CRIME ANALYSIS IN BOSTON

Sushrut Kerulkar

Priyanka Jadhav

Umanng Kolhe

Kaushik Vapiwala

TABLE OF CONTENTS

ABSTRACT:

The Boston Council is of the most active councils in the States, we decided to proceed with analyzing their datasets provided by Boston Police Department from August 2015 to 2018. The social security of the citizen is one of the major reasons any government would bring as their priority. The primary concern for the council and the governing party is the social safety of their citizens. According to a reliable source, there is 1 in 37 chance of being a victim in violent crime in Boston. We wanted to analyze the historical data, to understand what has happened over the years. The version of the dataset we started working with had 303371 stances with 17 attributes from 2015 to 2018, extracted the dataset from Kaggle Website. Another dataset we worked with was Fire Incidents in Boston, extracted from Government Website. The dataset consists of 191822 instances and 17 columns of attributes from 2015 to 2018.

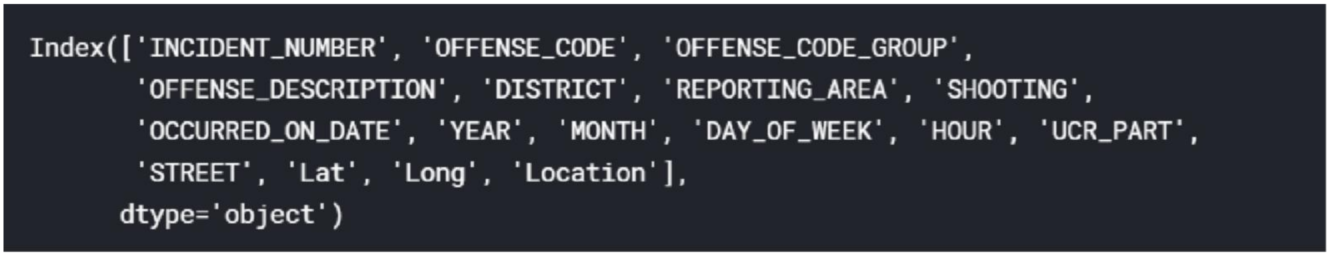
Boston has been economically developed, since our research we wanted to call attention on all the factors that could have been a major factor in crime. Initially, we focused on various other factors like food, geographic areas, latitude and longitude affecting the crime scenes. Find a linear relation between criminal activities, and fire incidents would be able to give us wider analysis and reasoning which we were able to present in our paper.

INTRODUCTION:

Boston is the capital of one of the most populous cities, of Massachusetts in the United States. As a combined statistical area, Boston is among the sixth-largest city in the United States. Crime in Boston generally occurs due to poverty in certain geographical areas. Boston has one of the oldest regulated Police Department in the United States. The BPD is 20th largest enforcement agency in the United States, the biggest help for us was the updated data provided by the Boston Council and BPD. They have a various dataset to gather, analyze and assist on the crime scenes. Our paper represents a fire-crime line-chart and heat map representing the crime caused in a geographical region from 2015 to 2018. We made pie-plot of crime in percentile format for every year presenting the crime conducted every year.

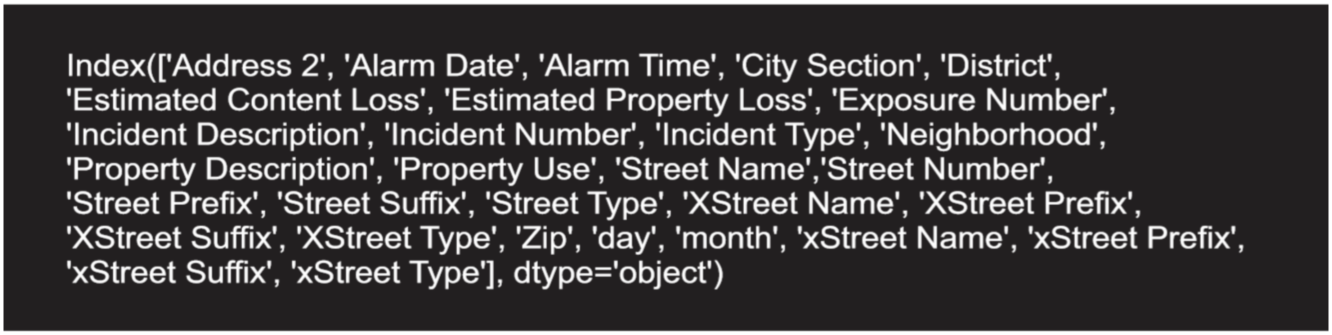
We used Python with Pandas, NumPy, Matplotlib, Seaborn, Base Map, Folium Library to help us with visualization. The common grounds we could find for correlating the Fire Dataset and Crime Dataset was Incident Number and Street Number.

The attributes in the crime datasets are as follows:



```
Index(['INCIDENT_NUMBER', 'OFFENSE_CODE', 'OFFENSE_CODE_GROUP',
       'OFFENSE_DESCRIPTION', 'DISTRICT', 'REPORTING_AREA', 'SHOOTING',
       'OCCURRED_ON_DATE', 'YEAR', 'MONTH', 'DAY_OF_WEEK', 'HOUR', 'UCR_PART',
       'STREET', 'Lat', 'Long', 'Location'],
      dtype='object')
```

The attributes for the Fire Dataset as follow:



```
Index(['Address 2', 'Alarm Date', 'Alarm Time', 'City Section', 'District',
'Estimated Content Loss', 'Estimated Property Loss', 'Exposure Number',
'Incident Description', 'Incident Number', 'Incident Type', 'Neighborhood',
'Property Description', 'Property Use', 'Street Name','Street Number',
'Street Prefix', 'Street Suffix', 'Street Type', 'XStreet Name', 'XStreet Prefix',
'XStreet Suffix', 'XStreet Type', 'Zip', 'day', 'month', 'xStreet Name', 'xStreet Prefix',
'xStreet Suffix', 'xStreet Type'], dtype='object')
```

BACKGROUND AND RELATED WORK

We used numerous techniques for analyzing the geographical density of Boston, one technique we used to be was of the heatmap. A heatmap is a geographical representation where we are using data values contained in a matrix, represented as colors. In one of our visualization, we used a heat map to display k-means clustering. It helped us with representing each cell of data with different color deciding its spectrum based on its statistical mean. If the visual representation of hierarchical clustering is required heat maps could be of great help.

