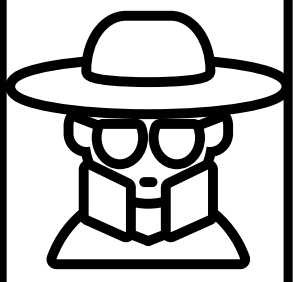
STAT 183

Introductory Data Analytics

THE DESCRIPTIVE ANALYTICS PROJECT

John Robert Torres



Carl Jacob Villar

Ma Andrea Villanueva

AB-1L

ORGANIZATION'S UNDERSTANDING

There is no other city that is filled with so much history as Boston. Apart from its history, it is a city renowned for its sports, universities, culture, and people. In relation, Boston is also one of the major cities of the USA and houses a population of diverse ethnic backgrounds. However, like the majority of cities, it also experiences a high crime rate (Cove, n.d.).

Boston's crime index ranking of 18 indicates that it is safer than just 18% of American cities. The figure is startling in light of the fact that 89.66 percent of Boston's workforce is working in white-collar professions, far higher than the national average (Boston Indicators, n.d.). However, the city's crime data are unusually high and contradict the city's economic significance within the country. Further, various statistics pertaining to Boston's crime incidents have been compared with the country's average figures in numerous instances. The studies specifically highlight the fact that the crime rate in Boston is considerably higher than the national average, and Boston has a higher proportion of violent crimes per 1000 crimes ("Boston, MA Real Estate & Demographic Data").

According to Crime Grade (n.d.), crime reports are separated into two categories: violent crime and property crime. These two groups account for the vast majority of crimes committed in cities and towns. Murder, homicide, assault and battery, manslaughter, rape, sexual assault, robbery, and abduction are all considered violent crimes. Meanwhile, theft, burglary, larceny, motor vehicle theft, arson, shoplifting, and vandalism are all considered property crimes. Although property crime is regarded as the less dangerous of the two, communities with a high rate of property crime should not be considered safe as well. Theft is the most frequently reported property crime, accounting for over 12,000 crimes per year. Burglary is also high, with more than 2,000 burglaries recorded. While property crime is lower in Boston than the national average, it remains significantly high. The Boston Globe also reported that crime in Boston climbed slightly in 2019. But at the same time, they also stated that the city has had the "lowest crime rate in history for four straight years." This means that even though there has been a rise in crime this year, it is still lower than it was in 2015.

Given the information, it is evident that there is indeed a high variability in the crime scene in Boston, which directs towards an important need to analyze crime trends. Accordingly, the city's crime data from 2015 to 2018 must be analyzed and evaluated to comprehend the prevalence of crime in Boston. Such analysis is critical since it will assist in evaluating the significance of underlying factors in determining the crime rates depending on the time and space of Boston and developing prevention strategies accordingly. For example, crime is hyper-localized to certain areas. The best way to ensure safety in Boston is to avoid visiting risky neighborhoods or traveling at night. But on the brighter side, Barresi (2021) stated that some areas of Boston that once had a reputation for crime have been completely transformed by revitalization and proper resource deployment. Understanding the perspectives of crime happenings by exploiting crime data helps early detection, and prevention of crime.

Overall, conducting a crime study in Boston will benefit both residents and the government in identifying and analyzing trends in the city's crime incidents. The findings of this research will not only aid law enforcement agencies in their resource deployment and in identifying and apprehending individuals but will also be beneficial for the residents of Boston to see and understand where and how often crime is happening in their neighborhood.

DATA UNDERSTANDING

BOSTON CRIME DATASET

The Boston Police Department (BPD) provided a crime incident report that was recorded from June 14, 2015, to September 3, 2018. The dataset includes 319,073 instances and 17 attributes in which data about the incident numbers, offense descriptions, districts, locations, and the exact date and time of occurrence of the crime incidents were recorded. The dataset was retrieved at the Kaggle website and can be accessed through the link: Crimes in Boston | Kaggle. Table 1 below shows the data type, level of measurement, and data description of the 17 variables.

Table 1. Data Dictionary of the Boston Crime Dataset

Variable Name	Variable Type	Level of Measurement	Variable Description
incident_number	qualitative	nominal	Unique number assigned to an offender
offense_code	qualitative	nominal	Numeric code for each offense classification
offense_code_group	qualitative	nominal	The grouped classification of offenses
offense_description	qualitative	nominal	More specific type of offenses in the grouped classification
district	qualitative	nominal	Code for districts in Boston
reporting_area	qualitative	nominal	Geographic area for which offenses are taken and recorded
shooting	qualitative	nominal	Incidence of shooting (Y - Yes, NA - No)
occured_on_date	qualitative	ordinal	Exact date and time of crime occurrence
year	qualitative	ordinal	Year when the offense was recorded (2015 - 2018)
month	qualitative	ordinal	Month of the year when the offense was recorded (1 - 12)
day_of_week	qualitative	ordinal	Day of the week when the offense was recorded

			(Sunday - Saturday)
hour	qualitative	ordinal	Hour when the offense was recorded (0 - 23)
ucr_part	qualitative	nominal	Stands for Uniform Crime Reporting that divides the offenses into parts or groups depending on the intensity of crime (Part One, Part Two, Part Three)
street	qualitative	nominal	Name of the streets in Boston
lat	quantitative	interval	Latitude of the location of crime
long	quantitative	interval	Longitude of the location of crime
location	quantitative	interval	Latitude and longitude of the location of crime

In the crime in Boston dataset, some of the attributes have missing data values as shown in table 2 below. It can be noticed that shooting has the highest number of missing data followed by reporting area, latitude, longitude, and lastly, district. However, in the case of the shooting variable, the column includes Y which means that the shooting incident happened as reported and Null or missing data is treated as no shooting incident happened in the crime report.

Table 2. Variables with Missing Data

Variable Name	Number of Missing Data
shooting	318,054
district	1,765
reporting_area	20,250
lat	19,999
long	19,999
street	10,871
ucr_part	90

DATA PREPARATION

Data profiling was performed to understand and identify patterns, relationships, and other attributes in the dataset. After identifying the issues, data preparation, including data cleansing, data transformation, and data enrichment, was conducted in order to address the inconsistencies, missing values, and other issues in the dataset. First and foremost, column names were converted to lowercase. Next, variables were converted to factors for better summary results. District code was also coded according to its district name since it is easier to analyze and interpret the results using the name rather than the code. Also, the values of the shooting variable were renamed for better analysis. Furthermore, as part of cleaning the Boston dataset, those with missing data points in the district variable were removed to create consistency within the dataset. Also since they have a missing observation of approximately 2000, which is significantly lower than the size of the dataset. Moreover, variables that were not used in the analysis were also dropped.

