Public school safety in New York City

Vidyasagar Kummarikunta
DSC 680 – Applied Data science

Professor Catie Williams
Bellevue University

Contents

Introduction --- 3

Data sources --- 3

Research Questions --- 4

Approach -- 4

Methods and Results -- 5

Q & A -- 18

Concluding Remarks -- 19

References --- 20

Appendix A -- 21

Appendix B -- 22

Introduction

As I am planning to move to New York city, parent's mission is to analyze the school system in New York city for kids. Naturally, I began to gather as much data and information as I can to learn about public, private and charter schools in the New York city (NYC) area. As a parent, we look for certain qualities while choosing the best school for our children. For some, it begins by buying or renting in a great school district. Other things we consider are students to teachers' ratio, size of the class, special programs, extracurricular activities and after school programs. There are many resources available to learn about such factors to help with decision making. During my research to find the best school that fit our needs, one such aspect that piqued my curiosity is 'School safety'. In recent years there is a sharp rise in mass shootings, especially in schools. While there are many resources citing which neighborhoods are safe in the New York City area, there are no clear numbers available for schools in particular. Being the cultural, financial and media capital of the world, how safe are the public schools in NYC? **(EDEN, 2017)** This study focuses on finding as much information as possible on school safety. I will also be trying to build a machine learning model with the available school safety data.

Data sources

The New York City Police Department (NYPD) (NYCLU, 2016)has been tasked with the collection and maintenance of crime data for incidents that occur in New York City Public schools. The NYPD has provided this data to the New York City Department of Education (DOE) and DOE has compiled this data by schools and locations for the information of the general public. Listed below are the datasets that I chose to work with (OpenData, n.d.)-

Dataset 1: NY 2010 - 2016 School safety report
This data set has 34 columns and 6310 rows

Dataset 2: 2016 - 2017 School safety report
This data set has 33 columns and 2046 rows

Dataset 3: 2017 – 2018 School safety report
This dataset has 25 columns and 1919 rows

Research Questions

Several news articles suggest that there is increase of crimes in the city's public schools. Especially in the school year 2017 **(Burke & Chapman, 2018)**. So, I will be testing the hypothesis that the school year 2016-17 has higher crime rates reported than previous years or the years after. The null hypothesis would state the opposite. In addition to testing the hypothesis, my aim for this study is to identify trends in the total crimes with each school year. Whether there is an increase or decrease in the crime rate. If there is a trend, what sort of crimes occur the most - violent crimes, non-criminal cases, property crimes or other crimes. Find any anomalies if any. Finally, I will try and attempt to build a model for forecasting crimes by using 2010-2016, 2016-2017 school safety reports as validation and train data and by using 2017-2018 school safety report as test data. I think this analysis will help in verifying the generalized conclusions and will help to shine light on the areas/crimes that require more attention.

In particular, the questions I would like to find answers for are –

- Which year had highest crimes reported?
- What kind of crimes are the most reported?
- Which Borough has the highest crimes reported?
- Is there any relation between the number of crimes and the number of students registered?
- Is there any relation between the number of schools in each building to number of crimes?
- Based on the data sets available, can we build a model to forecast crimes?
- Which model performs better?

Approach

For this project, I will be using Python program and Jupyter notebook. First step is to understand the datasets. Next, I will be performing exploratory analysis. The dataset has lot of missing values and exploratory analysis is key to understand how to handle these missing values. Next step will be feature engineering and selecting dependent and independent variables to build a model. I will be using linear logistic regression, decision tree classifier and random forest classifier models to fit the data and evaluate performance.

Methods and Results

Load datasets and printing the number of columns and rows for each dataset.

```
Number of Rows in 2010_-_2016_School_Safety_Report.csv:  6310
Number of Columns in 2010_-_2016_School_Safety_Report.csv:  34

Number of Rows in 2016_-_2017_School_Safety_Report.csv:  2046
Number of Columns in 2016_-_2017_School_Safety_Report.csv:  33

Number of Rows in 2017-2018_NYPD_Crime_Data.csv:  1954
Number of Columns in 2017-2018_NYPD_Crime_Data.csv:  24
```

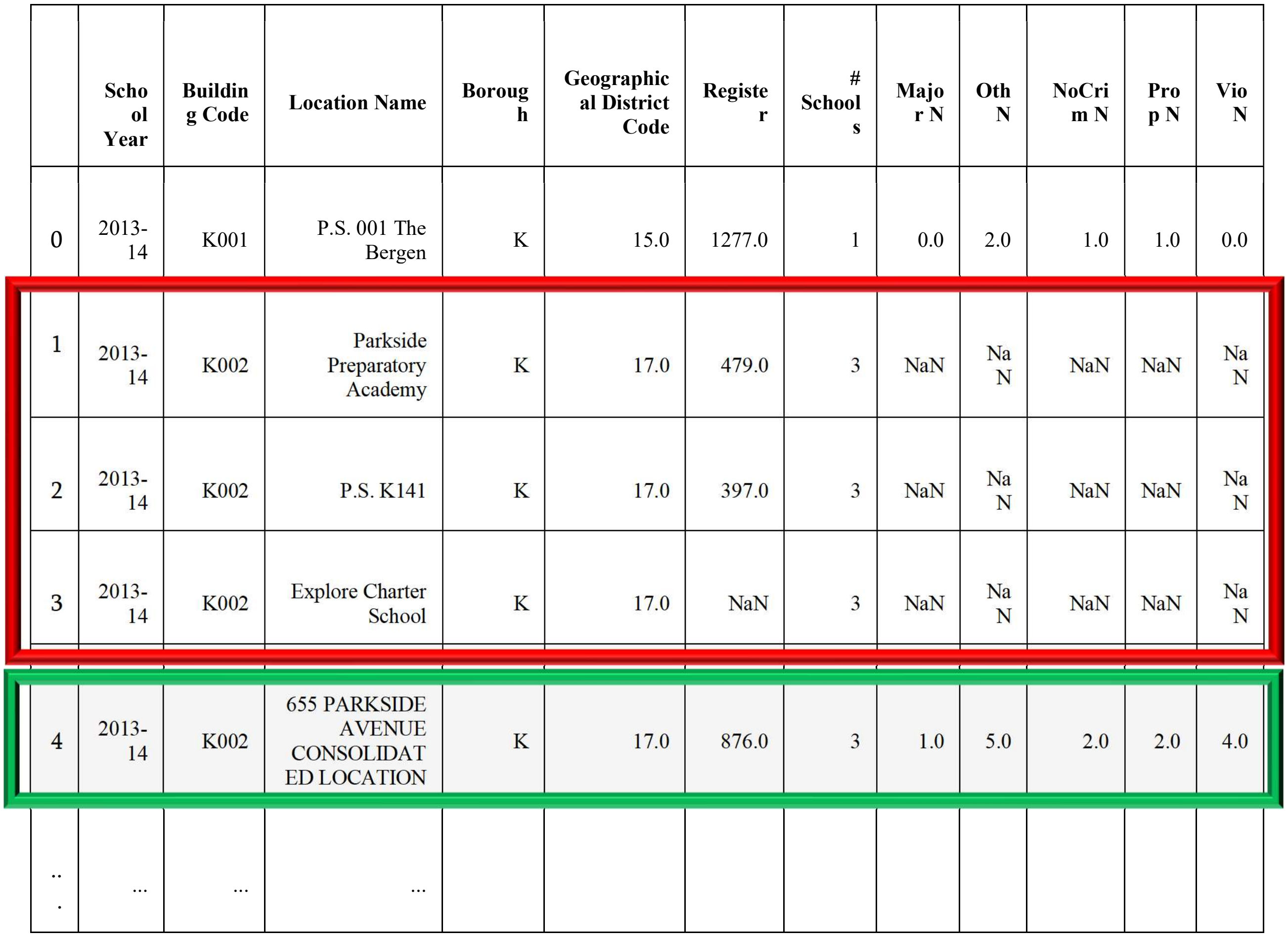
Exploring the column headers for each data set.

```
Column names of 2010-2016 dataframe
--------------------------------------------------------------------------
--
Index(['School Year', 'Building Code', 'DBN', 'Location Name', 'Location
        Code', 'Address', 'Borough', 'Geographical District Code',
        'Register', 'Building Name', '# Schools', 'Schools in Building',
        'Major N', 'Oth N', 'NoCrim N', 'Prop N', 'Vio N', 'ENGroupA',

        'RangeA', 'AvgOfMajor N', 'AvgOfOth N', 'AvgOfNoCrim N',
        'AvgOfProp N', 'AvgOfVio N', 'Borough Name', 'Postcode',
'Latitude', 'Longitude', 'Community Board', 'Council District ',
        'Census Tract', 'BIN', 'BBL', 'NTA'],dtype='object')
--------------------------------------------------------------------------
--
Column names of 2016-2017 dataframe
--------------------------------------------------------------------------
--
Index(['School Year', 'Location Name', 'Location Code', 'Borough',
       'Geographical District Code', 'Register', 'Building Name',
       '# Schools', 'NYPD Site Code', 'NYPD Site Name', 'Schools in Buildi
       ng', 'Major N', 'Oth N', 'NoCrim N', 'Prop N', 'Vio N', 'ENGroupA',
       'RangeA', 'AvgOfMajor N', 'AvgOfOth N', 'AvgOfNoCrim N', 'AvgOfProp
       N', 'AvgOfVio N', 'Geocode', 'Postcode', 'Latitude', 'Longitude',
       'Community Board', 'Council District', 'Census Tract', 'BIN',
       'BBL', 'NTA'], dtype='object')
--------------------------------------------------------------------------
--
Column names of 2017-2018 dataframe
--------------------------------------------------------------------------
--
Index(['ID', 'Building Code', 'DBN', 'Location Name', 'Location Code',
       'Address', 'Borough', 'Geographical District Code', 'Register',
       'Building Name', '# Schools', 'Schools in Building', 'Major N',
       'Oth N', 'NoCrim N', 'Prop N', 'Vio N', 'ENGroupA', 'RangeA', 'AvgO
       fMajor N', 'AvgOfOth N', 'AvgOfNoCrim N', 'AvgOfProp N', 'AvgOfVio
       N'], dtype='object')
---------------------------------------------------------------------------
--
```

