FDNY Violations Prediction

Predictive Modeling of Bureau of Fire Prevention (FDNY) Violation Orders to Prioritize Fire Inspections

Gurpreet Singh
DATA 698 Capstone Project
Professor Arthur O’Connor

05/04/2020

Table of Content

ABSTRACT

Violations Prediction utilizing Machine Learning will assist the Bureau of Fire Prevention in streamlining the inspections process to reduce the fire risk to people and property. High Risk Zip Codes in New York City were predicted using Machine Learning Algorithms Random Forests, K Nearest Neighbors, and Recursive Partitioning were built. Three Random Forest models were built using different tuning parameters providing similar results. The Random Forest model with best Accuracy was selected. The zip codes with greater than median number of violations were selected as high risk zones for violations.

INTRODUCTION

The Fire Department of the City of New York (FDNY) is the largest Fire Department in the United States[1]. The New York City Fire Department is utilizing Machine Learning for predicting Fires. Firecast is an example of this approach[2]. Fire Prevention is using preventive measures to reduce the damage of fires. The Bureau of Fire Prevention FDNY is utilizing the preventive measures to minimize fire risks. The non-fire related inspections is a strategy used by the Fire Department to reduce the risk and minimize the damage to life and property. As per Mayor's Management Report, there were 209,410 complete inspections performed by FDNY Bureau of Fire Prevention[3] (civilian fire personnel) in FY2019. Prioritizing the inspections through machine learning will provide BFP an extra edge for future strategies.

BACKGROUND

FDNY conducts inspections that examine buildings, structures, facilities, vehicles and other locations in New York City. The inspections are conducted by FDNY's Bureau of Fire Prevention. The purpose of an inspection is to ensure that fire codes, laws, and regulations are enforced to eliminate the fire risk in order to keep New Yorkers safe. These inspections include examining the locations for regulations of equipment such as range hoods, explosives, fire suppression systems, fire alarms systems, sprinklers systems, bulk fuel, hazardous materials, chemicals (for labs/hospitals) and other equipment that either poses a fire hazard or are required by City Administrative code to prevent fires.

[1] https://www1.nyc.gov/site/fdny/about/overview/overview.page
[2] https://apolitical.co/en/solution_article/new-york-city-saving-lives-predicting-fires-will-break
[3] https://www1.nyc.gov/assets/operations/downloads/pdf/pmmr2020/fdny.pdf

Violation: An FDNY violation is an official notice that a property is not in compliance with the New York City Fire Code and/or Fire Department Rules. A Violation is issued in case of non-compliance. Major infractions can create dangerous situations and immediate action is required resulting in issuance of a criminal summons. The following types of violations are issued in New York City Link.

1. Notice of Violation
2. Violation Order
3. Criminal Summons
4. Vacate Order

Reinspections are performed on the premises that have been issued the violations. The frequency of inspections and reinspection vary depending on type of equipment. Once the violation is removed or there are no additional violations, the premise owner/account holder is issued a permit for a certain amount of time depending up on the type of inspection.

This deployment of resources appropriately is the most important part in any organization. If the resources are related to public safety, then it becomes more important to organize and utilize them efficiently. FDNY's BFP is equipped with Fire Inspectors for performing the inspections. These inspections also include specialized inspections like Bulk Fuel Safety, Explosives handling and many more. The inspections are performed daily. The research is an effort to organize and streamline the inspections process by tracking certain zip codes with a high amount of violations. We will try to predict the zip codes with a higher amount of violations in New York City. This will assist the BFP to plan the inspections process strategically to ensure compliance and public safety.

LITERATURE REVIEW

FIREBIRD

Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta

Firebird is a framework developed using machine learning algorithms, geocoding and information visualization. It is utilized by Atlanta Fire Rescue Department (AFRD) to improve AFRD's inspections process and to ensure public safety. It assists in resolving two of the major issues.

1. Identifying New Properties for Inspections – AFRD was using the existing paper-based inspections for 2,573 annually inspected commercial properties[4]. The information for properties was based on pre-existing permits which lacked the robust procedures for identification of new

[4] http://firebird.gatech.edu/KDD16_Firebird.pdf

commercial properties for inspections. With the data collection from a variety of data sources and with the manipulation of data appropriately, AFRD was able to identify 19,397 new commercial properties to inspect[5].

2. Fire Risk prediction – With the increase in workload due to these newly identified commercial properties to be inspected, AFRD's next challenge was to manage the resources appropriately. To overcome this issue, the machine learning model was developed to prioritize the inspections based on the fire risk.

Machine Learning models Logistic Regression, Support Vector machines, Gradient Boosting and Random Forest were built for the fire risk prediction.

PREDICTIVE MODELING OF BUILDING FIRE RISK

Designing and evaluating predictive models of fire risk to prioritize property fire inspections

Metro21 Smart Cities Initiative

The research study is a collective effort of Carnegie Mellon University and The City of Pittsburgh Bureau of Fire to build a model to predict fires and prioritize non- fire inspections to ensure fire prevention and public safety.

The study utilizes machine learning to develop predictive models for Fire Risk to the buildings. The output of this predictive model is a set of risk scores from 1-10. This risk factor assists in providing the likelihood of fire occurring to a property. These machine learning models predict the likelihood of a fire for a given address for a 6-month window, by "training" the model on the first 7.5 years of data on historical fire incidents, and evaluating the model on a test set which was not used for training[6]. This research also analyzes the fire risk scores to identify the high risk neighborhoods, fire districts, or property types. These high risk areas are targeted for prioritizing the inspections. Datasets from Fire Incidents, non-fire inspection violations and property assessments were combined using the addresses and four different machine learning models were created for the prediction. These models include Logistic Regression, Ada Boost, Random Forest and XG Boost. XGBoost Classification model was selected to predict the fires due to better performance of this model compared to other models created. The study used the Atlanta Firebird model as a benchmark.

HYPOTHESIS

[5] https://www.kdd.org/kdd2016/papers/files/adf0511-madaioA.pdf

[6] http://michaelmadaio.com/Metro21_FireRisk_FinalReport.pdf

This research is an effort to analyze the relationship of violation orders issued by FDNY with the zip codes. We hypothesize that there is strong correlation between the violation orders issued in a specific zip code. Based on this hypothesis, we will be building machine learning models to predict the zip codes of high intensity violations in future. Intuitively this theory is supported due to the fact that some zip codes with specific permits will have more violations as compared to others.

DATA AND VARIABLES

The data is collected from NYC Open Data Portal. The studies done on the topic are related to creating a risk matrix based on the fires incidents and non-fire inspections. We will solely focus on the non-fire inspections.

