PROJECT 1
CLEANING AND ANALYZING CRIME DATA

IE6400 Foundations Data Analytics Engineering Fall Semester 2023

GROUP 22

Divyia Venkat Eachampatti Thirunavukarasu
eachampattithiruna.d@northeastern.edu

Pramoth Guhan
guhan.p@northeastern.edu

Rohith Adhitya Chinnannan Rajkumar
chinnannanrajkumar.r@northeastern.edu

Submission Date: 11/03/2023

INTRODUCTION

Crime data analysis and availability are important in numerous aspects of public life. The foundation for maintaining public safety and stopping criminal activity is crime data. This information is used by local governments and law enforcement to identify crime patterns and identify problem regions, which facilitates the effective use of resources and the early detection and resolution of new problems.

Crime statistics facilitate analysis and research while providing scholars, criminologists, and researchers with insightful information. It aids in understanding the underlying causes of crime, assessing the effectiveness of interventions, and examining the influence of social and economic factors. Moreover, businesses and individuals utilize crime statistics to make informed decisions about where to establish operations or choose a place to reside. Areas with low crime rates often become more attractive for investment and real estate development.

The dataset we are working with, "Crime Data from 2020 to Present," offers a comprehensive view of incidents of crime within the City of Los Angeles. The dataset provides a valuable resource for understanding and analyzing crime trends and patterns in the region. It is essential to acknowledge that the information present in this dataset has been transcriptions extracted from initial criminal reports that were handwritten on paper. Consequently, there may be some inaccuracies within the data. In addition, address fields have been expanded to the closest hundred blocks in order to protect privacy.

Furthermore, to gain deeper insights into the dataset, we will proceed to analyze its various aspects, including seasonal patterns, crime frequencies, demographic factors, and even predictive modeling. The subsequent sections of this report will provide a comprehensive exploration of the "Crime Data from 2020 to Present," offering valuable perspectives on crime trends, public safety, and the real-world applications of data-driven decision-making.

DATA SOURCES

The dataset, "Crime Data from 2020 to Present," offers a comprehensive overview of crime incidents in the City of Los Angeles from 2020 to the present day. With 28 columns, each providing specific information about these incidents, it serves as a foundational resource for gaining insights into crime trends, patterns, and demographic factors, all of which are critical for enhancing public safety and supporting informed decision-making.

The dataset's columns provide a wealth of data, ranging from internal reference numbers for crime reports to critical details about each incident, such as dates, times, locations, and descriptions of crimes. While the dataset is rich in information, it's important to note that some columns contain missing data. The analysis's completeness is impacted by these missing values. In order to guarantee the precision and dependability of any insights or reports generated from the dataset, it is essential to address this missing data.

Despite these challenges, the dataset plays a pivotal role in advancing our understanding of crime incidents, contributing to enhanced public safety measures, research efforts, and data-driven decision-making processes. It is an invaluable tool that enables the community, researchers, legislators, and law enforcement to collaborate in the goal of making the City of Los Angeles safer and more secure.

TASKS

1. Data Acquisition:

Download the dataset from the provided link and load it into your preferred data analysis tool

```
[176] import pandas as pd
      import matplotlib.pyplot as plt
      import seaborn as sns
      import numpy as np
```

```
#Loading the dataset
import pandas as pd
df = pd.read_csv('/content/Crime_Data_from_2020_to_Present.csv')
```

2. Data Inspection:

- **Display the first few rows of the dataset.**

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Crm Cd 4	LOCATION	Cross Street	LAT	LON	Year	Month	Crime Category	Day We
0	220120712.5	2020-01-08	2020-01-08	1900-01-01 22:30:00	3.0	Southwest	377.0	2.0	624.0	BATTERY - SIMPLE ASSAULT	...	0.0	1100 W 39TH PL	Not Available	34.0141	-118.2978	2020	1	Violent Crimes	Wednesd
1	220120712.5	2020-01-02	2020-01-01	1900-01-01 03:30:00	1.0	Central	163.0	2.0	624.0	BATTERY - SIMPLE ASSAULT	...	0.0	700 S HILL ST	Not Available	34.0459	-118.2545	2020	1	Violent Crimes	Wednesd
2	200110444.0	2020-04-14	2020-02-13	1900-01-01 12:00:00	1.0	Central	155.0	2.0	845.0	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	...	0.0	200 E 6TH ST	Not Available	34.0448	-118.2474	2020	2	Other Crimes	Thursd
3	220120712.5	2020-01-01	2020-01-01	1900-01-01 17:30:00	15.0	N Hollywood	1543.0	2.0	745.0	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	...	0.0	5400 CORTEEN PL	Not Available	34.1685	-118.4019	2020	1	Other Crimes	Wednesd
4	220120712.5	2020-01-01	2020-01-01	1900-01-01 04:15:00	19.0	Mission	1998.0	2.0	740.0	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	...	0.0	14400 TITUS ST	Not Available	34.2198	-118.4468	2020	1	Other Crimes	Wednesd

5 rows × 33 columns

We can see that the dataset contains 33 columns (DR_NO, Date Rptd, DATE OCC, TIME OCC, AREA, AREA NAME, etc.,).

- **Check the data types of each column.**

```
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
```

We can see the data types of each column. The data types in this data set are, int, float and object.

- **Review column names and descriptions, if available.**

```
Index(['DR_NO', 'Date Rptd', 'DATE OCC', 'TIME OCC', 'AREA', 'AREA NAME',
       'Rpt Dist No', 'Part 1-2', 'Crm Cd', 'Crm Cd Desc', 'Mocodes',
       'Vict Age', 'Vict Sex', 'Vict Descent', 'Premis Cd', 'Premis Desc',
       'Weapon Used Cd', 'Weapon Desc', 'Status', 'Status Desc', 'Crm Cd 1',
       'Crm Cd 2', 'Crm Cd 3', 'Crm Cd 4', 'LOCATION', 'Cross Street', 'LAT',
       'LON'],
      dtype='object')
```

All the 33 columns are being displayed. The meaning for each column has been described above in the report.

3. Data Cleaning:

- **Identify and handle missing data appropriately.**

```
Missing Data:
 DR_NO                0
Date Rptd             0
DATE OCC              0
TIME OCC              0
AREA                  0
AREA NAME             0
Rpt Dist No           0
Part 1-2              0
Crm Cd                0
Crm Cd Desc           0
Mocodes          114148
Vict Age              0
Vict Sex         108529
Vict Descent     108537
Premis Cd            10
Premis Desc         488
Weapon Used Cd   537498
Weapon Desc      537498
Status                0
Status Desc           0
Crm Cd 1             10
Crm Cd 2         764505
Crm Cd 3         823173
Crm Cd 4         825151
LOCATION              0
Cross Street     693343
LAT                   0
LON                   0
dtype: int64
```

```
DR_NO             0
Date Rptd         0
DATE OCC          0
TIME OCC          0
AREA              0
AREA NAME         0
Rpt Dist No       0
Part 1-2          0
Crm Cd            0
Crm Cd Desc       0
Mocodes           0
Vict Age          0
Vict Sex          0
Vict Descent      0
Premis Cd         0
Premis Desc       0
Weapon Used Cd    0
Weapon Desc       0
Status            0
Status Desc       0
Crm Cd 1          0
Crm Cd 2          0
Crm Cd 3          0
Crm Cd 4          0
LOCATION          0
Cross Street      0
LAT               0
LON               0
dtype: int64
```

We can see that the columns, 'Mocodes', 'Vict Sex', 'Vict Descent', 'Pemid Cd', 'Premis Desc', 'Weapon Used Cd', 'Weapon Desc', 'Crm Cd 1-4', and 'Cross Street' have missing values. Which were handled using fillna() a]since all these data were necessary for future analysis. The right side image depicts no missing values after handling.

- **Check for and remove duplicate rows.**

```
Duplicate Rows except first occurrence: Empty DataFrame
Columns: [DR_NO, Date Rptd, DATE OCC, TIME OCC, AREA, AREA NAME, Rpt Dist No, Part 1-2, Crm Cd, Crm Cd Desc, Mocodes, Vict Age, Vict Sex, Vict Descent, Pre
Index: []

[0 rows x 28 columns]
```

Our dataset did not have any duplicate rows.