In relation to data transformation and enrichment, some of the variables were collapsed into categories. For instance, time intervals were generated wherein six points of the day were generated to bin the day into four equal segments, namely,

early morning (1 AM to 6 AM), morning (7 AM to 12 PM), afternoon (1 PM to 6 PM), and evening (7 PM to 12 AM). The crime count per quarter of the year was also determined by generating a variable where months were collapsed to form the four seasons of the year, namely, summer (June to August), fall (September to November), winter (December to February), and spring (March to May). After data cleaning and transformation, the resulting dataset now has 317,308 rows and 13 variables.

MODELING/TOOLS

Following the data preparation, various data visualization techniques were applied to comprehend the data and identify the trends portrayed by the crime data.

DISTRIBUTION OF THE TYPES OF OFFENSES

The final dataset contains 317,308 total offenses from 2015 to 2018. These offenses were committed by 280,934 unique offenders. This results in an average of 1.13 offenses per offender in Boston from 2015 to 2018. The maximum number of offenses per offender is 13.

The majority of the offenses recorded were related to Motor Vehicle Accident Response (15.92%), Larceny (11.23%), and Medical Assistance (10.20%). On the other hand, some of the offenses that were recorded less than 100 times include offenses related to Manslaughter, Gambling, Human Trafficking, Arson, Home Invasion, and Biological Threats. Figure 1 shows the top 15 major offenses recorded.

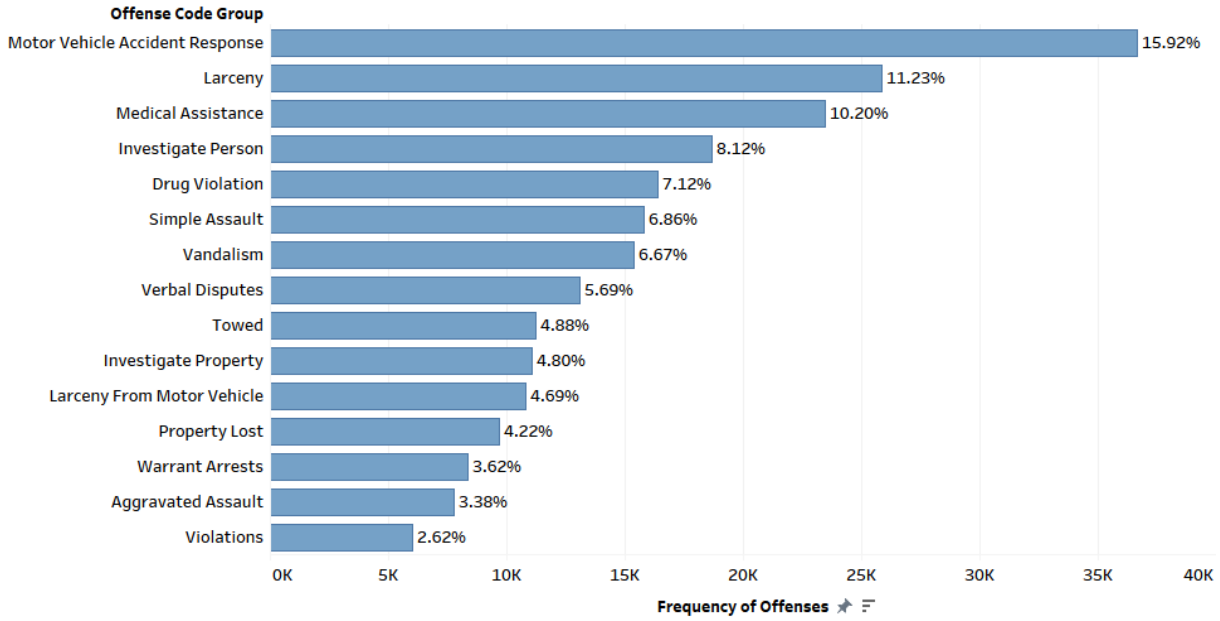


Figure 1. Distribution of the Top 15 Major Offenses in Boston

DISTRIBUTION OF THE NUMBER OF OFFENSES BY DISTRICT

To have an idea about which part of Boston the offenses were coming from, the recorded offenses were grouped according to the district where the offense took place. The districts of Roxbury (15.74%), Dorchester (13.40%), and South End (13.21%) recorded the highest number of offenses out of the 12 districts of Boston. Meanwhile, the districts that recorded the least number of offenses were Charlestown (2.05%), West Roxbury (4.17%), and East Boston (4.27%). Looking at the map of Boston, the districts with the highest number of offenses were the districts at the center of the city with a higher number of people, while the districts with the lowest number of recorded offenses were from the city borders lying beside other city or the Atlantic Ocean.

