Bridging the Gap between High School & Colleges

Data-Driven Policy Suggestions for High Schools in New York City

Data Science in the Wild: Final Project
Jo-Anne Loh (jl3839), Anirban Poddar (ap2296), Noel Konagai (nk639)

Introduction

NYC Gov has a rich collection of data across the 410 public high schools registered in New York City. We wanted to study and analyze this data to understand what are the key factors that contribute to students' success and *specifically,* play an important role in helping them enroll and prepare for college.

A few key highlights of the problem that we looked to analyze:

- College is widely seen as a tool that promotes social mobility
- However, only 33% of NYC's lowest-income public high school graduates obtained a college degree
- New York's Fair Funding Act for public high schools hopes to help students in need, but does it work?
- What are the key contributing factors to SAT Scores & Graduation Rates?
- Having a juvenile criminal record strongly reduces the chance of college admissions, how can we better identify schools at-risk of higher crime rates and what are some of the key associated factors?

Based on our analyses and findings, we aim to propose some important policy recommendations that can enable student success and help bridge the gap to college in a more effective manner.

The results of our project are accessible on GitHub: https://github.com/Joloh13/Data-Science-Final

Dataset Selection and Data Cleaning

Datasets

Our team's combined a total of six datasets. As the number of public high schools did not exceed 500 in NYC, we expected our dataset to contain less than 500 observations. These are listed below.

- 2014 Average SAT Scores per School
- AY 2012-13, AY 2013-14 Funding per student (this was scraped)
- School Safety Report, a report on crimes committed in a given school
- Graduation Rate, a report on enrollment number and number of students graduating

- School Quality Index, a survey conducted by students assessing various aspects of school life such as inclusivity
- Demographic Snapshot, i.e. the percentage of students with disabilities, belonging to a given race, etc.

Each of these databases contained a District Borough Number (DBN), which we used as our primary way of identifying the schools. We wrote a script that scraped data on the funding given to a school by the City Government of New York. After concatenation and removing of rows without meaningful values, our final dataset contained 409 public high schools with the following columns.

- AY12-13 budget
- AY13-14 budget
- Economic Need Index
- Percent Asian
- Percent White
- Percent Hispanic
- Percent Black
- Student Attendance Rate
- Percent of Students Chronically Absent
- Supportive Environment - Percent Positive
- Percent of English Language Learners
- Percent Students with Disabilities
- AvgOfNoCrim N
- Total Grads
- Dropout
- Still Enrolled
- Average Score (SAT Math)
- Average Score (SAT Reading)
- Average Score (SAT Writing)
- Percent Tested

Data Cleaning

Upon further inspection, it was discovered that the average number of crimes column contained many NaN values. These NaN values were missing systematically, as only positive numbers were reported. We deduced that NaN values represent zero crime in the given school. When appropriate data whitening methods were used, standardizing the dataset. Further details on data cleaning are reported for each of the model training.

Dealing with Confounding Factors

We understand that the field of education is affected by many confounding factors and we have tried to control for that. However, many of these factors are external and have not/cannot be accurately measured by numbers eg lack of self-confidence and social pressure.

Assumptions

1) Although we had a relatively small set of data samples, we made the assumption that it would be better to stick with what we have instead of employing techniques like SMOTE. This is because education has so many confounding factors and we did not want to introduce any more bias than necessary.
2) We assumed that data from across two-years (2013-2014) is generally reflective of the public school situation overall.

Part 1 - SAT Score Prediction

Exploratory Analysis

First, a total of 36 observations were removed, as they had no SAT score data. Second, the columns of the cleaned dataset were analyzed, looking into whether there was any correlation between them. The below report was generated using Pandas Profiling and indicates any columns that are highly correlated with another, as well as columns that have a large number of zero values.

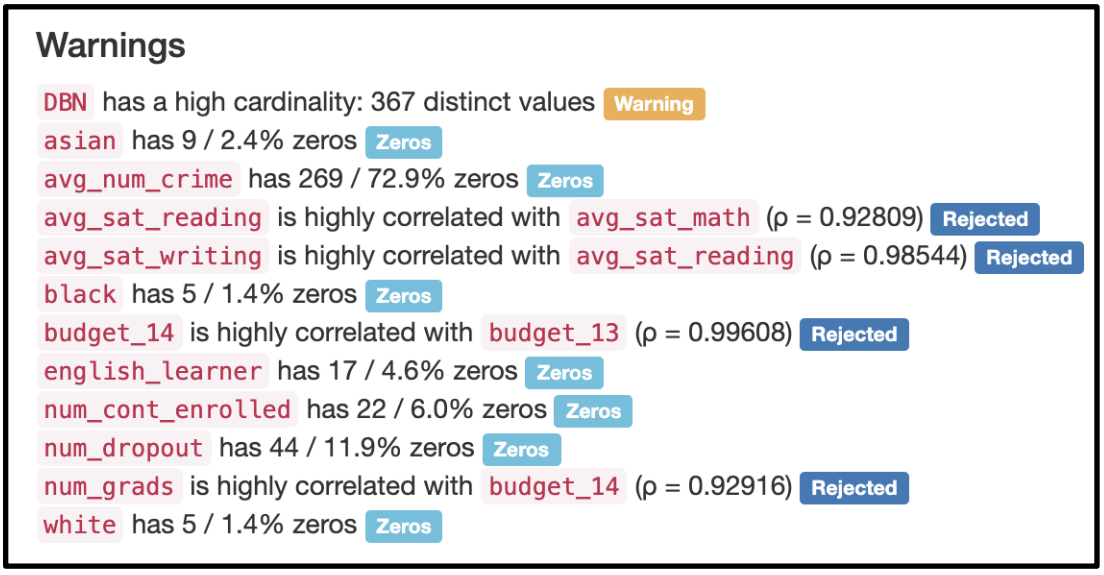


Figure 1. Pandas Profiling report of the dataset

This suggests that combining SAT scores of the reading, writing, and math section is highly desirable as they are correlated. The zero values are not a concern for our study. After this analysis, the average SAT scores were summed into a new column. Next, a Pearson correlation matrix was created from this refined dataset.

