# Analyzing the Spatial Distribution of Crimes in Boston

Sebastian Eide Aas

April 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". On the background of this, we formulate the following research aims:

- To analyze the spatial distribution of crime across different areas in Boston.
- To identify whether certain areas exhibit a higher prevalence of specific crime type.
- To determine the presence of distinct crime hotspots using multivariate statistical methods.

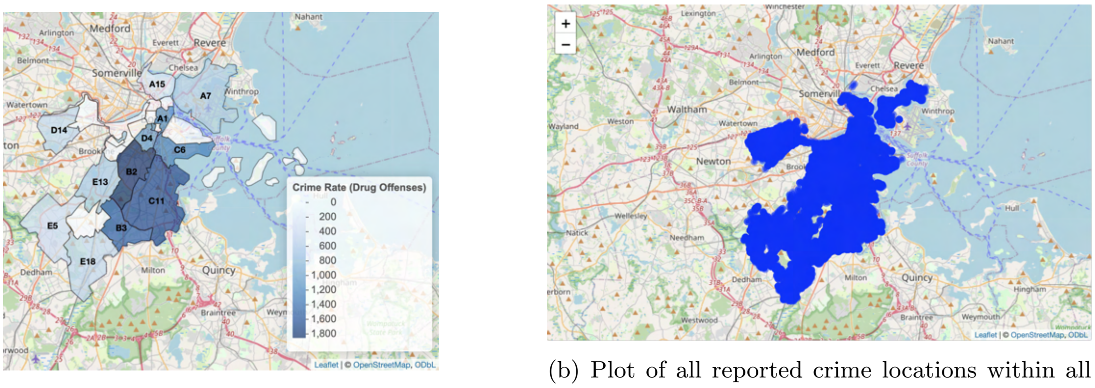
In many large cities, crime is concentrated in specific areas due to, for example, socioeconomic factors. We hypothesize that Boston will exhibit similar patterns, with larceny crimes concentrated in business areas and violent crimes concentrated in residential areas.

## 2 Exploratory Data Analysis

The crime data set was retrieved 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 where it was downloaded [3]. Furthermore, the data is dated from June 2015 to August 2020. We will also use data to define the boundaries of the districts of Boston. This data is retrieved from Boston Maps, where the boundaries are only an approximation as the data itself is not released by The Census Bureau (a bureau providing information about people and economy in America) [4].

The data set provides us with an unique identifier for each incident. Each incident number corresponds to when an officer or a patrol has responded, so there will be duplicates in the data set. These are removed as they represent the same incident multiple times. Furthermore, we have the offense code, offense code group, and offense description. This allows us to create different categories of crime offenses, which we will use to group the data. Table 2 in the appendix shows how we have grouped different offenses. In total, we find 279 unique crime offenses across 419692 incidents of crime. The category "Violent Crimes" is the largest with 43787 registered incidents. Moreover, we have "Larceny" with 41701 incidents, "Drug Offences" with 11508 incidents, and "Burglary" with 10046 incidents. Finally, we have "Sex Offences" with 295 incidents. The rest of the crimes are classified as "Other", as we want to focus on the most dangerous crimes. Each incident is associated to a position measured in latitude and longitude. We see from Figure 1b that the location of the crimes across all categories except "Other", are evenly distributed across the city of Boston.

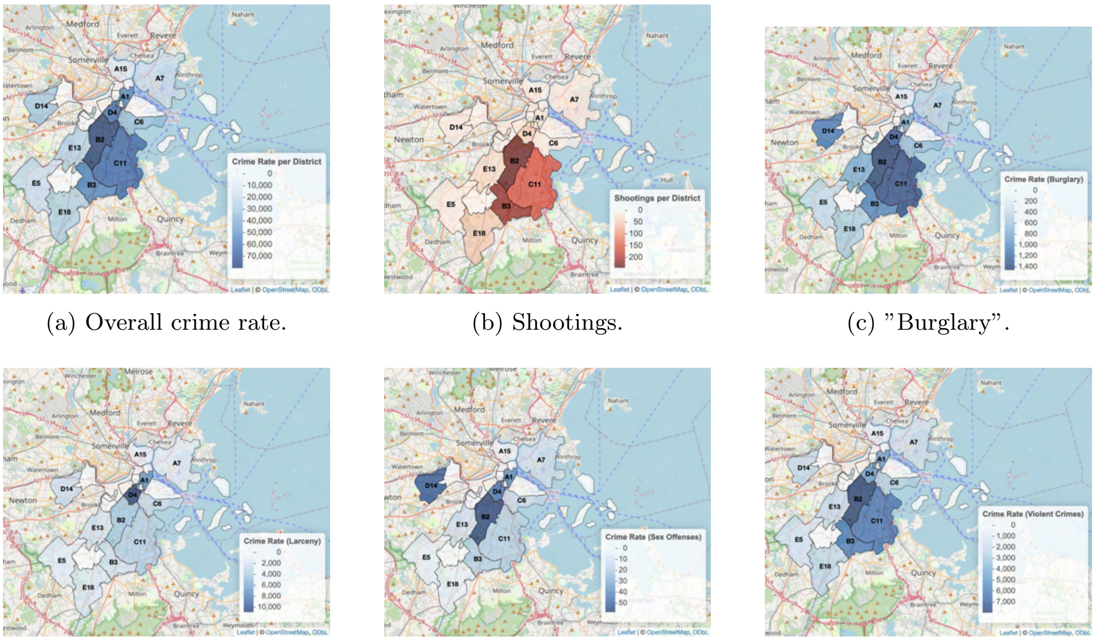
Other features includes the year, month, day, and hour of the incident. We see from Figure 3 that the crimes are about evenly distributed, when separated into morning, midday, afternoon, and evening. Furthermore, we have plotted the crime categories by day of the week, by month, and per year. We observe that the distribution is fairly even, with the exception of categories “Sex Offenses” and “Drug Offenses”. For the distribution of the category “Sex Offenses” per day of the week, there is an increase in the weekend. We also see that there is no data until the year of 2019, with an even higher number in 2020. The fact that there are no sex crimes until 2019 is a weird anomaly in the data, for which we can not find an explanation. There are also fewer incidents in 2015 and 2020, which is explained by the dating of the data, as mentioned earlier. There seem to be no potential in separating the data with respect to time, as the distribution of crime is mostly even. The incidents are therefore aggregated across the entire time period into their respective categories.



