IT-462

Exploratory Data Analysis

On Police Dataset

Mentored by : Dr. Gopinath Panda	Group 22

Group 22

NAME	STUDENT ID
ANIKET PANDEY	202411001
BHAVYA BODA	202203067
KRISHIL JAYSWAL	202203040

Table of Contents

Introduction

- The dataset that we have taken for this project is the crime dataset record in the year 2024 in the city of Washington, D.C.

- Washington, D.C., formally the District of Columbia and commonly known as Washington or D.C., is the capital city and federal district of the United States. The city is on the Potomac River, across from Virginia, and shares land borders with Maryland to its north and east.

Motivation

Why this project?

- Enhancing Public Safety
- Data-Driven Policing
- Community Awareness and Engagement
- Understanding Crime Patterns
- Resource Optimization

This project is not just about Washington, DC—it sets a framework for police to analyze crimes in other cities. It highlights the power of data analytics to solve complex societal challenges and demonstrates the impact of technology in improving community safety and well-being.

Problem Statement

The rising concerns over public safety and crime management in urban areas like Washington, DC, demand effective use of data-driven approaches to improve policing strategies. With crimes ranging from theft and assault to gun-related violence, identifying crime patterns and hotspots can help allocate law enforcement resources more effectively. However, the lack of granular, actionable insights hinders the ability of the police to target high-risk areas and prevent future incidents.

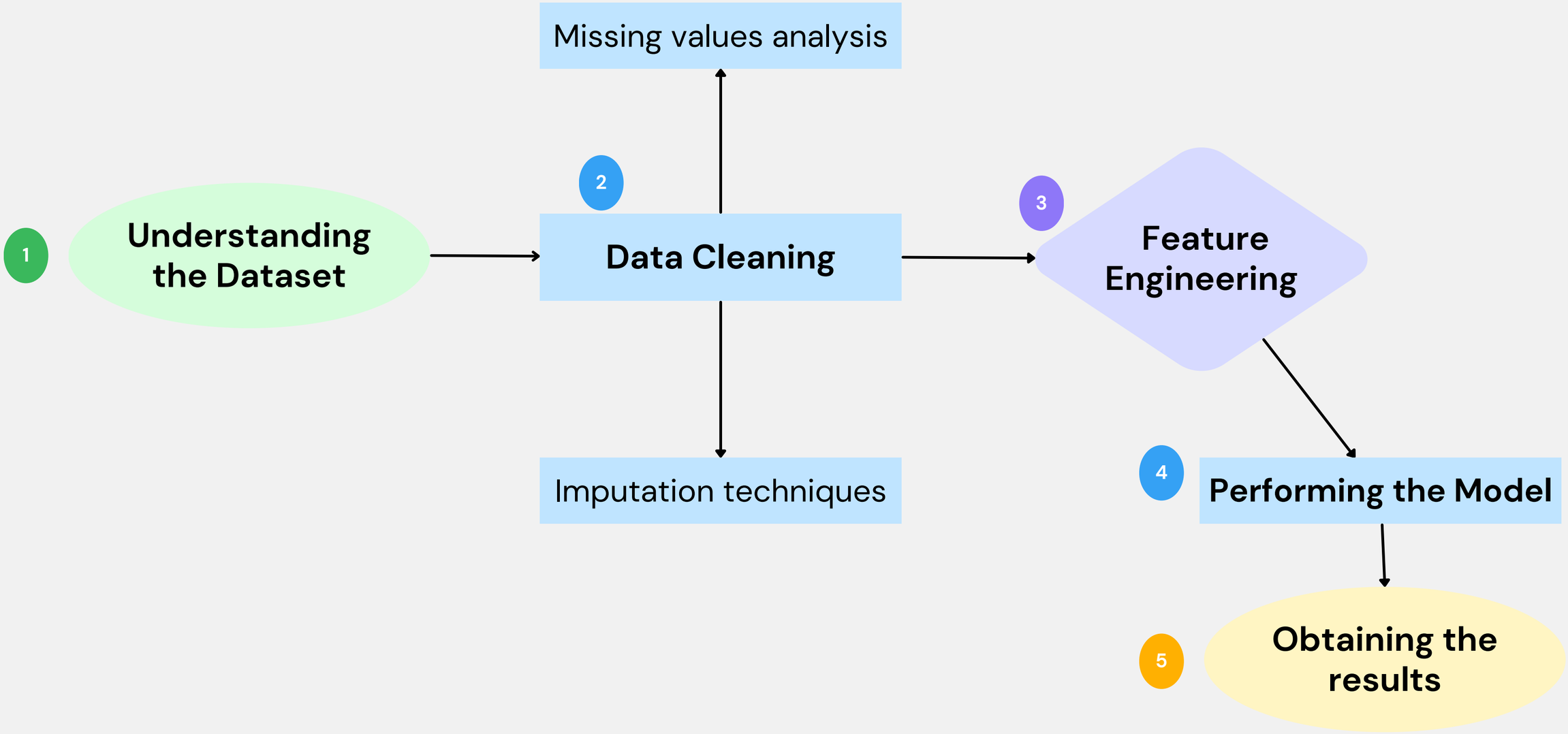
Objectives:

1. Provide detailed geospatial visualizations of crime hotspots to help law enforcement focus on high-risk areas.
2. Identify crime-based trends to aid in scheduling and deploying police resources effectively.
3. Offer insights into the types of crimes prevalent in different neighborhoods, helping design tailored prevention strategies.

