IE6400 Foundations Data Analytics Engineering
Fall Semester 2024
Project 1

Project Overview

This project focuses on cleaning and analyzing a crime dataset from 2020 to the present. The primary objectives include preparing the dataset, performing exploratory data analysis (EDA), and investigating trends and factors influencing crime rates. Key steps involved data acquisition, inspection, and cleaning by addressing missing values, duplicates, and outliers. Following the cleaning process, various visualizations were created to analyze crime trends over time, across regions, and by crime type. Advanced analyses, such as predictive modeling, were also considered. The findings are summarized in this report, supported by visualizations, descriptive statistics, and insights into crime patterns and factors.

Objectives

The primary objectives of this project are:

1. *Data Preparation*: Clean and preprocess the crime dataset from 2020 to the present by handling missing values, duplicates, and outliers, ensuring data quality for analysis.
2. *Exploratory Data Analysis (EDA)**:*** Analyze crime trends, patterns, and factors influencing crime rates through visualizations and statistical summaries, focusing on time, regional, and categorical variations.
3. *Trend and Pattern Identification*: Investigate relationships between crime and other factors, such as day of the week, regional differences, economic conditions, and significant events or policies.
4. *Advanced Analysis*: Optionally explore predictive modeling to forecast future crime trends and address additional hypotheses related to the dataset.

Data Preparation

The crime dataset from 2020 to the present was loaded and inspected for preliminary analysis. The following steps were taken to prepare the data for deeper exploration:

1. Data Loading and Inspection: The dataset was loaded into a Pandas DataFrame, and the first few rows were printed for initial review to understand the structure and content of the data.

```
data = pd.read_csv("Crime_Data_from_2020_to_Present.csv")
dataframe = pd.DataFrame(data)
first_data = dataframe.head()
print(first_data)
```

2. Column Selection and Data Types: A subset of columns, specifically the first 29 columns, was selected for further analysis. The data types of each column were checked to identify any inconsistencies or necessary transformations.

```
df_selected_range = dataframe.iloc[:, 0:29]
print(df_selected_range.dtypes)
for columns in df_selected_range:
    print(columns)
```

3. Handling Missing Values: The dataset was examined for missing values, and rows with all missing data were removed. This ensured that the data was clean and ready for analysis without unnecessary gaps.

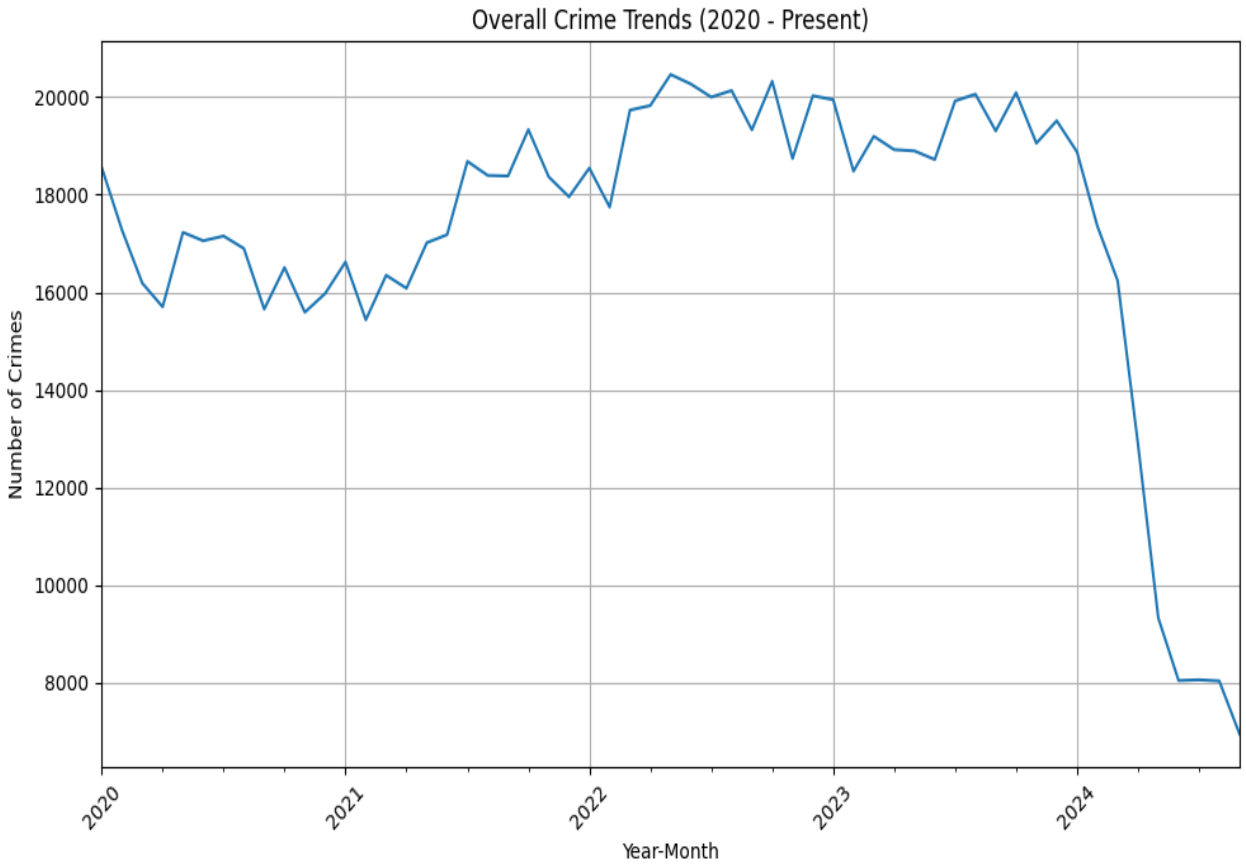
```
missing_data = dataframe.isnull().sum()
print(missing_data)
drop_all_missing_data = dataframe.dropna(how="all")
```

