Boston Incidents Data Analysis Project

Authors:

Asta Adhira Anggono

Christina Tian

Keni Ding

Shuyan Wu

Jiheng Li

Yilin Li

Research Topic

In today's society, incident rates and public safety issues have always been a matter of great concern. With the development of society and population growth, criminal activities have become increasingly complex and diverse, thus requiring more data analysis to understand and address these issues. This makes incident statistics a vital tool in aiding the government and criminal justice professionals in understanding the incident situation and coming up with a corresponding strategy to prevent incidents from occurring, ultimately protecting social and human capital.

Boston is one of the most historic cities in the United States, known for its famous universities, rich cultural heritage, significant landmarks, and diverse population. Boston also has a comprehensive and reliable database of incidents, which includes information on the date, time, and location of each incident, which makes Boston a perfect choice for us and makes it easier to conduct a detailed analysis of the data and to identify any patterns or trends that may be present.

This report will focus on analyzing incident data in the city of Boston, the data collected by the Boston Police Department which explores all incident records from June 15, 2015, to December 31, 2022. Through analysis of the dataset, we will examine trends in incidents in Boston, incident occurrences at different times and locations. The aim of this report is to provide a comprehensive data analysis perspective, helping policymakers and public safety experts better understand the incident situation in Boston and develop more effective public safety policies and measures.

Research Questions

RQ1: How does the month of year, day of week, and hour of day affect the incident rate in Boston?

RQ2: Is there a clear trend or seasonal change in the number of incidents and the number of shootings over the past few years? Are there any significant changes on specific dates (e.g., holidays, hot events, etc.)?

RQ3: Is there a significant association between the occurrence of shooting incidents and different district locations and hour ranges of incidents?

RQ4: Are there certain types of incidents that tend to occur together with specific time or location of Boston?

Data Description

Our dataset is provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond. This dataset contains records from the new incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. After cleaning the dataset, it now has 614826 rows recording every reported incident from June 15th 2015, to December 31st 2022, as well as 23 columns describing different details of incidents (19 columns from the original dataset, 3 columns we created for further analysis) as following:

Columns from original dataset:

INCIDENT_NUMBER: unique ID given to the incident once it was reported
OFFENSE_CODE: digital representation of offense type
OFFENSE_CODE_GROUP: character representation of offense type
OFFENSE_DESCRIPTION: character string to describe the offense in detail
DISTRICT: digital representation of district in which the offense took place
DISTRICT_NAME: name of district in which the offense took place
REPORTING_AREA: area where the offense was reported
SHOOTING: either ‘Shooting reported’ or ‘Not reported’
DATE: date on which each incident occurs, in "%Y-%m-%d %H:%M:%S" format
YEAR: year component of date (from 2015 to 2022)
MONTH: month component of date (from 1 to 12)
DAY_OF_WEEK: day component of date (from Monday to Sunday)
HOUR: hour component of date (from 0 to 23)
UCR_PART: offense part (Part I, Part II, Part III)
STREET: name of street where offense took place
Lat: Latitude of where offense took place

Long:Longitude of where offense took place

Location: Longitude and latitude coordinates of where offense took place

OFFENSE_NAME: name of offense

Columns we create:

HOUR_RANGE: ranges for hour component of date (Midnight-4:00am; 4:01 am-8:00 am; 8:01 am-Noon; Noon-4:00pm; 4:01 pm-8:00pm; 8:01 pm-Midnight)

DISTRICT_PERCENTAGE: percentage of incident cases of each district *100

DOWNTOWN: whether incident occurs in 'Central Boston' (1, 0)

Rationale of Techniques & Analytical Procedures

RQ1

The first research question that we are trying to address is how does the month of year, day of week, and hour of day affect the incident rate in Boston. Given the data that was obtained and cleaned from the Boston Police Department includes daily to hourly occurrences of each incident, we have decided to go first with a time series analysis. Though the data of observations were recorded at regular intervals over time and were in chronological order with each observation corresponding to a specific time period, the data wasn't exactly in a time series structure. Changing the data to a time series structure was the first thing that we did. With time series, we can use it to make predictions about future values of a variable based on historical data. Moreover, we will be able to identify trends and patterns in the data, such as seasonal fluctuations of long-term trends. The reason why we chose time series analysis for solving the first question is to help identify patterns and trends in incidence rates at different times of the day, week, and year. It allows for better understanding of the factors that influence incident in Boston. We applied seasonality analysis and forecasting models of time series techniques to know cyclical patterns over time and do future predictions with data visualizations.

RQ2

Time series analysis is suitable for investigating if there is any clear trend or seasonal change in the number of incidents and the number of shootings over the past few years, as it is a statistical method used to analyze time-based data. It can help us to identify patterns or trends that change

over time, such as seasonal variations or long-term trends. By plotting the monthly total incident count and monthly total shooting count from 2015 to 2022, we can visually identify any patterns or trends in the data over time. We used the stl() function to decompose our time series into three components: seasonal, trend, and remainder. We then used the autoplot() function to create visualizations of our time series data, including the seasonal decomposition produced by stl(). This helped us understand and identify whether there is a clear trend or seasonal change in the number of incidents and shootings over the past few years. Additionally, we used different time-series models to make predictions of future incident and shooting distributions based on historical patterns.

To investigate if there was any significant change in incident and shooting occurrence on holidays compared to normal days, it was most efficient and visually clear to build a time-series dataframe with incident/shooting count during holidays and incident/shooting count during normal days and make a visual plot. We first created a date list of holidays (including Independence Day, Labor Day, Martin Luther King Day, Halloween, Thanksgiving, Christmas, New Year's Eve) and then extracted the incident count and shooting count on these holidays and combined them as a data frame. We then plotted the monthly average incident count and monthly average shooting count as line charts and overlapping bar charts of the incident count and shooting count on holidays on top of the line charts, from which we could see whether there were any spikes or dips in incident and shooting incidents on specific dates. This could help us identify whether certain holidays or events are associated with higher or lower levels of incident and shootings, thus giving BPD some insights on the incident situations these days.

