PROJECT REPORT

Vy Nguyen
nguyen.vy7@northeastern.edu

Dec 14, 2020

1. Introduction and Research Questions

1.1. Introduction

There is no other city that is filled with so much history than the city of Boston. More than history, it is a city known for its sports, its colleges, its culture, and its people. However, there is one problem that always bothers us, especially those who just moved to Boston: Is Boston safe? Are there any specific places in Boston that are dangerous that we should avoid when we look for accommodation?

For this analysis, we will be exploring police report data from 2018 to 2021 for the city of Boston. The analysis will try to pinpoint times of year and day that are particularly dangerous to help both residents and visitors stay safer, and Boston Police identify crime hotspots, better anticipate staffing needs and improve security in menacing areas. Besides, the analysis will also examine where police incident reports are more likely by district and street as well as the most common types of incidents.

1.2. Research Questions

- Determine the most common crimes in Boston
- Determine the most dangerous time of year and in a day in Boston
- Identify the places (District and Street) where crimes are most likely to be committed
- Crimes related to Shooting analysis
- Incidents in Boston during Covid pandemic analysis
- Crimes at Northeastern University area analysis

1.3. Hypotheses for Research Questions

- The most common incidents in Boston may be related to motor vehicle and burglary as Boston is a densely populated city.
- The most common time for a crime to occur could be from midnight until dawn. While months having major holidays like November and December, may witness a high crime rate.
- Crime is significantly more likely to happen on major districts such as Downtown, and busy streets with high number of pedestrians like Boylston Street.
- Shootings are more likely to happen at nighttime, in major districts and in most serious crimes such as aggravated assault.

- The crime rate in Boston might decrease during the Covid pandemic because of the quarantine.

2. Summary of Results

Boston is the fourth most densely populated city in the USA, resulting in a high number of motor vehicle accidents and larceny-theft crimes, similar to our beginning hypothesis. During the evening rush hour, from 4PM to 6PM, the highest number of crimes is recorded. However, incidents where shooting takes place become more common during 9PM to 12PM period.

Downtown is the economic hub and most densely populated area of the city, thereby the number of crimes reported are high in the area. Also, the low economically backward communities such as Roxbury and Dorchester are the hotspots for all types of incidents. The streets, such as Boylston Street, Washington Street and Blue Hill Avenue, connecting those areas, experience a high number of incident reports.

Aggravated Assault crimes have the most probability of shootings associated with them, agreeing with our hypothesis. Besides, Roxbury, Dorchester and Mattapan are the most dangerous districts where shootings are more likely to happen. Blue Hill Avenue and Washington Street, which stretch through those neighborhoods, also witness high shooting rate.

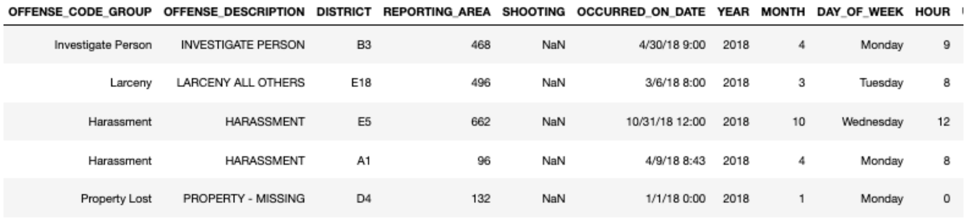
The incident reports during Covid-19 pandemic saw a big increase, contrary to our hypothesis. Covid-19 has deeply impacted the crime environment of the city. Many people were under quarantine and have lost their jobs. Therefore, there was an increase in crimes like larceny and vandalism. Also, the number of simple assaults has gone up because people were forced to stay at home during the pandemic.

The neighborhood near Northeastern University is a rather safe area. The major crimes that occur around Northeastern university are Fire, Burglary, Harassment, Assault and Drug-related crimes.

3. Data Set

3.1. Data Source

The dataset is obtained from Crime Incident Reports, provided by Boston Police Department (BPD) Source: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system. It contains records from the new crime incident report system, which includes a set of fields focused on capturing the type of incident as well as when and where it occurred. For this project, we got the incident reports from 2018 to 2021 (Figure 1)



OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR
Investigate Person	INVESTIGATE PERSON	B3	468	NaN	4/30/18 9:00	2018	4	Monday	9
Larceny	LARCENY ALL OTHERS	E18	496	NaN	3/6/18 8:00	2018	3	Tuesday	8
Harassment	HARASSMENT	E5	662	NaN	10/31/18 12:00	2018	10	Wednesday	12
Harassment	HARASSMENT	A1	96	NaN	4/9/18 8:43	2018	4	Monday	8
Property Lost	PROPERTY - MISSING	D4	132	NaN	1/1/18 0:00	2018	1	Monday	0

Figure 1: Sample of the data set for the project

3.2. Data Description

The full dataset for the project contains 328,687 observations and 17 variables (6 quantitative and 11 categorical) as following:

Variable	Description
INCIDENT_NUMBER (quantitative)	Internal Boston Police Department (BDP) report number
OFFENSE_CODE (quantitative)	Numerical code of offense description
OFFENSE_CODE_*GROUP* (categorical)	Internal categorization of offense description
OFFENSE_DESCRIPTION (categorical)	Primary description of incident
DISTRICT (categorical)	What district the crime was reported in
REPORTING_AREA (categorical)	Reporting area number associated with where the crime was reported from
SHOOTING (categorical)	Indicated a shooting took place
OCCURRED_ON_DATE (quantitative)	Earliest date and time the incident could have taken place
YEAR (categorical)	Which year the crime occurred
MONTH (categorical)	Which month the crime took place
DAY_OF_WEEK (categorical)	Which day in the week the offense occurred
HOUR (categorical)	What time of a day the incident happened
UCR_PART (categorical)	Universal Crime Reporting Part number (1,2,3)
STREET (categorical)	Which Street the crime was reported on
Lat (quantitative)	The Latitude of the place where the incident was reported in
Long (quantitative)	The Longitude of the place where the incident was reported in
Location (quantitative)	Location (Lat, Long) of the place where the crime happened

4. Methods and Results

4.1. Methods

In this project, we followed the workflow of five main stages.

