LUISS

Report: Python and R for Data Science

Boston Crimes Data Analysis

Module Co-ordinators: Valerio Rughetti, Macro Picone, Fabio Angeletti

Deadline: 28th November 2023

Student Names and Numbers: Andrea Marcoccia (775451), Aurora Menegatto (783301), Giulia Scalas (772601), Irene Fellin (785421)

MSc of Data Science and Management 1st Year

Word count: 3400

Abstract:

The aim of this report is to provide a full description of the analysis that was conducted for the chosen dataset. The following paragraphs will provide information on the exploration of the dataset, the data cleaning process, and the specific analysis that were conducted. Relevant graphs and explicatory visual images have been included within the report.

Data:

Our analysis utilizes a dataset titled "Crimes in Boston", sourced from Kaggle. This dataset is provided by the Boston Police Department (BPD) and is intended to document initial details surrounding various incidents to which BPD officers have responded.

It is structured to emphasize capturing the nature of the crime incidents, including the type of offense, the timing of the occurrence, and the location details. This information is crucial for understanding crime patterns and trends within the city of Boston.

The dataset encompasses records starting from June 14, 2015, up to September 3, 2018 and includes the following features:

- **INCIDENT_NUMBER**: A unique identifier for each crime incident.
- **OFFENSE_CODE, OFFENSE_CODE_GROUP**: Codes and corresponding group names categorizing the type of incident.
- **OFFENSE_DESCRIPTION**: Detailed description of the offense.
- **DISTRICT**: Police district where the incident occurred.
- **REPORTING_AREA**: Specific area within the district of the incident report.
- **SHOOTING**: Indicator of whether the incident involved a shooting.
- **OCCURRED_ON_DATE**: Date and time when the incident was reported to have occurred.
- **YEAR, MONTH, DAY_OF_WEEK, HOUR**: Time-related details of the incident.
- **UCR_PART**: The Uniform Crime Reporting (UCR) Part (I, II, III) categorizing the incident.
- **STREET, Lat, Long, Location**: Location details including street name and geographical coordinates.

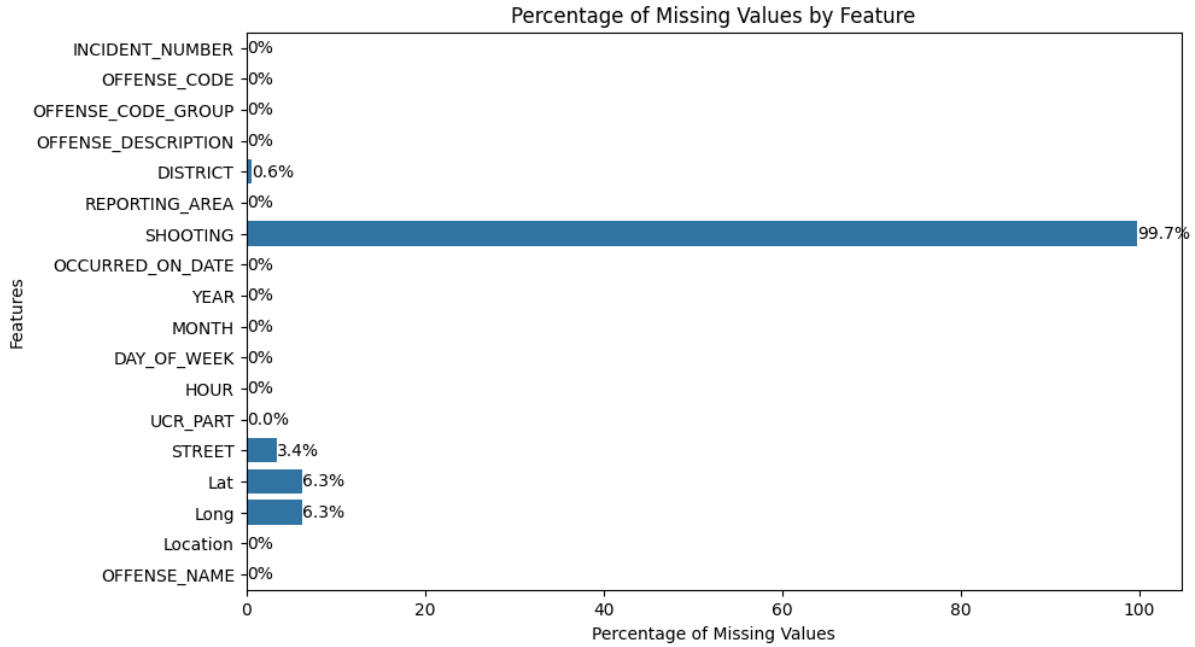
Furthermore, the dataset is complemented by an additional dataset that provides a mapping between **OFFENSE_CODE** and their respective **names.**

Data preparation:

In the initial phase of preparing our dataset for analysis, we incorporated additional information into our primary dataset. Utilizing the mapping available, we were able to enrich our dataset with the **OFFENSE_NAME** corresponding to each **OFFENSE_CODE**.

Subsequently, we conducted a thorough analysis of missing values within the dataset.

Missing values:



The visualization provided in the graph indicates that most features have a complete or near-complete set of values, except for the **SHOOTING** variable. It is notable that the **SHOOTING** variable exhibits a significant number of missing values, accounting for 99.7% of its entries. This high percentage of missing data suggests that the field may be structured to record an incident only when a shooting has occurred, leaving other entries blank. To rectify this and enhance data integrity, we have replaced missing values with 'N' to denote instances where a shooting did not occur.

Given the scope of our analysis, which is focused on exploratory data visualization and understanding crime trends rather than predictive modelling, we have decided to retain other missing values within the dataset. This decision was made under the consideration that the absence of these values does not significantly detract from the overall patterns and insights we seek to derive. Consequently, during the creation of plots and visualizations, any missing values will be excluded to ensure accuracy and clarity in our representations.

Outliers:

In our dataset, the geographical location of each crime incident is a critical feature for analysis. Accurate location data allows for a precise spatial analysis which can yield insights into crime patterns across different neighbourhoods. To capture the crime location, we focus on two primary features: **Lat** (Latitude) and **Long** (Longitude).

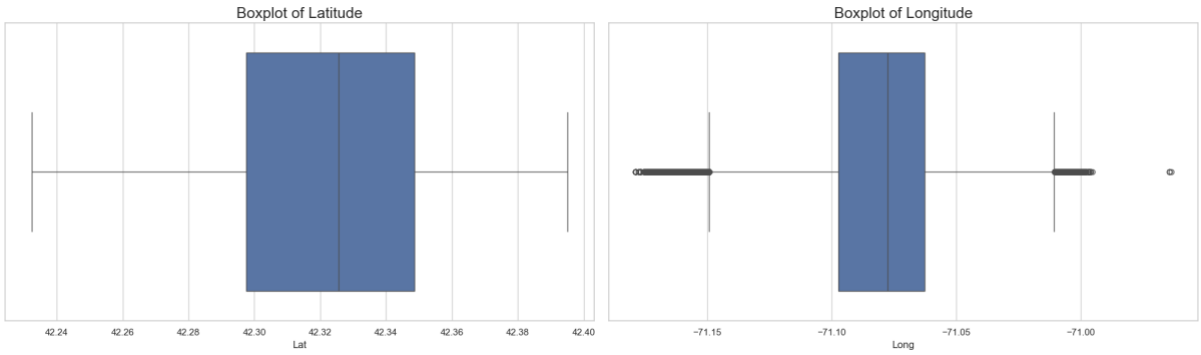
To ascertain the accuracy of our location data, we initiated our investigation by examining the descriptive statistics for both **Lat** and **Long**.

