IE6400 Foundations Data Analytics Engineering

Fall Semester 2024

Project 1

Cleaning and Analyzing Crime Data

Group 17
Nithin Arumugam Raja
Harrish Ebi Francis Peter Joshua
Harish Padmanabhan

arumugamraja.n@northeastern.edu
peterjoshua.h@northeastern.edu
padmanabhan.h@northeastern.edu

Submission Date: 10/15/2024

Abstract:

The project focuses on the cleaning and analysis of real-world crime data from 2020 to the present date. The goal of this project is to find patterns and trends that has been occurring from 2020 in California. The dataset went through a series of data cleaning processes, addressing missing values, formatting issues and its readiness for analysis. Exploratory Data Analysis (EDA) was conducted to identify visualize overall crime trends by each year, seasonal patterns by each month, most common crime that has been occurring, crime rates by different areas and cities, crime by economic factors, crimes that occur by the day of the week and the frequency of certain types of crimes.

Report on Data Cleaning and Analysis:

1) Data Acquisition:

The dataset was downloaded and loaded using the Python 'pandas' library. It contains crime data from 2020 to the present, which includes various fields such as crime type, dates, time, locations, victim details etc.

2) Data Inspection:

Initial steps included displaying the first few rows to understand the structure and inspecting data types to ensure consistency.

- The dataset had over 980,000 records with 28 columns.
- Columns included categorical, numerical, and date fields that needed further processing for analysis.

3) Data Cleaning Process:

3.1 Handling Missing Data

We identified missing data in many columns, such as victim sex and locations. To handle the missing values, we did the below.

- Replaced invalid or unknown values in categorical columns.
- Removed rows where all values were missing.

3.2 Removing Duplicates

No duplicate rows were found in the dataset after checking.

3.3 Converting Data Types

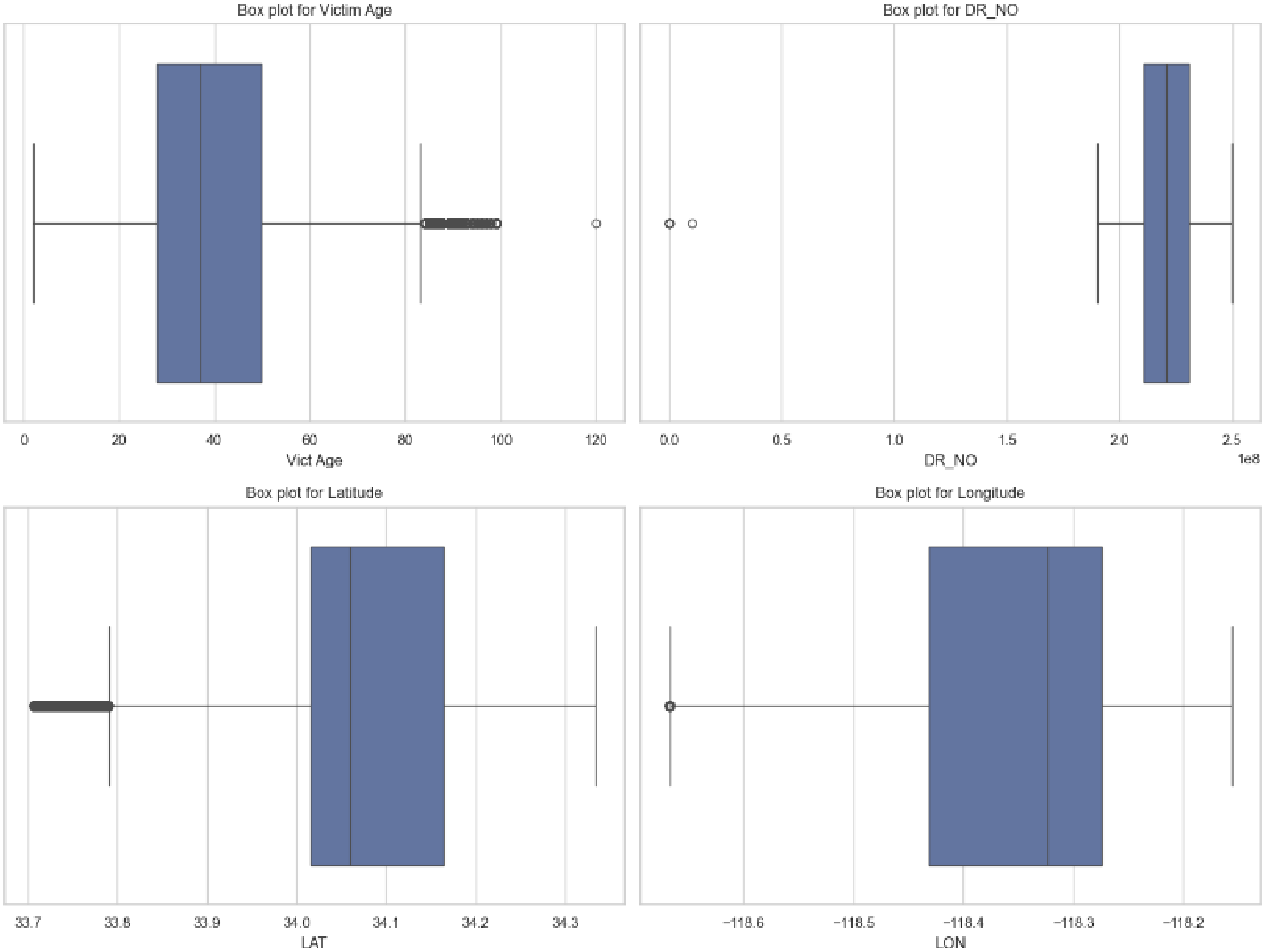
No Certain columns representing the dates and time were converted to appropriate formatting for analysis.

