Coursework 1 - Research Report & Visualisation

Submitted By Group 10

Group Members and Participation

Student Name	Student ID	Group Assignment Development Contribution Percentage(%)
Pratikraj Pavankumar Mugade	**2374911**	**35**
Srushti Parshuram Kore	**2375621**	**35**
Nikhil Sushil Muneshwar	**2372424**	**20**
Jay Dugad	**2344616**	**10**

Abstract

This project examines crime patterns through data science techniques, focusing on spatial analysis and identifying areas that could benefit from optimized policing resources. After thorough data preprocessing to address missing values, coordinate errors, and incorrect resolution times, we imputed values based on relationships between features like block name, city, and zip code to ensure data integrity.

Key analyses included spatial segmentation of crime incidents by regional quadrants and correlation assessments between fields like city, police district, and block name to uncover dependencies. Using visualizations, we explored crime distribution across areas with high incidents and varying police station densities. Guided by ten research questions, our approach provides insights into crime density, resource gaps, and prevalent incident types, contributing actionable findings for effective resource allocation and enhanced community safety.

1. Table of Contents

1. Introduction

Data science is a field that combines statistical analysis, machine learning, and computational methods to derive insights from data, supporting informed decision-making across sectors like finance, healthcare, and public safety. In law enforcement, data science is pivotal for understanding crime trends, identifying hotspots, and optimizing resource allocation. **(Cai, 2021)**

This project leverages a crime dataset capturing various details such as crime types, locations, police district information, and resolution times across a city. Analysing this data helps urban planners and public safety agencies assess crime patterns, high-risk areas, and police resource distribution. By applying data science methodologies, we aim to uncover spatial and temporal crime trends, offering insights to support data-driven crime prevention strategies and resource allocation for city safety.

1.1. Problem Statement:

In Maryland, crime rates and types vary widely across neighbourhoods, creating uneven demands on public safety resources. Some areas experience disproportionately high crime incidents, necessitating a detailed analysis of spatial and temporal crime patterns to better understand and address community safety needs.

1.2. Research Questions:

Table 1 Overall Research Questions

SR NO	Research Questions
1	Which **sectors** handle incidents most effectively in terms of time in which crime has been resolved?
2	Are there **specific times of day** when incidents occur more frequently?
3	Which **police districts or agencies** have the most **missing Resolution Time** entries(start-date time?
4	How does the **time required** to **respond** (Dispatch Date/Time to Start Date/Time) vary by Beat?
5	How is crime distributed **geographically** across various **citie**s, and what patterns emerge when examining **incidents by type, frequency, and seasonal variations**?
6	How do crime incident distributions vary across different **directional regions of the city,** and how does the **frequency** of these incidents differ by **geographic region**?
7	What are the key **crime hotspots in different cities**, and how do they vary **by crime type**?

8	How do **response times** vary across **cities and crime types**, and what are the underlying patterns in these variations?
9	How do **high-density crime areas**, identified **geographically by Zip Codes**, correlate with the types of crimes occurring within those regions?
10	How do **'Not a Crime' incidents** vary across different **Zip Codes and Place Types**, and what are the trends in incident **frequency** within these areas?
11	How does the distribution of crime incidents across **cities** compare to the **number of police stations** allocated, and do any **cities** exhibit a **need for additional police resources** based on crime rates?
12	Which **streets** experience the **highest frequency** of **crime incidents**, and what are the most **common types of crimes** on these **high-frequency streets**?

1.3. Objectives

- **Identify Crime Hotspots**: Pinpoint high-crime areas within Maryland's neighbourhoods to inform law enforcement resource allocation.
- **Analyse Spatial-Temporal Crime Patterns:** Examine crime variations over time and across regions to identify seasonal or geographic trends.
- **Evaluate Police Resource Distribution:** Assess the balance between police station density and crime frequency to spot areas potentially underserved by law enforcement.
- **Detect Anomalous Crime Records:** Identify and address anomalies in the dataset, such as unusual resolution times or missing data points, to improve data integrity.
- **Formulate Crime Prevention Insights:** Generate insights that may support crime prevention and community safety initiatives.

1.4. Intended Research Methodology

- **Data Preprocessing:**

  Cleaning: Handle missing values, correct erroneous entries, and validate geographic coordinates.

  Feature Engineering: Create new columns such as 'Directional Region' based on spatial coordinates to enable region-specific analysis.

- **Exploratory Data Analysis (EDA):**

  Visualization: Use scatter plots, box plots, and heatmaps to explore the dataset, identify patterns, and investigate relationships among variables.

  Geospatial Analysis: Visualize regional crime distributions using hexbin and directional maps.

- **Statistical Analysis:**

  Correlation & Chi-Square Tests: Use statistical tests to analyse relationships between features, such as between Block Name and Zip Code.

  Comparative Analysis: Examine the distribution of police stations against high-crime areas to assess resource distribution effectiveness.

- **Modelling & Interpretation:**

  Conduct spatial clustering if applicable, to understand crime densities better.

  Summarize findings and support them with visual and statistical interpretations.

- **Question Framing:**

  Define research questions focused on crime density, spatial distribution, and neighbourhood-specific insights to shape targeted analyses.

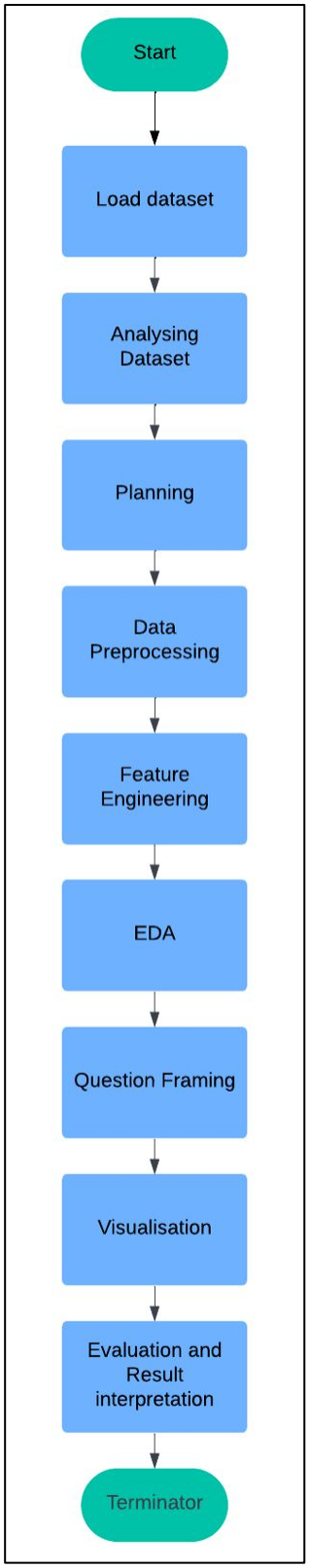

Figure 1 Workflow

1.5. Expected Outcome

- Detailed Mapping of Crime Hotspots: Visual identification of areas with high crime density, categorized by neighbourhood and region.
- Resource Allocation Insights: Identification of high-crime, low-police neighbourhoods, supporting targeted resource allocation.
- Temporal and Spatial Crime Patterns: Patterns that reveal the influence of time and location on crime rates, aiding predictive insights.
- Statistical Evidence of Feature Relationships: Validated correlations between variables, like Block Name and Zip Code, demonstrating the utility of imputed data for crime analysis.

2. Preliminary Data Analysis.

2.1 Dataset

The dataset utilized in this project, the USCrimesDataset from Kaggle, offers a comprehensive view of crime incidents across the United States. With 306,094 entries across 30 columns, it encapsulates crucial details such as incident ID, offense type, location specifics, date, time, and geographic coordinates. Each column highlights a unique dimension of crime data, enabling in-depth, multi-faceted analysis of patterns and trends within crime incidents. This structured dataset serves as a valuable resource for exploring spatial and temporal crime dynamics at a granular level across various U.S. regions. **(Bayoumi et al., 2018)**

Table 2 Brief Feature Info

Feature Name	*Description*	*Sample data*	*Unique Count*
Incident ID	A unique identifier for each reported crime incident, used to track and reference specific cases.	201202980	280928
Offence Code	A code corresponding to the type of offense or crime. These codes are used to categorize different types of crimes.	1115	619
CR Number	A code corresponding to the type of offense or crime. These codes are used to categorize different types of crimes.	190016809	280893
Dispatch Date / Time	The date and time when the police were dispatched to respond to the incident	08/23/2018 09:52:08 PM	235077
NIBRS Code	National Incident-Based Reporting System code, which classifies specific details about the crime for standardized reporting.	35A	58
Victims	National Incident-Based Reporting System code, which classifies specific details about the crime for standardized reporting.	1	10
Crime Name1	National Incident-Based Reporting System code, which classifies specific	Crime Against Society	5

	details about the crime for standardized reporting.		
Crime Name2	An additional description or subcategory of the crime, if applicable.	Shoplifting	59
Crime Name3	An additional description or subcategory of the crime, if applicable.	Drugs – marijuana – possess	336
Police District Name	An additional description or subcategory of the crime, if applicable.	CITY OF TAKOMA PARK	9
Block Address	An additional description or subcategory of the crime, if applicable.	8000 BLK BARRON ST	21503
City	An additional description or subcategory of the crime, if applicable.	GERMANTOWN	66
State	An additional description or subcategory of the crime, if applicable.	MD	11
Zip Code	The postal code of the location where the crime took place.	20874.0	186
Agency	The law enforcement agency that handled the case.	TPPD	8
Place	A more specific location description, like a neighbourhood or public place where the incident occurred.	Street – Residential	99
Sector	A geographic subdivision of the police district for resource allocation and tracking.	T	15
Beat	A smaller geographic area within a sector, assigned to specific patrol units.	3I2	56
PRA	"Patrol Reporting Area" designation within the police system.	702	1557
Address Number	The specific number of the location where the incident occurred, if available.	11100.0	396
Street Prefix	Directional information (e.g., N, S, E, W) associated with the street address.	N	4
Street Name	The name of the street where the incident took place.	RISING SUN	7777

Street Suffix	Additional information related to the street (e.g., alley, boulevard, road).	W	7
Street Type	The specific type of the street (e.g., Avenue, Lane, Street).	AVE	32
Start_Date_Time	The starting date and time of the incident, indicating when it began or was first reported.	08/23/2018 09:52:00 PM	221547
End_Date_Time	The ending date and time of the incident, indicating when it was resolved or ended.	06/15/2018 03:00:00 AM	101809
Latitude	The geographical latitude coordinates of the incident's location.	39.08429	44681
Longitude	The geographical longitude coordinates of the incident's location.	-77.2635	40035
Police District Number	A numeric identifier for the police district handling the case.	5D	15
Location	This could be a general description of the location of the incident, potentially encompassing fields like street and coordinates.	(39.1845, -77.2635)	33635

2.2 Data Quality Initial Assessment:

- Data Overview:
    - Using .info() to check the dataset's structure, confirming column names, data types, and counts of non-null entries.
    - Assessing memory usage and data type appropriateness (e.g., numeric for continuous data, datetime for date-related features).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306094 entries, 0 to 306093
Data columns (total 30 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Incident ID            306094 non-null  int64
 1   Offence Code           306094 non-null  object
 2   CR Number              306094 non-null  int64
 3   Dispatch Date / Time   257065 non-null  object
 4   NIBRS Code             306094 non-null  object
 5   Victims                306094 non-null  int64
 6   Crime Name1            305822 non-null  object
 7   Crime Name2            305822 non-null  object
 8   Crime Name3            305822 non-null  object
 9   Police District Name   306000 non-null  object
 10  Block Address          279888 non-null  object
 11  City                   304818 non-null  object
 12  State                  306094 non-null  object
 13  Zip Code               302915 non-null  float64
 14  Agency                 306094 non-null  object
 15  Place                  306094 non-null  object
 16  Sector                 304564 non-null  object
 17  Beat                   304564 non-null  object
 18  PRA                    305855 non-null  object
 19  Address Number         279985 non-null  float64
 20  Street Prefix          13631 non-null   object
 21  Street Name            306093 non-null  object
 22  Street Suffix          5432 non-null    object
 23  Street Type            305755 non-null  object
 24  Start_Date_Time        306094 non-null  object
 25  End_Date_Time          144436 non-null  object
 26  Latitude               306094 non-null  float64
 27  Longitude              306094 non-null  float64
 28  Police District Number 306094 non-null  object
 29  Location               306094 non-null  object
```

Figure 2 info() snippet

- Statistical Summary:
    - Applying .describe() to examine basic statistics for numeric columns, identifying ranges, means, and extreme values.
    - Looking for outliers or unusual values, such as zeros or negatives in non-logical columns (e.g., resolution time).