1. Question or problem definition and hypotheses
2. Acquire dataset
3. Wrangle, prepare and cleanse the data
4. Analyze, identify patterns, and explore the data.
5. Visualize, report, and present the problem-solving steps and final solution

As we discussed the research questions and hypotheses in Part 1 of this report. We will focus more on how we acquire and clean the dataset, as well as analyze and visualize the data.

Acquire dataset

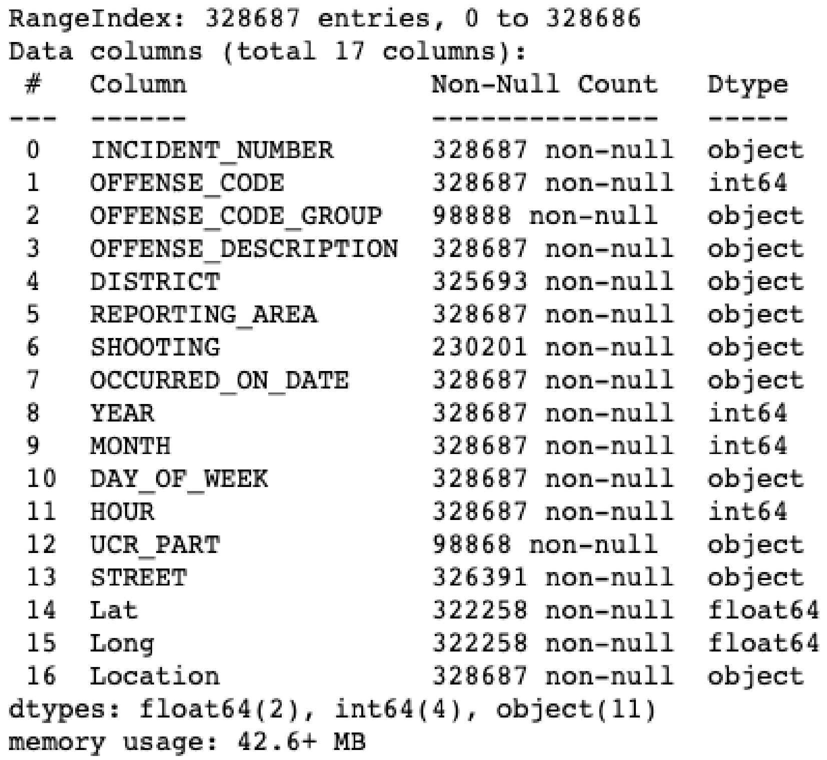
We download four different files of ‘Crime Incident Reports’ from 2018 to 2021 (each file for one year), provided by **Boston Police Department (BDP)** (from the link: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system).

Next, we combine four separate files into a final dataset, which contains all incidents reported to Boston Police Department (BPD) between 2018 and 2021.

The Python Pandas package helps us work with our dataset. We start by acquiring the dataset into Pandas Data Frame.

Wrangle, prepare and cleanse the data

Before cleaning the data, we need to get some major information about the dataset (Figure 2). Info() method in Pandas helps us to do that. There are *328,687 entries (*or *observations)* and *17 columns (*or *attributes)* in the dataset.



```
RangeIndex: 328687 entries, 0 to 328686
Data columns (total 17 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   INCIDENT_NUMBER      328687 non-null  object
 1   OFFENSE_CODE         328687 non-null  int64
 2   OFFENSE_CODE_GROUP   98888 non-null   object
 3   OFFENSE_DESCRIPTION  328687 non-null  object
 4   DISTRICT             325693 non-null  object
 5   REPORTING_AREA       328687 non-null  object
 6   SHOOTING             230201 non-null  object
 7   OCCURRED_ON_DATE     328687 non-null  object
 8   YEAR                 328687 non-null  int64
 9   MONTH                328687 non-null  int64
 10  DAY_OF_WEEK          328687 non-null  object
 11  HOUR                 328687 non-null  int64
 12  UCR_PART             98868 non-null   object
 13  STREET               326391 non-null  object
 14  Lat                  322258 non-null  float64
 15  Long                 322258 non-null  float64
 16  Location             328687 non-null  object
dtypes: float64(2), int64(4), object(11)
memory usage: 42.6+ MB
```

Figure 2: Information of the dataset

Moreover, we can get the number of non-null values in each column. From Figure 1, there are no missing values in *OFFENSE_DESCRIPTION* column, while some values in *OFFENSE_CODE_GROUP* are missing (only 98,888 out of 328,687 are non-null values). From that, we can expect that we need to impute missing values in *OFFENSE_CODE_GROUP* in data cleansing process.

The data types for each feature as described as below: Six features are integer or float, whereas the other eleven features are strings (object).

Now, we will delve into each step of the data cleansing process.

- **Drop unused columns**

Among 17 variables from the original dataset, there are three variables, namely *INCIDENT_NUMBER*, *OFFENSE_CODE* and *REPORTING_AREA*, not related to our research questions. Therefore, we use the *drop ()* function to remove these three columns from our dataset.

- **Rename column names**

Feature names in the original dataset are not in good format for doing analysis. Many of them are long, capitalized, and include underscores. We, therefore, rename all columns (except Lat, Long and Location) in order to make them shorter, simpler but still comprehensive by using *rename ()* function (Figure 3).