The datasets included are BFP Inspections and BFP Open Violations. The Open Violations data is a subset of the Inspections data. Once the inspection is performed on a premise, if the inspection status is not approved and has a violation order associated with it, that record from inspections data will be a part of Violations dataset. Although the dataset Inspections have more records as compared to Violations dataset, we will be focusing solely on the Violations.

Caveats We will utilize the Open Violations dataset. One drawback of using that data is that once the violation order is corrected, it will no longer be an open violation order and will not appear in the dataset. The number of violation orders in the dataset will be less due to released violation orders being dropped from the dataset. FDNY data for the violation orders is more robust and detail oriented. This research will be a prototype for the implementation of the inspection prioritization for BFP.

EXPLORATORY DATA ANALYSIS

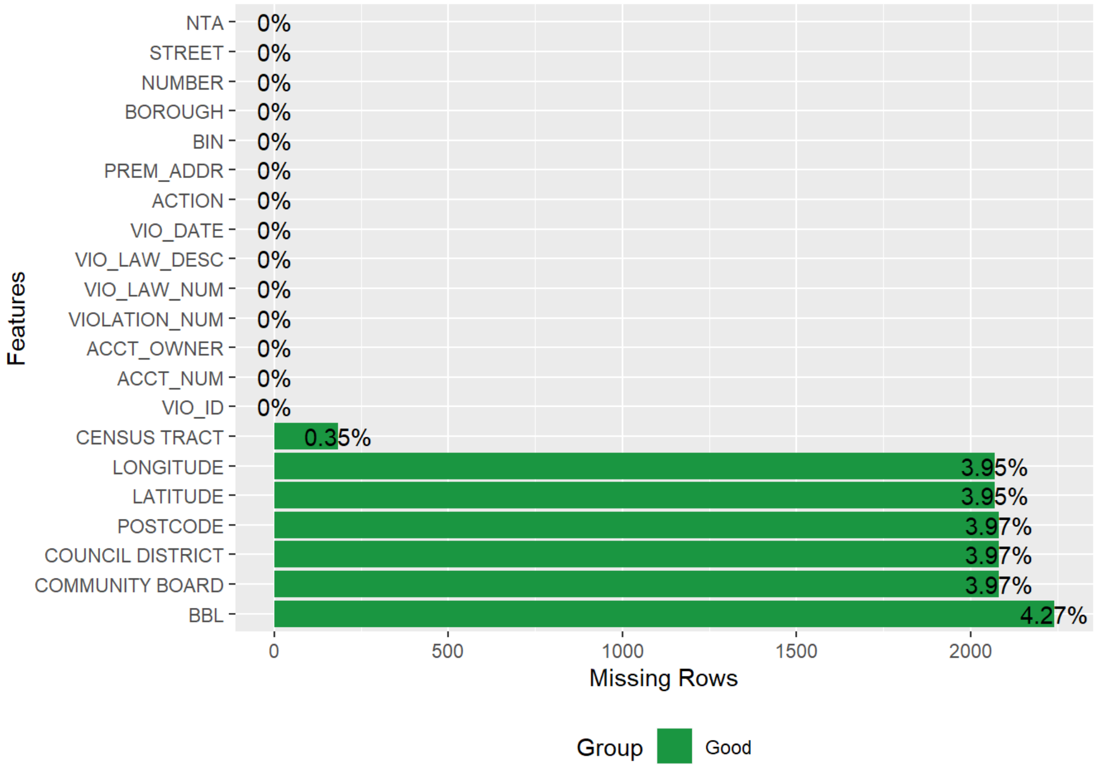
The data fields from the Violations data set will be Account Number, Violation Order Number, Violation Issue Date, Premise Address, Zip Code, Borough, Latitude and Longitude.

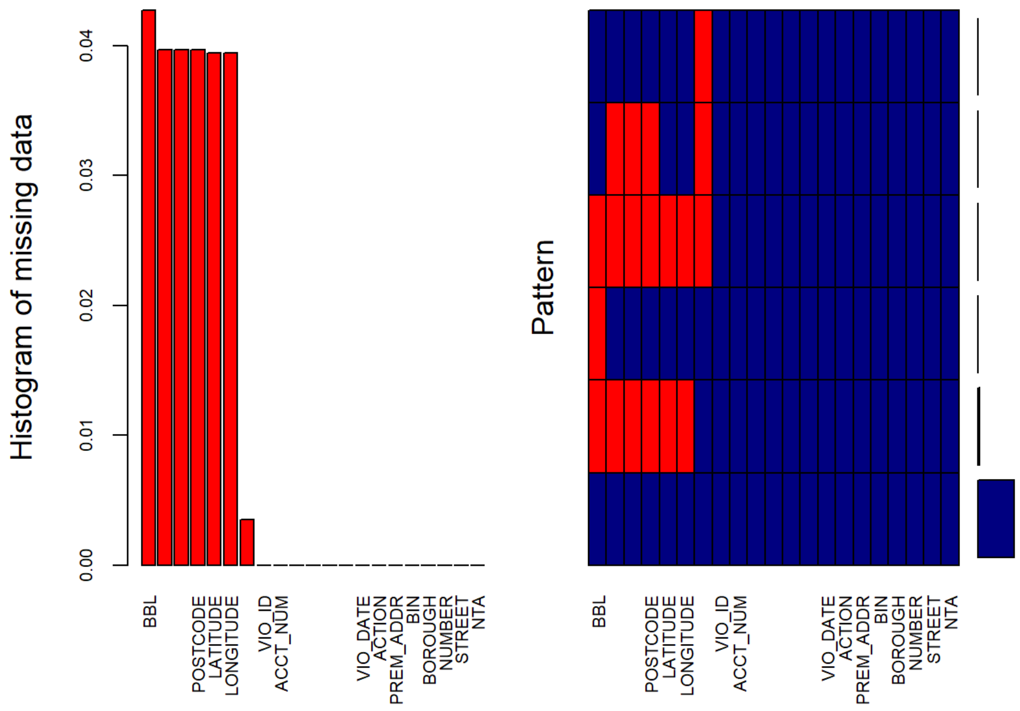
Variable Name	Description	Structure
ACCT_NUM	Account Number associated with permit	int
ACCT_OWNER	Account Owner	chr
VIOLATION_NUM	Violation Order Number	chr

VIO_LAW_NUM	Violation Law Number	chr
VIO_LAW_DESC	Violation Issue Date	chr
VIO_DATE	Violation Issued Date	Date
PREM_ADDR	Premise Address	chr
BIN	Building Information Number	int
LATITUDE	Latitude	num
LONGITUDE	Longitude	num
POSTCODE	Zip Code	int
BOROUGH	Borough	chr

