
**Boston Crimes: Milestone 1 EDA**

Aayushi Gupta
Atharva Shantanu Kulkarni
Dhwani Patel

College of Professional Studies, Northeastern University

ALY 6040: Data Mining Applications

Mr. Joe Reilly

October 4, 2022

## Business Problem

Crime and drug use remains to be one of the top issues in the US as they continue to rise at an alarming rate. According to the Major Cities Chiefs Association report, overall violent crime increased 4.2% from January 1 to June 30, 2022, compared to the same period in 2017 (*Contreras, 2022*). Although the number of homicides and rapes have shown a significant decrease since last year, the robberies and aggravated assaults have spiked up by 4.2% from January 2022 to June 2022.

Traditional crime-solving methods are ineffective in the current climate of rapidly rising crime because they are labor-intensive and inefficient. Therefore, it would lessen the strain on police and aid in crime prevention if we could develop ways to accurately forecast crime before it happens or create a "machine" to support police officers (*Shah, 2021*). So our goal of the project is to come up with a data mining technique which can help forecast crimes, especially violent crimes at a particular region and time in the Boston area. Once successfully trained and tested, this model can be scaled up to include the bigger regions and states as well.

## Tools and Technologies Used

This project will be developed in Jupyter notebook using python programming language. We will be using Tableau alongside Python to create visualizations and come up with insights that can benefit the model.

## Data Description

The Boston Police Department (BPD) provides crime event reports to record the preliminary information around an occurrence to which BPD officers respond. This dataset contains data from the new crime event report system, which has fewer fields designed to capture the incident's type as well as its timing and location (Crime Incident Reports (August 2015 - to

Date) (Source: New System) - Analyze Boston, n.d.). Records contain data from the year 2019-2022.

The data source is Analyze Boston. The dataset has 283588 observations recorded over the period of almost 4 years from 2019-2022. The number of variables present are 17 and have mixed datatypes- int, object and float. A few columns of interest might be the offense description, time and day of crime, district and latitude longitude.

## Data Preprocessing

The first step in any data mining technique is the cleaning of data. In order to clean the data, we analyzed the null values (missing values) in the data. On checking for missing values, it was found out that the offense_code_group and ucr_part consist of only null values. So, we removed the columns from our data. The columns District and street had 2844 and 683 missing values respectively. Since the dataset is quite large and these values are less than 5% of the entire dataset, we were able to drop the rows with null in either district or street. No duplicates were found in the database. The new shape of the dataframe came out to be 282304 rows and 15 columns. The target variable would be our offense description which would be simplified at a later stage to group the crimes into violent and non-violent. There are no outliers to be seen in the moment but as we progress to group our crimes in the next stage and create cluster regions, we will tackle the outliers if they arise.

## Exploratory Data Analysis

On taking an overall look at the summary statistics (Refer Appendix 1 Fig 2), it can be seen that the minimum number of crimes occur in January around 12 am. The maximum number of crimes seemed to occur in December around 11 pm.

Data visualization is is an important step in any data mining technique By providing it with a visual context via maps or graphs, data visualization helps us understand what the data means. As a result, it is simpler to spot trends, patterns, and outliers in enormous data sets since the data is easier for the human mind to understand.

The first question we tried to answer with this data on Boston PD is to understand ***what type of crimes are the most common?***. For which we obtained the count of unique values in the column *Group in a variable top_10,* we made use of *value_count()* function in pandas which gives back a series with count of distinct values. The resulting object will be arranged in descending order with the first element being the one that appears the most frequently. By default, it excludes NA values.

To visualize the crimes in Boston we made use of bar plot using *sns.barplot* function from seaborn package and feeding the *top_10 described earlier variable* a *x* to plot a bar graph. (Refer Appendix 1 Fig 3). Most common crime was registered as "investigate person", followed by "property damage- leaving scene" and "Sick assist".

Next, we tried to answer the question ***whether the frequency of crime changes over different months? To visualize*** this data we made use of *sns.countplot* available in seaborn package, this function use bars to visually represent the numbers of observations in each categorical bin. A count plot resembles a histogram over a categorical variable as opposed to a quantitative one (Refer Appendix 1 Fig

As we can see in the above plot the distribution of crimes registered monthly in years from 2019 – 2022, looking at the plot we can see that most of crimes have taken place in the months of August followed by July and September suggesting that majority are in the summer months.

