CLEANING AND ANALYZING CRIME DATA

IE6400 FOUNDATIONS DATA ANALYTICS ENGINEERING

Project Report

Group-9

Himabindu Peramala(002493479)

Swathi Baba Eswarappa(002373257)

Rohan Prakash Krishna Prakash(002317798)

Contents

Introduction:

As part of the IE6400 Foundations of Data Analytics Engineering course, this assignment concentrates on the cleansing and analysis of real-world criminal data from 2020.

The dataset, which is derived from public crime records, contains comprehensive information regarding crime incidents that have been reported in a variety of regions. The initiative aims to prepare the data for analysis by cleansing, investigating, and identifying key trends and patterns within the crime data.

The project's responsibilities encompass the acquisition, cleansing, and exploratory data analysis (EDA) of data in order to reveal insights regarding regional differences, seasonal variability, and crime trends. Furthermore, predictive modelling and other sophisticated analyses will be implemented to anticipate future criminal trends. This initiative offers valuable experience in the application of data analytics techniques to real-world challenges in criminal analysis and the management of large datasets.

Problem Definition:

The goal of this project is to find significant insights into crime trends, patterns, and variables influencing crime rates by efficiently analyzing large-scale crime data. Crime data presents issues with data quality, missing information, and heterogeneity between geographies and time periods. It is often defined by its complexity and volume. Inadequate analysis often results in the overlooking of important findings.

This issue is important because it has the ability to enhance resource allocation and public safety. Law enforcement organizations may more effectively allocate resources by recognizing patterns of criminal activity and predicting future trends, and legislators can create well-informed programs for preventing crimes. Furthermore, knowing how crime rates relate to outside variables like the state of the economy may assist communities in putting interventions in place that deal with the underlying causes of crime. Thus, this project intends to give actionable insights that might result in safer communities in addition to solving the technological hurdles associated with data analysis.

Objective:

This project's goal is to clean, examine, and extract insights from a 2020 crime dataset using data analytics approaches. The project includes getting the data ready for analysis and investigating patterns, trends, and important variables that affect crime rates. Students will use thorough exploratory data analysis (EDA) to look into the frequency of crimes over time, find regional differences, and evaluate the influence of outside variables. In order to predict future crime trends and provide actionable insights that can guide efforts for crime prevention and public safety enhancements, the project also uses predictive modelling.

Data Collection:

we collected the dataset from the U.S. Government's data catalog website which contains "Crime Data from 2020 to Present" available at the URL "https://catalog.data.gov/dataset/crime-data-from-2020-to-present." This dataset was selected due to its comprehensive and up-to-date information on crime incidents across various jurisdictions, allowing for a thorough examination of crime trends and patterns. The data spans multiple years, providing a robust foundation for longitudinal analysis. By utilizing a publicly available and authoritative source, we ensure the credibility and relevance of our analysis, enabling us to draw meaningful insights into crime dynamics during recent years. The 'pd.read_csv' function was utilized to load the CSV file's contents into a DataFrame, which is a tabular structure widely employed in data analysis. This DataFrame, referred to as 'cp', forms the basis for our upcoming exploratory data analysis.

```
import pandas as pd
cp = pd.read_csv(r"c:\ms\IE6400\Crimedate_till_sep29.csv")
print(cp)
```

Data Size:

The dataset provides a thorough picture of crime statistics from many areas and years, with nearly 980,000 rows and 28 columns. The data's scale and breadth allow for a variety of analysis, such as anticipating crime trends, comparing areas, and correlating with outside variables like economic indicators.

Data Set:

The dataset used for this study includes real-world crime statistics from public records starting in 2020. It offers comprehensive details regarding different criminal occurrences, such as the time and date of the incident, the location (latitude and longitude), the kind of crime that occurred, and other pertinent details.

Key Features of the Dataset:

- **DR_NO**: individualized number for every criminal complaint.
- **Date Reported (Date Rptd)**: The date on which the offense was formally reported.
- **Date Occurred (DATE OCC)**: The precise time and day when the offense was committed.
- **Time Occurred (TIME OCC)**: the moment the offense was committed.
- **Area Name (AREA NAME)**: the area or police jurisdiction in which the offense took place.
- **Crime Code (Crm Cd)**: a number that indicates the kind of crime.
- **Crime Description (Crm Cd Desc)**: a thorough explanation of the kind of crime (such as vehicle theft or burglary).
- **Victim Details**: The victim's age, gender, and ethnicity, if relevant.
- **Weapon Used**: if available, gives a description of the weapon and indicates whether one was used in the crime.
- **Location Details**: details about the incident scene, such as the latitude and longitude coordinates and the street address.
- **Status**: The state of the inquiry or its result (e.g., Arrest of an Adult, Investigation Pending).

Data Inspection:

Before beginning analysis, the 2020 crime dataset—which has approximately 982,000 rows and 28 columns—was carefully examined to ascertain its quality and structure.

Key Insights from Data Inspection:

1. **Displaying first few rows:**

- The dataset contains comprehensive records of criminal episodes, including victim information, crime type, location, and time of occurrence.
- Date Reported, Date Occurred, Area Name, Crime Description, and Victim Demographics (e.g., Age, Sex, and Descent) are significant columns.
- The first and simplest step is to show the initial rows of the dataset. This enables us to visually examine the data and understand its structure and contents. The

Pandas library offers a handy function called head() for this purpose. Here's how it can be used:

```
#Display the first 10 rows of the dataframe
cp.head(10)
```

Here we can get an overview of the data

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
0	190326475	03-01-2020 00:00	03-01-2020 00:00	2130	7	Wilshire	784	1	510	VEHICLE - STOLEN	...	AA	Adult Arrest	510.0	998.0	NaN	NaN	1900 S LONGWOOD AV	NaN	34.0375	-118.3506
1	200106753	02-09-2020 00:00	02-08-2020 00:00	1800	1	Central	182	1	330	BURGLARY FROM VEHICLE	...	IC	Invest Cont	330.0	998.0	NaN	NaN	1000 S FLOWER ST	NaN	34.0444	-118.2628
2	200320258	11-11-2020 00:00	11-04-2020 00:00	1700	3	Southwest	356	1	480	BIKE - STOLEN	...	IC	Invest Cont	480.0	NaN	NaN	NaN	1400 W 37TH ST	NaN	34.0210	-118.3002
3	200907217	05-10-2023 00:00	03-10-2020 00:00	2037	9	Van Nuys	964	1	343	SHOPLIFTING-GRAND THEFT ($950.01 & OVER)	...	IC	Invest Cont	343.0	NaN	NaN	NaN	14000 RIVERSIDE DR	NaN	34.1576	-118.4387
4	220614831	08/18/2022 12:00:00 AM	08/17/2020 12:00:00 AM	1200	6	Hollywood	666	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1900 TRANSIENT	NaN	34.0944	-118.3277
5	231808869	04-04-2023 00:00	12-01-2020 00:00	2300	18	Southeast	1826	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	9900 COMPTON AV	NaN	33.9467	-118.2463
6	230110144	04-04-2023 00:00	07-03-2020 00:00	900	1	Central	182	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1100 S GRAND AV	NaN	34.0415	-118.2620
7	220314085	07/22/2022 12:00:00 AM	05-12-2020 00:00	1110	3	Southwest	303	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	2500 S SYCAMORE AV	NaN	34.0335	-118.3537
8	231309864	04/28/2023 12:00:00 AM	12-09-2020 00:00	1400	13	Newton	1375	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1300 E 57TH ST	NaN	33.9911	-118.2521
9	211904005	12/31/2020 12:00:00 AM	12/31/2020 12:00:00 AM	1220	19	Mission	1974	2	624	BATTERY - SIMPLE ASSAULT	...	IC	Invest Cont	624.0	NaN	NaN	NaN	9000 CEDROS AV	NaN	34.2336	-118.4535

10 rows × 28 columns

2. Checking Data Types:

It's important to understand the data types for each column, as they influence how data is stored and can impact later data manipulation and analysis. Pandas provides a method to check the data types of each column in the dataset.

```
#Display the datatypes of each column in the dataframe
cp.dtypes
```

Through this we can see the data types of all columns in our dataset.

```
DR_NO              int64
Date Rptd         object
DATE OCC          object
TIME OCC           int64
AREA               int64
AREA NAME         object
Rpt Dist No        int64
Part 1-2           int64
Crm Cd             int64
Crm Cd Desc       object
Mocodes           object
Vict Age           int64
Vict Sex          object
Vict Descent      object
Premis Cd        float64
Premis Desc       object
Weapon Used Cd   float64
Weapon Desc       object
Status            object
Status Desc       object
Crm Cd 1         float64
Crm Cd 2         float64
Crm Cd 3         float64
Crm Cd 4         float64
LOCATION          object
Cross Street      object
LAT              float64
LON              float64
dtype: object
```

3.Reviewing column names and Description:

The "Description" section is a valuable addition to a report as it delivers a brief yet thorough overview of the dataset. It offers insights into data types, missing values, memory usage, and data quality, facilitating efficient data exploration and helping to make informed decisions during data analysis and preparation supports informed decision-making during data analysis and preparation.

```
#Display summary of the dataframe including column names, non-null counts, and data types
cp.describe()
```

