Crime Statistics Analysis

IE6400 – Foundations Data Analytics

Project Report 1

Group Number 09

Vikas Nijaguna Kagawad (002294765)

Deepikasai Mettu (002289270)

Kuchibhotla Anirudha Raghava Sarma (002207582)

Sri Reshma Paramel (002821717)

Siddhant Pandit (002838340)

1. Introduction:

We are going to explore the fields of data analysis and public safety in this project. We are going to explore a real-world dataset that is full of crime statistics from 2020 to the present. Our main goal is to carefully clean and prepare this dataset so that it may be used as an invaluable tool for research and data analysis.

Our compass will direct us toward a diverse purpose as we traverse this dataset. First, to make sure that our dataset is devoid of flaws and abnormalities, we will start the crucial process of data purification. This includes dealing with data types, removing duplicates, controlling outliers, and handling missing numbers.

1.1 Methods

We are going to venture out into the realm of exploratory data analysis (EDA) with our cleansed dataset in hand. Here, we'll lay out a plan for identifying patterns and trends in crime that have developed over time. We will look at the temporal dimension, looking at how crime rates have changed over time and looking for trends in monthly, seasonal, and annual rates. We will also investigate the spatial component, utilizing maps and geographic visualizations to decipher the spatial distribution of crimes.

But our journey doesn't end with identifying trends and patterns. We'll also investigate the underlying causes of crime rates. Our work will culminate in practical insights and conclusions that can guide policy decisions in the fields of law enforcement, community development, and government. By the time this project is through, we expect to have solved the enigmas surrounding crime statistics and shed light on the variables influencing the security and safety of our society.

To create a safer and more informed future, let's set out on this journey into the world of crime data, cleaning, exploring, and analyzing along the way.

The journey generally starts with data inspection in the field of data analysis. This critical stage lays the groundwork for comprehending and turning unprocessed input into insightful knowledge. We delve into the essential steps of data inspection and cleaning, which are the foundation of any data analysis project.

As we proceed with data cleaning, we need to take care of a few things to get the data ready for analysis. Finding and effectively handling missing data is one of the initial jobs. Making the decision of whether to impute the missing values or, in certain situations, eliminate the associated rows or columns is crucial since missing data can skew our results. Redundant rows have the potential to distort our analysis and add redundancy. Consequently, one of the most important phases in data cleansing is locating and eliminating duplicate entries.

Another part of data cleansing is data type conversion. For an analysis to be valid, dates must be formatted correctly, and numerical quantities must be represented effectively. Outliers can affect the results of statistical analysis if they exist. As a result, we must choose how to deal with outliers—whether it is by eliminating them or using transformational tactics. To bring numerical data to a uniform scale, standardization or normalization may be necessary, particularly when utilizing specific machine learning techniques. Lastly, many analytical procedures depend on the encoding of categorical data into numerical values.

1.2 Aim

To sum up, the process of inspecting and cleansing data paves the way for effective data analysis. These crucial actions guarantee that the dataset is prepared to produce insightful findings, revealing the undiscovered treasures within the unprocessed data. With the availability of large-scale statistics, such crime data from 2020 to the present, we have a rare chance to comprehend and tackle some of the most important problems affecting our society. This essay delves deeply into the importance and possibilities of using such a dataset, highlighting the opportunity to get priceless insights and strengthen communities.

This dataset, which includes crime statistics from 2020 to the present, is a storehouse of information that has the potential to have a substantial impact on community well-being and public safety. It is more than just a collection of figures and facts.

Understanding crime trends over time is one of the main purposes of using this dataset. We can obtain a thorough understanding of the changes in crime rates over time by closely examining the statistics, whether it is on a monthly, seasonal, or annual basis. With the use of these insights, we can locate prospective issue spots that need to be addressed and regions where crime is on the rise. On the other hand, by recognizing the effects of effective laws and tactics, we can also recognize regions where crime rates have dropped. Furthermore, the dataset gives us the ability to see beyond the specifics and consider the wider picture.

2.Data cleaning

Data cleaning is a critical foundational step in our project, focusing on analyzing crime data from 2020 to the present. Ensuring that our dataset is clean and well-structured is paramount to obtaining accurate and reliable insights. This process involves several key aspects. First, we address missing data points and outliers, which can skew our analysis. By carefully handling missing values through imputation or removal, we aim to maintain the integrity of our dataset. Simultaneously, identifying and managing outliers is essential, as they can distort the overall trends and patterns. Furthermore, data cleaning involves standardizing date formats, dealing with redundant or irrelevant columns, and unifying categories to enhance the dataset's consistency and usability. By performing these cleaning tasks, we lay the groundwork for more effective exploratory data analysis and insightful conclusions regarding crime trends and influencing factors.