(a) ”Drug Offenses”.

(b) Plot of all reported crime locations within all categories except the category ”Other”.

Figure 1: Heatmap displaying the number of the crimes in the category ”Drug Offenses” per district, and a plot of all reported crime locations within all categories except the crime category ”Other”. Districts with no symbol are not part of the data set retrieved from [5].



(a) Overall crime rate.

(b) Shootings.

(c) ”Burglary”.

(d) ”Larceny”.

(e) ”Sex Offenses”.

(f) ”Violent Crimes”.

Figure 2: Heatmaps displaying crime rates by category for each district, the overall crime rate, and the number of shootings per district. Districts with no symbol are not part of the data set retrieved from [5].

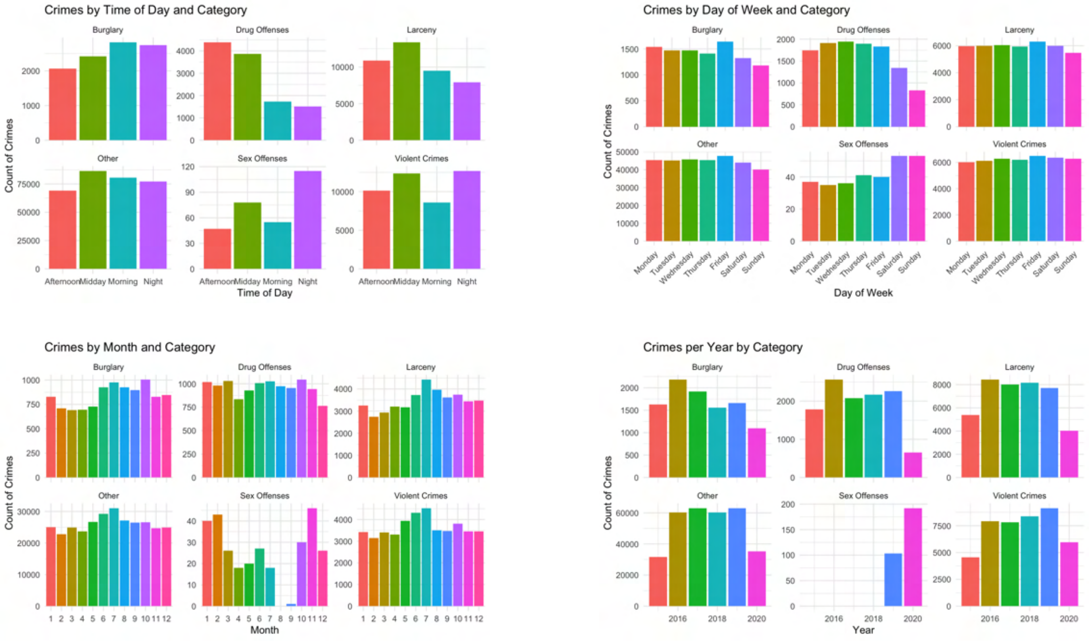


Figure 3: Distribution of the crimes by time of the day separated into morning (5-11), midday (12-16), afternoon (17-20), and night (21-4). The rest is separated by day of the week, month, and year.

Initially, we observe from Figure 1a and 2 that the crime rates for each category are very concentrated in the B2, B3, and C11 districts. Table 1 shows that B2 and C11 are the largest and next-largest districts, respectively, with B3 as the fifth largest district. The size of the districts will influence the number of crimes in each district, as a larger population tends to have a higher potential of crime. Furthermore, we see that number of shootings is the highest in these districts, which might indicate that the most violent crimes tends to occur there. This is confirmed by the plot in Figure 2f, from which we see that the plot matches with the plot of shootings in Figure 2b. For the other crime categories, we see that the amount of "Drug Offenses" and "Burglary" crimes are the largest in the same districts, while "Sex Offenses" has the highest amount of incidents in B2 and D14. "Larceny" crimes have the highest occurrence of incidents in D4.

Table 1: Population, in thousands, for each district [6]. Districts and corresponding neighborhood names were retrieved from BPD [5].

| District | Neighborhood | Population |
|---|---|---:|
| A1 | Downtown | 1.98 |
| A15 | Charlestown | 16.44 |
| A7 | East Boston | 40.51 |
| B2 | Roxbury | 52.53 |
| B3 | Mattapan | 34.39 |
| C6 | South Boston | 33.69 |
| C11 | Dorchester | 60.79 |
| D4 | South End | 30.36 |
| D14 | Brighton | 45.98 |
| E5 | West Roxbury | 30.44 |
| E13 | Jamaica Plain | 41.26 |
| E18 | Hyde Park | 31.85 |

# 3 Methodology

This report will use longitude and latitude as the key variables for the clustering. We will split the data into the different crime categories as grouped earlier, and use HDBSCAN for density-based clustering. Furthermore, we will use the Local Getis–Ord $G_i$ and the Local Moran's $I_i$ statistic for a spatial analysis.

## 3.1 Density-Based Clustering

The initial idea was to use DBSCAN for the clustering. However, the dense nature of the data makes it difficult for DBSCAN to cluster the locations properly. As shown in Figure 4, the crime locations for the category "Violent Crimes" are extremely dense. When tuning the parameters of DBSCAN according to the KNN distance plot, we obtain one very large cluster and several small ones. These clustering results are not very informative, and we therefore chose to use HDBSCAN instead.

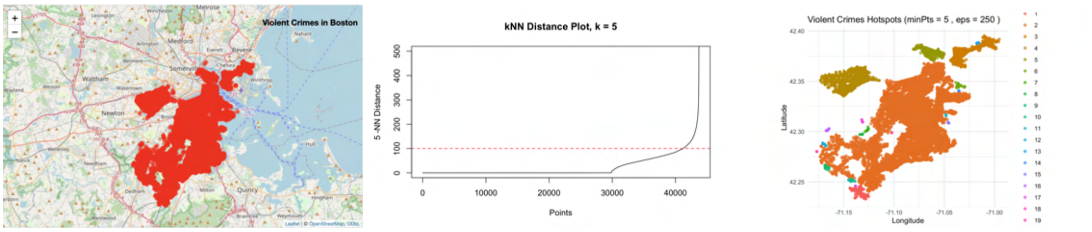


Figure 4: Visualization of the locations of the crimes in the category ”Violent Crimes” in the data set, along with a KNN distance plot used to estimate the minimum distance `eps` for the DBSCAN clustering. The resulting DBSCAN clustering is shown in the right-hand plot.

