Los Angeles - Crime Analysis

IE6400 – FOUNDATIONS OF DATA ANALYTICS

Project Report 1

Group Number 17

Janani Karthikeyan (002830003)

Milan Gurumurthy (002833029)

Prathyusha Adhikam (002835277)

Saathvika Kethineni (002893814)

Shreyas Sreenivas (002825934)

Submitted to: Sivarit Sultornsanee

Submitted Date: November 3rd, 2023

Contents:

Part 1: Introduction and Research Questions

For more than a century, policing and crime documentation have been essential to the administration of Los Angeles. The city, which is well-known for its diverse neighborhoods and dynamic urban landscape, faces a complex crime landscape. Los Angeles, like many major metropolitan areas, faces varying levels of crime in different neighborhoods and communities. The city's overall crime rate fluctuates and, at times, exceeds the national average, particularly in relation to certain types of offenses. Los Angeles has faced distinct challenges in addressing crime, grappling with issues ranging from gang-related incidents to property crimes and more, contributing to its dynamic nature.

The goal of our project is to understand and analyze crime patterns using a dataset derived from reported crimes in Los Angeles. This dataset contains a wide range of crime-related information, including crime type, geographical location, and timestamps of occurrence.

Our goal is to delve into this dataset using Python's exploratory data analysis (EDA) techniques, identifying hidden trends and patterns. Using Python, we hope to create visual maps of crime occurrences in various areas of Los Angeles, allowing for quick detection of crime hotspots and areas most affected by crime. In addition, our analysis will include the creation of time series plots that show the evolution of crime over time.

Our aim is to create models that can predict potential future criminal activities. This involves employing these models to forecast the likelihood of specific types of crimes occurring in specific areas at various times. We want to provide an in-depth analysis of Los Angeles crime patterns from the dataset provided to us.

Part 2: Summary of Results

We initiated a project that aimed to comprehensively examine a dataset containing crime data. Our objective was to understand the underlying structure of the data and ensure its cleanliness and integrity for potential future analyses. The dataset contained various attributes such as dates of crime occurrence and reporting, location details, crime codes, and victim demographics.

We began by importing the data into a Pandas DataFrame and subsequently conducted an initial exploratory data analysis (EDA). During this phase, we meticulously reviewed the data types of each attribute to ensure consistency and alignment with our expectations. To familiarize ourselves with the dataset, we examined the column names and inspected the first few rows of data.

Our next step was to identify and handle missing data within the dataset. We calculated the sum of null values in each column, which provided us insight into the completeness of the data. Recognizing the importance of maintaining a clean dataset, we strategically dropped certain columns that were deemed unnecessary for the scope of our project. These columns either contained redundant information or were not aligned with our focus.

In addressing the missing values in the 'Vict Sex' and 'Vict Descent' columns, we chose to impute these with a placeholder value 'X'. This approach allowed us to maintain the integrity of the dataset while dealing with missing data in a systematic manner. We then performed an additional check for duplicate rows and ensured that any null values were appropriately dealt with by dropping such rows.

Finally, we previewed the cleaned dataset to validate the effectiveness of our data cleaning and preprocessing steps. The dataset was now in a refined state, ready for any subsequent analyses that may be required.

1. Crime Types: Analyze the distribution of different types of crimes.
2. Area-wise Crime: Evaluate the frequency of crimes in different areas.
3. Time-based Analysis: Examine the occurrence of crimes based on the date and time.
4. Victim Demographics: Summarize the data based on the age, sex, and descent of the victims.
5. Weapons Used: Understand the distribution of weapons used in the crimes.

Summary based on the analysis of the dataset:

1. *Crime Types:*
- The most common type of crime is "VEHICLE - STOLEN" with 88,892 instances.
- This is followed by "BATTERY - SIMPLE ASSAULT" and "THEFT OF IDENTITY" with 66,149 and 52,321 instances respectively.

2. *Area-wise Crime:*
- The area with the highest number of reported crimes is "Central" with 55,923 instances.
- This is followed by "77th Street" and "Pacific" with 52,362 and 48,582 instances respectively.

3. *Time-based Analysis:*
- Yearly Crime: The highest number of crimes were reported in the year 2022 with 234,289 instances.
- Monthly Crime: The month of July has the highest crime rate with 75,333 instances.
- Hourly Crime: The data contains time of occurrence, but further processing is needed to extract meaningful hourly trends.

4. *Victim Demographics:*
- Victim Age: The most common age recorded is 0, with 205,156 instances. This might require further investigation to understand if it's a placeholder or actual data.
- Victim Sex: The majority of victims are Male (342,350) followed by Female (305,479).
- Victim Descent: The majority of victims are of Hispanic (H) descent (254,505), followed by White (W) and Black (B) with 168,980 and 118,097 instances respectively.

5. *Weapons Used:*
- The most common method of assault is "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)" with 155,040 instances.
- This is followed by "UNKNOWN WEAPON/OTHER WEAPON" and "VERBAL THREAT" with 30,760 and 21,201 instances respectively.

This summary provides an overview of the dataset across various dimensions.

Part 3: Data Sources

We employed Python as the primary coding language for our analysis and visuals. The data loading process, interpretation of data, and pictorial representations will be done via Jupiter studios. The statistics and analysis will be based on the data set provided to us.

Using EDA tools and other analytical approaches, the dataset was analyzed, visualized, interpreted, and represented.

Sources:

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

https://www.cfr.org/blog/ten-most-significant-world-events-2020

https://www.cfr.org/blog/ten-most-significant-world-events-2021

https://www.cfr.org/blog/ten-most-significant-world-events-2022

https://www.onthisday.com/events/date/2023

https://fred.stlouisfed.org/series/GDP

Part 4: Results and Methods

We used the following approach to prepare, analyze, visualize the given dataset:

- **Data Acquisition and inspection:**

We downloaded the data set from the given source link. We then displayed the first few rows of the dataset and checked the data types of each column to understand the data we are working with. We then reviewed the column names and descriptions.

```
In [1]: import pandas as pd
        import numpy as np

In [2]: df = pd.read_csv(r"D:\Documents\Prof_Docs\FDA\Projects\Project 1\Crime_Data_from_2020_to_Present.csv")

In [3]: pd.set_option('display.max_columns', None)
        df.head(5)
```

Out[3]:

	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 I
0	10304468	01/08/2020 12:00:00 AM	01/08/2020 12:00:00 AM	2230	3	Southwest	377	2	624	BATTERY - SIMPLE ASSAULT	0444 0913	36	F	B	501.0	SIN FAI DWELI
1	190101086	01/02/2020 12:00:00 AM	01/01/2020 12:00:00 AM	330	1	Central	163	2	624	BATTERY - SIMPLE ASSAULT	0416 1822 1414	25	M	H	102.0	SIDEW
2	200110444	04/14/2020 12:00:00 AM	02/13/2020 12:00:00 AM	1200	1	Central	155	2	845	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	1501	0	X	X	726.0	PO FACI
3	191501505	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	1730	15	N Hollywood	1543	2	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	0329 1402	76	F	W	502.0	MULTI-I DWELI (APARTM DUP I
4	191921269	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	415	19	Mission	1998	2	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	0329	31	X	X	409.0	BEA SUI ST

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 825212 entries, 0 to 825211
Data columns (total 28 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   DR_NO           825212 non-null  int64
 1   Date Rptd       825212 non-null  object
 2   DATE OCC        825212 non-null  object
 3   TIME OCC        825212 non-null  int64
 4   AREA            825212 non-null  int64
 5   AREA NAME       825212 non-null  object
 6   Rpt Dist No     825212 non-null  int64
 7   Part 1-2        825212 non-null  int64
 8   Crm Cd          825212 non-null  int64
 9   Crm Cd Desc     825212 non-null  object
 10  Mocodes         711064 non-null  object
 11  Vict Age        825212 non-null  int64
 12  Vict Sex        716683 non-null  object
 13  Vict Descent    716675 non-null  object
```

In [5]:
```
df.dtypes
```

Out[5]:
```
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
```

In [6]:
```
df.columns
```

Out[6]:
```
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')
```