3.Exploratory Data Analysis

In this section, we'll explore the trend in crime data from the dataset, specifically looking at the count of crimes over time. We'll create a time series visualization to better understand how crime rates have evolved. The code takes the original dataset and groups it by date, counting the number of crimes for each date. It then converts the 'Date' column to a datetime format and sets it as the index. This allows us to work with the data

EDA

In [529]:
```
# 1. Crime Trend
crime_count = crime_df.groupby(crime_df.index.date).size().reset_index()
```

In [530]:
```
crime_count.rename(columns={'index':'Date',0:'crime_count'}, inplace=True)
crime_count['Date']= pd.to_datetime(crime_count['Date'],format='%Y-%m-%d')
crime_count.set_index('Date',inplace=True)
```

In [531]:
```
crime_count.head()
```

Out[531]:

Date	crime_count
2020-01-01	1105
2020-01-02	526
2020-01-03	595
2020-01-04	540
2020-01-05	505

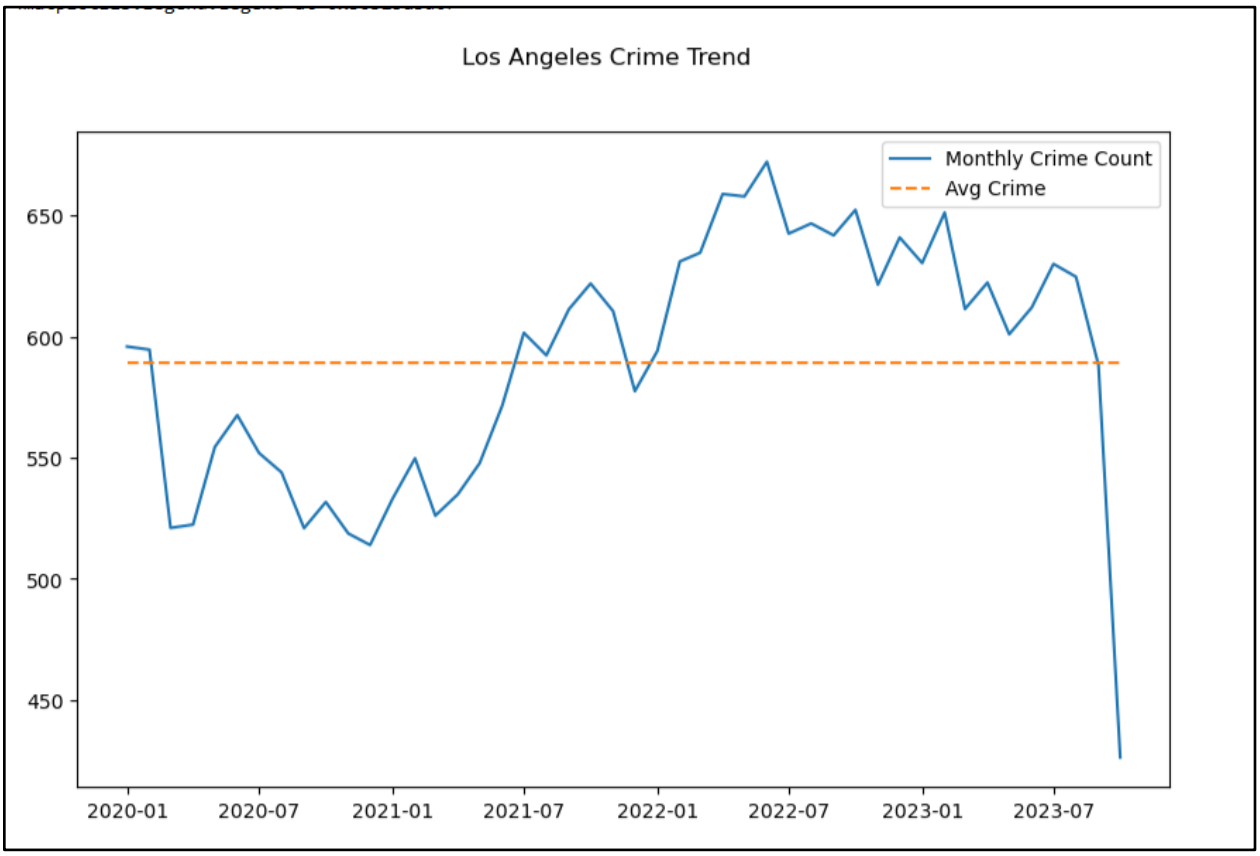
In [532]:
```
monthly_crime_count = crime_count.resample('MS').mean()
```

In [533]:
```
monthly_crime_count.head()
```

Out[533]:

Date	crime_count
2020-01-01	595.903226
2020-02-01	594.620690
2020-03-01	521.000000
2020-04-01	522.366667
2020-05-01	554.419355