Real-World Impact:

- Improved Safety.
- Resource Optimization.
- Policy Support.

Methodology



Implementation

1. Understanding the Dataset

- The project sourced the data from reputable government official website (data.gov)

- Dataset contains 25 features with 26339 instances.

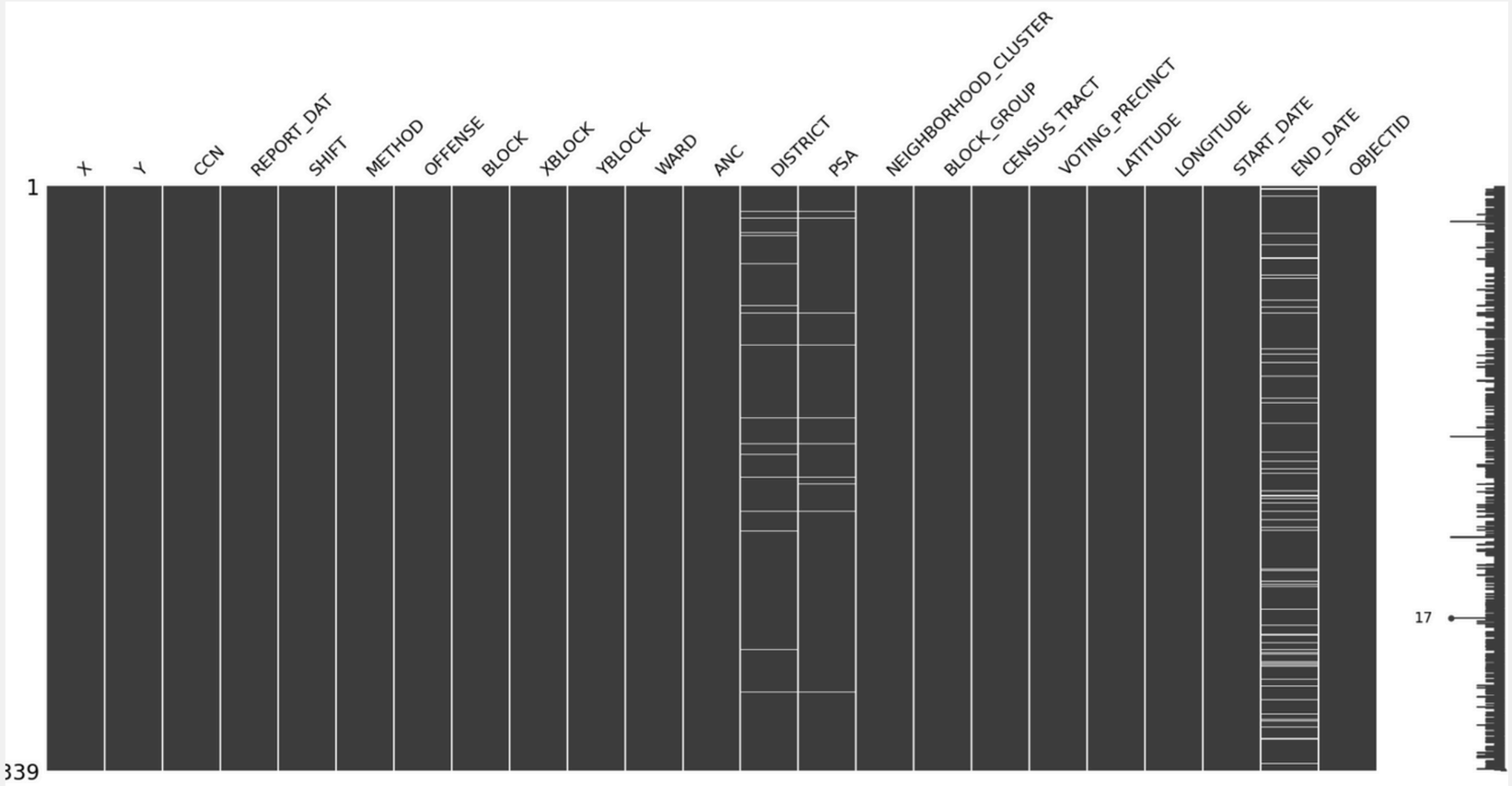
- Detailed crime statistics across different coordinates in the city.

- Major information about the crimes involves the shift (the time period of the day the crime was commited in), method of offense and type of crime, which provides crucial information about the distribution of crimes across the city.

2. **Cleaning the Data**

- Missing Value Analysis

- Highlights missing values across dataset columns.
- Helps identify features requiring imputation or exclusion during preprocessing



Little’s MCAR test and data imputation

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

- The test determines whether missing data is random or follows a pattern.
- A p-value of 1.0 confirms the data is Missing Completely at Random (MCAR) and that imputation can be applied without concern for bias, ensuring the integrity of subsequent analysis.
- The dataset was divided into numerical and categorical columns.
- Mean imputation was applied to numerical columns, while mode imputation was used for categorical columns to fill missing values effectively.