We also used k-means, which is the simplest form of unsupervised learning. Social security and personal safety have been the primary concern of our life. Boston is a city of 20 distinctive neighborhoods, housing 589,000 residents. More than 200,000 are filled with college students in these areas every year. According to Neighborhood Scout website, the chance of being a victim of either a violent or property crime in Boston is 1 in 34. Violent offenses tracked include rape, murder and non-negligent manslaughter, armed robberies and assaults with a deadly weapon. Boston being an education hub and a home to a large number of local as well as international residents it is pivotal to consider the safety of the people.

We conducted research on various datasets and tried to infer insights by connecting the dots between the offenses committed in Boston and the results we received because of data analysis. In this project, we have tried to connect various crimes in Boston and the Fire incidents in Boston districts. We have extrapolated the connection between the crime happening in Boston and the fire incidents happening in Boston at the same time.

METHODOLOGY

Crime data is extracted from a public dataset providing website Kaggle. The dataset had over 300K instances with17 columns. Although the data was of high quality, it had many instances which had vital attributes values missing. For the analysis, the location attribute is the key attribute. The location attribute is divided into three parts viz. Street Name, latitude, and longitude. There are cases of instances where the only street name was available but latitude and longitude missing. Moreover, there are also instances where the street name was missing but both latitude and longitude available. There are no cases where either one of the latitudes and longitude missing but the street name available. Crime instances are the key aspects of the analysis. Hence the goal of the data cleaning process is to keep as much data possible. To resolve the issue of the missing street names, all the street names are grouped by average with respect to latitude and longitude. Hence, with the help of grouped data, the nearest latitude and longitude are taken into consideration with the help of the street name. For the vice versa cases, Bing Map API is used to retrieve street names from the given latitude and latitude. Cases, where all the three values are missing, are deleted as the location is a vital aspect of the analysis for this project.

The core aspect of the project relied on an analysis of crime rate based on time factor. Hence, simple statistical mathematical charts like a bar graph, pie charts, and line charts are created for the type and count of crimes against time factor. In addition to that, the attribute of location is defined visually with the help of the map. The key observation here was that the data is very rich and diverse in terms of location. When the data is plotted as a map, it proved to give a rough map of Boston city. It was noted that the central part of Boston is safe in terms of crimes. Besides that, there is a gradual reduction in the crime rate from 2015 - 2018.

o further the analysis, the crime data were compared to fire data in Boston. The fire data is found to be of high quality as there were hardly any missing or corrupted values in the dataset. Hence there was no real cleaning required. To compare fire and crime instances, both the data is plotted using a line chart against the days of the week. A subtle trend of a matching number of crime and fire instances are noticed at the start of the week and weekend. It can be inferred roughly that fire and crime are related somehow.

To strengthen the analysis and derive a further understanding of the data, the clustering method of K-Means is used to understand the type of crime against the location. The offense type is clustered into 5 types and plotted against latitude and longitude. It is noticeable that although the central part of Boston is safe, a further part in the northeast is equally more criminally active.

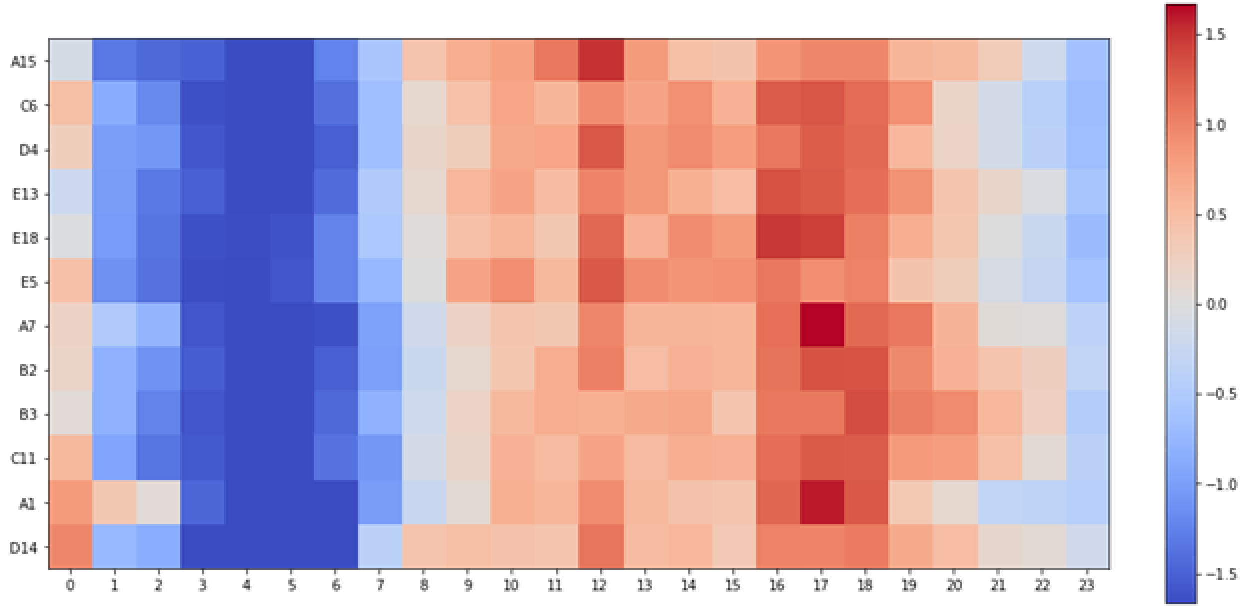
The overall analysis helped in deriving the conclusion that Boston is changing but at a slow phase. The city health in the year 2016 and 2017 was quite critical. Presumably, the active government noticed the deteriorating city health and might have pushed safety policies as it is seen from the improvement of city health in the year 2018. The analysis showcased a few of the safest and dangerous parts of Boston City.

