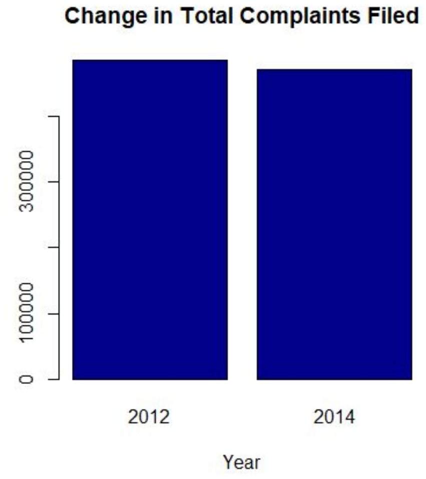
Caleb Ralphs, Vital Tavares and David Larson
MA 463X Final Report

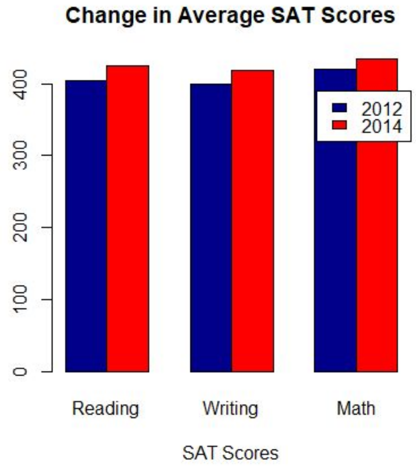
Introduction

In this project we tried to evaluate possible relationships between standardized test scores results, more precisely the SATs, and the crime rates around a determined school radius. This is a problem that not only involve data analysis but the social implications associated with education access and criminality. If relationship is proven, a diagnosis will be provided in order to minimize the public security issues surrounding the pre-established regions. Moreover, it provides precedent for students to boost their scholarly work due to projections linked to their respective futures.

The Data

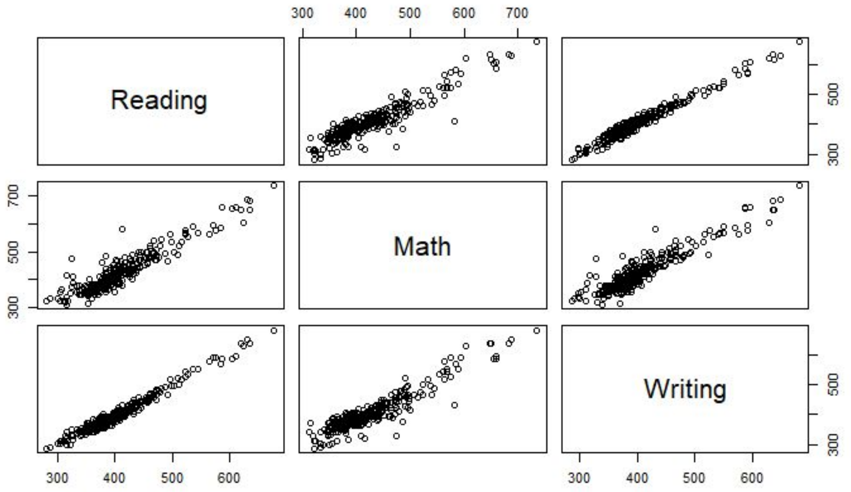
For this study we chose to look at data containing SAT scores for about 400 schools in New York City. We have two separate datasets, one containing those scores from 2012 and the other containing those from 2014. We then found a dataset containing all complaints filed to the New York Police Department (NYPD) between 2006 and 2015. This dataset also contains predictors including type of crime, date, time, location, and latitude and longitude. The csv file for the NYPD complaints is over a gigabyte in size and has several million entries. Before we began any data cleaning or feature engineering, we devised a few simple plots to better understand if there was any potential correlation to our data that was worth looking into.