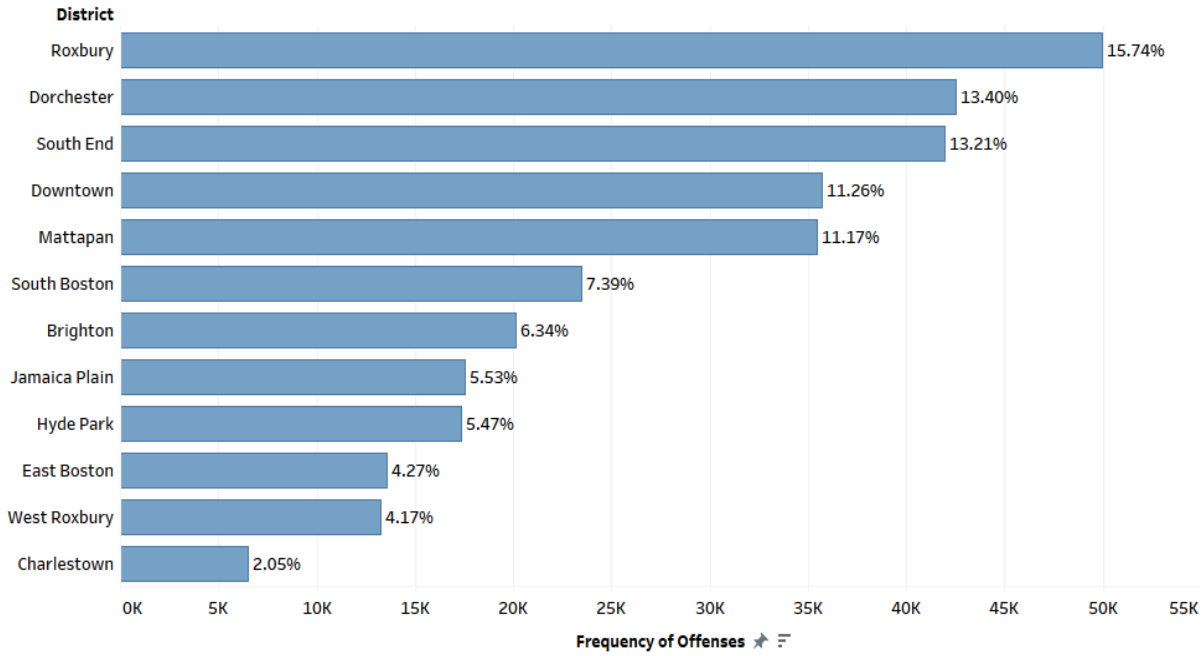


Figure 2. Distribution of the Number of Offenses by District

Furthermore, the next figure shows the distribution of offenses by district broken down by the top 15 offenses identified in the previous section. This allows the identification of the offenses with the highest number of records per district. For example, offenses related to Motor Vehicle Accident response, the most frequent offense in Boston, occur more frequently in the districts of Roxbury and Dorchester compared to other districts. Most cases of Larceny, however, were recorded from the districts of Southend and Downtown.

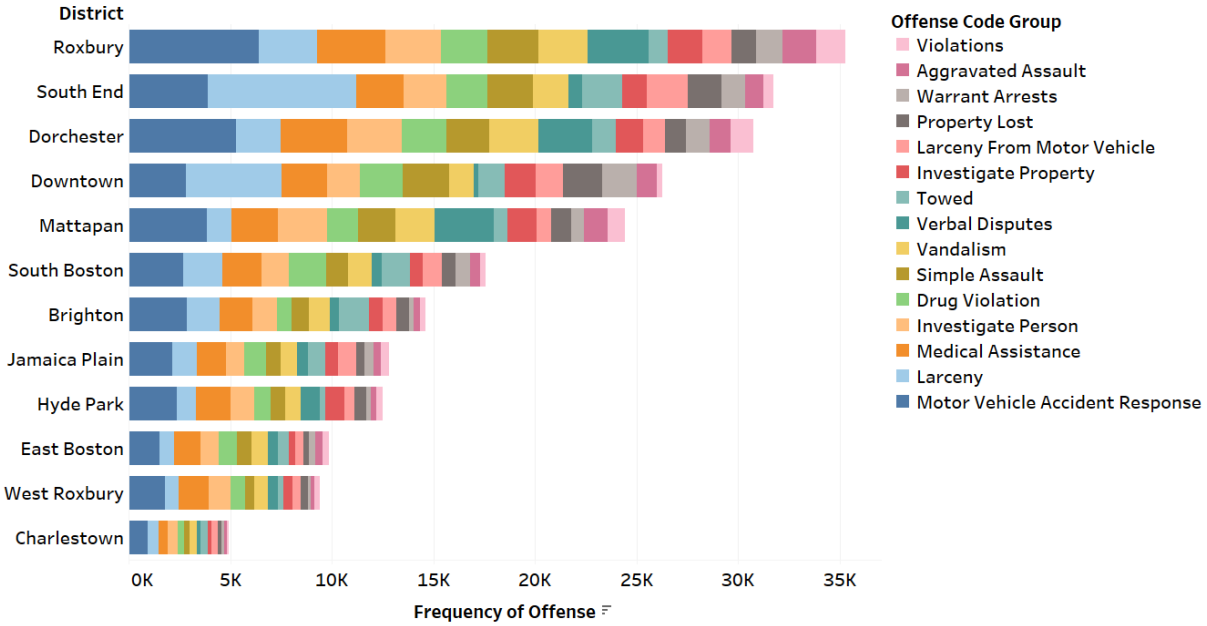


Figure 3. Distribution of the Top 15 Major Offenses by District

DISTRIBUTION OF OFFENSES BY DATE AND TIME OF OCCURRENCE

The occurrence of offenses was also broken down per year from 2015 to 2018. About 31% of the cases were recorded in 2017 and 2016, while about 20% and 16% were recorded in 2018 and 2015, respectively. However, it is important to take note that the distribution per year is like this because the dataset does not contain entries earlier than June 2015 and later than September 2018. If full-year records will be available, a more balanced distribution can be expected.

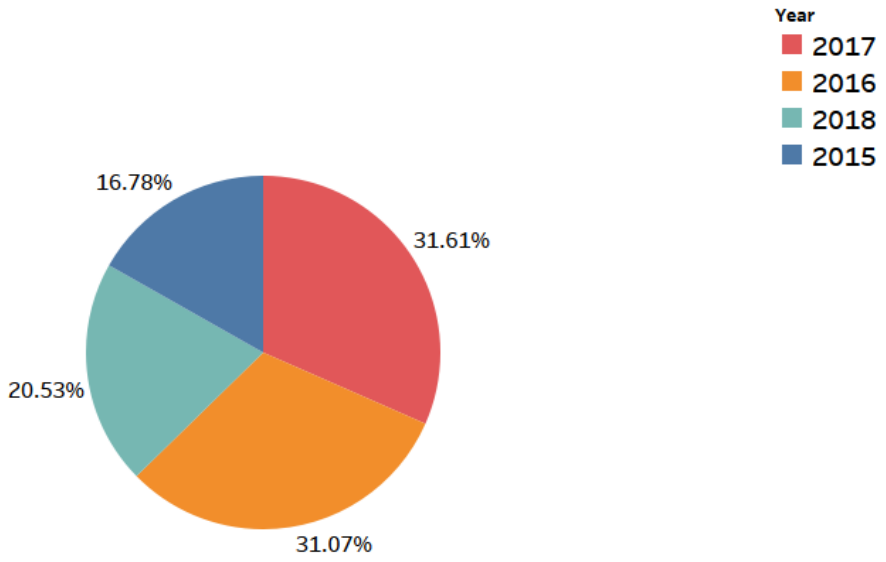


Figure 4. Distribution of offenses recorded per year

In addition to the yearly distribution, the top 15 major offenses were also broken down into yearly data to visualize the yearly distribution of each offense. It can be observed that the majority of each offense was recorded from 2016 to 2017 at which the same justification still applies.

