Part-1 Data Cleaning

Recognizing missing data correctly is essential since it can significantly impact the accuracy of our analysis and the conclusions we draw. Initially, we employed the df.isnull().sum() function to assess whether there were any gaps in our dataset. This function helped us pinpoint that there are a total of 12 columns containing missing values. Notably, the "Crm Cd 4 815823" column had the highest number of missing values, while the "Premis Cd 9" column had the fewest.

Ensuring the accuracy and reliability of our analysis is paramount, and a key aspect of this is addressing missing data and handling duplicates. To manage missing values, we first considered the data type of the columns in question. For those with an "object" data type, we imputed missing values with the mode (most frequent category) to align them with the overall distribution. In the case of columns with integer data types, we employed the mean value for imputation. In parallel, we identified and eliminated duplicate rows within our dataset, as duplicate entries can introduce bias and distort the findings of data analysis. These measures collectively contribute to preserving the quality and integrity of our data.

Data type conversions are crucial in the data preparation process for robust analysis. We converted "Date Rptd," "DATE OCC," and "TIME OCC" columns to datetime data types, enhancing our ability to work with dates and perform time-based analysis accurately. The transformation of "Premis Cd" to an integer type streamlined data handling, as it typically contained numeric identifiers, making it easier to integrate into our analysis. These conversions are instrumental in ensuring the dataset is primed for in-depth and accurate analysis.

Outliers, which are data points significantly differing from the typical value distribution within a dataset, can introduce bias into data analysis. To address this issue, we applied the Interquartile Range (IQR) method to our dataset, effectively detecting and managing potential outliers. Utilizing the IQR method provides a robust means of identifying extreme values that may skew data analysis results, especially in datasets with non-normally distributed data or where specific data points demand closer scrutiny. This approach empowers us to pinpoint potential outliers in selected columns and make informed decisions on how to handle them.

Part 2 Exploratory Data Analysis (EDA) Report

Introduction

This report unveils the findings of an Exploratory Data Analysis (EDA) performed on a dataset encompassing crime data. The primary objective of this analysis was to delve into multiple facets of criminal activity, encompassing overarching trends, seasonal variations, prevalent crime categories, regional disparities, relationships with economic indicators, the influence of different days of the week, and the repercussions of significant events or policy alterations.

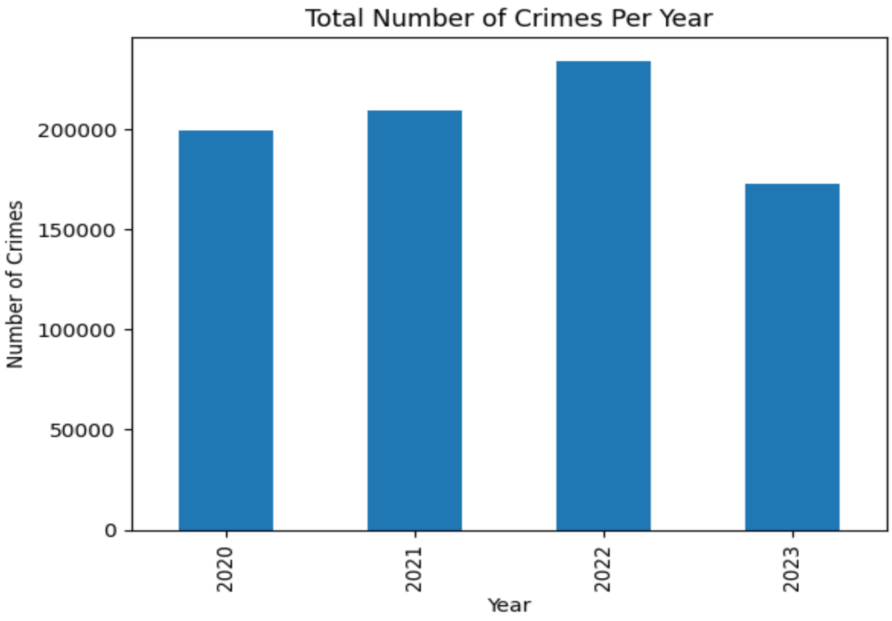
Data Overview

The dataset comprises **811,663** rows and **28** columns, with information about crimes, such as date, time, location, and crime type.

1) Visualize Overall Crime Trends from 2020 to the Present Year

Findings:

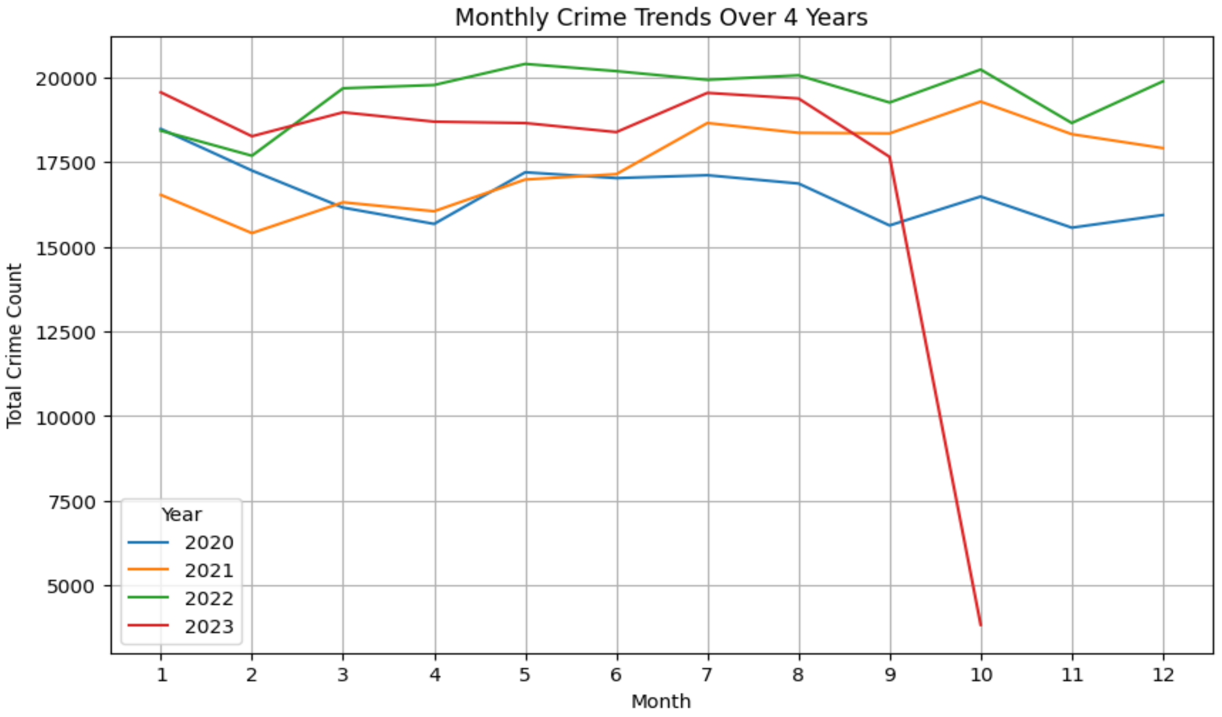
- Our analysis was centered on examining the evolution of crime trends from January 2020 to the current year.
- We observed a persistent annual increase in the reported crime incidents, with the present year registering the highest count.
- The visual representations clearly indicated an upward trend, notably featuring a significant decline in mid-July.