	DR_NO	TIME OCC	AREA	Rpt Dist No	Part 1-2	Crm Cd	Vict Age	Premis Cd	Weapon Used Cd	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LAT	LON
count	9.786280e+05	978628.000000	978628.000000	978628.000000	978628.000000	978628.000000	978628.000000	978613.000000	325959.000000	978617.000000	68816.000000	2309.000000	64.00000	978628.000000	978628.000000
mean	2.196564e+08	1338.802627	10.702561	1116.686084	1.404785	500.810635	29.122904	306.181502	363.815372	500.564847	958.156344	984.192724	991.21875	33.995399	-118.081108
std	1.290395e+07	651.622947	6.107280	610.836054	0.490851	206.309796	21.961531	218.908131	123.673988	206.107451	110.251477	51.506344	27.06985	1.640056	5.684520
min	8.170000e+02	1.000000	1.000000	101.000000	1.000000	110.000000	-4.000000	101.000000	101.000000	110.000000	210.000000	310.000000	821.00000	0.000000	-118.667600
25%	2.106073e+08	900.000000	5.000000	589.000000	1.000000	331.000000	0.000000	101.000000	311.000000	331.000000	998.000000	998.000000	998.00000	34.014600	-118.430500
50%	2.208116e+08	1420.000000	11.000000	1141.000000	1.000000	442.000000	30.000000	203.000000	400.000000	442.000000	998.000000	998.000000	998.00000	34.058900	-118.322500
75%	2.309110e+08	1900.000000	16.000000	1617.000000	2.000000	626.000000	44.000000	501.000000	400.000000	626.000000	998.000000	998.000000	998.00000	34.164900	-118.273900
max	2.499253e+08	2359.000000	21.000000	2199.000000	2.000000	956.000000	120.000000	976.000000	516.000000	956.000000	999.000000	999.000000	999.00000	34.334300	0.000000

Data Cleaning:

Accuracy and consistency of the data utilized in subsequent phases of the project are ensured by data cleaning, which is an essential step in getting our crime dataset ready for analysis. The main steps involved in the data cleansing process are listed below:

1. **Identifying & Handling Missing Data:**

- The.isnull() function was used to find missing data in the dataset. There were blanks in a number of columns, such as those with location coordinates, criminal descriptions, and other data.

```
: #Check for null values in the dataframe
  cp.isnull()
```

- We can see true in the place of missing values in our dataset

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	True	True	False	True	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	True	True	False	True	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
978623	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False
978624	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False
978625	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False
978626	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	False	False	False
978627	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	True	True	False	True	False	False

978628 rows × 28 columns

- We used dropna( ) function to handle the missing data and the poutput can be seen below:

```
#Drop rows with null values from the dataframe (handling missing data)
cp.dropna()
```

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
66026	201904032	01-02-2020 00:00	01-01-2020 00:00	2135	19	Mission	1924	1	761	BRANDISH WEAPON	...	AA	Adult Arrest	761.0	930.0	997.0	998.0	ASTORIA ST	SAN FERNANDO RD	34.2949	-118.4571
86496	200613424	08-02-2020 00:00	08-02-2020 00:00	2030	6	Hollywood	657	1	761	BRANDISH WEAPON	...	AO	Adult Other	761.0	920.0	930.0	998.0	WESTERN	ROMAINE	34.0885	-118.3092
363643	210617136	10-08-2021 00:00	10-07-2021 00:00	1950	6	Hollywood	659	1	121	RAPE, FORCIBLE	...	IC	Invest Cont	121.0	210.0	910.0	998.0	NORMANDIE	DE LONGPRE	34.0966	-118.3005
372408	210209196	05-08-2021 00:00	05-08-2021 00:00	230	2	Rampart	279	1	210	ROBBERY	...	AO	Adult Other	210.0	510.0	910.0	998.0	JAMES M WOOD	GREEN	34.0503	-118.2720
489920	220600626	04/27/2022 12:00:00 AM	04/23/2022 12:00:00 AM	2300	6	Hollywood	646	1	821	SODOMY/SEXUAL CONTACT B/W PENIS OF ONE PERS TO...	...	IC	Invest Cont	230.0	821.0	910.0	998.0	SELMA	LAS PALMAS	34.0997	-118.3363
537636	221718232	12/25/2022 12:00:00 AM	12/25/2022 12:00:00 AM	1150	17	Devonshire	1797	1	122	RAPE, ATTEMPTED	...	AA	Adult Arrest	122.0	230.0	910.0	998.0	PARTHENIA ST	HAYVENHURST	34.2285	-118.4939
585780	221401314	11-10-2022 00:00	11-10-2022 00:00	2117	14	Pacific	1452	2	910	KIDNAPPING	...	IC	Invest Cont	812.0	860.0	910.0	998.0	WASHINGTON	SPEEDWAY	33.9792	-118.4666
728192	231717599	11/15/2023 12:00:00 AM	11/15/2023 12:00:00 AM	400	17	Devonshire	1738	1	210	ROBBERY	...	IC	Invest Cont	210.0	230.0	761.0	998.0	HASKELL AV	SAN FERNANDO BL	34.2692	-118.4789
809005	231915572	10/21/2023 12:00:00 AM	10/21/2023 12:00:00 AM	1	19	Mission	1902	1	210	ROBBERY	...	AA	Adult Arrest	210.0	250.0	761.0	998.0	POLK	BORDEN	34.3103	-118.4467
922659	241905348	02-04-2024 00:00	02-03-2024 00:00	1100	19	Mission	1983	1	820	ORAL COPULATION	...	AO	Adult Other	761.0	820.0	910.0	998.0	BURNET	PARTHENIA	34.2282	-118.4633

10 rows × 28 columns

- Rows lacking data for critical columns, such as Date Rptd and DATE OCC, were removed since they are required for time-based analysis.

2. Removing Duplicates:

- In the data analysis process, it is vital to ensure that the dataset contains unique and non-repetitive information. Duplicate rows in a dataset can skew analysis results and potentially lead to inaccurate insights. To address this concern, we applied the following data cleaning step:
- The code "cp.drop_duplicates()" was executed to eliminate duplicate rows from the dataset. Duplicate rows are rows that have identical values across all fields.
- The resulting dataset contains only unique records, ensuring that each row represents distinct information, enhancing data accuracy and analytical validity.
- The "head()" function was used to display the first few rows of the cleaned dataset, offering a glimpse of the data's structure and content after removing duplicates

```
cp.drop_duplicates().head()
```

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	L
0	190326475	03-01-2020 00:00	03-01-2020 00:00	2130	7	Wilshire	784	1	510	VEHICLE - STOLEN	...	AA	Adult Arrest	510.0	998.0	NaN	NaN	1900 S LONGWOOD AV	NaN	34.03
1	200106753	02-09-2020 00:00	02-08-2020 00:00	1800	1	Central	182	1	330	BURGLARY FROM VEHICLE	...	IC	Invest Cont	330.0	998.0	NaN	NaN	1000 S FLOWER ST	NaN	34.04
2	200320258	11-11-2020 00:00	11-04-2020 00:00	1700	3	Southwest	356	1	480	BIKE - STOLEN	...	IC	Invest Cont	480.0	NaN	NaN	NaN	1400 W 37TH ST	NaN	34.02
3	200907217	05-10-2023 00:00	03-10-2020 00:00	2037	9	Van Nuys	964	1	343	SHOPLIFTING-GRAND THEFT ($950.01 & OVER)	...	IC	Invest Cont	343.0	NaN	NaN	NaN	14000 RIVERSIDE DR	NaN	34.15
4	220614831	08/18/2022 12:00:00 AM	08/17/2020 12:00:00 AM	1200	6	Hollywood	666	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1900 TRANSIENT	NaN	34.09

5 rows × 28 columns

3. Data Type Conversion:

The Date Rptd and DATE OCC columns were converted to a standard datetime format using Pandas' pd.to_datetime() function, ensuring consistent handling of various date formats and coercing invalid entries to NaT.

The TIME OCC column was processed to ensure all time entries were represented in a standardized HH:MM format. This involved padding times with leading zeros and converting them to time objects using datetime.strptime() to maintain uniformity.

The conversion processes improved the accuracy and consistency of the dataset, facilitating efficient sorting, filtering, and analysis of temporal trends, which are essential for effective crime trend analysis and forecasting.