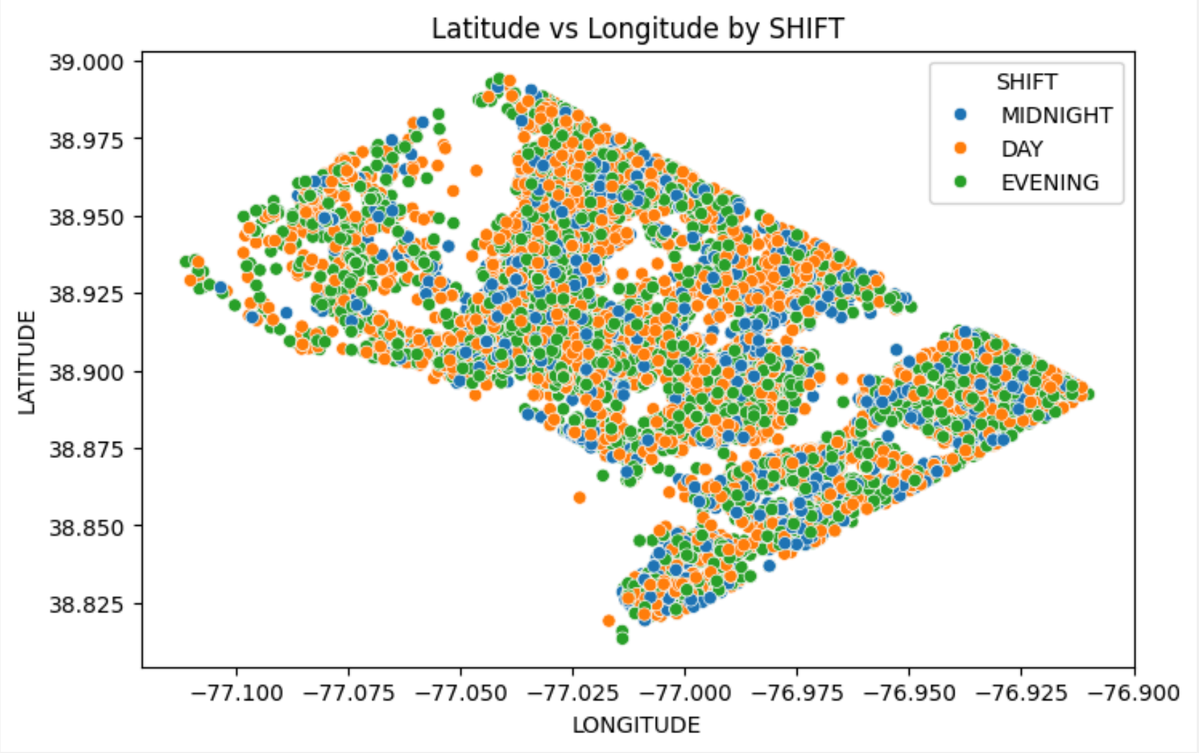
3. **Feature Engineering**

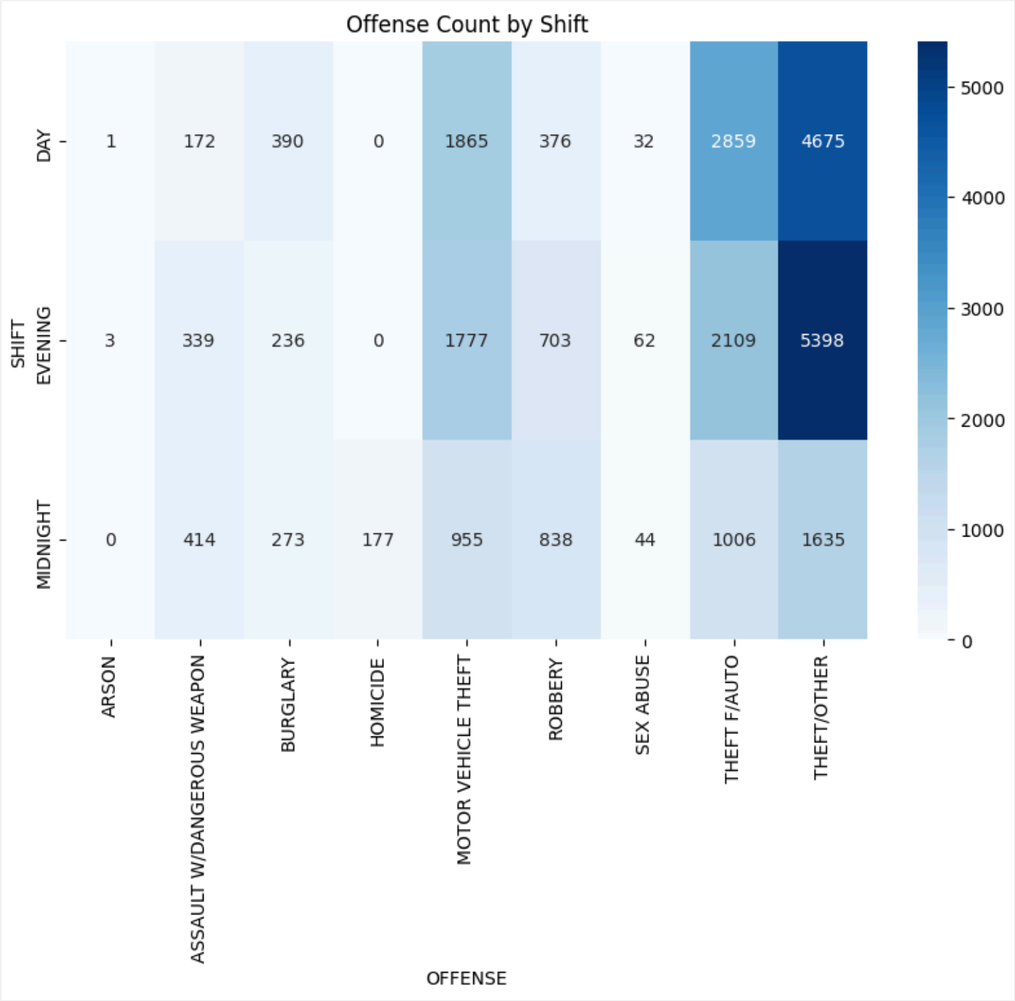
We done the Data Visulization using the following techniques.