- Date Rptd and DATE OCC were converted to datetime objects.
- TIME OCC was formatted as HH

3.4 Handling Outliers

There were certain columns which had outliers that were identified and replaced with NaN for any such outliers for relevant values.

- Vict Age which had values less than or equal to 0 were replaced with Nan



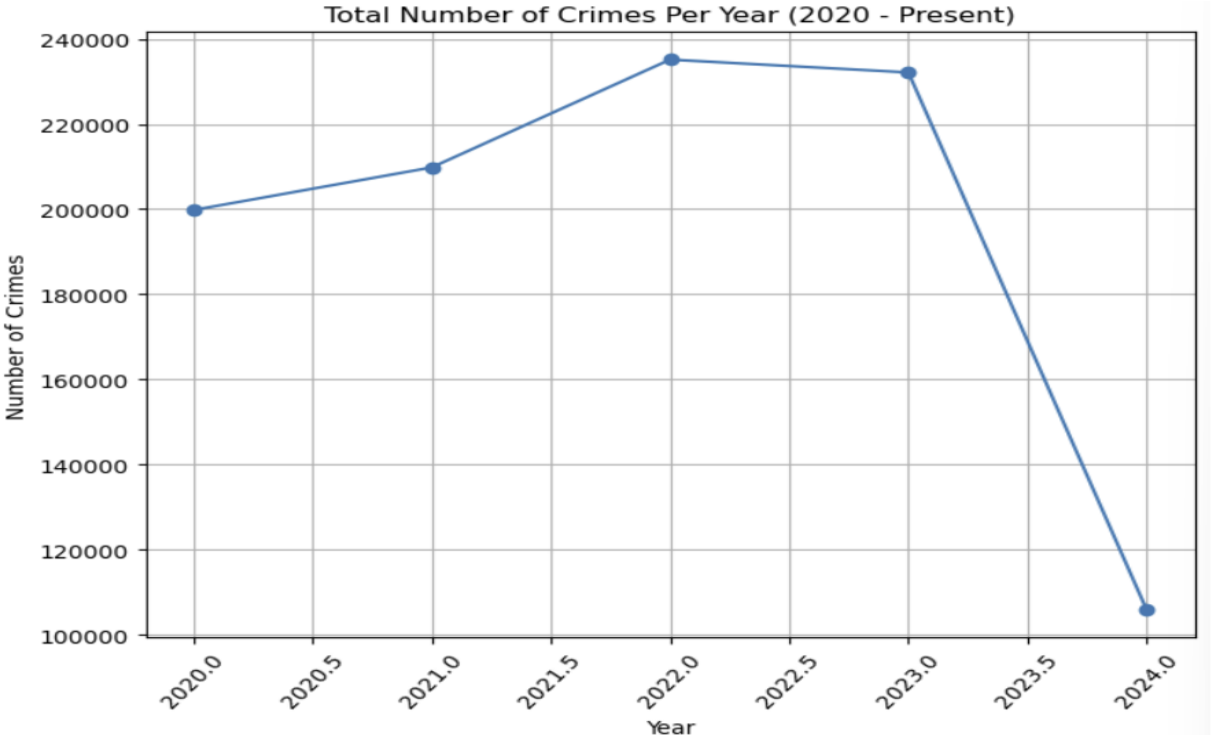