```
In [15]: df.head()
```

Out[15]:

	Date Rptd	DATE OCC	TIME OCC	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Vict Age	Vict Sex	Vict Descent	Premis Desc	Status Desc	LAT	LON
0	01/08/2020 12:00:00 AM	01/08/2020 12:00:00 AM	2230	Southwest	377	2	624	BATTERY - SIMPLE ASSAULT	36	F	B	SINGLE FAMILY DWELLING	Adult Other	34.0141	-118.2978
1	01/02/2020 12:00:00 AM	01/01/2020 12:00:00 AM	330	Central	163	2	624	BATTERY - SIMPLE ASSAULT	25	M	H	SIDEWALK	Invest Cont	34.0459	-118.2545
2	04/14/2020 12:00:00 AM	02/13/2020 12:00:00 AM	1200	Central	155	2	845	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	0	X	X	POLICE FACILITY	Adult Arrest	34.0448	-118.2474
3	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	1730	N Hollywood	1543	2	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	76	F	W	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	Invest Cont	34.1685	-118.4019
4	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	415	Mission	1998	2	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	31	X	X	BEAUTY SUPPLY STORE	Invest Cont	34.2198	-118.4468

- **Data cleaning and preprocessing:**

We verified for and confirmed that there were no duplicate records in our dataset. Next, we assessed the data to remove any missing values. If the percentage of missing values is minor, we simply removed the rows or columns with missing values using the dropna() method and we converted data types wherever needed.

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
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
```

In [8]:
```
#Checking for duplicated rows
df.duplicated().sum()
```

Out[8]:
```
0
```

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
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
```

In [10]:
```
df.drop(['DR_NO', 'AREA', 'Mocodes', 'Premis Cd', 'Weapon Used Cd', 'Crm Cd 1', 'Weapon Desc', 'Status', 'Crm Cd 2', 'C
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
Date Rptd            0
DATE OCC             0
TIME OCC             0
AREA NAME            0
Rpt Dist No          0
Part 1-2             0
Crm Cd               0
Crm Cd Desc          0
Vict Age             0
Vict Sex        108529
Vict Descent    108537
Premis Desc        488
Status Desc          0
LAT                  0
LON                  0
dtype: int64
```

In [12]:
```
df["Vict Sex"].fillna("X")
df["Vict Descent"].fillna("X")
```

Out[12]:
```
0         B
1         H
2         X
3         W
4         X
         ..
825207    H
825208    H
825209    B
825210    H
825211    H
Name: Vict Descent, Length: 825212, dtype: object
```

In [13]:
```
df.dropna(inplace=True)
```

In [14]:
```
df.isnull().sum()
```

Out[14]:
```
Date Rptd       0
DATE OCC        0
TIME OCC        0
AREA NAME       0
Rpt Dist No     0
Part 1-2        0
Crm Cd          0
Crm Cd Desc     0
Vict Age        0
Vict Sex        0
Vict Descent    0
Premis Desc     0
Status Desc     0
LAT             0
LON             0
dtype: int64
```

Out[19]:

	Date Rptd	DATE OCC	TIME OCC	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Vict Age	Vict Sex	Vict Descent	Premis Desc	Status Desc
0	01/08/2020 12:00:00 AM	01/08/2020 12:00:00 AM	2230	Southwest	377	2	624	BATTERY - SIMPLE ASSAULT	36	F	Black	SINGLE FAMILY DWELLING	Adul Othe
1	01/02/2020 12:00:00 AM	01/01/2020 12:00:00 AM	330	Central	163	2	624	BATTERY - SIMPLE ASSAULT	25	M	Hispanic/Latin/Mexican	SIDEWALK	Inves Con
2	04/14/2020 12:00:00 AM	02/13/2020 12:00:00 AM	1200	Central	155	2	845	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	0	X	Unknown	POLICE FACILITY	Adul Arres
3	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	1730	N Hollywood	1543	2	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	76	F	White	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	Inves Con
								VANDALISM -					

In [17]:
```
df['Vict Descent'].value_counts()
```

Out[17]:
```
Hispanic/Latin/Mexican            253094
White                             168047
Black                             117524
Unknown                            79364
Other                              65339
Other Asian                        18050
Korean                              4391
Filipino                            3435
Chinese                             3167
Japanese                            1145
Vietnamese                           851
American Indian/Alaskan Native       772
Asian Indian                         412
Pacific Islander                     219
Hawaiian                             167
Cambodian                             62
Guamanian                             58
Laotian                               50
Samoan                                46
Name: Vict Descent, dtype: int64
```

In [18]:
```
groups = [
    (df['Vict Age'] <= 12),
    (df['Vict Age'] >= 13) & (df['Vict Age'] < 18),
    (df['Vict Age'] >= 18) & (df['Vict Age'] < 65),
    (df['Vict Age'] >= 65)
]

labels = ['Child', 'Teen', 'Adult', 'Old']