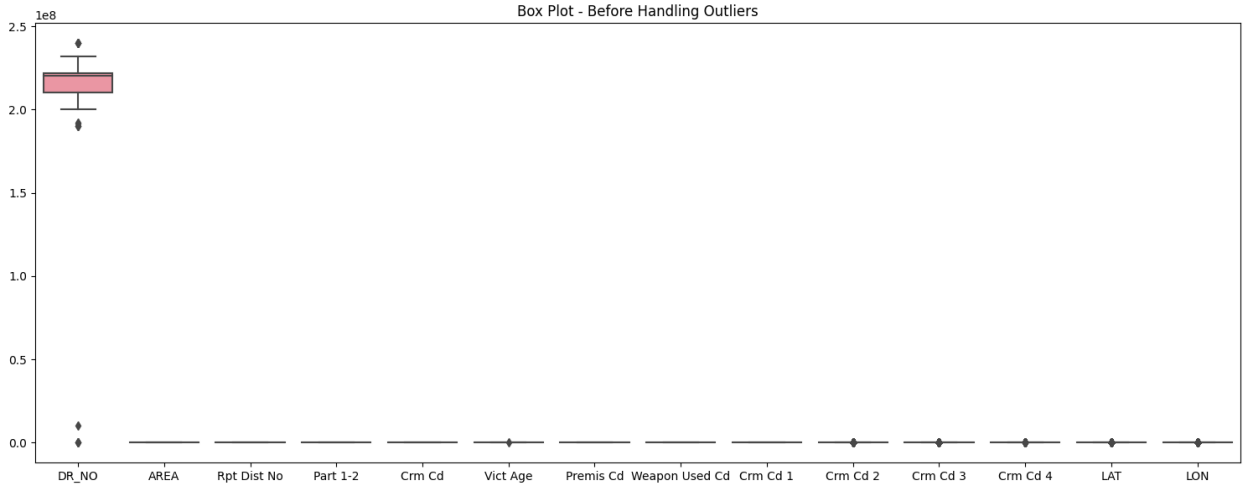
- **Convert data types if needed (e.g., dates to date format, numerical values to appropriate numeric types).**

```
[188] #3.3 Converting data types if needed (e.g., dates to date format, numerical values to appropriate numeric types).

      df['Date Rptd'] = pd.to_datetime(df['Date Rptd'])
      df['DATE OCC'] = pd.to_datetime(df['DATE OCC'])
      df['TIME OCC'] = pd.to_datetime(df['TIME OCC'], format='%H%M', errors='coerce')
      df['Premis Cd'] = df['Premis Cd'].astype(int)
      df['Weapon Used Cd'] = df['Weapon Used Cd'].astype(int)
```

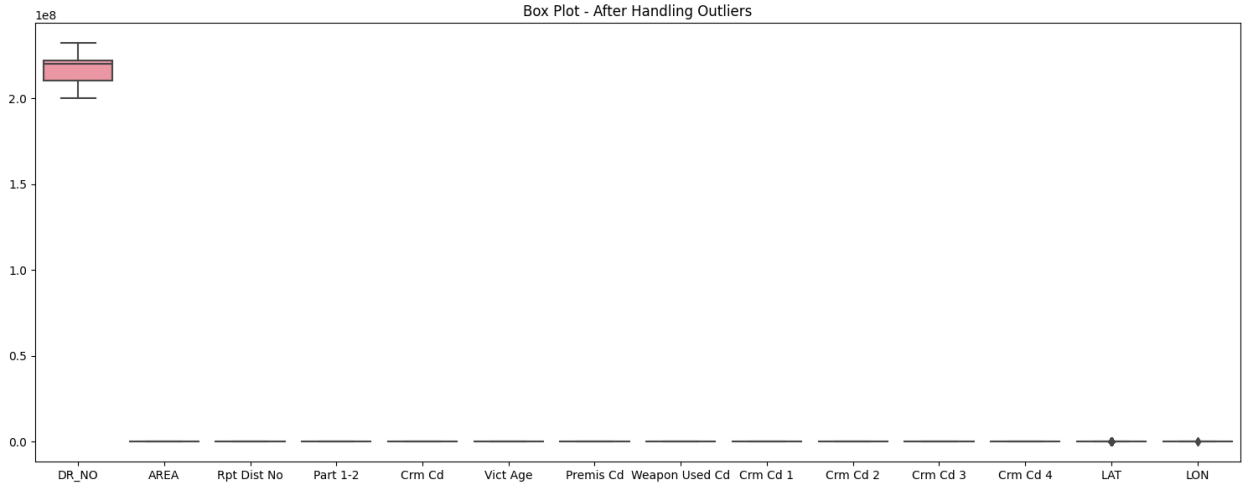
We have converted Date Rptd, DATE OCC and TIME OCC to DateTime. Premis Cd, and Weapon Used Cd to INT data types.

- **Deal with outliers if relevant to your analysis.**

This graph shows the outliers before handling them.



This graph shows after handling the outliers.



- **Standardize or normalize numerical data as necessary.**

```
[192] #Standardize or normalize numerical data as necessary.

      #from sklearn.preprocessing import StandardScaler

      #numerical_cols = df.select_dtypes(include=[np.number]).columns
      #scaler = StandardScaler()
      #df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
```

We normalized the data initially, but we felt it was not necessary for our analysis, so we commented the normalization.

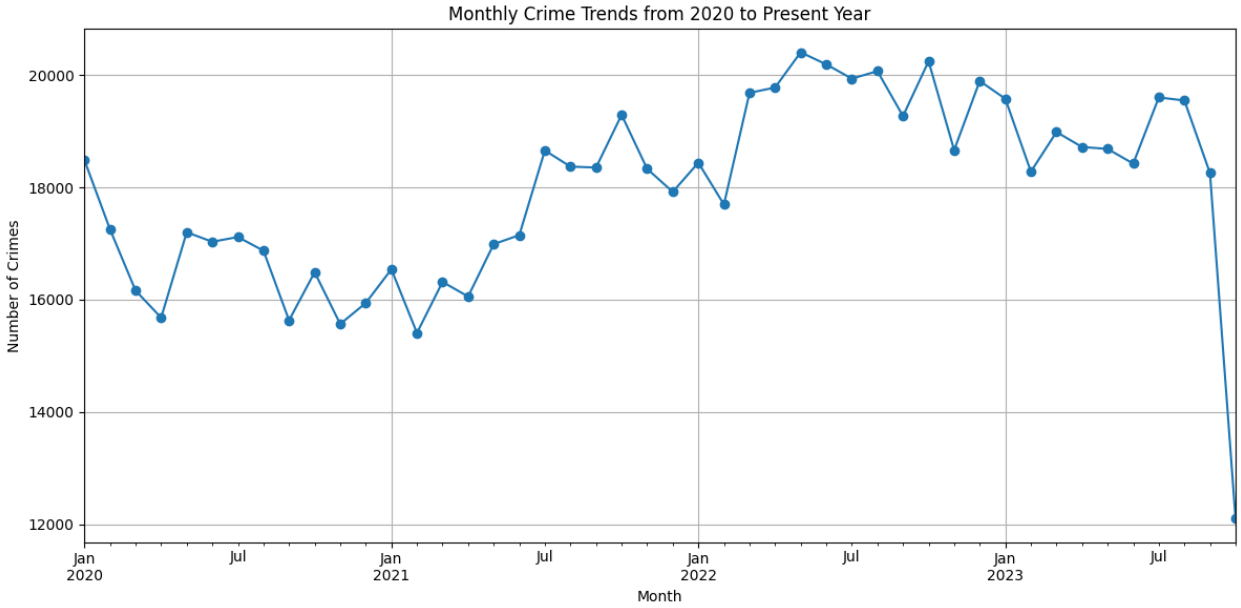
- **Encode categorical data if present.**

```
#Encode categorical data if present.

#from sklearn.preprocessing import LabelEncoder

#l = LabelEncoder()

# Apply label encoding to a specific ordinal column
#df['AREA NAME'] = l.fit_transform(df['AREA NAME'])
#df['Crm Cd Desc'] = l.fit_transform(df['Crm Cd Desc'])
#df['Vict Sex'] = l.fit_transform(df['Vict Sex'] )
#df['Vict Descent'] = l.fit_transform(df['Vict Descent'])
#df['Premis Desc'] = l.fit_transform(df['Premis Desc'])
#df['Weapon Desc'] = l.fit_transform(df['Weapon Desc'])
#df['Status'] = l.fit_transform(df['Status'])
#df['Status Desc'] = l.fit_transform(df['Status Desc'])
#df['LOCATION'] = l.fit_transform(df['LOCATION'])
#df['Cross Street'] = l.fit_transform(df['Cross Street'])
```

We used LabelEncoder for encoding initially, but for our analysis we were not able to acquire desired outputs, hence we commented it.