Consolidating the datasets by using only relevant variables. The major focus of this study is the crime data. Major N, Oth N, NoCrim N, Prop N and Vio N are the five columns of interest and also the target variable.

If we observe Dataset1 as shown below (**Fig.1**), it appears that multiple schools in an area share one building. For example, crime data for the building codes K002 in row 1, 2 and 3 (highlighted in red) is represented in row 4 (highlighted in green) as a single consolidated location. If I were to use the data as such, the observations will be misrepresented. The data for such group of schools is represented multiple times which makes the data redundant. After exploratory data analysis, the best method to handle missing values is to delete all the rows that have atleast one missing value. Keeping the redundant data will skew the model, and I think this is the best way to handle these missing values. The same is observed in datasets 2 and 3. So, I will be deleting all the NaN values to eliminate redundancy.



	School Year	Building Code	Location Name	Borough	Geographical District Code	Register	# Schools	Major N	Oth N	NoCrim N	Prop N	Vio N
0	2013-14	K001	P.S. 001 The Bergen	K	15.0	1277.0	1	0.0	2.0	1.0	1.0	0.0
1	2013-14	K002	Parkside Preparatory Academy	K	17.0	479.0	3	NaN	NaN	NaN	NaN	NaN
2	2013-14	K002	P.S. K141	K	17.0	397.0	3	NaN	NaN	NaN	NaN	NaN
3	2013-14	K002	Explore Charter School	K	17.0	NaN	3	NaN	NaN	NaN	NaN	NaN
4	2013-14	K002	655 PARKSIDE AVENUE CONSOLIDATED LOCATION	K	17.0	876.0	3	1.0	5.0	2.0	2.0	4.0
...	...	...	...									

Fig.1

Another important observation to be noted here is that, the data file from which Dataset 1 is extracted is labeled as ‘2010 -2016’ which led me to believe that the data in the csv file is from the year 2010 until 2016. However, upon checking the snapshot of the dataframe, it is evident that this data file has data only from 2013 – 2016.

To compensate for the misleading data, datasets 1 and 2 are joined (referred to as train dataset from here on) to get a comprehensive dataset that has information from years 2013 – 2017. Dataset 3, which has data for the school year 2017 – 2018 is left as is for testing purposes. After joining Datasets 1 and 2, a new column is included “All_crimes” which is the sum of Major N, Oth N, NoCrim N, Prop N and Vio N. The same was done for Dataset 3.

I also replaced special characters and empty spaces in the column names. Other initial data clean-up was done on dataset 3 in which the ‘Register’ column has string values. Those have to be replaced with NaN and the datatype was changed to float64, because any analysis on such data will not be truly represented.