4. Data Type Conversion: Key columns were converted to their appropriate data types for accurate analysis. The 'DR_NO' column was converted to an integer, while the 'DATE OCC' column was converted to a datetime format to enable time-based analysis.

```
dataframe['DR_NO'] = dataframe['DR_NO'].astype(int)
dataframe['DATE OCC'] = pd.to_datetime(dataframe['DATE OCC'])
```

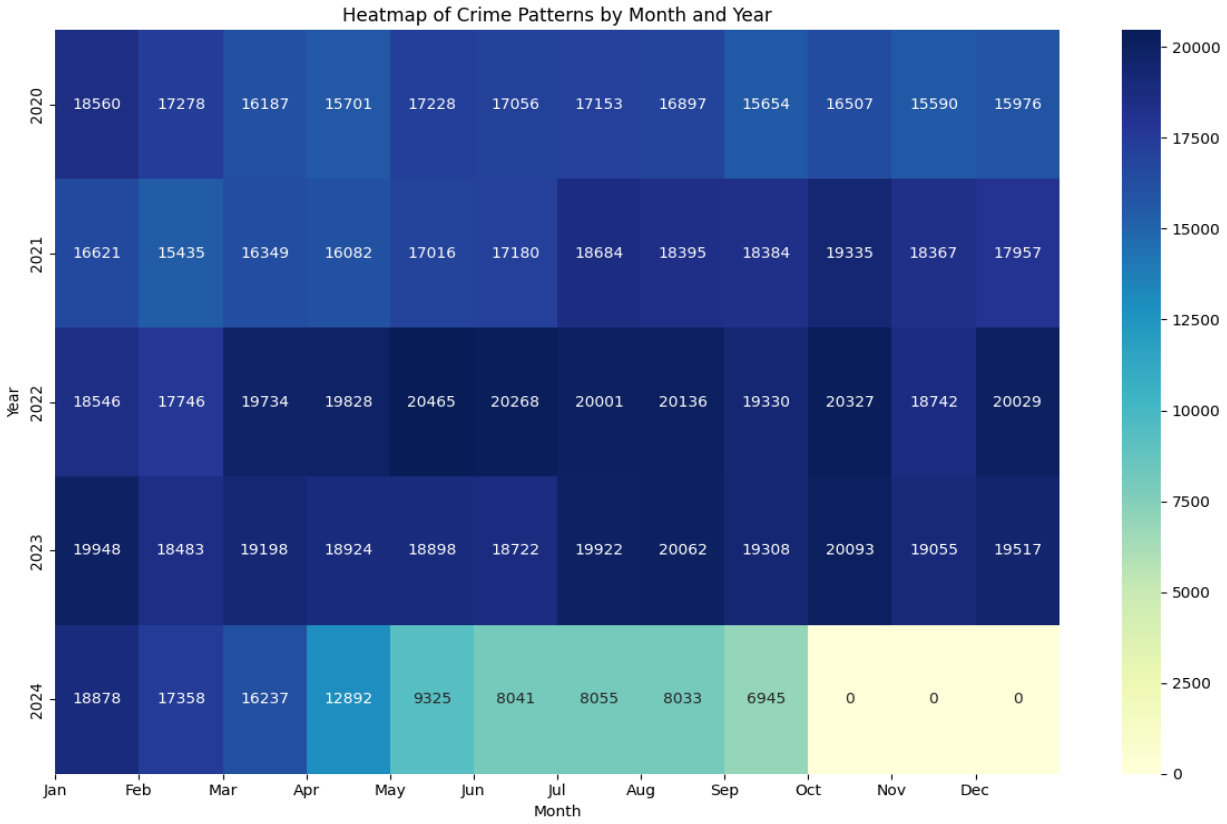
Exploratory Data Analysis (EDA):