HDBSCAN is an extension of DBSCAN. The algorithm uses a hierarchical approach which extracts a flat clustering based on the stability of the clusters [7]. The algorithm first constructs a mutual reachability graph, where the edges reflects the mutual reachability distance between the vertices. The algorithm creates a minimal spanning tree and converts it into a hierarchy by sorting the edges of the tree by distance, creating a new cluster for each edge [7]. HDBSCAN extends DBSCAN by creating a condensed tree, a tree which is essentially just a smaller tree with more data attached to each node. The algorithm takes in the parameter minimum cluster size, which is used to mark the most stable clusters. The stable clusters becomes the final output [7]. The reason for choosing this algorithm is that it can identify clusters of different shapes and sizes better than DBSCAN. In addition, we only have to tune the parameter minimum cluster size, which makes the practical implementation easier. However, we have to be aware of the computational complexity that follows with the algorithm [7].

To evaluate the cluster results, we use the Density-Based Clustering Validation Index (DBCV), which is designed to work for arbitrarily shaped clusters [8]. This fits well with the results of HDBSCAN, as we get clusters of varying sizes and densities. This algorithm evaluates clusters based on the relative density connection between pair of objects. In short, we can describe the algorithm as follows. Considering a cluster, it first calculates the core distance between the objects. Using the mutual reachability distances, it builds a minimal spanning tree. Based on all minimal spanning trees built, it calculates an index based on density sparseness and density separation. The index varies between -1 and 1, and clustering with more noise provides a lower index. In general, the higher the index, the better the cluster [9].

For the implementation, we have take into consideration that the algorithm uses the euclidean distance metric in the calculations by default. Since the HDBSCAN algorithm in the `dbscan` library in `R` does not take a distance metric as a parameter, we project our data to a coordinate reference system (CRS). This converts the coordinates into a linear unit so that we are consistent with the units in the calculations. We will use the `R` library `sf` and the CRS 26986 for Massachusetts Mainland [10].

## 3.2 Spatial Methods

The Local Getis–Ord $G_i$ statistic was developed to be used as a measure of spatial association [11]. It can be used to identify characteristics of patterns which global statistics may not capture. The $G_i$ statistic measure the spatial association from weighted points and all other weighted points within a radius of distance $d$ from the original weighted point [11]. For an area subdivided into $n$ regions, with $i = 1, 2, ..., n$, the $G_i$ statistic is defined as

$$G_i(d) := \frac{\sum_{j=1}^{n} w_{ij}(d) x_j}{\sum_{j=1}^{n} x_j}, \quad \text{for } i \neq j.$$

Here, $\{w_{ij}\}$ is a weight matrix between regions $i$ and $j$. Under the assumption that $G_i(d)$ is normally distributed, the significance of the cluster is tested by the $Z$-value:

$$Z_i = \frac{G_i(d) - \mathrm{E}\left(G_i(d)\right)}{\sqrt{\mathrm{var}\left(G_i(d)\right)}}.$$

Following the discussion in [11], we will study the Local Moran's $I_i$ statistic and compare it with the $G_i$ statistic. The reason for this is that while the Local Getis–Ord $G_i$ is based on covariances, the Local Moran's $I_i$ statistic calculates the product of deviations from the mean. Thus, the statistics reflect two different aspects of spatial association, giving a more complete understanding of the spatial dependence structure [11]. The Local Moran's $I_i$ statistic for an observation $i$ is defined as

$$I_i := \frac{z_i}{m_2} \sum_{j=1} w_{ij} z_j,$$

where $z_i$, $z_j$ are the deviations from the mean, $\{w_{ij}\}$ is the row-standardized weight matrix, and $m_2$ is a constant for all locations which follows from the row-standardization [12]. As for the $G_i$ statistic, the significance of the cluster is tested by the $Z$-value:

$$Z_i = \frac{I_i - \mathrm{E}(I_i)}{\sqrt{\mathrm{var}\left(I_i\right)}}.$$

While Local Moran's $I_i$ statistic measures how similar (or dissimilar) an observation is from its neighbor, the Getis-Ord $G_i$ statistic measures the concentration of values within a distance $d$. This is important to keep in mind when interpreting the results.

To compare and evaluate the two statistics, we will make use of five metrics as described by Zhanjun He et al. [13]. The HitRate measures the proportion of crime cases in the hotspot area to the total of cases in the study area. Similarly, AreaRatio measures the percentage of the area the hotspot area is compared to the entire study area. The density contrast ratio (DCR) metric calculates the ratio of the case density in the hotspot to the outside the hotspot. Furthermore, we calculate the standardized shape index (SSI), which captures the geometry of the hotspot compared to a circle [13]. The metrics are defined as

$$\text{HitRate} := \frac{n}{N}, \qquad \text{AreaRatio} := \frac{a}{A}, \qquad \text{DCR} := \frac{n/a}{(N-n)/(A-a)}, \qquad \text{SSI} := 1 - \frac{2\sqrt{\pi a}}{P},$$

where $n$ is the number of cases in the hotspot regions, $N$ is the total number of crimes in the study area, $a$ is the area of the detected hotspots, $A$ is the area of the study region, and $P$ is the perimeter of the hotspot area. Finally, we have the metric PAI which is the HitRate divided by AreaRatio. This measures the ratio of the case density in the hotspot to the entire study area.