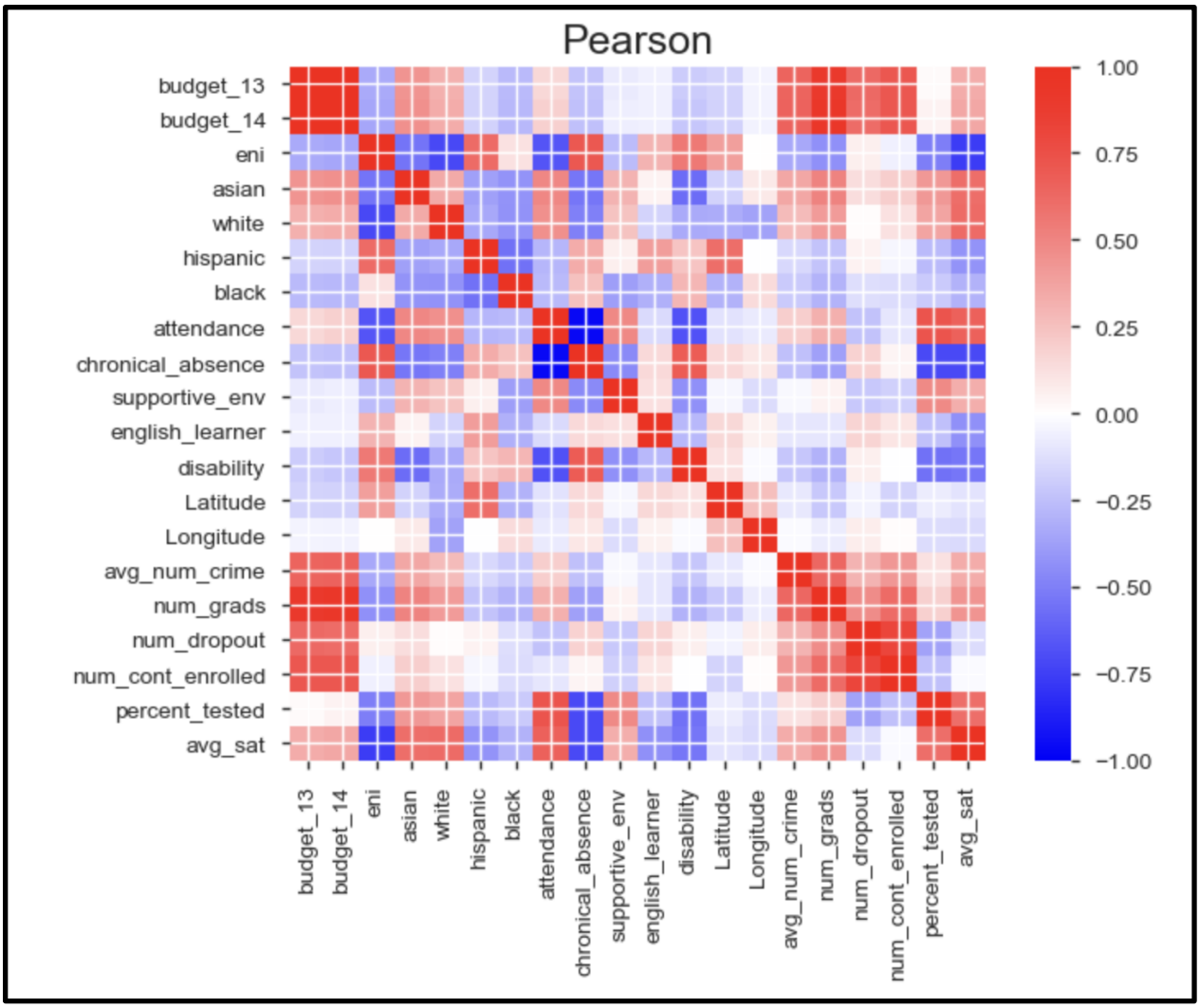
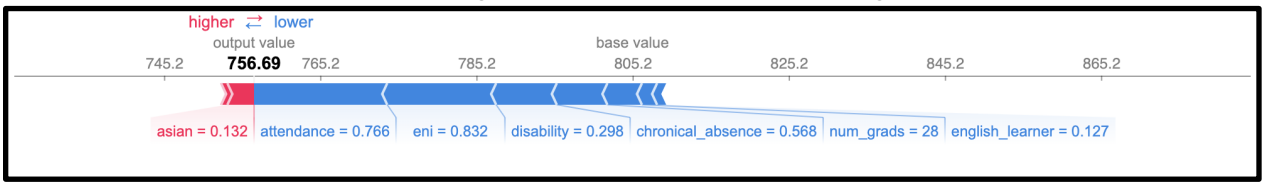


Figure 2. Pearson correlation matrix of our dataset

From this visualization, it can be seen that economic need index (ENI) is strongly and inversely correlated with the average SAT scores. The ENI indicates the probability that the children in that school are living in poverty. The higher the economic need, the lower the SAT score. This could indicate various issues. The SAT tests and SAT test preparations do not come with a low price, and therefore the price could be a barrier for students'. These students may not be able to afford to retake or to hire professional tutors.

