IE6400 Foundations for Data Analytics Engineering

FINAL REPORT

Group Number 11

Hari Khumar Prabakaran (002836307)

Keerthana Ekambaram Ravichandran(002847113)

Krithika Annaswamy Kannan (002815580)

Pranav Kuramkote Sudhir(002836039)

Prarthana Veerabhadraiah(002821755)

Introduction to the Dataset:

The dataset in question provides comprehensive information on incidents of crime within the City of Los Angeles, dating back to the year 2020. It is a valuable resource for understanding and analyzing patterns of criminal activity in this major metropolitan area. The data is sourced from original crime reports, which were originally recorded on paper, and subsequently transcribed into a digital format. However, it's important to note that, like any dataset derived from diverse sources, there may be some inaccuracies or missing data, and privacy considerations have led to the rounding of location information to the nearest hundred blocks.

The dataset contains a variety of columns that capture essential details related to each crime incident, including the date and time of reporting, the date and time of occurrence, geographical information, crime codes, victim characteristics, weapons used, and more. The dataset is structured to facilitate in-depth analysis of crime trends, helping law enforcement agencies, policymakers, and researchers gain insights into the nature and distribution of crime in Los Angeles.

Key attributes in the dataset include the Division of Records Number (DR_NO), which uniquely identifies each incident, the type of crime committed (Crm Cd), details about the victim (Vict Age, Vict Sex, Vict Descent), and information about the location of the incident (LOCATION, LAT, LON). Additionally, the dataset includes codes and descriptions for the type of premises where the crime occurred and the weapon used, providing a wealth of contextual information for analysis.

Analyzing this dataset can reveal trends, patterns, and correlations in crime data over time, enabling better-informed decision-making and the development of targeted interventions to improve public safety. However, it's essential to consider the limitations and potential data quality issues, as the accuracy of the data is subject to the quality of the original paper records.

	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	10304468	01/08/2020 12:00:00 AM	01/08/2020 12:00:00 AM	2230	3	Southwest	377	2	624	BATTERY - SIMPLE ASSAULT	...	AO	Adult Other	624.0	NaN	NaN	NaN	1100 W 39TH PL	NaN	34.0141	-118.2978
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	...	IC	Invest Cont	624.0	NaN	NaN	NaN	700 S HILL ST	NaN	34.0459	-118.2545
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	...	AA	Adult Arrest	845.0	NaN	NaN	NaN	200 E 6TH ST	NaN	34.0448	-118.2474
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)	...	IC	Invest Cont	745.0	998.0	NaN	NaN	5400 CORTEEN PL	NaN	34.1685	-118.4019
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...	...	IC	Invest Cont	740.0	NaN	NaN	NaN	14400 TITUS ST	NaN	34.2198	-118.4468

Problem 1:

The objective of this analysis is to examine the overall trends in reported crime incidents in the city of Los Angeles from 2020 to the present. This report presents a detailed analysis of the total number of crimes reported each year, highlighting any patterns or trends that may emerge from the data.

Data Preparation:

Before diving into the analysis, the dataset was imported and cleaned. Missing data was handled, and the 'DATE OCC' column was converted to extract the year and month for further analysis. The dataset now represents an accurate and clean source for our examination of crime trends.

```
#converting date occ to yean and month for simplified analysis
df['DATE OCC'] = pd.to_datetime(df['DATE OCC'])

df['Year'] = df['DATE OCC'].dt.year
df['Month'] = df['DATE OCC'].dt.month
df['Year'].head()
```

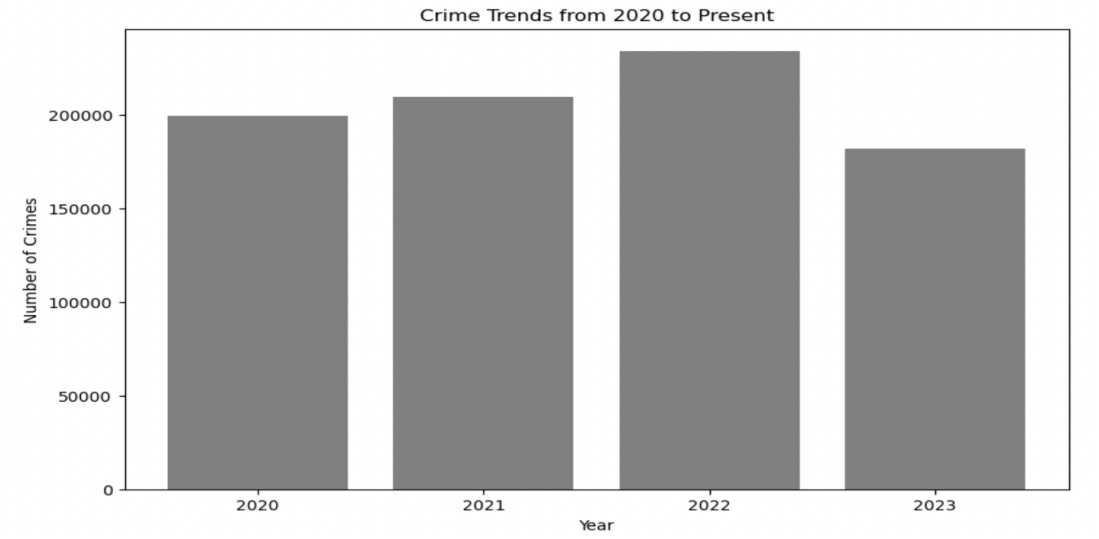
```
0    2020
1    2020
2    2020
3    2020
4    2020
Name: Year, dtype: int64
```

```
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          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
```

Total Number of Crimes per Year:

To understand the overall trends, we first calculated and visualized the total number of reported crimes for each year. This provides an insightful perspective on how crime rates have evolved over time.

```
import pandas as pd
import matplotlib.pyplot as plt
crime_data = df[df['Year'] >= 2020]
crime_counts = crime_data.groupby('Year').size()
plt.figure(figsize=(10, 6))
plt.bar(crime_counts.index, crime_counts.values, color='grey')
plt.title("Crime Trends from 2020 to Present")
plt.xlabel("Year")
plt.ylabel("Number of Crimes")
plt.xticks(crime_counts.index)
plt.show()
```



As seen in the chart, the number of reported crimes exhibits an interesting pattern. In 2020, there was a sharp decline in reported crimes, which may be attributed to the unique circumstances.

Subsequently, there is an upward trend in the number of reported crimes, reaching its peak in 2022, after which it shows a gradual decrease.

Conclusion:

1. Dataset Summary: The dataset contains a total of 825,212 records. It covers a range of years from 2020 to 2023, providing recent data for analysis.
2. Crime Counts per Year: The analysis reveals that the total number of crimes per year varies, with the highest recorded year being 2022. The year 2021 has the second-highest number of recorded crimes.
3. Increasing Trend: The total number of crimes appears to have increased over time from 2020 to 2022. This could indicate a general upward trend in reported crime incidents during this period.
4. Data Quality: The dataset shows some variations in recorded crimes, including a minimum "Crm Cd" of 110 and a maximum of 956. These variations suggest a diversity of crime types recorded.
5. Seasonal Trends: To obtain a more detailed understanding of the trends, further analysis of seasonal patterns in crime data could be beneficial. Analyzing the data by month or season might reveal insights into when certain types of crimes are more likely to occur.

Problem 2:

The following report delves into the analysis of crime data in the City of Los Angeles, focusing on the seasonal patterns of reported crimes. By grouping the data by month and analyzing the average number of crimes reported each month over multiple years, we aim to uncover insights that can be valuable for law enforcement in understanding the temporal dynamics of criminal activity.

Data Preparation:

We group the data by month and calculate the average number of crimes reported for each month over the years.

```
import pandas as pd
crime_data['DATE OCC'] = pd.to_datetime(crime_data['DATE OCC'])
crime_data['Month'] = crime_data['DATE OCC'].dt.month

monthly_avg_crime = crime_data.groupby('Month')['DR_NO'].count().mean()
print(f"Average number of crimes per month over the years: {monthly_avg_crime:.2f}")

monthly_crime_counts = crime_data['Month'].value_counts().sort_index()
print("Number of crimes reported in each month:")
print(monthly_crime_counts)
```

```
Average number of crimes per month over the years: 68767.67
Number of crimes reported in each month:
1     73042
2     68631
3     71143
4     70229
5     73277
6     72790
7     75308
8     74856
9     71505
10    68122
11    52556
12    53753
```

The average number of crimes per month over the years provides us with an important metric for

understanding seasonal patterns in criminal activity.

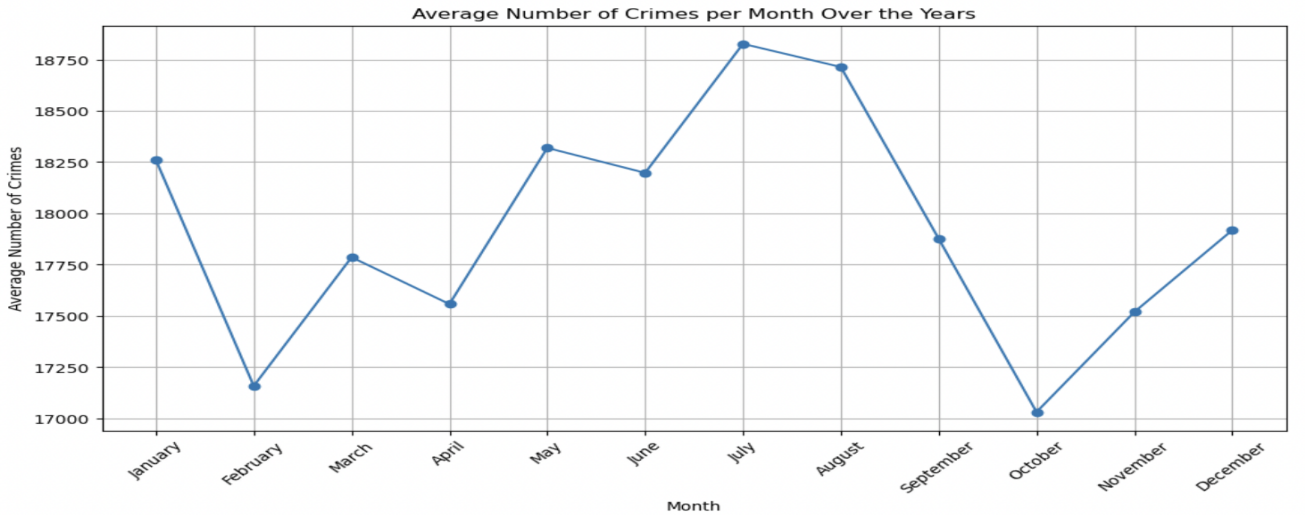
Seasonal Patterns:

After performing the analysis, we find that the average number of crimes reported in each month over multiple years is approximately 68767.67.