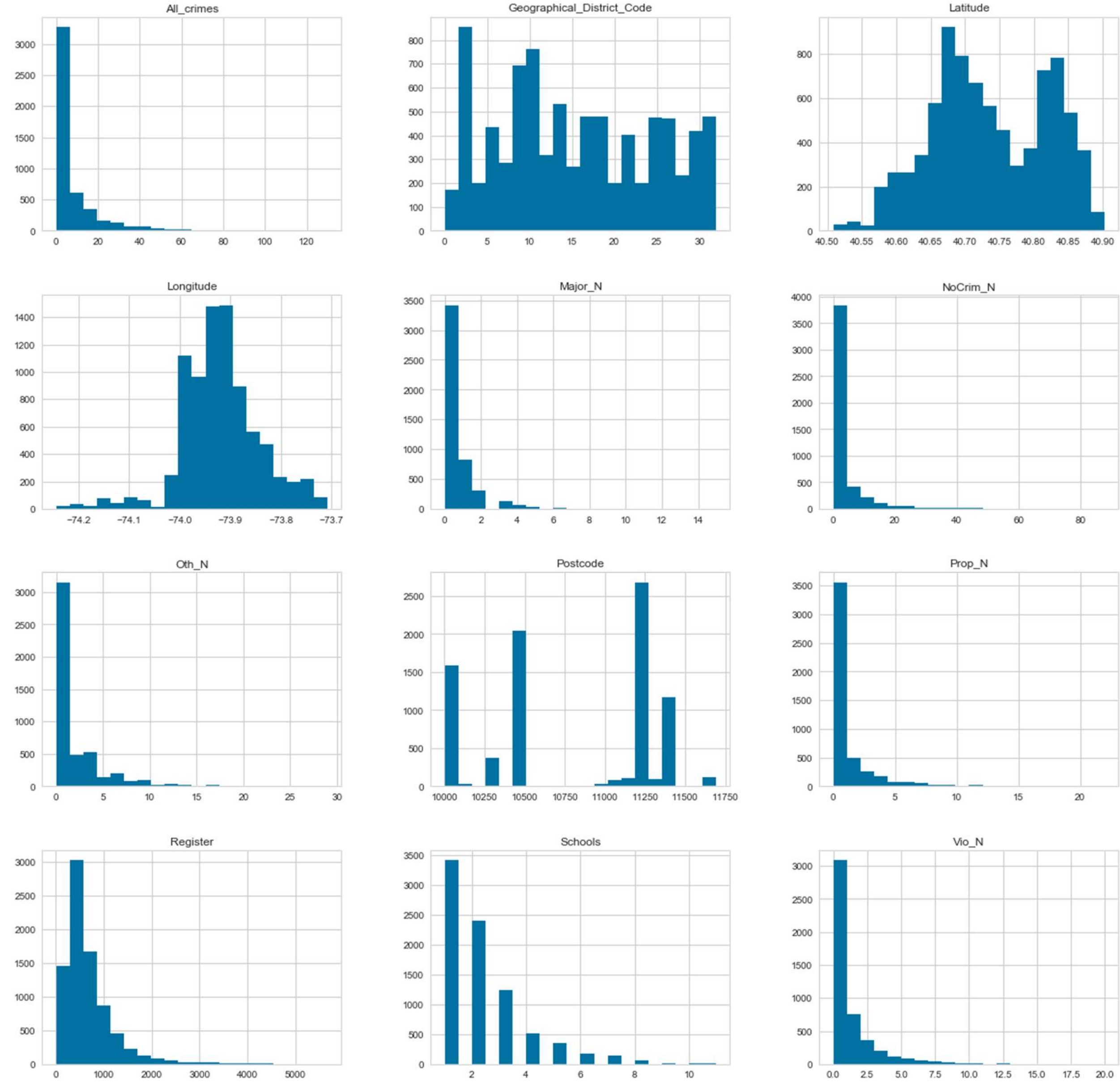


Fig.2 : Histograms of numerical columns in Train dataset

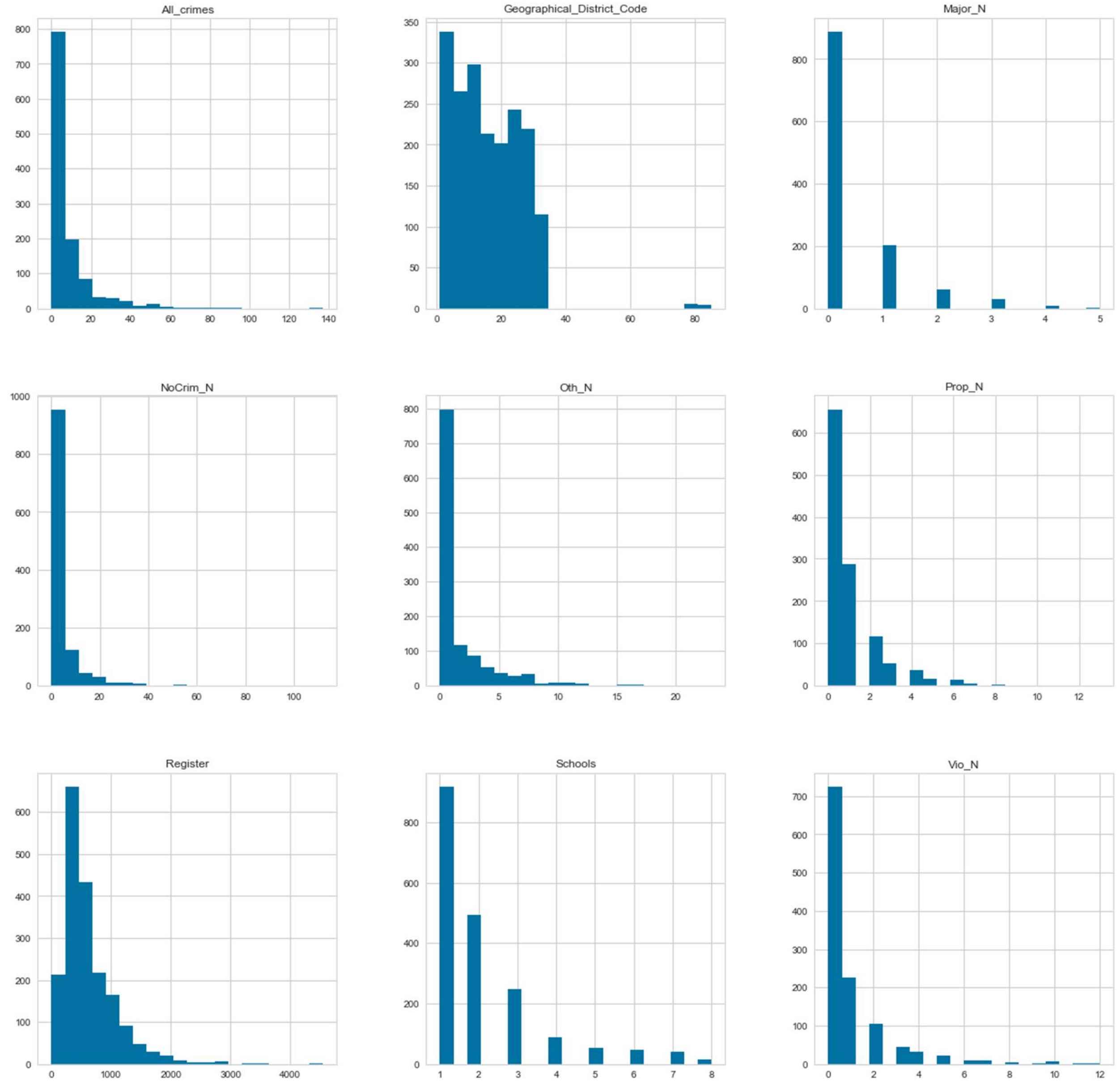


Fig.3: Histograms of numerical columns in Dataset 3

Comparing the histograms of Train set to Dataset 3:

- The column Geographical_District_Code in Train dataset has values that range from 1 – 32 and Dataset 3 has values from 1 – 32 and then 71 - 85. Upon brief research (http://www.newyorkschools.com/nyc-schools/), it is found that NYC has 1 – 31 school districts and there are special school districts that range from 70 – 85. It is clear that the distribution of Geographical_District_Code is significantly different in both datasets.
- Distribution of number of 'Schools' in each building is also noticeably different among the two datasets. In the train dataset, values of the number of schools in each building range from 1 – 11, while the Dataset 3 has a range from 1 – 8.
- There are slight differences in the distribution of columns - Major N, Oth N, NoCrim N, Prop N and Vio N in the two datasets. The distribution of sum of these columns ('All_crimes') is similar in both datasets.
- Distribution of number of students registered ('Register') in each school building is also similar in both datasets.
- Finally, the columns 'Postal_code', 'Latitude' and 'Longitude' are only available for Train dataset. They were not included in Dataset 3. These three columns will only be used for visualizations and will not be included in building a model.

Shown below (**Fig.4**) is the distribution of Borough counts for Train dataset. Brooklyn has the highest counts followed by Bronx, Manhattan, Queens and Staten Island. There is some anomaly in the distribution of Boroughs, as we can see the 'Other' borough category. This is probably an outlier.

