

# Exploratory Data Analysis on

POLICE

by

## Group 22



Krishil Jayswal  
ID: 202203040  
Course:  
BTech(MnC)



Bhavya Boda  
ID: 202203067  
Course:  
BTech(MnC)



Aniket Pandey  
ID: 202411001  
Course:  
MTech(ML)

Course Code: IT 462  
Semester: Autumn 2024

---

Under the guidance of

**Dr. Gopinath Panda**



Dhirubhai Ambani Institute of Information and Communication Technology

December 2, 2024

# ACKNOWLEDGMENT

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout the duration of my project titled "Police." Your invaluable assistance has played a pivotal role in shaping the successful completion of this endeavor.

I am extremely fortunate to have had the opportunity to work under your mentorship. Your expertise, encouragement, and willingness to share your knowledge have been instrumental in elevating the quality and scope of my project. Your constructive feedback and insightful suggestions have helped me overcome challenges and develop a deeper understanding of the subject matter.

Furthermore, I would like to extend my appreciation to the entire team at Dhirubhai Ambani Institute of Information and Communication Technology for fostering an environment of collaboration and innovation. The resources and facilities provided have been crucial in conducting comprehensive research and analysis.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience, and I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future endeavors.

Once again, thank you for your unwavering guidance and belief in my abilities. Your mentorship has been invaluable, and I am truly grateful for the opportunity to work with you.

Sincerely,

Krishil Jayswal, 202203040  
Bhavya Boda, 202203067  
Aniket Pandey, 202411001

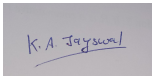
# DECLARATION

We, [202203040, 202203067, 202311001] hereby declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.


We acknowledge that the data used in this project is obtained from the [data.gov](https://data.gov) site. We also declare that we have adhered to the terms and conditions mentioned in the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project, except for the guidance provided by our mentor Prof. Gopinath Panda. We declare that there is no conflict of interest in conducting this EDA project.

We hereby sign the declaration statement and confirm the submission of this report on 2nd July, 2023.



Krishil Jayswal  
ID: 202203040  
Course:  
BTech(Mnc)



Bhavya Boda  
ID: 202303067  
Course:  
BTech(Mnc)



Aniket Pandey  
ID: 202411001  
Course:  
MTech(ML)

# CERTIFICATE

This is to certify that Group 22 comprising Krishil Jayswal, Bhavya Boda, and Aniket Pandey has successfully completed an exploratory data analysis (EDA) project on the Police, which was obtained from data.gov

The EDA project presented by Group 8 is their original work and has been completed under the guidance of the course instructor, Prof. Gopinath Panda, who has provided support and guidance throughout the project. The project is based on a thorough analysis of the NYPD\_Arrest\_Data dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the Police, which demonstrates the analytical skills and knowledge of the students of Group 22 in the field of data analysis.

Signed,  
Dr. Gopinath Panda,  
IT 462 Course Instructor  
Dhirubhai Ambani Institute of Information and Communication Technology  
Gandhinagar, Gujarat, INDIA.

December 2, 2024

# Contents

<b>List of Figures</b>	<b>5</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Project Idea and Defining the Problem . . . . .	1
1.2 Data Collection . . . . .	1
1.3 Dataset Description . . . . .	1
1.4 Packages required . . . . .	2
<b>2 Data Cleaning</b>	<b>3</b>
2.1 Initial Dataset Analysis . . . . .	3
2.1.1 Loading the Dataset . . . . .	3
2.2 Missing data analysis . . . . .	7
2.2.1 Visualization of Missing Values . . . . .	8
2.2.2 Type of Missingness in Our Dataset . . . . .	10
2.3 Imputation . . . . .	11
2.3.1 Imputation Strategy . . . . .	11
<b>3 Visualization</b>	<b>13</b>
3.1 Univariate analysis . . . . .	13
3.1.1 Pie Chart . . . . .	13
3.1.2 Histogram . . . . .	15
3.1.3 Bar Chart . . . . .	18
3.1.4 Box Plot . . . . .	21
3.2 Multivariate Analysis . . . . .	21
3.2.1 Scatter Plot . . . . .	22
3.2.2 Correlation Heatmap . . . . .	23
<b>4 Feature Engineering in Clustering Models</b>	<b>25</b>
4.1 Introduction to Feature Engineering . . . . .	25
4.2 Feature Selection Using Correlation . . . . .	25
4.3 Data Transformation . . . . .	28
4.3.1 Reasoning for Not Removing Outliers . . . . .	28
4.3.2 Reasoning for Not Scaling Latitude and Longitude . . . . .	28
4.4 Conclusion . . . . .	28
<b>5 Model Fitting</b>	<b>30</b>
5.1 Insight into Model Fitting in EDA . . . . .	30
5.2 Types of Models Used for Fitting . . . . .	30

5.2.1	Regression Models	30
5.2.2	Classification Models	31
5.2.3	Clustering Models	31
5.3	Clustering Models in Our Project	31
5.4	Performance Metrics for Clustering Models	31
5.4.1	Elbow Method	32
5.4.2	Silhouette Score	32
5.5	Model Fitting and Cluster Evaluation	33
5.5.1	Elbow Method for Optimal Clusters	33
5.5.2	Hotspots of Agglomeric	35
5.5.3	Hotspots of K-Means	36
5.5.4	Silhouette Score Vs No. of Clusters	37
5.5.5	Score Comparison	37
5.6	Conclusion	37
6	Conclusion & future scope	38
6.1	Findings/observations	38

# List of Figures

2.1	Sample snapshot of the dataset showing a few entries and columns. . . . .	4
2.2	List of the columns with non-null values and their data-types . . . . .	4
2.3	Count of unique values for each column in dataset . . . . .	6
2.4	Column wise count of missing values and percentage (%) of missing values . .	7
2.5	Bar Plot: Completeness of dataset features, illustrating the count of non-missing values for each attribute. . . . .	8
2.6	Matrix Plot: Visual summary of missing values in the dataset, highlighting patterns and gaps in the data. . . . .	9
2.7	Heatmap: Correlation matrix showing relationships between numerical features in the dataset. . . . .	9
2.8	Dendrogram: Hierarchical clustering of data attributes to identify groupings based on similarities. . . . .	10
2.9	Caption . . . . .	11
3.1	Pie Chart for METHOD: Proportion of crime methods used, illustrating the distribution of different techniques such as guns, knife and others. . . . .	14
3.2	Pie Chart for OFFENSE: Distribution of crime types like theft, robbery, assault etc, with smaller offense categories grouped under 'Other' to ensure clear visualization. . . . .	14
3.3	Pie Chart for SHIFT: Percentage of crimes occurring during different shifts like day, evening and midnight, highlighting temporal crime patterns. . . . .	15
3.4	Histogram for X: Distribution of X-coordinates, representing spatial data points on the map. . . . .	16
3.5	Histogram for Y: Distribution of Y-coordinates, illustrating the spread of data points along the vertical axis. . . . .	17
3.6	Histogram for LATITUDE: Frequency distribution of latitude values, highlighting the geographic spread of crime locations. . . . .	17
3.7	Histogram for LONGITUDE: Frequency distribution of longitude values, showing the east-west geographic variation of crime locations. . . . .	18
3.8	Bar Chart for SHIFT: Frequency distribution of crimes across different shifts, indicating the time periods with higher crime occurrences. . . . .	19
3.9	Bar Chart for METHOD: Frequency of different crime methods, showing the prevalence of specific techniques used in criminal activities. . . . .	19
3.10	Bar Chart for OFFENSE: Frequency of various crime types, highlighting the most common offenses in the dataset. . . . .	20
3.11	Bar Chart for WARD: Crime frequency across different wards, revealing the areas with higher or lower crime rates. . . . .	20

3.12	Box Plot for First Set of 6 Columns: Box plot illustrating the distribution, variability, and potential outliers in the first set of six numerical columns, providing insights into their statistical properties. . . . .	21
3.13	Box Plot for Second Set of 6 Columns: Box plot showing the distribution, variability, and potential outliers for the second set of six numerical columns, helping to understand their data characteristics. . . . .	21
3.14	Geospatial distribution of crimes based on latitude and longitude, categorized by shifts of the day (Midnight, Day, and Evening), highlighting temporal crime patterns across locations . . . . .	22
3.15	Heatmap showing the correlation between spatial features (X, Y, Latitude, and Longitude), providing insights into the relationships and linear dependencies among these geographical variables. . . . .	23
3.16	Offense counts by time of day, illustrating the distribution of various crime types (e.g., Theft, Burglary) across shifts (Day, Evening, Midnight) with color intensity reflecting the frequency of occurrences. . . . .	24
4.1	Heatmap showing the correlation between spatial features (X, Y, Latitude, and Longitude), providing insights into the relationships and linear dependencies among these geographical variables. . . . .	26
4.2	Pearson correlation coefficients between pairs of variables in the dataset, indicating strong linear relationships (values close to 1) across all computed pairs. . . . .	27
5.1	Elbow method plot showing the optimal number of clusters (K) for the dataset, where the point of inflection indicates the ideal value for K, balancing within-cluster variance and cluster count . . . . .	33
5.2	Caption . . . . .	34
5.3	Visualization of the top 6 crime hotspots throughout the year, highlighting areas with the highest frequency of criminal activities and showing seasonal patterns in crime distribution, through agglomerative . . . . .	35
5.4	Visualization of the top 6 crime hotspots throughout the year, highlighting areas with the highest frequency of criminal activities and showing seasonal patterns in crime distribution, through K means . . . . .	36
5.5	Caption . . . . .	37
5.6	Caption . . . . .	37
5.7	Caption . . . . .	37
6.1	Hotspots Areas . . . . .	38
6.2	Hotspots Areas . . . . .	39
6.3	Hotspots Areas . . . . .	39
6.4	Hotspots Areas . . . . .	40
6.5	Hotspots Areas . . . . .	41
6.6	Hotspots Areas . . . . .	41
6.7	Hotspots Areas . . . . .	42
6.8	Hotspots Areas . . . . .	42



## **Abstract**

Creating safe urban spaces is essential for building sustainable and thriving communities. This project uses Exploratory Data Analysis (EDA) and advanced Machine Learning (ML) techniques to study crime data, pinpoint areas with high crime rates, and predict common types of crimes in specific locations. The goal is to provide law enforcement and policymakers with valuable insights to enhance crime prevention efforts and allocate resources more effectively.

The project starts by carefully collecting and preparing the data, addressing issues like missing values and inconsistencies to ensure its accuracy. Using Exploratory Data Analysis (EDA), we examine crime patterns across locations and time, uncovering trends and relationships that help explain why certain areas see more crime. Visual tools like heatmaps, geospatial plots, and clustering techniques are used to identify crime hotspots and understand how crimes are distributed.

Feature engineering is a key step in turning raw data into useful insights by creating attributes like location details, time-based variables, and crime severity measures. These features are then used to train machine learning models, including methods like logistic regression, decision trees, ensemble techniques, and neural networks. These models are evaluated for their ability to accurately predict high-risk areas and common crime types.

This report focuses not only on statistical accuracy and making models easy to understand but also on using geospatial analytics to gain a deeper understanding of crime patterns. The insights from this project aim to support evidence-based decisions, helping local authorities and urban planners develop proactive strategies to improve public safety and enhance the quality of life for residents. By combining EDA with predictive analytics, this project contributes to building safer and smarter cities.

# Chapter 1. Introduction

## 1.1 Project Idea and Defining the Problem

The project focuses on analyzing police and crime datasets through exploratory data analysis (EDA) to determine whether a given latitude and longitude corresponds to a crime hotspot based on historical trends. **This analysis is crucial for identifying areas prone to criminal activity, enabling policymakers to allocate resources effectively, design targeted interventions, and implement preventive measures.** By leveraging historical crime data, this study aims to contribute to proactive crime management, improve public safety strategies, and foster safer communities.

## 1.2 Data Collection

The dataset used for this project, "**Crime\_Incidents\_In\_2024**", was sourced from [Data.gov](#), an official government-managed platform that provides access to a wide range of public datasets. This dataset was downloaded in CSV format and contains detailed records of reported crime incidents for the year 2024, making it an essential resource for analyzing crime patterns and trends.

## 1.3 Dataset Description

The dataset **Crime\_Incidents\_In\_2024** provides detailed information about reported crime incidents across various locations. It contains 26,339 entries and 24 columns, each representing different aspects of the incidents and their surroundings. Key columns in the dataset include:

- **X and Y:** Coordinates of the incident location.
- **LATITUDE and LONGITUDE:** Geographical location of the crime incident.
- **REPORT\_DATE:** Date when the incident was reported.
- **SHIFT:** Indicates the time of the incident (e.g., Day, Evening, Midnight).
- **METHOD:** The method or means by which the offense was carried out.
- **OFFENSE:** Type of crime reported (e.g., Theft, Assault, etc.).
- **WARD, DISTRICT, ANC, PSA:** Administrative divisions where the crime occurred.
- **BLOCK and BLOCK\_GROUP:** The block and its group where the crime was reported.

- **NEIGHBORHOOD\_CLUSTER**: Neighborhood classification.
- **START\_DATE** and **END\_DATE**: Start and end times associated with the incident.
- **OBJECT ID**: Unique identifier for each record.

The dataset captures jurisdictional, geographical, and administrative details to provide a comprehensive view of crime patterns. Most columns serve to pinpoint the location and context of the incidents, while others help in identifying trends and categorizing crimes for deeper analysis. The dataset, with its diversity of columns and details, is essential for analyzing spatial and temporal crime trends.

## 1.4 Packages required

Below is the list of Python libraries used in the project, along with their purposes:

- **numpy**: Used for numerical computations and array operations, providing fast and efficient data manipulation.
- **pandas**: Essential for data manipulation and analysis, including handling tabular data structures like DataFrames.
- **matplotlib.pyplot**: Used for creating static, animated, and interactive visualizations to explore and present data.
- **seaborn**: A data visualization library built on top of matplotlib, used for creating informative and attractive statistical graphics.
- **missingno**: Utilized for visualizing and handling missing data to ensure data quality and consistency.
- **sklearn.impute.KNNImputer**: Implements K-Nearest Neighbors imputation to fill missing values based on similarities between data points.
- **sklearn.cluster.DBSCAN**: A clustering algorithm that groups points based on density, often used for identifying spatial crime hotspots.
- **collections.Counter**: A utility for counting and aggregating occurrences of elements, useful for analyzing categorical data.
- **sklearn.cluster.KMeans**: A clustering algorithm used to group data into a predefined number of clusters.
- **sklearn.metrics.silhouette\_score**: A metric used to evaluate the quality of clustering by measuring how well-separated clusters are.
- **geopy.geocoders.Nominatim**: A geocoding library used to convert addresses or place names into geographic coordinates.

# Chapter 2. Data Cleaning

## 2.1 Initial Dataset Analysis

Before delving into the process of cleaning and preparing the dataset, we begin with an initial analysis to understand its structure and characteristics. This includes loading the dataset, inspecting its sample entries, and analyzing the types of each column.

### 2.1.1 Loading the Dataset

The dataset was loaded into a Python environment using the following code:

```
# Loading the dataset
df = pd.read_csv('Crime_Incidents_In_2024.csv')
```

Our dataset, **Crime\_Incidents\_In\_2024.csv**, comprises **26,339 rows and 24 columns**. To determine the shape of the dataset, we used the `df.shape` attribute of the Pandas library, which revealed its dimensions.

To get a quick overview of the dataset, we utilized the `sample()` method to extract 10 random rows. This random sample provides a glimpse into the type of data contained within the dataset and helps in understanding its structure.

The `df.info()` method plays a crucial role in offering a detailed overview of the dataset. It provides critical information about the columns, including:

- **Column Names:** A list of all column headers.
- **Non-Null Counts:** The number of non-null values present in each column.
- **Data Types:** The type of data stored in each column.

The data types of the columns in the dataset are as follows:

- **int64:** Represents integer values.
- **float64:** Represents floating-point numbers.
- **object:** Represents string or categorical values.

Additionally, the `df.dtypes` attribute of Pandas provides a quick reference to the data types of all columns. This information is essential for preprocessing and analyzing the dataset effectively.