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



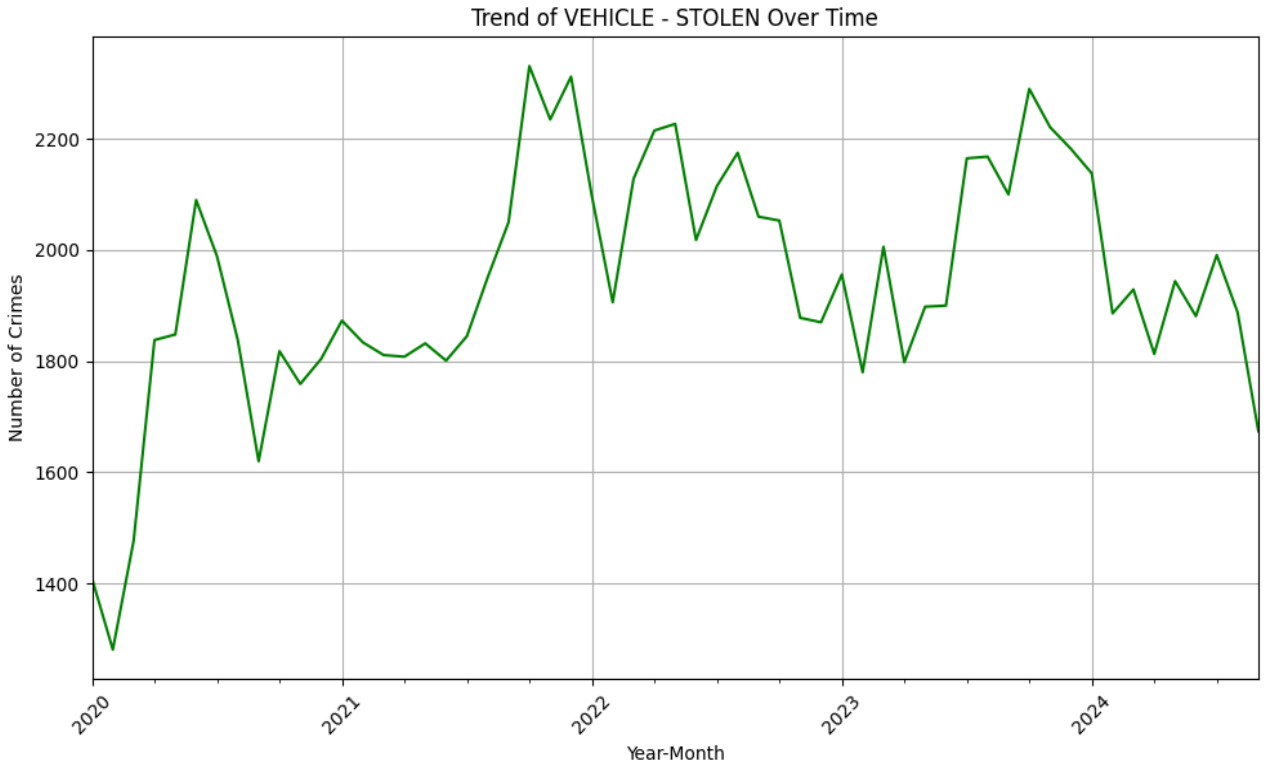
This line graph shows the overall crime trends from 2020 to the present. There is a noticeable decline in crime rates after a relatively stable period, with a significant drop starting around early 2024. This indicates a sharp reduction in reported crimes during that time frame.