RQ3

To investigate if there is a significant association between the occurrence of shooting incidents and different district locations and hour ranges of incidents, logistic regression is a suitable statistical method since it is specifically designed for analyzing binary outcomes, such as the presence or absence of shooting in this case. The dependent variable in the model is binary (‘Shooting reported’ or ‘Not reported’). The independent variables are the district locations and hour ranges of the incidents, which are categorical variables. Logistic regression can be used to model the strength and direction of relationship between these independent variables and the

binary outcome variable, as indicated by the magnitude and sign of the regression coefficients. The results of the logistic regression model can also provide information on which district locations and hour ranges are most strongly associated with the occurrence of shooting incidents. By evaluating the statistical significance of the coefficients, the model can determine whether the observed associations between district locations and hour ranges with shooting incidents are statistically significant, and not just due to chance.

RQ4

Basket analysis is relevant and suitable for the problem being addressed because it aims to identify associations or relationships between items or variables in a dataset. In this case, the code performs basket analysis on the "boston" dataset, particularly focusing on the "OFFENSE_NAME" column. This analysis can provide insights into the co-occurrence of offenses and potentially help in understanding the underlying factors or relationships among different offenses.

The code preprocesses the "Boston" dataset, selecting rows with no missing values and converting the relevant columns to the desired format (e.g., converting the "Date" column to date format and the "OFFENSE_NAME" column to a factor variable). In this case, the "basket" data box is created by grouping the data by "DATE" (an identifier for a specific time period). Each basket represents a combination of incidents reported during every ten minutes period. These rules will point to combinations of incidents that frequently occur together, as well as statistical measures like support and confidence. Support represents the proportion of the basket containing the antecedent and consequence of the rule, while confidence measures the likelihood of the consequence occurring given the antecedent.

At the same time, we wanted to find out whether there is a specific occurrence of incident type in different districts. We analyzed the frequency of different incident types in the dataset. By examining the offense code groups, we identified the top incidents that occurred most frequently in Boston. This allowed us to focus on the most significant incident types for further analysis.
We further explored the relationship between incident types and specific districts in Boston. We calculated the frequency of incidents by district, identifying the districts where certain incident types were more prevalent. We prefer to find the incident type that occurs more frequently than

300 in each district to see the association between district and incident type. Since some of the incident names occur less than two digits, we consider dropping these low frequency data for a better display of heatmap. By filtering out low-frequency incidents, we focused on the most prominent incident types in each district.

Visualizations were created to better understand the patterns of incident types across districts. A heatmap was generated to visualize the frequency of different incident types in each district. The heatmap provided insights into the distribution of incidents across Boston and highlighted any areas with higher incident rates or specific incident types.

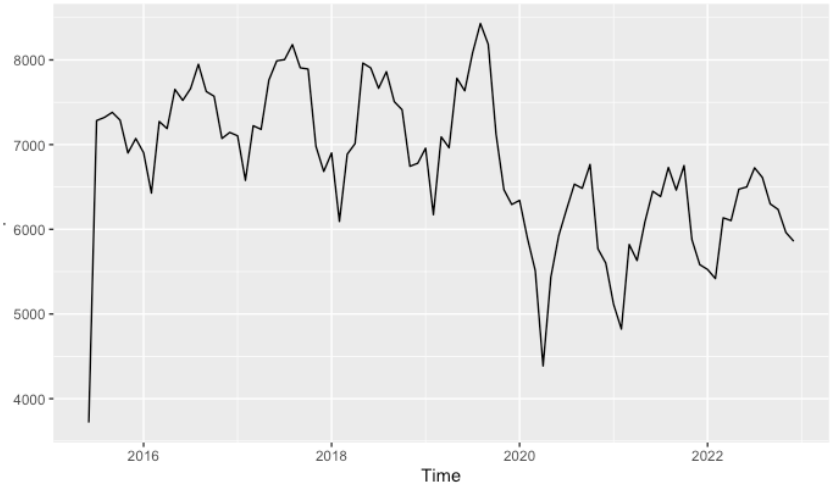
Continuing the analysis, we further examined the top incidents overall and investigated their frequency by district. By identifying the top 10 most frequent incidents, we gained a comprehensive understanding of the incident landscape in Boston. Additionally, we selected the top 5 incidents in each district based on frequency and visualized the results to identify patterns and variations in incident occurrence across districts.

By conducting this analysis, we have gained valuable insights into the occurrence of specific types of incidents with respect to time and location in Boston. This information can be used to inform law enforcement agencies, policymakers, and community organizations in their efforts to address and prevent incidents effectively.

Analytical Results

RQ1

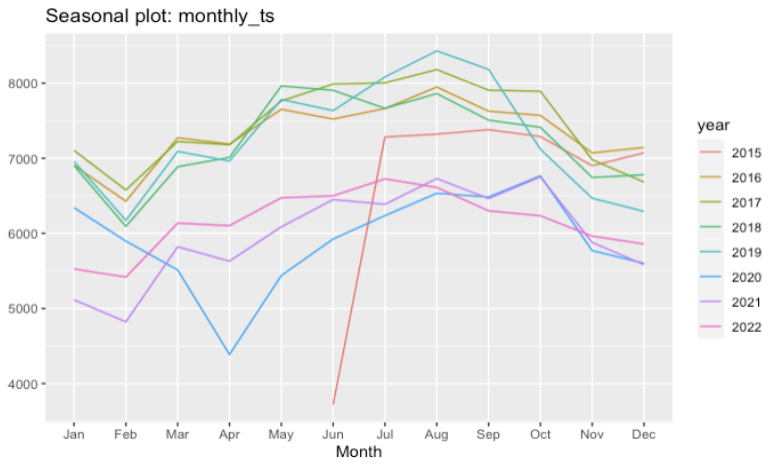
The first thing that we ran was to see the trend that spans over the historical data with simplifying the assumption that consecutive observations are equally spaced. Using the autoplot () function, we are given the graph 1 below.