## Sample Dataset Snapshot

The following figure shows a snapshot of the first few rows of the dataset:

Index	X	Y	CCN	REPORT_DATE	SHIFT	METHOD	OFFENSE	BLOCK	XBLOCK	YBLOCK	WARD	ANC	DISTRICT	PSA	NEIGHBORHOOD_CLUSTER	BLOCK_GROUP	CENSUS_TRACT	VOTING_PRECINCT	LATITUDE	LONGITUDE
11869	398885.960000001	138792.469999999	24093871	2024/06/20 17:11:32+00	DAY	OTHERS	THEFT F/AUTO	160 - 142 BLOCK OF U STREET NW	398885.96	138792.47	5.0	SE	3.0	306.0	Cluster 21	003301.2	3301.0	Precinct 135	38.9169995582	-77.0128459293
10586	398010.149999999	138760.59	24115876	2024/07/29 08:42:34+00	MIDNIGHT	OTHERS	ROBBERY	FLORIDA AVENUE NW AND 8TH STREET NW	398010.15001211	138760.5900142	1.0	1B	3.0	305.0	Cluster 3	004402.2	4402.0	Precinct 137	38.9167108262	-77.0229447497
420	402808.880000003	132407.5	24134117	2024/08/31 18:00:56+00	DAY	OTHERS	THEFT/OTHER	2700 - 2845 BLOCK OF ALABAMA AVENUE SE	402808.88	132407.5	7.0	7B	6.0	606.0	Cluster 35	007603.2	7603.0	Precinct 113	38.8594775631	-76.9676370731
2750	403259.9463	131832.6677	24088099	2024/06/10 16:19:01+00	DAY	OTHERS	THEFT F/AUTO	2800 - 2899 BLOCK OF 31ST STREET SE	403259.946336978	131832.667671771	7.0	7B	6.0	606.0	Cluster 35	007603.4	7603.0	Precinct 113	38.8542976731	-76.9624427708
1773	398747.163000003	134091.609999999	24076231	2024/05/21 16:40:23+00	DAY	OTHERS	THEFT/OTHER	1290 - 1399 BLOCK OF DELAWARE AVENUE SW	398747.16300479	134091.609113868	6.0	6D	1.0	105.0	Cluster 9	011001.3	11001.0	Precinct 127	38.8746523013	-77.0144378163
5075	399074.829999998	138550.809999999	24133244	2024/08/29 20:41:46+00	EVENING	OTHERS	THEFT/OTHER	1 - 99 BLOCK OF SEATON PLACE NW	399074.83	138550.81	5.0	SE	3.0	308.0	Cluster 21	003302.1	3302.0	Precinct 19	38.9148228237	-77.0106677549
4245	399364.75	138021.27	24010049	2024/01/21 07:23:50+00	MIDNIGHT	OTHERS	MOTOR VEHICLE THEFT	21 - 72 BLOCK OF FLORIDA AVENUE NE	399364.75	138021.27	5.0	5F	5.0	502.0	Cluster 21	006701.1	6701.0	Precinct 75	38.9100528057	-77.0073243167
12639	399996.229999997	137052.260000002	24424058	2024/11/23 22:00:13+00	EVENING	OTHERS	THEFT F/AUTO	400 - 499 BLOCK OF I STREET NE	399996.23	137052.26	6.0	6C	1.0	104.0	Cluster 25	010602.3	10602.0	Precinct 83	38.9013238548	-77.0000434621
22767	400437.009999998	135275.25	24031036	2024/02/29 01:51:29+00	EVENING	OTHERS	ROBBERY	300 - 339 BLOCK OF 6TH STREET SE	400437.01	135275.25	6.0	6B	1.0	107.0	Cluster 26	006700.3	6700.0	Precinct 89	38.8853157841	-76.9949631013
12222	397163.079999998	142085.399999999	24101321	2024/07/03 02:47:33+00	EVENING	OTHERS	THEFT/OTHER	4600 - 4699 BLOCK OF 14TH STREET NW	397163.08	142085.4	4.0	4E	4.0	404.0	Cluster 18	002501.2	2501.0	Precinct 48	38.946594034	-77.0327259674

Figure 2.1: Sample snapshot of the dataset showing a few entries and columns.

## Dataset Overview

Below is the detailed overview of the dataset, including the column names, non-null counts, and data types:

#	Column	Non-Null Count	Dtype
0	X	26339 non-null	float64
1	Y	26339 non-null	float64
2	CCN	26339 non-null	int64
3	REPORT_DATE	26339 non-null	object
4	SHIFT	26339 non-null	object
5	METHOD	26339 non-null	object
6	OFFENSE	26339 non-null	object
7	BLOCK	26339 non-null	object
8	XBLOCK	26339 non-null	float64
9	YBLOCK	26339 non-null	float64
10	WARD	26335 non-null	float64
11	ANC	26335 non-null	object
12	DISTRICT	25605 non-null	float64
13	PSA	25886 non-null	float64
14	NEIGHBORHOOD_CLUSTER	26335 non-null	object
15	BLOCK_GROUP	26326 non-null	object
16	CENSUS_TRACT	26326 non-null	float64
17	VOTING_PRECINCT	26335 non-null	object
18	LATITUDE	26339 non-null	float64
19	LONGITUDE	26339 non-null	float64
20	BID	4917 non-null	object
21	START_DATE	26334 non-null	object
22	END_DATE	24447 non-null	object
23	OBJECTID	26339 non-null	int64

dtypes: float64(10), int64(2), object(12)

Figure 2.2: List of the columns with non-null values and their data-types

## Float Data Types

The following columns contain floating-point values, representing numbers with decimal points:

- **X** – X-coordinate of the location (float64)
- **Y** – Y-coordinate of the location (float64)

- **XBLOCK** – X-coordinate of the block (float64)
- **YBLOCK** – Y-coordinate of the block (float64)
- **WARD** – Ward number, some missing values (float64)
- **DISTRICT** – Police district, some missing values (float64)
- **PSA** – Police service area, some missing values (float64)
- **CENSUS\_TRACT** – Census tract identifier, some missing values (float64)
- **LATITUDE** – Latitude of the incident location (float64)
- **LONGITUDE** – Longitude of the incident location (float64)

These columns represent spatial and geographic information, such as coordinates of crime locations and administrative district boundaries.

## Integer Data Types

The following columns contain integer values:

- **CCN** – Crime Case Number (int64)
- **OBJECTID** – Unique identifier for the record (int64)

These columns are identifiers and numerical references that uniquely identify records or crime cases.

## Object (String) Data Types

The following columns contain string (object) values, representing categorical or textual information:

- **REPORT\_DAT** – Date and time when the crime was reported (object)
- **SHIFT** – Shift during which the crime occurred (object)
- **METHOD** – Method of the crime (object)
- **OFFENSE** – Type of offense (object)
- **BLOCK** – Block where the crime occurred (object)
- **ANC** – ANC district, some missing values (object)
- **NEIGHBORHOOD\_CLUSTER** – Neighborhood cluster, some missing values (object)
- **BLOCK\_GROUP** – Block group identifier (object)
- **VOTING\_PRECINCT** – Voting precinct (object)
- **BID** – Business improvement district (object), some missing values

- **START\_DATE** – Start date of the incident (object)
- **END\_DATE** – End date of the incident (object)

These columns represent categorical and descriptive data, such as crime types, locations, times, and identifiers for various administrative regions.

	0
<b>X</b>	6733
<b>Y</b>	6758
<b>CCN</b>	26332
<b>REPORT_DAT</b>	26287
<b>SHIFT</b>	3
<b>METHOD</b>	3
<b>OFFENSE</b>	9
<b>BLOCK</b>	6837
<b>XBLOCK</b>	6760
<b>YBLOCK</b>	6780
<b>WARD</b>	8
<b>ANC</b>	46
<b>DISTRICT</b>	7
<b>PSA</b>	57
<b>NEIGHBORHOOD_CLUSTER</b>	45
<b>BLOCK_GROUP</b>	569
<b>CENSUS_TRACT</b>	206
<b>VOTING_PRECINCT</b>	144
<b>LATITUDE</b>	6847
<b>LONGITUDE</b>	6847
<b>BID</b>	12
<b>START_DATE</b>	23724
<b>END_DATE</b>	21959
<b>OBJECTID</b>	26339
<b>OCTO_RECORD_ID</b>	0

**dtype:** int64

Figure 2.3: Count of unique values for each column in dataset

This overview highlights the number of entries, column types, providing a foundation for further cleaning and analysis.

## 2.2 Missing data analysis

Missing data is a frequent issue in datasets, and it significantly affects the quality of our analysis. Before any data processing, it is crucial to understand the amount and proportion of missing data. In the **Crime Incidents in 2024** dataset, several columns have missing values, with varying proportions of missing data.

Missing Values:		Missing Values:	
X	0	X	0
Y	0	Y	0
CCN	0	CCN	0
REPORT_DAT	0	REPORT_DAT	0
SHIFT	0	SHIFT	0
METHOD	0	METHOD	0
OFFENSE	0	OFFENSE	0
BLOCK	0	BLOCK	0
XBLOCK	0	XBLOCK	0
YBLOCK	0	YBLOCK	0
WARD	4	WARD	4
ANC	4	ANC	4
DISTRICT	734	DISTRICT	734
PSA	453	PSA	453
NEIGHBORHOOD_CLUSTER	4	NEIGHBORHOOD_CLUSTER	4
BLOCK_GROUP	13	BLOCK_GROUP	13
CENSUS_TRACT	13	CENSUS_TRACT	13
VOTING_PRECINCT	4	VOTING_PRECINCT	4
LATITUDE	0	LATITUDE	0
LONGITUDE	0	LONGITUDE	0
BID	21422	BID	21422
START_DATE	5	START_DATE	5
END_DATE	1892	END_DATE	1892
OBJECTID	0	OBJECTID	0
dtype: int64		dtype: int64	

Figure 2.4: Column wise count of missing values and percentage (%) of missing values

The column **BID** has the highest percentage of missing values, accounting for **81.33%**. Other columns with significant missing data include **END\_DATE** (7.18%), **DISTRICT** (2.79%), and **PSA** (1.72%). On the other hand, columns such as **X**, **Y**, **CCN**, and **REPORT\_DAT** have no missing values. A few columns, like **WARD**, **ANC**, and **VOTING\_PRECINCT**, have less than 0.02% missing values, indicating that most of the data is intact.



## 2.2.1 Visualization of Missing Values

To better understand the distribution of missing values in the dataset, we utilized several visualization techniques provided by the `missingno` library. Visualizations are an effective way to identify patterns or structures in missing data, which can help inform our data cleaning strategies. We employed the following visualizations:

- **Barplot:** A barplot was used to visualize the total number of missing values for each column. This plot helps to quickly identify columns with significant amounts of missing data, which could be crucial for deciding whether to impute, drop, or retain these columns.
- **Matrix:** The missing data matrix provided a clear, compact view of the missing values in each row and column. The matrix displayed rows as horizontal lines and columns as vertical lines, with missing values represented by black bars. This helps in understanding how missingness is distributed across both rows and columns.
- **Dendrogram:** The dendrogram illustrated the hierarchical relationships between columns based on their missing value patterns. This visualization can be helpful for identifying clusters of columns with similar missingness profiles, which could guide decisions on how to handle missing data across related columns.
- **Heatmap:** The heatmap displayed the missing data as a grid, with missing values highlighted in a distinct color. This allows for a detailed view of missing values across the dataset and can highlight patterns or correlations between missingness and other variables.

Below are the plots showing the missing values in our dataset. These visualizations offer valuable insights into the structure and extent of missing data, providing a foundation for deciding on appropriate data imputation or handling techniques.

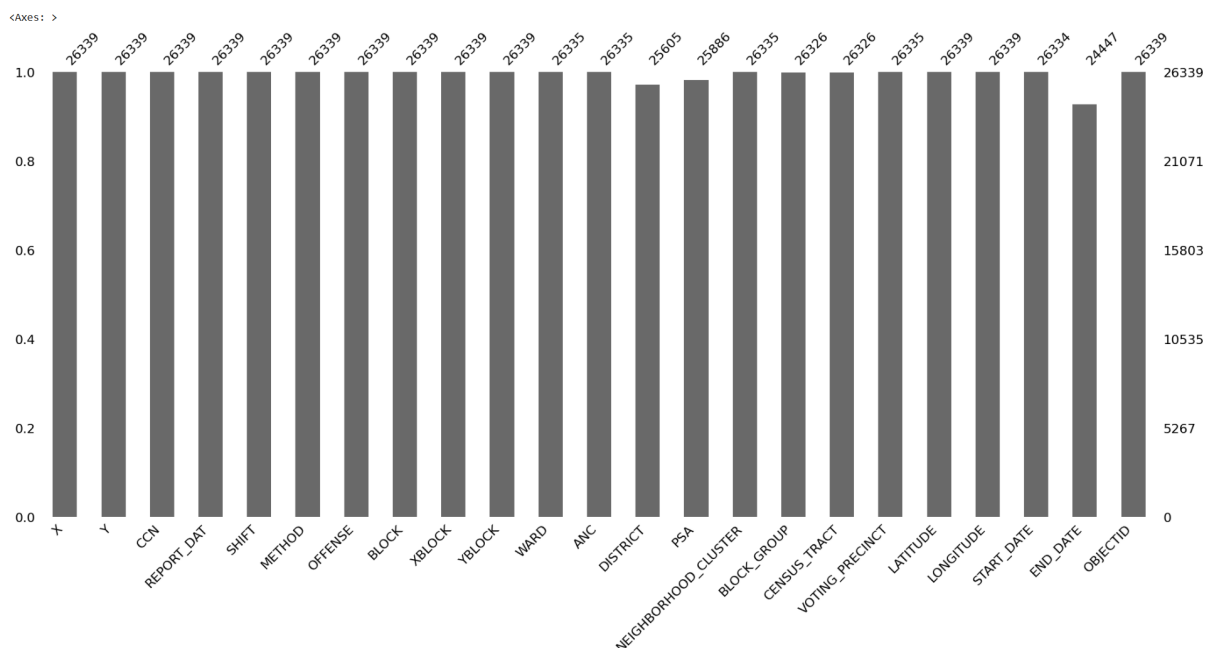


Figure 2.5: Bar Plot: Completeness of dataset features, illustrating the count of non-missing values for each attribute.

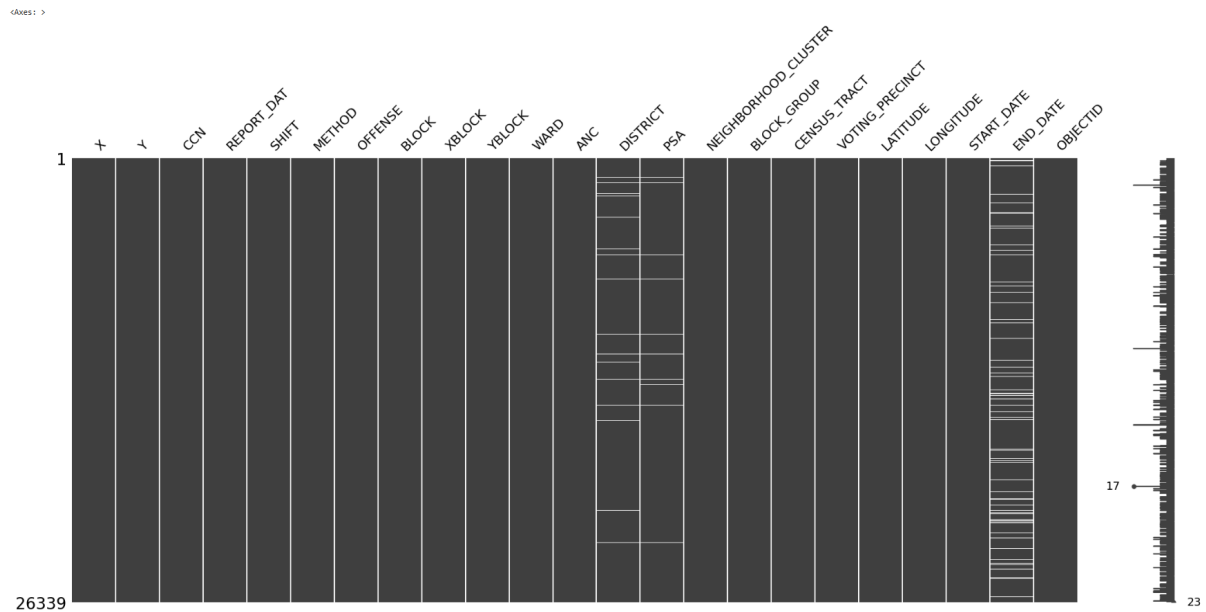


Figure 2.6: Matrix Plot: Visual summary of missing values in the dataset, highlighting patterns and gaps in the data.