To see if that's the case we divided the months into seasons, such as if month is between 12(Dec) - 2(Feb) its winter, if its 3(Mar) - 5(May) its Spring, if its 6(Jun) -8(Aug) its summer and rest to be Fall (reference Fig4.). Output can be seen in Appendix 1 Figure 6. With this plot we can surely say that most of the crimes have taken place in Summer months, the reasoning could vary from area to area but the frequency of crimes in summer months is highest.

The dataset also includes information about districts, this helped us answer the question ***Which places did crime mostly occur?*** We can see that most of the crimes took place in district B2 and the least in A15. (Refer Appendix 1 Figure 7)

Here we also try to answer the question as to ***what time of day*** the crime took place, for which we made use of the column "Hour" and divide it into if greater and equal to 12 its evening and anything less and equal to 5 is Night else morning (Appendix Fig.5) to get the output. This graph suggests that most crimes happened during Evening time of the day and Morning time of the day. (Refer Appendix 1 Figure 8)

The graphical representation of data in exhibit 1(Refer Appendix 1 Fig 9) provides an overview of crime rates in the city of Boston during each month beginning from January 2019 till September 2022. Each line shown in the area represents type of crime with the details of average shooting occurred at the monthly rate. July –September 2019 seems to have the highest reported crimes including aircraft incidents, assaults, drugs class b violations, property investigation and ballistic evidence found. In the month of September 2019, less weapon crimes have been reported with average shooting as 0.000 and reporting area being sum of 700. To dig deeper, looking at exhibit 1.1 (refer appendix 1 fig 10) will give a clear picture about specific offense-based progression each year.

The bar chart (Fig 11 exhibit 3) helps understanding the district vice crime for from 2019-2022. District codes have been considered for ease and availability of attributes. That would help the police department and other social services departments identify if in specific areas the crime rate is increasing or decreasing. The highest crime was committed in B2 district, which is Roxbury, then the second highest is D4 i.e. South End and the third highest was captured in C11 i.e. Dorchester(bpdnews,2022).

Exhibit 4 (Refer Appendix 1 Fig 12) represents detailed information about crimes increasing or decreasing on specific streets in the districts. Through this a safety analysis can be conducted which would help in getting an idea if the street would be safe & accessible around a specific time. For example, on 25 Brinsley st, Dorchester, MA 02121 it might not be safe around 11PM to roam from the previous data of crimes happened. The size of the bar indicates the average number of instances that have taken place. So, if the bar is thin which indicates that less cases have been reported as well as the street is safer than the one that has thicker bar.

## Proposed Next Steps

The next step would be to do an even deeper dive into the data to find more interesting insights which can help the modeling technique. We aim to deploy classification models to forecast crimes at a particular region and time. The first step would be to cluster the regions and identify the outliers (regions beyond the considered scope). Next the crimes would be grouped as violent and nonviolent. Then we aggregate the data and classify each row as 0 or 1 depending on whether a violent crime occurred at that location (*Uzel, 2021*). We also wish to integrate the historical weather data to add as another feature. We will then move on to dividing our dataset into training and testing sets and build a model whose focus is to predict positive labels.

## References

Contreras, R. (2022, September 10). *Survey: Homicides down midyear as overall violent crime jumps*. Axios. Retrieved October 4, 2022, from https://www.axios.com/2022/09/10/homicides-down-midyear-overall-violent-crime-up

*Crime Incident Reports (August 2015 - To Date) (Source: New System) - Analyze Boston*. (n.d.). Retrieved October 4, 2022, from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

Shah, N. (2021, April 29). *Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention - Visual Computing for Industry, Biomedicine, and Art*. SpringerOpen. Retrieved October 4, 2022, from https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z

Uzel, A. (2021, December 13). *Demand Forecasting: Boston Crime Data - Towards Data Science*. Medium. Retrieved October 4, 2022, from https://towardsdatascience.com/demand-forecast-boston-crime-data-64a0cff54820