Original column name	***Amended column name***
OFFENSE_CODE_GROUP	Group
OFFENSE_DESCRIPTION	Description
DISTRICT	District_number
SHOOTING	Shooting
OCCURRED_ON_DATE	Date
YEAR	Year
MONTH	Month
DAY_OF_WEEK	Day
HOUR	Hour
UCR_PART	UCR_Part
STREET	Street

Figure 3: Table of original and amended column names

- **Reformatting values in *Description* column**

Values in *Description* attribute are in different formats. Some are uppercase, some are lowercase, and some have leading and trailing whitespaces. Therefore, it is necessary to reformat all of them to lower case and remove leading and trailing whitespaces.

- **Filling missing values for *Group* variable**

As a high percentage of values (nearly 70%) in the *Group* column is missing values, we need to fill missing values for this variable. We use the *map* function to map the values of the *Description* column to values of *Group* column. After imputation, the number of missing values in *Group* feature dropped from 229,799 to 41,823 values.

Among the remaining missing values in *Group* column, there are 10,232 values which should be "sick assist". However, there are no “sick assist” values in *Group* attribute from the original dataset. So, they cannot be matched when we use the map function. We therefore fill "sick assist" values for *Group* feature manually.

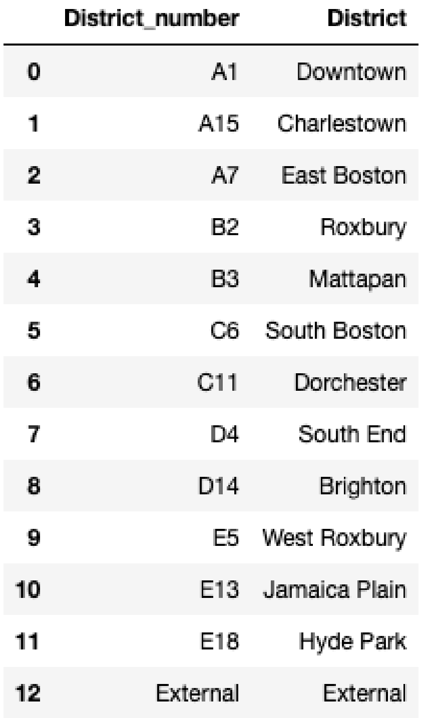
Finally, there are still 31,591 missing values in *Group* column, accounting for 9.61% of total values. Hence, we just drop all records having missing values in the *Group* column.

- **Create *Class* attribute**

To capture more high-level features from *the Group*, we have made a new column called *Class***.** Specifically, we group similar crime categories into a new category. For example, we group *Auto theft*, *License Plate Related Incidents,* and *Larceny from Motor vehicles* under a new category named *Motor* in *Class* column. The entire table is provided at link.

- **Cleaning District attribute**

Looking at the *District_number* feature, the number of missing values in this column accounts for only 0.86% of the dataset. We decided to drop all missing values records in *District_number* column and are left with 295,530 observations in total.



	District_number	District
0	A1	Downtown
1	A15	Charlestown
2	A7	East Boston
3	B2	Roxbury
4	B3	Mattapan
5	C6	South Boston
6	C11	Dorchester
7	D4	South End
8	D14	Brighton
9	E5	West Roxbury
10	E13	Jamaica Plain
11	E18	Hyde Park
12	External	External

Figure 4: District name table

The original dataset uses the district code (such as *A1*, *C6*) to represent for District where the incidents were reported in. However, readers may find it difficult to imagine which district by its code. Hence, we replace district codes with their names to be more readable (Figure 4).

Next, we drop *District_number* column. The current dataset has 15 attributes, consisting of: *Class, Group, Description, District, Street, Shooting, Date, Year, Month, Day, Hour, UCR_Part, Lat, Long and Location*.

- **Convert value type in *Date* column to datetime**

We apply *to_datetime()* method to convert value type in *Date* attribute to *datetime*.

- **Change values in *Shooting* column to 0 (No Shooting) and 1 (Shooting)**

There are five different types of value in *Shooting* feature: “Y”, “0” and “1” in string, and 0 and 1 in integer. In the dataset, 33% of values in the *Shooting* column are missing, which we assume those values mean no shooting. We therefore apply map method to change “Y” and “1” in string to integer 1 (means Shooting occurred), 0 in string and missing values to integer 0 (means Shooting did not occur).

- **Filling missing values for *UCR_Part* column**

Like the method of filling missing values for the *Group* attribute, we map values of the *Description* feature to those of *UCR_Part* column.

- **Filling missing values for *Lat* and *Long* attributes**

Besides 6,354 missing values in *Lat* and *Long* columns, there are some values of 0 and -1, which are considered as missing values. So, we fill all those missing values by the average values of Lat and Long (except 0 and –1) for each District.

- **Mapping values from *Lat, Long* features to *Location* feature**

Values of the *Location* attribute are derived directly from *Lat* and *Long’s* values by simply combining the two together. Thus, *Location* feature carries over the missing values from *Lat* and *Long* features. We also impute the Location feature by the new *Lat* and *Long* features updated from the previous step.