# create new column 'Age Group'
df['Age Group'] = np.select(groups, labels)
```

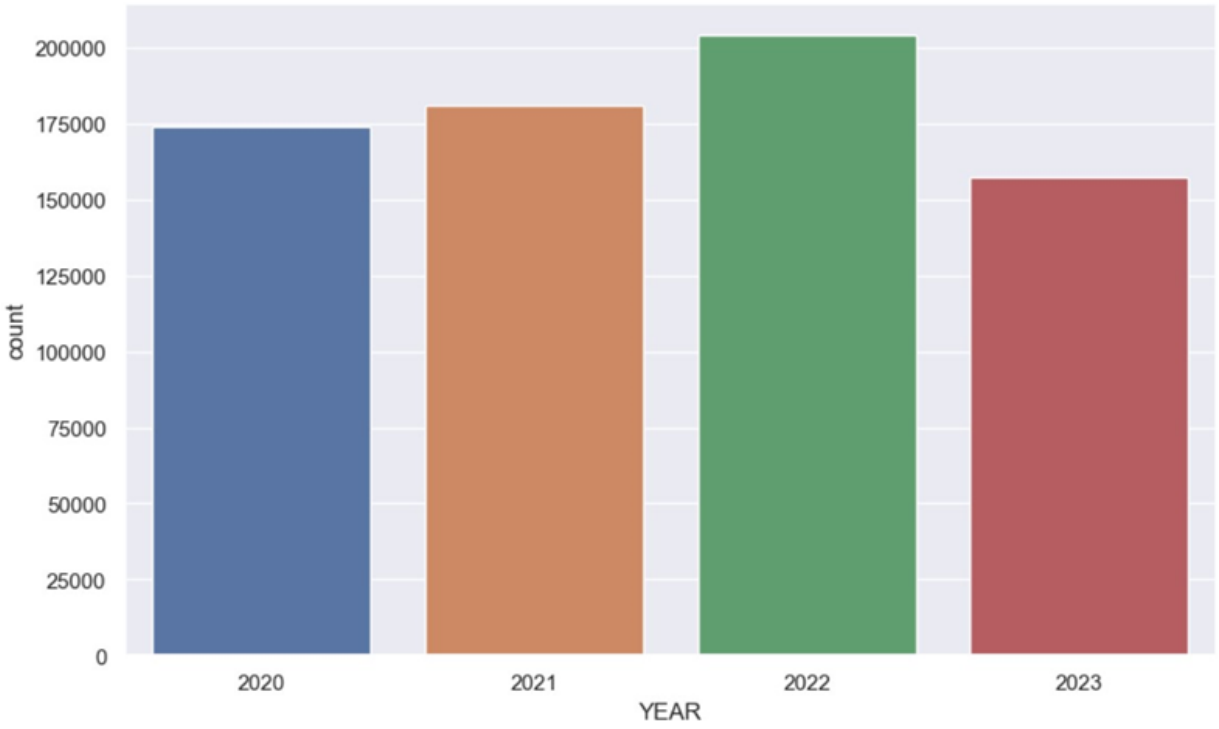
In [29]:
```
df.head()
```

Out[29]:

	Date Rptd	DATE OCC	TIME OCC	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Vict Age	Vict Sex	Vict Descent	Premis Desc	Status Desc	LAT	LON	G
0	2020-01-08	2020-01-08	2230	Southwest	377	2	624	BATTERY - SIMPLE ASSAULT	36	F	Black	SINGLE FAMILY DWELLING	Adult Other	34.0141	-118.2978	
1	2020-01-02	2020-01-01	330	Central	163	2	624	BATTERY - SIMPLE ASSAULT	25	M	Hispanic/Latin/Mexican	SIDEWALK	Invest Cont	34.0459	-118.2545	
2	2020-04-14	2020-02-13	1200	Central	155	2	845	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	0	X	Unknown	POLICE FACILITY	Adult Arrest	34.0448	-118.2474	
3	2020-01-01	2020-01-01	1730	N Hollywood	1543	2	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	76	F	White	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	Invest Cont	34.1685	-118.4019	
4	2020-01-01	2020-01-01	415	Mission	1998	2	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	31	X	Unknown	BEAUTY SUPPLY STORE	Invest Cont	34.2198	-118.4468	

Exploratory Data Analysis (EDA):

1. **Overall crime trends from 2020 to the present year:**



The total number of crimes have increased throughout 2020 – 2022. We can see a slight dip in the crime rates in 2023.

Statistics:

The total number of crimes in 2020 is: 173866

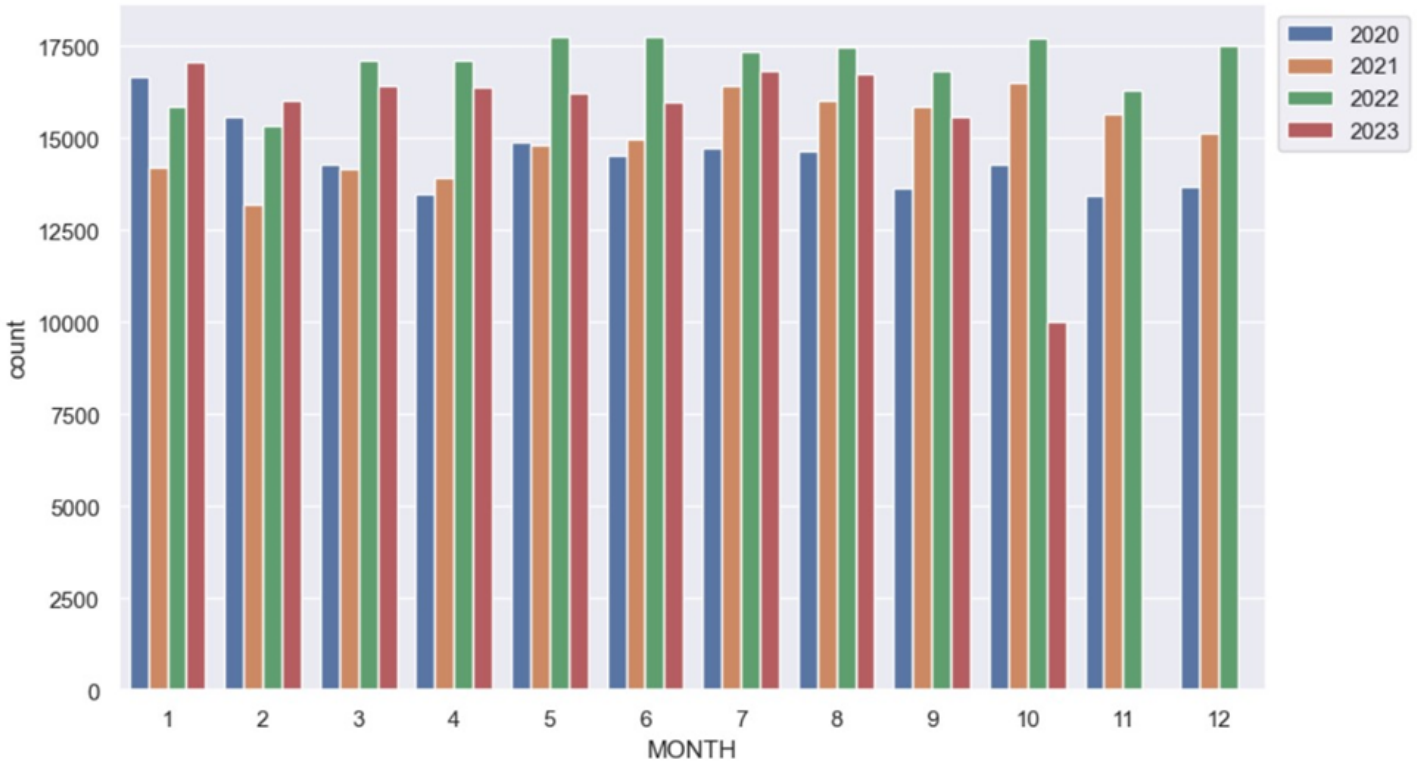
The total number of crimes in 2021 is: 180939

The total number of crimes in 2022 is: 204033

The total number of crimes in 2023 is: 153331

The highest number of crimes occurred in the year 2022 and the least crimes occurred in 2023.

2. Seasonal patterns in crime data:



We can observe that there is not much of a difference in the seasonal pattern of the crime. The crime frequencies follow a similar flow throughout the year.

3. Most common type of crime:

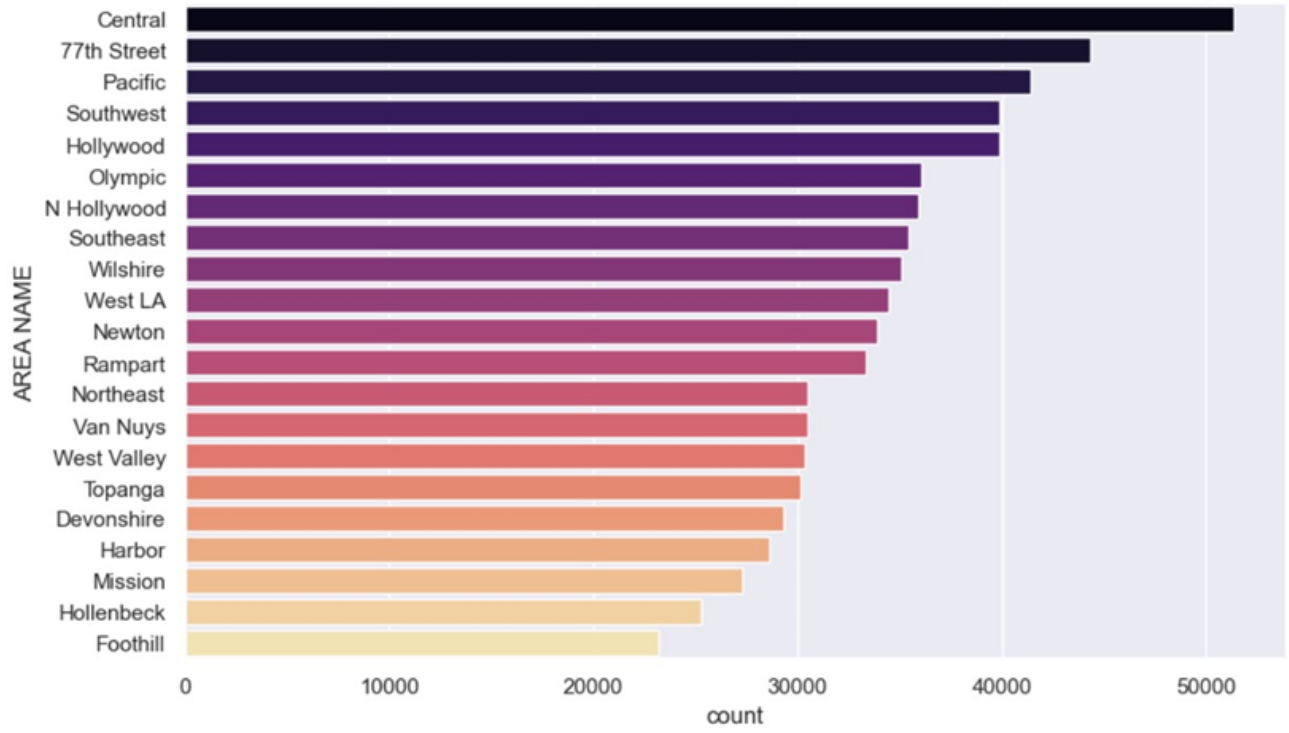
The top 5 most common types of crimes are Battery, theft of identity and burglary from vehicles, Vandalism, and Burglary.

Battery is the most common crime.

The last common crime is inciting riots.

