IS 733 DATA MINING (01.1958) SP2025

GROUP PROJECT FINAL REPORT

ON

CRIME PATTERN ANALYSIS ON NYPD ARREST DATA

GROUP – 11:

SUSHANTH REDDY MANDAPURAM
SAI SUMANTH REDDY KACHI
VAGBHAT DWIBHASHYAM
SHASHANK PUPPALA

1. Problem Statement / Objective:

Analyzing arrest data that uses unsupervised ML models to detect and understand crime hotspots in New York City is the primary objective of this project. In the project, geospatial clustering mechanisms like the **K-Means, DBSCAN and Hierarchical Clustering** will be leveraged to identify spatial patterns in the arrest data which can reveal crime hubs. Through this, we can obtain insights into designing smarter public safety strategies, better inform the Department of Law enforcement and additionally assist urban planners in addressing crime hotspots with improved efficiency.

2. Project Motivation:

In the case of urban neighborhoods, specifically in cities like **New York that boast huge population and experience socio-economic diversity**, crime happens to be one of the most persistent problems. The conventional techniques of mapping crime may not be very effective as they often tend to be static and inadaptable to real-time analysis. But as ML advances and public data becomes more accessible, opportunities might emerge to harness data-driven models to better understand crime patterns. Therefore, the project was inspired by the need to apply a data-driven approach to understand important trends using complex datasets, which would in-turn enhance resource allocation, community safety and long-term policy planning.

3. Abstract:

The project puts forth a data-driven approach to analyzing urban crime patterns using the openly available NYPD arrest data. To this, we applied the unsupervised ML algorithms such as K-Means, BDSCAN and Hierarchical Clustering to identify crime hubs in New York City using geospatial coordinates. The dataset obtained consisted of over 260,000 arrest records with demographic, temporal and location-based details. After processing the data extensively and conducting exploratory data-analysis, the clustering models were trained and compared using shadow scores to examine performance. The K-Means that was applied produced a high clarity segmentation of high-crime areas with a shadow score of 0.4776. The insights gained reiterate the importance of ML in understanding the dynamics of urban crime, thereby enabling law enforcement and public policy agencies to make data-driven decisions for crime prevention and resource management.

4. Introduction:

In New York, which houses densely populated neighborhoods and diverse socio-economic conditions, urban crime often possesses complex challenges that contribute to multiple criminal activities. Therefore, for public safety, it becomes essential to move beyond the responsive strategies and adopt proactive and data-driven crime prevention methods. The project aims to analyze New York City's arrest data by employing ML to unravel hidden patterns and spatial clusters that indicate potential crime hotspots. With

the help of clustering algorithms, this study puts down its objective to identify crucial zones with prominent crime activity, to interpret demographic influences and to lay a foundation for generating predictive modeling in public safety applications.

5. Background and Related work:

Crime in urban areas has been a major concern of public safety efforts, and with greater availability of open government data, crime pattern analysis is also becoming more data driven. The New York Police Department (NYPD) publishes comprehensive arrest data, including geographic coordinates, offense categories, and times of day, which enable researchers to examine spatial and temporal patterns. This frequency has been a major driver of the use of machine learning techniques to identify hidden patterns in crime data. Unsupervised learning algorithms, especially clustering algorithms, are commonly used to determine areas of high crime and discover how crimes are geographically dispersed throughout the city. City planners (Who build the government related buildings) and law enforcement can more effectively allocate resources and discover emerging hotspots in real-time through this data.

Previous research methodologies have also indicated the effectiveness of various clustering techniques in crime detection. As it is easy and well-liked cause, K-Means clustering performs badly with non-convex shapes (Which are irregular) and noise present in real-world datasets. DBSCAN (Density-Based Spatial Clustering of Applications with Noise), however, can identify arbitrarily shaped clusters and also get rid of sparse, unrelated points, and noise in better way than hierarchical and K-means. Thus, better suited to New York City's complicated urban environment.

Hierarchical clustering offers a solution without pre-specifying the number of groups, with exploratory analysis flexibility. Global and local usage of these models has been implemented, including in research on projects involving NYPD data, for the identification of burglary, assault, house break and drug offense trend detection. Our project is an expansion of such work by comparing a range of clustering algorithms to identify important spatial crime patterns from actual NYPD arrest data.

6. Methodology:

6.1. Dataset Description:

https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date/resource/c48f1a1a-5efb-4266-9572-769ed1c9b472

The dataset is acquired from data.gov (an official website of the United States Government) with a title 'NYPD_Arrest_Data__Year_to_Date_'. The dataset contains 19 attributes and 260503 instances. Out of these 19 attributes 9 are numerical, and 10 are categorical. There are no redundant or duplicate rows in the dataset. But the dataset has 1442 (<0.1%) missing cells.

Total attributes:

ARREST_KEY, ARREST_DATE, PD_CD, PD_DESC, KY_CD, OFNS_DESC, LAW_CODE, LAW_CAT_CD, ARREST_BORO, ARREST_PRECINCT, JURISDICTION_CODE, AGE_GROUP, PERP_SEX, PERP_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, New Georeferenced Column.

Important attributes:

OFNS_DESC, LAW_CAT_CD, ARREST_BORO, ARREST_PRECINCT, AGE_GROUP, PERP_SEX, PERP_RACE, Latitude, Longitude.