Missing Data:

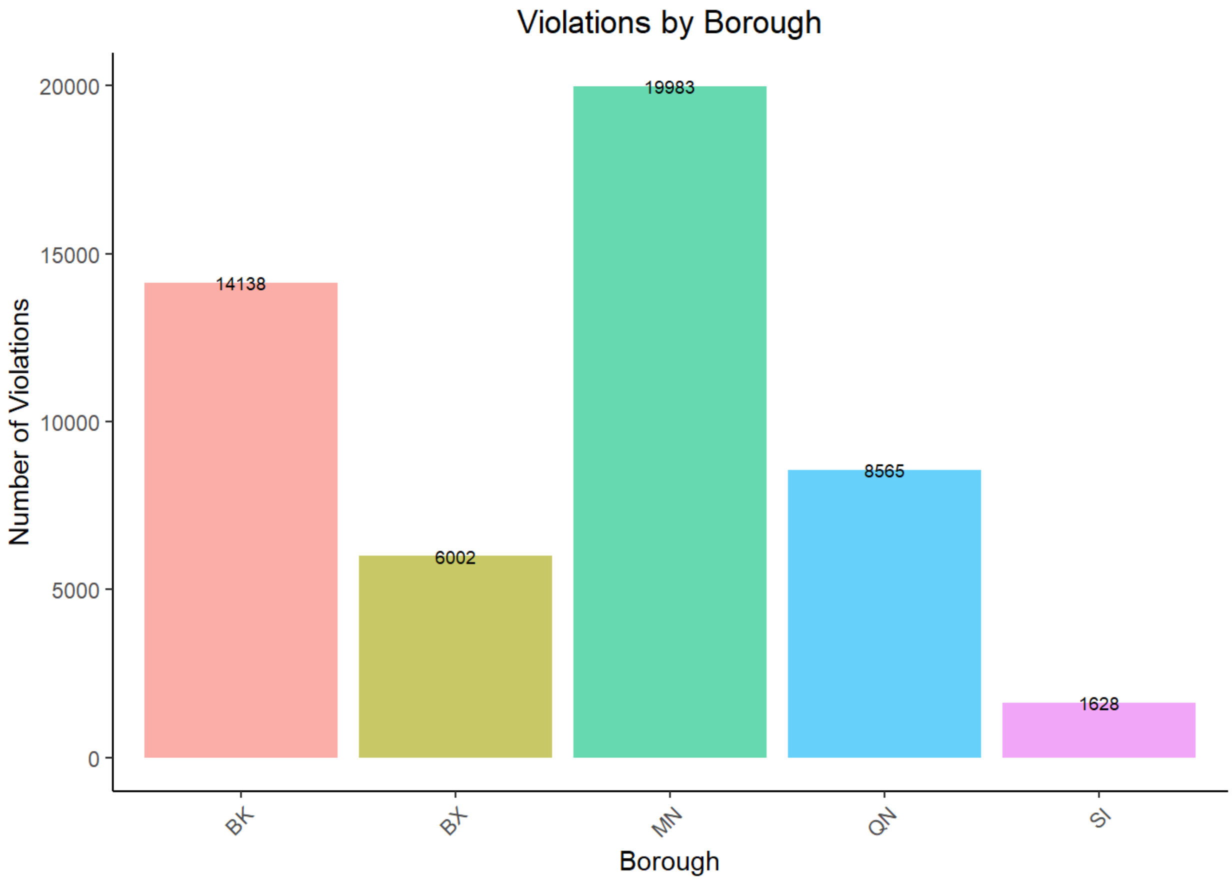
The missing data records were plotted and the percentage of missing data is below.

