IE6400 Foundation of Data

Analytics Project- 1

Topic: Cleaning and Analyzing Crime Data

Group 7

Name	NUID
Saisree Pothu	002294485
Sreevarshan Sathiyamurthy	002311854
Rithika Sankar Rajeswari	002311800

Introduction

Data-driven decision-making fundamentally depends on the quality and interpretability of available data. This report offers an in-depth exploration of the processing and analysis of a real-world dataset: Crime Data from 2020 to the present. In our dynamic and complex society, crime manifests in various forms. Gaining insights into crime trends, patterns, and underlying factors is not only of academic interest but also vital for law enforcement and public policy development

Background

Crime data, with its complexities and subtleties, presents distinct challenges and opportunities for data analysts. Through the processes of cleaning, analyzing, and interpreting this data, we can extract insights that may significantly influence and improve public safety strategies. The dataset being analyzed is publicly available and offers a detailed record of reported crimes, including information on incident times, locations, and victim demographics.

Task 1: Data Acquisition

The first step of our analysis involved acquiring the dataset. The crime data, covering the period from 2020 to the present, was downloaded from a Data.gov repository. After obtaining the dataset, it was imported into our data analysis environment, Google Colaboratory. For this project, we chose Python, a versatile and powerful language well-suited for data manipulation and analysis. The dataset was in CSV (Comma-Separated Values) format, which is highly compatible with Python. Using the pandas library, we successfully loaded the data for further analysis.

TASK 2 : Data Inspection

After examining the dataset, which contains 811,663 entries and 28 columns, we found a diverse mix of data types, with both numerical and categorical data well-represented. The `.info()` method showed that several columns had complete data, while others, like 'Mocodes' and 'Vict Descent,' had significant missing values that would require careful handling. Additionally, date-related columns were misclassified as objects, indicating the need to convert them to datetime format for accurate analysis. This initial inspection is essential, as it sets the foundation for the subsequent data cleaning process and helps ensure the integrity of our analysis.

TASK 3 : Data Cleaning

Handling Missing Values:

In our data cleaning process, we tackled a substantial amount of missing values, converting a dataset with significant information gaps into a complete one, ready for analysis. Before our intervention, columns like 'Mocodes,' 'Vict Sex,' 'Vict Descent,' and 'Weapon Desc' had considerable data omissions that could have affected our results. For example, 'Mocodes' had 112,024 missing entries, while both 'Weapon Used Cd' and 'Weapon Desc' had an overwhelming 528,880 missing values each.

Addressing Null Values:

Our approach to null values was categorical and context-specific. For 'Mocodes', 'Premis Cd', and 'Weapon Used Cd', we replaced missing values with 0 to indicate the absence of such details, a method often used when the data is numerical or categorical with an implicit 'none' category.

For textual categorical data such as 'Vict Sex', 'Vict Descent', 'Premis Desc', and 'Weapon Desc', we filled the voids with descriptive placeholders like 'NA' or 'Not Available'. This approach preserves the data structure while clearly documenting the absence of certain information, allowing for unimpeded analysis that acknowledges these gaps

Removing Redundant Columns:

We removed the columns 'Crm Cd 1,' 'Crm Cd 2,' 'Crm Cd 3,' and 'Crm Cd 4' from our analysis, as these likely provided detailed breakdowns or subclassifications of crimes that were not necessary for our high-level analysis. This simplification helps reduce complexity, allowing us to concentrate on the most impactful variables.

Dealing with Sparse Columns:

For 'Cross Street', a field with numerous missing values, we opted to fill the gaps with 'Not Available'. It's a strategic choice that mitigates the loss of entire rows of otherwise valuable data and aids in maintaining a comprehensive dataset for more generalized insights.

Through these targeted cleaning actions, we aimed to preserve data integrity and prepare the dataset for nuanced analysis, ensuring that our interpretations are grounded in the most complete and accurate information possible.

```
df.isnull().sum()
```

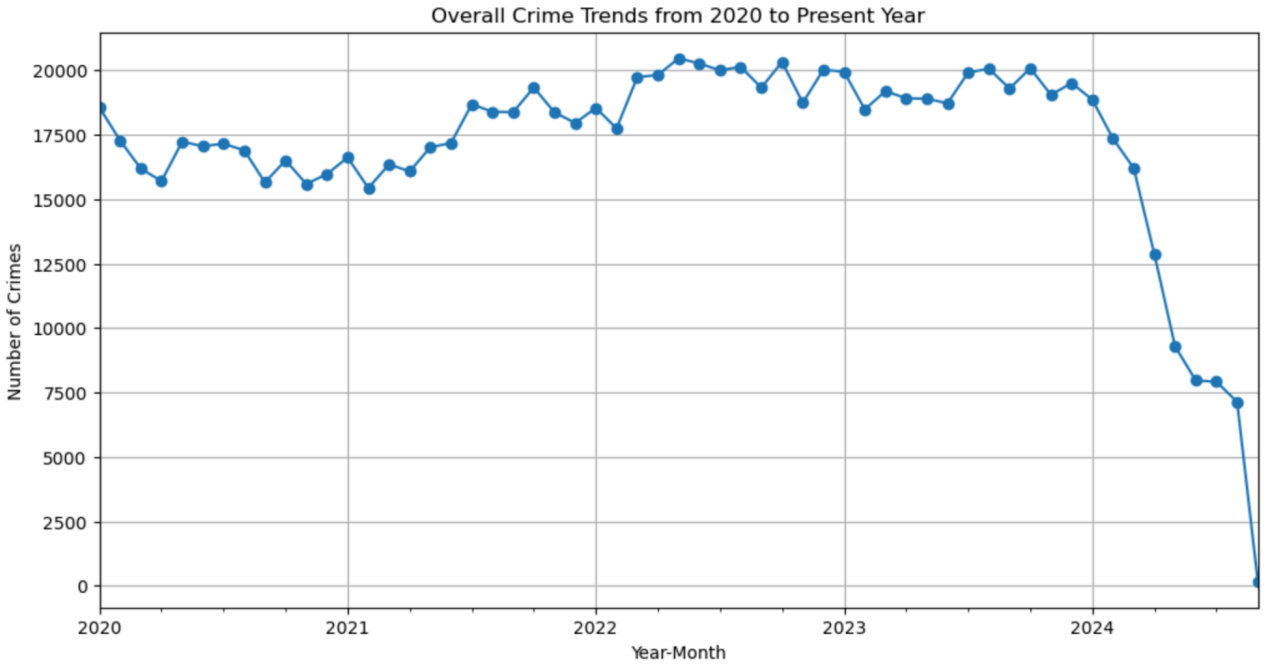
```
DR_NO                 0
Date Rptd             0
DATE OCC              0
TIME OCC              0
AREA                  0
AREA NAME             0
Rpt Dist No           0
Part 1-2              0
Crm Cd                0
Crm Cd Desc           0
Mocodes          142776
Vict Age              0
Vict Sex         136003
Vict Descent     136013
Premis Cd            14
Premis Desc         584
Weapon Used Cd   648695
Weapon Desc      648695
Status                1
Status Desc           0
Crm Cd 1             11
Crm Cd 2         905717
Crm Cd 3         972168
Crm Cd 4         974413
LOCATION              0
Cross Street     823461
LAT                   0
LON                   0
dtype: int64
```

```
df.isnull().sum()
```

```
DR_NO            0
Date Rptd        0
DATE OCC         0
TIME OCC         0
AREA             0
AREA NAME        0
Rpt Dist No      0
Part 1-2         0
Crm Cd           0
Crm Cd Desc      0
Mocodes          0
Vict Age         0
Vict Sex         0
Vict Descent     0
Premis Cd        0
Premis Desc      0
Weapon Used Cd   0
Weapon Desc      0
Status           1
Status Desc      0
LOCATION         0
Cross Street     0
LAT              0
LON              0
dtype: int64
```

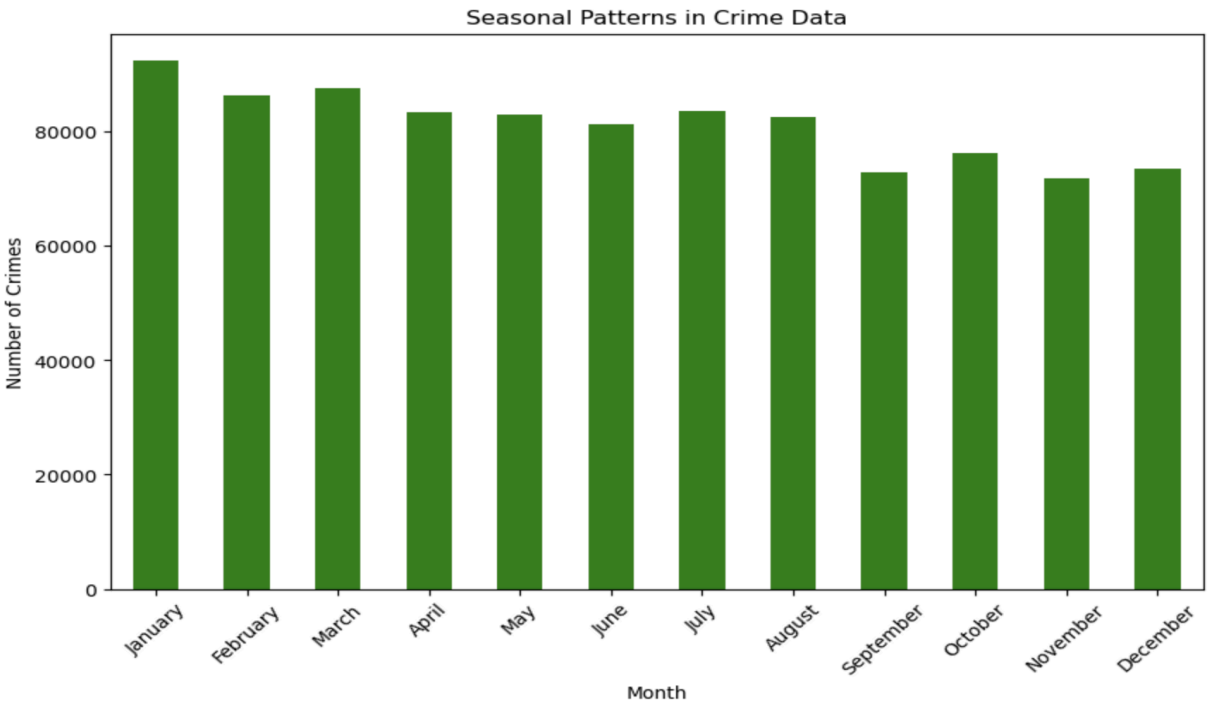
TASK 4 : Exploratory Data Analysis (EDA)