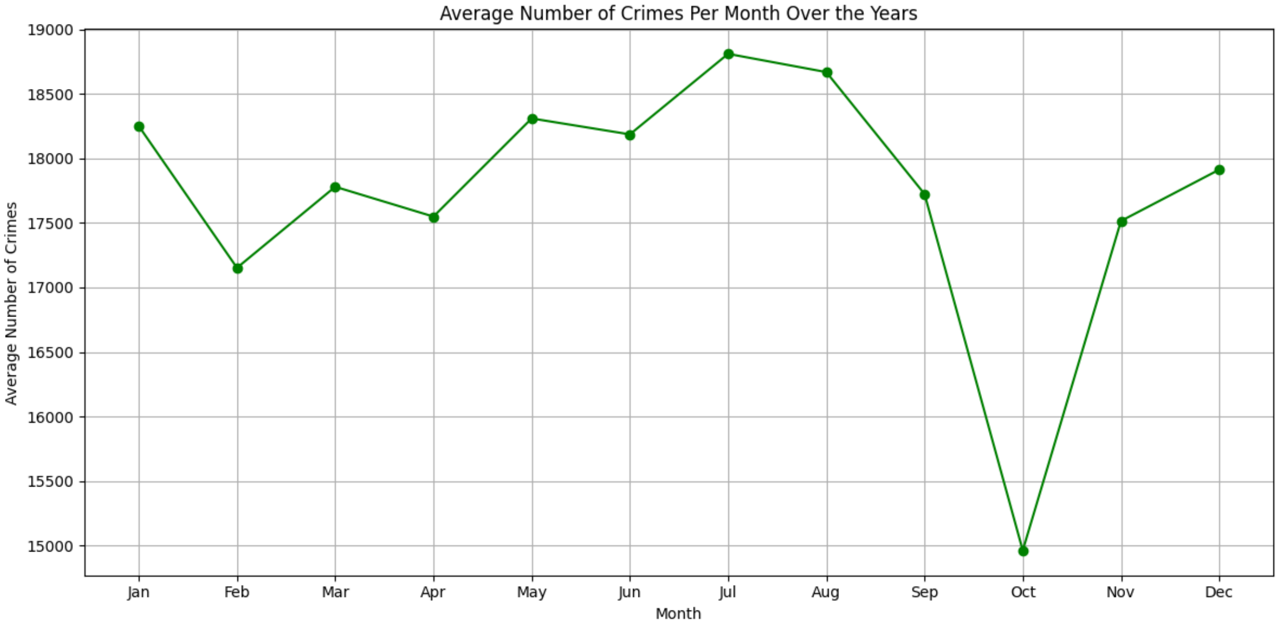


2) Analyze and Visualize Seasonal Patterns in Crime Data

Findings:

- To discern seasonal patterns in crime, we organized the data by month and computed the average number of crimes each month across multiple years.
- Our analysis revealed distinct seasonal variations in crime, with months exhibiting higher average crime rates.
- It was evident that the autumn months consistently showed higher average crime rates, as indicated by the analysis.