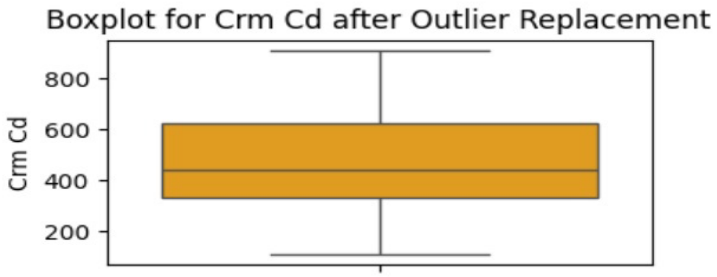
```python
import pandas as pd
from datetime import datetime
# Sample data creation for demonstration (replace this with your actual data loading)
cp = pd.read_csv(r"c:\ms\IE6400\Crimedate_till_sep29.csv")
# Convert the 'Date Rptd' column to datetime format, coercing invalid entries to NaT
cp['Date Rptd'] = pd.to_datetime(cp['Date Rptd'], errors='coerce')
# Convert the 'DATE OCC' column to datetime format, coercing invalid entries to NaT
cp['DATE OCC'] = pd.to_datetime(cp['DATE OCC'], errors='coerce')
# Function to process and convert 'TIME OCC' into HH:MM format
def process_time(x):
    try:
        # Ensure the value is treated as a string, fill with leading zeros if necessary
        time_str = str(int(x)).zfill(4)  # Convert to string and pad
        # Convert the string to time format
        return (datetime.strptime(time_str, "%H%M").time()).strftime("%H:%M")
    except (ValueError, TypeError):
        # Return a default value if conversion fails
        return '00:00'
# Apply the function to 'TIME OCC'
cp['TIME OCC'] = cp['TIME OCC'].apply(process_time)
# Display the updated columns to verify conversion
print("Converted Dates and Times:")
print(cp[['Date Rptd', 'DATE OCC', 'TIME OCC']].head())# Display the first 5 rows
print(cp[['Date Rptd', 'DATE OCC', 'TIME OCC']].tail())# Display the last 5 rows
```

```
Converted Dates and Times:
   Date Rptd   DATE OCC TIME OCC
0 2020-03-01 2020-03-01    21:30
1 2020-02-09 2020-02-08    18:00
2 2020-11-11 2020-11-04    17:00
3 2023-05-10 2020-03-10    20:37
4        NaT        NaT    12:00
       Date Rptd   DATE OCC TIME OCC
978623       NaT        NaT    14:00
978624       NaT        NaT    01:00
978625       NaT        NaT    07:57
978626       NaT        NaT    15:00
978627       NaT 2024-08-12    23:00
```

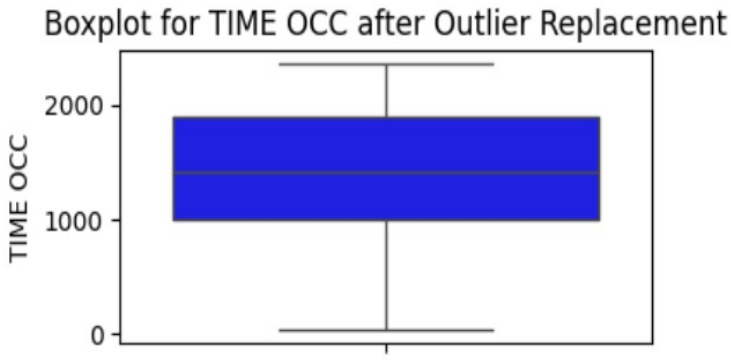
4. Outlier Detection and Removal:

using Z-Score Method:
The Z-score method is used to identify outliers in a field. Outliers represent data points significantly deviating from the typical range of values. These outliers were then replaced with the median value of the column. This imputation approach is robust to extreme values and helps maintain the integrity of the dataset. To visually confirm the effectiveness of this treatment, we created a boxplot, which provides a graphical representation of the central tendency and spread of the values post-outlier imputation. By implementing this approach, we aimed to ensure that extreme values in the field do not unduly influence subsequent analyses. This process serves to enhance the reliability and accuracy of our data for further exploration and interpretation.

```
import numpy as np
from scipy.stats import zscore
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
cp = pd.read_csv(r"c:\ms\IE6400\Crimedate_till_sep29.csv")

# Z-score outlier detection and replacement for 'Crm Cd'
z_scores_crm_cd = zscore(cp['Crm Cd'])
outliers_crm_cd = (np.abs(z_scores_crm_cd) > 2)
cp['Crm Cd'] = np.where(outliers_crm_cd, cp['Crm Cd'].median(), cp['Crm Cd'])

# Plot boxplot for 'Crm Cd'
plt.figure(figsize=(4, 2))
sns.boxplot(data=cp['Crm Cd'], color='orange')
plt.title('Boxplot for Crm Cd after Outlier Replacement')
plt.show()
```




```
# Z-score outlier detection and replacement for 'TIME OCC'
z_scores_time_occ = zscore(cp['TIME OCC'])
outliers_time_occ = (np.abs(z_scores_time_occ) > 2)
cp['TIME OCC'] = np.where(outliers_time_occ, cp['TIME OCC'].median(), cp['TIME OCC'])

# Plot boxplot for 'TIME OCC'
plt.figure(figsize=(4, 2))
sns.boxplot(data=cp['TIME OCC'], color='blue')
plt.title('Boxplot for TIME OCC after Outlier Replacement')
plt.show()
```



```
# Z-score outlier detection and replacement for 'Rpt Dist No'
z_scores_rpt_dist_no = zscore(cp['Rpt Dist No'])
outliers_rpt_dist_no = (np.abs(z_scores_rpt_dist_no) > 2)
cp['Rpt Dist No'] = np.where(outliers_rpt_dist_no, cp['Rpt Dist No'].median(), cp['Rpt Dist No'])

# Plot boxplot for 'Rpt Dist No'
plt.figure(figsize=(4, 2))
sns.boxplot(data=cp['Rpt Dist No'], color='green')
plt.title('Boxplot for Rpt Dist No after Outlier Replacement')
plt.show()
```

5. Normalization and Standardization:

- To improve the comparability of various attributes, numerical columns were standardized to have a mean of 0 and a standard deviation of 1. For certain machine learning models and statistical studies that rely on normalized data, this step is crucial.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Load the dataset
cp = pd.read_csv(r"c:\ms\IE6400\Crimedate_till_sep29.csv")

# Select only the numerical columns for standardization
numerical_columns = cp.select_dtypes(include=['float64', 'int64']).columns

# Clean the dataset to remove NaN values before standardization
cp_cleaned = cp.dropna(subset=numerical_columns)

# Standardization: Scale the numerical data to have a mean of 0 and std of 1
scaler = StandardScaler()
cp_standardized = cp_cleaned.copy()  # Make a copy to preserve the original data
cp_standardized[numerical_columns] = scaler.fit_transform(cp_cleaned[numerical_columns])

print("Standardized Data:")
print(cp_standardized[numerical_columns].head())
```

```
Standardized Data:
          DR_NO  TIME OCC      AREA  Rpt Dist No  Part 1-2    Crm Cd  \
2198  -1.282787 -1.623306 -0.983972    -0.954631 -0.388514  0.836245
4127  -1.166433  1.274569  1.356730     1.277183 -0.388514  1.040095
36672 -1.283900  1.303119 -0.983972    -0.937938 -0.388514  0.836245
37652 -1.284359 -2.037288 -0.983972    -0.996363 -0.388514  0.836245
39344 -1.283625 -1.723233 -0.983972    -0.969654 -0.388514  0.836245

       Vict Age  Premis Cd  Weapon Used Cd  Crm Cd 1  Crm Cd 2  Crm Cd 3  \
2198   0.924495   0.536197       -1.238294  1.223923  0.462166  0.646686
4127  -0.409332  -0.326574        0.869554 -0.883803  0.714340  0.403706
36672 -0.462685   0.536197       -1.238294  1.223923  0.462166  0.646686
37652 -0.409332   0.536197        1.654691  0.730956  0.462166  0.646686
39344  0.070846   0.531905       -0.538036  0.730956  0.462166  0.646686

       Crm Cd 4       LAT       LON
2198   0.233138 -0.370817  0.662651
4127   0.233138 -0.510100  0.786173
36672  0.233138 -0.494117  0.772449
37652  0.233138 -0.325150  0.173553
39344  0.233138 -0.360542  0.793659
```

6. Categorical Data Encoding:

- Using Label_Encoder, categorical variables, including AREA NAME (names of regions where crimes happened), were transformed into numeric format so they could be used in correlation analysis and prediction models.