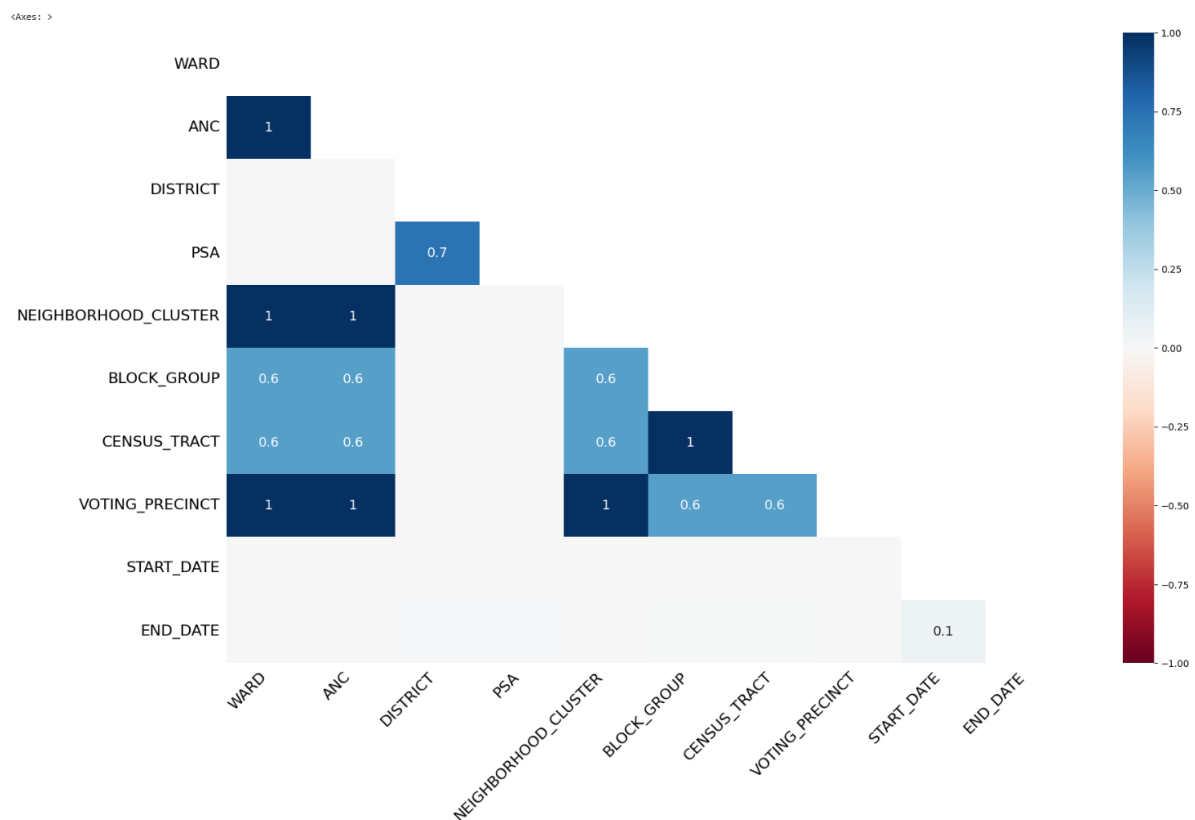


Figure 2.7: Heatmap: Correlation matrix showing relationships between numerical features in the dataset.

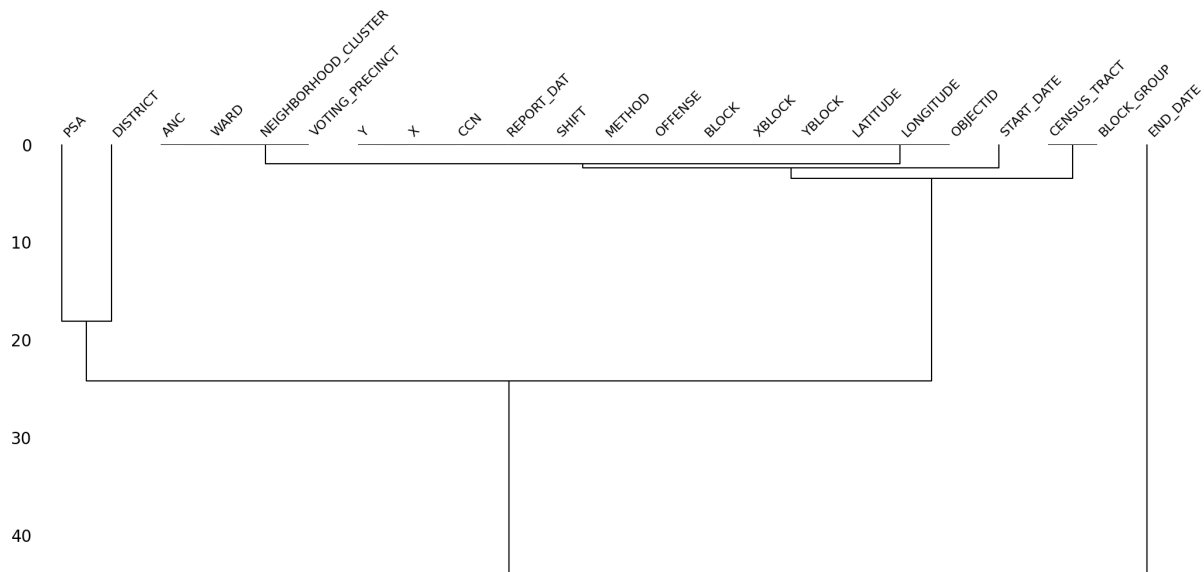


Figure 2.8: Dendrogram: Hierarchical clustering of data attributes to identify groupings based on similarities.

### 2.2.2 Type of Missingness in Our Dataset

When dealing with missing data, it is important to understand the type of missingness, as this influences how we handle the missing values. There are generally three types of missingness mechanisms in data:

- **Missing Completely at Random (MCAR):** Missingness is unrelated to any observed or unobserved data in the dataset. In other words, the missing values are distributed randomly across all data points, and there is no systematic pattern to the missingness.
- **Missing at Random (MAR):** Missingness is related to observed data but not to the missing data itself. In this case, the presence of missing values can be explained by other observed variables in the dataset. However, the probability of missingness does not depend on the unobserved values.
- **Missing Not at Random (MNAR):** Missingness is related to the unobserved data itself. That is, the probability of a value being missing depends on the value of the variable that is missing. This type of missingness is the most challenging to handle because the missing data could introduce bias into the analysis.

To determine the type of missingness in our dataset, we performed the Little's MCAR test, which is a statistical test used to assess whether the missing data mechanism is MCAR. The null hypothesis of the Little MCAR test is that the data is missing completely at random, while the alternative hypothesis suggests that the data is not MCAR (i.e., the missingness is at least partially explained by other variables in the dataset).

The Little MCAR test provides a p-value as the result of the hypothesis test. If the p-value is large (typically greater than 0.05), we fail to reject the null hypothesis, indicating that the missing data mechanism is MCAR. If the p-value is small (typically less than 0.05), we reject the null hypothesis and conclude that the missing data mechanism is not MCAR.

**Test Parameters:**

- **Null Hypothesis ( $H_0$ ):** The missing data is missing completely at random (MCAR).
- **Alternative Hypothesis ( $H_a$ ):** The missing data is not missing completely at random (i.e., it could be MAR or MNAR).
- **Test Statistic:** The test statistic is based on a comparison of the observed missing data patterns to the expected patterns under the assumption that the data is MCAR.
- **p-value:** A p-value greater than 0.05 supports the null hypothesis, indicating that the missing data is MCAR.

#### Results of the Little MCAR Test:

Upon performing the Little MCAR test on our dataset, the result returned a p-value of 1, which is greater than the typical threshold of 0.05. This means we fail to reject the null hypothesis and conclude that the missing data in our dataset is **Missing Completely at Random (MCAR)**.

```
Little's MCAR Test Result:
{'chi_square_stat': 1.2524175722706717e-25, 'degrees_of_freedom': 65, 'p_value': 1.0, 'is_mcar': True}
```

Figure 2.9: Caption

Since the data is MCAR, we can confidently proceed with standard techniques for handling missing data, such as imputation or deletion, without introducing significant bias into the analysis.

## 2.3 Imputation

After identifying and analyzing the missing data, the next crucial step in the data cleaning process is imputation, which refers to filling in the missing values with appropriate substitutes. Imputation helps ensure that our analysis can proceed smoothly without being hindered by the absence of critical information.

As previously mentioned, the missingness in our dataset is classified as **Missing Completely at Random (MCAR)**. This type of missingness suggests that the missing data points are not related to any specific variable or pattern, making the imputation process more straightforward. Given this, we can confidently apply imputation techniques that will not introduce significant bias or distortion into the analysis.

### 2.3.1 Imputation Strategy

We employed different strategies for handling missing data based on the type of variable:

- **Dropping the 'BID' Column:** The 'BID' column has a substantial proportion of missing data—approximately 81% of its values are missing. Such a high percentage of missingness makes it impractical to impute values, as it would introduce a lot of uncertainty. Therefore, we chose to drop the 'BID' column entirely, as its inclusion would not add significant value to our analysis.

- **Imputing Numeric Columns Using the Mean:** For the numerical columns in the dataset (such as 'WARD', 'DISTRICT', 'PSA', etc.), we applied **mean imputation**. Mean imputation involves replacing missing values in a column with the mean (average) of the existing values in that column. This method is simple and effective, especially when the data is MCAR, as it doesn't introduce any bias. However, it assumes that the missing values are randomly distributed and that the data is approximately normally distributed. For example, missing values in columns like 'WARD' or 'DISTRICT' were replaced with the mean value of those columns.
- **Imputing Categorical Columns Using the Mode:** For categorical variables (such as 'SHIFT', 'METHOD', 'OFFENSE', etc.), we used **mode imputation**, where the missing values are replaced with the most frequent category (the mode) in the column. This method is appropriate for MCAR data because the mode represents the most common or likely value, and substituting missing values with the mode helps maintain the consistency of the dataset. For instance, missing values in the 'SHIFT' column were replaced with the most frequent shift type in the dataset.

By applying these imputation methods, we were able to handle missing data effectively without losing valuable information from the dataset. This ensures that our analysis remains robust and that the imputed values are logically consistent with the rest of the data.

In conclusion, after imputation, our dataset is now complete and ready for further analysis and modeling.

# Chapter 3. Visualization

Visualization plays a crucial role in understanding and interpreting data. It allows us to identify patterns, trends, and outliers that may not be immediately apparent in raw data. By transforming complex datasets into graphical representations, we can effectively communicate insights, highlight relationships between variables, and support data-driven decision-making. In this chapter, we focus on using various visualization techniques to explore the cleaned dataset, uncover underlying structures, and present the findings in an easily digestible format. Through these visualizations, we aim to gain a deeper understanding of the dataset and provide a clear foundation for further analysis and modeling.

## 3.1 Univariate analysis

Univariate analysis involves the examination of the distribution of a single variable. It helps us understand the individual characteristics and frequency distribution of each variable in the dataset. Various graphical representations, such as histograms, bar charts, and pie charts, can be used to summarize and explore the data. In this section, we begin by visualizing some of the key columns of the dataset using the following methods:

1. Pie chart
2. Histogram
3. Bar chart
4. Box plot

### 3.1.1 Pie Chart

A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions. Each slice represents a category's contribution to the whole, making it easy to visualize the relative sizes of different categories. Pie charts are particularly useful when we want to show the percentage distribution of a categorical variable.

Below is the pie chart representing the distribution of values in some of the columns of our dataset:

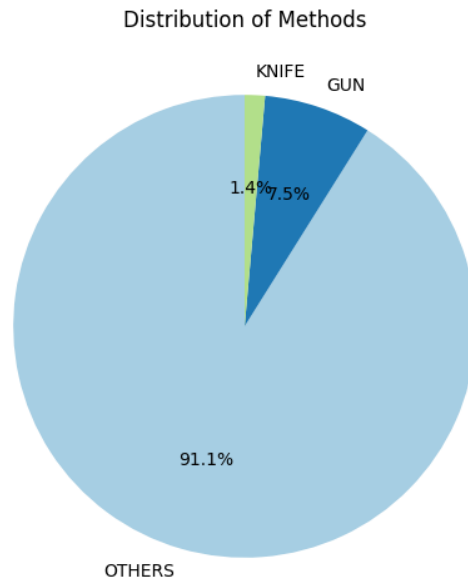


Figure 3.1: Pie Chart for METHOD: Proportion of crime methods used, illustrating the distribution of different techniques such as guns, knife and others.

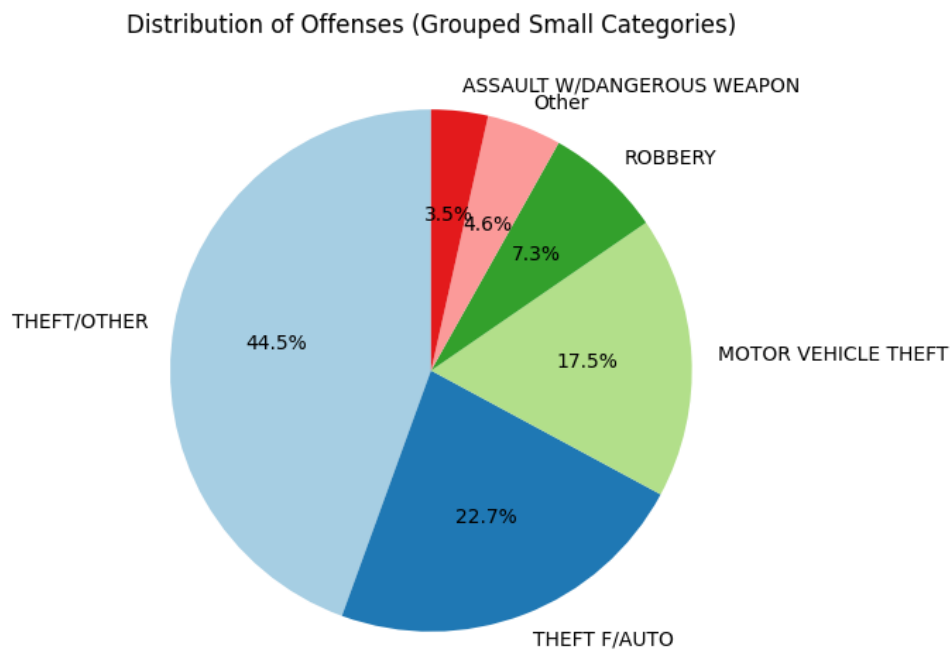


Figure 3.2: Pie Chart for OFFENSE: Distribution of crime types like theft, robbery, assault etc, with smaller offense categories grouped under 'Other' to ensure clear visualization.

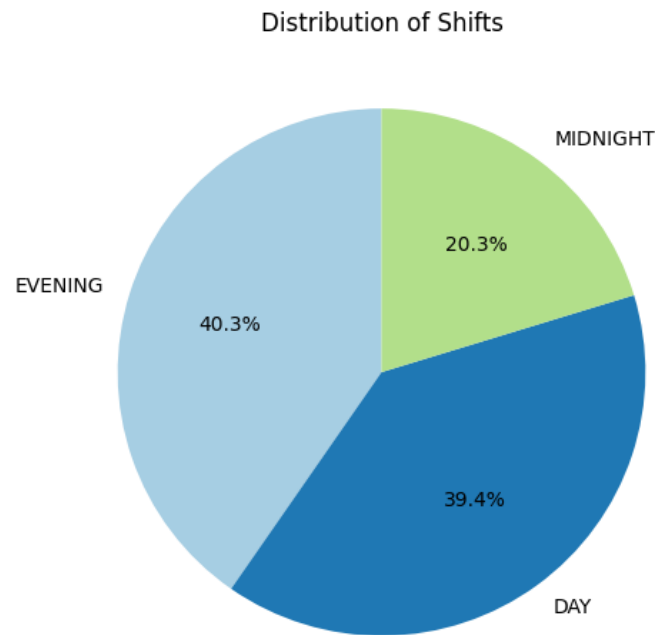


Figure 3.3: Pie Chart for SHIFT: Percentage of crimes occurring during different shifts like day, evening and midnight, highlighting temporal crime patterns.

### 3.1.2 Histogram

A histogram is a graphical representation used to visualize the distribution of a continuous variable. It divides the data into intervals, known as bins, and displays the frequency of data points within each bin using vertical bars. The height of each bar represents the count of data points in that range, allowing us to easily identify patterns, such as skewness, spread, and the presence of any peaks (modes) in the data.

Below are histograms that represent the distribution of values for selected continuous variables in our dataset:



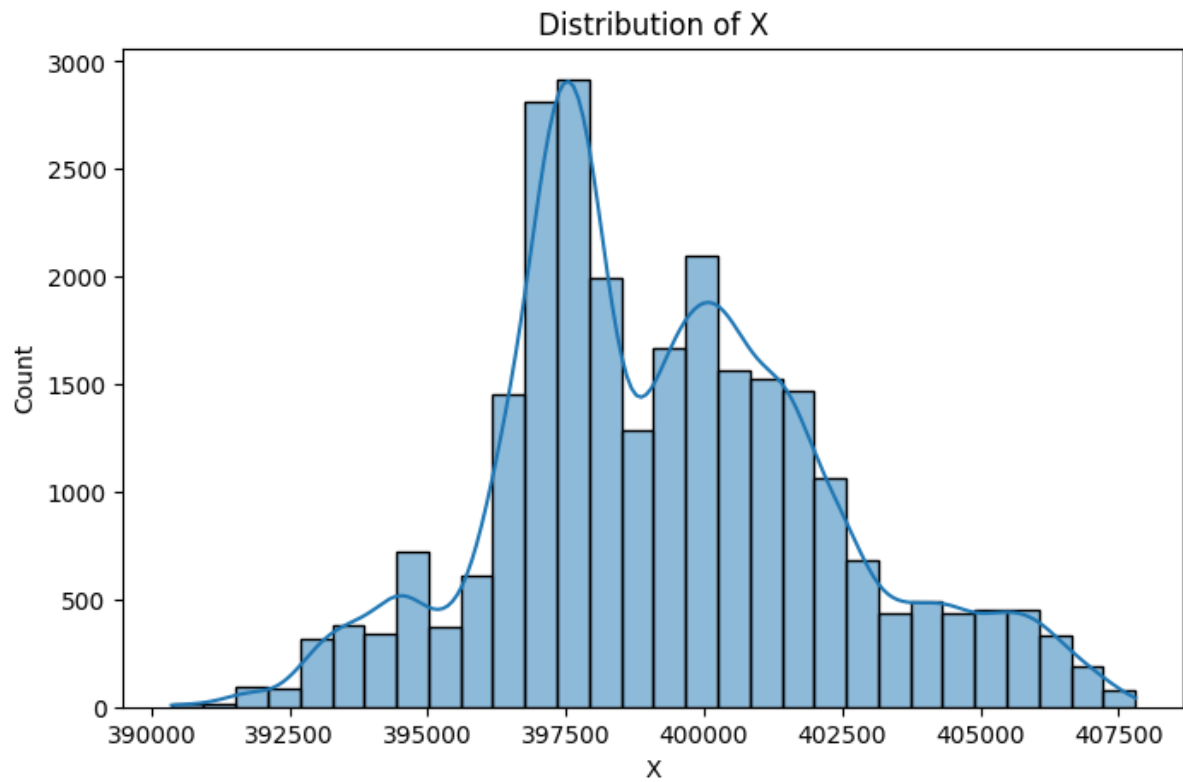


Figure 3.4: Histogram for X: Distribution of X-coordinates, representing spatial data points on the map.