3.5 Normalization and Standardization

Numerical columns like Rpt Dist No and Premis Cd were standardized and normalized as needed to ensure uniformity in scaling.

4) Exploratory Data Analysis (EDA):

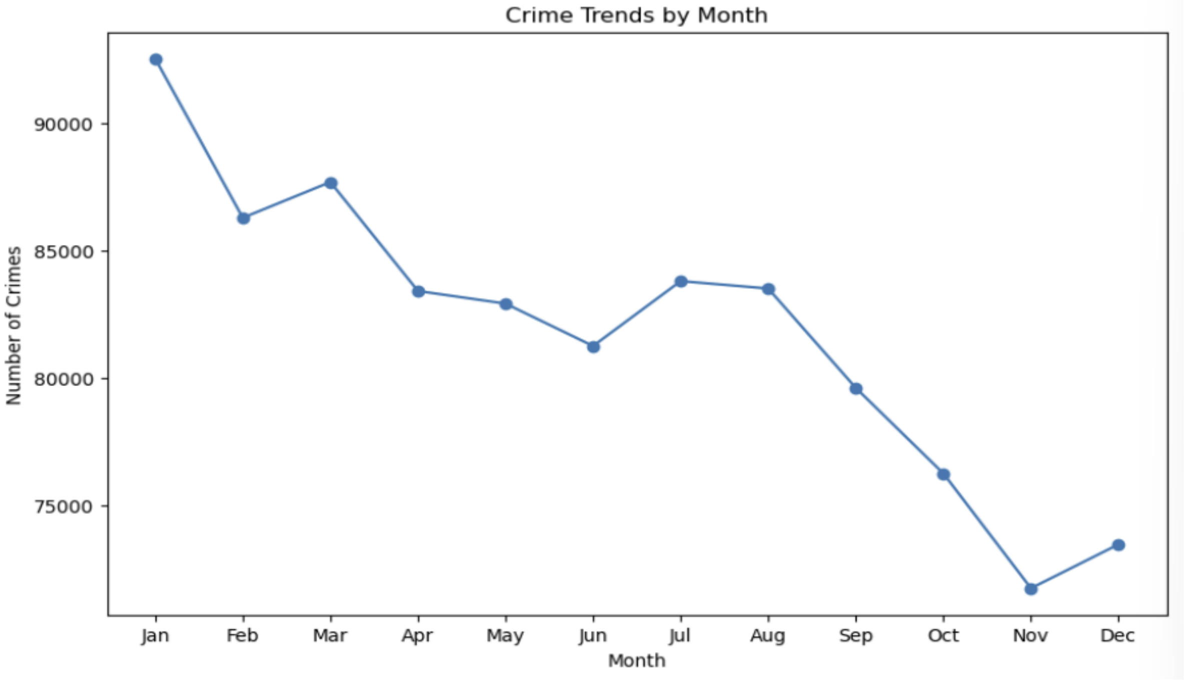
4.1 Crime Trends Over Time:

This analysis revealed crime fluctuations in crime over the years.