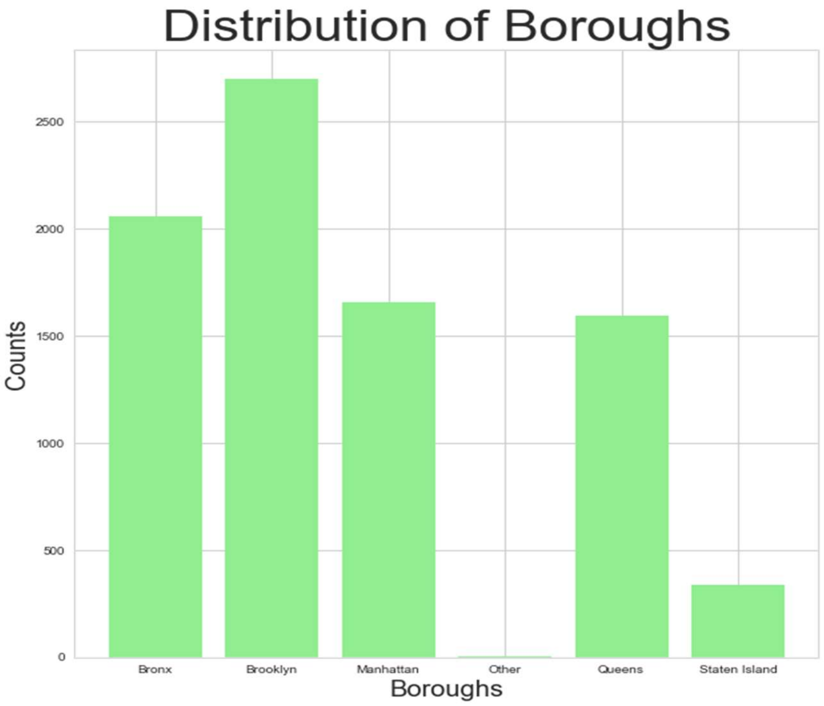


Fig.4

Shown below (**Fig.5**) is the distribution of Borough counts for Dataset 3. Dataset 3 also follows the same trend of Borough counts like the Test dataset.

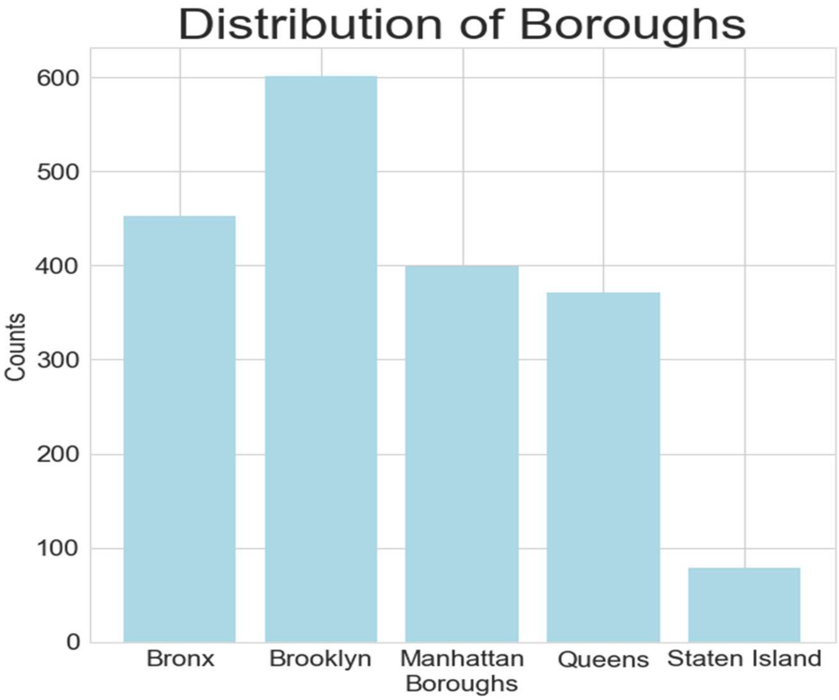


Fig.5

Bivariate analysis - Visualizing stacked bar plots to compare total number of crimes in each Borough from year 2013-2017 combined of the Train dataset (**Fig.6**). If we recall the distribution of Borough counts in Fig. 1, Brooklyn has the highest counts followed by Bronx, Manhattan, Queens and Staten Island. In this visualization, the total number of crimes follows the same trend as Borough count. So, the greater number of schools in a Borough, the greater number of crimes.

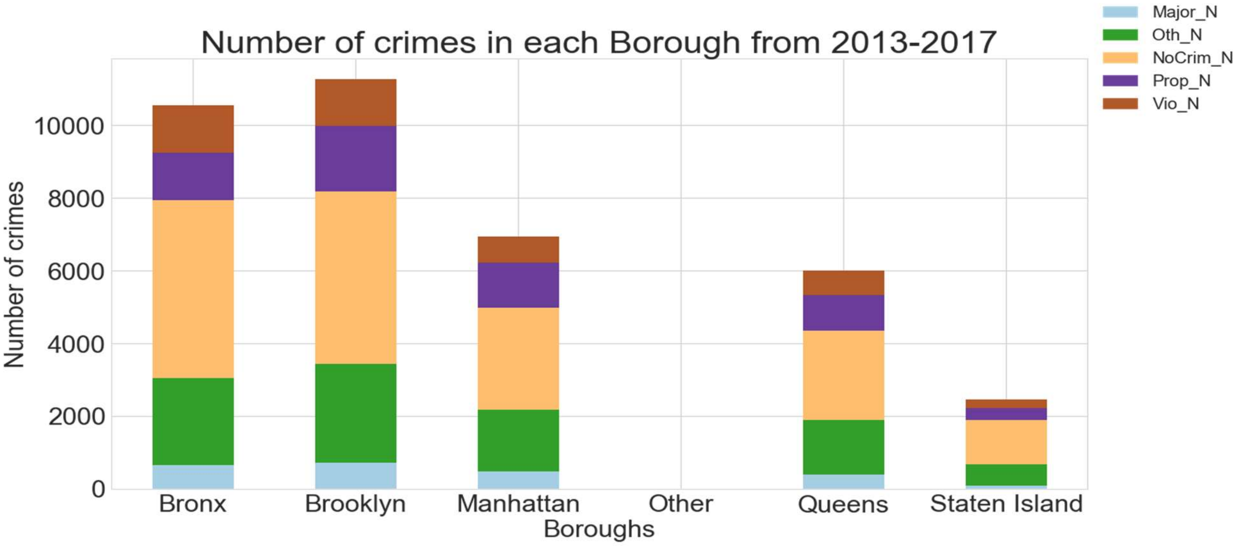


Fig.6

Bivariate analysis - Visualizing stacked bar plots to compare total number of crimes in each Borough from year 2017-2018 (**Fig.7)**. The total crimes in each Borough follow similar trend as in 2013-2017 and also are in proportion to the distribution of Borough counts.

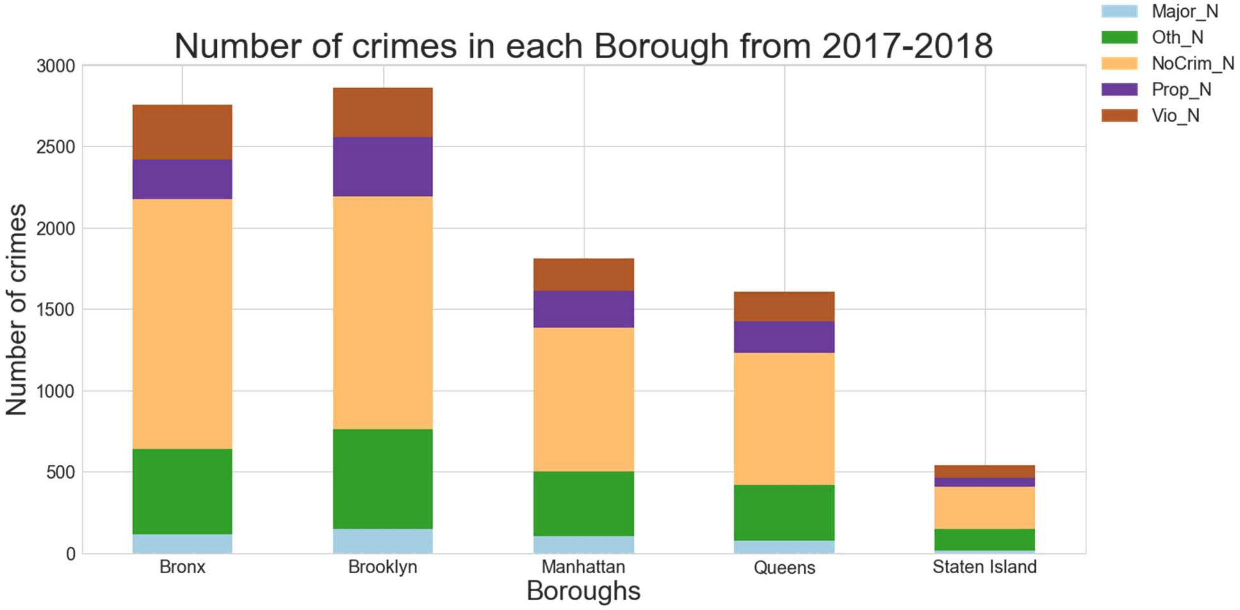


Fig.7

Comparing total number of crimes of the Train dataset **(Fig.8)** and Dataset 3 **(Fig.9)**. It is evident that the number of crimes each year are decreasing from 2013 to 2017 but there is an increase in the year 2017-2018.