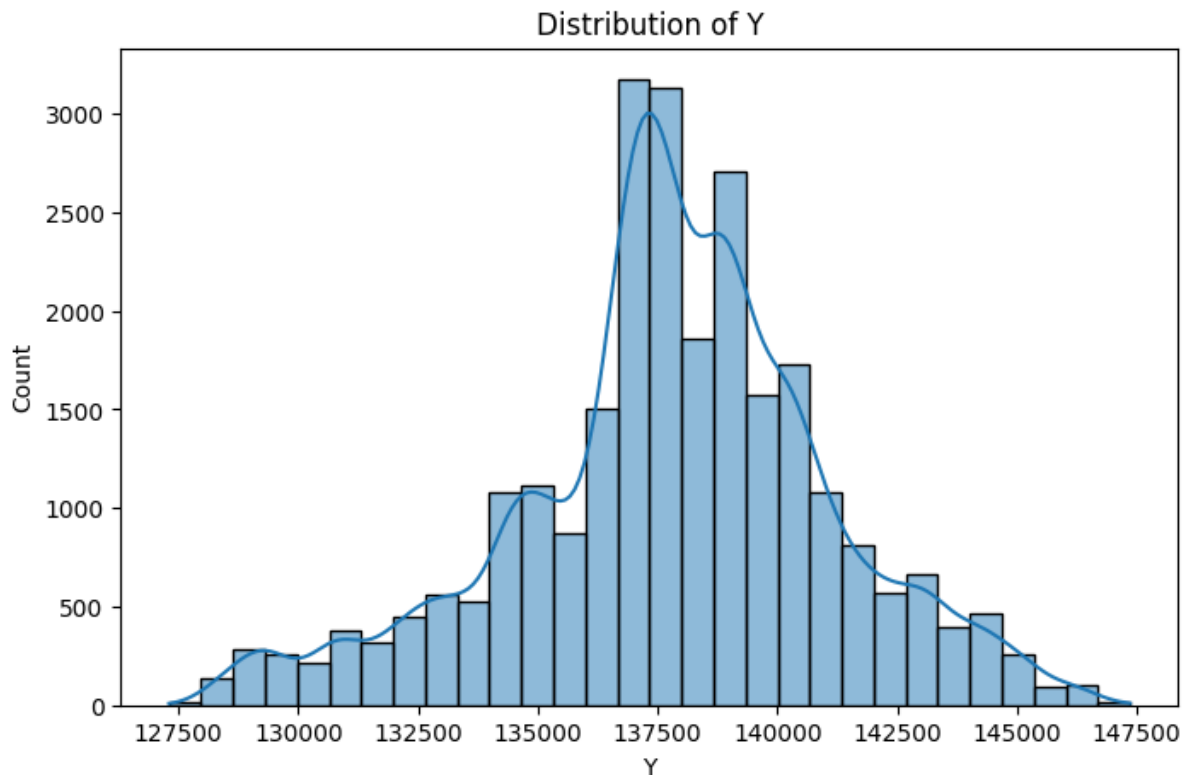


Figure 3.5: Histogram for Y: Distribution of Y-coordinates, illustrating the spread of data points along the vertical axis.

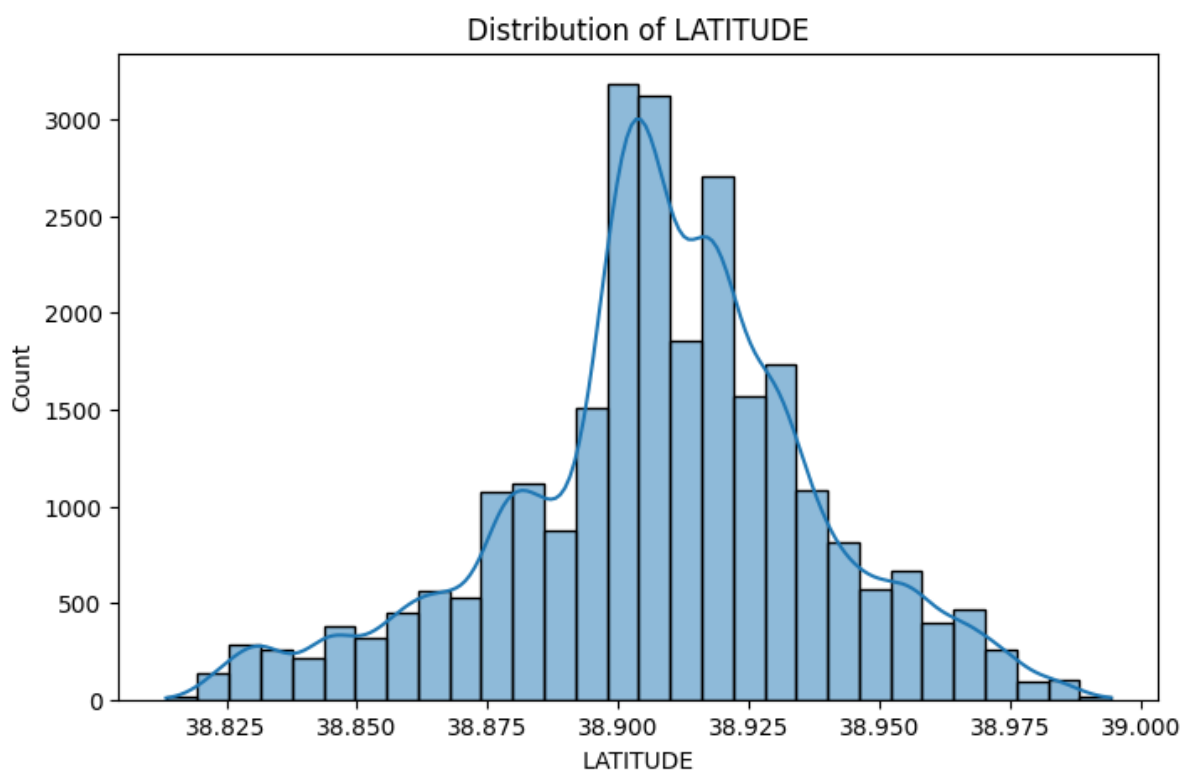


Figure 3.6: Histogram for LATITUDE: Frequency distribution of latitude values, highlighting the geographic spread of crime locations.

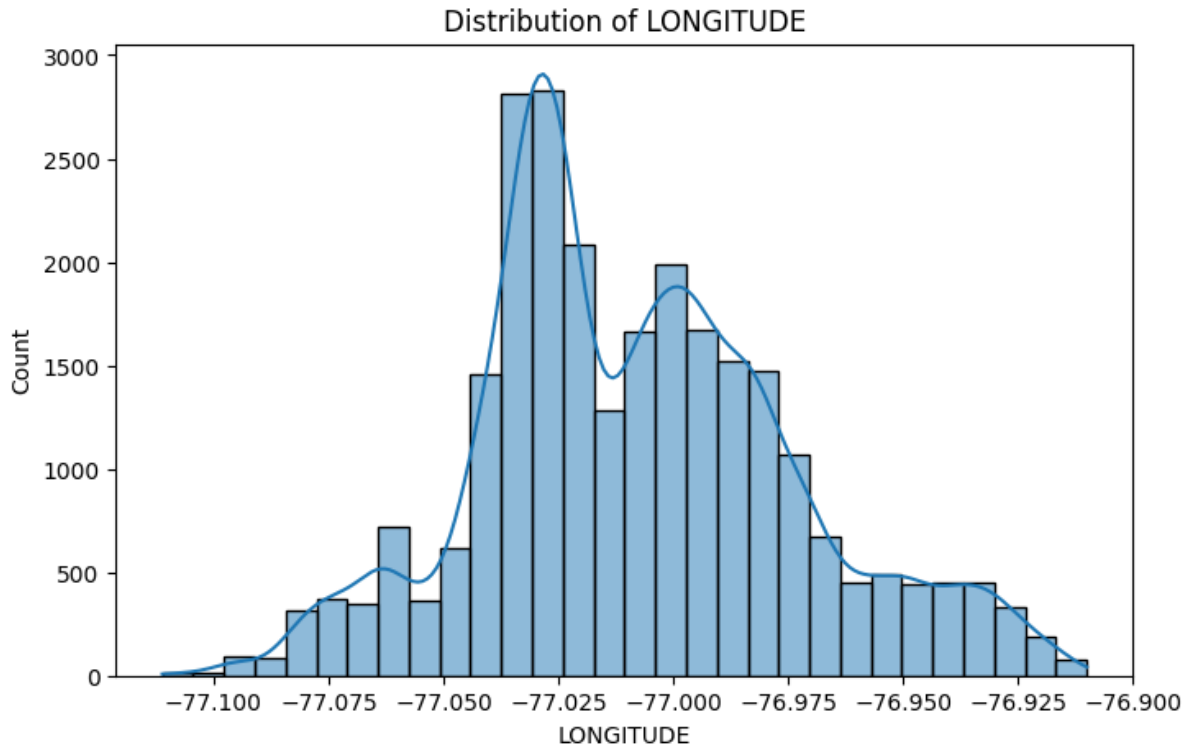


Figure 3.7: Histogram for LONGITUDE: Frequency distribution of longitude values, showing the east-west geographic variation of crime locations.

### 3.1.3 Bar Chart

A bar chart is a graphical representation used to display and compare the frequency or count of categorical data. It consists of rectangular bars, where the length of each bar is proportional to the value it represents. Bar charts make it easy to compare different categories and identify trends or patterns in the data.

Below are bar charts representing the frequency distribution of selected categorical variables in our dataset:

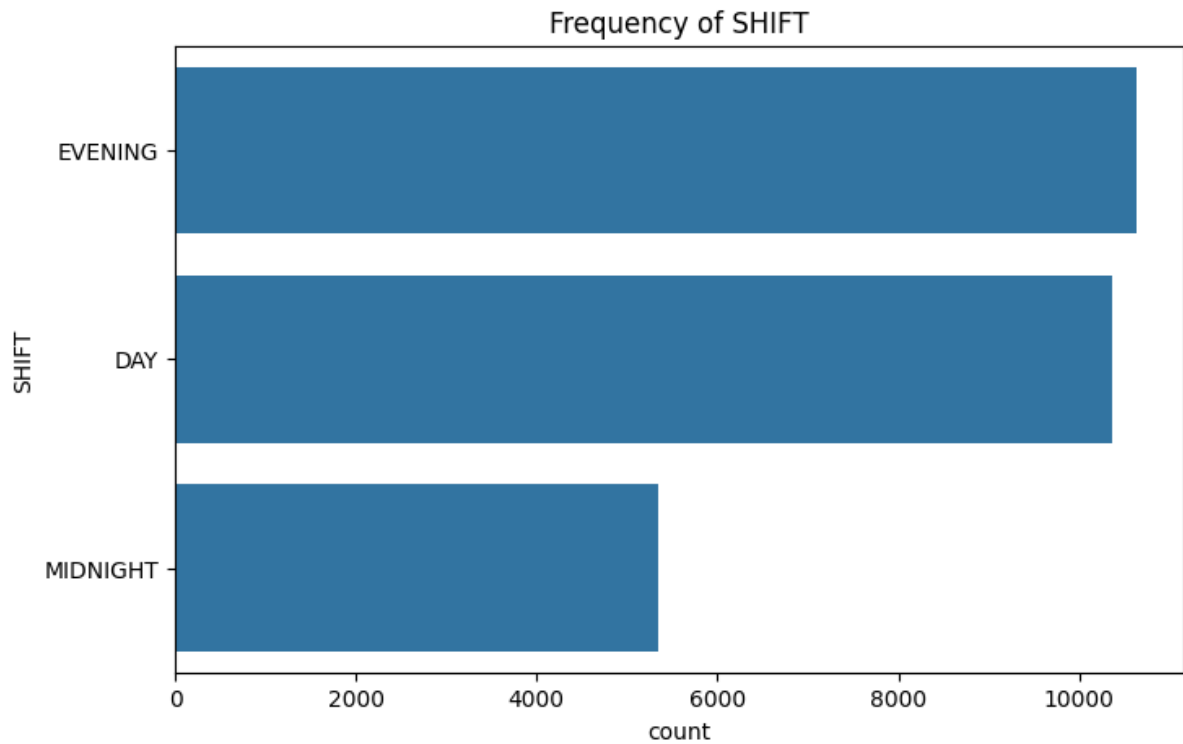


Figure 3.8: Bar Chart for SHIFT: Frequency distribution of crimes across different shifts, indicating the time periods with higher crime occurrences.

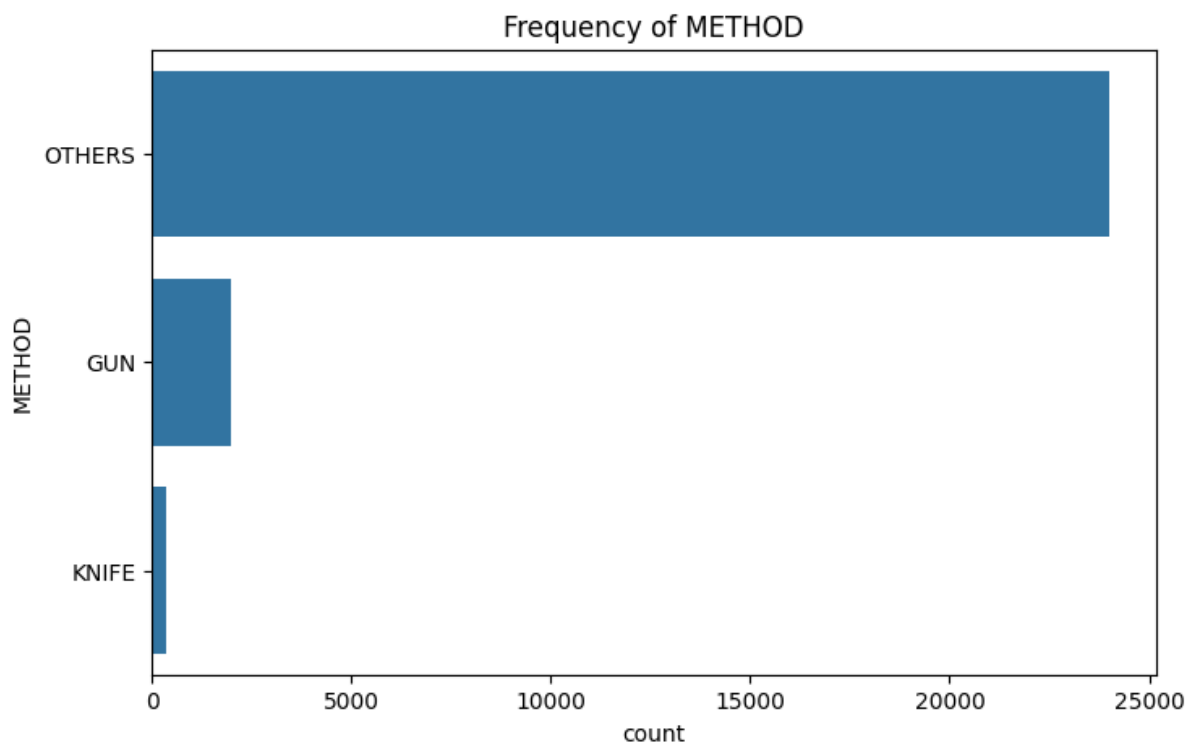


Figure 3.9: Bar Chart for METHOD: Frequency of different crime methods, showing the prevalence of specific techniques used in criminal activities.