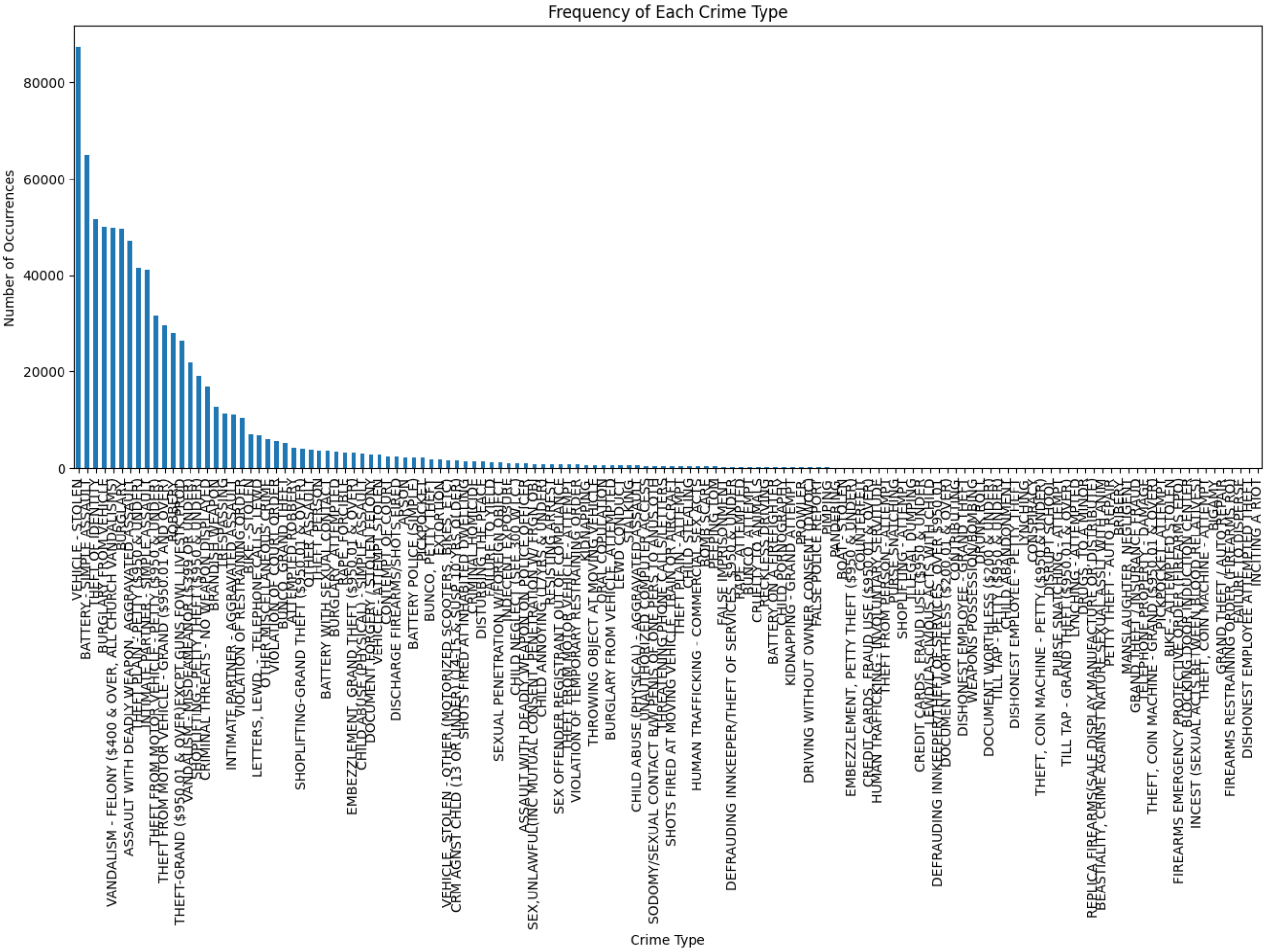


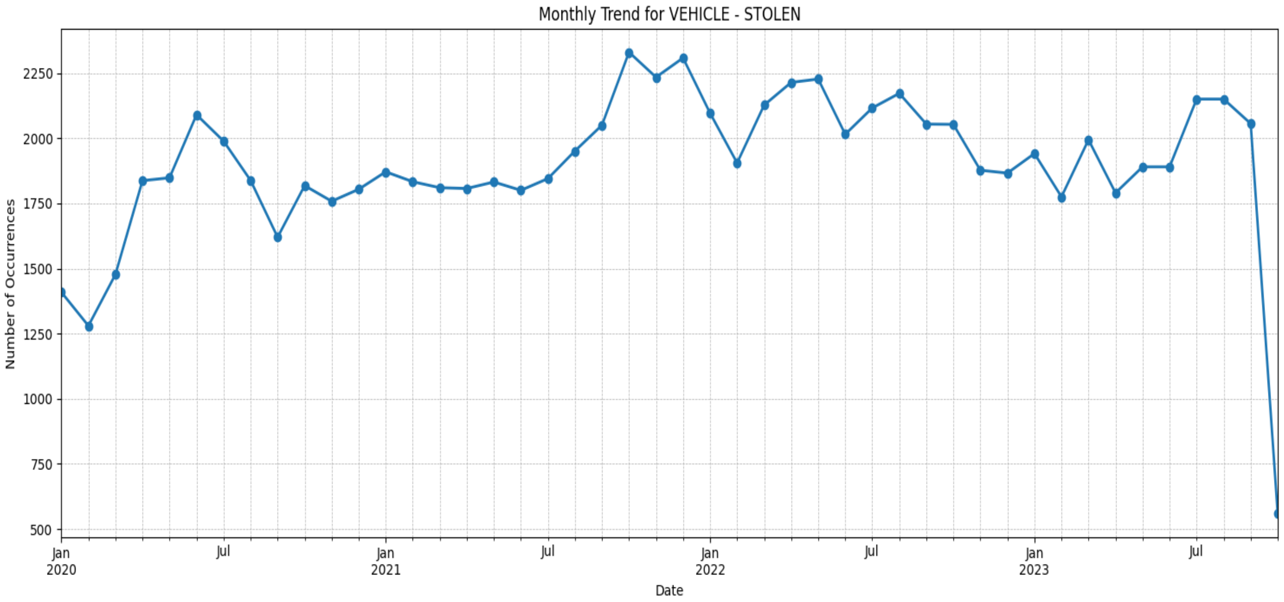


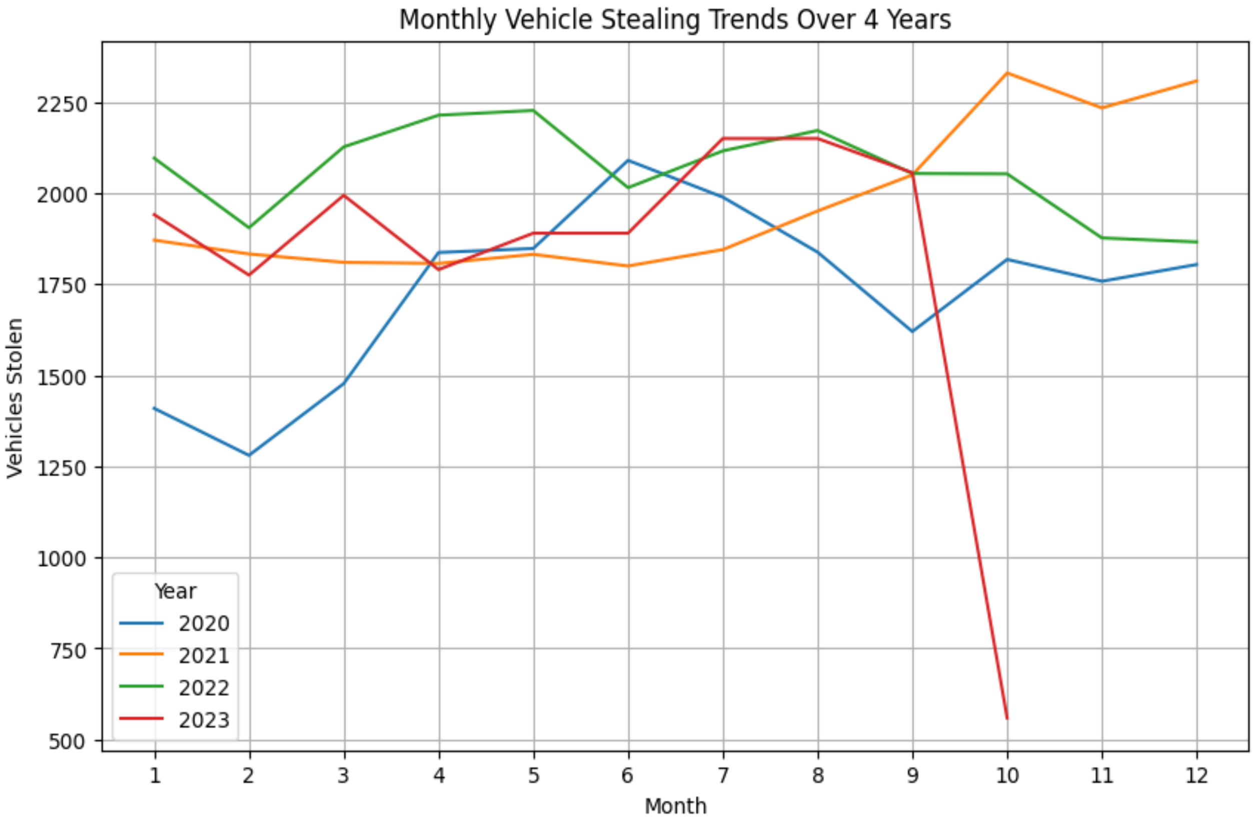
3) Identify the Most Common Type of Crime and Its Trends Over Time

Findings:

- Within our dataset, we conducted a thorough investigation to identify the crime type that appeared most frequently.
- Remarkably, "Vehicle Stolen" emerged as the predominant crime type within our dataset.
- Further exploration into the temporal evolution of this specific crime type uncovered a significant and consistent