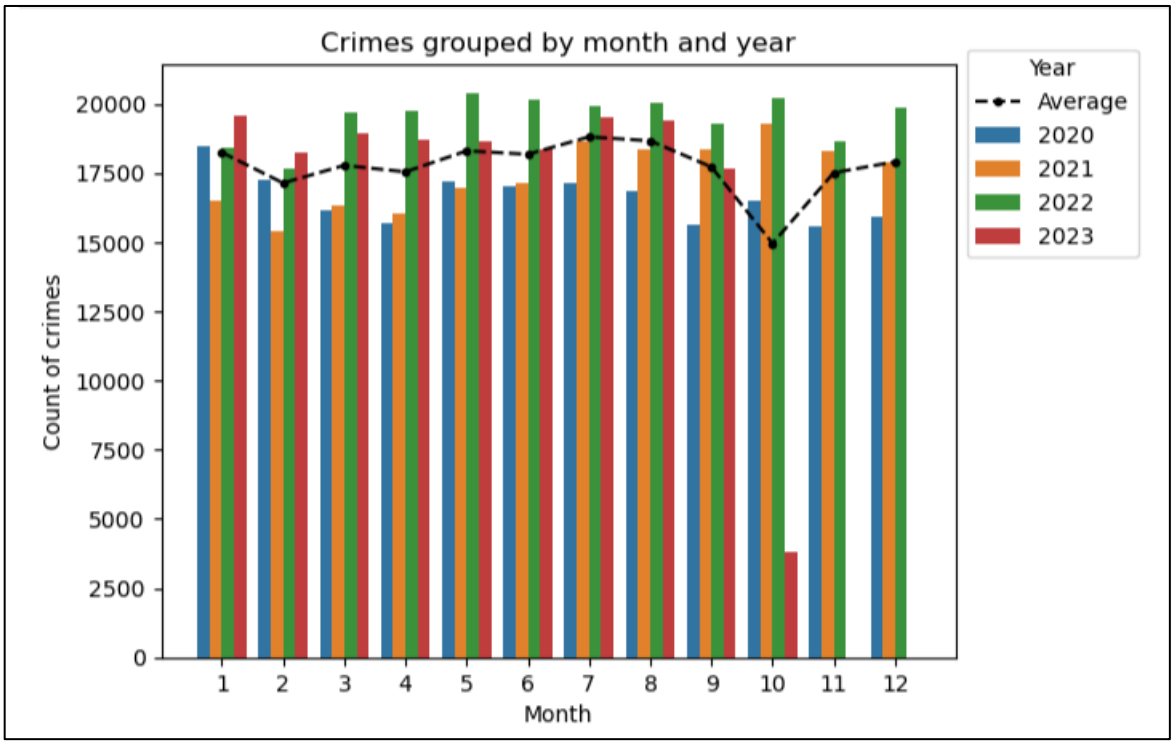
over time. We'll dive deeper into the crime data by analyzing monthly trends. We'll calculate the average monthly crime count and create a plot to visualize the variations.



The spike in crime rates in 2021 to 2023 can be attributed to various factors, and one significant contributor is the impact of the COVID-19 pandemic on the mental health and well-being of the population. According to LAPD (Los Angeles Police Department), the trauma of the past three years and the associated mental health crisis have played a role in the increase in crime rates.

	month occured	average crimes per month
0	1	18239.500000
1	2	17135.250000
2	3	17771.750000
3	4	17538.750000
4	5	18298.000000
5	6	18174.000000
6	7	18802.250000
7	8	18659.500000
8	9	17717.250000
9	10	14955.500000
10	11	17507.000000
11	12	17901.333333

We delve into the dataset to identify and analyze any potential seasonal patterns in crime occurrences. By aggregating the data based on the year and month of each crime, we can gain insights into how crime rates fluctuate over the months.

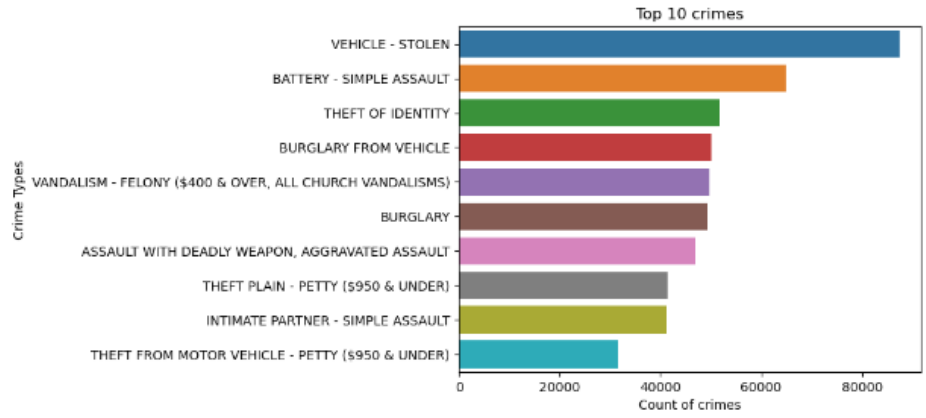


The graph above depicts the count of crimes by month, categorized by year. A black dashed line represents the average number of crimes per month.

Observing the data from 2020 to the present (2023), we can see that crime rates exhibit some fluctuations. The summer months, particularly July, tend to have higher crime rates, while October appears to have the lowest crime rates. These observations can provide insights into the seasonality of crime in the dataset, which can be further explored and analyzed to understand the underlying factors driving these patterns.