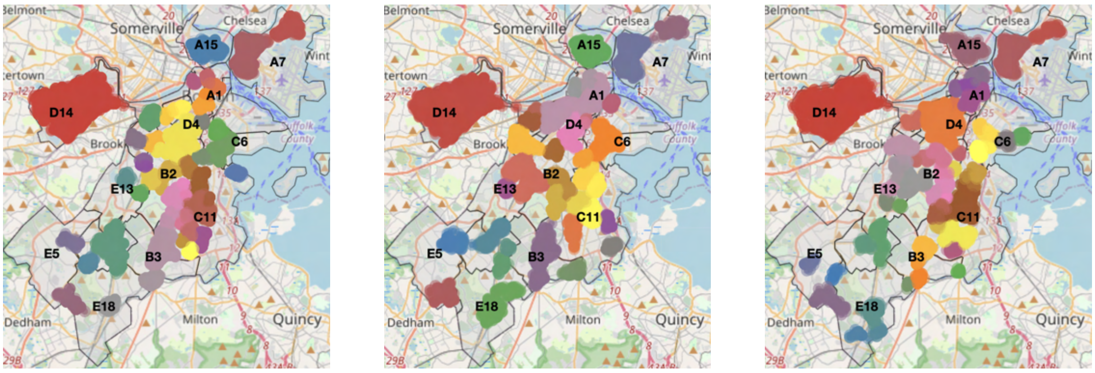
For the implementation, we will divide the area into a grid. The size of each cell will be chosen according to a sensitivity analysis, where we investigate the different values of the metrics with respect to the cell size. We will define the relationships between the grid cells using the queen criterion. This defines neighbors as cells sharing one common vertex [14]. Furthermore, the number of crimes per category will be aggregated and be used as the value for each cell. The significance tests will be carried out with $p = 0.05$.

# 4 Results and Discussion

## 4.1 Results From HDBSCAN

An overview over the detected hotspots by HDBSCAN are presented in Figure 5 and 6. We observe that there is a large overlap in hotspots between the crime categories “Violent”, “Larceny”, and “Burglary”, in Figure 5. We also notice from the plotting in Figure 5 and 6 that there are a lot of crimes that appear in districts which are not marked. Roslindale, a district bordering to B3, E5, E18, and E13, is a neighborhood covered by the police department in E5 Roxbury. As mentioned earlier, we only have the approximate boundaries for the districts of Boston. Thus, crimes outside of districts may show as “outliers” in our data, while they might actually be inside the districts. Furthermore, we observe that district A1 has a larger cluster for the crime categories “Larceny” and “Burglary”, than for the category “Violent Crimes”. The district A1, or Downtown, is a district that has a lot of tourist attractions and corporate headquarters, condos, and apartments [15]. Thus, it makes sense that we have what looks like a hotspot in this area.

There are also large clusters of crimes in A7, A15, and D14, which we can see in Figure 5. A reason for this could be that the data in these areas are dense and uniformly distributed. As a result of this, the clusters remain stable when HDBSCAN “prunes” the hierarchical tree. With `minpts` = 100, the algorithm merges many points into one large cluster. It does not manage to filter out local varieties within the districts, and this result can be interpreted as the areas being hotspots for the crime categories “Larceny” and “Burglary”.



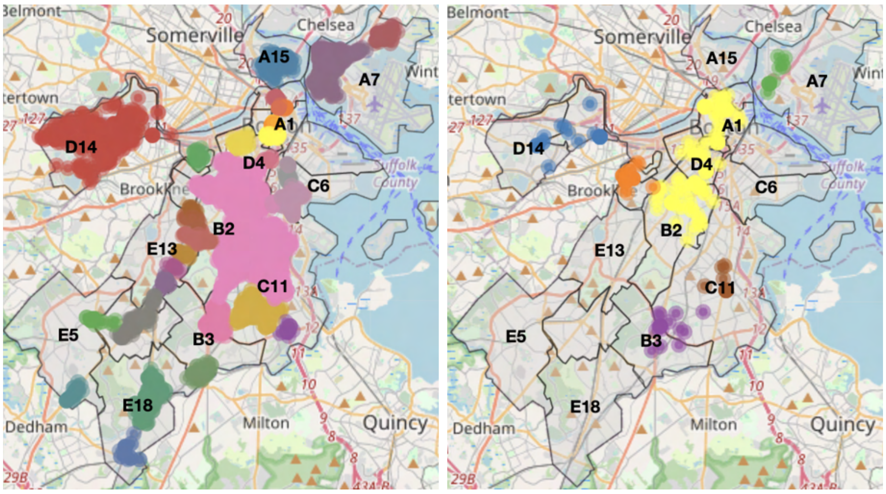
(a) ”Violent Crimes”, $\texttt{minpts} = 100, \texttt{DBCV} = 0.251$.

(b) ”Larceny”, $\texttt{minpts} = 100, \texttt{DBCV} = 0.281$.

(c) ”Burglary”, $\texttt{minpts} = 18, \texttt{DBCV} = 0.265$.

Figure 5: Detected hotspots by HDBSCAN of the crime categories ”Violent Crimes”, ”Larceny”, and ”Burglary”. Due to limitations of the color palette of Leaflet, colors may appear several times, even though they represent distinct clusters.

We can see from Figure 6 that there is a significantly large cluster for “Drug Offenses” between districts B2, B3, C11, and D4. The district D14 again appears with a large cluster, indicating that this district is heavily influenced by crimes in general. For the category “Sex Offenses”, we observe distinct hotspots in all districts except for the cluster within B2, D4, and A1. The number of sex crime incidents that we have in the data set might explain the large, yellow cluster in the middle. The cluster might form because the algorithm uses `minpts` to

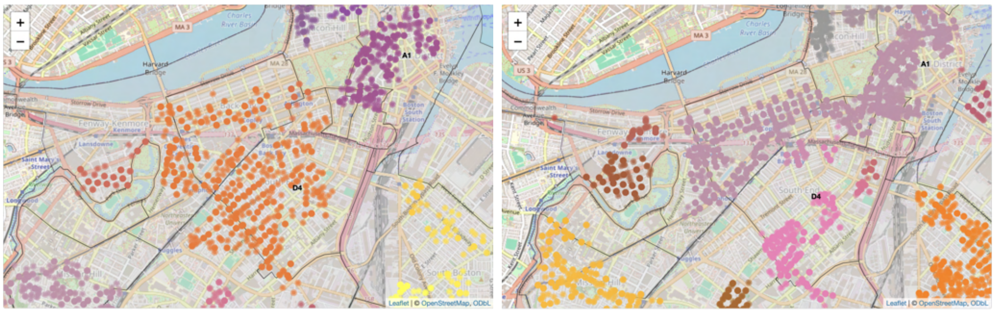


(a) ”Drug Offenses”, $\texttt{minpts} = 19, \texttt{DBCV} = 0.248$.

(b) ”Sex offenses”, $\texttt{minpts} = 11, \texttt{DBCV} = 0.185$.

