Kingston
University
London

Crime Analysis and Visualisation:

Montgomery County Case Study

Group Members and Participation

Student Name
ASHLEY DAUD
SAMRUDDHI SANJAY MENGE
FANNY NAMONDO NGOMBA
DEVASHISH SHARAD VIDHATE

Table of Contents

List of Figures

List of Tables

Keywords

- EDA - Exploratory Data Analysis
- IDA – Initial Data Analysis
- NIBRS - National Incident-Based Reporting System
- CJIS - Criminal Justice Information Services
- UCR - Division Uniform Crime Reporting
- ECC - Emergency Communications Centre
- MC – Montgomery County

Abstract

This research report aims to provide a comprehensive analysis of crime data in Montgomery County from 2018 until 2022. Through examining the dataset obtained from National Incident-Based Reporting System (NIBRS) of the Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program, this study identifies patterns and trends in criminal activities. The analysis focuses on determining common crimes, crime rates over time, geographical distribution of crimes, and correlations between crime types and time. The report also investigates response times, victim counts, and the impact of crime locations on public safety. By leveraging structured data in a CSV file format, this study utilises Python pandas, Matplotlib, Seaborn, Numpy, Missingno library etc for data manipulation and visualisation to derive valuable insights for enhancing public safety and preventing future crimes in Montgomery County.

1. Introduction

Crime, defined as an action or omission that constitutes an offense and is punishable by law, can result in physical or psychological harm, or damage to property. Data science refers to a set of principles, problem definitions, algorithms, and processes used to extract non-obvious and useful patterns from large data sets (Kelleher & Tierney, 2018). These large data sets are analysed across various industries such as finance, marketing, healthcare, and law to uncover patterns that can inform decision- making (data.montgomerycountymd.gov, 2024). The application of data science in crime analysis facilitates crime prevention , detection and investigation using surveillance, social media, and public records. In addition, it optimises resource allocation by analysing the crime patterns enabling law enforcement to be more effective with resources (Delgado et al., 2021).

According to the International Association of Crime Analysts (IACA), crime analysis is a type of analysis conducted within a police organisation, excluding any evidence analysis (IACANET, 2024). Crime data analysis is used to analyse reports and crimes based on historical data and information reported at the time e.g., the type of crime, the number of victims and the time stamp of the crime. This analysis aids various stakeholders, including law enforcement, government, and policymakers, in preventing and responding to crime effectively based on the gathered and analysed data. Through crime data analysis, patterns and trends are identified which then usually leads to strategies being developed, monitor policy effectiveness and crimes being solved using multiple sources.

The results of these data analysis are often presented through various visualisations, including images, graphics, and charts, which help decision-makers understand large data sets and identify hidden patterns. It is easier for humans to comprehend complex and large amounts of data when they are visualised using images or graphics rather than when they are presented in a tabular or written format (Setiawan & Suprihanto, 2021).

The integration of data science into crime analysis not only improves the efficiency and effectiveness of law enforcement agencies but also makes a substantial contribution to public safety. Stakeholders can make informed decisions that result in the development of proactive strategies and the optimal allocation of resources like community engagements and community programmes (Haslett et al., 2012). Thereby reducing crime rates and improving community well-being, by utilising advanced analytical techniques and visualisations.

1.1 Literature Review

Montgomery County(MC), located in Maryland, USA, is a significant region known for its diverse population and proximity to Washington, D.C - a mix of city and suburban areas with a diverse population. The dataset from the Montgomery Country represents the crime statistics over several years which was extracted from reported crimes classified by National Incident-Based Reporting System (NIBRS) of the Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program.

According to the 2023 annual report on crime and safety in MC:

> *The Montgomery County Emergency Communications Center (ECC) received 862,472 calls for service, approximately 4% more calls than 2022. Sixty-six percent (66%) of the calls received by the ECC were emergency calls, an average of 1,554 emergency calls per day which represents an increase of 4% from 2022. There were 273,114 non-emergency calls which is up 6% from 2022.*
>
> (The Policy and Planning Division, 2024).

In addition to the rise in call volume, the report identified several significant trends in crime and public safety. Overall, there was a 10% increase in crime rate in MC compared to the previous year. These findings underscore the critical need for continuous efforts to enhance public safety and address the evolving challenges faced by the community. The data collected and analysed by the ECC, and law enforcement agencies are influential in developing effective crime prevention strategies and strategic allocation of resources to areas with the greatest need (The Policy and Planning Division, 2024).

The dataset from MC, analysed in this report, provides valuable insights into the safety issues faced by its residents. Across narrowing down and analysing the data between the years 2018 to 2022, patterns and trends in criminal activities can be identified. This information can be used to enhance public safety and support local government efforts to reduce crime. The analysis and visualisations aim to understand the pattern of crimes occurring in the county and may help suggest the to the local government, and potentially on a national level, to mitigate crimes and predict future crime patterns.

The capacity of data science to resolve intricate criminal cases has been illustrated by the utilisation of genealogy websites and DNA services. The GEDMatch database was employed by the police in May 2019 to resolve a murder case that had been ongoing for decades. Researchers were able to identify the murderer, who also confessed to three additional murders, by examining genetic information and extended family trees (Kennedy & Hilling, 2023).This case underscores the importance of leveraging large databases and advanced analytical techniques to uncover critical connections and resolve ongoing investigations

Correspondingly In 2016 , Pittsburgh, Pennsylvania had the highest murder rate amongst all large US cities. The police department decided to work with data scientists to make a difference and started a ‘predictive policing’ program. The program involved data scientists analysing past crime reports and considering factors such as demographics and social factors to identify trends and patterns. This initiative resulted in a 10% decrease in violent crimes in Pittsburgh (Hvistendahl, 2016).

These studies show the vital role of data science in crime prevention and public safety. Continuous research is key to tackling evolving crime challenges. Insights will enhance academic understanding and provide practical recommendations for effective crime prevention strategies.

1.2 Problem Statement and Research Questions

The aim is to understand and mitigate crime in MC using a detailed crime dataset. By analysing this data, the aim is to identify patterns and trends, determine crime types and frequencies, and develop strategies to reduce crime rates. This analysis will aid local government efforts to enhance public safety and predict future crime trends, contributing to a safer community.

The following research questions will be investigated using the dataset provided, focusing on the years 2018 to 2022. The answers will be presented through visualisations:

1. What are the ten most common crimes in the Montgomery County crime dataset?
2. During which times of the year are crimes most and least frequent in Montgomery County?
3. How have crime rates evolved from 2018 to 2022?
4. Which ten cities in Montgomery County have the highest crime rates?
5. Is there a direct correlation between the top 10 types of crimes and the time they occur?
6. Which agencies report the highest and lowest crime counts, and what insights can be drawn from these variations?
7. How do crime patterns change depending on the day of the week?
8. What are the top 10 locations where crimes frequently occur?
9. What is the relationship between crimes and response times?
10. What is the top 10 annual number of victims impacted by specific types of crimes?

By addressing these research questions, this analysis aims to provide a comprehensive understanding of crime patterns in MC.

1.3 Methodology

Data Science is an interdisciplinary field that combines statistics, computer science, mathematics, and domain expertise to analyse large data sets and solve complex problems. It requires tools, skills, algorithms, and concepts from various academic frameworks (Setiawan & Suprihanto, 2021).

The research methodology follows these steps: Business Understanding, Data Understanding, Data Preparation, and Exploratory Data Analysis (EDA).

- **Business Understanding**: Identifies key variables, sets metrics for success, and aligns the project with business goals.
- **Data Understanding**: Involves identifying dataset variables, using methods like *df.shape* and *df.info()* to gain insights.
- **Data Preparation**: Includes data integration, cleaning, imputation of missing values, handling noisy data, and reduction.
- **EDA**: Explores data to identify pattern (Mayernik, 2023).

Upon completing the initial four steps, the visualisations will facilitate the creation of a data model. This model can subsequently be evaluated and deployed in alignment with the business objectives.

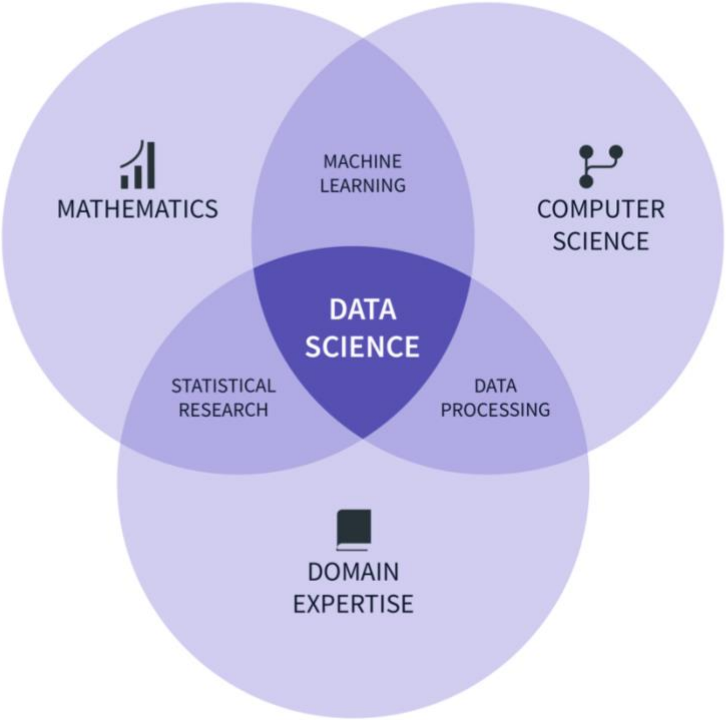


Figure 1 : The interdisciplinary field of Data Science

These steps ensure the quality of data and help shape research questions, as shown in Figures 1 and 2.

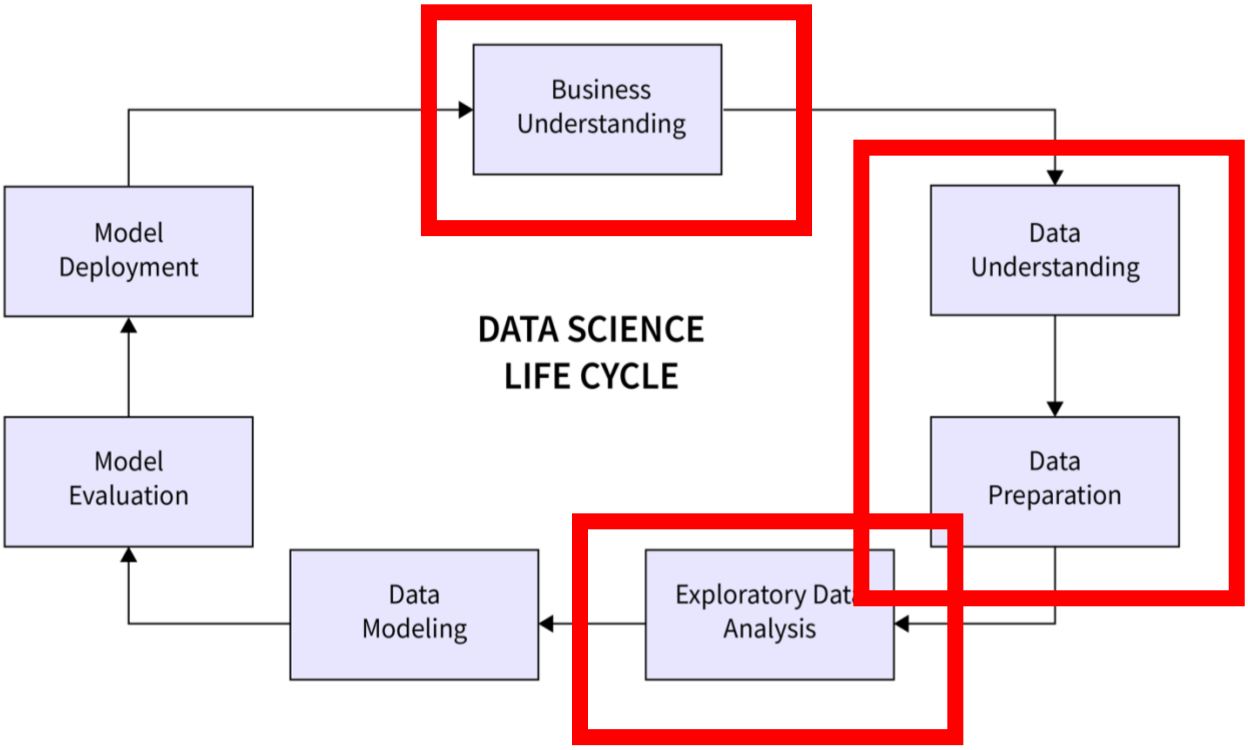


Figure 2 : Data Science Process Life Cycle

1.4 Python Libraries and Packages

Python is the chosen programming language for data visualisation in this analysis due to its numerous benefits:

- **User-Friendly Syntax**: Simplifies development and reduces errors.
- **Rich Library Ecosystem**: Offers robust tools for diverse visualisations.
- **Strong Community Support**: Ensures continuous enhancement of libraries and tools.
- **Versatile Integration**: Seamlessly integrates with other languages and technologies.

These attributes make Python ideal for creating efficient, accurate, and visually appealing data representations (see table 1).