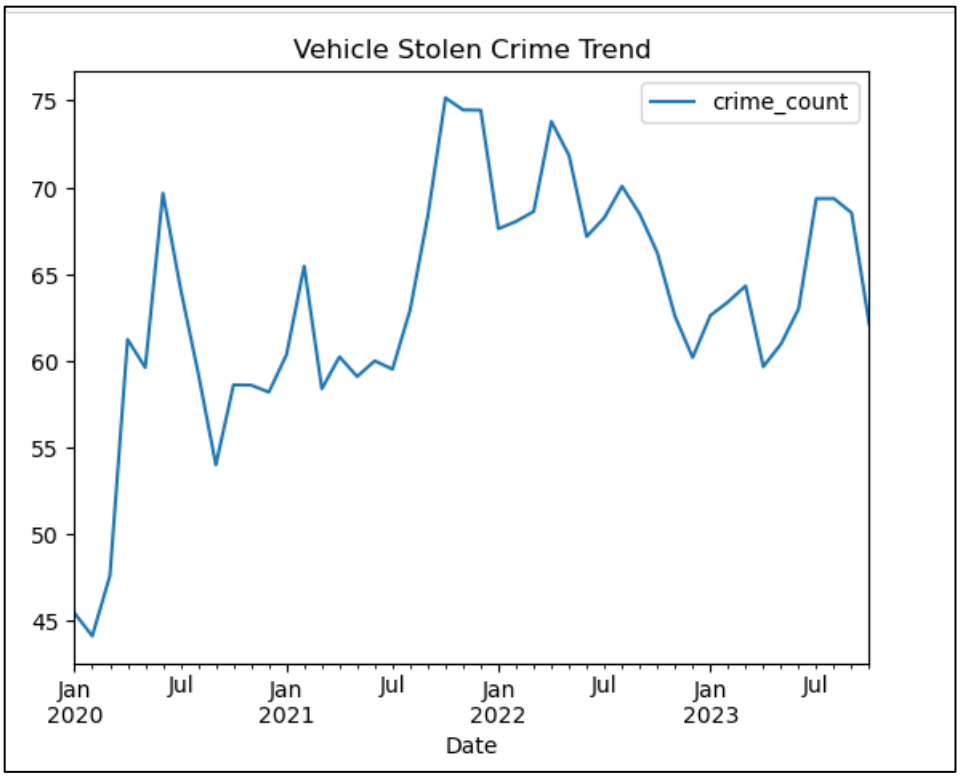
```
Common crime type
(crime_df[['Crm Cd Desc','Crm Cd']].value_counts())
rimes=crime_df['Crm Cd Desc'].value_counts().head(10)
ountplot(data=crime_df[crime_df['Crm Cd Desc'].isin(top_crimes.index)], y='Crm C
label("Count of crimes")
label("Crime Types")
itle("Top 10 crimes")
("\033[1mMost common crime type observed is:  " , top_crimes.head(1)," \033[0m")
```

```
Crm Cd Desc                                          Crm Cd
VEHICLE - STOLEN                                     510       87354
BATTERY - SIMPLE ASSAULT                             624       64929
THEFT OF IDENTITY                                    354       51723
BURGLARY FROM VEHICLE                                330       49995
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)  740   49748
                                                               ...
GRAND THEFT / AUTO REPAIR                            349           5
FIREARMS RESTRAINING ORDER (FIREARMS RO)             906           4
FAILURE TO DISPERSE                                  884           3
DISHONEST EMPLOYEE ATTEMPTED THEFT                   445           2
INCITING A RIOT                                      882           1
Length: 138, dtype: int64
Most common crime type observed is:   VEHICLE - STOLEN    87354
Name: Crm Cd Desc, dtype: int64
```



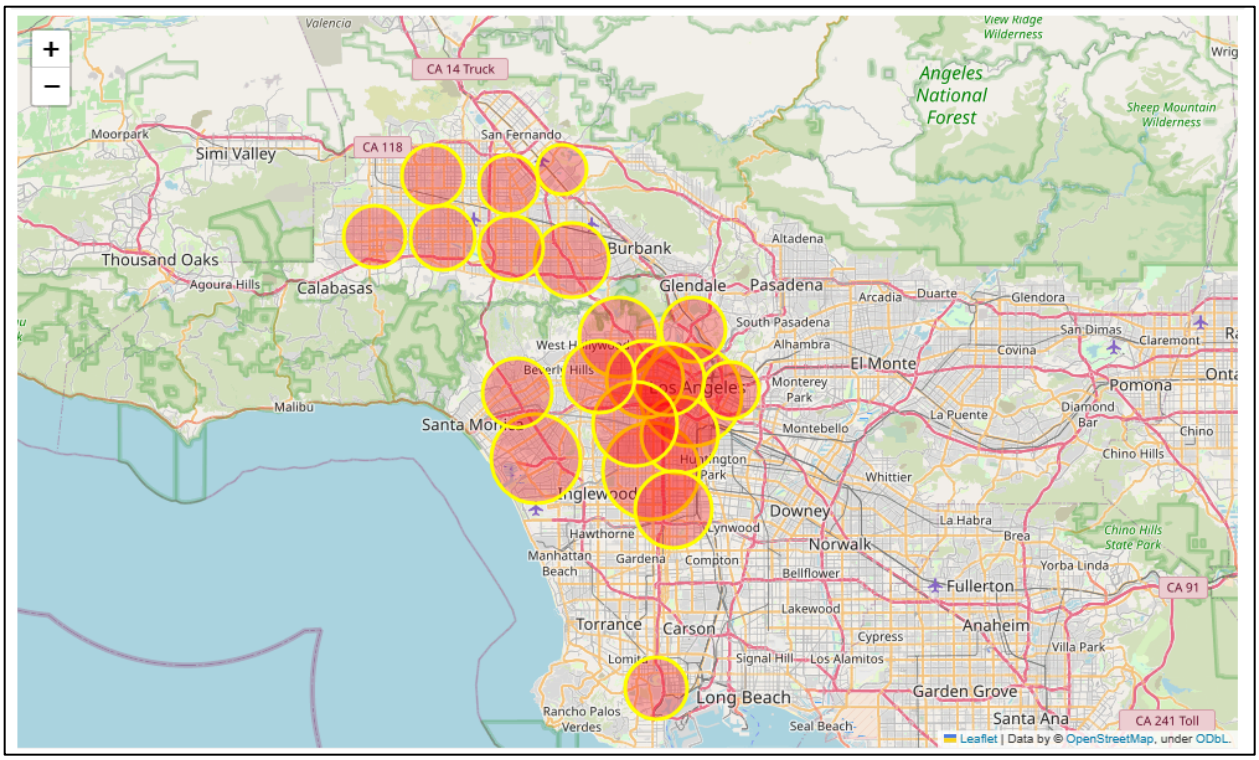
Prevalent Crime Types: The dataset indicates that "VEHICLE - STOLEN" is the most common crime type, with a staggering count of 87,354 cases. "BATTERY - SIMPLE ASSAULT" and "THEFT OF IDENTITY" are also highly frequent, with counts of 64,929 and 51,723, respectively.