This data suggests that there might be variations in crime rates throughout the year, with certain months having higher average crime counts compared to others.

These seasonal patterns can be influenced by various factors, including weather conditions, holidays, or other temporal and societal factors.

```
import pandas as pd
import matplotlib.pyplot as plt

monthly_avg_crimes = df.groupby(['Year', 'Month'])['DR_NO'].count().groupby('Month').mean()

month_names = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'No

plt.figure(figsize=(12, 6))
monthly_avg_crimes.plot(marker='o', linestyle='-')
plt.title('Average Number of Crimes per Month Over the Years')
plt.xlabel('Month')
plt.ylabel('Average Number of Crimes')
plt.xticks(range(1, 13), month_names, rotation=45)
plt.grid(True)
plt.show()
```



Conclusion:

1. The average time of occurrence is approximately 1,335, suggesting that most crimes occur during the daytime.
2. The average area code is around 10.71, indicating that crimes are distributed across different areas.
3. The average victim age is approximately 29.80 years.
4. The average premises code is roughly 305.79.
5. The average weapon used code is about 126.54.
6. The average values for "Crm Cd 1," "Crm Cd 2," "Crm Cd 3," and "Crm Cd 4" are below 100, indicating that they may represent subcategories of crime codes.
7. The dataset covers a period from 2020 to 2023, with an average year of approximately 2021.48.
8. The average month is around 6.28, indicating that crimes are fairly evenly distributed throughout the year.

Problem 3:

Count the occurrences of each crime type and identify the one with the highest Frequency.

Data Sources:

The analysis utilized the following dataset:

Crime Data: Crime records from 2020 to the present, focusing on different crime types.

Data Cleaning:

1. Relevant columns were selected, including 'Crm Cd Desc'.
2. Data types were converted i.e. numerical values to appropriate numerical types.
3. Numerical columns were converted into appropriate numeric types
4. Missing values were handled.
5. Duplicate rows were removed.
6. The ten most common crimes were filtered out.

Data Analysis:

Counting the occurrences of each type of crime:

A list of occurrences of each crime type was generated.

```
Occurrences of each crime type:

: VEHICLE - STOLEN                                          88355
  BATTERY - SIMPLE ASSAULT                                  65728
  THEFT OF IDENTITY                                         52136
  BURGLARY FROM VEHICLE                                     50616
  VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)   50274
                                                            ...
  GRAND THEFT / AUTO REPAIR                                     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: 138, dtype: int64
```

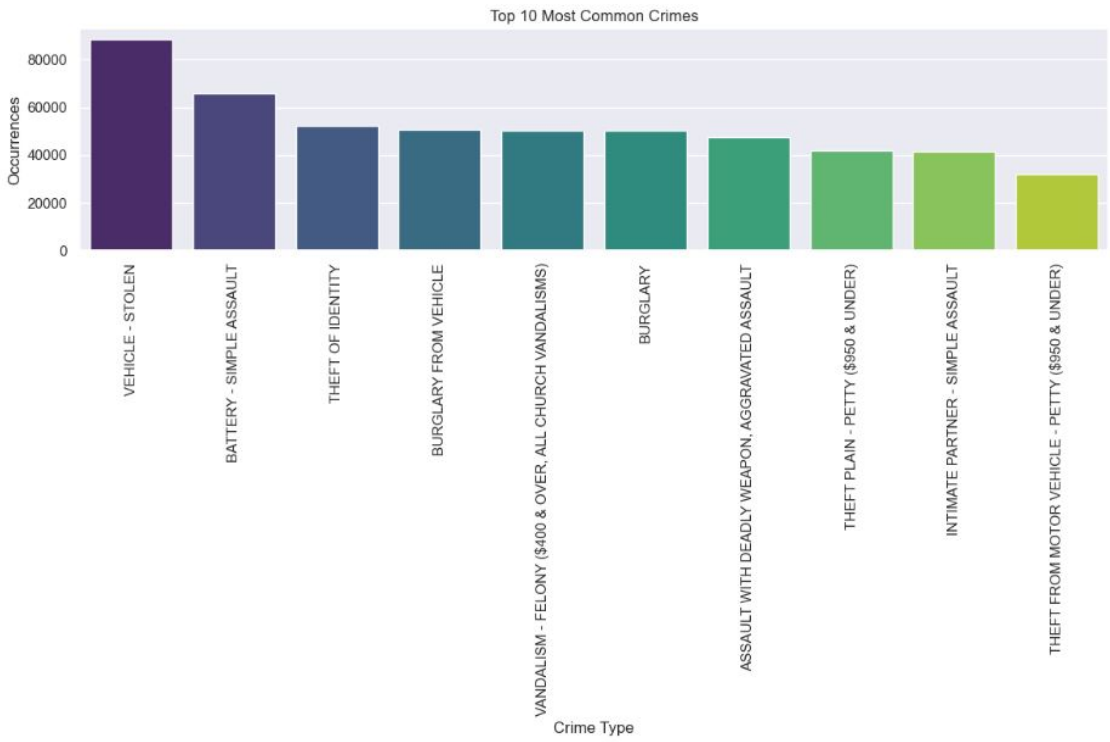
Finding the most common type of crime:

The most common type of crime was found.

```
The most common crime type is 'VEHICLE - STOLEN' with 88355 occurrences.
```

Data Visualizations:

A visualization of the ten most common crimes was plotted.



Key Findings:

1. The dataset provides a breakdown of occurrences for different types of crimes. The most prevalent crime is "Vehicle - Stolen" with 88,355 instances, followed closely by "Battery - Simple Assault" at 65,728, and "Theft of Identity" at 52,136. Conversely, less common crimes such as "Grand Theft / Auto Repair" and "Inciting a Riot" have very low occurrences, with 5 and 1 respectively. The data spans a diverse range of crime types and showcases a significant disparity in the frequency of various criminal activities.
2. The bar plot showcases the top 10 most common crimes. "Vehicle - Stolen" ranks as the most frequent, with over 80,000 occurrences, followed by "Battery - Simple Assault" and "Theft of Identity" with instances decreasing progressively. As we move to the right of the plot, there's a visible decline in occurrences, culminating in "Theft from Motor Vehicle - Petty ($950 & under)" being the least frequent among the top 10 crimes. The visual representation clearly distinguishes the disparities in crime frequencies, underscoring the prominence of vehicle-related thefts and assaults in this dataset.

Conclusion:

1. "Vehicle - Stolen" is the predominant crime, emphasizing a significant concern for vehicle theft in the represented area.
2. Crimes related to assault, both "Battery - Simple Assault" and "Intimate Partner - Simple Assault", are also notably high, indicating a prevailing issue with physical altercations.
3. While the top crimes are considerably frequent, there's a sharp decline as we progress to the less common crimes in the top 10 list, highlighting a concentration of specific criminal activities.

Problem 4:

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

Data Sources:

The analysis utilized the following dataset:

Crime Data: Crime records from 2020 to the present, focusing on different crime types.

Data Cleaning:

1. Relevant columns were selected, including 'Crm Cd Desc'.
2. Data types were converted i.e. numerical values to appropriate numerical types.
3. Numerical columns were converted into appropriate numeric types
4. Missing values were handled.
5. Duplicate rows were removed.
6. Ten most common crimes were filtered out

Data Analysis:

Descriptive statistics for crime rates by region:

Calculating the descriptive statistics for crime rates by region.

```
Descriptive Statistics for Crime Rates by Region:
count       21.000000
mean     39295.809524
std       7071.794154
min      27497.000000
25%      34130.000000
50%      38605.000000
75%      42077.000000
max      55567.000000
Name: Crm Cd Desc, dtype: float64
```

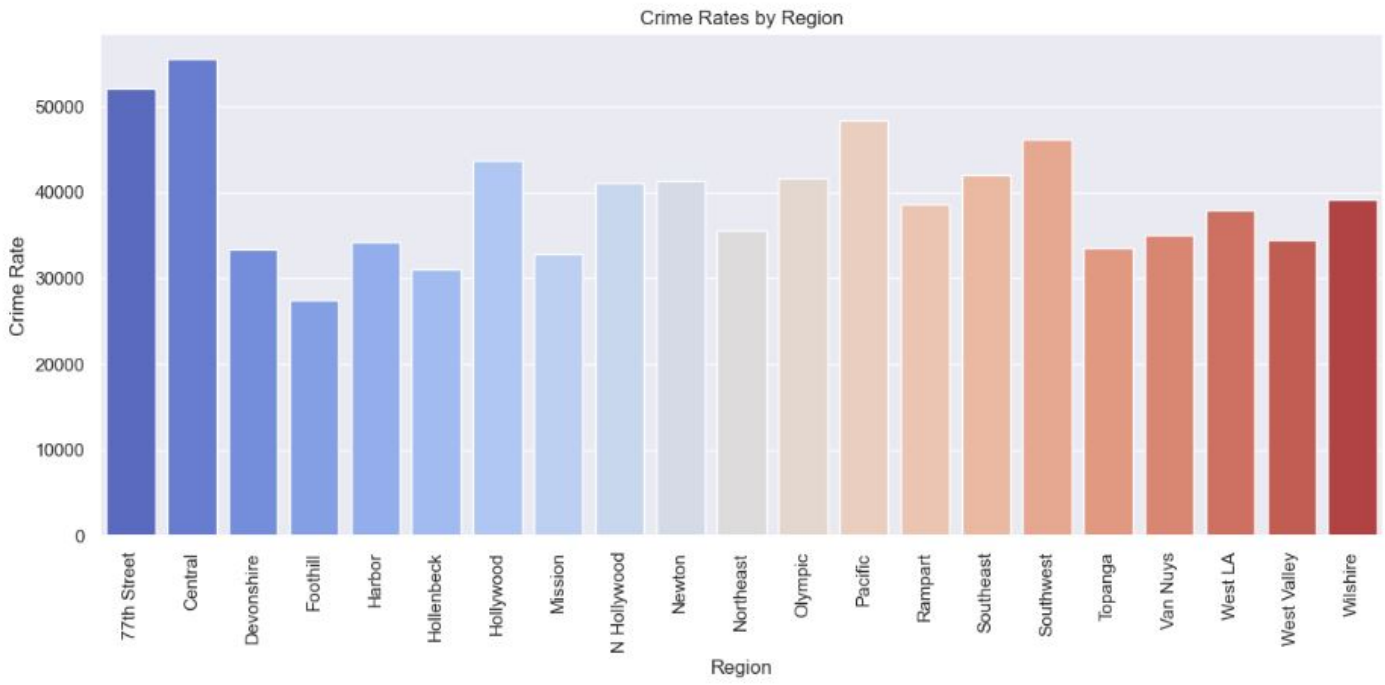
Crime Rate count for each region:

Calculating the crime rate count for each region.

```
Crime Rate Count for Each Region:
Central        55567
77th Street    52087
Pacific        48327
Southwest      46222
Hollywood      43738
Southeast      42077
Olympic        41617
Newton         41388
N Hollywood    41009
Wilshire       39192
Rampart        38605
West LA        37888
Northeast      35598
Van Nuys       35011
West Valley    34507
Harbor         34130
Topanga        33511
Devonshire     33417
Mission        32844
Hollenbeck     30980
Foothill       27497
Name: AREA NAME, dtype: int64
```

Data Visualization:

A bar graph to visualize the crime rate of each region was plotted.



