Crime Data Analysis

Group 35

Project 1
Final Report

Course Code: IE6400
Course Name: Foundation of Data Analytics
Fall 2023

Team Members:
Ginnegolla, Hema Venkata Sai Teja
Gopanwar, Anuj
Nalam, Lokhi
Rasineni, Neha
Sanapureddy, Sree Dharani Reddy

Part 1: Data Acquisition and Inspection

We initiated our data analysis by downloading the dataset from the provided source and loading it into a Jupyter Notebook. Using Python's pandas library, we executed the `.info()` method to review the dataset's structure, gaining insights into the various data types. This preliminary inspection is a fundamental step to assess data quality and to plan for any necessary preprocessing.

Part 2: Data Cleaning

Handling Missing Values:

The initial step in our data-cleaning process involved addressing missing values. We began by examining the number of missing values in each column and identified a few rows with a significant number of missing values. Specifically, we found that the following rows: 'Crm Cd 2,' 'Crm Cd 3,' 'Crm Cd 4,' and 'Cross Street,' had an excessive number of missing values. Consequently, we made the decision to remove these four rows from the dataset.

Subsequently, for the remaining columns with missing values, we categorized them into two groups: numeric columns and categorical columns. For the columns containing numeric data, we opted to replace the missing values with the mean of the respective columns. In the case of columns with categorical values, we chose to replace the missing values with the mode (most frequently occurring value) of the respective columns. This comprehensive approach enabled us to effectively handle missing values in the dataset, ensuring that it was more robust and suitable for further analysis.

	Missing Count	Missing Percentage
Mocodes	114856	13.841775
Vict Sex	109299	13.172077
Vict Descent	109307	13.173041
Premis Cd	10	0.001205
Premis Desc	492	0.059293
Weapon Used Cd	540459	65.132963
Weapon Desc	540459	65.132963
Crm Cd 1	10	0.001205
Crm Cd 2	768750	92.645262
Crm Cd 3	827720	99.751982
Crm Cd 4	829717	99.992649
Cross Street	697270	84.030909

	Missing Count	Missing Percentage
DR_NO	0	0.0
Date Rptd	0	0.0
DATE OCC	0	0.0
TIME OCC	0	0.0
AREA	0	0.0
AREA NAME	0	0.0
Rpt Dist No	0	0.0
Part 1-2	0	0.0
Crm Cd	0	0.0
Crm Cd Desc	0	0.0
Mocodes	0	0.0
Vict Age	0	0.0
Vict Sex	0	0.0
Vict Descent	0	0.0
Premis Cd	0	0.0
Premis Desc	0	0.0
Weapon Used Cd	0	0.0
Weapon Desc	0	0.0
Status	0	0.0
Status Desc	0	0.0
Crm Cd 1	0	0.0
LOCATION	0	0.0
LAT	0	0.0
LON	0	0.0

Handling Duplicate Values:

As part of our data cleaning process, we conducted a thorough examination to identify and address duplicate values within the dataset. Fortunately, our analysis revealed that there were no duplicates present, ensuring that the dataset was free from this data quality issue. This absence of duplicate values contributes to the overall data integrity and reliability, facilitating more accurate and meaningful analysis.

```
# Remove duplicate rows
cleanx_dtx = crimex_datasetx.drop_duplicates()

# Check if any duplicates remain
dupx = cleanx_dtx.duplicated().sum()
dupx
```

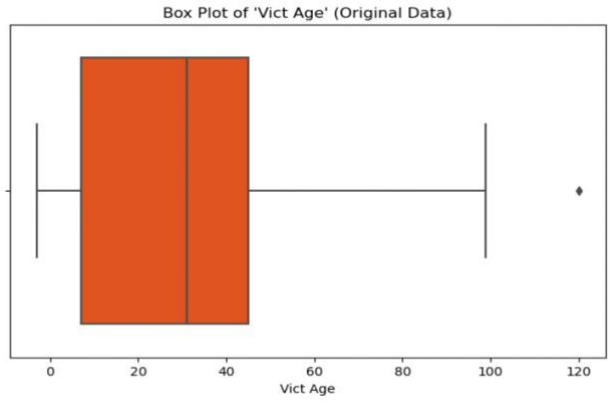
```
0
```

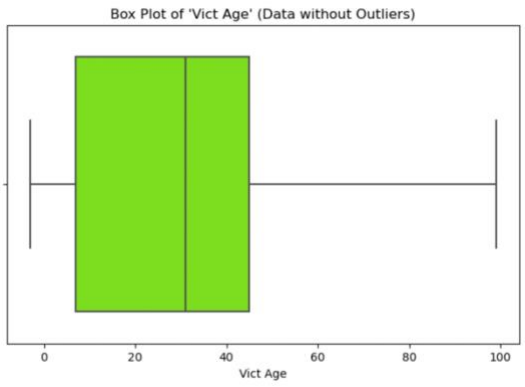
Conversion of Data Types:

In this data cleaning step, we focused on converting the data type of the 'date' column. Originally, the data type for this column was set as an object. To enhance the usability and consistency of the dataset, we transformed this column's data type to datetime. This conversion allows for better handling and analysis of date-related information, enabling more accurate and meaningful insights from the data.

Dealing with Outliers:

In our data-cleaning process, we employed a box plot analysis to identify potential outliers within the dataset. Subsequently, we identified outliers in the columns related to 'Vict age,' 'LAT,' and 'LON'. However, after careful consideration, we determined that the 'LAT,' and 'LON' columns were not essential for our analysis and decided not to address the outliers in these columns. Notably, only one outlier was detected in the 'Vict age' column, which we removed to maintain data integrity. This decision was made based on the understanding that these outliers did not significantly impact the specific analysis or objectives of our project, allowing us to focus on the more critical aspects of the data.





Standardization or Normalization of the numeric data:

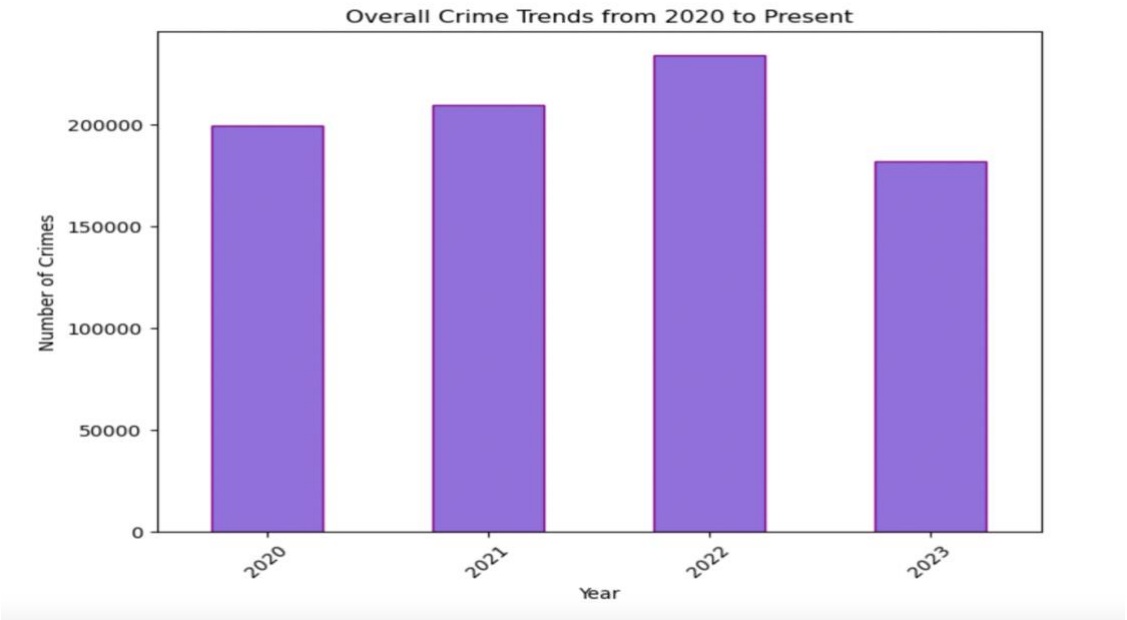
In the next phase, we employed two key techniques to transform the 'Vict Age' column of the crime dataset: standardization and normalization. Utilizing scikit-learn's StandardScaler, we standardized the 'Vict Age' column, adjusting its values to have a mean of 0 and a standard deviation of 1. The standardized values were subsequently stored in a new column titled 'Vict Age Standardized'. Following this, we implemented normalization using the MinMaxScaler. This process adjusted the 'Vict Age' values to lie within a 0 to 1 range, with these normalized values saved in another new column, 'Vict Age Normalized'. By executing these transformations, we effectively scaled our dataset, ensuring that the 'Vict Age' values are now suitable for algorithms sensitive to feature scales, enhancing the overall robustness and efficacy of any subsequent analyses.

Encode categorical data if present:

In this phase, we addressed the categorical 'AREA NAME' column in the crime dataset using the OneHotEncoder from scikit-learn. This encoder transformed each unique area into a distinct binary column, ensuring compatibility with machine learning algorithms. The binary-encoded data, represented in the area_encoded_df, was then merged with our primary dataset, resulting in the comprehensive crime_data_encoded data frame, enriched with one-hot encoded categorical data for enhanced analytical readiness.