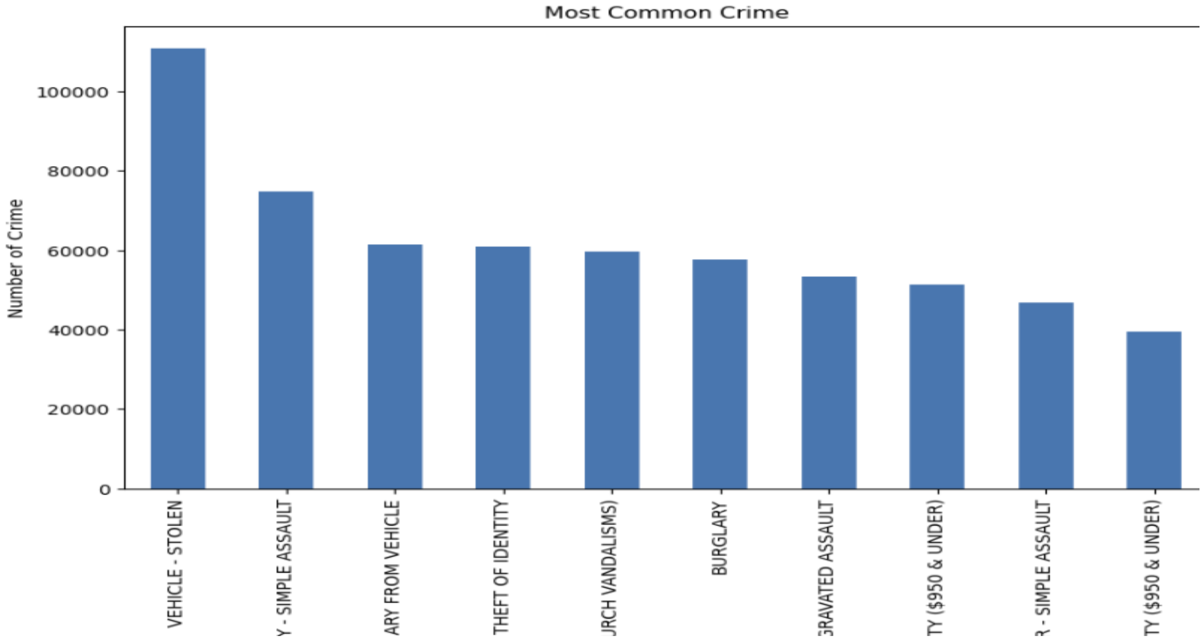
4.2 Seasonal Crime Patterns:

This analysis showed us that crime rates tend to fluctuate throughout the year, with the peak being observed during the month of January.