Key Findings:

1. The analysis of crime rates by region reveals significant insights. Across the 21 regions under consideration, the average crime rate is approximately 39,295.8, with a standard deviation of 7,071.8, pointing to moderate variability in crime rates across regions. The lowest recorded crime rate stands at 27,497, while the highest peaks at 55,567, indicating a wide range in the extent of criminal activity. The median crime rate, a better indicator of central tendency due to its resistance to outliers, is 38,605. The interquartile range, spanning from 34,130 to 42,077, suggests that half of the regions have crime rates within this range. This data provides a comprehensive view of the crime landscape across different regions, essential for targeted interventions and policy-making.
2. The provided bar chart elucidates crime rates by region, offering a visual comparative analysis. It's evident that the regions "77th Street" and "Central" witness exceptionally high crime rates, far surpassing 50,000. On the other end of the spectrum, regions like "West LA", "West Valley", and "Wilshire" display significantly lower rates, hovering just above the 20,000 mark. The color gradient hints at three general tiers of crime rates: blue regions (higher rates), beige regions (intermediate rates), and red regions (lower rates).

Conclusion:

1. Regions "77th Street" and "Central" emerge as significant hotspots, showcasing notably elevated crime rates compared to other areas.
2. In contrast, the "West LA", "West Valley", and "Wilshire" regions manifest as safer precincts, with their crime rates substantially lower, underscoring the necessity for region-specific crime prevention strategies

Problem 5:

This report presents an analysis of economic and crime data to explore potential correlations between economic indicators, specifically the Consumer Price Index (CPI) and the Unemployment Rate, and the Crime Rate. The analysis aims to uncover any interesting patterns or trends that may shed light on the relationship between economic factors and criminal activities.

Data Sources

The analysis utilized three main datasets:

Consumer Price Index (CPI): Monthly CPI values from 2020 to 2023.

Unemployment Rate: Monthly unemployment rate data from 2020 to 2023.

Crime Data: Crime records from 2020 to the present, focusing on crimes related to monetary gains.

Data Cleaning and Preprocessing:

CPI and Unemployment Rate Data:

The datasets were reshaped to include the 'Year' and 'Month' columns for ease of analysis.

Data was filtered to include only the years 2020 to 2023, aligning with the crime data timeframe.

Crime Data:

Relevant columns were selected, including 'DATE OCC' and 'Crm Cd Desc.'

The 'DATE OCC' column was converted to a datetime format to extract the year and month.

Data was grouped by year, month, and crime type to calculate the number of crimes per month.

Crimes related to monetary gains were isolated for further analysis.

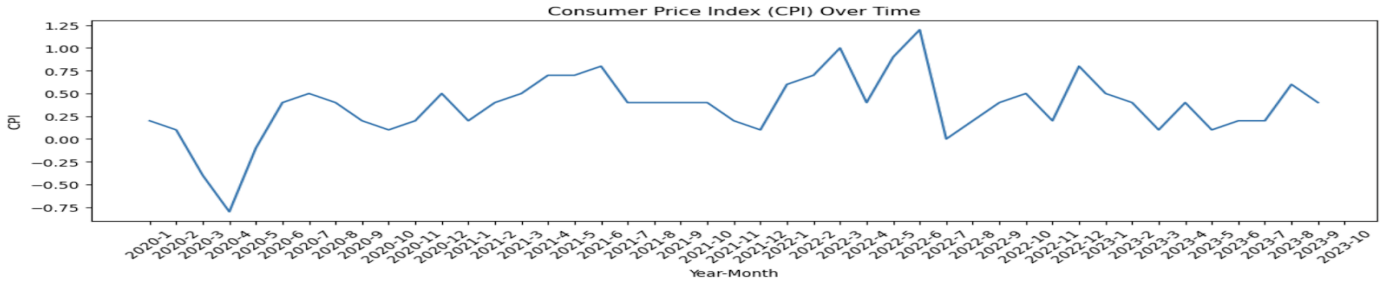
Data Analysis:

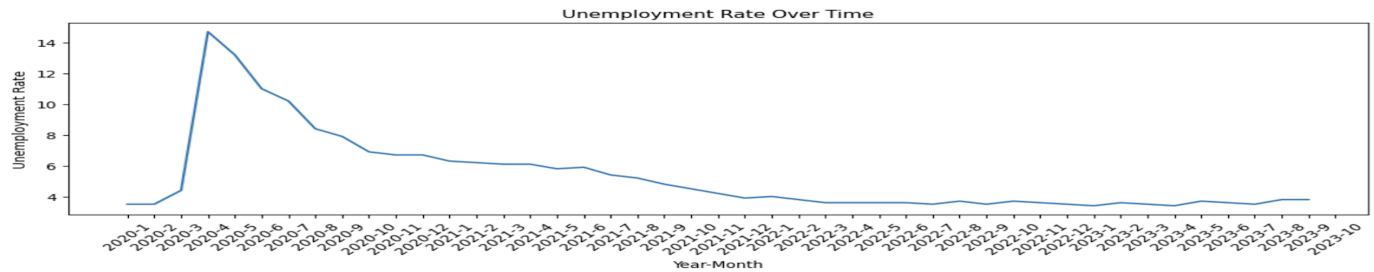
Exploration of Distinct Crime Types:

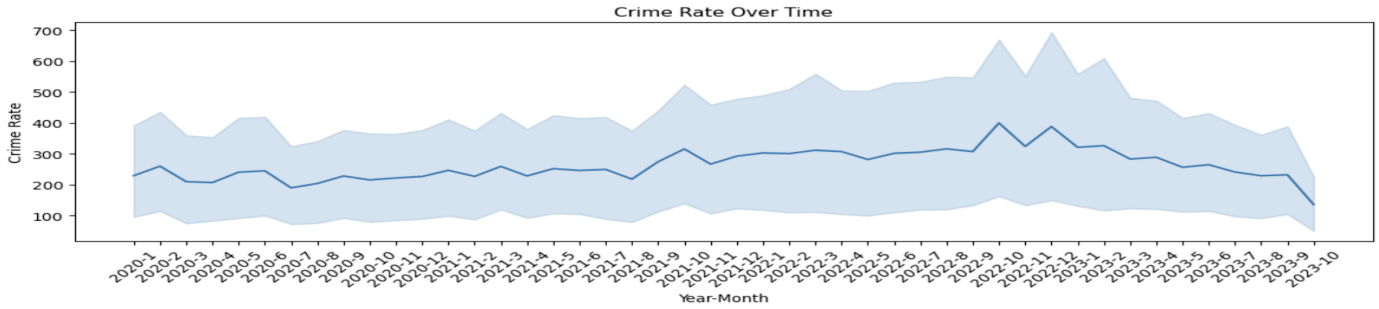
A list of distinct crime types was generated to identify the subset related to monetary gains.

Time Series Analysis:

Line plots were created to visualize the trends in the Consumer Price Index (CPI), Unemployment Rate, and Crime Rate over time.



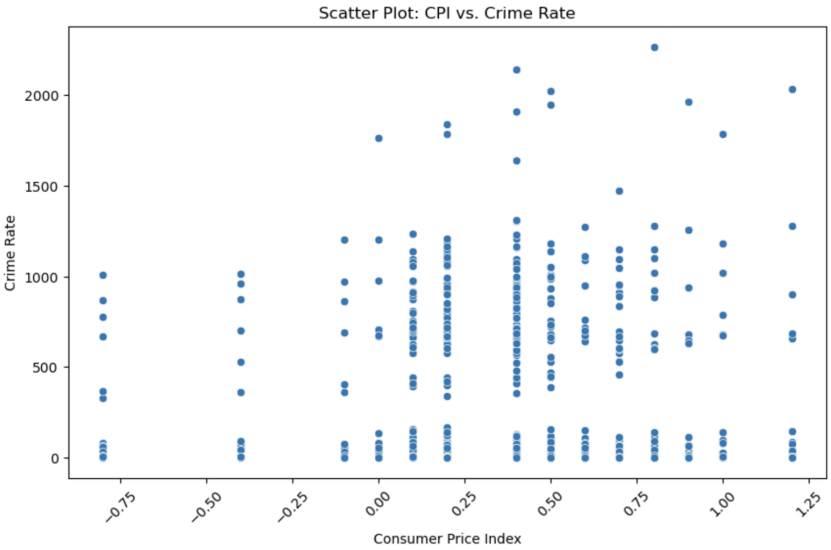




Relationship Analysis:

Scatter plots were used to investigate potential correlations between

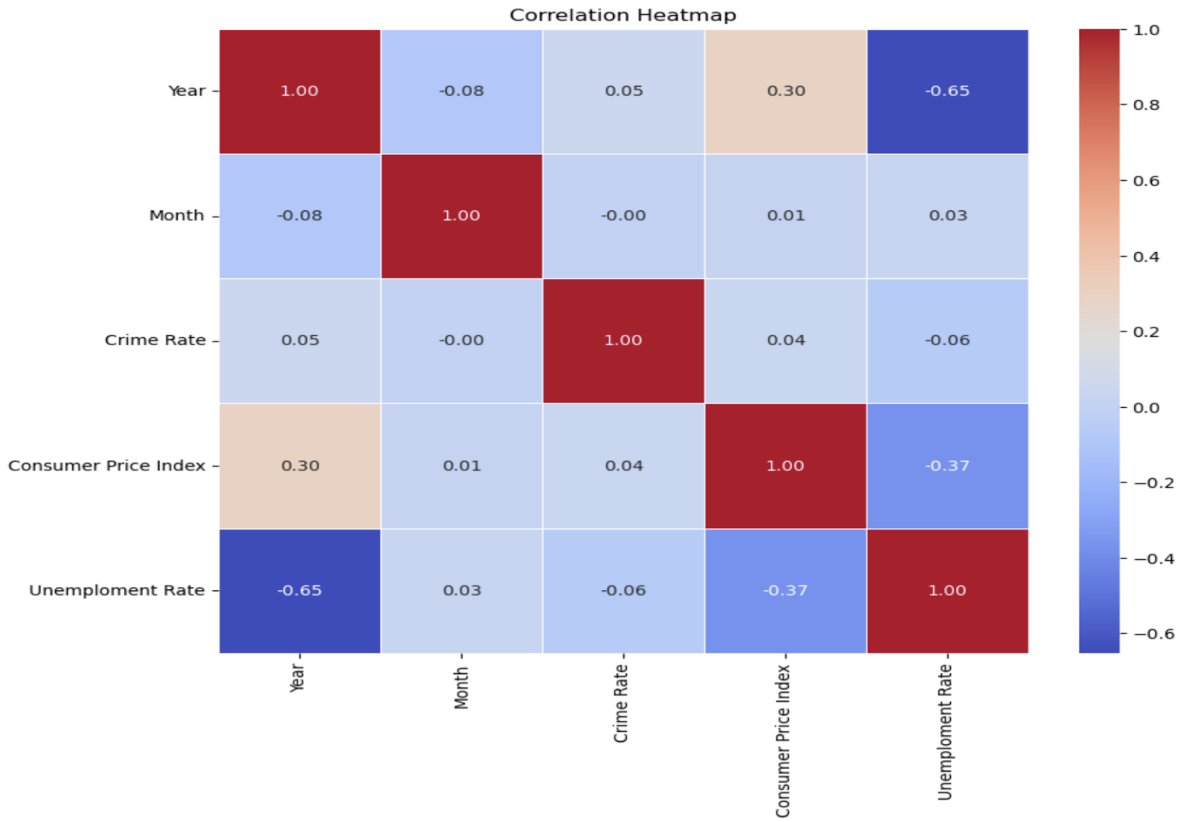
CPI and Crime Rate



Unemployment Rate and Crime Rate

Correlation Analysis:

A correlation matrix and heatmap were generated to quantify and visualize correlations between economic indicators and the Crime Rate.