Figure 6: Detected hotspots by HDBSCAN of the crime categories ”Drug Offenses” and ”Sex Offenses”. Due to limitations of the color palette of Leaflet, colors may appear several times, even though they represent distinct clusters.

“mark” the stable clusters. When data points are sparsely distributed and sufficient in number with respect to `minpts`, it is marked as a large cluster rather than multiple small ones. We also see a distinct overlap in the categories “Larceny” and “Burglary” in the D4 district. We can see in Figure 7 where the clusters are located. Back Bay and South End are two large residential areas, with a lot of shopping stores, restaurants, and vintage homes [16, 17]. This might explain the overlap of these types of crimes in these areas.


+
−
A1
D4
Leaflet | © OpenStreetMap, ODbL
+
−
A1
D4
Leaflet | © OpenStreetMap, ODbL

(a) ”Burglary”, $\texttt{minpts} = 18, \texttt{DBCV} = 0.265$.

(b) ”Larceny”, $\texttt{minpts} = 100, \texttt{DBCV} = 0.281$.

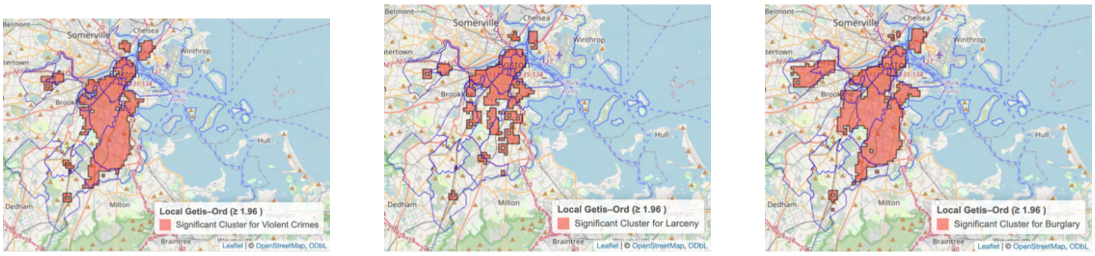
Figure 7: Clustering results for the crime categories ”Burglary” and ”Larceny”, respectively, near Back Bay and South End.

A difficulty with the clustering result of the violent crimes was that we obtained the same value of DBCV for 100 and 160 `minpts`. The only difference was the amount of clusters that was found, where a smaller value of `minpts` gave a larger number of clusters. We decided to use `minpts` = 100, to see if we could highlight more distinct areas. In general, the DBCV scores was not very high. We interpret scores in the area between 0.25 and 0.3 as moderate values. The reason for this is that the score also takes the noise into account, and due to the large amount of data, we will naturally have a lot of data categorized as noise. The noise in this context is data points that are not considered to be a part of a hotspot.

## 4.2 Results From the Spatial Methods

For the spatial methods, we first have to determine a cell size. We calculated the five different metrics for a cell size of $100, 250, 500$, and 1000 meters, for which the results can be viewed in Table 3 and 4 in the appendix. For the results with a cell size of 100 meters, we observe very high PAI values for the category “Sex Offenses”, as well as a very small value for AreaRatio. In general for this cell size, the HitRates was not very large for the crime categories. This is because more cells gives more areas without crimes. It seems as PAI and HitRate somewhat contradicts each other, as we would expect a high HitRate for a large PAI and a small AreaRatio. Thus, we will value the other metrics more for determining the cell size. The metrics for a cell size of 1000 meters seems fine, but visually, the clusters are too large to provide any distinct hotspots. The decision to use a cell size of 250 meters was decided on the basis of the metrics DCR and SSI. This cell size gave high DCR values for all categories. Furthermore, the SSI was higher for this size, indicating an ability to capture complex clusters better. We disregard the metric results for the category “Sex Offenses”, due to the small amount of data.

The results of the Local Getis-Ord $G_i$ statistic are presented in Figure 8 and 9. Visually, we see that the clustering of the category “Violent Crimes” in Figure 8a shares similar features to the one of HDBSCAN in Figure 5a. However, we observe that in district D14, we have more distinct clusters rather than a cluster of the entire district. We get more distinct clusters in districts A7, E18, and Roslindale. We observe that we get no cluster in E5. The same holds for the category “Larceny Crimes” in Figure 8b, compared to Figure 5b. They share mainly the same patterns in the center of B2 and D4, but the spatial approach provides more distinct clusters. For the crime category “Burglary”, however, we note that the HDBSCAN provides more distinct clusters, as we can see in Figure 5c compared to Figure 8c. The perhaps largest difference is visualized when comparing the categories “Drug Offenses” and “Sex Offenses” in Figure 11a and 11b to the ones HDBSCAN produces in Figure 6a and 6b. Notably, there are no large clusters in this case compared to the one with HDBSCAN. We get several distinct hotspots, but this could also be due to the lack of data for the category “Sex Offenses”. We also get some clusters with a void in the center, which is a result that we are not able to explain.


Local Getis-Ord (≥ 1.96)
Significant Cluster for Violent Crimes
Local Getis-Ord (≥ 1.96)
Significant Cluster for Larceny
Local Getis-Ord (≥ 1.96)
Significant Cluster for Burglary

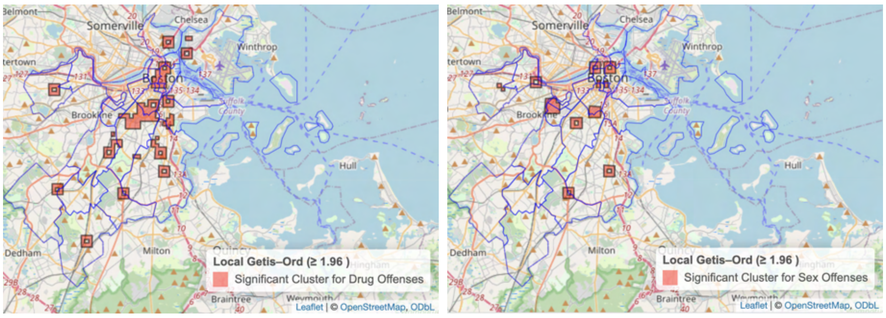
(a) ”Violent Crimes”.

(b) ”Larceny”.

(c) ”Burglary”.