4.2. Results

- **Determine the most common crimes in Boston**

Motor Vehicle Accident Response and *Larceny* are the two most popular incidents in Boston (Figure 5). Boston has a high population density, which means there is a lot of traffic and pedestrians moving around in the city. Therefore, motor accidents are inevitable, and the number of larceny-theft records should be high

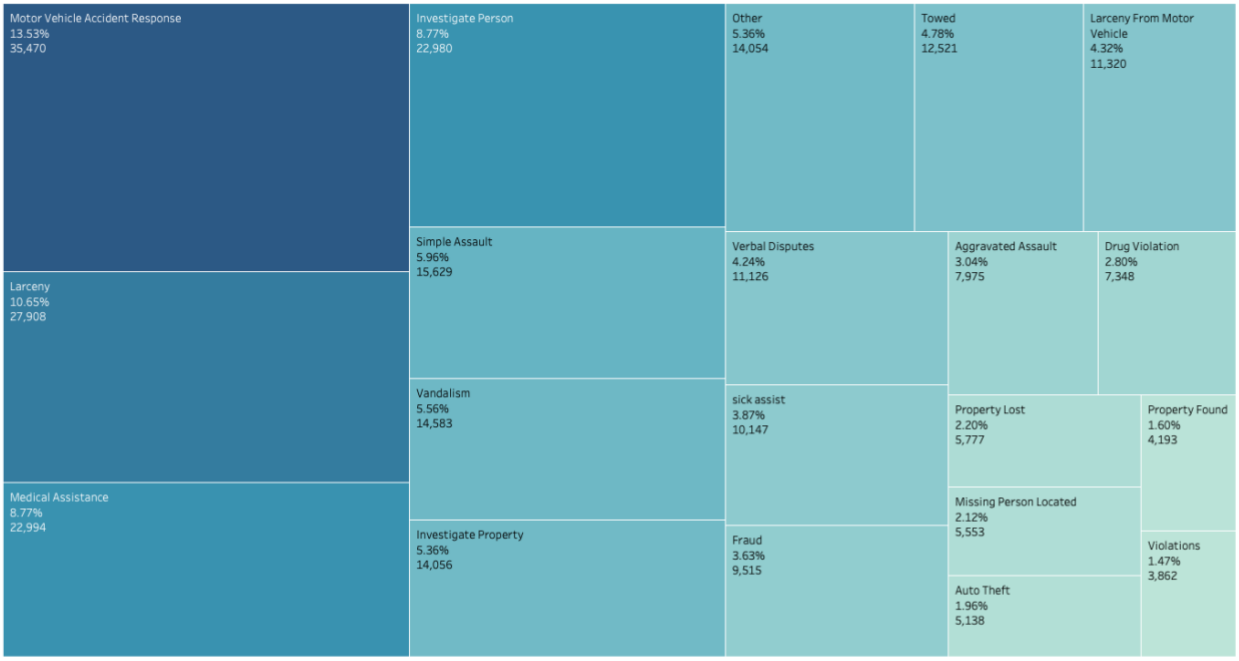


Figure 5: Tree map of top crimes in Boston

- **Determine the most dangerous time of year and in a day in Boston**

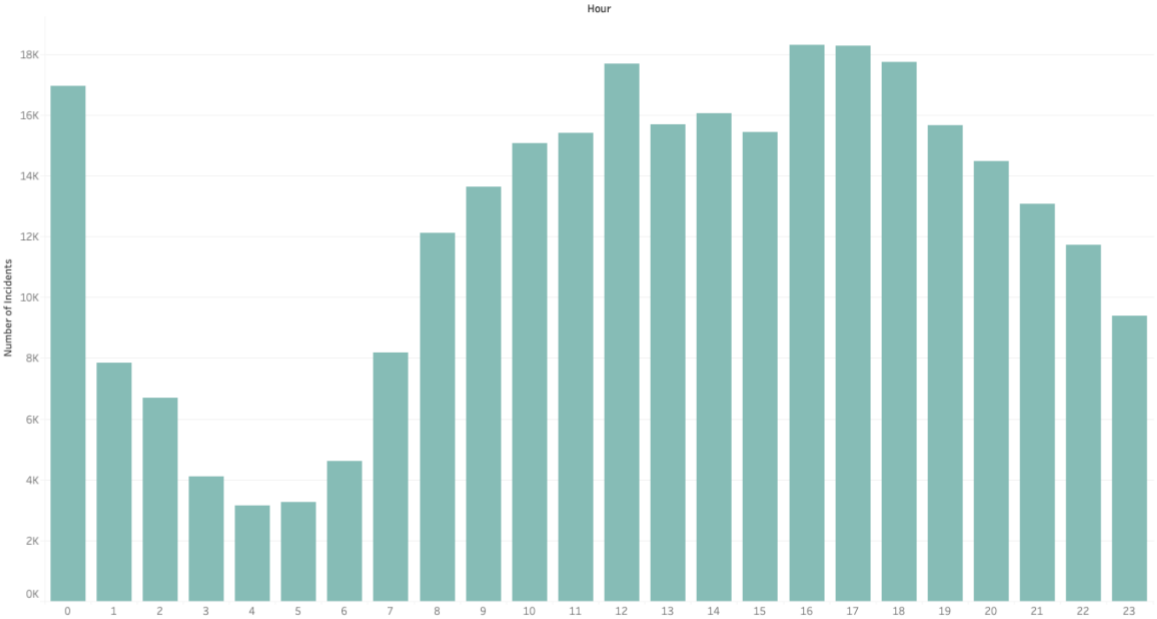


Figure 6: The number of incidents in Boston by hour

The most popular time for an incident to happen is right after work (from 4 PM to 6 PM) and at noon around lunch break, not late at night (Figure 6). This is contrary to popular belief and our beginning hypothesis. There may be some reasons for this result. It is considered that the most common incident is *Motor Vehicle Accident Response*, which is more likely to happen at peak hours when the traffic is hectic. Hence, vehicle accidents are unavoidable. As for the second most common crime, *Larceny (*such as theft of personal property), the same rationale also makes sense. There are a high number of people walking around the city who are likely careless and just trying to get home, leaving them more susceptible to being a victim of larceny.

Examining the season and month in a year in which a crime is most likely to occur helps to strengthen this theory. The months in summer and fall have high crime rates in Boston, peaking in August (Figure 7), contrary to our first hypothesis of the final months of the year experiencing a high number of crimes. In summer, people often spend time together outside as the children and students are off from school. In addition, the weather is nice, and it gets dark later, which are good opportunities for people to do outdoor activities. People are usually involved in these activities and are more careless, leaving them more vulnerable to theft.

