Kingston University London

Montgomery County Crime Data EDA

Submitted By Group 18

Group Members and Participation

Student Name	Student ID	Group Assignment Development Contribution Percentage (%)
D.N.I Chandrasoma	**K2417093**	
Isuru Pathirana	**K2422824**	
Sasheena Mohomed	**K2455092**	

Abstract

This report analyzes crime trends in Montgomery County using detailed crime data from 2016 to 2022. The study focuses on uncovering patterns related to the frequency, nature, and distribution of crimes across temporal, geographic, and categorical dimensions. Key findings reveal a clear downward trend in crimes and victim counts over the years, with Fridays experiencing the highest crime rates and Sundays the lowest. Midnight emerges as the peak time for criminal activity, and the Silver Spring neighborhood is identified as the region with the most reported crimes.

Further analysis highlights that crimes predominantly target property, with vehicles being the most affected. Weekdays witness higher crime rates compared to weekends, suggesting a correlation with weekday activities. A strong relationship between drug-related and prostitution crimes underscores the interconnected nature of these offenses.

These insights provide actionable recommendations for law enforcement, including heightened surveillance during late-night hours, targeted interventions in high-crime areas, and addressing drug-prostitution linkages through coordinated social and enforcement strategies. This study emphasizes the value of data-driven approaches in enhancing community safety and guiding policy decisions.

Table of Contents

1. Introduction

Data Science is the field of study that involves collecting, analysing and interpreting large sets of data to uncover insights, patterns and trends that can be used to make informed decisions and solve real world problems. (Biswal, 2024) It involves processes such as data collection, cleaning, exploration, visualization, and modelling. With advancements in computational power and the availability of large datasets, data science is increasingly applied across various industries, including healthcare, finance, retail, and public safety.

In an era of rapid technological advancements, incorporating digitalization into crime research has transformed the way law enforcement organizations and legislators approach public safety. In the context of crime analysis, data science provides tools to uncover patterns, and identify trends offering a data-driven approach to public safety and policy-making. The dataset provided for this research represents Montgomery County's crime statistics from 2016 to 2022. It is based on reports classified under the National Incident-Based Reporting System (NIBRS) and adheres to the Uniform Crime Reporting (UCR) rules. This dataset includes multiple offenses linked to single incidents, potentially involving multiple victims, providing a detailed picture of crime patterns within the county. The dataset provides a robust foundation for applying data science methodologies to analyse, interpret, and derive actionable insights from real-world crime data. Such datasets are critical in the law enforcement domain as they allow better understanding of crime trends, frequency and distribution. Insights derived from these datasets can support decision-making to reduce crime rates and allocate resources effectively.

a. Problem Statement

A crime can be defined as an intentional commission of an act, usually deemed socially harmful or dangerous and specifically defined, prohibited, and punishable under criminal law. (Ian David Edge, 2024) The existence of crime in every society is not news anymore. So we need effective strategies to minimize its occurrences and impact. The challenge is to analyse the Montgomery County crime dataset to identify patterns and trends in criminal activity. This analysis aims to answer specific research questions, reveal actionable insights, and propose strategies for crime mitigation. Key issues include handling data quality, exploring correlations between crime types, and visualizing geographical and temporal patterns. The research aims to equip local authorities with data-driven insights to enhance safety measures and optimize resource allocation.

Objectives

- Data Understanding: Analysing the dataset's structure, types of variables, and identifying gaps or inconsistencies.
- Trends and Patterns: Identifying yearly and monthly trends, most frequent crime types, and correlations between different variables.
- Policy Recommendations: Suggesting strategies for crime mitigation based on findings.

b. Research Questions

1. What are the observable year-over-year trends in the frequency of crimes and victim counts?
2. Which day of the week records the highest and lowest number of crimes or victims?
3. What is the temporal distribution of crimes across different times of the day?
4. Which Neighbourhood / Police District has the greatest number of crimes in the county?
5. Which type of locations (street types) mostly affected by crimes?
6. What is the comparative distribution of crimes between weekdays and weekends?
7. Which category is most impacted by crimes: persons, property, or society?

Proposed Research Methodology

1. Data Exploration:
    - Examine dataset structure and features (columns, data types, etc.).
    - Identify missing values and inconsistencies.
    - Pre-process data using techniques like imputation and normalization.
2. Exploratory Data Analysis (EDA):
    - Use descriptive statistics to summarize the data.
    - Create visualizations (e.g., heatmaps, bar charts, line graphs) to highlight key patterns.
3. Statistical Analysis:
    - Identify correlations and relationships between variables (e.g., crime types, locations).
4. Visualization and Reporting:
    - Generate 20 visualizations (2 per research question).
    - Document findings in a well-structured report.

Scope

- A comprehensive understanding of crime patterns in Montgomery County.
- Insights into the relationships between various crime types and contributing factors.
- A set of visualizations that effectively communicate findings.
- Practical recommendations for reducing crime and improving public safety.

Workflow

The workflow for this research is as follows:

1. Data Loading and Initial Inspection:
    - Import and examine the dataset using Python libraries like Pandas and NumPy.
    - Summarize the structure and key statistics.
2. Data Cleaning and Pre-processing:
    - Handle missing or inconsistent data.
    - Standardize formats for dates and geographical data.
3. EDA and Visualization:

- Create visualizations for trends, correlations, and geographic distributions.

4. Statistical analysis:
   - Conduct in-depth analyses to support research questions.
5. Report Writing and Recommendations:
   - Document findings and present actionable insights.

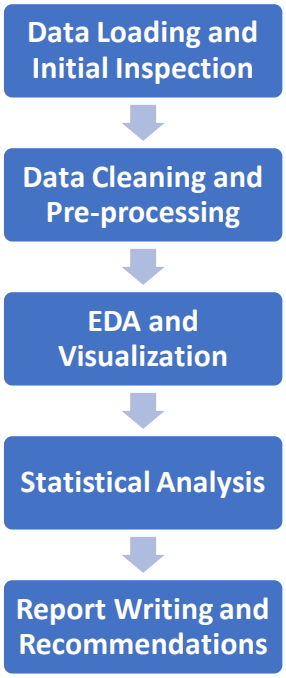


Figure 1: Research workflow

1. Preliminary Data Analysis

a. Dataset

Source

The dataset for this research comes from Montgomery County's crime statistics, which are reported under the National Incident-Based Reporting System (NIBRS). It follows Uniform Crime Reporting (UCR) guidelines and includes data collected between 2016 and 2022. This dataset is based on police incident reports and contains preliminary information provided by the Montgomery County Police Department.

Format and Type of Data

- Format: The dataset is structured and stored in a tabular format, commonly as a CSV (Comma-Separated Values) file, suitable for computational analysis

- Type of Data:

  Categorical: Columns like NIBRS Code, Crime Name1, Police District Name, etc. These columns often contain limited and specific sets of values, such as crime categories or Police District Name

  Numerical: Columns such as Victims, Police Report Number, etc., which contain numbers.

  Date time: Columns such as Start_Date_Time or End_Date_Time help identify temporal patterns (e.g., trends over time, seasonality of crimes). You may need to parse and manipulate these columns to extract specific information like day of the week, month, or year.

