IE6600: Data Analysis and Visualization Project Report

Project 2: Boston Crime Dataset

Madison Rodriguez, Adam Luk, Ameya Patil, Leela Maunika Talluri

Due Date: 6/9/2024

1. Introduction

1.1. Objective

The primary objective of this project is to analyze the Boston crime dataset to identify **trends** and **patterns**. We will evaluate various aspects such as the distribution of incidents across different districts, the timing of incidents by day/month/year, and specific types of offenses like harassment and property loss. This analysis aims to provide insights on Boston Crime from 2015 to 2024. Ultimately, leveraging data-driven insights can contribute to proactive crime prevention efforts and improve overall safety within communities.

1.2. Dataset Description

The dataset was obtained from Boston.data.gov and includes crime data from 2015 to 2024. The dataset fields are described as follows:

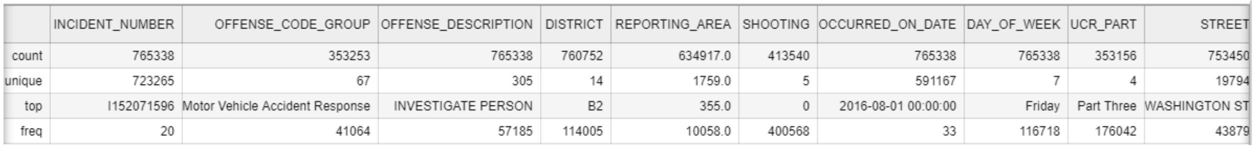
Field Name	Data Type	Description
Incident_num	varchar	Internal BPD report number
Offense_code	varchar	Numerical code of offense description
Offense_Code_Group_Description	varchar	Internal categorization of [offense_description]
Offense_Description	varchar	Primary descriptor of incident
District	varchar	What district the crime was reported in
Reporting_area	varchar	RA number associated with the where the crime was reported from.
Shooting	char	Indicated a shooting took place.
Occurred_on	datetime	Earliest date and time the incident could have taken place
UCR_Part	varchar	Universal Crime Reporting Part number (1,2, 3)
Street	varchar	Street name the incident took place

2. Data Acquisition and Preparation

2.1. Initial Inspection

The initial step involved reading the dataset as ‘df’, as shown by the following code: df = pd.read_csv('merged_data.csv'), followed by using df.head() to preview the first few rows. The data was labeled as ‘merged_data’ because the original data was separated on Boston.data.gov and we used merge procedures to put it in one csv file.

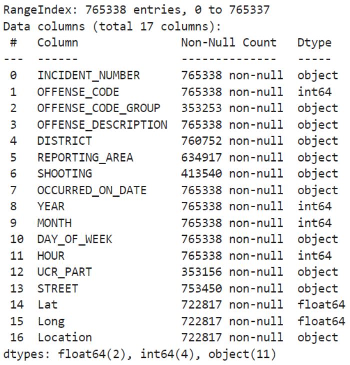
To determine the dataset's size, we used the **df.shape** command, revealing a total of 765,338 rows and 17 columns. For a more detailed summary, we used **df.describe(include='object')** to provide descriptive statistics for categorical columns, shown in Figure 1. This analysis enabled us to explore the distribution of values, including counts, unique values, top values, and their frequencies, aiding in identifying any irregularities or patterns. Notably, the most prevalent incidents were classified under the Offense Code 'Motor Vehicle Accident Response', predominantly occurring in District B2, particularly on Fridays, and frequently happening on Washington St. Additionally, it was noted that the 'SHOOTING' column contained five distinct values, rather than the expected two (either 1 or 0). After investigation, these values consisted of 'nan', 'Y', '0', '1', 0, and 1, necessitating cleaning in the subsequent step.