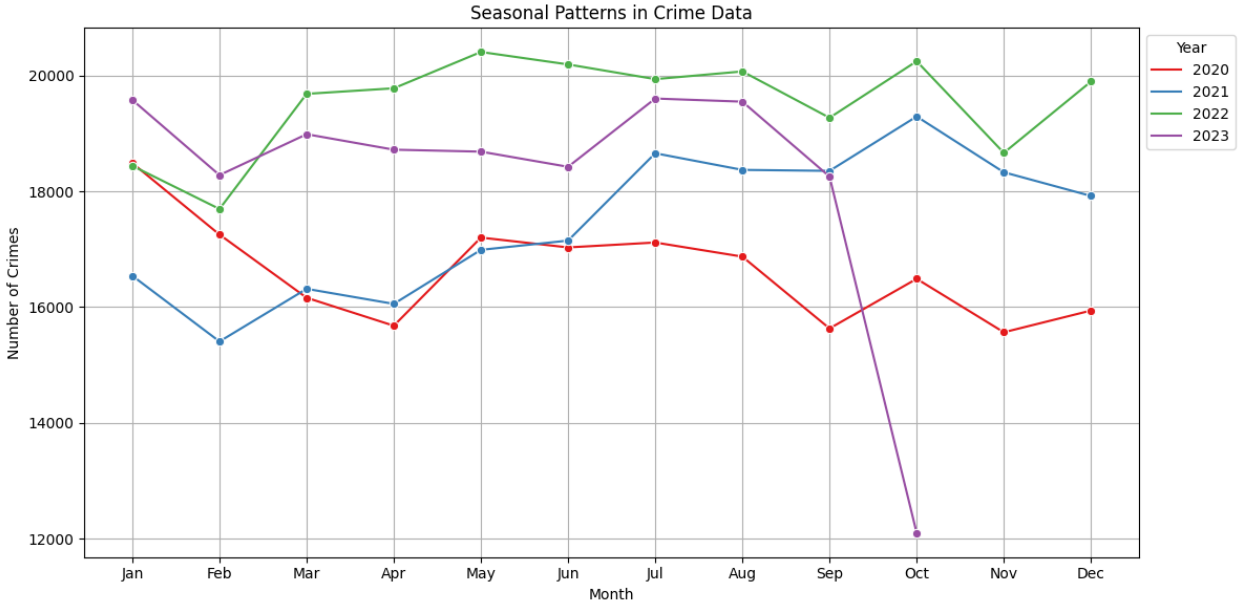
4. Exploratory Data Analysis (EDA):

- **Visualize overall crime trends from 2020 to the present year.**



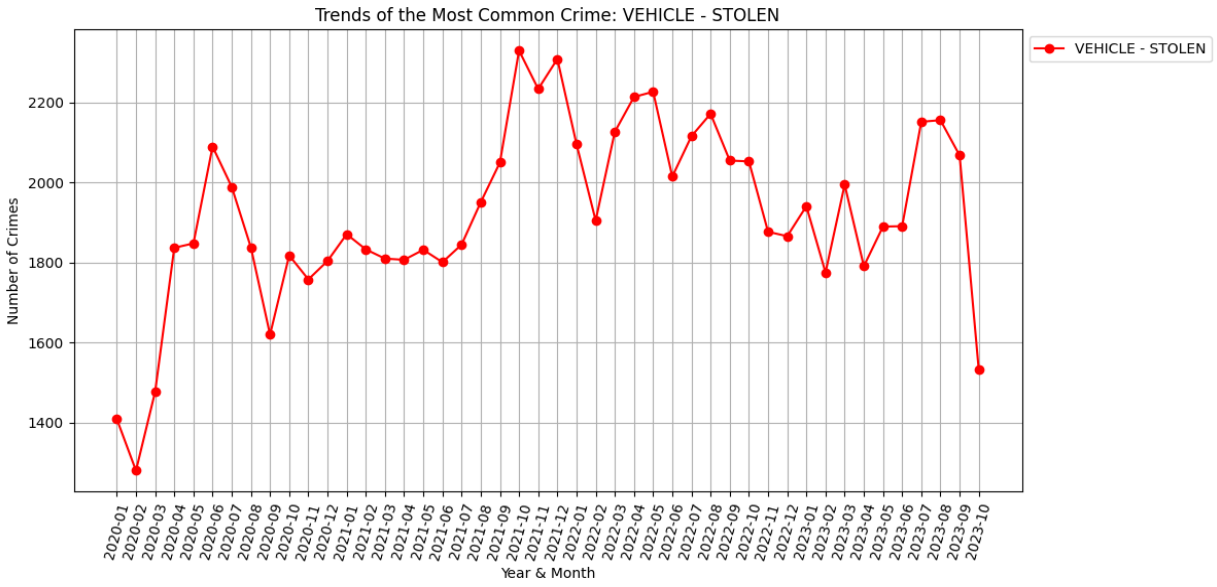
The above line chart represents the overall monthly average crime trends from Jan 2020 to Oct 2023 (present year), from the plot we can observe there is decline in crime rates from Jan 2020 -it was the subsequent effect of initial stages of Covid-19 lockdown, then the crime rates started increasing from May 2020. In May 2022 the crime rate attained its peak and from then there is a gradual decrease in crime rate. It is expected that the crime rate will decline in the upcoming future by observing the trend in the line chart.

- **Analyze and visualize seasonal patterns in crime data.**



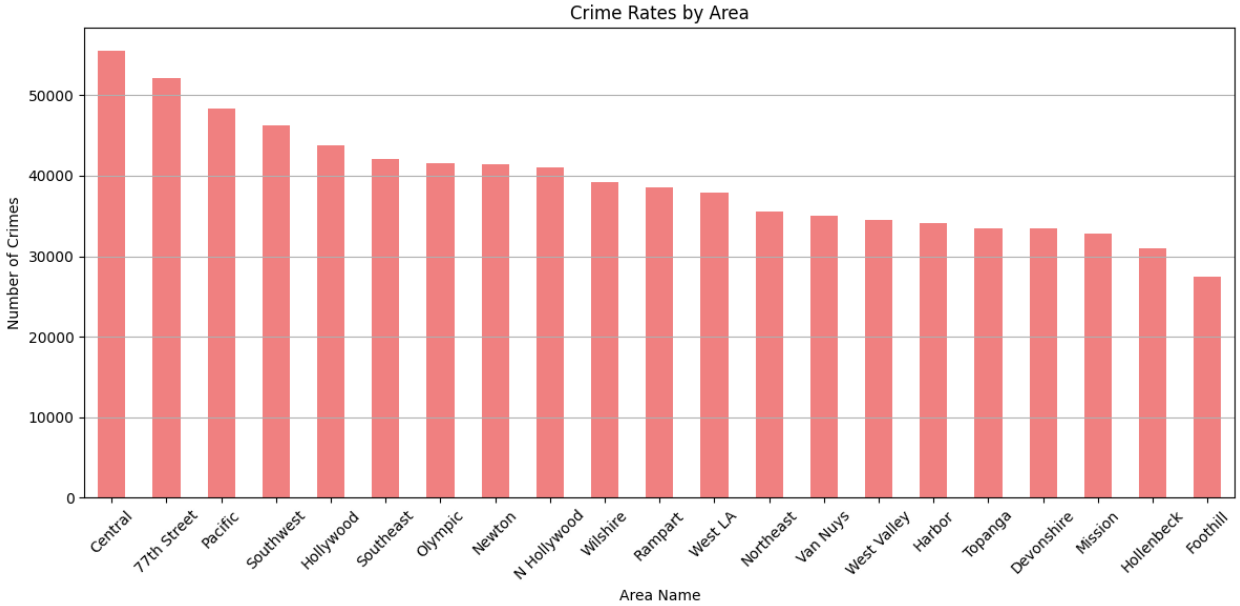
In the above line chart monthly pattern of crime rate is marked based on the calendar year from January to December. From the chart we can observe that the crime rates were at peak in the year 2022 where the crime rates were higher in 10 months out of 12 months.
It can be noted that in the month April to May there is a higher increase in crime rate in all the years and marginal decrease in months of February, September when compared to the respective previous months. It can be noted that crimes are both increasing and decreasing in the same pattern across months in all years from 2020 to 2023.

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



We have marked the highest crime reported Vehicle-Stolen's total on month wise from Jan 2020 to Oct 2023. In the plot we can observe that the Vehicle Stolen crime has a steep increase in two time periods. First is from Feb 2020 to June 2020 and second steep increase is from June 2021 to Oct 2021 where it attained its peak crime rate.
Vehicle theft is one of the major crimes occurring in the California region, in the city of Los Angeles from the dataset it can be observed that the highest recorded crime which can be seen started declining in the last few months based on the action by LAPD.

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



We can observe from the bar plot that the Central area has the highest crime rates and the Foothill area has the lowest crime rates in Los Angeles city of California. There is a major difference in crime rate between the Central and the Foothill area, but from highest to lowest crime rate area we can note that there is gradual decline in crime rates distributed among other areas in the city.