Furthermore, we conducted a SHAP analysis. Below are two visualizations where the pink color indicates that the specific variable pushes higher our predicted variable, the average SAT score, and the blue color indicates that the given variable drives the average SAT score lower.



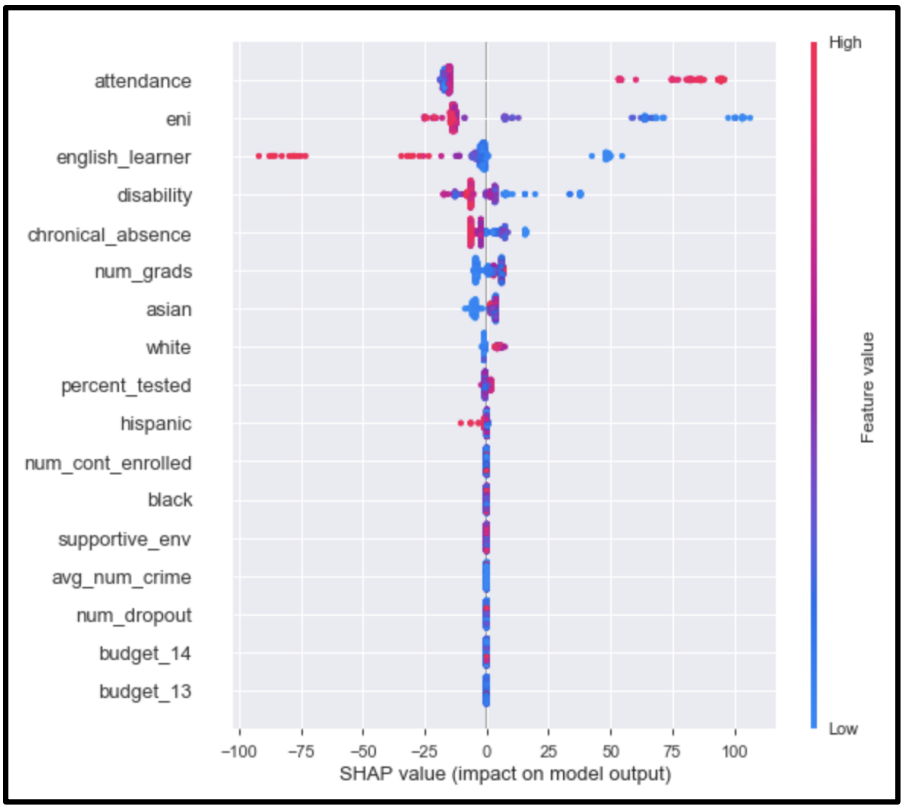


Figure 3. SHAP analysis of our dataset

From this visualization, it is apparent that a higher Asian student population drives the SAT scores higher, while the percentage of the English language learner drives the SAT score lower. As a report by Brookings shows, Black students tend to score lower on the SAT, while Asian students tend to score higher. We believe that here, in fact, we are dealing with a confounding factor that goes back to the complicated history of racial segregation, which resulted in deep economic disparity. While in 1954 Brown vs. Board of Education combat against racial segregation, its effects are still visible today. There is also a number of articles, such as the one written by The Conversation, which argues that Asian students are socialized into test preparation through after-school activities. In short, we believe here we encountered a confounding factor that our dataset did not capture.

On the other hand, the effect of a higher number of English language learners has a straightforward explanation as the SAT has a reading and writing component. This means, the larger the number of students whose mother tongue is not English, the lower the average SAT score of that school is. Again, ENI also appears in our SHAP analysis as strongly and negatively correlated with the average SAT score of a school. The percentage of disabled students also seems to drive the average SAT score lower. While our dataset does not have an explanation of what disabilities were considered when calculating the percentage of the student body that is disabled, this could also include dyslexia and dyscalculia. Both are known to hinder performance on standardized tests.