	Lat	Long
count	299074	299074
mean	42.214381	-70.908272
min	-1	-71.178674
25%	42.297442	-71.097135
50%	42.325538	-71.077524
75%	42.348624	-71.062467
max	42.395042	-1

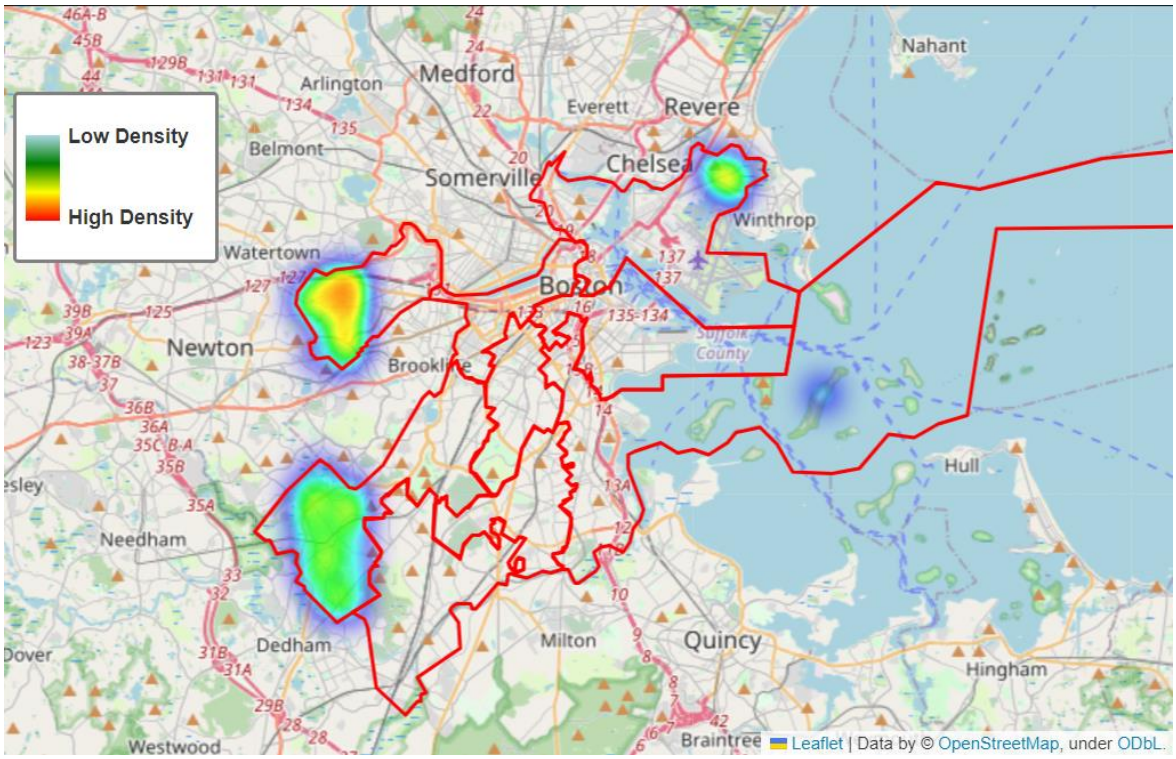
The descriptive analysis revealed that the minimum value of **Lat** and the maximum value of **Long** were both -1. Such values are evidently incorrect as they fall outside the known geographical coordinates of

Boston, which are close to 42 in latitude and close to -71 in longitude. To rectify this, we made the decision to remove these anomalous entries from our dataset.

Following the data correction, we employed boxplots for both features to further scrutinize the dataset for any additional anomalies.



The boxplots provided us with a visual assessment of potential outliers. Our analysis confirmed that while the latitude data exhibited no outliers, the longitude data did present some deviations from the norm. To determine the validity of these outlier values, we proceeded to visualize them using a heatmap.



Overlayed on the heatmap, we plotted the recognized Boston City Council Districts, obtained from the official data repository of the city of Boston (https://data.boston.gov/dataset/new-city-council-districts-may-2023-approved-plan). This allowed us to see the spatial distribution of our data in relation to the official boundaries of the city.

From the combined visualization of our outliers and the official city boundaries, it becomes evident that the outlier values are indeed within the confines of Boston's borders. Therefore, we concluded that these outliers are not erroneous but represent valid data points. As a result, we have opted to retain these outliers in our dataset for subsequent analysis.

The meticulous process of validating our data ensures the integrity of our spatial analysis and supports the reliability of our subsequent findings. This approach lays a solid foundation for exploring the spatial dynamics of crime incidents within Boston.

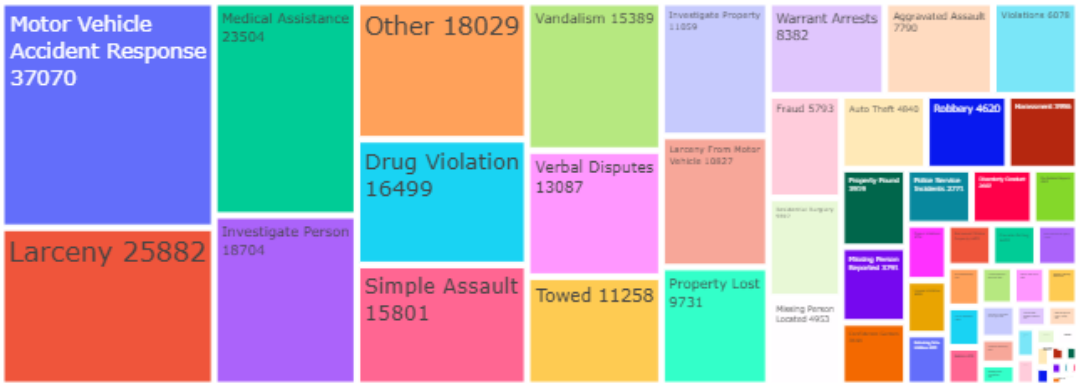
Exploratory Data Analysis (EDA)

In our exploratory data analysis (EDA), we concentrated on a subset of features from the "Crimes in Boston" dataset. These features provide a comprehensive view of crime incidents in terms of their nature, timing, and location. The key features analyzed include:

- **OFFENSE_CODE_GROUP**: The crime group attributed to each specific crime.
- Temporal variables: **YEAR, MONTH, DAY_OF_WEEK, HOUR,** and **OCCURRED_ON_DATE**.
- Spatial variables: **Lat** (latitude) and **Long** (longitude).

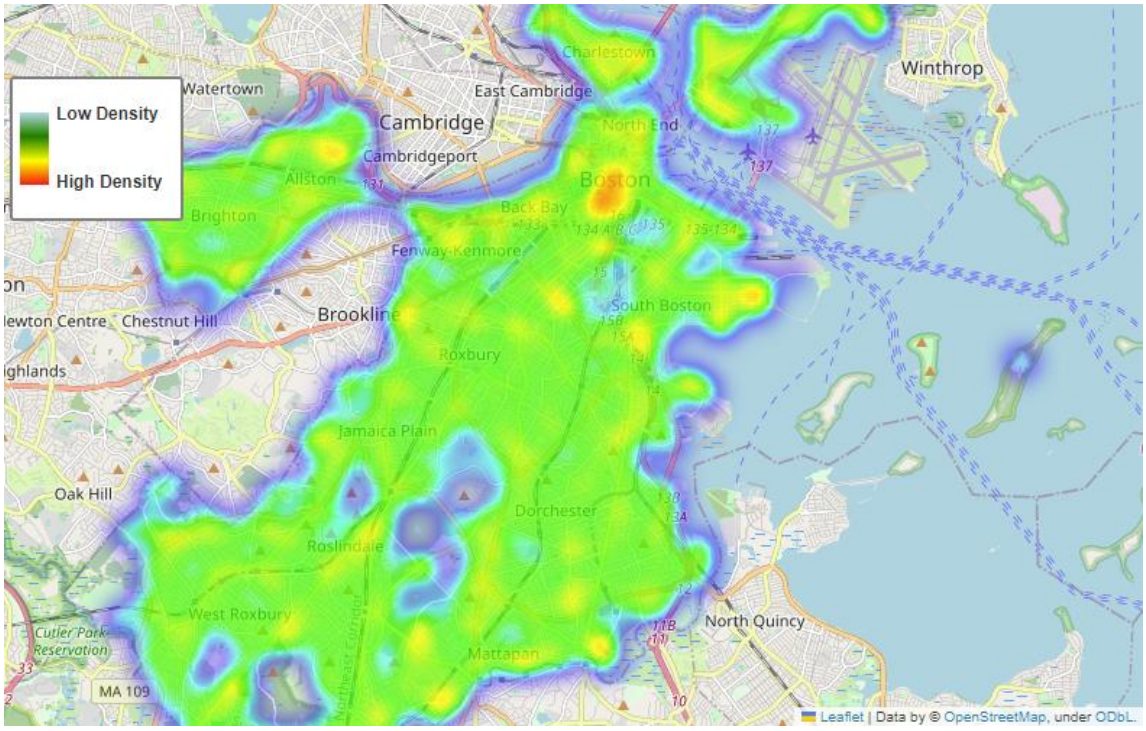
Other features in the dataset were not explored in this phase, allowing us to focus our analysis on these critical dimensions.