# Appendix 1

| | INCIDENT_NUMBER | OFFENSE_CODE | OFFENSE_CODE_GROUP | OFFENSE_DESCRIPTION | DISTRICT | REPORTING_AREA | SHOOTING | OCCURRED_ON_DA | YEAR | MONTH | DAY_OF_WEEK | HOUR | UCR_PART | STREET | Lat | Long | Location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 102091671 | 2647 | NaN | THREATS TO DO BODILY HARM | B3 | 417 | 0 | 2019-11-12 12:00 | 2019 | 11 | Tuesday | 12 | NaN | MORA ST | 42.282082 | -71.073649 | (42.28208197671972, -71.07364874515646) |
| 1 | 102095489 | 3115 | NaN | INVESTIGATE PERSON | E18 | 520 | 0 | 2019-11-25 16:30 | 2019 | 11 | Monday | 16 | NaN | POYDRAS ST | 42.256216 | -71.124019 | (42.256215920402155, -71.12401947329023) |
| 2 | 102096818 | 2905 | NaN | VAL - VIOLATION OF AUTO LAW | A1 | | 0 | 2019-11-30 21:00 | 2019 | 11 | Saturday | 21 | NaN | SUDBURY ST & CAMBRIDGE ST\nBOSTON MA 02108\nU... | 42.360866 | -71.061316 | (42.360866027118476, -71.0613160019785) |
| 3 | 129092894 | 3201 | NaN | PROPERTY - LOST/ MISSING | NaN | 503 | 0 | 2019-11-16 13:30 | 2019 | 11 | Saturday | 13 | NaN | AMERICAN LEGION HWY | 42.284467 | -71.111831 | (42.28446742674232, -71.11183088758158) |
| 4 | 129099920 | 3301 | NaN | VERBAL DISPUTE | B2 | 330 | 0 | 2019-12-12 07:50 | 2019 | 12 | Thursday | 7 | NaN | COLUMBIA ROAD | 0.000000 | 0.000000 | (0, 0) |
| ... | ... | ... | ... | ... | ... | ... | ... | | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 285383 | 149210376 | 3110 | NaN | SERVICE TO OTHER AGENCY | B2 | | 0 | 2022-01-02 00:00 | 2022 | 1 | Sunday | 0 | NaN | WASHINGTON ST & ROXBURY ST\nBOSTON MA 02119\n\n... | 42.329600 | -71.084620 | (42.32959998129787, -71.08462000476281) |
| 285384 | 122207734 | 3831 | NaN | M/V - LEAVING SCENE - PROPERTY DAMAGE | B3 | 465 | 0 | 2022-02-01 18:00 | 2022 | 2 | Tuesday | 18 | NaN | BLUE HILL AVENUE | 0.000000 | 0.000000 | (0, 0) |
| 285385 | 122201046 | 2101 | NaN | OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL | B2 | | 0 | 2022-02-12 21:54 | 2022 | 2 | Saturday | 21 | NaN | BLUE HILL AVENUE | 0.000000 | 0.000000 | (0, 0) |
| 285386 | 122056850 | 3801 | NaN | M/V ACCIDENT - OTHER | D4 | 271 | 0 | 2022-07-28 19:40 | 2022 | 7 | Thursday | 19 | NaN | MASSACHUSETTS AVE & HARRISON AVE\nBOSTON MA 0... | 42.334910 | -71.075170 | (42.33490995083834, -71.07517004893332) |
| 285387 | 22031068 | 3114 | NaN | INVESTIGATE PROPERTY | C11 | 249 | 0 | 2022-05-02 08:00 | 2022 | 5 | Monday | 8 | NaN | SAXTON ST | 42.313591 | -71.054324 | (42.3135909105722, -71.05432387045848) |

Fig 1: Initial Dataset

| | OFFENSE_CODE | SHOOTING | YEAR | MONTH | HOUR | Lat | Long |
|---|---|---|---|---|---|---|---|
| count | 282304.000000 | 282304.000000 | 282304.000000 | 282304.000000 | 282304.000000 | 282304.000000 | 2.823040e+05 |
| mean | 2357.052362 | 0.011923 | 2020.336014 | 6.312256 | 12.862418 | 40.786949 | -6.850487e+01 |
| std | 1206.503032 | 0.108541 | 1.107375 | 3.276033 | 6.425524 | 7.913779 | 1.329174e+01 |
| min | 100.000000 | 0.000000 | 2019.000000 | 1.000000 | 0.000000 | 0.000000 | -7.134947e+01 |
| 25% | 1102.000000 | 0.000000 | 2019.000000 | 4.000000 | 9.000000 | 42.293989 | -7.109757e+01 |
| 50% | 3005.000000 | 0.000000 | 2020.000000 | 6.000000 | 14.000000 | 42.324922 | -7.107626e+01 |
| 75% | 3201.000000 | 0.000000 | 2021.000000 | 9.000000 | 18.000000 | 42.348152 | -7.105976e+01 |
| max | 99999.000000 | 1.000000 | 2022.000000 | 12.000000 | 23.000000 | 42.461410 | 5.249691e-08 |