Figure 8: Detected hotspots by the Local Getis-Ord $G_i$ statistic of the crime categories ”Violent Crimes”, ”Larceny”, and ”Burglary”. Every hotspot is a significant cluster, with a $Z$-value above 1.96, corresponding to $p = 0.05$.

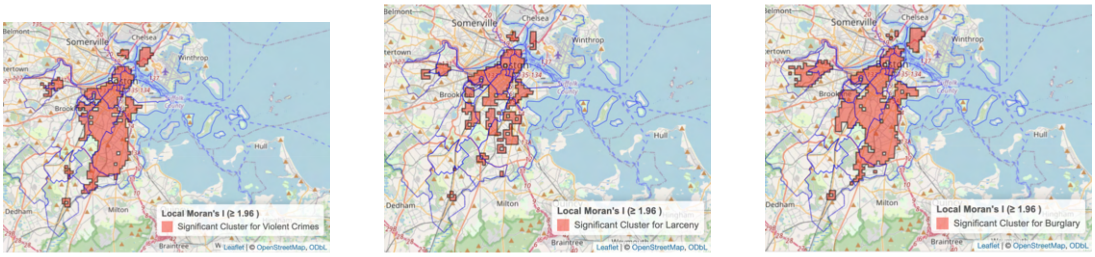
When studying the results of Local Moran’s $I_i$ statistic in Figure 10 and 11, we notice that it is quite similar to the results produced by the Local Getis-Ord $G_i$ statistic, with an exception of the category “Sex Offenses”. We see that we have a smaller amount of clusters in Figure 11b. We note that from Table 3 in the appendix that all crime categories had a high SSI and DCR. For the category “Sex Offenses”, we recall that there are only 295 crime incidents. This makes the metrics hard to interpret, as there are few data points to consider. When considering the metrics from the results using the Local Getis-Ord $G_i$ statistic, we observe from Table 4 in the appendix, that the DCR is the highest for cell size of 250. Furthermore, the HitRate is also high compared to the area ratio, indicating that the hotspots are compact and captures a large proportion of crime in the area. The same observations holds for the category “Sex Offenses”


Local Getis–Ord (≥ 1.96 )
🟥 Significant Cluster for Drug Offenses
Local Getis–Ord (≥ 1.96 )
🟥 Significant Cluster for Sex Offenses
Leaflet | © OpenStreetMap, ODbL

(a) "Drug Offenses". (b) "Sex Offenses".

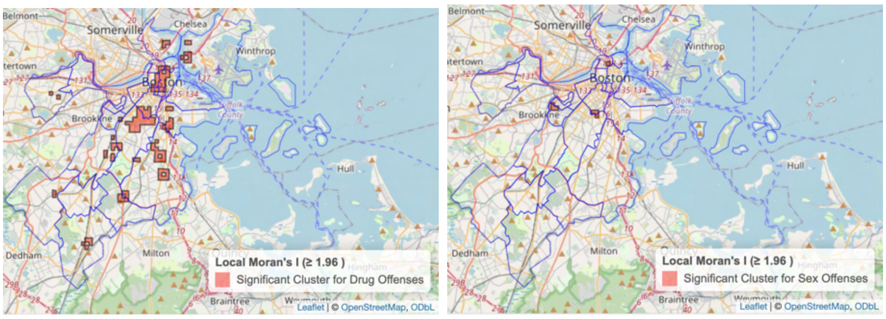
Figure 9: Detected hotspots by the Local Getis-Ord $G_i$ statistic of the crime categories "Drug Offenses" and "Sex Offenses". Every hotspot is a significant cluster, with a $Z$-value above 1.96, corresponding to $p = 0.05$.

for this statistic, as it did for the Local Getis-Ord $G_i$ statistic.


Local Moran's I (≥ 1.96 )
🟥 Significant Cluster for Violent Crimes
Local Moran's I (≥ 1.96 )
🟥 Significant Cluster for Larceny
Local Moran's I (≥ 1.96 )
🟥 Significant Cluster for Burglary
Leaflet | © OpenStreetMap, ODbL

(a) "Violent Crimes". (b) "Larceny". (c) "Burglary".

Figure 10: Detected hotspots by the Local Moran's $I_i$ statistic of the crime categories "Violent Crimes", "Larceny", and "Burglary". Every hotspot is a significant cluster, with a $Z$-value above 1.96, corresponding to $p = 0.05$.


Local Moran's I (≥ 1.96 )
🟥 Significant Cluster for Drug Offenses
Local Moran's I (≥ 1.96 )
🟥 Significant Cluster for Sex Offenses
Leaflet | © OpenStreetMap, ODbL

(a) "Drug Offenses". (b) "Sex Offenses".

Figure 11: Detected hotspots by the Local Moran $I_i$ statistic of the crime categories Violent, Larceny, and Burglary. Every hotspot is a significant cluster, with a $Z$-value above 1.96, corresponding to $p = 0.05$.

## 4.3 General Discussion

We notice that the two different approaches provides similar clusterings of the data, despite being two fundamentally different methods. We observe that there are certain areas which are more prone to the crime categories “Larceny” and “Burglary” than others, for example, the areas Back Bay and South End. Both approaches clustered this area for these crime categories, indicating that this is a hotspot. Furthermore, we observe that we have an overlap of distinct clusters from HDBSCAN and the Local Getis-Ord $G_i$ statistic in

- District E18 for “Violent Crimes”,
- District E13 for “Larceny”,
- District E5 for “Drug Offenses”,
- District C11, the area between districts E13, B3, and D14, and some overlap in district D14 for “Sex Offenses”.

The methods gave somewhat different results in other districts such as D14, A15, and A7. We argue that because HDBSCAN categorized the entire districts as one cluster, while the spatial methods did not, that it is difficult to single out a hotspot in these areas. However, it is clear that these areas are prone to high levels of crime.

Using the methods together, one can identify hotspots that are also statistically significant, providing an extra source of confirmation. The fact that the Local Moran’s $I_i$ and the Local Getis-Ord $G_i$ statistic provided similar results, makes us confident that we were able to capture most of the spatial patterns in the data. It is important to take into account that both methods are in general very sensitive to the input parameters. For the spatial methods, the grid size was the main factor influencing the result. Using the five different metrics helped the tuning process, where we valued the DCR and SSI the most. Tuning HDBSCAN’s parameter `minPts` based on DBCV proved to be efficient, but using other metrics could enhance the tuning process and possibly provide better results.