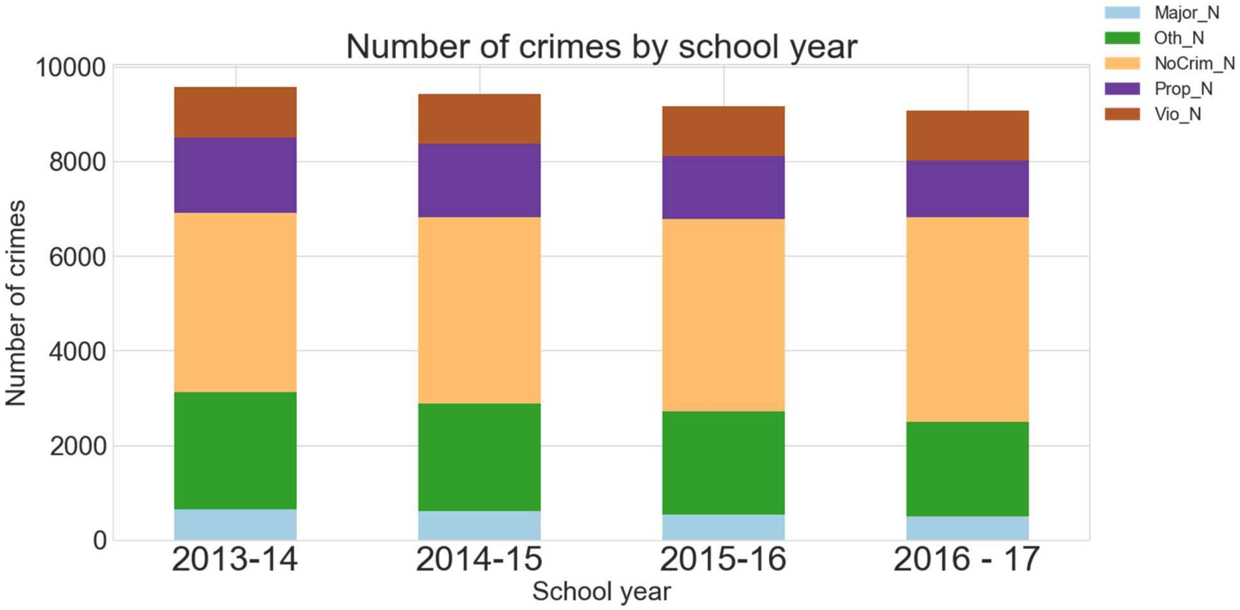


Fig.8

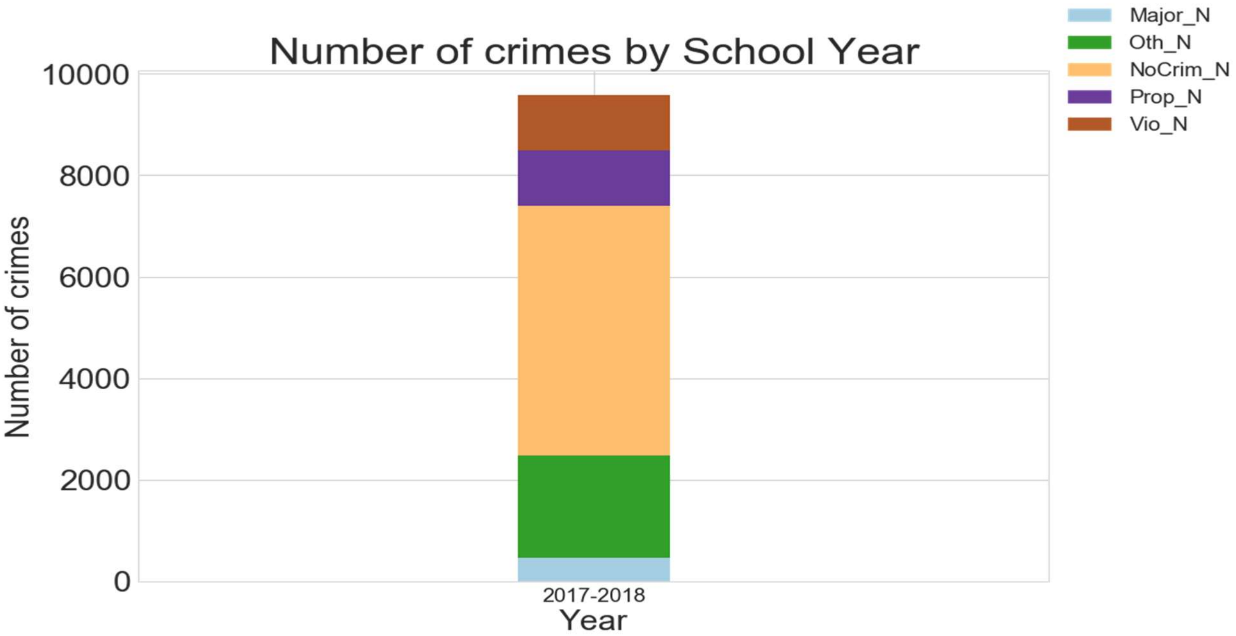


Fig.9

Comparing total number of crimes each year to the total number of students registered of the Train dataset **(Fig.10)** and Dataset 3 **(Fig.11)**. There is a sharp increase in the number of students registered in the school year 2015-2016 compared to 2014-2015 and at the same time, total crimes decreased. The number of registrations (indicated by grey arrow) is significantly low for the year 2017-2018.