- Univariate Analysis
- Multivariate Analysis

The feature selection in Clustering Model is different from Regression or Classification Model. It is primarily based on two things

- Spatial Coordinates
- Labels used for labelling the point


Latitude vs Longitude by SHIFT
SHIFT
MIDNIGHT
DAY
EVENING
LATITUDE
LONGITUDE
39.000
38.975
38.950
38.925
38.900
38.875
38.850
38.825
−77.100
−77.075
−77.050
−77.025
−77.000
−76.975
−76.950
−76.925
−76.900


Offense Count by Shift
SHIFT
DAY
EVENING
MIDNIGHT
1
172
390
0
1865
376
32
2859
4675
3
339
236
0
1777
703
62
2109
5398
0
414
273
177
955
838
44
1006
1635
ARSON
ASSAULT W/DANGEROUS WEAPON
BURGLARY
HOMICIDE
MOTOR VEHICLE THEFT
ROBBERY
SEX ABUSE
THEFT F/AUTO
THEFT/OTHER
OFFENSE
5000
4000
3000
2000
1000
0

For the Spatial Coordinates, we have the following variables from our dataset.

- Latitude
- X
- XBlock
- Longitude
- Y
- YBlock

```
Pearson Correlation: 1.0 (X vs XBLOCK)
Pearson Correlation: 1.0 (Y vs YBLOCK)
Pearson Correlation: 0.9999999463012285 (X vs longitude)
Pearson Correlation: 0.999999972503696 (Y vs latitude)
Pearson Correlation: 0.9999999463012184 (longitude vs XBLOCK)
Pearson Correlation: 0.9999999725036924 (latitude vs YBLOCK)
```

From the Correlation Coefficient, we chose Latitude and Longitude.

And for the Labelling the points, we used the following variables which describes the characterstic of a particular point

- Shift
- Type of Offence
- Method

4. **Performing the Model**

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. The two clustering algorithms selected for this project were:

1. Agglomerative Clustering

2. K-Means Clustering

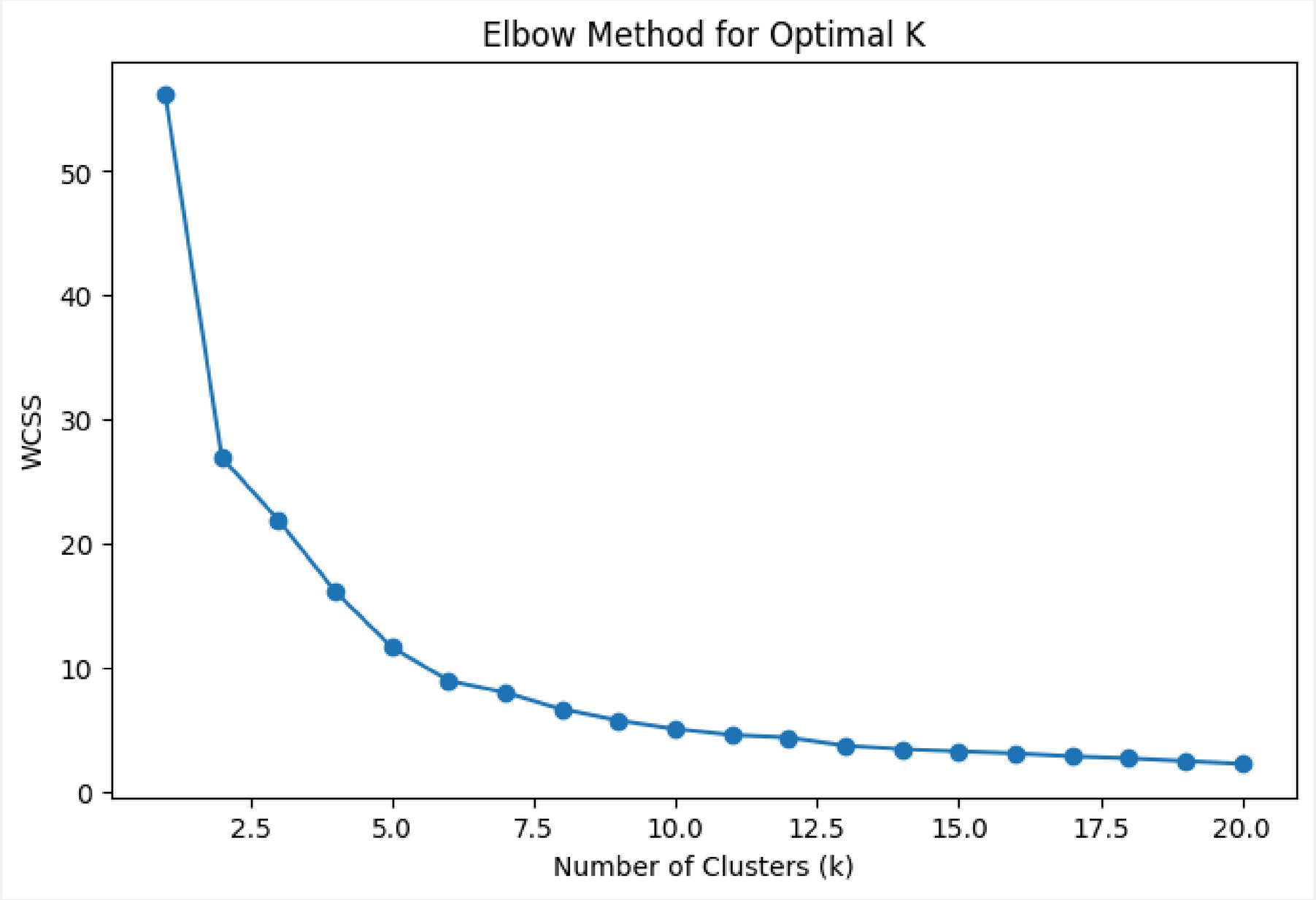
Performance Metrics for Clustering Models

Elbow Method:

The Elbow Method is a technique used to determine the optimal number of clusters, K, for KMeans
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,
- μk is the centroid of the k-th cluster,
- xi is a data point.


Elbow Method for Optimal K
WCSS
0
10
20
30
40
50
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of Clusters (k)

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:

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

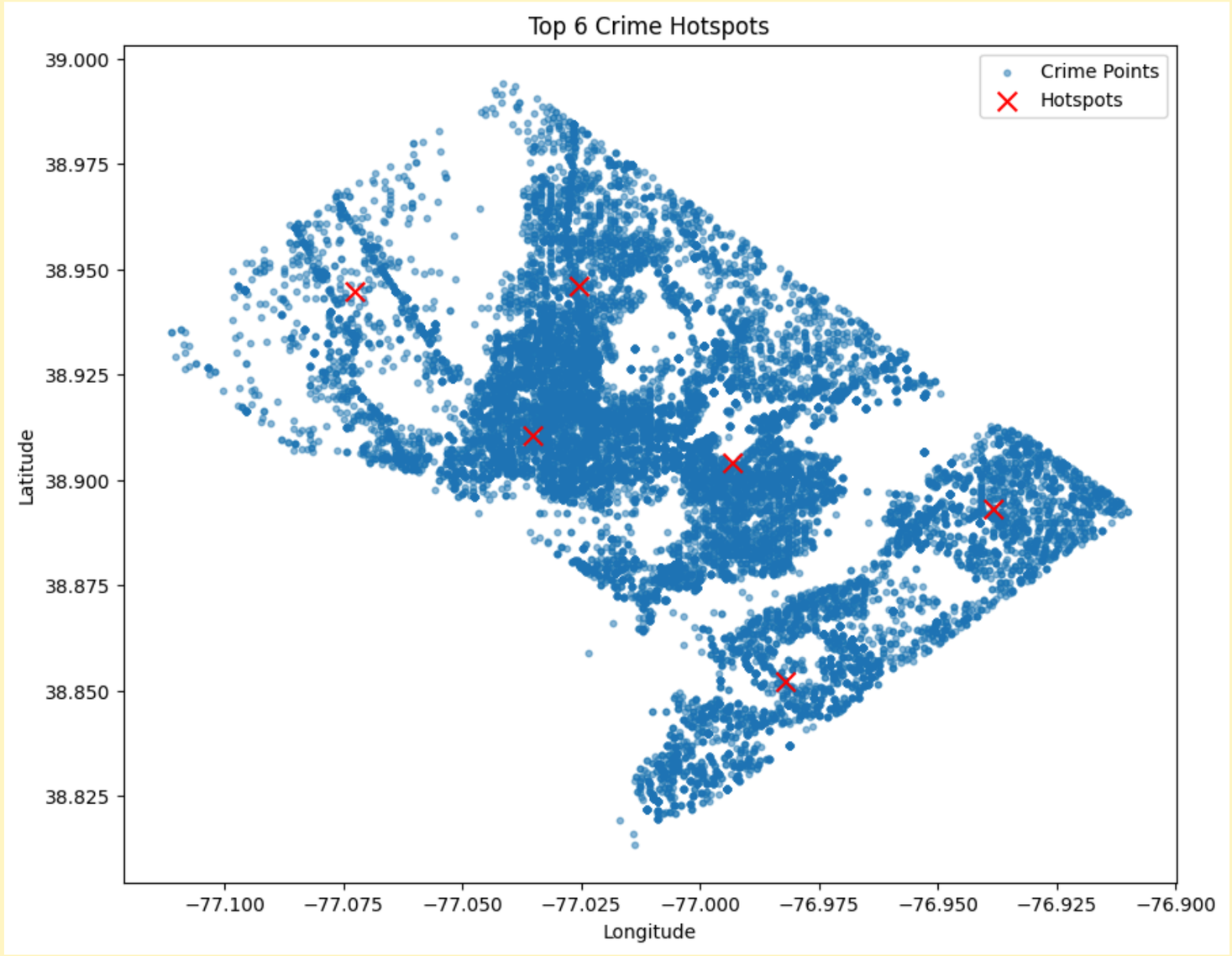
Where:

- 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.

Results

The silhouette score for Agglomerative Clustering was found to be 0.39

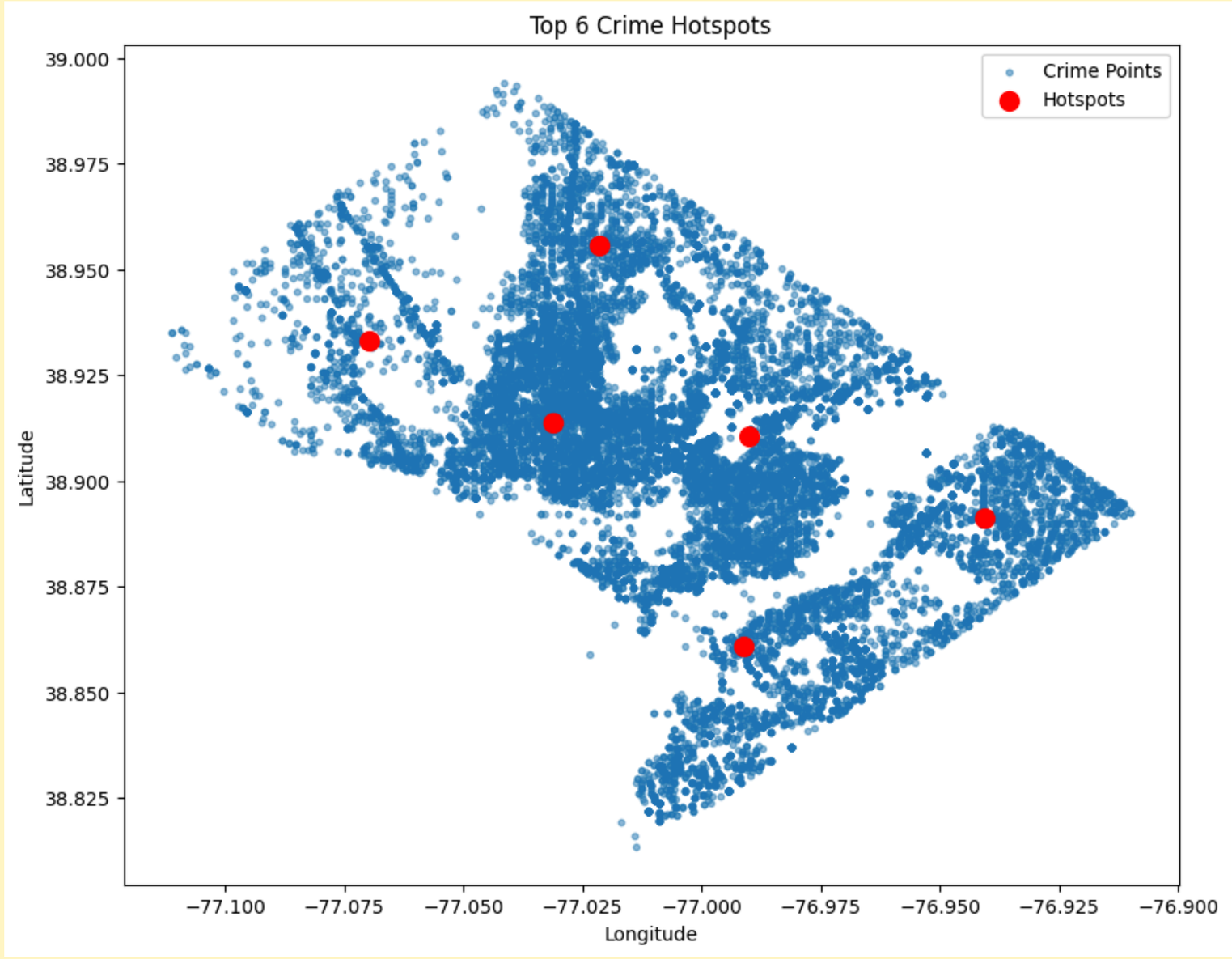


Top 6 crime hotspots using agglomerative clustering

Results

The silhouette score for K-Means Clustering was found to be almost 0.45

We can see that K-Means performs better clustering compared to Agglomerative Clustering and hence it will be taken into further analysis and agglomerative clustering will be dropped.