Part 3: Exploratory Data Analysis

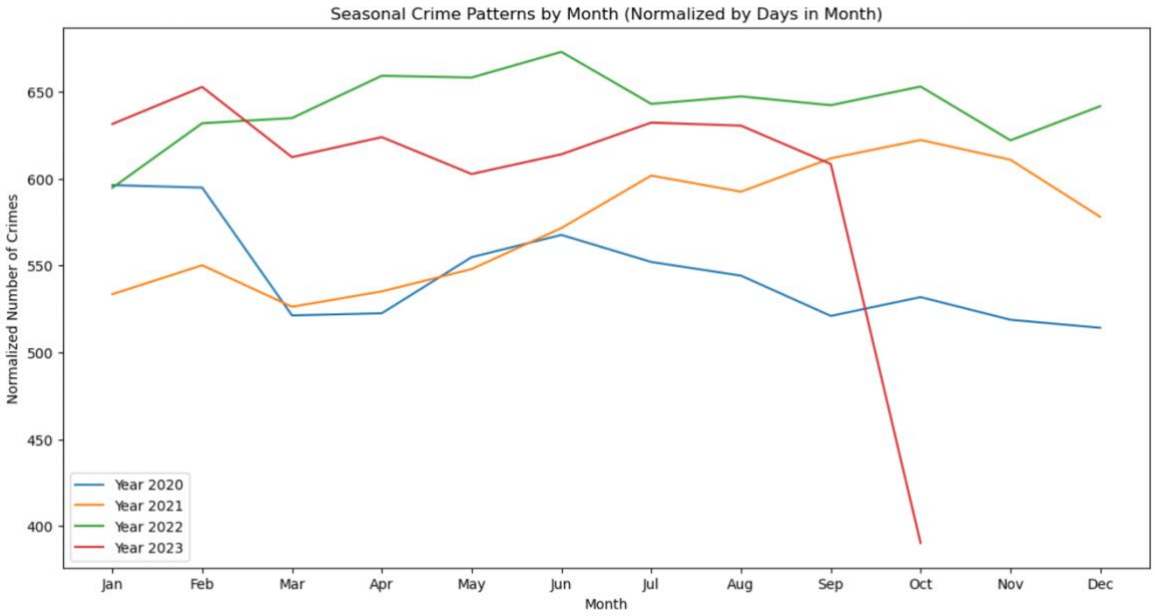
1. Visualize overall crime trends from 2020 to the present year.



The visualization offers insight into crime trends from 2020 to the present year. In 2020, crime rates commenced at 200K, witnessing a steady ascent, peaking at 230K by the start of 2022.

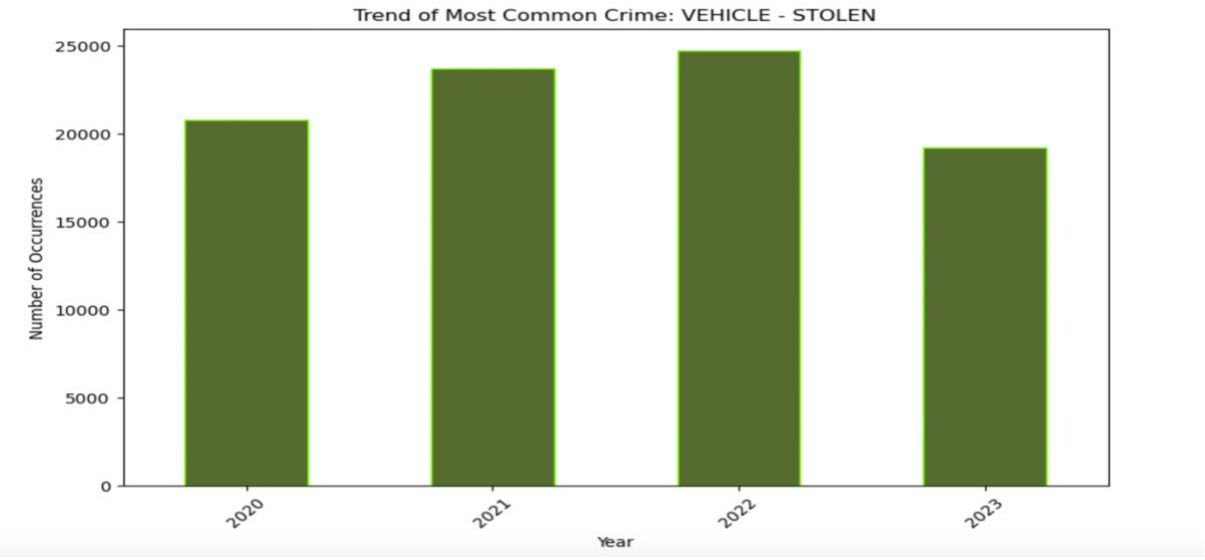
Subsequently, a gradual decline is observed, with numbers descending to 170K at the onset of 2023. This pattern underscores the variability in crime occurrences over this period.

2. Analyze and visualize seasonal patterns in crime data.



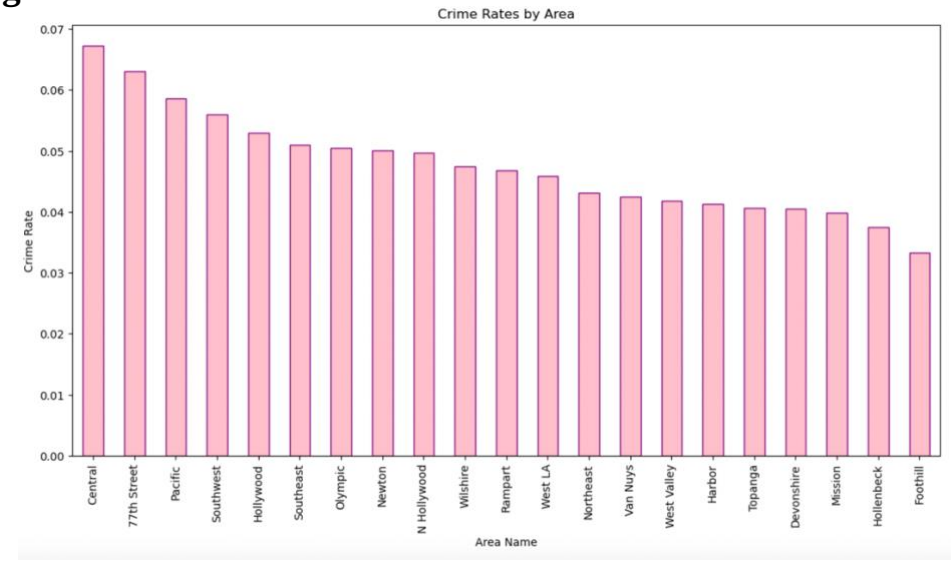
This graph, normalized by days in each month, delineates the fluctuations in crime rates from 2020 to 2023. Across all years, a discernible trend is the peak in crime rates during mid-year, specifically from May to August, with 2021 showcasing a pronounced spike in September. Conversely, the latter part of the year, particularly from October to December, typically witnesses a downturn in criminal activity. While these overarching patterns persist, individual years also exhibit unique variations, underscoring the multifaceted influences on crime occurrences.

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



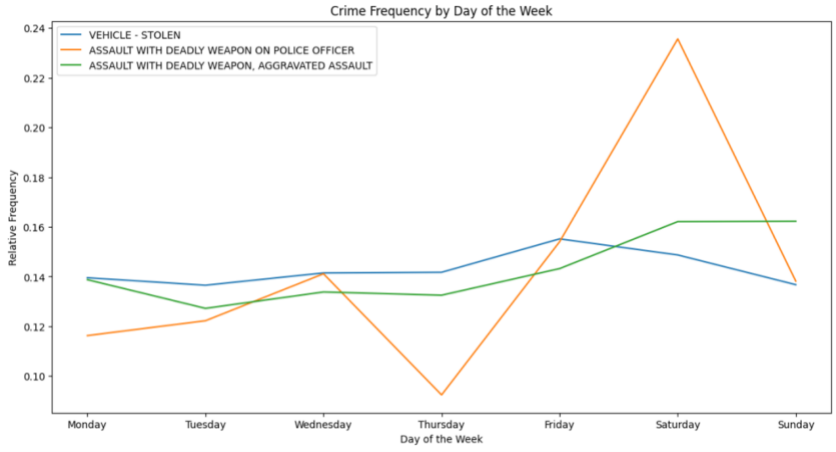
The bar chart depicts the trend of vehicle thefts, identified as the most common crime, from 2020 to 2023. In 2020, the occurrences were just above 20,000, and it saw a dip in 2021. However, the number of thefts surged in 2022, reaching a level comparable to 2020, before experiencing a slight decline in 2023. Overall, the graph illustrates a fluctuating trend in vehicle thefts over the four-year period, with 2021 having the fewest incidents and 2022 seeing a significant resurgence.