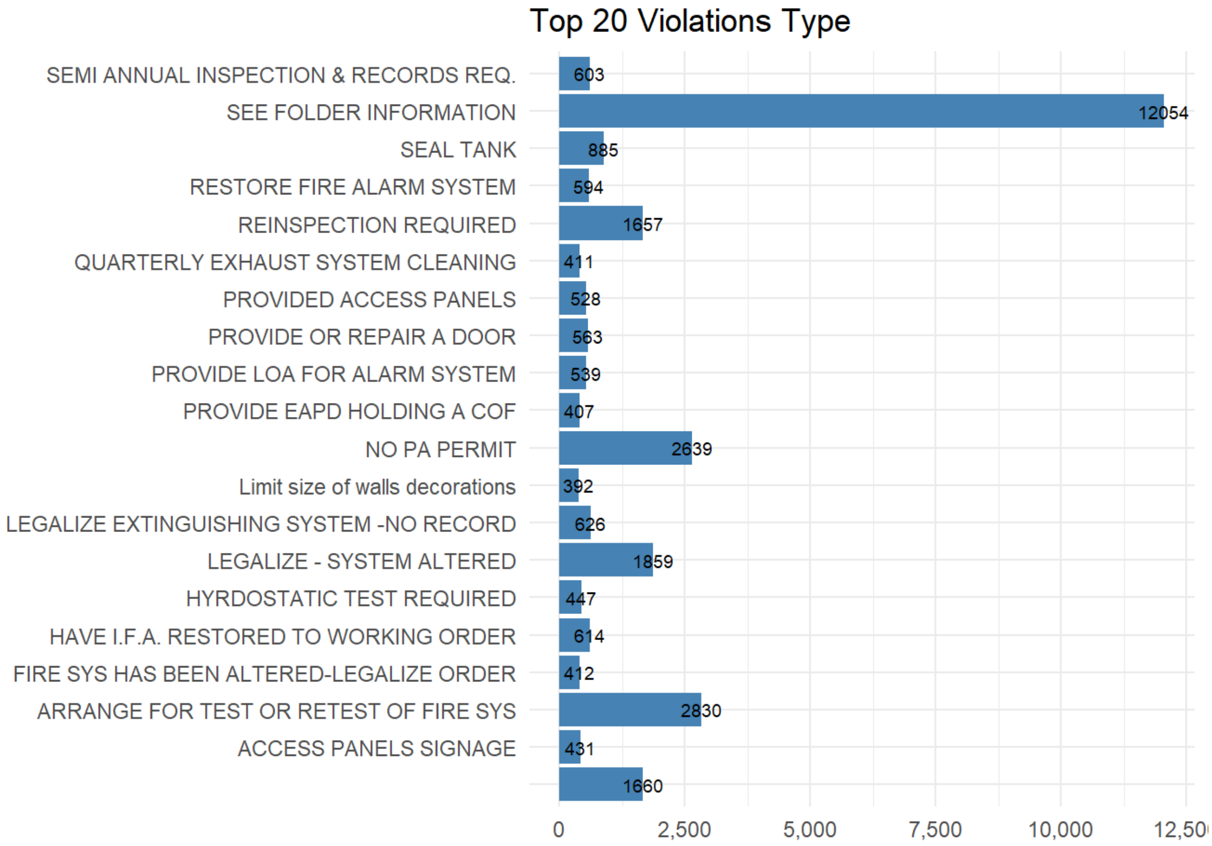


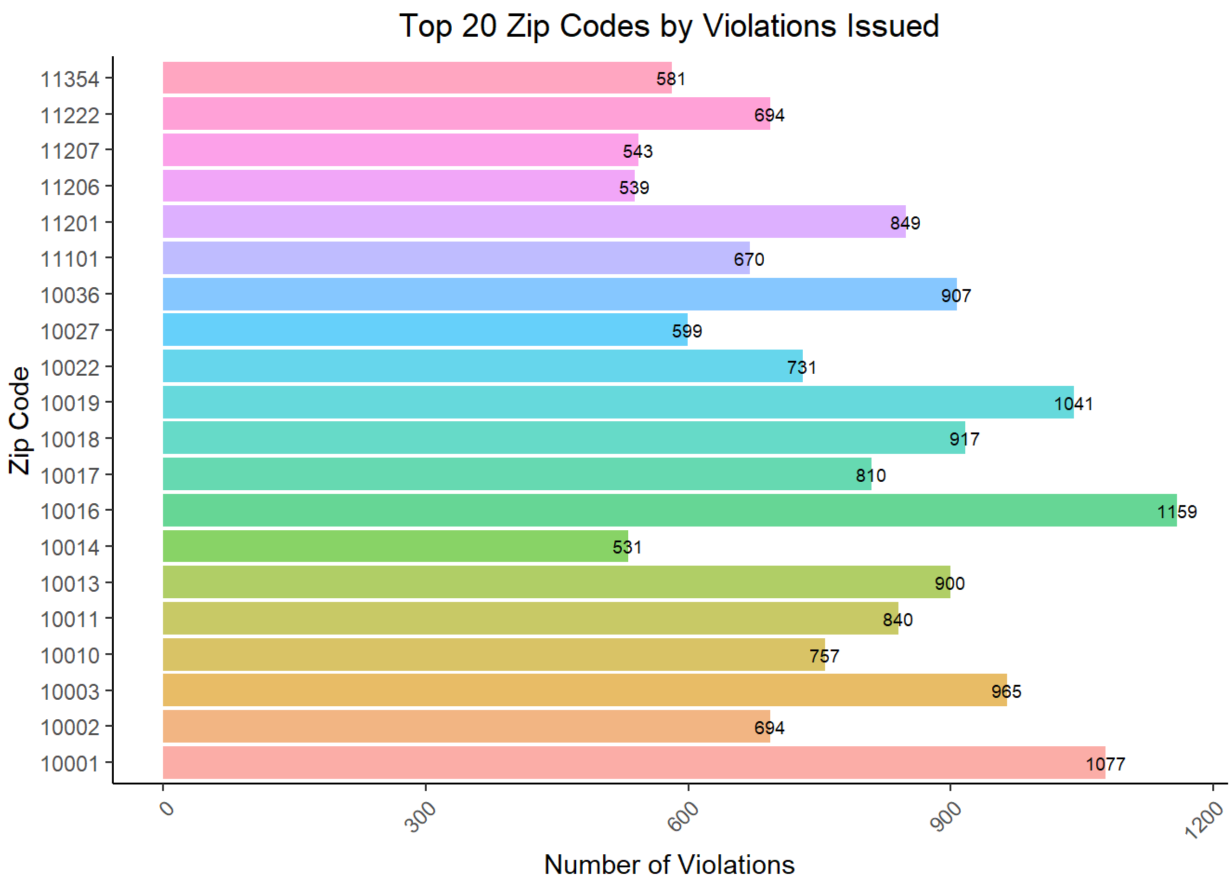


The open violations data was summarized by different categorical variables to get the summary of violations by the category. The categories include borough, postcode, violation law (code) and month.

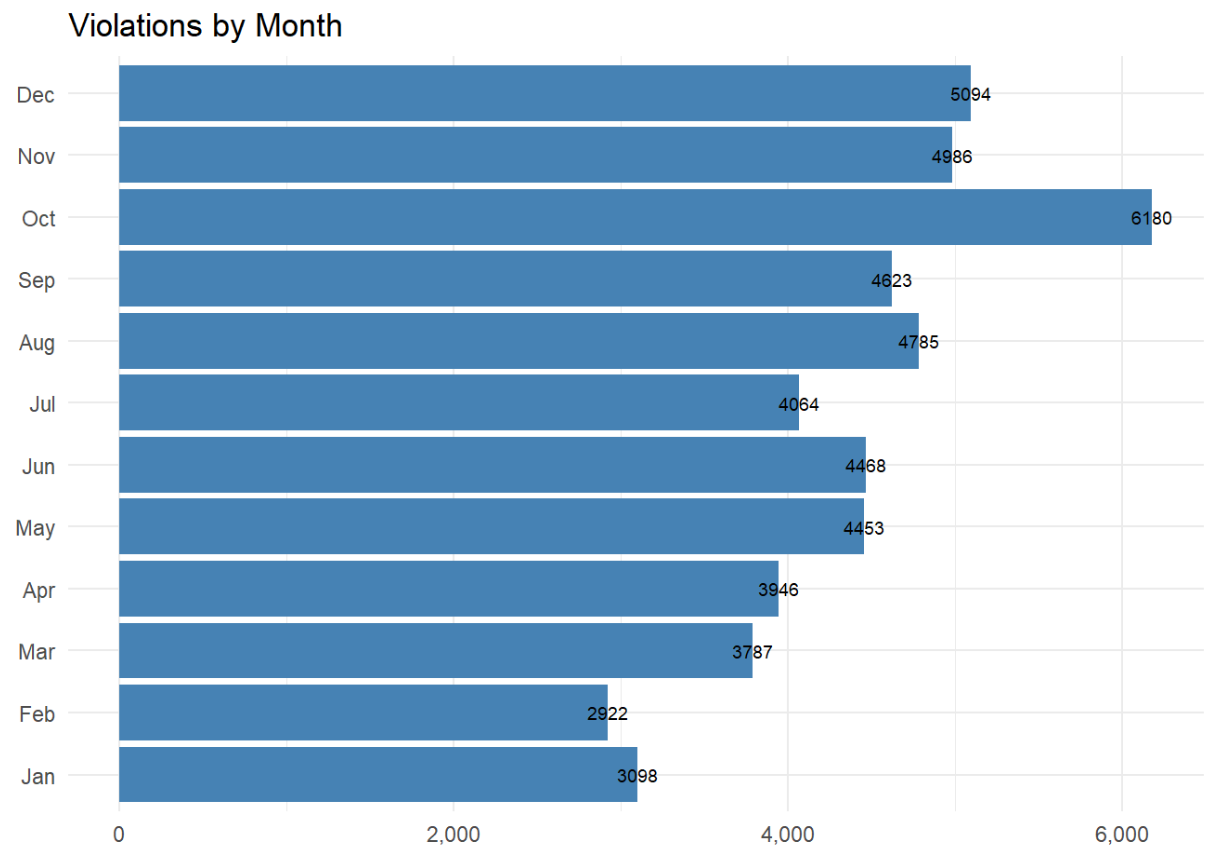
Most common breakdown for the data in New York City is by five boroughs. Manhattan has the most violations, followed by Brooklyn.