Graph 1

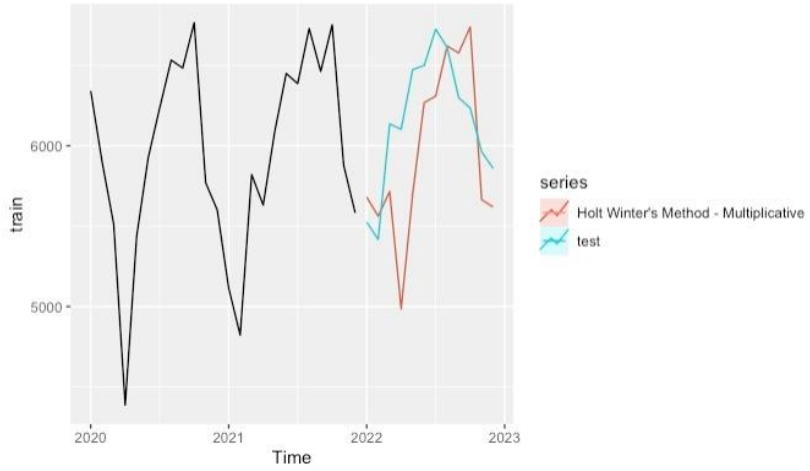
The graph shows a trend from June 2015 to December 2022 that there is a dominant seasonal pattern happening. It gives a fixed frequency from 2015 to the end of 2019 and drops a little bit from 2020 to 2022. It shows the insight that Covid-19 might influence the total occurrence of incidents because most people were quarantined at home.

Due to seasonality not being shown clearly from the graph attached we have decided to use gg seasonplot() to spot seasonality. Shown below is the seasonal plot (graph 2). This seasonal plot presents a relative fixed pattern for the changes of each period of years. During the summer months each year, incidents hit a high peak especially in August. On the contrary, incidents drop to the lowest point in February each year. It gives the results that incidents are more likely to happen in summer and less likely to happen in winter.



Graph 2

Next, we did forecasting. We split the data into train and test such that train data start from Jan 2020 to Dec 2021 and test data begins on Jan 2022 to Dec 2022. We tried several models and it turns out that Holt_Winter's Multiplicative Model gives the lowest RMSE. The visualized forecast is shown below (graph 3). Even though the RMSE still looks large in the train set, we believe that it is because of the decrease during Covid-19. Based on our modeling, the government can use the forecast on incident numbers to do planning and budgeting for the next few years and allocate resources effectively.

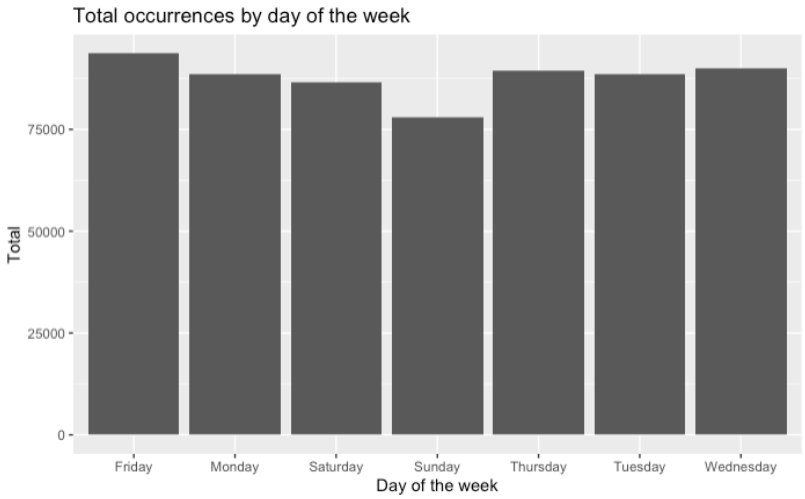


Graph 3

The second part of our first question is to look at how the day of week affects the incident rate. By aggregating the total occurrences and using 'weekdays ()' function and in R, it returns a dataframe (table 1) with the total number of occurrences corresponding to the specific weekdays from 2015 to 2022. Then we plot a bar chart (graph 4) to see individuals more clearly. It gives the result that Sunday has the lowest number of occurrences. Besides Sunday, other days of week show average numbers of incident occurrences compared with each other. We assume that people may prefer to stay at home and there are reduced business activities on Sunday. Sunday is also traditionally a day of reflection in many cultures and religions. People may be more inclined to avoid activities that involve risk or danger.

DayOfWeek <chr>	TotalOccurrences <int>
Friday	93636
Monday	88656
Saturday	86650
Sunday	77929
Thursday	89402
Tuesday	88554
Wednesday	89999

Table 1



Graph 4

The third part of our first research question is to look at how hours of days affect incident rates. We used the 'ifelse' function to assign the value "Day" if the hour is between 8AM and 8PM and " Night" if the hour is outside of that range. Then we aggregated the data by time of day and summarized the total number of occurrences of incidents during the day and night. The result is shown in table 2. It shows that the incidents occurred more than two times during the day than at night. People are more active during the daytime and create more opportunities for incidents especially in the busy public places.

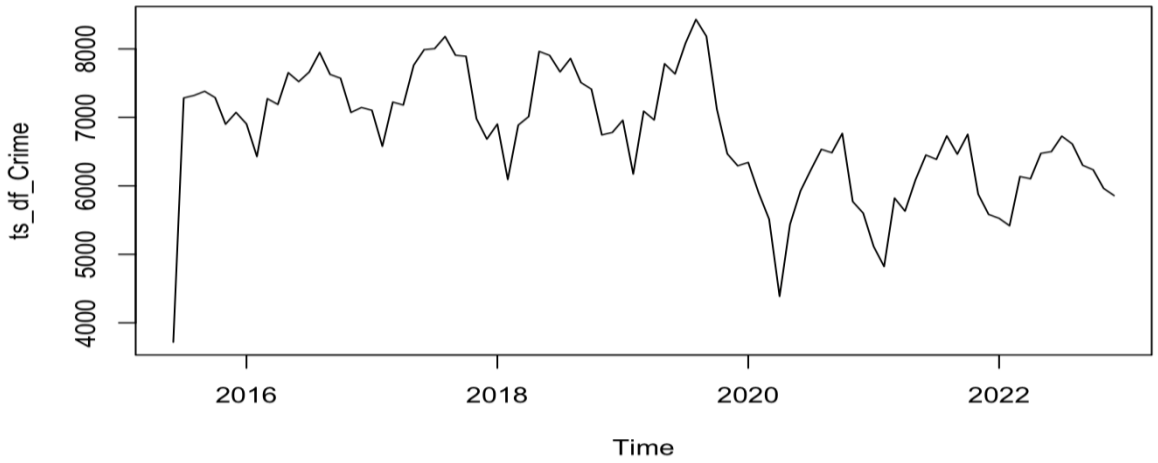
TimeOfDay <chr>	total_occurrences <int>
Day	430010
Night	184816