Table 1 : Python libraries used for the data visualisation

Python Libraries	Category	Description	Pros	Cons
PANDAS	Manipulation and Visualisation	Data manipulation library with basic plotting capabilities.	▪ Easy to use	▪ Basic plotting compared to specialised libraries.
NUMPY	Scientific Computing	For scientific computing with the support for arrays	▪ Supports mathematical operations ▪ Efficient handling of large data sets	▪ Not primarily a visualisation library, needs integration with others.
MISSINGNO	Visualisation	Library for visualising missing data patterns.	▪ Easy to use	▪ Limited to missing data visualisation.
MATPLOTLIB	Visualisation	A library for static, animated, and interactive visualisations.	▪ Customisable ▪ Supports various plots	▪ Requires more codes for customisation
SEABORN	Statistical	An interface for drawing attractive statistical graphs	▪ Easy to use ▪ Integrates with pandas	▪ Dependant on Matplotlib : limits statistical visualisations
SQUARIFY	Visualisation	Library used to create treemaps	▪ Useful for hierarchical data	▪ Limited to Treemaps visualisations

Adapted from (Tableau, 2024)

Figure 3 and Figure 4 show different python libraries that are used worldwide.

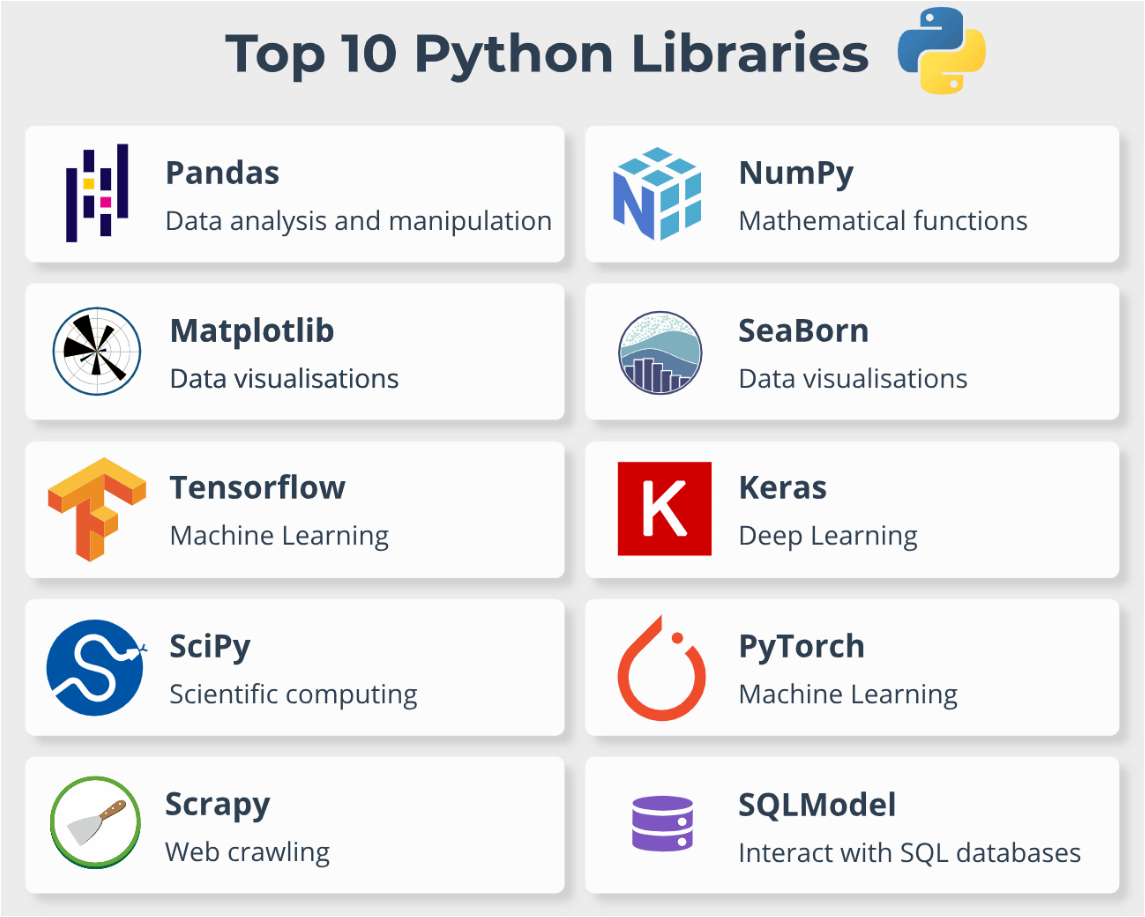


Figure 3 : Most popular python libraries used worldwide

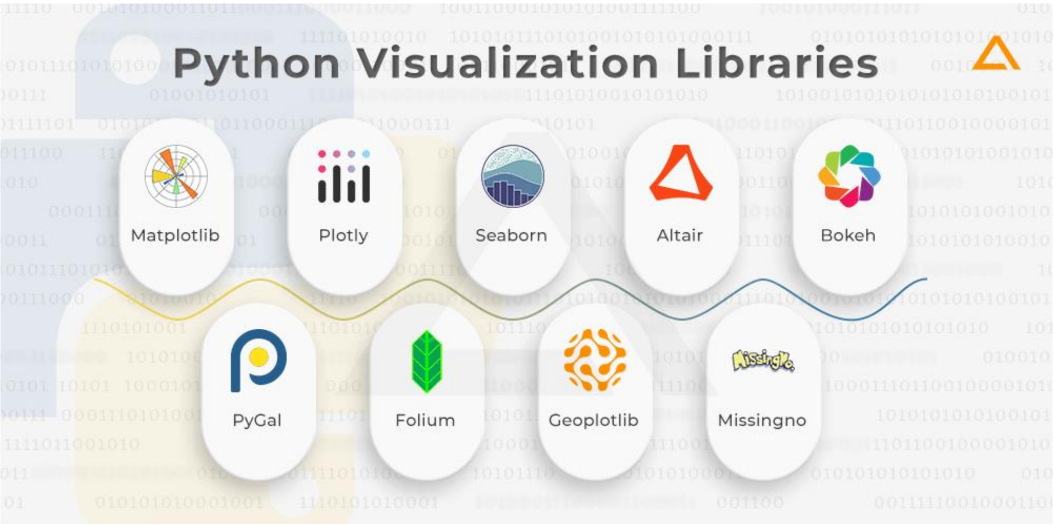


Figure 4: Most popular python visualisation libraries used worldwide

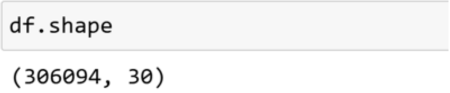
2. Preliminary Data Analysis

2.1 Dataset

The dataset that represents the MC originated from the county’s open data website (dataMontgomery) which allows the public to have direct access to the crime statistics database. The identities of the victims and offenders are not released in this database due to privacy protection reasons however all founded crimes from July 1st 2016, are entered into the database. The crime reports available are based on preliminary information obtained by the police departments in the reporting parties and are updated on a quarterly basis to show any changes based on the on-going police investigations (data.montgomerycountymd.gov, 2024).

The data was downloaded in the structured data format of a csv file with a memory usage of 70.1+ megabytes, in October 2024. A Comma Separated Value (CSV) file is a text file that stores structured data in a table format. Each line of the file represents a data record, which corresponds to a new row in the table and each value in the line is separated by a comma, representing a value in a separate column (Python Software Foundation, 2024). The dataset contains 30,6094 rows and 30 columns, which were confirmed using the .shape property in the Python pandas library. Each column referred to a new name (see Table 2)This property confirms the number of rows and columns in a Dataframe (see figure 5). The column names were displayed in the original data source, with some being abbreviated. However, the definitions of the columns were provided (see Table 2).

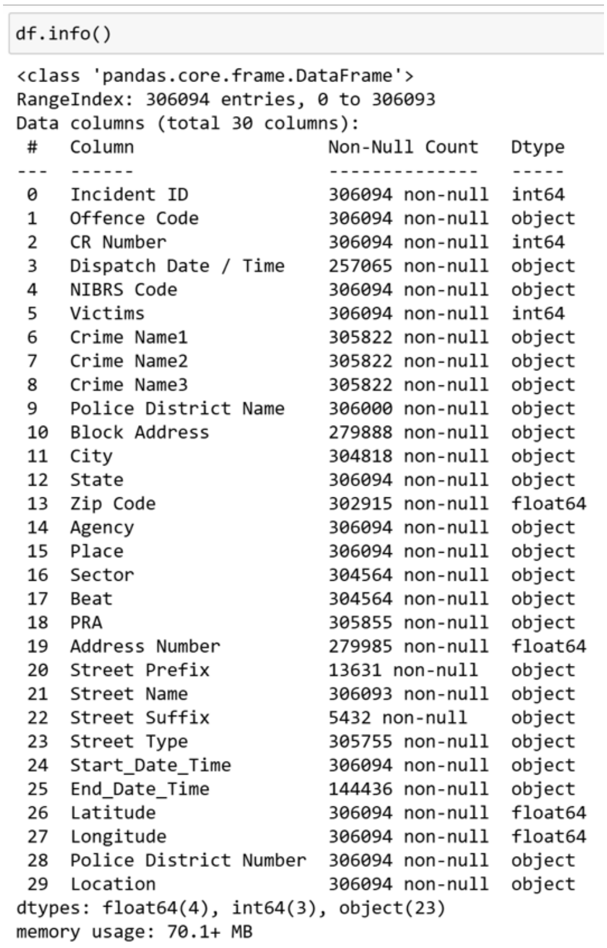
To obtain detailed information about the columns in the DataFrame, we used the ***.info()*** method from the Pandas library. This method provided important details such as the index and column names, the count of non-null values and the data types of each column of the DataFrame. The columns had 3 datatypes identified: float64 , int64 and object (see figure 6).



```
df.shape

(306094, 30)
```

Figure 5 : .shape method being used on the data frame



```
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
dtypes: float64(4), int64(3), object(23)
memory usage: 70.1+ MB
```

Figure 6: .info method shows information about the MC Dataframe including the datatype of each column

Table 2: The description of the column names in the Dataframe

Column Name	Description	Column Name	Description
Incident ID	Police Incident Number	Zip Code	Zip code
Offence Code	Offense Code is the code for an offense committed within the incident as defined by the National Incident-Based Reporting System (NIBRS) of the Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program.	Agency	Assigned Police Department
CR Number	Police Report Number	Place	Place description
Dispatch Date / Time	The actual date and time an Officer were dispatched	Sector	Police sector name, a subset of District
Start_Date_Time	Occurred from date/time	Beat	Police patrol area, a subset of Sector
End_Date_Time	Occurred to date/time	PRA	Police Response Area, a subset of Beat
NIBRS Code	FBI NIBRS codes	Address Number	House or Business Number
Victims	Number of Victims	Street Name	Street Name
Crime Name1	Crime against Society/Person/Property or Other	Street Suffix	Quadrant (NW, SW, etc)
Crime Name2	Describes the NIBRS_CODE	Street Type	Ave, Drive, Road, etc
Crime Name3	Describes the OFFENSE_CODE	Street Prefix	North, South, East, West
Police District Name	Name of District	Latitude	Latitude
Block Address	Address in 100 block level	Longitude	Longitude
City	City	Location	Location
State	State	Police District Number	Major Police Boundary

Table 2 : Adapted from Source : (Montgomery County, MD, 2015)

2.2 Data Quality Initial Assessment

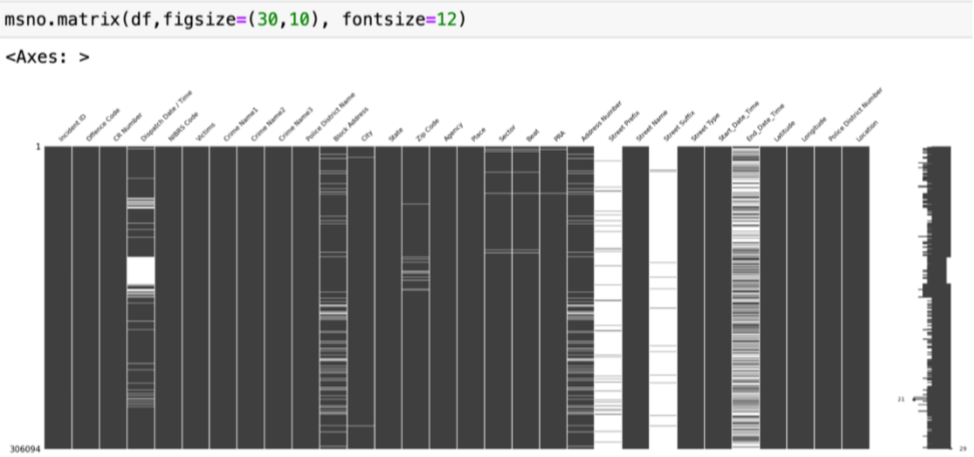
When the data was assessed, several quality issues were discovered that needed to be addressed to ensure the reliability and accuracy of our analysis also different columns present had different datatypes (see Table 3 and Table 4).

Table 3: Data Quality Issues and Resolutions

Issues	Description	Resolution
Missing and Null Values	Many missing or null values that could lead to incomplete or biased results.	Removed rows and columns with missing values; filled critical missing values with specific values or imputation methods.
Data Inconsistencies	Columns with different formats and unexpected data types complicating analysis.	Standardised data types across columns to ensure consistency and accuracy.