Crime Distribution in the Dataset:



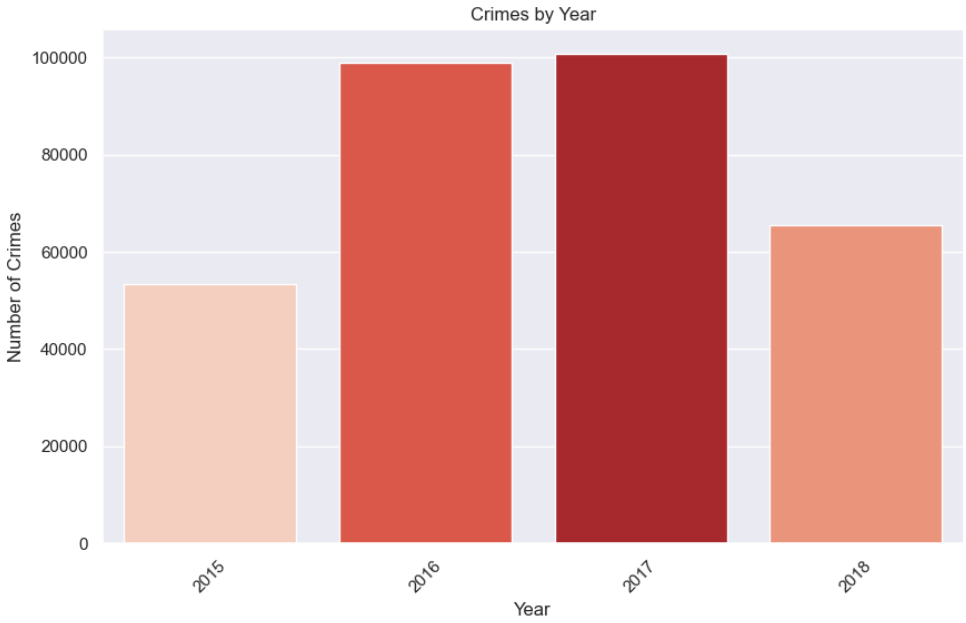
The distribution of crimes in the dataset reveals a wide range of crime types and frequencies. We observe a spectrum from common incidents like "Motor Vehicle Accidents," which occurred 37,070 times over the four years, to rare occurrences like "Human Trafficking" with only 7 incidents reported. This diversity in crime frequency underscores the varied nature of criminal activities in Boston.

Spatial distribution of crimes:

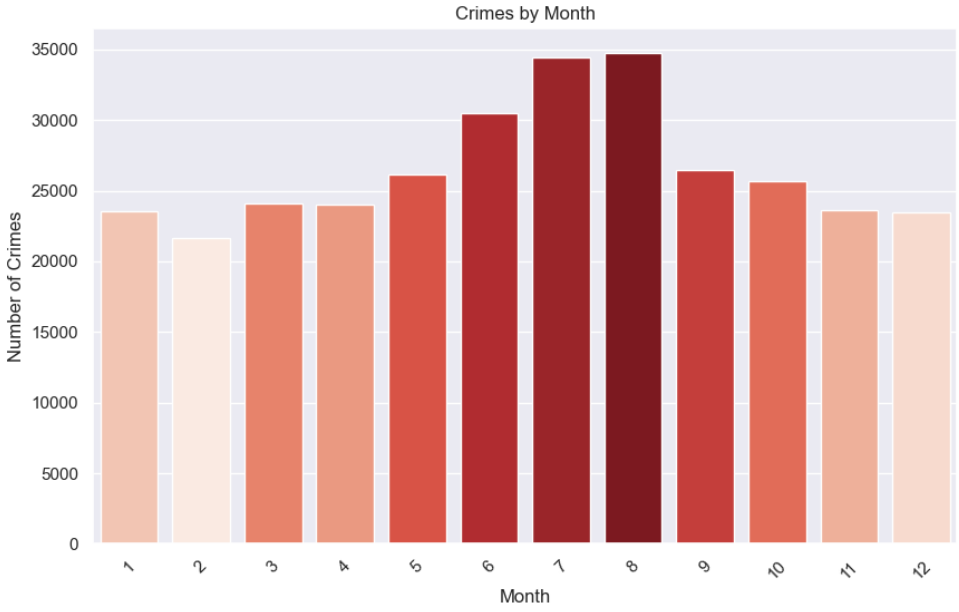


The spatial distribution of crimes in the Boston area, as depicted in the heatmap, indicates a widespread presence of criminal activity across the city. However, certain areas exhibit notably higher crime density. A significant concentration of crimes is observed in the Chinatown area, suggesting it as a hotspot for criminal incidents. Additionally, South Boston, particularly near the coast, emerges as another area with an elevated number of crimes.

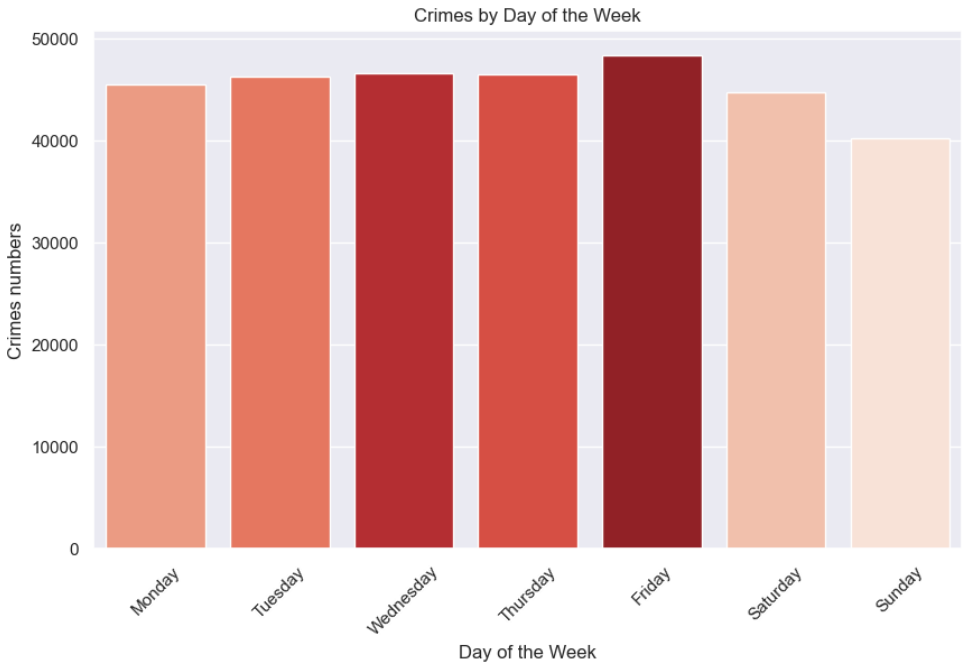
Temporal distribution of crimes:



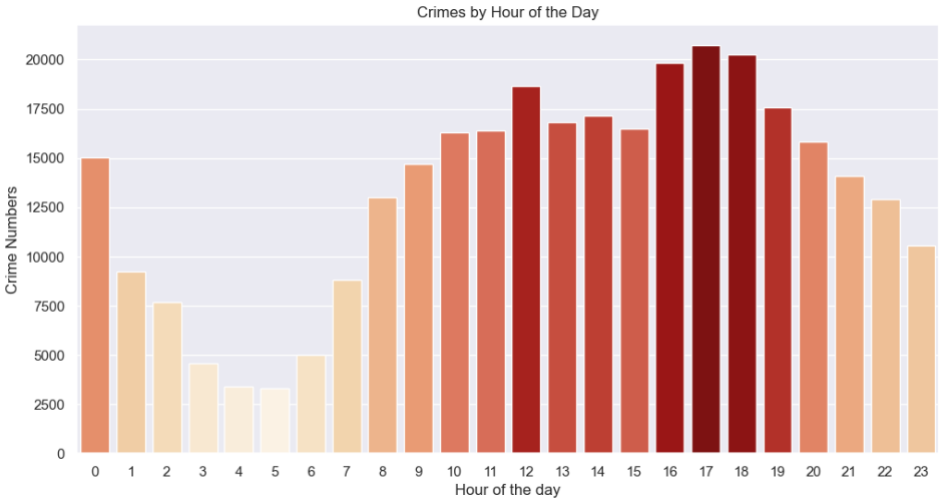
The yearly count-plot of crimes indicates an uneven distribution across the years. The years 2017 and 2016 have almost double the number of crimes compared to 2015 and 2018. This disparity is attributable to the dataset's timeframe, which starts from June 14, 2015, and ends on September 3, 2018.



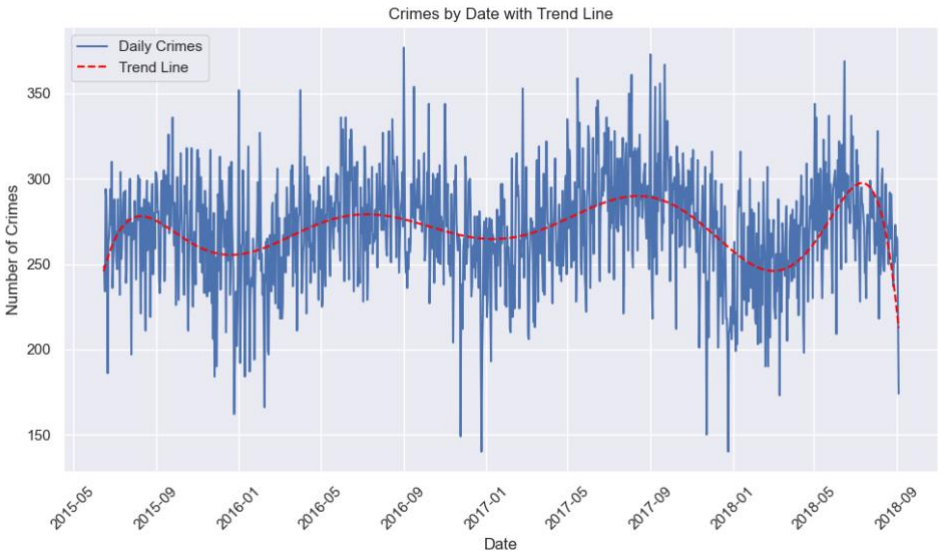
Examining crimes by month, we find that while the count hovers around 25,000 for most months, there is a noticeable peak in the summer months of June, July, and August, with counts exceeding 30,000. This pattern, however, is influenced by the dataset's timeframe, which includes an additional year of data for these months, starting from June 14, 2015, and ending on September 3, 2018.