- Dataset Columns and Their Types
  The dataset contains the following columns

Column	**Description**	**Type**
Incident ID	Police Incident Number	Integer
Offence Code	The code defined by NIBRS	String
CR Number	Police Report Number	Integer
Dispatch Date / Time	The actual date and time an Officer was dispatched	String
NIBRS Code	FBI NIBRS codes	String
Victims	Number of Victims	Integer
Crime Name1	Crime against Society/Person/Property or Other	String
Crime Name2	Describes the NIBRS_CODE	String
Crime Name3	Describes the OFFENSE_CODE	String
Police District Name	Name of District (Rockville, Weaton etc.)	String
Block Address	Address in 100 block level	String
City	City	String
State	State	String
Zip Code	Zip Code	Float
Agency	Assigned Police Department	String
Place	Place description	String
Sector	Police sector name, a subset of District	String
Beat	Police patrol area, a subset of Sector	String
PRA	Police Response Area, a subset of Beat	String
Address Number	House or Business Number	Float
Street Prefix	North, South, East, West	String
Street Name	Street Name	String
Street Suffix	Quadrant (NW, SW, etc.)	String
Street Type	Ave, Drive, Road, etc.	String
Start_Date_Time	Occurred from date/time	String
End_Date_Time	Occurred to date/time	String
Latitude	Latitude	Float
Longitude	Longitude	Float
Police District Number	Major Police Boundary	String
Location	Location	String

Table 1: Dataset columns and Data types

b. Data Quality Initial Assessment

Data quality initial assessment is a critical part of the analysis, as it helps you evaluate the state of the dataset and decide how to proceed with cleaning and transforming the data.

Initial data analysis show that we have 306094 records, 0 duplicate rows. The incident id column have duplicate values because there can be multiple offences claimed for one incident.

- **Missing Values**

The following table shows the columns with missing values and the count.

Column Name	**Missing Values**
Street Suffix	300662
Street Prefix	292463
End_Date_Time	161658
Dispatch Date / Time	49029
Block Address	26206
Address Number	26109
Zip Code	3179
Sector	1530
Beat	1530
City	1276
Street Type	339
Crime Name1	272
Crime Name2	272
Crime Name3	272
PRA	239
Police District Name	94

Table 2: Columns with missing values

Based on the analysis of missing values and other details, the following decisions were made to ensure data quality for the exploratory data analysis (EDA):

- Columns with Missing Values:

Address Related Columns:
The dataset includes the columns Street_Suffix and Street_Prefix, which have 95% and 98% missing values, respectively. Given the high percentage of missing data, imputing or using these columns for analysis would not be practical. Additionally, the dataset provides precise location information through latitude and longitude, which makes these columns redundant. Therefore, we decided to drop Street_Suffix and Street_Prefix to streamline the dataset and focus on relevant features.

Temporal Columns:
The dataset includes three columns with temporal values: Start_date_time, Dispatch_date/Time, and End_date_time. Among these, the Start_date_time column has no missing values and provides sufficient temporal information for extracting the day, month, and year required for our analysis. The Dispatch_date/Time column has 49,029 missing values, and the End_date_time column has 161,658 missing values, making them unsuitable for reliable imputation. Since these columns are not essential for the research objectives, they were dropped to simplify the dataset and focus on meaningful features

Crime Name Columns (Crime Name1, Crime Name2, Crime Name3):
These columns are crucial for EDA as they describe the type of crime. For missing values in these columns, we will merge the dataset with another dataset that maps NIBRS codes to crime names. This will allow us to reconstruct the missing values accurately. The additional data set we used is from https://ucr.fbi.gov/nibrs/2011/resources/nibrs-offense-codes
The data was extracted and stored in a csv file before the mapping. The csv file is available in the appendix section of this report.

- **Data Transformation**

Datetime Transformation

The column Start_date_time in the dataset was originally in string format, which limited its usability for temporal analysis. To address this, we converted it to a datetime format using Python's pandas library. This transformation enabled efficient extraction of temporal features such as year, month, and day, which are crucial for understanding temporal patterns in the dataset.

- **Data Wrangling**

Extraction: Extracting relevant information from complex data types.

After converting Start_date_time column to datetime format, various temporal components were extracted to gain deeper insights into the data. These features included:

Year: The year of the incident.
Month: The month of the incident (1–12).
Day: The day of the month when the incident occurred (1–31).
Time: The specific time of the incident in hours.

By extracting these features, we transformed the temporal data into multiple granular components that are instrumental for detailed analysis. This process enhances the dataset's usability for identifying trends and patterns in crime occurrences over different time intervals.

Merging: We can merge two datasets using a common key.

As we explained handling missing values for Crime Name Columns we used another data set and merge it with this dataset using NIBRS code.

2. Exploratory Data Analysis

a. Introduction to EDA

Exploratory Data Analysis is a critical step in the data science process. It is the foundation for understanding and interpreting complex data sets. EDA helps data scientists identify patterns, spot anomalies, test hypotheses, and check assumptions through various statistical and graphical techniques. Practitioners can uncover underlying structures, detect outliers, and determine the relationships between variables, which is essential for developing accurate predictive models by thoroughly exploring the data. (Biswal, 2024) It not only provides insights into the data but also helps in determining the next steps in the analysis.

In the context of this report, EDA is pivotal for exploring the Montgomery County crime dataset. The dataset contains detailed crime statistics recorded between 2016 and 2022. Using EDA, we aim to answer key research questions such as identifying the most common types of crimes, analysing temporal and spatial patterns. This process will enable us to uncover trends and relationships in the data, which can inform actionable insights.

- **Further Data Cleaning or Transformation Needs**

While the dataset has been pre-processed to handle major issues like missing values and inconsistencies, further cleaning and transformations are necessary for effective EDA:

Handling Residual Missing Values:
The Police District Name column is important for EDA. Missing values in this column will be filled with a placeholder value, such as "UNKNOWN", to maintain data consistency.

Parsing and Utilizing Temporal Data:
The Start_Date_Time column has been transformed into a datetime format, enabling the extraction of features like day, month, year, and time of day. Furthermore, we extracted more components which will help to enhance the depth of the analysis.
Weekday: The day of the week as an integer (0 = Monday, 6 = Sunday).
Day Name: The full name of the weekday (e.g., "Monday").

Normalization
As part of data transformation, normalization techniques were applied to text and categorical data to ensure consistency and prepare the data for analysis. For example, all text in the Crime_name columns were converted to uppercase, ensuring uniformity across entries. This transformation was crucial for accurate grouping and comparison during analysis

Feature Engineering:
Deriving new features such as Day of the Week or Day Type will enhance the depth of analysis.

Geospatial Data Validation:
Latitude and longitude coordinates will be validated against Montgomery County's geographic boundaries, and grouping by police districts or other areas will be applied for meaningful spatial analysis.

b. Approaches for Exploratory Data Analysis

To explore this dataset effectively, the following EDA techniques will be employed:

- Univariate Analysis:

Examining individual variables to understand their distributions and detect anomalies. For instance, bar charts and histograms will reveal the frequencies of different crime types and victim demographics.

- Bivariate and Multivariate Analysis:

Analysing relationships between variables, such as crime types and time of day, to identify correlations or dependencies.
Tools: Scatter plots, correlation matrices, and pair plots.

- Geospatial Analysis:

Using latitude and longitude data to generate crime density maps and identify hotspots.
Tools: Geographic heatmaps and cluster plots.

- Temporal Analysis:

Exploring patterns over time to identify trends, such as seasonal variations in crime rates.
Tools: Line graphs and time-series visualizations.

c. Descriptive Statistics

Descriptive statistics methods such as min, max, mean, count, unique, top, freq, and std provide valuable insights into the dataset by summarizing key characteristics of each column. These methods allow us to understand how individual features influence the overall analysis and help identify potential issues or trends.