```
]: crime_type = df['Crm Cd Desc'].value_counts()
   crime_type

]: BATTERY - SIMPLE ASSAULT                                    65689
   THEFT OF IDENTITY                                           52117
   BURGLARY FROM VEHICLE                                       50589
   VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)     50240
   BURGLARY                                                    49916
                                                               ...
   THEFT, COIN MACHINE - ATTEMPT                                   5
   FIREARMS RESTRAINING ORDER (FIREARMS RO)                        4
   FAILURE TO DISPERSE                                             3
   DISHONEST EMPLOYEE ATTEMPTED THEFT                              2
   INCITING A RIOT                                                 1
   Name: Crm Cd Desc, Length: 137, dtype: int64

]: print('The highest crime type is:',crime_type.head(1))

   The highest crime type is: BATTERY - SIMPLE ASSAULT    65689
   Name: Crm Cd Desc, dtype: int64
```

4. Notable differences in crime rates between regions:



The above graph shows the distribution of crime within different areas of Los Angeles in a descending pattern. We can observe the density of crimes in different areas.

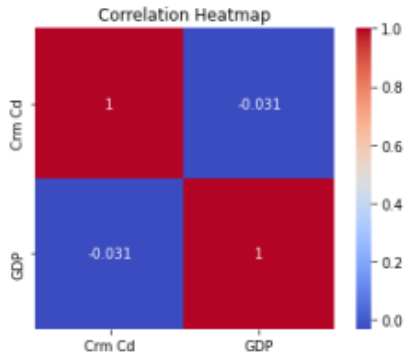
Central area has the highest crime rate, followed by 77th street and Pacific. Foothill has the least crime rates in the city.

The top 5 places of attacks and their corresponding number of attacks are:

1. Single family dwelling - 139736
2. Street - 125890
3. Multi-unit dwelling (Apartment, complex etc) - 101289
4. Parking lot - 43636
5. Other business premises - 37095

5. Correlations between economic factors and crime rates:

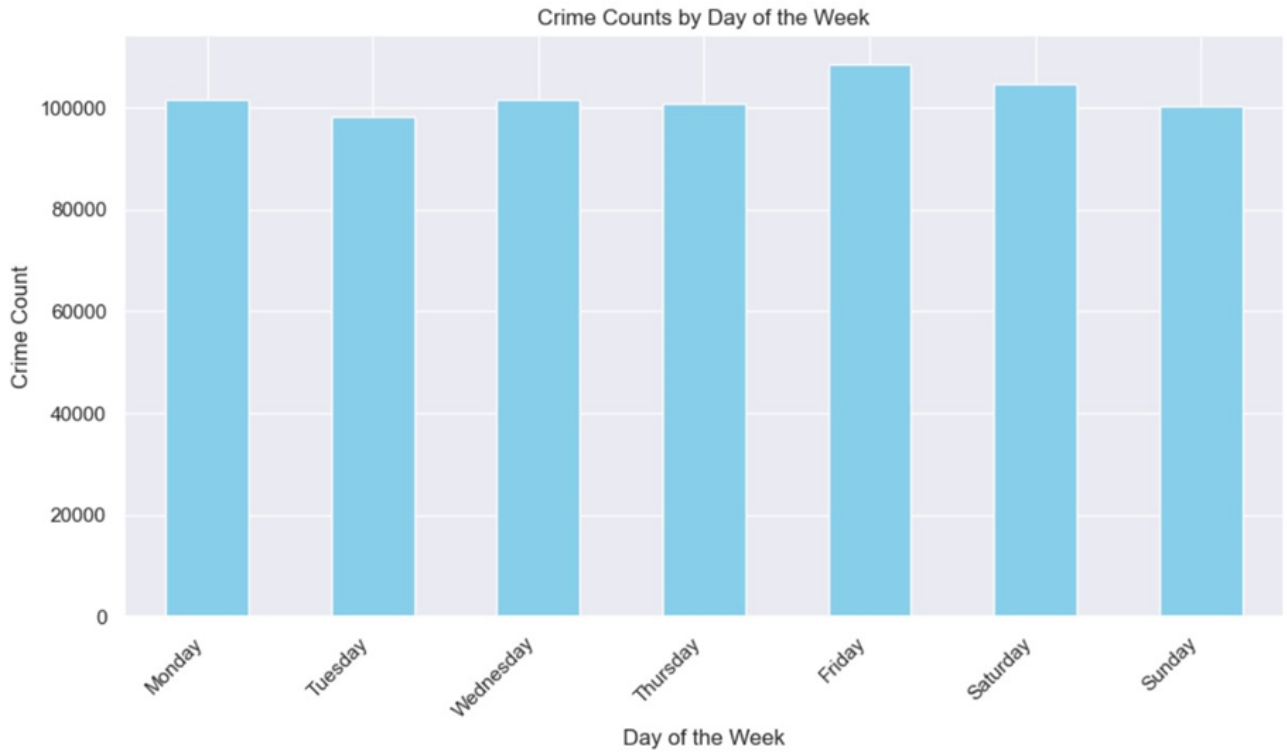
```
In [78]: columns_to_correlate = ['Crm Cd', 'GDP']
         correlation_matrix = data[columns_to_correlate].corr()
         plt.figure(figsize=(6, 4))
         sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', square=True)
         plt.title('Correlation Heatmap')
         plt.show()
```



We have a negative correlation here, Thus we can conclude that as the GDP increases the crime rates tends to drop.

The correlation heatmap visualizes the relationships between different variables in the dataset. The heatmap includes annotations to display correlation values. Variables are depicted on both axes. The colormap "RdBu" is used to represent positive and negative correlations. This heatmap serves as a valuable tool for quickly identifying patterns and relationships between variables in the dataset.

6. Relationship between the day of the week and the frequency of crimes:



Crime Frequencies by Day of the Week:

- Friday- 108658
- Saturday - 104765
- Wednesday - 101676
- Monday - 101557
- Thursday - 100906
- Sunday - 100359
- Tuesday - 98272

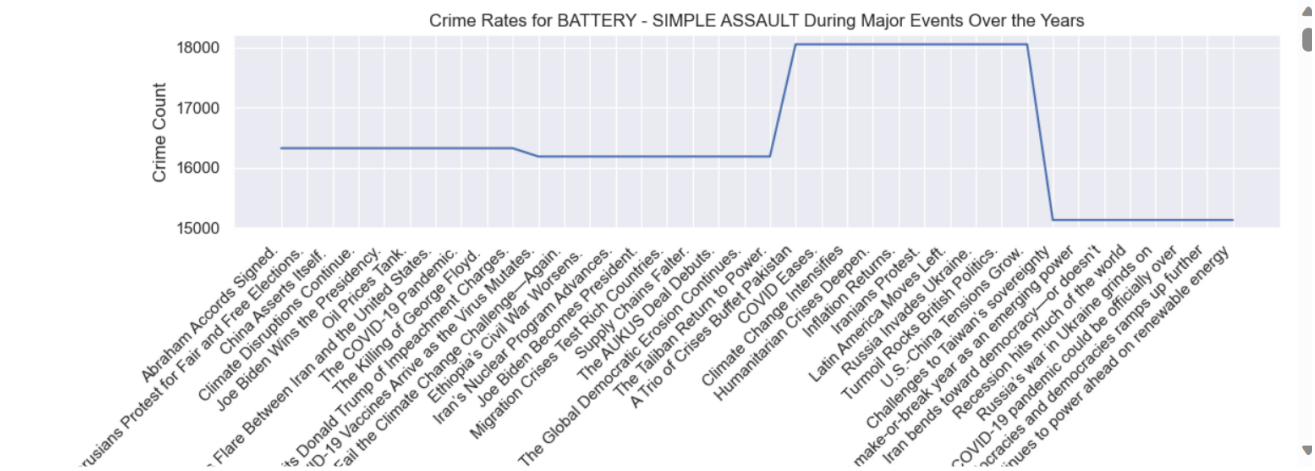
Total Crimes: 716193

Average Crimes per Day: 102313.28571428571

Day with the Most Crimes is Friday and the day with the Fewest Crimes is Tuesday.

7. Investigating any impact of major events or policy changes on crime rates:

```
plt.ylabel('Crime Count')
plt.xticks(rotation=45, ha='right')

plt.tight_layout()
plt.show()
```



From the above graph, we have a series of line plots, each illustrating how specific crime types have evolved in response to major events or policies over the years. These plots, tailored to unique crime categories, allow for a detailed examination of crime rate fluctuations. The x-axis represents major events or policies, offering context for the trends, while the y-axis quantifies crime counts. With event labels rotated for readability and a well-organized layout, these line plots provide a comprehensive visual overview of the impact of major events and policies on different types of crimes, aiding in informed decision-making for law enforcement and policymakers.

In [43]:

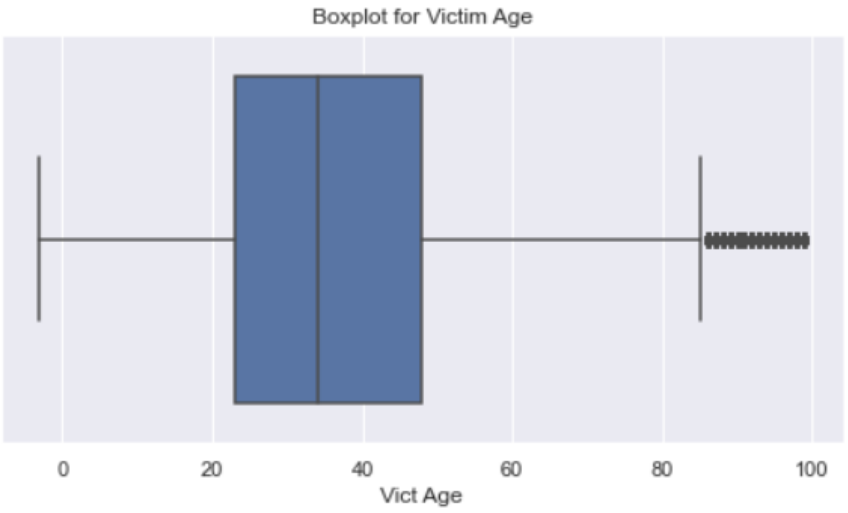
```python
unique_crime_types = crime_count_by_year_event['Crm Cd Desc'].unique()

average_rates = {}
percentage_changes = {}

for crime_type in unique_crime_types:
    crime_data = crime_count_by_year_event[crime_count_by_year_event['Crm Cd Desc'] == crime_type]

    for major_event in crime_data['Major event/policy'].unique():
        subset_data = crime_data[crime_data['Major event/policy'] == major_event]

        crime_rates_before = subset_data[subset_data['YEAR'] < 2023]['Count']
        crime_rates_after = subset_data[subset_data['YEAR'] >= 2023]['Count']

        average_rate_before = crime_rates_before.mean()
        average_rate_after = crime_rates_after.mean()

        percentage_change = ((average_rate_after - average_rate_before) / average_rate_before) * 100

        average_rates[(crime_type, major_event)] = (average_rate_before, average_rate_after)
        percentage_changes[(crime_type, major_event)] = percentage_change

for (crime_type, major_event), (average_rate_before, average_rate_after) in average_rates.items():
    print(f'Crime Type: {crime_type}, Major Event: {major_event}')
    print(f'Average Crime Rate Before: {average_rate_before:.2f}')
    print(f'Average Crime Rate After: {average_rate_after:.2f}')
    print(f'Percentage Change in Crime Rates: {percentage_changes[(crime_type, major_event)]:.2f}%')
    print()
```

```
Crime Type: ARSON, Major Event: Abraham Accords Signed.
Average Crime Rate Before: 664.00
Average Crime Rate After: nan
Percentage Change in Crime Rates: nan%

Crime Type: ARSON, Major Event: Belarusians Protest for Fair and Free Elections.
Average Crime Rate Before: 664.00
Average Crime Rate After: nan
Percentage Change in Crime Rates: nan%

Crime Type: ARSON, Major Event: China Asserts Itself.
Average Crime Rate Before: 664.00
Average Crime Rate After: nan
Percentage Change in Crime Rates: nan%

Crime Type: ARSON, Major Event: Climate Disruptions Continue.
Average Crime Rate Before: 664.00
Average Crime Rate After: nan
Percentage Change in Crime Rates: nan%
```

From the above code we could examine the impact of major events on different crime types by calculating average crime rates before and after each event and determining the percentage change. The provided output showcases this analysis for "ARSON" during various major events, revealing the average rates before and after, as well as the percentage change. However, some combinations have missing data, possibly due to a lack of post-event data. This analysis offers insights into how specific events affect crime rates.

8. Outliers and Anomalies:

```
In [175]: # Checking for outliers and patterns
          plt.figure(figsize=(8, 4))
          sns.boxplot(x=df['Vict Age'])
          plt.title('Boxplot for Victim Age')
          plt.show()
```

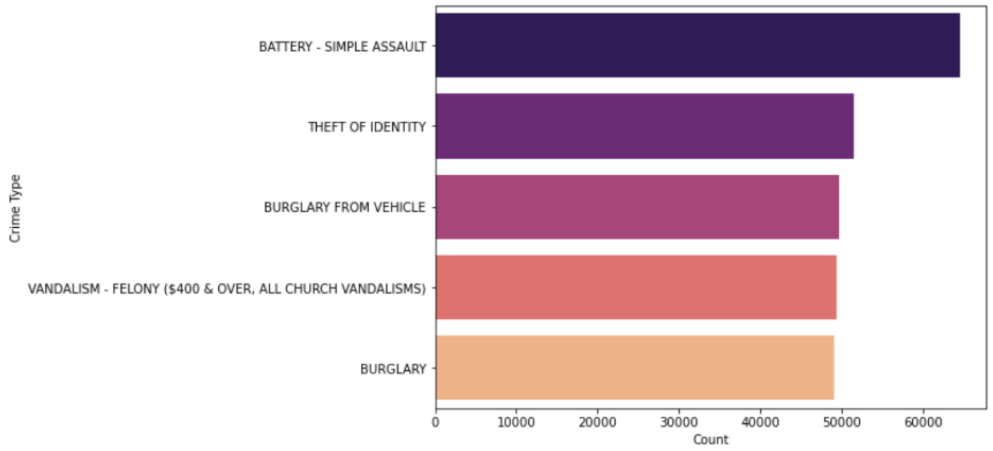


The above visuals present box plots for four different datasets: Victim Age, DR_NO, Latitude (LAT), and Longitude (LON). Here are the main takeaways:

- Victim Age: The majority of victims fall within a narrow age range, with some outliers.
- DR_NO: There's a well-defined spread of data with the majority clustered around the middle. A few outliers can be observed.
- Latitude (LAT): Data points for latitude are fairly concentrated, with a few outliers. There's a thin range where most data lie.
- Longitude (LON): The data spread is narrow for longitude, similar to latitude, with a couple of outliers noticeable.

Overall, while there's a consistent data range for each box plot, there are outliers in all datasets, especially evident in the Victim Age and LAT graphs.

