JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

jiit

विद्या तत्व ज्योतिसमः

Analyzing Crime Data Patterns using Self-Organizing Maps

Team:

Hritik Aggarwal - 17104024
Naman Vashistha - 17104063
Sarthak Agarwal - 17104027

Certificate

This is to certify that Hritik Aggarwal, Sarthak agarwal and Naman Vashistha, students of Information Technology Engineering, have successfully completed the project on the topic “Analyze Crime Data Pattern with S.O.M.” under the guidance of our supervisor during the year 2019.

This project is absolutely genuine and does not indulge in plagiarism of any kind. The reference taken in making this project have been declared at the end of the report.

Signature
(Supervisor)

Signature

(External Examiner)

Acknowledgement

It gives us immense pleasure to express our deepest sense of gratitude and sincere thanks to our highest respected and esteemed guide of our supervisor for his valuable guidance, encouragement and help for completing this work. His useful suggestions for this whole work and co-operative behaviour are sincerely acknowledged. We would like to express our sincere thanks to her for giving us this opportunity to undertake this project. We would also like to express our indebtedness to our parents as well as our family members whose blessings and support always helped us to face the challenges ahead.

Place: Sector 62, Noida
Date: 26th November 2019

Candidate’s Declaration

We hereby declare that the work presented in this report entitled “Analyze Crime Data Pattern with S.O.M.”, in fulfilment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, submitted in Computer Science and Engineering Department, affiliated to Jaypee Institute of Information Technology, Sector 62 Noida is an authentic record of my own work carried out during my degree. The work reported in this has not been by us for award of any other degree or diploma.

Place: Sector 62, Noida
Date: 26th November 2019

Introduction

In recent years, volumes of crime had brought serious problems to many countries in the world. For example, in the united states, the crime volumes have increased more than 71% in a decade, which may not only bring physical harms but also serious mental injuries for the victims. In order to avoid the emergence of the calamity, the National Police Agency, US must invest more In recent years, volumes of crime had brought serious problems to many countries in the world. For example, in Taiwan, the crime volumes have increased more than 71% in a decade, which may not only bring physical harms but also serious mental injuries for the victims. In order to avoid the emergence of the calamity, the National Police Agency, We must invest more human resources in criminal investigation and increase the law enforcement duties, such as patrolling, raids and guarding for maintaining public order, preventing all kinds of hazards and promoting the welfare of citizens. Traditional law enforcement strategies for crime prevention focus on preventive police patrol which is the most general duty for the police. Since 2002, the Ministry of Internal Affairs in the US had been working on a project to examine the public security index (PSI) with red, yellow, purple, green, blue five lights that represent five kinds of status of public security, which are “very” bad”, ‘‘bad”, ‘‘intermediate”, ‘‘good”, and ‘‘very good,” in order to strengthen the degree of people’s impression of public security. These lights are linguistic in nature in order to more effectively attract people’s attentions to the public security. Indeed, the formal enforcement of preventive laws is seen as an important means of preventing crime and ensuring public safety in modern societies. In addition, domain experts believed that criminal history model can be used to identify other analogous pattern (Kaza, Wang, & Chen, 2007). Therefore, more research focused on using various intelligent approaches to analyze different types of crime characteristics and proceeded specific programs.

Consequently, it is imperative to develop novel approaches for handling such types of data. In this paper, we propose a framework of intelligent decision support model in order to identify crime trend patterns for different criminal activities, conduct temporal rule extraction to uncover their shift around effect, and provide a reference for experts when analyzing the different types of crime. Furthermore we analyse the clusters of crime pattern on the basis of volume of its density which could help law enforcement,policy makers,government agencies.

Problem Statement

In the recent era of increasing volume crimes, crime prevention is now one of the most important global issues, along with the great concern of strengthening public security. Government and community officials are making an all-out effort to improve the effectiveness of crime prevention. Numerous investigations addressing this problem have generally employed disciplines of behavior science and statistics.

Recently, the data mining approach has been shown to be a proactive decision-support tool in predicting and preventing crime. However its effectiveness is often limited due to different nature of crime data, such as linguistic crime data evolving over time. In this paper, we propose a framework of decision-support model based on a self-organizing map (SOM) network to analyze crime data using SOM patterns from temporal crime activity data. In addition,we tried to visualise our crime cluster by SOM on open street map and heat map. In contrast to most present crime related studies,As a case study we analyze two states of the United States of America i.e. Chicago and Washington D.C. . The resultant model can support police managers in assessing more appropriate law enforcement strategies, as well as improving the use of police duty deployment for crime prevention.

Literature

Because crime has drawn much attention over time, the public has become increasingly concerned about the government's response to it. The International Journal of Forecasting published a special issue of crime forecasting in 2003 (Gorr & Harries, 2003). Numerous modern researches of crime certainly demonstrate that the sharply rising crime volume has blossomed into a high-priority problem that needs to be promptly solved. In order to prevent criminal acts and discover crime trend, researchers have developed a number of methods to support law enforcement activities over the last two decades. These works originate from disciplines of behavior and psychology, statistics, and artificial intelligence.

The approach of behavior science and psychology has been aimed at preventing individuals from committing crimes since the 1960s (Visher & Weisburd, 1998). This traditional approach relies on the expertise and tacit knowledge of specialists, which easily leads to criminal experts' fatigue, misjudgment, and slow response. Moreover, it usually lacks the real data for verification, thus making it difficult to apply the findings to the actual policy (Visher & Weisburd, 1998). The second approach deals with the problem of predicting the crime volume at a specific time and place using various statistical models (Brown & Oxford, 2001; Gorr, Olligschlaeger, & Thompson, 2003; Greenberg, 2001; Harries, 2003; Osgood, 2000; Palocsay, Wang, & Brookshire, 2000; Ratcliffe,2005). Gorr et al. (2003) utilized the Naïve exponential model to forecast crime one month ahead in Pittsburgh, US. The result displays that practically any model-based forecasting approach is vastly more accurate than current police practices. Brown and Ox-ford (2001) presented baseline models, normal regression, and lognormal regression to predict the number of breaking and enterings (B&Es) in the city of Richmond, Virginia and found that log-normal regression is the best model. Unfortunately, Palocsay et al. (2000)pointed out that the performance of these statistical models has been considered weak due to their high error rates and limited explanatory power. Researchers have thus continued to look for new improved methods in this arena. pointed out that the performance of these statistical models has been considered weak due to their high error rates and limited explanatory power. Researchers have thus continued to look for new improved methods in this arena.