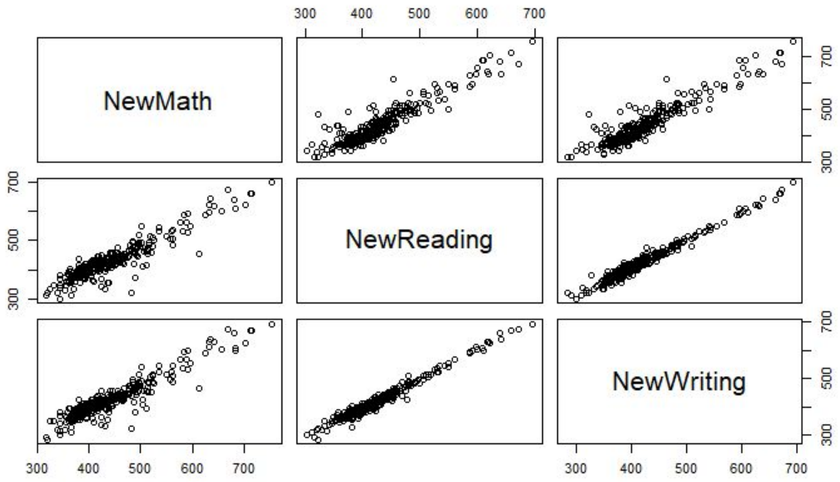

Change in Total Complaints Filed


Change in Average SAT Scores

We created two very simple graphs showing the general trend in the data. The chart on the left shows the total number of complaints filed to the NYPD in the years 2012 and 2014. As we see there is about a 2.7% decrease in crime between 2012 and 2014, and about a 2.5% increase in average SAT scores between the same years. These numbers seem to point to a small yet significant correlation between the two sets of data.

The 2012 scores contains the average reading, writing, and math scores, the number of test takers, the Borough, and the District Borough Number (DBN) for each school in New York City. The dataset from 2014 contains a few additional predictors which we did not use, such as demographic information and percentage of students that took the SAT. More importantly, the 2014 dataset contained the DBN of each school as well as the latitude and longitude. This allowed us to essentially use the 2014 dataset as a mapping between DBN and latitude and longitude, which we will discuss further in depth in the Feature Engineering section. As we began to look at using the data to predict crime rates, we examined the correlation between scores in each of the sections of the SAT. The following graphs reflect our findings.



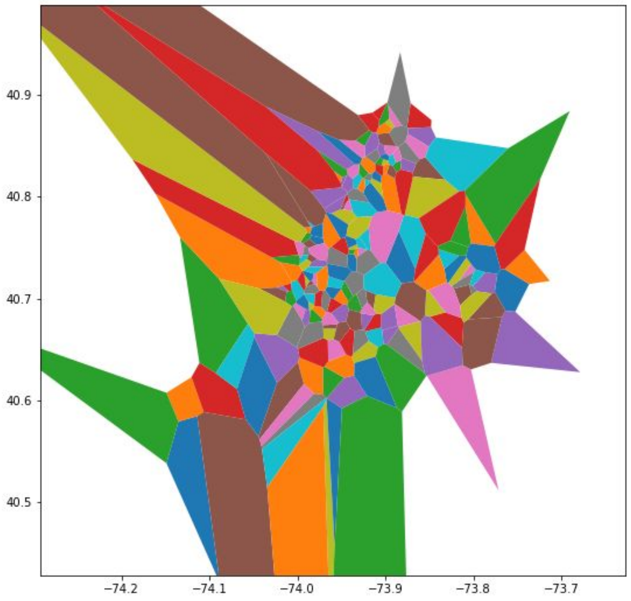


The first graph set of graphs show the correlation between sections for 2012 test results, and the second shows those from 2014. As one can clearly see, the results are very heavily correlated. As such, we decided that we could use the total score as a predictor without losing any significant data. A test result who scored very high in reading was very likely to have scored similarly high in writing and math, so we could essentially extrapolate the score of one section to a total score and maintain reasonable accuracy.