Crime Diversity: While some crime types dominate the statistics, there is a wide range of less common crime categories present in the dataset. These less common crimes collectively represent a diverse array of criminal activities.

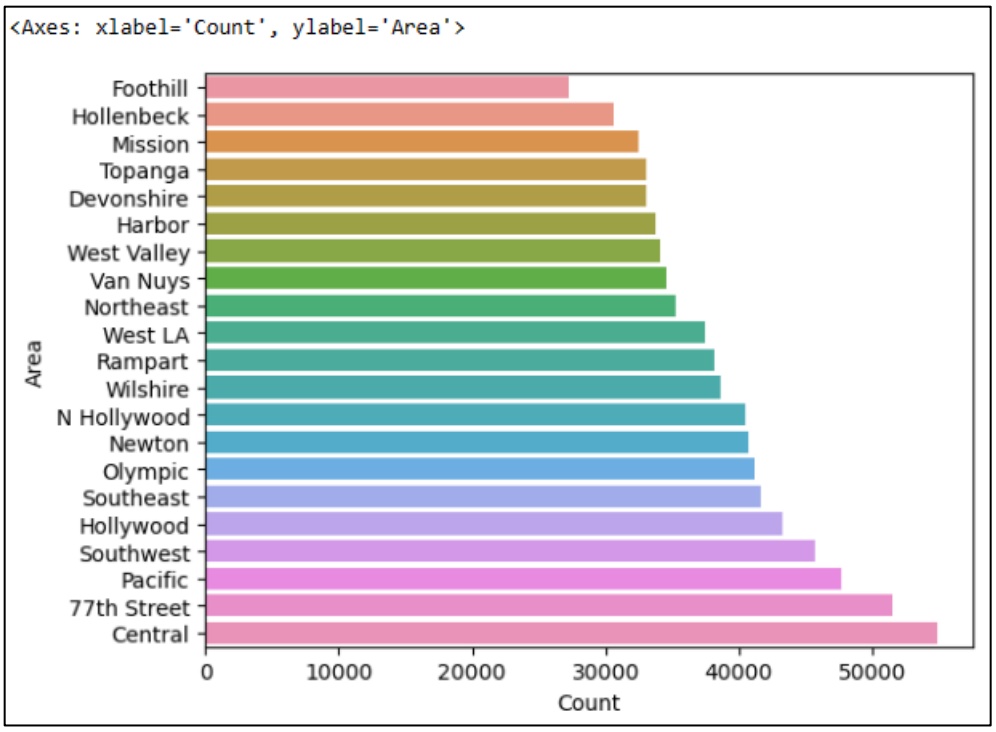


Yearly Variation: The trend in vehicle theft shows fluctuations over the years. While the exact numbers may vary, there is a clear recurring pattern in vehicle theft occurrences.

Seasonal Patterns: The data suggests that vehicle theft exhibit seasonal patterns. As we can see spikes during July of every year. These patterns could be influenced by factors such as weather, holidays, and economic conditions.



Geographical Distribution of Crime in Regions: We have created a map that shows the geographical distribution of crime occurrences in various regions. Each circle marker represents an area, with the circle's size proportional to the number of reported crimes in that area. The map provides a visual overview of where crimes are concentrated within the region.



Bar Chart: The bar chart displays the count of reported crimes for each area, ordered in ascending order. It allows for a quick comparison of crime rates across different regions. The highest bars indicate areas with the highest crime counts, while the lowest bars represent areas with fewer reported crimes.