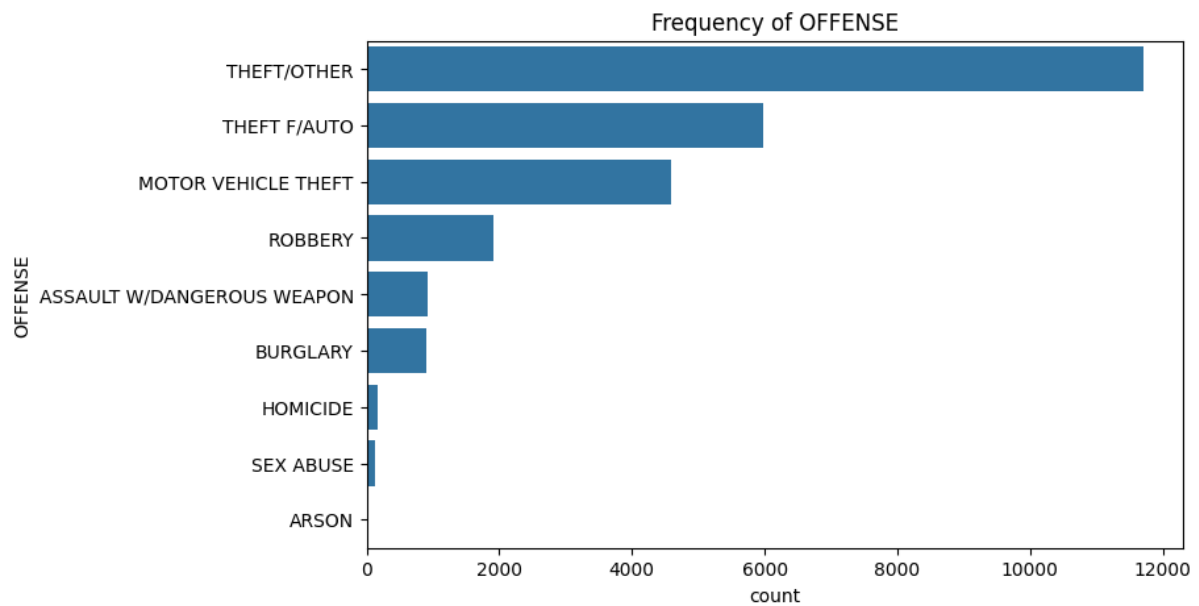


Figure 3.10: Bar Chart for OFFENSE: Frequency of various crime types, highlighting the most common offenses in the dataset.

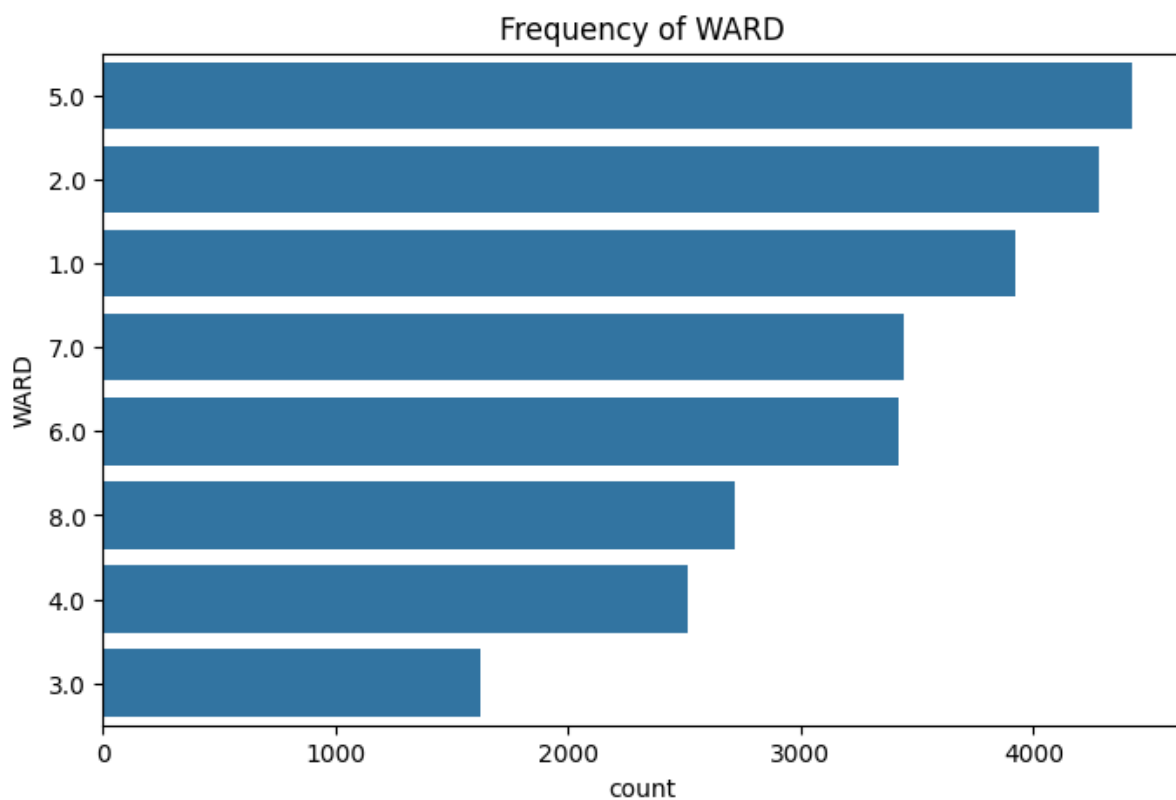


Figure 3.11: Bar Chart for WARD: Crime frequency across different wards, revealing the areas with higher or lower crime rates.

### 3.1.4 Box Plot

A box plot, also known as a box-and-whisker plot, is a statistical graphic used to summarize the distribution of a dataset. It displays the dataset's minimum, first quartile (Q1), median, third quartile (Q3), and maximum, providing a clear visualization of the data's spread and central tendency. The "box" represents the interquartile range (IQR), while the "whiskers" extend to the minimum and maximum values within a specified range. Outliers are often plotted as individual points. Box plots are particularly useful for identifying skewness, variability, and outliers in the data.

Below are box plots illustrating the distribution of selected continuous variables in our dataset:

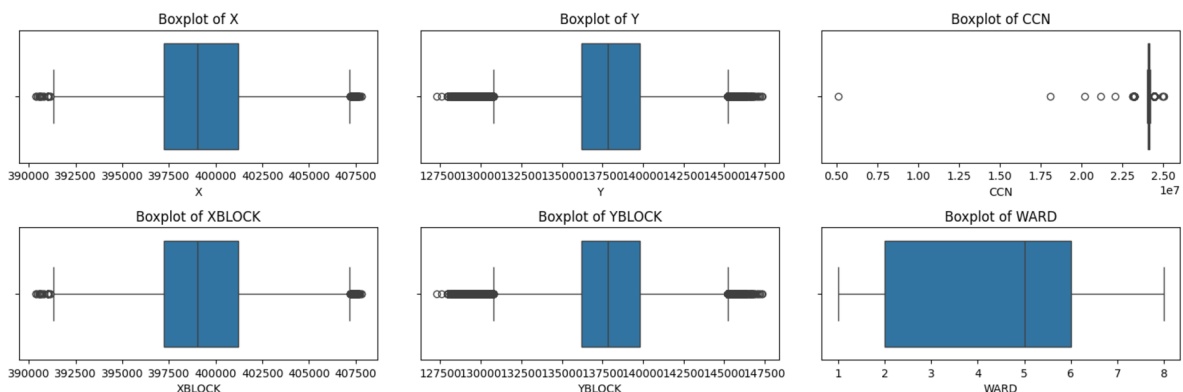


Figure 3.12: Box Plot for First Set of 6 Columns: Box plot illustrating the distribution, variability, and potential outliers in the first set of six numerical columns, providing insights into their statistical properties.

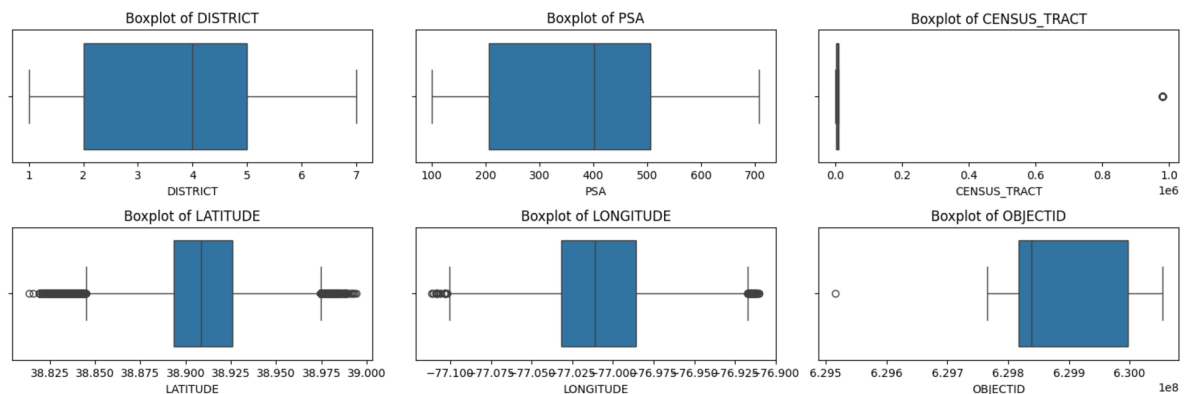


Figure 3.13: Box Plot for Second Set of 6 Columns: Box plot showing the distribution, variability, and potential outliers for the second set of six numerical columns, helping to understand their data characteristics.

## 3.2 Multivariate Analysis

Multivariate analysis involves the examination of relationships and interactions between two or more variables. This type of analysis is crucial for understanding how variables influence each other and can help uncover patterns, correlations, or trends within the dataset. It allows us

to explore the combined effect of multiple variables, which is often essential for building more accurate models and making better predictions.

In this section, we focus on visualizing the relationships between various pairs of variables using different methods such as scatter plots, pair plots, and correlation heatmaps. These techniques help in detecting any significant interactions, patterns, or correlations in the dataset.

The key techniques for multivariate analysis include:

- Scatter Plot
- Correlation Heatmap

### 3.2.1 Scatter Plot

A scatter plot is a type of data visualization used to represent the relationship between two continuous variables. Each point on the plot corresponds to one observation in the dataset, with the position determined by its values for the two variables. Scatter plots are useful for identifying correlations, trends, and outliers in the data.

In the context of our analysis, the scatter plot shown below visualizes the relationship between **Latitude** and **Longitude**. The points are colored based on the **Shift** attribute, which represents different time periods (Midnight, Day, Evening). This allows us to visually inspect if there are any spatial patterns or clustering of incidents based on the time of occurrence. The plot also helps in observing any potential correlations between the geographical locations and the time of day when crimes occurred.

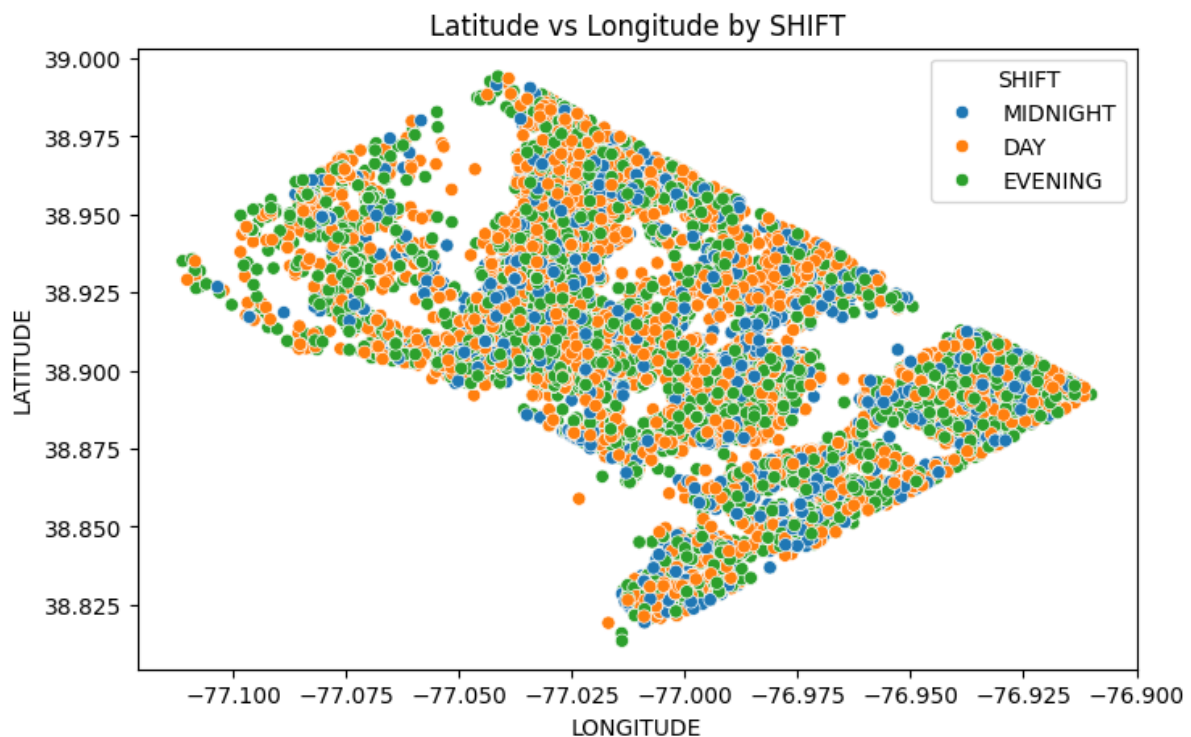


Figure 3.14: Geospatial distribution of crimes based on latitude and longitude, categorized by shifts of the day (Midnight, Day, and Evening), highlighting temporal crime patterns across locations

### 3.2.2 Correlation Heatmap

The Correlation Heatmap visualizes the correlation matrix between different variables in the dataset. It represents the strength of the linear relationships between variables through color intensity. In this heatmap, a red color indicates a strong positive correlation (with a value of 1.00), while a blue color indicates a negative correlation (with values closer to -1). The heatmap is useful for identifying relationships between variables such as X, Y, Latitude, and Longitude, and can help in understanding which features are highly correlated. Below is the correlation heatmap for the selected variables.

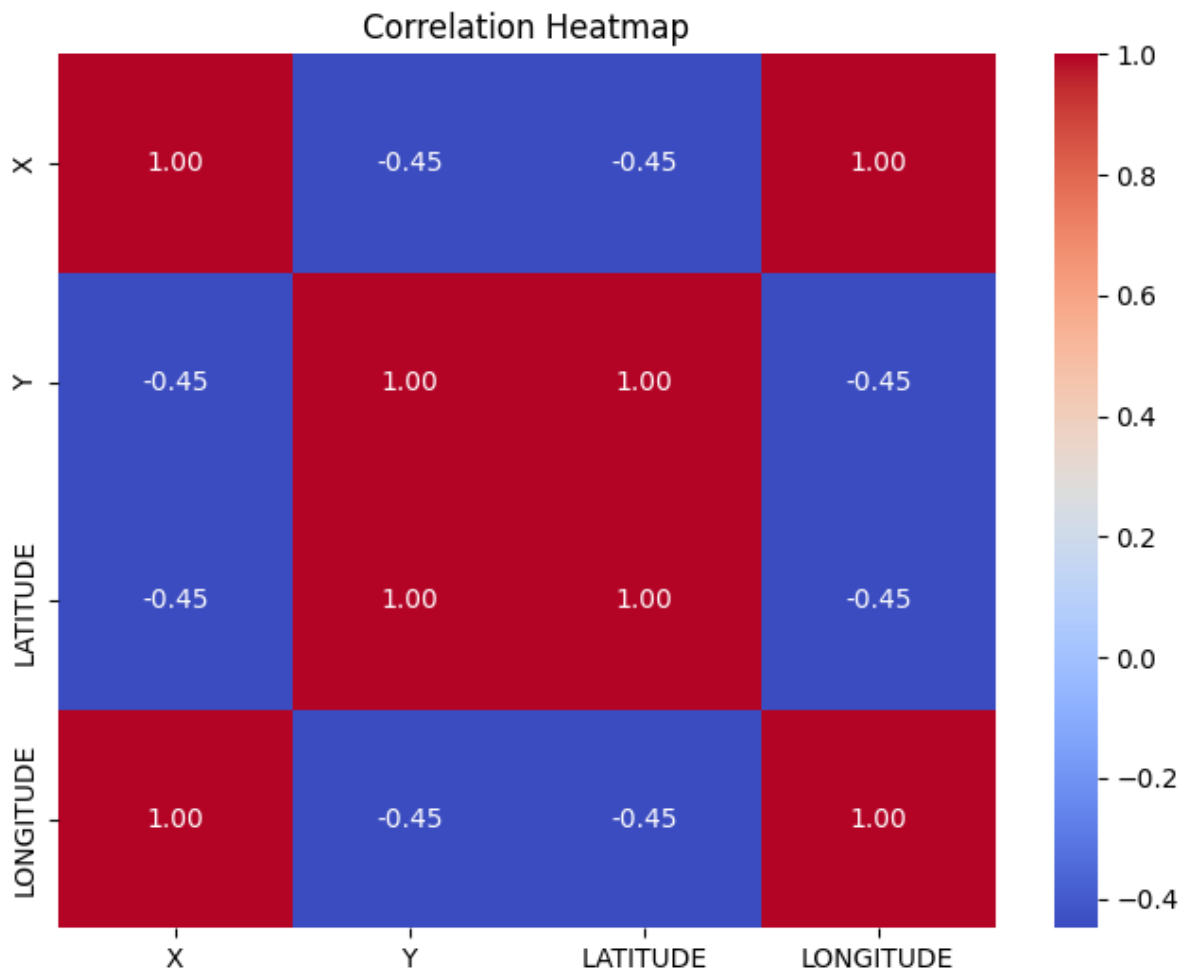


Figure 3.15: Heatmap showing the correlation between spatial features (X, Y, Latitude, and Longitude), providing insights into the relationships and linear dependencies among these geographical variables.