Sample (first 10 rows) of the dataset:

	ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO
0	281369711	01/30/2024	177.0	SEXUAL ABUSE	116.0	SEX CRIMES	PL 1306501	F	M
1	284561406	03/30/2024	105.0	STRANGULATION 1ST	106.0	FELONY ASSAULT	PL 1211200	F	B
2	284896016	04/06/2024	105.0	STRANGULATION 1ST	106.0	FELONY ASSAULT	PL 1211200	F	M
3	285569016	04/18/2024	105.0	STRANGULATION 1ST	106.0	FELONY ASSAULT	PL 1211200	F	K
4	287308954	05/22/2024	464.0	JOSTLING	230.0	JOSTLING	PL 1652501	M	M
5	286793332	05/13/2024	155.0	RAPE 2	104.0	RAPE	PL 1303001	F	Q
6	279892607	01/03/2024	153.0	RAPE 3	104.0	RAPE	PL 1302503	F	Q
7	280263905	01/10/2024	157.0	RAPE 1	104.0	RAPE	PL 1303501	F	B
8	288072319	06/06/2024	808.0	TAX LAW	125.0	OTHER STATE LAWS	TAX18140B3	F	M
9	288408753	06/12/2024	105.0	STRANGULATION 1ST	106.0	FELONY ASSAULT	PL 1211200	F	B

ARREST_PRECINCT	JURISDICTION_CODE	AGE_GROUP	PERP_SEX	PERP_RACE	X_COORD_CD	Y_COORD_CD
25	0	25-44	M	BLACK	1000558	231080
44	0	25-44	M	BLACK	1004297	242846
19	0	25-44	M	BLACK	997304	222853
69	0	25-44	M	BLACK	1010576	175628
18	0	18-24	M	WHITE	991530	217373
112	0	18-24	M	BLACK HISPANIC	1025401	202586
113	0	25-44	M	BLACK	1046315	187088
42	0	25-44	M	BLACK	1008690	238862
13	0	45-64	M	BLACK	987373	210805
52	0	45-64	M	BLACK	1012026	253649

Latitude	Longitude	New Georeferenced Column
40.800930	-73.941098	POINT (-73.9410982410066 40.8009303727402)
40.833209	-73.927554	POINT (-73.927554 40.833209)
40.778348	-73.952863	POINT (-73.952863 40.778348)
40.648698	-73.905128	POINT (-73.905128 40.648698)
40.763313	-73.973717	POINT (-73.973717 40.763313)
40.722641	-73.851542	POINT (-73.8515418216779 40.7226409964758)
40.679981	-73.776234	POINT (-73.7762339071953 40.6799807384666)
40.822271	-73.911698	POINT (-73.911697780277 40.8222710411331)
40.745287	-73.988729	POINT (-73.98872939424497 40.7452870263689)
40.862840	-73.899580	POINT (-73.89958 40.86284)

6.2. Exploratory Data Analysis (EDA):

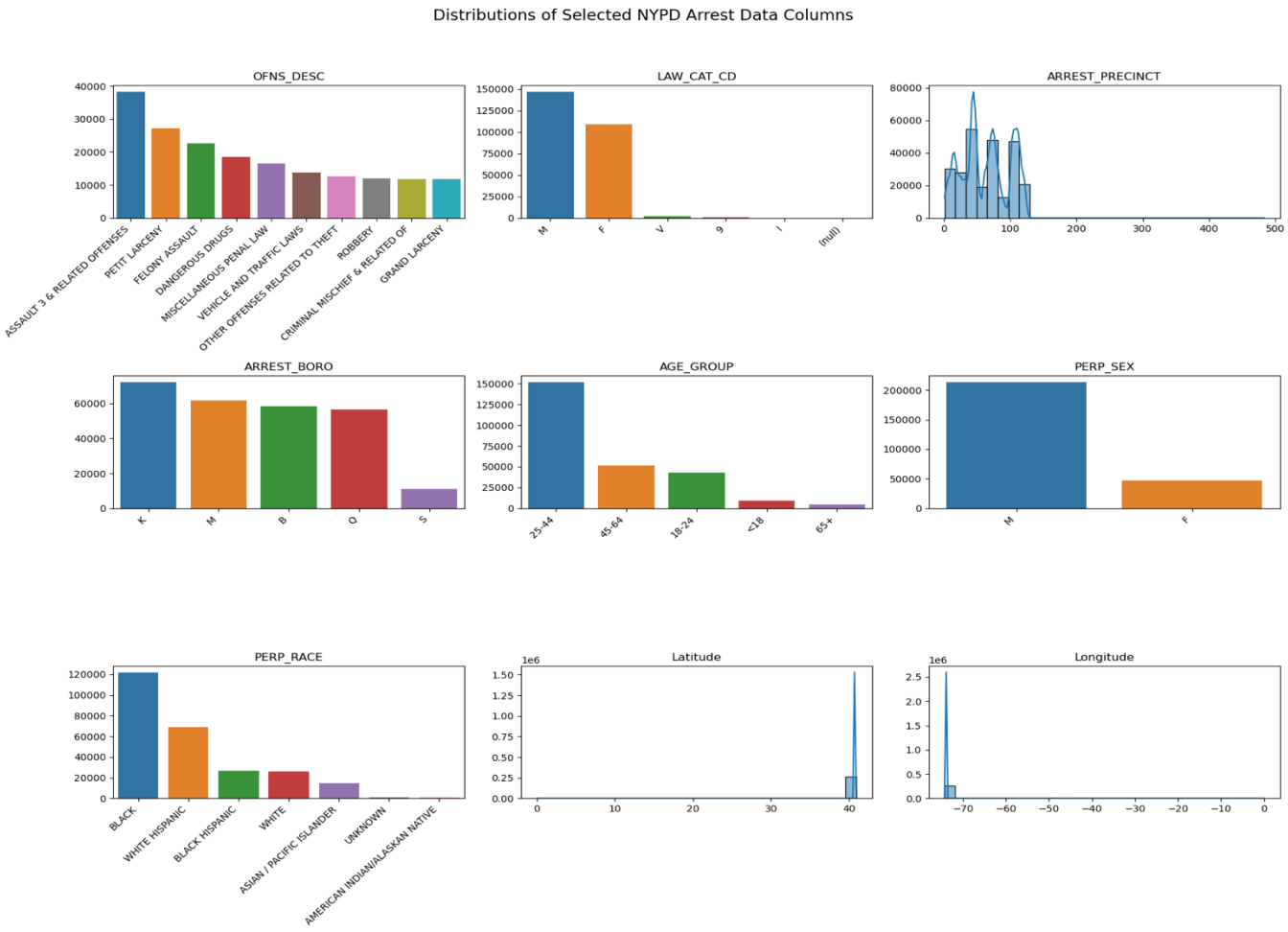
EDA played a crucial role in understanding the dataset and preparing it for modeling.