We explore the relationship between crime rates and various economic factors, including income levels, emigration rates, homelessness, and unemployment rates. The objective is to understand whether there is a correlation between these economic variables and crime rates over the years.

```
# 5. Correlation with economic factors
yearly_economic_data_la = pd.DataFrame({'year1':['2020','2021','2022','2023'],
                                        'mean_income':[112834,106931,120272,129000],
                                        'median_income':[76367,69778,82516,98200],
                                        'emmigration':[20259,40786,33235,30200],
                                        'homeless':[66436,80436,81235,75518],
                                        'unemployement_rate':[6.1,6,4.2,5.1],
                                        'inflation':[1.2,3.8,7.4,3.8]})
yearly_economic_data_la
```

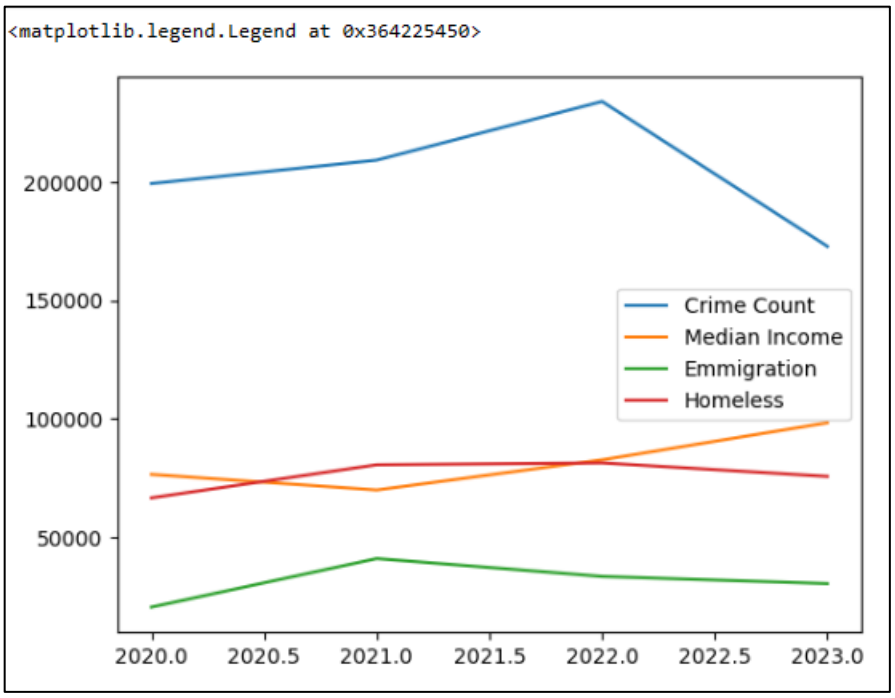
	year1	mean_income	median_income	emmigration	homeless	unemployement_rate	inflation
0	2020	112834	76367	20259	66436	6.1	1.2
1	2021	106931	69778	40786	80436	6.0	3.8
2	2022	120272	82516	33235	81235	4.2	7.4
3	2023	129000	98200	30200	75518	5.1	3.8

```
yearly_crime = crime_count.resample('Y').sum().reset_index()
yearly_crime.columns = ['year','crime_count']
yearly_crime['year']=yearly_crime['year'].dt.year
yearly_crime
```

	year	crime_count
0	2020	199328
1	2021	209240
2	2022	234049
3	2023	172775

```
economic_crime_data = pd.concat([yearly_crime,yearly_economic_data_la],axis = 1)
economic_crime_data.drop('year1',axis=1,inplace=True)
economic_crime_data['crime_rate']= (economic_crime_data['crime_count']/economic_crime_data['crime_count'].sum())*10
economic_crime_data
```

	year	crime_count	mean_income	median_income	emmigration	homeless	unemployement_rate	inflation	crime_rate
0	2020	199328	112834	76367	20259	66436	6.1	1.2	2.444567
1	2021	209240	106931	69778	40786	80436	6.0	3.8	2.566128
2	2022	234049	120272	82516	33235	81235	4.2	7.4	2.870386
3	2023	172775	129000	98200	30200	75518	5.1	3.8	2.118919

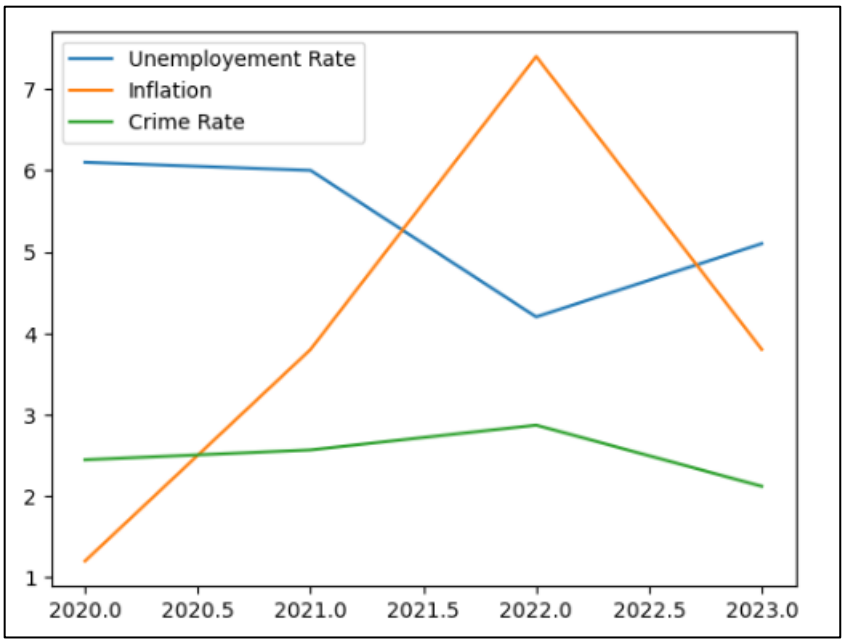


The lines represent different variables, including crime count, median income, emigration, and homelessness. From the graph, we can observe the following insights:

Median Income and Crime: There appears to be a negative correlation between median income and crime rates. As median income increases, crime rates tend to decrease, and vice versa.

