Boston Crimes: Final Report

Aayushi Gupta
Atharva Shantanu Kulkarni
Dhwani Patel

College of Professional Studies, Northeastern University

ALY 6040: Data Mining Applications

Mr. Joe Reilly

October 25, 2022

Business Problem

Crimes seem to be everywhere in the United States including Boston. Due to the alarming rate at which crimes are increasing, crime continues to be among the worst problems in the US. In comparison to the same time period in 2017, the Major Cities Chiefs Association study states that overall violent crime increased 4.2% from January 1 to June 30, 2022 (Contreras, 2022). While homicides and rapes have significantly decreased since previous year, the number of robberies and violent assaults has increased by 4.2% between January 2022 and June 2022.

Because traditional methods of solving crime require a lot of work and are ineffectual, they are ineffective in the contemporary environment of fast escalating crime. Therefore, finding techniques to effectively predict crime before it occurs or creating a "machine" to assist police personnel will relieve pressure on the force and promote crime prevention (Shah, 2021).

The goal of the project was to take a deep dive into the factors which might be affecting the type and frequency of crimes in the Boston region. While the area of focus was Boston here, the long-term goal is to make a scalable model which can be extended to multiple areas.

After importing and processing the data, models were constructed which can learn the relationships between the predictors and the type of crime from a training set and then evaluate performance on unseen data.

In the process of trying to come up with models and the approach, we figured that there were a number of ways we could go about it. We could either focus on just the timestamps and build a time series model which would be focused on the frequency of crimes and their seasonality. Then we could group the areas geographically and make clusters with the prominent crimes in that area or we could treat every data point individually and use timestamps and location coordinates as separate features which is the approach we decided to move forward with. Reason behind this was that we wanted to build something scalable, and which can be extrapolated into future and to different areas which would have been challenging with either time series or clustering model. We also want to incorporate more variables later on as therefore we decided to go with the third approach.

Data Description

The Boston Police Department (BPD) provides crime event reports to record the preliminary information around an occurrence to which BPD officers respond. This dataset contains data from the new crime event report system, which has fewer fields designed to capture the incident's type as well as its timing and location (Crime Incident Reports (August 2015 - to Date) (Source: New System) - Analyze Boston, n.d.). Records contain data from the year 2015-2018.

The data source is Analyze Boston. The dataset has 353253 observations recorded over the period of almost 4 years from 2015-2018. The number of variables present are 17 and have mixed datatypes- int, object and float. A few columns of interest are offense description, time and day of crime, district and latitude longitude. The target variable for this data would be 'OFFENSE_CODE_GROUP' which has values such as larceny, property damage, motor vehicle accident response etc. One more column called 'UCR_PART' is taken in conjunction as the target variable which has four unique values- Part One, Part Two, Part Three and Other. Part I Offenses are ten serious crimes that occur on a regular basis and are likely to be reported to law enforcement and include murder, rape, robbery, assault, motor vehicle theft, burglary etc. Part I Offenses are generally referred to as the "Crime Index" measurement. Part II Offenses represent "less serious" crime classifications. Part I and Part II crimes are defined by the FBI Uniform Crime Reporting

Program. Part Three and Other are not defined and represent the least serious crimes. (*Custom404 • Roseville • CivicEngage, n.d.)*

We initially began with Crime dataset for Atlanta but later on switched to Boston crime data as that dataset had more variables which could be taken into account for our modeling. We began working on the Boston crime data from the years 2019-2022 but the biggest issue we ran into was that the column that we were targeting was entirely null for the chosen years. Therefore, for research purposes we moved our timeframe to 2015 up to 2018.

Bias in the data: The records that we have are the ones that have only been updated by the public officials and does not include crimes which go unreported or untracked. One major drawback is places equipped with better laws and rules are likely to have a higher percentage of crimes which have been logged. This makes it difficult to analyze the correlation between the law enforcement of that area and the crimes. Other confidential and unstructured data sources like CCTV footages might be helpful to get better quality data to make informed decisions.