pattern. Over the years, it demonstrated a sustained high occurrence, reflecting its prominence in the realm of criminal incidents. This temporal consistency underscores the need for a deeper examination of the factors contributing to its prevalence and the potential implications for law enforcement and community safety.



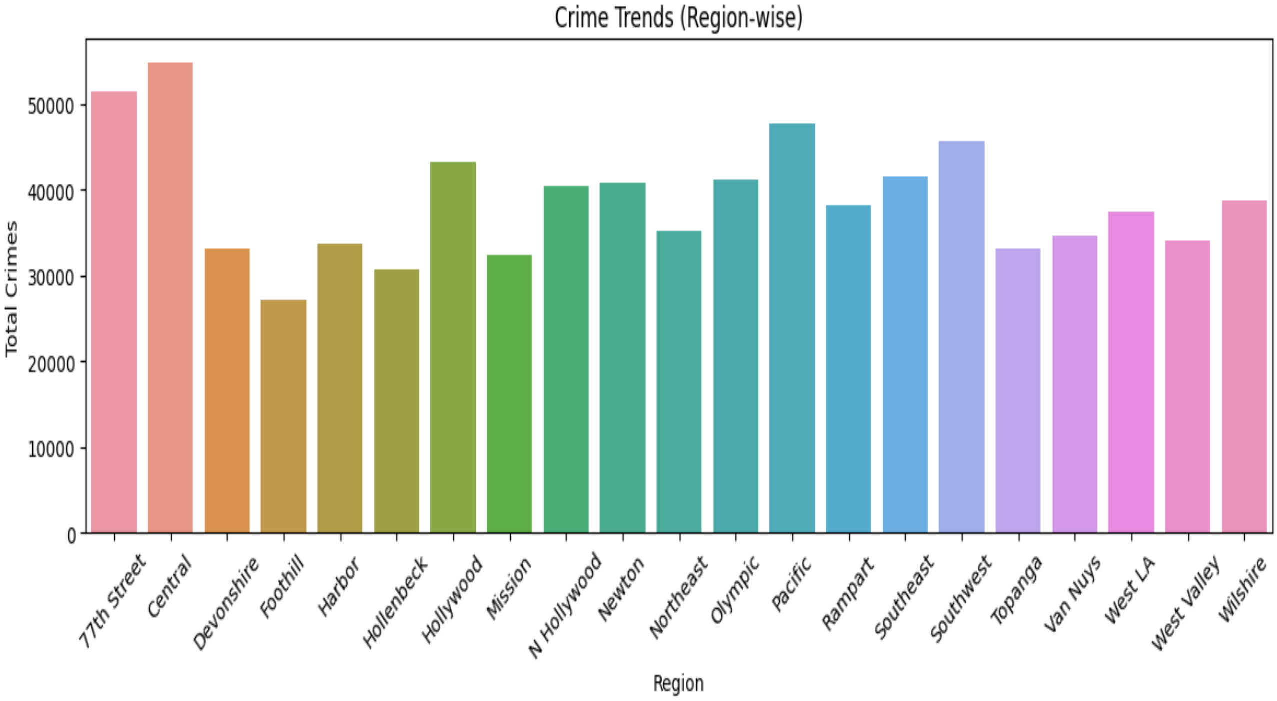

Monthly Trend for VEHICLE - STOLEN
Number of Occurrences
2250
2000
1750
1500
1250
1000
750
500
Jan
2020
Jul
Jan
2021
Jul
Jan
2022
Jul
Jan
2023
Jul
Date