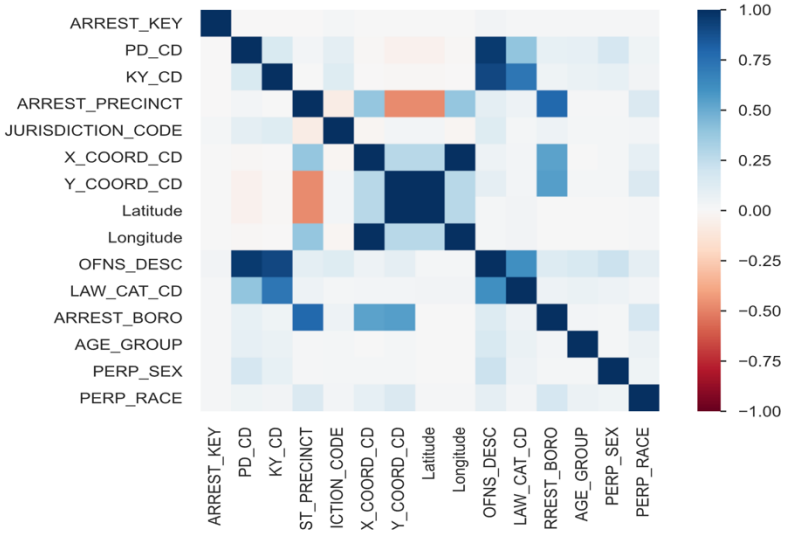
Key techniques included:

Correlation Heatmap: These helped identify relationships between features, highlighting attributes with strong positive or negative correlations to the target variable.

Distribution Plots: Used to examine the spread of features, detect outliers, and understand patterns in data distributions. The distribution plots of latitude and longitude helped in defining the scale for the clusters plotted by the models. All the latitude points are around +40 and all the latitude points are around –74. This ensures that the x – axis and y – axis values of the cluster plot should be around –74 and +40 respectively.

The Distribution plots are as shown below:





6.3. Data Pre-processing:

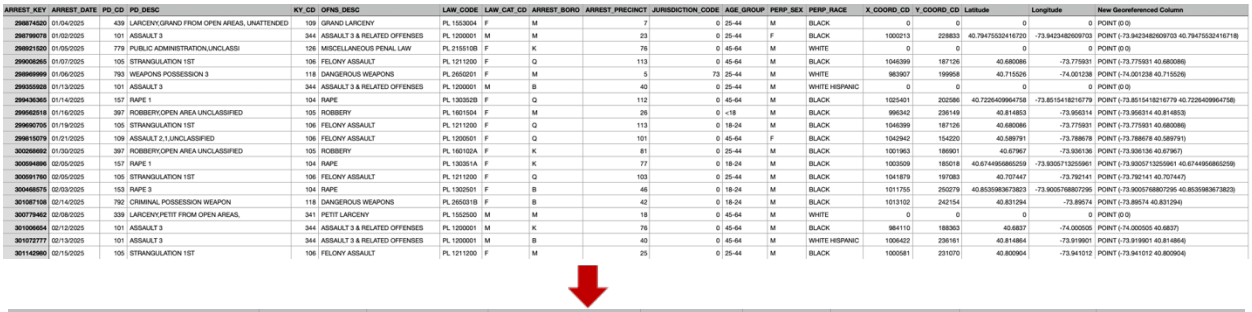
6.3.1. Dimensionality reduction:

A total of 10 attributes (ARREST_KEY, ARREST_DATE, PD_CD, PD_DESC, KY_CD, LAW_CODE, X_COORD_CD, Y_COORD_CD, JURISDICTION_CODE, and New Georeferenced Column) which are not used to develop or build the model have been dropped from the dataset.

- The ARREST_KEY is dropped as it is the unique key which is only used to access the row, and it is not useful to build the model.
- The ARREST_DATE is dropped as the model is not built using temporal analysis.
- The PD_CD, PD_DESC and KY_CD are dropped as they are highly correlated with OFNS_DESC.
- The LAW_CODE is dropped due to its high cardinality.
- The X_COORD_CD and Y_COORD_CD are dropped are they are highly correlated with Longitude and Latitude respectively.
- The JURISDICTION_CODE is dropped as it does not provide any significant information useful to develop the model.
- The New Georeferenced Column is dropped as it is just a new column combining the Latitude and Longitude columns.

6.3.2. Handling missing values:

The entire dataset has only 1442 missing cells which is less than 0.1% cells of the dataset. Hence, after dropping the unnecessary attributes from the dataset, the rows which contain the missing cells are dropped from the dataset.



ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO	ARREST_PRECINCT	JURISDICTION_CODE	AGE_GROUP	PERP_SEX	PERP_RACE	X_COORD_CD	Y_COORD_CD	Latitude	Longitude	New Georeferenced Column
298874520	01/04/2025	439	LARCENY,GRAND FROM OPEN AREAS, UNATTENDED	109	GRAND LARCENY	PL 1553004	F	M	7	0	25-44	M	BLACK	0	0	0	0	POINT (0 0)
298799078	01/02/2025	101	ASSAULT 3	344	ASSAULT 3 & RELATED OFFENSES	PL 1200001	M	M	23	0	25-44	F	BLACK	1000213	228833	40.794755324167 20	-73.942348260970 3	POINT (-73.942348260970 40.794755324167 18)
298921520	01/05/2025	779	PUBLIC ADMINISTRATION,UNCLASSI	126	MISCELLANEOUS PENAL LAW	PL 215510B	F	K	76	0	45-64	M	WHITE	0	0	0	0	POINT (0 0)
299008265	01/07/2025	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F	Q	113	0	45-64	M	BLACK	1046399	187126	40.680086	-73.775931	POINT (-73.775931 40.680086)
298969999	01/06/2025	793	WEAPONS POSSESSION 3	118	DANGEROUS WEAPONS	PL 2650201	F	M	5	73	25-44	M	WHITE	983907	199958	40.715526	-74.001238	POINT (-74.001238 40.715526)
299355928	01/13/2025	101	ASSAULT 3	344	ASSAULT 3 & RELATED OFFENSES	PL 1200001	M	B	40	0	25-44	M	WHITE HISPANIC	0	0	0	0	POINT (0 0)
299436365	01/14/2025	157	RAPE 1	104	RAPE	PL 1303528	F	Q	112	0	45-64	M	BLACK	1025401	202586	40.722640996475 8	-73.851541821677 9	POINT (-73.851541821677 9 40.722640996475 8)
299565518	01/16/2025	397	ROBBERY,OPEN AREA UNCLASSIFIED	105	ROBBERY	PL 1601504	F	M	28	0	<18	M	BLACK	996342	236149	40.81483	-73.956314	POINT (-73.956314 40.81483)
299690705	01/19/2025	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F	Q	113	0	18-24	M	BLACK	1046399	187126	40.680086	-73.775931	POINT (-73.775931 40.680086)
299815079	01/21/2025	109	ASSAULT 2,1,UNCLASSIFIED	106	FELONY ASSAULT	PL 1200501	F	Q	101	0	45-64	F	BLACK	1042942	154220	40.589791	-73.788678	POINT (-73.788678 40.589791)
300268692	01/30/2025	397	ROBBERY,OPEN AREA UNCLASSIFIED	105	ROBBERY	PL 160102A	F	K	81	0	25-44	M	BLACK	1001963	186901	40.67967	-73.936136	POINT (-73.936136 40.67967)
300594896	02/05/2025	157	RAPE 1	104	RAPE	PL 130351A	F	K	77	0	18-24	M	BLACK	1003509	185018	40.674495865259	-73.930571325996	POINT (-73.930571325996 40.674495865259)
300591760	02/05/2025	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F	Q	103	0	25-44	M	BLACK	1041879	197083	40.707447	-73.792141	POINT (-73.792141 40.707447)
300468575	02/03/2025	153	RAPE 3	104	RAPE	PL 1302501	F	B	46	0	18-24	M	BLACK	1011755	250279	40.853598367382 3	-73.900576807295	POINT (-73.900576807295 40.853598367382 3)
301067108	02/14/2025	792	CRIMINAL POSSESSION WEAPON	118	DANGEROUS WEAPONS	PL 265031B	F	B	42	0	18-24	M	BLACK	1013102	242154	40.831294	-73.89574	POINT (-73.89574 40.831294)
300779462	02/08/2025	339	LARCENY,PETIT FROM OPEN AREAS,	341	PETIT LARCENY	PL 1552500	M	M	18	0	45-64	M	WHITE	0	0	0	0	POINT (0 0)
301006654	02/12/2025	101	ASSAULT 3	344	ASSAULT 3 & RELATED OFFENSES	PL 1200001	M	K	76	0	45-64	M	BLACK	984110	188363	40.6837	-74.000505	POINT (-74.000505 40.6837)
301072777	02/13/2025	101	ASSAULT 3	344	ASSAULT 3 & RELATED OFFENSES	PL 1200001	M	B	40	0	45-64	M	WHITE HISPANIC	1006422	236161	40.81486	-73.919901	POINT (-73.919901 40.81486)
301142980	02/15/2025	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F	M	25	0	25-44	M	BLACK	1000581	231070	40.800904	-73.941012	POINT (-73.941012 40.800904)

OFNS_DESC	LAW_CAT_CD	ARREST_BORO	ARREST_PRECINCT	AGE_GROUP	PERP_SEX	PERP_RACE	Latitude	Longitude
SEX CRIMES	F	M	25	25-44	M	BLACK	40.8009303727402	-73.9410982410066
FELONY ASSAULT	F	B	44	25-44	M	BLACK	40.833209	-73.927554
FELONY ASSAULT	F	M	19	25-44	M	BLACK	40.778348	-73.952863
FELONY ASSAULT	F	K	69	25-44	M	BLACK	40.648698	-73.905128
JOSTLING	M	M	18	18-24	M	WHITE	40.763313	-73.973717
RAPE	F	Q	112	18-24	M	BLACK HISPANIC	40.7226409964758	-73.8515418216779
RAPE	F	Q	113	25-44	M	BLACK	40.6799807384666	-73.7762339071953
RAPE	F	B	42	25-44	M	BLACK	40.8222710411331	-73.911697780277
OTHER STATE LAWS	F	M	13	45-64	M	BLACK	40.7452870263689	-73.98872939424500
FELONY ASSAULT	F	B	52	45-64	M	BLACK	40.86284	-73.89958

6.4. Modelling:

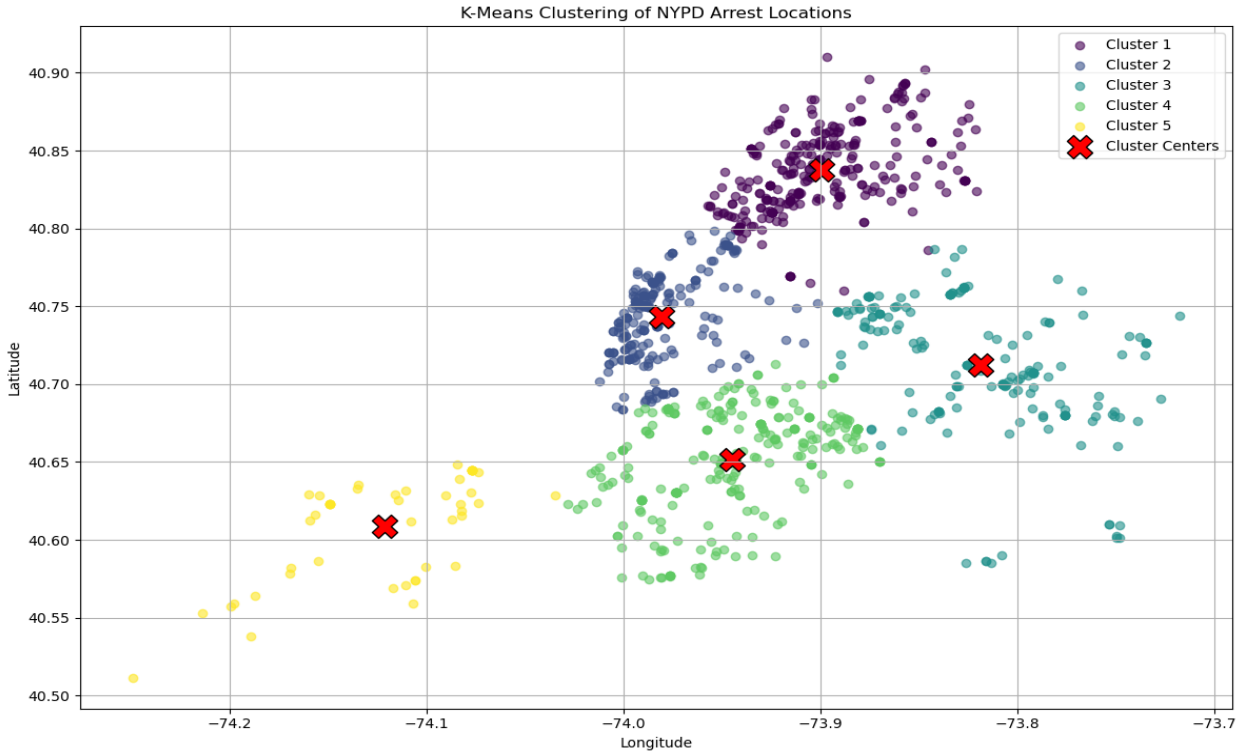
The modelling approaches used for the analysis of NYPD Crime data are unsupervised Learning Techniques such as K-Means Clustering, Hierarchical Clustering, and DBSCAN. The First Method which is selected is K-Means Clustering after the data is properly pre-processed and selected required parameters for analysis.