Top 6 crime hotspots using K-Means clustering

Results

```
Top 6 Crime Hotspot Coordinates:
Hotspot 1: Latitude = 38.913608, Longitude = -77.031320
Hotspot 2: Latitude = 38.910403, Longitude = -76.990030
Hotspot 3: Latitude = 38.860789, Longitude = -76.991199
Hotspot 4: Latitude = 38.955590, Longitude = -77.021520
Hotspot 5: Latitude = 38.891175, Longitude = -76.940657
Hotspot 6: Latitude = 38.933282, Longitude = -77.069863
```

Coordinates of the top 6 crime hotspots

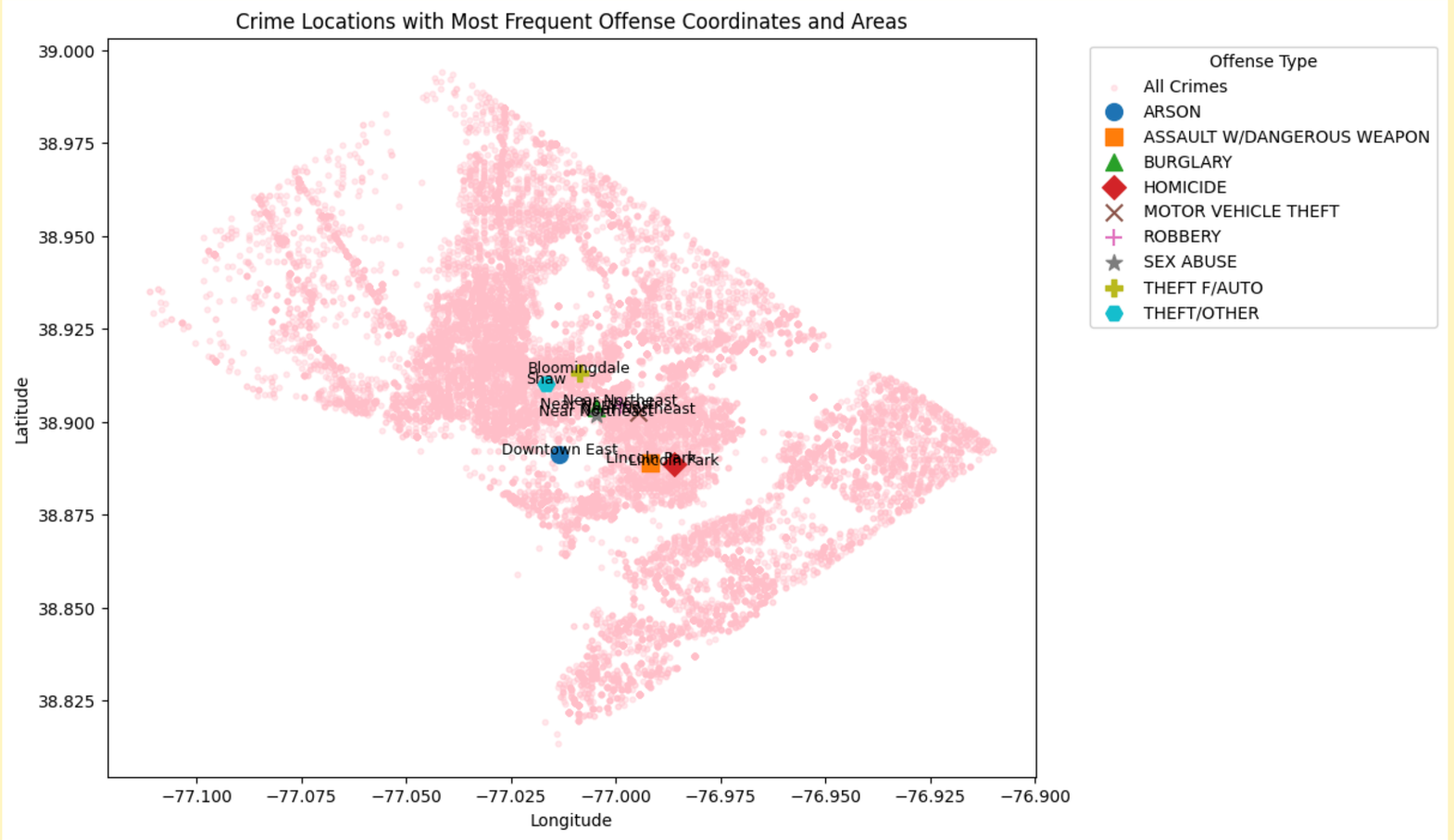
```
Top 6 Crime Hotspots with Area Names:
Hotspot 1: Latitude = 38.913608, Longitude = -77.031320, Area = 1339, Riggs Street Northwest, Logan Circle/Shaw, Ward 2, Washington, District of Columbia, 20009, United States
Hotspot 2: Latitude = 38.910403, Longitude = -76.990030, Area = Gallaudet University, Brentwood Parkway Northeast, Ivy City, Ward 5, Washington, District of Columbia, 20002, United S
Hotspot 3: Latitude = 38.860789, Longitude = -76.991199, Area = 1338, Talbert Terrace Southeast, Bridge District, Hillsdale, Ward 8, Washington, District of Columbia, 20020, United S
Hotspot 4: Latitude = 38.955590, Longitude = -77.021520, Area = 621, Jefferson Street Northwest, Brightwood Park, Ward 4, Washington, District of Columbia, 20011, United States
Hotspot 5: Latitude = 38.891175, Longitude = -76.940657, Area = 4262, East Capitol Street Northeast, Ward 7, Washington, District of Columbia, 20019, United States
Hotspot 6: Latitude = 38.933282, Longitude = -77.069863, Area = 3512, Macomb Street Northwest, Cleveland Park, Ward 3, Washington, District of Columbia, 20016, United States
```