Table 3 : Issues identified in the dataframe

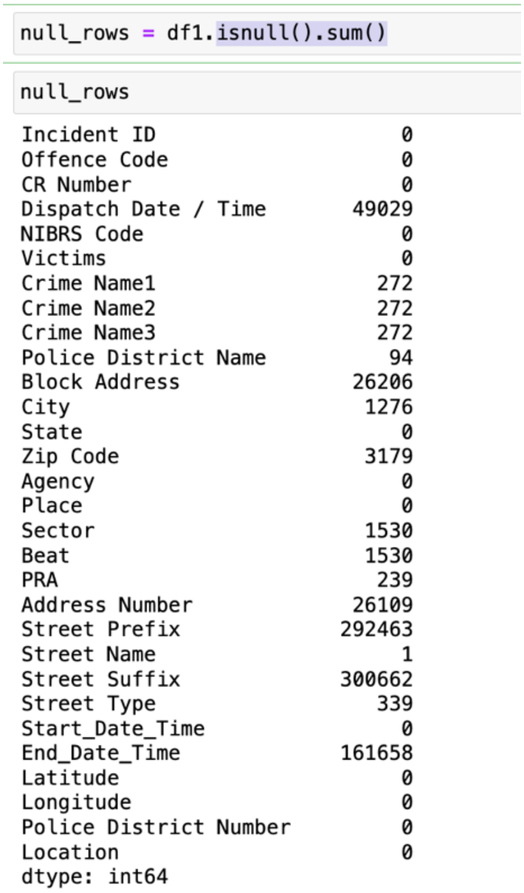
Missing and null values can lead to incomplete or biased results, compromising analysis reliability. Data inconsistencies, such as different formats and unexpected data types, complicate processing and may cause errors. These issues underscore the need for comprehensive data quality assessment. The Python library Missingno was used to visualise missing data and analyse patterns , by the white lines representing missing data in the dataframe (McDonalds, 2021). The .matrix method helped visualise these patterns, while the .isnull().sum() method in pandas returned the exact number of missing values per column, aiding in data cleaning(see figure 8).



```
msno.matrix(df,figsize=(30,10), fontsize=12)
```

```
<Axes: >
```

Figure 7: Missingno matrix used to visualise pattern of missing data



```
null_rows = df1.isnull().sum()
```

```
null_rows
```

```
Incident ID                   0
Offence Code                  0
CR Number                     0
Dispatch Date / Time      49029
NIBRS Code                    0
Victims                       0
Crime Name1                 272
Crime Name2                 272
Crime Name3                 272
Police District Name         94
Block Address             26206
City                       1276
State                         0
Zip Code                   3179
Agency                        0
Place                         0
Sector                     1530
Beat                       1530
PRA                         239
Address Number            26109
Street Prefix            292463
Street Name                   1
Street Suffix            300662
Street Type                 339
Start_Date_Time               0
End_Date_Time            161658
Latitude                      0
Longitude                     0
Police District Number        0
Location                      0
dtype: int64
```

Figure 8: .isnull().sum() used within the data cleaning process

Table 4: Datatypes in columns

KEY

Int = Represents integer data type, which includes whole numbers without decimal points.

Float64 = Represents a floating-point data type with double precision, used for numbers that require decimal points.

Object = Represents a general data type in pandas, often used for text or mixed data types.

Column Name	Data Type	Column Name	Data Type
Incident ID	Int	Zip Code	Float64
Offence Code	Object	Agency	Object
CR Number	Int	Place	Object
Dispatch Date / Time	Object	Sector	Object
Start_Date_Time	Object	Beat	Object
End_Date_Time	Object	PRA	Object
NIBRS Code	Object	Address Number	Float
Victims	Int	Street Name	Object
Crime Name1	Object	Street Suffix	Object
Crime Name2	Object	Street Type	Object
Crime Name3	Object	Street Prefix	Object
Police District Name	Object	Latitude	Float64
Block Address	Object	Longitude	Float64
City	Object	Location	Object
State	Object	Police District Number	Object

Table 4 : Adapted from Source : (Montgomery County, MD, 2015)

2.3 Data Pre-Processing and Cleaning

Pre-processing and cleaning data are essential steps in getting a dataset ready for analysis. To guarantee that the data is reliable, consistent, and usable, this procedure entails several crucial steps. To start, the dataset is streamlined by eliminating unnecessary observations, such as duplicates and irrelevant data.

The process of altering data's format or structure to make it compatible with a target system or suitable for analysis is known as data transformation. It is frequently utilised in data integration, migration, and warehousing and is an essential component of data management (see Table 5 and Table 6).

Table 5: Data Cleaning Steps and Descriptions

STEPS	DESCRIPTION
STEP 1: Remove duplicate or irrelevant observations	Unwanted observations, including duplicates and irrelevant data, should be removed from the dataset to ensure efficient analysis and create a more manageable, performant dataset
STEP 2: Fix structural errors	Structural errors, such as strange naming conventions, typos, or incorrect capitalisation, can cause mislabelled categories, requiring standardization to ensure consistent analysis.
STEP 3: Filter unwanted outliers	One-off observations that do not fit the data should be evaluated for validity, and if deemed irrelevant or erroneous, they should be removed to improve data performance.
STEP 4: Handle missing data	Missing data must be addressed because many algorithms cannot handle it; options include dropping observations, imputing missing values, or altering data usage to navigate null values, each with potential drawbacks.
STEP 5: Data Validation and Quality Control	At the end of the data cleaning process, basic validation should ensure the data makes sense, follows appropriate rules, supports, or refutes theories, reveals trends, and addresses any data quality issues to avoid false conclusions and poor decision-making.

Table 5 : Steps for Data Cleaning in Research Analysis, Adapted from Source (Tableau, 2024)

Table 6 : Data Attributes and Cleaning Actions

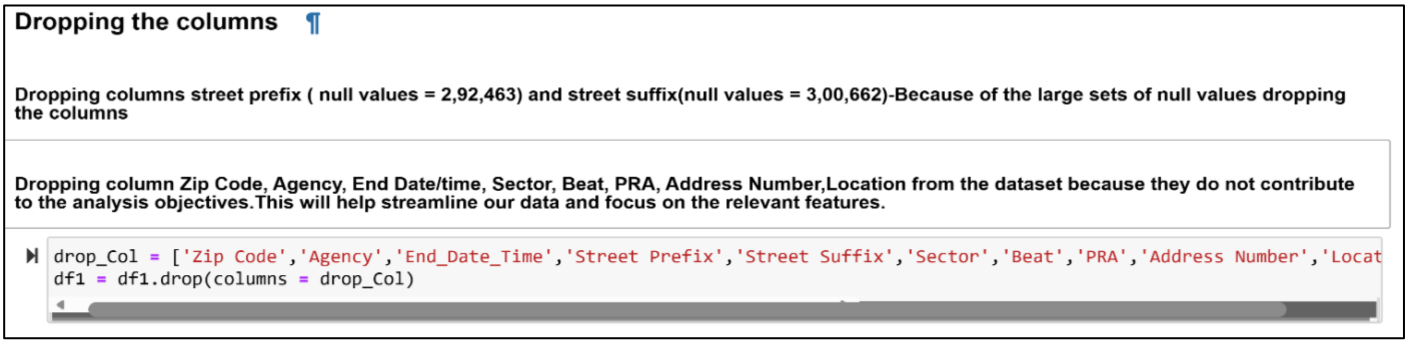
Column Name	Data Type	Null Values	Action Completed	Reason
Incident ID	int	0	-	
Offence Code	object	0	-	
CR Number	int	0	-	
Dispatch Date /Time	object	49029	Start_Date_Time*	
Start_Date_Time	object	0	Split into separate columns : Hour, Day, Month, Year, Day of the week	
End_Date_Time	object	0	-	
NIBRS Code	object	0	-	
Victims	int	0	-	
Crime Name1	object	272	Null values replaced with 'Unknown'	
Crime Name2	object	272	Null values replaced with 'Unknown'	
Crime Name3	object	272	Null values replaced with 'Unknown'	
Block Address	object	26206	Null values replaced with 'Unknown'	
City	object	1276	Null values replaced with 'Unknown'	
State	object	0	-	
Zip Code	float	3179	-	Column Dropped
Agency	agency	0	-	
Place	object	0	-	Column Dropped
Sector	object	1530	-	Column Dropped
Beat	object	1530	-	Column Dropped
PRA	object	239	-	Column Dropped
Address Number	float	26109	-	Column Dropped
Street Prefix	object	292463	-	Column Dropped
Street Name	object	1	Null values replaced with 'Unknown'	
Street Suffix	object	300662	-	Column Dropped
Street Type	object	339	Null values replaced with 'Unknown'	
Location	object	0	-	Column Dropped
Latitude	float	0	-	
Longitude	float	0	-	
Police District Name	object	94	Null values replaced with 'Unknown'	

**Dispatch date/time was replaced with start_date_time instead of being set to 0. Since the Dispatch date/time values are the same as start_date_time, we assumed that the missing values for Dispatch Date/Time would match those of start_date_time, which has no missing values*

Table 6 : Data cleaning per column in the dataframe

The following details the specific data pre-processing, cleaning, and transformation steps undertaken to prepare the dataset for effective data visualisation.

1. **Dropping columns**: Several columns were removed from the dataset for the following reasons (see figure 9):
   - **Irrelevant Data**: Columns like "zip_code" and "Agency" were excluded as they did not support the analysis objectives.
   - **High Percentage of Missing Values**: Columns such as "street prefix" and "street suffix" had over 50% missing data.
   - **Redundancy**: The "location" column was redundant, duplicating information in the "longitude" and "latitude" columns.



Dropping the columns ¶

Dropping columns street prefix (null values = 2,92,463) and street suffix(null values = 3,00,662)-Because of the large sets of null values dropping the columns

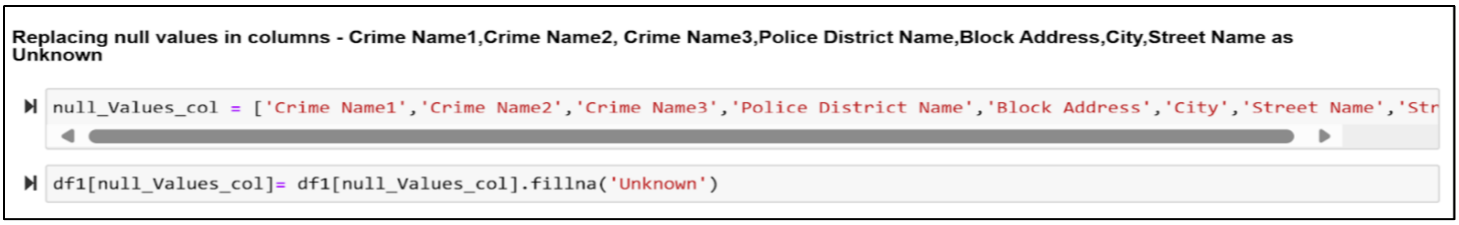
Dropping column Zip Code, Agency, End Date/time, Sector, Beat, PRA, Address Number,Location from the dataset because they do not contribute to the analysis objectives.This will help streamline our data and focus on the relevant features.

```
drop_Col = ['Zip Code','Agency','End_Date_Time','Street Prefix','Street Suffix','Sector','Beat','PRA','Address Number','Locat
df1 = df1.drop(columns = drop_Col)
```

Figure 9 : the code used to drop columns

2. **Replacing Null Values :** Null values were replaced with 'Unknown' for the following reasons(see figure 10):

   - **Data Integrity**: Ensures the dataset is complete and maintains its integrity.
   - **Consistency**: Using 'Unknown' as a placeholder maintains uniformity, simplifying handling and interpretation.



Replacing null values in columns - Crime Name1,Crime Name2, Crime Name3,Police District Name,Block Address,City,Street Name as Unknown

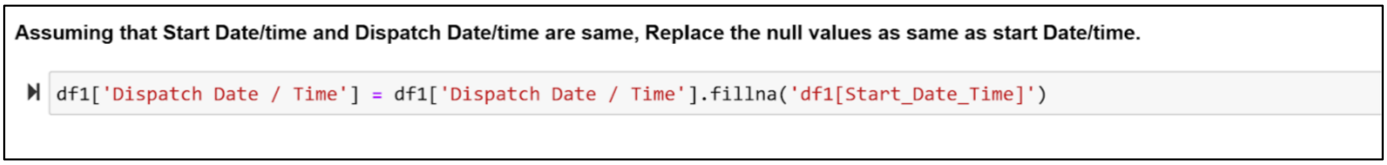
```
null_Values_col = ['Crime Name1','Crime Name2','Crime Name3','Police District Name','Block Address','City','Street Name','Str
```

```
df1[null_Values_col]= df1[null_Values_col].fillna('Unknown')
```

Figure 10 : the code used to replace null values

3. **Replacing a Column with Another Column**(see figure 11):

- The "Dispatch date/time" column was replaced with the "Start_Date_Time" column because their values were identical. It was assumed that the missing "Dispatch date/time" values would match the complete "Start_Date_Time" values.