- Minimum and Maximum Values (min, max):

The min and max of the Victims column reveal the smallest and largest number of victims per incident. This is useful for detecting potential outliers, such as incidents with unusually high victim counts, which may require further investigation or handling.
For temporal columns, such as Year, the min and max values indicate the range of recorded incidents, from the earliest to the latest, providing context for time-based analyses.

- Count (count):

The count method helps identify the total number of non-missing values in a column. This is particularly useful for understanding the completeness of each variable and prioritizing columns for further analysis.

- Unique Values (unique):

The unique method provides the number of distinct values in a column. For categorical variables like Crime Name or Police District Name, this helps in assessing the diversity of categories and identifying any anomalies, such as unexpected or inconsistent entries.

- Most Frequent Values (top, freq):

The top method identifies the most frequently occurring value in a column, while the freq method shows how often it appears. For instance:
In the Crime Name column, top can indicate the most common crime type, and freq will show how prevalent it is compared to other crime types.
In temporal columns, such as Month, these methods can help identify peak crime periods.

- Central Tendency and Variability (mean, std):

The mean and std methods summarize the average value and variability in numeric columns, such as Victims. For example, the standard deviation (std) of the Victims column helps understand the spread of incident severity, while the mean provides an overall average.

These descriptive statistics enable us to summarize large datasets efficiently and draw meaningful insights. For instance, combining these methods allows us to:

Detect outliers in numeric columns like Victims and assess their impact on the analysis.
Understand temporal trends using date-related columns such as Year and Start_Date_Time.
Evaluate the diversity and consistency of categorical columns like Crime Name and Police District Name.

By leveraging these methods, we can gain a comprehensive understanding of the dataset and ensure a robust foundation for further analysis.

d. Data Visualization

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with information graphics, information visualization and statistical graphics. (Cameron Hashemi-Pour, 2024)

Why is Data Visualization Important?

Simplifying Complexity:
Large and complex datasets can be challenging to interpret in tabular form. Visualizations simplify these datasets, making trends and anomalies evident at a glance.

Supporting Decision-Making:
Decision-makers rely on visual insights to understand key metrics and make informed choices.

Identifying Trends and Patterns:
Visual tools make it easier to observe trends over time, relationships between variables, and outliers in the data.

Effective Communication:
Visualizations help communicate findings to stakeholders who may not have a technical background.

Best Practices for Data Visualization

To create effective visualizations, the following best practices should be followed:

Clarity:
Avoid clutter and ensure that the visualization clearly conveys the intended message.
Use appropriate labels, titles, and legends.

Audience Focus:
Tailor the visualization to suit the audience's level of expertise and the purpose of the analysis.

Choosing the Right Chart:
Select the type of visualization that best represents the data and supports the narrative.

Consistency:
Use consistent colors, scales, and formatting to avoid confusion.

Accessibility:
Ensure the visualization is easy to interpret, even for colorblind viewers (e.g., use patterns or contrasting colors).

Common Types of Visualizations and Their Uses

Bar Charts:
Used to compare categorical data, such as crime counts by type or year.
Example: Crime type frequencies in Montgomery County.

Line Graphs:
Used for analysing trends over time.
Example: Yearly crime trends or seasonal patterns.

Pie Charts:
Represent parts of a whole, but are less preferred for detailed analysis due to difficulty in comparing segments.
Example: Proportions of different crime categories.

Scatter Plots:
Show relationships between two variables, often highlighting correlations.
Example: Relationship between time of day and specific crime types.

Histograms:
Visualize the distribution of a single variable.
Example: Distribution of victim counts per incident.

Heatmaps:
Represent data intensity in a specific area using colors.
Example: Crime density across Montgomery County.

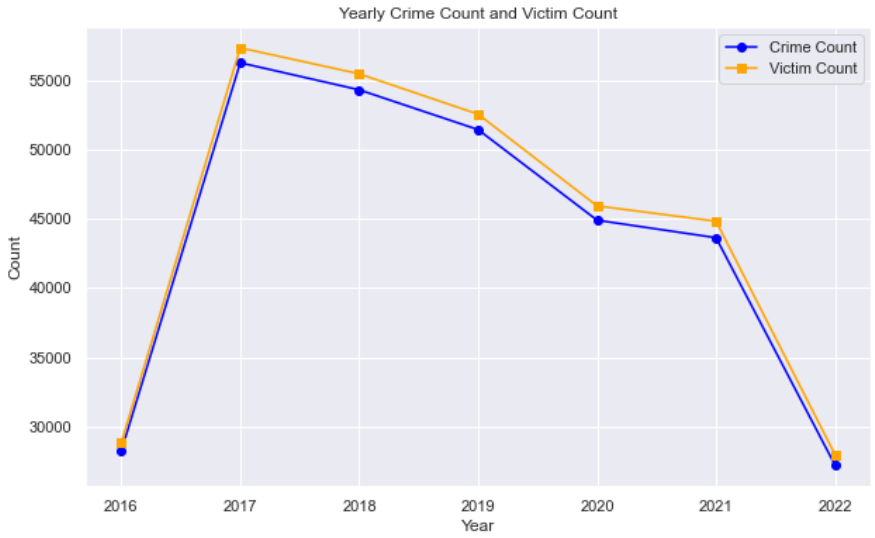
Box Plots:
Display data distribution, variability, and outliers.
Example: Victim counts per crime type.

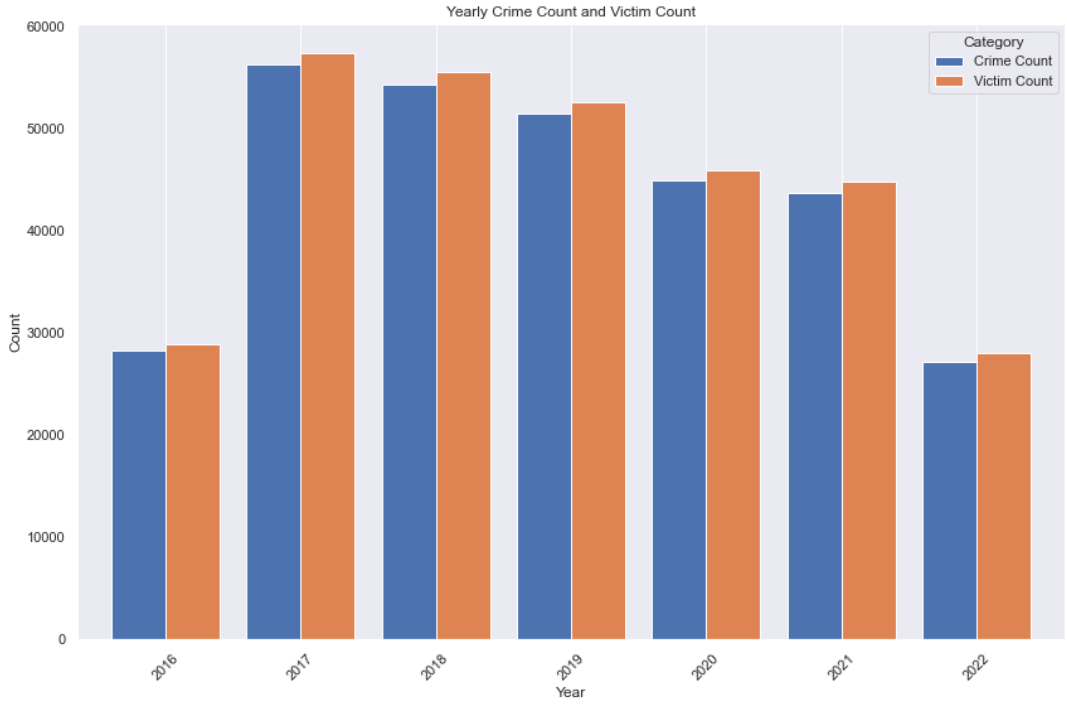
Geospatial Maps:
Show geographic distribution of data points.
Example: Crime hotspots using latitude and longitude data.