1. Visualize overall crime trends from 2020 to the present year



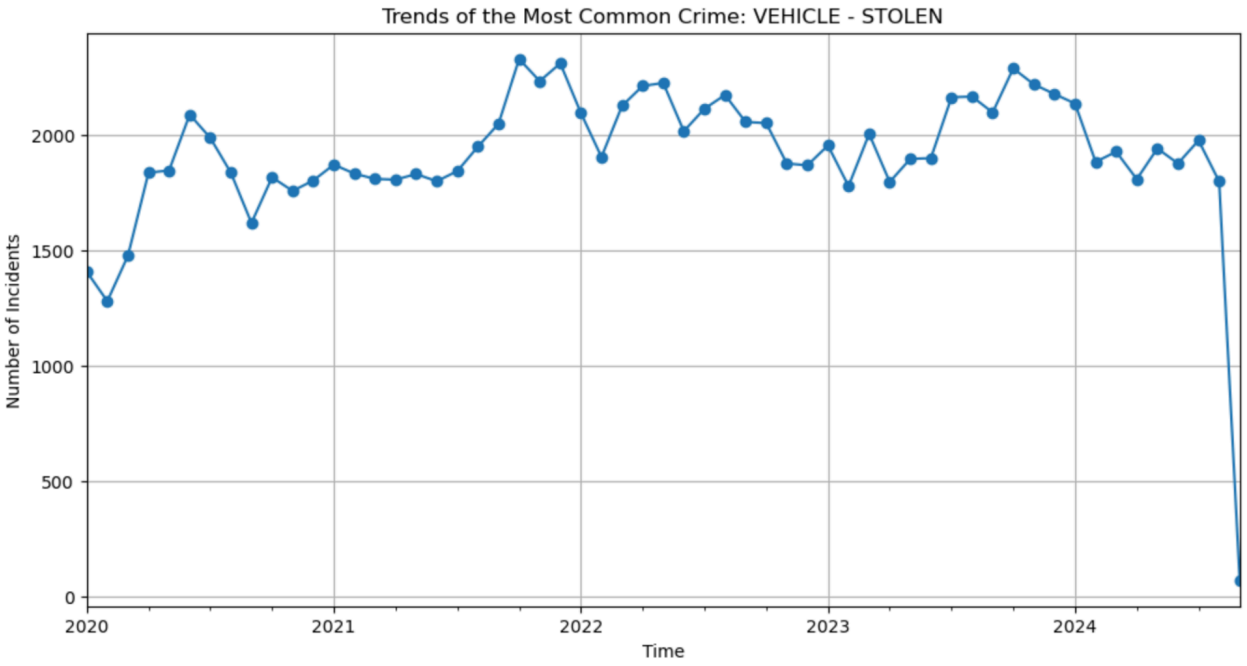
The graph delineates the fluctuation in the number of crimes from January 2020 through end-2024. For the majority of this period, crime rates have remained relatively steady, oscillating between approximately 15,000 to 20000 incidents per month. These figures reflect a consistent pattern with mild peaks and troughs, hinting at potential seasonal variations. However, a significant deviation is observed around end-2024, where there's a precipitous drop in the crime rate, suggesting a potential major event or impactful policy change that might have led to this drastic reduction in crime incidents.

2. Analyze and visualize seasonal patterns in crime data



The bar graph illustrates the monthly distribution of crimes throughout the year, highlighting potential seasonal patterns in crime data. Notably, January records the highest number of crimes, with subsequent months maintaining a relatively high but slightly decreasing level through the summer months. Crime rates remain fairly steady from February to August, with only minor variations. However, starting in September, there is a noticeable decline in crime incidents, which continues through to December. This trend suggests that crime rates are generally higher at the beginning of the year and tend to decrease as the year progresses, possibly indicating seasonal influences or external factors affecting criminal activity across different months.

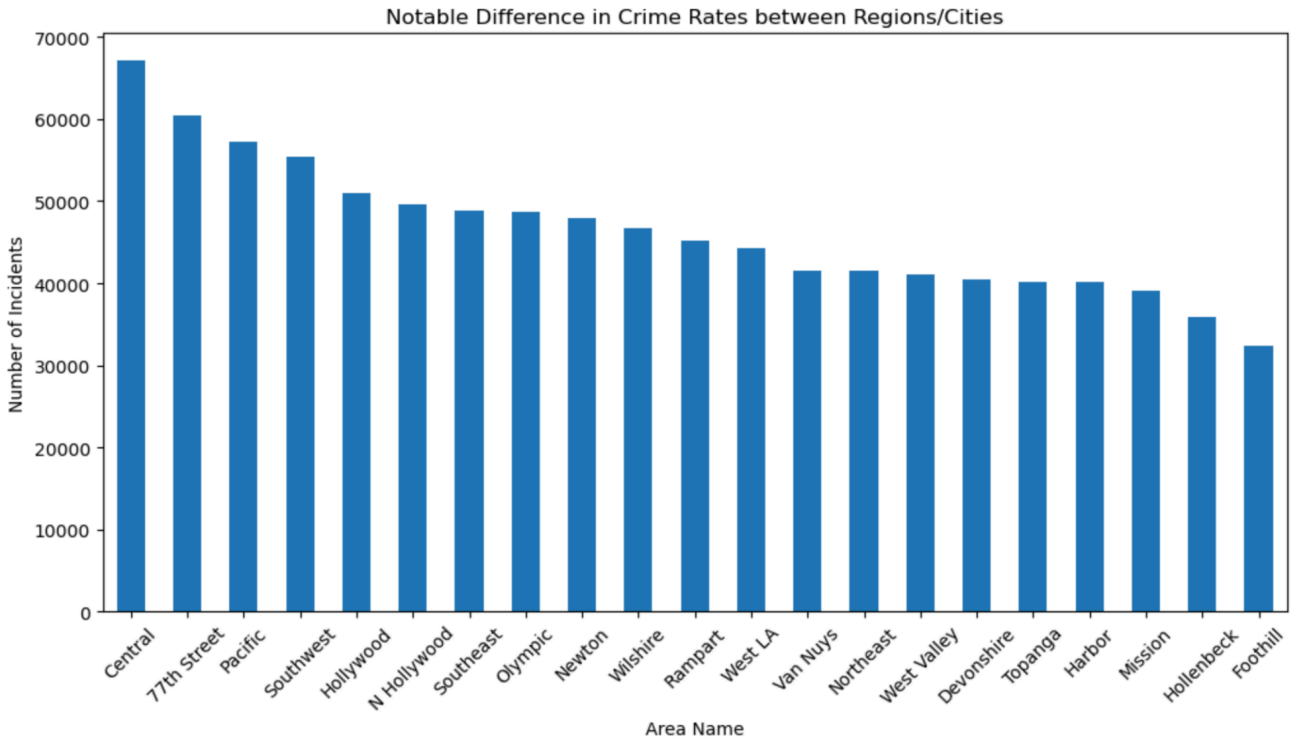
3. Identify the most common type of crime and its trends over time.



The line graph depicts the monthly trends for the most common crime type, "Vehicle - Stolen," from 2020 to 2024. Initially, there is an increase in incidents from early 2020, reaching peaks around 2,000 incidents per month by mid-2021. This elevated level persists with minor fluctuations through to early 2023, indicating a generally high prevalence of vehicle thefts over the period.

Toward the end of 2023, a sharp decline in vehicle theft incidents is observed, with the trend continuing into 2024. By mid-2024, the number of incidents approaches zero, suggesting a dramatic reduction in vehicle thefts. This significant drop could be indicative of a policy change, enhanced vehicle security measures, or shifts in reporting practices. The pattern reflects a notable decrease in this crime type, marking a departure from previous years' consistently high rates.

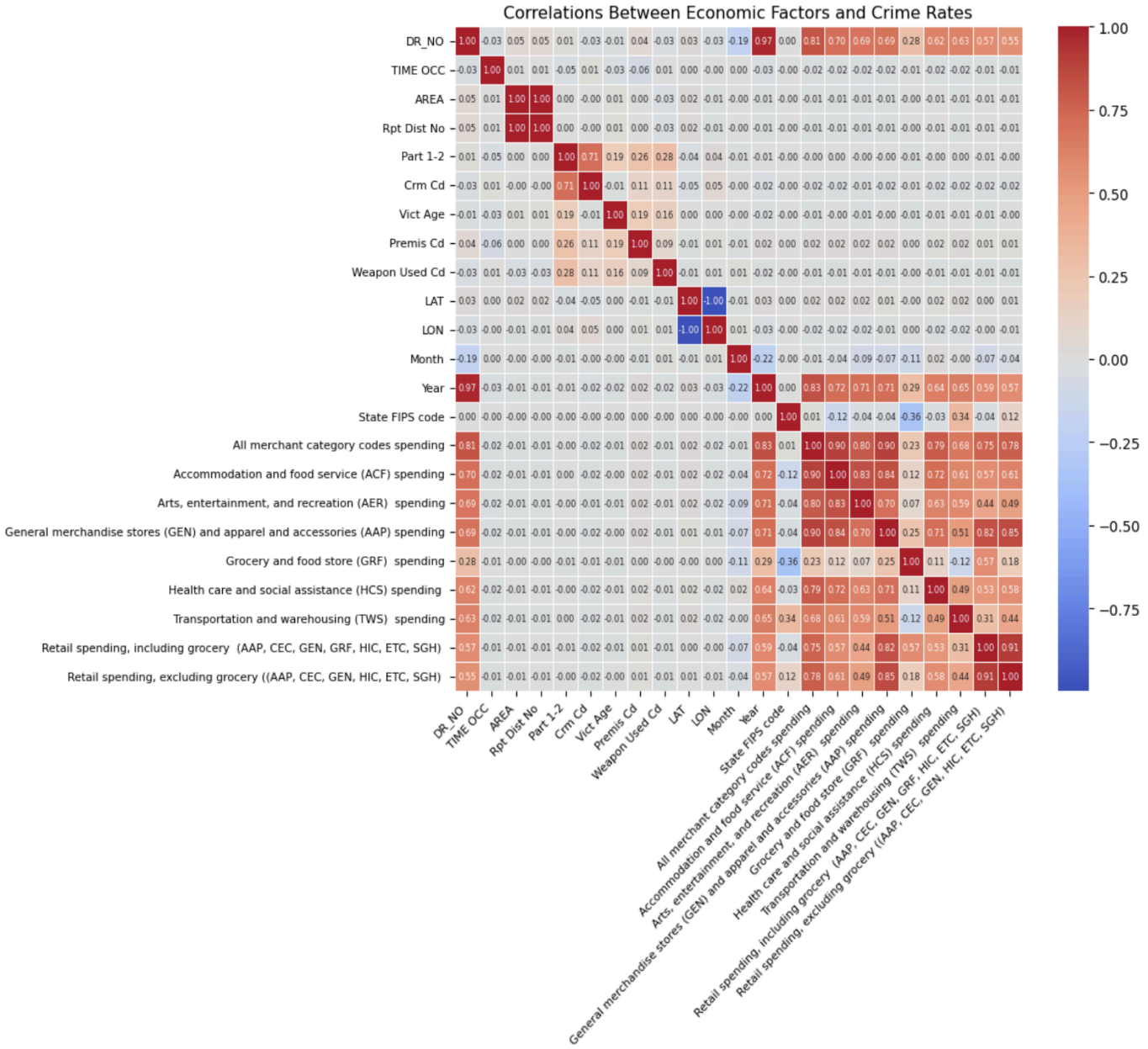
4. Investigate if there are any notable differences in crime rates between regions or cities.