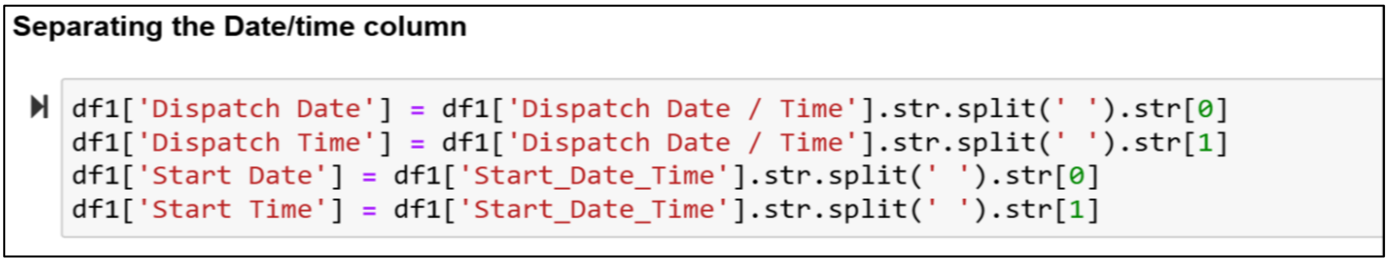
Assuming that Start Date/time and Dispatch Date/time are same, Replace the null values as same as start Date/time.

```
df1['Dispatch Date / Time'] = df1['Dispatch Date / Time'].fillna('df1[Start_Date_Time]')
```

Figure 11: the code used to replace a column with another

4. Date/time Column Transformation (see figure 12):

- **Separated Date/Time :** By separating date and time into separate columns granular analysis can be performed effectively

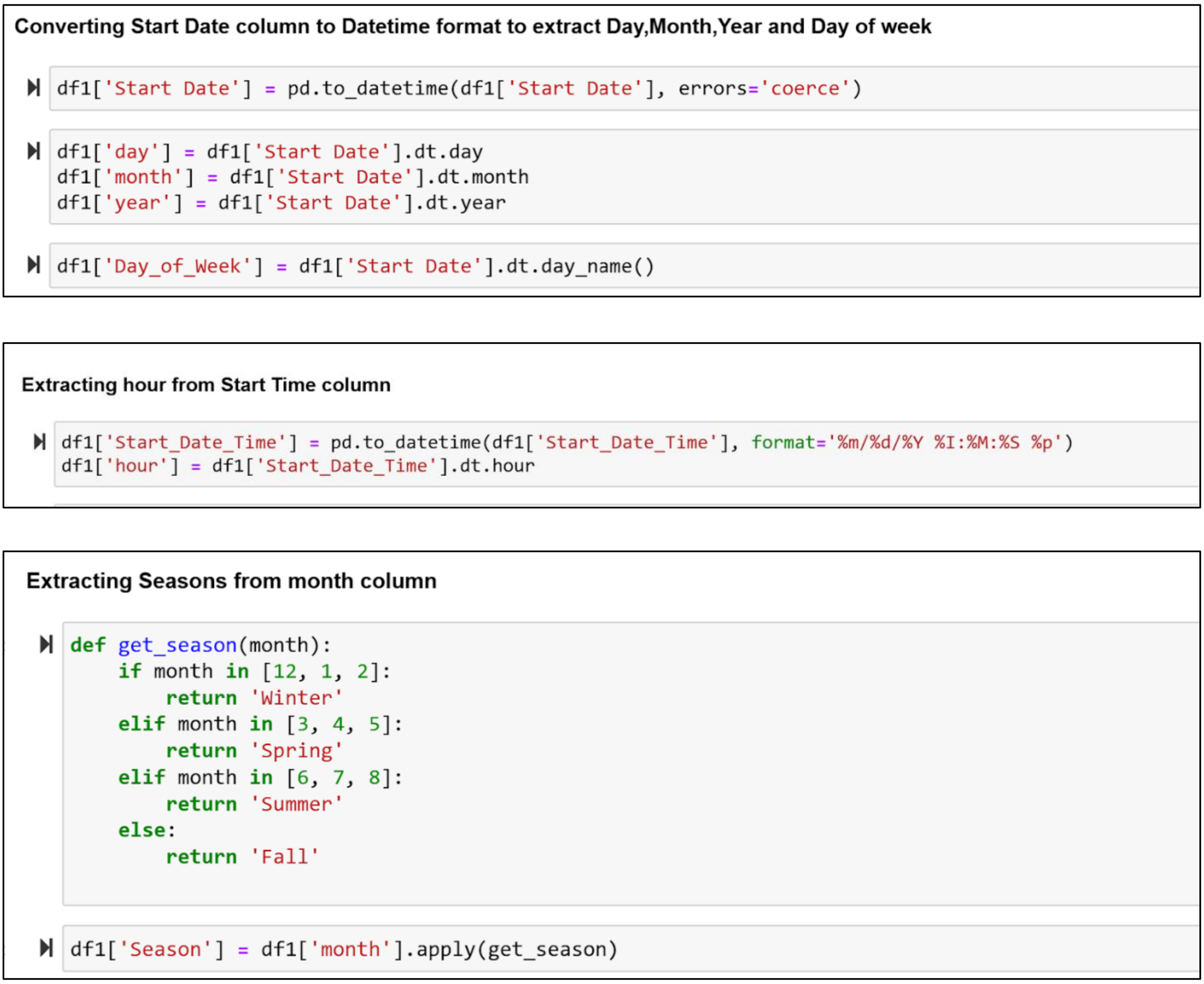


Separating the Date/time column

```
df1['Dispatch Date'] = df1['Dispatch Date / Time'].str.split(' ').str[0]
df1['Dispatch Time'] = df1['Dispatch Date / Time'].str.split(' ').str[1]
df1['Start Date'] = df1['Start_Date_Time'].str.split(' ').str[0]
df1['Start Time'] = df1['Start_Date_Time'].str.split(' ').str[1]
```

Figure 12 : the code used for transformation

- **Converting Start Date column to Datetime format** allows for precise analysis and easy extraction of components like season, day, month, year, hour, minute, and second (see figure 13).



Converting Start Date column to Datetime format to extract Day,Month,Year and Day of week

```
df1['Start Date'] = pd.to_datetime(df1['Start Date'], errors='coerce')
```

```
df1['day'] = df1['Start Date'].dt.day
df1['month'] = df1['Start Date'].dt.month
df1['year'] = df1['Start Date'].dt.year
```

```
df1['Day_of_Week'] = df1['Start Date'].dt.day_name()
```

Extracting hour from Start Time column

```
df1['Start_Date_Time'] = pd.to_datetime(df1['Start_Date_Time'], format='%m/%d/%Y %I:%M:%S %p')
df1['hour'] = df1['Start_Date_Time'].dt.hour
```

Extracting Seasons from month column

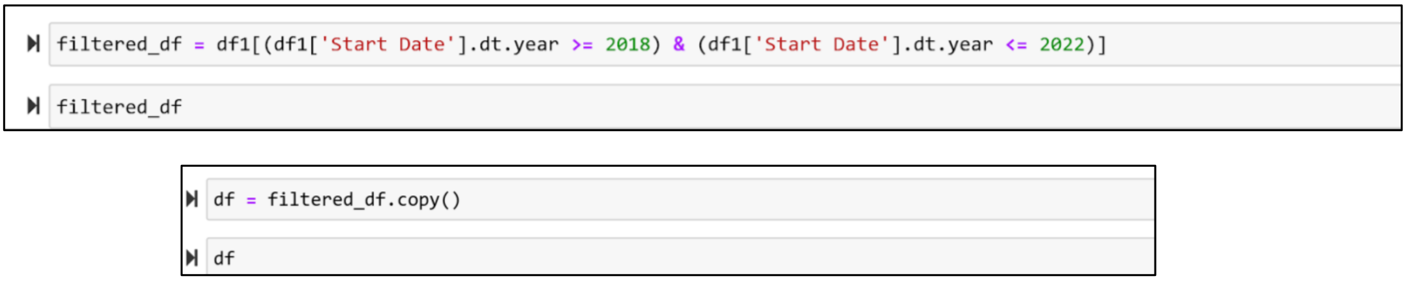
```
def get_season(month):
    if month in [12, 1, 2]:
        return 'Winter'
    elif month in [3, 4, 5]:
        return 'Spring'
    elif month in [6, 7, 8]:
        return 'Summer'
    else:
        return 'Fall'
```

```
df1['Season'] = df1['month'].apply(get_season)
```

Figure 13 : the codes used for further transformation

5. Filtered Data

- **Filtering the data to include only the years 2018-2022 :** This ensures the analysis focuses on relevant information, eliminating unnecessary data. Additionally, a dataset copy was made to preserve data integrity and prevent tampering (see figure 14).



```
filtered_df = df1[(df1['Start Date'].dt.year >= 2018) & (df1['Start Date'].dt.year <= 2022)]
```

```
filtered_df
```

```
df = filtered_df.copy()
```

```
df
```

Figure 14: the codes used for filtering and creating a copy

3 Exploratory Data Analysis

3.1 Introduction to EDA

Exploratory Data Analysis (EDA) is a crucial phase in any data analysis process, as it involves the study and visualisation of data to better understand its key characteristics, patterns, and relationships within the dataframe. Implementing EDA, insights are revealed that will update future analysis and modelling choices. In the context of MC dataframe, EDA will help us comprehend the distribution of crimes across various dimensions such as time, location, and type. Additionally, it will enable us to identify potential relationships between different variables and uncover any data quality issues that may need addressing.

Table 7 : EDA approaches and examples

EDA APPROACH	DESCRIPTION	EXAMPLES
Univariate Statistics	Analysis involving observations on a single variable at a time to understand each attribute's behaviour without considering relationships to other variables.	▪ Calculate mean and standard deviation of response times. ▪ -Determine the frequency of top 10 crimes.
Bivariate Statistics	Quantitative analysis of the empirical relationships between two variables simultaneously.	▪ Explore correlation between crimes and time of day. ▪ Analyse relationship between response times and crime types. ▪ -Compare locations and days of the week to uncover patterns like specific crime-time correlations.
Multivariate Statistics	Analysis involving three or more variables to understand complex interactions and relationships, often used to explore higher-dimensional correlations and their combined effects.	▪ Study relationships between seasonality, crime types, and response times. ▪ Analyse correlation between crime categories, locations, and response times ▪ Examine correlations of crimes across time of day and days of week.
Graphical Representation	Focuses on presenting the dataset through visualisations like graphs, tables, diagrams, and charts to make insights easy to understand and interpret.	▪ Use scatter plots, histograms, bar plots, line plots, heatmaps etc. to visualise data. ▪ Clear visualisations help stakeholders quickly grasp results.

Table 7 : Adapted from Setiawan & Suprihanto, 2021 and Mukhiya, 2020

3.2 Descriptive Statistics

Descriptive statistics focuses on describing and organising data to provide a clear understanding of its key characteristics. It summarises a sample without making probability-based predictions, highlighting central tendencies, variability, and overall distribution (see Tables 7, 8, and 9).

Table 8: Types of Descriptive Statistics

Measure	Description
Measures of Central Tendency	These measures represent the "average" or most representative values within the dataset.
Mean	The average value, determined by dividing the total of all values by the number of data points.
Median	The central value in a sorted dataset, offering a more accurate measure when the data is unevenly distributed.
Mode	The value that appears most frequently, particularly useful for categorical information.
Measures of Variability	These measures illustrate the extent of spread or variability in a dataset, showing how much the values differ from one another and from the mean.
Range	The difference between the largest and smallest values in the dataset.
Variance and Standard Deviation	Variance indicates the average of the squared deviations from the mean, while standard deviation reflects the average dispersion in the original measurement units.
Interquartile Range (IQR)	The range that encompasses the middle 50% of data points, minimizing the influence of outliers.
Measures of Distribution Shape	These offer insights into the structure of the data distribution and how the data points are arranged within a dataset.
Skewness	Refers to the degree of asymmetry in the data. Right-skew suggests a predominance of higher values, whereas left-skew indicates a prevalence of lower values.
Kurtosis	Assesses the "peakedness" or flatness of the data distribution. High kurtosis signifies heavier tails, while low kurtosis represents lighter tails.
Measures of Position	These metrics offer an understanding of how a particular value rank within a dataset.
Percentiles	Values that indicate the percentage of observations that fall below a certain point.
Quartiles	Metrics that split the dataset into four equal segments, providing additional insight into its distribution.
Frequency Distribution	Outlines the number of times each value occurs, typically represented in tables, histograms, or bar charts to visually illustrate the distribution of data.

Table 8 : A summary of the key types of descriptive statistics

Table 9: Descriptive Statistics for Crime Response Times used

Descriptive Statistic	Description	Purpose
Mean (mean)	Each category of the Crime Name2 column has its average response time calculated using the mean.	Understand the average response time for different types of crimes, providing insight into how long it typically takes to respond to various crime categories.
Standard Deviation (std)	The standard deviation of response times for each crime in Crime Name2 is calculated and stored in df5. A smaller standard deviation indicates consistent response times, whereas a larger value suggests variability	Provide information on the variability or consistency of response times for different crimes.
Using the Mean	The mean helps identify which crime categories have the longest and shortest average response times.	Guide operational improvements or resource allocation.
Assessing Response Time Consistency	Evaluating the standard deviation helps in identifying crime categories with irregular response times.	Provide insights into areas of investigation (e.g., why some crimes are responded to with a longer response time).
Improving Data Quality	Data quality is improved by removing outliers in response time by filtering out impractically long response times.	Enhance the quality of data analysis by focusing on realistic data.
Stakeholder Decision-Making	These metrics allow stakeholders to make informed, data-driven decisions for improving response efficiency.	Enable stakeholders to make decisions based on accurate and relevant data insights.

Table 9 : A summary of the descriptive statistics used within the visualisations

3.3 Data Visualisation

Data visualisation transforms raw data into visual formats like charts, graphs, and maps, making complex data more accessible and useful Assam & Evergreen, 2013). It reveals patterns, trends, and insights that might be missed in text-based data, enabling informed, data-driven decisions. Effective data visualisation requires clear goals, appropriate visual types, simplified design, functional colour schemes, and up-to-date data accuracy (Berinato, 2016).

The following section presents visualisations produced based on the research questions.

Q1. What are the 10 most common crime categories in the Montgomery County crime dataset?