Table 2

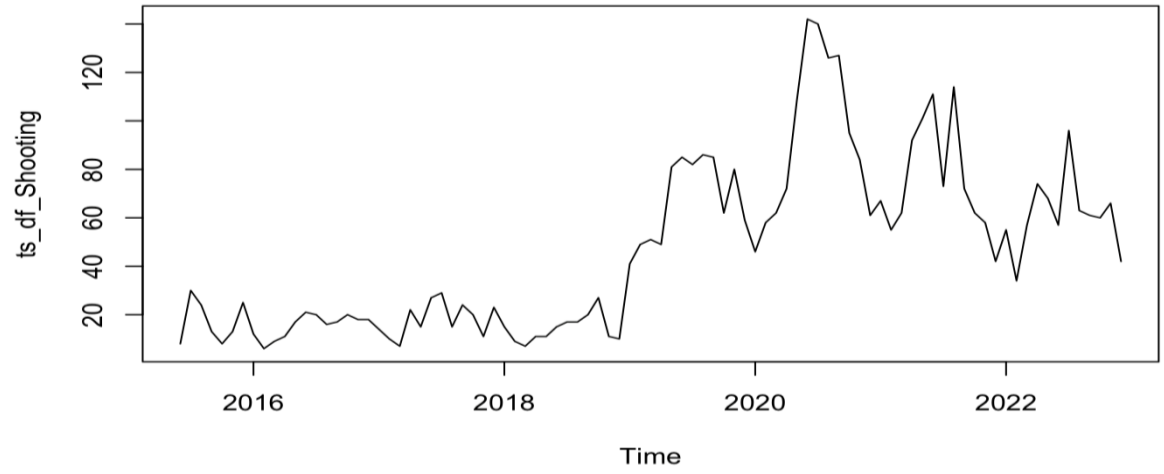
RQ2

We first calculated the monthly sums of incident cases and made them as a list, and then transformed the list into a ts object with time-series data format. For the monthly sums of shooting cases, we performed the same procedure. Graph 5 visualizes the monthly sums of incident cases and Graph 6 visualizes the monthly sums of shooting cases across all time periods. As shown in Graph 5, the count of incident cases had a huge drop in the 2nd quarter of 2020, followed by a slight rebound in the 3rd quarter of 2020 and then maintained at a lower level compared to before 2020. As shown in Graph 6, the count of shooting cases remained low before 2019, but there was a significant increase starting from 2019 and incurred a sudden rise and reached the peak in the

2nd and 3rd quarters of 2020, after which it had a slight drop and remained at the same level as 2019.