Emigration and Crime: The emigration rate does not show a clear correlation with crime rates.

Homelessness and Crime: There may be a positive correlation between homelessness and crime rates, suggesting that areas with higher homeless populations may experience higher crime rates.



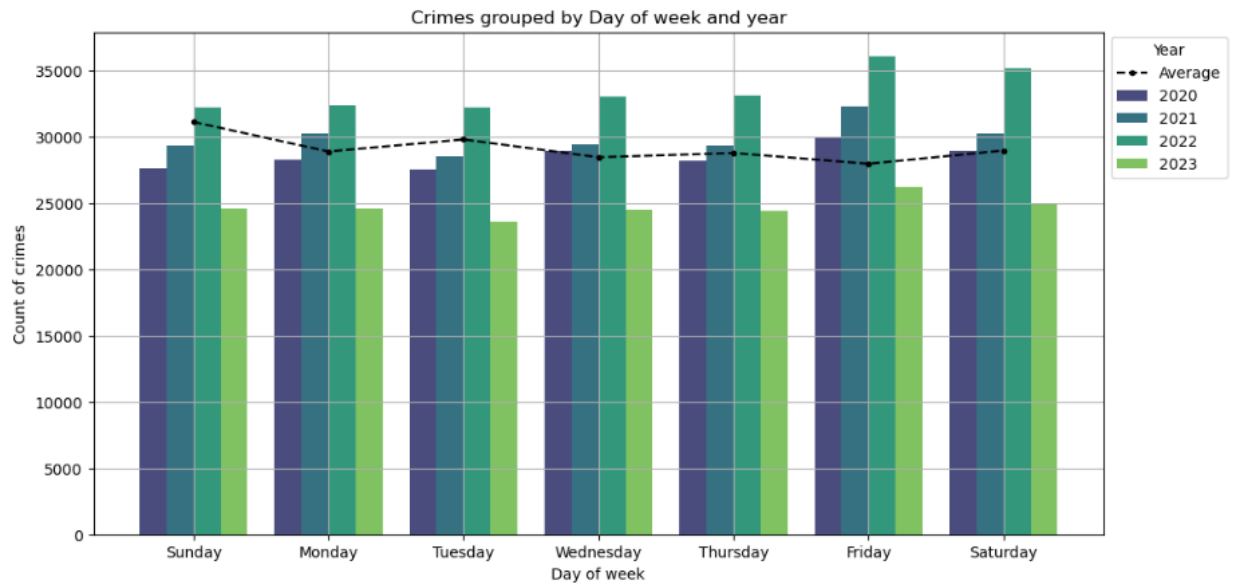
Unemployment Rate and Crime: There doesn't appear to be a clear linear correlation between the unemployment rate and crime rates. The relationship between these two variables may be more complex, and other factors could influence crime rates.

Inflation and Crime: Inflation and crime rates do not seem to exhibit a strong linear correlation. However, it's worth considering that the impact of inflation on crime can be multifaceted, and further analysis may be needed to understand any potential relationships.

```
# 6.Day of week:
dup = crime_df.reset_index()
dup['Day of Week'] = dup['DATE OCC'].dt.day_name()
crimes_DOW = dup.groupby(['year occured','Day of Week']).size().groupby('Day of Week').mean().reset_index()
crimes_DOW.columns=['Day','Average crimes per day of week']
crimes_DOW
```

	Day	Average crimes per day of week
0	Friday	31100.75
1	Monday	28873.25
2	Saturday	29780.75
3	Sunday	28438.75
4	Thursday	28760.75
5	Tuesday	27947.25
6	Wednesday	28946.50

```
plt.figure(figsize=(12, 6))
sns.countplot(data=dup, x='Day of Week', hue='year occured', palette='viridis',
              order=['Sunday','Monday','Tuesday','Wednesday','Thursday','Friday','Saturday'])
y=crimes_DOW['Average crimes per day of week']
plt.plot(y, marker='.', color='black', linestyle='--', label='Average')
plt.legend(title='Year', loc='center left', bbox_to_anchor=(1, 0.85))
plt.grid()
plt.xlabel("Day of week")
plt.ylabel("Count of crimes")
plt.title("Crimes grouped by Day of week and year")
plt.show()
```