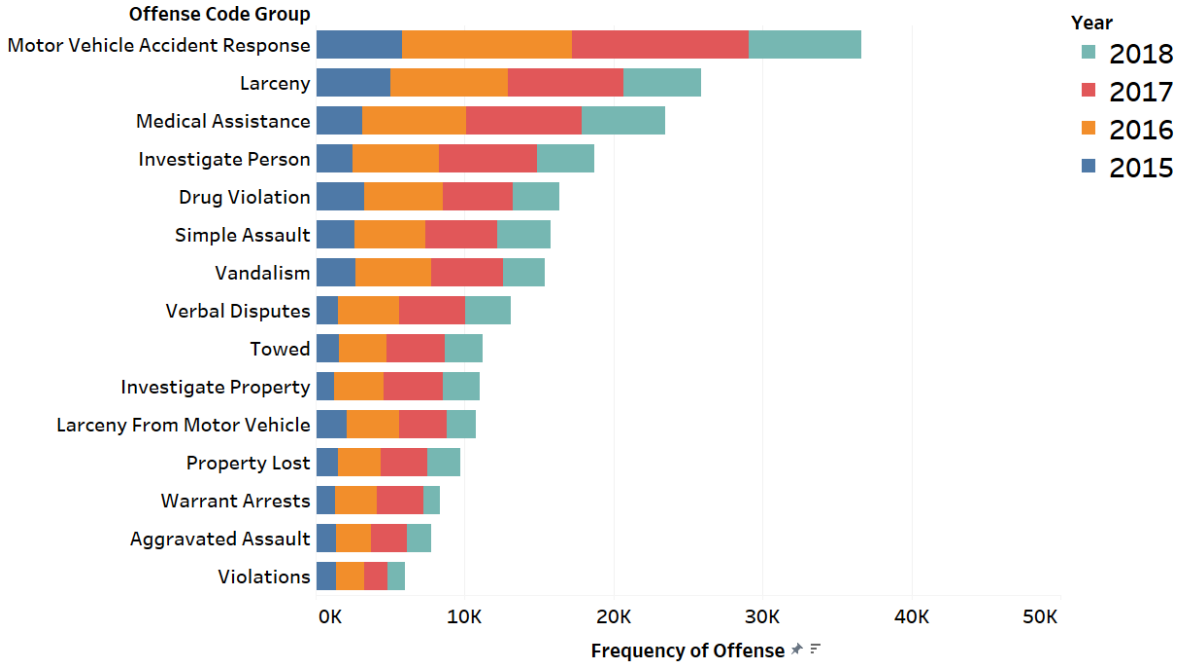


Figure 5. Distribution of the Top 15 major offenses by year of occurrence

Figure 6 shows the frequency distribution of offenses throughout each year. It can be observed that those four years almost have the same pattern all throughout the year except for the unrecorded offenses in 2015 and 2018.