```
#Import labelencoder from sklearn's preprocessing module
from sklearn.preprocessing import LabelEncoder
#Create an instance of labelEncoder
label_encoder = LabelEncoder()
#Transform the 'AREA NAME' column into numerical format
#Each unique category in 'AREA NAME' will be assigned a unique integer
cp['AREA NAME'] = label_encoder.fit_transform(cp['AREA NAME'])
#Display the first 10 rows of the dataframe to see the change in 'AREA NAME' column
cp.head(10)
```

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT
0	190326475	03-01-2020 00:00	03-01-2020 00:00	2130	7	20	784	1	510	VEHICLE - STOLEN	...	AA	Adult Arrest	510.0	998.0	NaN	NaN	1900 S LONGWOOD AV	NaN	34.0375
1	200106753	02-09-2020 00:00	02-08-2020 00:00	1800	1	1	182	1	330	BURGLARY FROM VEHICLE	...	IC	Invest Cont	330.0	998.0	NaN	NaN	1000 S FLOWER ST	NaN	34.0444
2	200320258	11-11-2020 00:00	11-04-2020 00:00	1700	3	15	356	1	480	BIKE - STOLEN	...	IC	Invest Cont	480.0	NaN	NaN	NaN	1400 W 37TH ST	NaN	34.0210
3	200907217	05-10-2023 00:00	03-10-2020 00:00	2037	9	17	964	1	343	SHOPLIFTING-GRAND THEFT ($950.01 & OVER)	...	IC	Invest Cont	343.0	NaN	NaN	NaN	14000 RIVERSIDE DR	NaN	34.1576
4	220614831	08/18/2022 12:00:00 AM	08/17/2020 12:00:00 AM	1200	6	6	666	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1900 TRANSIENT	NaN	34.0944
5	231808869	04-04-2023 00:00	12-01-2020 00:00	2300	18	14	1826	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	9900 COMPTON AV	NaN	33.9467
6	230110144	04-04-2023 00:00	07-03-2020 00:00	900	1	1	182	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1100 S GRAND AV	NaN	34.0415
7	220314085	07/22/2022 12:00:00 AM	05-12-2020 00:00	1110	3	15	303	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	2500 S SYCAMORE AV	NaN	34.0335
8	231309864	04/28/2023 12:00:00 AM	12-09-2020 00:00	1400	13	9	1375	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1300 E 57TH ST	NaN	33.9911
9	211904005	12/31/2020 12:00:00 AM	12/31/2020 12:00:00 AM	1220	19	7	1974	2	624	BATTERY - SIMPLE ASSAULT	...	IC	Invest Cont	624.0	NaN	NaN	NaN	9000 CEDROS AV	NaN	34.2336

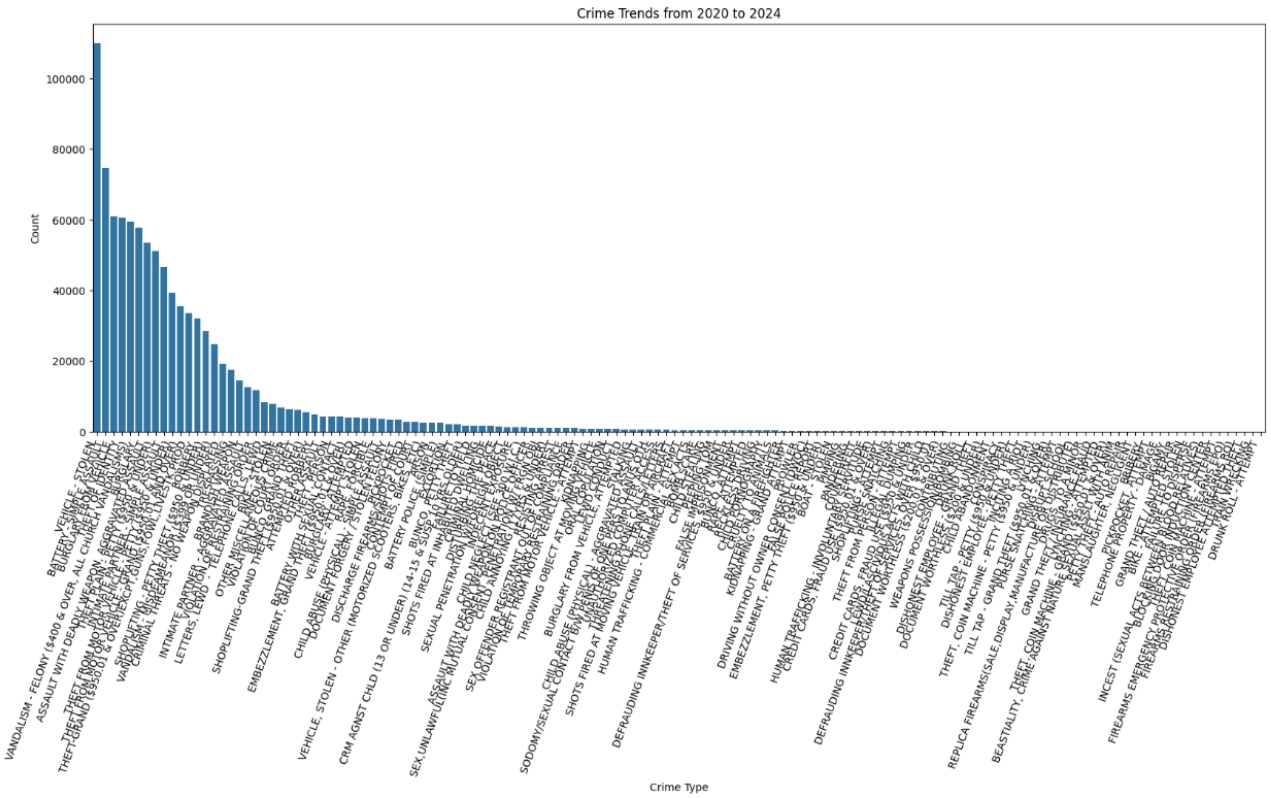
10 rows × 28 columns

Exploratory Data Analysis (EDA):

Finding patterns, trends, and anomalies in the dataset is the main goal of the project's exploratory data analysis (EDA) phase. We learn more about the data and its structure by using a variety of statistical and visual tools. The following are the main conclusions drawn from the EDA: The outcome was

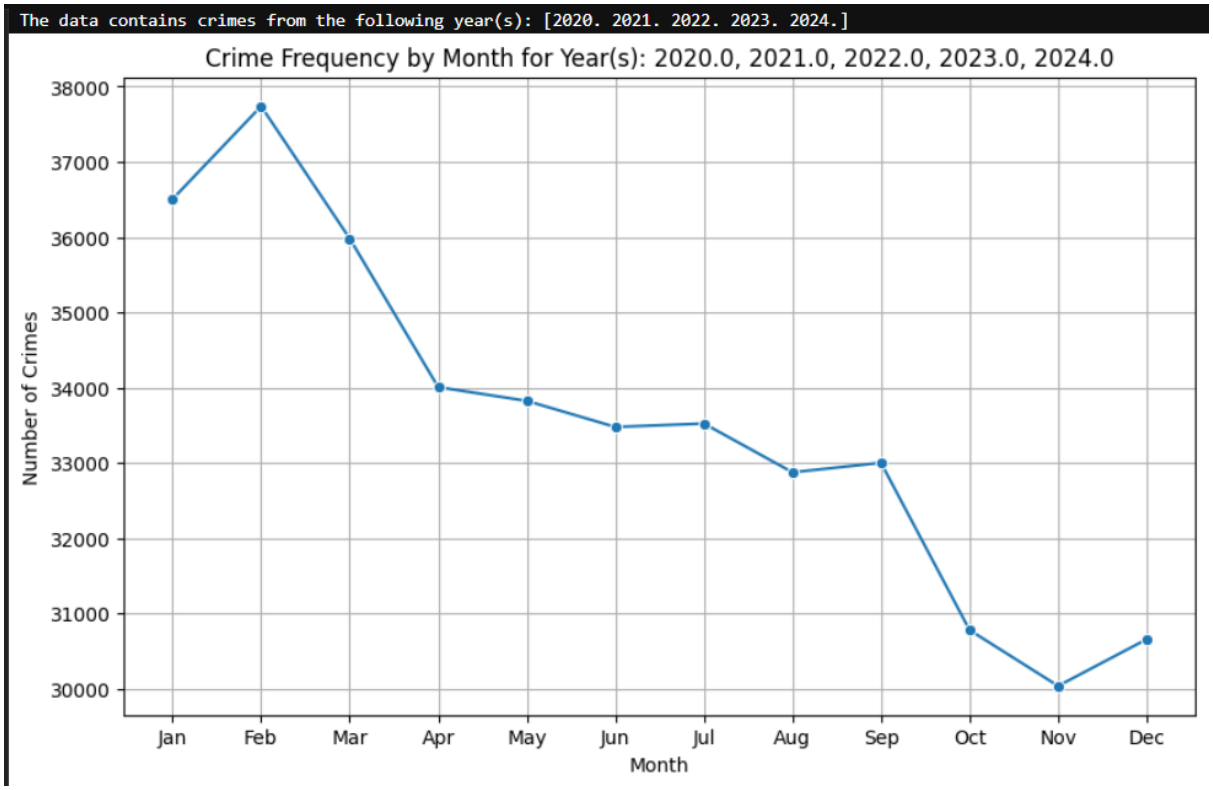
1.Overall Crime Trends

- To display the general distribution of crime types over time, a thorough count of crimes by type was shown. Car theft, vehicle burglaries, and identity theft were the most common crimes during this time. The graphic assisted in emphasizing the most urgent crime categories.