EXPERIMENTAL RESULTS

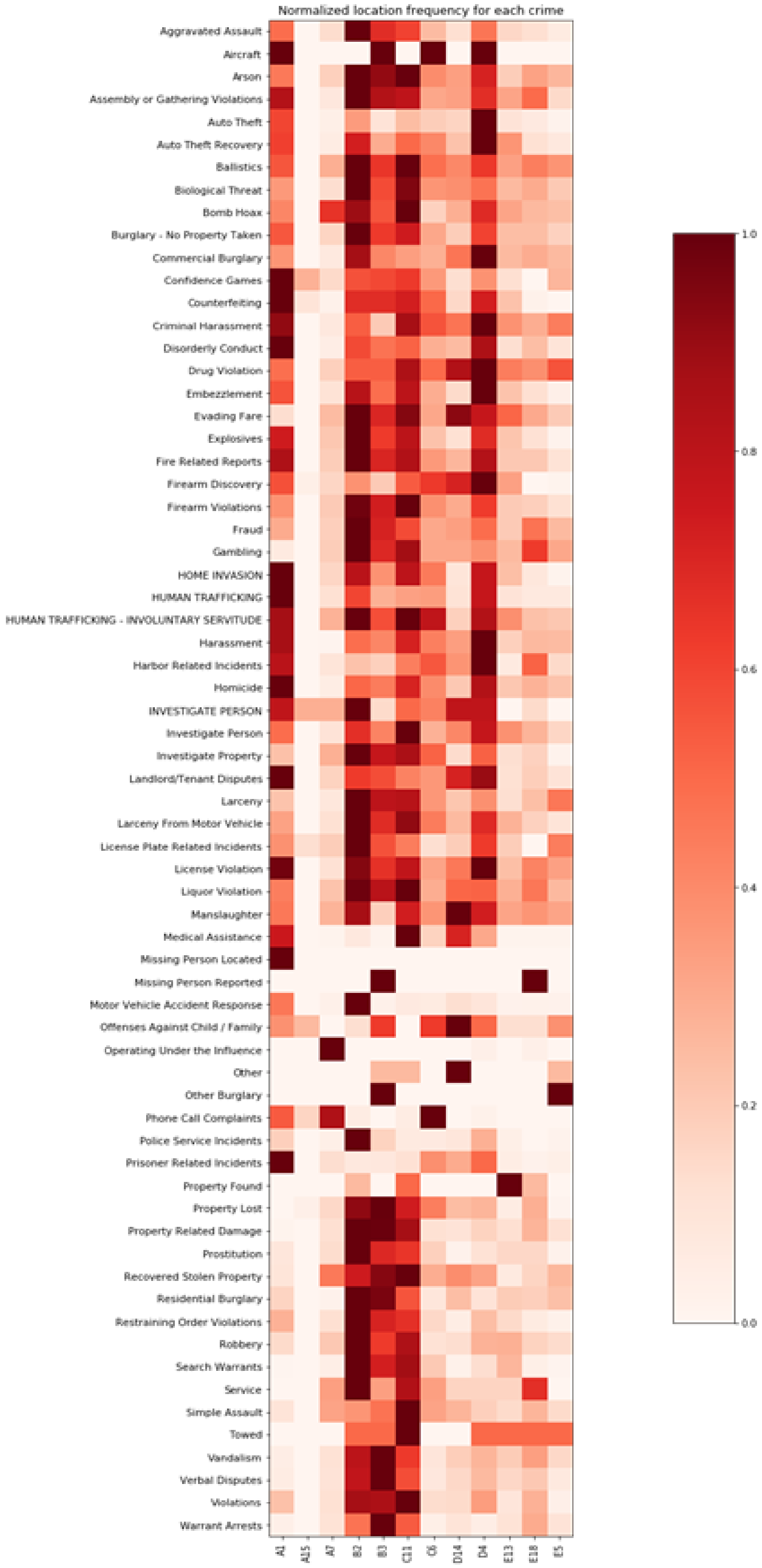
To arrive at results, we made a preliminary exploratory data analysis:

Here is what we found:

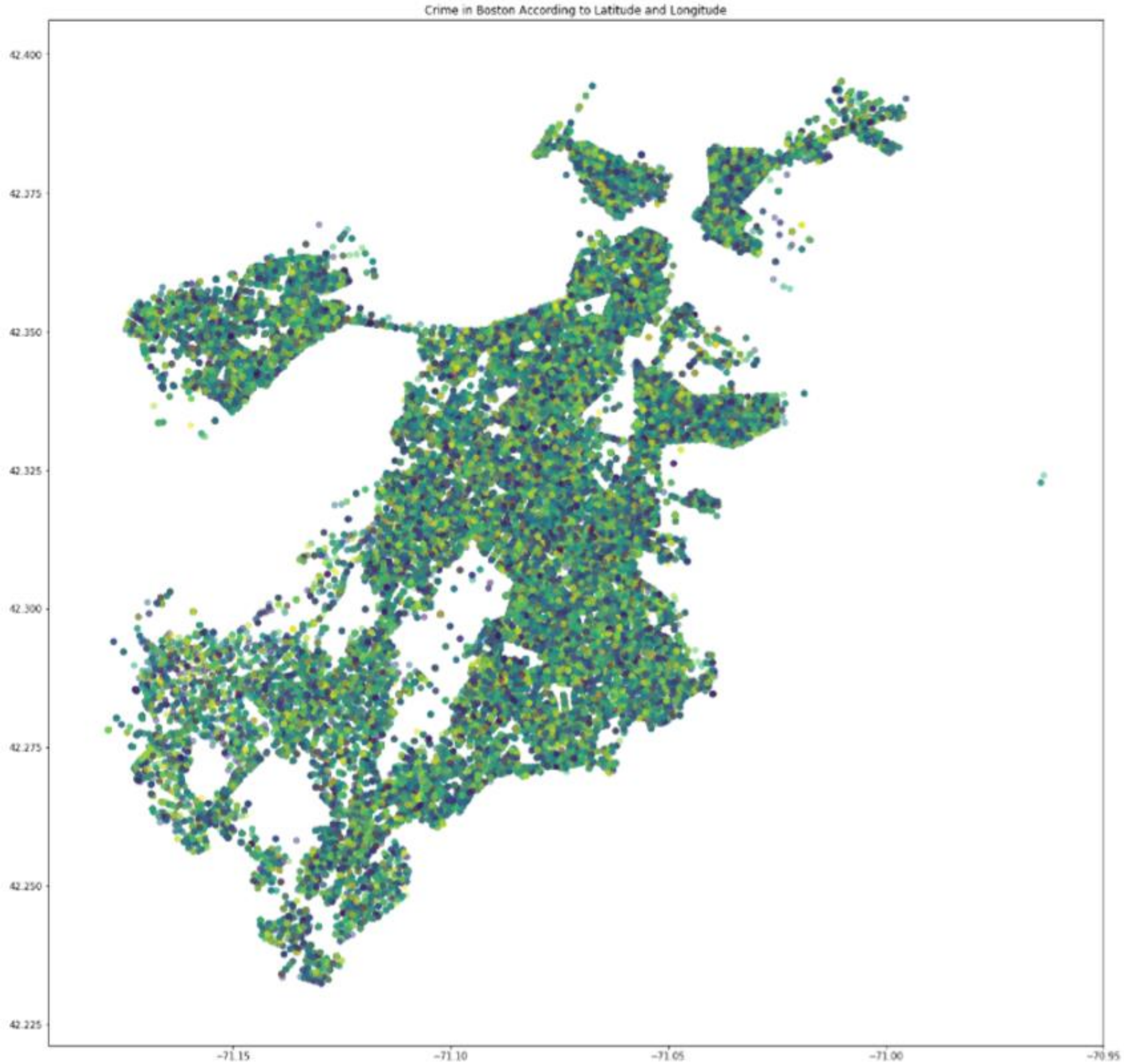
The graph below shows the time at which hours the crime has occurred in that district. The red color indicating the highest number of crimes and blue color representing the lowest number of crimes. We used a heatmap representation for presenting the crime conducted every hour of the day, against the street number in the datasets. We took 24 rows matrix against the 12 columns. The 24 hours are represented in X-axis of the dataset and District is represented in Y- axis of the heatmap.

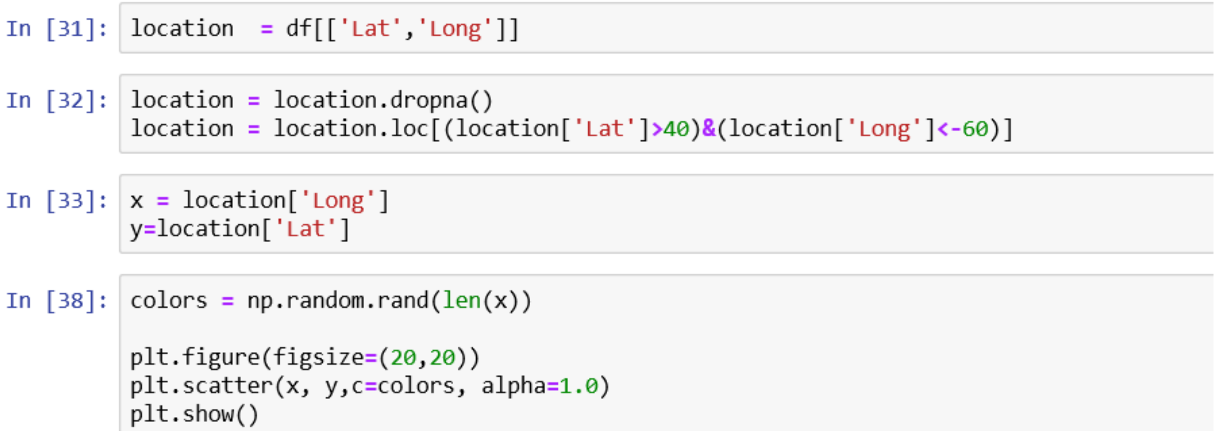


The below-represented graph we wanted to bring a details representation of heatmap letting us provide causes of crime at the hour. We concluded in the previous graph that 17:00 is when the most crime is conducted, we also wanted to know the kind of crimes is conducted during that hour. As the scale rises to 1 represented aside the heatmap, we can conclude that the crime is conducted extensively. We have used Agglomerative clustering algorithm to sort the rows into meaningful groups and use group labels to re-sort our matrix. The only change in the represented in this heat map is that we have drilled down towards the District vs the Type of the Crime conducted District.