Research Questions and Visualizations

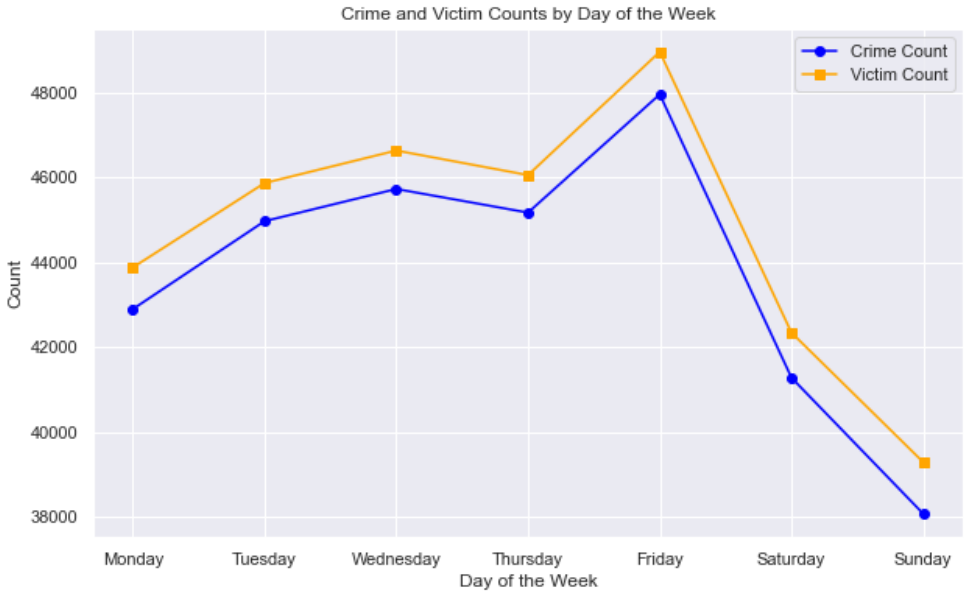
Note:
We observed that in the provided dataset, data for the year 2016 is available only from July to December, while data for the year 2022 is available only up to June.

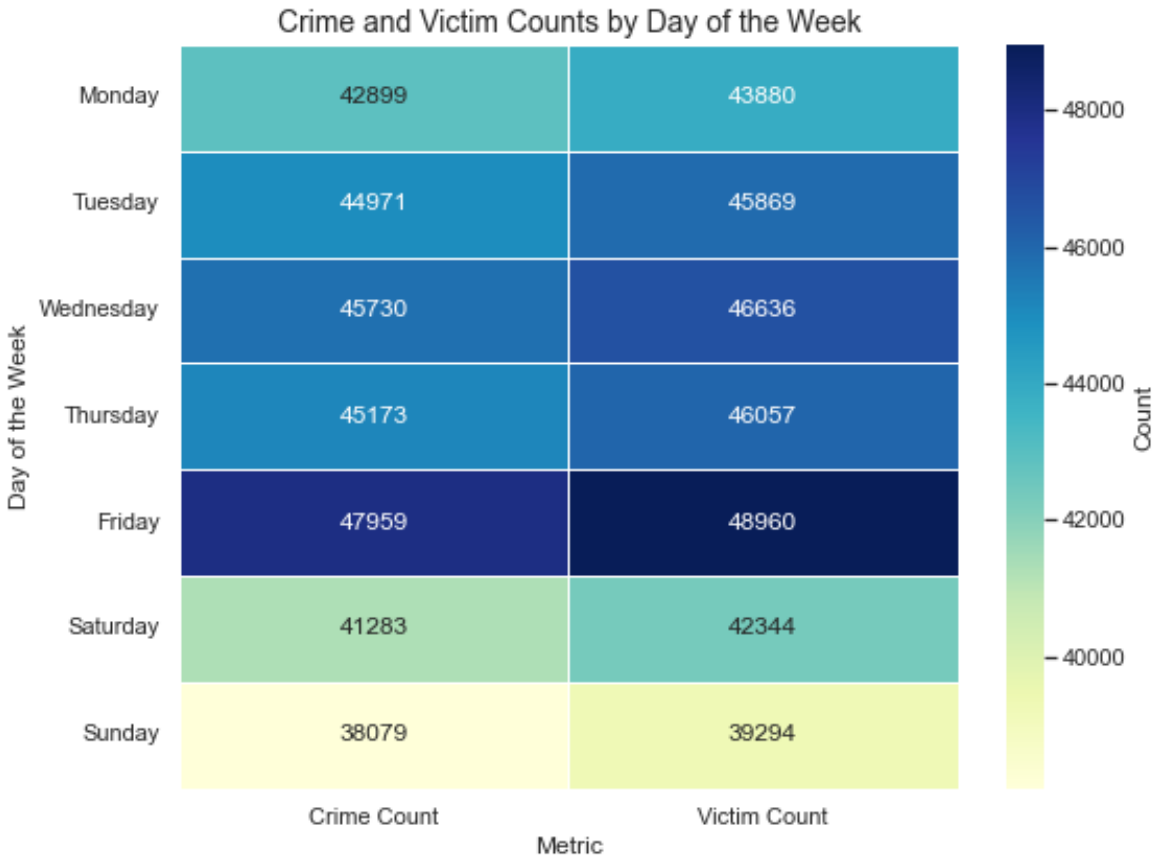
1. What are the observable year-over-year trends in the frequency of crimes and victim counts?