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



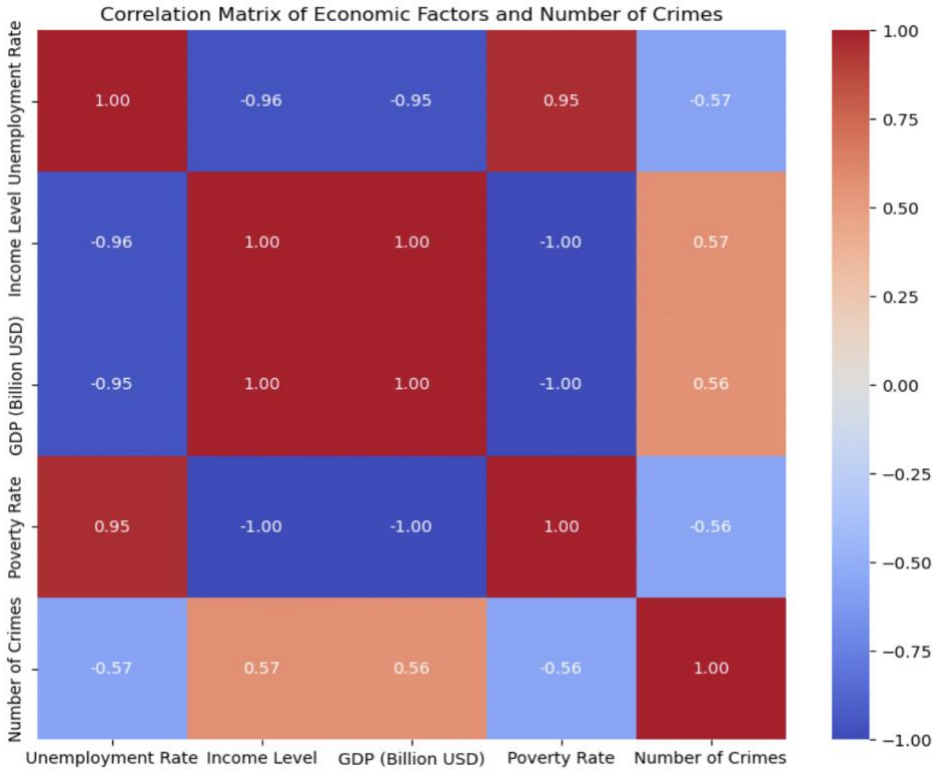
The bar chart represents crime rates across various areas or regions. At first glance, "Central" has the highest crime rate, closely followed by "77th Street" and "Pacific". However, as we move rightward on the graph, the crime rates generally show minimal variation, with most areas having rates between 0.04 and 0.05. The only exception to this consistency is "Foothill", which exhibits the lowest crime rate among all the areas listed. This data suggests that while there are some discrepancies in crime rates between regions, most of the areas have somewhat comparable crime levels, with "Central" being a notable outlier on the higher end and "Foothill" on the lower end.

5. Analyse the relationship between the day of the week and the frequency of certain types of crimes.



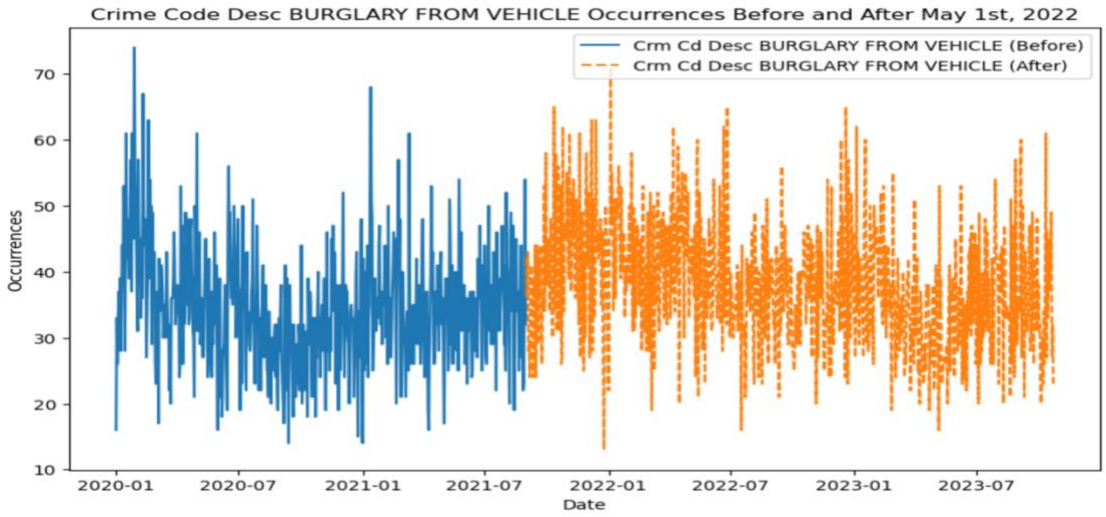
This graph presents an intriguing analysis of the correlation between specific crime types and the days on which they predominantly occur. Stolen vehicles, depicted in blue, maintain a relatively consistent frequency throughout the week. In contrast, the assault with a deadly weapon on a police officer, represented in orange, exhibits a stark spike on Sundays, while the other days remain notably lower. The green line, signifying assault with a deadly weapon resulting in an aggravated assault, reveals a gradual increase from Monday to Friday, followed by a slight decline during the weekend. This data underscores the importance of understanding temporal patterns when strategizing preventive measures.

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

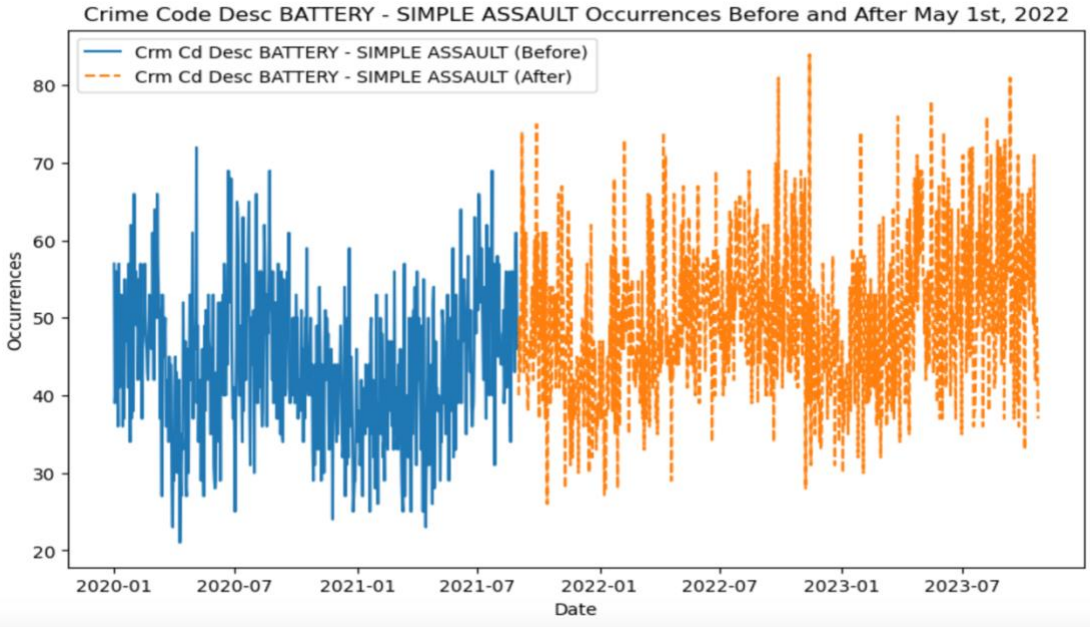


The provided correlation matrix elucidates the interrelationships between various economic factors and their association with crime rates. There is a pronounced negative correlation between the unemployment rate and both income level (-0.96) and GDP (-0.95), suggesting that regions with higher unemployment tend to have lower income levels and GDP. A perfect positive correlation between income level and GDP (1.00) indicates their concurrent movement. Interestingly, while higher unemployment correlates with an increase in the number of crimes (-0.57), regions with a higher GDP witness a contrasting trend, correlating with fewer crimes (0.56). This data underscores the intertwined nature of economic well-being and its potential influence on crime dynamics.

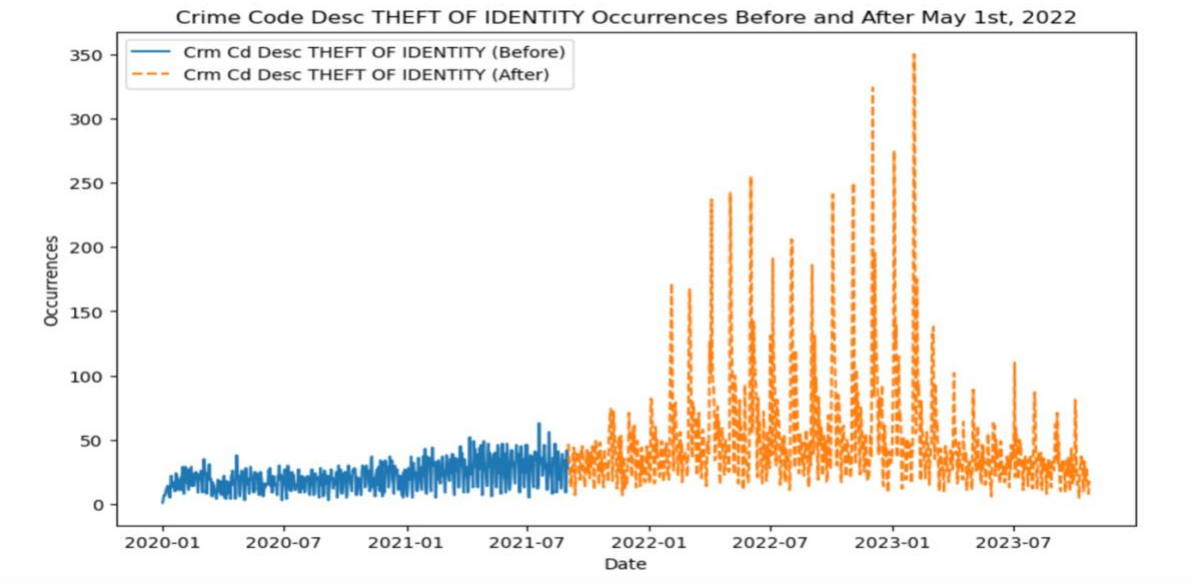
7. Investigate any impact of major events or policy changes on crime rates.