```
df.describe()
```

	Incident ID	CR Number	Victims	Zip Code	Address Number	Latitude	Longitude
count	3.060940e+05	3.060940e+05	306094.000000	302915.000000	2.799850e+05	306094.000000	306094.000000
mean	2.012369e+08	1.692787e+08	1.022692	20876.535939	8.393001e+03	38.146328	-75.269490
std	8.626185e+04	6.210304e+07	0.192311	170.157722	1.526296e+04	5.974082	11.786118
min	2.010871e+08	1.001107e+07	1.000000	6.000000	1.000000e+00	0.000000	-77.516753
25%	2.011625e+08	1.705431e+08	1.000000	20853.000000	1.600000e+03	39.020392	-77.197117
50%	2.012361e+08	1.900183e+08	1.000000	20878.000000	8.100000e+03	39.072844	-77.099464
75%	2.013109e+08	2.000466e+08	1.000000	20904.000000	1.250000e+04	39.142072	-77.029046
max	2.013872e+08	2.204211e+08	22.000000	29882.000000	2.090600e+06	90.000000	0.000000

Figure 4 describe() snippet

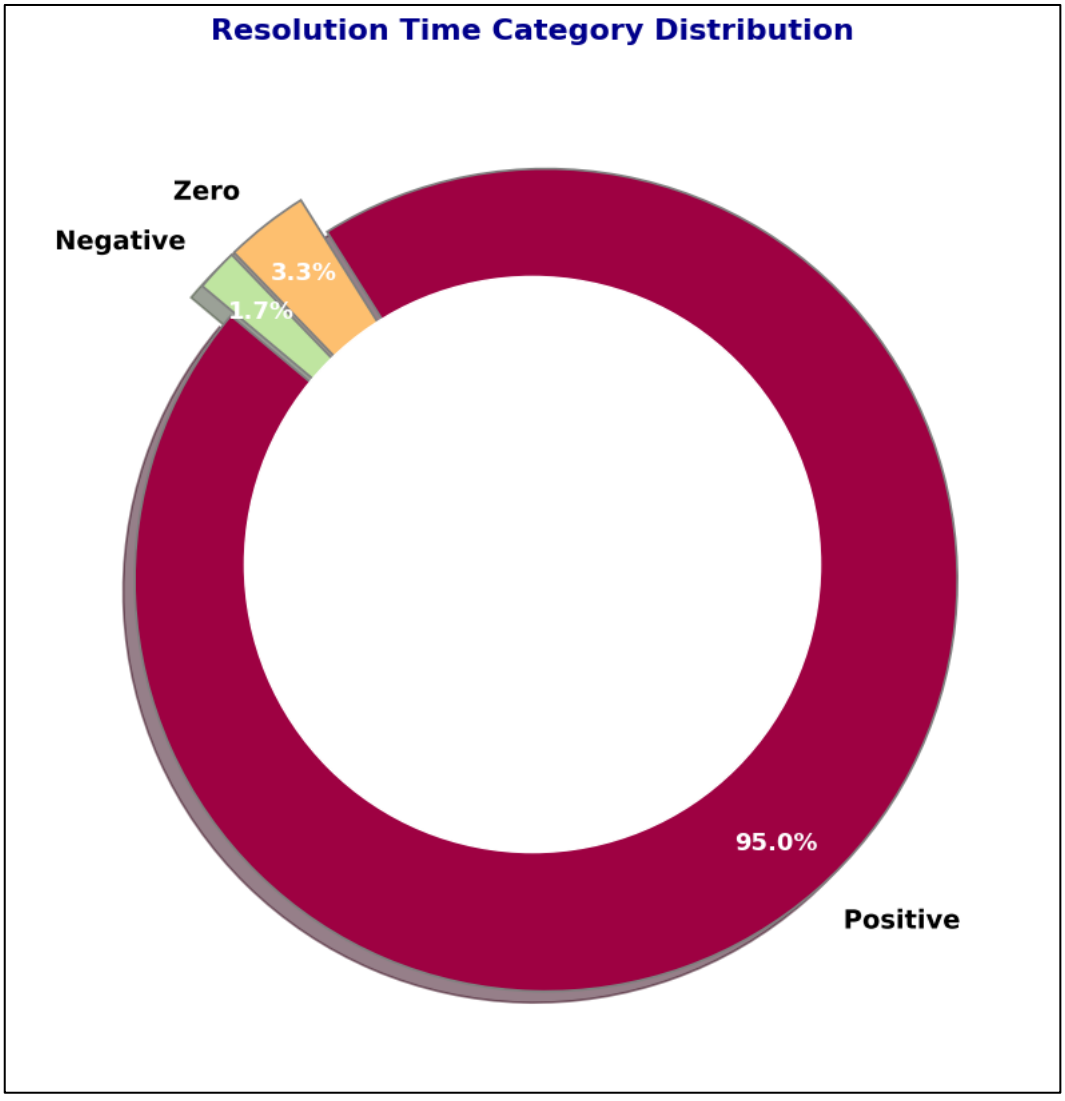

Figure 3 Pie chart of Resolution time

- Missing Values:
    - Counting the missing values in each feature to identify any data gaps.

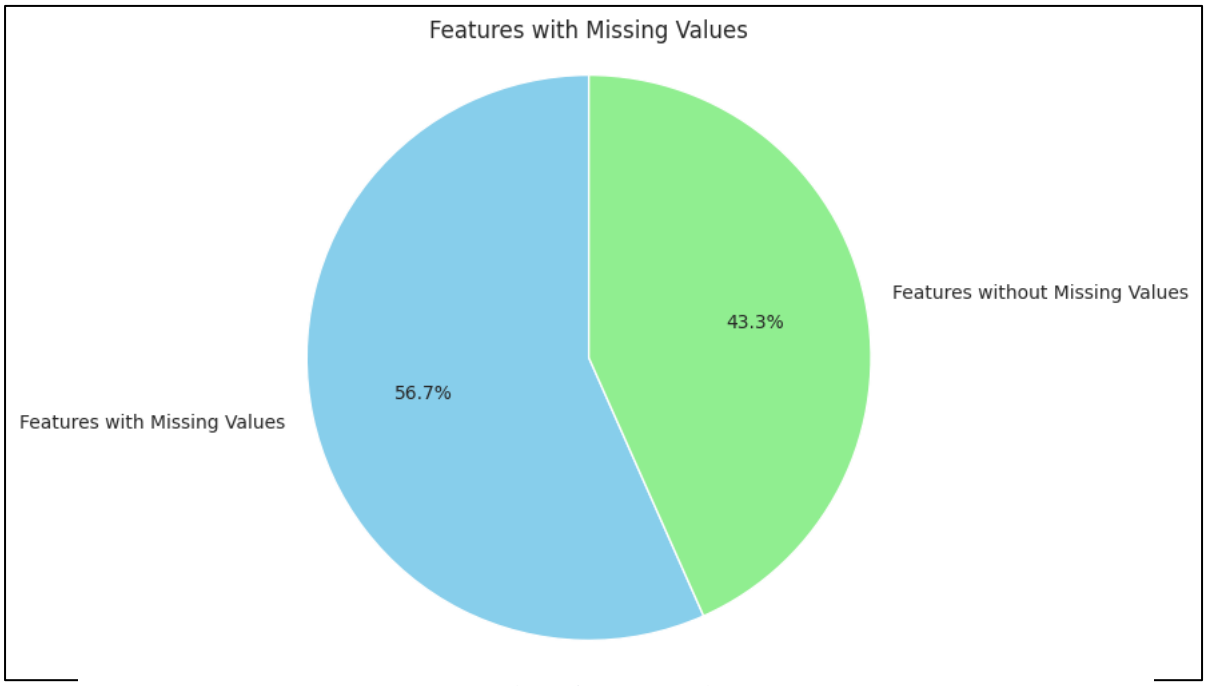

Figure 5 Pie chart of Features with missing status

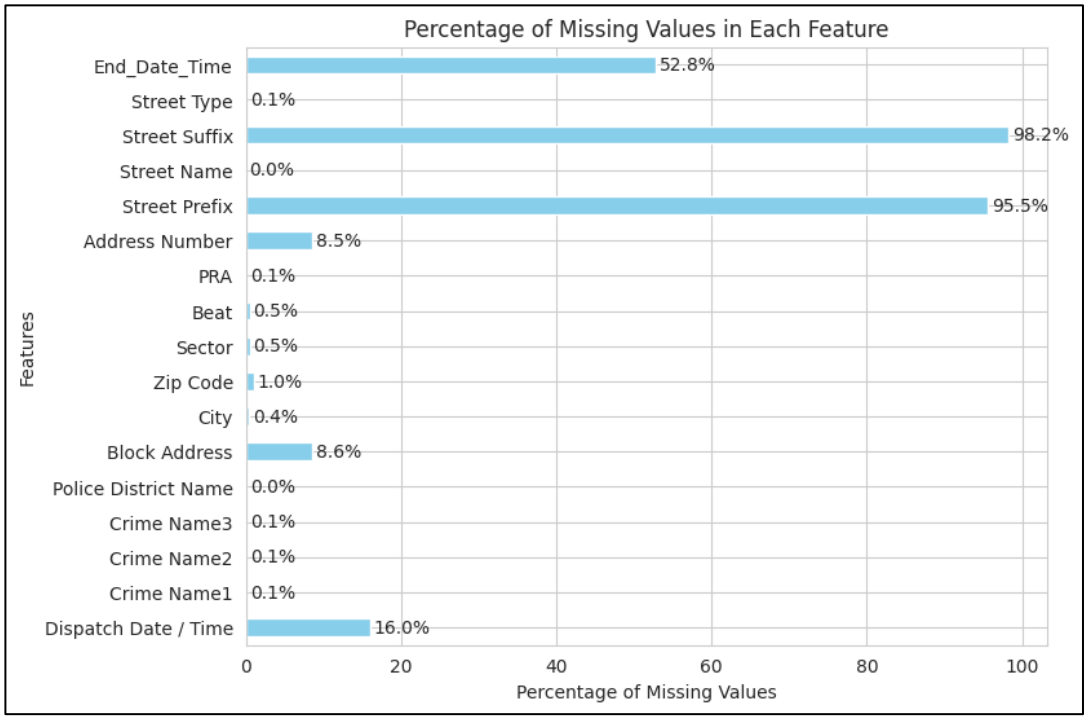

Figure 6 Bar graph representing missing values in each feature

- Unique Values:
    - Checking for the number of unique values in categorical features to verify consistency.

```
# Get the count of unique values in each column
unique_counts = df.nunique()

# Display the unique counts for each feature
print(unique_counts)
```

```
Incident ID               280928
Offence Code                 619
CR Number                 280893
Dispatch Date / Time      235077
NIBRS Code                    58
Victims                       10
Crime Name1                    5
Crime Name2                   59
Crime Name3                  336
Police District Name           9
Block Address              21503
City                          66
State                         11
Zip Code                     186
Agency                         8
Place                         99
Sector                        15
Beat                          56
PRA                         1557
Address Number               396
Street Prefix                  4
Street Name                 7777
Street Suffix                  7
Street Type                   32
Start_Date_Time           221547
End_Date_Time             101809
Latitude                   44681
Longitude                  40035
Police District Number        15
Location                   33635
```

Figure 7 unique_counts snippet

2.3 Main issues identified within the dataset:

- Features with data type mismatch
- Abundant missing/Null values
- Improper entries
    - Mistaken start date time and end date time which caused negative response time (newly derived feature)
    - Some entries with (0,0) co-ordinates
    - No proper formatting for time and date columns

2.4 Proposed Solutions:

Table 3 Overview of Proposed Solutions

Feature Name	*Null Count (in %)*	*Action Taken*
Street Suffix	98.2	Merged in the Street Name
Street Prefix	95.5	Merged in the Street Name
Street Type	0.1	Deleted null entries
Address Number	8.5	Deleted null entries
PRA	0.1	Deleted null entries
Beat & Sector	0.5	Deleted null entries
City	0.4	Imputed by using Block Address
Zip code	1.0	Imputed by using Block Address
Block address	8.6	Deleted null entries
Police District Name	0.0	Imputed by using City
Crime Name1& CrimeName2 & Crime Name 3	0.1	Kept as it is for special type of visualisation
End_Date_Time	52.8	Kept as it is for special type of visualisation
Dispatch Date / Time	16.0	Deleted null entries

2.5 Data Wrangling Skills

- Newly Derived features:
    - Resolution Time = (Start_Date_Time – End_Date_Time)

```
df['Start Date'] = df['Start Date'].astype(str)
df['Start Time'] = df['Start Time'].astype(str)
df['End Date'] = df['End Date'].astype(str)
df['End Time'] = df['End Time'].astype(str)

# Combine date and time for Start and End
df['Start_Date_Time'] = pd.to_datetime(df['Start Date'] + ' ' + df['Start Time'])
df['End_Date_Time'] = pd.to_datetime(df['End Date'] + ' ' + df['End Time'], errors='coerce')

# Calculate the duration (time required to resolve the case)
df['Resolution Time'] = df['End_Date_Time'] - df['Start_Date_Time']

# Display the updated DataFrame
print(df[['Start_Date_Time', 'End_Date_Time', 'Resolution Time']])
```

Figure 8 Resolution Time snippets

- Block Name= (Extracted from Block Address)

```
# Extract the block name part after 'BLK' from the Block Address
df['Block Name'] = df['Block Address'].str.extract(r'BLK\s+(.+)', expand=False)

# Display the updated DataFrame to verify the Block Name extraction
print(df[['Block Address', 'Block Name']].head())
```

```
                Block Address         Block Name
0   12800 BLK  MIDDLEBROOK RD     MIDDLEBROOK RD
1      8300 BLK  WOODMONT AVE       WOODMONT AVE
2      8300 BLK  WOODMONT AVE       WOODMONT AVE
3  400 BLK  QUINCE ORCHARD RD  QUINCE ORCHARD RD
4      4800 BLK  FALSTONE AVE       FALSTONE AVE
```

Figure 9 Block Name snippet

- Response Time= (Start DateTime – Dispatch DateTime)

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df_cleaned = pd.read_csv('/content/drive/MyDrive/cleaned_dataset (1).csv')

# Ensure that date columns are in datetime format and time columns are treated as strings
df_cleaned['Dispatch Date'] = pd.to_datetime(df_cleaned['Dispatch Date'], errors='coerce').dt.date
df_cleaned['Start Date'] = pd.to_datetime(df_cleaned['Start Date'], errors='coerce').dt.date
df_cleaned['Dispatch Time'] = pd.to_datetime(df_cleaned['Dispatch Time'], errors='coerce').dt.time
df_cleaned['Start Time'] = pd.to_datetime(df_cleaned['Start Time'], errors='coerce').dt.time

# Convert dates and times to strings before concatenating
df_cleaned['Dispatch DateTime'] = pd.to_datetime(df_cleaned['Dispatch Date'].astype(str) + ' ' + df_cleaned['Dispatch Time'].astype(str), errors='coerce')
df_cleaned['Start DateTime'] = pd.to_datetime(df_cleaned['Start Date'].astype(str) + ' ' + df_cleaned['Start Time'].astype(str), errors='coerce')

# Calculate response time in minutes
df_cleaned['Response Time'] = (df_cleaned['Start DateTime'] - df_cleaned['Dispatch DateTime']).dt.total_seconds() / 60

# Filter out rows with negative response times and drop rows with null values if any
df_cleaned = df_cleaned[(df_cleaned['Response Time'] > 0) & df_cleaned['Response Time'].notna()]
```

Figure 10 Response Time snippet

- Separated Features from existing features.
    - Separated 'Start_Date_Time' into **'Start Date'** and **'Start Time'**
    - Separated 'Dispatch Date / Time' into **'Dispatch Date'** and **'Dispatch Time'**
    - Separated 'End_Date_Time' into **'End Date'** and **'End Time'**
    - Dropped the original combined date-time columns as no longer needed.

```
# Convert 'Start_Date_Time', 'Dispatch Date / Time', and 'End_Date_Time' columns to datetime format
df['Start_Date_Time'] = pd.to_datetime(df['Start_Date_Time'])
df['Dispatch Date / Time'] = pd.to_datetime(df['Dispatch Date / Time'])
df['End_Date_Time'] = pd.to_datetime(df['End_Date_Time'])

# Separate 'Start_Date_Time' into 'Start Date' and 'Start Time'
df['Start Date'] = df['Start_Date_Time'].dt.date
df['Start Time'] = df['Start_Date_Time'].dt.strftime('%H:%M')  # 24-hour format

# Separate 'Dispatch Date / Time' into 'Dispatch Date' and 'Dispatch Time'
df['Dispatch Date'] = df['Dispatch Date / Time'].dt.date
df['Dispatch Time'] = df['Dispatch Date / Time'].dt.strftime('%H:%M')  # 24-hour format

# Separate 'End_Date_Time' into 'End Date' and 'End Time'
df['End Date'] = df['End_Date_Time'].dt.date
df['End Time'] = df['End_Date_Time'].dt.strftime('%H:%M')  # 24-hour format

# Drop the original combined date-time columns if no longer needed
df = df.drop(columns=['Start_Date_Time', 'Dispatch Date / Time', 'End_Date_Time'])

# Display the updated DataFrame
print(df)
```

Figure 11 Date & Time Separating Features snippets

- Merging Highly corelated features:
    - It is observed that both 'Street Prefix' and 'Street Suffix' are mutually corelated to the 'Street Name' Feature. Also, they do exhibit some importance so we can't just delete them.

```
# Combine 'Street Prefix', 'Street Name', and 'Street Suffix' intelligently
# Where 'Street Suffix' is null, use 'Street Prefix', otherwise use 'Street Suffix'
df['Street Name'] = df['Street Prefix'].fillna('') + ' ' + df['Street Name'] + ' ' + df['Street Suffix'].fillna('')

# Drop 'Street Prefix' and 'Street Suffix' columns after merging
df = df.drop(columns=['Street Prefix', 'Street Suffix'])

# Remove any extra spaces that might have been introduced
df['Street Name'] = df['Street Name'].str.strip()

# Display the updated DataFrame
print(df)
```

Figure 12 Merging features snippet

- Formatting all the date and time related features:

  As per above fig. 10.

