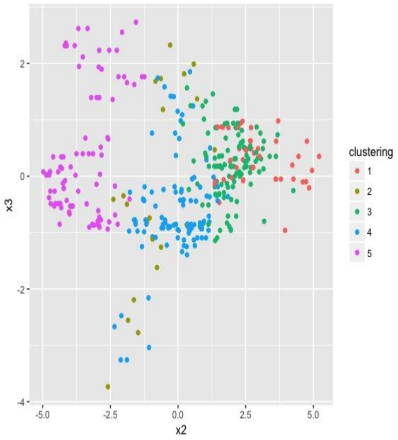
- **Compare K**



K = 3



K = 4



K = 5