- To see how different sorts of crimes changed over time, a time series line plot was made. This figure showed distinct surges in crime at certain times, which might be related to noteworthy occasions or seasonal trends. The outcome was



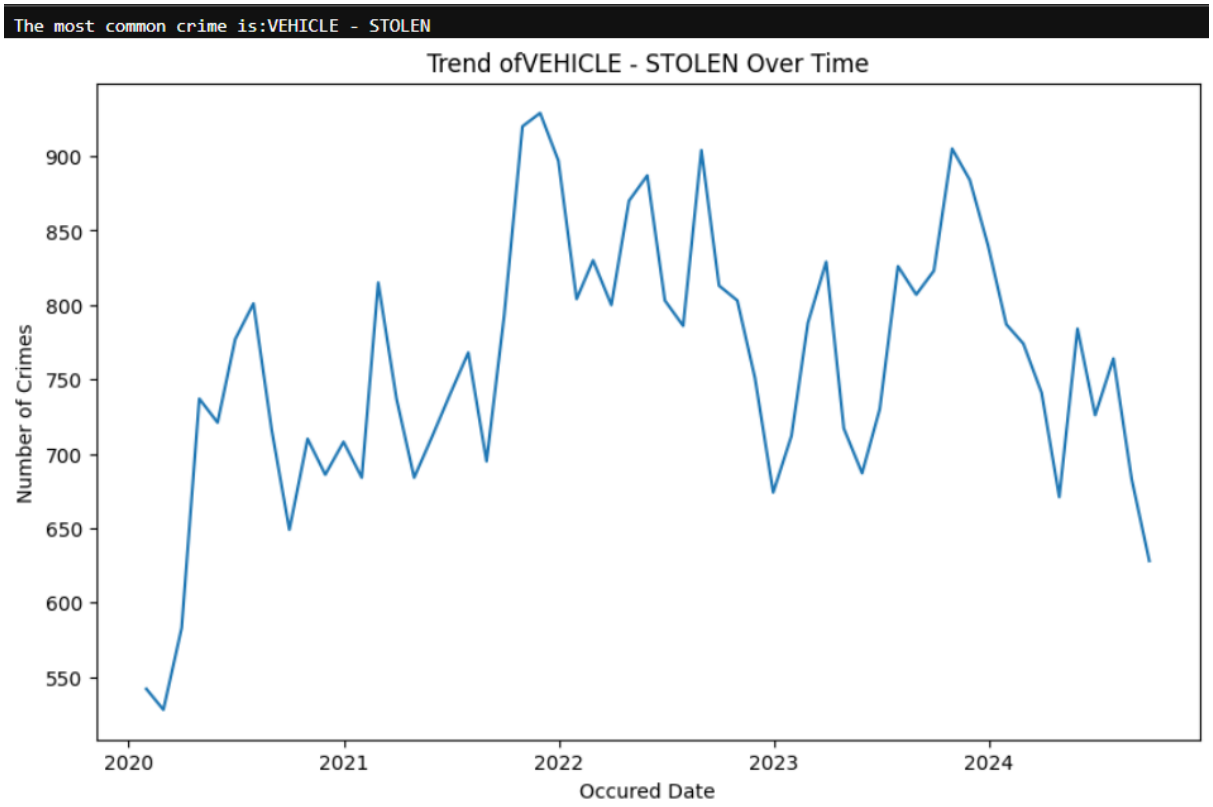
2.Seasonal Patterns:

- In order to investigate any seasonal differences, the dataset was divided into months. To determine if particular seasons of the year saw more crime than others, a line chart was created to illustrate the frequency of crimes by month. This investigation revealed variations, with greater crime rates occurring during particular months (such the summer).
- The data allows for the investigation of seasonal changes over time since it covers a number of years (2020 to 2024). This helps in determining if certain crime categories are more common throughout specific seasons. It's that simple.



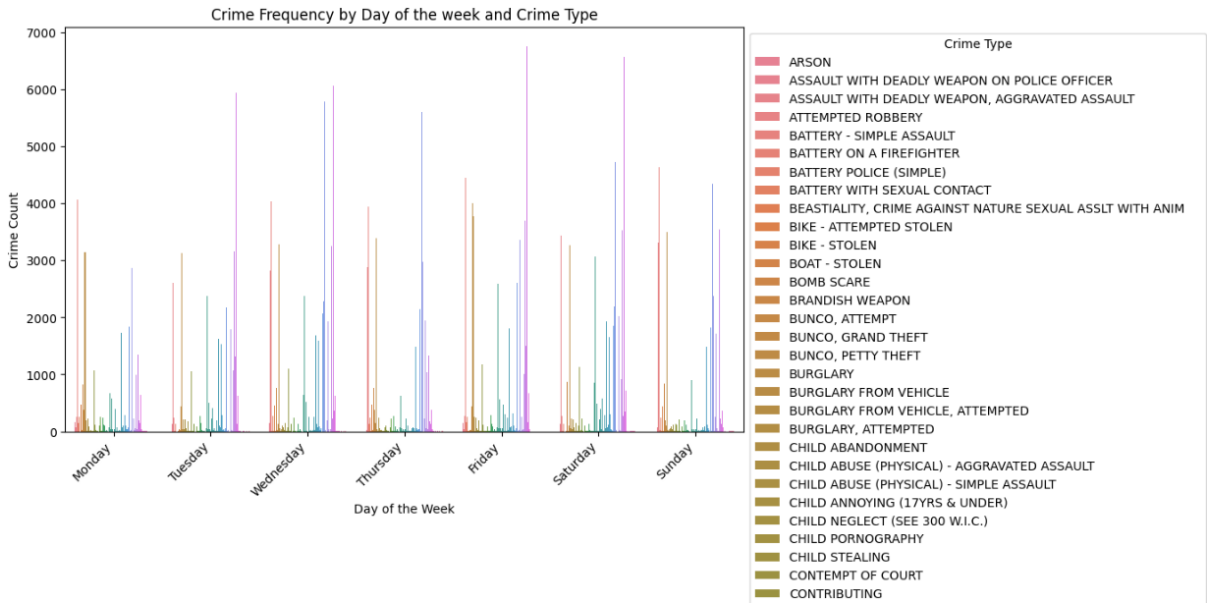
3.Most Common Crime Types:

- The most common crime type across the dataset was Vehicle Theft. Further analysis focused on its trend over time. A line plot illustrated fluctuations in vehicle thefts, showing peaks at certain times, possibly correlating with external factors such as economic downturns or changes in law enforcement policies.
- The frequency and trend analysis of this crime over time provided a clearer picture of how this specific crime evolved, helping to identify any preventative measures that could be taken to curb this trend.



Day of the Week Analysis:

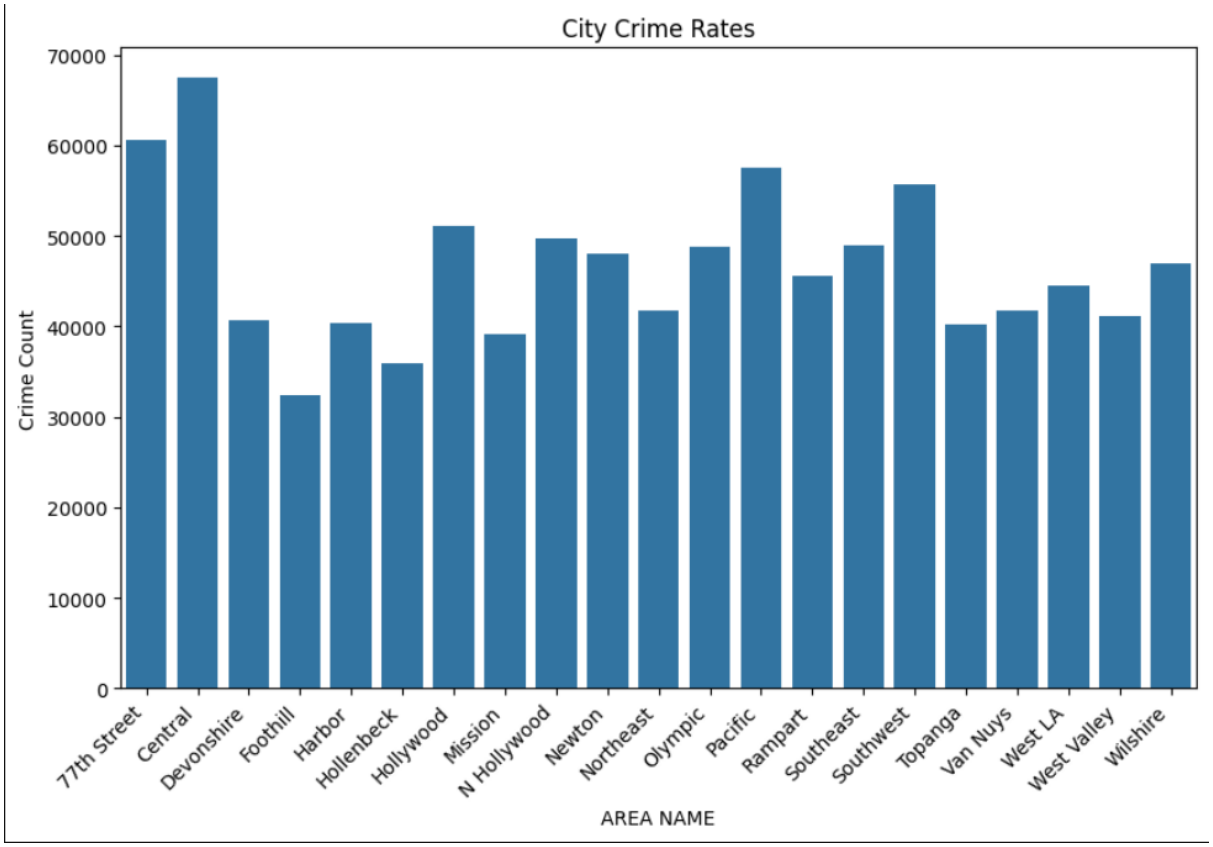
- The analysis of crime statistics by day of the week showed that there were greater crime rates on certain days, especially on the weekends. The distribution of crimes by type over a given period of time was shown using a bar plot, which made it possible to identify trends in criminal activity according to the day of the week.
- This information may help law enforcement use resources more effectively by, for example, stepping up patrols on days when crime is particularly high.



4.Regional Differences:

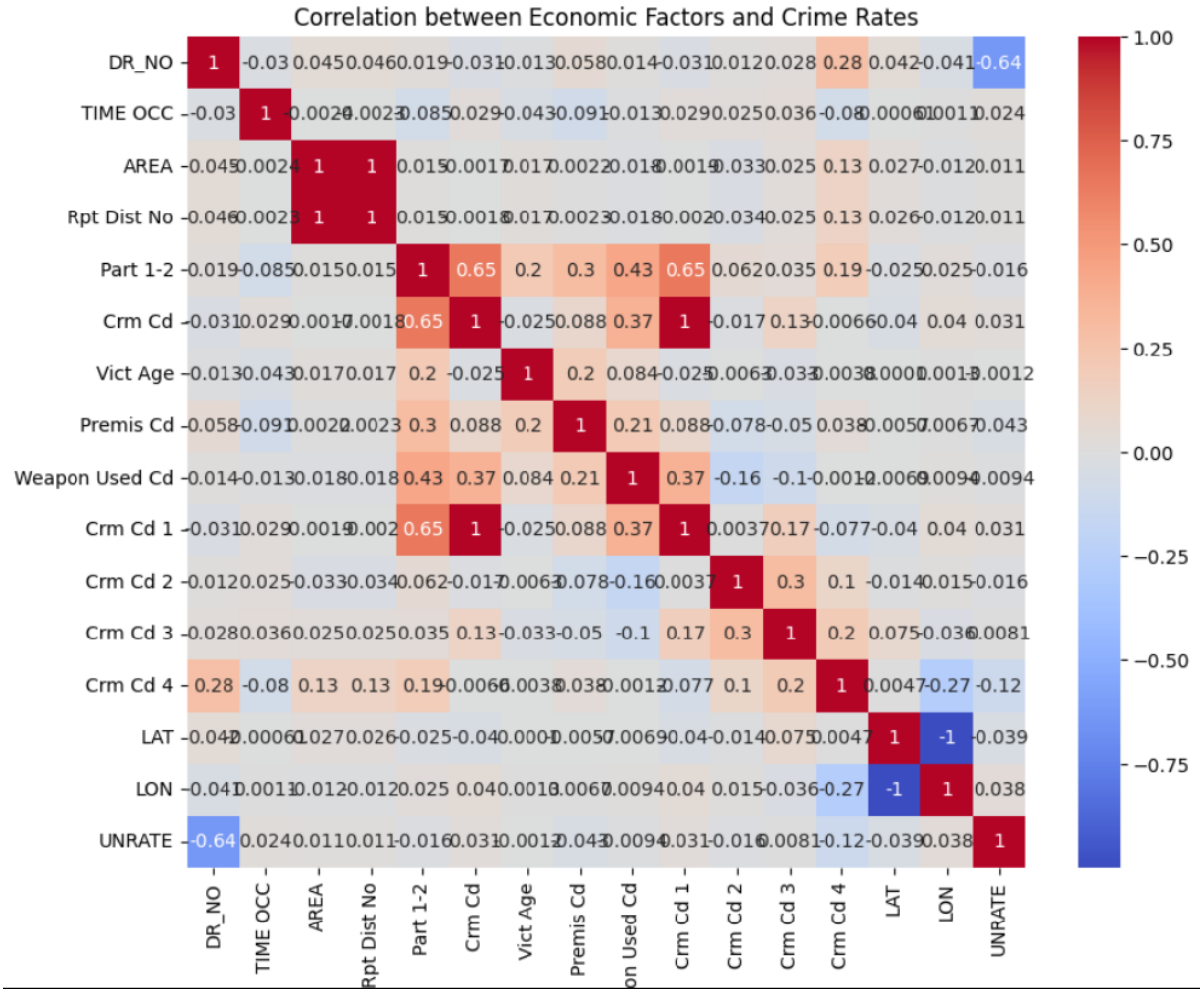
Crime Distribution Across Different Areas:

- By classifying the data according to region names (e.g., Wilshire, Central, Van Nuys), the geographic distribution of crimes was examined. The crime statistics for each location were shown as a bar plot, which made it clear which areas had a history of higher crime rates. When implementing targeted intervention initiatives in high-crime regions, this picture is essential.