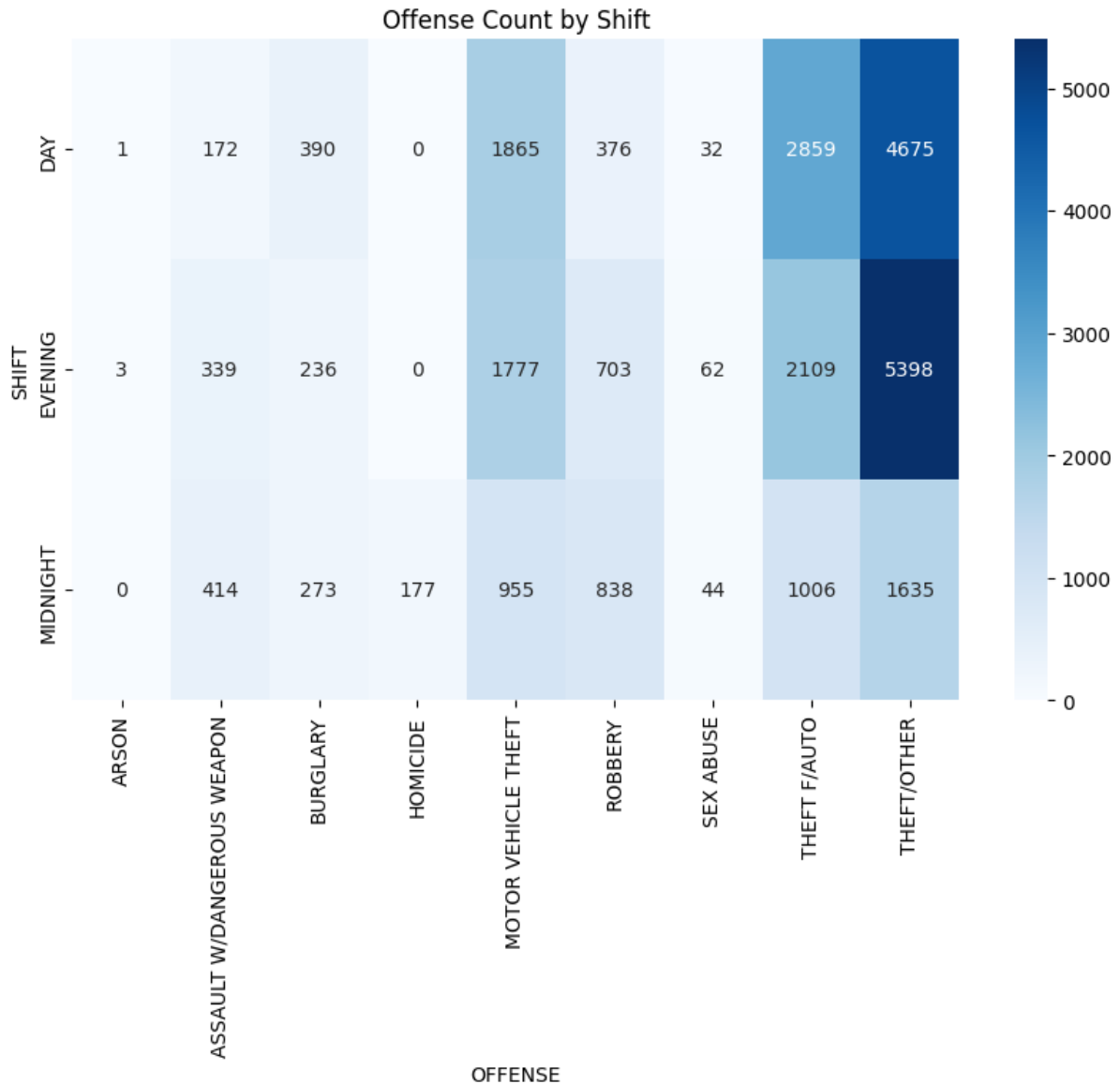


Figure 3.16: Offense counts by time of day, illustrating the distribution of various crime types (e.g., Theft, Burglary) across shifts (Day, Evening, Midnight) with color intensity reflecting the frequency of occurrences.

The figure above presents a heatmap of offense counts categorized by the time of day shift (Day, Evening, and Midnight). The rows represent different shifts, while the columns correspond to various offense types, such as Arson, Burglary, Homicide, Motor Vehicle Theft, Robbery, Sex Abuse, Theft, and others. The intensity of the color indicates the count of each offense type during the respective shift. Darker blue shades indicate higher offense counts, allowing us to easily identify trends in criminal activity across different times of the day. For example, the Evening shift shows a significantly higher number of offenses, especially in categories like Theft and Motor Vehicle Theft.

# Chapter 4. Feature Engineering in Clustering Models

Feature engineering plays a crucial role in the performance of machine learning models, especially in clustering tasks where the goal is to identify meaningful patterns or groups in data. In clustering, unlike supervised models, the lack of labels requires careful feature selection, extraction, and transformation to ensure the model can identify relevant patterns. This chapter focuses on the approach used in feature selection, data transformation, and the reasoning behind key decisions, such as the choice to not remove outliers and the decision not to scale certain features like geographic coordinates.

## 4.1 Introduction to Feature Engineering

Feature engineering is the process of selecting, modifying, or creating new features to improve the performance of machine learning algorithms. In the case of clustering models, feature engineering is vital for transforming raw data into a set of meaningful and informative features that enable the model to uncover hidden patterns effectively. For clustering algorithms such as K-means, the selection and transformation of features significantly influence the model's ability to detect meaningful clusters.

The main aspects of feature engineering in clustering include:

- **Feature Selection:** Identifying the most relevant features that will help the model form better clusters.
- **Feature Extraction:** Creating new features that capture more information from the data.
- **Data Transformation:** Applying transformations to the data, such as scaling, to make it more suitable for the algorithm.

## 4.2 Feature Selection Using Correlation

In our clustering model, **feature selection** was performed based on the correlation between different features. The features chosen—latitude, longitude, shift, method, and type of offense—are all relevant to detecting crime hotspots and patterns.

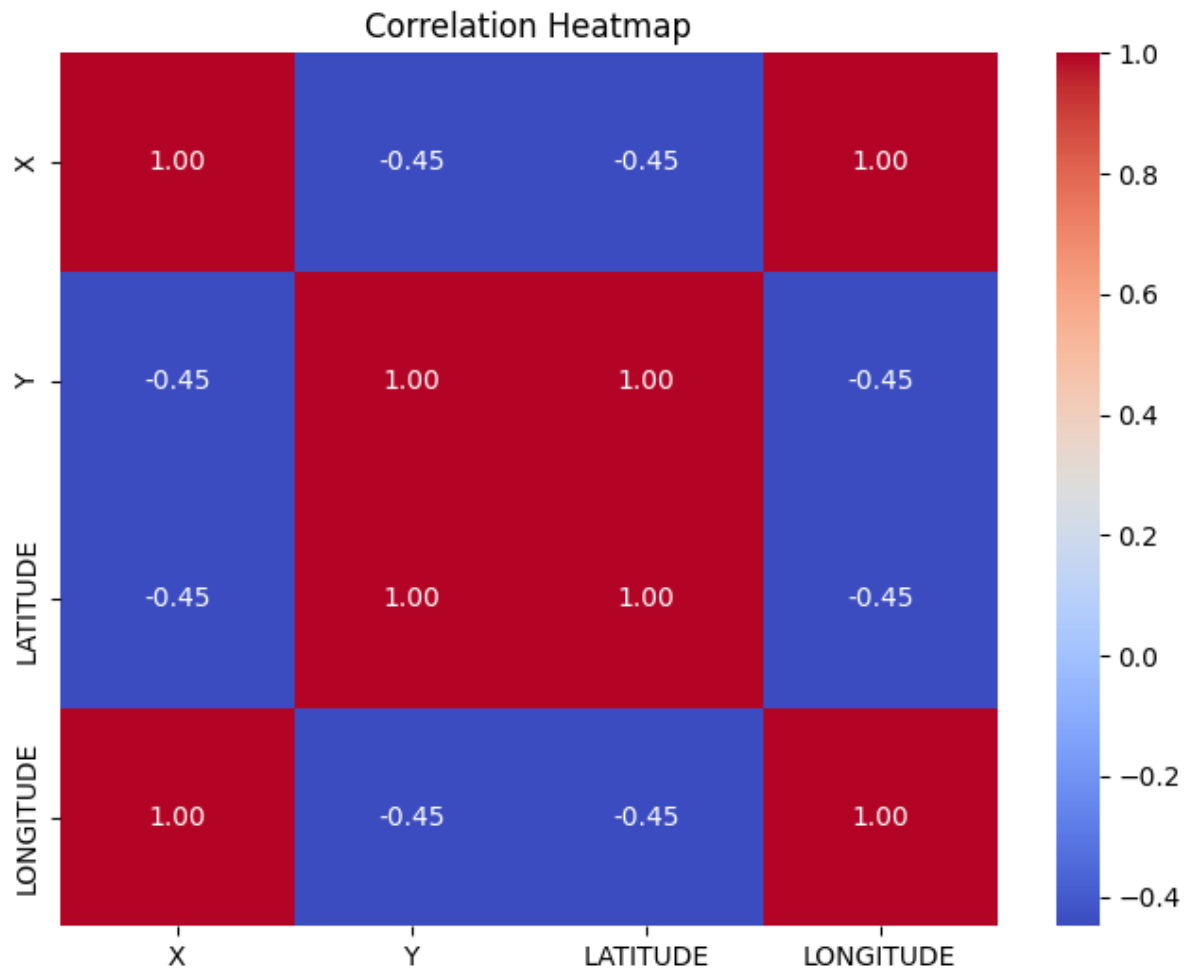


Figure 4.1: Heatmap showing the correlation between spatial features (X, Y, Latitude, and Longitude), providing insights into the relationships and linear dependencies among these geographical variables.

```
correlation = df2['X'].corr(df2['XBLOCK'], method='pearson')
print(f"Pearson Correlation: {correlation}")

correlation = df2['Y'].corr(df2['YBLOCK'], method='pearson')
print(f"Pearson Correlation: {correlation}")

correlation = df2['X'].corr(df2['LONGITUDE'], method='pearson')
print(f"Pearson Correlation: {correlation}")

correlation = df2['Y'].corr(df2['LATITUDE'], method='pearson')
print(f"Pearson Correlation: {correlation}")

correlation = df2['LONGITUDE'].corr(df2['XBLOCK'], method='pearson')
print(f"Pearson Correlation: {correlation}")

correlation = df2['LATITUDE'].corr(df2['YBLOCK'], method='pearson')
print(f"Pearson Correlation: {correlation}")
```

➡ Pearson Correlation: 1.0  
Pearson Correlation: 1.0  
Pearson Correlation: 0.9999999463012285  
Pearson Correlation: 0.999999972503696  
Pearson Correlation: 0.9999999463012184  
Pearson Correlation: 0.9999999725036924

Figure 4.2: Pearson correlation coefficients between pairs of variables in the dataset, indicating strong linear relationships (values close to 1) across all computed pairs.

1. **Latitude and Longitude:** These spatial features are crucial for detecting geographical hotspots, allowing us to group crimes based on their locations. As geographical coordinates are inherently important, we ensured that the correlation between latitude and longitude (X and Y) was considered for clustering. These features were found to have strong correlations, which was expected as crimes that occur in close proximity will have similar latitude and longitude values.
2. **Shift:** This feature allows the model to uncover time-based patterns of crime, identifying whether certain types of crimes are more likely to occur during specific hours of the day. By including the shift feature, we are incorporating the temporal dimension into the clustering analysis.
3. **Method and Type of Offense:** These categorical features provide additional context, enabling the model to distinguish between different types of crimes. For example, the method can reveal insights into weapon-related crime trends, while the type of offense allows the model to differentiate between violent and property-related crimes. Including these features aids in refining the clusters, offering more specific crime prevention insights.

The correlation between these features was considered to ensure that only the most relevant and non-redundant features were retained for the clustering process.

## 4.3 Data Transformation

Data transformation is an essential step in preparing data for clustering algorithms. This process involves adjusting the features to improve the clustering results and ensure the algorithm can accurately group similar data points together.

### 4.3.1 Reasoning for Not Removing Outliers

Outliers are often removed in supervised learning tasks because they can distort model predictions. However, in clustering, outliers can carry valuable information. Specifically, outliers may represent rare but important events, such as crimes occurring in unique locations or unusual time periods. Removing these outliers could lead to the loss of significant data points that help identify less common but potentially critical crime hotspots.

Since the goal of clustering is to find meaningful groups of data, including outliers ensures that the model remains sensitive to these rare occurrences. Outliers may not form their own cluster but could still contribute valuable information about the distribution of crime or unique patterns, such as anomalies or trends in specific locations or times.

### 4.3.2 Reasoning for Not Scaling Latitude and Longitude

Scaling features is a common preprocessing step in many machine learning tasks to ensure that each feature contributes equally to the model. However, for geographic data like **latitude** and **longitude**, scaling is unnecessary and can distort the model's ability to identify clusters based on real-world geographic distances.

1. **Geographic Coordinates' Intrinsic Meaning:** Latitude and longitude represent real-world distances on the Earth's surface. A degree of latitude corresponds to a fixed distance (approximately 111 kilometers), and a degree of longitude changes depending on the latitude. Scaling these coordinates would distort their natural distance relationships, leading to incorrect clustering.
2. **K-means Clustering and Euclidean Distance:** K-means clustering uses Euclidean distance to measure proximity between data points. Since latitude and longitude already provide meaningful geographic distances, scaling these coordinates could alter the interpretation of the distances between crimes, leading to inaccurate clustering.

By not scaling these coordinates, we preserve the inherent spatial relationships between points, ensuring that the model's clusters are based on actual geographic distances.

## 4.4 Conclusion

Feature engineering is a critical step in the clustering model, directly impacting the ability to detect crime hotspots and uncover meaningful patterns. By carefully selecting relevant features such as latitude, longitude, shift, method, and type of offense, and by transforming the data appropriately, we can ensure that the model detects the most relevant crime patterns.

In this chapter, we have:

- Chosen **latitude** and **longitude** as essential spatial features that provide valuable insights into geographical hotspots.

- Incorporated the **shift** feature to uncover time-based patterns of crime.
- Included the **method** and **type of offense** to add contextual information, helping differentiate between violent and property-related crimes.
- Decided against removing outliers, as they could represent valuable rare events in the crime dataset.
- Justified not scaling the latitude and longitude, as their natural spatial meaning is crucial for accurate clustering.

By combining spatial, temporal, and categorical features, this approach enhances the clustering model's ability to uncover richer insights into crime patterns, offering a more comprehensive view of crime in different areas and providing a foundation for targeted crime prevention strategies.

# Chapter 5. Model Fitting

Model fitting is a critical step in the data analysis process, particularly when dealing with unsupervised learning tasks like clustering. Unlike supervised learning, where the model is trained on labeled data, clustering seeks to uncover the underlying structure of the data by grouping similar data points together. In exploratory data analysis (EDA), model fitting allows us to identify patterns, trends, and structures that can inform decision-making or further analysis.

In this chapter, we will explore different types of models used for fitting, provide an insight into clustering models, and discuss the specific algorithms chosen for this project. Additionally, we will dive into performance metrics used to evaluate clustering results, including the Elbow Method and Silhouette Score.

## 5.1 Insight into Model Fitting in EDA

In exploratory data analysis, the goal of model fitting is to identify relationships within the data and validate any hypotheses or patterns that arise during analysis. Since EDA often involves unsupervised learning, the model fitting process is used to detect hidden structures without predefined labels. This process allows for the discovery of groups, trends, and anomalies that could guide further analysis or lead to actionable insights.

In clustering, model fitting does not involve training a model with labeled data but rather identifying the optimal grouping of data points. The process involves selecting the appropriate clustering algorithm, tuning hyperparameters (such as the number of clusters), and assessing the model's performance.

## 5.2 Types of Models Used for Fitting

There are several types of models used for fitting in data analysis. These models can be broadly classified into three categories:

### 5.2.1 Regression Models

Regression models are used in supervised learning tasks to predict continuous numerical values based on input features. Common types of regression include:

- **Linear Regression:** Predicts a continuous target variable based on a linear relationship with input features.
- **Logistic Regression:** Used for binary classification tasks, where the outcome is a probability value between 0 and 1.

### 5.2.2 Classification Models

Classification models are used for supervised learning tasks where the goal is to predict discrete class labels. Examples include:

- **Decision Trees:** A tree-like model used for both classification and regression tasks.
- **Random Forests:** An ensemble of decision trees used to improve accuracy by reducing overfitting.
- **Support Vector Machines (SVMs):** Used for classification tasks by finding an optimal hyperplane that separates the classes.

### 5.2.3 Clustering Models

Clustering is a type of unsupervised learning used to group data points into clusters based on similarities. The key clustering models include:

- **K-Means Clustering:** A partitioning method that divides the data into K clusters by minimizing the sum of squared distances between points and their corresponding cluster centroids.
- **Hierarchical Clustering:** Builds a tree of clusters by iteratively merging or splitting clusters based on a distance metric.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A density-based algorithm that groups closely packed points and identifies noise points.