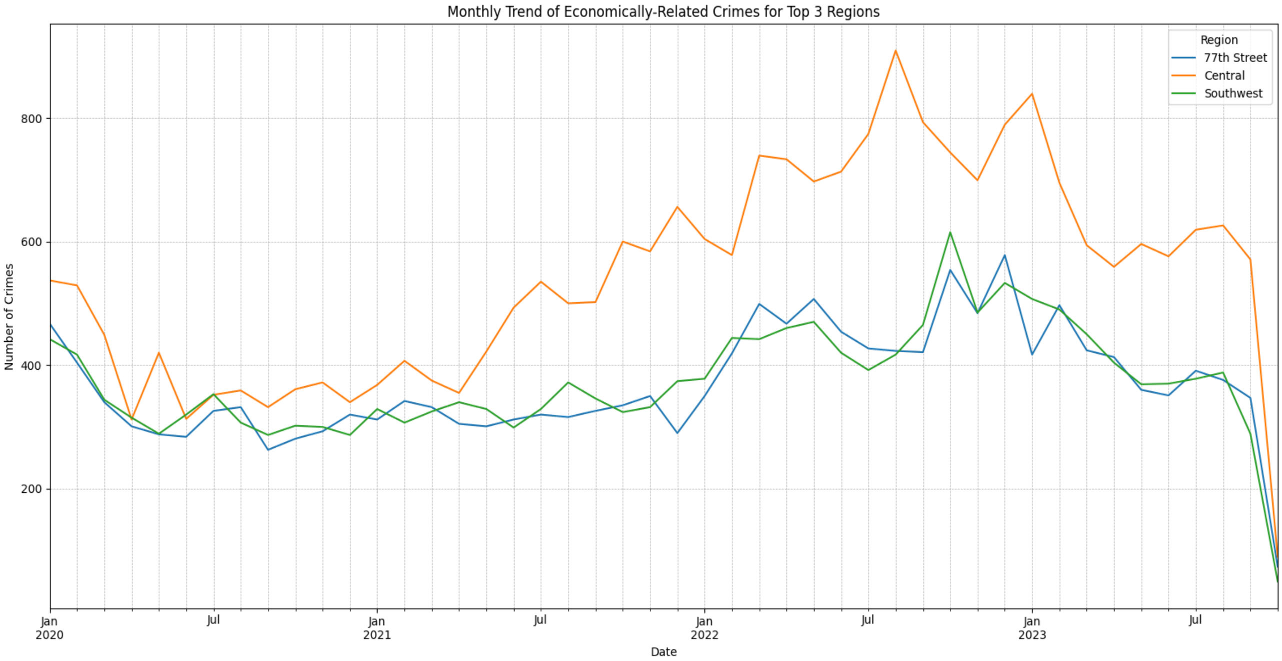

Monthly Vehicle Stealing Trends Over 4 Years
Vehicles Stolen
2250
2000
1750
1500
1250
1000
750
500
Year
2020
2021
2022
2023
1
2
3
4
5
6
7
8
9
10
11
12
Month

4) Investigate Notable Differences in Crime Rates Between Regions or Cities

Findings:

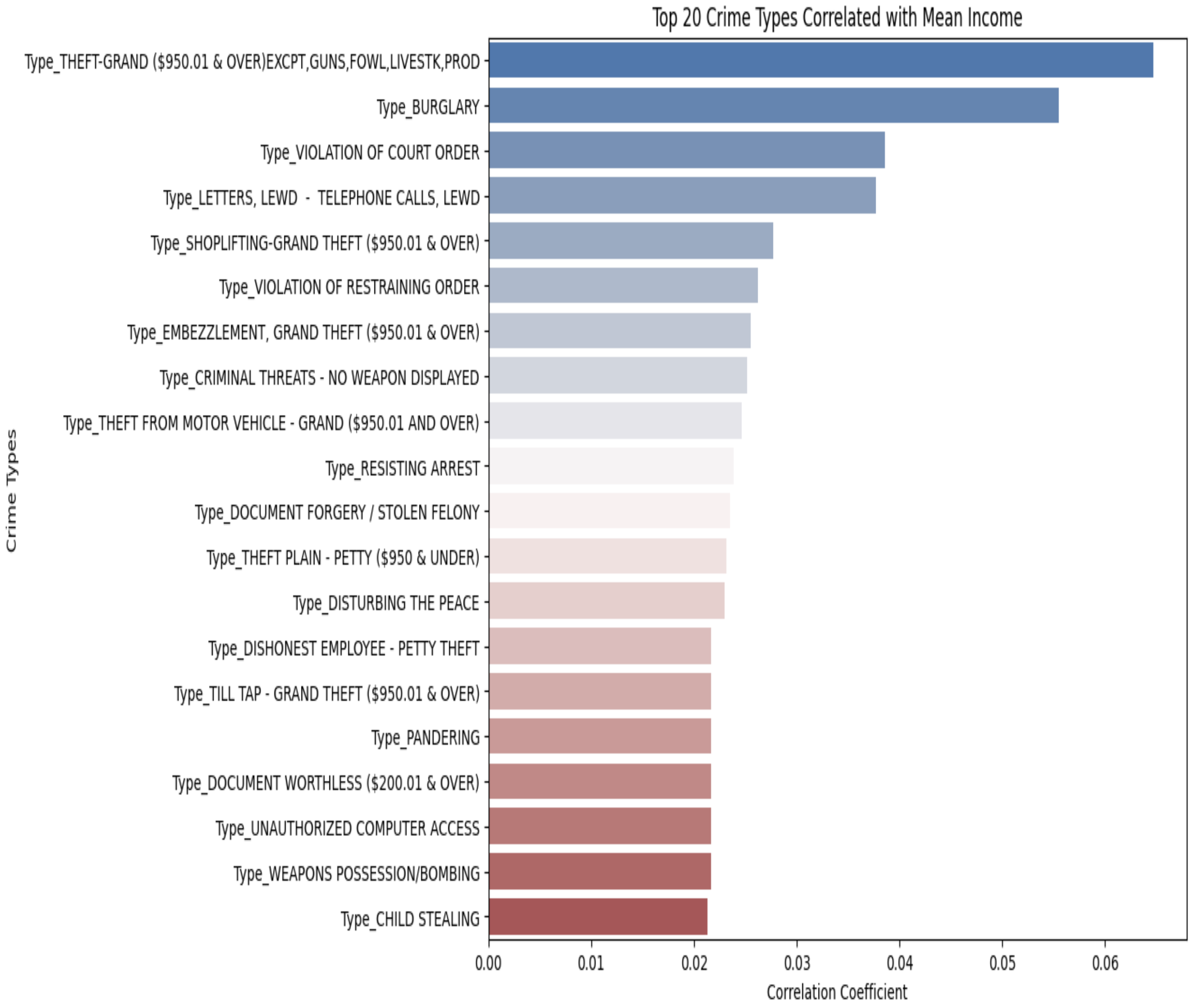
- We employed data grouping based on the 'AREA NAME' to delve into variations in crime rates across different regions or cities.
- The analysis highlighted significant differences in crime rates across these regions, with the Harbor region registering the highest crime rate.