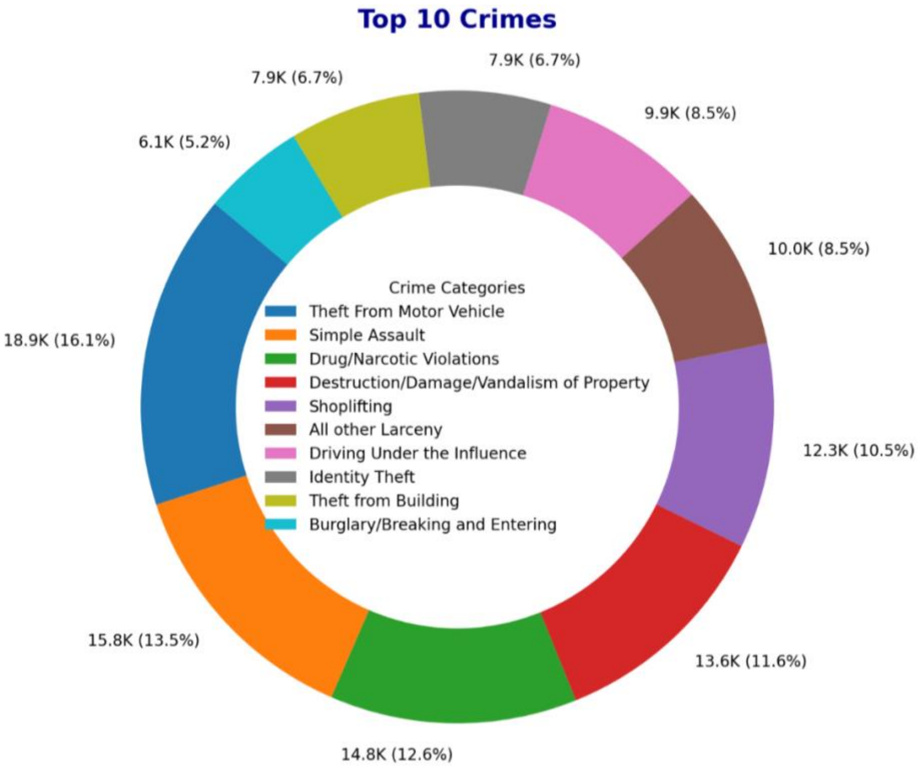


Figure 15 : A pie chart illustrating the ten most common crimes in the Montgomery crime dataset.

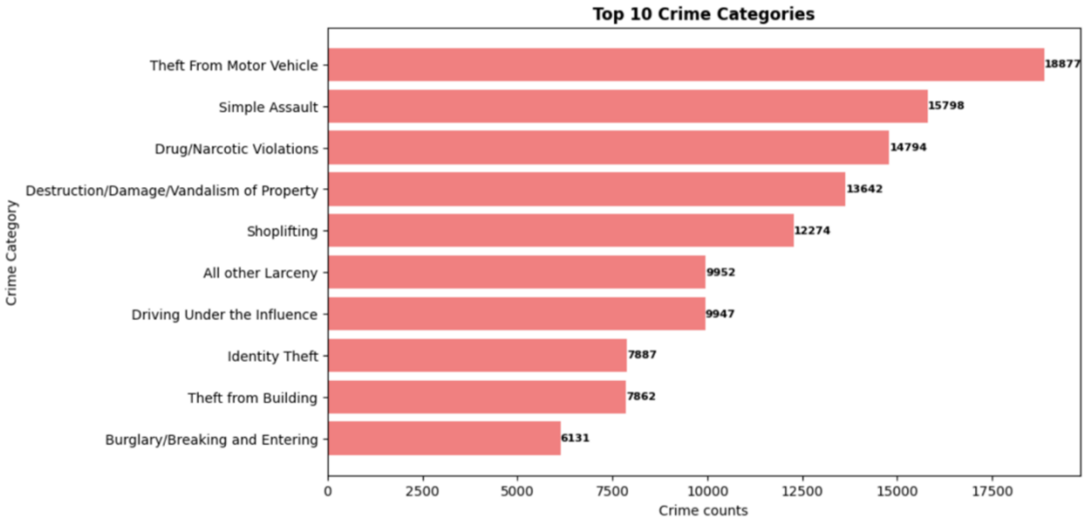


Figure 16 : A bar chart illustrating the ten most common crimes in the Montgomery crime dataset.

Figures 15 and 16 illustrate crime frequency, with "All Other Offenses" being the most prevalent at 54,154 incidents. "Theft From Motor Vehicle" (18,877) and "Simple Assault" (15,798) follow, while "Identity Theft" (7,887) and "Theft from Building" (7,862) are the least common, aiding resource allocation and prevention strategies.

Q2. During which times of the year are crimes most and least frequent in Montgomery County?

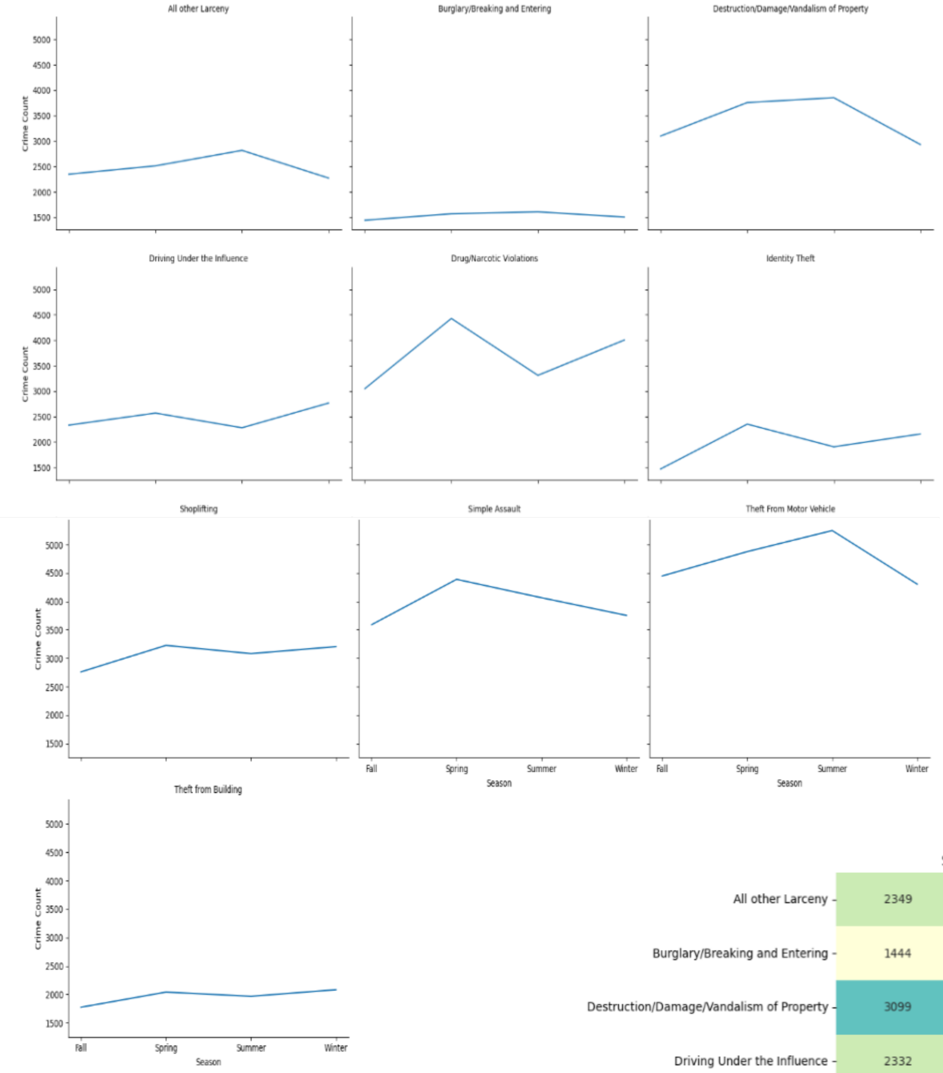


Figure 17 : A Facet-grid of seasonal variation in Crime Frequency in MC

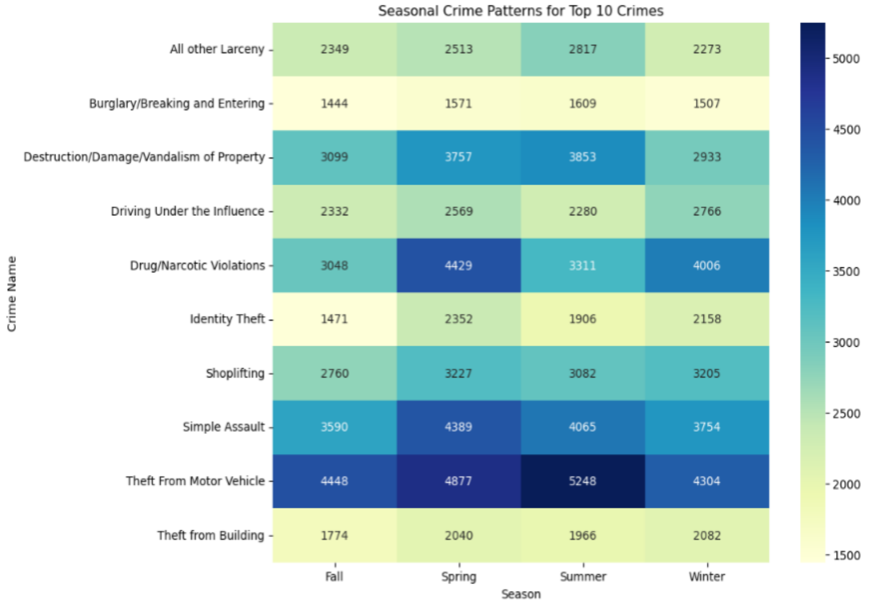


Figure 18 : A heatmap of seasonal variation in Crime Frequency in MC

The heatmap highlights periods needing increased vigilance for top ten crimes, with a high concentration of Theft from Motor Vehicles in summer guiding resource allocation and patrolling. The facet-grid displays the yearly distribution of these crimes, supporting informed decision-making(see Figure 17 and Figure 18).

Q3. How have crime rates evolved from 2018 to 2022?

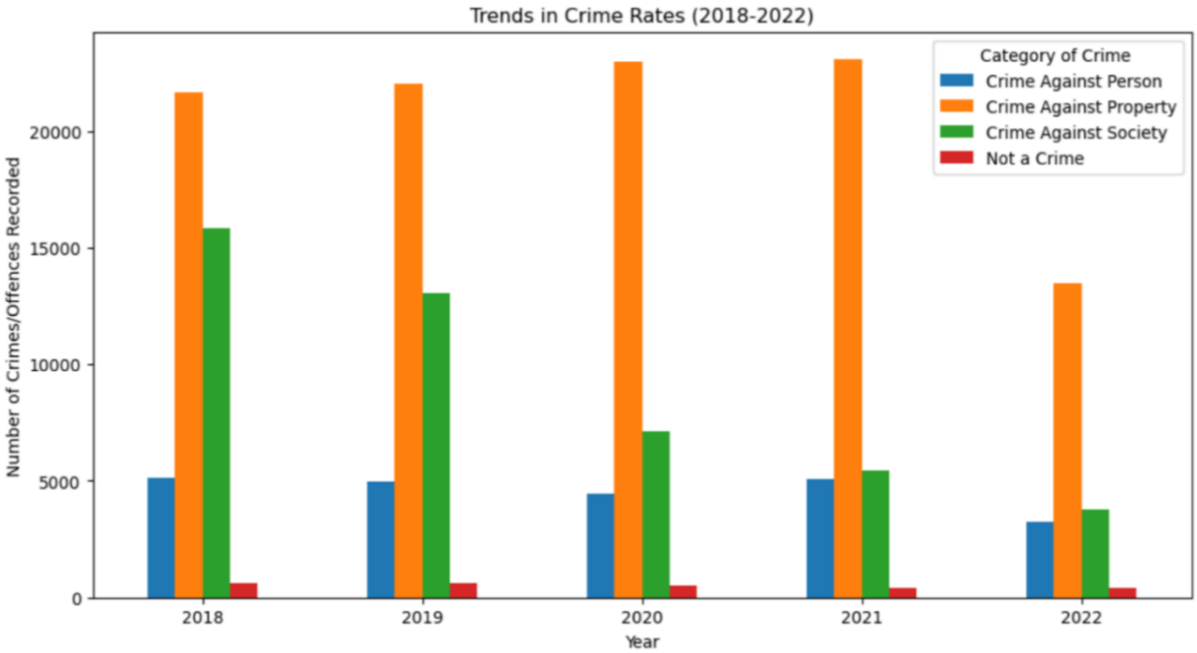


Figure 19 : A bar graph of the trends in crime rates in Montgomery Country (2018 – 2022)

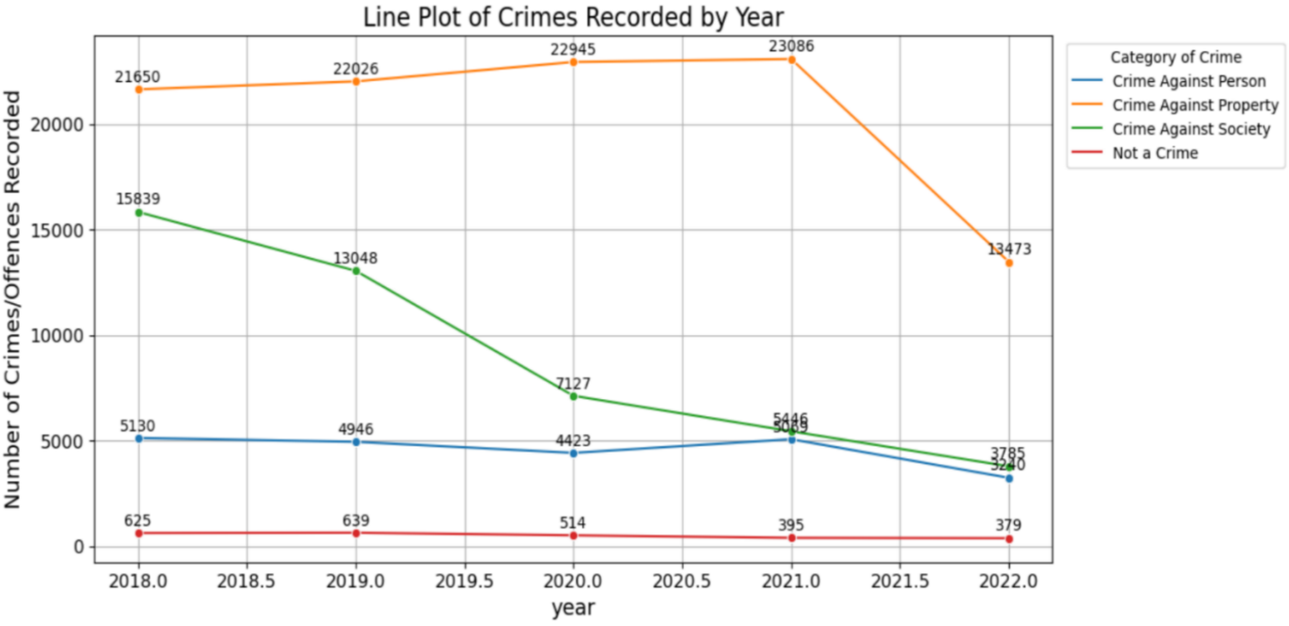


Figure 20 : A line plot of the trends in crime rates in Montgomery Country (2018 – 2022)

The visualisations indicate a general decline in most crime categories, particularly crimes against society and individuals, with a notable spike in property-related crimes in 2021, potentially due to specific events or reporting changes.(see Figure 19 and Figure 20).