	INCIDENT_NUMBER	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	DAY_OF_WEEK	UCR_PART	STREET
count	765338	353253	765338	760752	634917.0	413540	765338	765338	353156	753450
unique	723265	67	305	14	1759.0	5	591167	7	4	19794
top	I152071596	Motor Vehicle Accident Response	INVESTIGATE PERSON	B2	355.0	0	2016-08-01 00:00:00	Friday	Part Three	WASHINGTON ST
freq	20	41064	57185	114005	10058.0	400568	33	116718	176042	43879

Figure 1: df.describe()

Next, we used **df.info()** to find the columns with null values which included DISTRICT, OFFENSE_CODE_GROUP, REPORTING_AREA, SHOOTING, UCR_PART, STREET, Lat, Long, and Location. Furthermore, the dataset consists of three different data types: object, int64, and float64. However, 'OCCURRED_ON_DATE' still needs to be converted into datetime format.



```
RangeIndex: 765338 entries, 0 to 765337
Data columns (total 17 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   INCIDENT_NUMBER      765338 non-null  object
 1   OFFENSE_CODE         765338 non-null  int64
 2   OFFENSE_CODE_GROUP   353253 non-null  object
 3   OFFENSE_DESCRIPTION  765338 non-null  object
 4   DISTRICT             760752 non-null  object
 5   REPORTING_AREA       634917 non-null  object
 6   SHOOTING             413540 non-null  object
 7   OCCURRED_ON_DATE     765338 non-null  object
 8   YEAR                 765338 non-null  int64
 9   MONTH                765338 non-null  int64
 10  DAY_OF_WEEK          765338 non-null  object
 11  HOUR                 765338 non-null  int64
 12  UCR_PART             353156 non-null  object
 13  STREET               753450 non-null  object
 14  Lat                  722817 non-null  float64
 15  Long                 722817 non-null  float64
 16  Location             722817 non-null  object
dtypes: float64(2), int64(4), object(11)
```

Figure 2: df.info()

2.2. Data Cleaning

The first step was to remove 'OFFENSE_DESCRIPTION' and 'Location' because the description provided specific details of individual incidents, which are not useful for trend analysis, and the location was already represented by the lat and lon columns. The next step involved converting 'OCCURRED_ON_DATE' to datetime format. Step three included converting all column names to lowercase for improved readability.

The next step involved removing all NULL values. We began with 'offense_code_group', which we handled by sorting the data by 'offense_code' and then applying the backfill method. This ensured that missing values were filled with contextually relevant data, improving accuracy. Next, we standardized the 'SHOOTING' column by replacing 'Y' and '1' with 1, and 'NAN' and '0' with 0, followed by filling any remaining NA values with 0. For the numerous missing values in 'reporting_area', ‘lat’,’long’, and ‘street’, we sorted the data by 'district' and 'street' and used the backfill method again, ensuring the filled values were consistent with the geographical context. Additionally, for rows where 'occurred_on_date' was empty but 'year' and 'month' were provided, we used a function to generate a date based on the month and year, assigning a random day from 1 to 28 to ensure a valid date. Also, for the rest of the Null values we used the mode function to fill in the rest of the values.

3. Exploratory Data Analysis (EDA)

3.1. Overview

Initially, we identified districts with high crime rates. Subsequently, we analyzed the time-related distribution of incidents, with a focus on daily and monthly trends. We then investigated the evolution of incident counts over time, with particular emphasis on the period before and after COVID-19. Finally, we conducted an examination of the primary types of crimes occurring in each district.

3.2. Visualizations

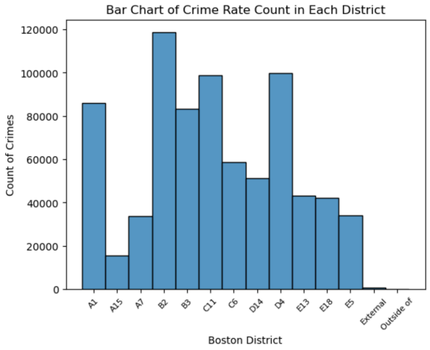


Figure 3: Bar Chart of Crime Rate Count in Each District

The bar chart displays the crime rate across various Boston districts. District B2 has the highest crime rate, with nearly 120,000 incidents, while District A15 has the lowest, with counts below 20,000. Other districts, such as A1, B3, and D4, also show relatively high crime rates, exceeding 80,000 incidents. In contrast, districts like C6 and E5 report lower crime rates, around 40,000 incidents. This chart presents the overall crime rate. To explore further, the following box plot chart depicts the distribution of the crime rate per year.