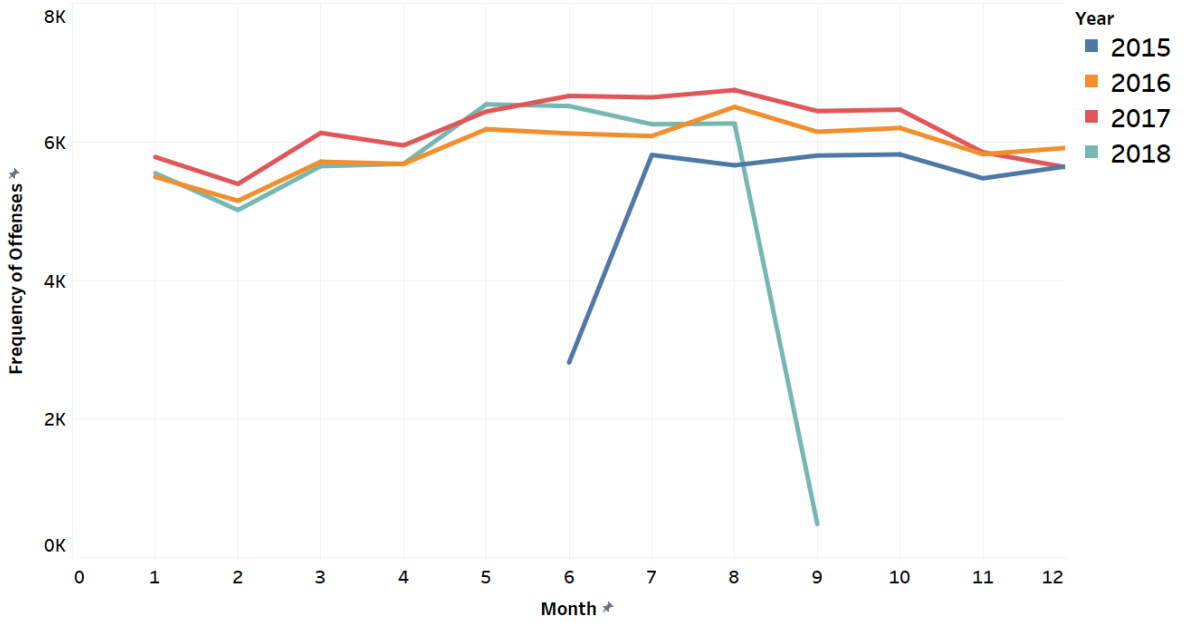


Figure 6. Distribution of offenses throughout one year

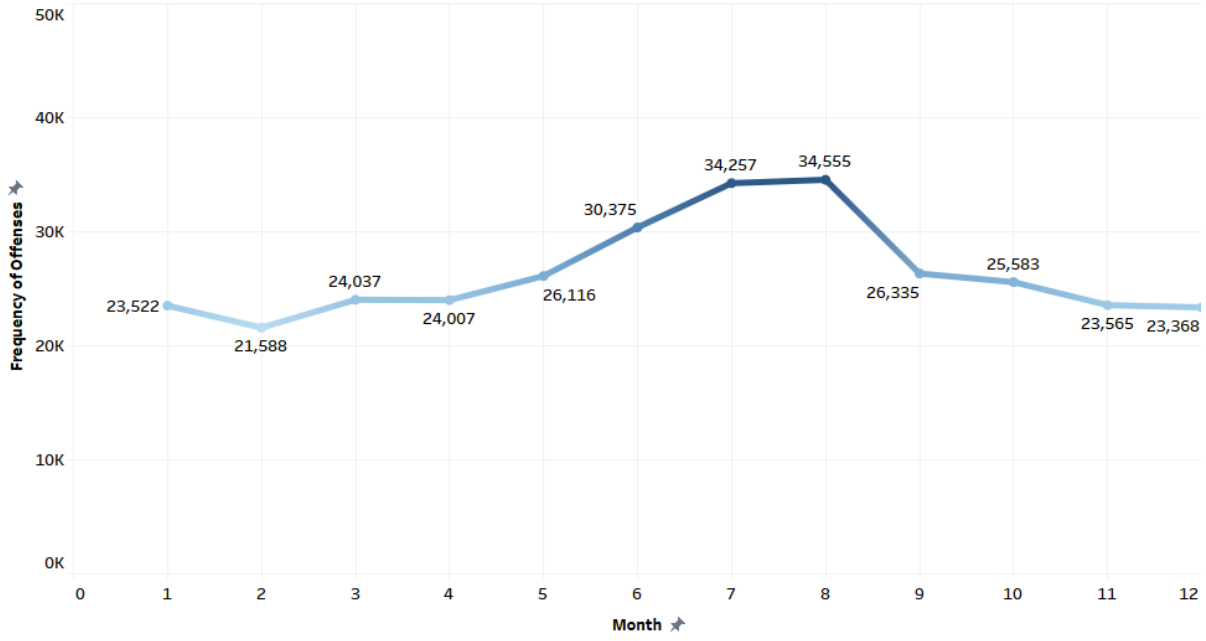


Figure 7. Distribution of offenses by month of occurrence

Figure 7 shows more details of the seasonality of offenses throughout the year. From the figure, it can be inferred that offenses are most frequently recorded in the months of June, July, and August (summer season) while the least frequent offenses were recorded in the cold months of December to February during the winter season. Figure 8 supports the previous figure that the most number of offenses occurred during summer, while the least occurred during winter.

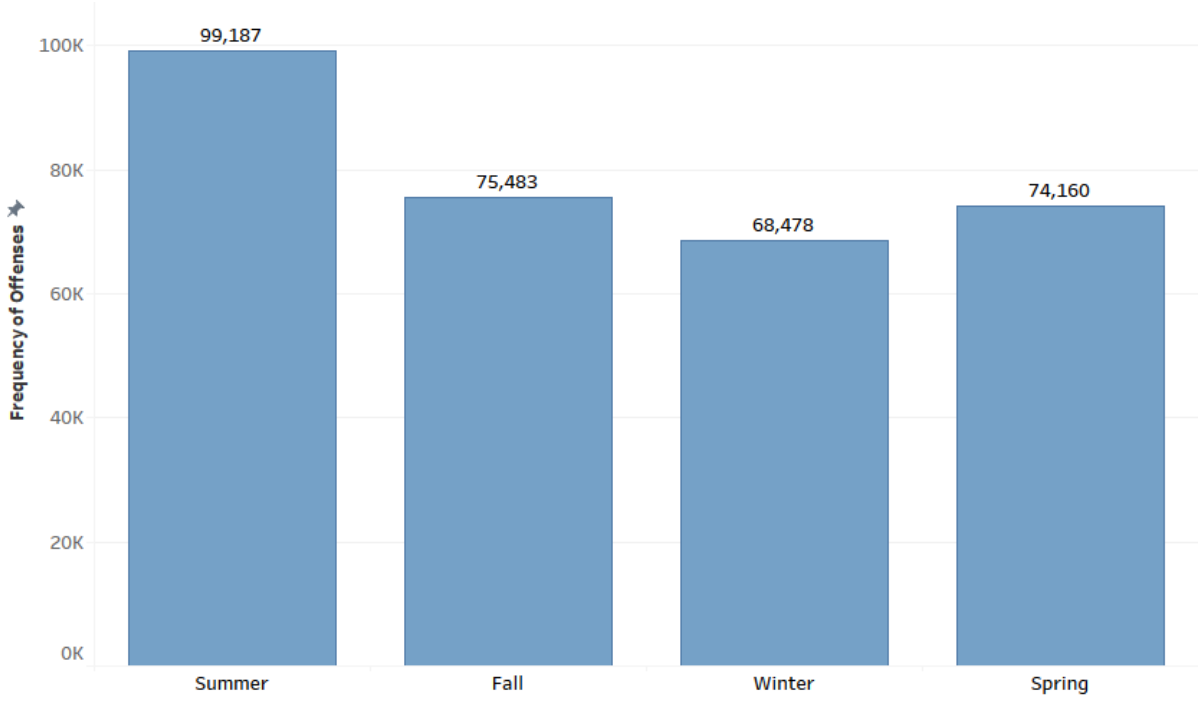


Figure 8. Distribution of offenses by season

Weekly and hourly data were also available in the records. This can be used to identify a possible relationship between a day within the week or a time within the day and the frequency of committed offenses. Figure 9 shows the general trend of the number of offenses committed hourly during all days of the week. All the days of the week show almost the same pattern. However, it can be noticed that the most number of offenses occur between 4 PM and 6 PM, while the least number of offenses occur between 4 AM and 6 AM.