Methodology

Methods Used

1. The Elbow Method

The **Elbow method** is a heuristic method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". We now define the following:-

1. **Distortion:** It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.
2. **Inertia:** It is the sum of squared distances of samples to their closest cluster center.

To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion.

References:

https://en.wikipedia.org/wiki/Elbow_method_(clustering)

https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

2. S.O.M. (Self Organising Map)

Self Organizing Map(SOM) by providing a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a map. SOM also represents clustering concept by grouping similar data together. Therefore it can be said that SOM **reduces data dimensions** and **displays similarities among data**.

With SOM, clustering is performed by having several units compete for the current object. Once the data has been entered into the system, the network of artificial neurons is trained by providing information about inputs. The weight vector of the unit is closest to the current object becomes the winning or active unit. During the training stage, the values for the input variables are gradually adjusted in an attempt to preserve neighborhood relationships that exist within the input data set. As it gets closer to the input object, the weights of the winning unit are adjusted as well as its neighbors.

Reducing Data Dimensions

Unlike other learning technique in neural networks, training a SOM requires no target vector. A SOM learns to classify the training data without any external supervision.

Data Similarity

Getting the Best Matching Unit is done by running through all wright vectors and calculating the distance from each weight to the sample vector. The weight with the shortest distance is the winner. There are numerous ways to determine the distance, however, the most commonly used method is the Euclidean Distance.

SOM Algorithm

Each data from data set recognizes themselves by competing for representation. SOM mapping steps starts from initializing the weight vectors. From there a sample vector is selected randomly and the map of weight vectors is searched to find which weight best represents that sample. Each weight vector has neighboring weights that are close to it. The weight that is chosen is rewarded by being able to become more like that randomly selected sample vector.

The neighbors of that weight are also rewarded by being able to become more like the chosen sample vector. From this step the number of neighbors and how much each weight can learn decreases over time. This whole process is repeated a large number of times, usually more than 1000 times.

In sum, learning occurs in several steps and over many iterations. :

1. Each node's weights are initialized.
2. A vector is chosen at random from the set of training data.
3. Every node is examined to calculate which one's weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).
4. Then the neighbourhood of the BMU is calculated. The amount of neighbors decreases over time.
5. The winning weight is rewarded with becoming more like the sample vector. The neighbors also become more like the sample vector. The closer a node is to the BMU, the more its weights get altered and the farther away the neighbor is from the BMU, the less it learns.
6. Repeat step 2 for N iterations.

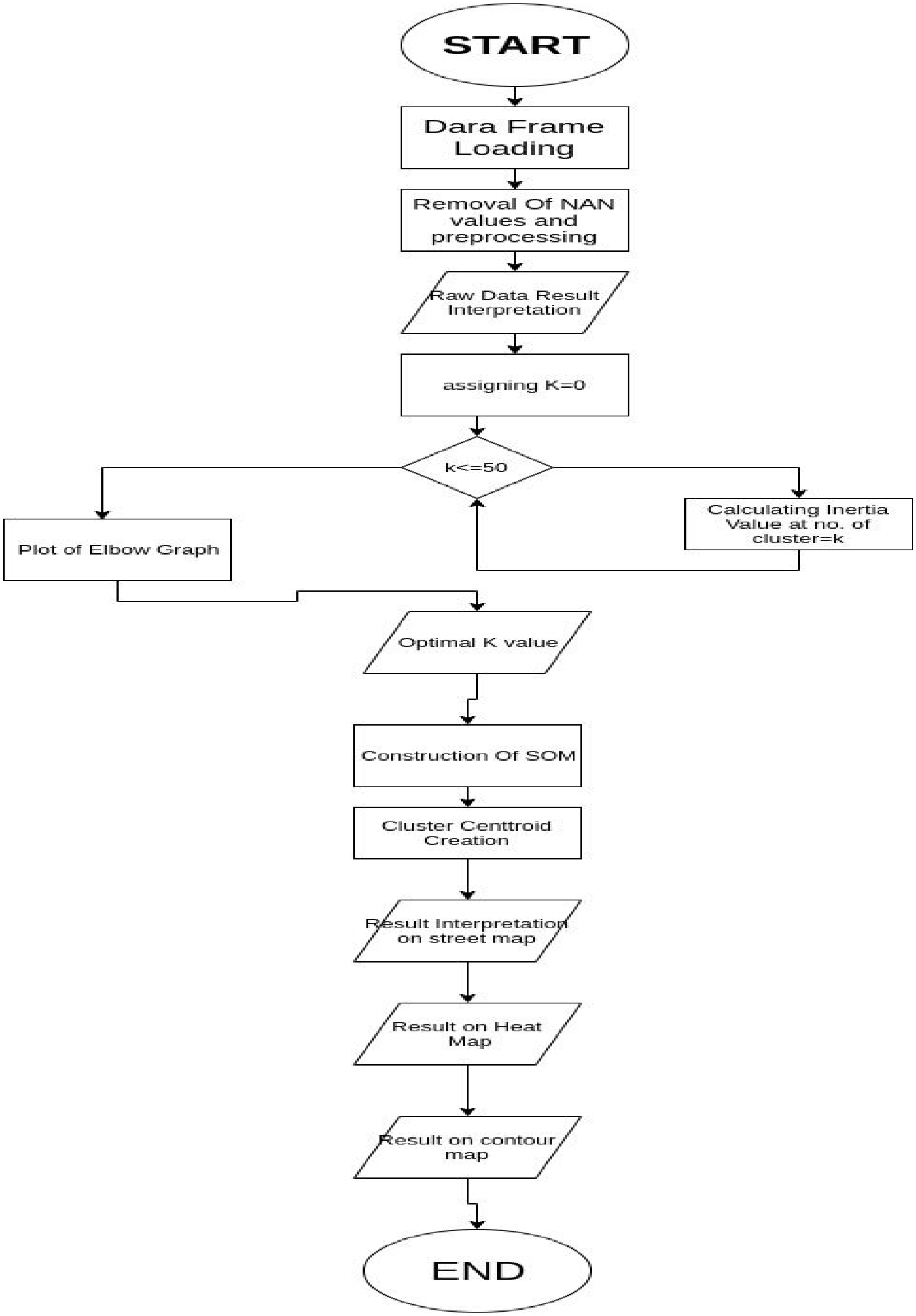
Result Interpretation

If the average distance is high, then the surrounding weights are very different and a dark color is assigned to the location of the weight. If the average distance is low, a lighter color is assigned. The resulting map shows that black is not similar to the white parts because there are lines of black representing no similarity between white parts. Looking at the map it clearly represents that the two not very similar by having black in between. It can be said that the white parts represent different clusters and the black lines represent the division of the clusters.