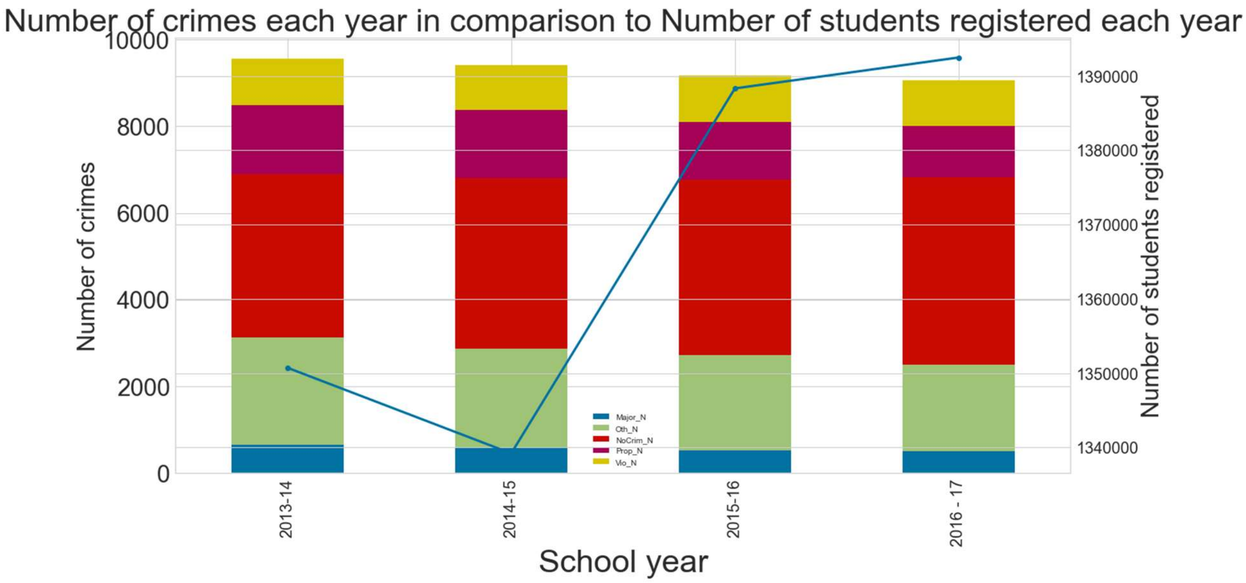


Fig.10

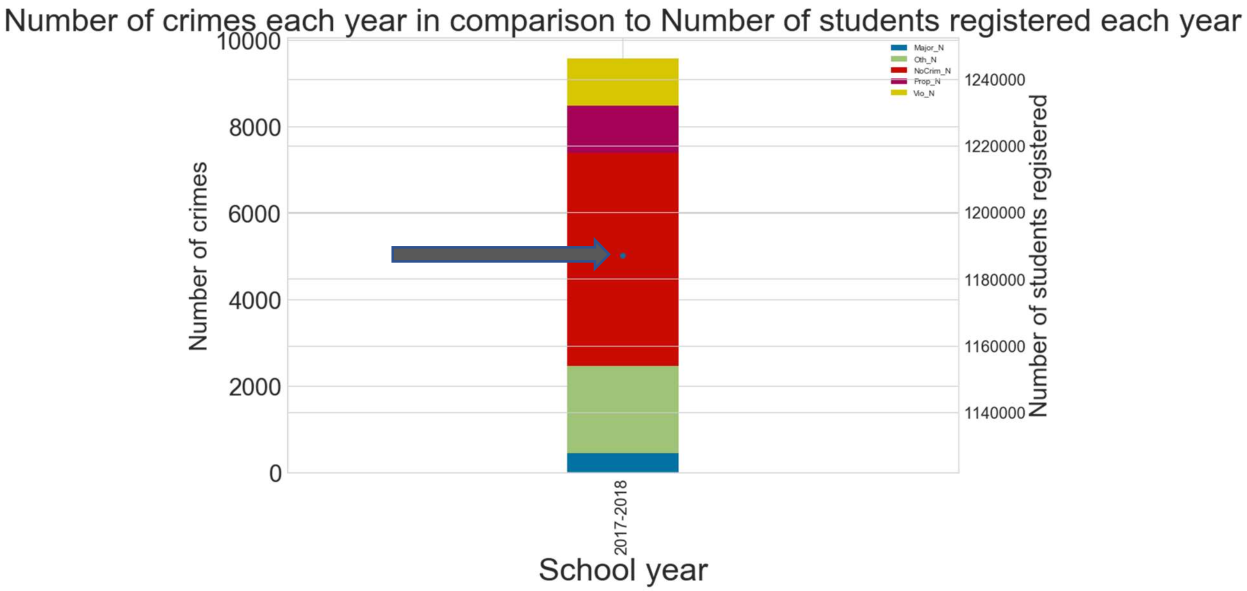

Number of crimes each year in comparison to Number of students registered each year
Number of crimes
10000
8000
6000
4000
2000
0
Number of students registered
1240000
1220000
1200000
1180000
1160000
1140000
Major_N
Oth_N
NoCrim_N
Prop_N
Vio_N
2017-2018
School year

Fig.11

A geographical representation of number of crimes from 2013 – 2017 grouped by Borough.

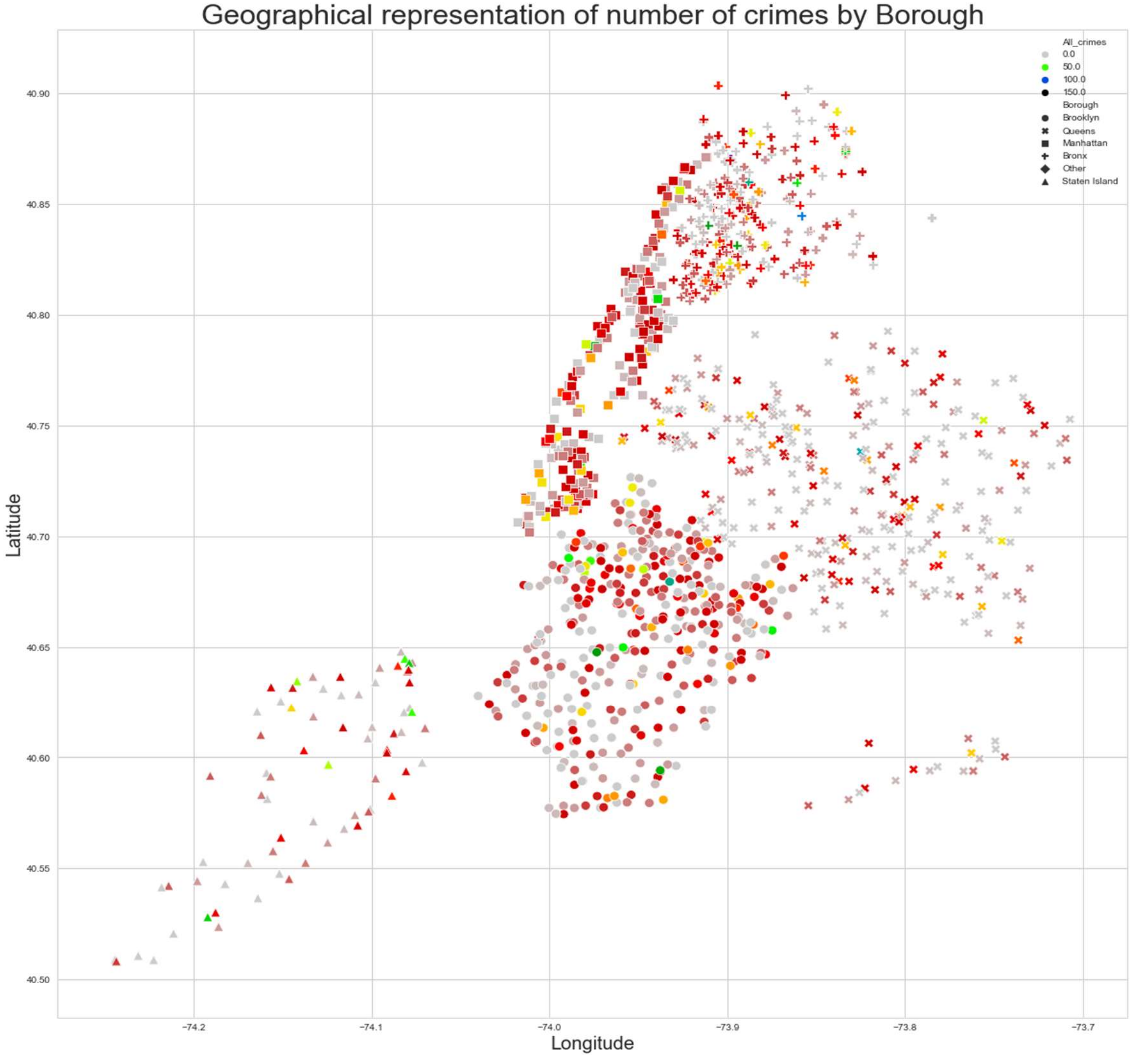


Fig.12

A reference map of the five Boroughs in New York City. Source: (Five Boroughs, n.d.)