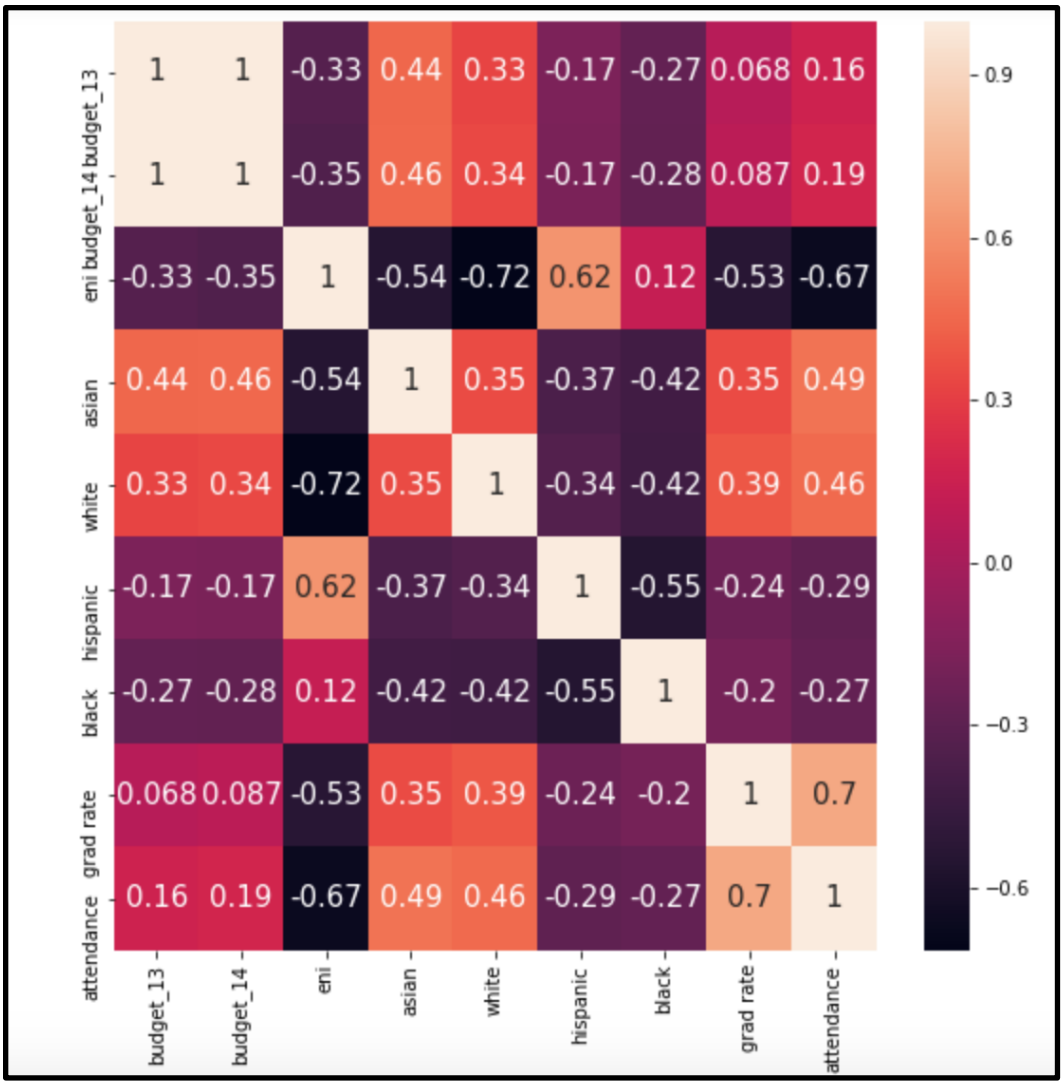
Model Choice

In order to predict the SAT score, a Logistic Regression model was trained. The Vanilla Logistic Regression was initialized using default parameters using sklearn. The dataset was split into a train-test dataset with a split of 33% of the data going into testing. As we are using logistic regression, we whitened our dataset by using a standardizer_scaler. Our Vanilla Logistic Regression returned a Mean Absolute Error (MAE) of approximately 104 SAT points. Afterward, a grid search was conducted to improve model performance. With the parameters of using an L2 penalty score and a maximum iteration of 5, our model was able to reduce its MAE to 98 SAT points.

Part 2 - Graduation Rates

High school graduation is a necessity for college admissions. However, many children do not have the chance to graduate from college. For data cleaning, we removed rows that had NaN values. We made a new column called grad_rates that were generated from (number of graduates)/(number of dropouts + number of still enrolled + number of graduates). Note: number of still enrolled indicates the students who failed that year's grade and had to retake the grade.

Based on the correlation matrix of our data overall, we selected some columns of data that had strong correlations with graduation rates. Below is a correlation matrix for a smaller subset of columns.



We can see that graduation rates are the most strongly negatively correlated with the ENI, which intuitively means that the higher the likelihood of poor children in a school, the lower the graduation rate of that school. For over a decade, New York City has the Fair Funding Act in place to help these children, but it shocked us to see that school budget funding (budget_13 and budget_14) barely affected the graduation rates. We hypothesize that maybe the schools are not getting the funding they were promised, or that the funds are not being used appropriately.

Graduation rates are most strongly positively affected by attendance rate, which intuitively means that the more the child attends school, the more likely he or she is to graduate. There is a correlation (not causation!) between the race of the child and their likelihood of graduating high school.

Model Choice

We ran a linear regression model to predict the effect of ENI on graduation rates. We focused on schools that had a graduation rate of less than 75%, which means that 1 in 4 kids don't graduate. Before this we did some data standardization to make the optimizer more stable. We were able to predict graduation rates accurately up to a mean absolute error of 0.6% and a median absolute error of 5%. We decided to report median absolute error because it is less resistant to outliers.

We also did a thought experiment where we wanted to see the effects of helping 1 in 5 children out of poverty. We found that graduation rates drastically improved.

Part 3: Crime in NYC Public High Schools

Any record of high school crime can have a major impact on a student's chances of admission to college. Moreover, having a safe and secure environment is a critical consideration for parents looking to select a high school for their children and for students looking to thrive.

In this section of the report, we dived into the data on crimes in NYC high schools to build a better understanding of the current situation, contributing factors and possible ways to detect 'at-risk' schools.