Fig 2: Summary Statistics

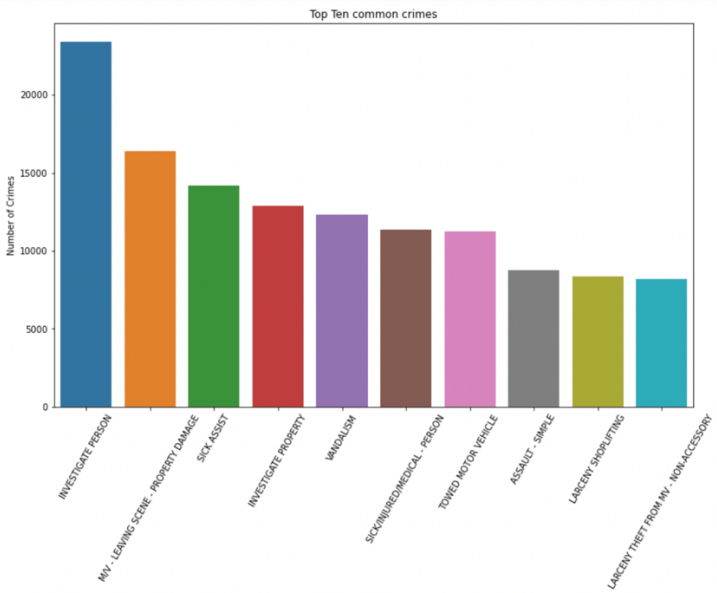


Fig 3: Top 10 Crimes in Boston

```
##getting seasons for visualization

def Seasons(mon):
    if (mon == 12 or mon == 1 or mon == 2):
        return "Winter"
    elif(mon == 3 or mon == 4 or mon == 5):
        return "Spring"
    elif(mon ==6 or mon==7 or mon == 8):
        return "Summer"
    else:
        return "Fall"
```