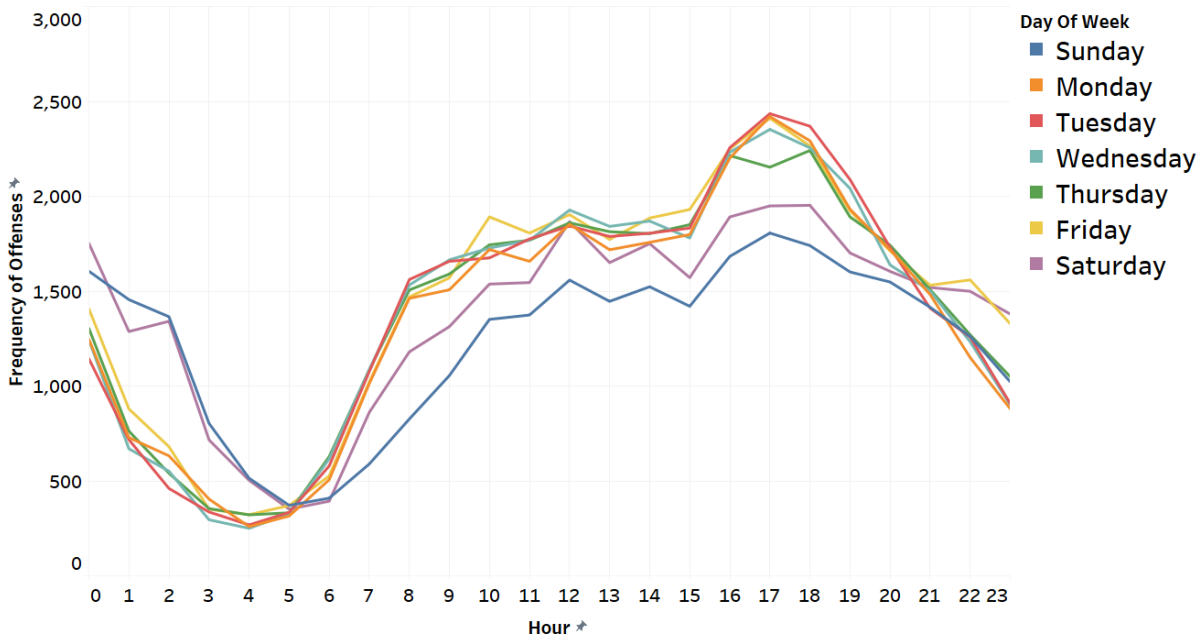


Figure 9. Distribution of offenses throughout one day

Figure 10 shows more detail on the trend of offense occurrence throughout one day. The figure peaks around noon (12 PM) and the afternoon (4 PM to 6 PM). Meanwhile, low values were observed during the early morning.

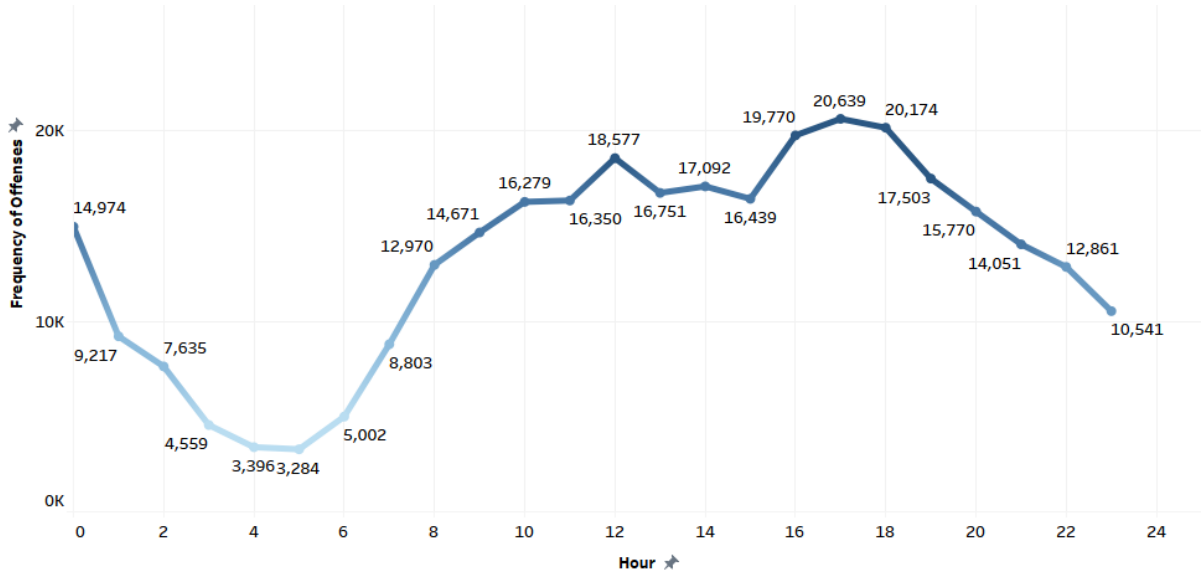


Figure 10. Distribution of offenses by the hour of occurrence

In addition, the distribution per time segment within one day is shown in Figure 11. It was interesting to observe that most of the towing incidences occurred in the Morning (7 AM to 12 PM) and most drug-related offenses occurred in the Afternoon (1 PM to 6 PM).

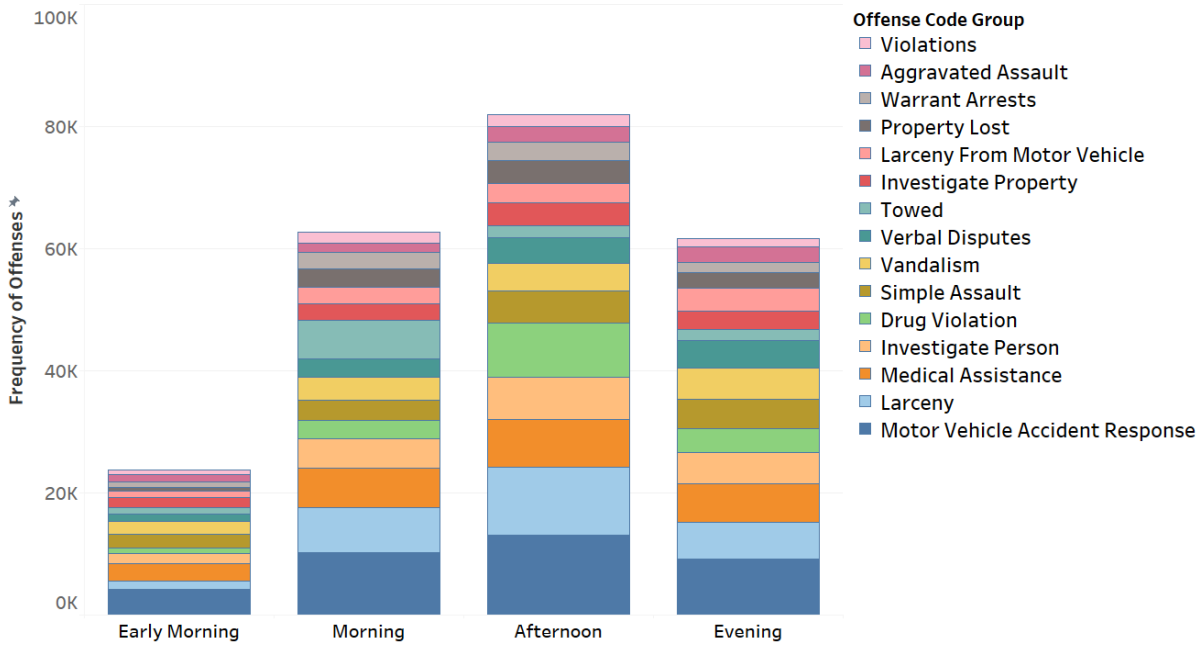


Figure 11. Distribution of the top 15 major offenses by the time of occurrence

SHOOTING INCIDENCES

From the 280,934 total offenses in the record, only 625 (0.22%) were recorded for having a shooting incidence during police operations. These incidences can be broken down into offense groups and districts to have an idea of which offenses are involved in frequent shooting incidences or the Boston districts related to a high number of shooting incidences.