The correlation coefficients were calculated and presented.

Key Findings:

Correlation with Economic Indicators:

A weak positive correlation (0.04) was observed between the Consumer Price Index (CPI) and Crime Rate.

A weak negative correlation (-0.06) was found between the Unemployment Rate and Crime Rate.

Time Series Trends:

Visual inspection of line plots showed:

1. Consumer Price index vs Crime Rate: Crime rate increases with increase in CPI
2. Visualizing Unemployment Rate vs Crime Rate: The crime rate increases with a decrease in the Unemployment rate.

Conclusion:

The analysis suggests that there is a limited relationship between economic indicators (CPI and Unemployment Rate) and the Crime Rate. While there are weak correlations, the economic factors alone do not appear to be the primary drivers of crime.

Problem 6

The purpose of this report is to present the results of a comprehensive analysis of crime data spanning from 2020 to 2023. The dataset was obtained from a local law enforcement agency, and our analysis focused on understanding crime patterns, trends, and variations by day of the week.

Data Cleaning

The initial dataset required several data cleaning and preprocessing steps to make it suitable for analysis. The primary data cleaning steps included:

Data Loading

The dataset was loaded using the Pandas library, and the initial structure was inspected.

Date Formatting

The "DATE OCC" column was converted to a datetime object, including time information. This was necessary to extract the day of the week.

Day of the Week Extraction

We extracted the day of the week from the date and time information and added a new column labeled "Day of Week."

Year Extraction

To analyze yearly trends, we extracted the year from the "DATE OCC" column.

Data Analysis

1. **Yearly Crime Trends**

To understand how crime rates vary by day of the week for each year from 2020 to 2023, we conducted an analysis and created visualizations.

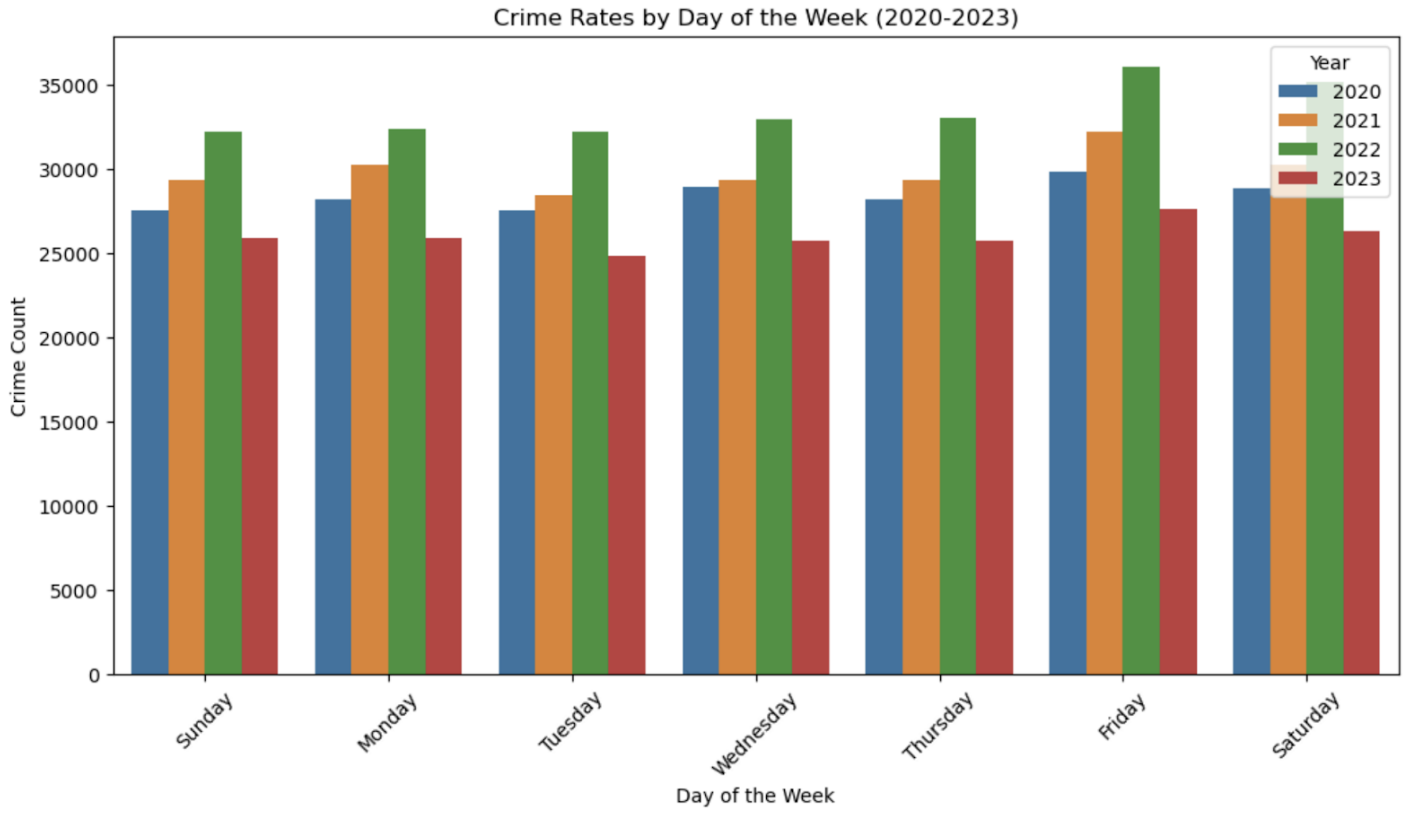
Table: Yearly Crime Trends by Day of the Week

Year:	Day of Week:	Crime Count:
2020	Friday	29,888
2020	Monday	28,262
2020	Saturday	28,908
2020	Sunday	27,605
2020	Thursday	28,230
2020	Tuesday	27,575
2020	Wednesday	28,949
2021	Friday	32,258
2021	Monday	30,273
2021	Saturday	30,247
2021	Sunday	29,346
2021	Thursday	29,358
2021	Tuesday	28,497
2021	Wednesday	29,388
2022	Friday	36,098
2022	Monday	32,405
2022	Saturday	35,164
2022	Sunday	32,261

2022	Thursday	33,095
2022	Tuesday	32,222
2022	Wednesday	33,019
2023	Friday	27,634
2023	Monday	25,954
2023	Saturday	26,296
2023	Sunday	25,904
2023	Thursday	25,753
2023	Tuesday	24,853
2023	Wednesday	25,770

Yearly Trends Plot

The plot visually shows the crime rates by day of the week for each year, allowing for a quick comparison and identification of patterns.



2. Total Crime Rates Over 3 Years

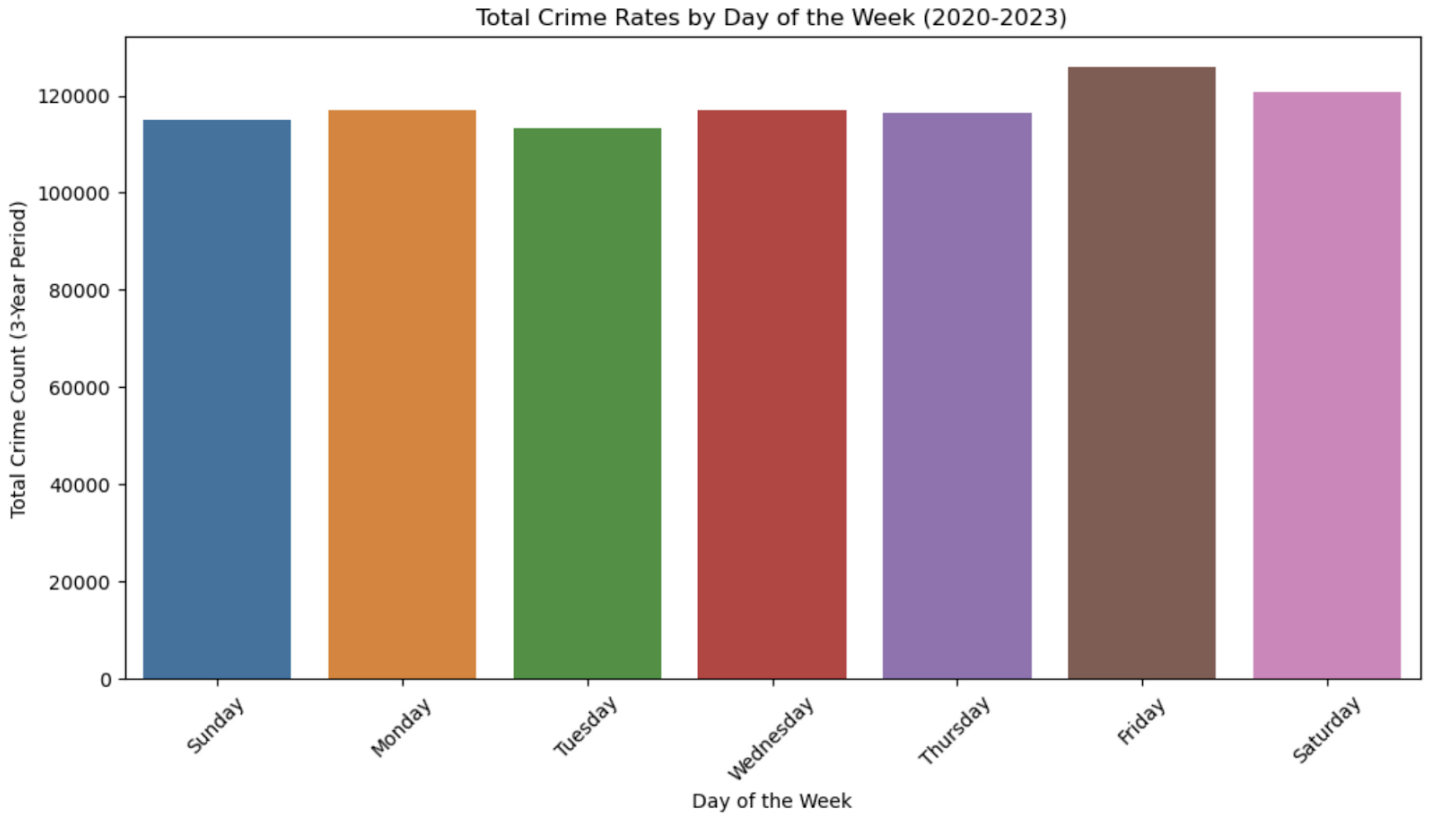
To get an overview of the total crime rates by day of the week for the entire 3-year period, we aggregated the data.