3. Exploratory Data Analysis:

3.1 Introduction

EDA helps uncover patterns **(Mukhiya and Ahmed, 2020)**, spot anomalies, test assumptions, and determine relationships within the dataset. It provides a foundational understanding that guides feature engineering, model selection, and further data processing.

3.2 Assumptions:

- **(0,0) Coordinates**:

  Entries with (0,0) coordinates likely represent missing location data or errors, which will need handling to avoid issues in spatial analysis**.**

- **Correlation of Geographic Features**:

  City and Block Name, Police District Name and City, as well as Zip Code and Block Name, are assumed to be strongly correlated and can assist in imputing missing values.

- **Retention of Street Prefix and Suffix**:

  Despite high missing rates, Street Prefix and Suffix are kept for potential street-level insights and spatial patterns.

- **Temporal Consistency**:

  End Date Time should always be after Start Date Time, ensuring that derived Resolution Time values are positive.

- **Positive Resolution Time**:

  Resolution Time must be positive, as zero or negative values are unrealistic and indicate data issues needing correction.

3.3 Descriptive Statistics: Handling Null Values

- **Null Value Removal**:

  Removed entries with null values where the percentage of missing data was low, as these minimal entries do not significantly impact the dataset's integrity.

  Null entry removed features list:

  1) Dispatch Date

2) Beat
3) Sector
4) PRA
5) Zip Code
6) Block Address
7) Address Number

- **Correlation-Based Imputation**:
  Based on the assumptions, correlation analysis was conducted to validate relationships among features, confirming the following strong associations based on correlation tests and Chi-square test:

1) City and Block Name

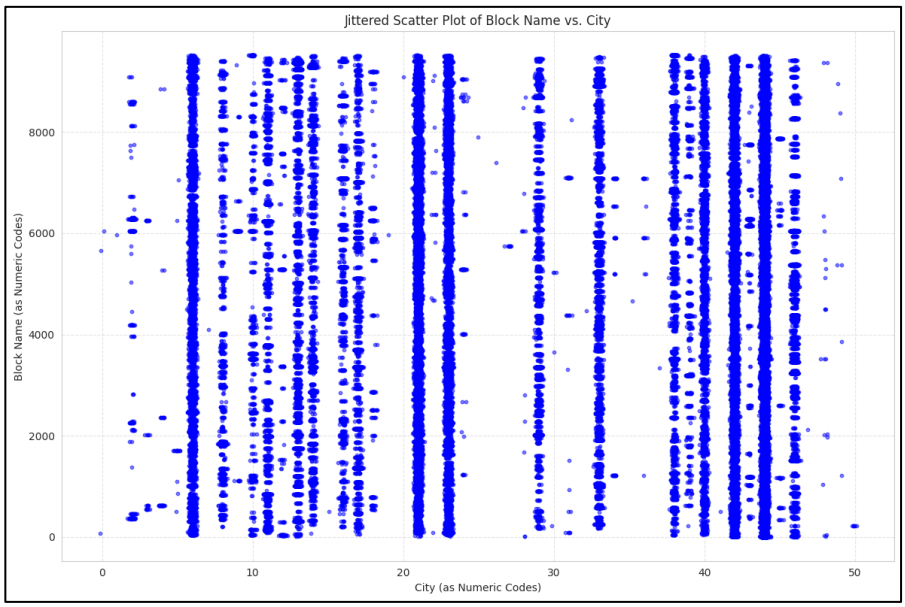

Figure 13 Jittered Scatter Plot of Block Name vs. City

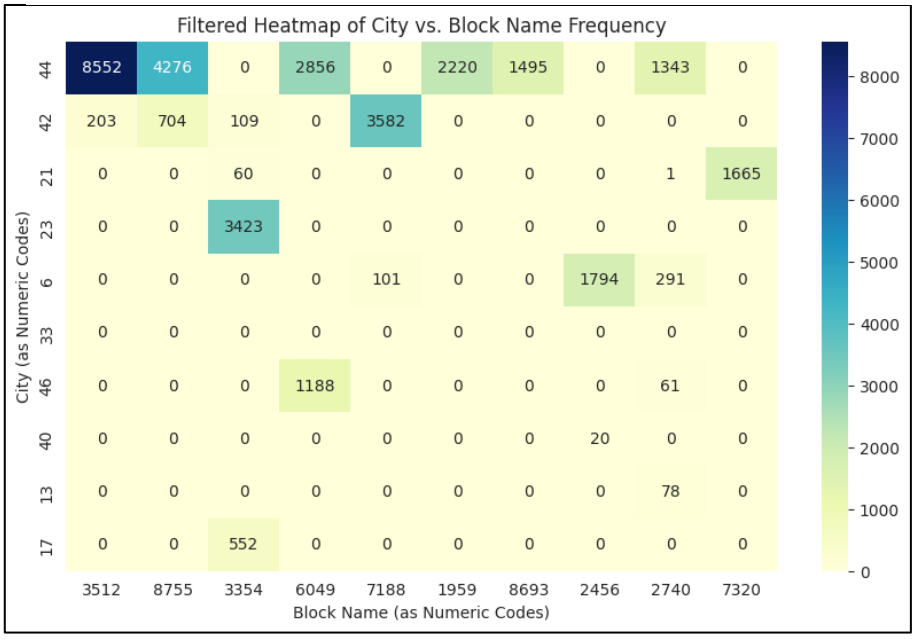

Figure 14 Heatmap of City vs. Block Name Frequency

2) Police District Name and City:

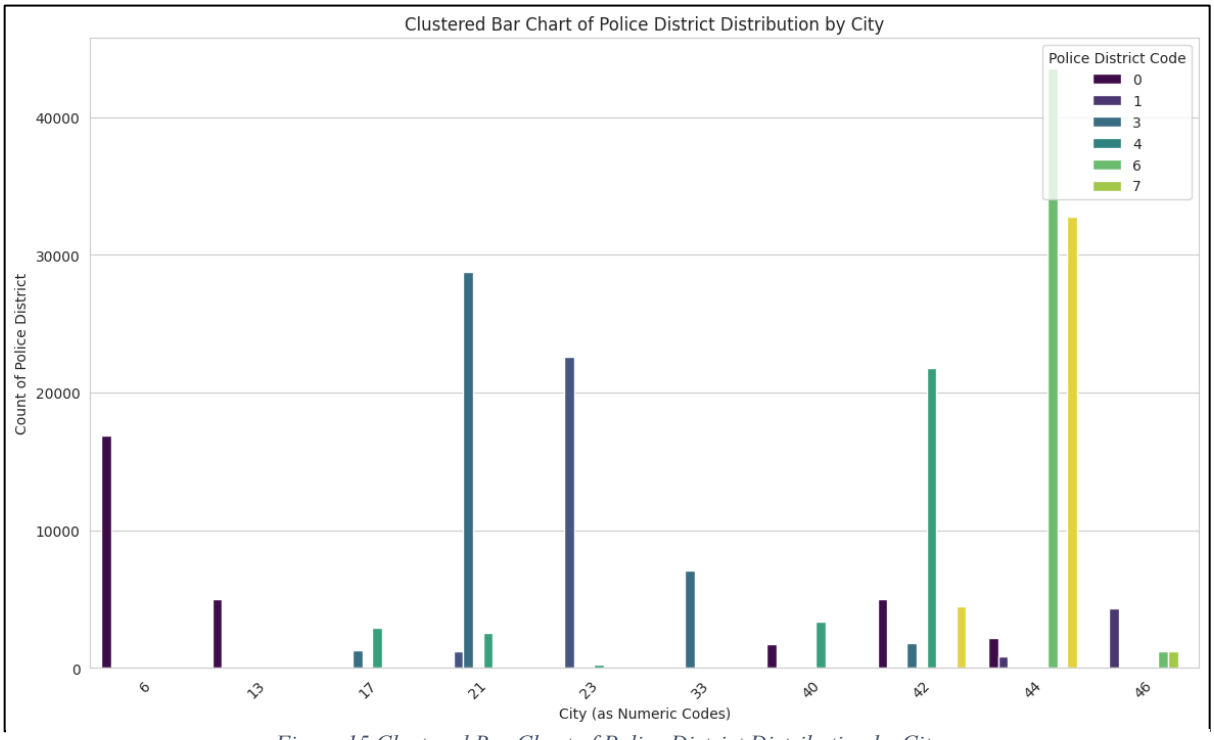

Figure 15 Clustered Bar Chart of Police District Distribution by City

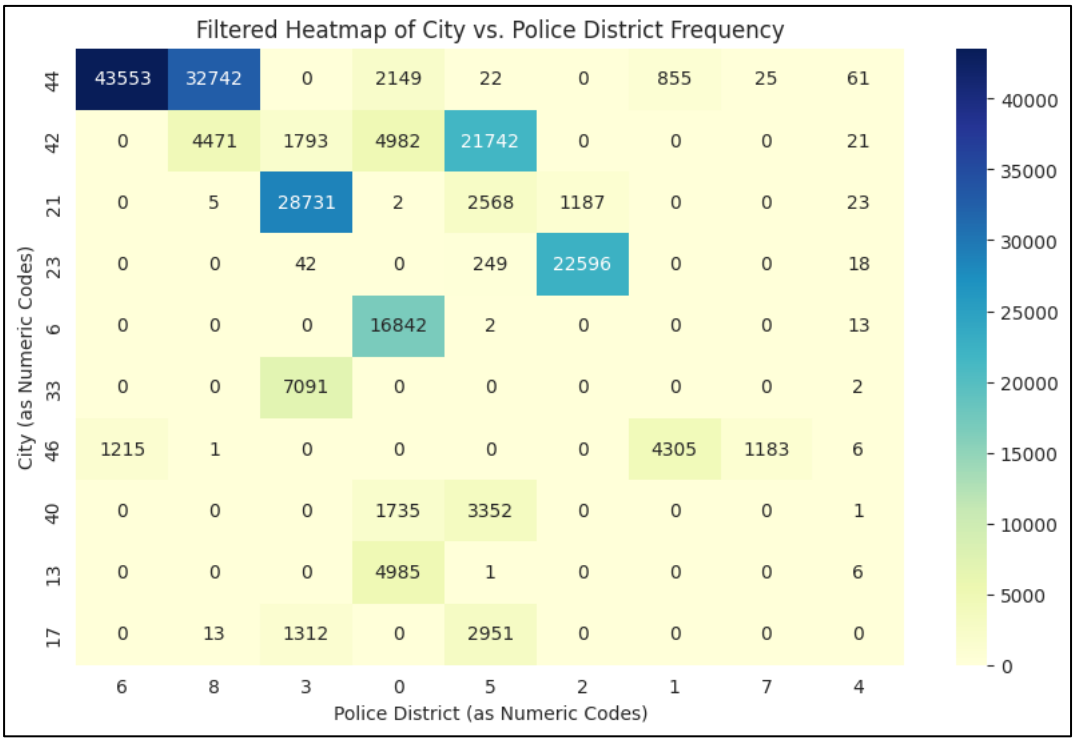

Figure 16 Heatmap of City vs. Police District Frequency

3) Zip Code and Block Address

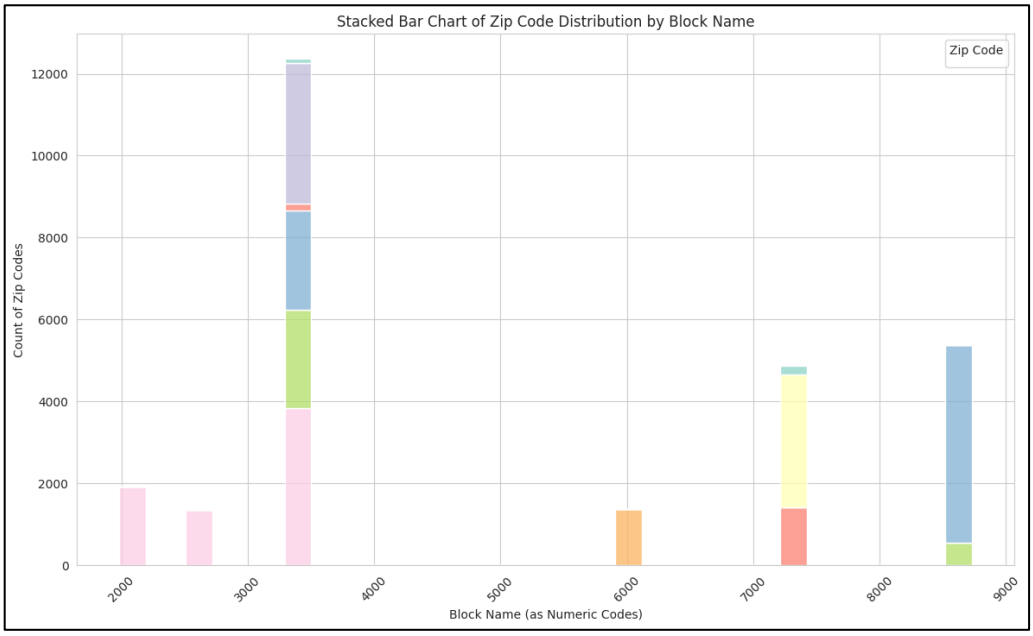

Figure 17 Stacked Bar Chart of Zip Code Distribution by Block Name

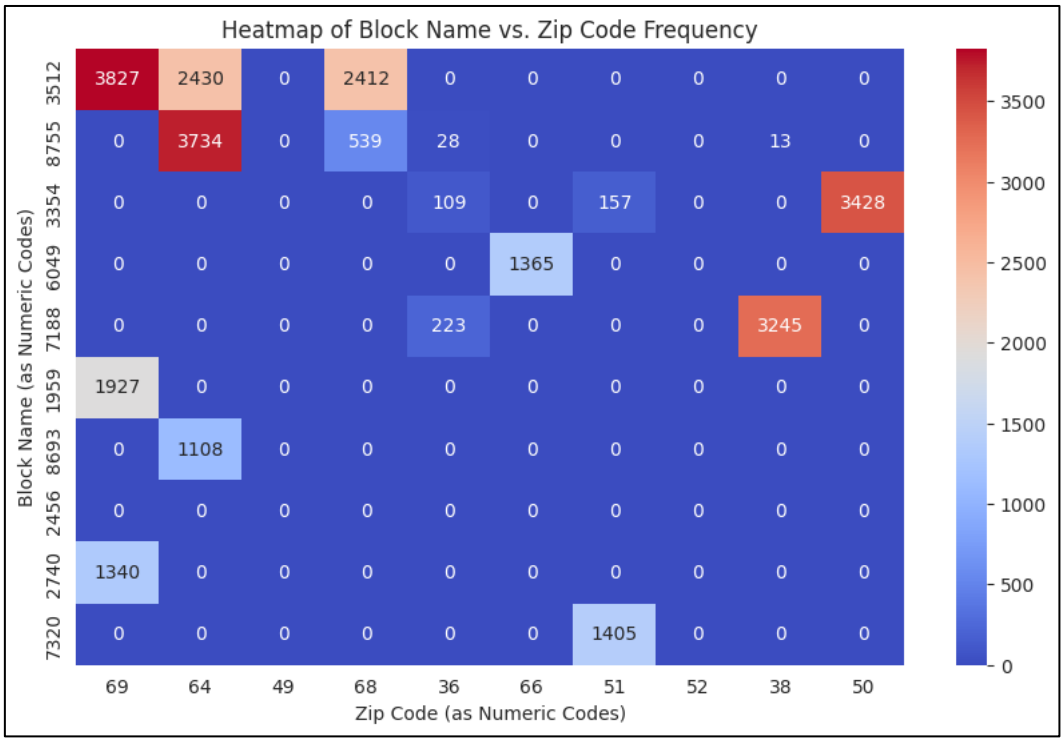

Figure 18 Heatmap of Block Name vs. Zip Code Frequency

Based on above all graphs of Jittered Scatter Plot, Cluster Bar chart and stacked bar chart from fig. 13, 15 and 17, we can conclude there is a positive correlation between these features. Even

the Heat maps from fig. 14, 16 and 18, even bolsters the evidence by showing high dense cluster of numbers.

- **Imputation Strategy**:

1) **City**: Missing values in City were filled based on Block Name, leveraging their observed correlation using mode and lambda for mapping these values.

```
# Create a dictionary to map Block Name to the most frequent City
block_to_city = df.groupby('Block Name')['City'].apply(lambda x: x.mode()[0] if not x.mode().empty else None).to_dict()

# Fill missing City values using the Block Name
df['City'] = df.apply(lambda row: block_to_city.get(row['Block Name']) if pd.isnull(row['City']) else row['City'], axis=1)

# Verify the imputation
print(df[['Block Name', 'City']].head())
```

```
          Block Name          City
0     MIDDLEBROOK RD    GERMANTOWN
1       WOODMONT AVE      BETHESDA
2       WOODMONT AVE      BETHESDA
3  QUINCE ORCHARD RD  GAITHERSBURG
4       FALSTONE AVE   CHEVY CHASE
```

Figure 19 City imputation snippet

2) **Police District Name**: Imputed using the City feature, as analysis showed a high dependency between these two attributes.

```
# Create a dictionary to map City to the most frequent Police District Name
city_to_district = df.groupby('City')['Police District Name'].apply(lambda x: x.mode()[0] if not x.mode().empty else None).to_dict()

# Display the mapping to verify
print(city_to_district)
```

```
SPRING', 'ALEXANDRIA': 'OTHER', 'ASHTON': 'WHEATON', 'BARNESVILLE': 'GERMANTOWN', 'BEALLSVILLE': 'ROCKVILLE', 'BELTSVILLE': 'SILVER SPRING', 'BETHESDA': 'BETHESD
```

```
# Fill missing Police District Name values using the most frequent district for each city
df['Police District Name'] = df.apply(lambda row: city_to_district.get(row['City']) if pd.isnull(row['Police District Name']) else row['Police District Name'],

# Verify the imputation
print(df[['City', 'Police District Name']].head())
```

```
           City Police District Name
0    GERMANTOWN           GERMANTOWN
1      BETHESDA             BETHESDA
2      BETHESDA             BETHESDA
3  GAITHERSBURG   MONTGOMERY VILLAGE
4   CHEVY CHASE             BETHESDA
```

Figure 20 Police District Name imputation snippet

3) **Zip Code**: Imputed using Block Address, supported by strong correlation evidence.

```
# Create a dictionary to map City to the most frequent Zip Code
city_to_zip = df.groupby('Block Name')['Zip Code'].apply(lambda x: x.mode()[0] if not x.mode().empty else None).to_dict()

# Display the mapping to verify
print(city_to_zip)
```

```
{'10TH AVE': 20903.0, '11TH AVE': 20903.0, '12TH AVE': 20903.0, '13TH AVE': 20912.0, '13TH PL': 20912.0, '13TH ST': 20910.0, '13TH ST NW': 20004.0, '14TH AVE':
```

```
# Fill missing Zip Code values using the most frequent Zip Code for each city
df['Zip Code'] = df.apply(lambda row: city_to_zip.get(row['Block Name']) if pd.isnull(row['Zip Code']) else row['Zip Code'], axis=1)

# Verify the imputation
print(df[['Block Name', 'Zip Code']].head())
```

```
          Block Name  Zip Code
0     MIDDLEBROOK RD   20874.0
1       WOODMONT AVE   20814.0
2       WOODMONT AVE   20814.0
```

Figure 21 Zip code imputation snippet

3.4 Remaining Missing Values After EDA

Despite extensive imputation, some features still contain missing values, which will be preserved for specific visualisations and analyses:

- **Resolution Time**: 51.7% of entries lack this data, and due to its dependency on accurate event start and end times, it could not be imputed through strategies like mode or forward/backward filling.
- **Crime Name1, Crime Name2, Crime Name3**: These features still have 0.1% missing values, and due to the variability in crime type classifications, these values couldn't be reliably imputed.

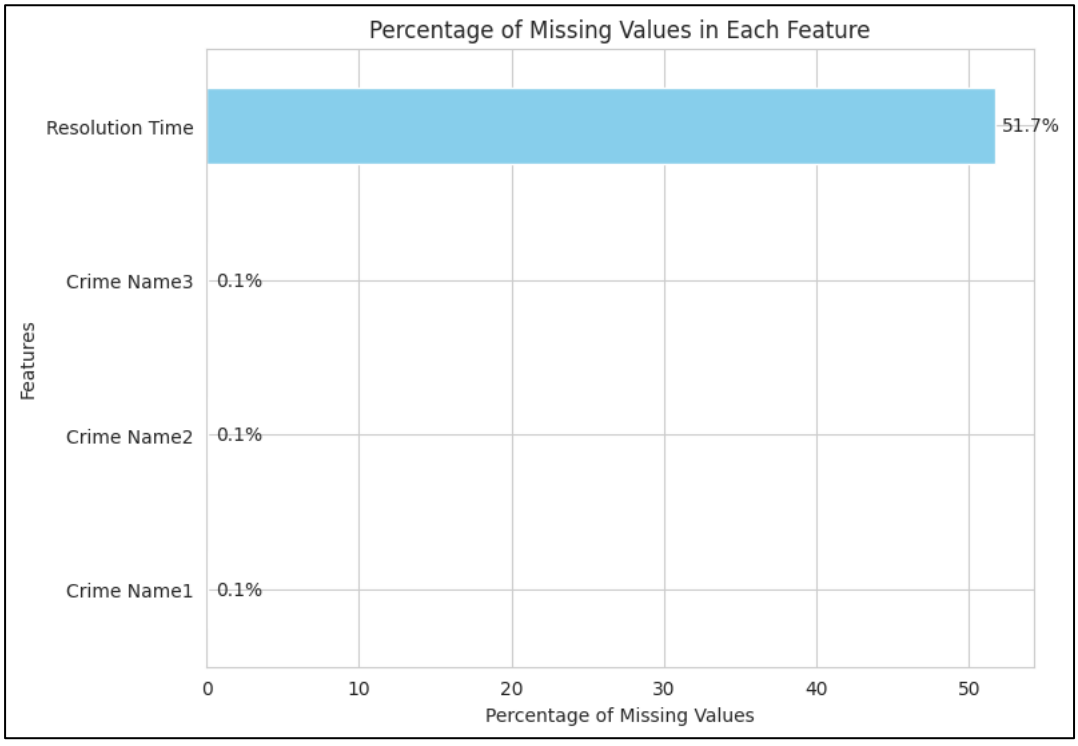

Figure 22 Bar of percentage of missing values in each feature after performing EDA

3.5 Justification for Retaining Missing Values

These unfilled values will be used in subsequent visualizations and analyses, reflecting the limitations within the dataset and offering insights into missing data patterns without compromising the validity of our approach.

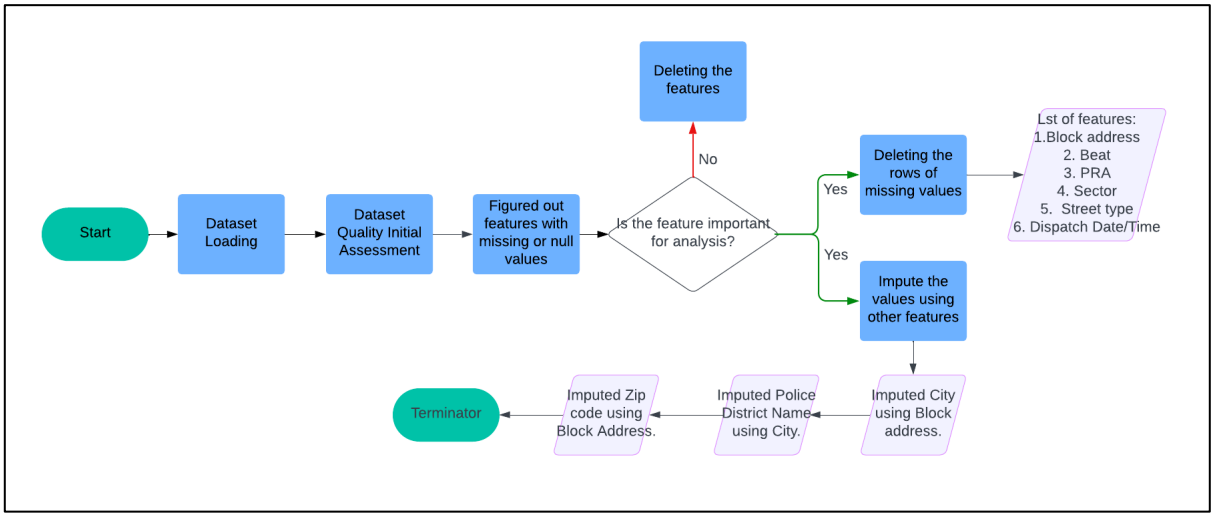

Figure 23 Flowchart of Overall EDA performed

4. Data Visualisations and Conclusions.

Table 4 Overview of Chart Types

Research Question No.	*Chart Type(s)*	**Purpose**	**Strengths**
1	Bar Plot, Box Plot	Compares the average and distribution of resolution times by sector.	Bar plot highlights averages, while box plot shows distribution, variability, and outliers.
2	Line Chart, Heatmap	Line chart shows overall frequency trends, while heatmap reveals time-of-day and day-of-week patterns	Line chart shows hourly trends, heatmap enables insight into complex time combinations.
3	Bar Chart, Heatmap	Identifies police districts and agencies with missing data patterns.	Bar chart for absolute counts, heatmap for district-agency combinations.
4	Box Plot, Bar Chart	Compares response times across different beats, showing both variability and averages.	Box plot highlights distribution by beat, bar chart shows average times for clearer comparison.
5	Scatter Plot (Geographic)	Maps incidents by latitude and longitude, identifying spatial distributions.	Geographic scatter plot enables spatial pattern analysis
6	Scatter Plot (Directional Regions)	Visualizes incidents by geographic region (e.g., northeast, southwest).	Scatter plot visualizes regional distribution, adding directional understanding.
7	Scatter Plot (with Hotspot Annotations)	Maps major hotspots by type, labelled for different cities.	Annotated scatter plot shows high-density areas, aiding hotspot identification by type.
8	Heatmap, Bar Chart	Compares average response times by city and type.	Heatmap enables pattern identification across cities, bar chart for specific response times
9	Hexbin Map, Stacked Bar Chart	Hexbin map for incident density by zip code, stacked bar chart for type distribution.	Hexbin highlights high-density areas, bar chart clarifies crime type proportions in these
10	Bar Chart	Shows the distribution of 'Not a Crime'	Highlights non-crime location patterns effectively with a bar chart.

		incidents by location type and zip code.	
11	Scatter Plot, Box Plot	Plots crime vs. police station counts by city, showing the allocation of resources.	Scatter plot for overall resource-crime relationship, box plot for city-level distribution of police resources.
12	Bar Chart, Heatmap	Identifies high-crime streets and their common crime types.	Bar chart for frequency by street, heatmap for crime type patterns on high-frequency streets.

1) Which sectors handle incidents most effectively in terms of Resolution Time?