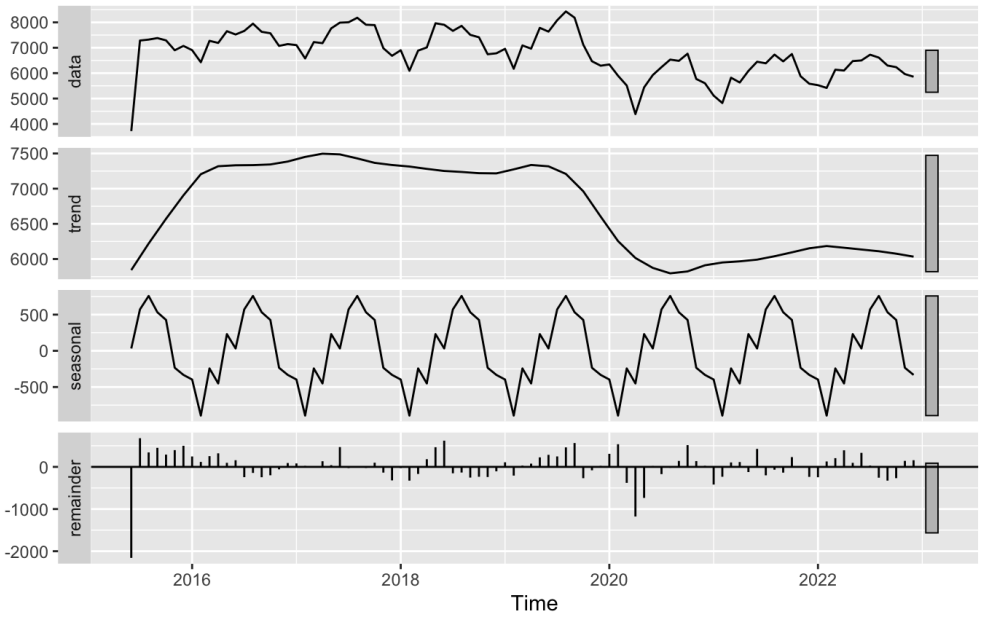


Graph 5

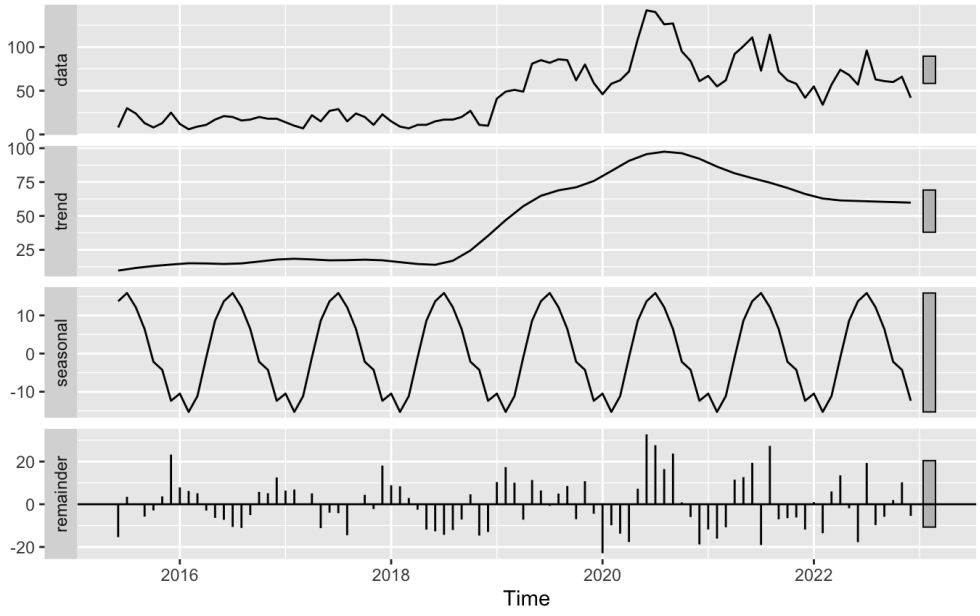


Graph 6

Graph 7 and Graph 8 show the time series components of the previously mentioned ts objects of monthly incident sums and monthly shooting sums. The trend components do not show a continuously increasing or decreasing trend, but rather display fluctuations over time, indicating that there is no clear direction of trends in these two time-series data. The seasonality components are in cyclic patterns that repeat at fixed intervals, indicating that these two time-series data have seasonality. This pattern can be due to various factors, which needs deeper analysis.



Graph 7

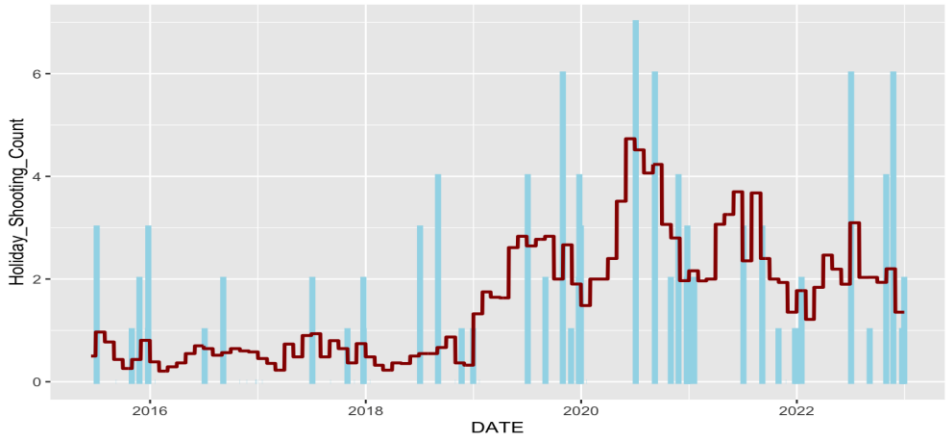


Graph 8

In Graph 9, the orange line represents the monthly average incident counts and the green bar represents the incident counts on special holidays. We can see the orange line generally overlaps the top of the green bar, indicating that the count of incidents happening on special holidays and on normal days does not have an obvious distinction. In Graph 10, the dark red line represents the monthly average shooting counts and the blue bar represents the shooting counts on special

holidays. Different from the previous case, we can see the dark red line generally sits much lower than the top of the green bars, indicating that the count of shooting cases happening on special holidays are much higher than on normal days.

Graph 9



Graph 10

RQ3

All 'DISTRICT_NAME' with p-value < 0.05

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
Central Boston	-0.73209	0.17660	-4.146	3.39e-05 ***
Charlestown	0.45951	0.21287	2.159	0.030879 *

Dorchester	1.77243	0.12322	14.384	< 2e-16 ***
East Boston	0.56039	0.16276	3.443	0.000575 ***
Hyde Park	1.46286	0.13498	10.838	< 2e-16 ***
Jamaica Plain	1.50579	0.13473	11.177	< 2e-16 ***
Mattapan	2.27489	0.12187	18.667	< 2e-16 ***
Roxbury	2.03218	0.12167	16.703	< 2e-16 ***
South Boston	0.60955	0.14494	4.206	2.60e-05 ***
South End/Kenmore	0.46614	0.13767	3.386	0.000709 ***
West Roxbury	0.64913	0.16230	4.000	6.35e-05 ***

All ‘HOUR_RANGE’ with p-value < 0.05

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
4:01 pm - 8:00 pm	-0.26904	0.07128	-3.774	0.000161 ***
8:01 am - Noon	-1.44987	0.09657	-15.013	< 2e-16 ***
8:01 pm - Midnight	0.75444	0.06670	11.310	< 2e-16 ***
Midnight - 4:00am	0.95046	0.06784	14.011	< 2e-16 ***
Noon - 4:00pm	-0.88584	0.08077	-10.967	< 2e-16 ***

Among all the districts in Boston, ‘Mattapan’, ‘Roxbury’, ‘Dorchester’ has the top 3 highest coefficients of 2.27489, 2.03218, 1.77243, indicating that compared to the other districts, a incident in these three districts is associated with a higher log odds of a shooting incident being reported. ‘Central Boston’ has the lowest coefficient of -0.73209, indicating that Boston’s central area is associated with the lowest log odds of a shooting incident.