References

https://en.wikipedia.org/wiki/Self-organizing_map

Design Flow-Chart


START
Dara Frame Loading
Removal Of NAN values and preprocessing
Raw Data Result Interpretation
assigning K=0
k<=50
Calculating Inertia Value at no. of cluster=k
Plot of Elbow Graph
Optimal K value
Construction Of SOM
Cluster Centtroid Creation
Result Interpretation on street map
Result on Heat Map
Result on contour map
END

Framework

1. Data Collection

The first dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. Should you have questions about this dataset, you may contact the Research & Development Division of the Chicago Police Department at PSITAdministration@ChicagoPolice.org. Disclaime.http://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

The second dataset reflects DC Index Crime incident locations for 2016. The dataset contains a subset of locations and attributes of incidents reported in the ASAP (Analytical Services Application) crime report database by the District of Columbia Metropolitan Police Department (MPD). Please visit http://crimemap.dc.gov for more information.

These data are shared via an automated process where addresses are batch matched (geocoded) to the District's Master Address Repository. Users may find that some data points will contain 0,0 for X,Y coordinates resulting in inconsistent spatial locations. Addresses for these data points could not be automatically geocoded and will need to be manually geocoded to 'best fit' locations in DC. Use the MAR Geocoder to help complete this.

Data retrieved from DC Data Catalog (http://data.dc.gov/)

2. Data Frame Creation

Data frame is created by using pandas library in python and retrieving each and every row of the dataset .csv file and loading it into the data frame variable to make the data-frame variable ready for preprocessing.

3. Data Pre-Processing

The used datasets have NaN (Not a Number) values at certain rows ,columns and different cells which cause failure of algorithms and throws error.

These NaN values has been omitted from the dataset to ensure proper working and functioning of the proposed framework.

The only required attributed in the processing of framework has been separated to avoid extra time consumption in processing of unnecessary data.

4. Initial Data Analysis

The data-frame now contains the required dataset which has been modeled graphically to give an overview how the data looking without the application of the used algorithms and clustering model.

The visualized data is unclustered and hence is much scattered.

5. Optimal Number Of Clusters

The method to be used is based on unsupervised learning model hence we are required to find the optimal number of clusters that will be formed on the used datasets.

To achieve this target the algorithm used is **elbow method** we looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion

6. Cluster Formation

Cluster refers to the data with similar traits hence the data passed through the S.O.M. algorithm and unsupervised learning hence forming clusters of data with maximum similarity and minimum differences and the two clusters have maximum differences and minimum similarity.

Providing different colour to different cluster help in better understanding and visualization.

7. Locating Cluster Centroid

The centroid is the point which has an approximately equal distance from all extremes.

The centroid is found at the point where the within sum of squares is minimum.

8. Final Visualization

8.1. Open Street Map

This shows a rendering of the map that we like to call a "slippy map". This is an online map, that can interactively be zoomed and paned. The slippy map on the site is nothing more than a display of map tiles, static images rendered from OpenStreetMap data as an example of what you can do with it. The real power of OpenStreetMap is not this default rendering, but the possibility to actually access the data behind this map rendering. And this is where it differs from Google Maps, Bing Maps and Yahoo! Maps and various others. OpenStreetMap

This shows a rendering of the map that we like to call a "slippy map". This is an online map, that can interactively be zoomed and paned. The slippy map on the site is nothing more than a display of map tiles, static images rendered from OpenStreetMap data as an example of what you can do with it. The real power of OpenStreetMap is not this default rendering, but the possibility to actually access the data behind this map rendering. And this is where it differs from Google maps, Bing Maps and Yahoo! Maps and various others. OpenStreetMap

is the only mapping service that allows you to do something more than just look at pretty map tiles. OpenStreetMap is a database project, with as its main purpose to have an exhaustive database of every street, city, road, building etc on the planet, and not a map display project.is the only mapping service that allows you to do something more than just look at pretty map tiles. OpenStreetMap is a database project, with as its main purpose to have an exhaustive database of every street, city, road, building etc on the planet, and not a map display project.

8.2. Heat Map

A heatmap is a graphical representation of data that uses a system of color-coding to represent different values.

Heatmaps are also a lot more visual than standard analytics reports, which can make them easier to analyse at a glance. This makes them more accessible, particularly to people who are not accustomed to analysing large amounts of data.

heatmaps need to have a large amount of data before they can be accurately analysed. Analysing heat-maps based on a small amount of data is similar to ‘calling’ A/B tests too early, based on too few visits or conversions. As heat-maps show trends, it is important to have enough information to ensure that any anomalies do not affect the overall heat-map picture.

8.3. Contour Map

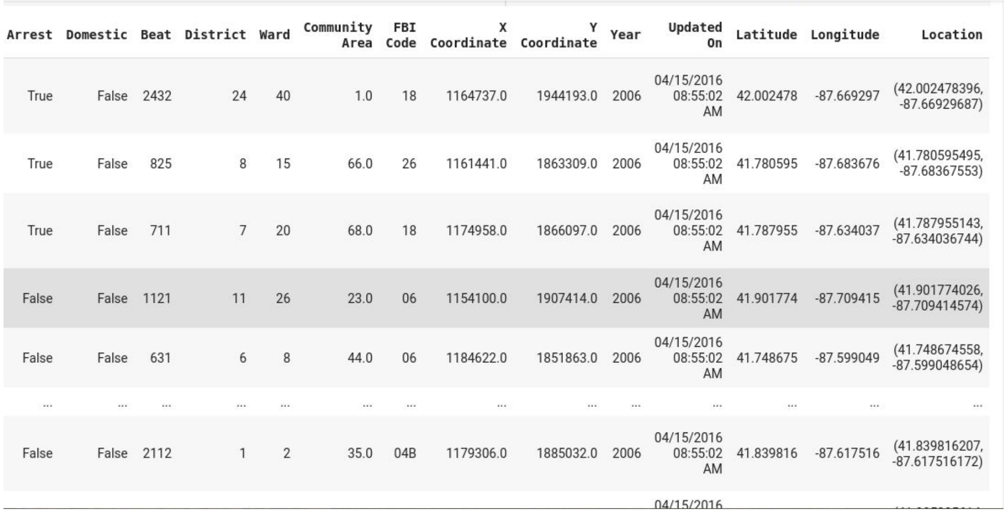
Contour mapping, the delineation of any property in map form by constructing lines of equal values of that property from available data points. A topographic map, for example, reveals the relief of an area by means of contour lines that represent elevation values; each such line passes through points of the same elevation. The method is not wholly objective because two investigators may produce somewhat different maps whenever interpolation between data points is necessary for construction of the contours. In addition to topography, there are scores of geophysical, geochemical, meteorological, sociological, and other variables that are mapped routinely by the method. The availability of plotting devices in recent years has permitted mapping by computer, which reduces the effect of human bias on the final product.