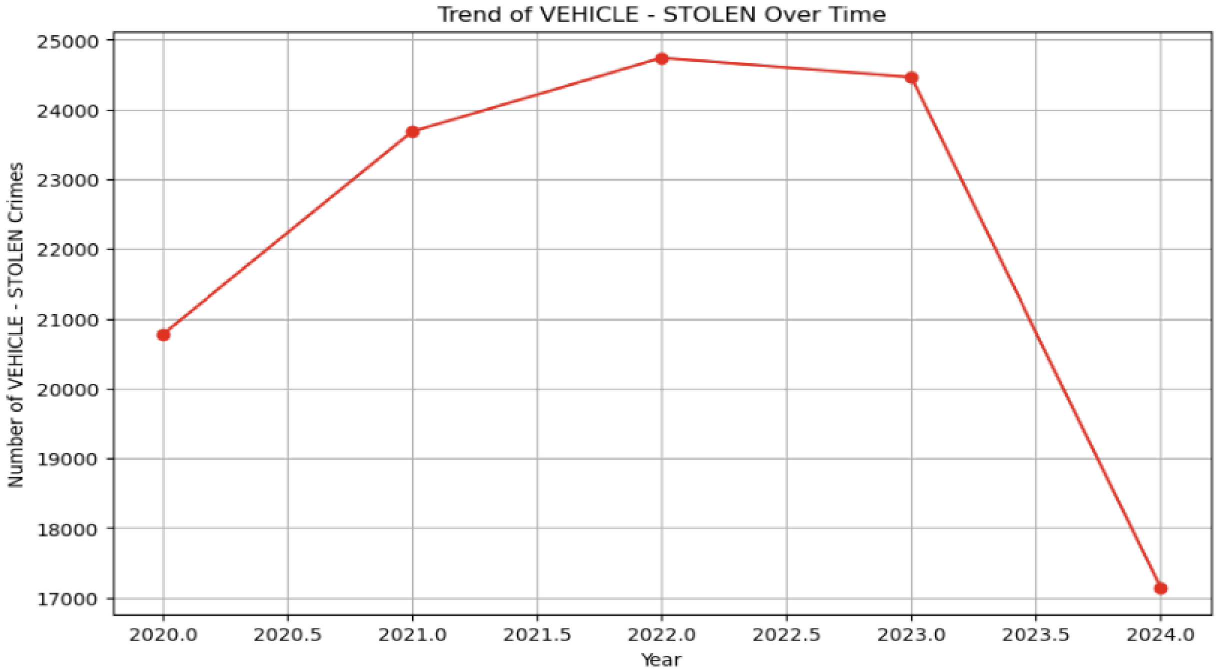


4.3 Most Common Types of Crimes:

Most common crime - This analysis showed us the top 10 crime that occurs most frequently in the dataset, the most common was ‘Vehicle - Stolen’ followed by simple assaults and burglaries.

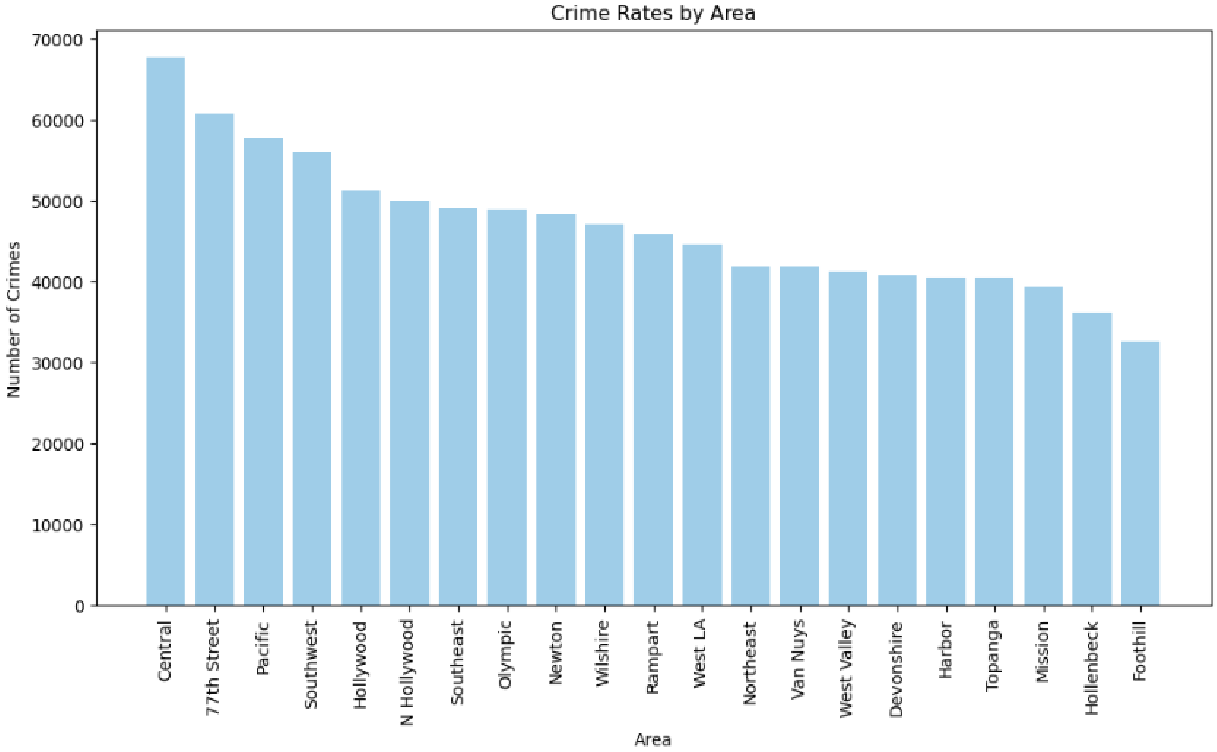


Trends in Vehicle Stolen - This analysis showed us the Vehicle Stolen crime over the years showed a steady pattern.