Q4. Which ten cities in Montgomery County have the highest crime rates?

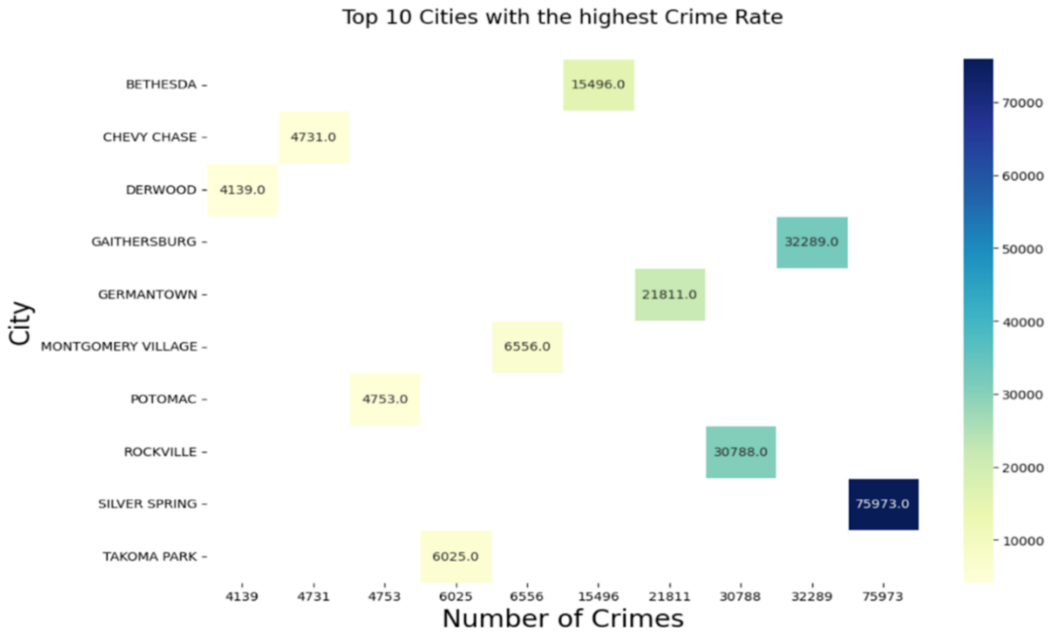


Figure 21 : A heat map of the Top 10 Cities with the Highest Crime Rates in MC

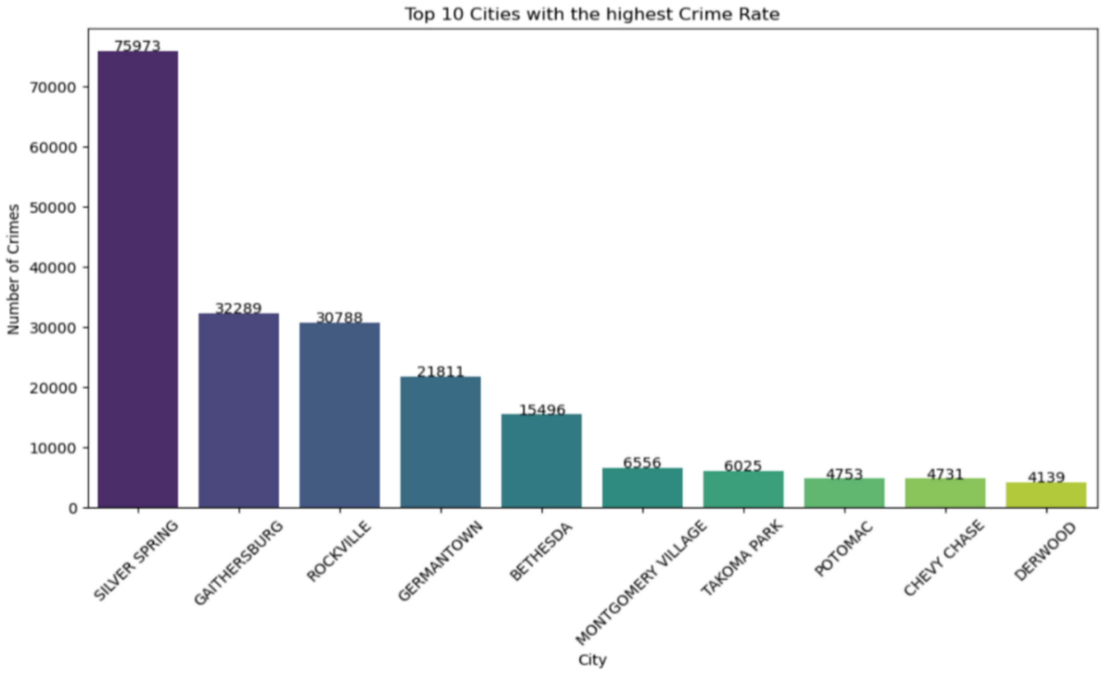
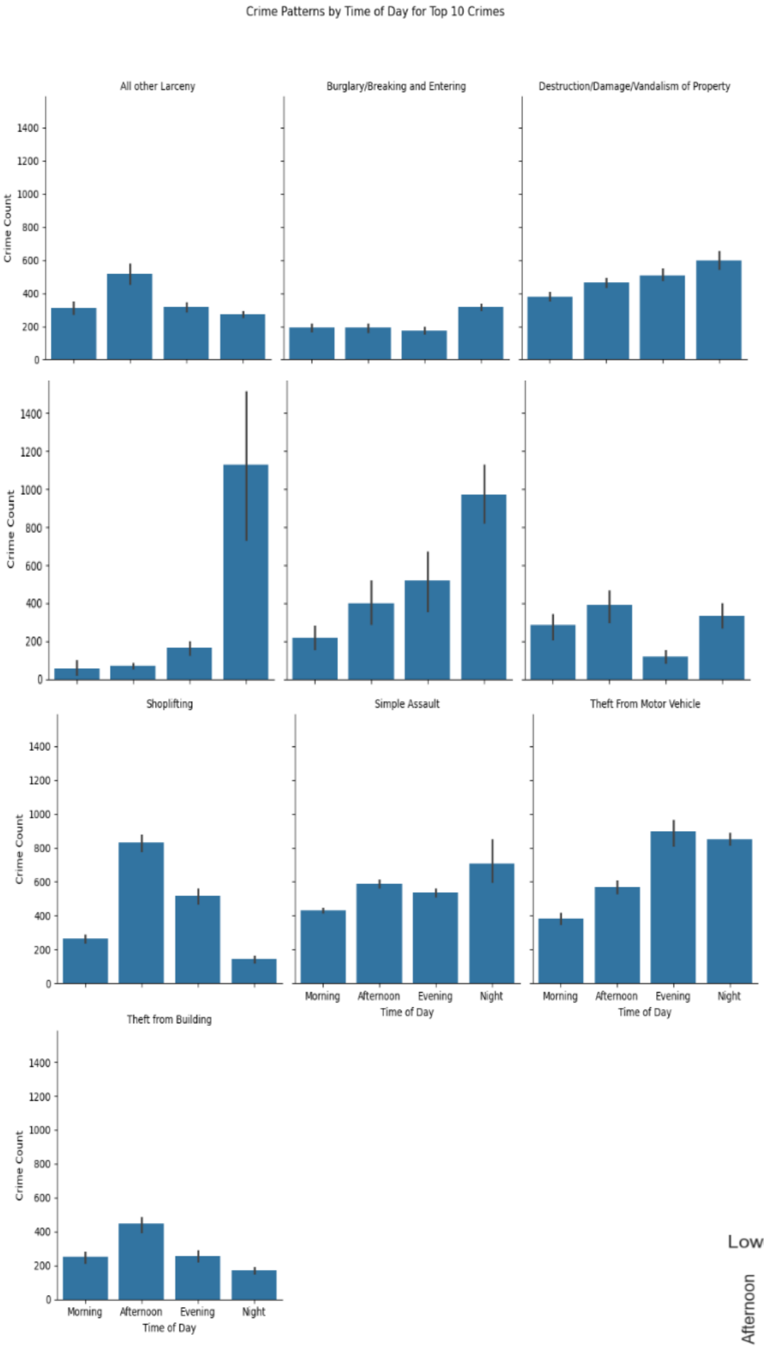


Figure 22 : A bar chart of the Top 10 Cities with the Highest Crime Rates in MC

Figures 21 and 22 highlight crime trends across cities, with Silver Spring having the highest reported crimes (75,973), indicating a need for increased safety measures. Gaithersburg (32,289) and Rockville (30,788) also face challenges, while Montgomery Village (6,556) and Takoma Park (6,025) report fewer crimes, suggesting greater safety.

Q5. Is there a direct correlation between the top 10 types of crimes and the time they occur?



The Correlation-Matrix reveals crime patterns across different times of the day, aiding in planning for recurring events or seasonal increases. The Facet Grid visualises the top 10 crimes' distribution throughout the day, helping authorities tailor their deployment and enhance security through collaboration with local entities.(see Figure 23 and Figure 24).

Figure 23 : A Facet- grid showing the correlation between the top 10 crimes and their occurrence time in the MC

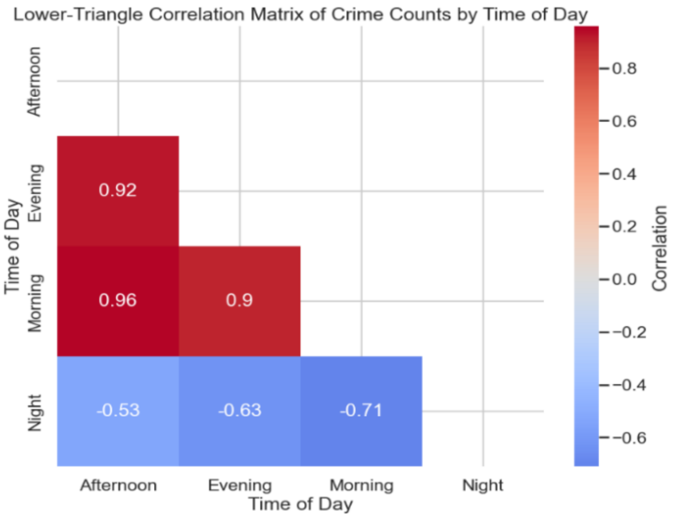


Figure 24 : A correlation matrix showing the correlation between the top 10 crimes and their occurrence time in the MC

Q6. Which agencies report the highest and lowest crime counts, and what insights can be drawn from these variations?

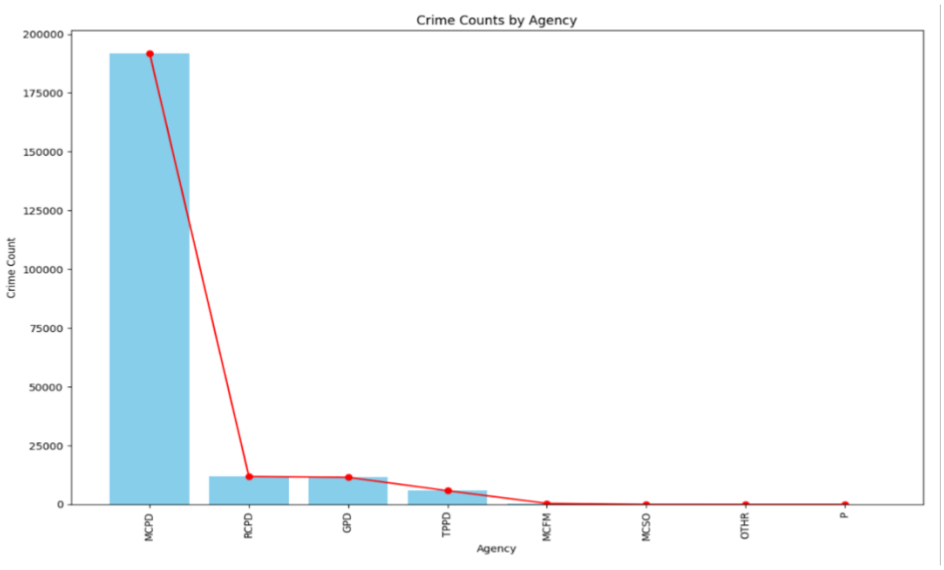


Figure 25 : A bar-line chart depicting the agencies with the highest and lowest reported crime counts.