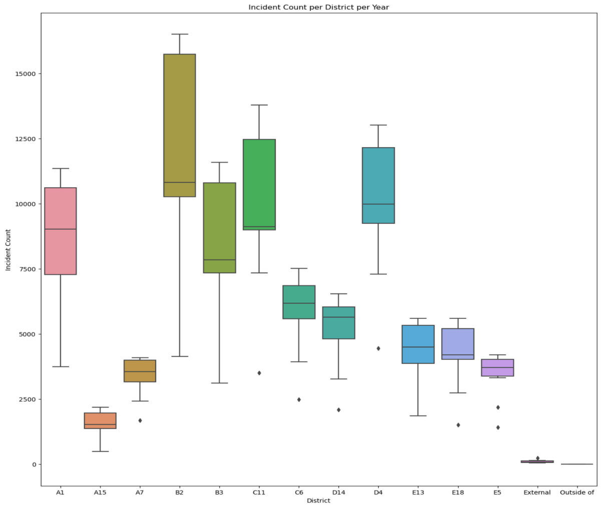


Figure 4: Incident Count Per District From 2015 - Present

The chart illustrates the incident count per district per year. District B2 remains the district with the highest incident count, with a median around 11,000 per year, while District A15 has the lowest, around 2,000. Districts like B2 show significant variability, indicating wide fluctuations in incident counts each year. In contrast, districts such as A15, A7, and E5 exhibit narrower ranges, suggesting more consistent incident counts annually. The "External" category shows very few incidents, highlighting its insignificance compared to other districts. Overall, districts B2, followed by D4, C11, and A1, have high crime rates, suggesting they may be less safe compared to other districts. It would be interesting to investigate the top crimes in these areas to further analyze why their incident rates are so high each year.

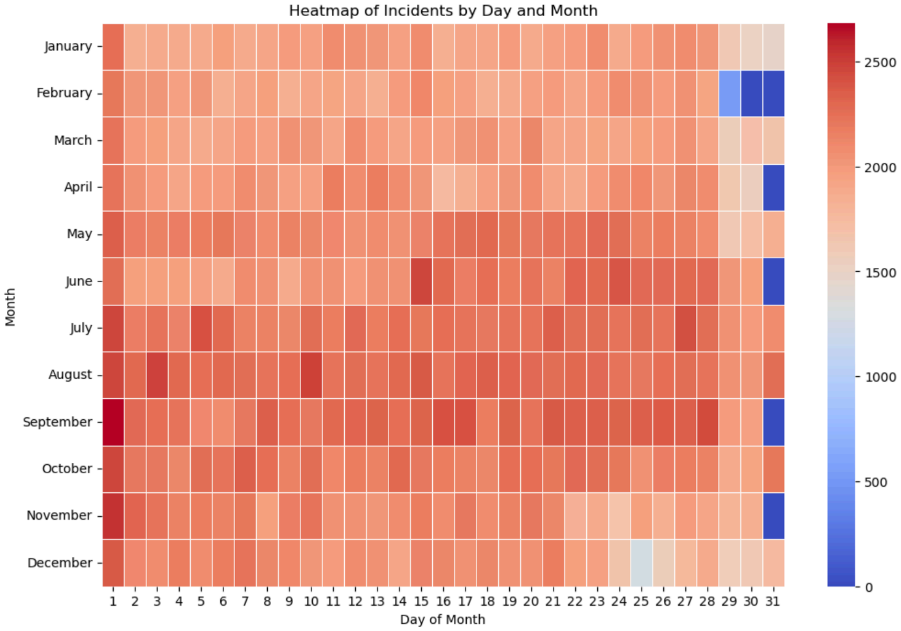


Figure 5: Heatmap of Incidents by Discrete Day and Month

The heatmap reveals a higher frequency of incidents during the warmer months, with an increase starting in May, peaking in September and declining after November. Additionally, there are notable spikes in crime at the beginning of each month, particularly on dates such as September 1st, November 1st, August 3rd, August 10th, June 15th, and September 28th. This pattern may be due to factors such as increased outdoor activities, school vacations, and specific social or economic events that occur during these times. However, September 1st is known as moving day so may contribute to increased activity (most notably motor vehicle accident responses) and November 1st falls right after Halloween. It would be interesting to see what are the most popular crimes that occur during these days.