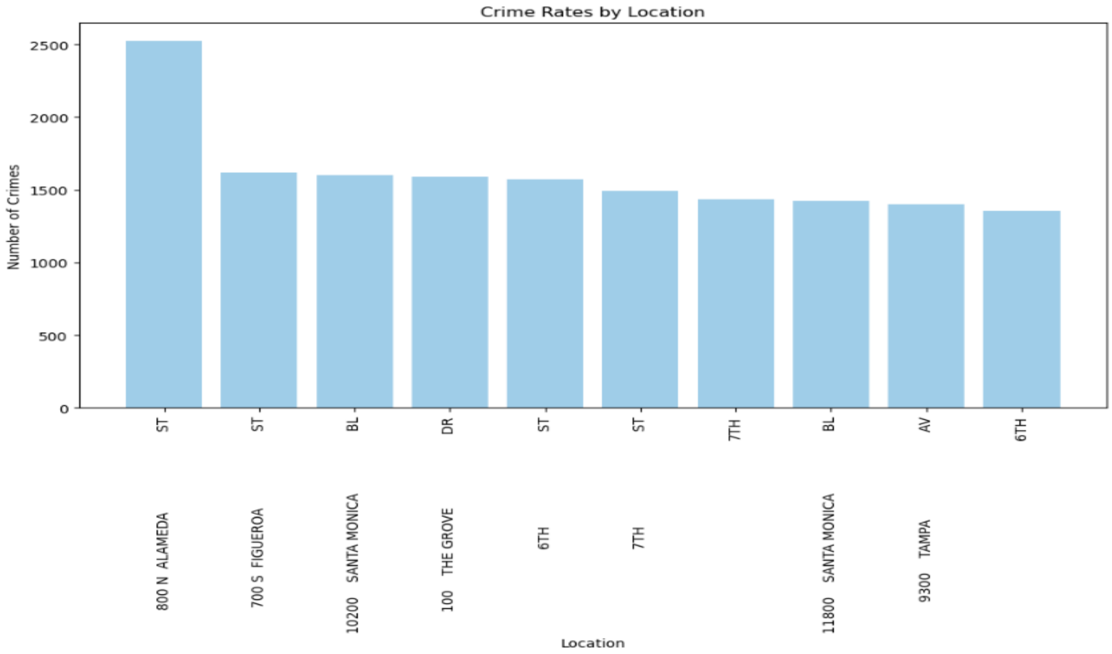
4.4 Crime Rates by Region

This analysis showed in which area the most number of crime occurs, in areas like Wilshire and Central had higher crime rates.



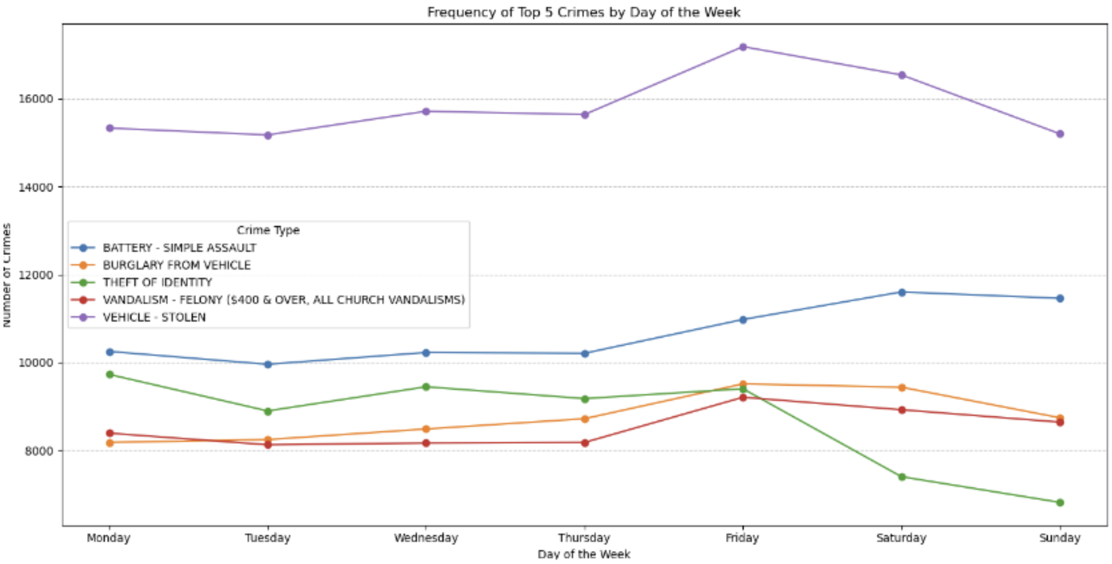
4.5 Crime Rates by Location:

This analysis showed in which location (i.e) streets were there is a higher concentration of crimes occuring.