Results and Discussions

1. Data Frame

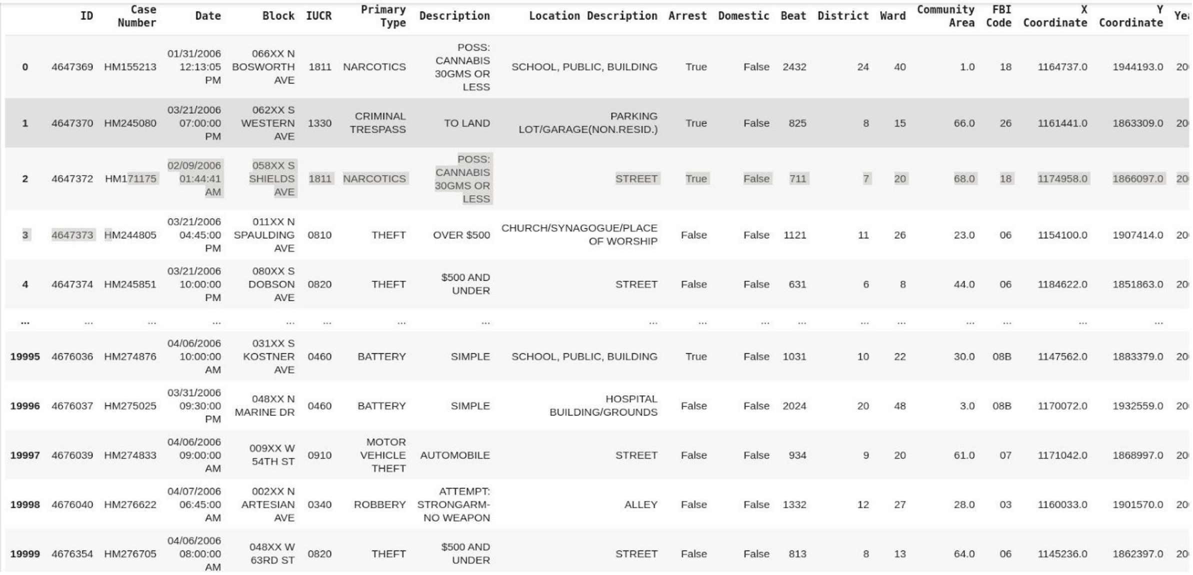
The desired data frame after preprocessing for the chosen datasets are shown below,

a. Chicago Data Frame



Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location
True	False	2432	24	40	1.0	18	1164737.0	1944193.0	2006	04/15/2016 08:55:02 AM	42.002478	-87.669297	(42.002478396, -87.66929687)
True	False	825	8	15	66.0	26	1161441.0	1863309.0	2006	04/15/2016 08:55:02 AM	41.780595	-87.683676	(41.780595495, -87.68367553)
True	False	711	7	20	68.0	18	1174958.0	1866097.0	2006	04/15/2016 08:55:02 AM	41.787955	-87.634037	(41.787955143, -87.634036744)
False	False	1121	11	26	23.0	06	1154100.0	1907414.0	2006	04/15/2016 08:55:02 AM	41.901774	-87.709415	(41.901774026, -87.709414574)
False	False	631	6	8	44.0	06	1184622.0	1851863.0	2006	04/15/2016 08:55:02 AM	41.748675	-87.599049	(41.748674558, -87.599048654)
...	...	...	...	...	...	...	...	...	...	...	...	...	...
False	False	2112	1	2	35.0	04B	1179306.0	1885032.0	2006	04/15/2016 08:55:02 AM	41.839816	-87.617516	(41.839816207, -87.617516172)
										04/15/2016			