```
In [93]: plt.figure(figsize=(8, 6))
sns.countplot(y='Crm Cd Desc',data=df,order=df['Crm Cd Desc'].value_counts().head(5).index,palette='magma')
plt.xlabel('Count')
plt.ylabel('Crime Type')
plt.show()
```



9. Predicting Future Trends

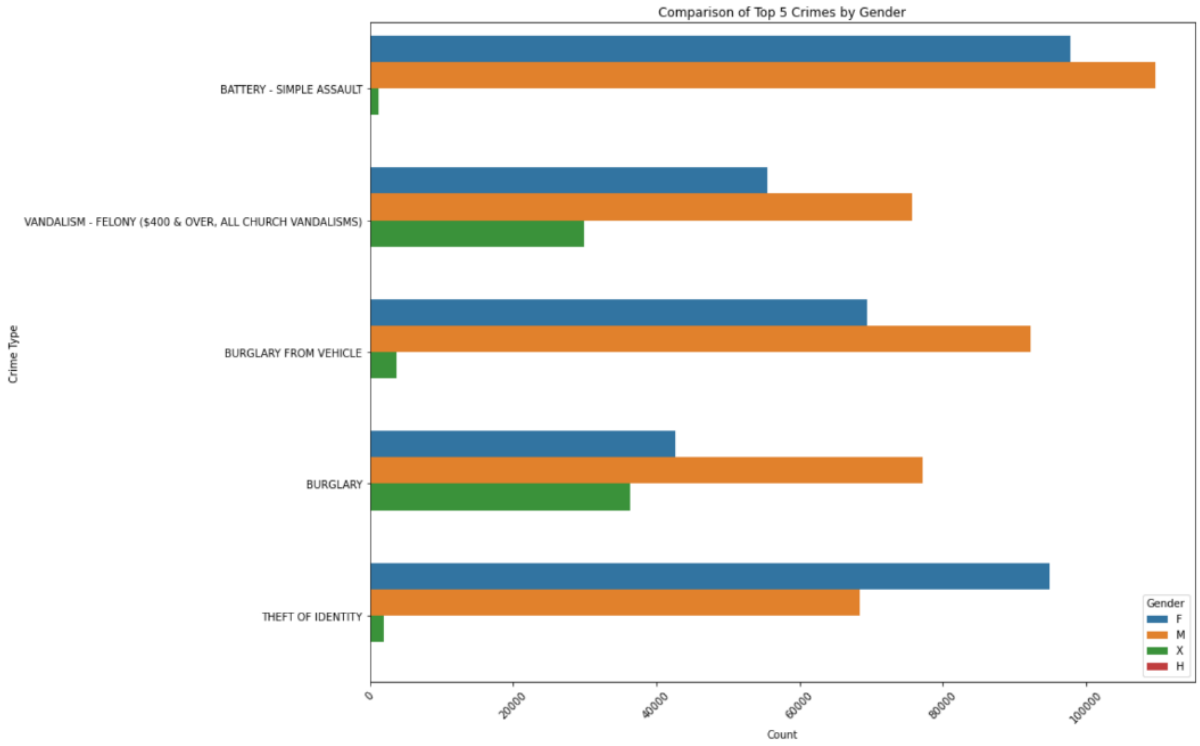
Correlation:

Crime type:

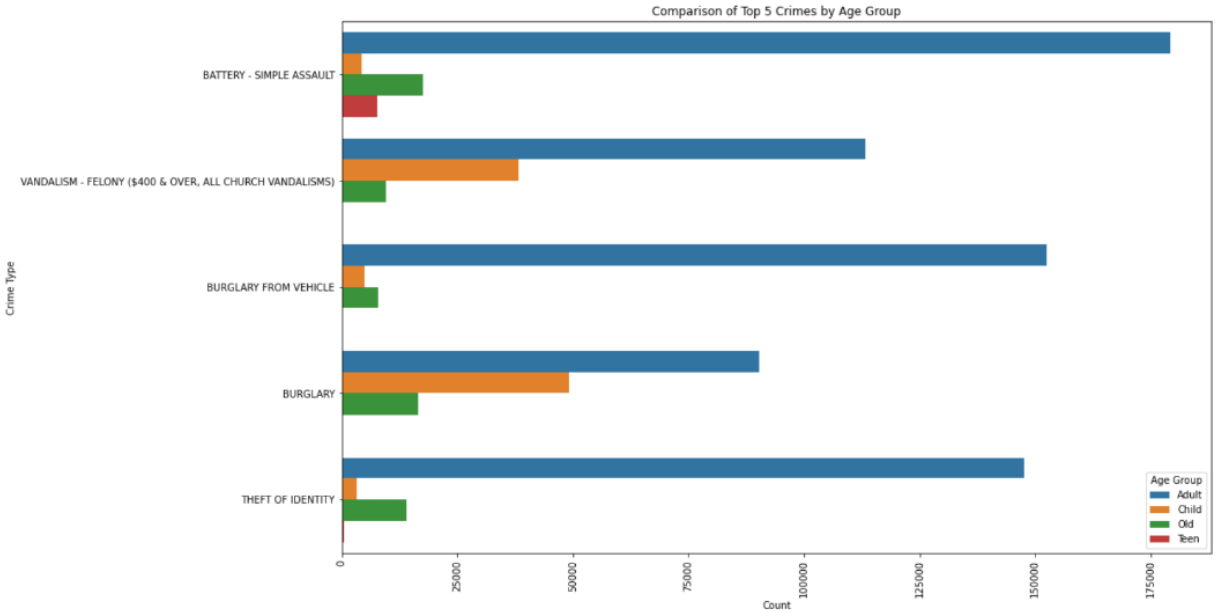
1) *Battery - Simple Assault*
    - The number of male victims are more than the number of female victims thus suggesting that males get assaulted slightly more than women.
    - The victims of this assault are mostly adults followed by senior citizens.
    - Central area has the most number of crimes when compared to the other areas.
2) *Theft of identity*
    - The number of female victims are more than the number of male victims thus suggesting that 30,000 women have their identity stolen than men.
    - The victims of this assault are mostly adults followed by senior citizens, while teens rarely get their identity stolen
    - 77th street has the most number of crimes when compared to the other areas followed by the southwest area.

3) *Vandalism*
    - The number of male victims are more than the number of female victims thus suggesting that men are more affected than women.
    - The victims of this assault are mostly adults followed by children.
    - Central has the most number of crimes when compared to the other areas.
4) *Burglary*
    - The number of male victims are more than the number of female
    - The victims of this assault are mostly adults followed by children.
    - West LA has the most number of crimes when compared to the other areas.
5) *Burglary from vehicle*
    - The number of male victims are more than the number of female
    - The victims of this assault are mostly adults followed by old people.
    - Central significantly has the most number of crimes when compared to the other areas.

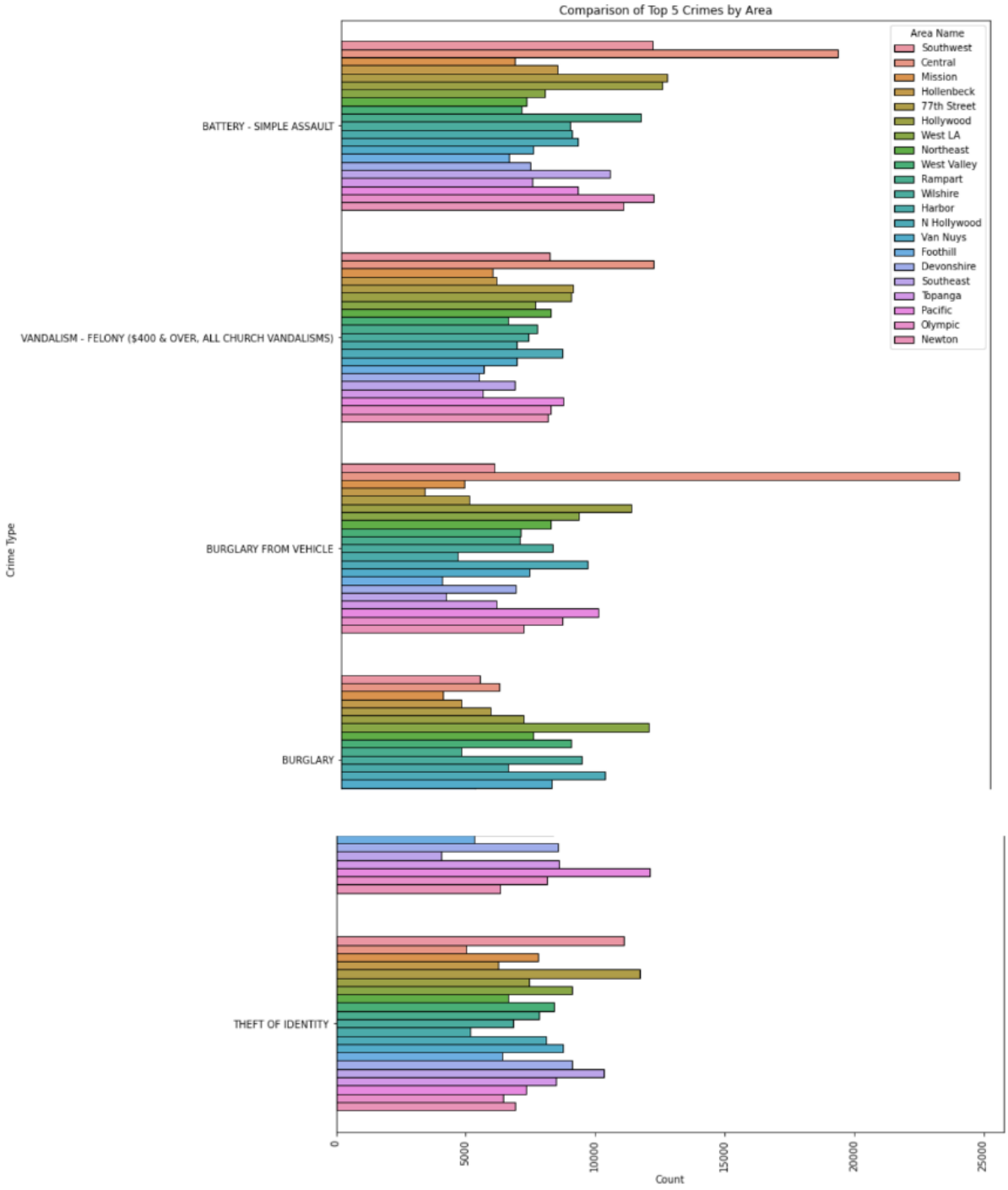
In [132]:
```
# Factor 1(Victim Gender)
top_5_crime_types = data['Crm Cd Desc'].value_counts().head(5).index
filtered_df = data[data['Crm Cd Desc'].isin(top_5_crime_types)]