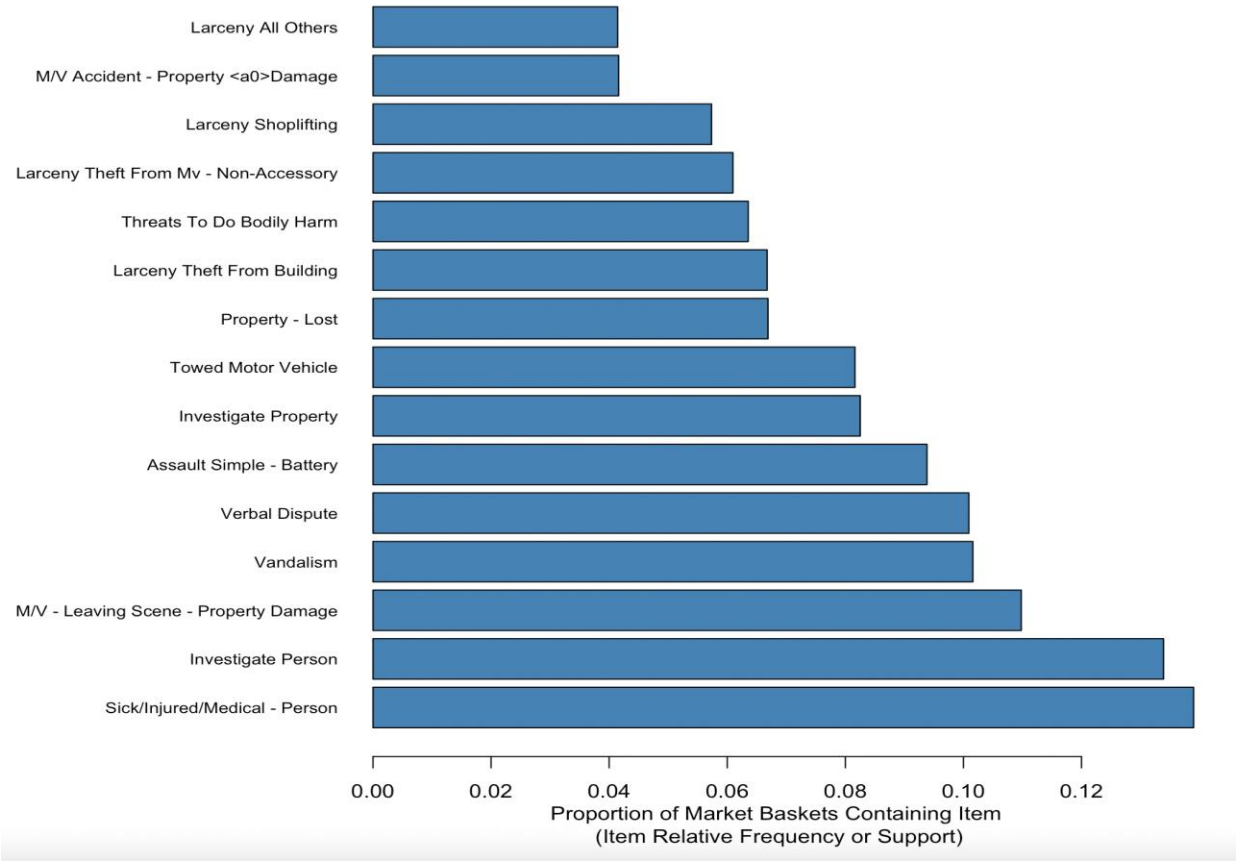
Turning to the hour range variable, the ‘8:01 am - Noon’ and ‘Noon - 4:00 pm’ hour ranges have low coefficients of -1.44987 and -0.88584, indicating that a incident during these time periods is

associated with a lower log odds of a shooting incident. The ‘Midnight - 4:00am’ and ‘8:00 pm - Midnight’ hour ranges have the high coefficients of 0.95046 and 0.75444, indicating that an incident during these time periods is associated with higher log odds of a shooting incident.

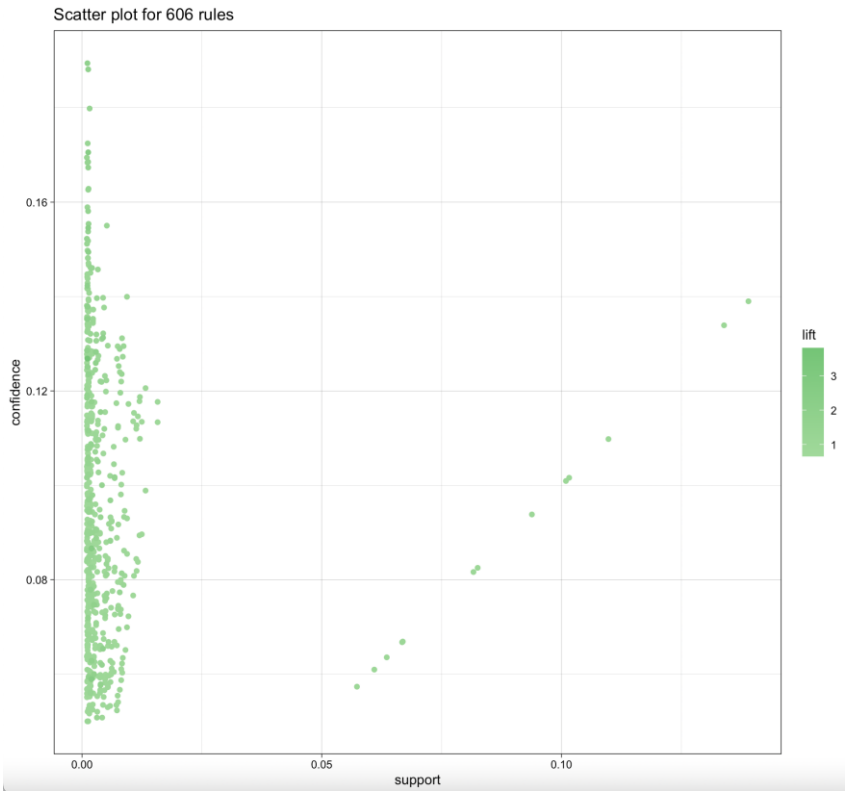
In general, the results of the logistic regression model suggest that both district location and hour range are important predictors of shooting incidents. This analysis result could be useful for law enforcement officials in directing resources and developing strategies to prevent shootings in high-risk areas and during high-risk time periods.