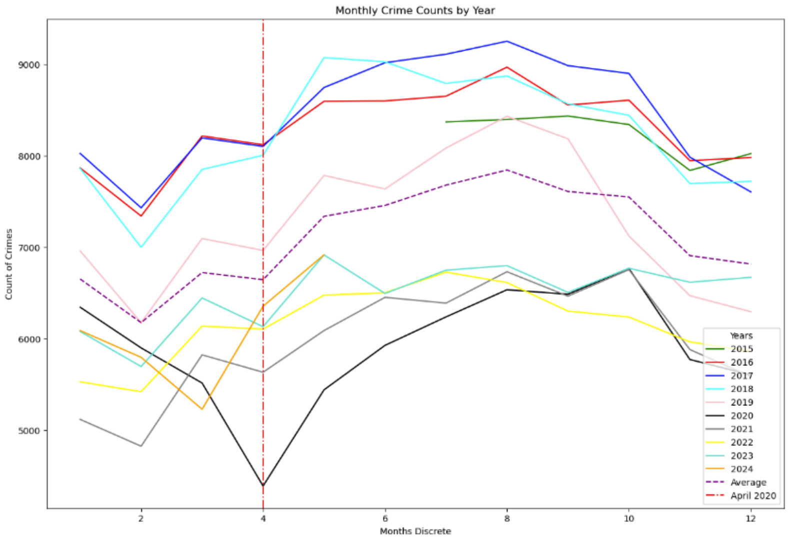


Figure 6: Line Graph of Monthly Crime Rates by Year

The dataset covers the period from July 2015 to May 2024, which accounts for the incomplete lines observed for both 2015 and 2024. Notably, 2017 recorded the highest amount of crime throughout the year, whereas 2020 exhibited the lowest. A noticeable trend is seen with a peak in the number of incidents from April to August and a decline from February to April, as depicted in the heatmap above. 2020 exhibited the lowest (reported and actual crimes committed) likely due to the quarantine mandate throughout the United States from the COVID-19 virus which makes sense with less people actually going out, less crimes are committed and reported. The purple line represents the average crime rate per year, while the red vertical dashed line marks April 2020, the first month of quarantine, which recorded the lowest crime counts.

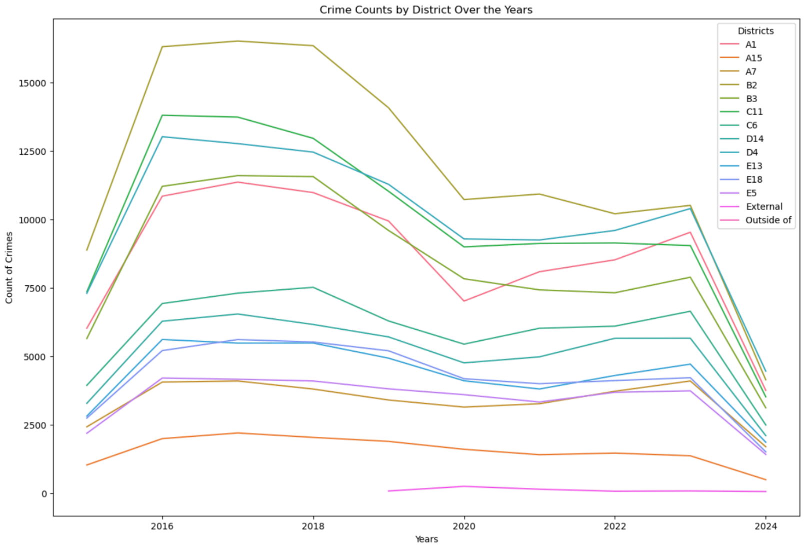


Figure 7: Line Graph of Crime Over the Years for Each District

The overall crime counts for each district show a recurring pattern of highs and lows. Notably, districts B2 and C11 consistently report the highest numbers over the years, while A15 and the "External" category consistently have the lowest crime rates.