Key Data Cleaning Steps: A few key steps taken to facilitate data analysis and modeling include:

1) Dropping null values using NaN
2) Dropping irrelevant columns and specifically fields which do not have any impact on crime such as Location Index and School Code
3) Converting string values to float/int values

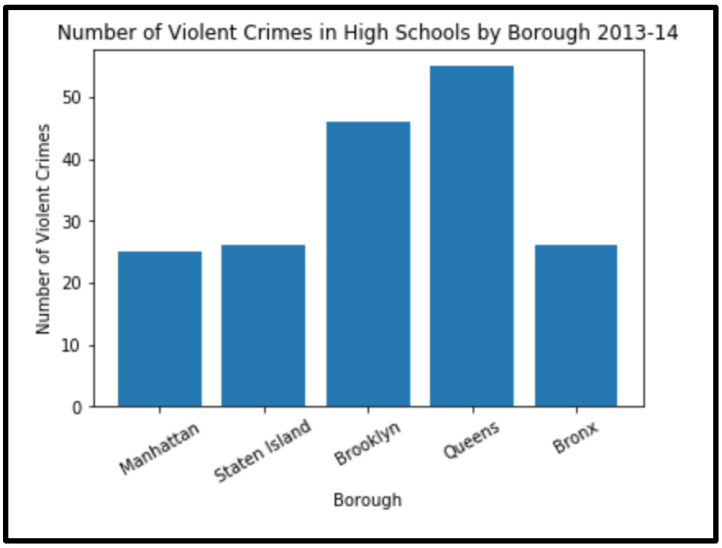
Visualizing Crimes by Location & Borough

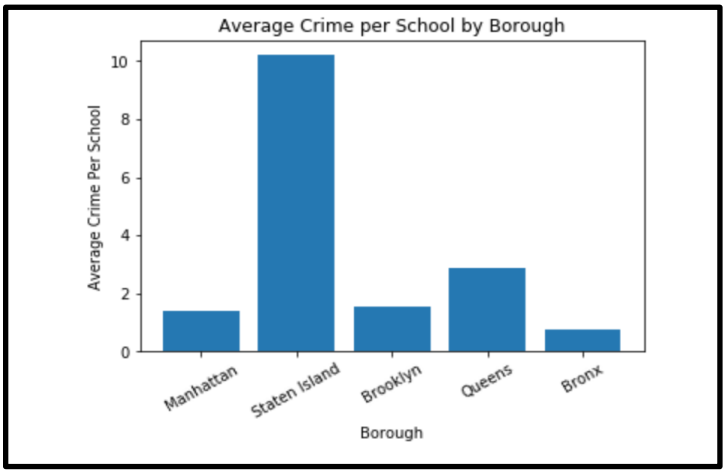
Students in public high-schools are assigned to a specific public high school based on the district of residence. This makes the location of high schools with crimes a very important piece of information.

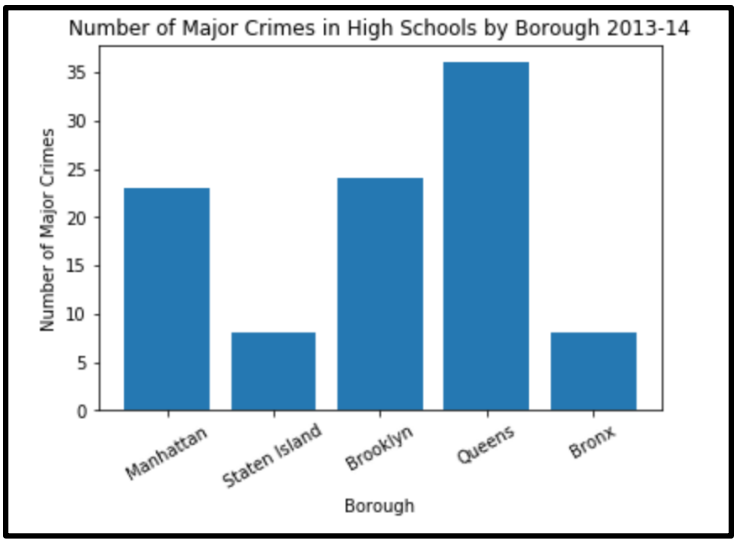
We ran analyses to visualize the number of a) Total Crimes by borough and b) Major or Violent crimes by borough. We also visualized the average number of crimes at public high schools by borough.

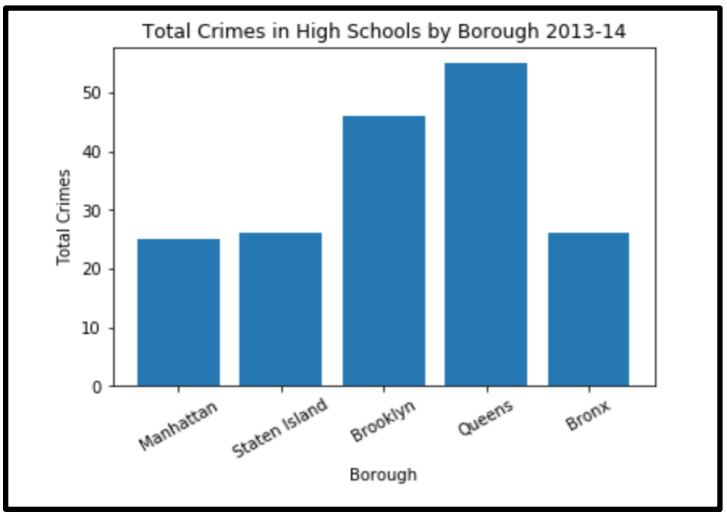
Lastly, we used the Folium package to build a heat map of a) Total Crimes b) Major Crimes and c) Violent Crimes on a map of NYC.