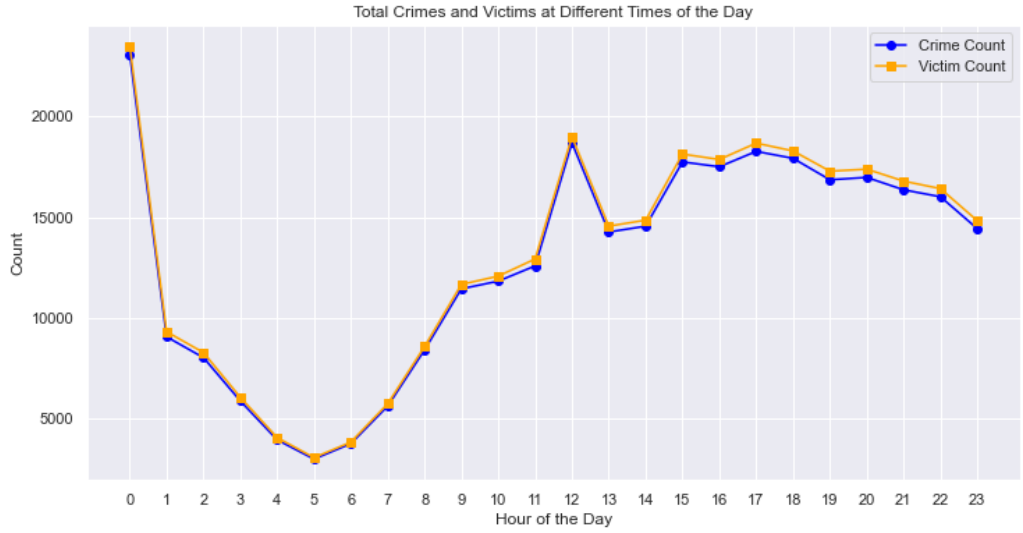


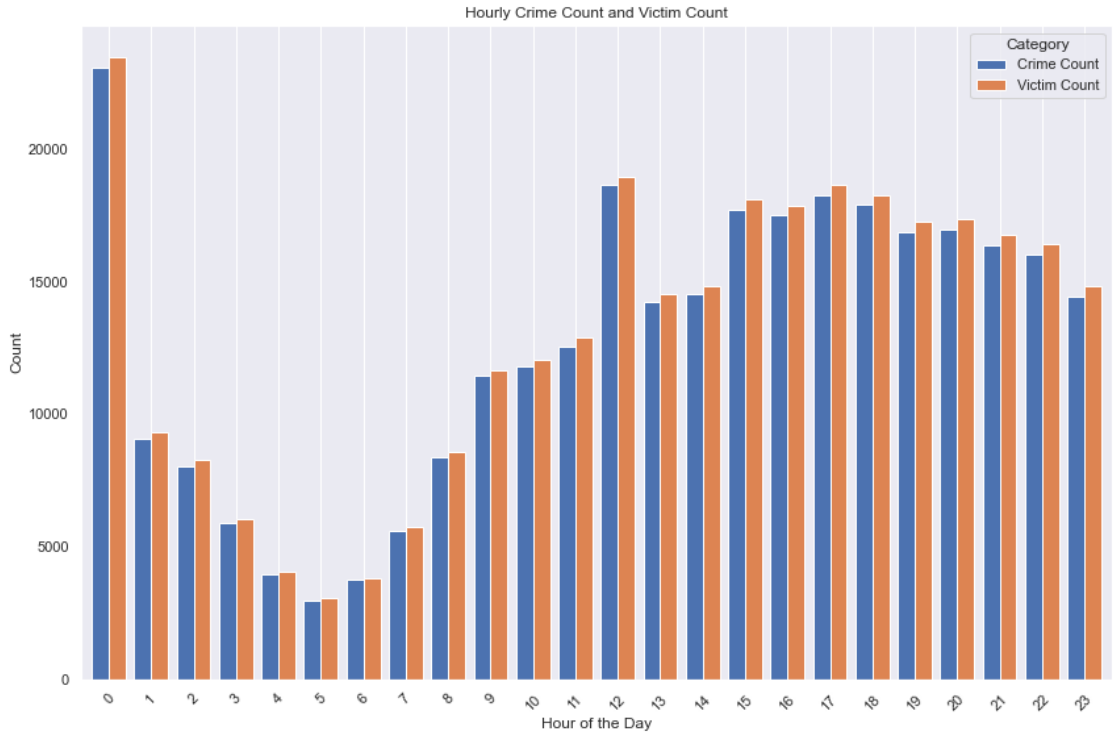
2. Which day of the week records the highest and lowest number of crimes or victims?



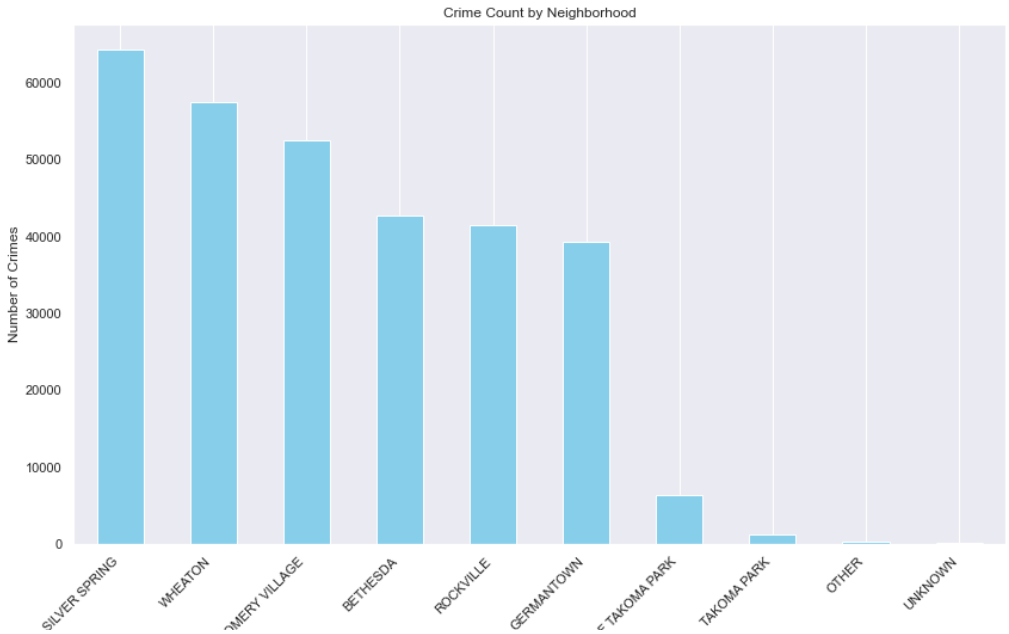


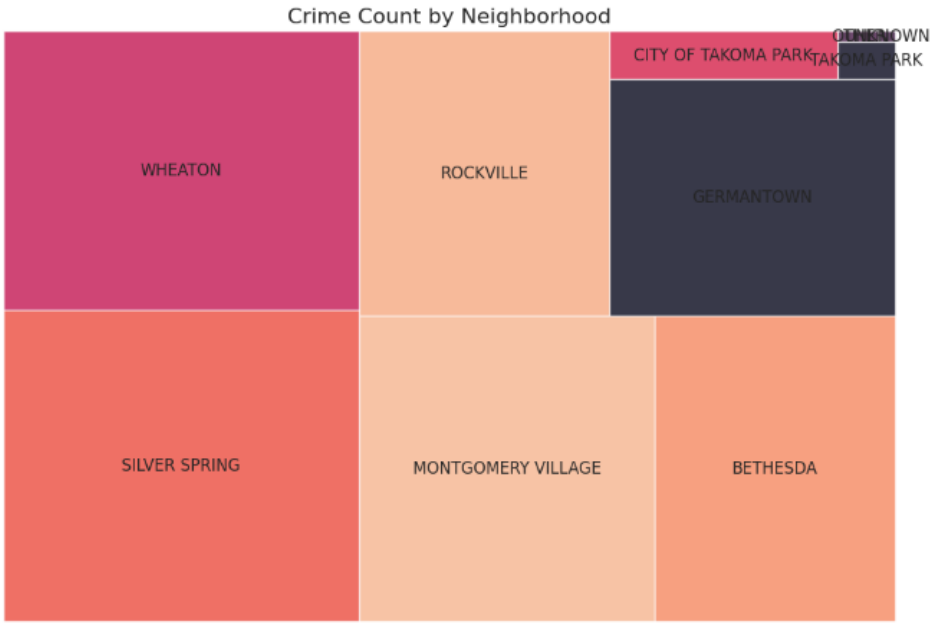
3. What is the temporal distribution of crimes across different times of the day?