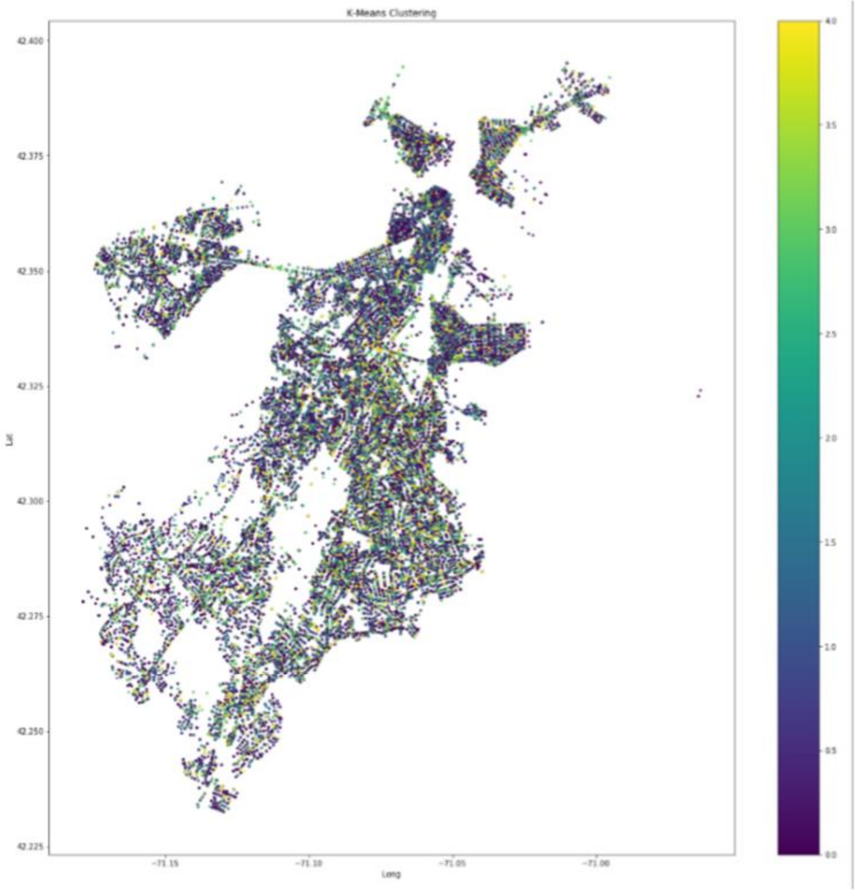

Normalized location frequency for each crime
Aggravated Assault
Aircraft
Arson
Assembly or Gathering Violations
Auto Theft
Auto Theft Recovery
Ballistics
Biological Threat
Bomb Hoax
Burglary - No Property Taken
Commercial Burglary
Confidence Games
Counterfeiting
Criminal Harassment
Disorderly Conduct
Drug Violation
Embezzlement
Evading Fare
Explosives
Fire Related Reports
Firearm Discovery
Firearm Violations
Fraud
Gambling
HOME INVASION
HUMAN TRAFFICKING
HUMAN TRAFFICKING - INVOLUNTARY SERVITUDE
Harassment
Harbor Related Incidents
Homicide
INVESTIGATE PERSON
Investigate Person
Investigate Property
Landlord/Tenant Disputes
Larceny
Larceny From Motor Vehicle
License Plate Related Incidents
License Violation
Liquor Violation
Manslaughter
Medical Assistance
Missing Person Located
Missing Person Reported
Motor Vehicle Accident Response
Offenses Against Child / Family
Operating Under the Influence
Other
Other Burglary
Phone Call Complaints
Police Service Incidents
Prisoner Related Incidents
Property Found
Property Lost
Property Related Damage
Prostitution
Recovered Stolen Property
Residential Burglary
Restraining Order Violations
Robbery
Search Warrants
Service
Simple Assault
Towed
Vandalism
Verbal Disputes
Violations
Warrant Arrests
A1
A15
A7
B2
B3
C11
C6
D14
D4
E13
E18
E5
1.0
0.8
0.6
0.4
0.2
0.0

K-Means Clustering of the Latitude & longitude Values:



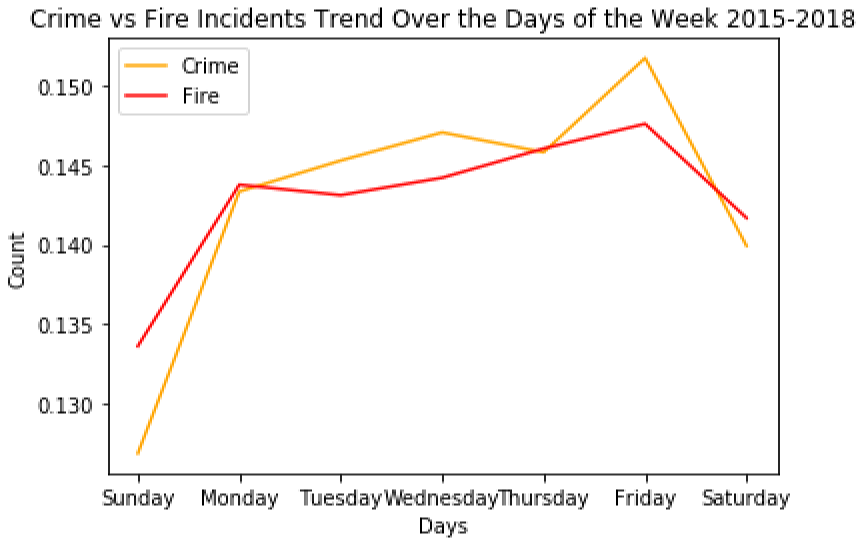


```
In [31]: location  = df[['Lat','Long']]

In [32]: location = location.dropna()
         location = location.loc[(location['Lat']>40)&(location['Long']<-60)]

In [33]: x = location['Long']
         y=location['Lat']

In [38]: colors = np.random.rand(len(x))

         plt.figure(figsize=(20,20))
         plt.scatter(x, y,c=colors, alpha=1.0)
         plt.show()
```