The bar chart illustrates the distribution of crime incidents across different regions or cities, highlighting notable differences in crime rates. The "Central" area records the highest number of incidents, with close to 70,000 reported cases, followed by the "77th Street" and "Pacific" areas, each with substantial crime figures exceeding 50,000 incidents.As we move down the chart, areas such as "Southwest," "Hollywood," "N Hollywood," and "Southeast" also report relatively high crime rates, indicating these regions as significant hotspots. In contrast, regions like "Mission," "Hollenbeck," and "Foothill" show comparatively lower crime rates, each with fewer than 30,000 incidents.

The data reveals a clear disparity in crime distribution across different regions, suggesting that certain areas experience higher crime concentrations. This could be due to various factors, such as population density, socioeconomic conditions, or the effectiveness of local law enforcement in these regions.

5. Explore correlations between economic factors (if available) and crime rates



Key Observations:

a. High Correlation Among Economic Categories:

Economic factors such as “All merchant category codes spending,” “General merchandise stores spending,” and “Arts, entertainment, and recreation spending” exhibit strong positive correlations with each other (over 0.6), suggesting these spending categories tend to increase or decrease together.

b. Economic Factors and Crime Counts:

The correlation between economic factors and the primary crime attributes (e.g., `Crime Count`) is relatively low, indicating a weak direct relationship between overall crime rates and these economic indicators.

However, specific categories like "Accommodation and food service spending" show moderate positive correlations with crime attributes like `DR_NO` and `Year`, which might reflect that certain types of spending are linked to fluctuations in crime reporting or incidents.

c. Geographical Correlations:

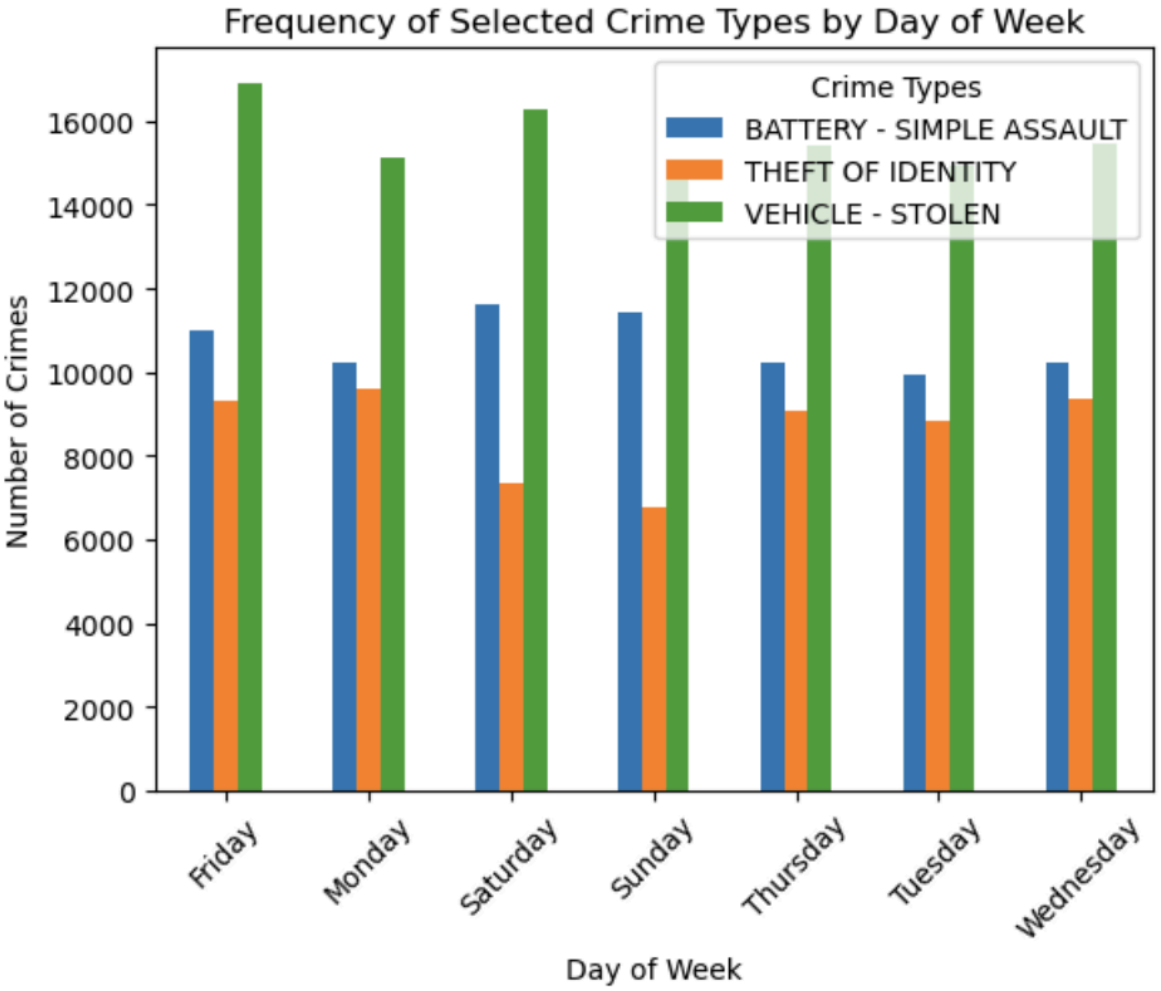
The variables `LAT` (latitude) and `LON` (longitude) have low correlations with economic factors, indicating minimal geographic influence in these economic attributes on crime distributions within the dataset.

d. Time-Related Correlations:

The `Year` attribute shows a positive correlation with several economic spending categories, suggesting that certain spending types increased over the period covered in the dataset.

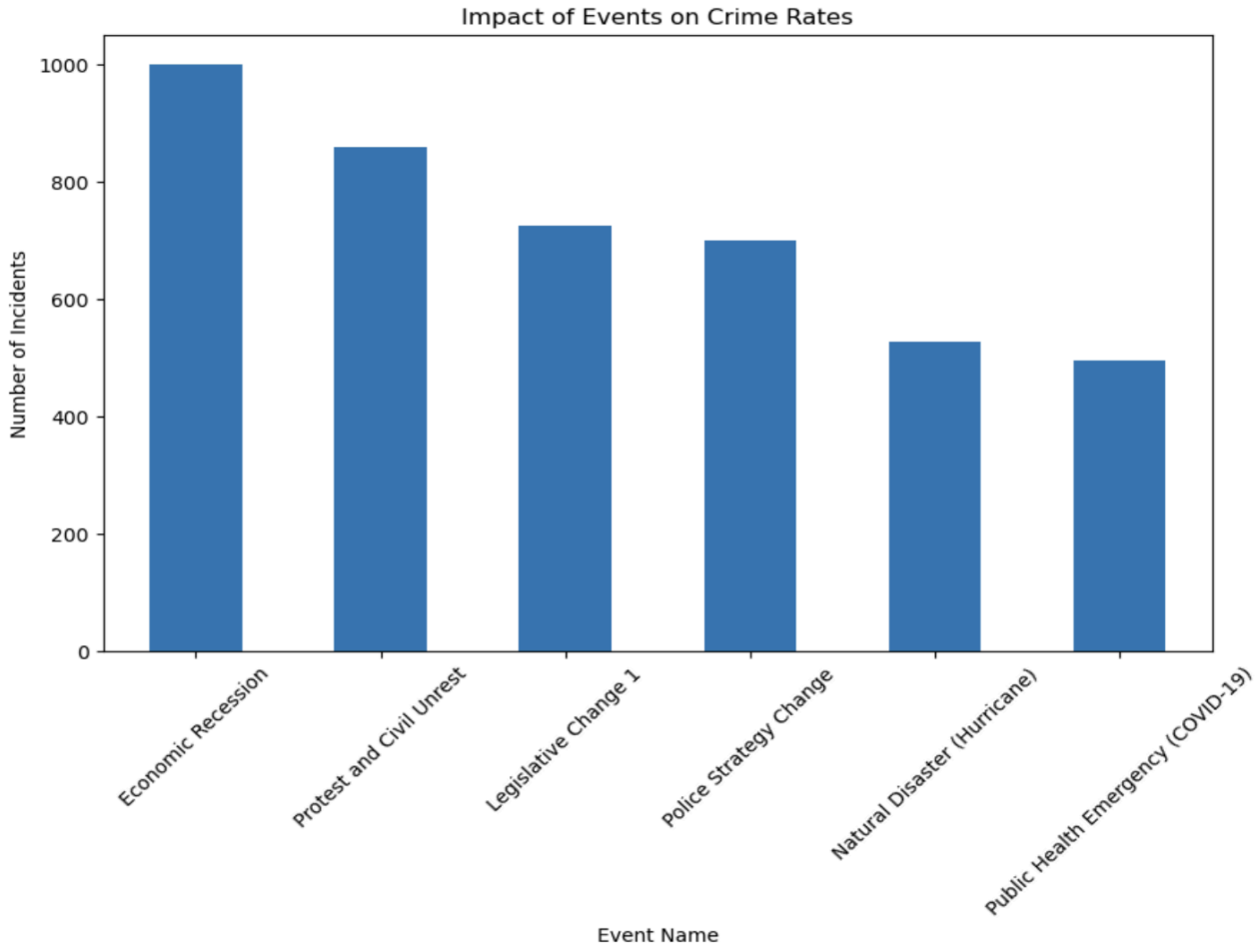
Overall, while there is a strong interconnection among economic factors, their direct correlation with crime rates remains low, suggesting that crime trends may not be heavily influenced by fluctuations in consumer spending alone. Other factors may play a more significant role in influencing crime rates.