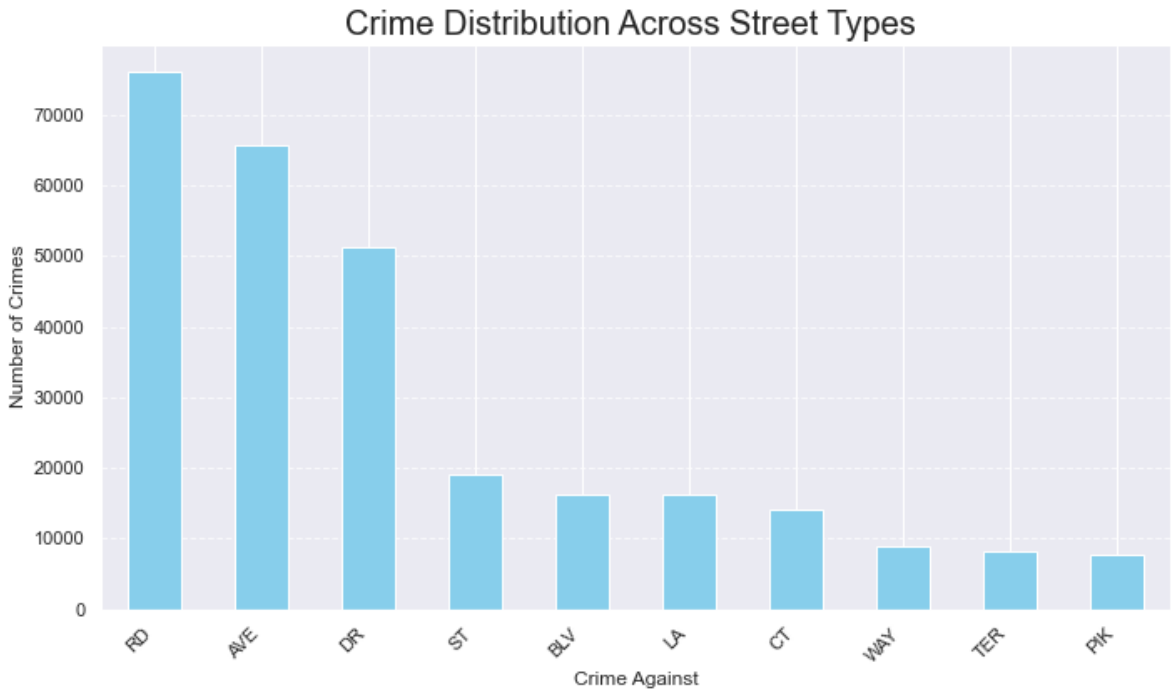


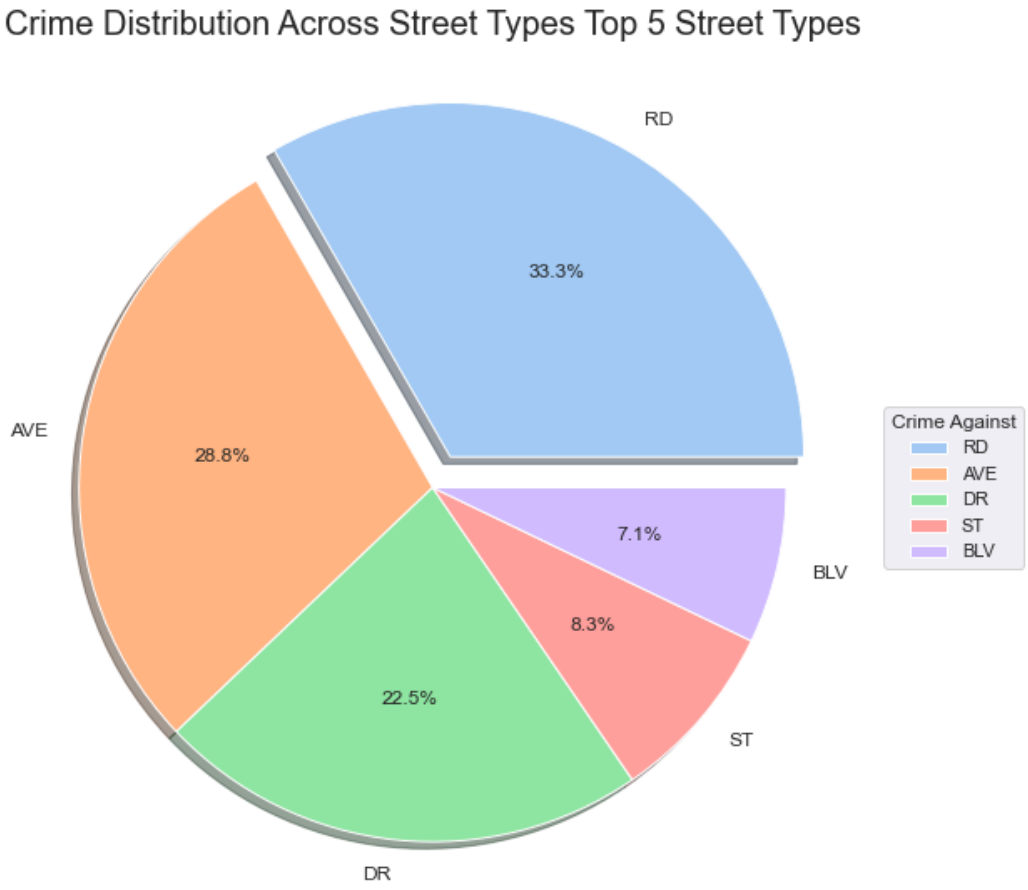
4. Which Neighbourhood / Police District has the greatest number of crimes in the county?