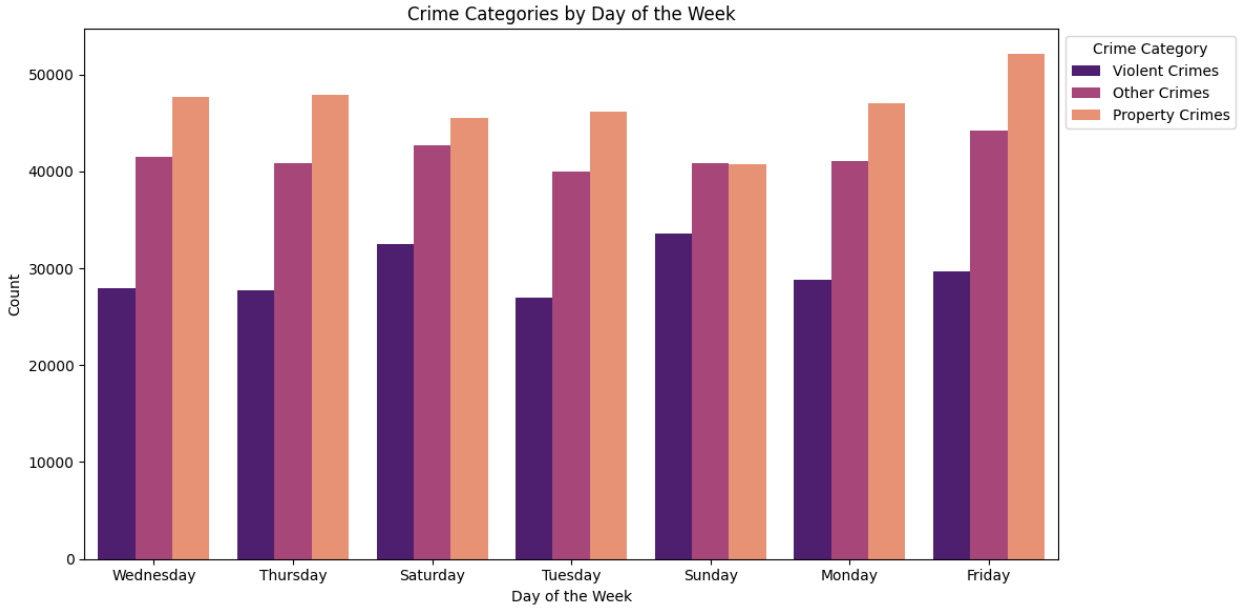
- **Explore correlations between economic factors (if available) and crime rates.**

```
#4.5 Exploring correlations between economic factors (if available) and crime rates.

economic_data = {
    'Year': ['2020', '2021', '2022', '2023'],
    'Median_Household_Income': [70643, 76800, 77339, 82545],
    'Median_Capita_Income': [63654, 65421, 68990, 73678 ]
}

economic_df  = pd.DataFrame(economic_data)

df['Year'] = df['Year'].astype(str)
economic_df['Year'] = economic_df['Year'].astype(str)
merged_df = pd.merge(df, economic_df, on='Year', how='left')
correlation = merged_df['Crm Cd'].corr(merged_df['Median_Household_Income'])
print(f'Correlation between "Crm Cd" and "Median_Household_Income": {correlation:.2f}')
```

```
Correlation between "Crm Cd" and "Median_Household_Income": -0.02
```

The Correlation between the economic factor median household income is compared with the crime rate in the above part of calculation and the correlation value derived is -0.02 which can be seen in the output.

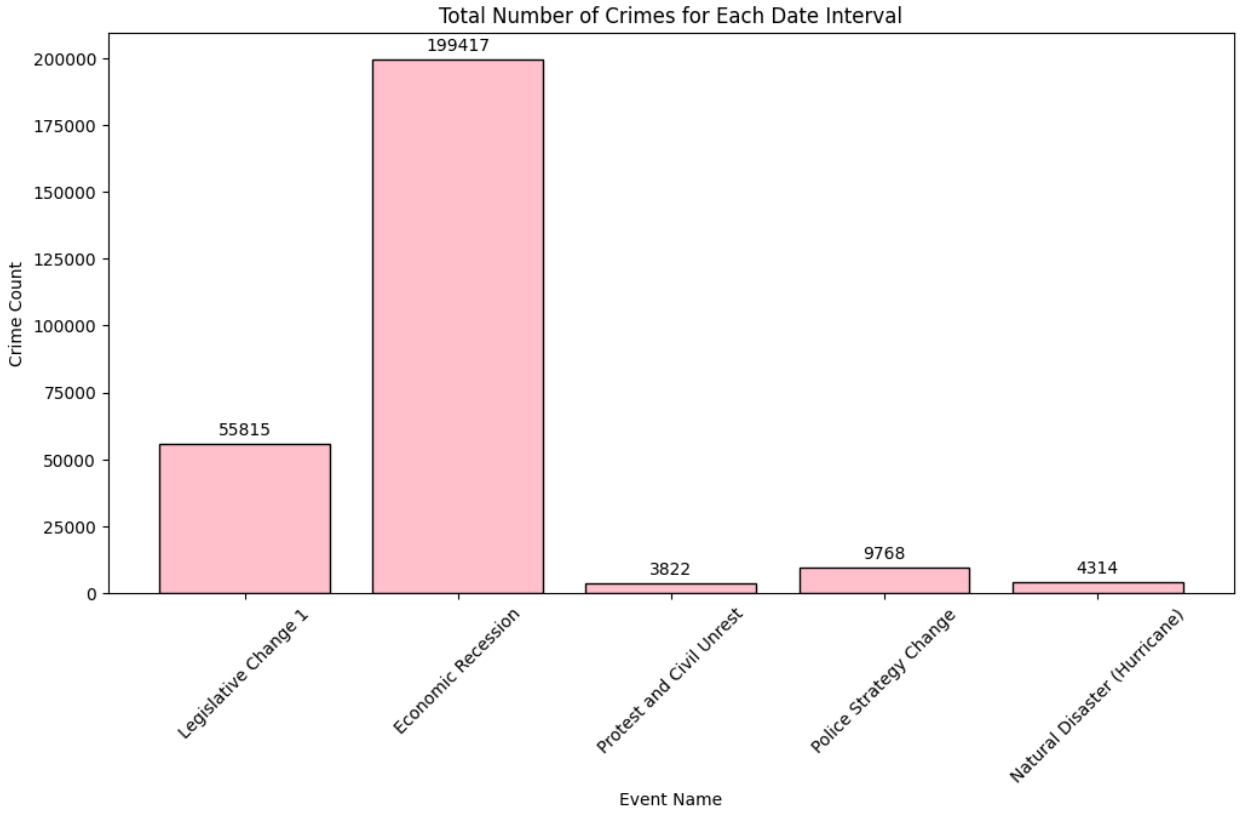
- **Analyze the relationship between the day of the week and the frequency of certain types of crimes.**



To visualize the crime rate among the days of the week, we have categorized the crimes into three major categories such as Violent crimes, Property crimes and Other crimes and then plotted their distributions among the week days. It can be interpreted that property crimes are the most occurring type of crime and violent crimes are the lowest crimes occurring in all week days. We can also

observe that other types of crimes category are approximately distributed in the same range of rate across all days.