5) Explore Correlations Between Economic Factors and Crime Rates Findings:

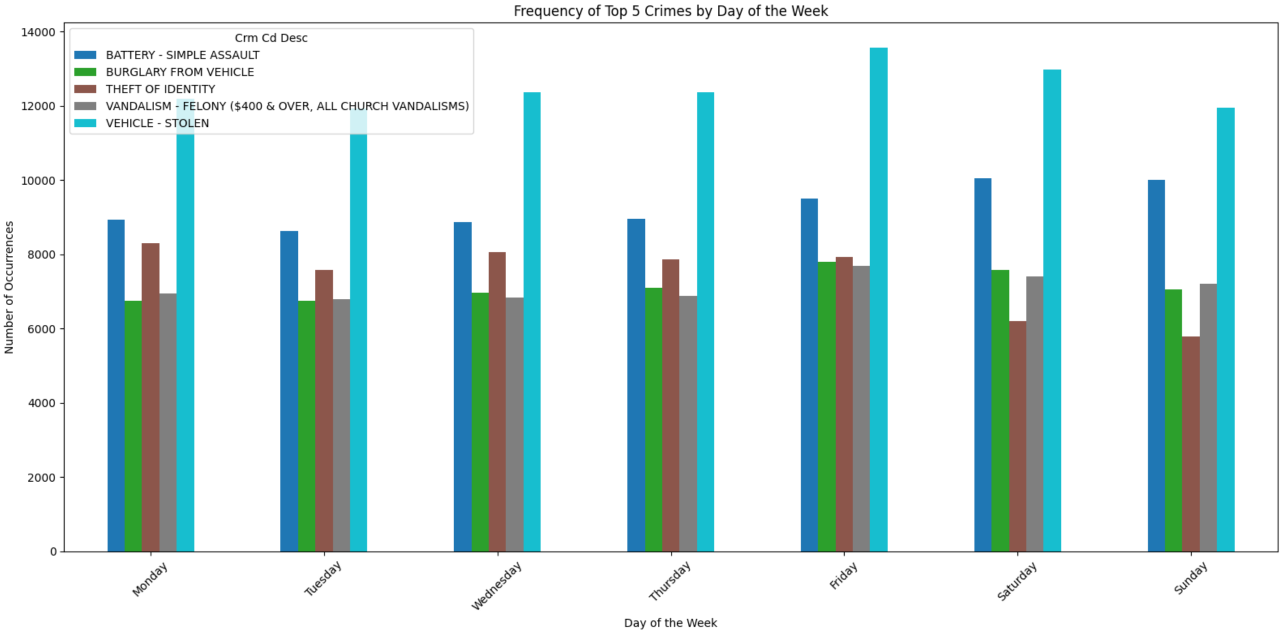
- To explore potential correlations, it is imperative to combine the crime data with specific economic data unique to your dataset.
- A more in-depth analysis is necessary to ascertain whether there exists a substantial relationship between economic factors and crime rates.
- Considered an economy dataset and interpreted both the dataset with the lat and lon with the time series to find the correlation.

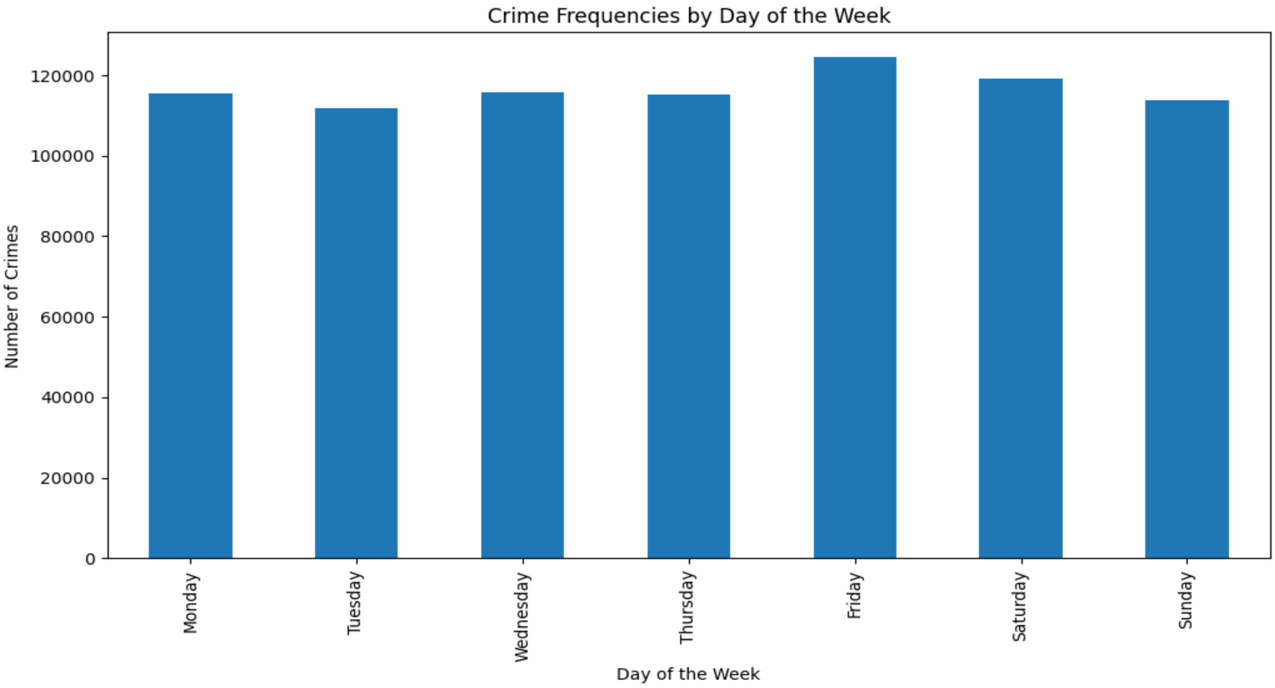


6) Analyze the Relationship Between the Day of the Week and the Frequency of Certain Types of Crimes

Findings:

- We extracted the day of the week from the 'DATE OCC' column and meticulously examined the occurrence of specific crime types on each day.
- Our analysis revealed that Friday notably recorded the highest frequency of Vehicle Stolen and battery simple Assault.