Data Preprocessing

To clean the data, it is important to look at the structure of data including its columns and their data types. *(Refer Appendix 1 Fig 1 and Fig 2)*. Using the functions describe(), head() and dtypes(), it was made clear that there are some columns which not be required. The column 'SHOOTING' was almost completely null, so we dropped that column. We also find missing values which were necessary to be dealt with before building our model. We first dropped the rows with missing values in STREET and UCR_PART. There were a lot of missing values (approximately 6%) in the 'Lat' and 'Long' column. So, dropping them would not have provided the best results and also would have resulted in loss of data. So, we substituted the null values in Lat and Long with the mean of the respective columns. This project is purely for academic purposes and in the real world it would not make a lot of sense to replace the 'lat' and 'long' values with the mean. If we had more time, we would try to figure out to fill the NA values in lat and long using some other location-based column like the street or district. Due to time constraint and pure academic purposes we stuck with the fill with mean method.

On checking for duplication, we found out that the data had 628 duplicated rows. So, we dropped the duplicates and kept the first occurrence of every duplicate as redundant data would cause bias. It is essential to categorize the crimes and generate the labels that I want to predict. The Offense groups and the UCR Part provided the information essential to do that. The rows with UCR_PART were labelled as '1' for violent crime and the others were labelled as '0' for nonviolent crimes. Since most of the models cannot work with data of datetime type, we decided to use day, month and year as separate features in the model. The features month and year were already present, so we just had to extract the day of the month from the date column.

Before deciding on the data mining techniques, we wanted to go one step further and try integrating the weather data as another independent variable in the model. The reason behind this was that weather also plays a big role in people's planning of the day in general. If the weather is too hot or there is a snowstorm, there is a high chance that people might not leave their homes. We wanted to consider this in the model to accurately predict crime and take into account the weather changes as well.

This was done using the wwo-hist package provided by Python. The package was first installed and then the required parameters were specified. The start date was given as Jan 1, 2015 and end date as Dec 31, 2018. Since we were fetching weather data by the day, the frequency was set as 24. By using the retrieve function, the historical weather data was fetched and had quite a few columns of interest (*Refer Appendix 1 Fig 10)* For this model we decided to use temperature, wind speed, cloud cover and visibility columns.

Before merging the weather data and the crime data into one dataframe, the primary task was to decide on a column on which the join is going to be performed. We did run into a few errors while performing the join using the date column. We later realized that the datatype and formatting of the date column in the dataframes was different. So, we converted both the date columns to a datetime datatype. The merge function could then be performed smoothly.

Exploratory Data Analysis

Data visualization is an important step in any data mining technique By providing it with a visual context via maps or graphs, data visualization helps us understand what the data means. As a result, it is simpler to spot trends, patterns, and outliers in enormous data sets since the data is easier for the human mind to understand.

Since we were trying to predict an occurrence of a violent crime, the first question we tried to answer was- What type of crimes are the most common in Boston? For this, we used the offense code group column and by using the unique_values function, obtained the top 10 most common crimes. The list was topped by Motor Vehicle Accident Response followed by Larceny and Medical Assistance *(Refer Appendix 1 Fig 3).*

The next visualization was created for explore the crimes with respect to the time variable. The number of crimes were analyzed in different months, seasons, time of the day and the day of the week. On digging deeper, it became clear that the mist crimes occur in evenings in the months of July-October. The difference between number of crimes on different days of the week was not very distinctive. *(Refer Appendix 1 Fig 4-7)*

Last and the most important question was where are crimes being committed the most? The number of crimes were plotted against the different districts, and it was found that B2, C11, D4, B3 and A1 were the most dangerous places. To take a closer look, I utilized the folium maps to plot the regions on a geographical map. *(Refer Appendix 1 Fig 8 and Fig 9)*

To proceed with the modeling, it is essential to categorize the crimes and generate the labels that I want to predict. The Offense groups and the UCR Part provided the information essential to do that. The rows with UCR_PART were labelled as '1' for violent crime and the others were labelled as '0' for nonviolent crimes.

Data Modeling

The three models that we chose for our dataset were Decision Tree, KNN and Random Forest. Each member of our team worked on one model individually and tried to optimize it in the best way possible. The project's aim is to investigate the relationships in different types of crimes and the corresponding features. Therefore, extracting the variables of the greatest importancefor the prediction is of the utmost importance. The features that we selected based on numbers and the business logic were 'Lat', 'Long', 'cloudcover' 'precipitation', 'windspeed', 'visibility', 'day', 'month' and the 'year'. A brief description about the models, reasons behind selecting them and their performance can be seen below-