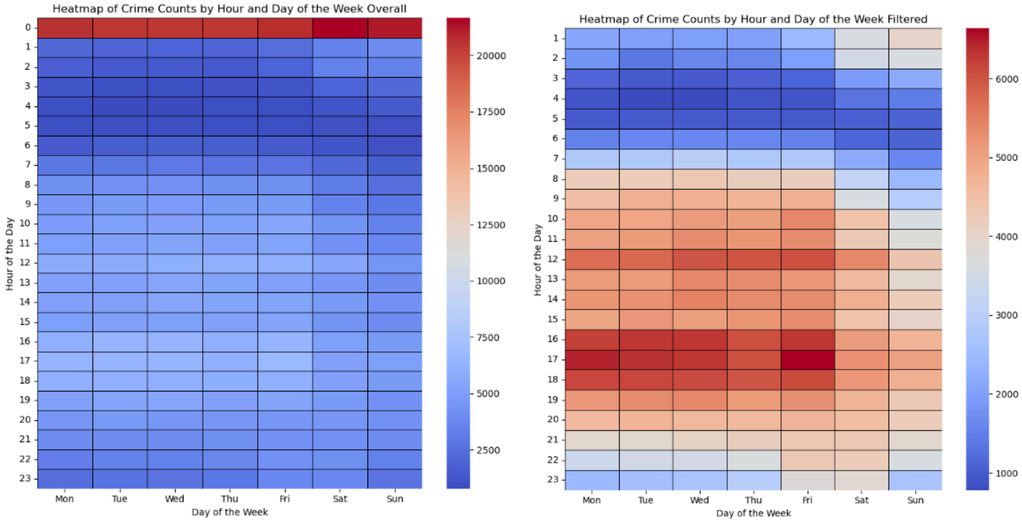


Figure 8a: Heatmap of Crime Counts by Hour and Day of the Week All Data (Left) and Filtered (Right)

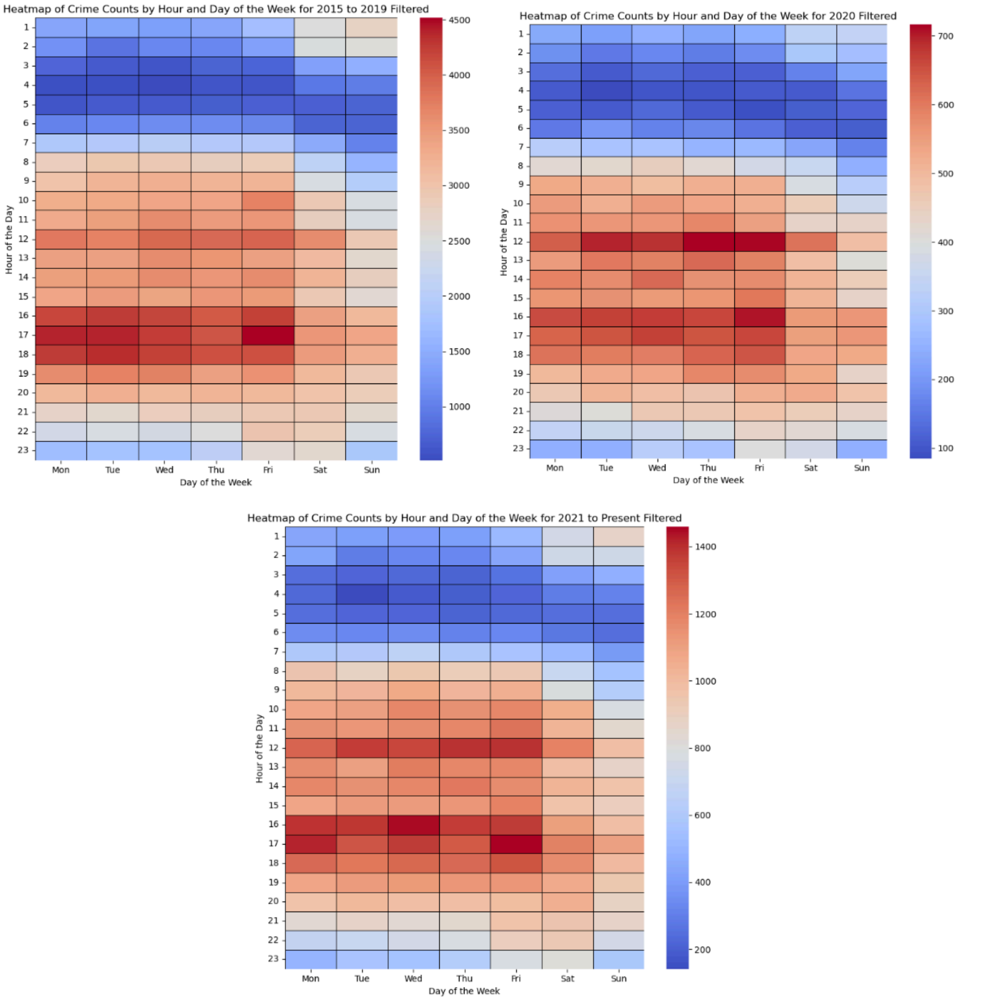


Figure 8b: Heatmap of Crime Counts by Hour and Day of the Week Filtered Data 2015-2019 (Top Left), 2020 (Top Right), and 2021 to Present (Bottom)

Figures 8a and 8b compile crime counts by the specific days and times they were reported. Figure 8a reveals that most crimes occurred from 12:00 AM to 1:00 AM, heavily skewing the data. To address this, the right figure in 8a filters out this time period, showing that most crimes were reported between 9:00 AM and 5:00 PM, with a peak at 5:00 PM on Fridays. This pattern is logical as many people are away at work during these hours, leading to more outdoor incidents or home-related crimes in their absence. The peak on Friday evenings is also reasonable since people often go out to celebrate the end of the week, increasing the likelihood of reported incidents.