RQ4

Examine top 15 items with support>0.04



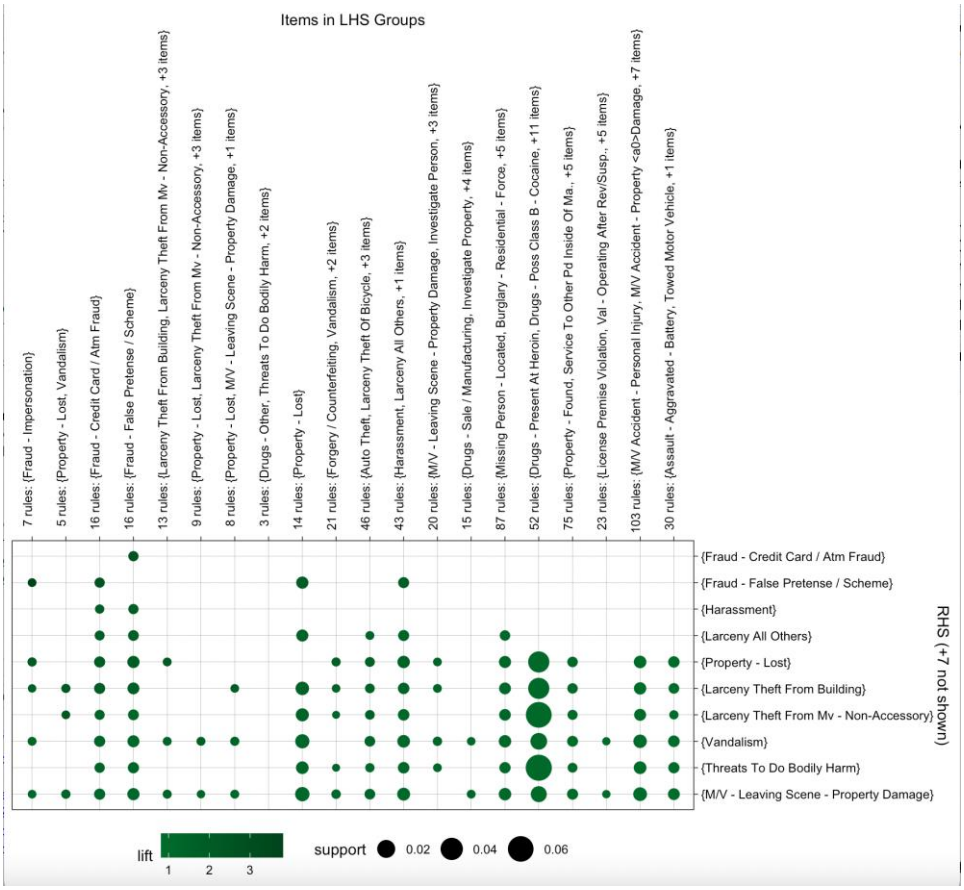
Graph 11



Graph 12

By identifying association rules with high support, confidence and lift, we can gain insights into the factors that contribute to criminal activity in Boston. In this scatterplot, the x-axis represents the support values, which measure the frequency of occurrence of the rule in the dataset. The y-axis represents the confidence values, which measure the conditional probability of the consequent given the antecedent. The shading of the plot is done using the lift measure, which compares the observed support of a rule to what would be expected if the rule's antecedent and consequent were independent. The darker shading indicates stronger lift values, suggesting that these rules are more significant.

The plot shows a distribution of points scattered throughout the plot, with some clustering towards the upper left-hand corner. This clustering suggests that there are some strong association rules with both high lift and high confidence. So, this plot can be used to identify interesting and potentially valuable association rules in the Boston incident type case, which could be further explored and analyzed.

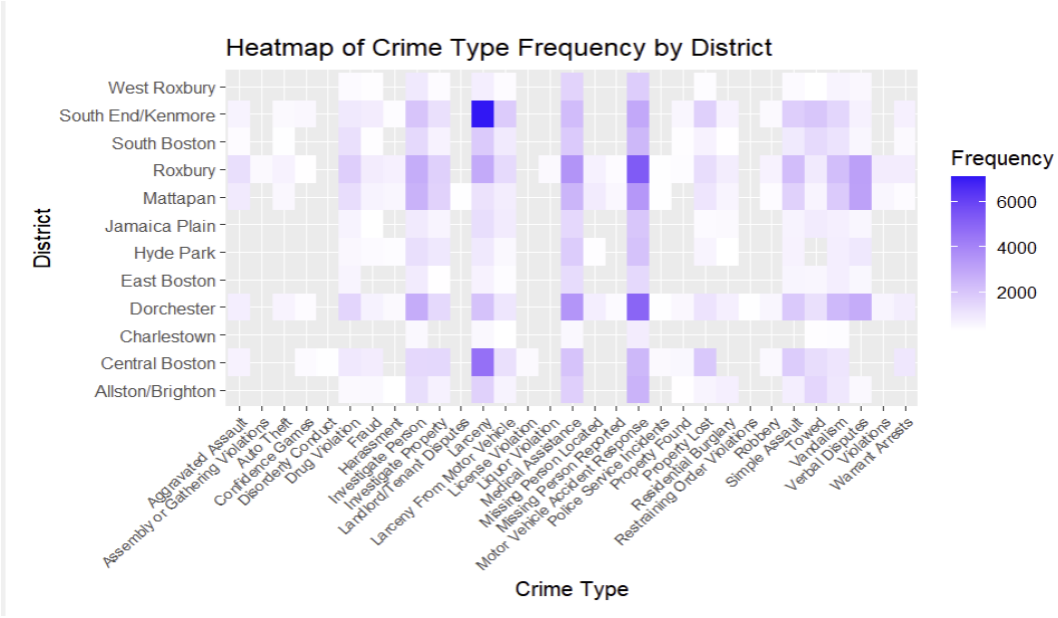


Graph 13

This is a grouped bar chart of association rules for a Boston incident type dataset. Each point in the chart represents a combination of antecedent and consequent itemsets, with the size of the circle indicating the support of the rule (i.e., the frequency of occurrence of the rule in the dataset). The different colors of the bars represent the life values of the rules, with darker shades indicating higher lift.