Top 20 Violations Type
SEMI ANNUAL INSPECTION & RECORDS REQ.
603
SEE FOLDER INFORMATION
12054
SEAL TANK
885
RESTORE FIRE ALARM SYSTEM
594
REINSPECTION REQUIRED
1657
QUARTERLY EXHAUST SYSTEM CLEANING
411
PROVIDED ACCESS PANELS
528
PROVIDE OR REPAIR A DOOR
563
PROVIDE LOA FOR ALARM SYSTEM
539
PROVIDE EAPD HOLDING A COF
407
NO PA PERMIT
2639
Limit size of walls decorations
392
LEGALIZE EXTINGUISHING SYSTEM -NO RECORD
626
LEGALIZE - SYSTEM ALTERED
1859
HYRDOSTATIC TEST REQUIRED
447
HAVE I.F.A. RESTORED TO WORKING ORDER
614
FIRE SYS HAS BEEN ALTERED-LEGALIZE ORDER
412
ARRANGE FOR TEST OR RETEST OF FIRE SYS
2830
ACCESS PANELS SIGNAGE
431
1660
0
2,500
5,000
7,500
10,000
12,50



Violations summary by month can also be an important factor as there might be some months with higher amounts of violations issued.

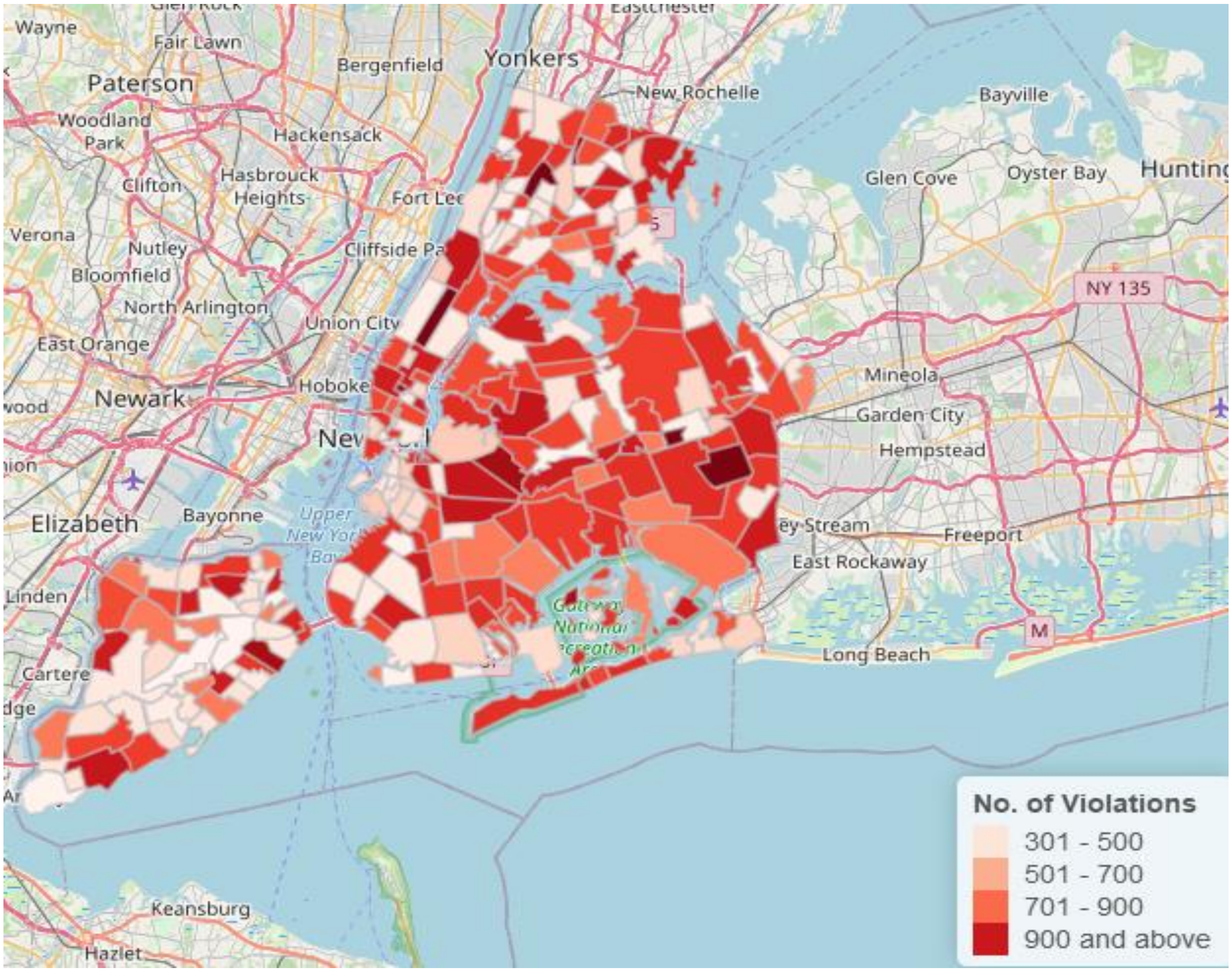


Splitting the Zip Codes into high, medium and low risk Violation Zones:

The violations dataset includes all the zip codes for all five (5) boroughs of New York City. We will use the quantile to split the zip codes into high, medium and low risk violations zones. Data is summarized by the zip codes and number of violations for each zip were calculated. Zip codes with a number of violations above the third quartile will be marked as a high violation zone[7]. The borough map below illustrates the zip codes above the third quartile in the high risk violations zone.[8]

[7] See Appendix A

[8] Leaflet Dynamic Map



DATA PREPARATION

Missing data: The missing data columns were eliminated from the analysis. Due to the fact the columns with missing data are postcode and location coordinates, we will not use imputation for the missing data treatment. Looking at the percentage of missing data, all the variables have less than 5% missing data which is considered as lower threshold[9].