6. Analyze the relationship between the day of the week and the frequency of certain types of crimes.



The bar chart shows the frequency of three selected crime types—"Battery - Simple Assault," "Theft of Identity," and "Vehicle - Stolen"—across different days of the week. "Vehicle - Stolen" incidents are consistently the most frequent, peaking on Saturday and Sunday with more than 16,000 incidents, indicating a higher propensity for vehicle theft on weekends."Battery - Simple Assault" and "Theft of Identity" display a relatively even distribution throughout the week, with a slight increase on Fridays and Saturdays. Overall, the chart suggests that weekends experience higher crime rates for these selected types, particularly vehicle theft, while weekdays show more balanced levels of assault and identity theft. This pattern could reflect increased outdoor activities and vehicle use during weekends, contributing to the rise in vehicle theft incidents.

7. Investigate any impact of major events or policy changes on crime rates.



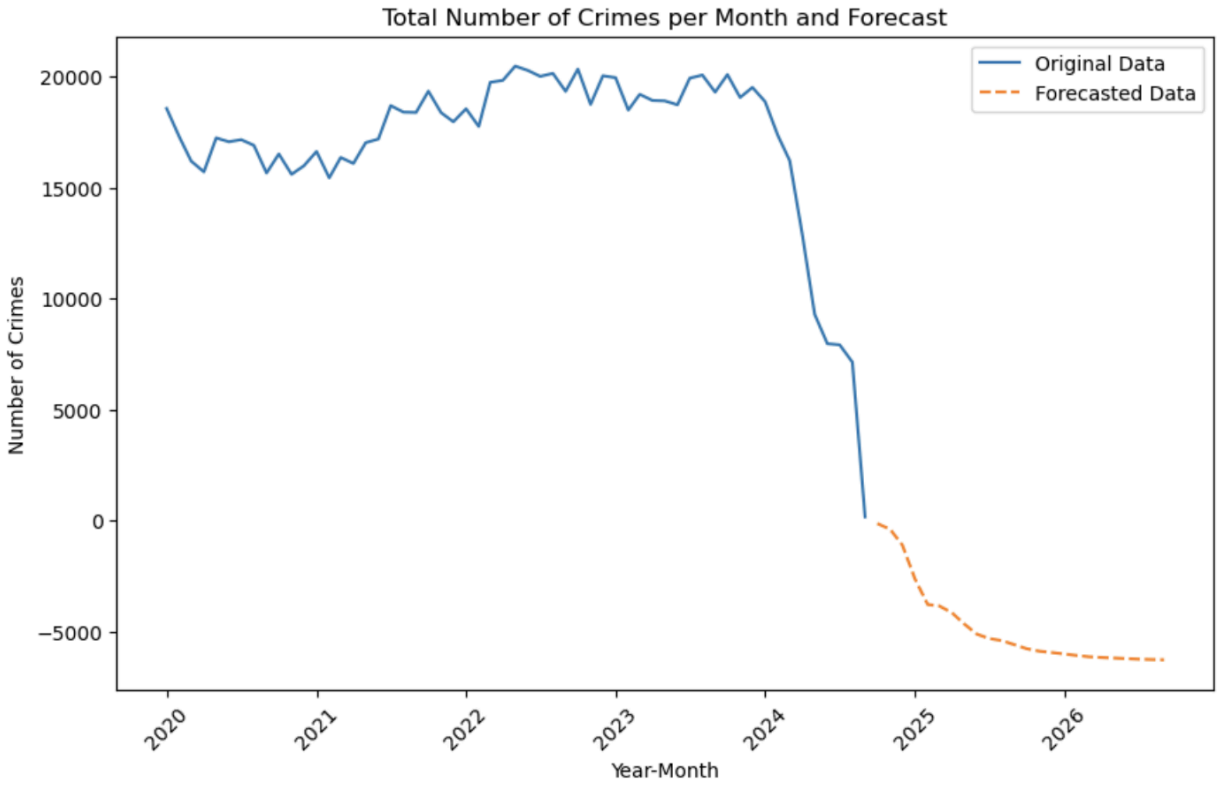
The bar chart illustrates the impact of various significant events on crime rates, showing the number of incidents associated with each event. The "Economic Recession" had the highest correlation with crime rates, recording nearly 1,000 incidents, which suggests that financial instability may contribute to increased criminal activity. "Protests and Civil Unrest" also correlated with a high number of incidents, indicating a potential link between social instability and elevated crime rates.

Other events, such as "Legislative Change 1" and "Police Strategy Change," show a moderate impact on crime, with each correlating with over 600 incidents. These might reflect the effects of policy shifts on criminal behavior or law enforcement activity. Conversely, "Natural Disaster (Hurricane)" and "Public Health Emergency (COVID-19)" have a somewhat lower correlation with crime rates, indicating that these events may have had a less direct impact

on crime or potentially led to reductions in certain types of criminal activity. Overall, the chart highlights that economic and social events appear to have a more substantial effect on crime rates compared to natural disasters and public health emergencie**s.**

TASK 5 : ADVANCED ANALYSIS :

We utilized the ARIMA (AutoRegressive Integrated Moving Average) model, a popular time series forecasting technique. By analyzing the historical data of crimes reported from 2020 to 2024 we aimed to project the possible trends for the subsequent years. ARIMA is particularly suited for datasets that exhibit patterns or trends over time, making it a fitting choice for our analysis. The model captures both the temporal structures in the data and the external factors impacting it. The following sections detail our interpretation of the results and the potential inferences drawn from them.



Interpretation:

The graph illustrates the total number of crimes per month from 2020 to mid-2024, followed by a forecast extending to 2026. Crime rates remain relatively high and stable from 2020 through early 2023, with monthly incidents fluctuating between 15,000 and 20,000. However, a sharp decline occurs in late 2023, with crime rates approaching near-zero levels by mid-2024.

Inference:

The ARIMA model forecast suggests that the declining trend in crime rates will continue into 2025, stabilizing at much lower levels compared to previous years. This indicates that the recent decrease in crime is not expected to reverse, and the low crime rate is anticipated to persist in the coming years. This trend can inform future planning by law enforcement agencies, which may need to adjust their resource allocation in response to a reduced crime landscape.

External Factors:

The abrupt decrease in crime rates in late 2023 could be attributed to significant external factors, such as major policy changes, increased law enforcement efforts, or shifts in social dynamics. Other possibilities include economic improvements, community crime prevention programs, or advancements in crime detection and deterrence technologies. These factors could have contributed to the sustained reduction in crime, but further investigation would be needed to pinpoint the exact causes.

Long-Term Stability:

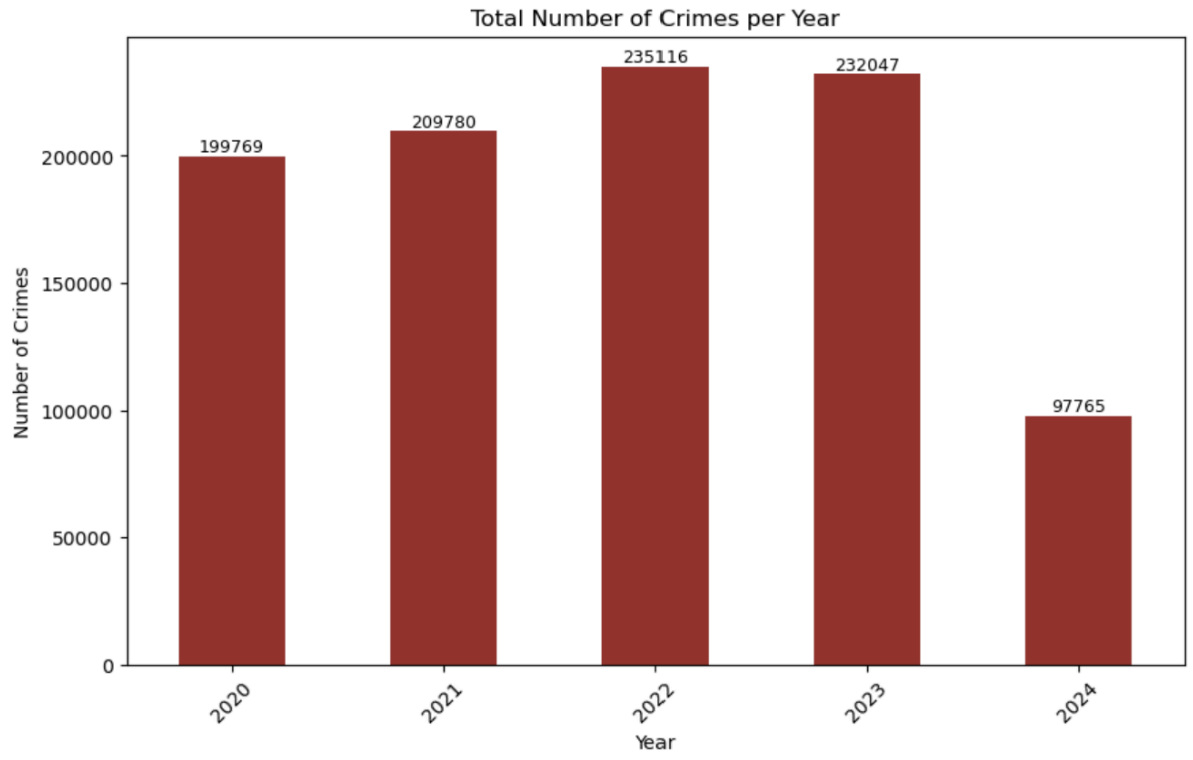
If the forecast is accurate, crime rates are likely to remain low and stable through 2026, indicating a period of long-term stability. This stability suggests that whatever factors contributed to the decline in crime may have had a lasting impact, potentially leading to a sustained reduction in criminal activity. This is a positive outlook, as it implies fewer crime-related challenges for communities and authorities.