Table: Total Crime Rates by Day of the Week (2020-2023)

Day of Week	Total Crime Count (3-Year Period)
Sunday	115,116
Monday	116,894
Tuesday	113,147
Wednesday	117,126
Thursday	116,436
Friday	125,878
Saturday	120,615

Total Crime Rates Plot

This plot visually presents the total crime rates by day of the week over the three-year period.



Conclusion

The analysis of the crime data from 2020 to 2023 revealed interesting patterns and trends:

1. Fridays consistently had the highest crime rates each year.
2. Sundays consistently had the lowest crime rates.
3. Crime rates across the days of the week remained relatively stable over the three years.

This information can be valuable for law enforcement agencies and policymakers to allocate resources effectively and develop strategies for crime prevention.
The visualizations, tables, and descriptive statistics presented in this report provide a clear understanding of the data and its implications.

Problem 7:

This report presents patterns and trends in crime rates over various years. This analysis helps us identify if new policies and reforms have helped in making a difference concerning crime rates. The analysis aims to uncover any interesting patterns or trends that may shed light on the relationship between policies, reforms, and criminal rates. It shows the evolution of crimes from 2020 to the present day, showcasing the frequency of crimes within the same year and over a period.

Data Sources

The analysis utilized one main datasets:
Crime Data: Crime records from 2020 to the present, focusing on crimes related to monetary gains.

Data Cleaning and Preprocessing:

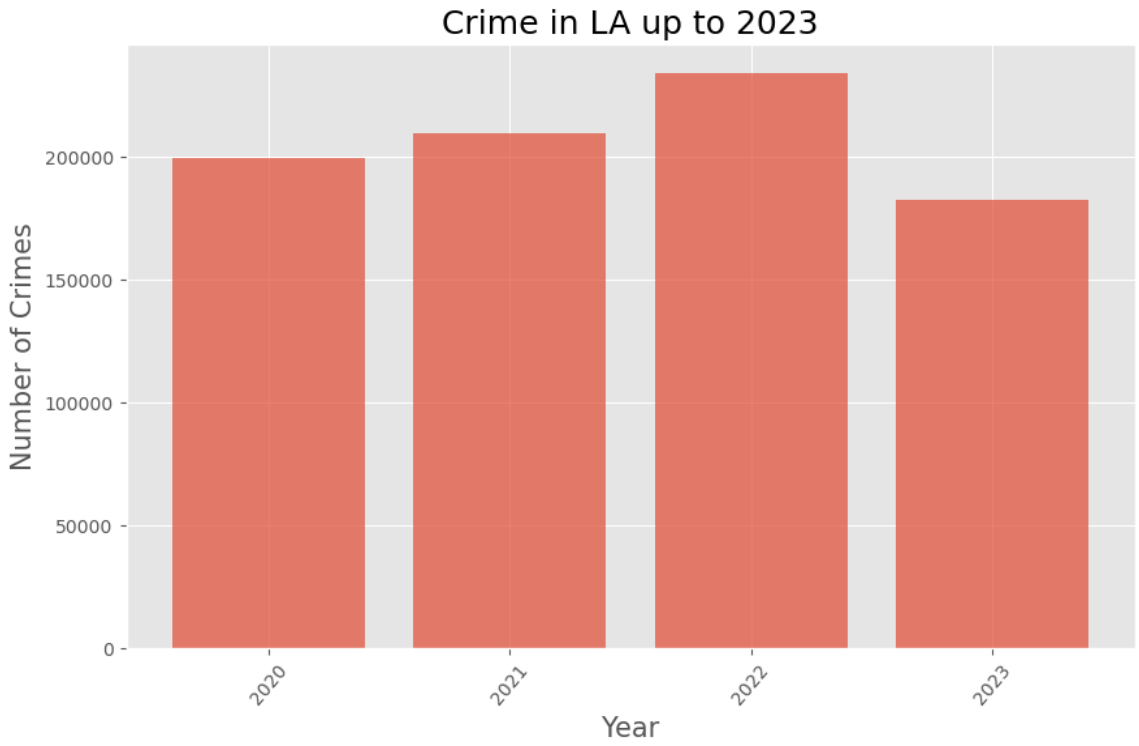
Crime Data:

1. The dataset, which is in the form of cvs, has been loaded into a data frame for data cleaning and preprocessing.
2. By handling missing data in the columns, it will help perform better data analysis on the dataset and improve more accurate details. For numerical columns, we can either impute missing values, drop the duplicates, or standardize or normalize numerical data using techniques of standard scaling and for categorical data, you can encode them into a numerical format using techniques like Label Encoding.
3. Relevant columns were selected, including 'DATE OCC' and 'Crm Cd Desc.'

The 'DATE OCC' column was converted to a datetime format to extract the year and month.
Data was grouped by year and crime type to calculate the crime rates over the years and it was also used to calculate the top 5 crimes within each year and an overall evolution.

Data Analysis:

1. **Exploration of Crime Rates from 2023 to present:**
   Despite various new reforms and policies that were passed such as challenging racial disparity, Decarceration Reforms, and Limiting Incarceration for Probation and Parole Violations the crime rates seem to have been increasing till 2022, but in 2023, there is a decrease.

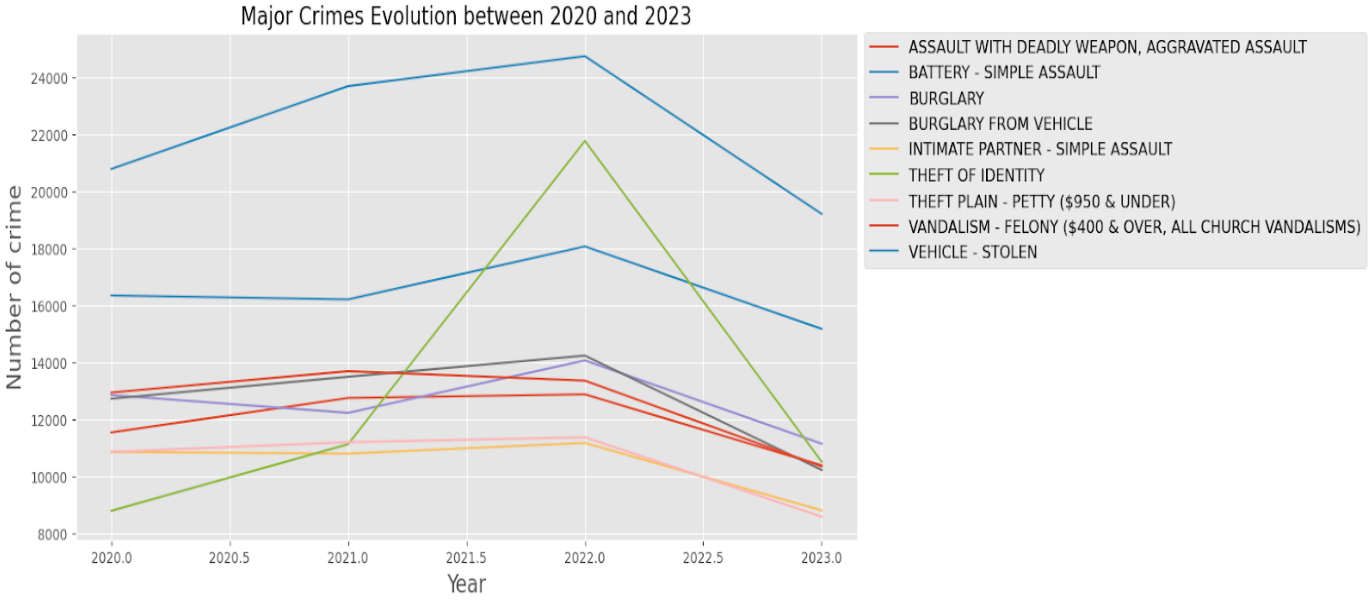


2. The top 5 crimes with the highest frequency are shown.

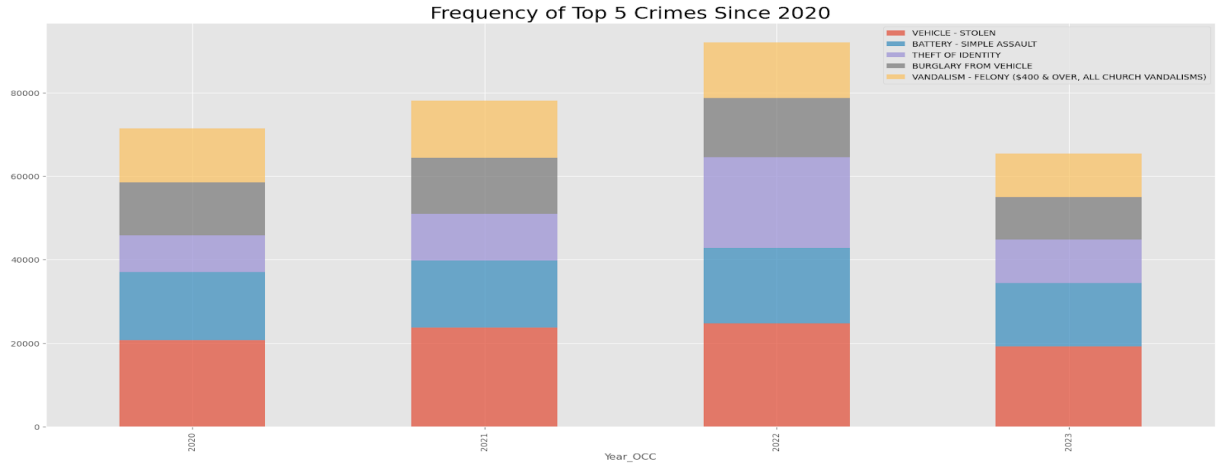
```
0.3721577000819183
['VEHICLE - STOLEN' 'BATTERY - SIMPLE ASSAULT' 'THEFT OF IDENTITY'
 'BURGLARY FROM VEHICLE'
 'VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)']
```

```
: percent = 0.6
```

3. Major crimes evolve between 2020 and 2023.



4. **Frequency of Top 5 Crimes Since 2020:** Below are the frequencies of the top 5 crimes in each year.