K-Means Clustering – K-means is simply a centroid based clustering algorithm which partitions the data into K-unique Clusters which do not overlap. Once the data is preprocessed using the StandardScalar to scale the co-ordinates, it is split into clusters. The Code block to execute K-Means is as follows:

Model 1 - K-Means Clustering Model

```
# Modelling:

# Model 1:

import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

data = pd.read_csv('modified_dataset.csv')
data = data.dropna(subset=['Latitude', 'Longitude'])
data = data[(data['Latitude'] != 0) & (data['Longitude'] != 0)]

sample_data = data.sample(n=1000, random_state=42)
coordinates = sample_data[['Latitude', 'Longitude']].values

scaler = StandardScaler()
scaled_coords = scaler.fit_transform(coordinates)

kmeans = KMeans(n_clusters=5, random_state=42)
kmeans.fit(scaled_coords)

cluster centers = scaler.inverse transform(kmeans.cluster centers )
```

The output is scatterplot which is shown below:



These are the Cluster Centers identified:

Cluster Centers (Latitude, Longitude): Cluster 1: (40.8374, -73.8996) Cluster 2: (40.7431, -73.9809) Cluster 3: (40.7122, -73.8190) Cluster 4: (40.6518, -73.9449) Cluster 5: (40.6090, -74.1215)

The plot is Scatter plot which is plotted using Latitudes and Longitudes. The Plot shows 5 Clusters (Considered **K=5** based on the dataset) and gives us that the major cities with **Crimes Hotspots are Brooklyn, Queens, Manhattan, Staten Island and Bronx.**

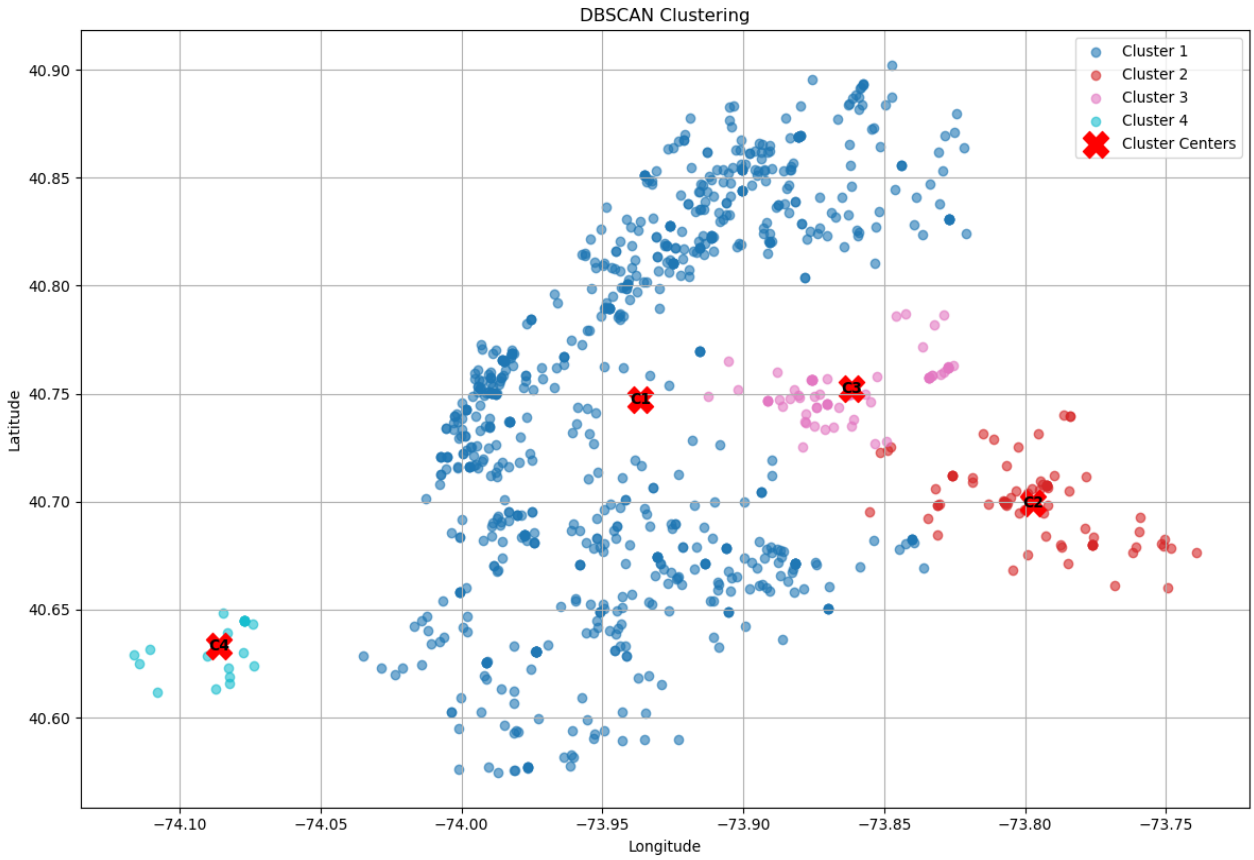
DBSCAN (Density Based Spatial Clustering of Applications with Noise) - The problem with the K-Means Clustering is that it started assuming the clusters to be spherical in shape and tried to fit the data points into a circle of vicinity. But unlike K-Means, DBSCAN is a density based and can take any irregular shapes. The model could even handle the noise and identified the cluster regions and Crime Hotspots accurately.