Data Reliability:

While the forecast provides valuable insights, it is essential to remember that predictive models rely on historical data and trends. Unforeseen events or significant changes in external factors could alter the future trajectory of crime rates. Therefore, it is crucial to update the model with new data regularly to maintain its accuracy and relevance. Continuous monitoring and data analysis will ensure that the model remains a reliable tool for predicting and responding to crime trends.

everage this information for strategic planning and ensuring community safety.

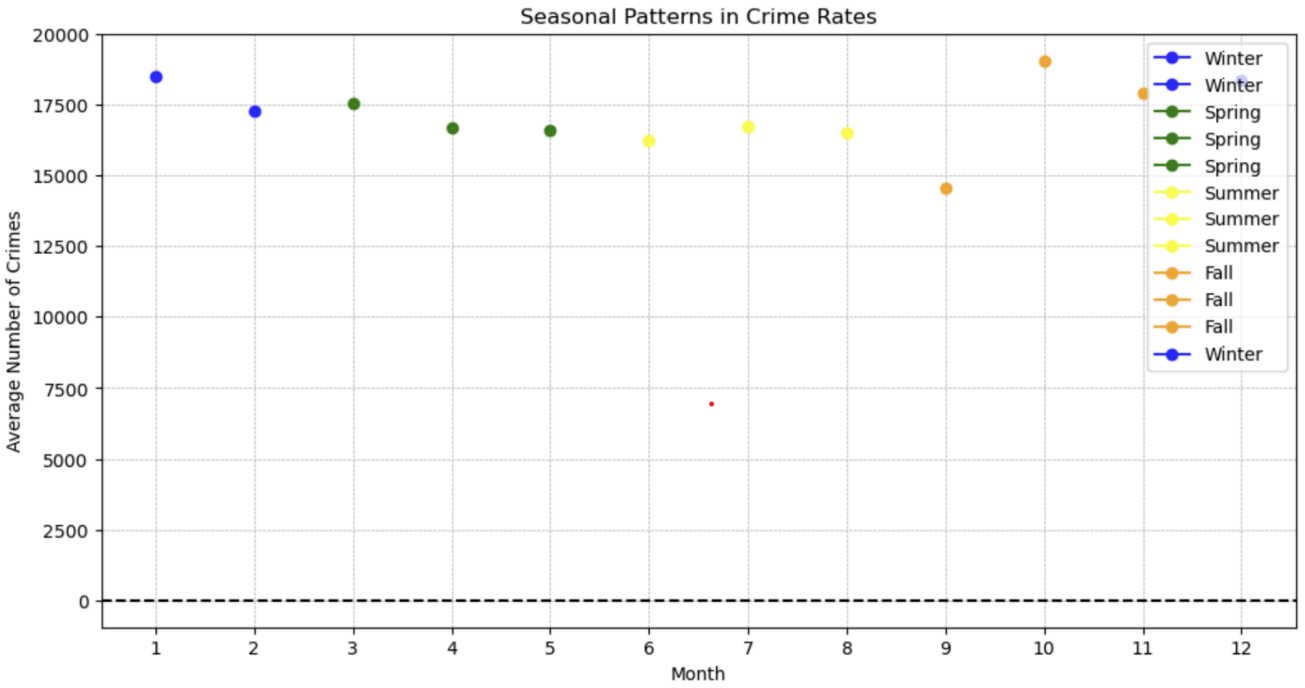
1. Overall Crime Trends:



The provided bar chart showcases the fluctuation in the total number of crimes between the years 2020 and 2024. Notably, there is an increasing trend in crime from 2020, which recorded 199,769 incidents, peaking in 2022 at 235,116 incidents. A slight decline follows in 2023, with 232,047 reported crimes, suggesting a minor drop but still a high level relative to previous years. The year

2024 shows a significant decrease, with 97,765 incidents recorded so far, though this figure likely reflects partial-year data as it's notably lower than in previous years. This downward trend in 2024 might indicate a potential reduction in crime rates, or it could simply be due to incomplete data for the current year. Overall, the chart highlights a relatively stable but high crime rate, with variations that warrant further investigation into specific factors influencing these yearly trends.

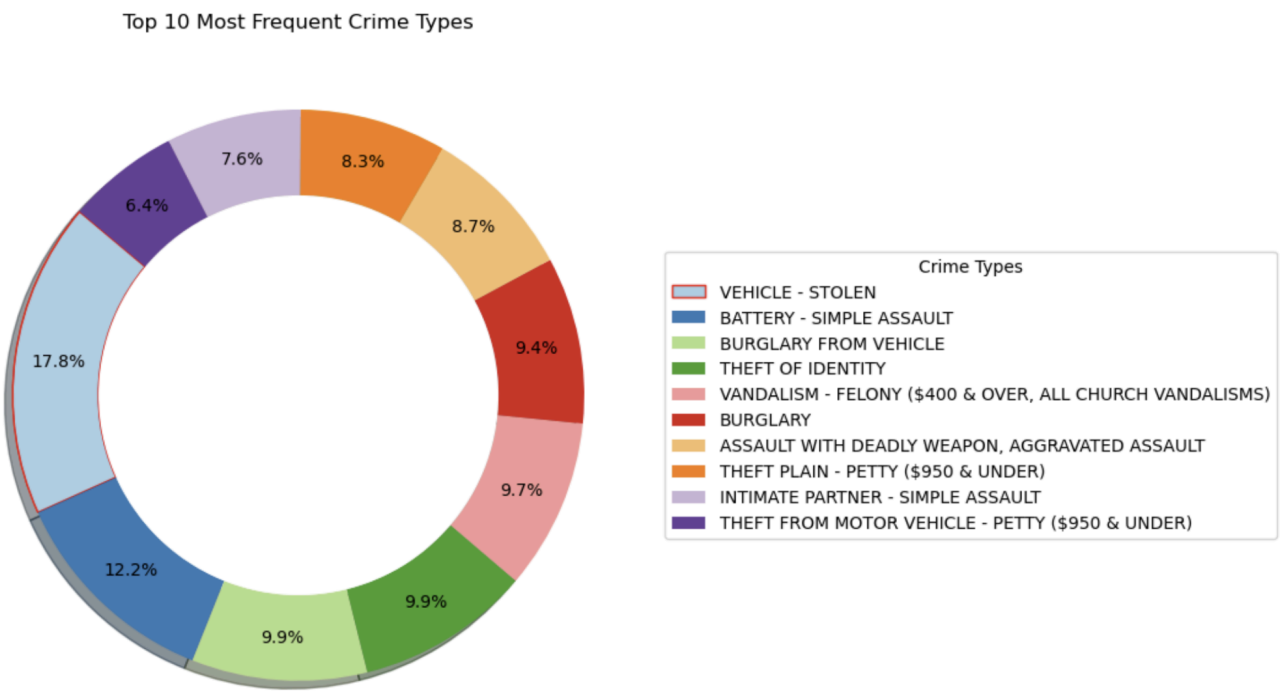
2. Seasonal Patterns:



The scatter plot illustrates the seasonal patterns in crime rates across different months, showing the average number of crimes per month categorized by season. Winter months (December, January, and February) and spring months (March, April, and May) consistently display higher average crime rates, around 17,500 to 19,000 incidents, indicating a peak in criminal activity during these times. In contrast, summer (June, July, and August) and fall (September, October, and November) show a slight decline in average crime rates, ranging between 15,000 and 17,000 incidents. This seasonal trend suggests that crime rates tend to be higher during the colder months and decrease slightly during warmer seasons. This pattern could be influenced by various factors, including seasonal behaviors, weather conditions, and social dynamics.

This data suggests that the crime rate might be influenced by seasonal variations. Winters seem to witness a heightened activity in crimes as compared to other seasons, especially fall, which shows a decline. Such patterns might be attributed to various socio-economic or environmental factors that vary across seasons. Further analysis would be required to pinpoint the exact reasons behind these fluctuations.

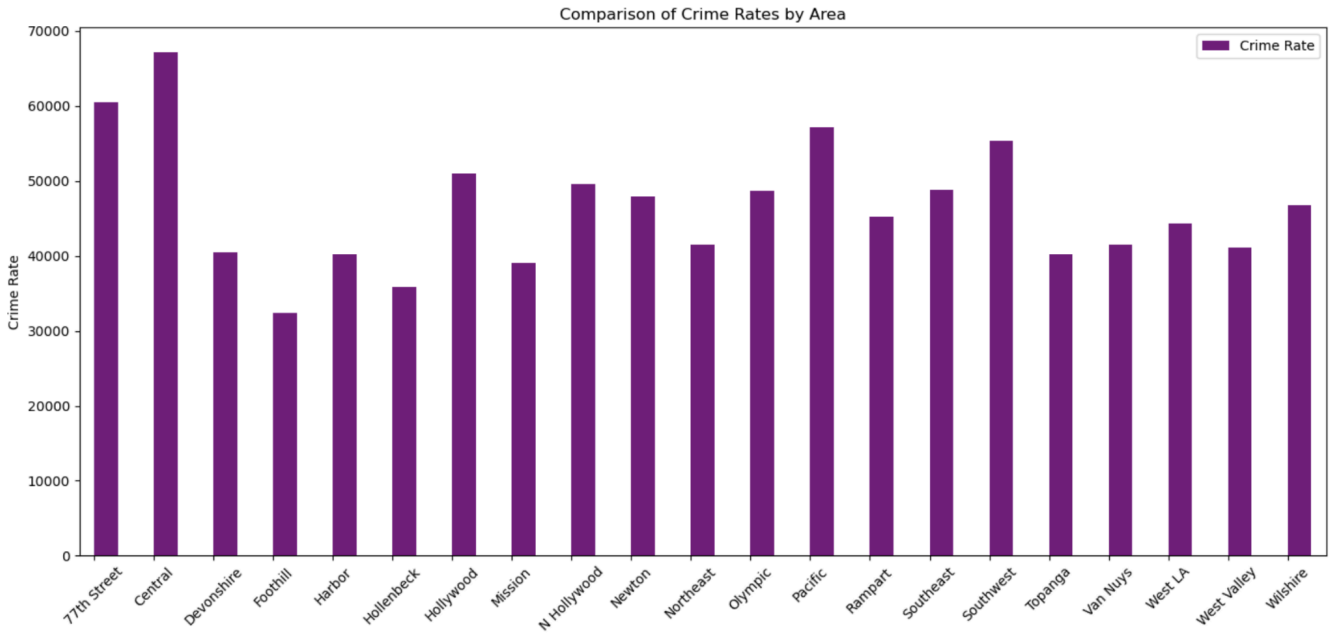
3. Most Common Crime Type:



The chart illustrates the top 10 most frequent types of crimes, with each segment representing a crime category's relative frequency. 'Vehicle - Stolen' is the most prevalent crime, comprising 17.8% of the total, followed by 'Battery - Simple Assault' at 12.2%. Other notable categories include 'Burglary from Vehicle' (9.9%) and 'Theft of Identity' (9.9%). Lower frequency crimes include 'Vandalism - Felony' at 9.7% and 'Theft Plain - Petty' ($950 & under) at 8.7%. The remaining categories, such as 'Assault with Deadly Weapon,' 'Intimate Partner - Simple Assault,' and 'Theft from Motor Vehicle - Petty' ($950 & under), make up smaller portions, ranging from 6.4% to 8.3%. This visualization highlights that theft-related offenses, particularly those involving vehicles, dominate the crime landscape, suggesting areas where public safety initiatives might focus to mitigate these issues.Each crime type is distinguished by a unique color, and their respective percentages provide insights into their

frequency relative to the top ten crime types in the dataset

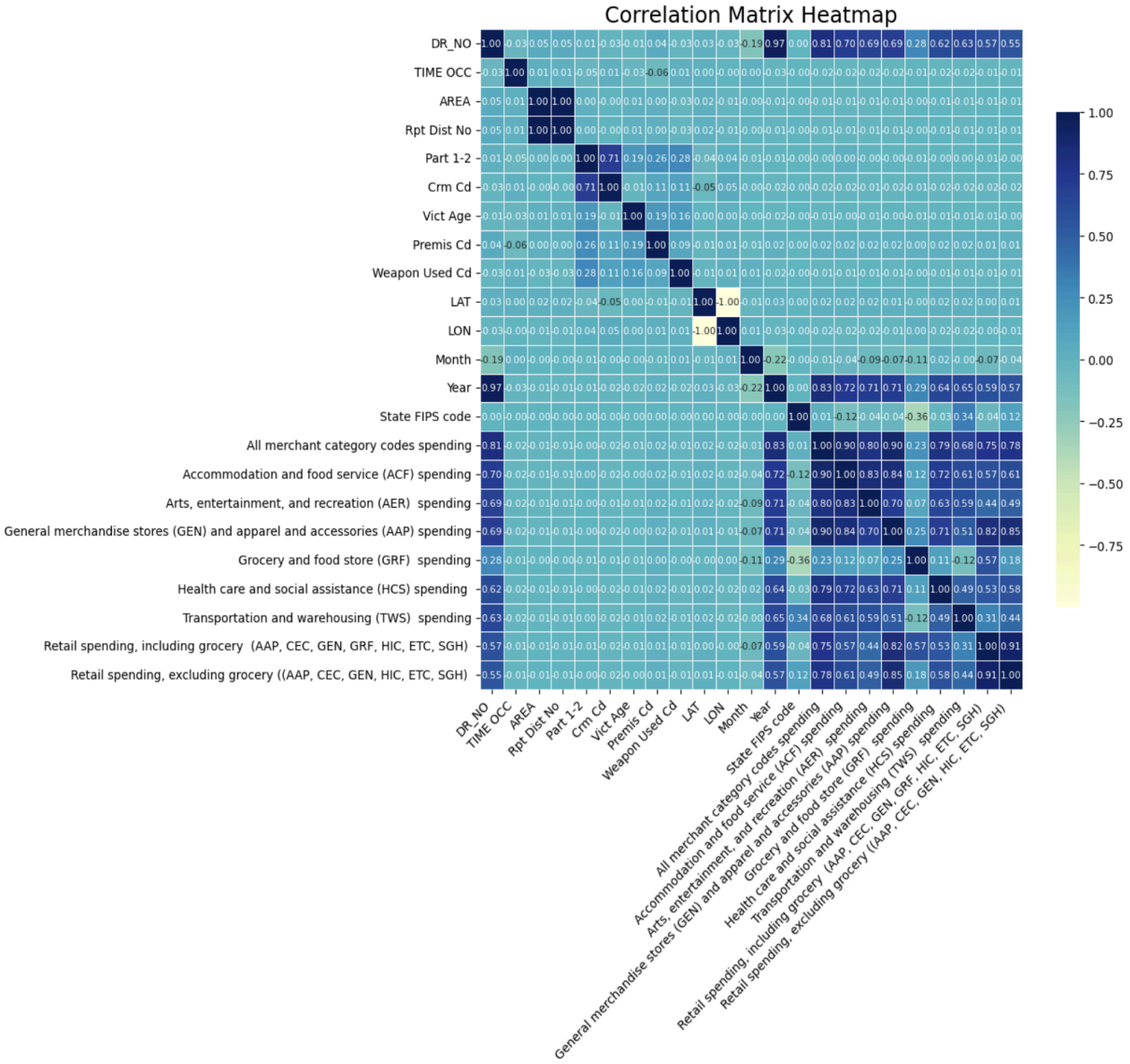
4. Regional Differences:



The bar chart provides a comparative view of crime rates across various regions, highlighting disparities in criminal activity. 'Central' and '77th Street' lead with the highest crime rates, each recording well above 60,000 incidents, underscoring these areas as significant crime hubs. Other regions such as 'Southwest,' 'Southeast,' and 'Newton' also demonstrate elevated crime levels, exceeding 50,000 incidents, which may signal a need for intensified public safety efforts and resource allocation.

On the other hand, areas like 'Topanga,' 'Van Nuys,' and 'Hollenbeck' reflect lower crime rates, staying below 40,000 incidents. These regions, while not devoid of crime, appear to experience comparatively fewer incidents, potentially allowing for a more preventive and community-focused approach. The variation in crime rates across different areas suggests that certain neighborhoods might benefit from targeted crime reduction strategies, tailored to address specific local challenges. Overall, this analysis offers valuable insights for prioritizing law enforcement resources and designing area-specific interventions to enhance public safety and crime prevention.

5. Correlation with Economic Factors:

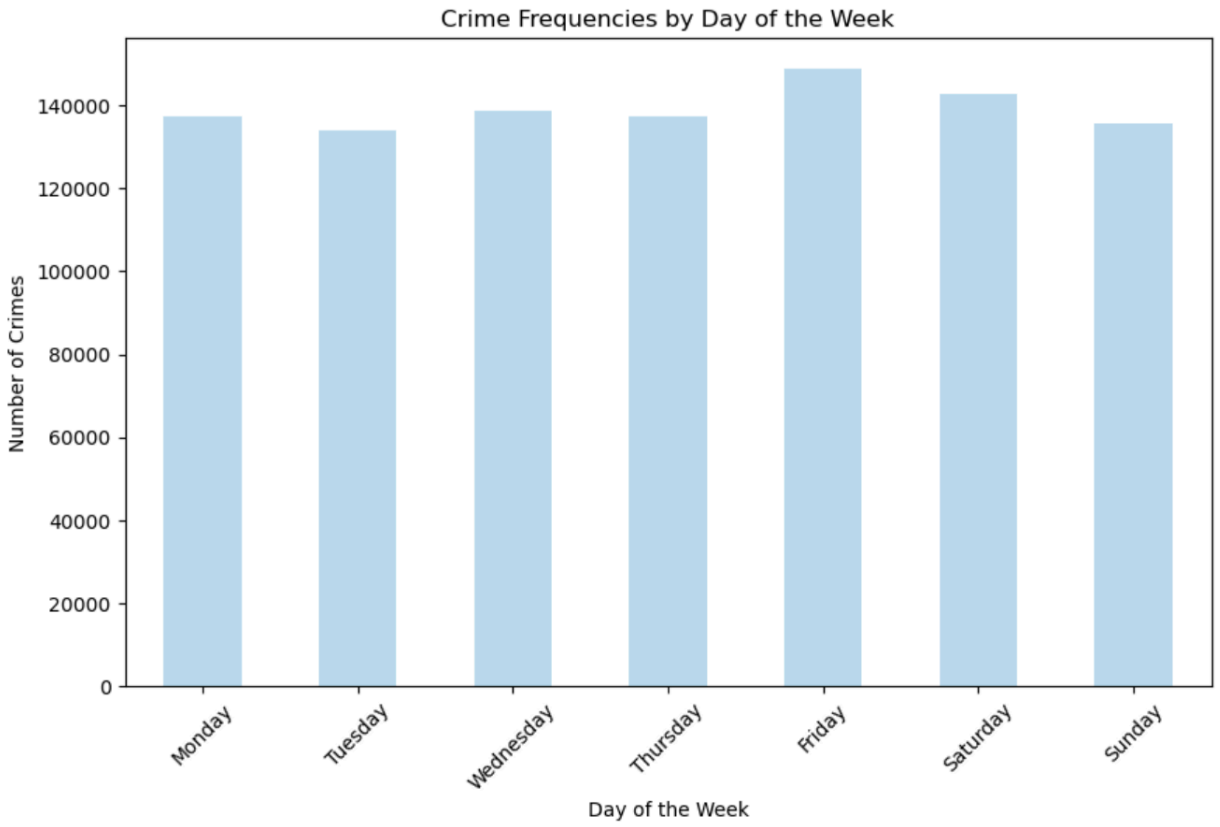


The correlation matrix heatmap reveals several insights into the relationships between crime-related attributes and economic factors. Notably, economic factors such as "All merchant category codes spending," "Accommodation and food service spending," "Arts, entertainment, and recreation spending," and "General merchandise stores spending" demonstrate strong positive correlations with each other. This suggests that these spending categories tend to fluctuate together. However, there is a weak correlation between crime-related attributes and economic factors, indicating that variations in consumer spending have a limited direct impact on crime rates within this dataset.

Time-related factors, such as the "Year," show moderate correlations with certain economic factors, implying shifts in spending over time, though not

necessarily influencing crime rates. Geographical factors, represented by latitude and longitude, also show minimal correlation with both crime and economic attributes, suggesting limited geographic influence on these economic categories. Additionally, categories like "Retail spending" (both including and excluding groceries) correlate strongly with other discretionary spending categories, reflecting interconnected consumer behavior. Overall, while economic factors are interrelated, they exhibit little direct correlation with crime trends, indicating that crime rates may be influenced more by other external factors beyond economic spending patterns.