b. Washington D.C. Data Frame

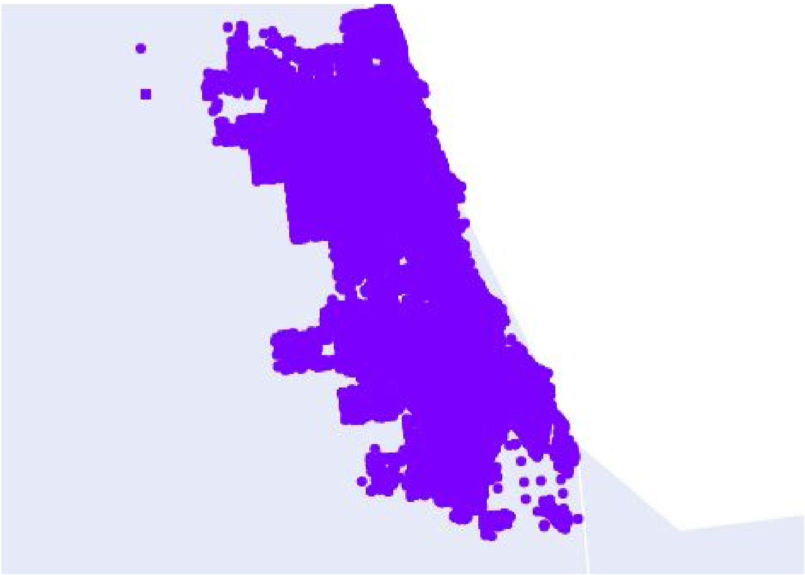


	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Ye
0	4647369	HM155213	01/31/2006 12:13:05 PM	066XX N BOSWORTH AVE	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	SCHOOL, PUBLIC, BUILDING	True	False	2432	24	40	1.0	18	1164737.0	1944193.0	20
1	4647370	HM245080	03/21/2006 07:00:00 PM	062XX S WESTERN AVE	1330	CRIMINAL TRESPASS	TO LAND	PARKING LOT/GARAGE(NON.RESID.)	True	False	825	8	15	66.0	26	1161441.0	1863309.0	20
2	4647372	HM171175	02/09/2006 01:44:41 AM	058XX S SHIELDS AVE	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	STREET	True	False	711	7	20	68.0	18	1174958.0	1866097.0	20
3	4647373	HM244805	03/21/2006 04:45:00 PM	011XX N SPAULDING AVE	0810	THEFT	OVER $500	CHURCH/SYNAGOGUE/PLACE OF WORSHIP	False	False	1121	11	26	23.0	06	1154100.0	1907414.0	20
4	4647374	HM245851	03/21/2006 10:00:00 PM	080XX S DOBSON AVE	0820	THEFT	$500 AND UNDER	STREET	False	False	631	6	8	44.0	06	1184622.0	1851863.0	20
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
19995	4676036	HM274876	04/06/2006 10:00:00 AM	031XX S KOSTNER AVE	0460	BATTERY	SIMPLE	SCHOOL, PUBLIC, BUILDING	True	False	1031	10	22	30.0	08B	1147562.0	1883379.0	20
19996	4676037	HM275025	03/31/2006 09:30:00 PM	048XX N MARINE DR	0460	BATTERY	SIMPLE	HOSPITAL BUILDING/GROUNDS	False	False	2024	20	48	3.0	08B	1170072.0	1932559.0	20
19997	4676039	HM274833	04/06/2006 09:00:00 AM	009XX W 54TH ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE	STREET	False	False	934	9	20	61.0	07	1171042.0	1868997.0	20
19998	4676040	HM276622	04/07/2006 06:45:00 AM	002XX N ARTESIAN AVE	0340	ROBBERY	ATTEMPT: STRONGARM- NO WEAPON	ALLEY	False	False	1332	12	27	28.0	03	1160033.0	1901570.0	20
19999	4676354	HM276705	04/06/2006 08:00:00 AM	048XX W 63RD ST	0820	THEFT	$500 AND UNDER	STREET	False	False	813	8	13	64.0	06	1145236.0	1862397.0	20

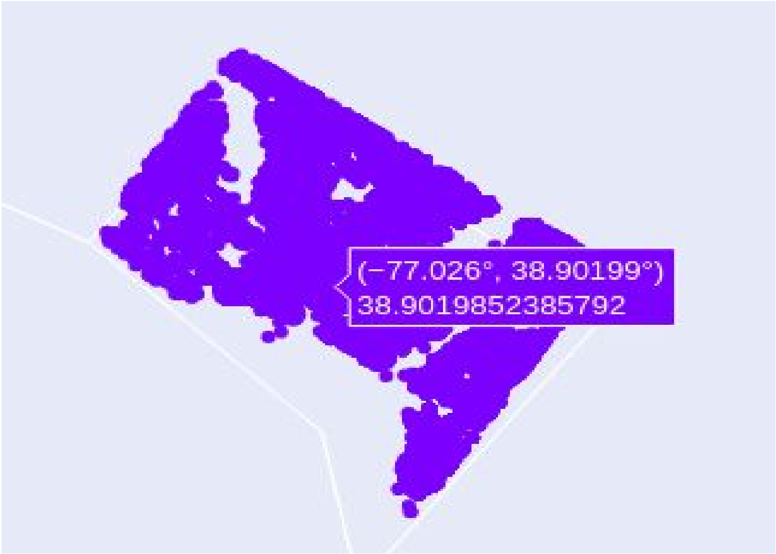
2. Initial Data Analysis

The plot of the raw scattered data of the chosen dataset are as follows,

a. Chicago Data Frame



b. Washington D.C. Data Frame

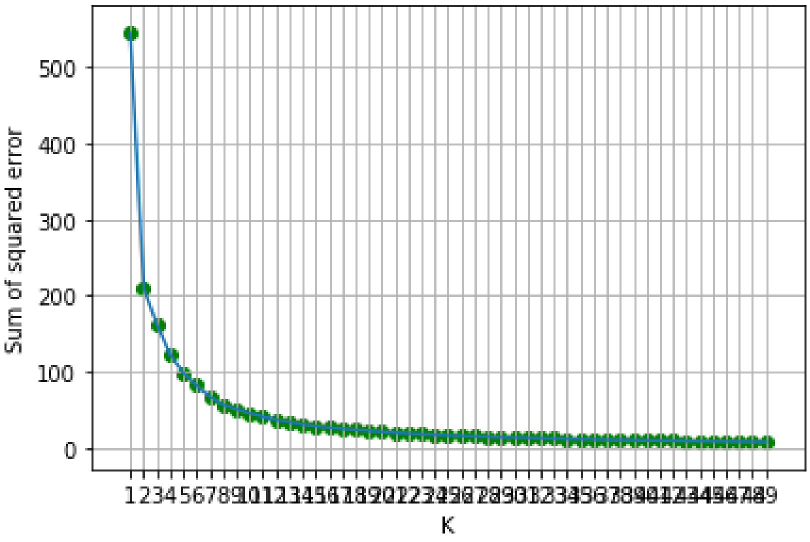


3. Optimal Number Of Clusters

The optimal value of total number of clusters formed for the datasets with the elbow method graphs indicating the values are shown below,

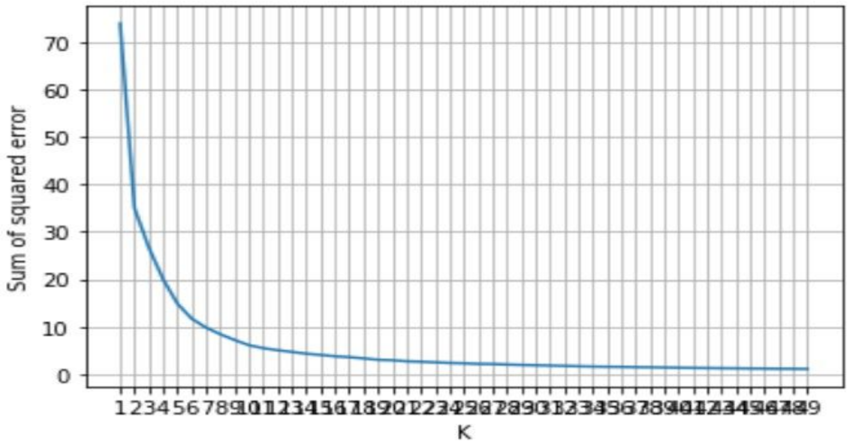
a. Chicago Data Frame

The number of clusters formed are 8.



b. Washington D.C. Data Frame

The number of clusters formed are 8.