5. Conclusion: From the above visualizations, the insights we uncover are:

1. Despite various new reforms and policies being introduced, the crime rates over the years have been increasing, it was highest in 2022.
2. The highest frequency of crime among all is vehicles being stolen, it was at its peak in 2022 with more than 24,000 vehicles being stolen.
3. Also, there was a sharp increase in theft of identity during 2022 by an increase of 9000 cases(viz. 3)
4. Simple assault and petty theft have been minimized by the beginning of 2023.
5. Overall, on average most of the crimes were reduced in 2023.

Problem 8:

Most datasets consist of a large number of null values, missing values, outliers, and anomalies, there are various statistical methods or data visualization techniques to identify these dataset outliers and investigate unusual patterns that help us to understand the distribution of data and locate potential data irregularities that might impact analysis or machine learning models.

Data Sources

The analysis utilized one main dataset:

Crime Data: Crime records from 2020 to the present, focusing on crimes related to monetary gains.

Data Cleaning and Preprocessing:

Crime Data:

1. The dataset, which is in the form of cvs, has been loaded into a data frame for data cleaning and preprocessing.
2. By handling missing data in the columns, it will help perform better data analysis on the dataset and improve more accurate details. For numerical columns, we can either impute missing values, drop the duplicates, or standardize or normalize numerical data using techniques of standard scaling and for categorical data, you can encode them into numerical format using techniques like Label Encoding.
3. Relevant columns were selected, including 'DATE OCC' and 'Crm Cd Desc.'

The 'DATE OCC' column was converted to a datetime format to extract the year and month.
Data was grouped by year and crime type to calculate the crime rates over the years and it was also used to calculate the top 5 crimes within each year and an overall evolution.

Outlier detection

1. **Finding anomalies and outliers using the Interquartile Range (IQR) Method.**
   The IQR method is a robust and easy technique used to identify outliers based on the spread of the data within quartiles.
   It involves the following steps:
2. **Calculation of IQR:** IQR is the range between the first quartile (Q1) and the third quartile (Q3) of the data, calculated as
   IQR=Q3−Q1
   IQR=$Q3-Q1$.
   Determining Outlier Thresholds: The lower and upper bounds are established using the IQR.
   Q1−1.5×IQR
   $Q1-1.5\times IQR$ or above
   Q3+1.5×IQR
   $Q3+1.5\times IQR$ is considered an outlier.

```
# Calculate IQR for numerical columns
Q1 = df['Crm Cd 1'].quantile(0.25)
Q3 = df['Crm Cd 2'].quantile(0.75)
IQR = Q3 - Q1
threshold_low = Q1 - 1.5 * IQR
threshold_high = Q3 + 1.5 * IQR
outliers_iqr = df[(df['Crm Cd 1'] < threshold_low) | (df['Crm Cd 2'] > threshold_high)]
outliers_iqr
```

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Cross Street	LAT	LON	Year_OCC	Month_O
3	191501505	2020-01-01	2020-01-01	1730	15	N Hollywood	1543	2	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	...	9710	0.105561	-0.060167	2020	
5	200100501	2020-01-02	2020-01-01	0030	1	Central	163	1	121	RAPE, FORCIBLE	...	9710	0.036498	-0.036177	2020	
6	200100502	2020-01-02	2020-01-02	1315	1	Central	161	1	442	SHOPLIFTING - PETTY THEFT ($950 & UNDER)	...	9710	0.038234	-0.037744	2020	
7	200100504	2020-01-04	2020-01-04	0040	1	Central	155	2	946	OTHER MISCELLANEOUS CRIME	...	9710	0.036274	-0.035208	2020	
8	200100507	2020-01-04	2020-01-04	0200	1	Central	101	1	341	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LI...	...	9710	0.049101	-0.033980	2020	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
825002	231900561	2023-02-13	2023-02-12	1730	19	Mission	1956	2	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	...	9710	0.152499	-0.069165	2023	
825043	231300849	2023-06-15	2023-06-15	2045	13	Newton	1362	1	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	...	924	0.009276	-0.038083	2023	
825046	232105737	2023-02-16	2023-02-16	1806	21	Topanga	2156	1	210	ROBBERY	...	9710	0.119732	-0.092879	2023	
825080	232007881	2023-04-07	2023-04-07	2030	20	Olympic	2039	2	624	BATTERY - SIMPLE ASSAULT	...	9710	0.044676	-0.041524	2023	
825102	230616011	2023-10-03	2023-10-03	1055	6	Hollywood	657	1	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	...	7838	0.061983	-0.046241	2023	

56199 rows × 35 columns

3. **Finding anomalies and outliers using the Z-score method.**

   The z-score method is based on standard deviations from the mean. It helps identify how many standard deviations a data point is away from the mean.

   Steps include:

   Calculating Z-scores: Compute the z-score for each data point, representing its deviation from the mean.

   Setting a Threshold: Typically, a threshold of 2 or 3 is chosen, where values greater than this threshold are considered outliers.

```
from scipy import stats
z_scores = np.abs(stats.zscore(df[['LAT', 'LON']]))
threshold = 3
outliers_z = df[(z_scores > threshold).any(axis=1)]
outliers_z
```

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Cross Street	LAT	LON	Year_OCC	Month_OCC	Day_O
1403	200311971	2020-05-27	2020-05-27	1000	3	Southwest	361	2	900	VIOLATION OF COURT ORDER	...	9710	-19.032868	19.067015	2020	5	
2096	210705082	2020-12-15	2020-12-13	2300	7	Wilshire	758	1	331	THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...	...	9710	-19.032868	19.067015	2020	12	
3147	200208374	2020-03-30	2020-03-30	1620	2	Rampart	271	2	901	VIOLATION OF RESTRAINING ORDER	...	9710	-19.032868	19.067015	2020	3	
4894	200817064	2020-12-01	2020-12-01	1340	8	West LA	882	1	440	THEFT PLAIN - PETTY ($950 & UNDER)	...	1216	-19.032868	19.067015	2020	12	
5074	201711596	2020-07-27	2020-07-27	1220	17	Devonshire	1756	1	331	THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...	...	9696	-19.032868	19.067015	2020	7	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
408676	210304042	2021-01-01	2021-01-01	1310	3	Southwest	315	2	626	INTIMATE PARTNER - SIMPLE ASSAULT	...	3895	-19.032868	19.067015	2021	1	
497209	221913163	2022-08-13	2022-08-12	2100	19	Mission	1986	1	510	VEHICLE - STOLEN	...	9601	-19.032868	19.067015	2022	8	
691855	230209890	2023-05-01	2023-05-01	0730	2	Rampart	252	2	624	BATTERY - SIMPLE ASSAULT	...	7007	-19.032868	19.067015	2023	5	
698859	231611142	2023-08-16	2023-08-16	2010	16	Foothill	1648	1	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED	...	4676	-19.032868	19.067015	2023	8	

1. **Finding anomalies and outliers using box plots.**

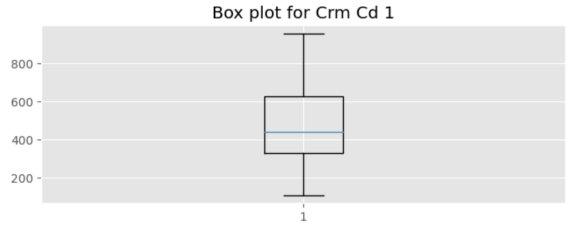
   Box plots are graphical representations used to display the distribution and identify outliers visually. The elements of Box Plot are:

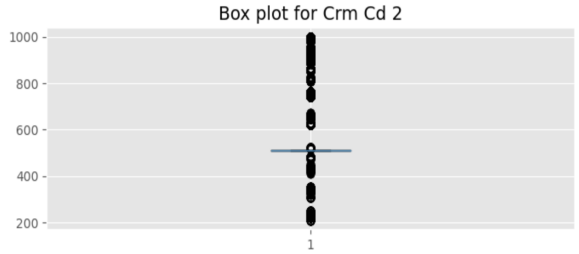
   Box: Represents the IQR, with the median marked by a line within the box.

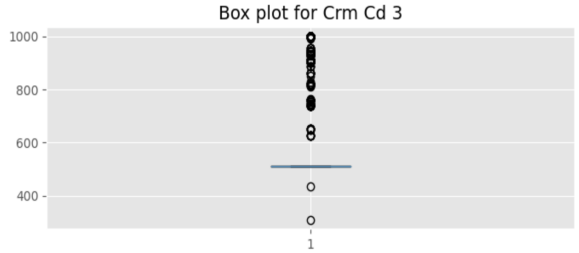
   Whiskers: Extend to the minimum and maximum values within a range determined by the IQR and can be used to identify potential outliers.

   Outliers: Individual data points beyond the whiskers, visualized as separate points.

```
In [63]: plt.figure(figsize=(8, 6))
         plt.subplot(2, 1, 1)
         plt.boxplot(df['Crm Cd 1'])
         plt.title('Box plot for Crm Cd 1')

         plt.figure(figsize=(8, 6))
         plt.subplot(2, 1, 1)
         plt.boxplot(df['Crm Cd 2'])
         plt.title('Box plot for Crm Cd 2')

         plt.figure(figsize=(8, 6))
         plt.subplot(2, 1, 1)
         plt.boxplot(df['Crm Cd 3'])
         plt.title('Box plot for Crm Cd 3')

         plt.figure(figsize=(8, 6))
         plt.subplot(2, 1, 1)
         plt.boxplot(df['Crm Cd 4'])
         plt.title('Box plot for Crm Cd 4')
Out[63]: Text(0.5, 1.0, 'Box plot for Crm Cd 4')
```







Conclusion:

1. Despite handling null and missing values, outliers and anomalies are found.
   The following are some typical causes of anomalies:
2. Human error in data entering: Errors in data entry can cause abnormalities.
3. Measurement errors: Outliers may be the consequence of inaccurate measurements or malfunctioning sensors.
4. Processing errors are mistakes that happen during the process of gathering or processing data.
5. Although managing missing values in columns is crucial for maintaining data cleanliness, it does not ensure that there are no abnormalities in the dataset.
6. The techniques I previously discussed are utilized to identify and display anomalies or outliers in the data, including z-score, IQR, and visualizations. Even when the values are absent.

Problem 9

This report presents correlations between demographic factors and specific types of crimes . The analysis aims to uncover any interesting patterns or trends that may shed light on the relationship between age, gender descent, and specific types of crimes.