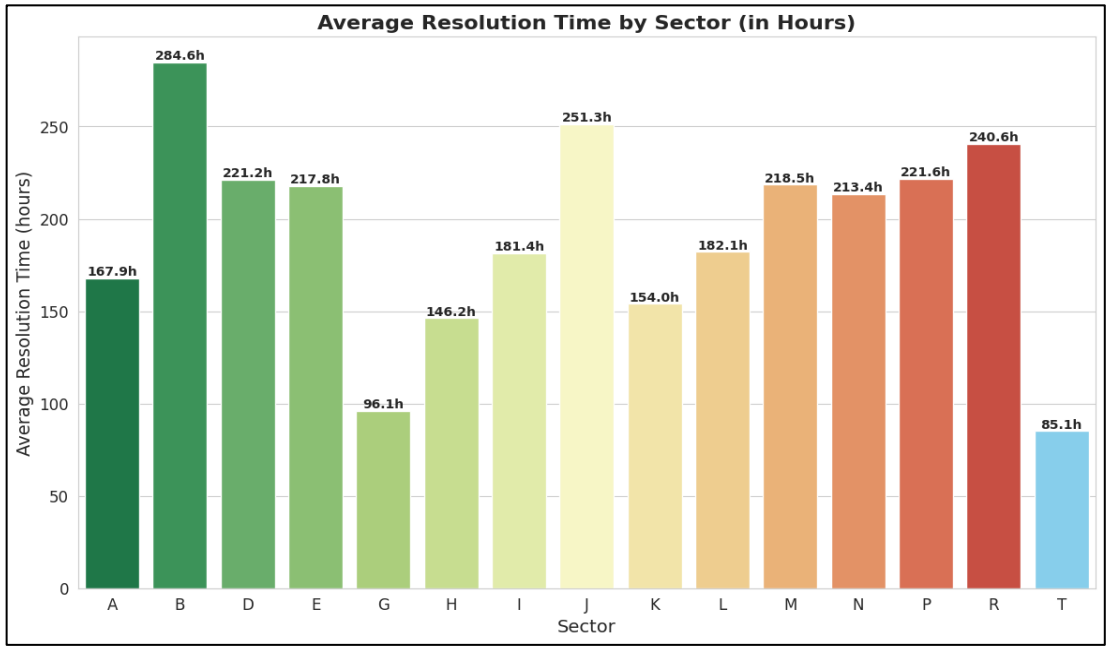

Figure 25 Bar graph Average resolution time by sector

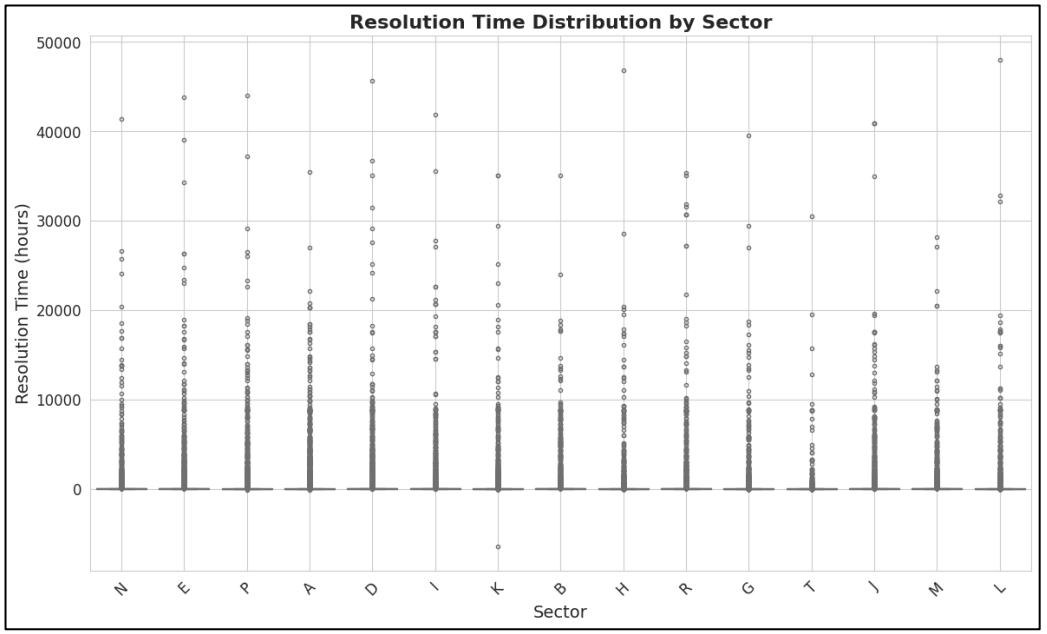

Figure 24 Box plot of Resolution Time Distribution by sector

Figures 24 and 25 show that Sector T has the lowest average resolution time at 85.1 hours. The box plot confirms a consistent distribution, supporting the conclusion of Sector T's efficient incident handling.

2) Are there specific times of day when incidents occur more frequently?

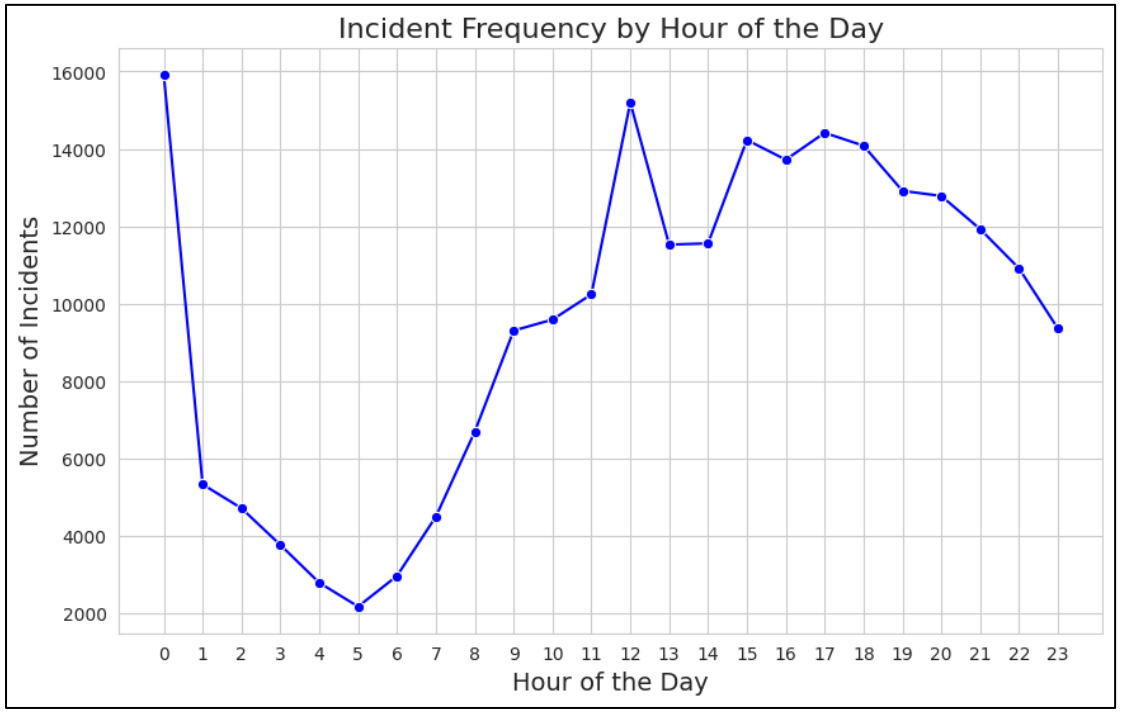

Figure 27 Line graph of Incident Frequency by Hour of the day

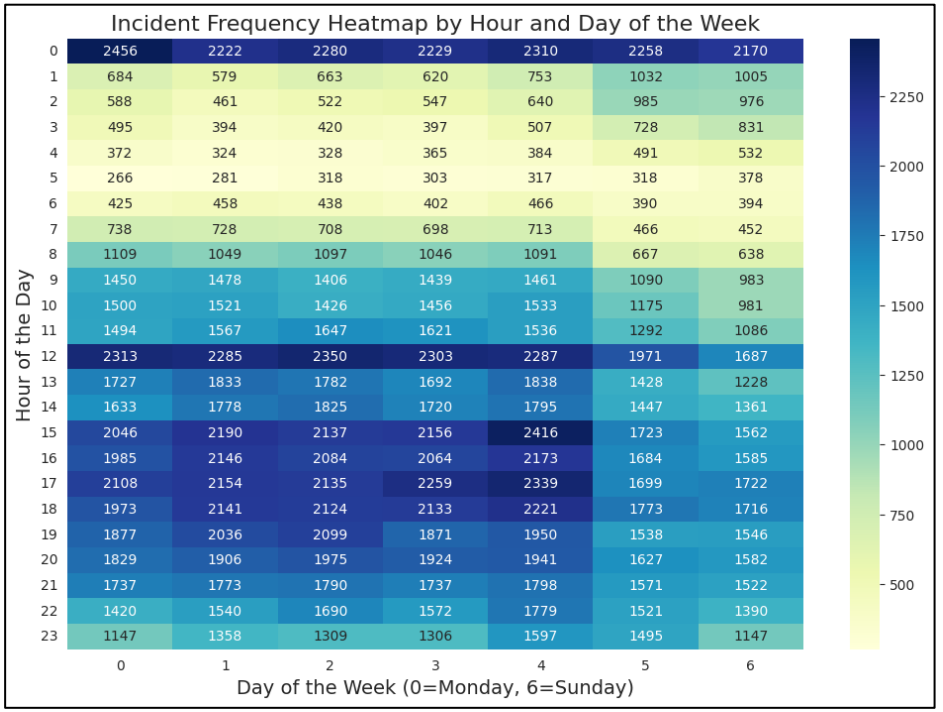

Figure 26 Heatmap of Incident Frequency by Hour & Day of week

Figure 26 reveals peak incident frequencies at midnight, followed by notable rises at 12:00, 15:00, and 17:00. Figure 27, a heatmap, highlights Friday as the day with the highest density of incidents, with elevated activity continuing into the weekend

3) Which police districts or agencies have the most missing Resolution Time entries?

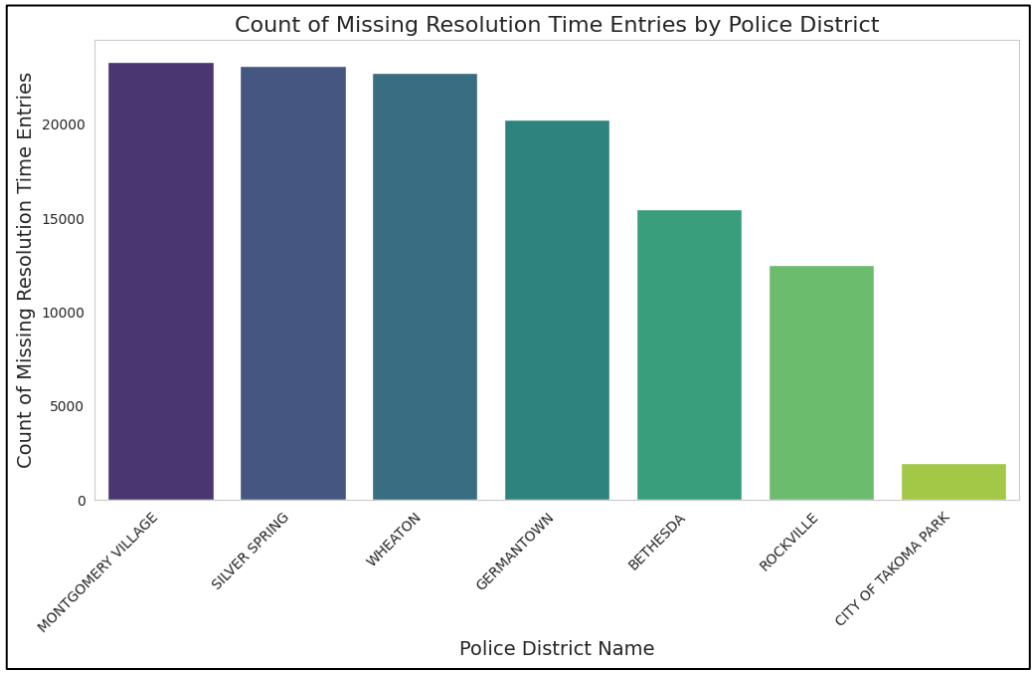

Figure 28 Bar graph of count of missing Resolution time entries by Police District

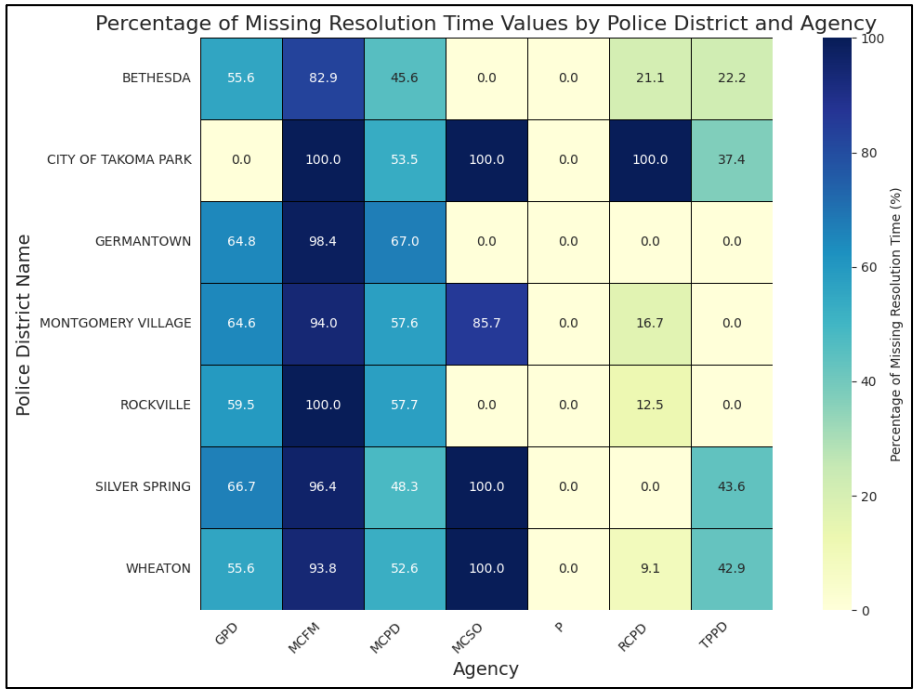

Figure 29 Heatmap of percentage of missing Resolution time by Police District and Agency

Figure 28 shows Montgomery Village, Silver Spring, and Wheaton with the highest counts of missing resolution time entries. Figure 29 further indicates that the MCFM agency is primarily associated with these significant missing values, highlighting a possible reporting gap in these areas.

4) How does the time required to respond (Dispatch Date/Time to Start Date/Time) vary by Beat?