The daily countplot shows that Friday is the most crime-prone day of the week, whereas Sunday experiences the least criminal activity. This insight could reflect societal patterns and routines that vary throughout the week.

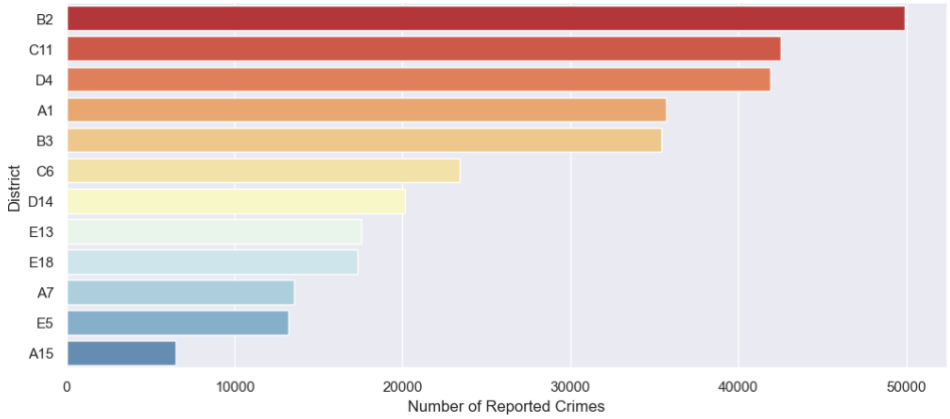


The hourly distribution of crimes reveals that the highest number of incidents is recorded at 5 p.m., while 5 a.m. is the quietest hour. Generally, night hours tend to have fewer reported crimes, which might be related to the lower activity levels during these times.



The time series plot of daily crimes highlights notable fluctuations in crime rates. Days vary considerably, ranging from 150 to over 350 crimes. A seasonal trend is apparent: crime rates decrease towards January, rise to a peak around July, and then diminish again. This pattern suggests a strong seasonal influence on crime trends in Boston, potentially linked to variations in social activities and environmental factors throughout the year.

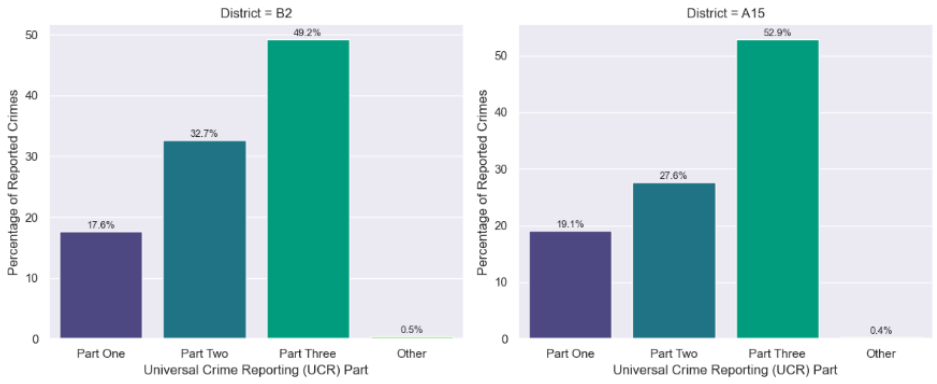
To dive deeper into the data to understand where these crimes are occurring, we analysed the distribution of reported crimes across different districts to gain insights into which areas of Boston experience the highest and lowest crime rates. We plotted an horizontal bar chart showing the number of reported crimes per district. Each bar represents a different district, labelled with what appears to be a unique code (ex: 'B2', 'C11', 'D4', etc.).



From this graph, the disparity between the districts in terms of the number of reported crimes is evident. District 'B2' shows the highest number of crimes, followed closely by 'C11' and 'D4', potentially indicating a higher level of criminal activity, higher population density, or higher effectiveness in the crime reporting process. In sharp contrast, district 'A15' has the lowest number of crimes reported, which could reflect a variety of factors including lower crime rates or demographic differences.

After observing the considerable difference in the number of reported crimes between the various districts but also between district B2, with the highest number, and district A15, with the lowest number, a further question arose: is this discrepancy also reflected in the seriousness of the crimes committed in these districts? It is important to examine not only the quantity but also the nature of the crimes to obtain a more complete picture of the security situation in these areas.

To investigate this aspect, we analysed the type of crimes reported in both districts (Using UCR_PART column, which represents a statistical system developed in the United States to collect and analyse crime data where crimes are classified according to seriousness and range from very serious in Part One to less serious in Parts Two and Three.). This allows us to understand whether, for example, district B2 not only has a higher number of crimes but also tends to report more serious incidents than district A15, or whether the numerical difference is mainly attributable to factors such as population density or the effectiveness of the crime reporting system.



From the graph, it turned out that district A15 had fewer 'Part Two' crimes than district B2.

While district B2 shows an overall higher volume of reported crimes in all categories, district A15 reveals a different proportion, with fewer 'Part Two' crimes. This difference could reflect a variety of factors, such as different levels of vigilance, police responses, or socio-economic characteristics of the districts.

To quantify and verify the actual difference in the distribution of crimes between districts B2 and A15, we applied the chi-square test to the frequencies of crimes reported as 'Part One', 'Part Two', 'Part

Three' and 'Other' according to the UCR classification. Completing this picture with a frequency distribution table, we observe that district B2 has about 5% more crimes reported as 'Part Two' according to UCR than district A15. Furthermore, when comparing crime categories, we see that district B2 has about 3.5% fewer crimes reported as 'Part Three' than A15.

The p-value obtained from the chi-square test is really close to 0, so it confirms with statistical significance the discrepancy between the two districts not only in terms of volume but also in terms of the severity of the crimes committed. This implies that the severity of crimes reported in these districts is not the same and varies considerably. These statistical results provide a solid basis for stating that the observed differences are not random or the result of natural fluctuations, but rather indicate a substantial variation in crimes between the two districts.

These observations provide an initial understanding of the crime landscape in Boston, highlighting the importance of temporal and spatial factors in crime analysis. The next sections of the report will delve deeper into these aspects, examining the interplay between time, location, and types of crime incidents.

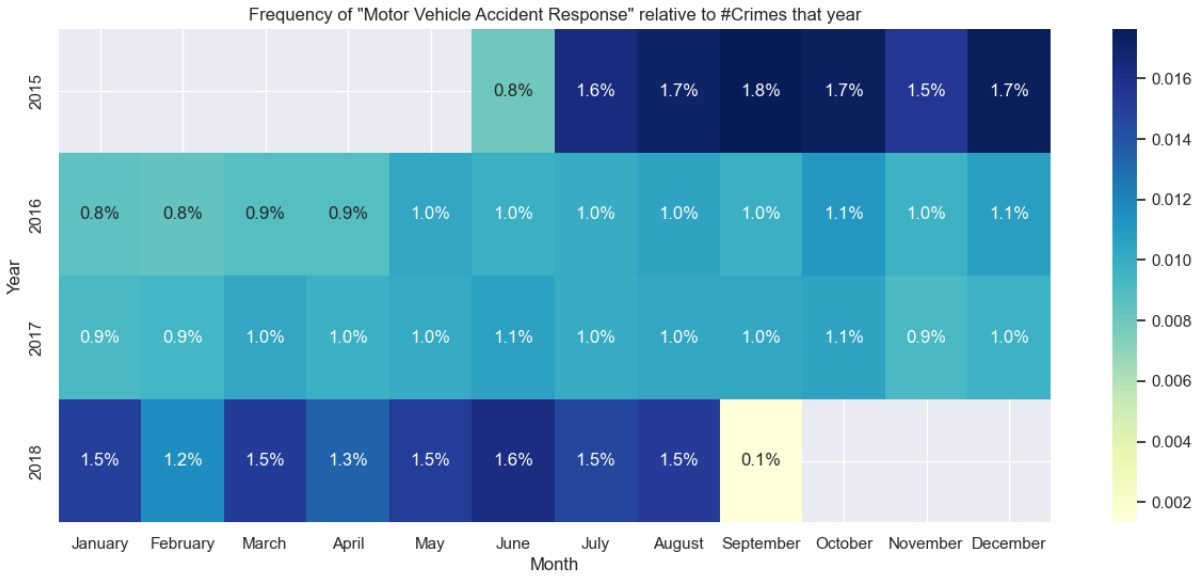
Specific Crimes – Data Analysis

After generically analysing the data and their spatio-temporal distribution, we focused our analysis on specific targeted crimes. The crimes we chose were chosen both for relevance and because we consider some crimes very serious and were interested in extracting interesting insights from them.