## 5.3 Clustering Models in Our Project

In this project, the primary goal was to identify crime hotspots based on spatial, temporal, and categorical features. Since our dataset involved geographical coordinates (latitude and longitude), a clustering model was ideal to identify regions with high crime densities and discern time-based patterns. The two clustering algorithms selected for this project were:

1. **Agglomerative Model:** This method was explored for its potential in identifying unique crime patterns based on the given features. It works by minimizing a distance metric across the dataset to group similar points.
2. **K-Means Clustering:** This algorithm was chosen for its simplicity and effectiveness in partitioning the dataset into clusters based on similarity in feature space. Specifically, K-Means was applied to the latitude and longitude data to find geographic clusters of crime events.

## 5.4 Performance Metrics for Clustering Models

Evaluating the performance of clustering models is more challenging than supervised models, as there are no ground-truth labels to compare predictions against. However, several methods can help evaluate the quality of the clusters. Here, we will focus on two widely used performance metrics for clustering:



### 5.4.1 Elbow Method

The Elbow Method is a technique used to determine the optimal number of clusters,  $K$ , for K-Means clustering. It involves plotting the **Within-Cluster Sum of Squares (WCSS)** against the number of clusters. As the number of clusters increases, the WCSS decreases, but after a certain point, the rate of decrease slows down significantly, forming an "elbow" in the plot. The optimal number of clusters is typically at the point where the curve begins to level off.

$$WCSS(K) = \sum_{i=1}^N \sum_{k=1}^K \|x_i - \mu_k\|^2$$

Where:

- $N$  is the number of data points,
- $K$  is the number of clusters,
- $\mu_k$  is the centroid of the  $k$ -th cluster,
- $x_i$  is a data point.

By plotting the WCSS for different values of  $K$ , we can visually identify the "elbow" point, which indicates the optimal number of clusters.

### 5.4.2 Silhouette Score

The Silhouette Score is a measure of how similar each point is to its own cluster compared to other clusters. It ranges from -1 to 1, with a higher score indicating better-defined clusters. A Silhouette Score close to 1 means that the data points are well-clustered, while a score close to -1 indicates that the points may have been assigned to the wrong clusters.

The formula for the Silhouette Score for a point  $i$  is given by:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where:

- $a(i)$  is the average distance from point  $i$  to all other points in the same cluster,
- $b(i)$  is the average distance from point  $i$  to all points in the nearest cluster.

The overall Silhouette Score is the average of the Silhouette Scores of all data points.

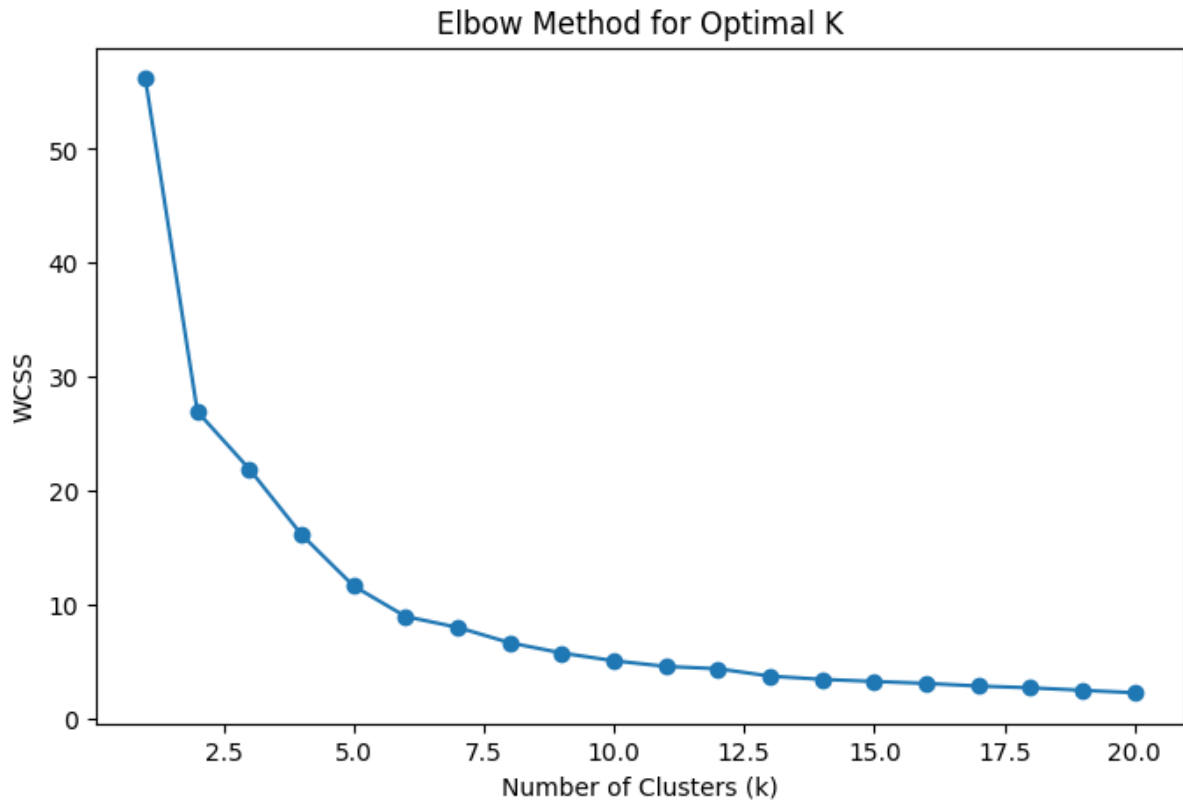


Figure 5.1: Elbow method plot showing the optimal number of clusters ( $K$ ) for the dataset, where the point of inflection indicates the ideal value for  $K$ , balancing within-cluster variance and cluster count

## 5.5 Model Fitting and Cluster Evaluation

### 5.5.1 Elbow Method for Optimal Clusters

The **Elbow Method** helps determine the optimal number of clusters by plotting the *Within-Cluster Sum of Squares (WCSS)* for various values of  $K$ . The point where the decrease in WCSS slows down and forms an "elbow" is considered the optimal number of clusters. In our case, the graph shows that the optimal number of clusters is  $K = 6$ . The WCSS decreases sharply until  $K = 6$  and then levels off, indicating that adding more clusters would not significantly improve the clustering quality. Thus, the optimal number of clusters is determined to be **6**.

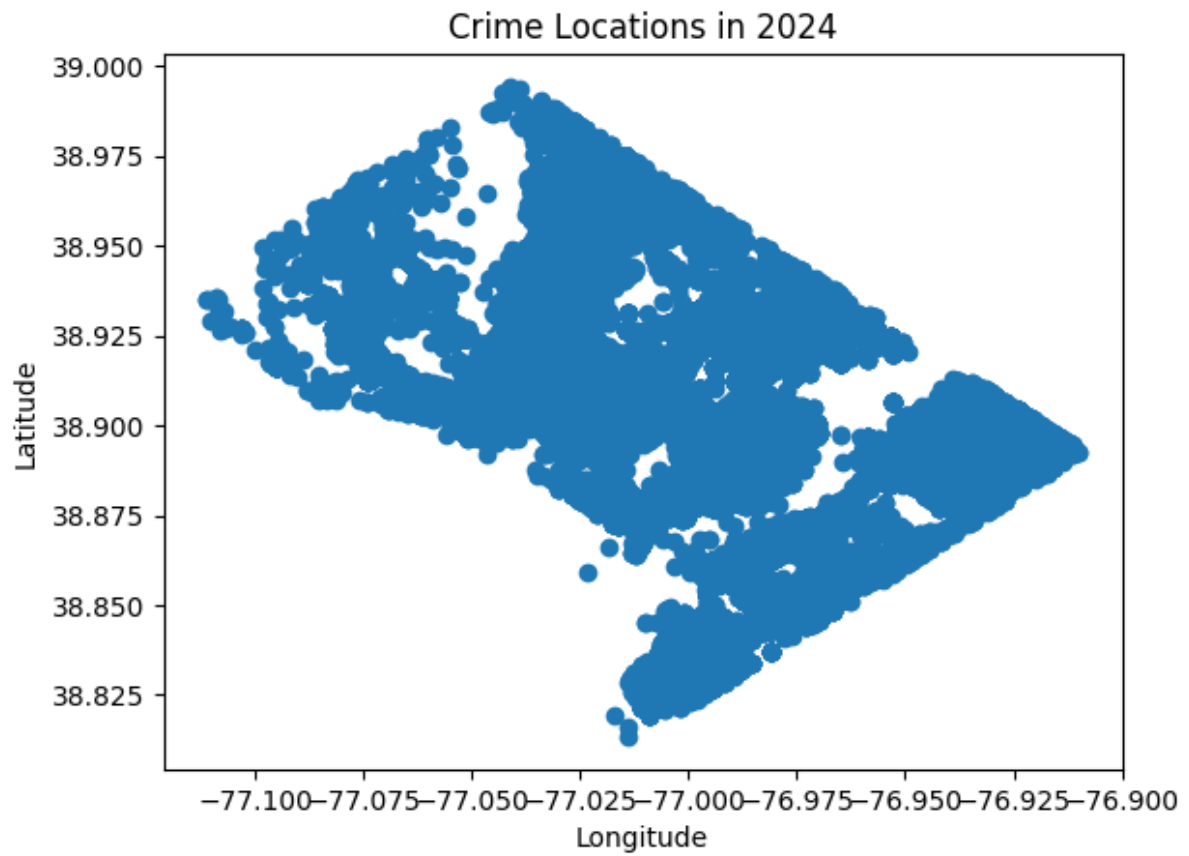


Figure 5.2: Caption

### 5.5.2 Hotspots of Agglomerative

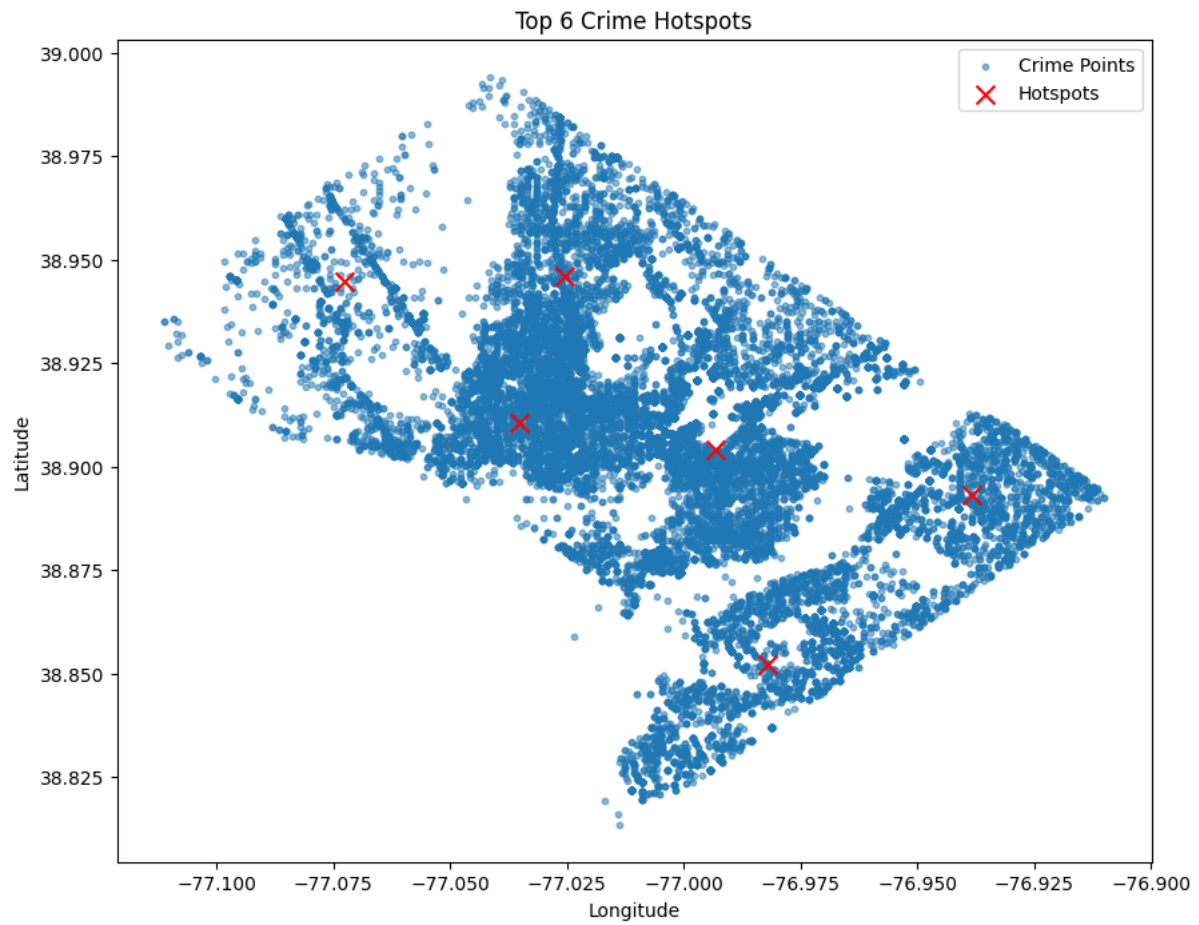


Figure 5.3: Visualization of the top 6 crime hotspots throughout the year, highlighting areas with the highest frequency of criminal activities and showing seasonal patterns in crime distribution, through agglomerative

### 5.5.3 Hotspots of K-Means

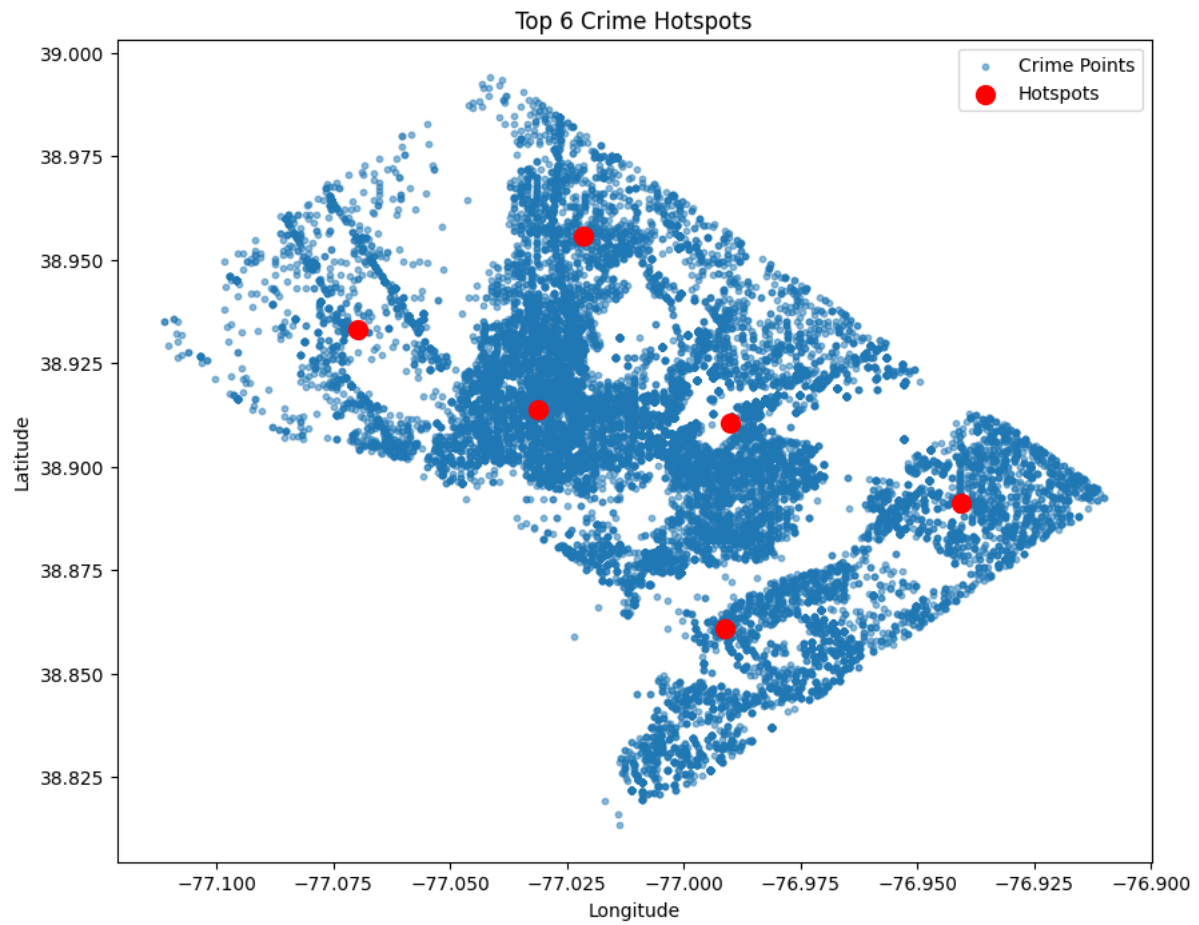


Figure 5.4: Visualization of the top 6 crime hotspots throughout the year, highlighting areas with the highest frequency of criminal activities and showing seasonal patterns in crime distribution, through K means