1. **Decision Tree:** The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label. *(GeeksforGeeks, 2022*). It is capable of obtaining great accuracy in a wide range of tasks and is also quite intuitive. Decision trees are unique in the world of ML models because of how clearly they express

information. A hierarchical structure is dynamically formulated from the "knowledge" that a decision tree gains via training. So non-experts may readily understand the information because of the manner it is held and presented in this framework. (Seif, 2022)

We chose this particular model because-:

A. Ability to using different feature subsets and decision rules at different stages of classification (Sun & Du, 2008).
B. The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree. (1.10. Decision Trees, n.d.)
C. They can handle both numerical and categorical data
D. Uses a white-box model. (instead of being totally opaque and acting like a black box, it provides some transparency to help the person understand why a particular datapoint is being categorized into a particular category)
E. Model validation is possible using statistical tests

Performance: The model achieved an accuracy of 69.26%. After the tuning of hyperparameters (same as Random forest) , the accuracy was increased to 71.23%.

2. **KNN:** K-Nearest neighbors, or KNN is a supervised classifier which as the tile entails uses the concept of proximity which in turn relates to neighbors, for classification or prediction of certain features fall near to a data point, this is widely used for classification based on the concept that similar features can be located in close proximity of one another. We chose this model because-:

   A. The calculation time is quick
   B. Fairly simple algorithm to interpret
   C. There is no training period and hence it is called a lazy learner.

D. Does not require training before predictions and hance the data can be easily added.

E. Easy to implement as there are only two parameters to choose- K and the distance metric

Performance: The model achieved an accuracy of 77.47% on the unseen data using the default values defined by scikit-learn as follow-:

```
{'algorithm': 'auto',
 'leaf_size': 30,
 'metric': 'minkowski',
 'metric_params': None,
 'n_jobs': None,
 'n_neighbors': 5,
 'p': 2,
 'weights': 'uniform'}
```

After tuning the hyperparameters, the accuracy on the test was brought up to 80.48%. The hyperparameters used were as follows-:

```
Best leaf_size: 1
Best p: 1
Best n_neighbors: 25
```

3. **Random Forest:** Like its name suggests, a random forest is made up of numerous independent decision trees that work together as an ensemble. Every individual tree in the random forest spits out a class forecast, and the classification that receives the most votes becomes the prediction made by our model. The key is the low correlation between models. Uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don’t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction (Yiu, 2021).

Reasons behind selecting this model:

A. By utilizing many trees, the random forest method avoids and prevents overfitting.

B. Results are known to be more accurate

C. Efficiently handle large datasets

Performance: The model when run with the default parameters gave an accuracy of 78.62% on the testing set. The values of the default parameters were checked and they were as below-

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

Tuning was performed on 10% of the dataset due to computational restrictions and these parameters were found to be the best-:

```
{'n_estimators': 1342,
 'min_samples_split': 30,
 'min_samples_leaf': 1,
 'max_depth': 100,
 'criterion': 'entropy',
 'bootstrap': True}
```

The model was run again using these parameters and the accuracy was increased to 80.63%.

Comparison:

Model	Accuracy	Precision	Recall
Decision Tree	71.23%	0.22	0.23
KNN	80.48%	0.32	0.01
Random Forest	80.63%	0.53	0.01

Conclusion, Recommendations and Next Steps

To conclude, we feel that some of the results that have been achieved are promising and further evaluation would be beneficial in yielding significant improvements. The best performing model on this dataset is the Random Forest Model with an accuracy of 80.63%. Hence, our recommendation would be to use to use this model after required accuracy is achieved. We would also recommend using some other complex but efficient models like XGBoost classifier to further improve performance. As a binary classification problem, the dataset that has been presented is highly unbalanced with only 19% of the data point belonging to the 'Violent' category and rest to the 'Non-Violent category'. To deal with this bias, data augmentation can be done using the SMOTE technique (Synthetic Minority Over Sampling Technique) which creates artificial samples for the minority class and then run the models. The other approach which can be followed is to perform a multiclass clustering on the Offense Code Group column and classify crimes into Road Accidents, Larceny, Drug Violation, Vandalism etc.

A next step in the future would also be to integrate more features into the dataset including street lighting, proximity to a police station, median household income of the residents in that area, race of the people living in the area, and the poverty rate. This would give us more features to examine, better predictions and would also be very helpful in making informed decisions to help prevent crime. As mentioned before, the bias with only tracked crimes being taken into consideration needs to be dealt with by incorporating a way to log untracked and unreported offenses. After all these steps have been performed, the objective would be to integrate this model into a software which would predict the crime type and probability when the values of the variables are given in any area and not just Boston.