2. Analyze and visualize seasonal patterns in crime data.



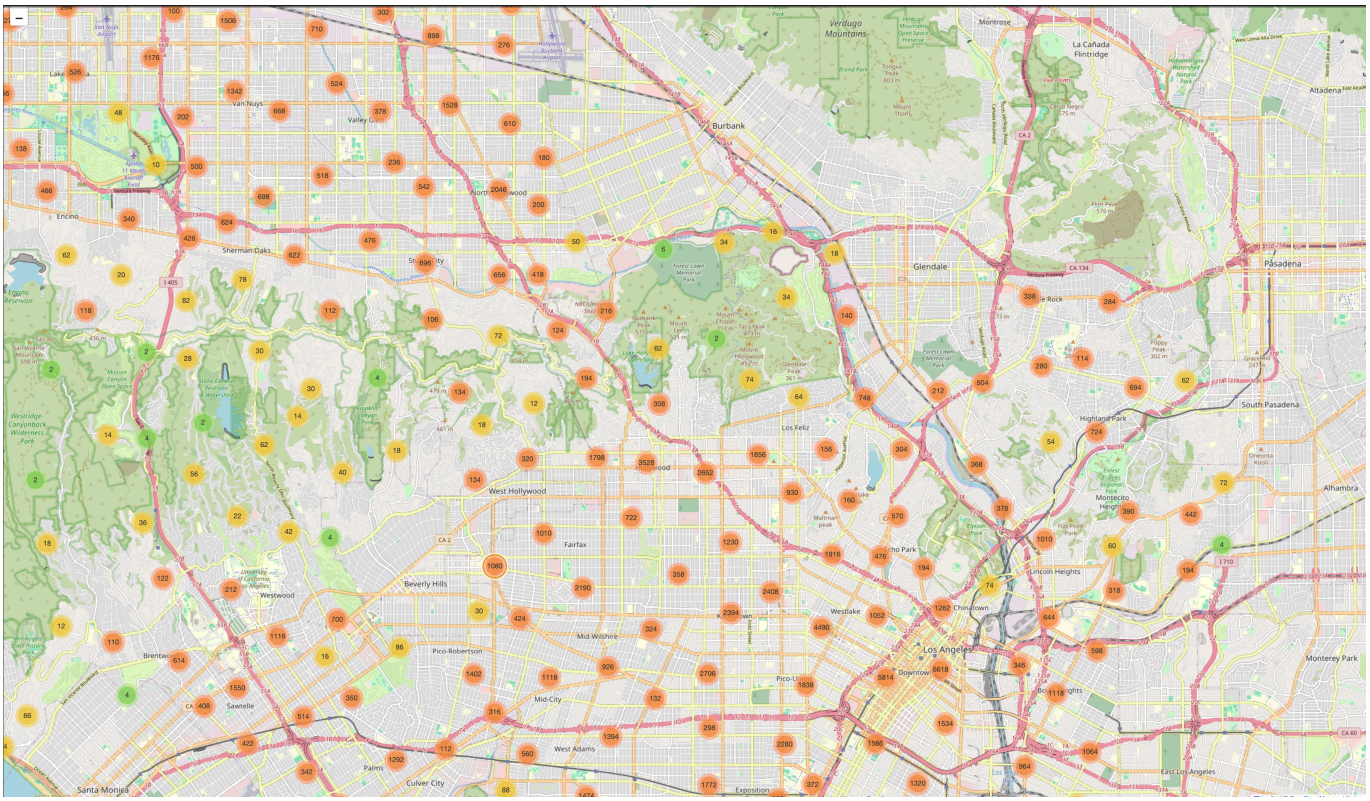
This heatmap shows crime patterns by month and year from 2020 to 2024. Darker shades indicate higher crime rates, while lighter shades represent lower rates. The heatmap reveals a consistent level of crime activity from 2020 to 2023, with peak months in mid-2022 and early 2023. However, there is a sharp decline in crime starting in mid-2024, with no recorded crimes in the last few months of that year. This suggests a significant reduction in crime activity towards the end of the dataset period.