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

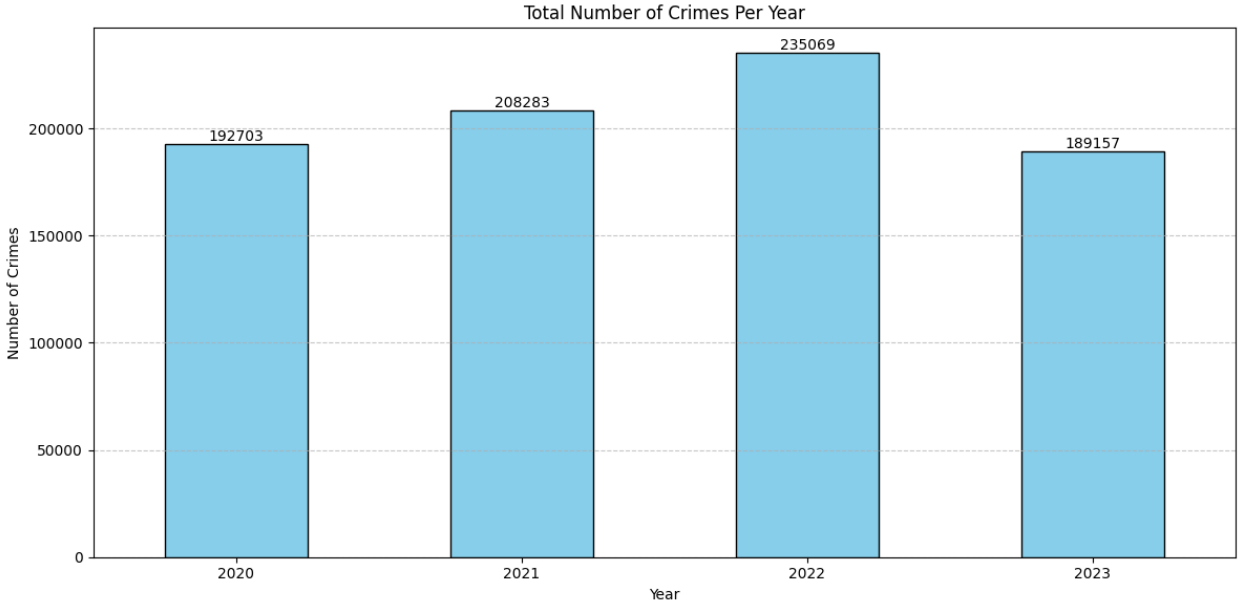


The above plot describes the crime rate distribution among a few major time intervals where these events may accompany the crime rate at that period of time. The events taken in this case were Legislative Change 1(Jan 2020 to Mar 2020), Economic Recession (Jan 2020 to Dec 2020), Protest and Civil unrest(Jun 1,2020 to Jun 7,2020), Police strategy change (Sep 15,2021 to Sep 30,2021) and Hurricane natural disaster (Aug 24,2020 to Aug 30,2020). It can be observed that during the economic recession time period in the year of 2020 that many crime records were reported and most of the crimes were directly or indirectly connected to the economic recession. All other periods reported a marginally higher crime rate even in the short period of time.

Find the solutions to these questions:

1. Overall Crime Trends:

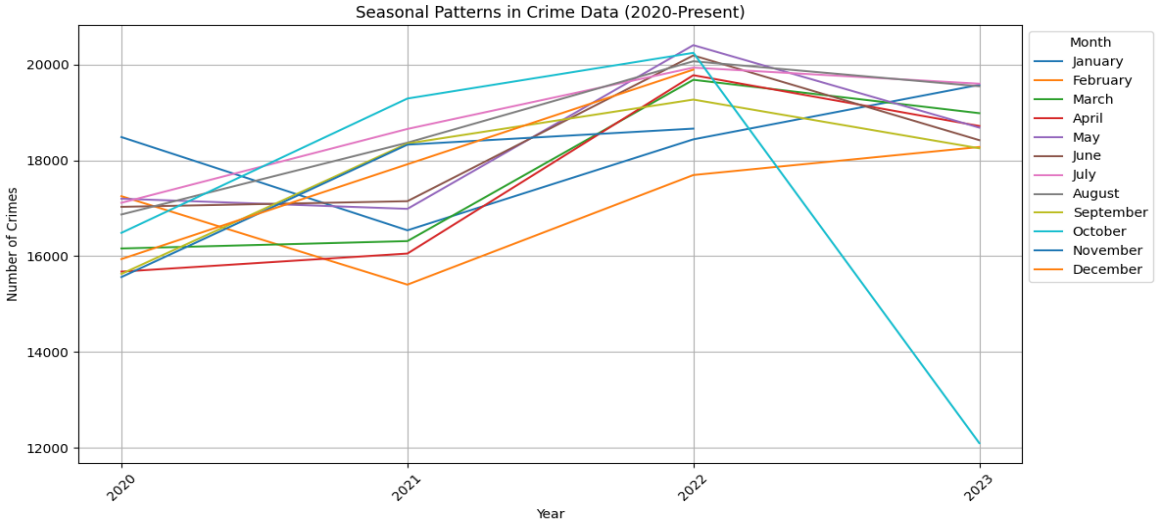
Calculate and plot the total number of crimes per year to visualize the trends.



This plot gives us an overview of the total number of crimes per year. We can see from the plot that the total number of crimes have been increasing from 2020. The year 2022 has been recorded highest in terms of total number of crimes with 2,35,069. The current year(2023) has a total number of crimes of 1,89,157. But since the dataset has values only till July 2023. Hence, we do not know whether it will surpass 2022.

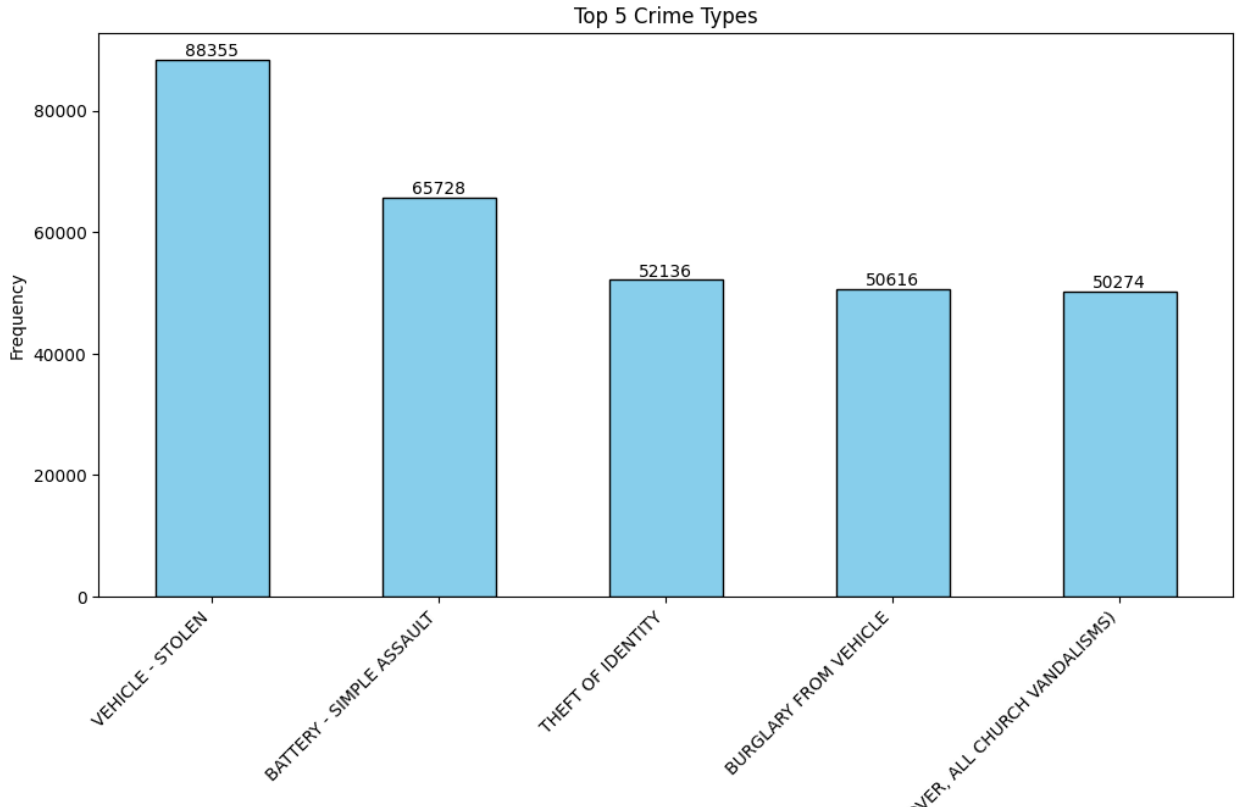
2. Seasonal Patterns:

Group the data by month and analyze the average number of crimes per month over the years.



This plot gives us an overview of seasonal patterns in crime data from 2020 to present. We can see various patterns from the graph, for example, January had the most number of crimes in 2020 but in the following year, October surpassed January with more number of crimes. In the year 2022, May had the most number of crimes. These are a few inferences from the graph. Similarly, we can get more insights from the plot.