The Code block to execute DBSCAN model is as shown below:

Model 3 - DBSCAN Model

```
# Model 3:

import pandas as pd
import numpy as np
from sklearn.cluster import DBSCAN
import matplotlib.pyplot as plt

df = pd.read_csv('modified_dataset.csv')
df = df.dropna(subset=['Latitude', 'Longitude'])
df = df[(df['Latitude'] != 0) & (df['Longitude'] != 0)]

if len(df) > 1000:
    df = df.sample(n=1000, random_state=42)

coords_deg = df[['Latitude', 'Longitude']].values
coords_rad = np.radians(coords_deg)

epsilon_km = 2.6
epsilon_rad = epsilon_km / 6371.0088
min_samples = 21

db = DBSCAN(eps=epsilon_rad, min_samples=min_samples, metric='haversine')
labels = db.fit_predict(coords_rad)
df['cluster'] = labels
```

The output of the DBSCAN model is as the scatterplot shown below:



The Epsilon radius is **2.6 KM** and identified minimum of **21 data samples** to become a dense region. The model successfully identified the dense clusters and eliminated the noise. The Cluster centers identified are:

Cluster Centers (Latitude, Longitude): Cluster 1 center: (40.7472, -73.9367) Cluster 2 center: (40.6993, -73.7970) Cluster 3 center: (40.7525, -73.8616) Cluster 4 center: (40.6332, -74.0861)

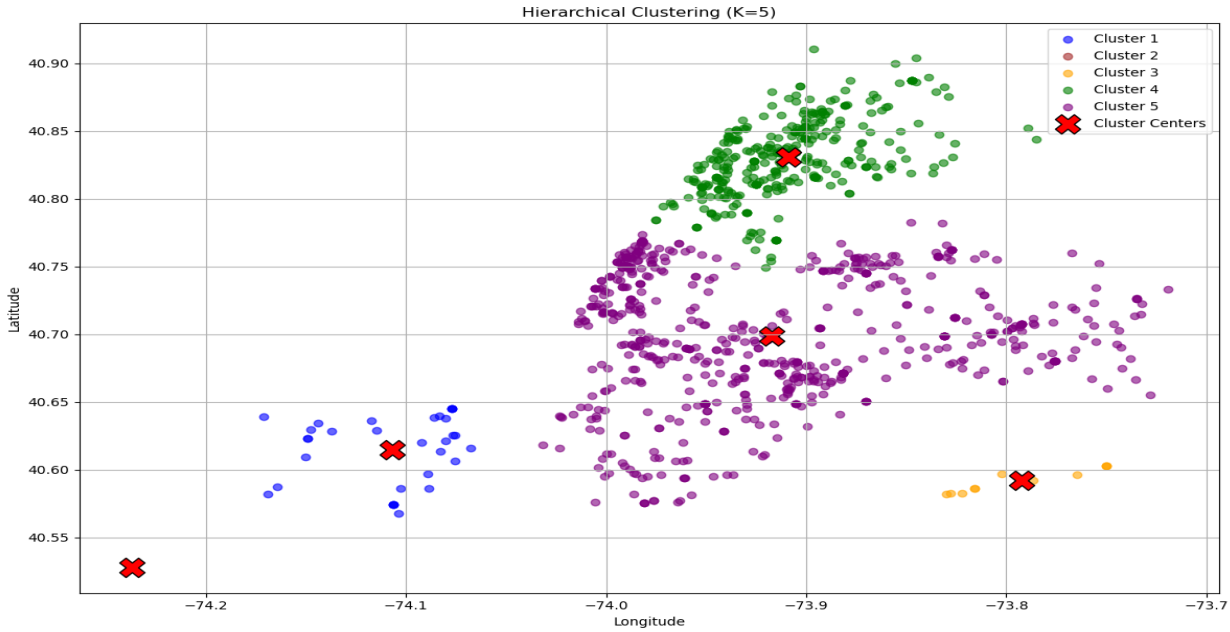
Hierarchical Clustering –

Though DBSCAN could find the high-density clusters, it got few flaws of identifying the less density clusters as noises instead of separate clusters. So, hierarchical clustering is introduced to identify the crime hotspots and segregate the data points into clusters. Unlike K-means clustering, in hierarchical there is no need to mention the k value before. It employs an Agglomerative Approach (Bottom-Up) in building up the clusters and locating the co-ordinates of hotspots. The only drawback with this sort of method is due to its computational complexity and sensitivity to noise which may affect the robustness of the system. The Code block executed for Hierarchical clustering is as follows:

Model 2 - Hierarchical Clustering Model

```
# Model 2:

import pandas as pd
import numpy as np
from scipy.cluster.hierarchy import linkage, fcluster
from scipy.spatial.distance import pdist
import matplotlib.pyplot as plt

df = pd.read_csv('modified_dataset.csv')

if len(df) > 1000:
    df = df.sample(n=1000, random_state=42)

coords_deg = df[['Latitude', 'Longitude']].values
coords_rad = np.radians(coords_deg)

def haversine(u, v):
    R = 6371.0088
    dlat = v[0] - u[0]
    dlon = v[1] - u[1]
    a = np.sin(dlat / 2)**2 + np.cos(u[0]) * np.cos(v[0]) * np.sin(dlon / 2)**2
    return 2 * R * np.arcsin(np.sqrt(a))
```

The scatterplot of the model is as follows:



The Cluster centers identified are as follows:

Cluster Centers (Latitude, Longitude): Cluster 1: (40.6146, -74.1070) Cluster 2: (40.5281, -74.2371) Cluster 3: (40.5921, -73.7923) Cluster 4: (40.8305, -73.9091) Cluster 5: (40.6985, -73.9170)