4.6 Crime and Days of the week:

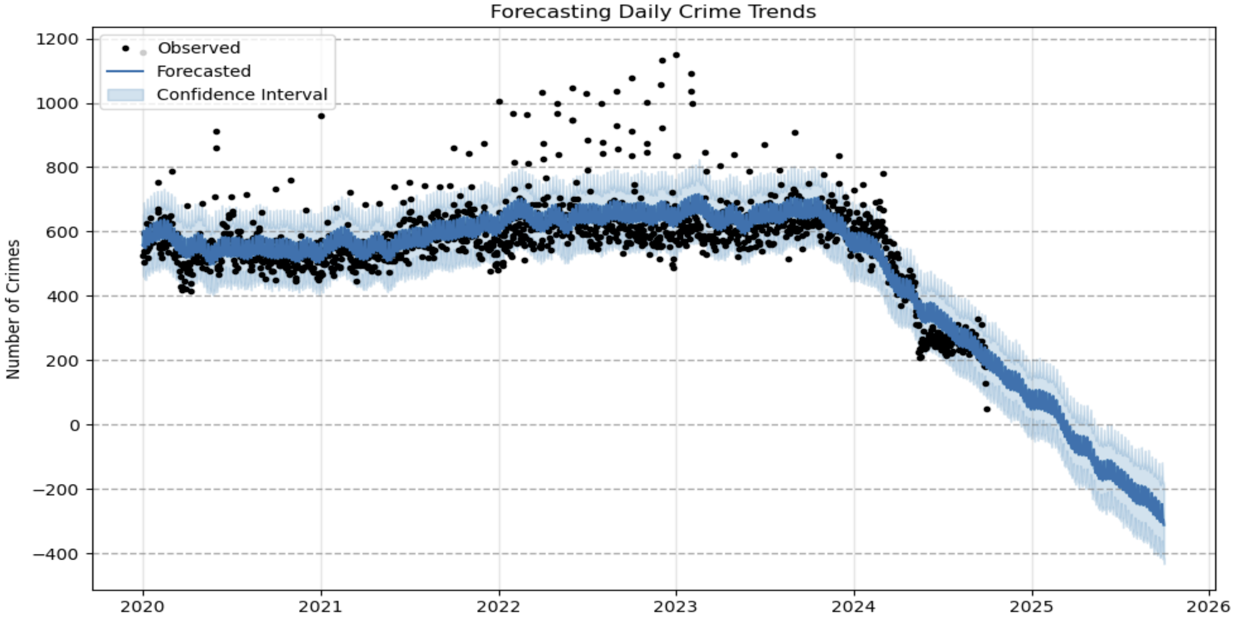
This analysis showed the most types of crime and the frequency of them occuring on day of the week. Weekends (Friday and Saturday) had higher occurrences of top crimes like Vehicle Stolen and Battery Simple Assault.



5) Advanced Analysis:

5.1 Use predictive modeling techniques:

This analysis showed us the future trends of the crime rate, based on the existing crime records. The model predicts that the number of daily crimes will significantly decrease, even dropping into negative values by 2025 – 2026.



6) Key Insights:

a. Vehicle-related crimes are a significant concern, constituting the highest volume of criminal activity.

b. Geographic focus: Certain areas, like Wilshire and Central, consistently report higher crime rates, and targeted interventions could reduce crime in these hotspots.

c. Weekly patterns: Crimes occur more frequently on weekends, particularly vehicle-related offenses.

d. The model predicts that the number of daily crimes will significantly decrease, even dropping into negative values by 2025 – 2026.

7) Conclusion

The data reveals trends and patterns that can inform law enforcement strategies. By focusing on high-crime areas, understanding seasonal fluctuations, and targeting specific types of offenses (e.g., vehicle theft), authorities can implement more effective crime prevention measures.