[9] https://www.sciencedirect.com/science/article/pii/S0895435618308710

“Many Machine Learning Models cannot operate on label data directly. They require all input variables and output variables to be numeric”[10]. Although some decision tree models can take categorical variables directly for prediction, converting these into binaries will provide the flexibility to build different models. We will use one hot encoding for this conversion.
Dates (Lubridate and One Hot Encoding)[11]: Violation Dates were formatted with lubridate. A new column was created by extracting months from dates. The categorical variable month was then converted to binary using dummy_columns for months

Borough (One Hot Encoding)[12] : Using the above approach utilizing the function dummy_cols, we created dummy variables to convert categorical column borough to dummy variables.

STATISTICAL METHODS

BUILD MODELS

Predictive Modeling - We will build Supervised Classification Machine Learning algorithms for the prediction. K-Nearest Neighbors, Random Forests and Recursive Partitioning were built for the zip code prediction.
Caret Package was used for building all the models. Accuracy metric was used for the model selection. The models KNN and RPART were comparatively low in the Accuracy. Random Forest models were built and tuned by increasing the number of trees each time. The results (Accuracy) from all three models were similar and comparatively higher than KNN and RPART models. The data was split into 75% train and 25% test sets for all the models.

Model-1: K NEAREST NEIGHBOURS (KNN)

```
k   Accuracy   Kappa
 5  0.5397602  0.5354189
 7  0.5033211  0.4986067
 9  0.4749651  0.4699465
11  0.4513821  0.4461290
13  0.4348333  0.4294140
15  0.4198014  0.4142108
17  0.4033123  0.3975437
19  0.3921918  0.3862792
```

[10] https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/
[11] https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/
[12] https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

```
  21  0.3830299  0.3769973
  23  0.3731058  0.3669656
```

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.

The accuracy of the predictions on test data set was 59%

Model-2: RANDOM FOREST 1 (RF)

First Random Forest model was with ntrees = 100. No additional tuning was applied.

```
  mtry  Accuracy   Kappa
  2     0.9033042  0.9024009
  3     0.9168216  0.9160477
```

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 3.

The Accuracy of predictions on the test data set was 94%.

Model-3: RANDOM FOREST 2 (RF)

Second Random Forest model was with ntrees = 200. We used repeatedcv for the tuning. 10 fold cross-validation was used

```
 mtry  Accuracy   Kappa
  2     0.9050551  0.9041654
  3     0.9177288  0.9169599
```

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 3.

The Accuracy of predictions on the test data set was 94%.

Model-4: RANDOM FOREST 3 (RF)

Third Random Forest model was with ntrees = 500. We used repeatedcv for the tuning. 10 fold cross-validation was used

```
 mtry  Accuracy   Kappa
```

```
2     0.9070411  0.906163
3     0.9195207  0.918763
```

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 3

The Accuracy of predictions on the test data set was 94%.

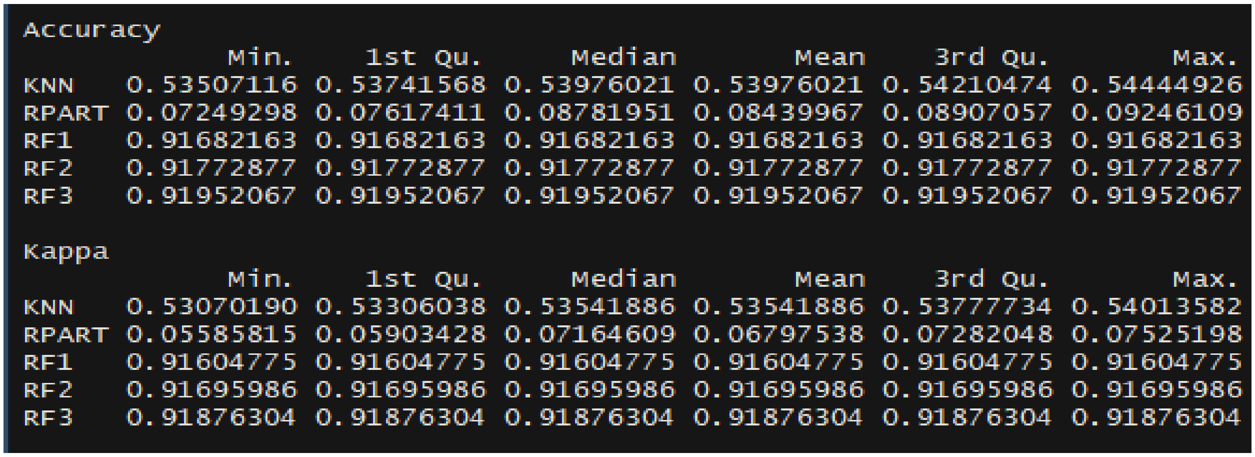
Model-5: RECURSIVE PARTITIONING (RPART)

```
cp              Accuracy       Kappa
0.01599329  0.08439967  0.06797538
0.01750873  0.05749032  0.03931038
0.01775228  0.04835305  0.02958221
```

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.01599329.
The Accuracy of predictions on the test data set was 7%.