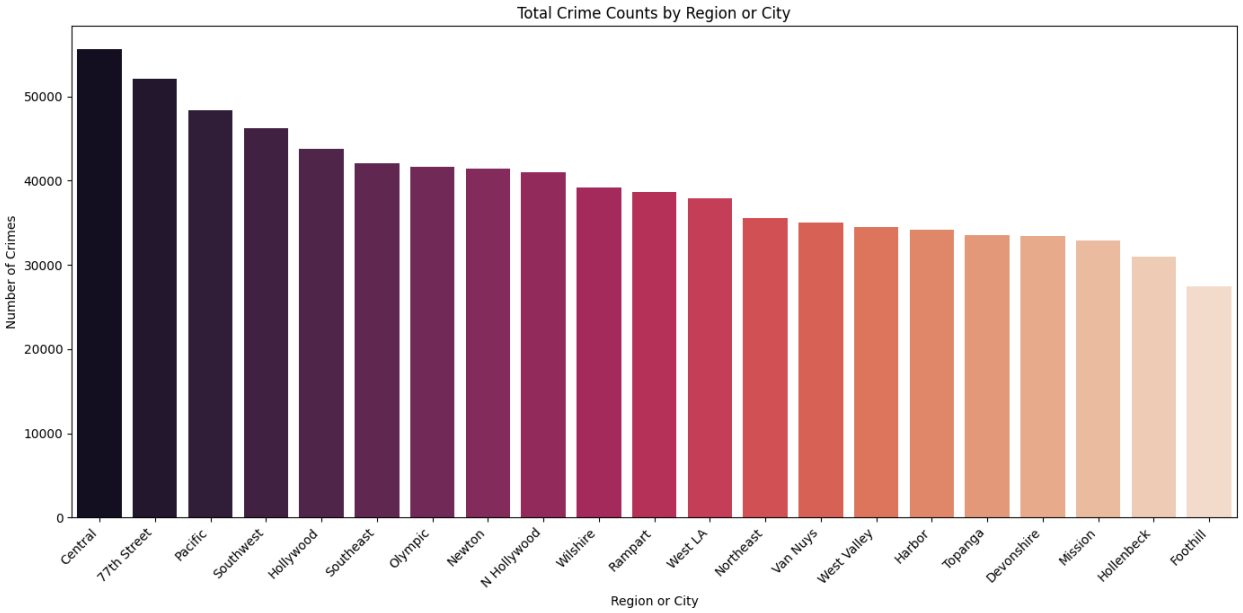
3. Most Common Crime Type:
Count the occurrences of each crime type and identify the one with the highest frequency.



This plot shows us the highest occurrence of crime type. This graph has been plotted for only the top 5 crime types for better visualization. We can see that the crime type ‘VEHICLE-STOLEN’ has the highest occurrence among all the other crime types present in the dataset.

4. Regional Differences:

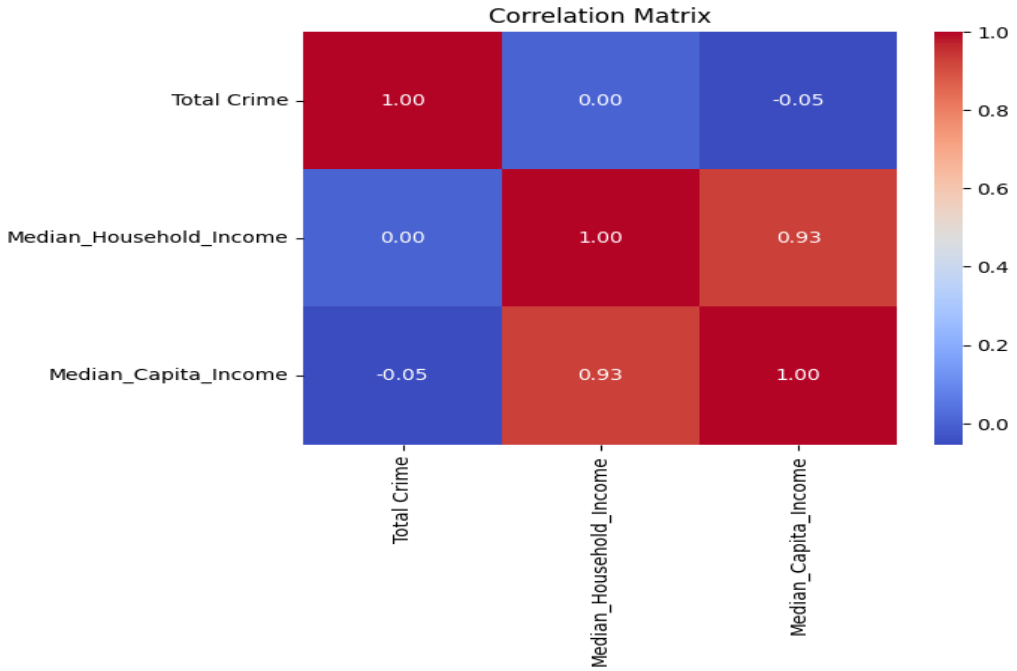
Group the data by region or city and compare crime rates between them using descriptive statistics or visualizations.



This plot shows us the crime rates by region/city. From the graph we can see that the 'Central' region has the most number of crimes among other regions and 'Foothill' region has the lowest number of crimes among the regions.

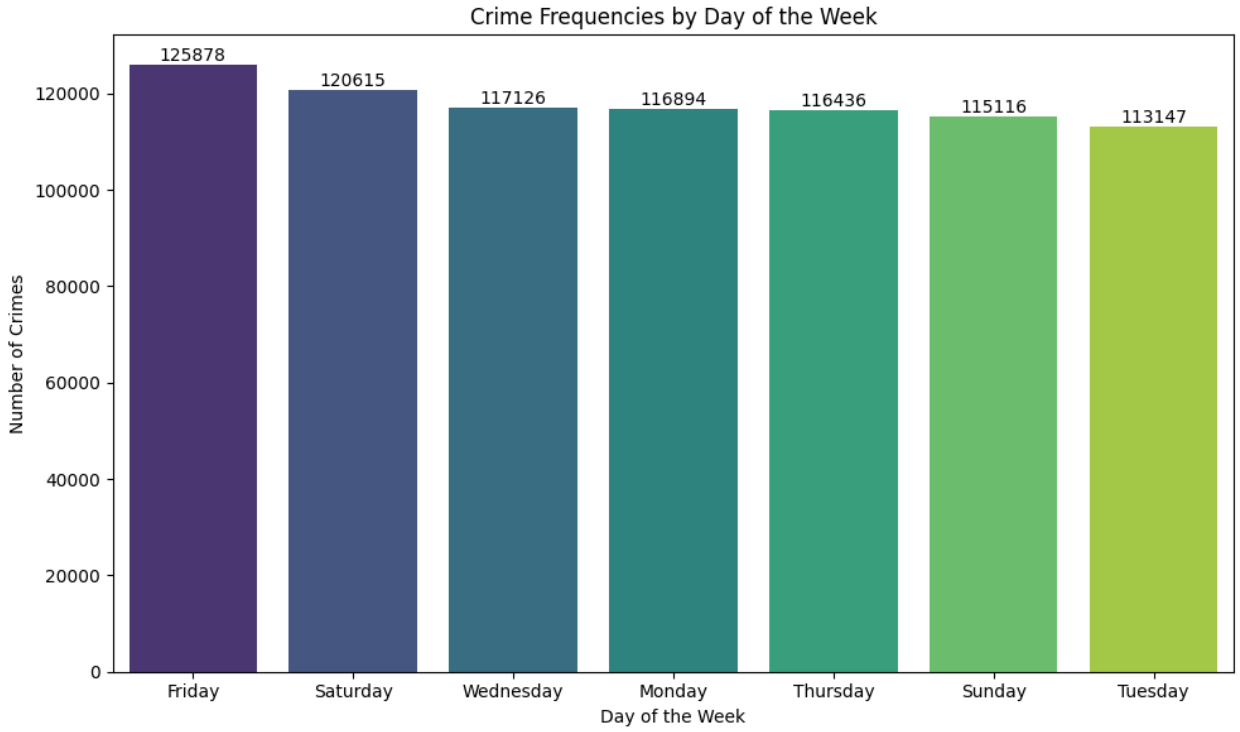
5. Correlation with Economic Factors:

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.



This plot represents the correlation between economic factors and the total number of crimes. We can get many insights from the above plot, whether the economic factors have affected the crime rates, has it increased the crime rates or decreased the crime rates.

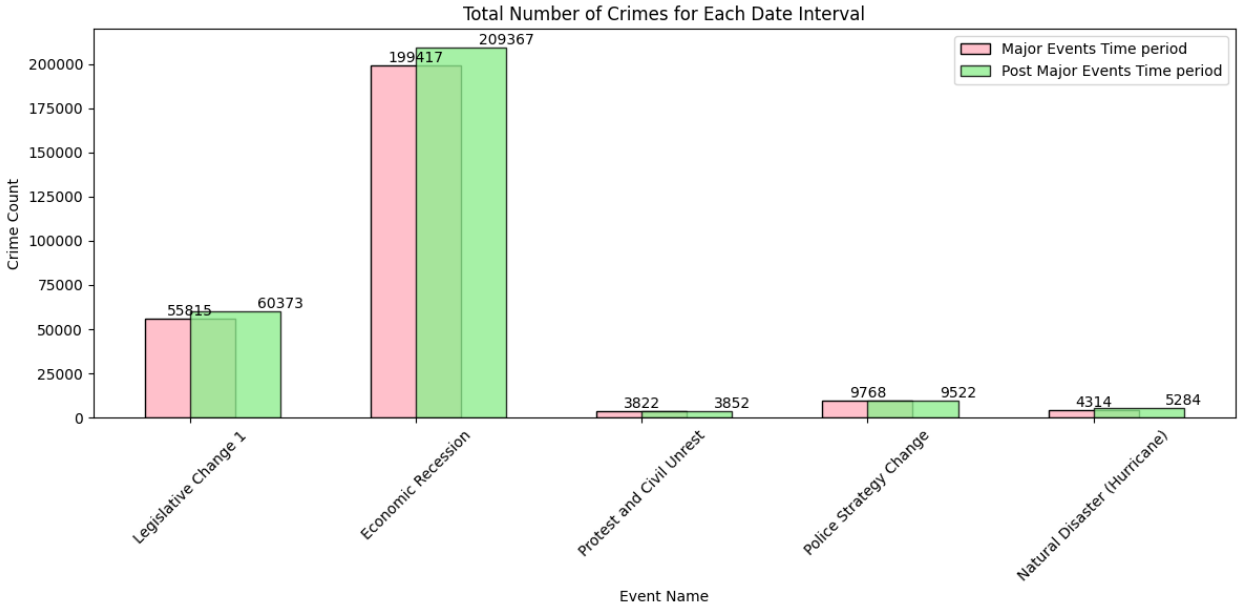
6. Day of the Week Analysis:
Group the data by day of the week and analyze crime frequencies for each day.