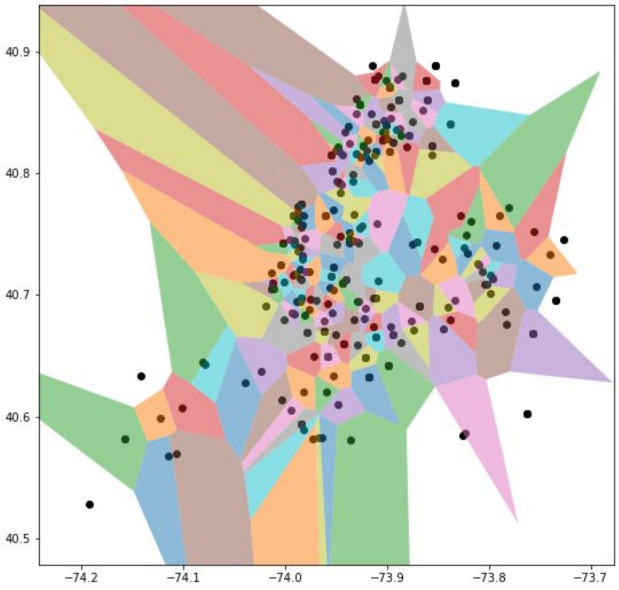
Feature Engineering

There were several things that we needed to change with our datasets in order to make them more easy to work with. Firstly, we needed to clean all of our datasets and remove entries with null values. We also removed predictor columns which we deemed irrelevant to our study in order to make the data file smaller and easier to work with. Once we had two cleaned datasets of SAT scores, we wanted to combine them. We decided to extrapolate the longitudes and latitudes from the 2014 dataset based on the DBN and add them to the 2012 dataset. Each school in both datasets had a unique DBN, so we could make this change with just a few lines of code. At this point, we had two datasets with the same predictors but a slightly different number of schools. We then took the intersect of the two datasets based on the DBNs, which yielded two datasets that had the exact same list of schools and the exact same predictors. From there we simply renamed the columns in the file to have a truly uniform group of datasets.

We now turn our attention to the dataset of NYPD complaints. One problem that we faced was that we had a list of SAT scores that were organized by school and a list of complaints that were organized based on latitudinal and longitudinal coordinates. In order to solve this, we used a Voronoi diagram to create our own “districts” for each school in which we could place crime complaints from the complaint dataset. At first, we were planning on using a dictionary to map those polygon regions from the Voronoi diagram to the DBNs and the DBNs to their respective regions and crime count, but there were complications with determining the “value” of a polygon object. Below is the Voronoi diagram which we created for the vertices corresponding the the coordinates of the different schools in NYC.



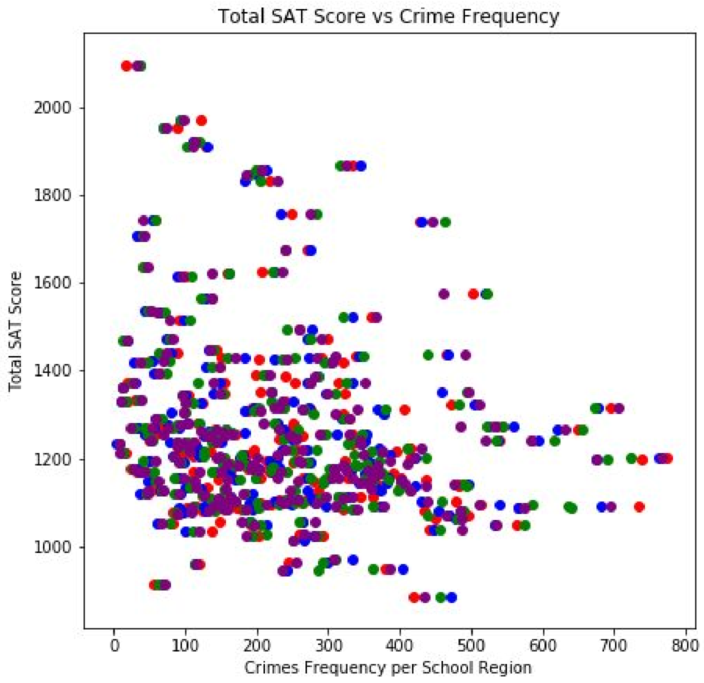
In a Voronoi diagram, each colored region consists of points that are closer to a predefined list of points, called “seeds,” than they are to any other point. Our seeds in this example were our list of latitudes and longitudes for each high school in New York City. Therefore, each colored region in the diagram represent all points that are closer to their respective high schools than they are to any other high school. The diagram was generated by passing all of locations of the schools, longitude and latitude, to the Voronoi function, acting as vertices on polygons overlayed on NYC. By use of an algorithm we developed which calculates the closest school to a complaint, we are now able to place each complaint into one of our defined regions, associating the complaint with the school based upon proximity. We used this algorithm to replicate the functions provided with locating coordinates inside given regions for the different schools. When we overlay the school locations on the regions created by the Voronoi diagram, we can see that it is an effective because it visualizes what our developed algorithm does.



Using the algorithm that we developed we were able to identify which crimes corresponded to which schools, generating a crime frequency relative to that school. We will base our regression models off of these frequencies and will attempt to predict the crime frequency of a district based off of the SAT scores from the given district. Overall, the feature engineering was a the bulk amount of work, coding wise, for the project and you can see exactly what was done in the annotations of the Jupyter Notebook attached to the project submission.