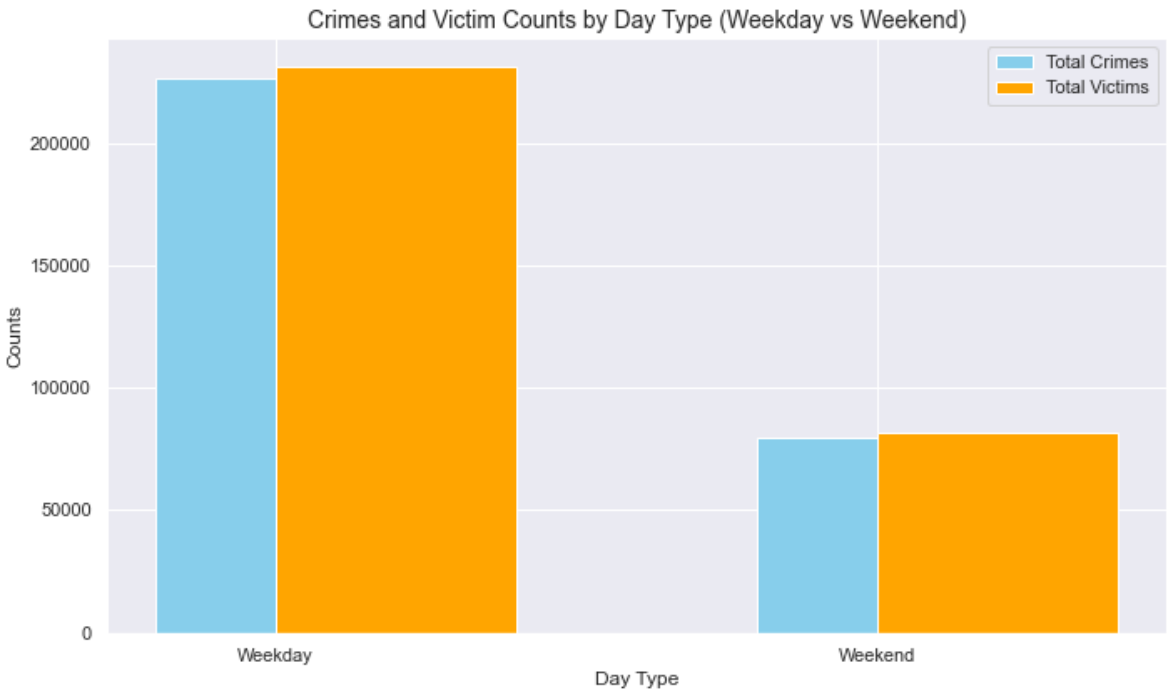


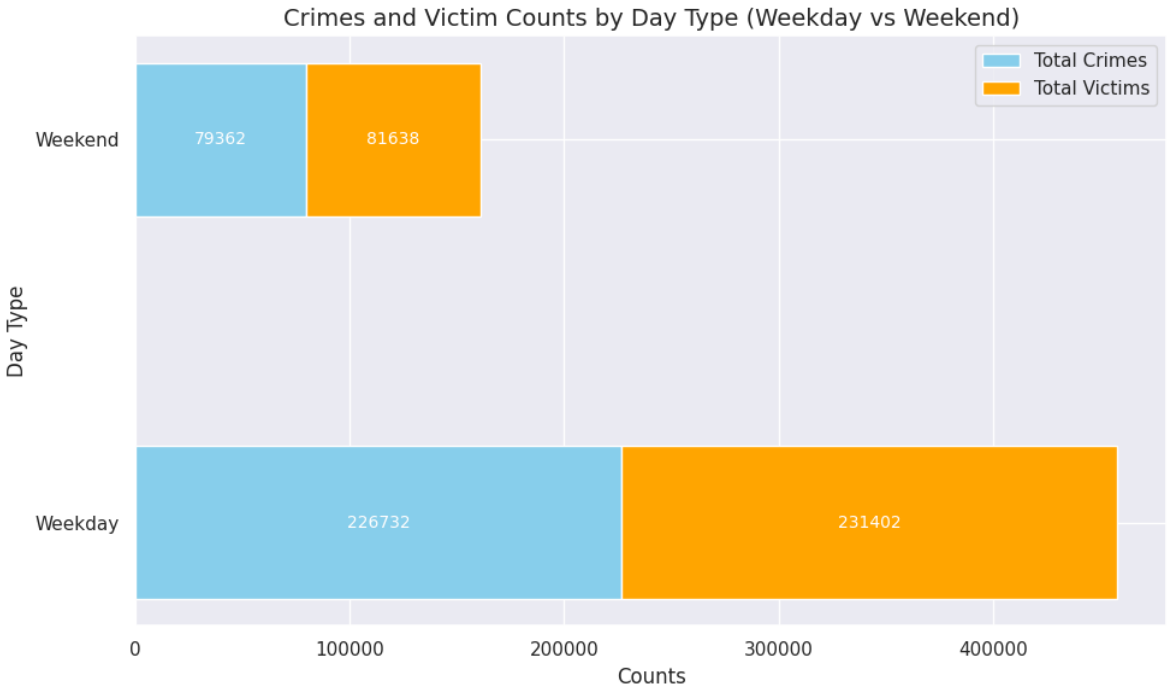
5. Which type of locations (street types) mostly affected by crimes?



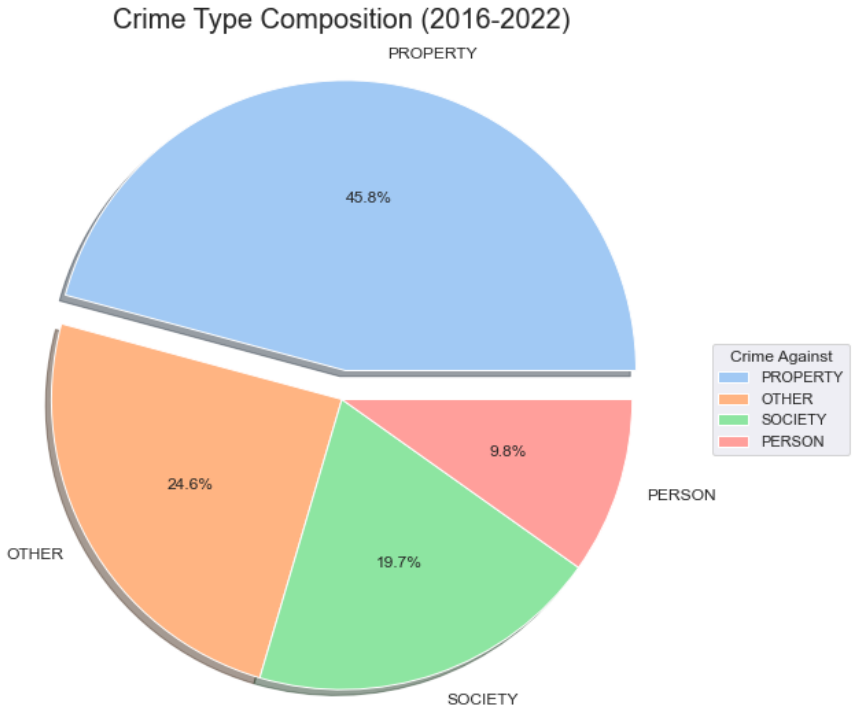


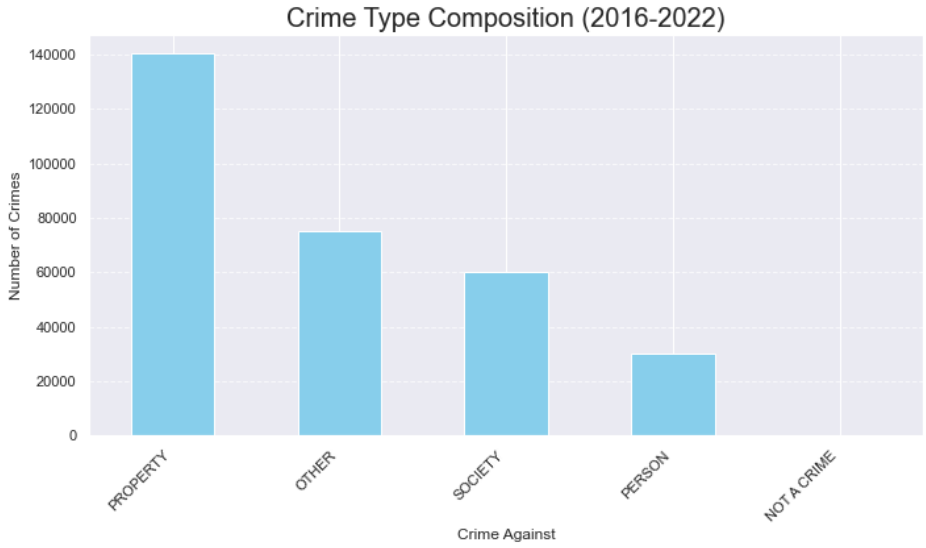
6. What is the comparative distribution of crimes between weekdays and weekends?