References

1.10. Decision Trees. (n.d.). Scikit-learn. Retrieved October 22, 2022, from https://scikit-learn.org/stable/modules/tree.html

Contreras, R. (2022, September 10). *Survey: Homicides down midyear as overall violent crime jumps*. Axios. Retrieved October 4, 2022, from https://www.axios.com/2022/09/10/homicides-down-midyear-overall-violent-crime-up

Crime Incident Reports (August 2015 - To Date) (Source: New System) - Analyze Boston. (n.d.). Retrieved October 4, 2022, from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

Custom404 • Roseville • CivicEngage. (n.d.). Retrieved October 10, 2022, from https://www.cityofroseville.com/404.aspx?aspxerrorpath=/DocumentCenter/View/26568/Description-of-Uniform-Crime-Offenses

GeeksforGeeks. (2022, October 4). *Decision Tree*. Retrieved October 18, 2022, from https://www.geeksforgeeks.org/decision-tree/

Seif, G. (2022, February 11). *A Guide to Decision Trees for Machine Learning and Data Science*. Medium. Retrieved October 22, 2022, from https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956

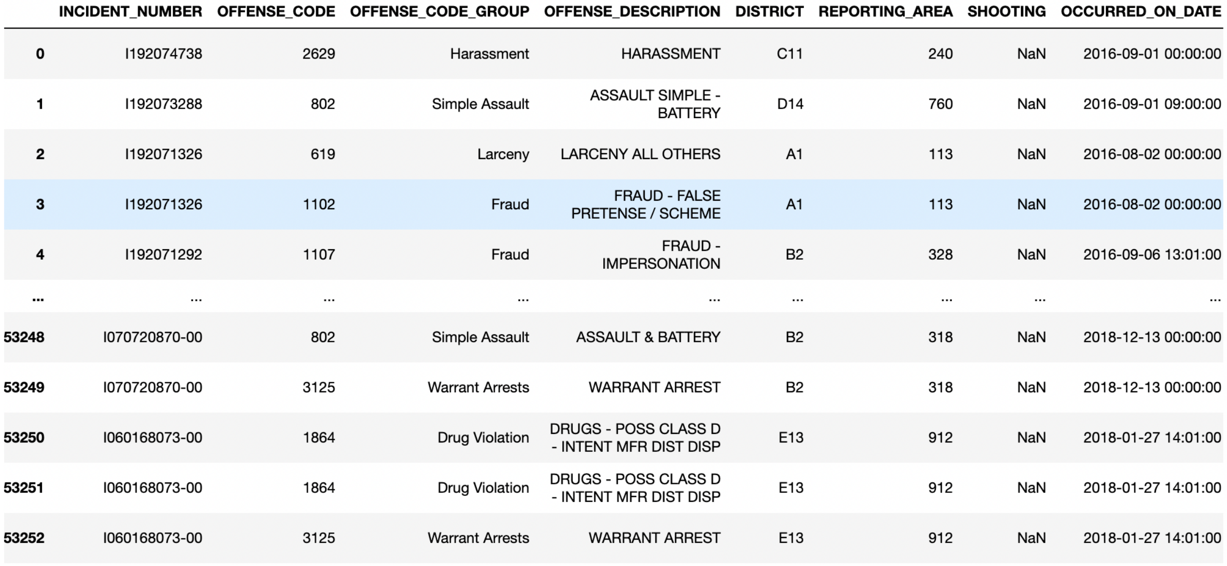
Shah, N. (2021, April 29). *Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention - Visual Computing for Industry, Biomedicine, and Art*. SpringerOpen. Retrieved October 4, 2022, from https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z

Sun, D. W., & Du, C.-J. (2008). Computer Vision Technology for Food Quality Evaluation. *Science Direct*. https://doi.org/10.1016/b978-0-12-373642-0.x5001-7

Uzel, A. (2021, December 13). *Demand Forecasting: Boston Crime Data - Towards Data Science*. Medium. Retrieved October 4, 2022, from https://towardsdatascience.com/demand-forecast-boston-crime-data-64a0cff54820