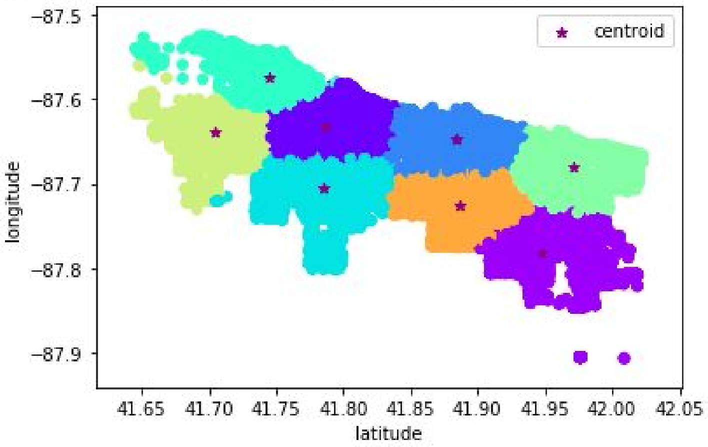


4. Cluster Formation and Locating Cluster Centroid

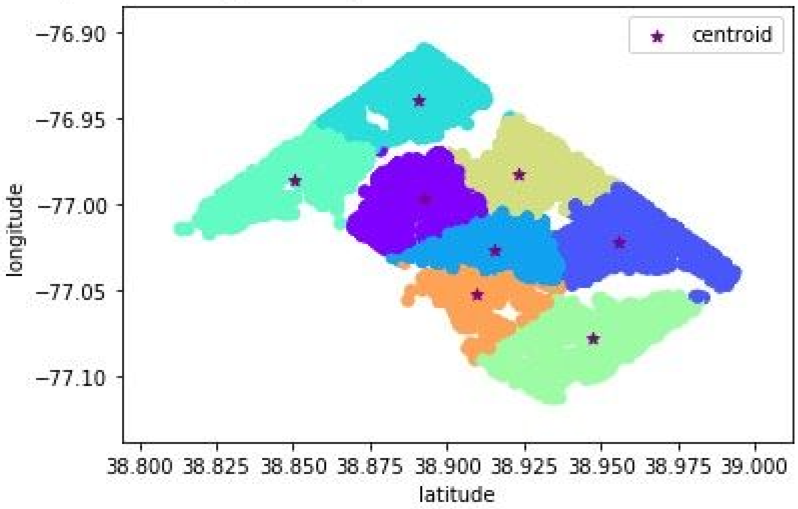
After the formation of clusters on different basis(for example density, type of crime e.t.c) the plot of centroid of different clusters are shown below, this will help the police and the law enforcement authorities to analyse the condition in the area and make rules, regulations and deployment of police or other units in the areas accordingly