plt.figure(figsize=(16, 10))
sns.countplot(y='Crm Cd Desc', data=filtered_df, hue='Vict Sex')
plt.title('Comparison of Top 5 Crimes by Gender')
plt.ylabel('Crime Type')
plt.xlabel('Count')
plt.xticks(rotation=45)
plt.legend(title='Gender')
plt.tight_layout()
plt.show()
```



In [131]:
```
# Factor 2(Victim Age Group)
top_5_crime_types = data['Crm Cd Desc'].value_counts().head(5).index
filtered_df = data[data['Crm Cd Desc'].isin(top_5_crime_types)]

plt.figure(figsize=(16, 10))
sns.countplot(y='Crm Cd Desc', data=filtered_df, hue='Age Group')
plt.title('Comparison of Top 5 Crimes by Age Group')
plt.ylabel('Crime Type')
plt.xlabel('Count')
plt.xticks(rotation=90)
plt.legend(title='Age Group')
plt.show()
```



In [130]:
```
# Factor 3(Area)
top_5_crime_types = data['Crm Cd Desc'].value_counts().head(5).index
filtered_df = data[data['Crm Cd Desc'].isin(top_5_crime_types)]

plt.figure(figsize=(12, 20))
sns.countplot(y='Crm Cd Desc', data=filtered_df, hue='AREA NAME', edgecolor='black')
plt.title('Comparison of Top 5 Crimes by Area')
plt.ylabel('Crime Type')
plt.xlabel('Count')
plt.xticks(rotation=90)
plt.legend(title='Area Name')
plt.show()
```


Comparison of Top 5 Crimes by Area
Area Name
Southwest
Central
Mission
Hollenbeck
77th Street
Hollywood
West LA
Northeast
West Valley
Rampart
Wilshire
Harbor
N Hollywood
Van Nuys
Foothill
Devonshire
Southeast
Topanga
Pacific
Olympic
Newton
BATTERY - SIMPLE ASSAULT
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
BURGLARY FROM VEHICLE
BURGLARY
THEFT OF IDENTITY
Crime Type
0
5000
10000
15000
20000
25000
Count

Q 10 Predicting Future Trends

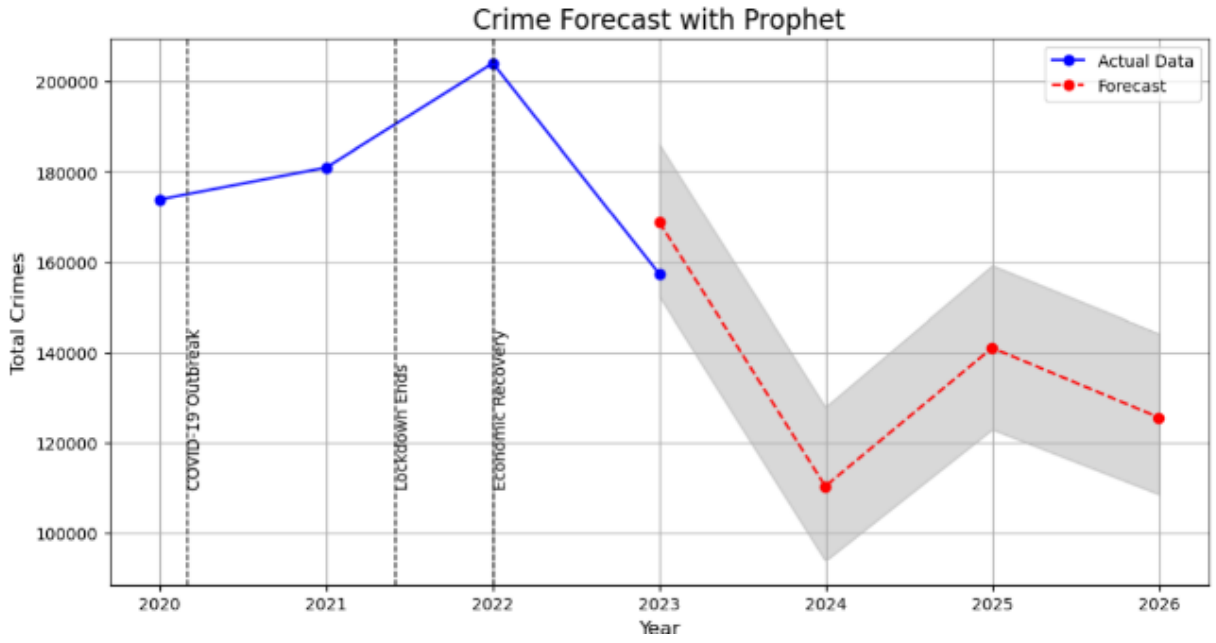
10. Predicting Future Trends:

In [80]:
```
print(total_crimes_by_year)
```