# 5 Conclusion

In this report, we have studied two different approaches to identify hotspots of crime in the city of Boston. The first approach was HDBSCAN, which gave distinct clusters of crime in districts such as E5 and E18, in addition to large clusters in the city center. The second approach was spatial methods. They provided similar results in the city center, and with some more distinct clusters for the crime categories “Drug Offenses” and “Sex Offenses”. The spatial methods also showed some smaller clusters in districts such as D14, but not the same distinct clusters in districts such as E5.

It is difficult to directly compare the two methods since they are fundamentally different. However, using the results from both methods, we may conclude that there are some areas such as Back Bay and South End that exhibit a higher prevalence of the crime categories “Larceny” and “Burglary”. Furthermore, there exists distinct hotspots in the different districts, but there are not many overlaps between the distinct hotspots from the two different approaches. Thus, it is difficult to conclude that there exists any other distinct hotspots by the use of these methods. The analysis could be improved by including non-spatial features such as demographic statistics. This could provide more insight, and perhaps improve the clustering to provide more distinct hotspots.

# References

[1] Emanne Khan. *Crime in Boston: A Look at the Trends Before and After COVID.* Accessed: 08-03-2025. 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).* Accessed: 08-03-2025. URL: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system.

[3] Sourin Roy. *Boston Crime Dataset (updated July 2020).* Retrieved: 08-03-2025. URL: https://www.kaggle.com/datasets/sourinroy/boston-crime-dataset-updated-july-2020.

[4] Boston Maps. *Boston Neighborhood Boundaries Approximated by 2020 Census Tracts.* Retrieved: 22-03-2025. URL: https://data.boston.gov/dataset/boston-neighborhood-boundaries-approximated-by-2020-census-tracts.

[5] Boston Police Department. *Districts.* Accessed: 21-03-2025. URL: https://police.boston.gov/districts/.

[6] Statistical Atlas. *Population of Boston, Massachusetts.* Accessed: 23-03-2025. Last updated: 17-09-2018. URL: https://statisticalatlas.com/place/Massachusetts/Boston/Population.

[7] Leland McInnes, John Healy, and Steve Astels. *How HDBSCAN Works.* Accessed: 21-03-2025. URL: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

[8] Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. *Density-Based Clustering Validation.* Accessed: 23-03-2025. URL: https://www.dbs.ifi.lmu.de/~zimek/publications/SDM2014/DBCV.pdf.

[9] Rdocumentation. *dbcv: Density-Based Clustering Validation Index (DBCV).* Accessed: 23-03-2025. URL: https://www.rdocumentation.org/packages/dbscan/versions/1.2.2/topics/dbcv.

[10] ArcGIS. *ArcGIS Pro 3.1 and ArcGIS Enterprise 11.1 Projected Coordinate System Tables.* Accessed: 23-03-2025. Page 140. URL: https://pro.arcgis.com/en/pro-app/3.1/help/mapping/properties/pdf/projected_coordinate_systems.pdf.

[11] Luc Anselin et al. *Perspectives on Spatial Data Analysis.* Accessed: 29-03-2025, Page 127-145. Berlin, Heidelberg: Springer, 2010. ISBN: 978-3-642-01976-0. URL: https://link.springer.com/book/10.1007/978-3-642-01976-0.

[12] Luc Anselin. "Local Indicators of Spatial Association—LISA". In: *Geographical Analysis* 27.2 (Apr. 1995), pp. 93–115. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1995.tb00338.x.

[13] Zhanjun He et al. "Comparative Study of Approaches for Detecting Crime Hotspots with Considering Concentration and Shape Characteristics". In: *International Journal of Environmental Research and Public Health* 19.21 (2022). Accessed: 29-03-2025. DOI: 10.3390/ijerph192114350. URL: https://doi.org/10.3390/ijerph192114350.

[14] Luc Anselin. *Contiguity-Based Spatial Weights.* Accessed: 06-04-2025. URL: https://geodacenter.github.io/workbook/4a_contig_weights/lab4a.html.

[15] Boston Government. *DOWNTOWN.* Accessed: 24-03-2025. URL: https://www.boston.gov/neighborhood/downtown.

[16] Boston Government. *BACK BAY.* Accessed: 24-03-2025. URL: https://www.boston.gov/neighborhood/back-bay.

[17] Boston Government. *SOUTH END.* Accessed: 24-03-2025. URL: https://www.boston.gov/neighborhood/south-end.

# A Additional Tables

Table 2: Crime offenses separated into categories.

| **Crime category** | **Conditions** | **Incidents** |
|---|---|---|
| **Violent Crimes** | `shooting == "1"` or one of:<br>assault simple - battery, threats to do bodily harm, assault - simple, assault - aggravated, assault - aggravated - battery, murder, non-negligient manslaughter, assault & battery, assault d/w - other, assault d/w - knife on police officer, kidnapping/custodial kidnapping, manslaughter - vehicle - negligence, manslaughter - non-vehicle - negligence, kidnapping - enticing or attempted. | 43,787 |
| **Burglary** | burglary - commerical - no force, burglary - commerical - force, burglary - residential, burglary - other - force, burglary - other - attempt, burglary - other - no force, burglary - commerical - attempt, burglary - residential - attempt, burglary - residential - no force, burglary - residential - force, breaking and entering (b&e) motor vehicle, home invasion, robbery - bank, robbery - car jacking, robbery attempt - knife - bank, robbery - unarmed - street, robbery. | 10,046 |
| **Larceny** | larceny theft from building, larceny theft of mv parts & accessories, larceny theft from mv - non-accessory, larceny shoplifting, larceny theft of bicycle, larceny pick-pocket, larceny purse snatch - no force, larceny non-accessory from veh. $50 to $199, larceny in a building $50 to $199, larceny shoplifting under $50, larceny in a building under $50, larceny theft from coin-op machine, larceny in a building $200 & over, larceny shoplifting $200 & over, larceny bicycle $200 & over, larceny other $200 & over. | 41,701 |
| **Drug Offenses** | drugs - other, drugs - sick assist - heroin, drugs - poss class a - heroin, etc., drugs - poss class a - intent to mfr dist disp, drugs - sale / manufacturing, drugs - poss class d - intent to mfr dist disp, drugs - poss class b - intent to mfr dist disp, drugs - sick assist - other narcotic, drugs - class b trafficking over 18 grams, drugs - class a trafficking over 18 grams, drugs - sick assist - other harmful drug, drugs - poss class e, drugs - poss class d, drugs - poss class c, drugs - poss class c - intent to mfr dist disp, drugs - consp to viol controlled substance, drugs - possession of drug paraphanalia, drugs - class d trafficking over 50 grams, drugs - poss class e - intent to mfr dist disp, drugs - possession/ sale/ manufacturing/ use, drugs - glue inhalation, drugs - poss class e intent to mf dist disp, drugs - possession. | 11,508 |
| **Sex Offenses** | sex offense - rape - forcible, sex offense - rape - other, sex offense - rape - sodomy, fondling - indecent assault, sexual assault kit collected. | 295 |
| **Other** | all other offenses | 313,424 |