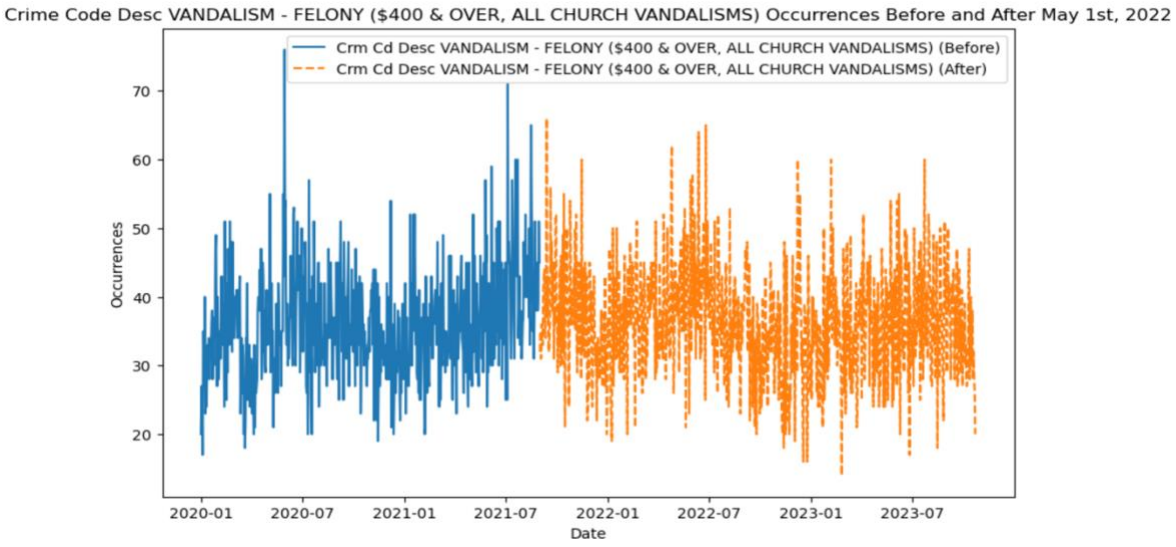
Burglary from Vehicle Trends: The graph illustrates the occurrences of "BURGLARY FROM VEHICLE" incidents before and after May 1st, 2022. Prior to May, as indicated by the blue line, occurrences fluctuated between roughly 10 to 70 daily incidents. However, after May 1st, represented by the orange dotted line, there's an observable rise in the frequency of such incidents, with numbers predominantly ranging from 30 to 70. The increase in the density of orange spikes after May suggests a higher consistency in burglary from vehicle occurrences during this period.



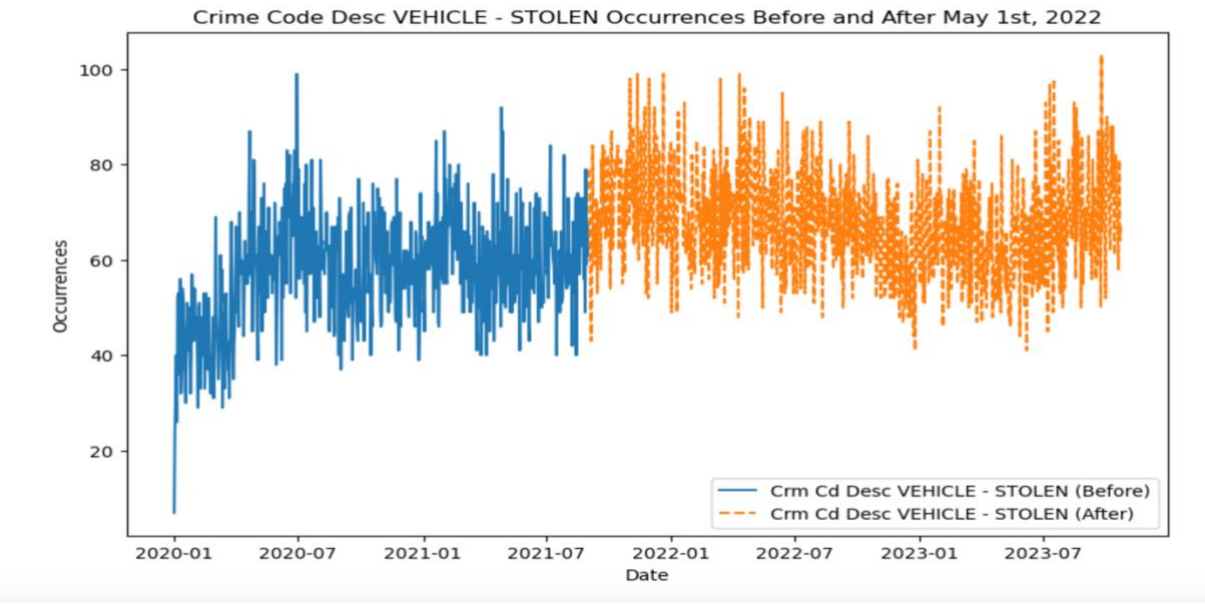
Simple assault: The graph shows the crime occurrences spanning from January 2020 to July 2023, a clear demarcation is observed before and after May 1st, 2022. Prior to this date, represented by the blue line, the occurrences fluctuated between approximately 20 to 80 daily incidents, with a relatively stable pattern. However, post-May 1st, 2022, delineated by the orange dashed line, there's a noticeable increase in the variability of occurrences, even though the range remains similar.



Theft of Identity Trends: The graph showcases the occurrences of "THEFT OF IDENTITY" crimes. The blue line represents the occurrences before May 1st, 2022, while the orange dotted line illustrates those after. Before May 1st, the daily incidents remain relatively low, oscillating below 50 occurrences. After this date, there's a noticeable increase with occurrences predominantly ranging from 150 to 350 daily incidents. The frequency of spikes post-May suggests a heightened and consistent pattern of identity.



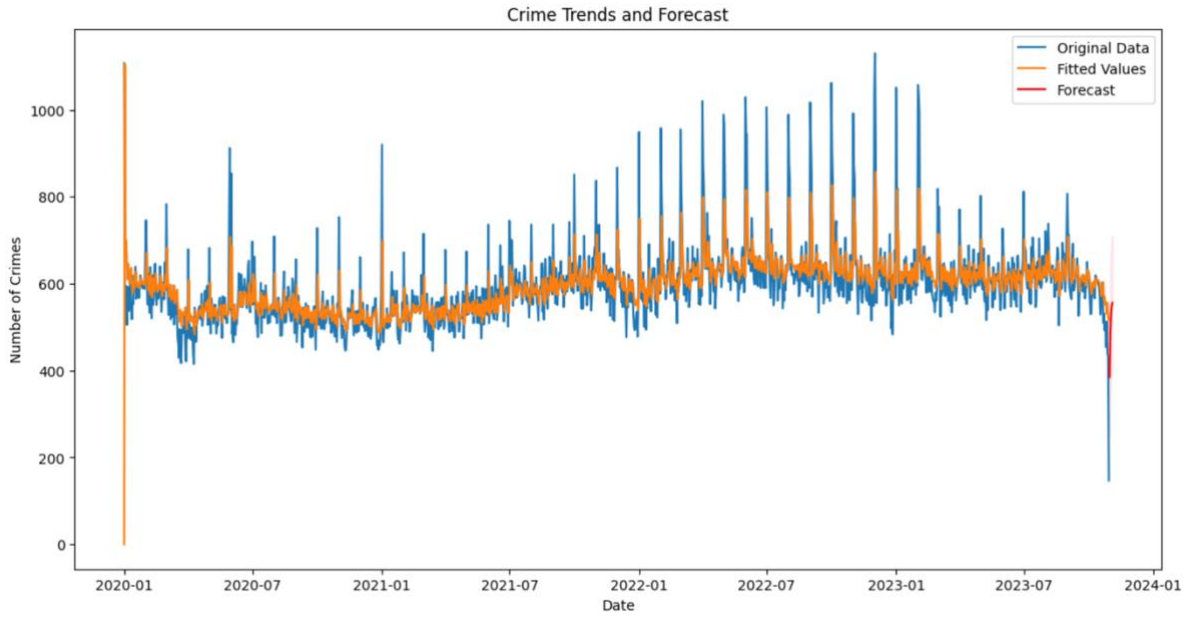
Vandalism in Churches: The above graph showcases the occurrences of felony vandalism specifically targeting churches. Similar to the first graph, the blue line shows the occurrences before May 1st, 2022, while the orange dotted line represents those after. Before May 1st, there are peaks and troughs, with an overall trend of declining occurrences nearing the cutoff date. After May 1st, the occurrences seem to have increased, but the frequency of spikes and drops appears to be more regular compared to the previous period.



Stolen Vehicles: This graph illustrates the occurrences of vehicle thefts. The blue line indicates the occurrences before May 1st, 2022, and the orange dotted line represents occurrences after that date. It's evident that the frequency of this type of crime was somewhat steady before May 1st, 2022, with occasional spikes. Post May 1st, there is a noticeable increase in the occurrences of stolen vehicles, although there are still some fluctuations.

Part 4: Advanced Analysis

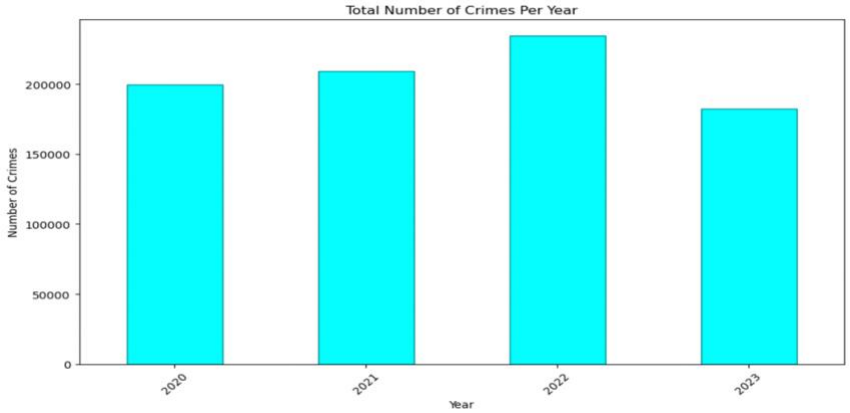
Use predictive modelling techniques (e.g., time series forecasting) to predict future crime trends.