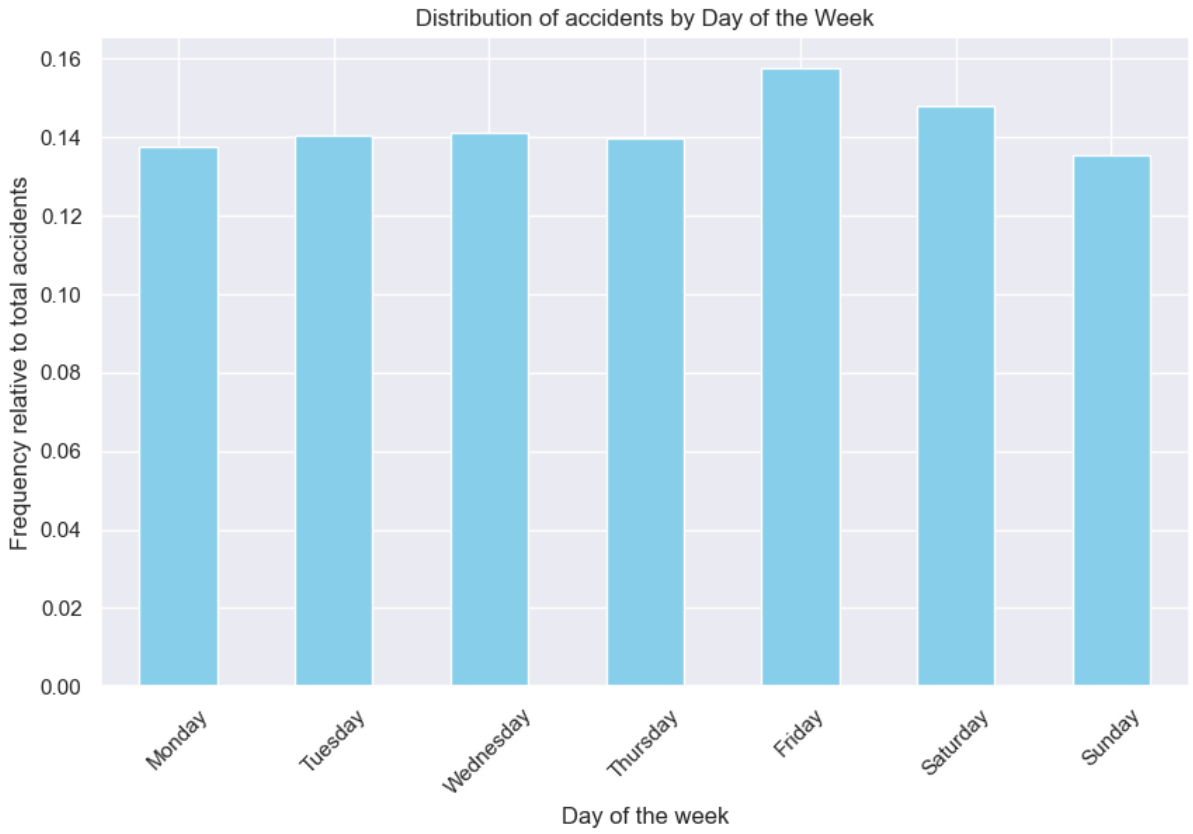
In order to avoid losing the focus of the analysis, I will only report below the views that actually showed particular trends.

Motor vehicle accidents

The detailed analysis of crimes began by investigating motor vehicle accidents, the crime with the most records within the dataset. Considering the timeframe bias of the data and to obtain the most precise analysis possible, we calculated the frequency relative to the number of crimes in each year for each crime and show timeseries of those frequencies.



Analysing the behaviour over time, we can denote that 2015 and 2018 show significantly higher relative frequencies than the other years. This suggests that there were relatively more road accidents in these years than in the other years depicted in the graph.



Analysing the accident distribution graph by day we noticed that weekend accidents are much more frequent, hypothetically due to more people on the roads given the days off.

Larceny

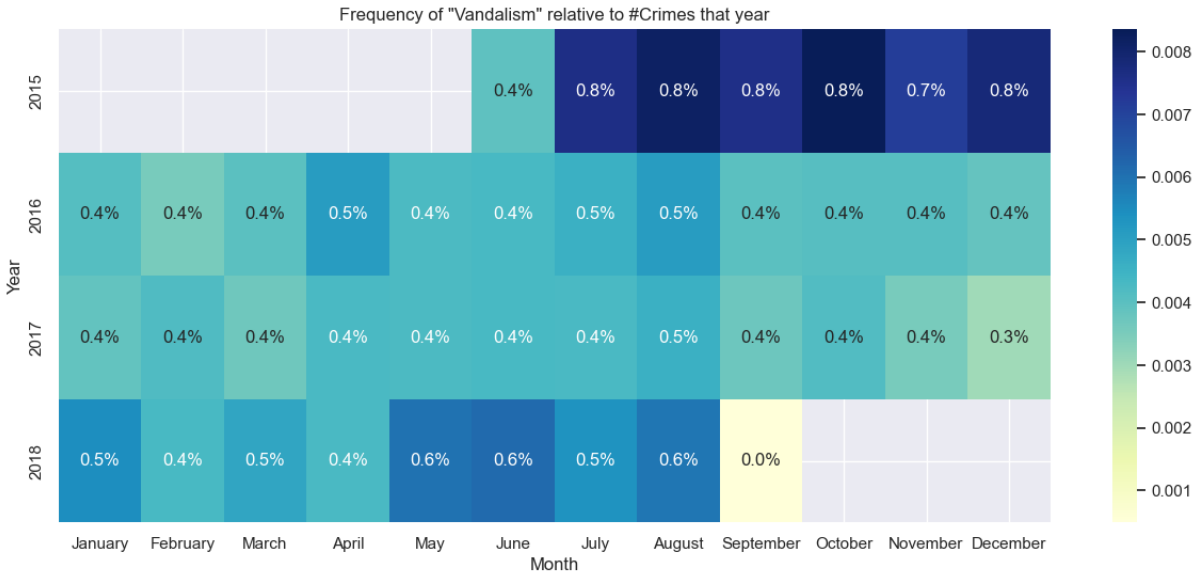
We decided to analyse the crime distribution by hour of the day:

From the graph breaking down the distribution by time of day, we can see that the relative frequency of Larceny crime increases significantly during daylight hours, peaking in the afternoon hours, around 5-6 p.m., which are the hours when one might expect more traffic of people and activity or a greater tendency to leave items unattended in public places during these hours.

In addition, some businesses close in the late afternoon or evening, which may provide opportunities for unobserved theft. The relative frequency then gradually decreases during the evening until it reaches its lowest levels in the early morning hours, when there is presumably less activity on the streets and opportunities for theft decrease.

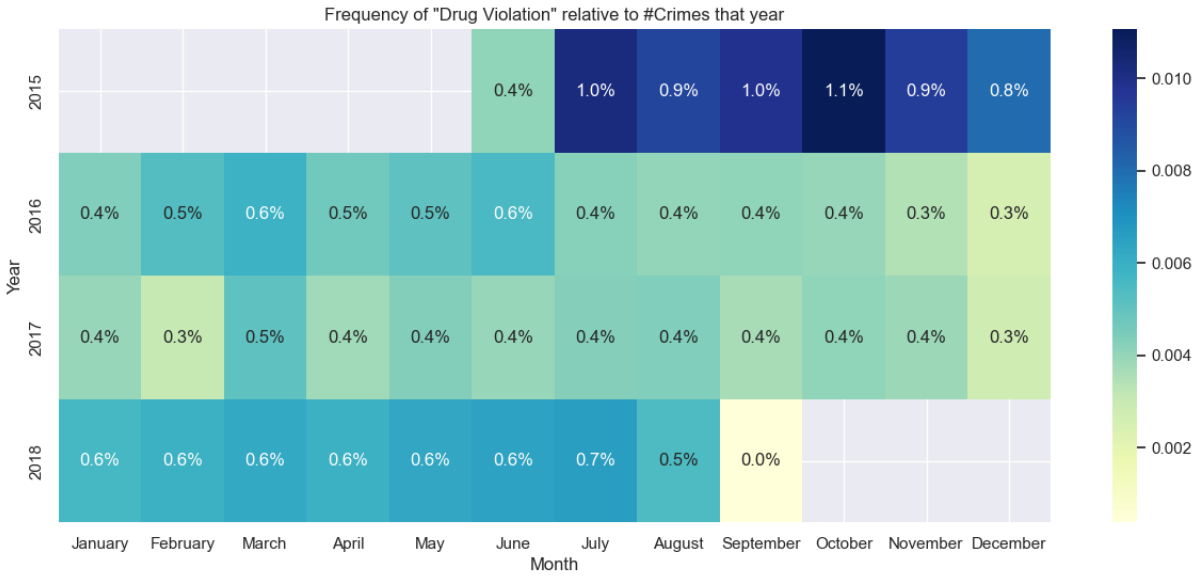
Vandalism

By analysing the behaviour over time (relative to crime numbers on these years) we can spot some trends.