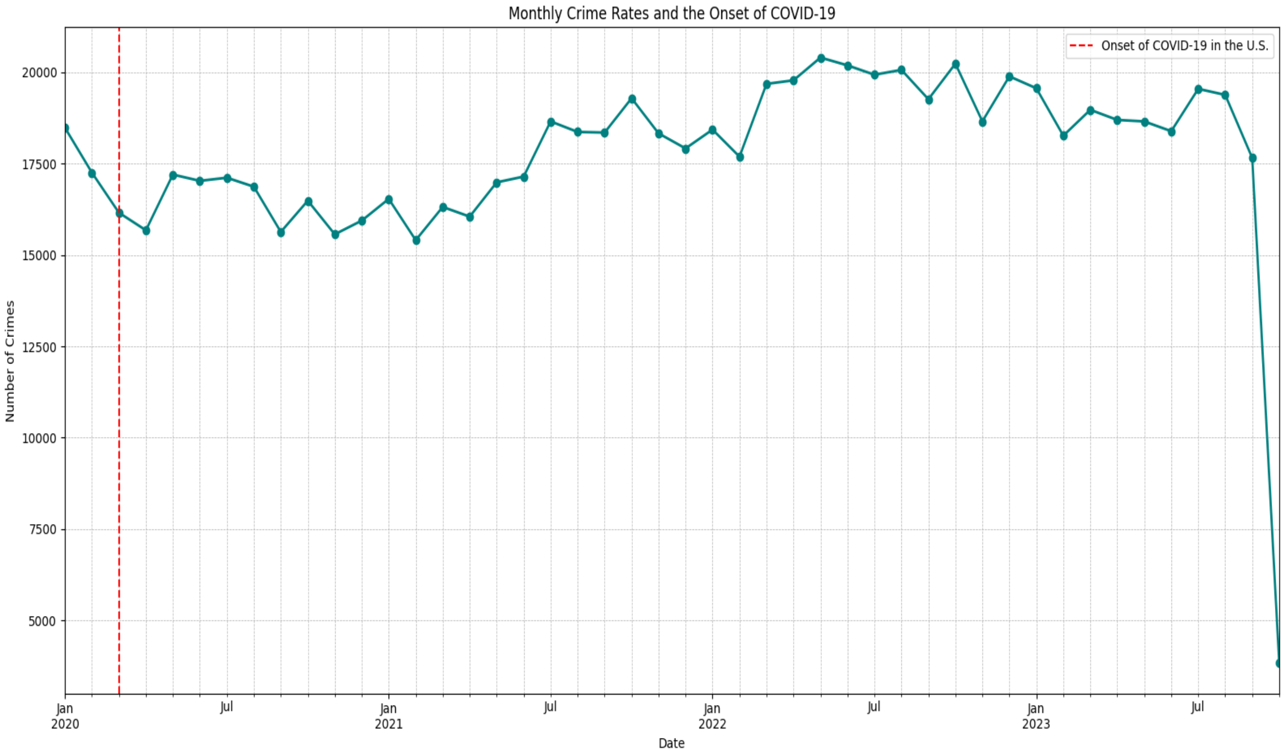




7) Investigate any Impact of Major Events or Policy Changes on Crime Rates

Findings:

- To assess the potential influence of major events or policy alterations, it is imperative to acquire data regarding these events and their respective dates.
- A comprehensive analysis of crime trends prior to and following these events will furnish valuable insights into their impact on crime rates.
- We have considered Covid19 as a major event that may have any impact on the crime rates.