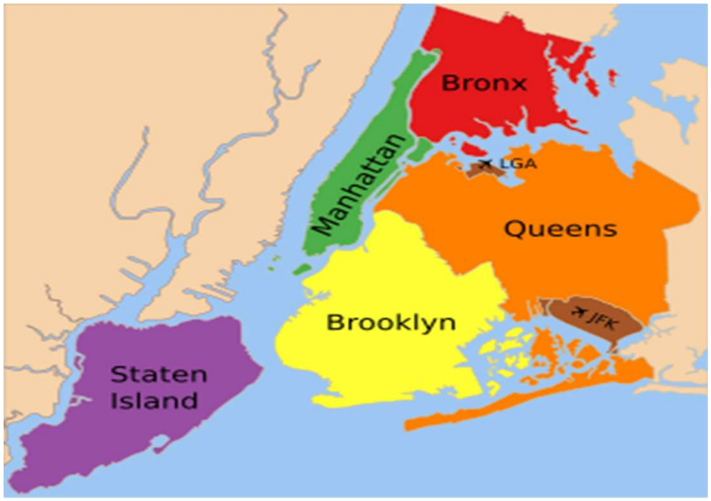


Fig.13

Shown below is the Correlation matrix of Train dataset. From this visualization, we can conclude that the total number of crimes (All_crimes) is positively correlated to the number of students registered (Register) and number of schools in each building (Schools).

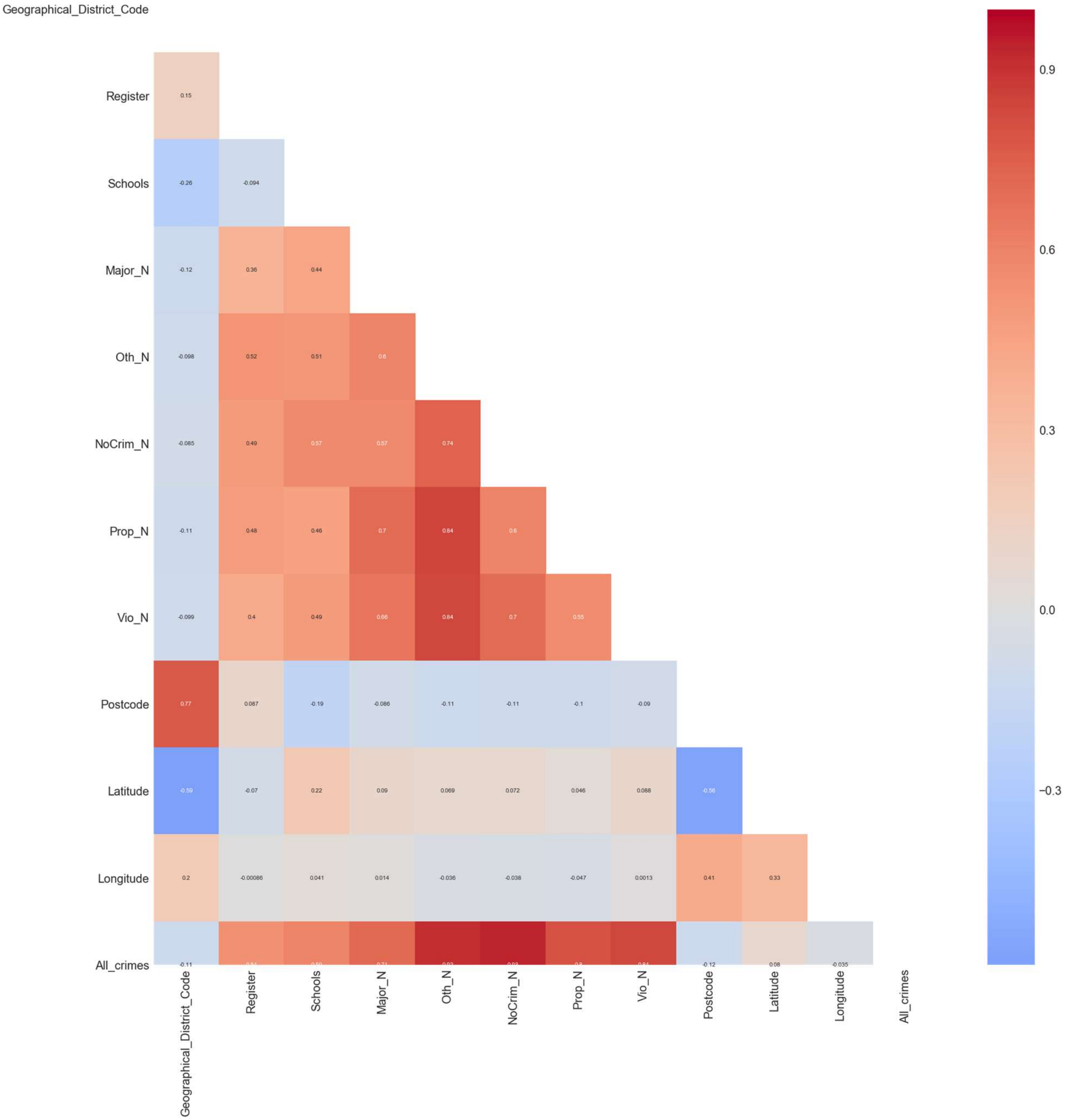


Fig.14

After the exploratory analysis, the next step is to handle missing values. As mentioned earlier, I think the best choice in this case is to eliminate the rows containing missing values. Here is the structure of Train dataset and Dataset 3 after dropping NaNs.

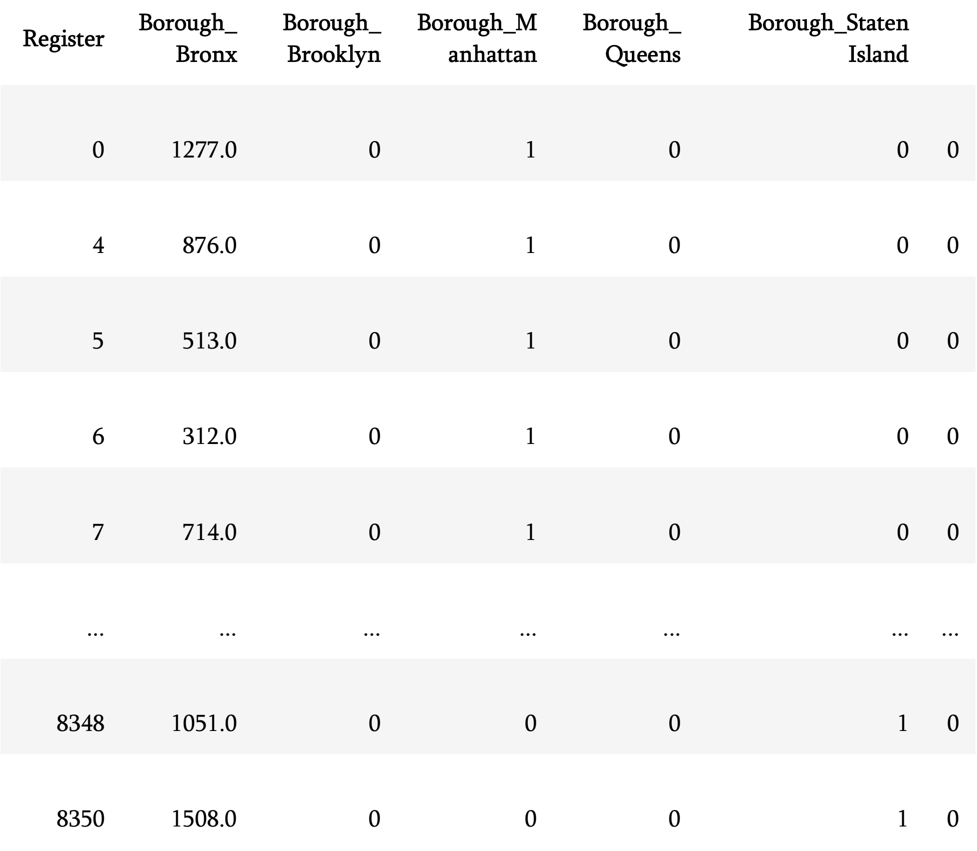
```
Number of Rows in Train dataset  4712
Number of Columns in Train dataset:  16

Number of Rows in Dataset 3:  921
Number of Columns in Dataset 3:  13
```