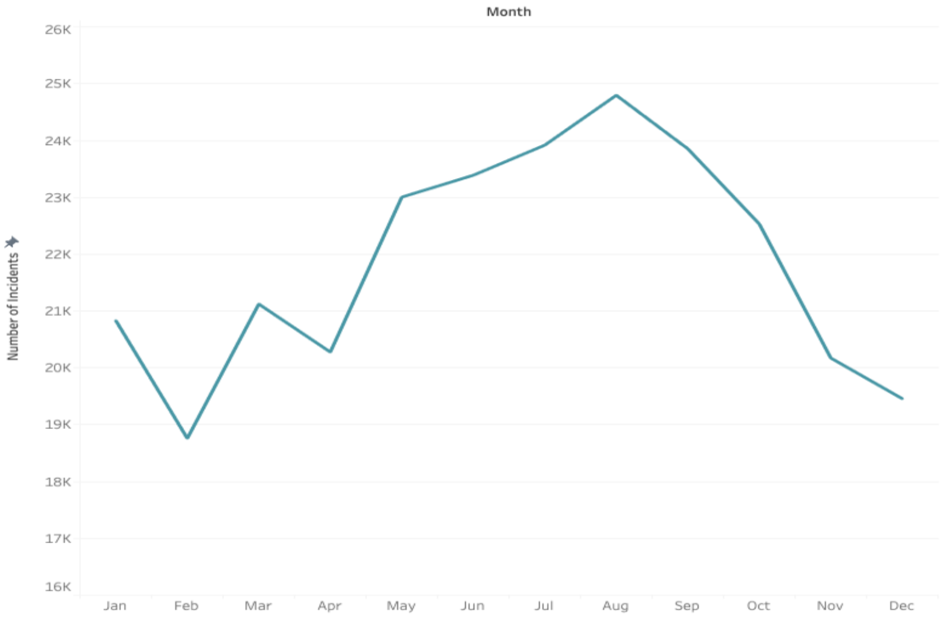


Figure 7: The number of incident reports in Boston by month

- **Identify the places (District and Street) where crimes are most likely to be committed**

The rate of police reports is high in low-income neighborhoods, which are Roxbury and Dorchester, and the major district, which is Downtown (Figure 8). This result is partially the same as our beginning hypothesis. According to [1], Roxbury and Dorchester, have been experiencing high poverty levels, significant unemployment and considerable racialized economic inequality in the form of homeownership rates and median income levels. Therefore, we conclude that those might be the main reasons for the high number of incident reports in these two districts. Regarding Downtown, this is a high foot traffic area leaving more people as targets of crimes like larceny, the second most common type of crime.