a. Chicago Data Frame

```
<matplotlib.legend.Legend at 0x7f246be5bfd0>
```



b. Washington D.C. Data Frame

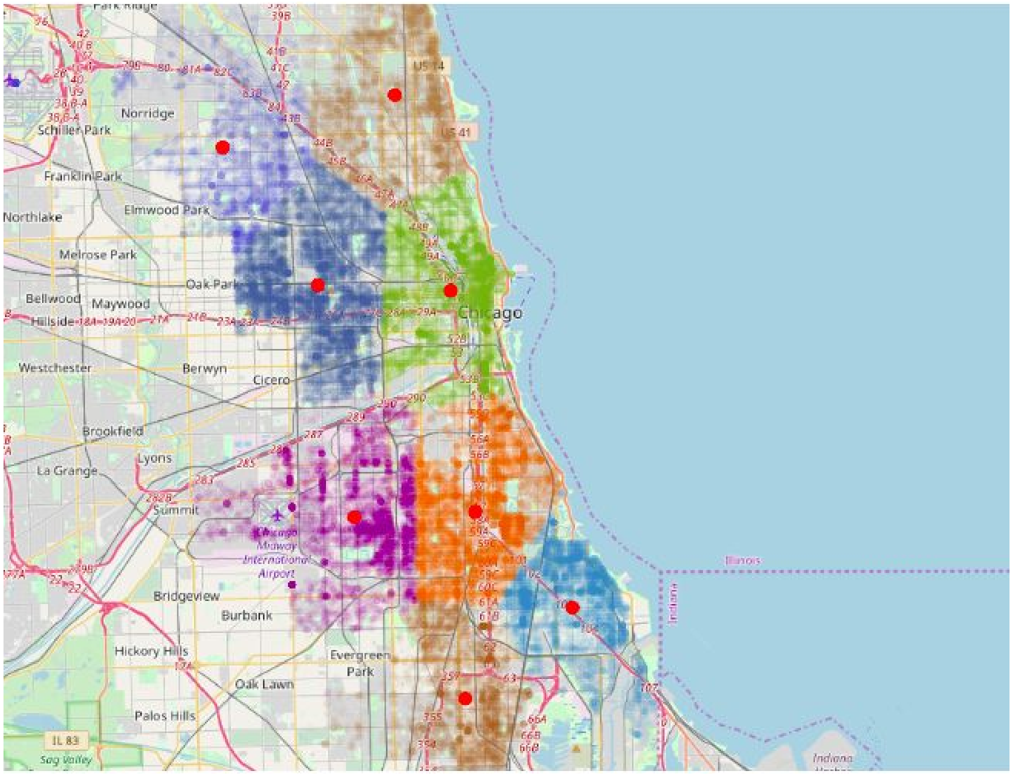


5. Final Visualization

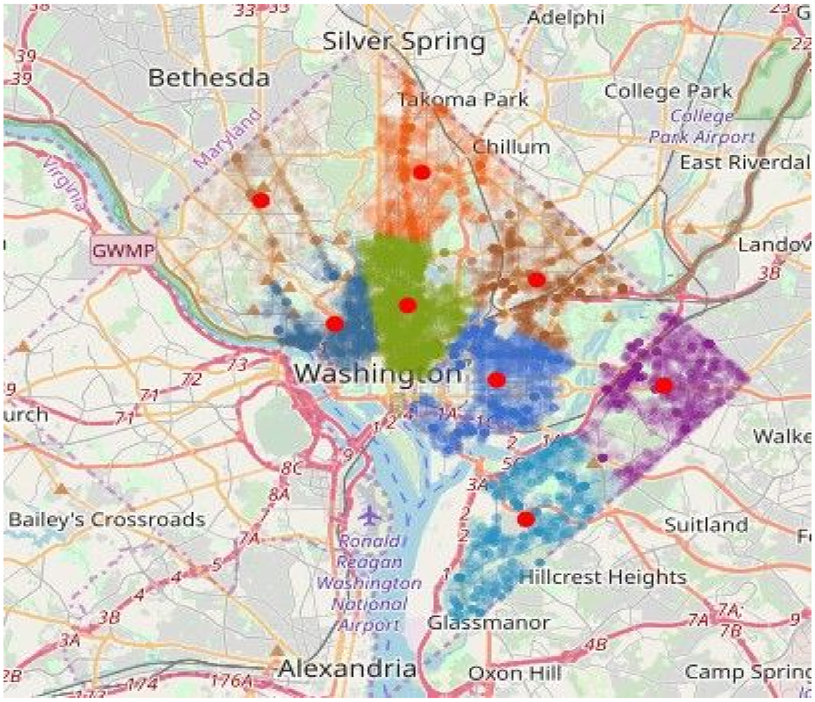
The final results that we have came through after the total processing are as follows,