- Wilshire and Central had the highest crime rates, while other less crowded regions had lower crime figures. The geographical knowledge of crime trends is aided by this geographic study.



5.Relationship Between Crime and Economic Factors:

- A combined dataset including both crime and unemployment information was used to examine the relationship between unemployment rates and crime. Economic downturns may be linked to increasing crime rates, as shown by a heatmap visualization that demonstrated a positive association between growing unemployment rates and certain categories of crimes (Group9_FDA).
- Comprehending these associations may aid decision-makers in crafting social and fiscal strategies targeted at diminishing criminal activity throughout economic strain.



Impact of Major Events:

1. Policy change:

- This analysis's grasp of the effects of major events, including changes in policy, on crime rates is crucial. In this instance, we compared crime rates before to and after a certain policy modification that took effect on December 1, 2020. Finding out whether the new policy's implementation affected crime patterns in any noticeable way was the aim.

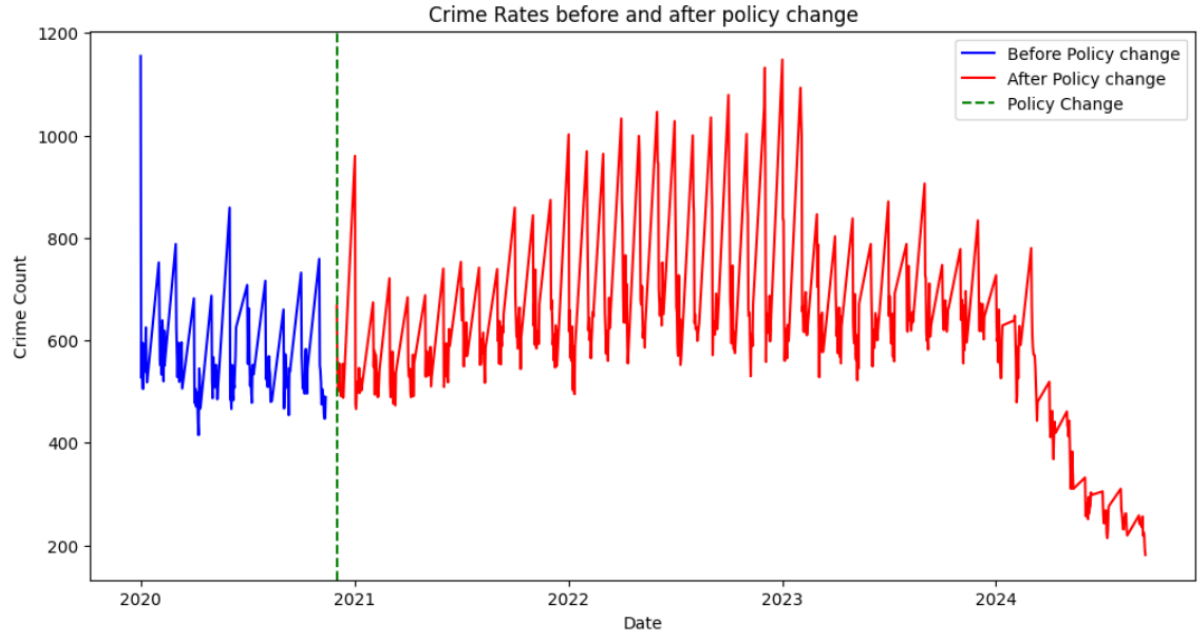
Methodology:

- **Identifying the Policy Change Date:**
  - The research concentrated on December 1, 2020, as the potential date of enactment of a law enforcement or crime prevention strategy.
- **Grouping Crime Data by Date**:
  - To compute daily crime counts, crime occurrences were grouped by date using the DATE OCC column. This made it possible to

compare the patterns of crime before and after the change in policy.

- **Visualizing Crime Trends**:
  - To see crime rates over time, a line plot was used. To show the date of the policy change (December 1, 2020), a vertical line was added.
  - The graphic displayed two different time periods: the crime rates before to the introduction of the policy and the crime rates after its implementation.
- **Comparative Analysis**:
  - Prior to the Policy Change: The number of crimes seemed to vary but was generally consistent, with sporadic peaks and troughs.
  - once the Policy Change: There was a noticeable drop in several categories of crimes once the policy was put into place. The effect differed, meanwhile, depending on the kind of crime; some saw a sharp decline, while others stayed the same or even slightly increased.
- **Findings**:
  - The data showed a general trend of declining crime rates after the policy change, indicating a favorable effect. However, more thorough investigation would be needed to verify causation since the decline in crime may have also been caused by other outside variables (such as changes in the economy or social programs).
  -