### 5.5.4 Silhouette Score Vs No. of Clusters

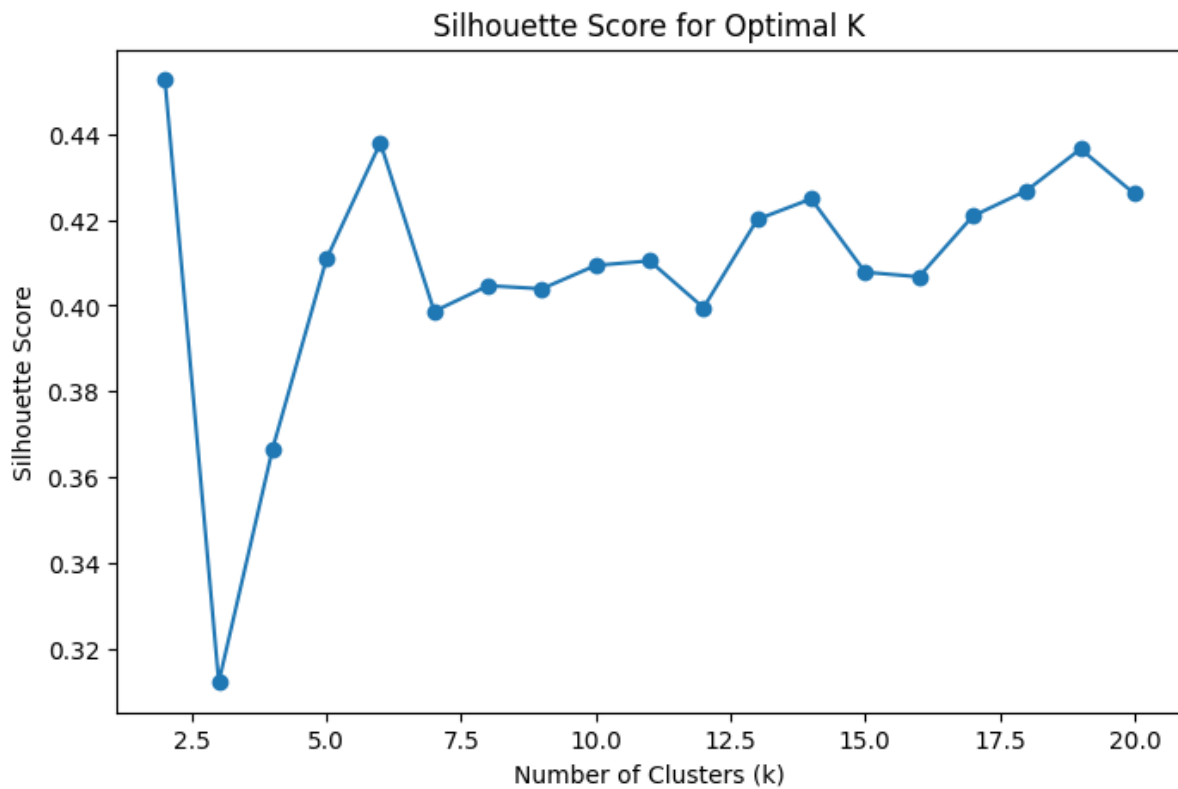


Figure 5.5: Caption

### 5.5.5 Score Comparison

```
Silhouette score for Agglomerative Clustering: 0.3990445426098079
```

Figure 5.6: Caption

```
Silhouette Score for the KMeans clustering with 6 clusters: 0.4386733560757238
```

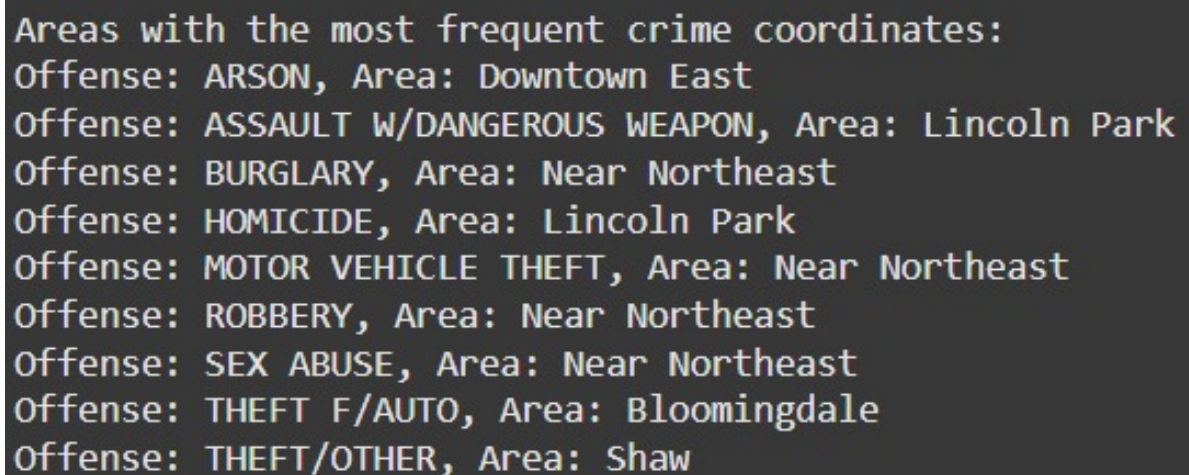
Figure 5.7: Caption

## 5.6 Conclusion

From the scores, KMeans outperforms the Agglomerative.

## Chapter 6. Conclusion & future scope

### 6.1 Findings/observations



```
Areas with the most frequent crime coordinates:  
Offense: ARSON, Area: Downtown East  
Offense: ASSAULT W/DANGEROUS WEAPON, Area: Lincoln Park  
Offense: BURGLARY, Area: Near Northeast  
Offense: HOMICIDE, Area: Lincoln Park  
Offense: MOTOR VEHICLE THEFT, Area: Near Northeast  
Offense: ROBBERY, Area: Near Northeast  
Offense: SEX ABUSE, Area: Near Northeast  
Offense: THEFT F/AUTO, Area: Bloomingdale  
Offense: THEFT/OTHER, Area: Shaw
```

Figure 6.1: Hotspots Areas

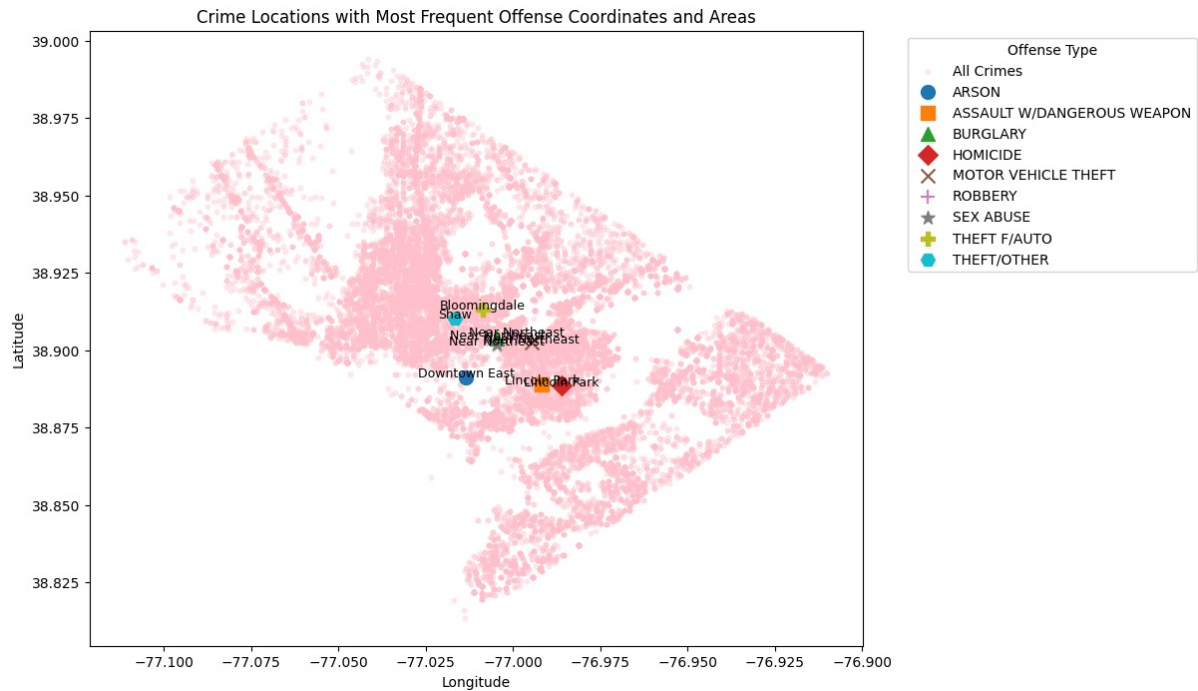


Figure 6.2: Hotspots Areas

```

Cluster 5: Most common offense is 'THEFT/OTHER' with 4740 occurrences.
Cluster 0: Most common offense is 'THEFT/OTHER' with 2076 occurrences.
Cluster 3: Most common offense is 'THEFT/OTHER' with 1551 occurrences.
Cluster 4: Most common offense is 'THEFT/OTHER' with 1184 occurrences.
Cluster 1: Most common offense is 'THEFT/OTHER' with 815 occurrences.
Cluster 2: Most common offense is 'THEFT/OTHER' with 1342 occurrences.

```

Figure 6.3: Hotspots Areas



```
Most common method of crime:
Method: OTHERS, Count: 23989

Top coordinates witnessing the most crimes via GUN:
Latitude: 38.8966558827, Longitude: -76.9476475094, Count: 14.0

Top 10 coordinates for gun crimes:
      LATITUDE  LONGITUDE  count
623  38.896656  -76.947648    14
680  38.899976  -76.982460     8
185  38.854172  -76.988424     8
189  38.855203  -76.989731     8
  2   38.821626  -77.011081     7
845  38.906662  -76.952819     6
330  38.874120  -76.972067     6
155  38.848137  -76.976139     6
210  38.858465  -76.990056     6
  6   38.823601  -77.008890     6
```

Figure 6.4: Hotspots Areas

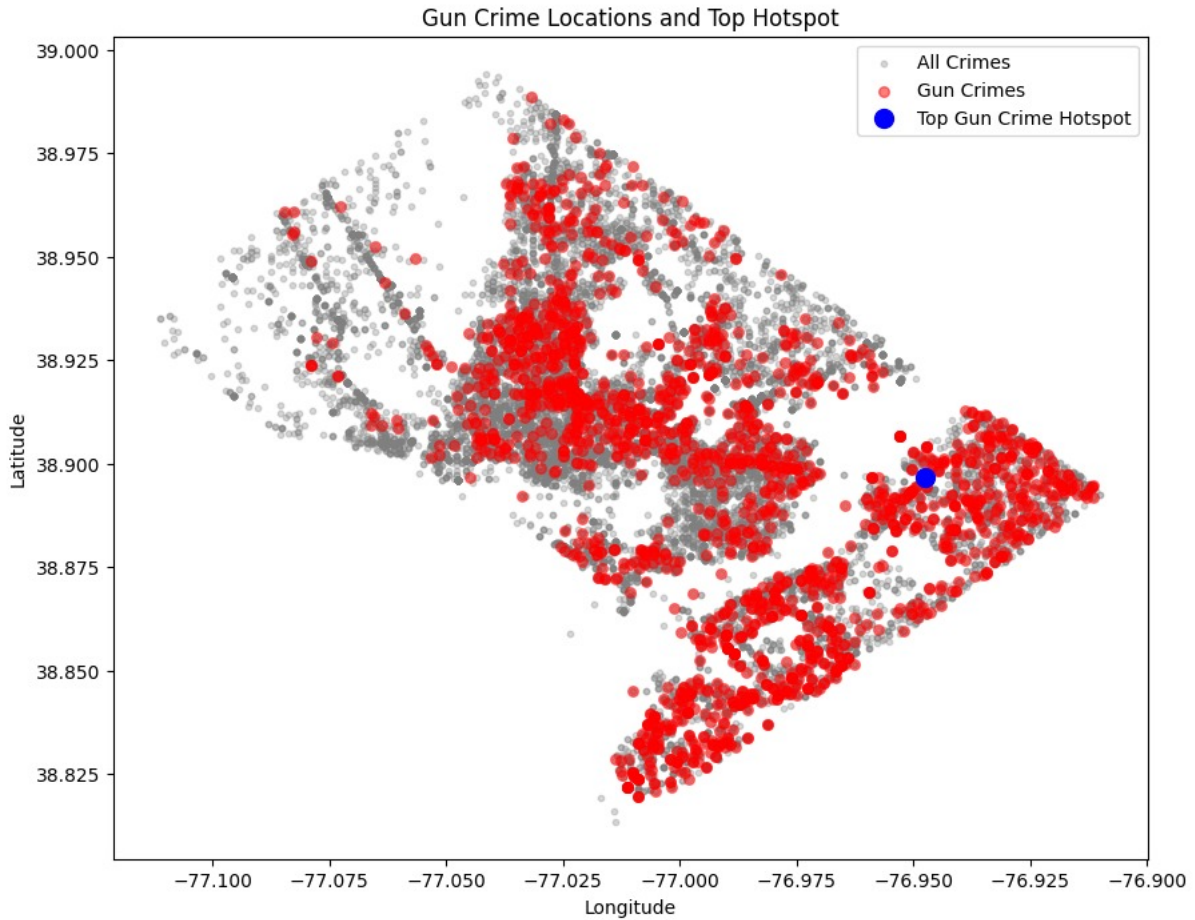


Figure 6.5: Hotspots Areas

Top Gun Crime Hotspot Area:  
Latitude: 38.8966558827, Longitude: -76.9476475094, Area: Central Northeast

Top 10 Gun Crime Hotspot Areas:

	LATITUDE	LONGITUDE	count	area
623	38.896656	-76.947648	14	Central Northeast
680	38.899976	-76.982460	8	Carver
185	38.854172	-76.988424	8	Hillsdale
189	38.855203	-76.989731	8	Hillsdale
2	38.821626	-77.011081	7	Washington
845	38.906662	-76.952819	6	Mayfair
330	38.874120	-76.972067	6	Washington
155	38.848137	-76.976139	6	Washington
210	38.858465	-76.990056	6	Bridge District
6	38.823601	-77.008890	6	Bellevue

Figure 6.6: Hotspots Areas

```
The shift with the most crimes is: EVENING with 10627 occurrences.
```

Figure 6.7: Hotspots Areas

```
Most common offense in each shift:
Shift: DAY, Most common offense: THEFT/OTHER with 4675 occurrences
Shift: EVENING, Most common offense: THEFT/OTHER with 5398 occurrences
Shift: MIDNIGHT, Most common offense: THEFT/OTHER with 1635 occurrences
```

Figure 6.8: Hotspots Areas

## Future Plans

Based on the clustering analysis identifying *crime hotspots*, particularly where **gun crimes** are prevalent, the following recommendations focus on improving crime prevention and public safety:

- **Increase Patrols in High-Risk Areas:** Deploy targeted patrols and surveillance in hotspots with frequent *gun crimes* to deter violent offenses and improve public safety.
- **Focus on Evening Shifts:** Enhance police presence during *evening hours*, which witness the highest crime rates, particularly for *theft* and gun-related incidents.
- **Address High-Theft Zones:** Implement measures like *security cameras*, *plainclothes patrols*, and *public awareness campaigns* in areas with high theft rates to reduce property crimes.
- **Leverage Data-Driven Policing:** Use insights from crime clustering to allocate resources effectively, focusing on high-risk areas and critical times with *predictive policing tools*.
- **Community Engagement:** Strengthen community collaboration through *neighborhood watch programs*, *community policing*, and regular *crime awareness meetings*.
- **Environmental Design Improvements:** Use *Crime Prevention Through Environmental Design (CPTED)* principles to address vulnerabilities in high-crime areas by improving lighting, visibility, and urban design.
- **Continuous Monitoring and Adjustment:** Regularly reassess crime patterns and adjust strategies based on new data to ensure proactive crime prevention.

**Conclusion:** These plans emphasize a *data-driven, community-oriented approach* to reducing crime, focusing on **hotspot areas**, **evening shifts**, and addressing **theft** and **gun crimes**. By continuously monitoring and refining strategies, the police can ensure public safety and resource efficiency.

# Group Contribution

## **Krishil Jayswal**

25 % code 40 % report 25 % ppt

## **Bhavya Boda**

25 % code 40 % report 25 % ppt

## **Aniket Pandey**

50 % code 20 % report 50 % ppt

# Short Bio

1. **Krishil Jayswal** I am krishil Jayswal from MnC batch 2022, student ID 202203040.

2. **Bhavya Boda** I am Bhavya Boda from MnC batch 2022, student ID 202203067. I am very much passionate and have a keen interest in the field of mathematics from my childhood and that's the reason I have taken the Mathematics and Computing branch in DAI-ICT. My hometown is in Jamnagar, Gujarat. I have interest in problem solving and competitive programming. I also like playing cricket, badminton, chess, and have a keen interest in listening music and mobic games too. I have participated in various Drama events as a part of the college drama club.

3. **Aniket Pandey** I am aniket Pandey from MTech ML batch 2024, with student ID 202411001. I was brought up in Vadodara, Gujarat but my origin is from Buxar, Bihar. I pursued my B.Tech in ICT from Pandit Deendayal Energy University, Gandhinagar. My technical skills include C , Python, Machine Learning and basics of MySQL. I am an avid sports fan with massive interest in cricket, football, combat sports and I like to analyze sports data. I also like watching movies and tv shows especially crime and mystery thrillers hence, this project is a good opportunity to put my crime investigation skills to a test.

# References

- [1] Website from where dataset is taken for study URL: <https://data.gov/>