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



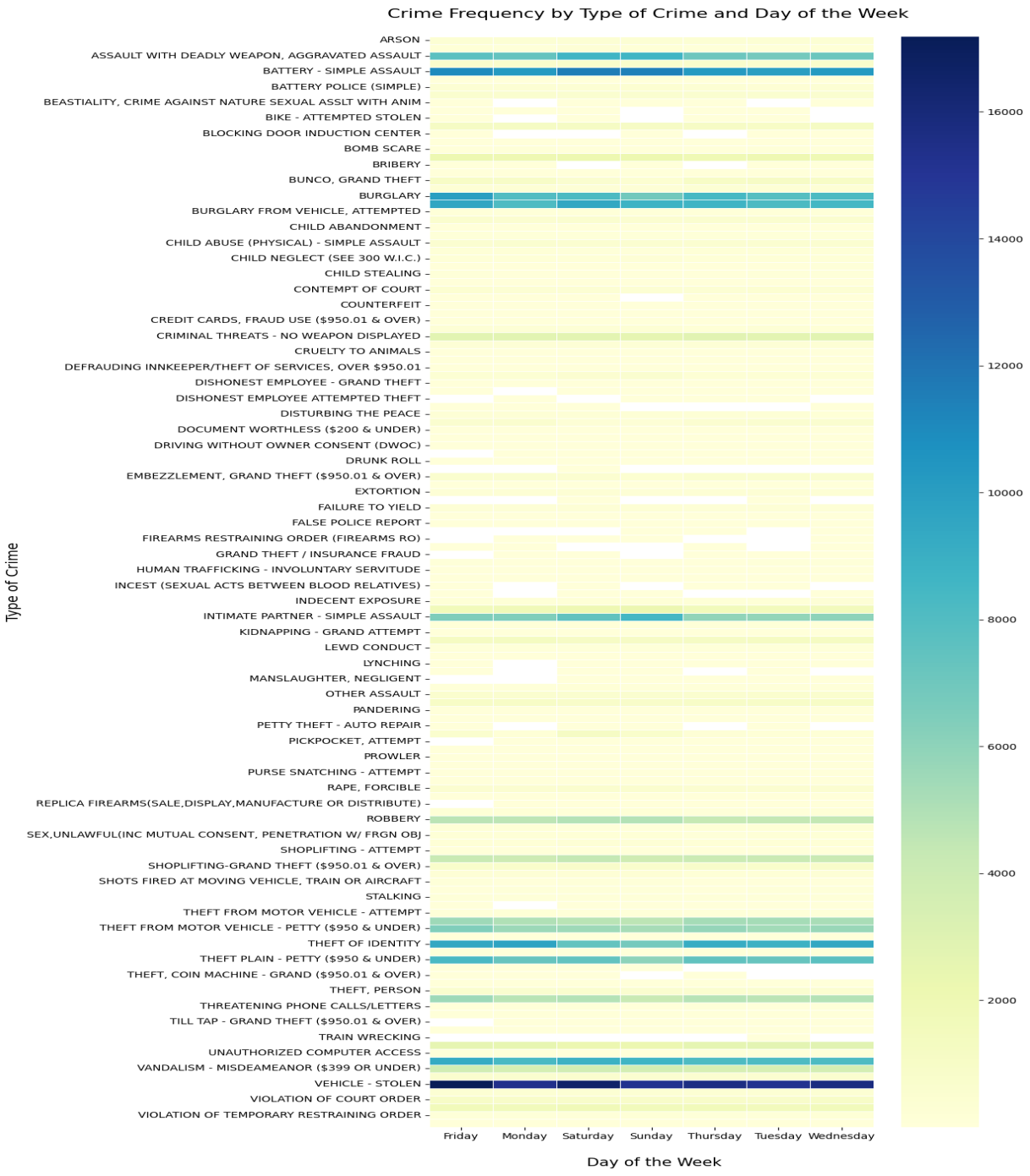
This line graph illustrates the trend of stolen vehicle crimes from 2020 to 2024. The number of vehicle thefts fluctuates over time, with notable peaks in early 2022 and late 2023. There is a general upward trend from mid-2021 to 2022, followed by periodic rises and falls in 2023. However, a significant decline is observed in mid-2024, suggesting a reduction in vehicle thefts toward the end of the period.