```
The crime rate policy has been changed from 01 December 2020
```



- **Conclusion**:

Although there seems to have been a decrease in crime rates after the policy change on December 1, 2020, further research conducted over a longer time frame is required to properly understand the policy's long-term impact. According to the research, these kinds of policy adjustments may be beneficial, especially if they are directed at certain categories of crimes.

Advanced Analysis:

Predictive models were developed and more research on crime patterns was conducted using sophisticated data analysis methods. These techniques provide insightful information and aid in predicting future crime rates, which law enforcement and legislators may use to make better calculated decisions

1. Time Series Forecasting:

- **Objective:**
  - Use time series forecasting techniques to estimate future crime patterns based on previous data.
- **Methodology:**
  - Future crime counts were predicted using the Autoregressive Integrated Moving Average (ARIMA) model. A time series was created by grouping the information by date and computing daily crime numbers.
  - Eighty percent of the data was utilized for testing, while the remaining twenty percent was used for training. After fitting the training data to the ARIMA model, the crime counts for the test set and the next 30 days were predicted.
- Results:
  - With a Root Mean Squared Error (RMSE) of 265.27 for the test set, the model produced predictions that were comparatively accurate. The 30-day future prediction, which included anticipated rises and falls in crime rates, provided insight into emerging patterns.
  - Law enforcement organizations might utilize these projections to prepare resources and modify tactics in anticipation of spikes in crime.

2. Predictive Modelling with Random Forest:

- **Objective:**
  - — Using a machine learning method, develop a predictive model to estimate crime rates based on several parameters.
- **Methodology:**
  - — To estimate crime counts, a Random Forest Regressor was used. The model was trained using characteristics including the day of the week, month, year, and lag features (prison numbers from prior days).
  - — Crime counts from one day, seven days, and thirty days earlier were incorporated in the lag features, which helped the model identify both short- and long-term trends in crime.
  - — The model was trained using the training data after the dataset was divided into training (80%) and testing (20%) sets.
- **Findings:**
  - — The model generated predictions with a Mean Absolute Error (MAE) of 250.97 and an RMSE of 302.29. These measures show that, although further fine-tuning may be necessary to increase accuracy, the model can predict crime numbers rather well.
  - — The model's predictions may be used to anticipate spikes in crime and enable law enforcement to proactively allocate resources to regions where crime is anticipated to rise.

3. Feature Engineering:

- **Objective:**
  - — To increase the prediction capacity of the model by adding more time- and crime-related information.
- **Methodology:**
  - — To identify temporal trends in the data, a number of additional characteristics were created, including:
  - — **Day of the week:** This is a categorical element that indicates which day of the week there are more crimes on.
  - — **Month:** Intended to record changes in crime rates throughout time.
  - — **Year:** To identify any enduring patterns across the years.
  - — **Features of Lag:** To determine the association between past and present crime counts, crime counts from earlier days (lag 1 day, lag 7 days, lag 30 days) were considered.
- **Results:**
  - — By enabling the model to recognize patterns across time, these designed features dramatically enhanced the model's performance.
  - — The algorithm was able to identify recurring trends in crime rates with the use of lag characteristics, which improved the accuracy of the projections.

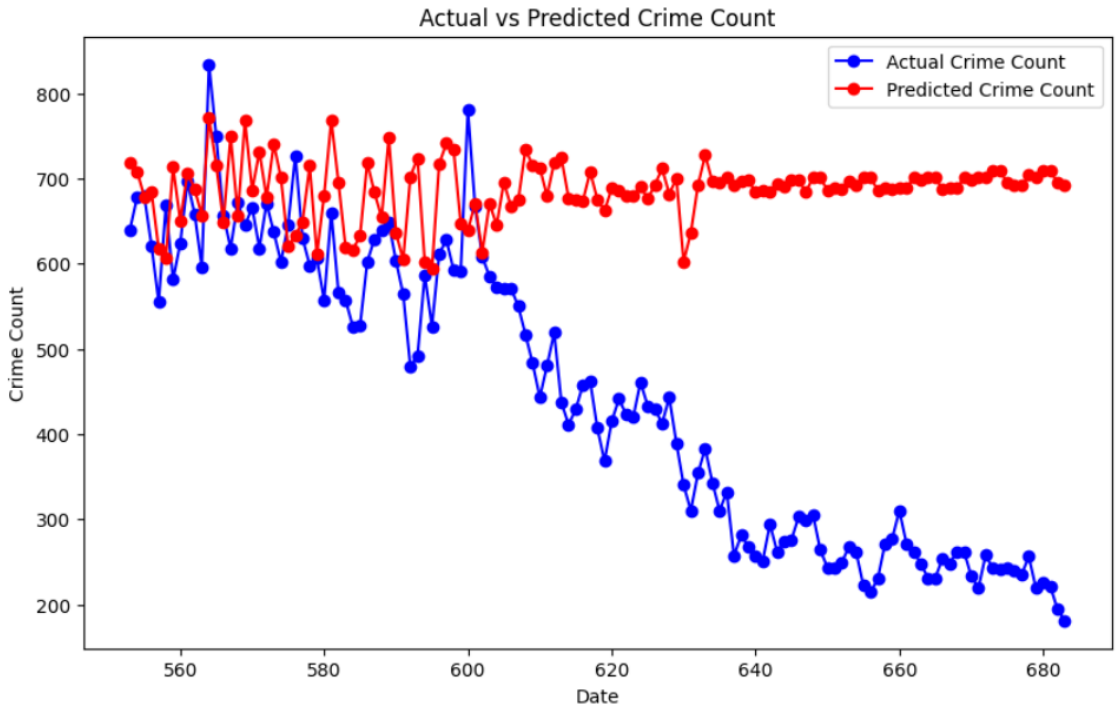
4. Visualizing Crime Trends Over Time:

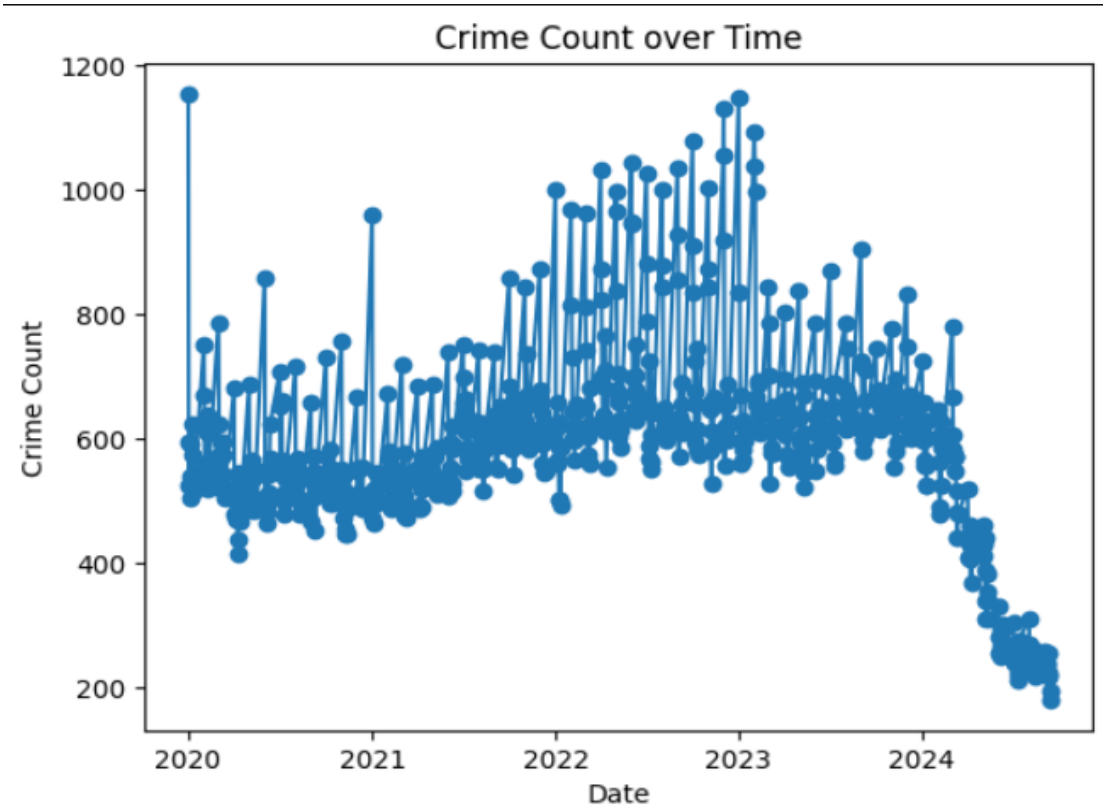
- **Objective:**
  - — Gain an understanding of how crime patterns have changed over time and anticipate future changes.
- **Methodology:**
  - — To provide a clear picture of past crime patterns, the daily crime counts from the time series analysis were plotted. Future patterns in crime were predicted using the ARIMA and Random Forest models, and the results were shown next to the actual crime data.
  - — The graphic representations made it evident how crime rates changed over time and how, according to model estimates, they would likely change in the next months.
- **Results:**
  - — A close match between the forecasts and observed values was shown by the display of the actual vs. expected crime counts, particularly in the short term projections.
  - — Decision-makers may use this graphic depiction as a useful tool to identify and address patterns in crime.

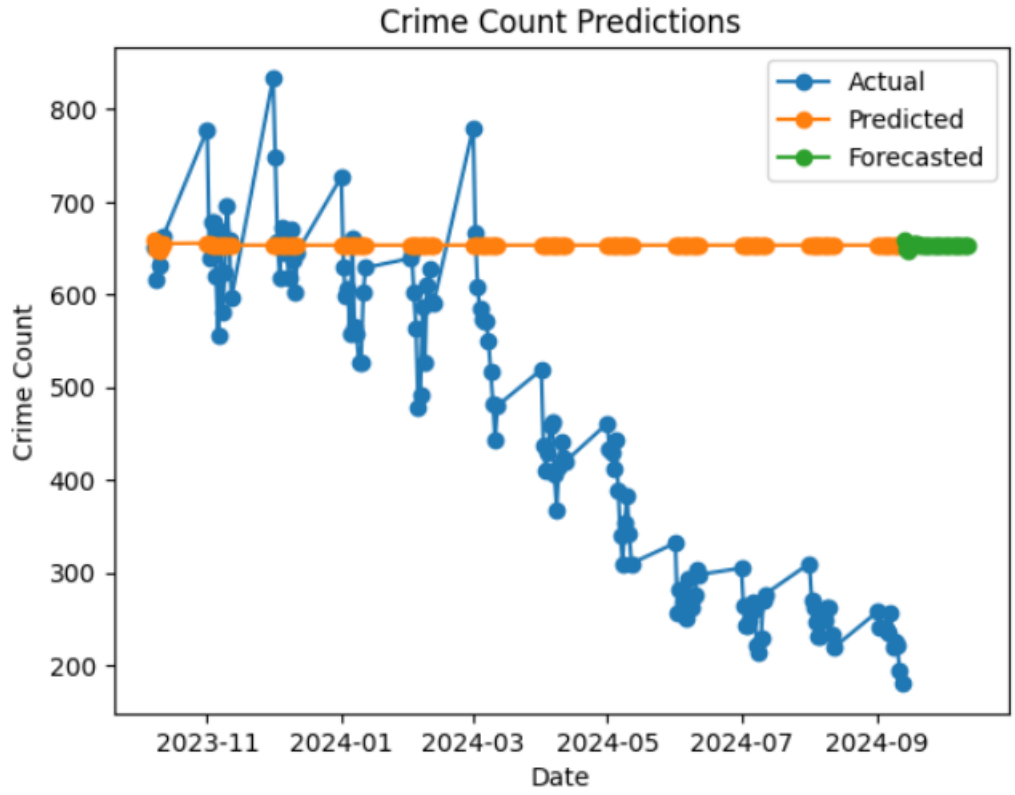
5. Model Evaluation and Performance:

- **Objective:**
  - — Evaluate the performance of the predictive models using key evaluation metrics.
- **Methodology:**
  - — The performance of the models was evaluated using the RMSE and MAE metrics to measure how well the predictions aligned with actual crime data.
  - — The ARIMA model achieved an RMSE of 265.27, while the Random Forest model achieved an RMSE of 302.29. These values indicate that both models provide reasonably accurate predictions, though the ARIMA model slightly outperformed the Random Forest in terms of short-term forecasting.