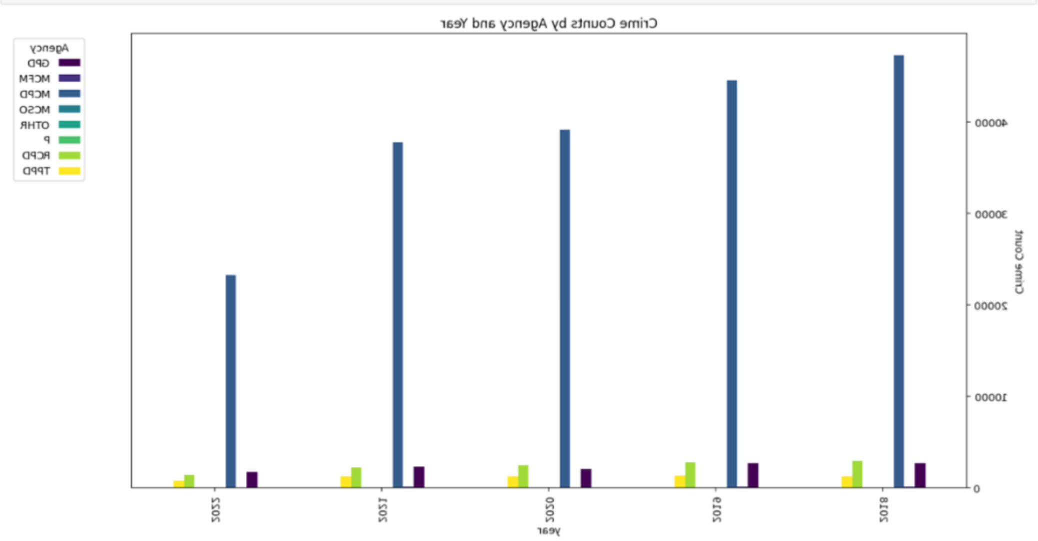


Figure 26: Grouped bar chart depicting the agencies with the highest and lowest reported crime counts.

Figures 25 and 26 show MCPD reports the highest number of crimes (191,962), likely due to its large or high-crime area coverage. RCPD (11,853) and GPD (11,490) have moderate counts, while MCFM and MCSO report fewer crimes, focusing on specific issues or rural areas. OTHR and P report very few incidents (4 and 1), indicating they serve small areas or handle rare crimes.

Q7. How do crime patterns change depending on the day of the week?

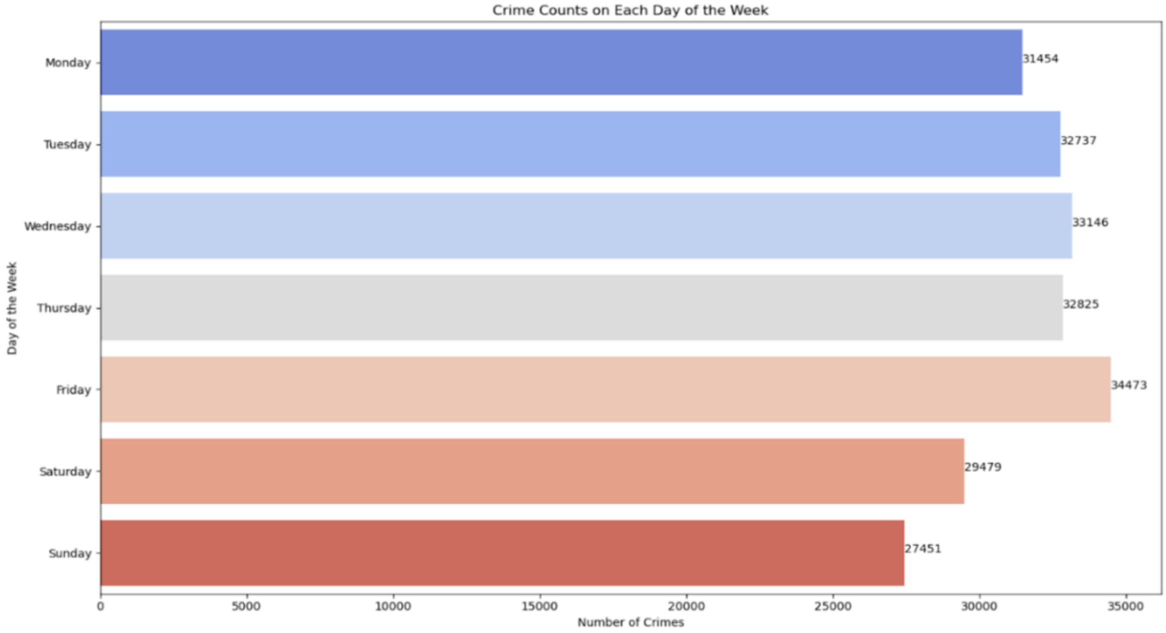


Figure 27 : A bar chart that shows how crimes change depending on the day of the week

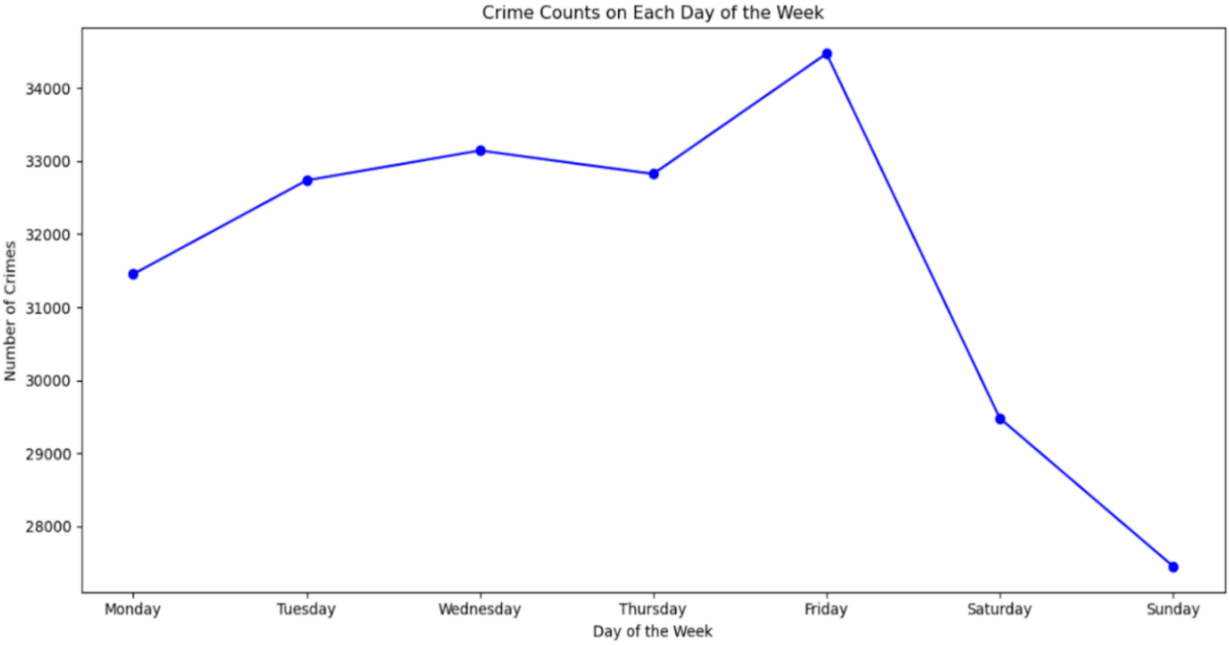


Figure 28 : A line plot that shows how crimes change depending on the day of the week

The visualisations highlight weekly crime trends, with a notable increase on Fridays, likely due to heightened social activities, alcohol consumption, and late-night outings, creating opportunities for various offenses(see figure 27 and Figure 28).

Q8. What are the top 10 locations where crimes frequently occur?

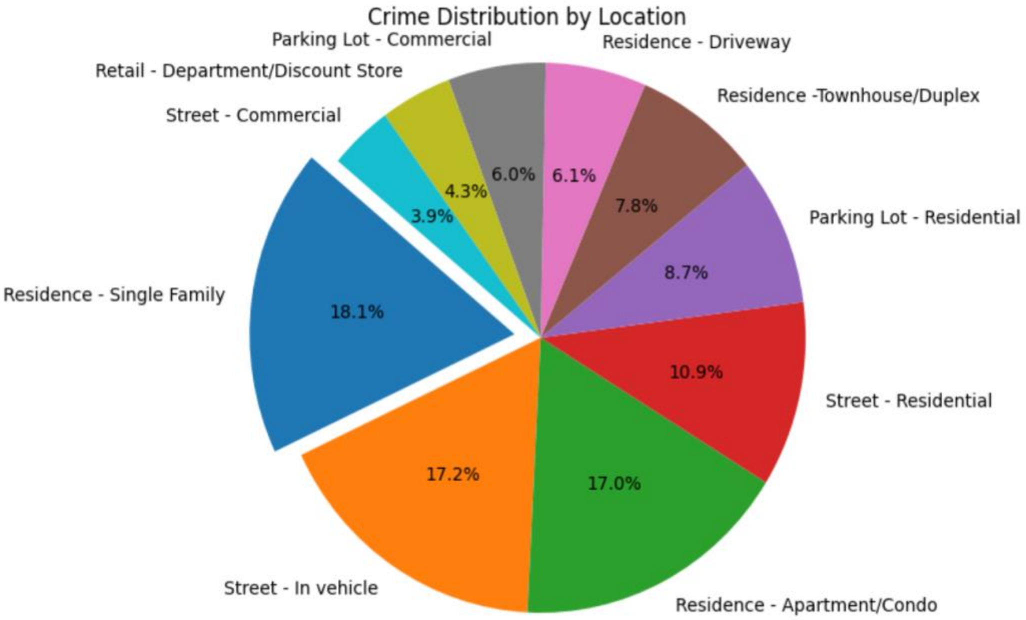


Figure 29 : A pie chart that shows where crimes mostly occur

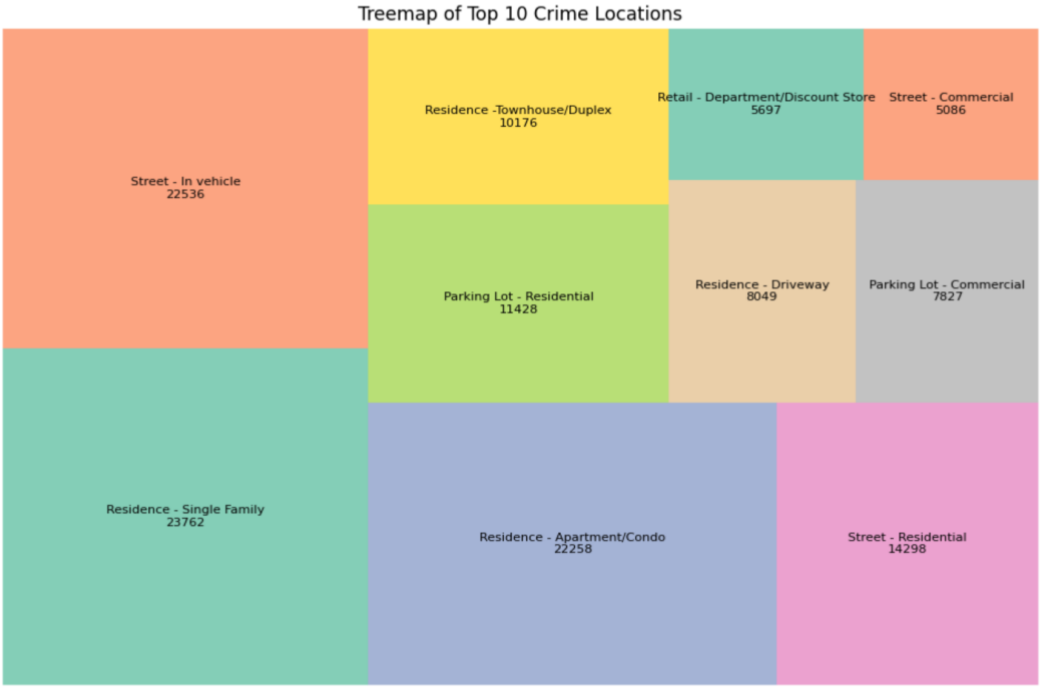


Figure 20 : A treemap illustrating the predominant locations of criminal activity.

Figures 29 and 30 show most crimes occur in single-family homes, followed by vehicles on the street and apartments or condos. Residential streets and parking lots also see significant incidents, while commercial areas have fewer cases but still require safety measures.

Q9. What is the relationship between crimes and response times?