To predict future crime trends, an ARIMA model with parameters (1, 1, 1) was applied to a time series of crimes aggregated by date. Preliminary exploration involved differencing the data to achieve stationarity and analysing the autocorrelation to determine model parameters. Upon fitting the ARIMA model, it was used to forecast crime rates for the next five periods. The visualization captures historical data, the model's fitted values, and the forecast, with a shaded region representing the confidence interval around the predictions. This approach allows stakeholders to understand potential future crime scenarios and make proactive decisions.

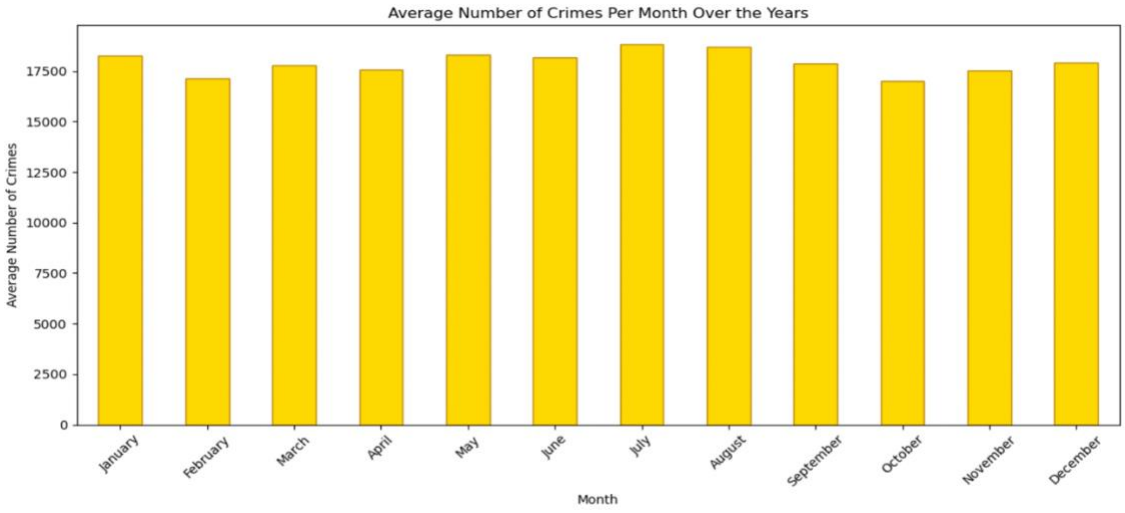
Part 5: Explore additional questions:

1. **Overall Crime Trends:** Calculate and plot the total number of crimes per year to visualize the trends.



The bar chart illustrates the total number of crimes from 2020 to 2022, indicating a consistent rise in reported incidents during this period. In contrast, 2023 shows a marked decline in the crime rate. This reversal in trend in 2023 is noteworthy, suggesting that while the previous years saw an upward trajectory, there was a shift or intervention in the recent year that contributed to the reduction in crime numbers.

2. **Seasonal Patterns:** Group the data by month and analyze the average number of crimes per month over the years.



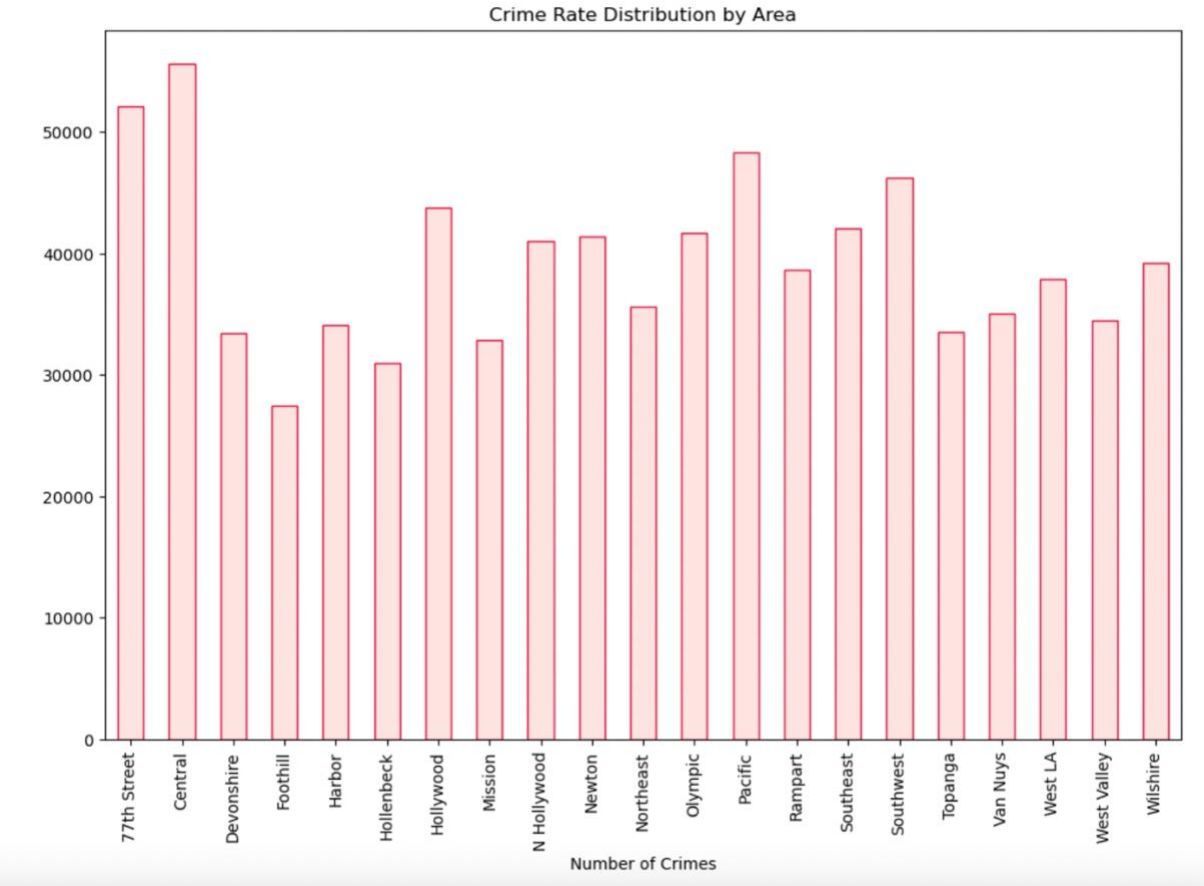
The bar chart displays the average number of crimes per month over an unspecified range of years. From January to December, the number of crimes remains relatively consistent, with only slight fluctuations. This suggests that, across the years, there are no pronounced seasonal patterns or specific months with significantly higher or lower crime rates. In other words, the monthly average of reported crimes appears to be stable throughout the year.

3. **Most Common Crime Type:** Count the occurrences of each crime type and identify the one with the highest frequency

```
('VEHICLE - STOLEN', 88892)
```

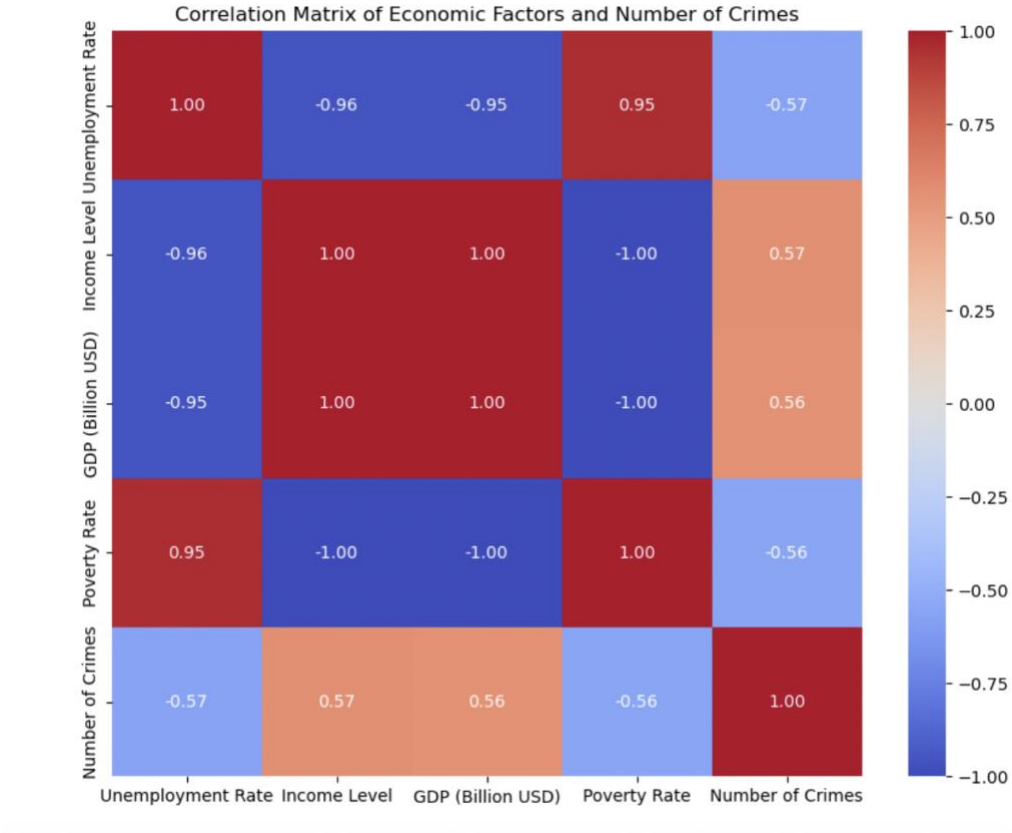
The data indicates that the most frequently reported crime type is "VEHICLE - STOLEN" with a count of 88,892 incidents. This suggests that vehicle theft is the predominant crime in the specified dataset or region over the given time period.