DISCUSSION

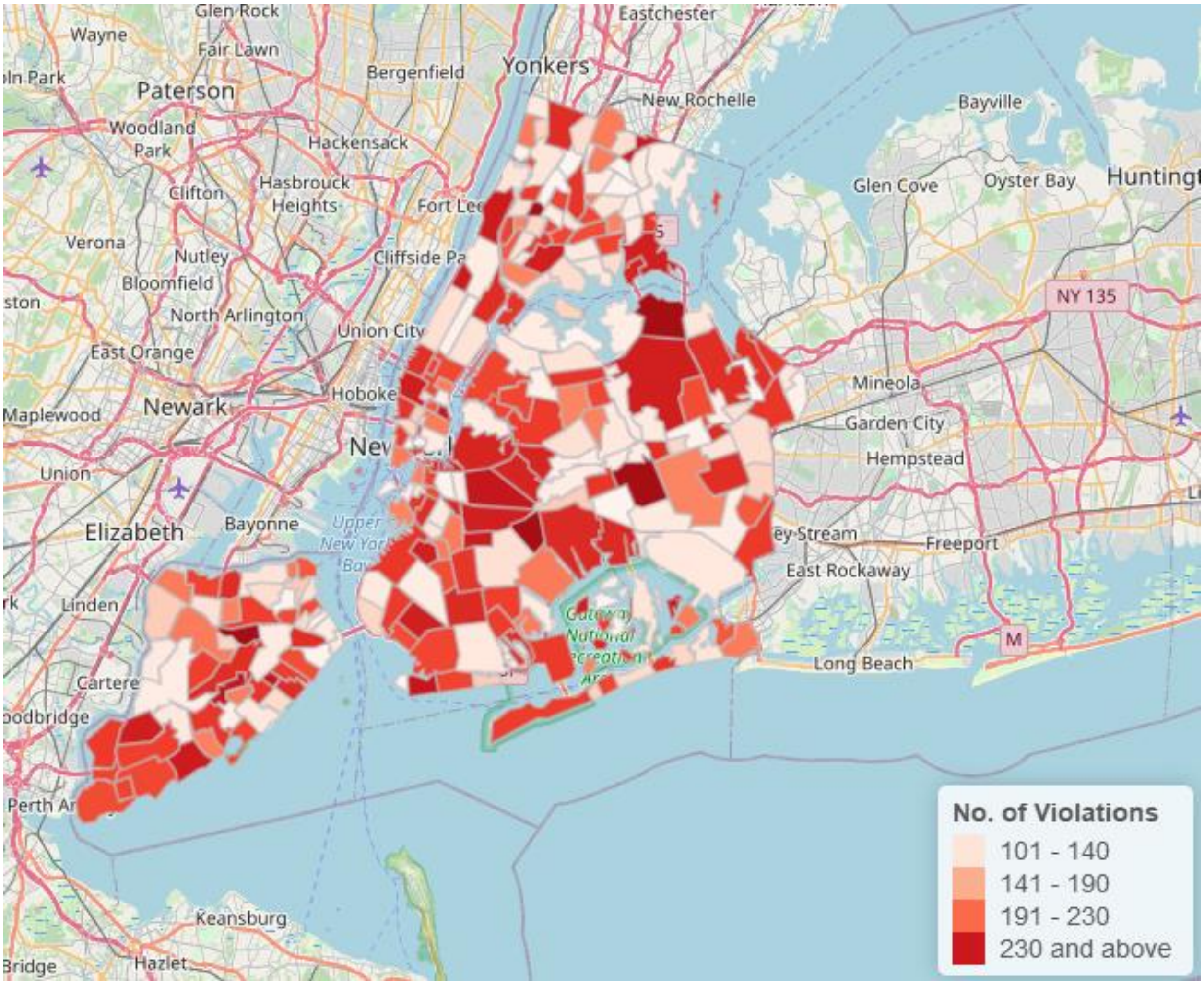
The Random Forest models created with different tuning parameters were similar and all three models were more accurate while comparing the Accuracy metric for all the models.



```
Accuracy
            Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
KNN   0.53507116 0.53741568 0.53976021 0.53976021 0.54210474 0.54444926
RPART 0.07249298 0.07617411 0.08781951 0.08439967 0.08907057 0.09246109
RF1   0.91682163 0.91682163 0.91682163 0.91682163 0.91682163 0.91682163
RF2   0.91772877 0.91772877 0.91772877 0.91772877 0.91772877 0.91772877
RF3   0.91952067 0.91952067 0.91952067 0.91952067 0.91952067 0.91952067

Kappa
            Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
KNN   0.53070190 0.53306038 0.53541886 0.53541886 0.53777734 0.54013582
RPART 0.05585815 0.05903428 0.07164609 0.06797538 0.07282048 0.07525198
RF1   0.91604775 0.91604775 0.91604775 0.91604775 0.91604775 0.91604775
RF2   0.91695986 0.91695986 0.91695986 0.91695986 0.91695986 0.91695986
RF3   0.91876304 0.91876304 0.91876304 0.91876304 0.91876304 0.91876304
```

Random Forest model 3 was selected as the final model with a test set Accuracy of 94%.
The results from this model were extracted and zip codes were summarized based on the number of violations issued in that specific zip code. The zip codes with greater above median Number of Violations were prone to violations and marked as high risk zones for violations.



The performance of the more complex Random Forest model and Classification Trees were better than that of basic K Nearest Neighbors and Recursive Partitioning models. Random Forest models performed better than on the evaluation set with an Accuracy of 94%.

Different Predictor variables were tried during the model building phase. Majority of the variables in the dataset were providing geographical information only. In addition, various dummy variables created during the data preparation were tested also during the models building phase. Due to the intensive processing time for predictive models, we built the models with few predictors.

CONCLUSION

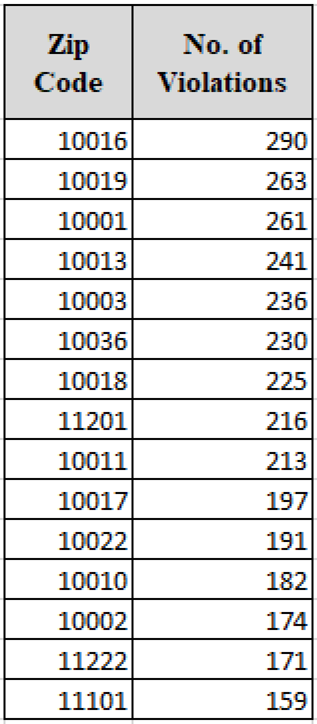
Future Expansion and modification of this research can result in more valuable insights. Availability of data sets and developing advanced models can cover existing gaps in the research.

1. Future research can include adding FDNY Fire Prevention Inspections Management System Data. The data will provide an opportunity to break the inspection and violations related with different inspection units and provide more insights about the violations.
2. Insights from this research can be enhanced effectively with the availability of the availability of data from across various city agencies and experimentation on various parameters to create risk factors for expansion of this research.
3. Further analysis in the area of violation issued dates can provide more insights. The time series forecasting models can provide insights with months with higher amounts of violations issued.
4. FDNY is already utilizing FIRECAST for predicting Fire Risk, however with this research being more focused on the Non Fire based Inspections and detailed research with internal inspections data can assist in deploying more Engine/Ladder companies in the areas of high risk.
5. BFP is in the developing phase of Active Violations Order dashboard, a plotly application built during CUNY Data 608 Final Project. Violations Prediction expansion will assist in additional functionality to the application[13].

