STA5069Z - Project Proposal

Sebastian Eide Aas

March 2025

1 Introduction

For those who are familiar with crime rates in the United States, it is well known that several cities, such as Boston and Chicago, experience high crime levels. In fact, the number of violent crimes per 100.000 residents in Boston exceeded the national average between 2010 and 2019 [1]. In this case, the FBI defines violent crimes as "offenses involving the threat of force". This raises the questions: Are certain areas in Boston more prone to a specific type of crime than others? Are there any distinct crime spots, or is the crime distributed over the city? This report aims to answer these questions through multivariate statistical methods, exploring how crime data can be used to identify high-crime areas. In many large cities, crime is concentrated in specific areas due to for example socioeconomic factors. We hypothesize that Boston will exhibit similar patterns, with for example property damage/theft likely concentrated in business areas and violent crimes concentrated in residential areas.

2 Data Description

The Boston crime data set was collected from the Boston Police Department (BPD). Its purpose was to document the details surrounding an incident that BPD responds to, and the data was published by the Department of Innovation and Technology [2]. The data is provided as different files for each year, but was merged by the publisher on Kaggle. Link to the data dictionary in the GitHub repository can be found here.

The data set provides us with an unique identifier for each incident. It is important to note that each incident number corresponds to when an officer or patrol has responded, so there might be some duplicates we have to take into consideration. We have the offense code, offense code group, and offense description. This would allow us to correctly create types of offenses, which can further be used to cluster. As we see from Figure 1, we could for example have one category of non-criminal activities including the first three variables. We could do one group of property damage/theft and vandalism, and one group of violent crimes. The variable shooting, which is binary, will also contribute to the offense type.

Additionally, we have the year, month, day, and hour of the incident, so we could probably also use these as factors to identify crime spots. As we can see from Figure 2, there appears to be some potential for further analysis, perhaps by grouping hours into morning, midday, afternoon, evening, and night. Finally, we have the police district where the incident occurred, the reporting area (area code), and the street for the incident alongside the coordinates.

From Table 1 we can observe that the values of latitude and longitude are within reason, as the coordinates of the port of Boston is $(42.364506, -71.038887)$ [3]. However, the minimum value of latitude and the maximum of longitude does not make sense in

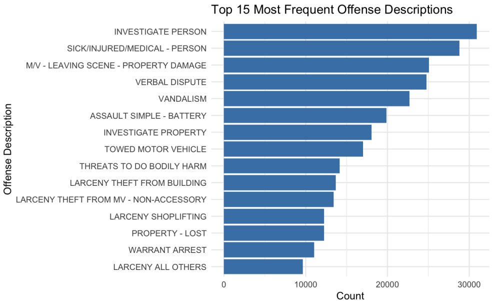

Figure 1: Top 15 crime descriptions in the data.

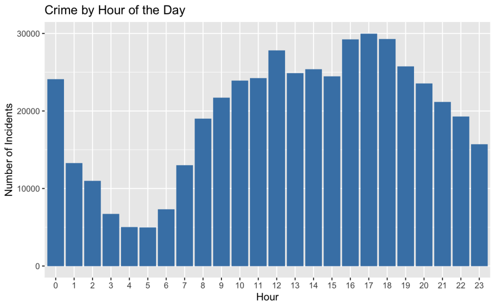

Figure 2: Distribution of crime by the hour.

this case, so we have to do some data cleaning to remove some outliers. Any NAs also have to be dealt with appropriately.

Statistic	NA	Mean	Median	Min	Max
Lat	29349	42.237	42.33	−1.000	42.450
Long	29349	−70.945	-71.08	−71.244	0.000

Table 1: Summary of descriptive statistics for latitude and longitude values in the dataset.

3 Analysis Approach

This report will use longitude and latitude as the key variables for the clustering. Observe from Figure 3 and 4 that the crimes “vandalism” and “threat to do bodily harm” occur at very similar places. This poses a challenge for clustering based purely on coordinates. We would have to incorporate different crimes type as discussed earlier, and perhaps the date and time for the incident as well. We will try to do principal component analysis to reduce the dimensionality, and then try to cluster with the new components that we get from the dimensionality reduction. The main idea is to use DBSCAN for density-based clustering.

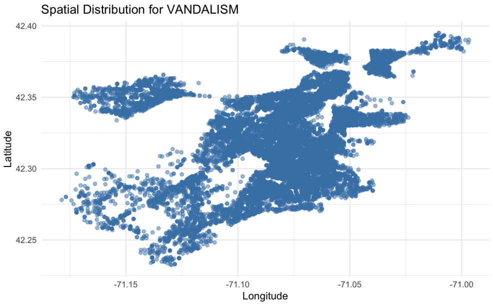

Figure 3: Spatial distribution of the crime VANDALISM.

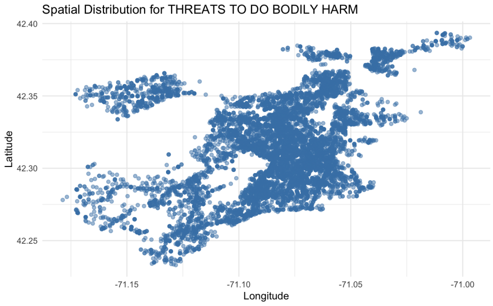

Figure 4: Spatial distribution of the crime THREAT TO DO BODILY HARM.

We note that from Table 2 that there is generally little correlation between the variables. The relationship between latitude and longitude is moderately positive, meaning that as latitude increase, so does longitude. This matches the geography of Boston.

Other than that, the correlations are weak. This suggests that linear PCA might not be very good in this case, so we seek to do other approaches. One approach is the Geographically Weighted PCA, as described in [4]. This model computes local principal components at each defined location, capturing local differences that a linear (and global) PCA might miss. This might work well considering that there might be differences in crime patterns in the city. Other potential methods that could be used is kernel PCA and t-SNE.

	Lat	Long	HOUR	CRIME_TYPE_NUM
Lat	1.00	0.38	-0.03	-0.09
Long	0.38	1.00	-0.01	0.02
HOUR	-0.03	-0.01	1.00	0.02
CRIME_TYPE_NUM	-0.09	0.02	0.02	1.00

Table 2: Correlation matrix for selected crimes (INVESTIGATE PERSON, VANDALISM, VERBAL DISPUTE, THREATS TO DO BODILY HARM) and latitude, longitude, and hour.

References

[1] Emanne Khan. *Crime in Boston: A Look at the Trends Before and After COVID.* Retrieved: 08-03-2025. 2020. URL: https://www.bostonpoliticalreview.org/post/crime-in-boston-a-look-at-the-trends-before-and-after-covid.

[2] Boston Police Department. *CRIME INCIDENT REPORTS (AUGUST 2015 - TO DATE) (SOURCE: NEW SYSTEM).* Retrieved: 08-03-2025. URL: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system.

[3] LatLong. *Port of Boston, MA, USA.* Retrieved: 08-03-2025. URL: https://www.latlong.net/place/port-of-boston-ma-usa-2420.html.

[4] Isabella Gollini et al. *GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models.* Retrieved: 09-03-2025. 2015. DOI: 10.18637/jss.v063.i17. URL: https://www.jstatsoft.org/article/view/v063i17.