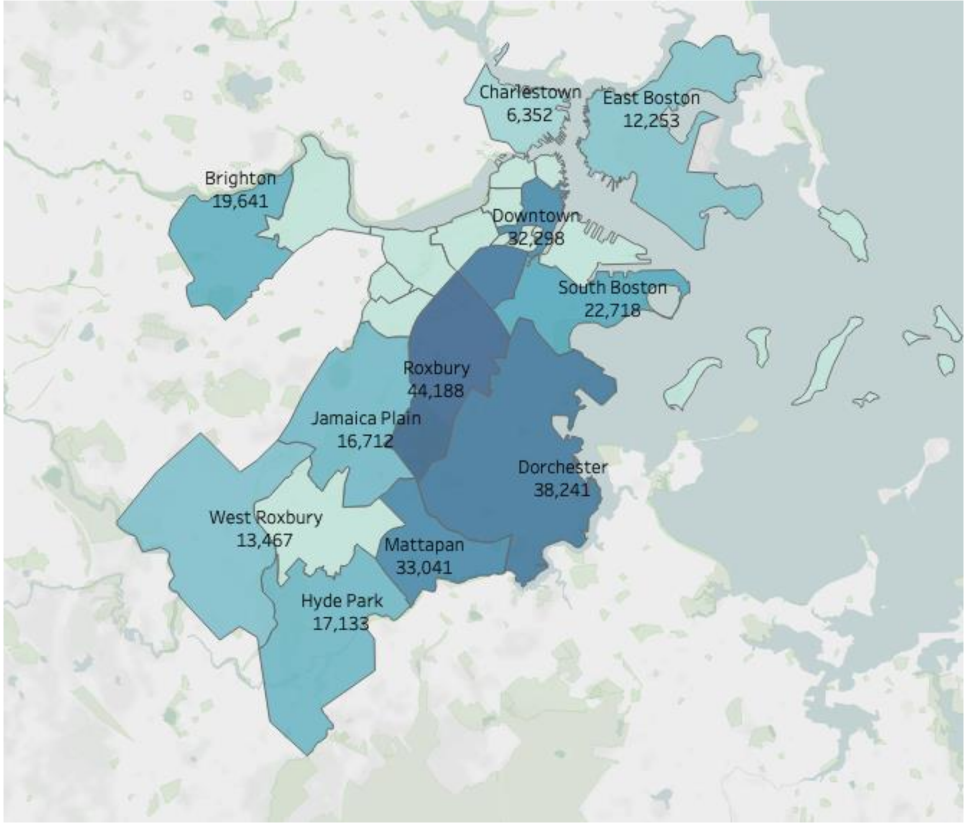


Figure 8: The number of Incident Reports by District

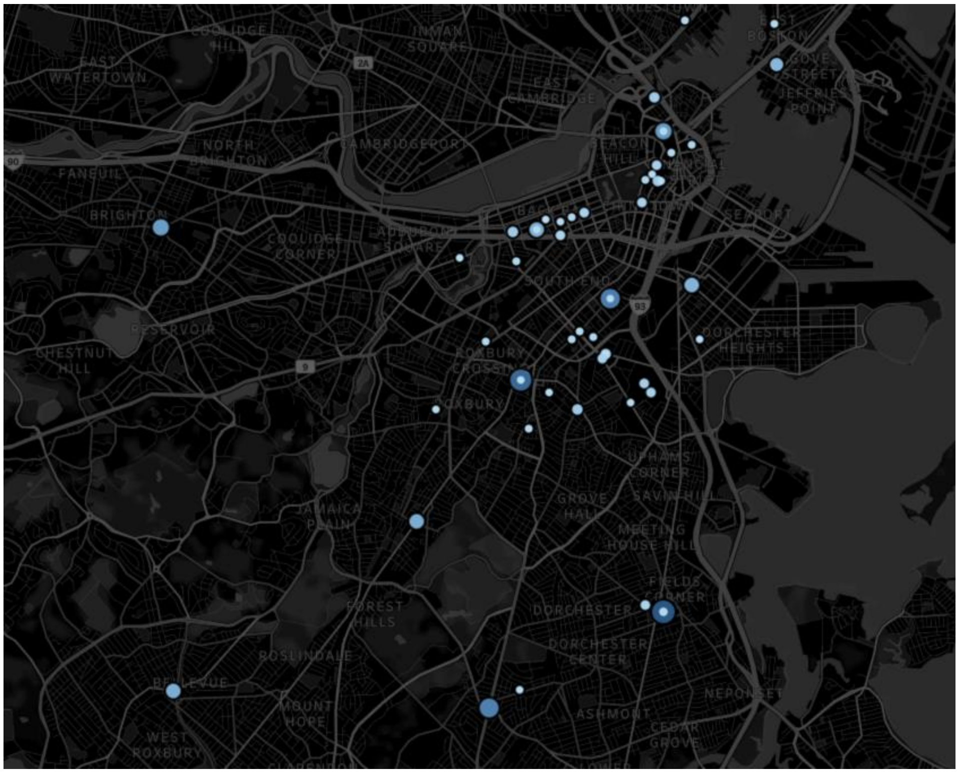


Figure 9: Incident Reports Analysis by Street

Crime is significantly more prevalent on major streets. The most common street for crime to occur is Washington Street, followed by Boylston Street, Blue Hill Avenue (Figure 9). We believe that each of these streets may have slightly different explanations for their high rate of police reports. Washington Street is a long street, which has a high number of intersections, on-street parking, and pedestrians. Boylston street is an incredibly long street that stretches through many neighborhoods. In addition, one part of this street is in Downtown, where there are many pedestrians and train stations, resulting in high number of crime reports, especially larceny. Blue Hill Avenue is the main road, connected to Dorchester and Roxbury. Therefore, the number of crime reports is high on this street is reasonable.

- **Incidents in Boston during Covid pandemic analysis**

During the peak time of the Covid pandemic (from April 2020 to October 2020), the number of incidents had a high increase (Figure 10), even though most of the population was under quarantine. That is contrary to our first hypothesis. We believe that there are many factors for this result. Because of Covid-19 pandemic, businesses shut down and many residents lost their jobs, resulting in a rise in the number of property crimes (Part 1) (42% increase in the number of larceny and vandalism crimes between May 2020 to August 2020). Furthermore, during the first couple months of the stay-at-home advisory, many people were unable to see their friends and loved ones or take part in everyday activities, taking a heavy toll on many residents' physical and mental health, which caused a double increase in simple assault (Part 2 crime) reports to Boston Police (from April 2020 to July 2020).

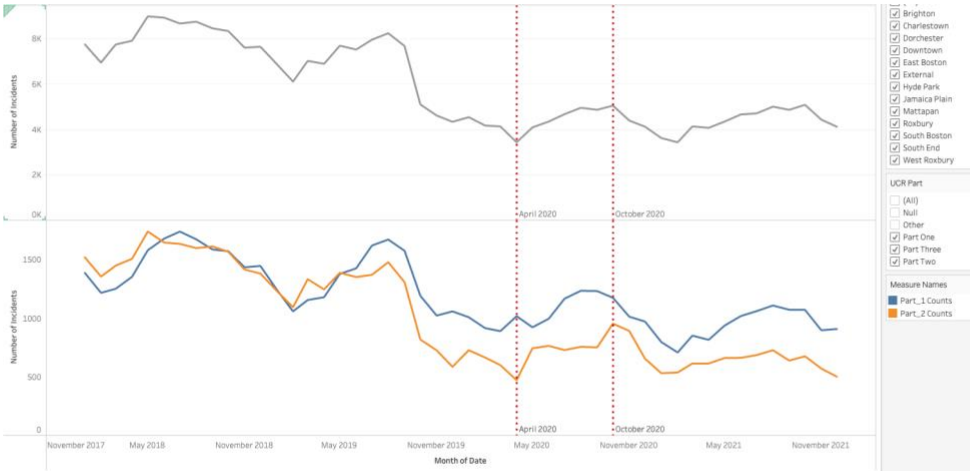


Figure 10: Incidents in Boston during Covid pandemic analysis

- **Crimes related to Shooting analysis**

Aggravated Assault is the most popular crime in which shootings occurred (Figure 11). Given the fact that the number of this crime reports is not high compared to other crimes, at only 3.04% of the total incident reports from 2018 to 2021, the number of aggravated assault crimes having shootings account for 32% of total number of crimes with shootings. We believe that aggravated assault is violent attacks between gangs and is in Part 1 (the most serious crimes), so shootings are more likely to happen.