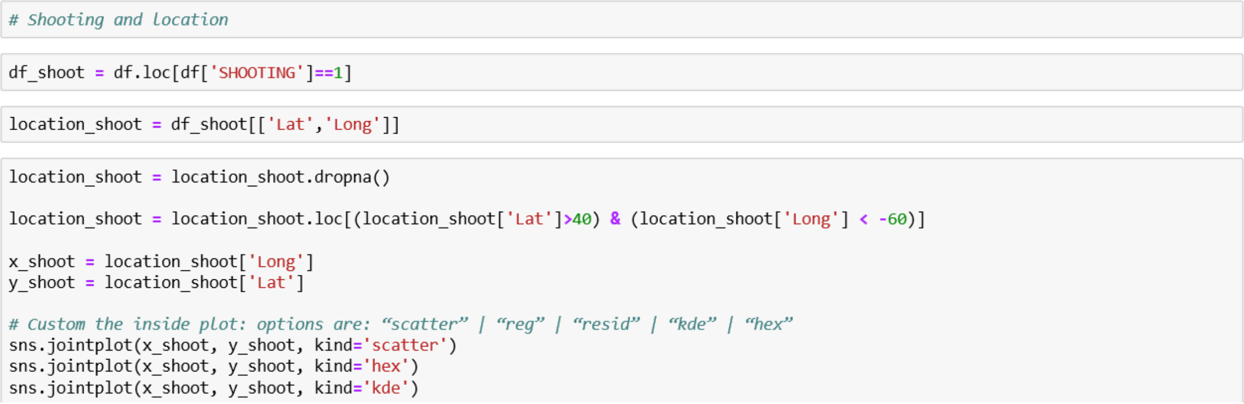
KMeans Clustering Using Location and Offense Code

The graph shows the crime vs fire incidents over the week.


Crime vs Fire Incidents Trend Over the Days of the Week 2015-2018

We tried to analyze the data in various representation, before concluding the prediction.

The below is the analysis in Scatter/KDE/HEX representation:

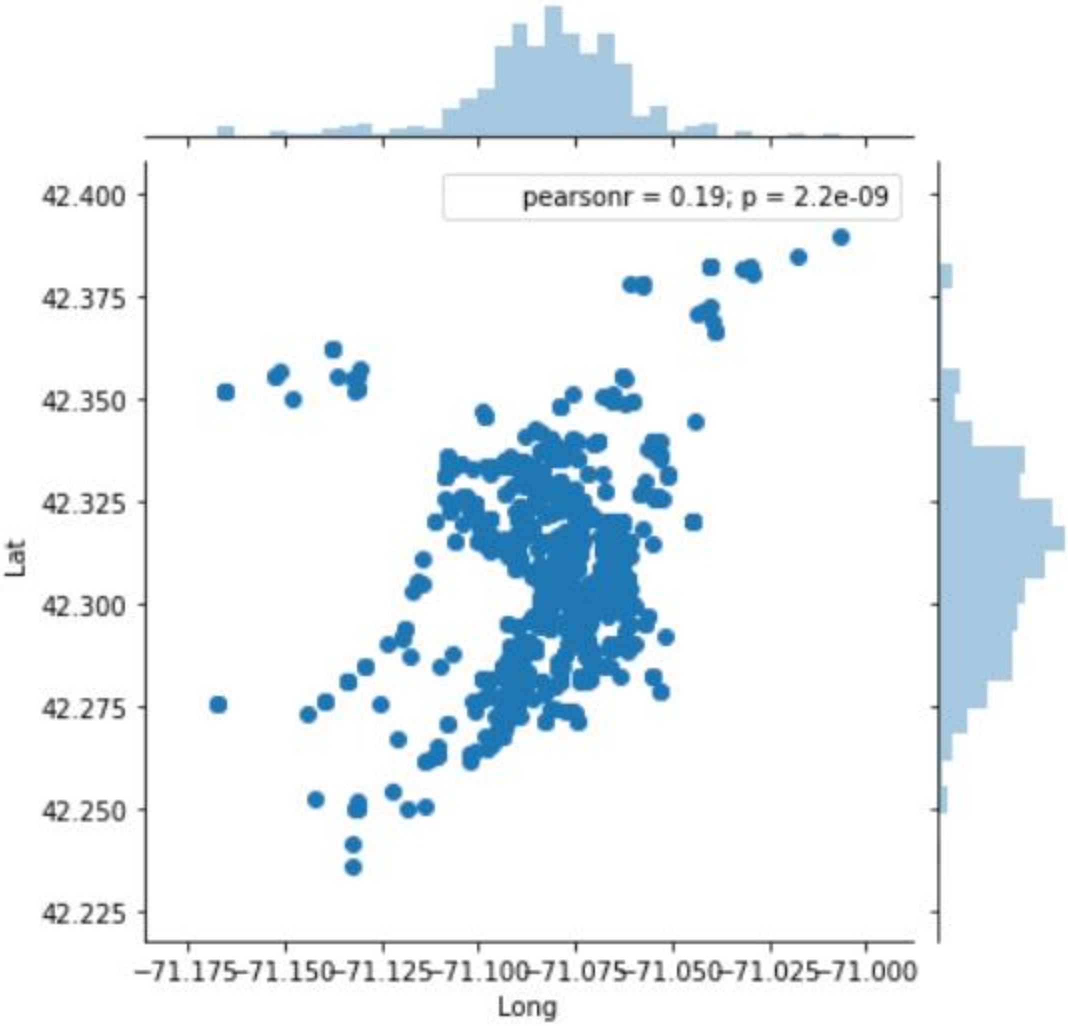


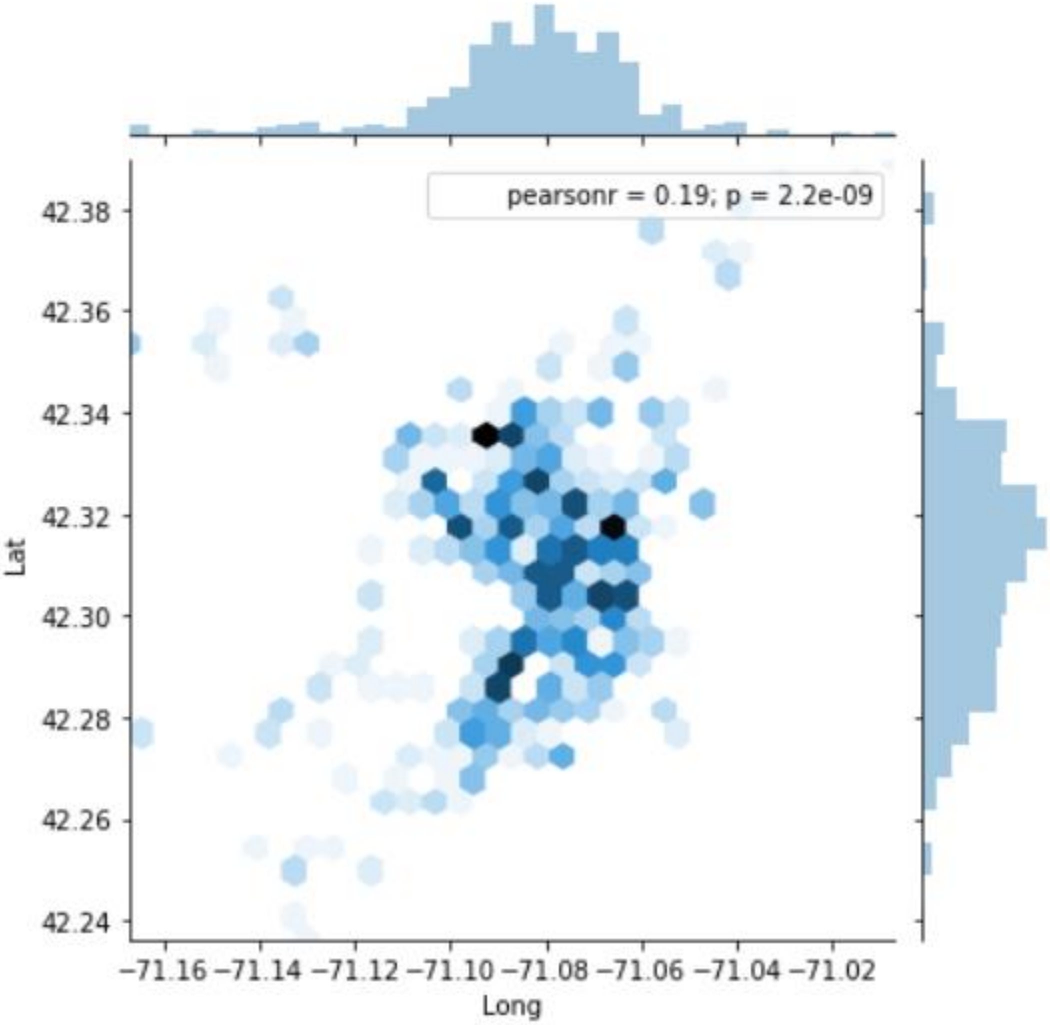
```
# Shooting and location
```

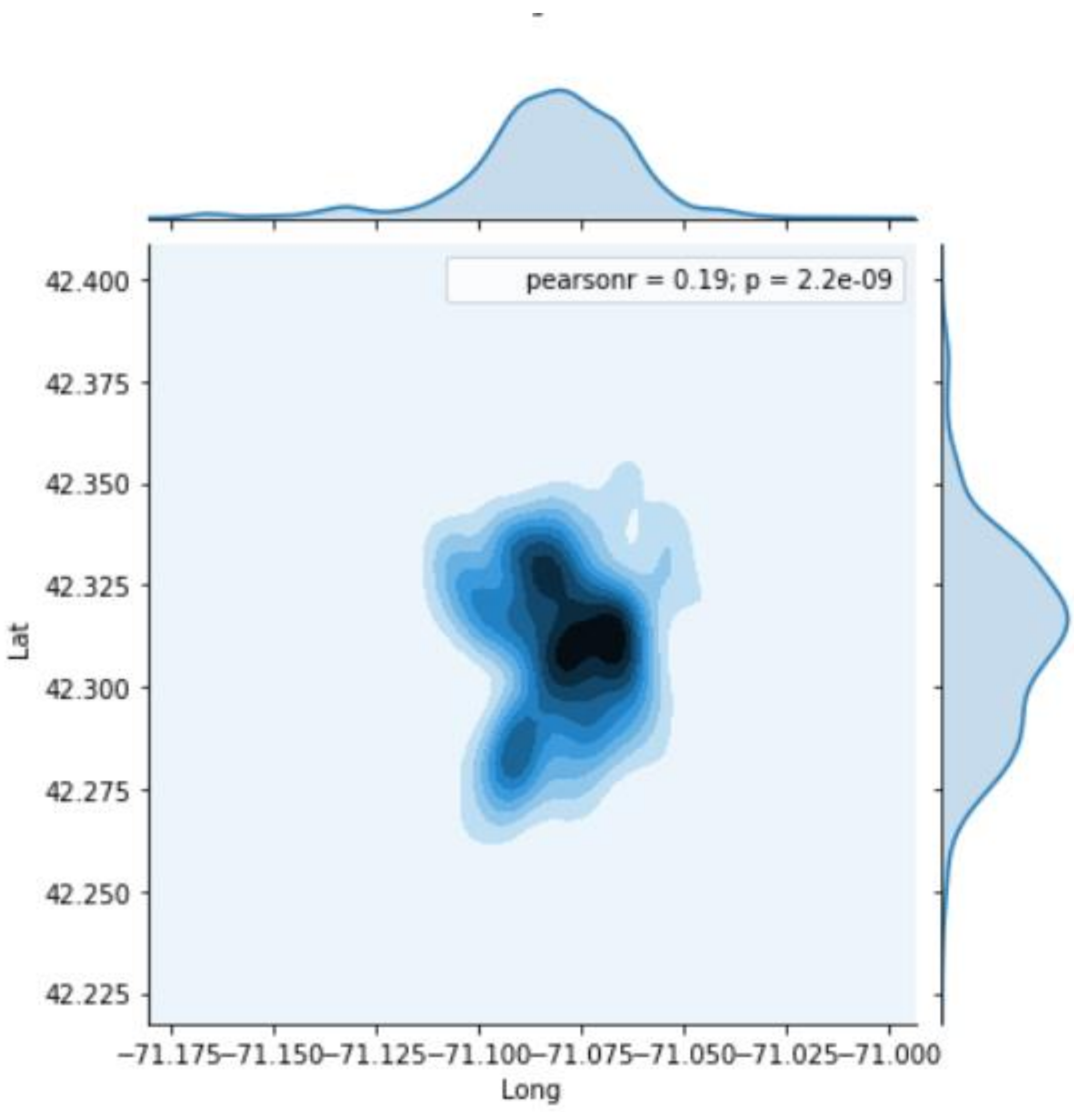
```
df_shoot = df.loc[df['SHOOTING']==1]
```

```
location_shoot = df_shoot[['Lat','Long']]
```

```
location_shoot = location_shoot.dropna()

location_shoot = location_shoot.loc[(location_shoot['Lat']>40) & (location_shoot['Long'] < -60)]

x_shoot = location_shoot['Long']
y_shoot = location_shoot['Lat']

# Custom the inside plot: options are: “scatter” | “reg” | “resid” | “kde” | “hex”
sns.jointplot(x_shoot, y_shoot, kind='scatter')
sns.jointplot(x_shoot, y_shoot, kind='hex')
sns.jointplot(x_shoot, y_shoot, kind='kde')
```


pearsonr = 0.19; p = 2.2e-09
42.400
42.375
42.350
42.325
42.300
42.275
42.250
42.225
Lat
−71.175
−71.150
−71.125
−71.100
−71.075
−71.050
−71.025
−71.000
Long


pearsonr = 0.19; p = 2.2e-09
42.38
42.36
42.34
42.32
42.30
42.28
42.26
42.24
Lat
−71.16
−71.14
−71.12
−71.10
−71.08
−71.06
−71.04
−71.02
Long



CONCLUSION

Boston, city of crime. You name crime and it must have been committed in Boston. Not just once but a couple of times. The crime rate in Boston is higher than many other communities in America, from smallest to largest, 30 crimes per thousand residents. The chance of being a victim of either violent or property crime is 1 in 37. Boston is not one of the safest cities in America. Boston has a crime rate that is higher than 97% of the state's cities and towns of all sizes.

To analyze these crime incidents and predict a safe zone we went through a couple of datasets resulting in crime in Boston. After careful consideration and analysis, we have some findings that state the year, month, week, day, and time of the day when the crime rate is higher. Along with crime data we also came across some fire incidents data and found conclusive evidence resulting in a common link between crime and fire incidents.

This link helped us to understand that whenever a fire incident occurs there is a crime occurred. After a deep analysis of crime and location of the crime along with the same for fire incidents, we found that people that advantage of the fire incidents and commit a crime on the same location.