The results were as follows:

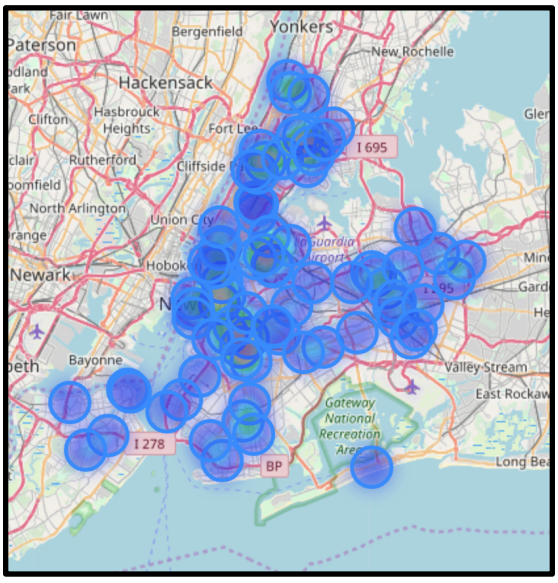




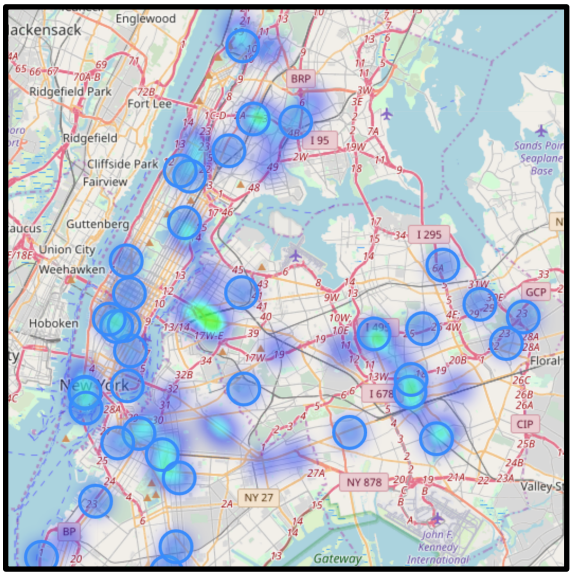

Number of Major Crimes in High Schools by Borough 2013-14
Number of Major Crimes
0
5
10
15
20
25
30
35
Manhattan
Staten Island
Brooklyn
Queens
Bronx
Borough


Total Crimes in High Schools by Borough 2013-14
Total Crimes
0
10
20
30
40
50
Manhattan
Staten Island
Brooklyn
Queens
Bronx
Borough

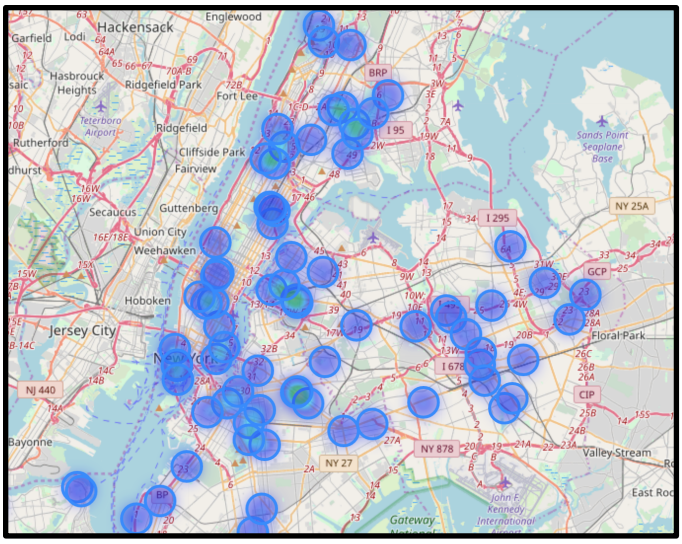
Heatmap Visuals of Total Crimes on NYC Map:



Heatmap Visual of Major Crimes in NYC High Schools:



Heatmap Visuals of Violent Crimes in NYC:

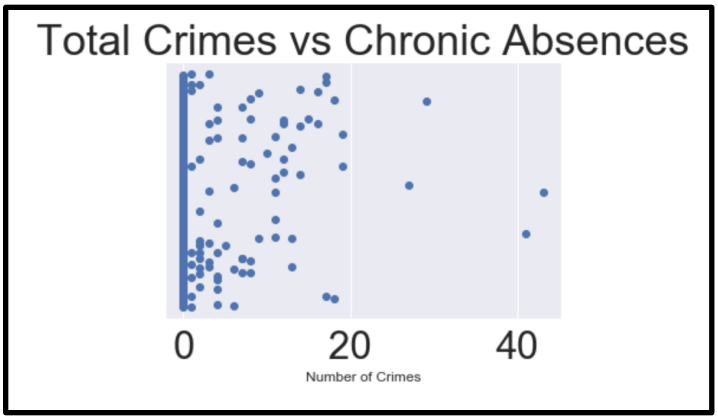


Analyzing Correlations & Relationships Between Variables & Recorded Crimes:

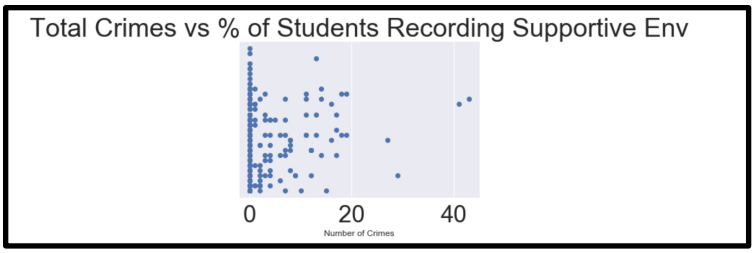
We then looked into which data variables most highly correlated with the number of recorded crimes and/or recorded major/violent crimes.