Data Sources

The analysis utilized the following dataset:

Crime Data: Crime records from 2020 to the present, focusing on Gender, age, Descent, and Specific Types of crimes.

Data Cleaning:

Crime Data:

Relevant columns were selected, including 'Vict Age.',' Vict Sex',' Vict Descent' and 'Crm Cd Desc'

Data types were converted i.e. numerical values to appropriate numerical types.

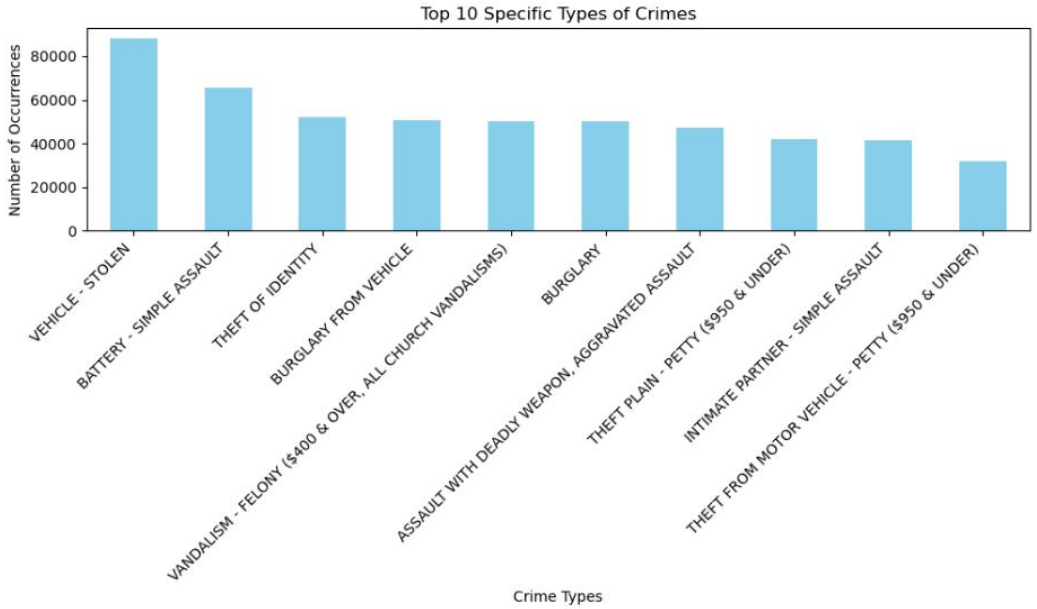
The top Ten most Specific types of Crimes were filtered out.

Five Top descents of victims were filtered out.

Data Analysis:

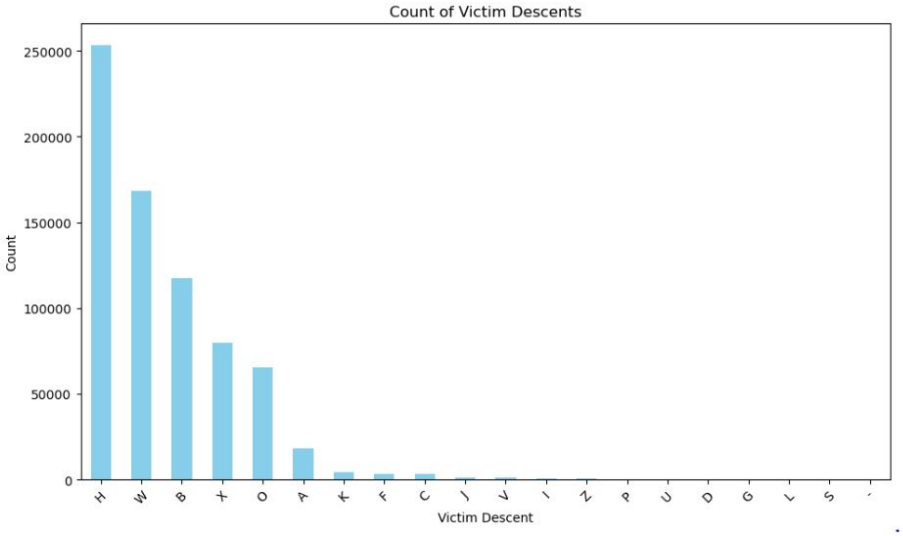
Exploration of Distinct Crime Type:

A list of distinct crime types was generated to identify the top ten specific types of crimes.



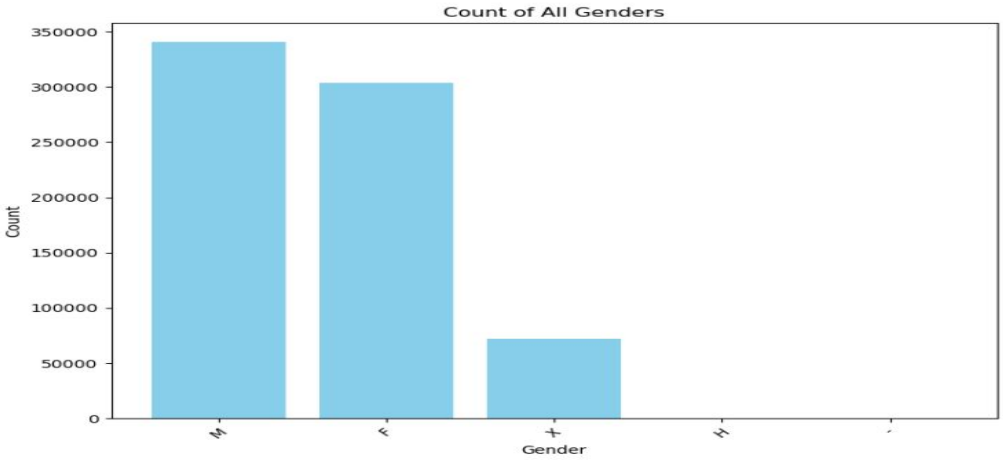
Exploration of Distinct Victim Descent:

A list of distinct victim descent was generated to identify the top ten descents with the most crimes.



Exploration of Distinct Victim Gender:

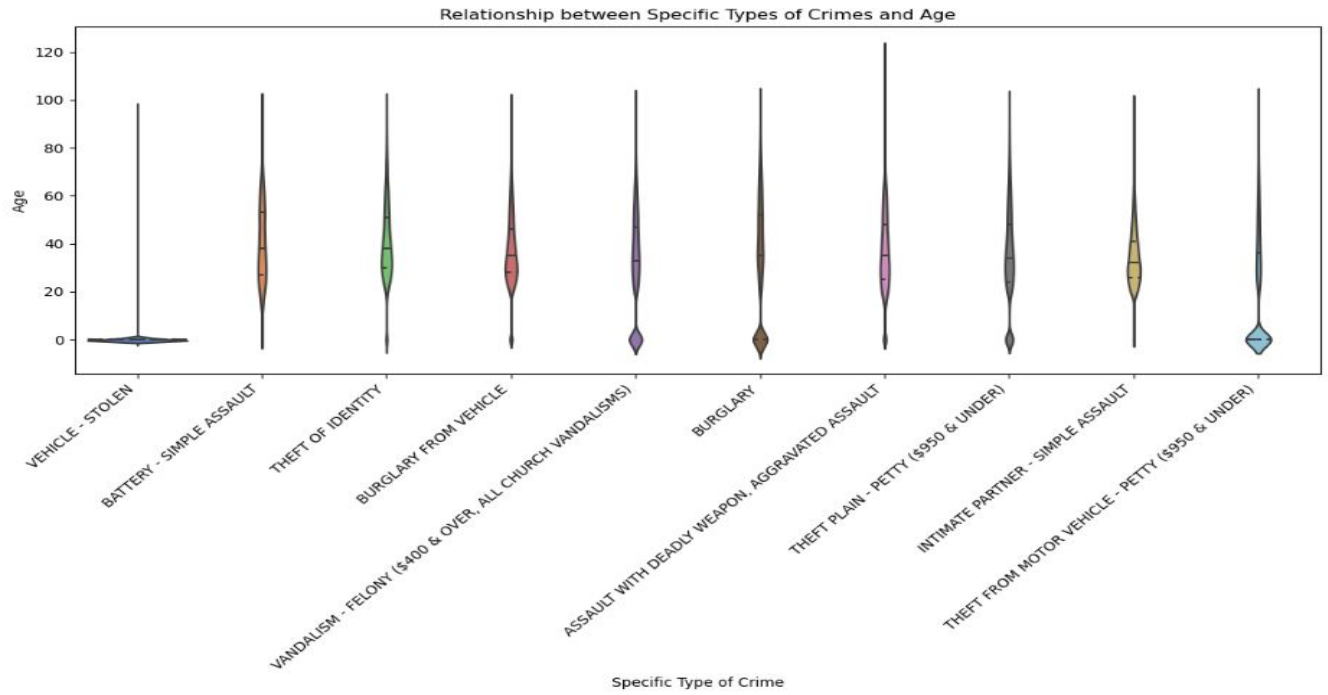
A list of distinct victim descent was generated to identify the genders of victims.



Relationship Analysis:

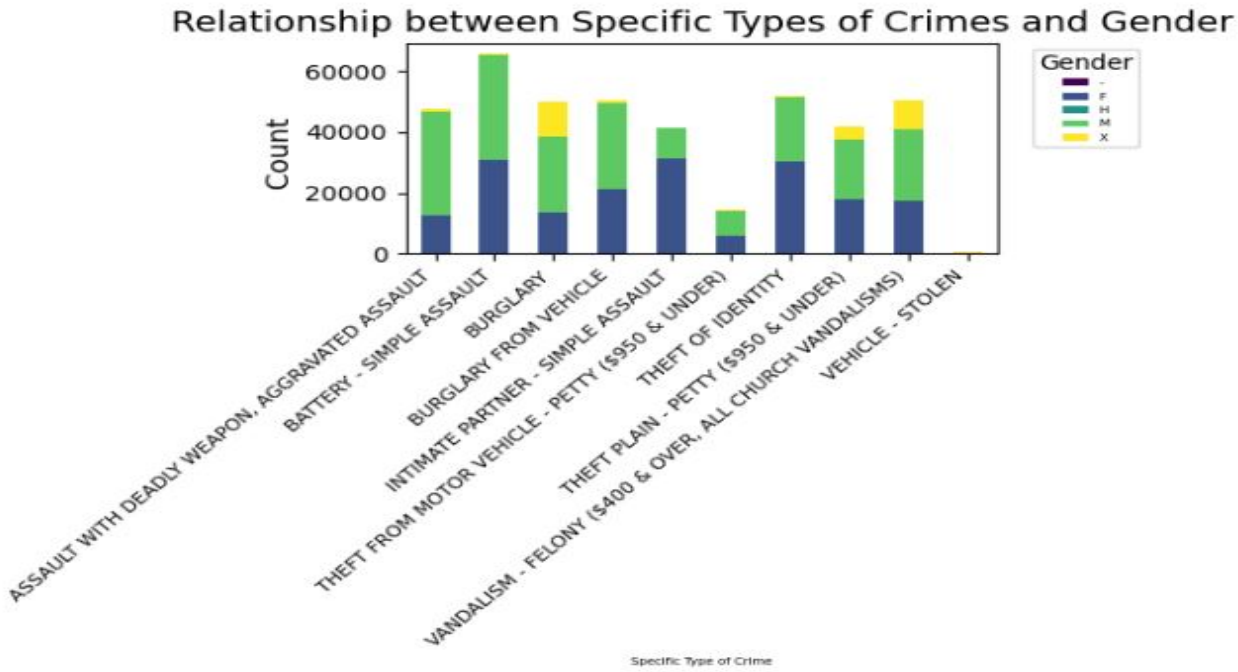
Relationship between Victim Age and Ten most specific types of crimes:

A violin plot was used to visualize the relationship between victim age and ten most specific types of crimes



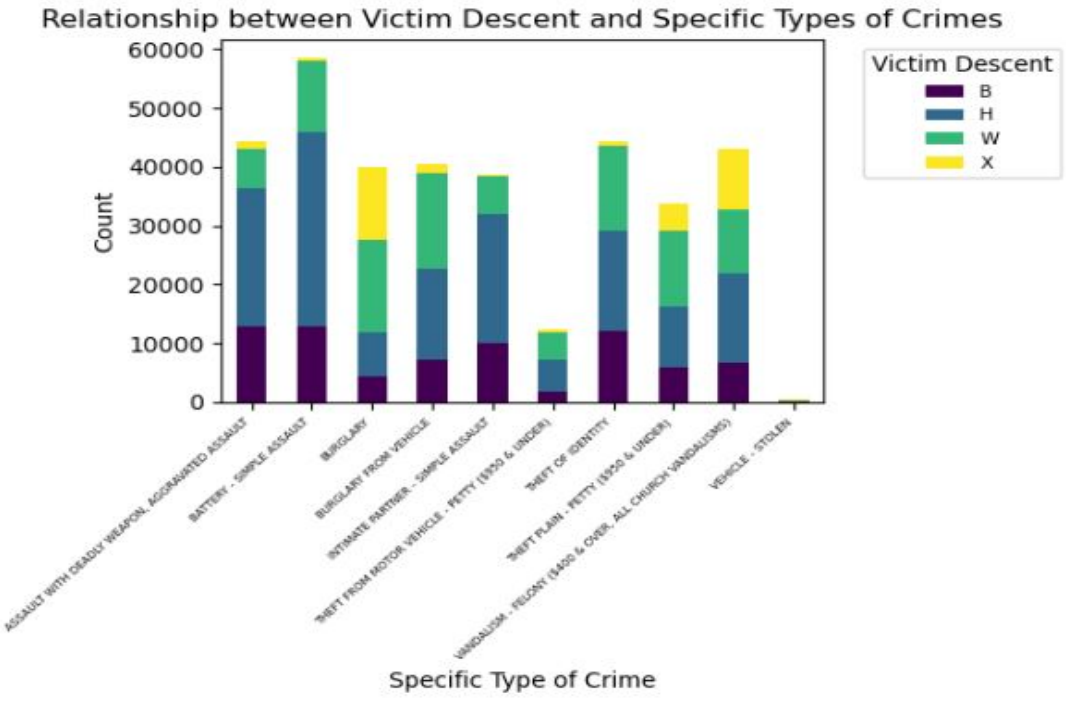
Relationship between Victim Gender and Ten most specific types of crimes:

Stacked bar chart was used to visualize the Relationship between Victim gender and Ten most specific types of crimes.



Relationship between Victim Gender and Ten most specific types of crimes:

A stacked bar chart was used to visualize the Relationship between Victim descent and the Ten most specific types of crimes.



Key Findings:

1. Relationship between Specific Types of Crimes and Age:

a. Crimes such as "Battery - Simple Assault" and "Theft of Identity" show a wide age range of perpetrators, with a noticeable concentration in the middle age groups.

b. "Vehicle - Stolen" crimes seem to be perpetrated mostly by a younger age group when compared to other types of crimes.

2. Relationship between Victim Descent and Specific Types of Crimes:

a. For crimes like "Burglary from Vehicle" and "Battery - Simple Assault", the victim descent "B" has the highest count, followed closely by "H".

b. "Vandalism - Felony ($400 & over, all church vandalism)" shows a near-equal distribution across "B", "H", and "W" descents.

3. Relationship between Specific Types of Crimes and Gender:

a. Males seem to be the dominant gender in most of the depicted crimes, with especially high counts in "Battery - Simple Assault", "Burglary", and "Theft from Motor Vehicle".

b. Females have a notable presence in "Intimate Partner - Simple Assault", though they are still outnumbered by males.

Conclusion:

Across all visualizations, it's evident that certain demographics, whether based on age, descent, or gender, have higher affiliations with particular types of crimes. This information can be crucial for law enforcement agencies and policymakers to tailor prevention and intervention strategies more effectively.

The data showcases discernible patterns across age, descent, and gender in relation to specific crimes, emphasizing the need for targeted intervention strategies based on demographic trends to effectively address and prevent criminal activities.

Problem 10

Employ time series forecasting methods, such as ARIMA or Prophet, to predict future crime trends based on historical data. Consider incorporating relevant factors into your models.

Data Sources

The analysis utilized the following dataset:

Crime Data: Crime records from 2020 to the present, focusing on crimes related to monetary gains.

Data Cleaning:

'Date Rptd' column was converted to DateTime

Crime count was calculated based on 'Crm Cd Desc'.

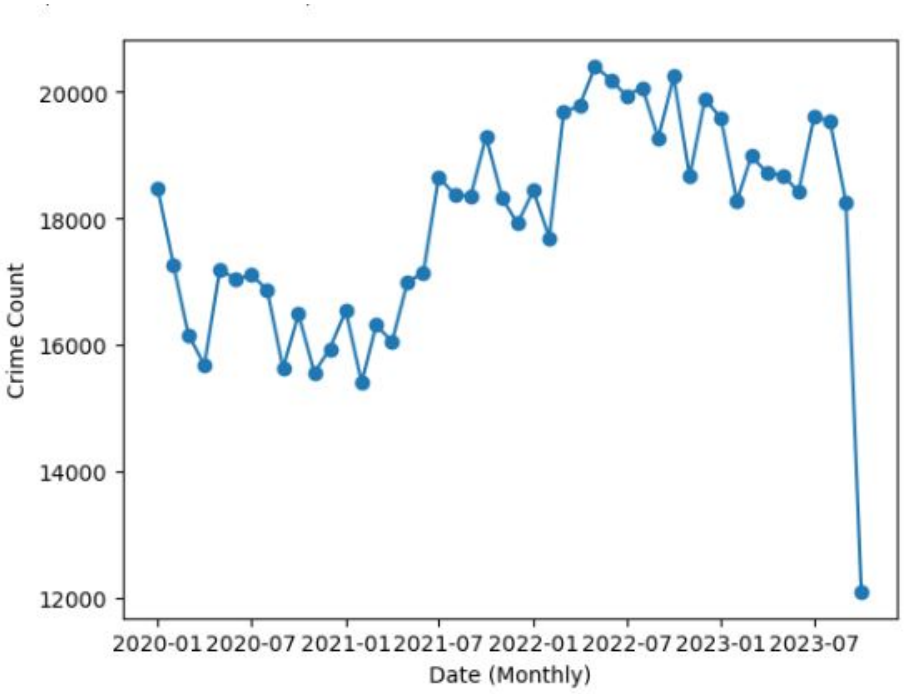
Data Analysis:

Historical Crime Data:

To understand how crime rates varied by each month of the year from 2020 to 2023, we conducted an analysis and created visualizations.

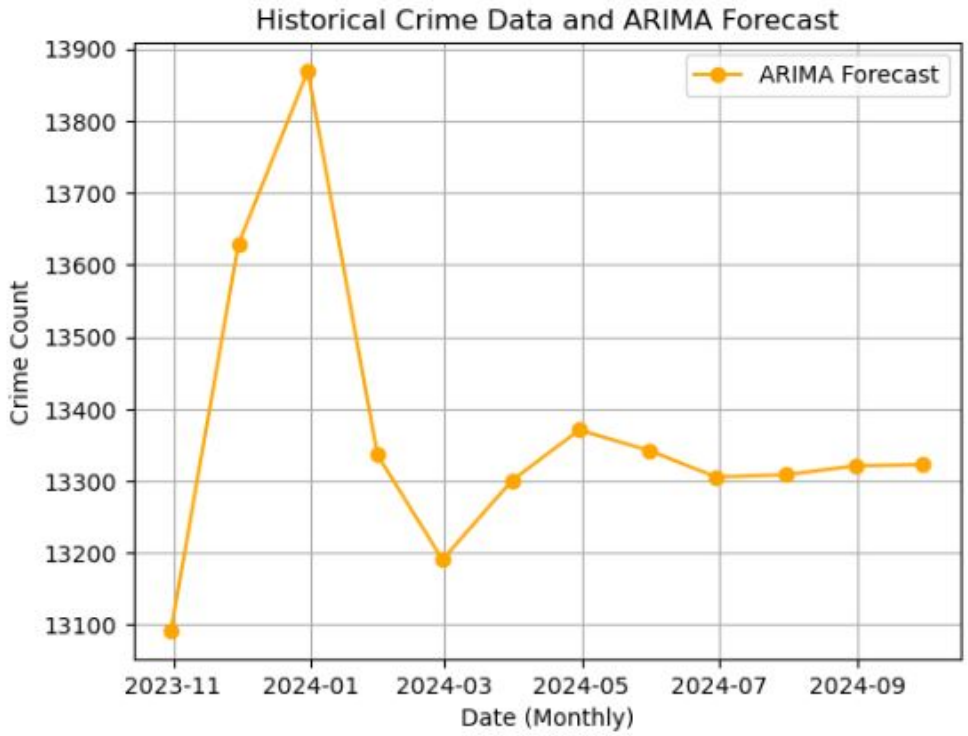
Monthly Trends Plot

The plot visually shows the crime rates by month for each year, allowing for a quick comparison and identification of patterns.



Prediction based on ARIMA Forecast:

This plot shows the forecast of crime counts predicted by ARIMA analysis for the future year.



Conclusion:

There are discernible variations in crime rates between 2020 and 2023 according to historical crime data. Peaks can be seen in early 2021 and mid-2022, with a notable decline in 2023. According to the ARIMA estimate, there should be little fluctuations in the crime rate between 2023 and late 2024, keeping it relatively steady. Based on historical trends, this estimate suggests that crime rates may stabilize in the following months. While external circumstances may still bring about unforeseen changes, policymakers and law enforcement organizations can use this prognosis to coordinate resources and plans.