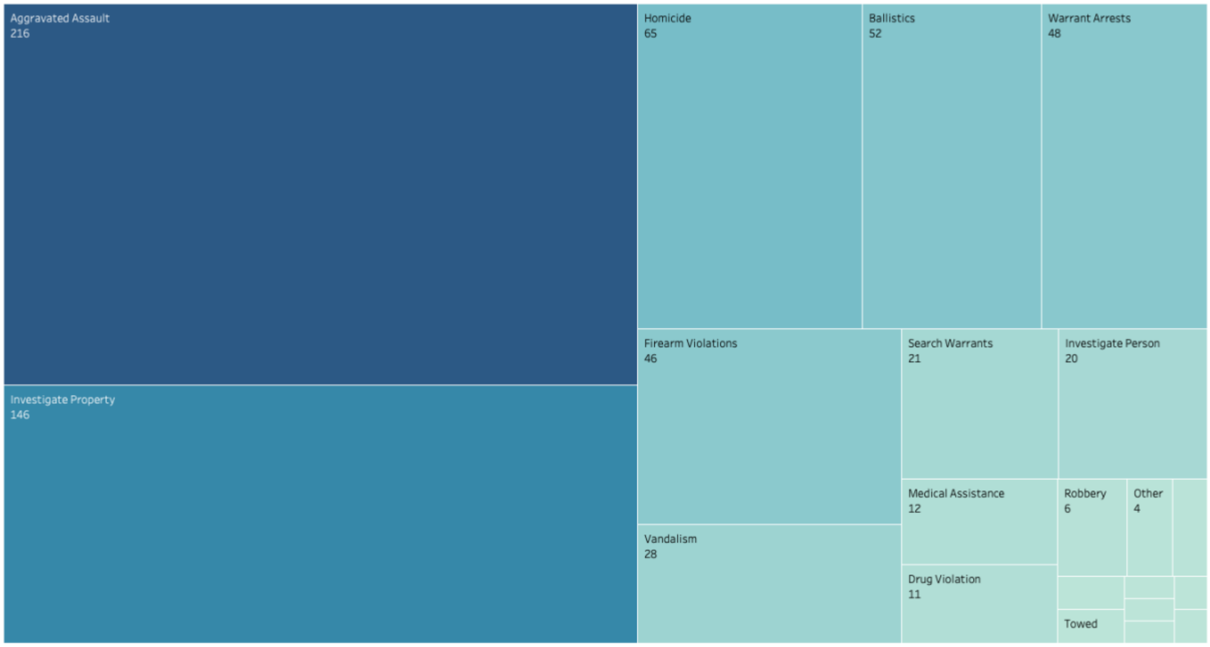


Figure 11: Shooting analysis by crime

As for time, the number of crimes with shooting are high from 9PM to 12PM, reaching a peak at 9PM (Figure 12) and in November and December, peaking in November.

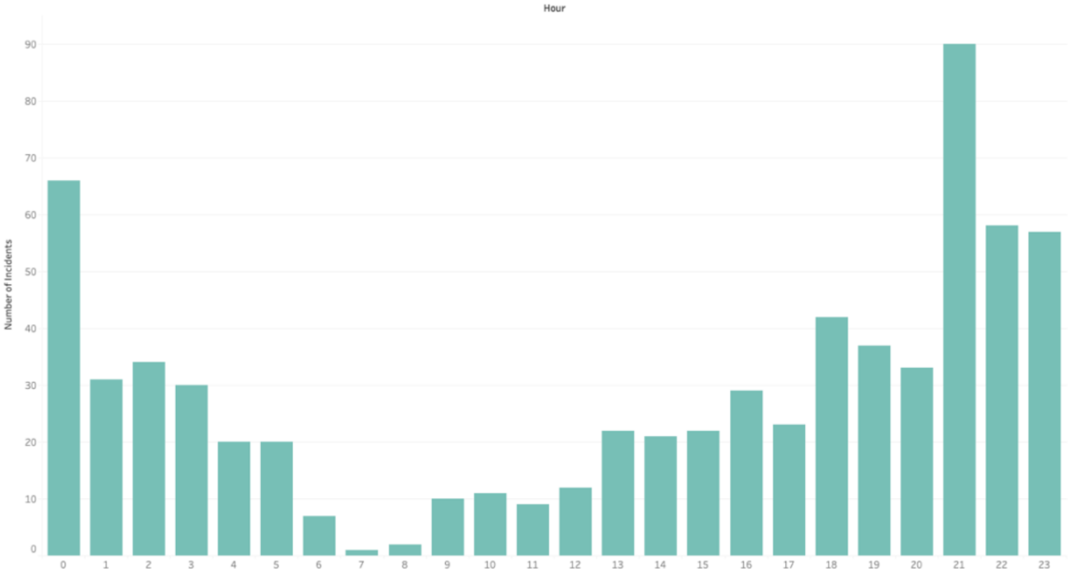


Figure 12: Shootings by Hour

The shooting rate is high in three low-income neighborhoods, consisting of Roxbury, Dorchester and Mattapan (Figure 13), where aggravated assault and larceny usually happen. Blue Hill Avenue and Washington street are two major hotspots of shooting. These are two long main streets, stretching through many neighborhoods, including Roxbury, Dorchester and Mattapan. Therefore, the high number of incidents having shooting on these streets is reasonable.