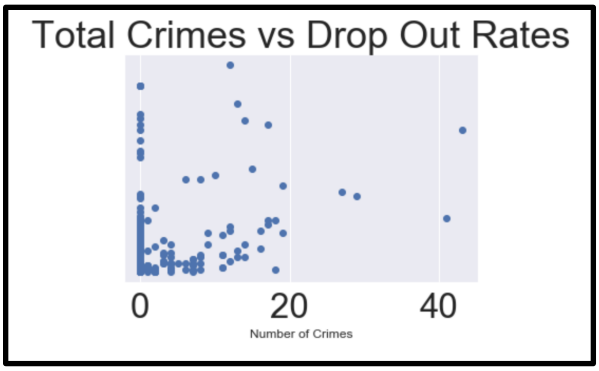
To do this, we first built 2-dimensional scatter-plots (crime vs selected factors) before building a more detailed Pearson's Correlation Heatmap Matrix.

The results were as follows

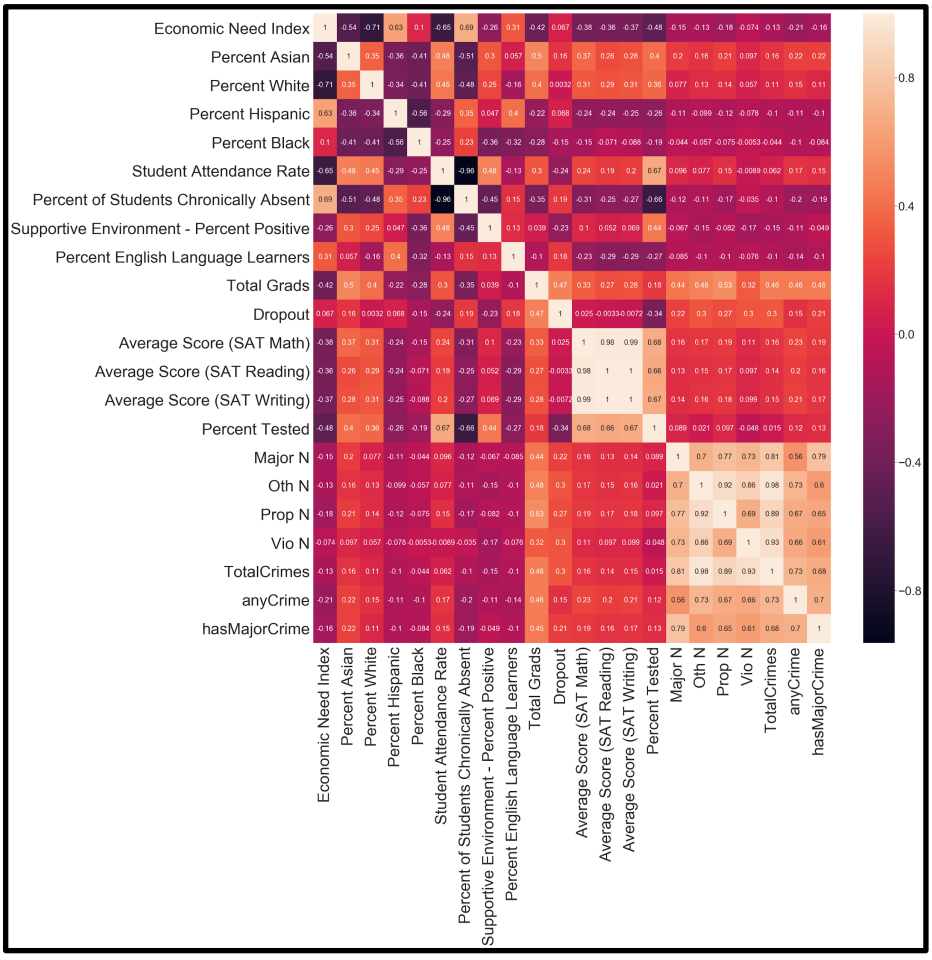


:





Pearson's Correlation Matrix:



Building a Logistic Regression Model to Predict Crime:

1) We first built a model to predict whether which schools may have any crime i.e. major, violent, property and others. We built a 2nd model to predict whether there were any major crimes at schools.
2) We chose the top 7 most important independent variables from the results of the Pearson's correlation and tried to ensure there were no correlated independent variables.

3) We split the data 75%-25% for training and testing

4) Running the results logistic regression models, we were able to then use the coefficient values to determine the most critical contributing factors to a prediction of crime in high schools.

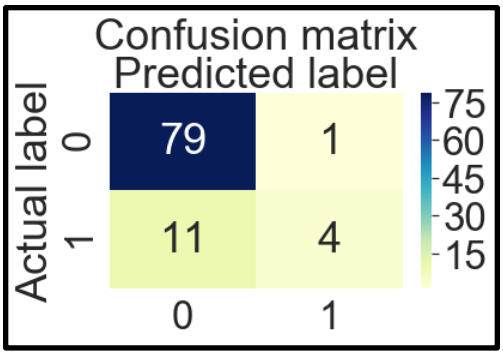
Important Note/Caveat:

Binary Logistic Regression requires that all Independent Variables are completely independent of each other. In other words, there should be little or no multicollinearity. Given the many possible confounding variables in the Education Industry, it is unlikely that all of the independent features chosen are completely independent. For example: Economic Need Index is likely to be higher at schools with a lower Student Attendance Rate.