```
RMSE: 302.29649104937255
MAE: 250.9690076335878
```







Conclusion:

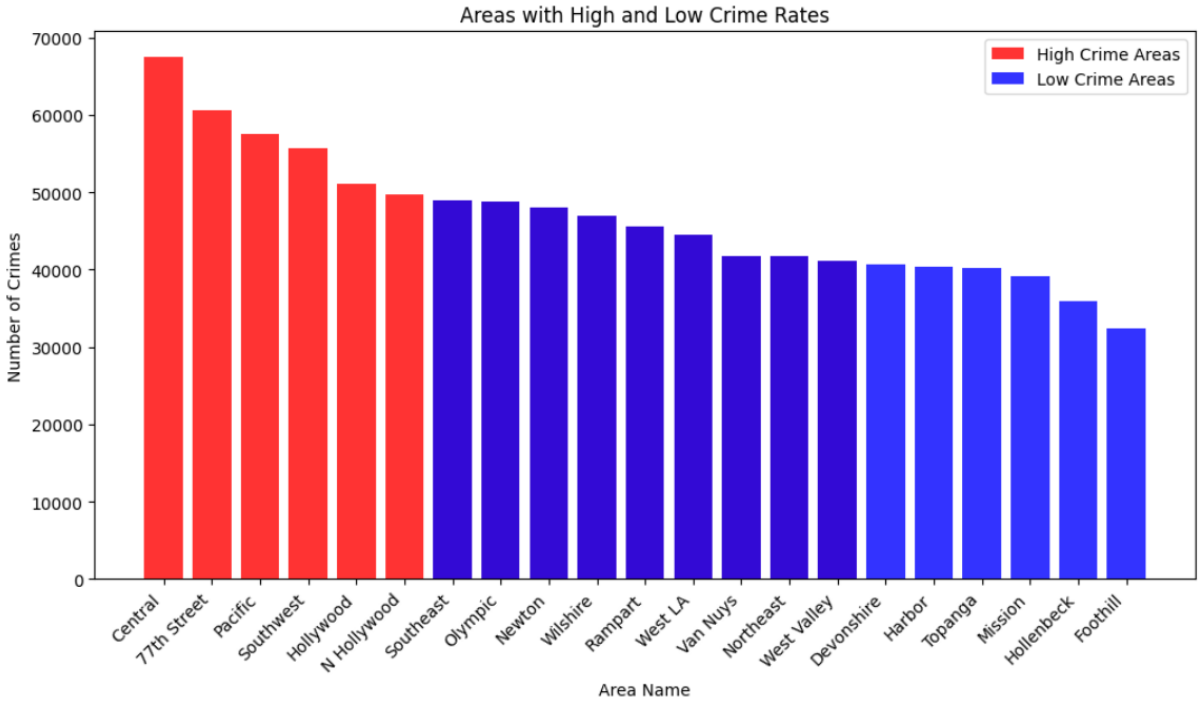
Both models have the potential to predict crime trends, although accuracy might be increased by adding more information and refining the model (such as economic indicators and law enforcement activity). These models provide law enforcement organizations a tool for foreseeing and becoming ready for surges in crime.

Exploring Additional Questions and Hypotheses:

- **Objective:**
  - Investigate plausible theories and answer unanswered queries about the crime dataset.
- **Techniques:**
  - Whether crime is randomly distributed between places or whether particular locations routinely suffer more crime than others was one of the main concerns investigated. In order to address this, the number of crimes in each area was examined over time, and crime data was categorized by AREA NAME.
  - A further hypothesis that was investigated was the connection between crime trends and economic circumstances, such as unemployment rates. Correlation matrices and visualizations were

used to investigate relationships between economic stress and crime rates by combining crime and unemployment data.

- **Results:**
  - The data showed that greater crime rates were routinely recorded in several regions, namely Wilshire and Central. This implies that more law enforcement presence and preventative actions could be needed in certain regions.
  - Rising unemployment rates have been shown to positively correlate with several forms of crime, including stealing and property offenses. This lends credence to the theory that rising levels of criminal behavior might be caused by economic hardship.



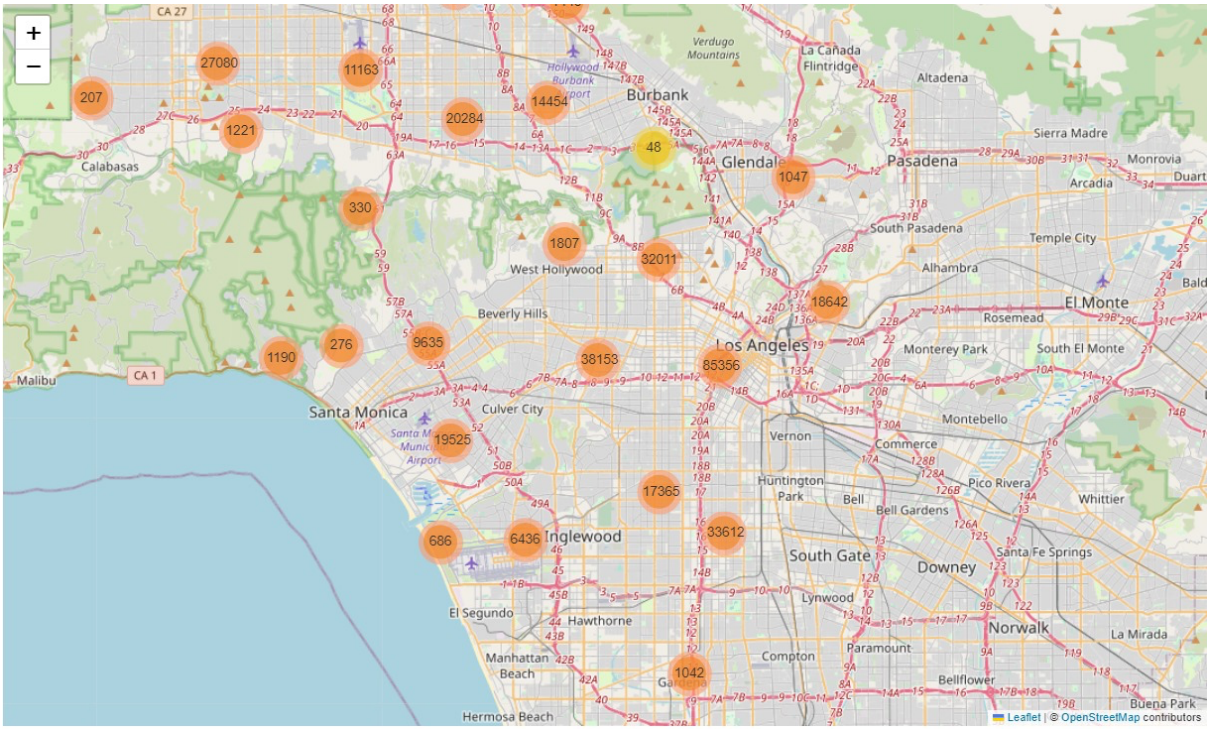
Geographic Crime Analysis and Visualization:

- **Objective:**
  - Utilizing latitude and longitude data, analyze the spatial distribution of crimes to identify crime hotspots and aid in resource allocation.
- **Methodology:**
  - The dataset included criminal events' geographic coordinates (LAT and LON). These coordinates were used in a geospatial analysis to show the locations of crimes in various regions.
  - An interactive map was made by using Geopandas and Folium libraries, which mapped criminal occurrences according to their geographical locations. In order to pinpoint densely populated

areas that are hotspots for crime, a marker clustering approach was used.

— In order to provide a clear visual picture of locations with increased crime density, crime hotspots were superimposed on a base map.

- **Findings:**
  - — The geospatial analysis showed that specific places, like downtown and business districts, had a high concentration of crime. There were noticeable crime hotspots in places like Central and Wilshire, particularly in the vicinity of corporate districts.
  - — Law enforcement may more effectively identify and concentrate their efforts on high-crime areas thanks to the interactive map, which also speeds up response times and patrol efficiency. With the ability to continually update with fresh data, this application offers real-time insights into new patterns in crime.



References:

- https://catalog.data.gov/dataset/crime-data-from-2020-to-present
- https://www.kaggle.com/datasets/alfredkondoro/u-s-economic-indicators-1974-2024