6.1. Open Street Map

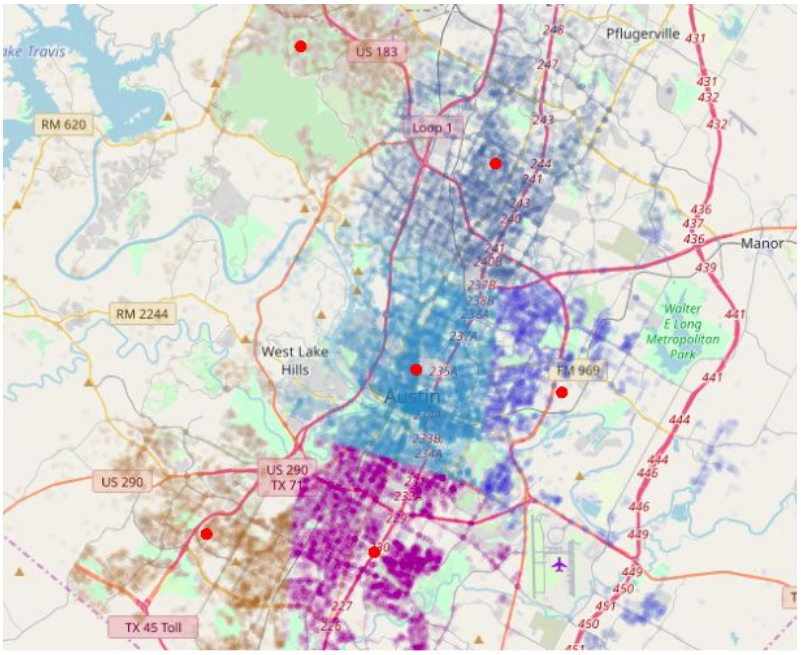
a. Chicago Data Frame



b. Washington D.C. Data Frame

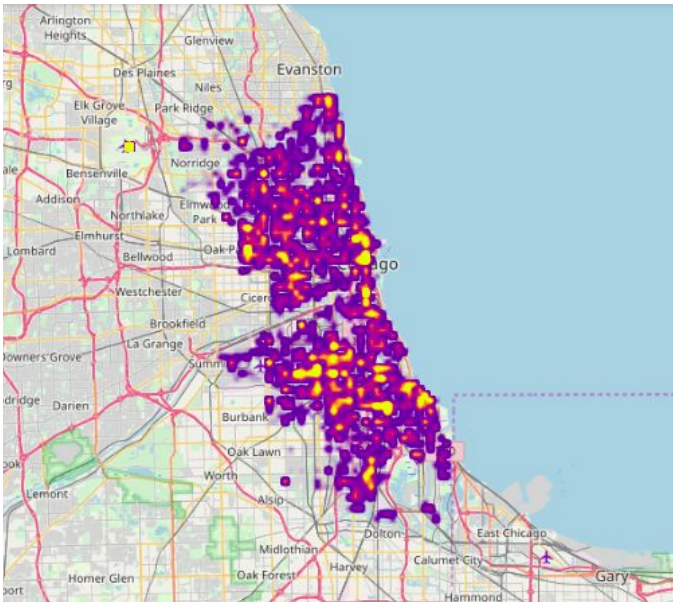


c. Austin

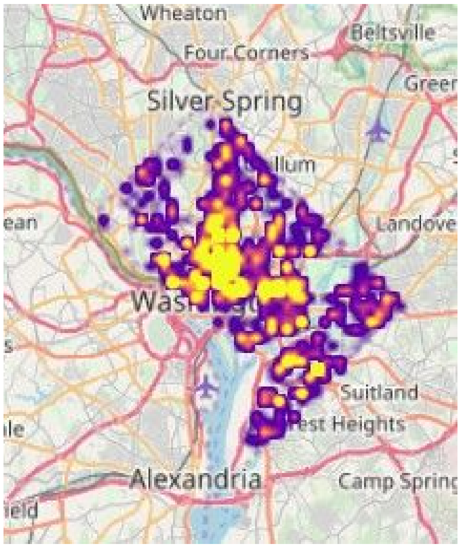


6.2. Heat Map

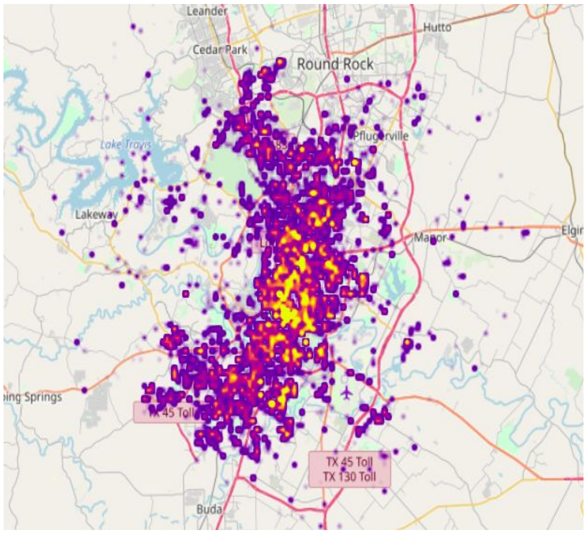
a. Chicago Data Frame



b. Washington D.C. Data Frame

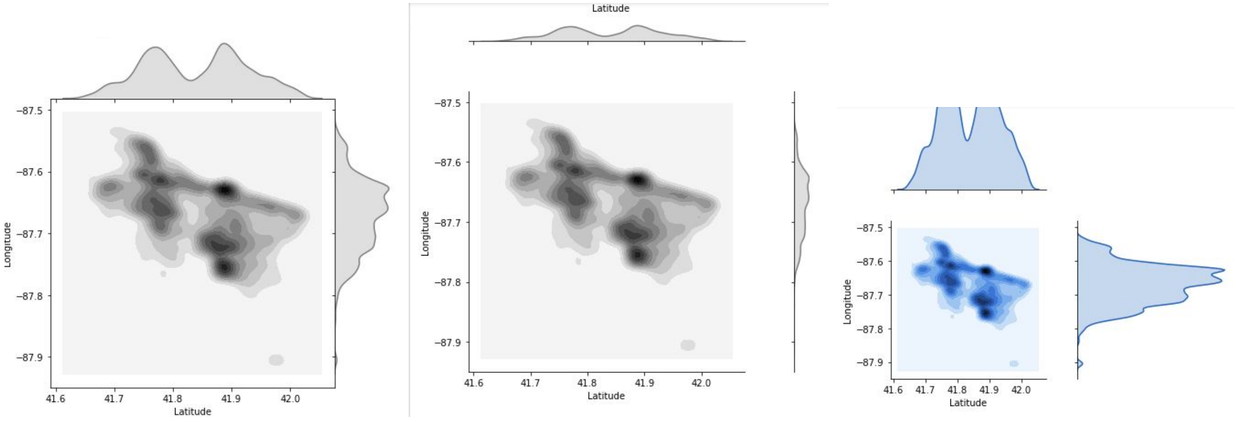


c. Austin

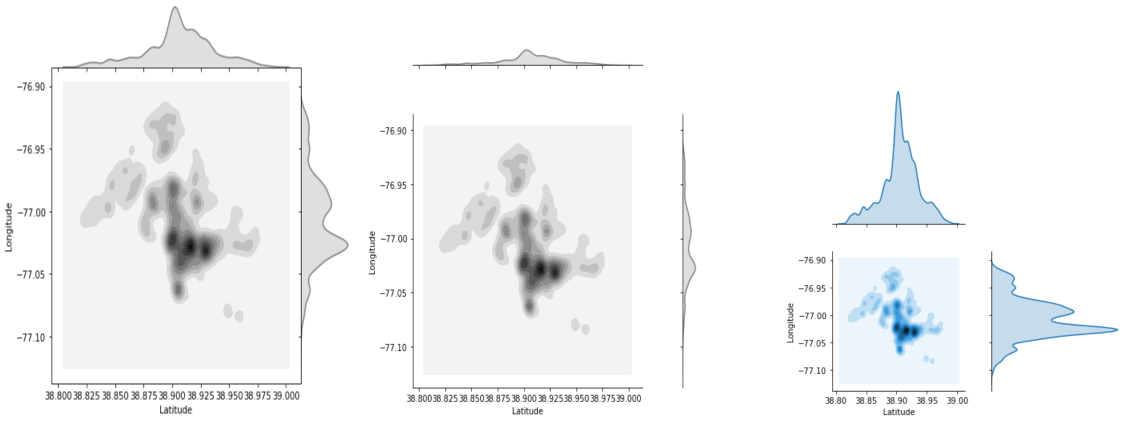


6.3. Contour Map

a. Chicago Data Frame



b. Washington D.C. Data Frame



Developments:

1. The paper “**Crime Analysis Using K-Means Clustering**” by Anant Joshi, A. Sai Sabitha; Tanupriya Choudhury analyzes, crime is essential for providing safety and security to the civilian population. Using data mining, we can discover critical information which can help local authorities detect crime and areas of importance. The main purpose of this paper is to analyze the crime which entails theft, homicide and various drug offenses which also include suspicious activities, noise complaints, and burglar alarm by using qualitative and quantitative approaches. Using the K-means clustering data mining approach on a crime dataset from the New South Wales region of Australia, crime rates of each type of crime and cities with high crime rates have been found.

2. The Paper “**Integration of the Kohonen’s self-organizing map and k-means algorithm for the segmentation of the AE data collected during tensile tests on cross-ply composites**” by S. Huguet, R. Gaertner. This study deals with the ability of a Kohonen’s map to classify recorded AE signals collected during tensile tests on cross-ply glass/epoxy composites in order to monitor the chronology of the damaging process. An unsupervised clustering analysis shows that AE signals are distributed into three clusters. The proposed two-stage procedure is a combination of the Self-Organising Map (SOM) and the k-means methods. In the present work, Kohonen’s map is applied as an unsupervised clustering method for the AE signals generated in cross-ply composite specimens during tensile tests. The input vectors of the signal descriptors used in the clustering procedure are calculated from the signal waveforms. The k-means method is then applied to the neurons of the map in order to delimit the clusters and to visualize the topology of the map.

Conclusion:

The proposed framework can perform better and can be used by different agencies like Law enforcement agencies, Police e.t.c in assessing more appropriate law enforcement strategies, as well as improving the use of police duty deployment for crime prevention.

References

https://ieeexplore.ieee.org/abstract/document/8307327

https://www.sciencedirect.com/science/article/pii/S0963869504001197

https://ieeexplore.ieee.org/abstract/document/1304160

https://ieeexplore.ieee.org/abstract/document/1187438/

https://ieeexplore.ieee.org/abstract/document/844265