The plot shows the number of crimes per day. The above graph is displayed from the day with the most number of crimes to the lowest. We can see that Friday has the most number of crimes with 1,25,878 crimes and Tuesday has the least number of crimes with 1,13,147 crimes.

7. Impact of Major Events:

Identify major events or policy changes during the dataset period and analyze crime rate changes before and after these events.

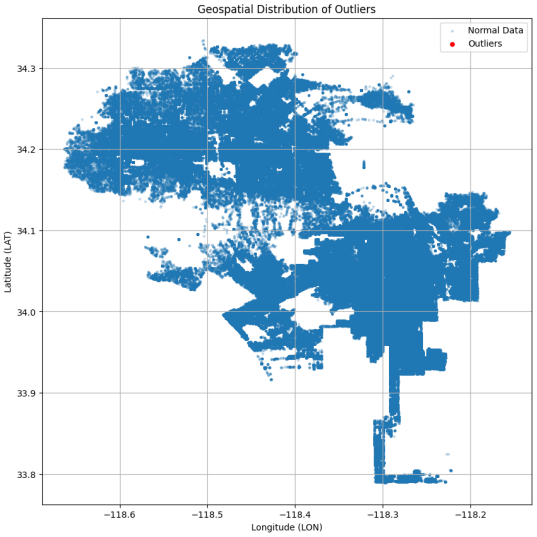


This plot shows us the comparison between the crime rates during major events in the year and the crime rates after the major events in the year. We have selected particular time frames for comparison. From the plot we can clearly see that there is an increase in crime rates after the major events. Especially during the economic recession the crime rates were too high compared to any other major event.

8. Outliers and Anomalies:

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

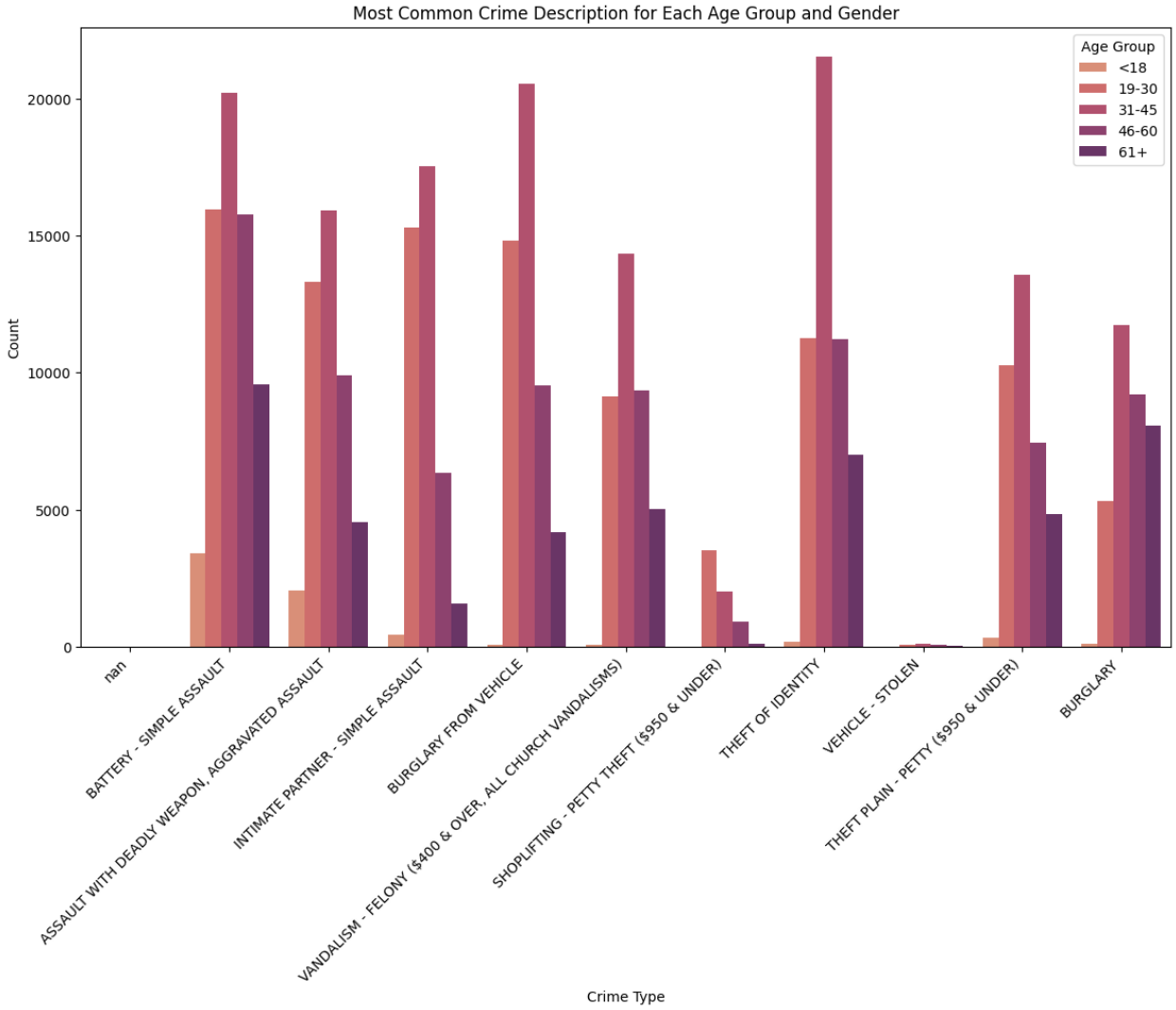
```
Number of outliers: 0
Summary statistics of outlier data:
       DR_NO  AREA  Rpt Dist No  Part 1-2  Crm Cd  Vict Age  Premis Cd  \
count    0.0   0.0          0.0       0.0     0.0       0.0        0.0
mean     NaN   NaN          NaN       NaN     NaN       NaN        NaN
std      NaN   NaN          NaN       NaN     NaN       NaN        NaN
min      NaN   NaN          NaN       NaN     NaN       NaN        NaN
25%      NaN   NaN          NaN       NaN     NaN       NaN        NaN
50%      NaN   NaN          NaN       NaN     NaN       NaN        NaN
75%      NaN   NaN          NaN       NaN     NaN       NaN        NaN
max      NaN   NaN          NaN       NaN     NaN       NaN        NaN

       Weapon Used Cd  Crm Cd 1  Crm Cd 2  Crm Cd 3  Crm Cd 4  LAT  LON  Year
count             0.0       0.0       0.0       0.0       0.0  0.0  0.0   0.0
mean              NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN
std               NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN
min               NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN
25%               NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN
50%               NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN
75%               NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN
max               NaN       NaN       NaN       NaN       NaN  NaN  NaN   NaN

       Month
count    0.0
mean     NaN
std      NaN
min      NaN
25%      NaN
50%      NaN
75%      NaN
max      NaN
```