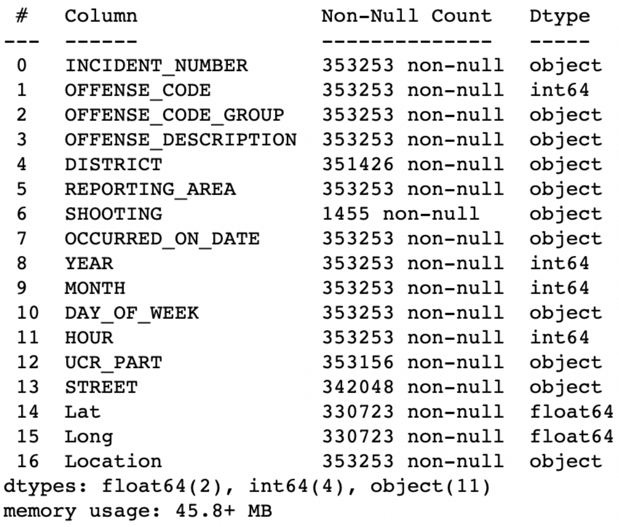
Yiu, T. (2021, December 10). *Understanding Random Forest - Towards Data Science*. Medium. Retrieved October 10, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Appendix 1



	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE
0	I192074738	2629	Harassment	HARASSMENT	C11	240	NaN	2016-09-01 00:00:00
1	I192073288	802	Simple Assault	ASSAULT SIMPLE - BATTERY	D14	760	NaN	2016-09-01 09:00:00
2	I192071326	619	Larceny	LARCENY ALL OTHERS	A1	113	NaN	2016-08-02 00:00:00
3	I192071326	1102	Fraud	FRAUD - FALSE PRETENSE / SCHEME	A1	113	NaN	2016-08-02 00:00:00
4	I192071292	1107	Fraud	FRAUD - IMPERSONATION	B2	328	NaN	2016-09-06 13:01:00
...	...	...	...	...	...	...	...	...
53248	I070720870-00	802	Simple Assault	ASSAULT & BATTERY	B2	318	NaN	2018-12-13 00:00:00
53249	I070720870-00	3125	Warrant Arrests	WARRANT ARREST	B2	318	NaN	2018-12-13 00:00:00
53250	I060168073-00	1864	Drug Violation	DRUGS - POSS CLASS D - INTENT MFR DIST DISP	E13	912	NaN	2018-01-27 14:01:00
53251	I060168073-00	1864	Drug Violation	DRUGS - POSS CLASS D - INTENT MFR DIST DISP	E13	912	NaN	2018-01-27 14:01:00
53252	I060168073-00	3125	Warrant Arrests	WARRANT ARREST	E13	912	NaN	2018-01-27 14:01:00

Fig 1: Dataframe



```
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   INCIDENT_NUMBER      353253 non-null  object
 1   OFFENSE_CODE         353253 non-null  int64
 2   OFFENSE_CODE_GROUP   353253 non-null  object
 3   OFFENSE_DESCRIPTION  353253 non-null  object
 4   DISTRICT             351426 non-null  object
 5   REPORTING_AREA       353253 non-null  object
 6   SHOOTING             1455 non-null    object
 7   OCCURRED_ON_DATE     353253 non-null  object
 8   YEAR                 353253 non-null  int64
 9   MONTH                353253 non-null  int64
 10  DAY_OF_WEEK          353253 non-null  object
 11  HOUR                 353253 non-null  int64
 12  UCR_PART             353156 non-null  object
 13  STREET               342048 non-null  object
 14  Lat                  330723 non-null  float64
 15  Long                 330723 non-null  float64
 16  Location             353253 non-null  object
dtypes: float64(2), int64(4), object(11)
memory usage: 45.8+ MB
```