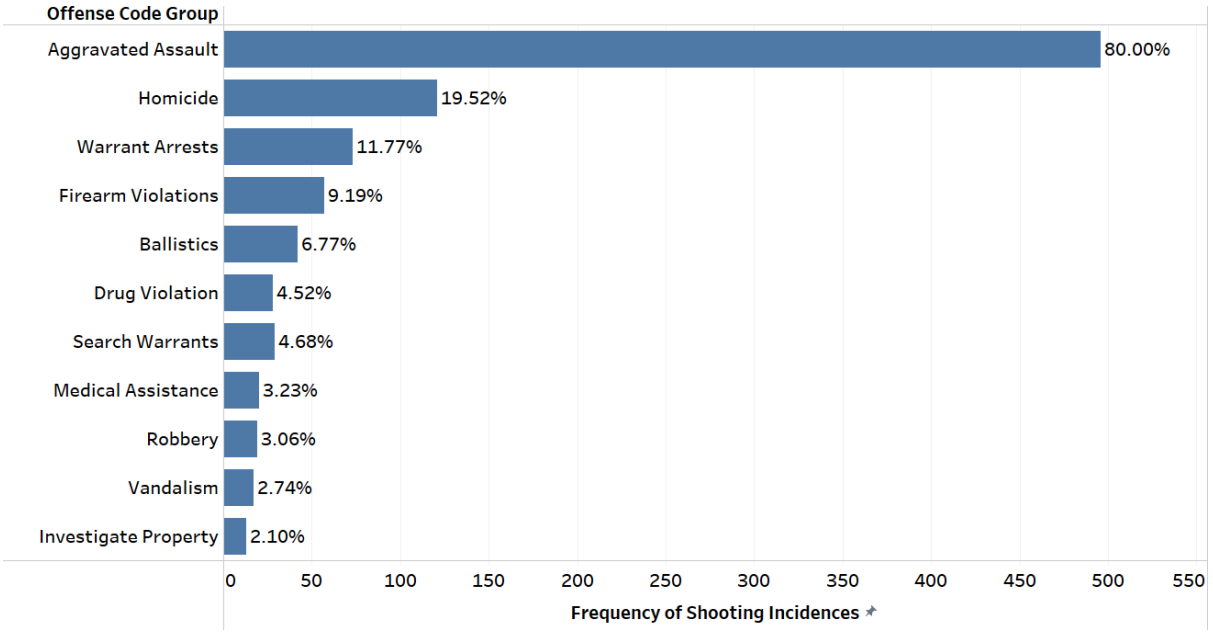


Figure 12. Types of offenses with more than ten shooting incidences

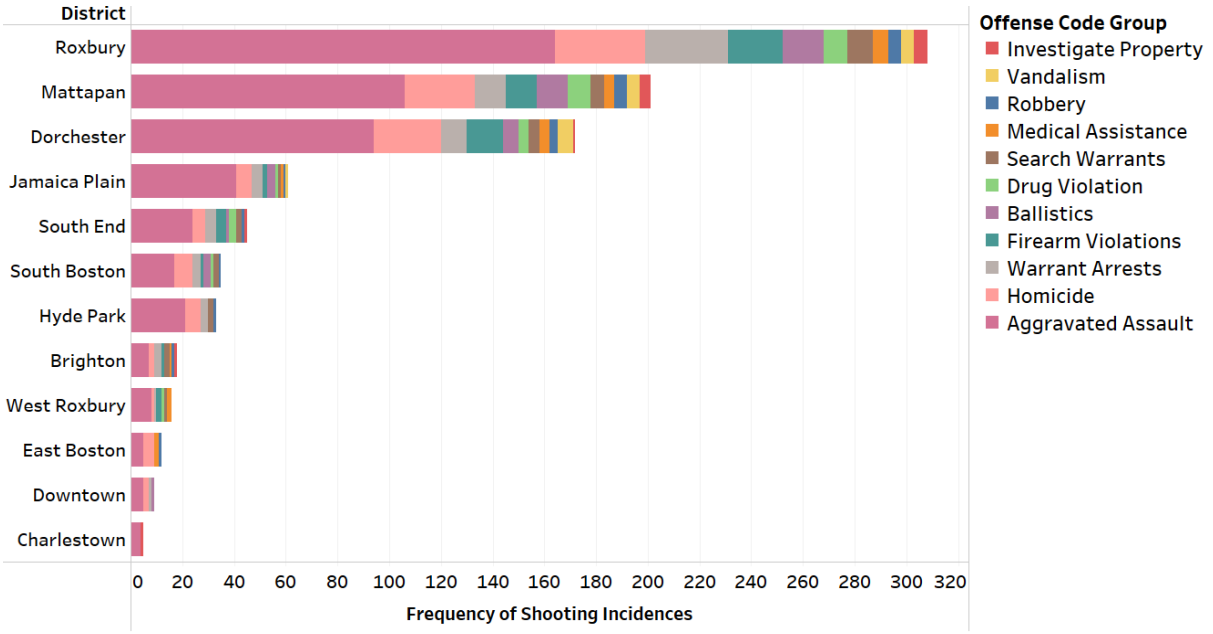


Figure 13. Distribution of offenses involving shooting per district

Figure 12 shows the offenses that have more than ten shooting incidences. The majority of the shooting incidents are linked to crimes operations related to Aggravated Assault (80.00%), or the unlawful attack by one person upon another for

the purpose of inflicting severe or aggravated bodily injury. It was followed by Homicide incidences (19.52%) and operations under Warrant of Arrest (11.77%).