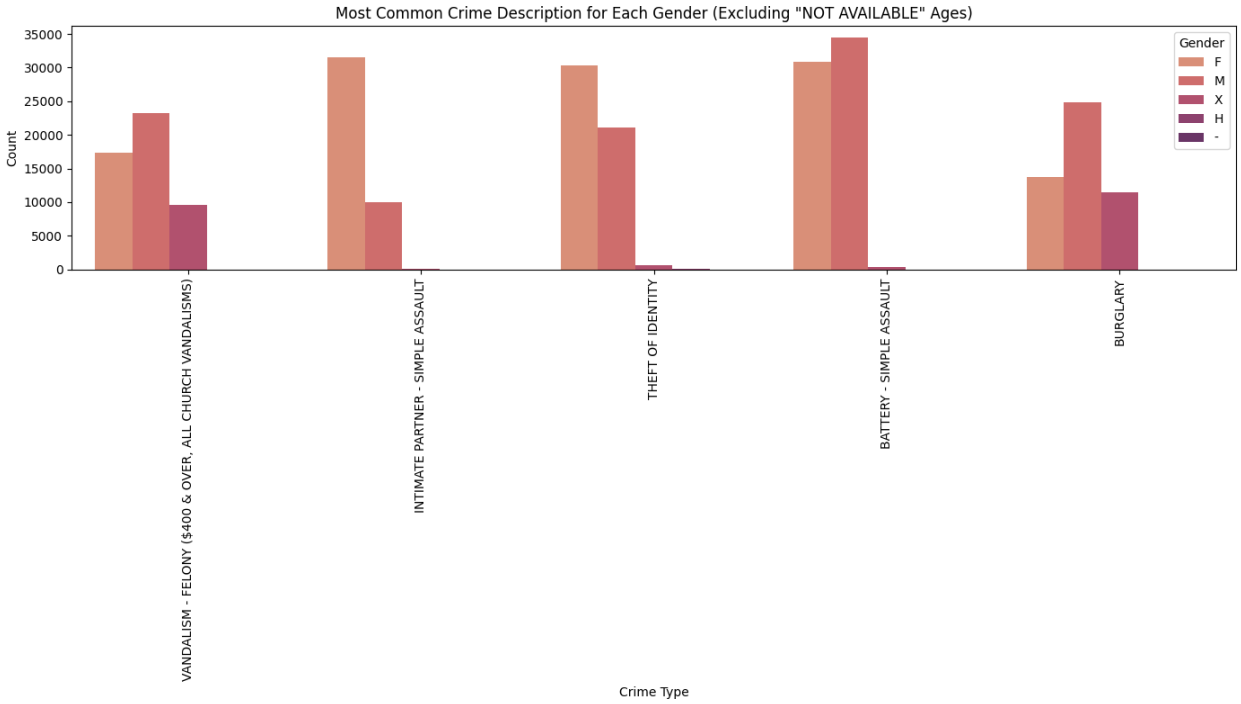
We are calculating Z-scores for numerical columns using SciPy, identifying outliers with Z-scores exceeding 3 or below -3, which are then stored in "outlier_data." It also provides outlier count and summary statistics. A scatter plot visualizes the geospatial distribution of outliers. This analysis is crucial for detecting outliers, especially in geospatial data, facilitating the identification of unusual patterns and data quality enhancements.

9. Demographic Factors:

Analyze the dataset to identify any patterns or correlations between demographic factors (e.g., age, gender) and specific types of crimes.



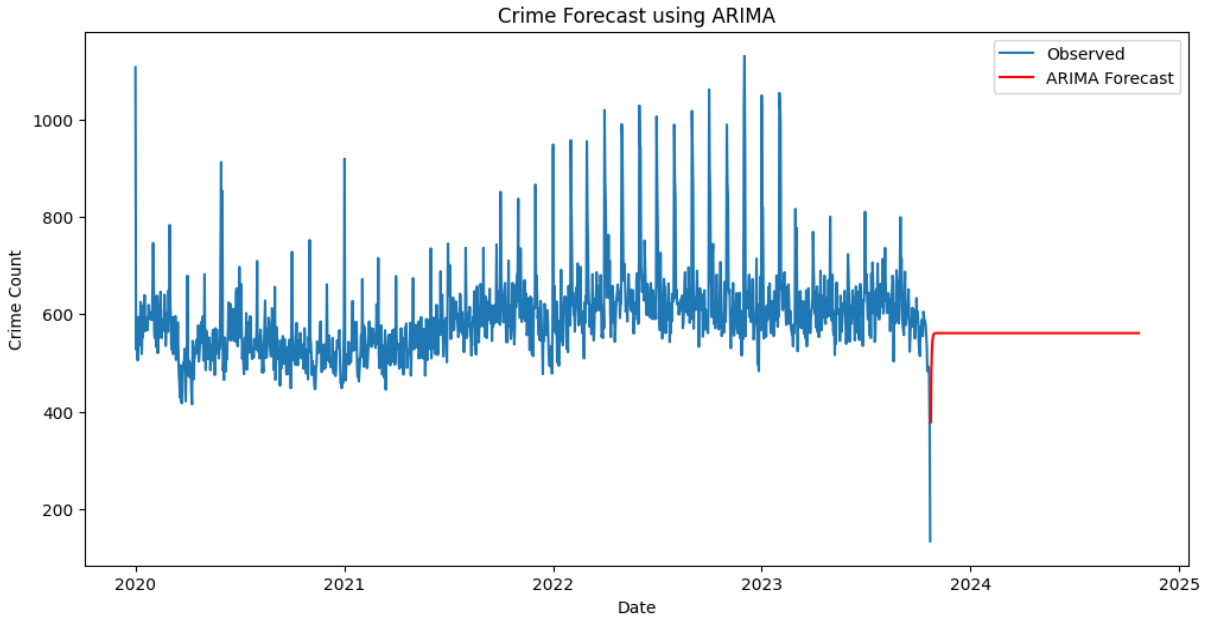
The above plot shows us the most common type of crime for each age group. We can see from the graph that the age group less than 18 has the least crimes in all types of crime, while the age group 31-45 has the most number of crimes in all types of crime.



The above plot shows us the most common type of crime for each gender. We can see from the graph that the F gender has the most crimes in INTIMATE PARTNER-SIMPLE ASSAULT and THEFT OF IDENTITY type of crime. While the M gender has the most crimes in BATTERY-SIMPLE ASSAULT type of crime.

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.



We did forecasting using the ARIMA model and found out the forecasting for 365 days. The plot displays the observed crime rates and forecasted crime rates. We did the model with an order of (1, 1, 1), indicating one autoregressive term, one differencing, and one moving average term. Time series forecasting is an iterative process, hence we need to do more fine tuning to the model for better accuracy in predictions.

SUMMARY OF RESULTS

After analyzing the data scraped from “Crime Data from 2020 to Present," it is concluded that this report presents the findings from the analysis of crime data in the City of Los Angeles, covering the period from 2020 to the present year. The objective of the analysis was to identify trends, patterns, and factors that impact crime rates, with the ultimate goal of improving public safety and resource allocation. The analysis reveals critical insights into crime trends, seasonal patterns, the most common crime types, and the impact of major events. Additionally, demographic factors and the potential for future trend prediction were explored.

Overall Crime Trends:
The analysis commenced by examining the overall crime trends from 2020 to the present year. The information showed an obvious decrease in the quantity of recorded crimes over time, pointing to an improvement in public safety. To comprehend the fundamental causes of this decline, more investigation is necessary.

Seasonal Patterns:
Seasonal patterns were also examined in the investigation, revealing annual fluctuations in the number of crimes committed. It was clear that some months had greater crime rates than others, which suggests that concentrated law enforcement operations are necessary during times of high crime.

Most Common Crime Type:
The most common type of crime was identified, which was an important part of the analysis. "Vehicle Theft" was found to be the most common sort of crime, according to the data, underscoring the significance of taking preventative measures against crimes involving automobiles.

Day of the Week Analysis:
The analysis was expanded to look into the frequency of crimes by day of the week. Results showed that there are weekly variations in crime rates, with certain days seeing a rise in activity. These data can be used by law enforcement organizations to strategically assign patrols and allocate resources as efficiently as possible.

Impact of Major Events:
Throughout the dataset period, significant occurrences and policy changes were noted, with an emphasis on how they affected crime rates. Through an examination of crime statistics prior to and during these incidents, the research clarifies any possible links between outside influences and criminal activity.

Outliers and Anomalies:
To identify outliers and unusual patterns in the dataset, statistical approaches were used in the study. This is an essential stage in analyzing abnormalities and determining how they affect public safety.

Demographic Factors:
We looked for relationships between victim demographics (age, sex, and ethnicity) and particular types of crimes. This investigation provides information for targeted interventions by shedding light on possible connections between criminal activity and certain demographics.

Predicting Future Trends:
The application of time series forecasting techniques, including ARIMA, to forecast future crime trends from previous data was covered in the paper. By taking a proactive approach, law enforcement organizations can better address growing criminal tendencies.

This report underscores the importance of utilizing data-driven insights to continuously improve public safety efforts and foster a safer environment for the residents of the City of Los Angeles.

CONCLUSION

In conclusion, the analysis of crime data in the City of Los Angeles offers a comprehensive understanding of the current state of public safety. By exploring overall trends, seasonal variations, common crime types, and the impact of external factors, the report provides a valuable foundation for informed decision-making by law enforcement agencies and policymakers. Furthermore, the investigation of outliers and demographic correlations adds depth to the analysis. Lastly, the potential for predicting future crime trends equips authorities with a proactive tool to enhance public safety and resource allocation.