```
YEAR
2020    173872
2021    180951
2022    204068
2023    157302
dtype: int64
```

In [86]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from prophet import Prophet
from sklearn.metrics import mean_absolute_error, mean_squared_error
import math

data1 = pd.DataFrame({
    'ds': pd.to_datetime(['2020-01-01', '2021-01-01', '2022-01-01', '2023-01-01']),
    'y': [173872, 180951, 204068, 157302]
})

model = Prophet()

model.fit(data1)

future = model.make_future_dataframe(periods=3, freq='Y')

forecast = model.predict(future)

forecasted_data = forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail(4)

print("Forecasted Crime Data for the Next 3 Years:")
print(forecasted_data)
print("\nds: This column represents the date for each forecasted year.\n In this case, the dates correspond to the end of each year (December 31st).")
print("\nyhat: This column represents the forecasted value for the total number of crimes.\n It's the central estimate of the forecasted value.")
print("\nyhat_lower: This column represents the lower bound of the forecasted value.\n It provides a lower estimate of the expected value. Values are not expected to fall
print("\nyhat_upper: This column represents the upper bound of the forecasted value.\n It provides an upper estimate of the expected value. Values are not expected to exce

fig, ax = plt.subplots(figsize=(12, 6))

plt.plot(data1['ds'], data1['y'], label='Actual Data', color='b', marker='o')

plt.plot(forecasted_data['ds'], forecasted_data['yhat'], label='Forecast', color='r', linestyle='--', marker='o')

plt.fill_between(forecasted_data['ds'], forecasted_data['yhat_lower'], forecasted_data['yhat_upper'], alpha=0.3, color='gray')

observations = {
    'COVID-19 Outbreak': '2020-03-01',
    'Lockdown Ends': '2021-06-01',
    'Economic Recovery': '2022-01-01'
}

for label, date in observations.items():
    plt.axvline(pd.to_datetime(date), color='k', linestyle='--', linewidth=1)
    plt.text(pd.to_datetime(date), forecasted_data['yhat'].min(), label, rotation=90)

plt.title("Crime Forecast with Prophet", fontsize=16)
plt.xlabel("Year", fontsize=12)
plt.ylabel("Total Crimes", fontsize=12)
plt.legend()

mae = mean_absolute_error(data1['y'][:-1], forecasted_data['yhat'][:-1])
mse = mean_squared_error(data1['y'][:-1], forecasted_data['yhat'][:-1])
rmse = math.sqrt(mse)
mape = (abs((data1['y'][:-1] - forecasted_data['yhat'][:-1]) / data1['y'][:-1])).mean() * 100

print(f"Mean Absolute Error (MAE): {mae:.2f}")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.2f}")
print(f"Mean Absolute Percentage Error (MAPE): {mape:.2f}%")

plt.grid()
```

```
Forecasted Crime Data for the Next 3 Years:
          ds           yhat     yhat_lower     yhat_upper
3 2023-01-01  168771.107706  152272.051104  186102.532745
4 2023-12-31  110365.413912   93814.823054  127939.001247
5 2024-12-31  141017.220740  122946.703731  159199.288931
6 2025-12-31  125625.084130  108614.309639  144194.719603

ds: This column represents the date for each forecasted year.
 In this case, the dates correspond to the end of each year (December 31st).

yhat: This column represents the forecasted value for the total number of crimes.
 It's the central estimate of the forecasted value.

yhat_lower: This column represents the lower bound of the forecasted value.
 It provides a lower estimate of the expected value. Values are not expected to fall below this lower bound.

yhat_upper: This column represents the upper bound of the forecasted value.
 It provides an upper estimate of the expected value. Values are not expected to exceed this upper bound.

Mean Absolute Error (MAE): 46245.75
Mean Squared Error (MSE): 2994581610.29
Root Mean Squared Error (RMSE): 54722.77
Mean Absolute Percentage Error (MAPE): nan%
```



Introduction:

The following report presents a crime data forecasting analysis using the Prophet model. Accurate crime data forecasting is essential for informed decision-making, resource allocation, and policy development.

Data Description:

The dataset used in this analysis covers the years 2020 to 2023. It includes columns for the date ('ds') and the number of reported crimes ('y').

Forecasted Crime Data for the Next 3 Years:
The Prophet model was applied to forecast crime data for the years 2023, 2024, and 2025. The results are as follows:

- **2023-01-01**: The forecasted crime count is 168,771.11, with a lower bound of 152,272.05 and an upper bound of 186,102.53.
- **2023-12-31**: The forecasted crime count is 110,365.41, with a lower bound of 93,814.82 and an upper bound of 127,939.00.
- **2024-12-31**: The forecasted crime count is 141,017.22, with a lower bound of 122,946.70 and an upper bound of 159,199.29.
- **2025-12-31**: The forecasted crime count is 125,625.08, with a lower bound of 108,614.31 and an upper bound of 144,194.72.

Interpretation of Columns:

- **ds**: This column represents the date for each forecasted year, corresponding to the end of each year (December 31st).
- **yhat**: This column represents the central estimate of the forecasted crime count.
- **yhat_lower**: This column provides a lower estimate of the expected value, and values are not expected to fall below this lower bound.
- **yhat_upper**: This column provides an upper estimate of the expected value, and values are not expected to exceed this upper bound.

Model Evaluation:

- Mean Absolute Error (MAE): 46,245.75
- Mean Squared Error (MSE): 2,994,581,610.29
- Root Mean Squared Error (RMSE): 54,722.77
- Mean Absolute Percentage Error (MAPE): Not Available (nan%)

Observations and Insights:
The forecasted crime data for the next three years shows varying trends and uncertainty intervals. The Prophet model provides central estimates, lower bounds, and upper bounds, enabling a comprehensive understanding of the expected crime counts.

Conclusion:
The crime data forecasting analysis indicates that while the Prophet model offers valuable insights into future crime trends, there is inherent uncertainty in the predictions. The provided metrics, including MAE, MSE, and RMSE, assist in assessing the model's performance and forecasting accuracy.

Recommendations:
Based on the analysis, it is recommended that policymakers and law enforcement agencies consider the forecasted crime data in their decision-making processes. The uncertainty intervals (yhat_lower and yhat_upper) should be taken into account when planning resource allocation and crime prevention strategies.

Future Work:
Future research can focus on improving the forecasting model's accuracy by incorporating additional data sources, considering external factors, and exploring advanced modeling techniques.

Part 5: Conclusions

Throughout this academic project, we rigorously engaged in data preprocessing and exploratory analysis with a professional approach. The dataset in question, centered around crime data, presented us with an opportunity to apply data cleaning techniques and ensure the reliability of the data for potential future use.

We meticulously dropped certain columns that were not pertinent to our focus, ensuring that the dataset remained streamlined and relevant. The handling of missing data was approached with precision; we strategically imputed missing values in certain columns and removed rows with null values, thereby ensuring data consistency.

Our work, although preliminary, laid a strong foundation for any future explorations or analyses that could be conducted on this dataset. By carefully cleaning and pre-processing the data, we have ensured that it is in an optimal state for further academic exploration. The skills and techniques applied in this project are reflective of a professional approach to data management and analysis.

In conclusion, this project served as a valuable exercise in data handling and exploratory analysis within an academic context. Our efforts were concentrated on preparing the dataset for potential future analyses while ensuring data integrity and cleanliness. The experience has been enriching and stands testament to our team's capability to approach data systematically and professionally.