Table 3: The values of each metric discussed for the spatial analysis, sorted by cell size. Each metric is calculated with the results from the Local Moran's $I_i$ statistic.

| **Cell Size (m)** | **Category** | **HitRate** | **AreaRatio** | **PAI** | **DCR** | **SSI** |
|---|---|---|---|---|---|---|
| 100 | Violent Crimes | 0.6554 | 0.1330 | 4.9271 | 12.3959 | 0.9346 |
| 100 | Larceny | 0.4461 | 0.0529 | 8.4305 | 14.4149 | 0.9167 |
| 100 | Burglary | 0.5389 | 0.0837 | 6.4383 | 12.7952 | 0.9526 |
| 100 | Drug Offenses | 0.2091 | 0.0090 | 23.1969 | 29.0666 | 0.8370 |
| 100 | Sex Offenses | 0.0068 | 0.0001 | 98.4251 | 99.0947 | 0.1138 |
| 250 | Violent Crimes | 0.7723 | 0.2484 | 3.1096 | 10.2652 | 0.7868 |
| 250 | Larceny | 0.6372 | 0.1582 | 4.0273 | 9.3432 | 0.8395 |
| 250 | Burglary | 0.7728 | 0.2731 | 2.8299 | 9.0523 | 0.7959 |
| 250 | Drug Offenses | 0.2927 | 0.0533 | 5.4908 | 7.3496 | 0.8334 |
| 250 | Sex Offenses | 0.1365 | 0.0030 | 44.9943 | 51.9500 | 0.5311 |
| 500 | Violent Crimes | 0.7457 | 0.2791 | 2.6715 | 7.5740 | 0.5757 |
| 500 | Larceny | 0.6447 | 0.2133 | 3.0233 | 6.6954 | 0.6291 |
| 500 | Burglary | 0.7810 | 0.3312 | 2.3585 | 7.2037 | 0.6550 |
| 500 | Drug Offenses | 0.4733 | 0.1526 | 3.1020 | 4.9907 | 0.7318 |
| 500 | Sex Offenses | 0.2116 | 0.0347 | 6.1024 | 7.4718 | 0.6697 |
| 1000 | Violent Crimes | 0.7484 | 0.3043 | 2.4596 | 6.8015 | 0.4782 |
| 1000 | Larceny | 0.6709 | 0.2263 | 2.9654 | 6.9727 | 0.4601 |
| 1000 | Burglary | 0.7846 | 0.3784 | 2.0735 | 5.9838 | 0.5742 |
| 1000 | Drug Offenses | 0.5601 | 0.2497 | 2.1244 | 3.3942 | 0.5746 |
| 1000 | Sex Offenses | 0.4881 | 0.1456 | 3.3511 | 5.5926 | 0.5488 |

Table 4: The values of each metric discussed for the spatial analysis, sorted by cell size. Each metric is calculated with the results from the Local Getis-Ord $G_i$ statistic.

| **Cell Size (m)** | **Category** | **HitRate** | **AreaRatio** | **PAI** | **DCR** | **SSI** |
|---|---|---|---|---|---|---|
| 100 | Violent Crimes | 0.4954 | 0.0931 | 5.3234 | 9.5688 | 0.9324 |
| 100 | Larceny | 0.3208 | 0.0314 | 10.2120 | 14.5634 | 0.9064 |
| 100 | Burglary | 0.3638 | 0.0542 | 6.7076 | 9.9710 | 0.9409 |
| 100 | Drug Offenses | 0.1437 | 0.0062 | 23.2761 | 27.0133 | 0.8100 |
| 100 | Sex Offenses | 0.0068 | 0.0001 | 98.4251 | 99.0947 | 0.1138 |
| 250 | Violent Crimes | 0.6740 | 0.1981 | 3.4025 | 8.3693 | 0.7758 |
| 250 | Larceny | 0.4674 | 0.0832 | 5.6166 | 9.6685 | 0.7864 |
| 250 | Burglary | 0.6567 | 0.2163 | 3.0364 | 6.9325 | 0.7894 |
| 250 | Drug Offenses | 0.2148 | 0.0399 | 5.3866 | 6.5866 | 0.8132 |
| 250 | Sex Offenses | 0.0853 | 0.0026 | 32.8084 | 35.7756 | 0.5658 |
| 500 | Violent Crimes | 0.6460 | 0.2185 | 2.9573 | 6.5296 | 0.5478 |
| 500 | Larceny | 0.4604 | 0.0988 | 4.6586 | 7.7800 | 0.5818 |
| 500 | Burglary | 0.6788 | 0.2687 | 2.5261 | 5.7519 | 0.6382 |
| 500 | Drug Offenses | 0.3599 | 0.0971 | 3.7064 | 5.2279 | 0.7293 |
| 500 | Sex Offenses | 0.0990 | 0.0191 | 5.1897 | 5.6499 | 0.6542 |
| 1000 | Violent Crimes | 0.6957 | 0.2843 | 2.4467 | 5.7540 | 0.4027 |
| 1000 | Larceny | 0.5609 | 0.1664 | 3.3700 | 6.3975 | 0.3321 |
| 1000 | Burglary | 0.6106 | 0.2843 | 2.1475 | 3.9472 | 0.4596 |
| 1000 | Drug Offenses | 0.3785 | 0.1734 | 2.1832 | 2.9038 | 0.5569 |
| 1000 | Sex Offenses | 0.4300 | 0.1179 | 3.6475 | 5.6450 | 0.4780 |