Fig 2: Columns and their datatypes

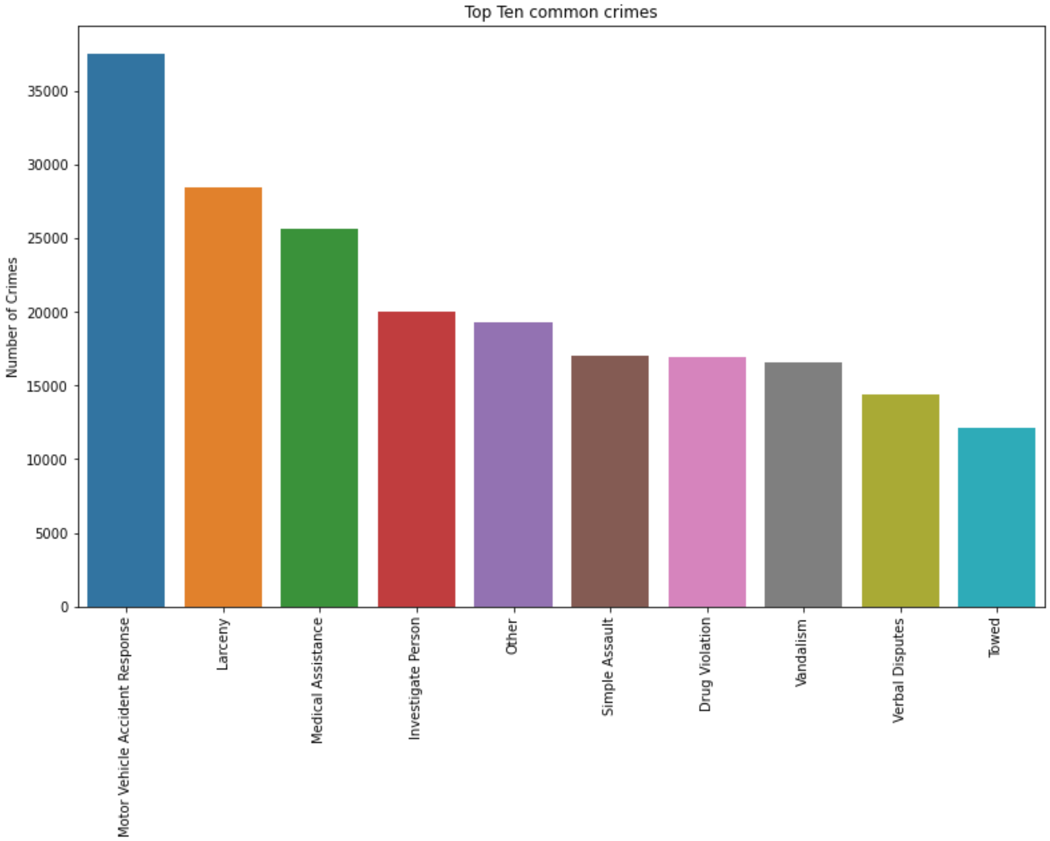


Fig 3: Top 10 Most Common Crimes

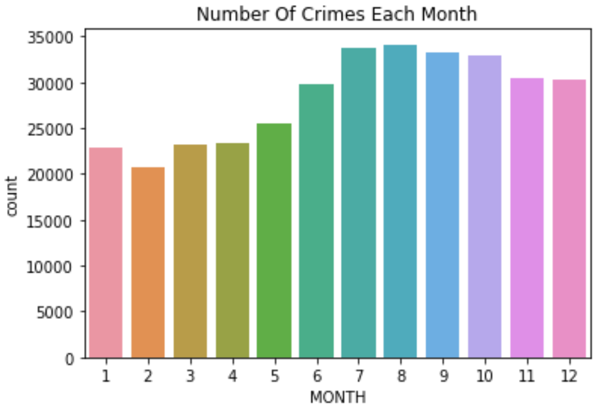


Fig 4: Crimes during different months

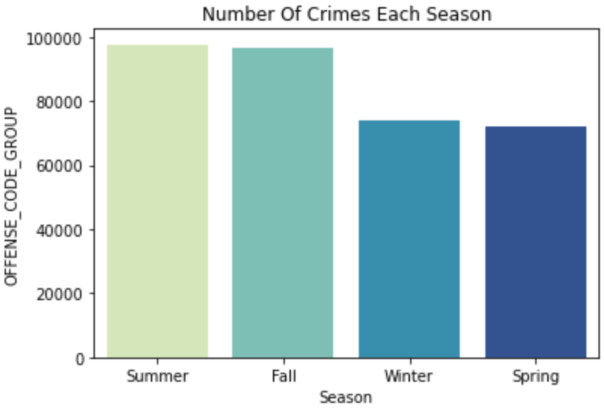


Fig 5: Crimes during different seasons

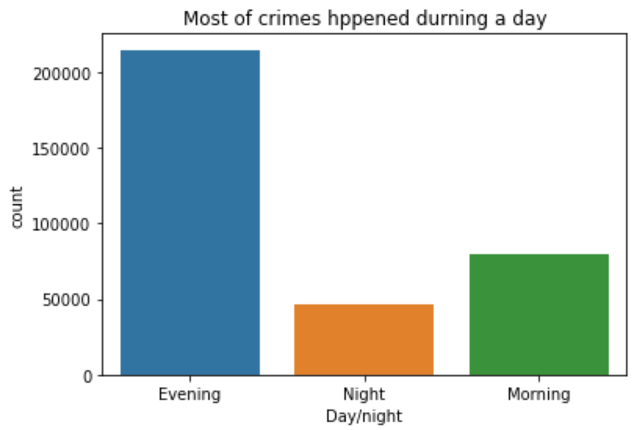


Fig 6: Crimes during different time of the day

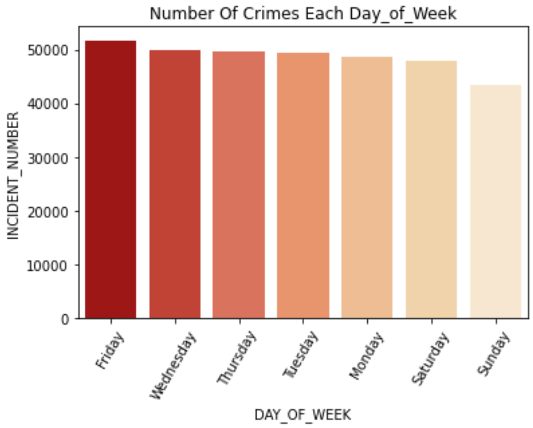


Fig 7: Crimes during different day of week

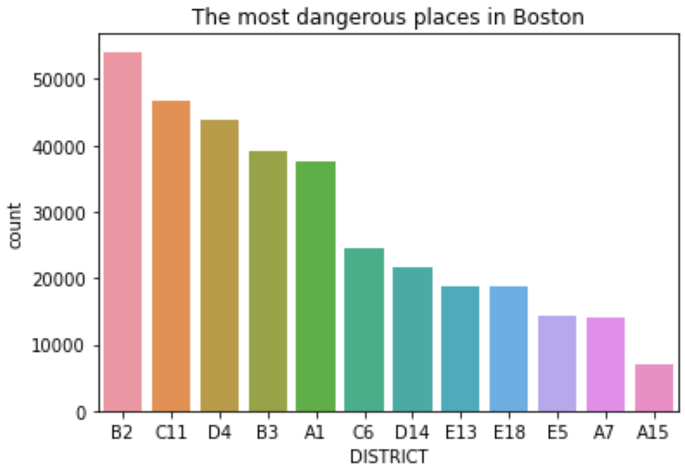


Fig 8: Crimes in different district