Fig.4: Formatting months into Seasons

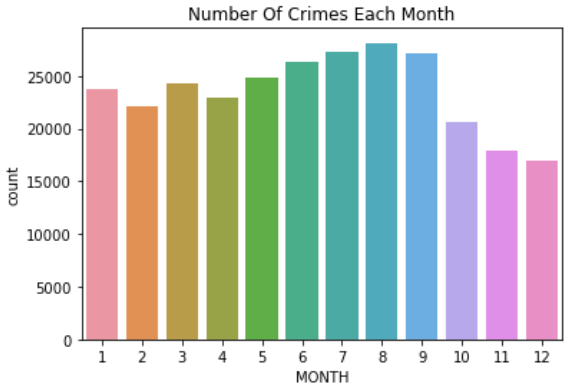


Fig 5: Crime rate in different months

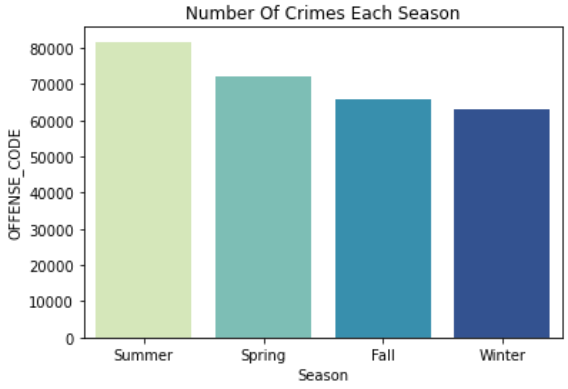


Fig 6: Crimes in different seasons

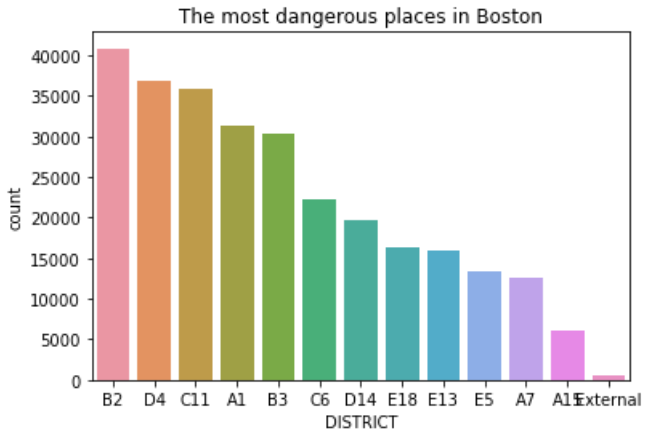


Fig 7: Most dangerous places in Boston

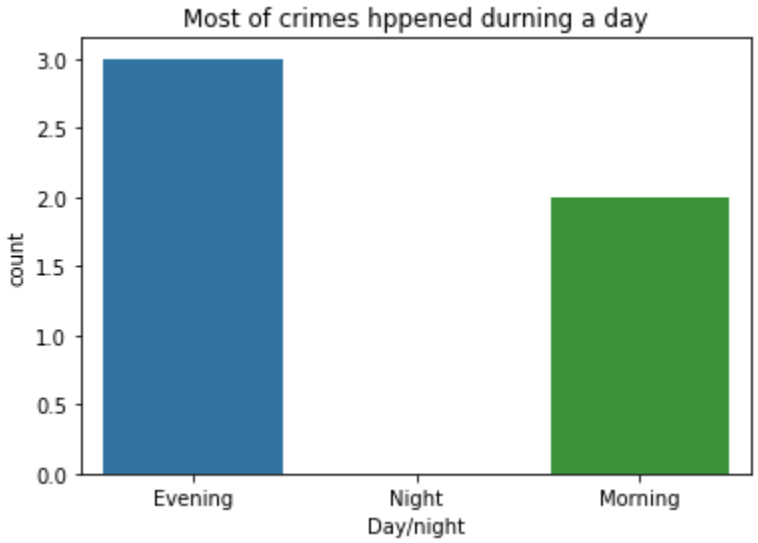