Training a model:

The aim of building a machine learning model is to predict the number of crimes. So, the target variable is ‘All_crimes’ – ‘y’ set. ‘Borough’ and ‘Register’ columns are used to train (‘X’ set) the model.

Before using the data to train model, dummy values are assigned to the ‘Borough’ column.



Register	Borough_Bronx	Borough_Brooklyn	Borough_Manhattan	Borough_Queens	Borough_Staten Island	
0	1277.0	0	1	0	0	0
4	876.0	0	1	0	0	0
5	513.0	0	1	0	0	0
6	312.0	0	1	0	0	0
7	714.0	0	1	0	0	0
...	...	...	...	...	...	...
8348	1051.0	0	0	0	1	0
8350	1508.0	0	0	0	1	0

Register	Borough_Bronx	Borough_Brooklyn	Borough_Manhattan	Borough_Queens	Borough_Staten Island	
8351	712.0	0	0	0	1	0
8354	334.0	0	0	0	1	0
8355	1287.0	0	0	0	0	1

Using the sklearn library, Train dataset is split into 70:30 – train:test ratio. I have chosen three types of models – Logistic regression, Decision tree classifier and Random forest classifier for this study. After training the models, Dataset 3 is used as a validation set. The accuracy scores of each model is shown below.

```
Logistic regression Train Accuracy ::  23.47949080622348
Logistic regression Test Accuracy ::  24.429967426710096

Decision tree classifier Train Accuracy ::  13.861386138613863
Decision tree classifier Test Accuracy ::  13.897937024972856

Random forest classifier Train Accuracy ::  19.943422913719942
Random forest classifier Test Accuracy ::  20.412595005428884
```

Q & A

With the results/findings in hand, I would like to revisit the questions to which I intend to find answers for –

Which year had highest crimes reported?

```
School_Year     Total crimes
   2013-14      9566.0
   2014-15      9420.0
   2015-16      9165.0
   2016-17      9060.0
   2017-18      9579.0
```

According to the bar plots and calculations, as we can see in the above table, year 2017-18 the highest reported total crimes. There was a decrease in crimes from years 2013 to 2017 academic years but it rose to high in 2017.

What kind of crimes are the most reported?

For all the academic years from 2013-2018, non-criminal crimes are the most reported (Fig.8 and Fig.9).

Which Borough has the highest crimes reported?

Brooklyn borough has the highest number of total crimes reported from academic years 2013-2018 (Fig.6 and Fig.7).

Is there any relation between the number of crimes and the number of students registered?

The number of crimes has stayed low from years 2013-17 but increased in 2017-18, on the other hand, number of students registered are higher in 2013-17 in comparison to 2017-18 (Fig.10 and Fig.11).

Based on the data sets available, can we build a model to forecast crimes?

The accuracies of Logistic regression, Decision tree classifier and Random forest classifier are very low. This concludes that the data as such is not enough for building any models.

Which model performs better?

None of the three models tested have performed better.

Concluding Remarks

School safety statistics, especially in New York city where students from all backgrounds come together is an interesting topic to me. This study has been helpful in understanding the trends of crime data with respect to Boroughs. This study shows that Brooklyn has highest number of schools, highest number of students registered and also highest number of crimes from 2013 – 2018. Also, in 2015-2016 and 2016-2017, total crimes are lowest compared to other years and had highest number of registered students.

One of the aims of this study is to train a model for ‘forecasting’ number of crimes for subsequent years. However, based on the accuracy scores, none of the models are performing well. I think the parameters that are used to train these models are not sufficient and additional data and features are required to build a better model.

References

Burke, K., & Chapman, B. (2018, March 02). *Crime is up in NYC public schools, NYPD data shows*. Retrieved from Daily News: https://www.nydailynews.com/new-york/education/crime-nyc-public-schools-nypd-data-shows-article-1.3851995

EDEN, M. (2017, June 7). *Exclusive: How Safe Are NYC’s Schools? New Interactive Map Compares What Teachers & Students Are Reporting*. Retrieved from the 74 million: https://www.the74million.org/article/exclusive-how-safe-are-nycs-schools-new-interactive-map-compares-what-teachers-students-are-reporting/

Five Boroughs. (n.d.). Retrieved from https://upload.wikimedia.org/wikipedia/commons/thumb/3/34/5_Boroughs_Labels_New_York_City_Map.svg/600px-5_Boroughs_Labels_New_York_City_Map.svg.png

NYCLU. (2016, July 21). *NYPD RELEASES COMPLETE SCHOOL SAFETY DATA FOR FIRST TIME*. Retrieved from nyclu: https://www.nyclu.org/en/press-releases/nypd-releases-complete-school-safety-data-first-time

OpenData, N. (n.d.). *NYC OpenData*. Retrieved from https://opendata.cityofnewyork.us

Appendix A – Data Variables

- School year – is the academic year of school
- Building code – the ID of each building where the school is located
- Location name – Address of school
- Borough – New York City borough the school is situated in
- # Schools – Number of schools located in the same building
- Geographical district code – The school’s geographical district as defined by the NYC Department of Education
- Register – Number of students on that are registered for the school year
- Major N - Number of major crimes
- Oth N – Number of ‘other’ crimes
- NoCrim N – Number of non-criminal crimes
- Prop N – Number of property related crimes
- Vio N – Number of violent crimes
- Postcode – the zip code of school’s location
- Latitude - school geographical information
- Longitude – school geographical information

Appendix B - Web resources

The data I obtained for this study is from NYC OpenData website (OpenData, n.d.). Here are some news articles and references about the project background.

- https://opendata.cityofnewyork.us – This is the source website that I will be obtaining data.
- https://www.nydailynews.com/new-york/education/crime-nyc-public-schools-nypd-data-shows-article-1.3851995 - News article that claims that school crimes have increased in the last quarter of 2017.
- https://www.nyclu.org/en/press-releases/nypd-releases-complete-school-safety-data-first-time - News article that discusses the school safety statistics in NYC and the measures taken to prevent them.
- https://www.the74million.org/article/exclusive-how-safe-are-nycs-schools-new-interactive-map-compares-what-teachers-students-are-reporting/ - This article shows visualizations of school safety statistics in the form of a geo map. Which I intend to implement a similar on in my final project report
- https://www.wnyc.org/story/school-safety-incidents-vary-depending-who-counts/ - Another news article
- https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4 - I will be using data cleaning techniques that seem appropriate for this project.
- Think stats – Exploratory data analysis by Allen Downey, - This text book is a great source for me to perform basic exploratory analysis to determine the nature of data.
- Story telling with data, a data visualization guide for business professionals by Coke Nussbaumer Knaflic – For ideas to present the project that appeals to audience