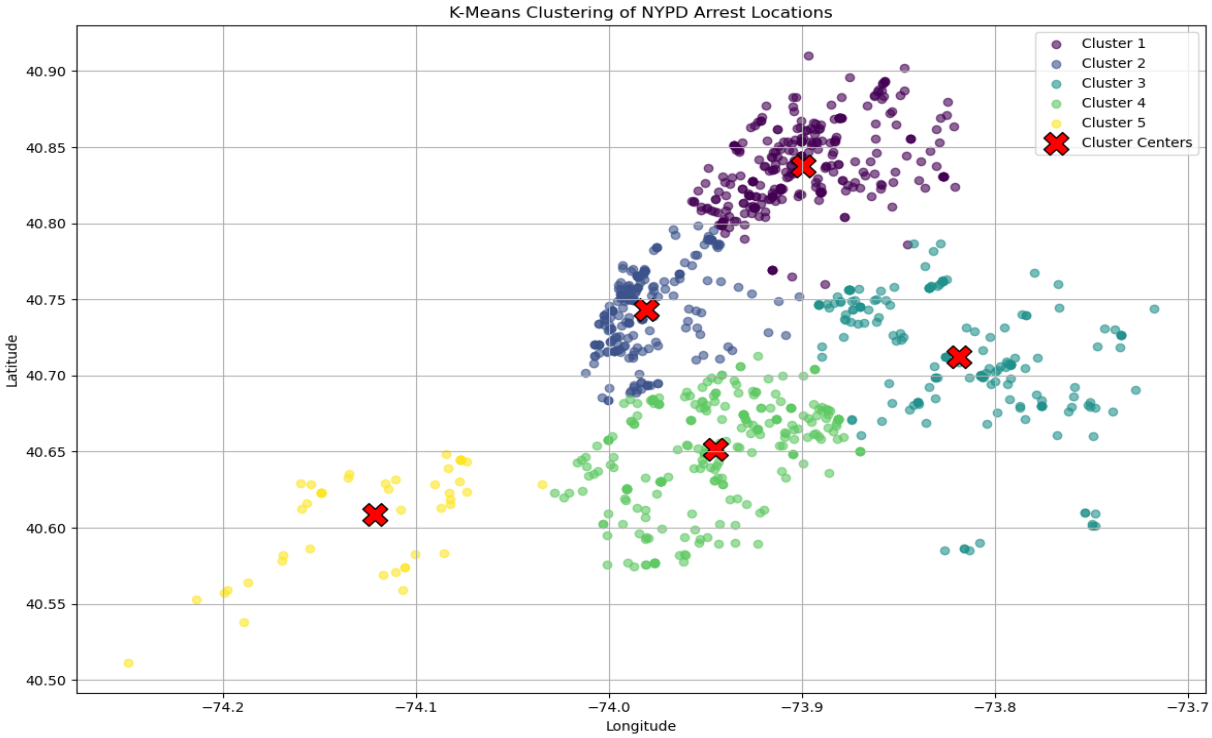
7. <u>Results:</u> To compare various clustering models, it is necessary to have comparison metric. The one used here is the "**Silhouette Score**", and when the score of all the 3 models is observed, it is K-means which performed better than DBSCAN and Hierarchical clustering models. The Silhouette score of all the 3 models is as follows:

Silhouette Scores:

K-Means: 0.4776 Hierarchical: 0.4568 DBSCAN: 0.0541

So, **K-means effectively identified the clusters and hotspots for crimes** in New York city. But just finding out the clusters isn't enough; it is more important to understand who the ones are committing the crime. Which means it is very much necessary to understand the user profiles to draw conclusions about who are the ones committing these crimes.

The scatterplot generated by K-means is as follows:



From analyzing the profiles using the demographics and various other attributes, these are the few profiles who have committed the crimes mostly in New York province:

1.BLACK Male aged 25-44 committed 7188 Misdemeanor crimes (ASSAULT 3 & RELATED OFFENSES)

2.WHITE HISPANIC Male aged 25-44 committed 4957 Misdemeanor crimes (ASSAULT 3 & RELATED OFFENSES)

3.BLACK Male aged 25-44 committed 4603 Felony crimes (FELONY ASSAULT)

4. BLACK Male aged 25-44 committed 4495 Misdemeanor crimes (OTHER OFFENSES RELATED TO THEFT)

5.BLACK Male aged 25-44 committed 4297 Misdemeanor crimes (PETIT LARCENY)

6.BLACK Male aged 25-44 committed 4110 Felony crimes (MISCELLANEOUS PENAL LAW)