K means clustering have helped us to find and predict safe and hurtful time zone in Boston. Also, we have found some safer areas in Boston where there is less chance of a crime being committed. This analysis led us to not down some precautions that a common citizen must take to avoid such crimes. It also led to choose an area to buy the safe property and live healthily.

FUTURE SCOPE

Hotspots:

The most common method of forecasting crime is Hotspots, hotspots of yesterday are hotspots of tomorrow. We can simply find hotspots from our analysis and provide them to the police department. These spots can be kept under surveillance all the time. The police department can also locate its police stations nearby to these spots or maintain a helpline in the immediate next zone.

Property:

People can avoid buying property in these hotspots. We can find data on property zones or districts in Boston. Analyze both the datasets. This analyzed data can be sold to different property buyers, who can buy and trade property deals. We can also set a fair amount of price for a particular property depending upon the frequency of crime in that locality.

Portable Crime Forecast system:

To expand the handiness of Crime Forecasting framework, one of the improvements that can be made is to build up the framework as a portable application. This would allow the user to report the crime remotely as soon as he observes one and eliminate the need for a desktop system to report a crime. To build the adequacy of the Crime Forecasting System, more procedures could be utilized for gauging wrongdoing.

REFERENCES

1. Groof E and Vigne N, “Forecasting the future of predicting crime mapping”. *Education.Psu web,* retrieved on May 09, 2019
2. Schneider S, (2002), “Predicting Crime: A review of Research”, Retrieved from Ryerson University
3. Boston City Guide, Retrieved from https://adventure.howstuffworks.com/boston-city-guide.htm
4. Neighborhood Scout – Boston Crime Data, Retrieved from https://www.neighborhoodscout.com/ma/boston/crime