Fig 8: Crimes during different times of the day

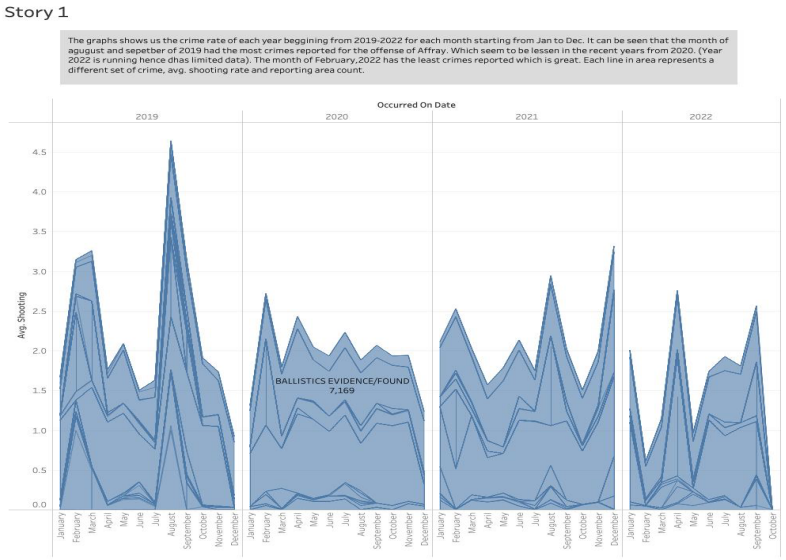


Fig 9: Exhibit 1

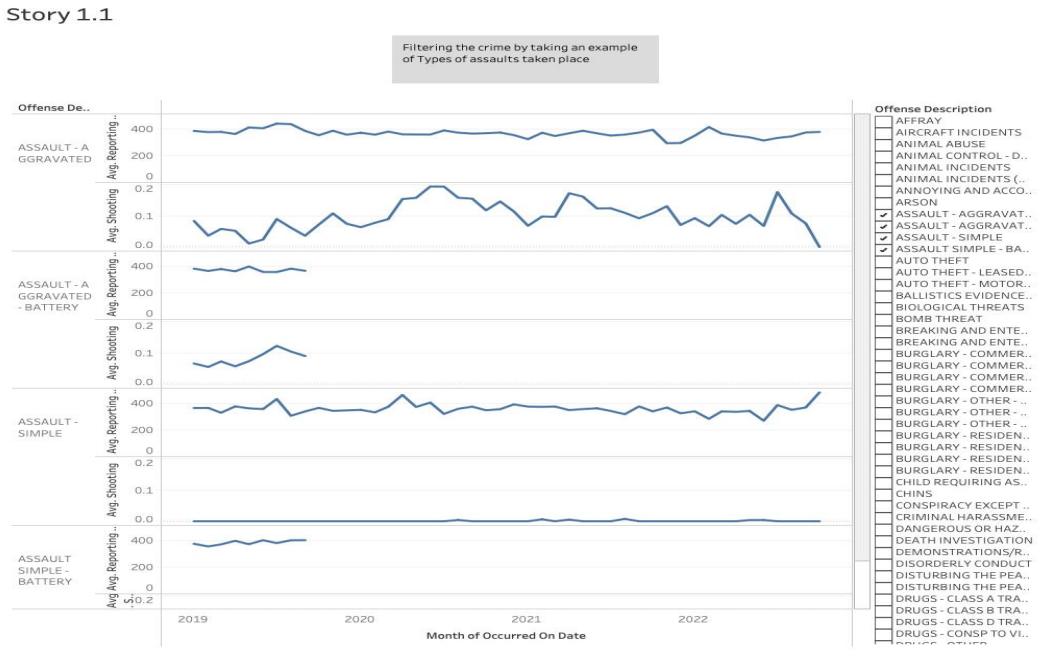


Fig 10: Exhibit 1.1

Story 3

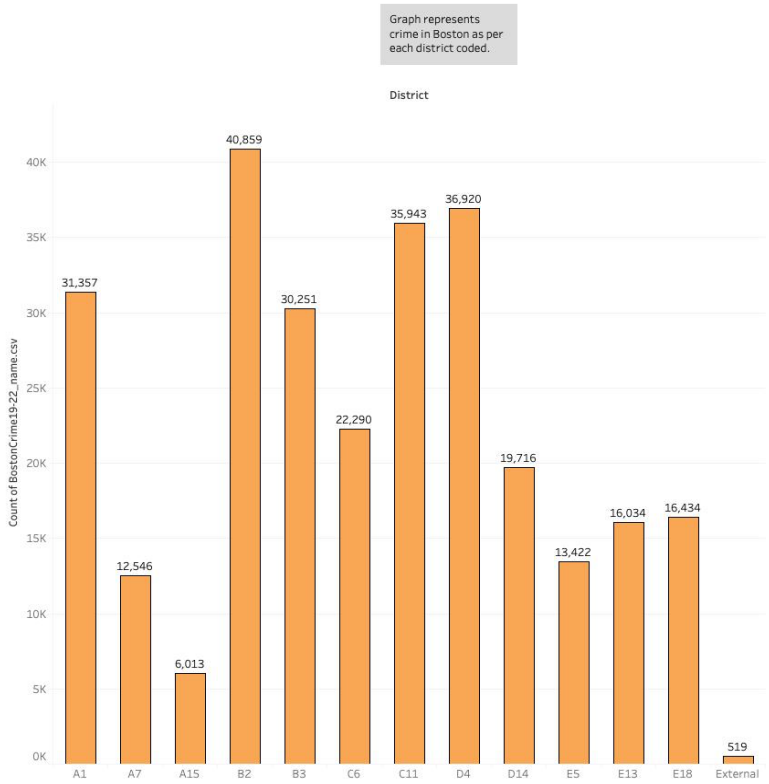


Fig 11: Exhibit 3

Story 4

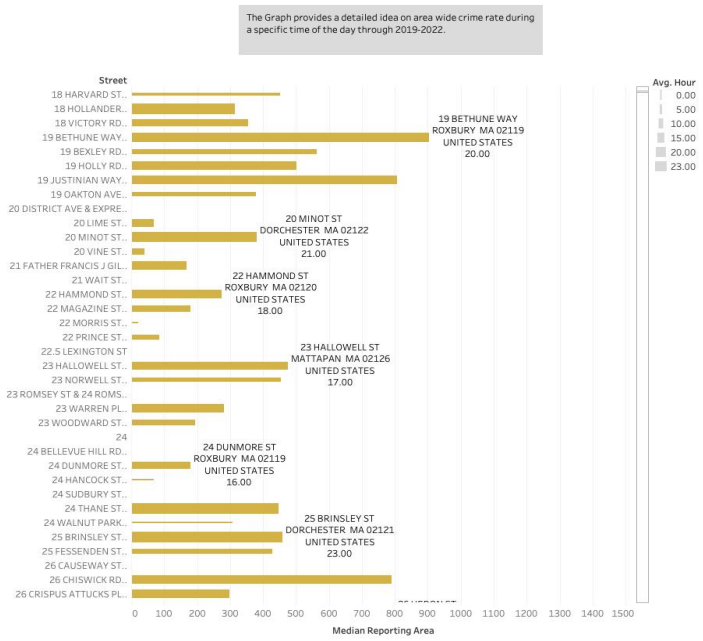


Fig 12: Exhibit 4