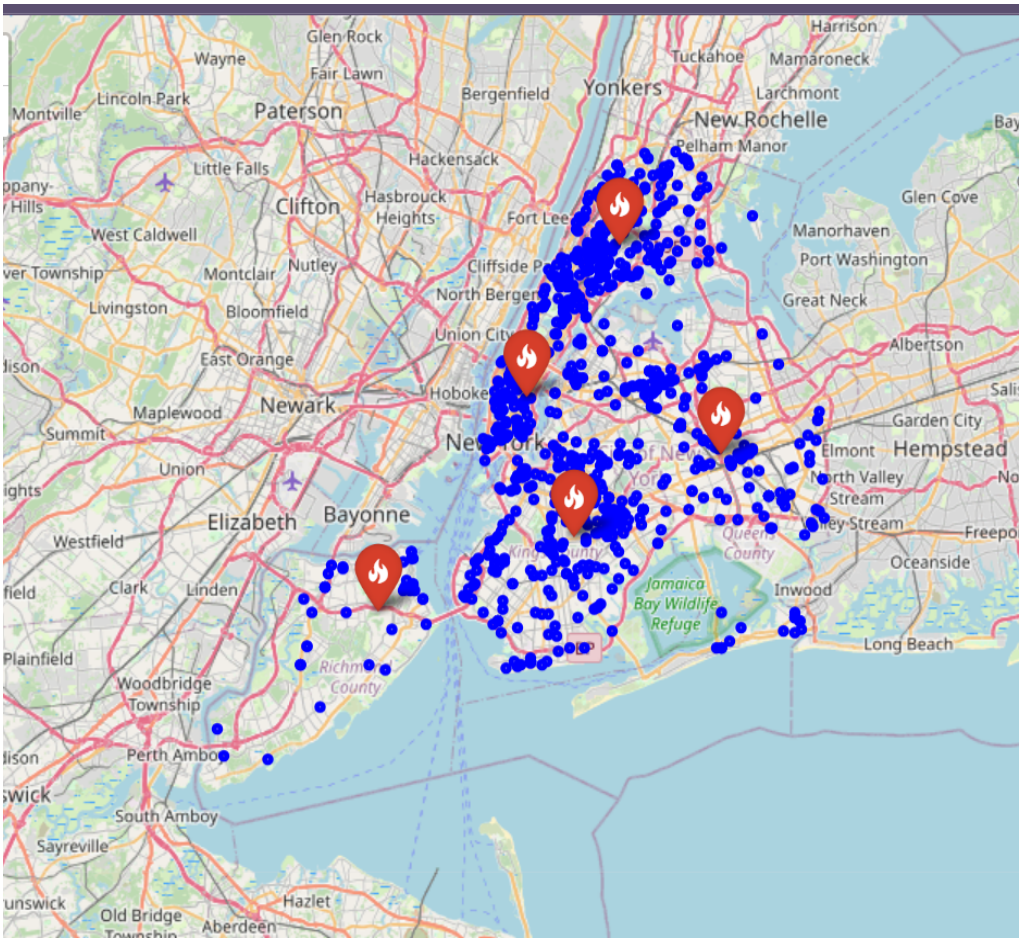
7.BLACK Male aged 25-44 committed 3664 Misdemeanor crimes (VEHICLE AND TRAFFIC LAWS)

8.WHITE HISPANIC Male aged 25-44 committed 3402 Misdemeanor crimes (PETIT LARCENY)

9.BLACK Male aged 45-64 committed 3097 Misdemeanor crimes (PETIT LARCENY)

10.BLACK Female aged 25-44 committed 3038 Misdemeanor crimes (ASSAULT 3 & RELATED OFFENSES)

The Crime Profile analysis helps law enforcement to understand who is committing the crime and make the decisions accordingly. With the Scatter plot shown above it is only possible to get the co-ordinates (Latitude, Longitude) of the place. It is very necessary for the cops or decision makers to know the exact place name instead of retrieving it via the co-ordinates. As a result, we imported the folium package required for retrieving the map and plotting the geospatial coordinates. The output of the K-means model algorithm with the folium package would be as follows:



Where the markers indicate the Crime hotspots and blue points indicate the places of crime.

8. Future Work:

For the future enhancements there are number of things which can be done such as:

(i) Expanding the Temporal Analysis – The crime pattern analysis can be expanded across several years to find long-term trends and seasonal patterns in crimes and arrest data.

(ii) Multivariate Clustering – Extending the clustering to have attributes and include data related unemployment rates, socio-economic backgrounds, education levels, and housing conditions could provide a deeper understanding and analysis of context for arrests and their distribution.

(iii) Real time Crime Heatmaps – By embedding real time data into the analysis and having interactive visualization dashboards can significantly improve monitoring of crimes and locate them more accurately.

(iv) Geospatial Techniques – By utilizing GIS (Geographic Information System) Software the precision of the spatial modeling can be improved, and zoning analysis can be done, to identify the crime zones more effectively.

9. Conclusion:

The primary aim of the project was to discover meaningful patterns in crime data in the city of New York. This has been achieved with the help of unsupervised learning techniques such as **K-Means, Hierarchical, and DBSCAN Clustering**. These models helped to Cluster the data into various Clusters and identify the hotspots of data. Each model had its own strength such as, K-Means provided distinct segmentation, Hierarchical Clustering uncovered nested spatial hierarchies, and DBSCAN handled noise and ir-regular clusters. Out of all the 3 models, **K-means had better silhouette score** and provided hotspots accurately.

Improved the model by finding the location names using the folium package and even analyzed the Crime profiles of users. This data analysis helps in better **urban planning, increased public safety, and allows decision makers to take better decisions**.

Please find below the GitHub Link to the Project Folder consisting of Outputs, Python Code in Jupyter Notebook, and Digital Poster: -

https://github.com/UB01976/is7332025/tree/main/data-mining-project-repo/Group_11_Project

10. References:

1. Kang, J., & Kang, S. (2021). A study on crime hotspot detection using spatio-temporal clustering. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 1830–1837). IEEE. https://doi.org/10.1109/BigData52589.2021.9671321

2. Wang, F., & Brown, D. (2012). The spatio-temporal modeling for criminal incidents. Computers, Environment and Urban Systems, 36(1), 1–11. https://doi.org/10.1016/j.compenvurbsys.2011.10.001

3. Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. *Communications of the ACM, 48*(6), 100–107. https://doi.org/10.1145/1047671.1047676

4. Nath, S. V. (2006). Crime pattern detection using data mining. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence (pp. 41–44). IEEE. https://doi.org/10.1109/WI.2006.68

5. Chainey, S., & Ratcliffe, J. (2005). GIS and crime mapping. Wiley.

6. U.S. Department of Justice, Bureau of Justice Assistance. (2005). Crime analysis in America: Findings and recommendations (NCJ 208994). https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/crime-analysis-in-america.pdf

7. Levine, N. (2004). CrimeStat III: A spatial statistics program for the analysis of crime incident patterns. In *Proceedings of the National Institute of Justice MAPS Conference* (pp. 1–17). https://nij.ojp.gov/library/publications/crimestat-spatial-statistics-program-analysis-crime-incident-patterns

8. Townsley, M. (2008). Visualising space time patterns in crime: The hotspot plot. *Crime Patterns and Analysis, 1*(1), 61–74. https://www.researchgate.net/publication/228564663

9. Liu, H., Brown, D. E., & Pitkin, E. (2011). Clustering criminal activity: A spatio-temporal analysis. International Journal of Intelligence and CounterIntelligence, 24(1), 1–18. https://doi.org/10.1080/08850607.2011.528302

10. Boba Santos, R. (2013). Crime analysis with crime mapping (3rd ed.). Sage Publications.