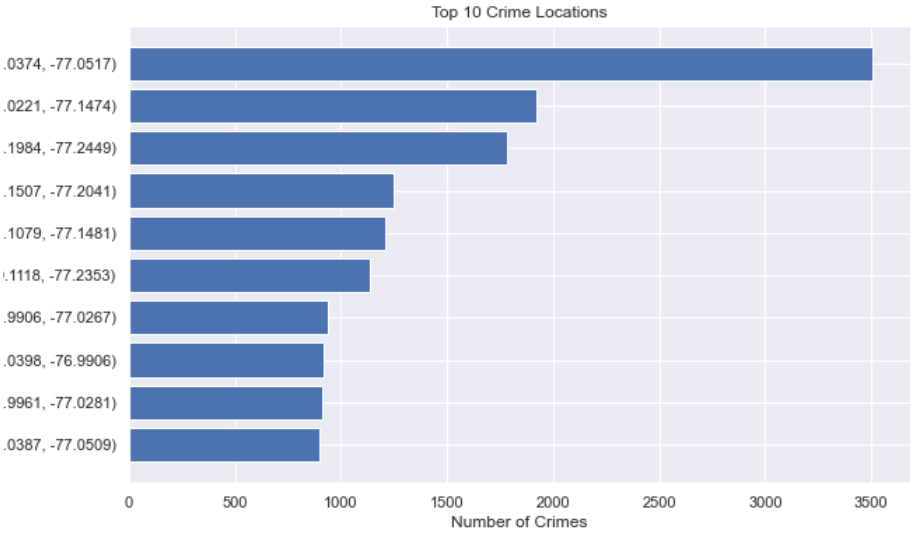


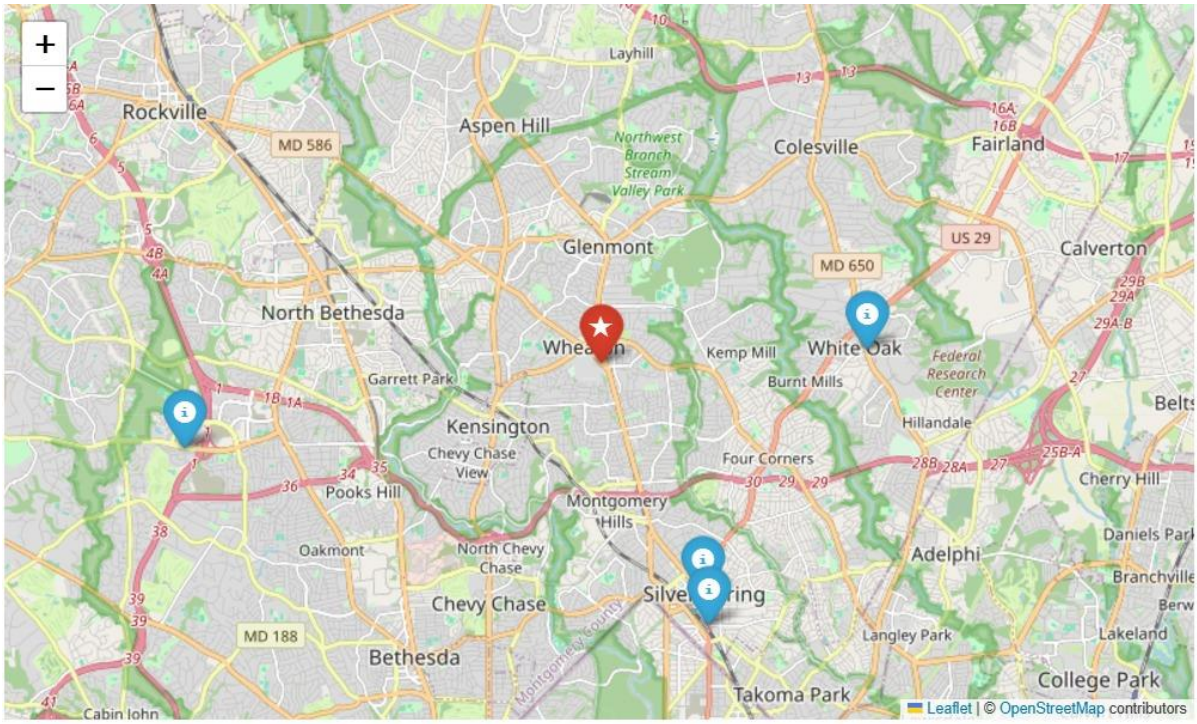
7. Which category is most impacted by crimes: persons, property, or society?





8. What is the location which has the highest crimes reported?





3. Summary and Conclusion

This study aimed to analyse crime trends in Montgomery County using the provided dataset and answer key research questions related to the frequency, nature, and patterns of crimes. The analysis covered various dimensions, including temporal trends, geographic hotspots, and relationships between different crime types. Below is a summary of the key findings:

a. Summary of Findings

1. Trends in Crime and Victims over the Years:

    - A clear downward trend in both crime incidents and victim counts has been observed since 2017, indicating potential improvements in public safety or reporting changes.

2. Day of the Week Patterns:

    - Highest Crimes: Friday experiences the highest crime and victim counts, likely due to increased activity as the weekend begins.
    - Lowest Crimes: Sunday has the lowest crime rates, reflecting reduced activity levels.

3. Temporal Patterns:

    - The majority of crimes occur around midnight, emphasizing the need for heightened surveillance during late-night hours.

4. Geographic Hotspots:

    - The Silver Spring neighbourhood has the greatest number of reported crimes, suggesting it is a significant area of concern for law enforcement.

5. Which type of locations (street types) mostly affected by crimes?

    - High concentration of crimes on certain street types (e.g., Road and Avenue) suggests that these areas may act as crime hotspots. Police resources and patrols might need to be allocated more effectively to these areas.
    - These locations may require additional manpower, surveillance, or community policing initiatives to deter criminal activities.

6. Weekday vs. Weekend Trends:
    - Most crimes occur during weekdays, which could be attributed to higher population movement, work-related activities, and urban interactions.

7. Most Affected Categories:

Crimes are predominantly committed against property, emphasizing the importance of implementing robust property protection strategies to mitigate such incidents effectively.

8. What is the location which has the highest crimes reported?

    - The location with the highest crime count, corresponding to the location code [39.0374, -77.0517]. This highlights the need for focused law enforcement efforts and resource allocation in this area to address the elevated crime rates effectively

b. Conclusion

The analysis provides valuable insights into the patterns and trends of crimes in Montgomery County. The findings can guide law enforcement and policymakers in implementing targeted interventions to reduce crime rates further and improve public safety.

c. Recommendations:

1. Enhance late-night patrols and monitoring in high-crime areas like Silver Spring.
2. Implement strategies to address weekday crime trends, such as workplace awareness programs and improved urban planning.
3. Focus on property crime prevention, particularly vehicle security, by increasing public awareness and deploying anti-theft measures.
4. Address the link between drug-related crimes and prostitution through coordinated social programs and enforcement strategies.

This study highlights the importance of data-driven decision-making in crime prevention and provides a strong foundation for future research and predictive modelling efforts to further improve community safety.

6. References

Biswal, A., 2024. *Simplilearn.* [Online]
Available at: https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20is%20a%20critical%20step%20in%20the%20data,various%20statistical%20and%20graphical%20techniques.

Biswal, A., 2024. *What is Data Science: Lifecycle, Applications, Prerequisites and Tools.* [Online]
Available at: https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science

Cameron Hashemi-Pour, 2024. *TechTarget.* [Online]
Available at: https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

Ian David Edge, D. A. T., 2024. *crime.* [Online]
Available at: https://www.britannica.com/topic/crime-law