On the other hand, Figure 13 shows the number of shooting incidents that occurred per district and the offenses that are related to it. The most frequent shooting incidents occurred in the districts of Roxbury (31.45%), Mattapan (21.77%), and Dorchester (19.52%) and the majority of which are related to Aggravated Assault (15% - 26%), followed by Homicide (4% - 6%), and Warrant Arrests (2% - 5%).

EVALUATION

Performing an in-depth descriptive analysis of the Boston Crime dataset enabled us to clearly see the trends and relations between the types of crimes, location, and the occurrence of the crime. In this analysis, the major crimes, statistics of UCR activities, and crime trends in specific locations and times were also analyzed.

In the first 3 figures, the distribution of the offenses by district was analyzed. The results show that the majority of the offenses were related to Motor Vehicle accidents, Larceny, and Medical Assistance. The districts of Roxbury, Dorchester, and South End in Boston recorded the highest number of offenses out of the 12 districts. Knowing the major crimes committed and the location of these offenses will be valuable in identifying target areas for crime prevention programs of the local enforcement agencies.

On the other hand, figures 4 to 8 show the date of occurrence of offenses by district. In the years 2016 and 2017, the top 15 major offenses were recorded. Although, the fact that the dataset does not include the offenses earlier than June 2015 and later than September 2018, can not be ignored. In addition, it can also be noted that the most number of offenses are recorded during the months of July and August or the summer season, and the lowest during December to February which is the winter season. The information that can be gathered from this analysis will be useful to learn if there is an indication of seasonal patterns or a relation between

crimes and community habits that will be beneficial in creating strategies to avoid or deal with the crime.

Meanwhile, figures 9 to 11 deal with the time of occurrence of crimes recorded. The outcome shows that the peak of offenses occurred between 4 PM to 6 PM or afternoon and the least during 4 AM to 6 AM or early morning. This information is crucial to Boston’s enforcement agencies in terms of making personnel allocation plans.

Figures 12 and 13 show the shooting incidences recorded during police operations. It can be noted that the majority or 80% of these incidences were related to Aggravated Assault. The knowledge gained from this analysis will be beneficial in the allocation of resources and preparedness of law enforcement in dealing with these types of offenses.

Overall, these findings provide a positive view on the issue at hand and assist in the development of different solutions that may be employed to reduce crime in Boston. This information may also be utilized for a variety of reasons, depending on the unique needs of the concerned stakeholders. For instance, the Boston Police Department would be interested in gaining insights in order to implement more effective crime control measures. Similarly, the insights may also support the business community in enhancing its security based on an area's crime statistics. Additionally, the information may assist real estate investors in determining which areas of Boston to invest in based on safety and cost considerations.

SUMMARY OF R CODES AND DATASET

```
# for cleaning
# libraries
library(tidyr)
library(dplyr)
library(janitor)
library(writexl)

# loading raw dataset
raw_crime =  read.csv('crime.csv', sep
    = ",", na.strings =
    c('','NA','na','N/A','n/a','NaN','nan'),
    strip.white = TRUE,
    stringsAsFactors = FALSE)
raw_crime = clean_names(raw_crime)
nrow(raw_crime)

# summary of missing data per column
sort(sapply(raw_crime, function(x)
    sum(is.na(x))), decreasing = TRUE)
summary(raw_crime)

# drop rows with missing district
crime = raw_crime %>%
    drop_na(district)
sort(sapply(crime, function(x)
    sum(is.na(x))), decreasing = TRUE)
nrow(crime)

# derive season from months
crime = crime %>% mutate(season =
                    ifelse(month %in%
    c(6,7,8), "Summer",
                    ifelse(month %in%
    c(9,10,11), "Fall",
                    ifelse(month %in%
    c(12,1,2), "Winter", "Spring"))))

# derive time of the day from hour
crime = crime %>% mutate(timeDay =
                    ifelse(hour %in% c(1,
    2, 3, 4, 5, 6), "Early Morning",
                     ifelse(hour %in% c(7,
    8, 9, 10, 11, 12), "Morning",
                    ifelse(hour %in% c(13,
    14, 15, 16, 17, 18), "Afternoon",
    "Evening"))))

#rename values of shooting
crime$shooting = crime$shooting
    %>% replace_na("No")
s_name = c(No = 'No', Y = 'Yes')
crime$shooting =
    as.factor(s_name[(crime$shooting)]
    )

# rename district name
district_name = c(A1 = 'Downtown',
    A15 = 'Charlestown',  A7 = 'East
    Boston',
     B2 = 'Roxbury', B3 = 'Mattapan',
    C6 = 'South Boston', C11 =
    'Dorchester',
     D4 = 'South End', D14 =
    'Brighton', E5 = 'West Roxbury',
     E13 = 'Jamaica Plain', E18 =
    'Hyde Park')

crime$district =
    as.factor(district_name[(crime$distr
    ict)])

# drop unnecessary columns
crime = subset(crime,
            select = -c(offense_code,
    offense_description,
    reporting_area, occurred_on_date,
    street, location))
summary(crime)

# save as excel file
write_xlsx(crime, "final_crime.xlsx")
```

REFERENCES

Barresi, S. (2021, August 22). *Boston MA crime rate: is Boston safe? [data, stats, reports, map]*. Mass Bay Movers. Retrieved April 25, 2022, from https://massbaymovers.com/blog/boston-crime-rate/

Boston Indicators. (n.d.). *Boston's more diverse than you may realize*. Boston Indicators. Retrieved April 25, 2022, from https://www.bostonindicators.org/article-pages/2018/september/boston-diversity

Boston Police Department. (n.d.). *The Boston Police Department's virtual community*. bpdnews.com. Retrieved April 25, 2022, from https://bpdnews.com/?category=Crime%2BStats

Cove. (2022, April 11). *Boston crime rate: Is Boston a safe city?* Self-Installed DIY Home Security Systems. Retrieved April 25, 2022, from https://www.covesmart.com/blog/boston-crime-rate-is-boston-a-safe-city/

Crime Grade. (n.d.). *The safest and most dangerous places in Boston Metro, MA* . Retrieved April 25, 2022, from https://crimegrade.org/safest-places-in-boston-ma-metro/

Neighborhood Scout. (n.d.). *Boston, MA crime rates*. NeighborhoodScout. Retrieved April 25, 2022, from https://www.neighborhoodscout.com/ma/boston/crime