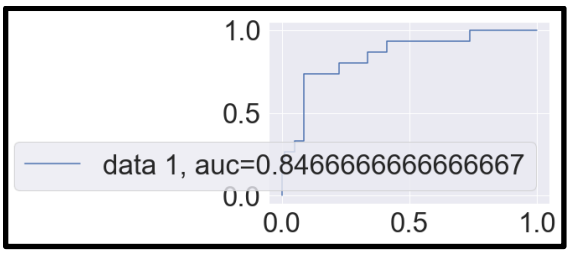
The other major caveat we want to call out is the fact that Logistic Regression usually requires the sample data set to be quite large. However, there are only 410 public schools in New York, and even fewer public schools with reported crimes or major crimes. We also looked for data on private high schools in NYC - however we found limited to no data.

Thus this model is built on a fairly small sample of data

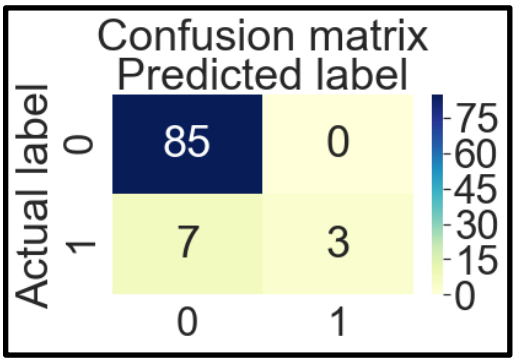
Results from 1st Logistic Regression Model: Predicting *Any Crime:*



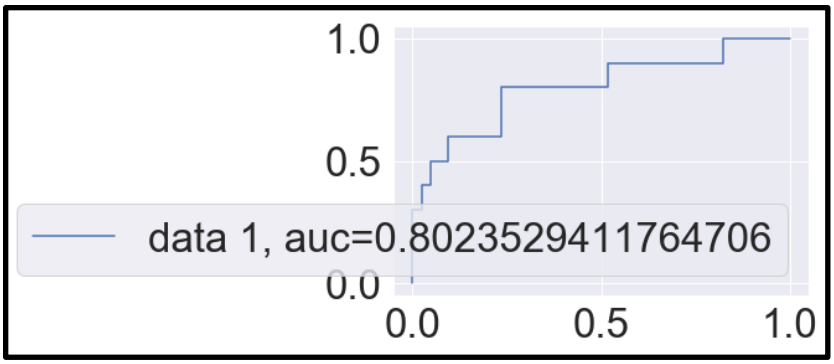
Accuracy: 0.8736842105263158
Precision: 0.8
Recall: 0.26666666666666666



Results from 2nd Logistic Regression Model: Predicting *Major Crimes at High Schools:*



Accuracy: 0.9263157894736842
Precision: 1.0
Recall: 0.3



Key Takeaways from Logistic Regression Analyses:

The independent variables sorted by the most impact on the prediction model to the least, are: 1) Supportive Environment - Percent Positive Recorded by Students 2) Percent English Language Learners 3) Student Attendance Rate

---

Part 4: Policy Suggestions

1. **Building Student Support Systems & Increasing Number of Guidance Counselors:**
   Based on our market research and findings on key contributors to crime/lower student graduation rates; building a more supportive environment for personalized student support is critical to reducing crime and enabling students to graduate.

   The current ratio of students to guidance counselors stands at an abysmal 400:1 Ratio. This is a key overlooked aspect of the student experience and focused initiatives on increasing the number of guidance counselors for personal and professional support can be important for student success.

2. **Identifying & Building a Strategy for Schools with Low Supportive Index as Reported by Students:**
   The most critical factor for juvenile crime rates in high schools (as revealed by our Logistic Regression analyses) was the % of students reporting a positive environment for support.

   As per the definitions of the ‘School Quality’ report, this term encompasses the resources for personal and professional development that students have at their

disposal in their high school environments. Some examples include mental health initiatives and support, professional development services and facilities for athletic development as well.

For potential policy solutions we would recommend a 3 pronged approach:

- Understanding the key drivers from a student's perspective for a 'positive supportive environment'
- Identifying which measures, facilities and factors have been most important at schools with a high number of students (above some threshold) reporting a positive environment
- Planning initiatives to help build these key facilities and support drivers in schools with a currently low/inadequate 'positive support environment'

3. **Subsidizing or Funding Standardized Test Preparation & Test Taking:**
   Based on our findings, we found a high correlation between the number of students that were tested and graduation rates/college enrollment. Moreover, we found a strong negative correlation between the number of students in low-income households and the total percentage of students tested. Additionally, the current prices & fees for standardized test preparation and test taking are quite high.

   Based on these three findings, we believe that policies aimed at increasing affordability for students to boost the number of students a) Being able to prepare for SATs and b) Taking SAT tests can go a long way in helping encourage college enrollment and student success in general. Subsidies and or complete funding options for students can perhaps be a more effective means of helping students rather than giving blanket funding for schools.

4. **Targeted Initiatives for Encouraging Student Attendance:**
   As many children in poverty are expected to help parents with things like manual labor, we can think of an initiative to encourage parents to send their children to school by giving incentives for attendance rates over a certain percentage. Studies have shown that providing incentives to children themselves might not work well in the future as they will be conditioned to only be motivated to do something when there is a reward for them.