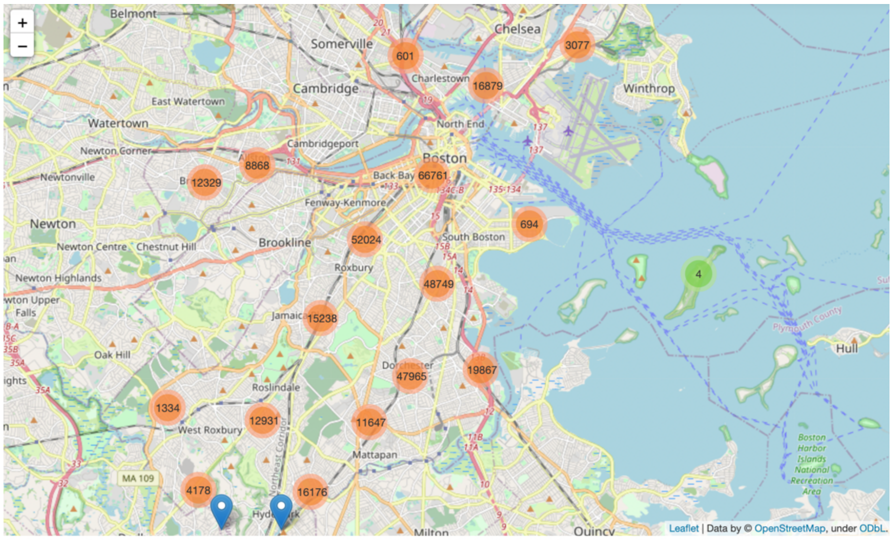
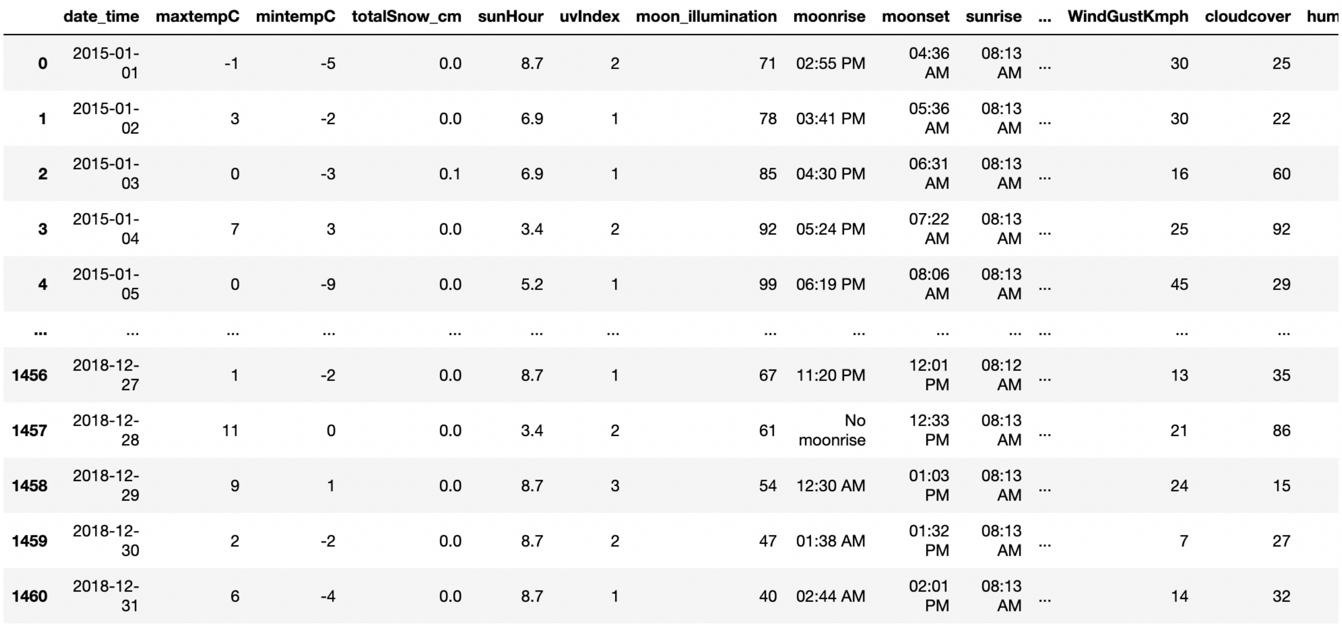


Fig 9: Crimes in Boston



	date_time	maxtempC	mintempC	totalSnow_cm	sunHour	uvIndex	moon_illumination	moonrise	moonset	sunrise	...	WindGustKmph	cloudcover	hum
0	2015-01-01	-1	-5	0.0	8.7	2	71	02:55 PM	04:36 AM	08:13 AM	...	30	25	
1	2015-01-02	3	-2	0.0	6.9	1	78	03:41 PM	05:36 AM	08:13 AM	...	30	22	
2	2015-01-03	0	-3	0.1	6.9	1	85	04:30 PM	06:31 AM	08:13 AM	...	16	60	
3	2015-01-04	7	3	0.0	3.4	2	92	05:24 PM	07:22 AM	08:13 AM	...	25	92	
4	2015-01-05	0	-9	0.0	5.2	1	99	06:19 PM	08:06 AM	08:13 AM	...	45	29	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1456	2018-12-27	1	-2	0.0	8.7	1	67	11:20 PM	12:01 PM	08:12 AM	...	13	35	
1457	2018-12-28	11	0	0.0	3.4	2	61	No moonrise	12:33 PM	08:13 AM	...	21	86	
1458	2018-12-29	9	1	0.0	8.7	3	54	12:30 AM	01:03 PM	08:13 AM	...	24	15	
1459	2018-12-30	2	-2	0.0	8.7	2	47	01:38 AM	01:32 PM	08:13 AM	...	7	27	
1460	2018-12-31	6	-4	0.0	8.7	1	40	02:44 AM	02:01 PM	08:13 AM	...	14	32	

Fig 10: Historical Weather Data