Figure 8b compares crime data from pre-COVID (2015-2019), during COVID (2020), and post-COVID (2021-present) periods. The pre- and post-COVID data generally mirror the patterns seen in Figure 8a, though post-COVID data shows a significant decrease in crime, aligning with the trends in Figure 6. Interestingly, in 2020, crime rates remained similar, but most incidents occurred on Tuesdays, Wednesdays, Thursdays, and Fridays at 12:00 PM.

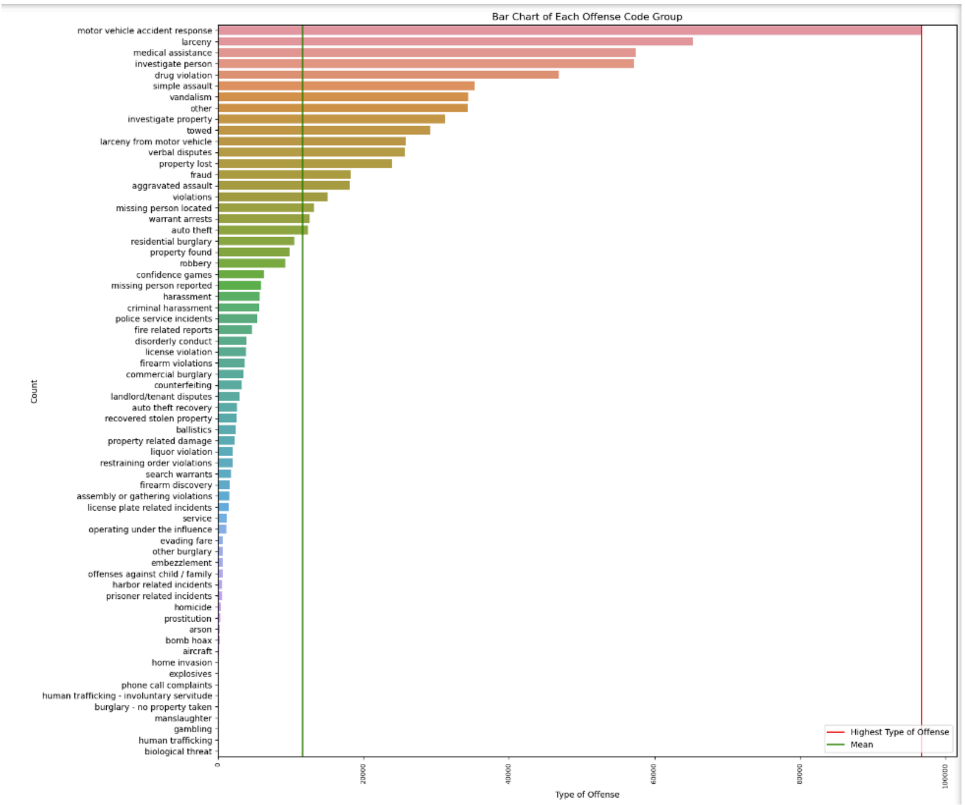


Figure 9: Bar Graph of Offense Group Type Count

The bar chart illustrates the distribution of offenses by offense group. The top 5 prevalent incidents occurred in categories such as Motor Vehicle Accident Response, Larceny, Medical Assistance, Investigate Person, and Drug Violation. Conversely, the least frequent offense groups comprise Biological Threat, Human Trafficking, Gambling, Manslaughter, and Burglary - No property taken. The green line shows the mean count of all types of offenses.