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



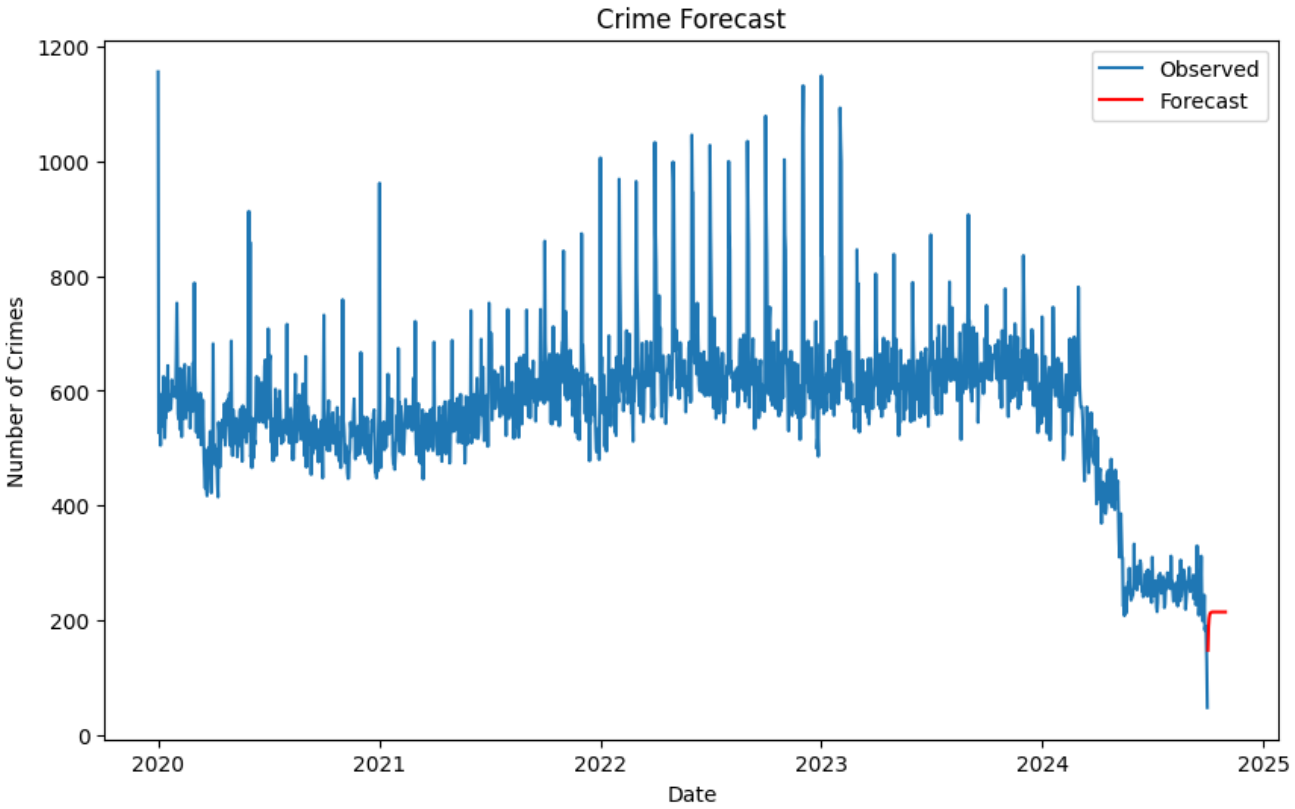
This image displays a heat map of crime incidents across Los Angeles, with crime intensity marked by colored circles. Red and orange areas represent high crime rates, particularly in urban centers, while green and yellow indicate lower crime rates in suburban and park regions.

5. Analyze the relationship between the day of the week and the frequency of certain types of crime.



This heatmap illustrates crime frequency by type and day of the week, with darker shades indicating higher occurrences. Offenses like "Battery - Simple Assault" and "Theft Plain" are among the most frequent, with peak incidences spread across multiple days. The color gradient reveals patterns in which certain crimes occur more often on specific days, providing insight into weekly crime trends.

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



The graph shows a time series analysis of crime data aggregated daily from 2020 to late 2024, with a 30-day crime forecast for early 2025. The data was analyzed using the ARIMA model to identify trends and project future crime counts. The observed data indicates periodic fluctuations in crime, with a notable decline in late 2023 continuing into 2024. The red line in the forecast suggests a continuation of the downward trend with slightly lower levels of criminal activity anticipated in the early part of 2025.

Summary of Findings

This report provides an analysis of crime trends from 2020 to the present using a dataset of crime reports. The data preparation involved loading, inspecting, and cleaning the dataset, including handling missing values and converting data types. Exploratory data analysis (EDA) was conducted to visualize crime trends, patterns, and relationships over time.

Key findings include:

- *Overall Crime Trends*: Crime rates remained stable until early 2024, after which there was a sharp decline.
- *Seasonal Crime Patterns*: A heatmap analysis revealed consistent levels of crime activity through most months, with noticeable reductions starting mid-2024.
- Vehicle Theft Trends: Stolen vehicle crime fluctuated over the years, peaking in late 2023 before declining sharply in 2024.

These insights provide a foundation for understanding the temporal dynamics of crime and the impact of external factors on crime rates.