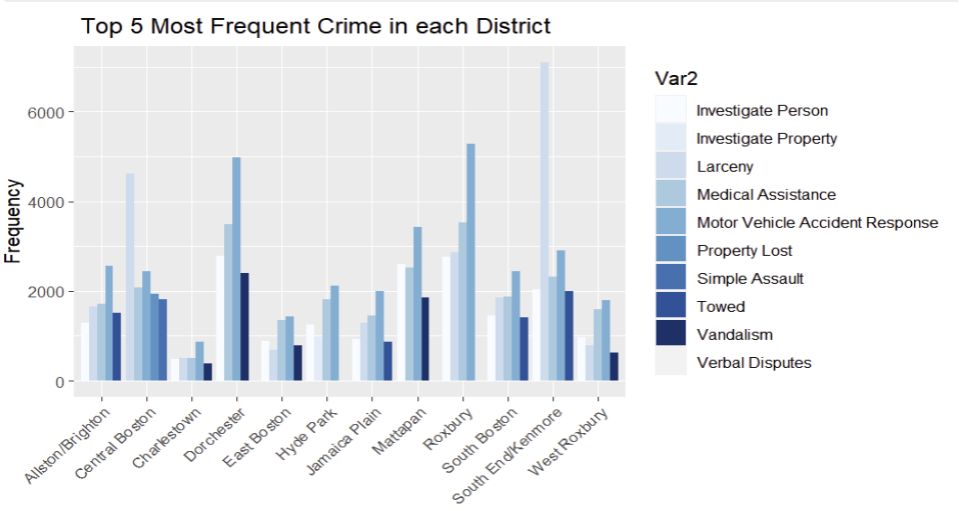
Rules with drugs have a high support value with property loss, larceny and threats to do bodily harm, which suggests that the combination of these items is frequent or popular in the dataset. In this case, the rule suggests that there may be a strong positive association between drug use and criminal behavior such as property theft, larceny, and violent acts. Moreover, fraud - impersonation -> fraud -false pretense / Scheme with lift of 4. The association rule suggests that fraudsters who use impersonation tactics are more likely to engage in fraud that involves false pretenses or schemes. This could mean that fraudsters who are successful at impersonation are more likely to use this success to convince victims to part.

We want to examine the relationship of Location and Incident Types by using heatmap and histogram.



Graph 14

As shown above in the heatmap, the darker the color represents a higher frequency of such incidents. It is not difficult to notice that Larceny and Motor Vehicle Accident Response have the darkest color in the map.



Graph 15

The areas of Central Boston and South End/Kenmore are among those more frequented by tourists, so Larceny is more likely to appear than other areas. Allston/Brighton, Dorchester, Mattapan, and Roxbury are residential areas, and Motor Vehicle Accident Response is the most frequent case.

Conclusion & Recommendations

From all the analysis shown above we can conclude that there is a clear seasonality pattern of incident shown from the trend of the graph. We would suggest especially tightening security and increasing enforcement during the summer months. Reducing incidents in an area requires a multifaceted approach that involves various stakeholders, including law enforcement agencies, government officials, community leaders, as well as residents. Firstly, we can do community policing, which involves community outreach, regular patrols, and incident prevention education programs. Secondly, we can increase social services. By providing access to affordable housing, education, healthcare, incident rates can be reduced by addressing underlying causes such as poverty, lack of opportunity, and social isolation. Since it's easier for people to walk around during the summer months due to the weather being more accessible as well, we can improve the street lighting. Poorly lit areas can provide cover for criminals, and therefore improving street lighting in public spaces and residential areas can help deter incidents and increase safety. Moreover, we would also recommend strengthening neighborhood watch programs, by increasing surveillance, reporting suspicious activity, and identifying potential risks in the community. Finally, enforcing existing laws and policies related to incidents, such as gun control laws, can help reduce incidents, as well as developing and implementing new policies that address emerging incident trends can be beneficial.

The result of ‘Special Date analysis suggests that the count of shooting cases on special holidays including Independence Day, Labor Day, Martin Luther King Day, Halloween, Thanksgiving, Christmas, New Year's Eve, is significantly higher than on normal days. This indicates that there may be a greater risk of gun violence during special holidays, which could pose a threat to public safety. It is recommended that the Boston Public Department take proactive measures to address this issue, such as increasing their level of staffing and preparedness during special holidays and implementing community policing programs to prevent and address gun-related incidents.

According to the result of the conducted logistic regression, 'Mattapan', 'Roxbury', and 'Dorchester' are associated with a higher likelihood of a shooting incident being reported and 'Central Boston' is associated with the lowest likelihood of a shooting incident. The hour ranges of 'Midnight - 4:00am' and '8:00 pm - Midnight' are associated with a higher likelihood of a shooting incident being reported and the hour ranges of '8:01 am - Noon' and 'Noon - 4:00pm'are associated with a lower likelihood of a shooting incident being reported. As a reference, BPD should direct more police force and resources and develop efficient strategies to the districts of 'Mattapan', 'Roxbury', and 'Dorchester', as these areas are identified as high-risk areas for shooting incidents. The department should also pay attention to the night hours, which are identified as high-risk time periods for shooting incidents. Additional resources such as increased patrols and targeted interventions may be necessary during 8:00 pm to 4:00 am to prevent and address shooting incidents.

As a famous tourist city in the United States, there is a distinction between residential areas and tourist areas in different parts of Boston. Therefore, it is easy to see that the areas with the most larceny are the ones most visited by tourists. We recommend that local police teams can erect signs to remind tourists to be careful with their belongings. For residential areas, residents can take more precautions after understanding our analysis of the data.

Based on market basket analysis, drugs are a major issue that can lead to property loss, theft, and threats to bodily harm. The government should enforce drug-related laws, such as possession, trafficking, and distribution laws, to reduce the prevalence of drugs in the city. The government can allocate more resources to law enforcement agencies to increase their capacity to detect, investigate, and prosecute drug-related incidents. In addition, both types of fraud can have a high lift value, meaning that the perpetrator can obtain a large amount of money or other valuable assets through the fraud. Fraud is a serious incident that involves intentional deception for personal gain, often at the expense of others. To address fraud, the Boston government can work with law enforcement agencies and community organizations to educate the public about the risks of fraud and how to protect themselves.

In conclusion, there is no one size fits all solution to reducing incidents, and a combination of strategies tailored to the specific needs of a community is likely to be most effective.