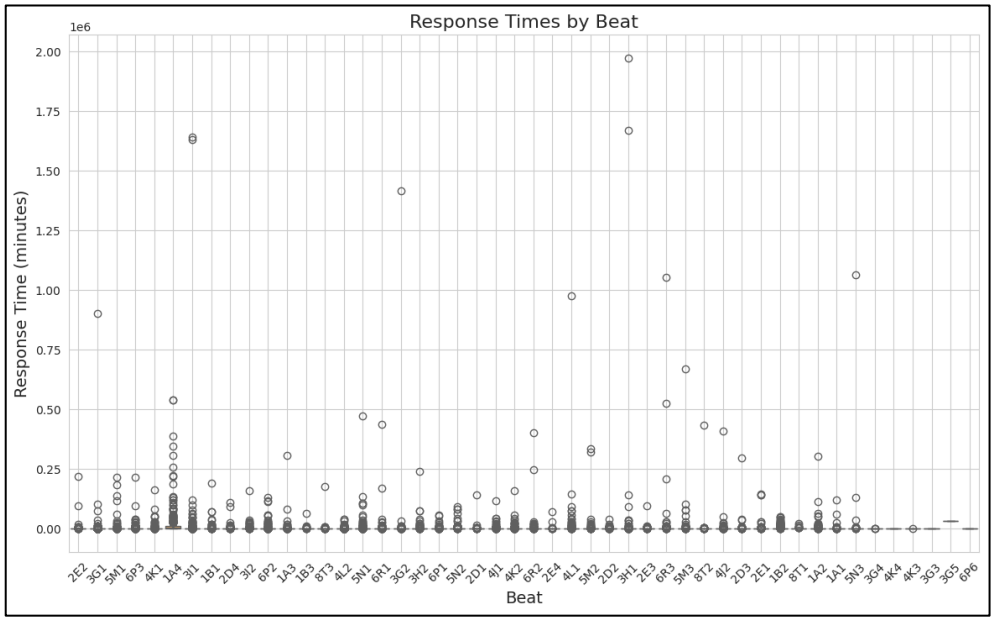

Figure 31 Box plot of Response Times by Beat

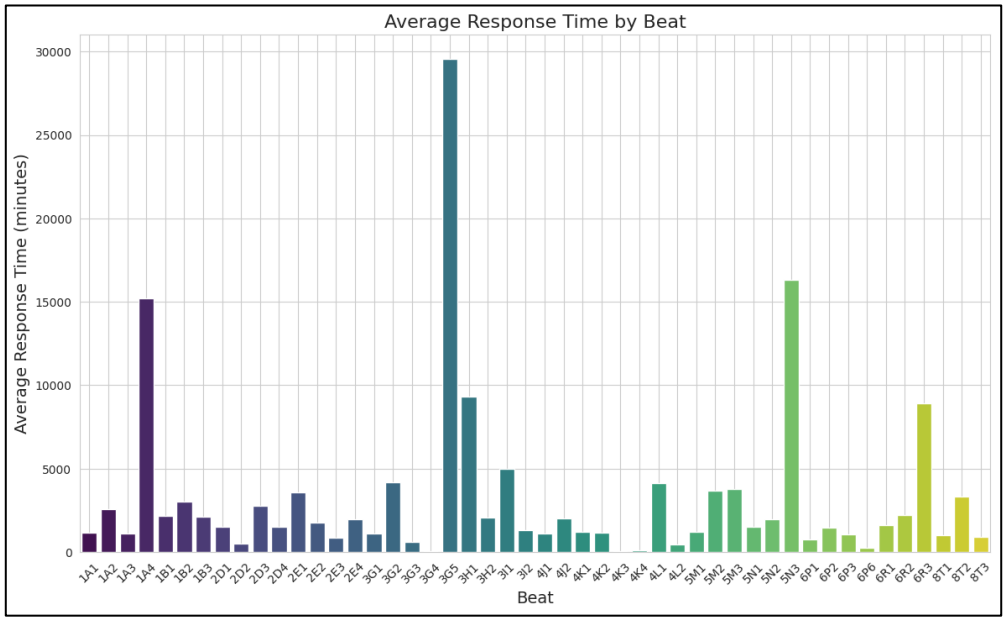

Figure 30 Avg Response Time by Beat

Figure 31 highlights that Beats 3G5, 5N3, and 1A4 have the highest average response times, with 3G5 showing the highest overall. In Figure 30, the box plot indicates that while 3G5 and 5N3 have fewer data points, suggesting a more limited or skewed distribution, Beat 1A4 displays a broader distribution with more data points, indicating a more balanced spread in response times.

5) How is crime distributed geographically across various cities, and what patterns emerge when examining incidents by type, frequency, and seasonal variations?

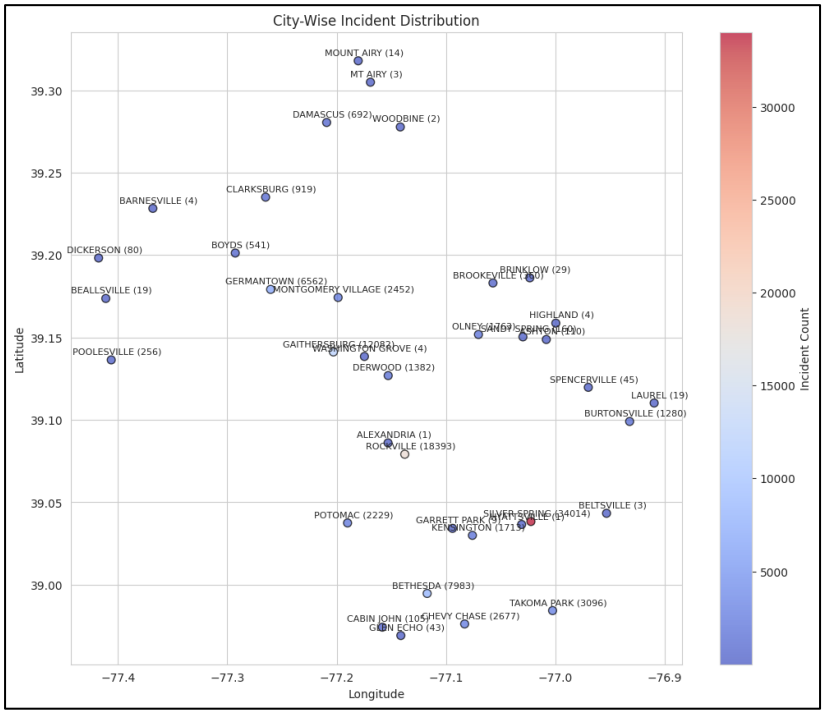

Figure 33 Geographical distribution of city-wise incident

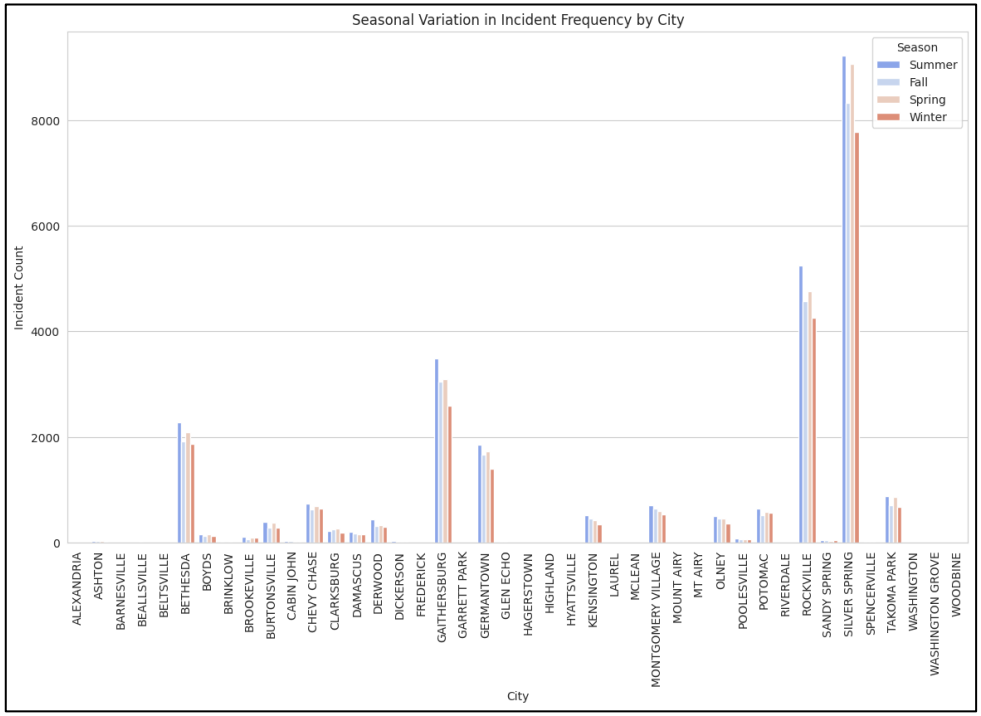

Figure 32 Stacked bar chart of grouping city and season for incident count

Figure 32 shows that most cities have moderate incident counts, with Silver Spring standing out with a significantly higher crime density. In Figure 33, the seasonal trend across all cities indicates that crime incidents peak in summer, followed by spring, fall, and winter, consistently showing higher crime rates during warmer months. **(Nöllenburg, 2007)**

6) How do crime incident distributions vary across different directional regions of the city, and how does the frequency of these incidents differ by geographic region?

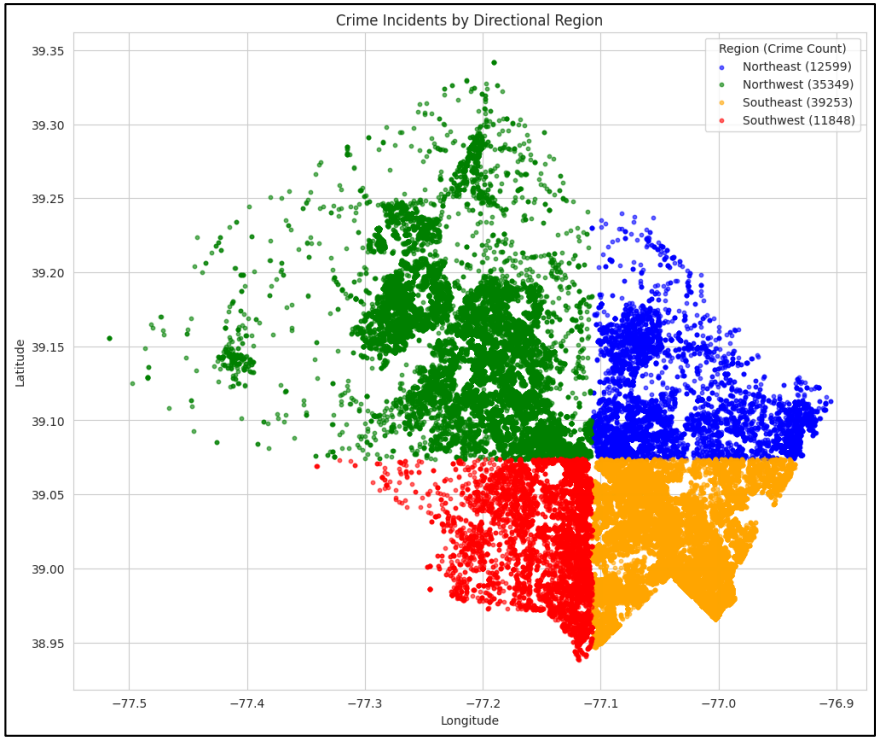

Figure 34. Geographical distribution of crime incidents by directional region

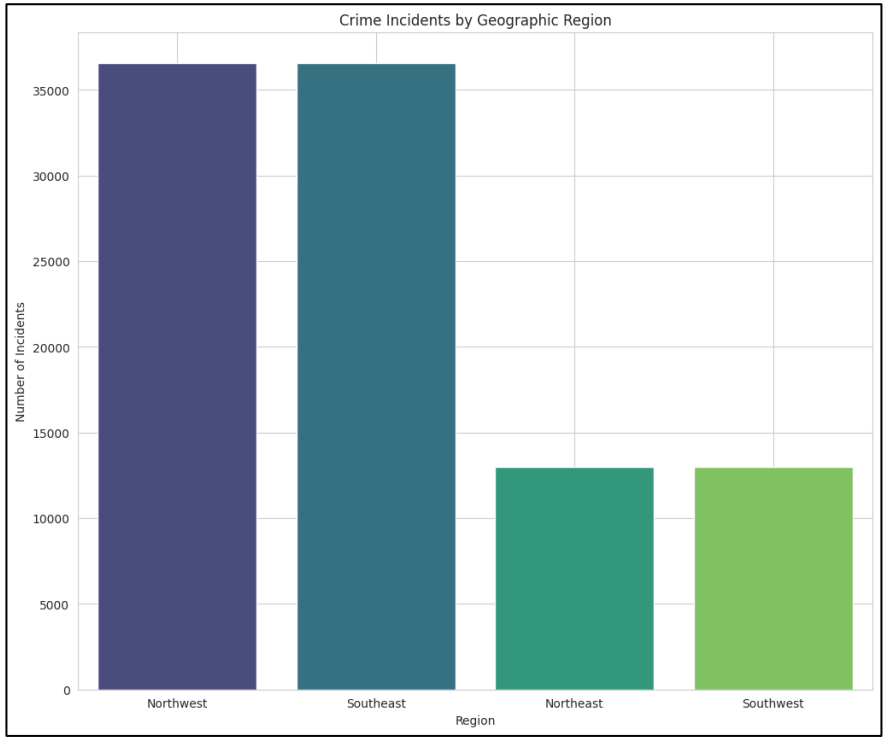

Figure 35 Bar chart representing crime incidents by geographic region

Figures 34 and 35 reveal that the northwest and southeast regions, though smaller in area, exhibit a high density of crime incidents compared to other parts of the city, indicating concentrated crime activity in these areas.

7) What are the key crime hotspots in different cities, and how do they vary by crime type?