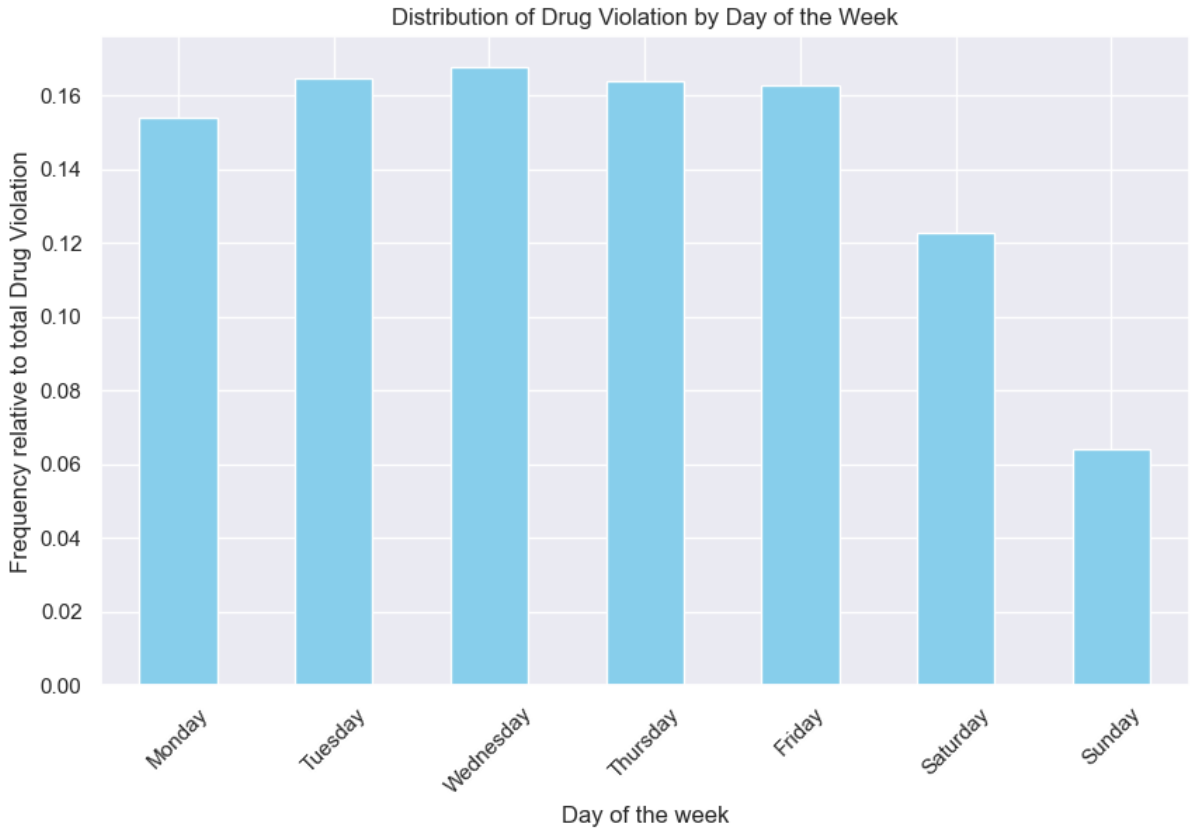


The plot above shown us a slight decreasing trend of vandalism over last years.

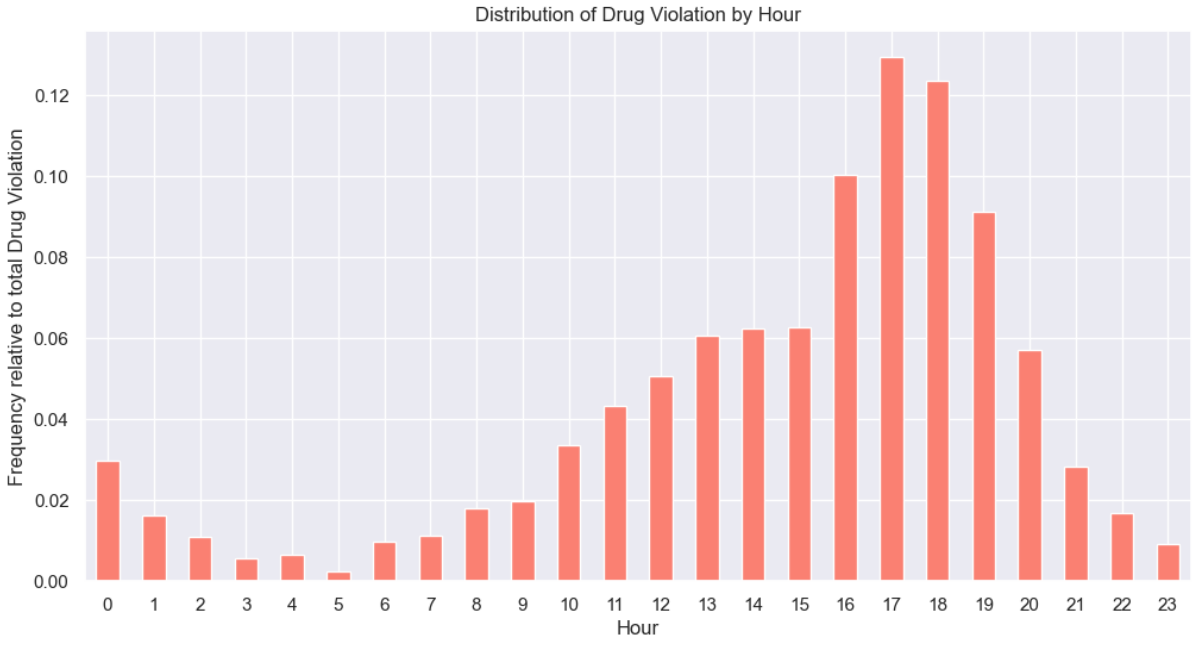
Drug Crime



Analysing the relative frequency across years and months, we can denote a general downward trend in the relative frequency of drug-related crimes from 2015 to 2017, with then a slight recovery in 2018 where frequency increased slightly compared to 2017.

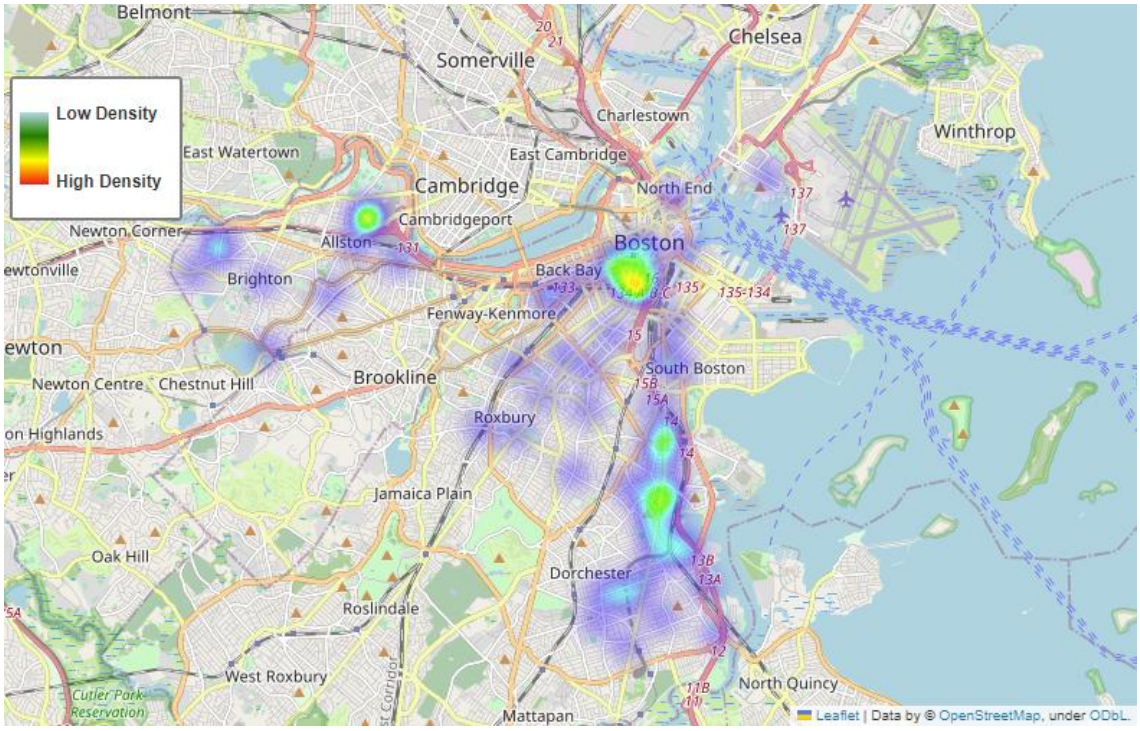


Proceeding with the analysis of the distribution by days of the week, it is interesting to note an opposite trend to the crimes analysed up to this point, in fact Saturday and Sunday show a significantly lower frequency than weekdays, where the incidence of drug-related crimes remains constant. This could suggest less drug-related criminal activity on weekends or reduced police attention on those days.