The locations of top 6 crime hotspots after using reverse geocoding

Results

Unique responses in Offense:

- HOMICIDE
- THEFT/OTHER
- MOTOR VEHICLE THEFT
- ROBBERY
- THEFT F/AUTO
- ASSAULT W/DANGEROUS WEAPON
- SEX ABUSE
- BURGLARY
- ARSON




```
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
```

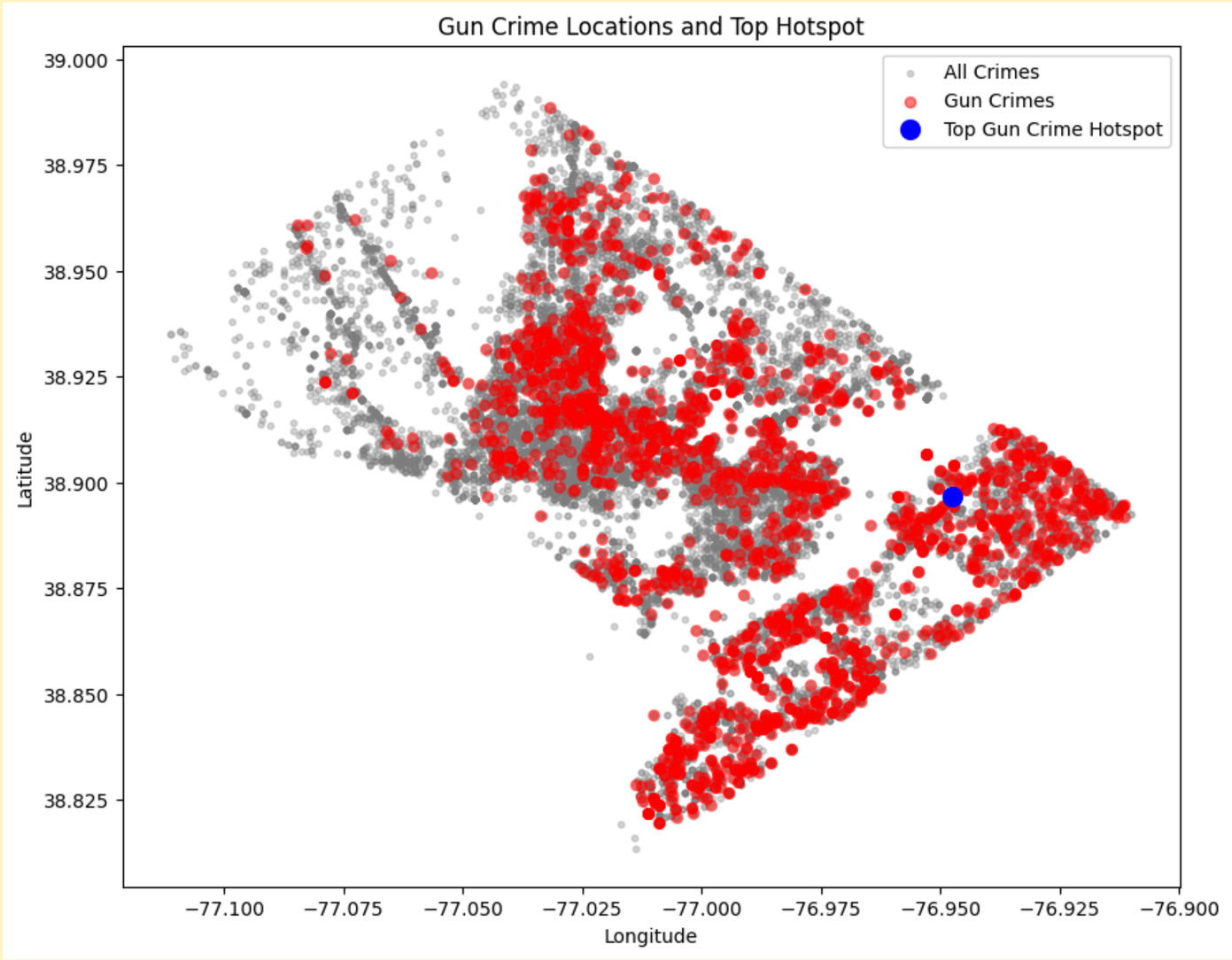
Results

```
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
```

```
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
```



Recommendations

These are the following recommendations that 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 the 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.

- **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.

References

Dataset Link: Click Here

Colab File Link: Click Here