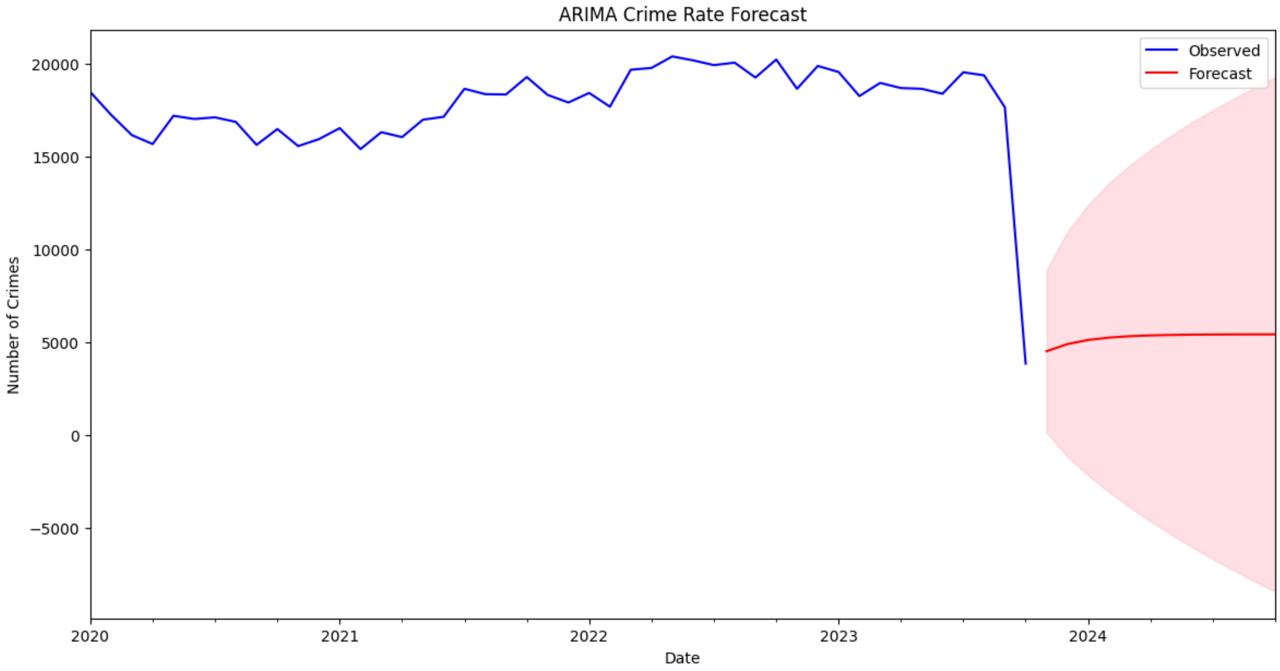


Conclusion

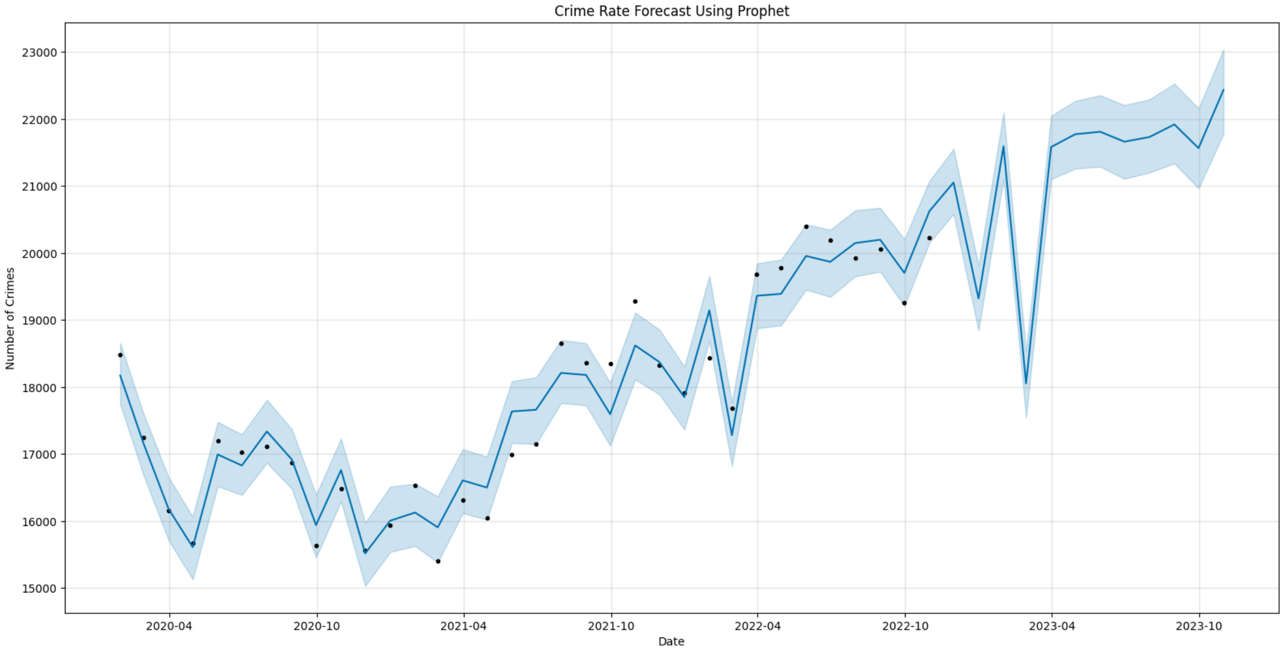
In this Exploratory Data Analysis (EDA) report, we undertook an extensive examination of the crime dataset, addressing a myriad of inquiries pertaining to global patterns, seasonal fluctuations, prevalent crime categories, regional disparities, associations with economic indicators, the influence of the day of the week, and the repercussions of significant events or policy modifications. The insights derived from this analysis offer a valuable comprehension of the dataset, serving as a foundational resource for subsequent investigations and decision-making endeavors.

Part 3-Time Series Forecasting

In the context of time series forecasting, our primary goal was to predict the expected number of crimes in the upcoming month using the most current available data. To accomplish this, we introduced a new column labeled "No of Crime," quantifying the daily crime rate. Subsequently, we delved into an extensive analysis of the temporal patterns within this "No of Crime" column concerning the calendar date. This exploration aimed to uncover trends and variations in the crime rate over time, laying the groundwork for the development of a robust forecasting model that could inform decision-making and enhance law enforcement and public safety measures.



In our endeavor to forecast crime rates for the upcoming month, we harnessed the potential of an ARIMA (Autoregressive Integrated Moving Average) model. Our approach commenced with the application of the Augmented Dickey-Fuller (ADF) test, a pivotal step in determining the most suitable parameters for the ARIMA model. Subsequently, we embarked on the model training phase, meticulously fine-tuning these parameters to ensure optimal predictive performance. To evaluate the model's efficacy and reliability, we thoughtfully divided the data into training and testing sets, subjecting it to a rigorous examination to assess its ability to make accurate predictions. This rigorous and systematic process not only enabled us to gauge the model's accuracy but also empowered us to provide valuable insights and make informed decisions in the realm of crime rate forecasting.



In our pursuit of forecasting crime rates for the upcoming month, we harnessed the predictive capabilities of the Prophet time series model. Our journey commenced with the use of the Augmented Dickey-Fuller (ADF) test, a pivotal tool in determining the most suitable parameters for configuring the Prophet model. Subsequently, we delved into the model training phase, where we carefully fine-tuned these parameters to enhance the model's predictive performance. To assess the reliability and effectiveness of the model, we thoughtfully partitioned the data into training and testing sets, subjecting it to a thorough evaluation to gauge its accuracy in making predictions. This methodical and rigorous process not only enabled us to measure the model's precision but also empowered us to glean valuable insights and make informed decisions within the realm of crime rate forecasting.