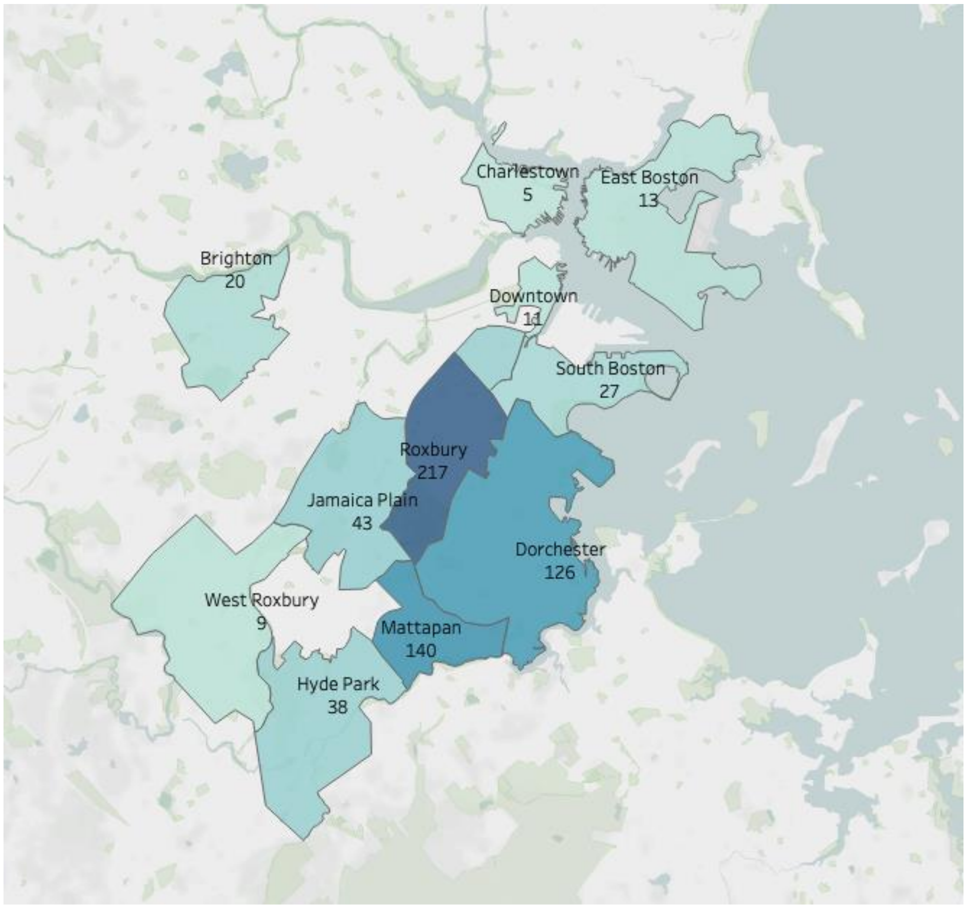


Figure 13: Shootings by District

- **Crime Analysis of Area Near Northeastern University (0.5 mile radius):**

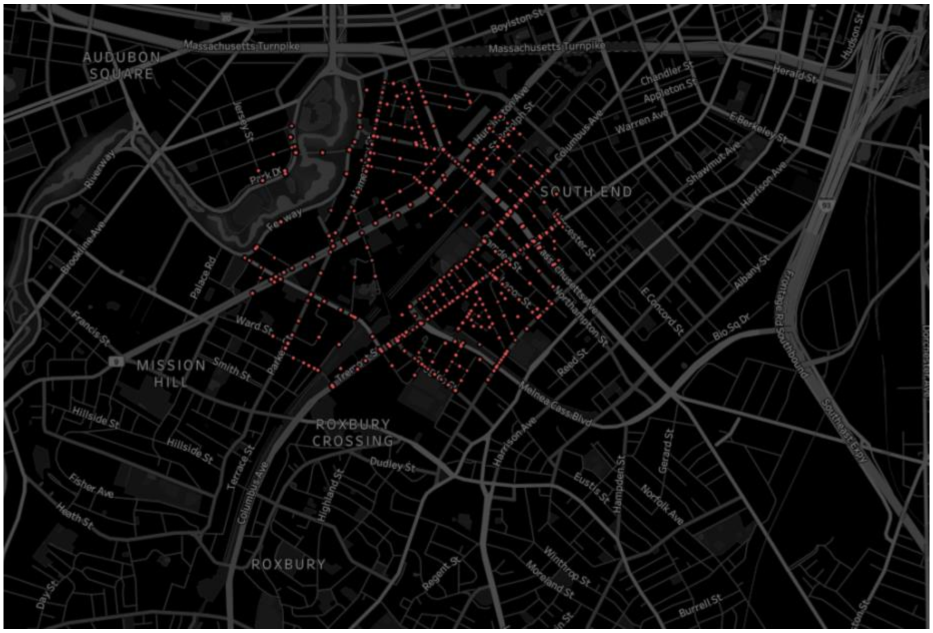


Figure 14: Incident reports in the area near NEU

The area near Northeastern University is quite safe, as only 2.78% of total number of incidents reported in this neighborhood. The major crimes that occur around Northeastern university are Fire, Burglary, Harassment, and Assault. Drug-related crimes are also high. The majority of crimes occur from 16PM to 18PM, when the traffic is busy and many pedestrians walk around the school, which has the similar pattern to the whole city. From 2018 to 2021, there were only 20 cases, accounting for 2.91% of total crimes where shooting occurred.

5. Limitations and Future Works

5.1. Limitations

- The number of missing values in the Shooting column is rather high (33%), and we assumed all these missing values mean no shooting. Therefore, our shooting analysis may not be fairly evaluated.
- A lot of Latitude and Longitude values were missing in the dataset, we had to impute those values by Neighborhood Criteria, making geographic analysis of street a little inaccurate.

5.2. Future Works

- Our current dataset is not the most updated, just from 2018 to 2021. Hence, we will update the 2022 dataset when we have all incident reports of 2022.

References

[1] James Jennings. *A Select Overview of Poverty in Three Boston Neighborhoods: Roxbury, Dorchester and Mattapan,* Report Prepared for The Center for Church and Prison, Inc. Dorchester, Massachusetts 2019