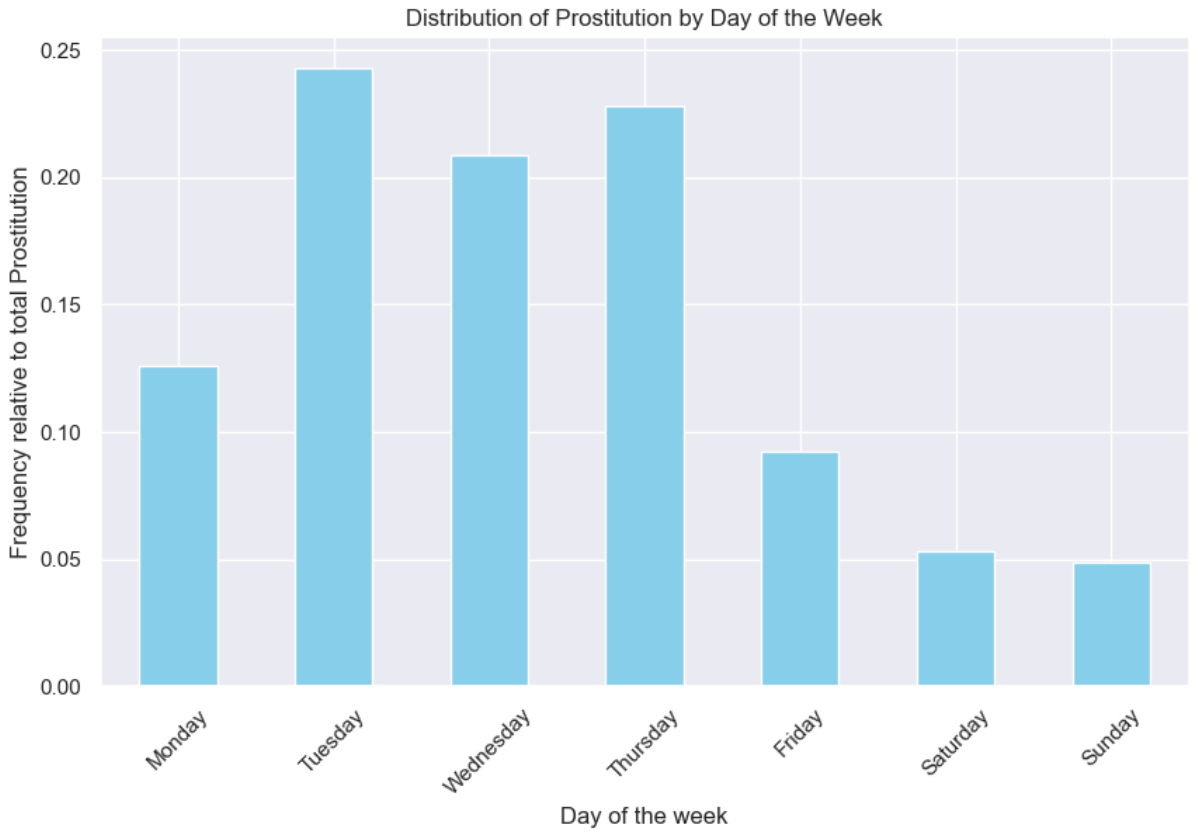
On the other hand, the graph broken down by daily time shows a concordance with the general data, in fact there is a clear peak in the evening hours, particularly between 6 p.m. and 10 p.m. This could be due to the increased social activity during these hours or could reflect the time when such crimes are more easily observed and therefore reported, whereas the hours from early morning until early afternoon show a much lower frequency of drug-related crimes, which could be due to less public activity or increased surveillance during these hours.

Prostitution



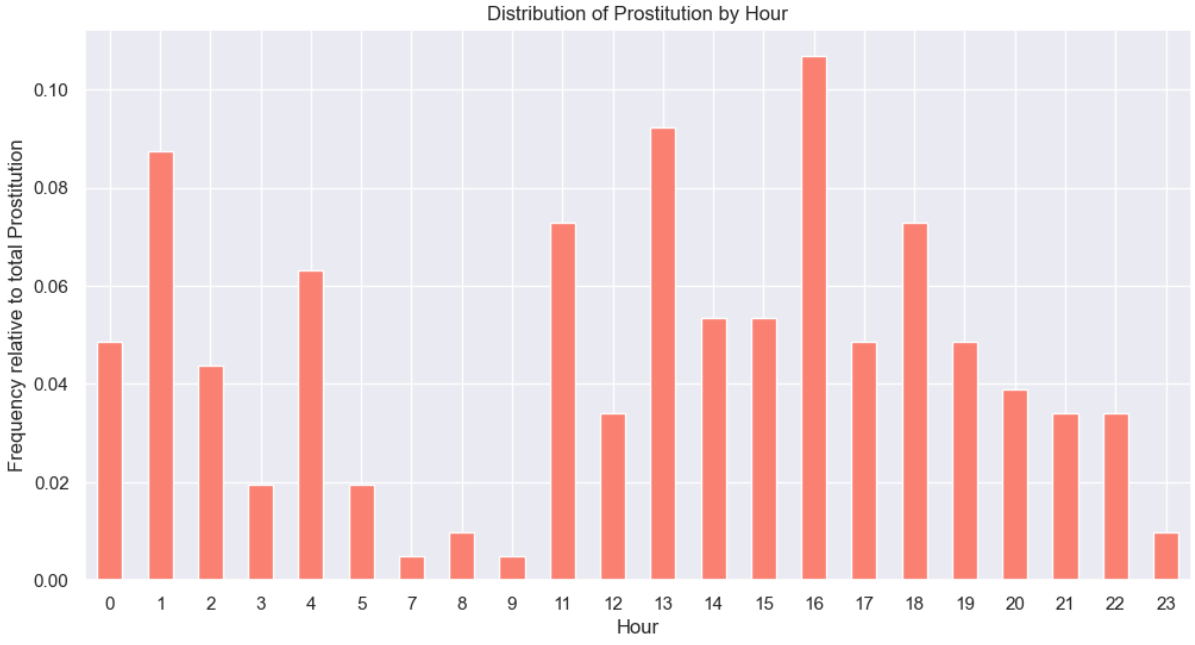
From the map we can see a lower density of reported crimes. In fact, this crime was chosen out of interest and given the recent events on violence against women, it was decided to also undertake an analysis not directly related but noteworthy.

There appear to be some hot spots, particularly in the central areas near Boston's urban core and in areas immediately adjacent to it. This could correspond to areas with increased nighttime activity or neighborhoods with socioeconomic characteristics that favor this type of crime. Also, areas along major road corridors appear to have higher densities, suggesting that the ease of access and anonymity provided by busy arteries may be factors contributing to the prevalence of crime.



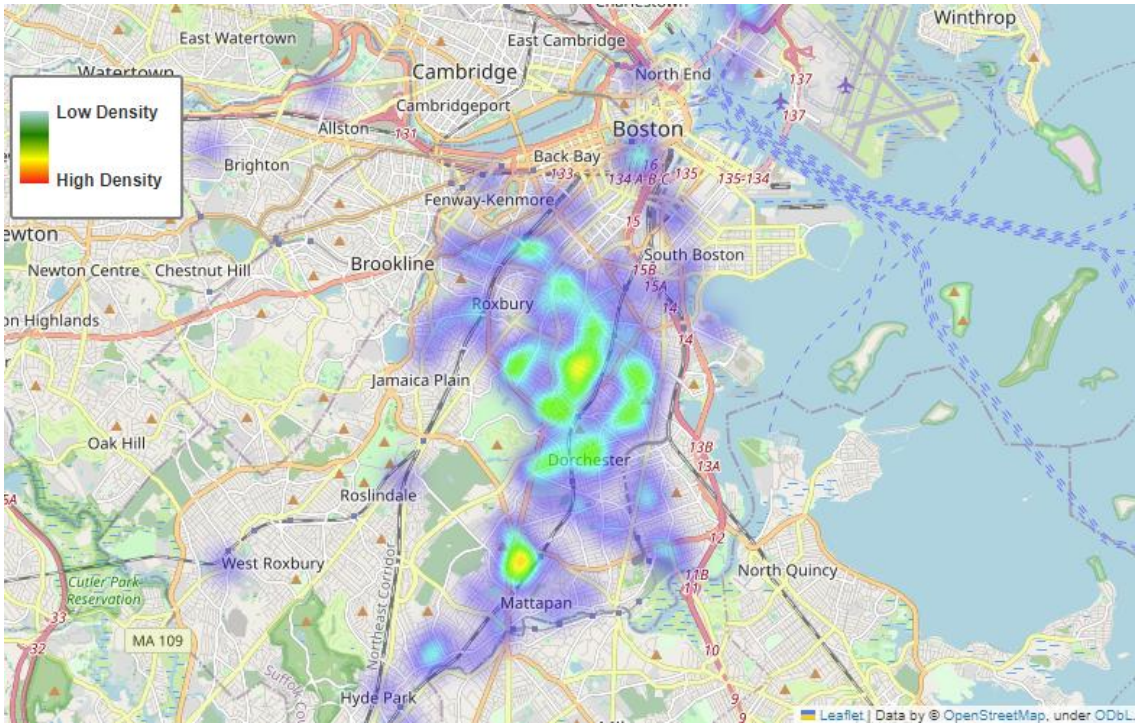
Furthermore, by analyzing the relative frequency of prostitution crimes for each day of the week, it can be deduced that:

- Wednesday shows the highest relative frequency compared to other days, which is interesting and could reflect social or law enforcement dynamics specific to that day.
- Saturday and Sunday show a significantly lower frequency, which might be unexpected given the common perception of more weekend activity. This could indicate less law enforcement attention or fewer reports on those days.