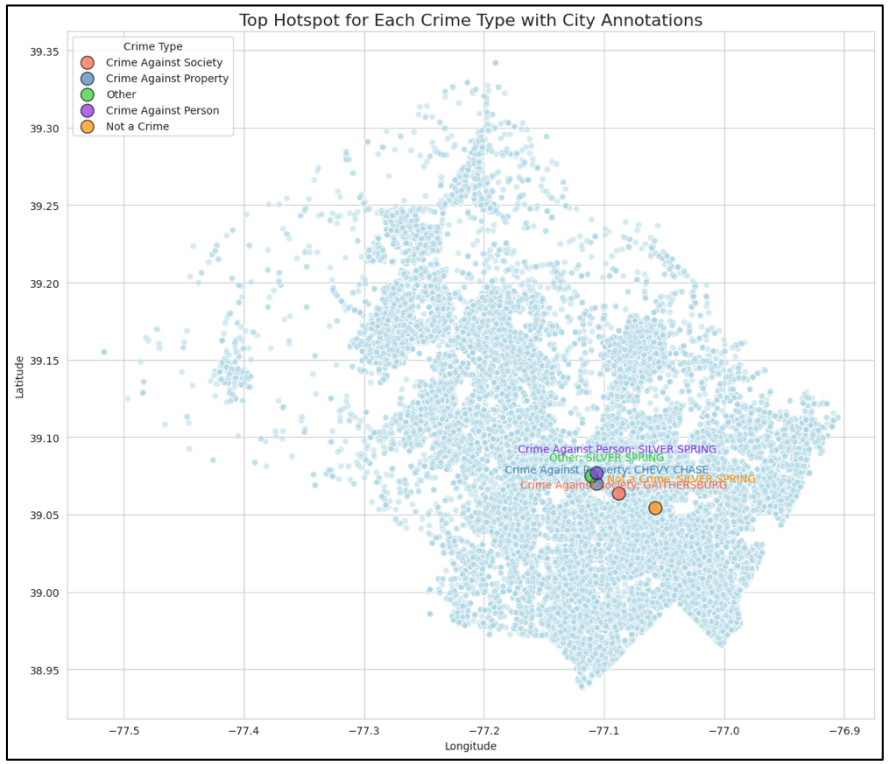

Figure 37 Geographical distribution of hotspot cities for each crime type

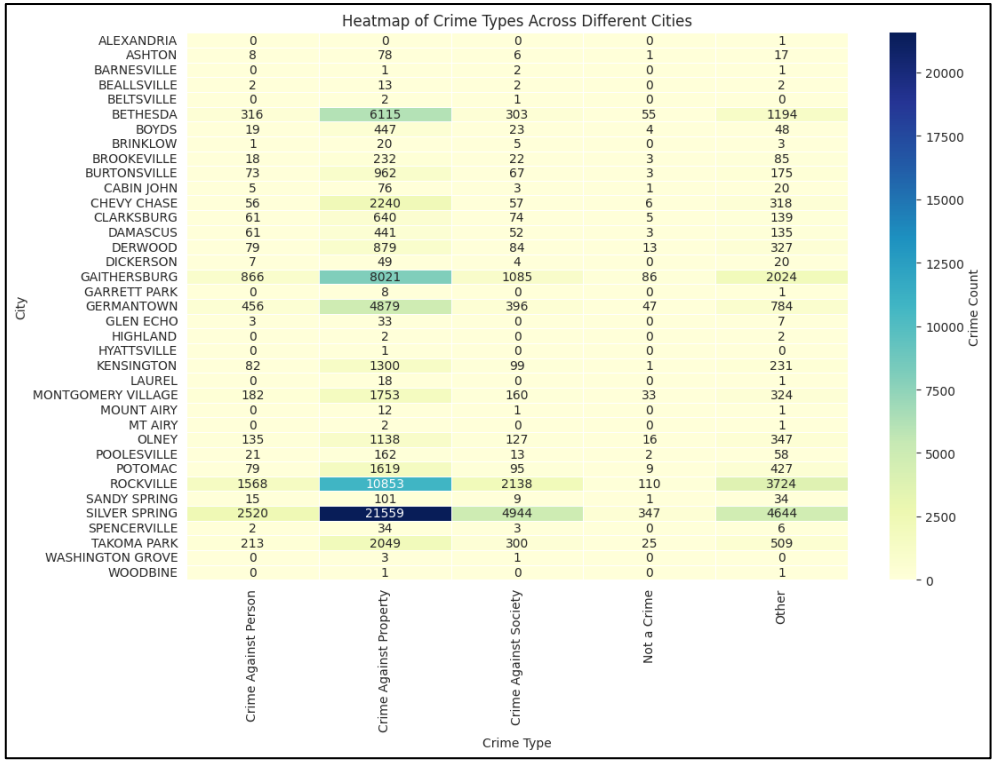

Figure 36 Heatmap of crime types across different cities

Figures 36 and 37 highlight Silver Spring as a primary hotspot across multiple crime categories, including crimes against persons, society, and other incidents. Chevy Chase emerges as a hotspot for property-related crimes, while Gaithersburg shows high occurrences of societal crimes. The heatmap further illustrates that Silver Spring experiences a broad spectrum of crime types, underscoring its prominence as a high-crime area. **(Thomas and Raja, 2024)**

8) How do response times vary across cities and crime types, and what are the underlying patterns in these variations?

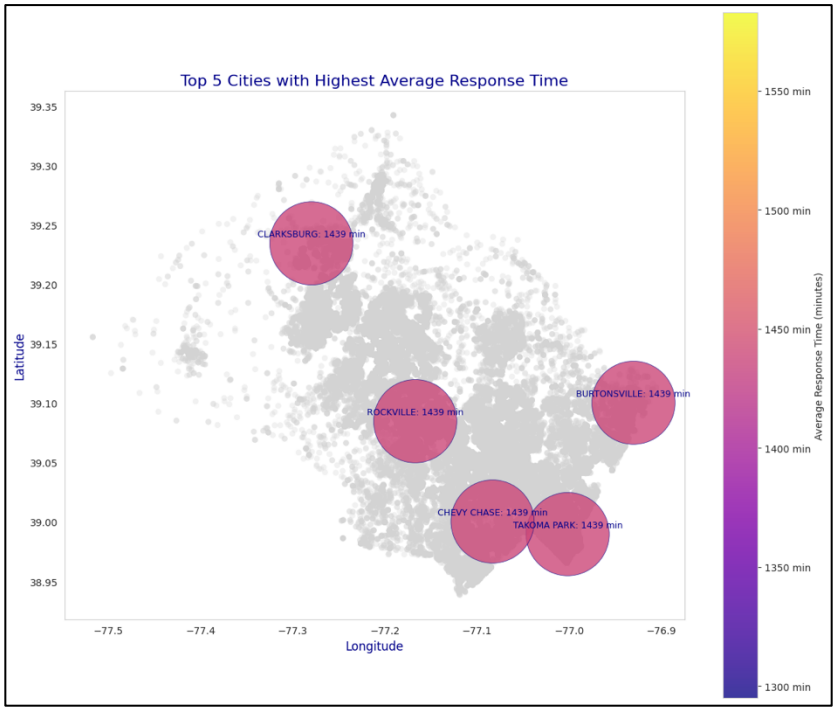

Figure 39 Geographical distribution of cities with highest average response time

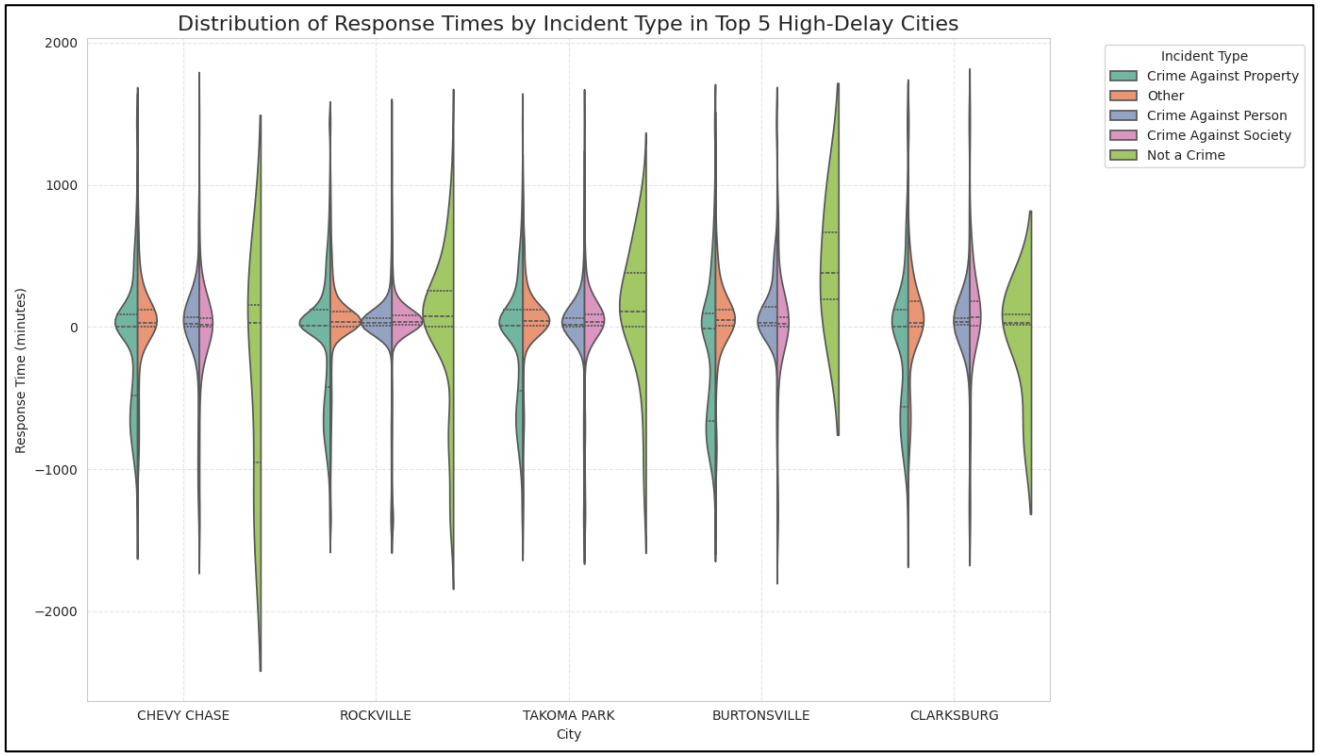

Figure 38 Representation using violin map of response time by incident type in top 5 high delay cities

Figures 38 and 39 highlight that cities such as Rockville, Burtonsville, Chevy Chase, Takoma Park, and Clarksburg show the highest average response times of 1439 minutes. The violin plot reveals that Crime Against Property exhibits the widest response time distribution across these cities, while Not a Crime incidents show faster resolution times. Cities like Rockville and Chevy Chase experience higher variability in response times, while Takoma Park shows faster responses for Not a Crime incidents. **(Molina et al., 2022)**

9) How do high-density crime areas, identified geographically by Zip Codes, correlate with the types of crimes occurring within those regions?

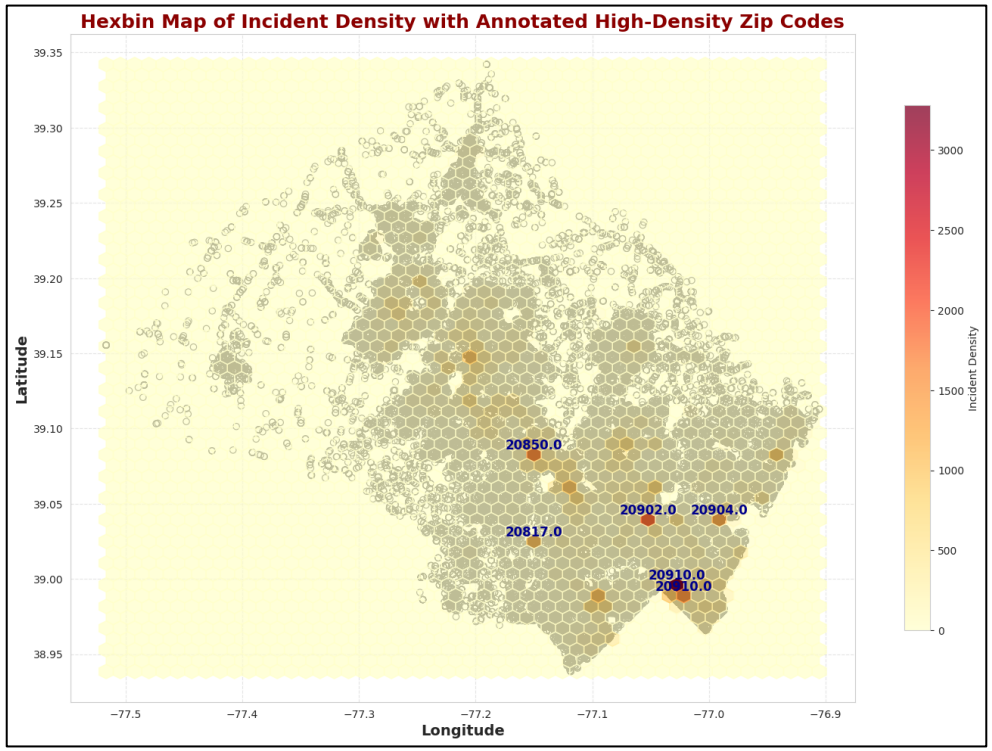

Figure 41 Hexbin map of incident density with annotated high density zip codes

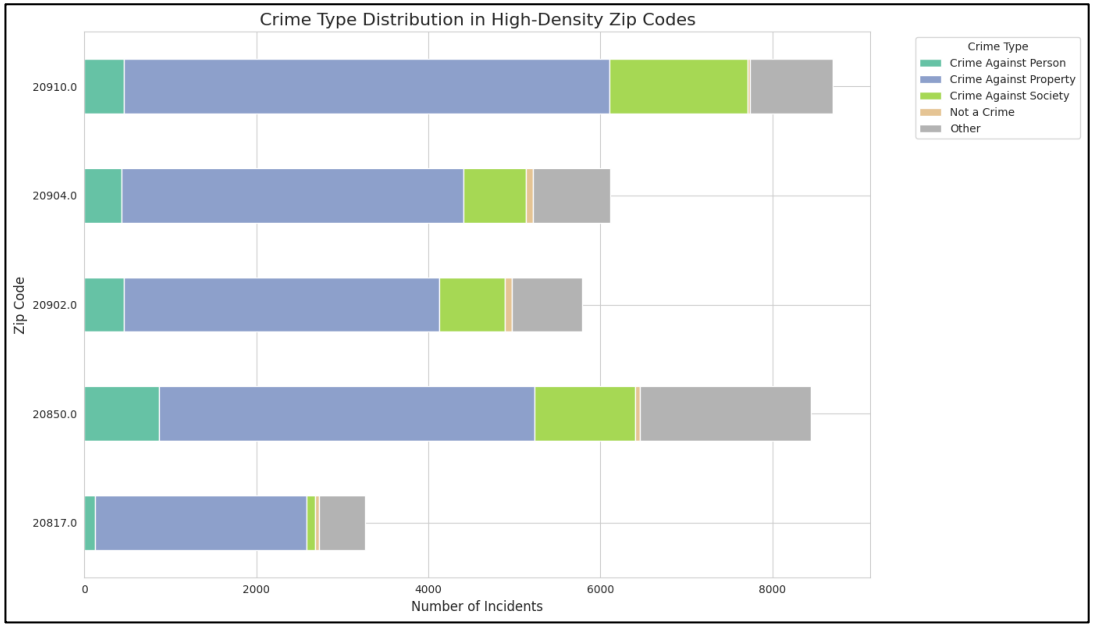

Figure 40 Stacked bar chart representing crime type in high density zip codes

Figures 40 and 41 show that Zip Codes 20910, 20904, 20902, 20850, and 20817 exhibit high crime incident densities, with these areas closely situated. The stacked bar chart in Figure 41 reveals that Crime Against Property is the most prevalent crime type in these high-density areas, followed by Crime Against Society and Other crimes. 'Not a Crime' incidents are rare across these zip codes.