4. **Regional Differences:** Group the data by region or city and compare crime rates between them using descriptive statistics or visualizations.



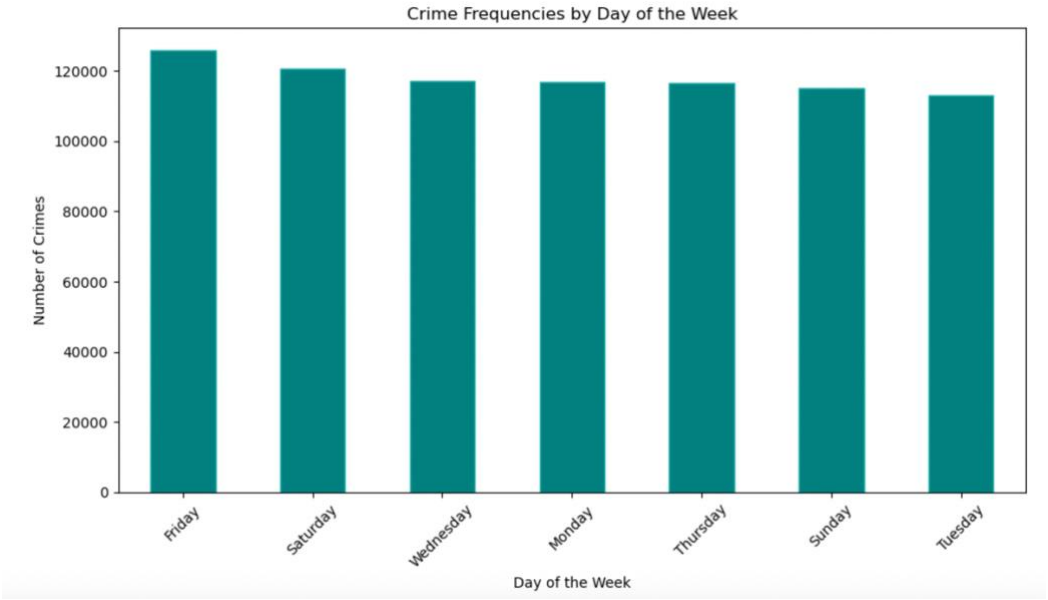
The bar chart illustrates the distribution of crime rates across various regions or areas. The highest crime rate appears to be concentrated in the "77th Street" region, closely followed by "Central" and "Southwest." Conversely, areas like "Harbor" and "Hollenbeck" have considerably lower crime occurrences in comparison. The data showcases a significant variance in the number of crimes across these regions. Such regional differences could be attributed to a myriad of factors, including socioeconomic conditions, population density, or law enforcement practices. It would be essential to delve deeper into these underlying causes to formulate effective crime prevention strategies tailored to each area.

5. **Correlation with Economic Factors:** Collect economic data for the same time frame and use statistical methods like correlation analysis to assess the relationship between economic factors and crime rates.



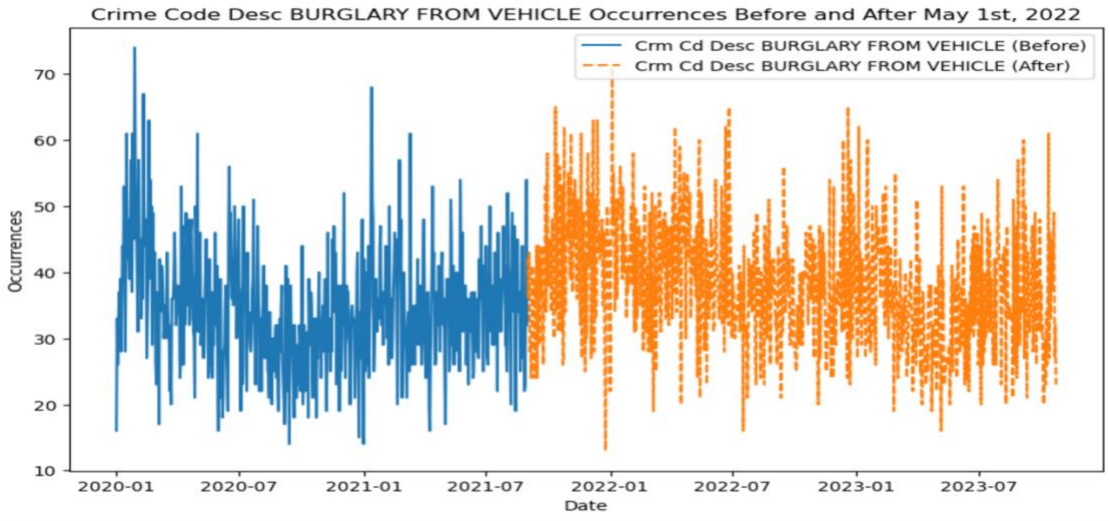
The correlation matrix visually represents the relationships between various economic factors and their potential impact on the number of crimes. Notably, there's a pronounced negative correlation between the unemployment rate and income level (-0.96), suggesting that areas with higher incomes tend to have lower unemployment. Similarly, as GDP increases, poverty rates appear to decline, indicated by a correlation of -1.00. However, the relationship between economic well-being and crime is more complex. The number of crimes displays a moderate positive correlation with GDP (0.56) and income level (0.57), implying that regions with economic prosperity might experience a slight uptick in crime rates. Conversely, areas with higher unemployment show a modest decrease in crime rates, evidenced by a -0.57 correlation. This matrix underscores the multifaceted interplay between economic indicators and crime prevalence, necessitating a deeper analysis to understand the underlying dynamics.

6. Day of the Week Analysis: Group the data by day of the week and analyze crime frequencies for each day.

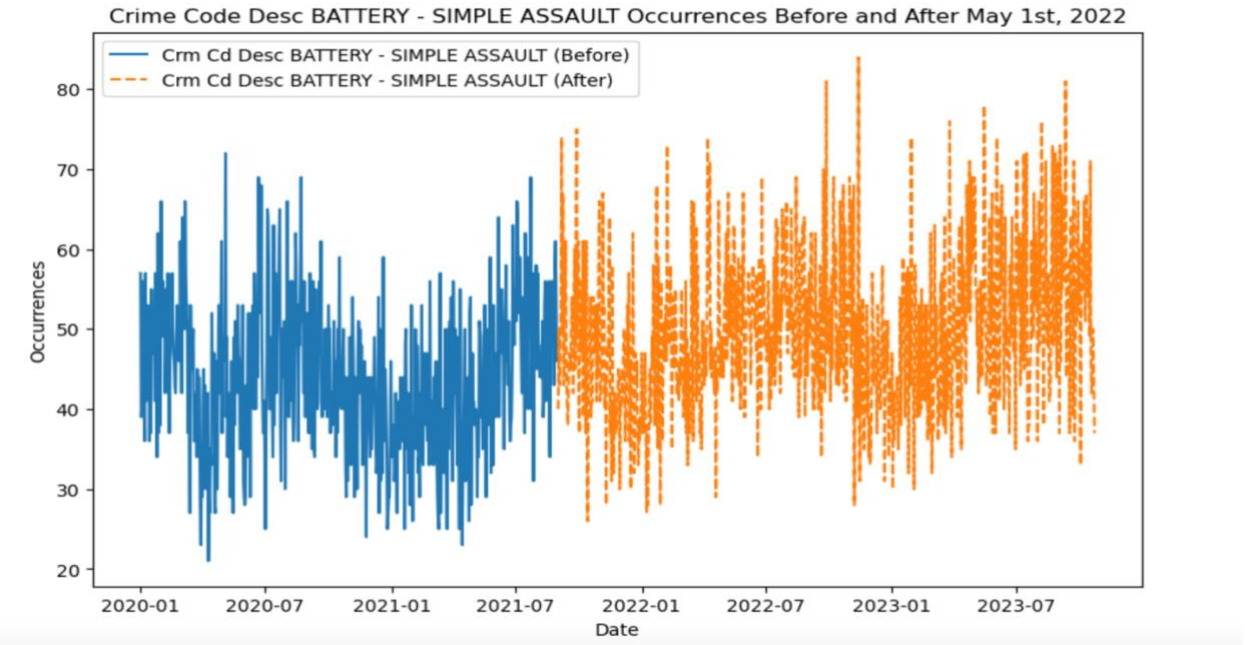


The chart titled "Crime Frequencies by Day of the Week" presents a comprehensive view of the distribution of crimes throughout a typical week. From the data, it's evident that Friday and Saturday witness the highest crime frequencies, surpassing the 110,000 mark, indicating a potential surge in criminal activities during weekends. However, the other days of the week—Wednesday, Monday, Thursday, Sunday, and Tuesday—show relatively consistent crime rates, each hovering around the 90,000 to 100,000 range. This uniformity suggests that aside from the weekend spike, criminal activities remain fairly stable throughout the weekdays. Such insights emphasize the need for heightened security measures, especially during weekends, to curtail these rising incidents.