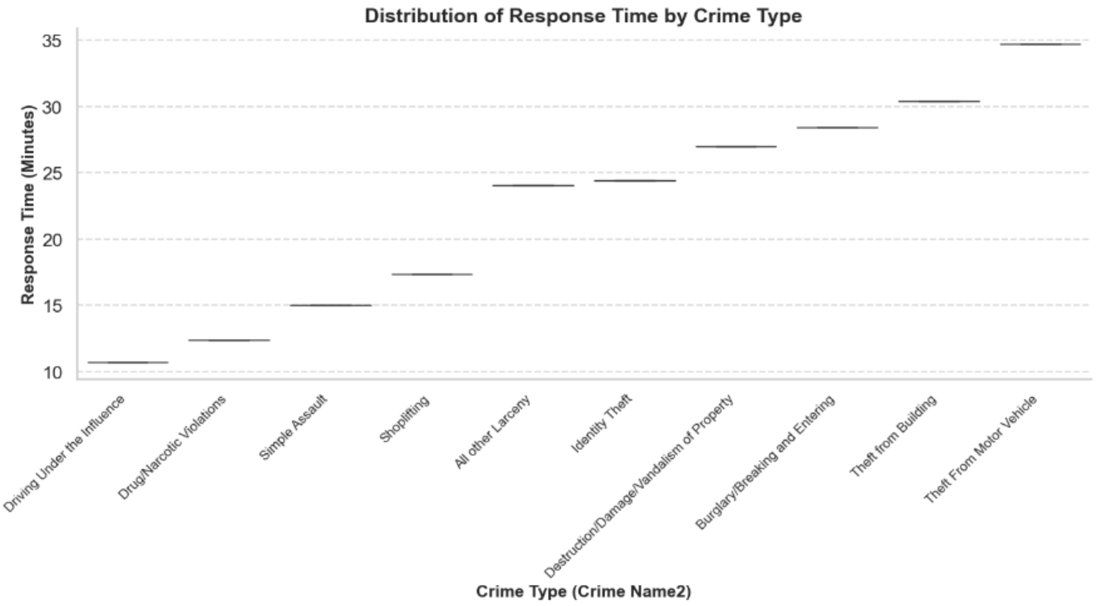


Figure 31 : Crime Type vs. Response Timeline plot with Error Chart

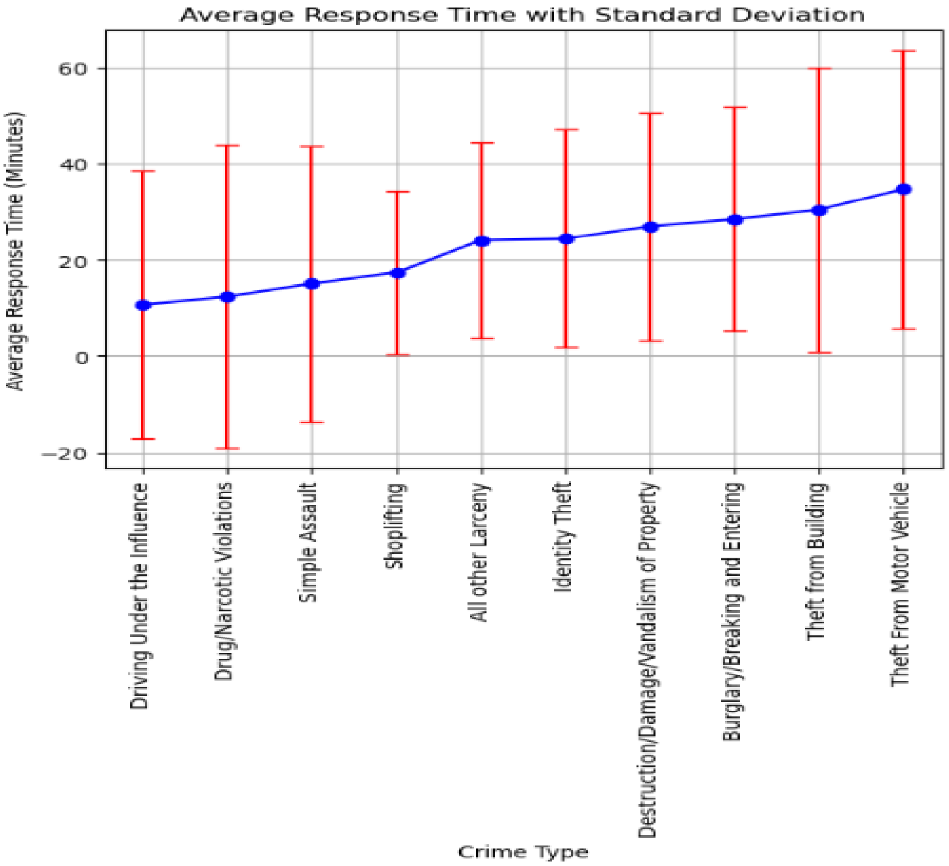


Figure 32 : Crime Type vs. Response Time with box plot

The line plot with error bars highlights average and variability in response times for each crime type, suggesting high averages may need strategy adjustments. The box plot shows median, quartiles, and outliers, indicating areas needing faster responses and identifying crimes requiring quicker responses (see Figure 31 and Figure 32).

Q10. What is the top 10 annual number of victims impacted by specific types of crimes?

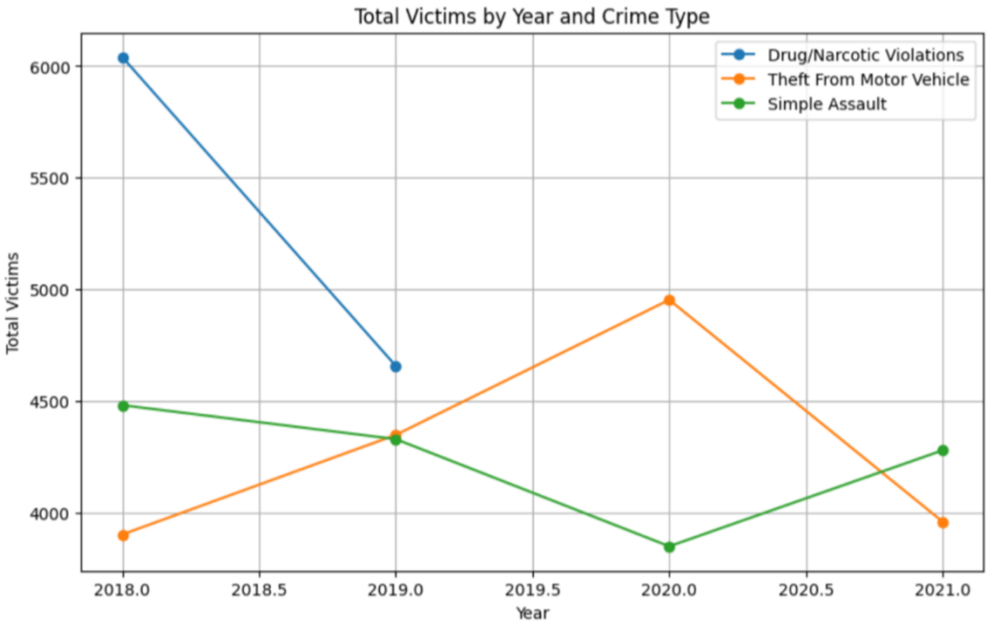


Figure 33 : A line graph which shows the Top 10 Crime Victims Annually

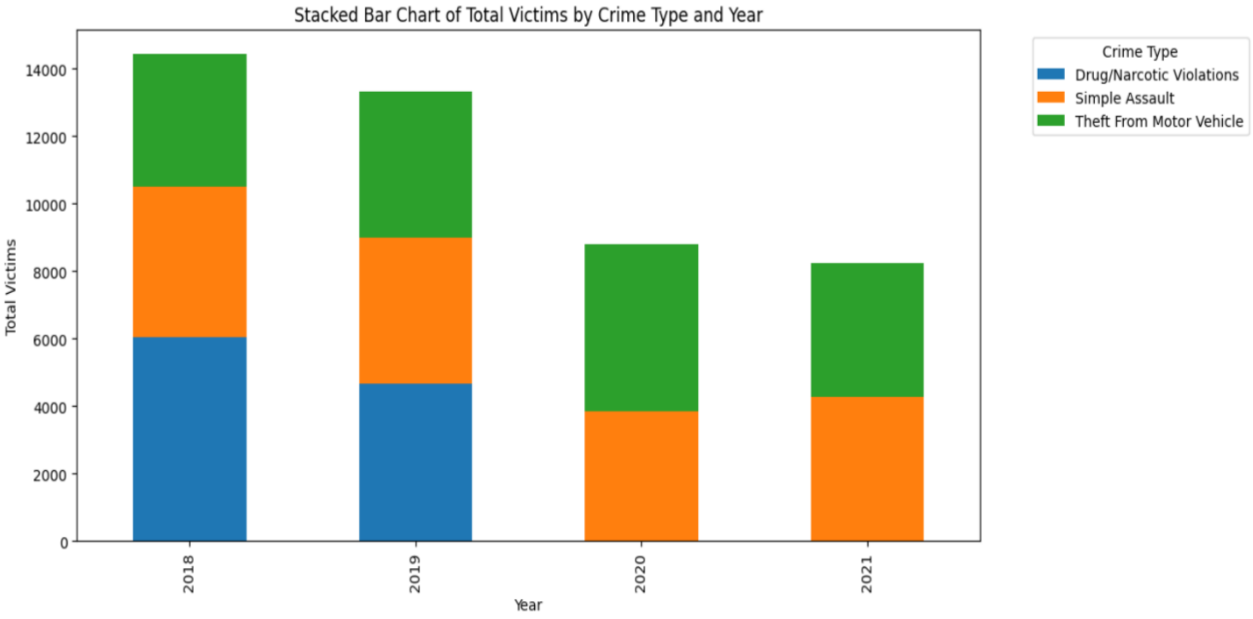


Figure 34: A bar chart which shows the Top 10 Crime Victims Annually

Figures 34 and 35 compare total victims across crime types and years, highlighting a significantly higher victim count for Drug/Narcotic Violations in 2018. The charts provide insights into yearly crime dynamics, with overall bar height and line graph segments indicating each crime type's contribution.

4. Conclusion

The analysis of Montgomery County's crime data from 2018 to 2022 reveals several key findings:

- **Most Common Crime**: "All Other Offenses" is the most prevalent.
- **Seasonal Spikes**: There is a significant increase in "Theft from Motor Vehicles" during summer.
- **High Crime Area**: Silver Spring reports the highest crime rates, requiring enhanced safety measures.
- **Weekly and Location Patterns**: Crimes are most frequent on Fridays and predominantly occur in single-family homes and vehicles on the street.

ACTION PLAN: To address these findings, it is recommended that the local government increase patrols and preventive measures during the summer months (June – August) and on Fridays, particularly in high-crime areas like Silver Spring.
Additionally, focusing on crime hotspots such as single-family homes and street-parked vehicles can help reduce incidents. Implementing community outreach programs and enhancing neighbourhood watch initiatives will further contribute to a safer community.

3. References

Assam, T. and Evergreen, S.D.H. (2013) *Data Visualization and Evaluation*. San Francisco, CA: Wiley.

Berinato, S. (2016) ‘Visualizations That Really Work’.

Data.montgomerycountymd.gov, data. montgomerycountymd. gov (2024) ‘Montgomery County of Maryland’.

Delgado, Y. *et al.* (2021) ‘Forensic intelligence: Data analytics as the bridge between Forensic Science and Investigation’, *Forensic Science International: Synergy*, 3, p. 100162. doi:10.1016/j.fsisyn.2021.100162.

Haslett, T. *et al.* (2012) *Framework for Development and evaluation of community engagement*.

Hvistendahl, M. (2016) ‘Can ‘predictive policing’ prevent crime before it happens?’

IACANET (2024) *About Crime Analysis, International Association of Crime Analysts*. Available at: https://www.iaca.net/ (Accessed: 29 October 2024).

Kelleher, J.D. and Tierney, B. (2018) *Data science*. Cambridge, MA, London, England: The MIT Press.

Kennedy, J.D. and Hilling, H. (2023) *Solving cold cases: Investigation techniques and Protocol*. Jefferson, NC, North Carolina: Exposit.

Mayernik, M.S. (2023) *Data Science as an interdiscipline: Historical parallels from information science, Data Science Journal*. Available at: https://datascience.codata.org/articles/10.5334/dsj-2023-016 (Accessed: 02 November 2024).

McDonalds, A. (2021) *Using the missingno python library to identify and visualise missing data prior to machine learning, Using the missingno Python library to Identify and Visualise Missing Data Prior to Machine Learning*. Available at: https://www.andymcdonald.scot/using-the-missingno-python-library-to-identify-and-visualise-missing-data-prior-to-machine-learning-34c8c5b5f009 (Accessed: 31 October 2024).

Montgomery County, MD, M.C., MD (2015) ‘Data Montgomery’.

Mukhiya, S.K. (2020) *Hands-on exploratory data analysis with python: Perform EDA techniques to understand, summarize, and investigate your data, Google Books*. Available at: https://books.google.com/books/about/Hands_On_Exploratory_Data_Analysis_with.html?id=GSR2zQEACAAJ (Accessed: 03 November 2024).

Office for national statistics, office for national statistics (2024) *Crime in England and Wales QMI, Crime in England and Wales QMI - Office for National Statistics*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/methodologies/crimeinenglandandwalesqmi (Accessed: 01 November 2024).

The Policy and Planning Division (2024) *MONTGOMERY COUNTY DEPARTMENT OF POLICE 2023 ANNUAL REPORT ON BIAS INCIDENTS*, 1 March.

Python Software Foundation, P.S.F. (2024) ‘csv — CSV File Reading and Writing’.

Setiawan, I. and Suprihanto, S. (2021) ‘Exploratory Data Analysis of Crime Report’, *Exploratory data analysis of crime report*, 11(2), pp. 71–80. doi:10.31940/matrix.v11i2.2449.

Setiawan, I. and Suprihanto, S. (2021) ‘Exploratory Data Analysis of Crime Report’, *Matrix : Jurnal Manajemen Teknologi dan Informatika*, 11(2), pp. 71–80. doi:10.31940/matrix.v11i2.2449.

Tableau, T. (2024) *Guide to data cleaning: Definition, benefits, components, and how to clean your data, Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data*. Available at: https://www.tableau.com/learn/articles/what-is-data-cleaning (Accessed: 01 November 2024).

Wyckoff, L. (2014) *Definition and Types of Crime Analysis* .