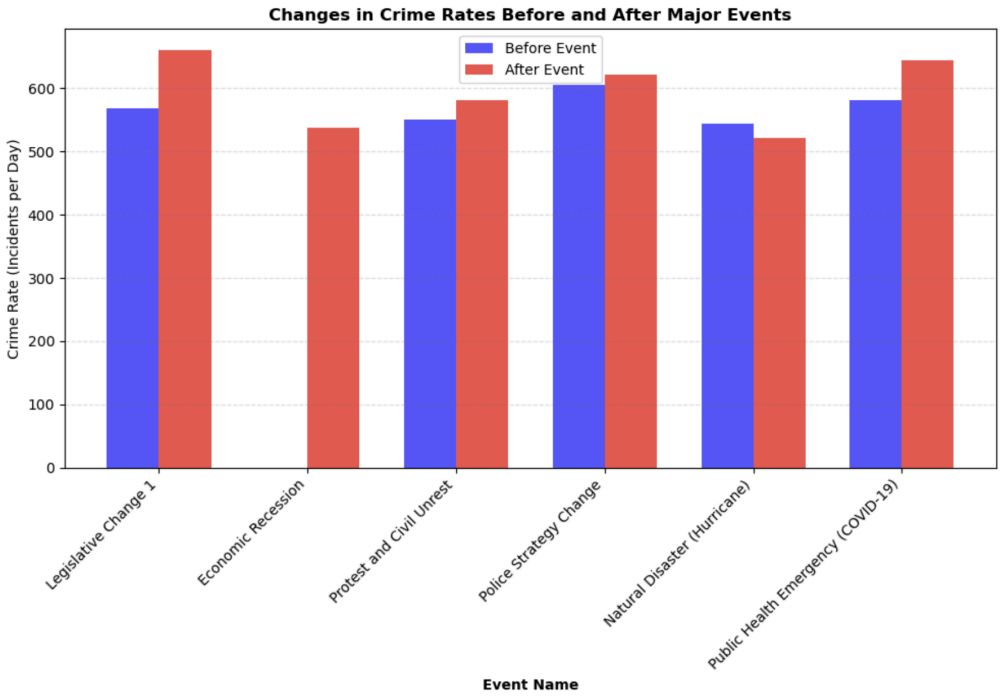
6. Day of the Week Analysis:



The bar chart displays crime frequencies by day of the week, indicating a fairly consistent distribution of crime across all days. Notably, Friday has the highest number of reported incidents, slightly exceeding 150,000 crimes, suggesting a marginal increase in criminal activity as the weekend approaches. Monday, Tuesday, and Wednesday maintain high crime rates as well, hovering around 145,000 incidents each, while Sunday shows a slight dip compared to other days.

The relatively even spread of crime throughout the week implies that criminal activity does not fluctuate drastically on any particular day, although the slight peak on Friday could reflect patterns associated with the start of the weekend, such as increased social activity or gatherings. This insight can help inform law enforcement agencies about resource allocation throughout the week, particularly around the weekend, to manage the slight uptick in criminal activity observed on Fridays.

7. Impact of Major Events:



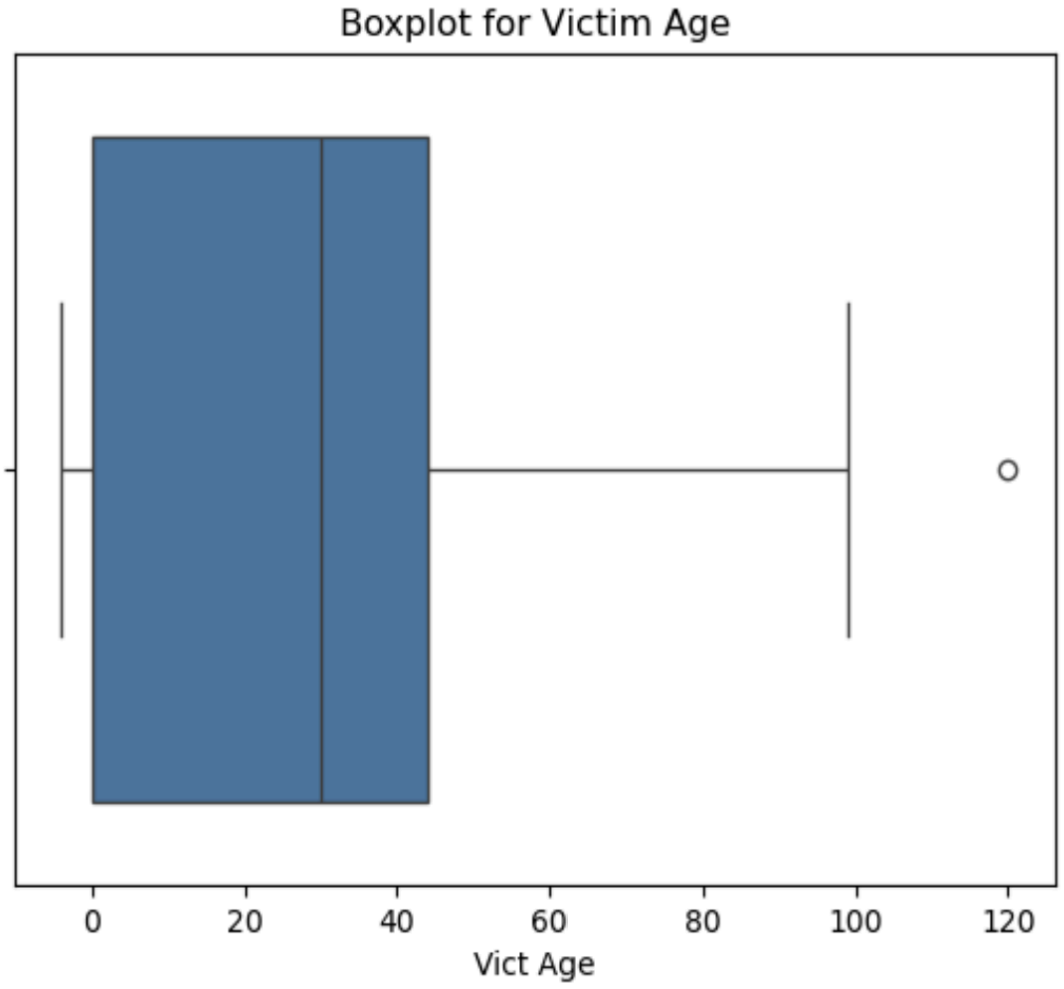
The bar chart compares crime rates before and after major events, illustrating how these incidents influenced daily crime rates. Each event shows variations in crime rates, with noticeable increases in some cases. For instance, after a legislative change (Event 1) and during a public health emergency (COVID-19), there was a significant rise in crime rates, with post-event rates reaching over 600 incidents per day.

The economic recession and protest periods saw moderate increases, while

crime rates following natural disasters and police strategy changes remained relatively stable. In the case of the natural disaster and police strategy change events, the crime rates before and after the events are nearly equal, suggesting minimal impact on daily crime incidents.

Overall, this chart reveals that certain events, particularly those with widespread societal implications like legislative changes and public health emergencies, can lead to substantial shifts in crime rates. In contrast, localized events, such as natural disasters, appear to have less pronounced effects on overall crime rates. This analysis underscores the need for adaptive crime prevention strategies tailored to different types of societal disruptions.

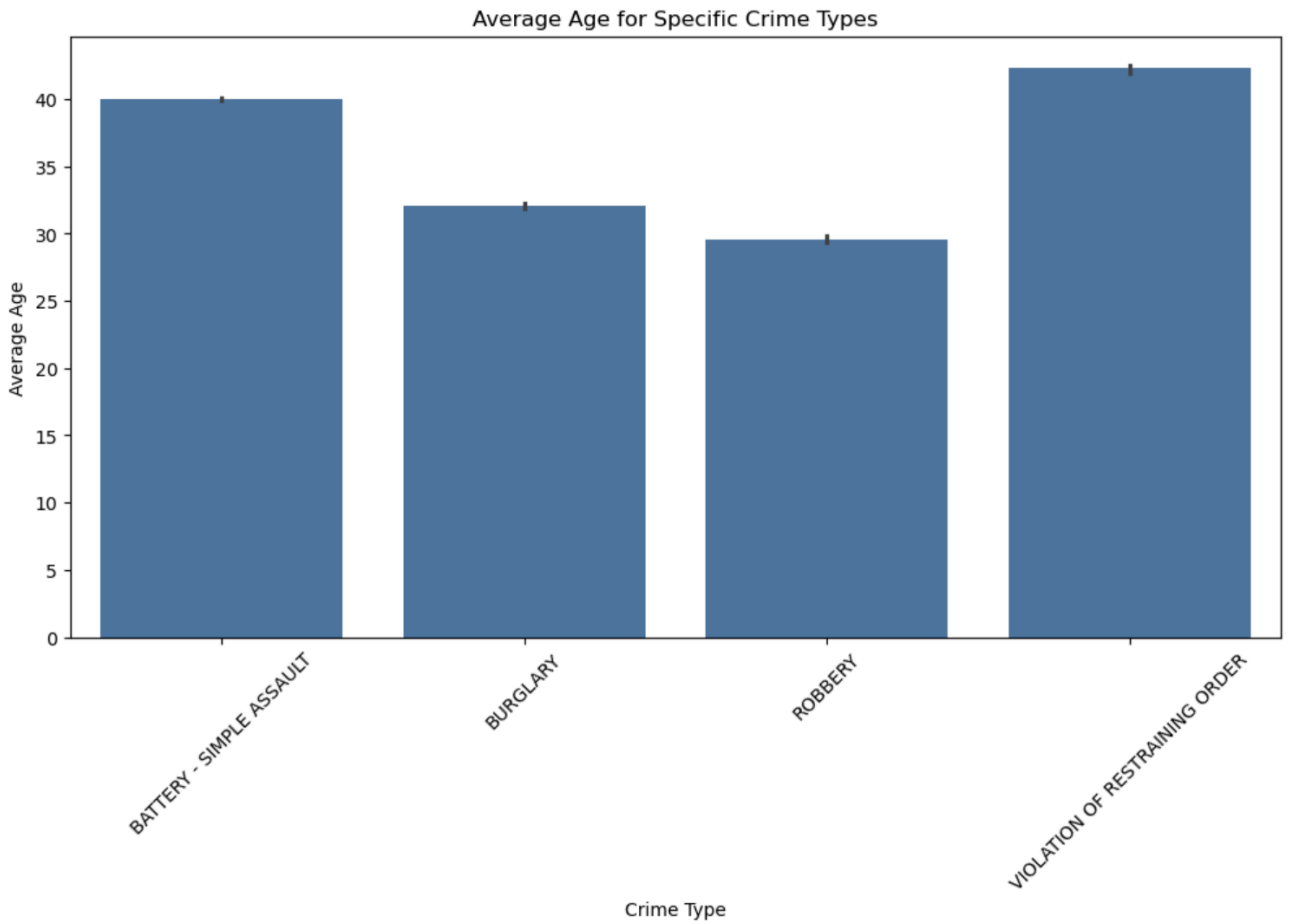
8. Outliers and Anomalies:



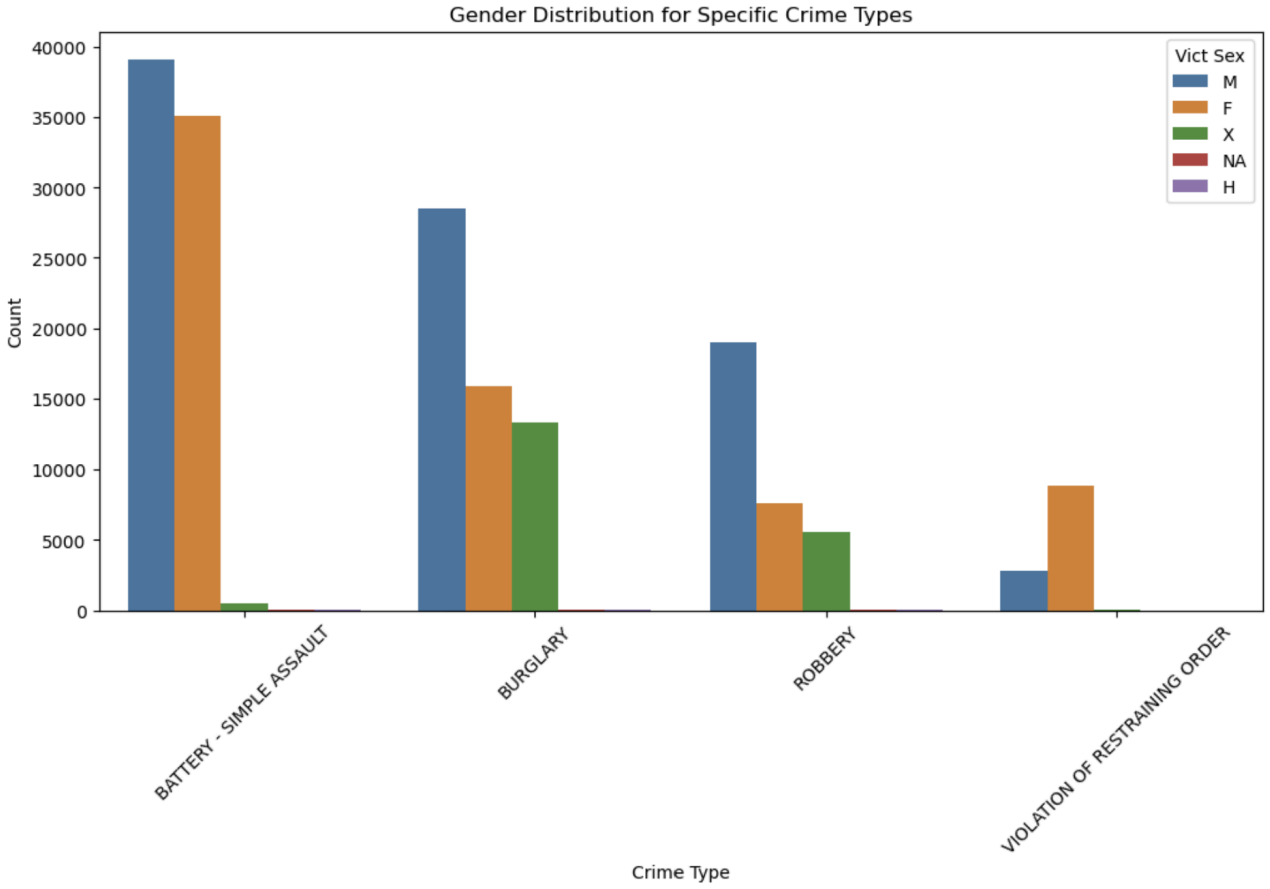
The box plot illustrates the distribution of victim ages, showing the spread and central tendency of this variable. The majority of victim ages fall within the interquartile range, which spans from approximately 20 to 40 years, indicating that the central 50% of victims are within this age range. The median victim age is also around this range, reflecting a concentration of cases among younger to middle-aged individuals.

The whiskers extend to roughly 0 and 100, highlighting the broader age range of victims, though extreme values are rare. Notably, there is an outlier beyond 100, suggesting a few cases involving significantly older victims. This outlier indicates that while most victims are relatively young, crimes do occasionally affect older age groups. Overall, the plot suggests that crime impacts a wide age spectrum, with a significant concentration among younger adults, underscoring the need for age-specific crime prevention and support measures.

9. Demographic Factors:



The bar chart shows the average age of victims for specific crime types, including "Battery - Simple Assault," "Burglary," "Robbery," and "Violation of Restraining Order." Victims of "Battery - Simple Assault" and "Violation of Restraining Order" tend to be older, with average ages around 40, indicating that these crimes are more frequently associated with adult victims. In contrast, "Burglary" and "Robbery" involve somewhat younger victims, with average ages in the mid-30s.



The bar chart displays the gender distribution for the same set of crime types. "Battery - Simple Assault" and "Burglary" have the highest number of male victims, followed closely by female victims, reflecting a similar trend in gender distribution for these crimes. "Robbery" also shows a higher incidence among male victims, while "Violation of Restraining Order" has a notably higher number of female victims compared to other crime types, suggesting that this crime predominantly affects women. The presence of other gender classifications such as "X," "NA," and "H" is minimal across all crime types, indicating they are less commonly recorded. These charts highlight the variations in victim demographics

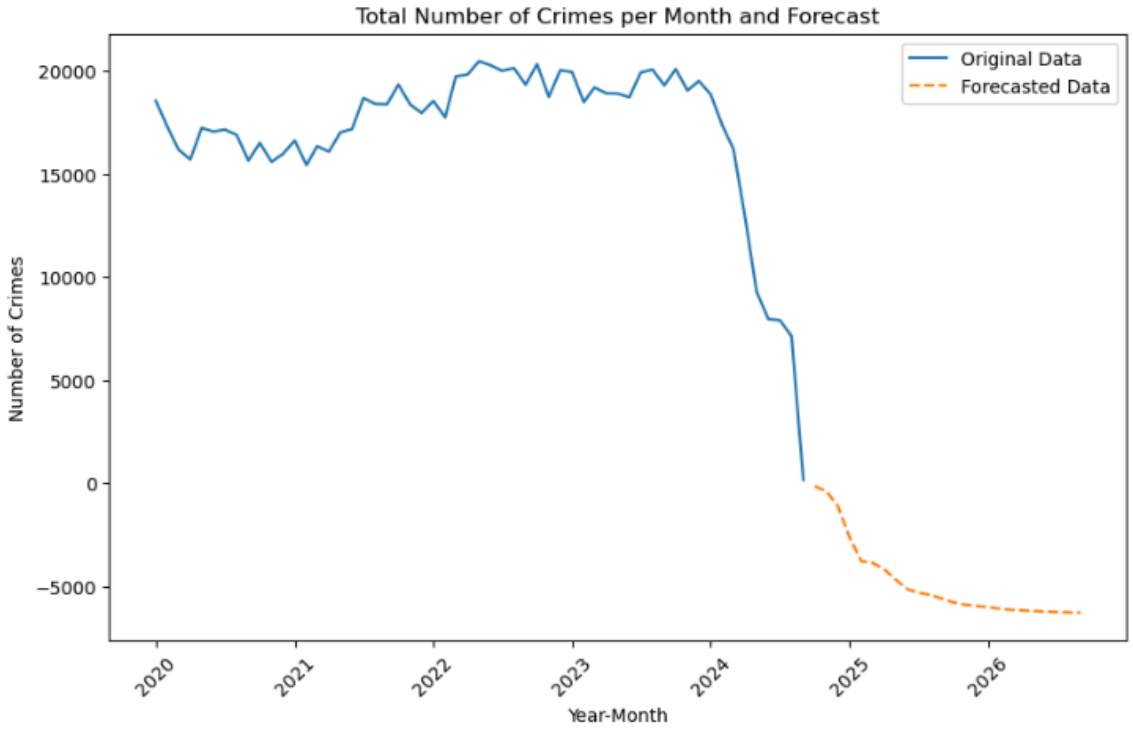
across different crime types, suggesting that certain crimes may be more prevalent among specific age groups and genders.

Gender Distribution for Specific Crime Types:

The chart showcases the gender distribution for victims of different crimes.

- "Battery/Simple Assault" sees a higher number of male victims than female victims.
- For "Robbery," both genders appear to be equally targeted.
- "Burglary" also has a roughly equal distribution, with a slight inclination towards male victims.
- "Violation of Restraining Order" predominantly affects female victims.
- Additionally, there's a category labeled "X" and another labeled "H" whose meanings aren't explicitly provided, as well as a "NA" category. To derive meaningful insights, it would be crucial to understand these categories better.
- Overall, gender plays a significant role in the type of crime, especially evident in cases like "Violation of Restraining Order."

10. Predicting Future Trends:



The line graph displays the total number of crimes per month from 2020 to 2024, alongside a forecast extending into 2026. The observed data indicates a stable crime rate from 2020 through early 2023, with monthly incidents typically ranging between 15,000 and 20,000. However, a sharp decline in crime rates begins in late 2023, dropping steeply until mid-2024 when crime rates approach zero.

The forecasted trend, shown by the dashed line, suggests that the downward trajectory is expected to continue, with projected values potentially dipping below zero by 2025. This unrealistic forecast of negative crime rates indicates limitations in the model and suggests that actual crime rates are likely to stabilize rather than continue to drop. The abrupt decline in 2024 might reflect data anomalies, changes in crime reporting practices, or policy interventions, highlighting the need for further investigation to understand the causes of this significant decrease.