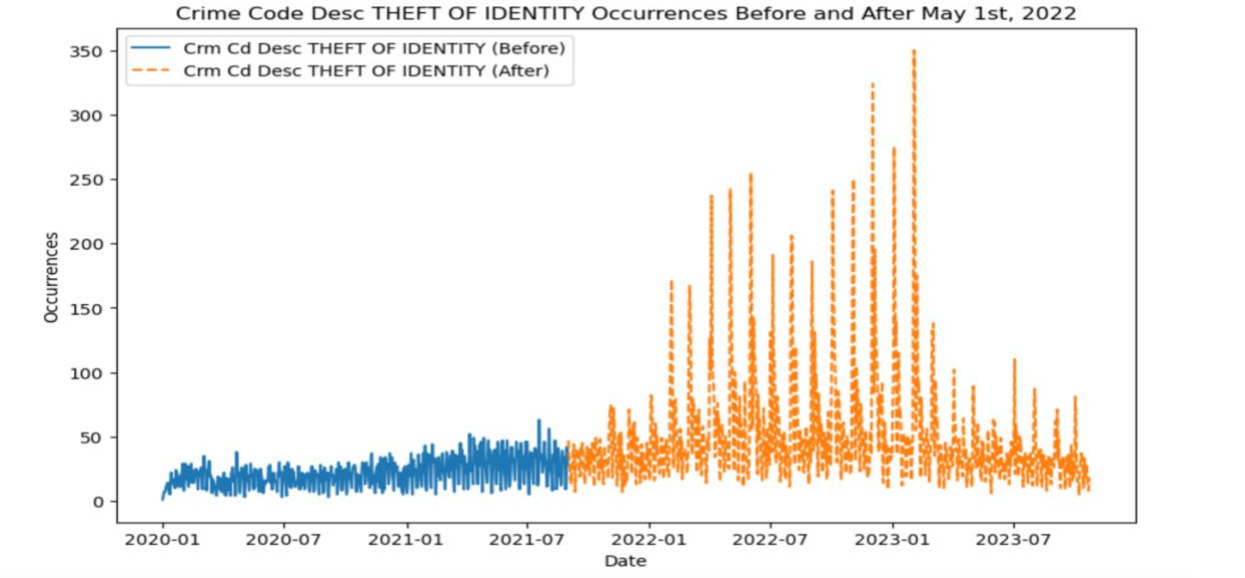
7. **Impact of Major Events:** Identify major events or policy changes during the dataset period and analyze crime rate changes before and after this event.



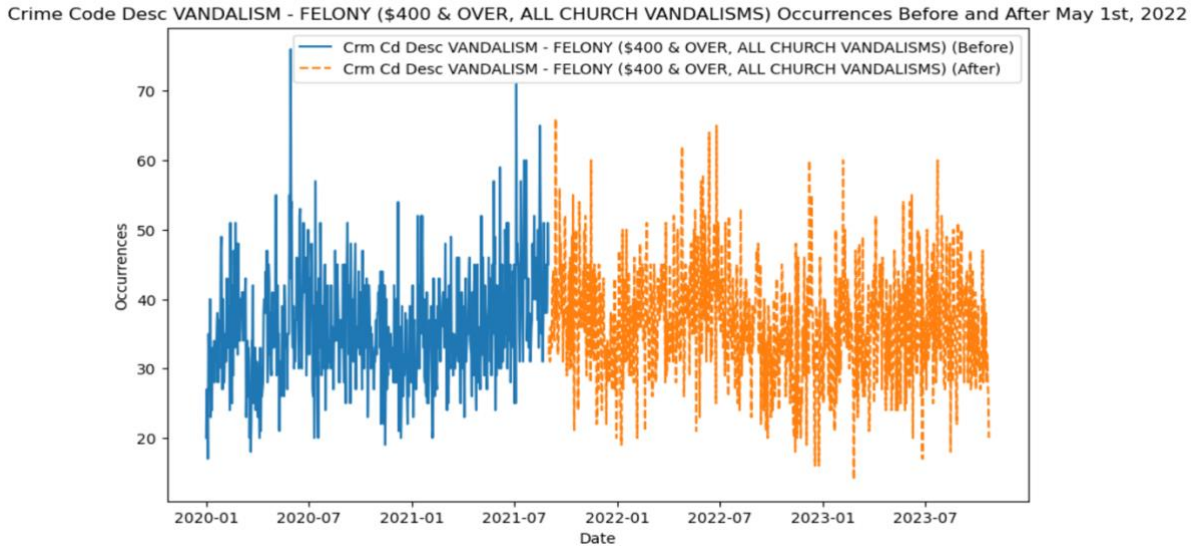
Burglary from Vehicle Trends: The graph illustrates the occurrences of "BURGLARY FROM VEHICLE" incidents before and after May 1st, 2022. Prior to May, as indicated by the blue line, occurrences fluctuated between roughly 10 to 70 daily incidents. However, after May 1st, represented by the orange dotted line, there's an observable rise in the frequency of such incidents, with numbers predominantly ranging from 30 to 70. The increase in the density of orange spikes after May suggests a higher consistency in burglary from vehicle occurrences during this period.



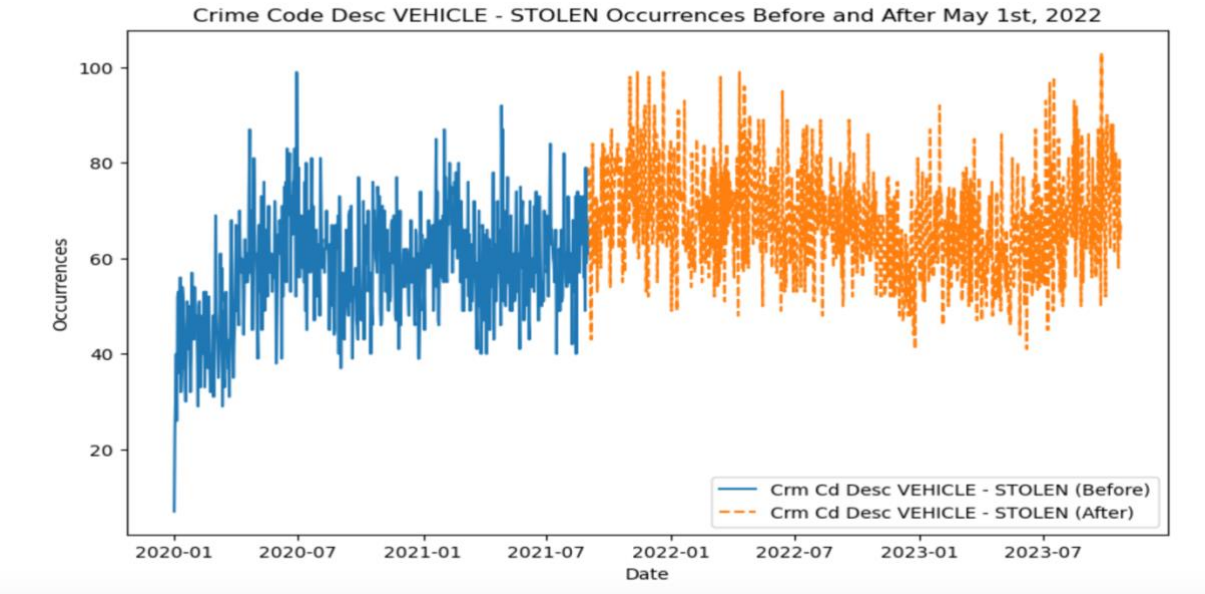
Simple assault: The graph shows the crime occurrences spanning from January 2020 to July 2023, a clear demarcation is observed before and after May 1st, 2022. Prior to this date, represented by the blue line, the occurrences fluctuated between approximately 20 to 80 daily incidents, with a relatively stable pattern. However, post-May 1st, 2022, delineated by the orange dashed line, there's a noticeable increase in the variability of occurrences, even though the range remains similar.



Theft of Identity Trends: The graph showcases the occurrences of "THEFT OF IDENTITY" crimes. The blue line represents the occurrences before May 1st, 2022, while the orange dotted line illustrates those after. Before May 1st, the daily incidents remain relatively low, oscillating below 50 occurrences. After this date, there's a noticeable increase with occurrences predominantly ranging from 150 to 350 daily incidents. The frequency of spikes post-May suggests a heightened and consistent pattern of identity.

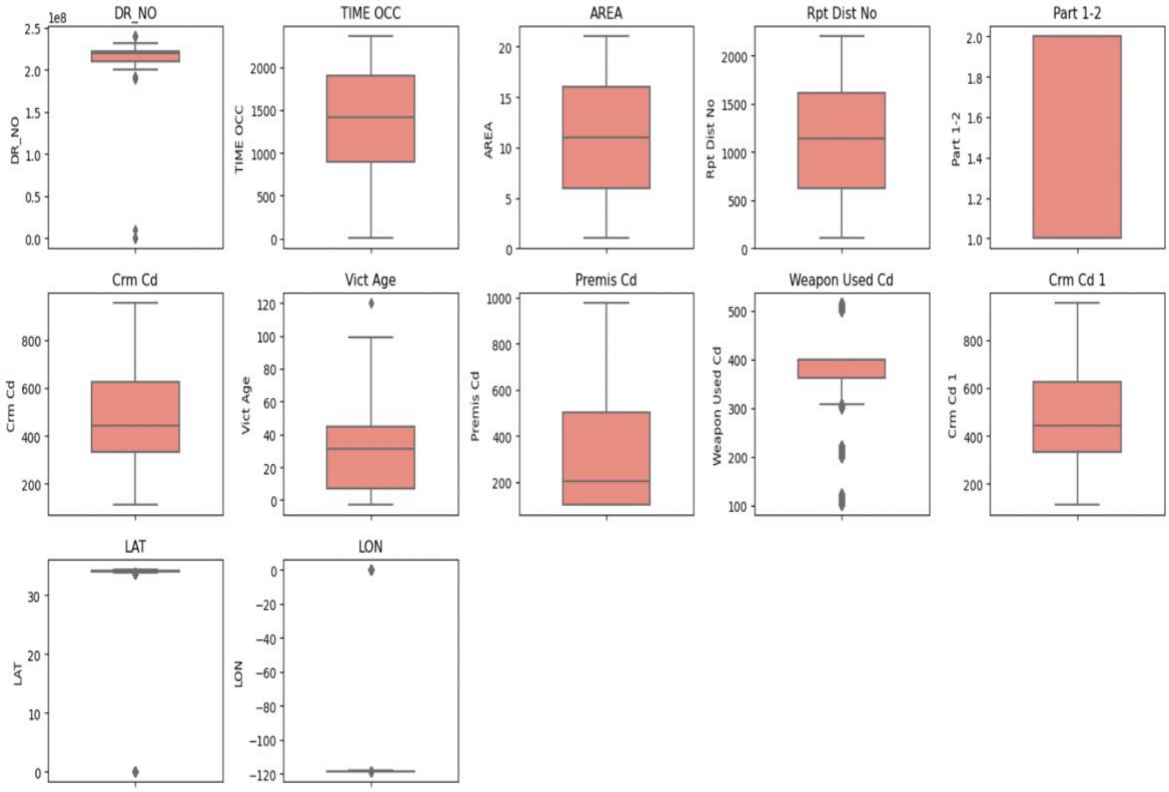


Vandalism in Churches: The above graph showcases the occurrences of felony vandalism specifically targeting churches. Similar to the first graph, the blue line shows the occurrences before May 1st, 2022, while the orange dotted line represents those after. Before May 1st, there are peaks and troughs, with an overall trend of declining occurrences nearing the cutoff date. After May 1st, the occurrences seem to have increased, but the frequency of spikes and drops appears to be more regular compared to the previous period.