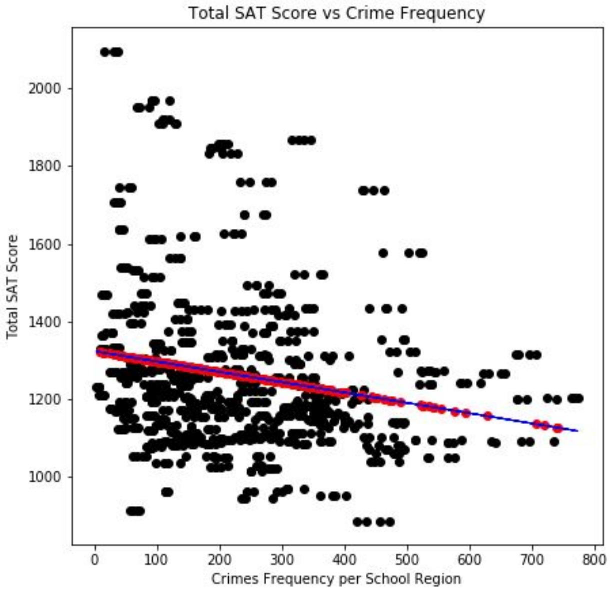
Methodology

Our first challenge was to clean the data sets provided. The SAT scores for both 2012 and 2014 did not prove to be a challenge as we eliminated NULL entries and renamed some columns. The crime complaints set was the most troublesome as it contained 5.6 million data entries that ranged from 2006 and 2015. Luckily we had access to both the latitude and longitude for each school and complaint made. This allowed to associate the proximity of the crime to a determined school. The rest of the feature engineering we did to get clean data is described in the previous section and in the annotations attached to this report. Of the data we sampled from raw data, which was just under 1,000,000 entries, we used around 20-25% of the data for training and around 5% for testing. Our training data is plotted below.



The four different colors (red, blue, green, purple) correspond to the four random uniformly distributed samples which we took from the crime data. As you can see, all of the samples are pretty much the same due to the nature of a uniform distribution. We chose to take multiple samples from the same dataset in order to minimize computational time because the algorithm that we developed for region classification of the crime location had time complexity growth of O(n^2) which posed a problem when we tried taking samples of size 100,000 and larger.

For our regression model we compared the average SAT score result from 2012 and the four training models obtained from random uniform distributions of our data set. Due to lack of computing power and time complexity conditions explained above, we were only able to work with a small portion of the set and hope that our models gave some correlation between them. Of course, we are aware of the high bias present in these relationships. Our linear fit figure is shown below.



Results

After wrestling with our dataset for longer than we would like to admit, we have finally come to a conclusion: we cannot make any conclusions! Unfortunately, our regression models did not yield errors that were small enough for us to make accurate predictions of crime rates based on SAT scores. Our R-squared errors fell in the range of 0.037 - 0.041. These values are very close to 0, suggesting that there is extremely limited correlation between our predictors and the values we were trying to estimate.

Although our relationship proved to be weak between the predictors, resulting in a weak inverse relationship between crime and SAT scores, the testing error seemed relatively low in comparison to the rest of our results. This does not come to a surprise because the testing sample from the crime data was extremely similar to the training data because all of them were taken as random uniformly distributed samples from the crime data; therefore, our mean squared error being just about 53 did not come to a surprise. The MSE being 53 is not statistically large because it is on the scale of SAT scores ranging from around 800 to around 2300.

After completing linear regression, we looked to begin using the lasso shrinkage method. However, after having spent tens of hours working with and editing our datasets, we realized that lasso would not be an appropriate method to use. We only had 3 predictors that we could attempt to relate to crime rates, all of which could be easily combined into one single predictor of total SAT score. This being the case, it didn't make much sense to use lasso, a method that focused on eliminating predictors with low correlation to the values we were interested in.

Although many would consider this attempt at regression a failure, we have taken away many successes from the experience. Our group took an outside the box approach to solving a problem that was quite difficult to begin with. We began with data that was not ready for regression "out-of-the-box" as a lot of datasets could be. We spent tens of hours on correcting our data such that we could relate data categorized based on a school to other data based on latitude and longitude. Although statistically the relationships were insignificant, in the domain of crime frequency and SAT scores, a weak inverse relationship between the crime frequency and sat scores may still prove somewhat meaningful for the social implications associated with this problem.

Comments

We did not obtain ideal R-squared errors as they tended to be close to between 0.037 and 0.041. As we moved on with our research, it came across us the act that human based actions tend to have low R-squared values. This is understood due to the complexity of the of the social relationships present in a metropolis such as New York City. Observing the correlations between 2012 and 2014, we came to the conclusion that the results for the last would be as unsatisfactory and decided not to move on with the regression analysis.