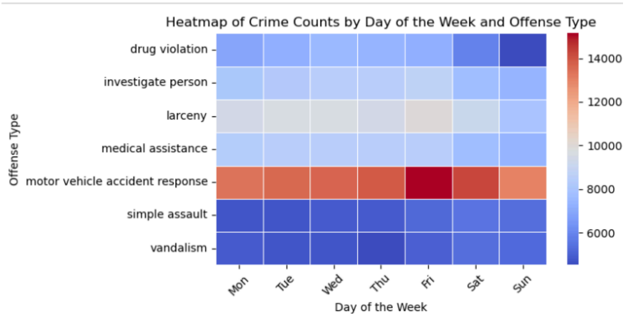


Figure 10: Heatmap of Crime Counts by Hour and Day of the Week Filtered Data (2021 to Present)

The heatmap displays the top seven offense types by day of the week. Motor Vehicle Accident Responses are the most frequently reported crime throughout the week, peaking on Fridays. Larceny is the second most prevalent crime, also highest on Fridays.

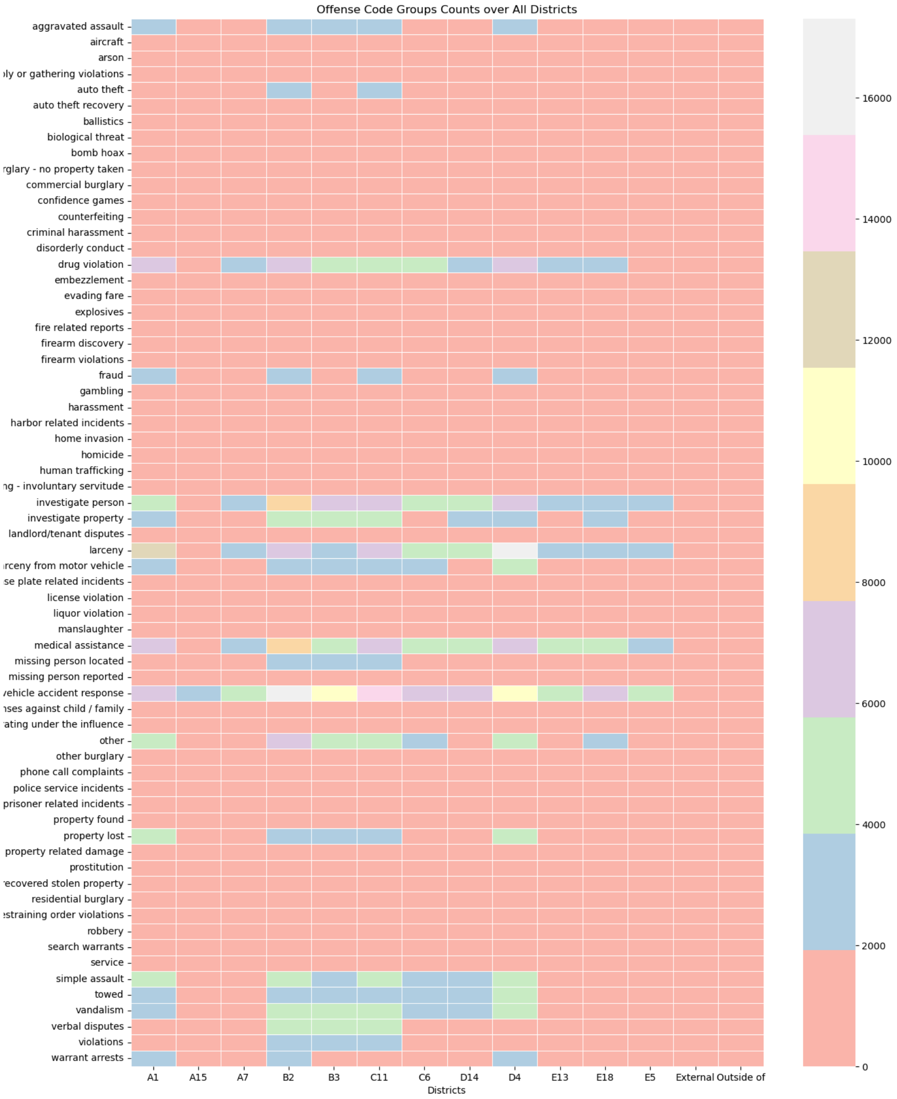


Figure 11: Heat Map of Offense Group Type Count for Different Districts

The heat map in Figure 11 shows that motor vehicle accidents, medical assistance, and investigations of persons are most frequent in district B2, while larceny is highest in district D4, and drug violations are most common in districts A1, B2, and D4. It also indicates that districts A15, external, and areas outside Boston are generally safer with the least amount of crime. Most crimes occur in districts A1, B2, B3, C11, and D4.