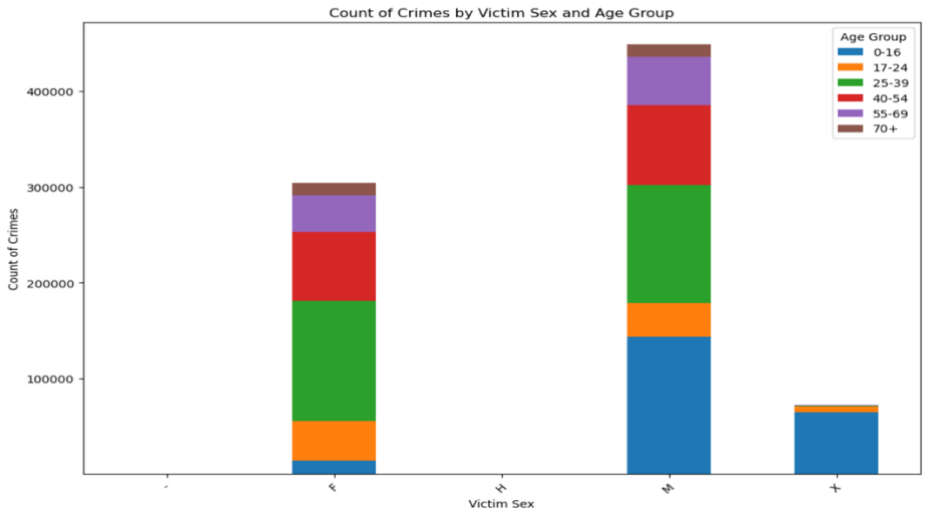
Stolen Vehicles: This graph illustrates the occurrences of vehicle thefts. The blue line indicates the occurrences before May 1st, 2022, and the orange dotted line represents occurrences after that date. It's evident that the frequency of this type of crime was somewhat steady before May 1st, 2022, with occasional spikes. Post May 1st, there is a noticeable increase in the occurrences of stolen vehicles, although there are still some fluctuations.

8. Outliers and Anomalies: Use statistical methods or data visualization techniques to identify dataset outliers and investigate unusual patterns.



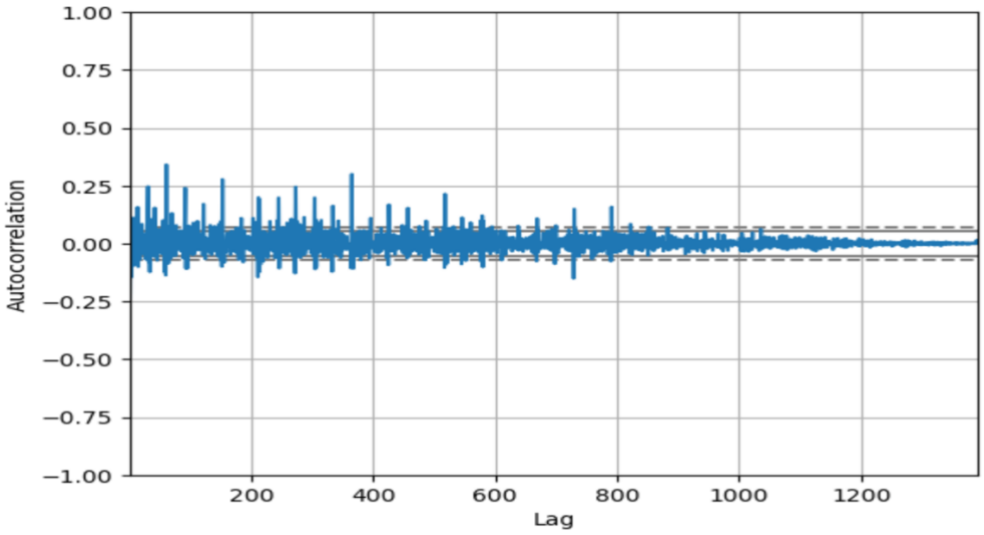
The collection of boxplots presented provides an overview of various crime-related data distributions. Key observations include a consistently distributed "AREA" and "TIME_OCC", while anomalies emerge in "DR_NO" with a noticeable outlier at its lower end and "LON" which displays an unusual negative outlier. "Vict Age" reveals an outlier towards higher age values, and both "Crm Cd" and "Crm Cd 1" showcase outliers predominantly in the upper ranges. "Weapon Used Cd" and "Rpt Dist No" also have specific data points that extend beyond their typical data range. These graphical representations highlight the need for a closer examination of the outliers to ensure data integrity and to discern potential patterns or insights within the crime dataset.

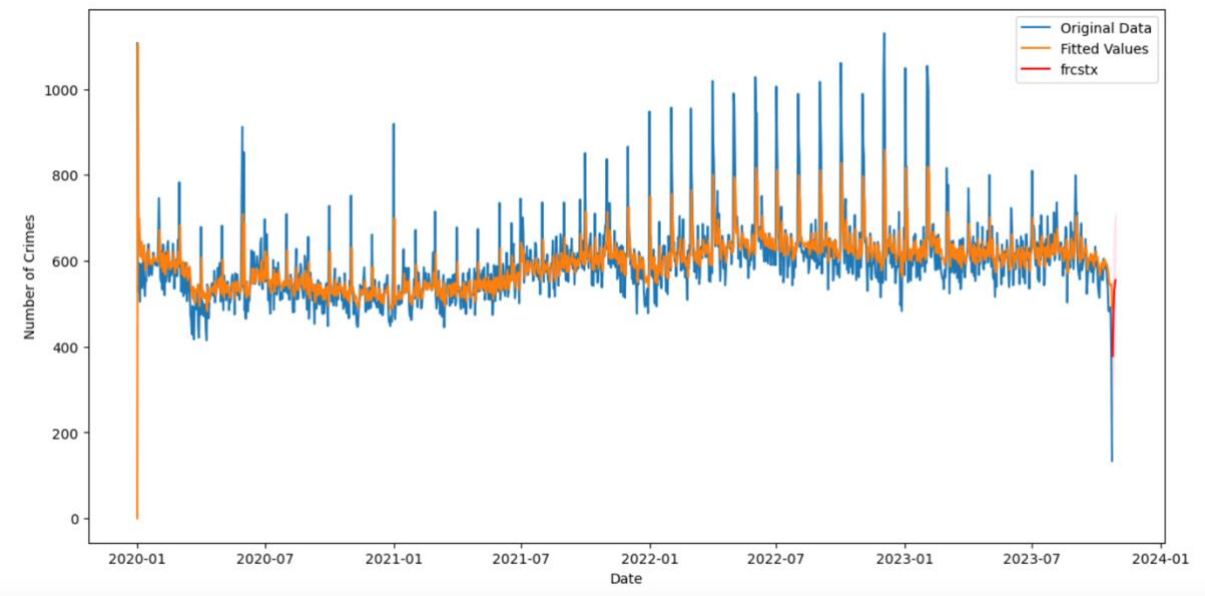
9. Demographic Factors: Analyse the dataset to identify any patterns or correlations between demographic factors (e.g., age, gender).



In the presented bar chart detailing the count of crimes by victim's sex and age group, it is evident that males (denoted as "M") are victimized at a substantially higher rate, especially within the 0-16 and 25-39 age brackets. The second category, potentially representing females, displays significant victimization primarily within the 25-39 and 17-24 age groups, with a decreasing trend as age advances. The third category, possibly indicating non-binary or unspecified genders, showcases a considerably lower overall crime rate, with a spike in the 55-69 age range. This data underscores the imperative to delve deeper into the nature and context of these crimes to derive comprehensive insights and strategies.

10. Predicting Future Trends: Employ time series forecasting methods, such as ARIMA or Prophet, to predict future crime trends based on historical data. Consider incorporating relevant external factors into your models.





The graph delineates the temporal progression of crime rates from January 2020 to January 2024. The original data, depicted in blue, exhibits noticeable fluctuations with periodic peaks. An overlaid fitted values line in orange demonstrates a smoothing of this data, capturing the general trend over time. The forecasted values, represented by the red line, indicate a downward trajectory in crime rates towards the latter half of 2023, culminating in a pronounced dip as 2024 begins. This information suggests a potential decline in criminal activities in the upcoming period, which can be pivotal for law enforcement and policy-makers in strategizing preventive measures and resource allocation.