10) How do 'Not a Crime' incidents vary across different Zip Codes and Place Types, and what are the trends in incident frequency within these areas?

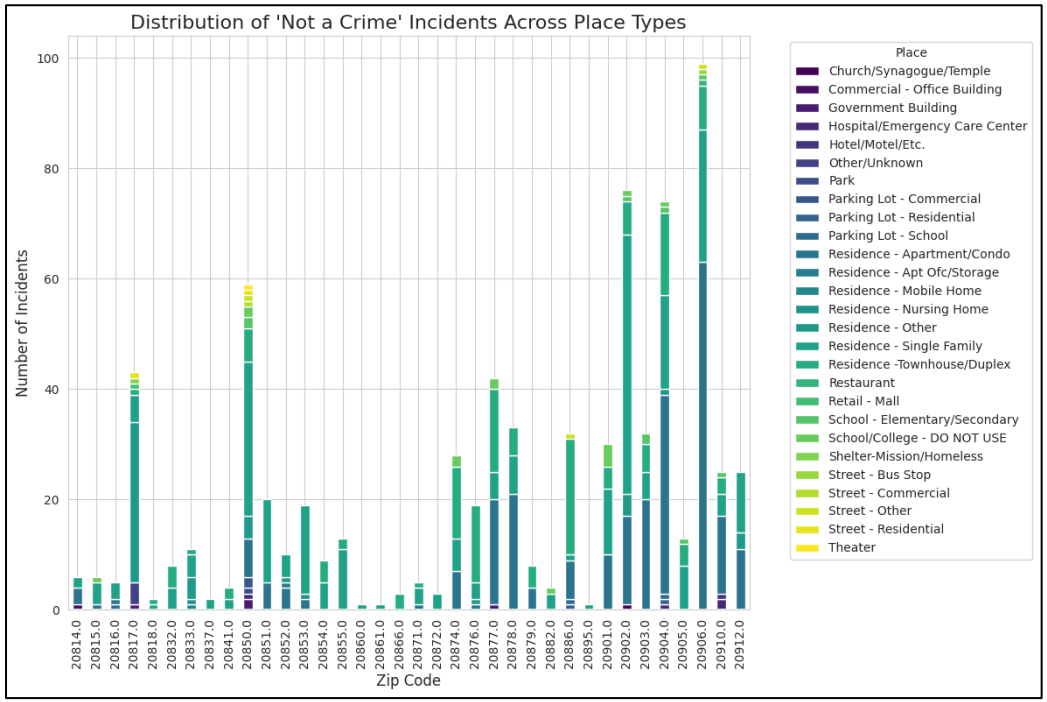

Figure 43 Stacked bar chart representing 'Not a crime' incidents across place types

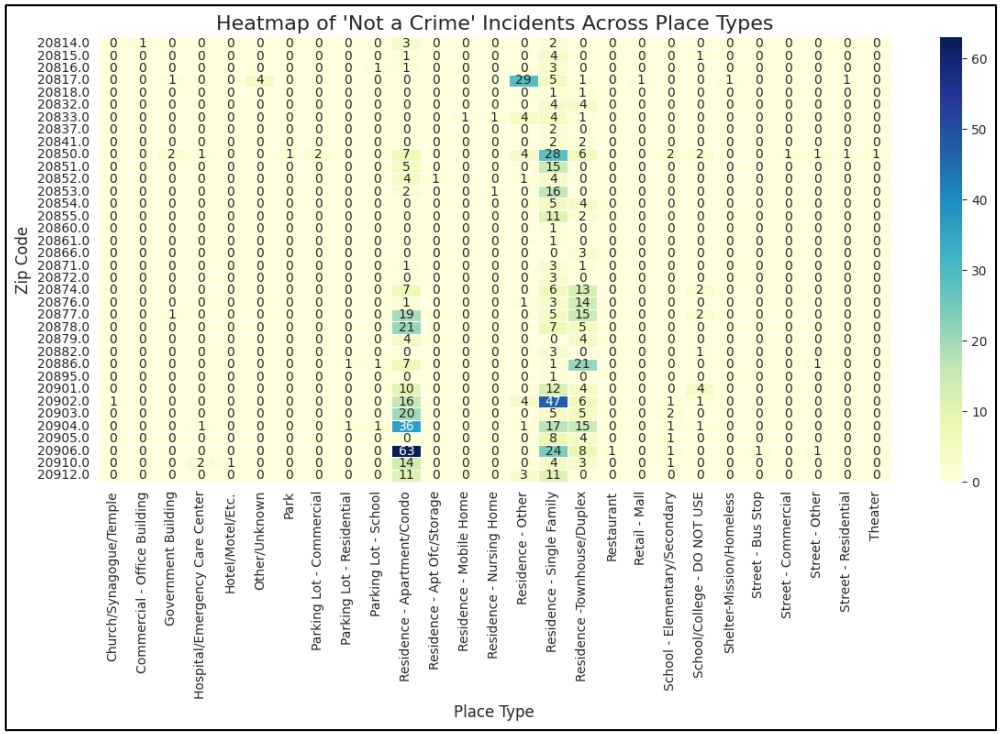

Figure 42 Heatmap representing 'Not a crime' incidents across place types

Figures 42 and 43 highlight that Zip Code 20906 reports the highest number of 'Not a Crime' incidents, followed by 20904. The majority of these incidents are concentrated in residential areas, particularly in apartment/condo, single-family homes, and townhouse/duplex places.

11) How does the distribution of crime incidents across cities compare to the number of police stations allocated, and do any cities exhibit a need for additional police resources based on crime rates?

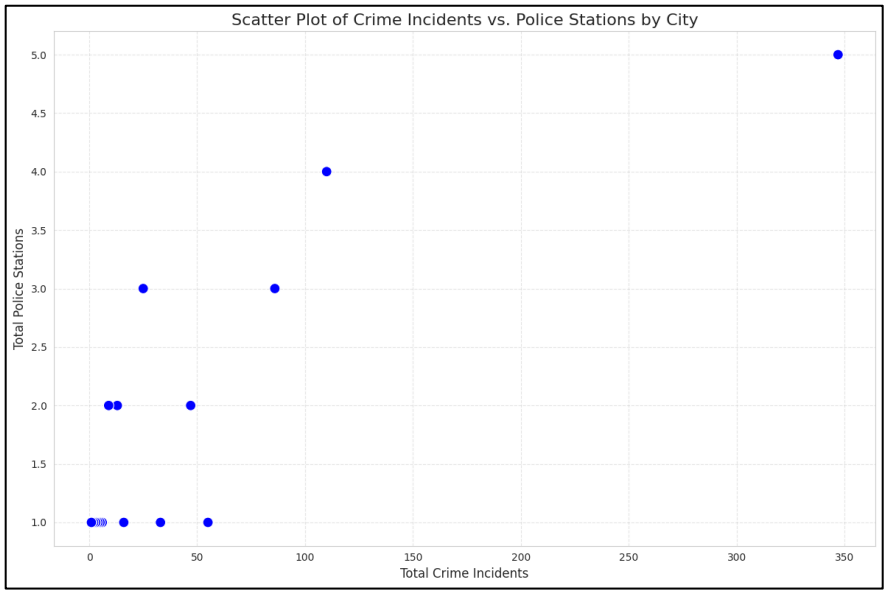

Figure 45 Scatter plot distribution of crime incidents vs police stations by city

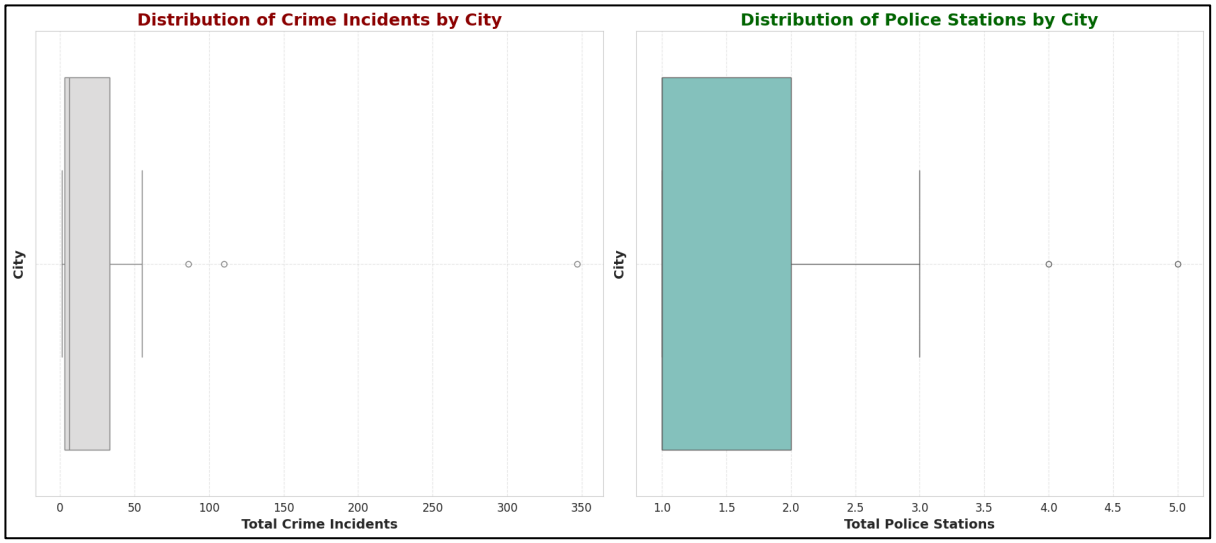

Figure 44 Box plot representation of distribution of crime incidents by city and police stations by city

The scatter plot (Figure 44) and box plot (Figure 45) reveal no clear correlation between crime incidents and the number of police stations across cities. Most cities have 1 to 3 police stations, with no significant outliers, indicating that the allocation of police stations is generally aligned with crime distribution. Therefore, there appears to be no immediate need for additional police resources in most cities.

12) Which streets experience the highest frequency of crime incidents, and what are the most common types of crimes on these high-frequency streets?

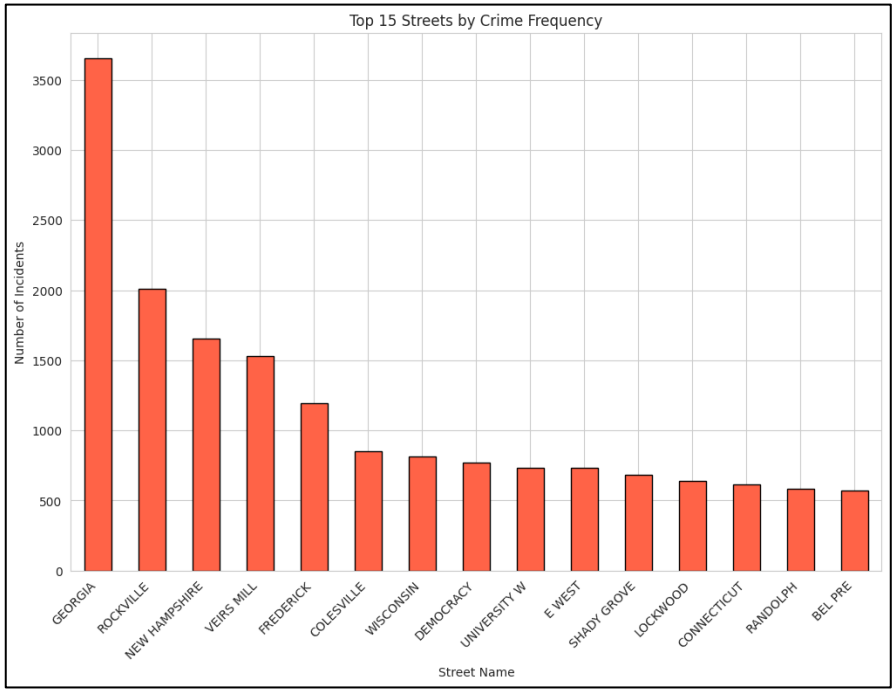

Figure 46 Bar chart representing top 15 streets by crime frequency

Figure 47 Heat map representing crime type on high frequency crime streets

The bar chart (Figure 46) shows that Georgia Street has the highest frequency of crime incidents, followed by Rockville Street. The heatmap (Figure 47) indicates that "Crime Against Property" and "Crime Against Society" are the most common crime types on these high-frequency streets.

5. Summary:

This analysis reveals key insights into crime patterns and police resource allocation. ***Sector T*** shows the most efficient incident resolution, while ***Montgomery Village, Silver Spring, and Wheaton exhibit*** significant gaps in resolution time data, highlighting the need for better reporting practices. ***Crime peaks*** at specific times, especially ***midnight and Fridays***, signalling the need for targeted policing during these periods.
Geographically, ***high-density crime*** areas in the ***northwest and southeast*** suggest potential for additional police presence. While ***Silver Spring*** is a ***hotspot***, the overall distribution of police stations appears sufficient, with ***no immediate need for new districts***. High-density zip codes like ***20910 and 20904***, where property and societal crimes are prevalent, ***require focused attention***.
'Not a Crime' incidents in residential areas emphasize the ***need*** for ***clearer data classification***. Lastly, streets like ***Georgia*** and ***Rockville***, with high ***crime rates***, highlight areas for ***potential intervention***. Overall, while police station distribution is generally adequate, improvements in reporting practices and resource deployment in hotspot areas could enhance public safety.

6. References:

1. **Bayoumi, S., AlDakhil, S., AlNakhilan, E., Al Taleb, E. and AlShabib, H.**, 2018, April. A review of crime analysis and visualization. Case study: Maryland state, USA. In *2018 21st Saudi Computer Society National Computer Conference (NCC)* (pp. 1-6). IEEE.
2. **Cai, L.**, 2021, September. Analysis of hate crime rates in the United States: statistical modeling of public safety issues based on socioeconomic factors. In *2021 International Conference on E-Commerce and E-Management (ICECEM)* (pp. 388-392). IEEE.
3. **Molina, E., Viale, L. and Vázquez, P.**, 2022, October. How should we design violin plots?. In *2022 IEEE 4th Workshop on Visualization Guidelines in Research, Design, and Education (VisGuides)* (pp. 1-7). IEEE.
4. **Mukhiya, S.K. and Ahmed, U.**, 2020. *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing Ltd.
5. **Nöllenburg, M.**, 2007. Geographic visualization. In *Human-centered visualization environments: GI-Dagstuhl research seminar, Dagstuhl Castle, Germany, March 5-8, 2006, revised lectures* (pp. 257-294). Berlin, Heidelberg: Springer Berlin Heidelberg.
6. **Thomas, B.A. and Raja, S.**, 2024, July. Crime mapping and predictive analysis of crimes in Maryland, USA. In *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)* (pp. 1-6). IEEE.