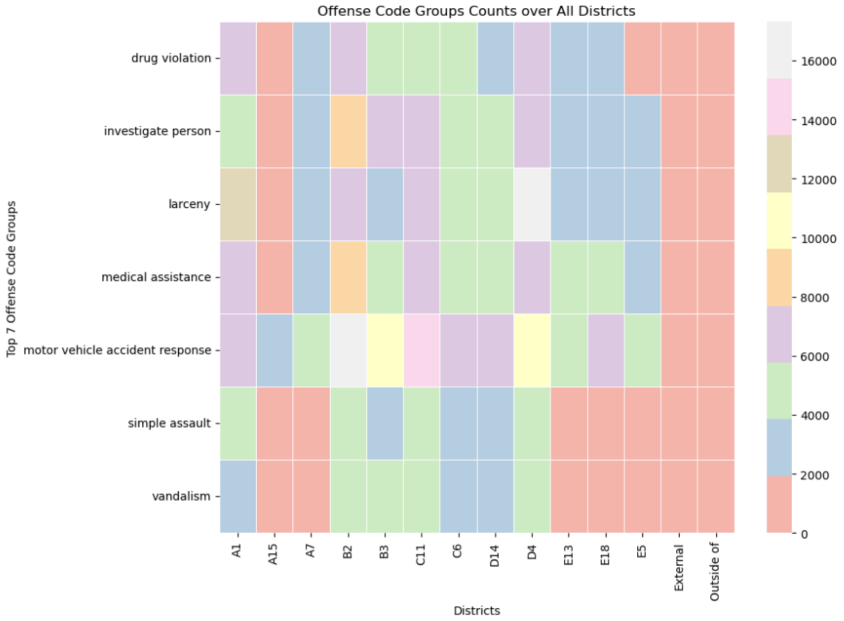


Figure 12: Heatmap of Crime Counts of Top 7 Offense Groups for All Districts

After analyzing the top seven offense types, we examined their distribution across all districts using a heatmap. Larceny was most prevalent in district D4, while district B2 had the highest rates of Motor Vehicle Accident Responses, followed closely by C11. District A1 had the second-highest counts of larceny.

4. Conclusion

4.1. Summary of Findings

Our analysis identified several key insights regarding crime patterns in Boston. District B2 consistently reported the highest crime rates, with nearly 120,000 incidents, while District A15 had the lowest, with fewer than 20,000 incidents. Other districts with high crime rates included A1, B3, and D4, each exceeding 80,000 incidents. Temporal trends showed significant variability throughout the year, with higher frequencies in warmer months (May to September) and noticeable spikes at the beginning of each month.

Crime rates decreased significantly during the COVID-19 pandemic in 2020, with a noticeable reduction in incidents compared to previous years. The patterns of crime shifted, with more incidents occurring during midday on weekdays. The most common offenses were Motor Vehicle Accident Responses, Larceny, Medical Assistance, Investigate Person, and Drug Violation. Motor Vehicle Accident Responses were the most frequently reported crime across the week, especially on Fridays. Larceny was most prevalent in district D4, while district B2 had the highest rates of Motor Vehicle Accident Responses, followed closely by C11. Drug violations were common in districts A1, B2, and D4.

4.2. Implications

The findings have several implications for policy and law enforcement strategies. Given the high crime rates in districts B2, A1, B3, C11, and D4, resources should be prioritized in these areas. Increased policing could help reduce incidents. Law enforcement should anticipate higher crime rates during warmer months and at the beginning of each month, allowing for targeted interventions during these periods. The reduction in crime during the COVID-19 pandemic suggests that certain public health measures can indirectly impact crime rates. Strategies to maintain lower crime levels post-pandemic should be explored. Specific attention should be given to the most prevalent crimes, such as motor vehicle accidents and larceny, to develop targeted prevention and intervention strategies. Additionally, certain date periods are indicative of when extra policing personnel should be deployed (i.e, September 1st has the most amount of incidents due to the hectic Boston moving schedule on that day).

5.3. Future Work

Future research should explore other datasets to identify additional factors that might affect crime rates beyond time of day and year. Examining the influence of weather conditions, economic status, and other socioeconomic variables could provide a more comprehensive understanding of crime dynamics.