APPENDIX

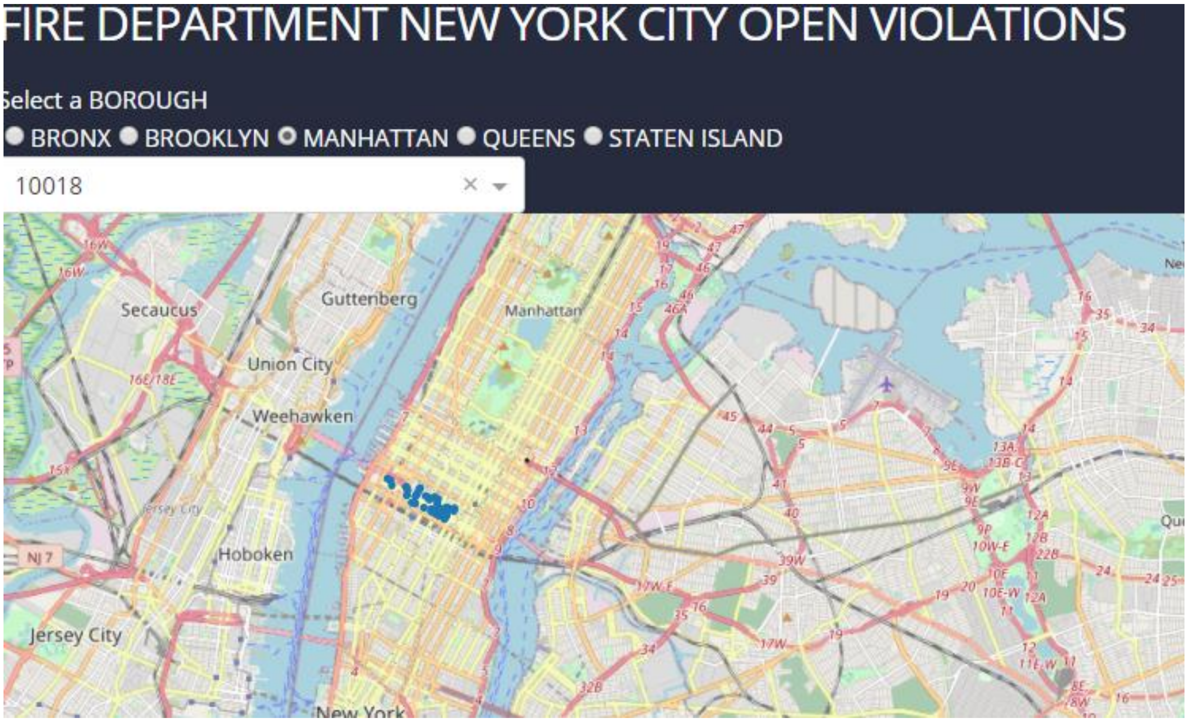
A. Prediction Results - Top 15 Zip Codes with Number of Violations Issued.

[13] See Appendix B and C



Zip Code	No. of Violations
10016	290
10019	263
10001	261
10013	241
10003	236
10036	230
10018	225
11201	216
10011	213
10017	197
10022	191
10010	182
10002	174
11222	171
11101	159

B.



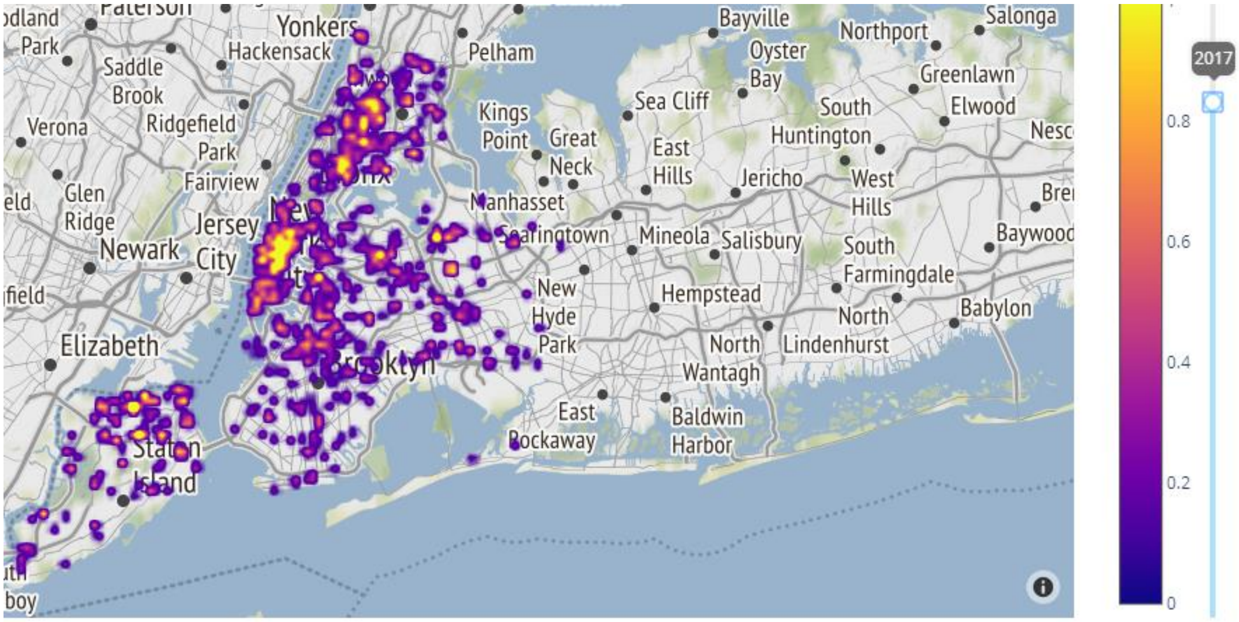
FIRE DEPARTMENT NEW YORK CITY OPEN VIOLATIONS

Select a BOROUGH

● BRONX ● BROOKLYN ○ MANHATTAN ● QUEENS ● STATEN ISLAND

10018 × ▾

C.



REFERENCES

https://www1.nyc.gov/site/fdny/about/overview/overview.page

https://apolitical.co/en/solution_article/new-york-city-saving-lives-predicting-fires-will-break

https://www1.nyc.gov/assets/operations/downloads/pdf/pmmr2020/fdny.pdf

http://firebird.gatech.edu/KDD16_Firebird.pdf

https://www.kdd.org/kdd2016/papers/files/adf0511-madaioA.pdf

http://michaelmadaio.com/Metro21_FireRisk_FinalReport.pdf

https://www.dnainfo.com/new-york/20110201/manhattan/tribeca-has-countrys-highest-number-of-restaurants-per-person/

https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c

https://www.statmethods.net/advstats/cart.html https://datascienceplus.com/random-forests-in-r/

http://rstudio-pubs-static.s3.amazonaws.com/309009_b057419cae50467a953c16b100f1a7ac.html

https://rpubs.com/jhofman/nycmaps

https://www.knime.com/knime-applications/forest-fire-prediction