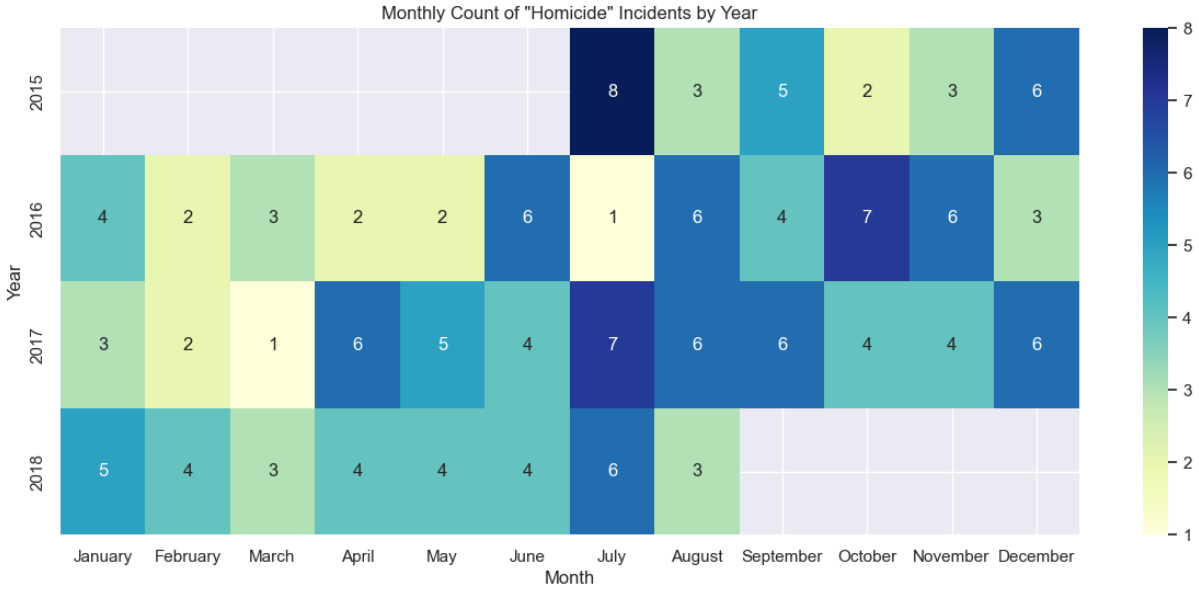
The hours graph also has an unusual distribution with a significant peak around 3 p.m., and contrary to what might be expected, there is a drop in the relative frequency of prostitution crime reports in the early evening hours and then renewed around 1 a.m.

Homicide

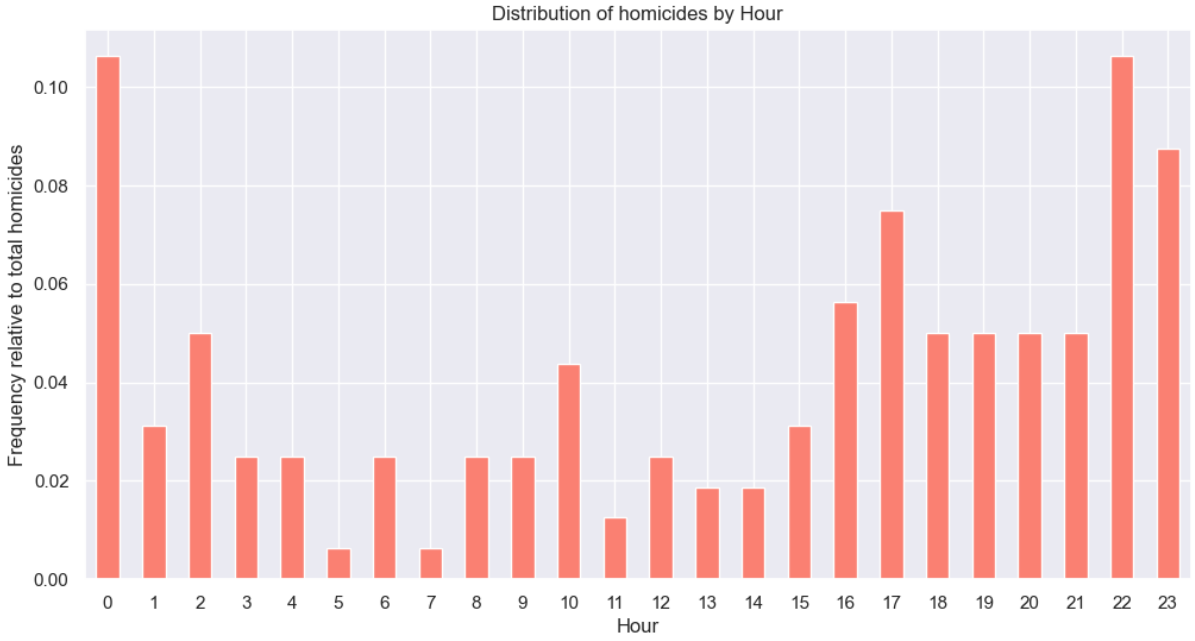


It can also be denoted from this map depicting the distribution of homicides, how there is a higher density of crime in the urban areas of Boston, indicating that the central areas are the scene of more criminal activity for both prostitution and homicide. Furthermore, comparing this map with the prostitution map, one can denote the presence of specific points, particularly in the central and southeastern areas of Boston, where crime density is particularly high in both maps, suggesting that these areas may have common risk factors for different types of crimes.

Then, since homicide are very few we will plot absolute values of them over time.



There are obvious high variations in the frequency of homicides. We can denote a great variance in month-to-month frequency given by the fact that murders occur infrequently. For example, 2016 shows a peak in October (7 homicides), while 2017 shows a peak in July (7 homicides), where instead 2016 have only 1 occurrence. In several years, first months of the year (January-march) tend to have a lower frequency of homicides, which could reflect a seasonal influence on criminal behavior. But analyzing years, there is no clear pattern of increase or decrease year by year, indicating that homicide frequency may be influenced by variable factors that change from year to year.



Clear peaks in relative frequency are noted in the nighttime and early morning hours, with the highest around 0 a.m. (midnight) and another significant peak around 10 p.m. (10 p.m.) and 11 p.m. (11 a.m.). This could suggest that homicides are more likely to occur during the hours of darkness, which could be related to lower witness attendance and reduced activity providing more opportunities for crimes of this severity. We can also detect a relatively lower frequency of homicides, with the lowest points around 9 a.m. (9 a.m.) and 1 p.m. (1 p.m.). This could reflect the greater movement of people and activity during daylight hours that could act as a deterrent. We can infer that there is a visible fluctuation in frequencies between different times of the day, suggesting that the risk of homicide may vary greatly by time of day.

Findings and Conclusion

In this analysis, we examined Boston crime data in detail, focusing on seasonal trends and geographic distribution, as well as the most common times and days when crimes occur. However, it is critical to emphasize that crime analysis requires a comprehensive understanding of the socioeconomic, demographic, and cultural context in which these events occur. Without additional data on population, income levels, education, unemployment rates and other socio-economic factors, it becomes difficult to draw accurate conclusions about the origin or causes of crimes. In addition, detailed demographic data, such as the age and gender of those involved in crimes, could offer additional keys to better understanding the crime landscape.

Therefore, this analysis is only a starting point for understanding the dynamics of crime in Boston. It emphasizes the importance of conducting further research that includes contextual data and a more in-depth assessment of the factors that contribute to crimes. This could include sociological, economic and demographic analyses to identify more significant correlations and trends.

In conclusion, the data collected so far provide a solid foundation for further research in the field of criminology and crime prevention in Boston. This work could serve as a springboard for future more detailed analyses and targeted intervention strategies to improve safety and quality of life in the city.