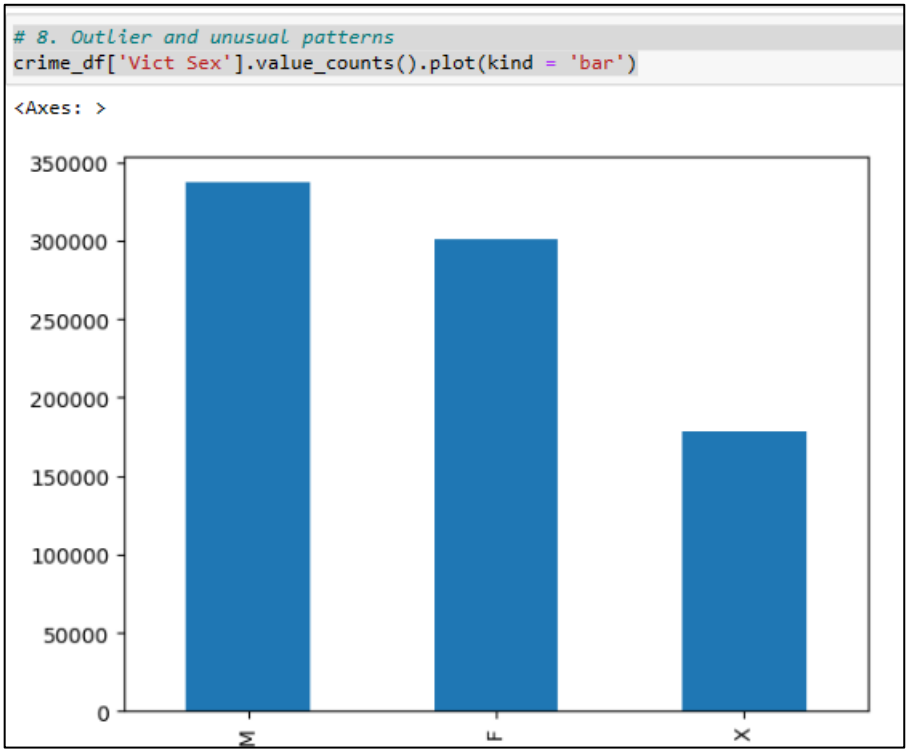
The graph visually represents the count of crimes by day of the week, categorized by the year. The dashed black line represents the average number of crimes per day of the week.

Key insights from the analysis include:

Consistent Patterns: Throughout the years, the patterns of crime occurrences by day of the week appear relatively consistent. There are slight variations, but no specific day stands out as significantly different from others.

Weekday Peak: Crime rates tend to be slightly higher on weekdays (Monday to Friday) compared to weekends (Saturday and Sunday).

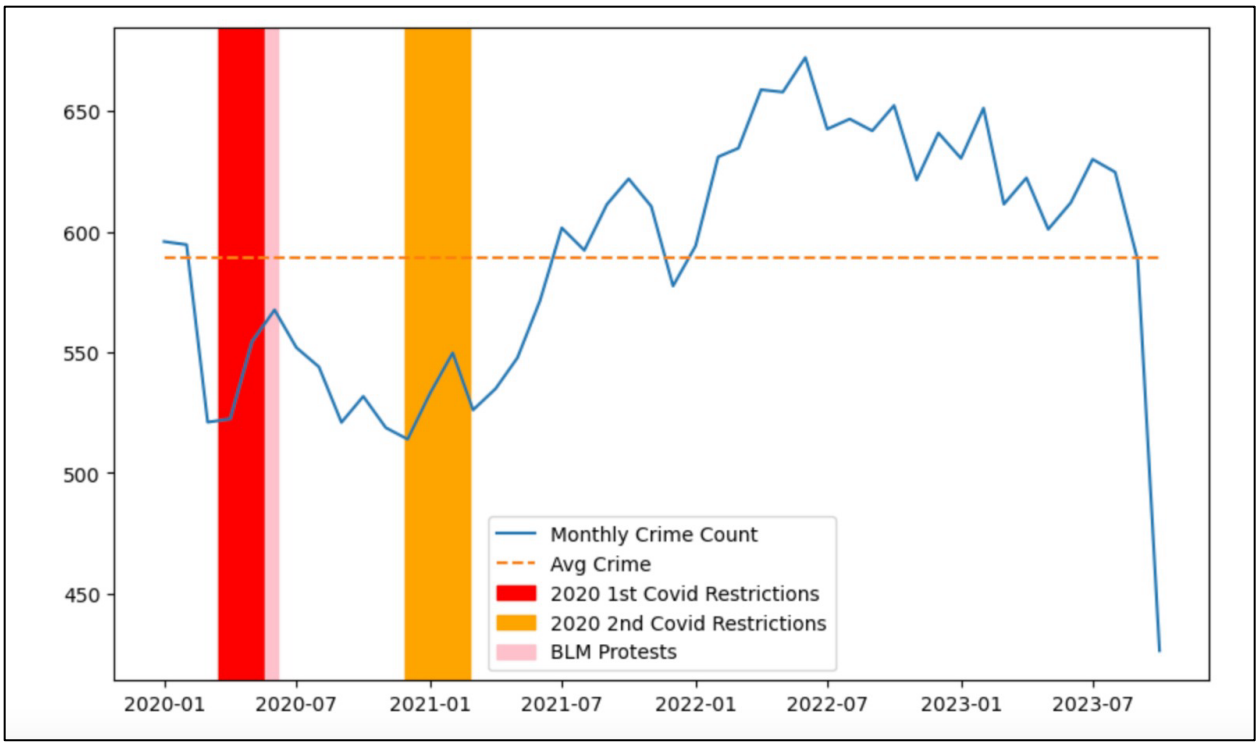
Friday: Friday appears to have the highest average crime count among weekdays.



Key Observations:

Male Victims: The bar chart shows that the highest number of victims in reported crimes are male. This suggests that males are more frequently victims of these crimes in the dataset.

Female Victims: Following male victims, the second most common category is female victims. While the count of female victims is lower than that of males, it still represents a substantial portion of the reported crimes.

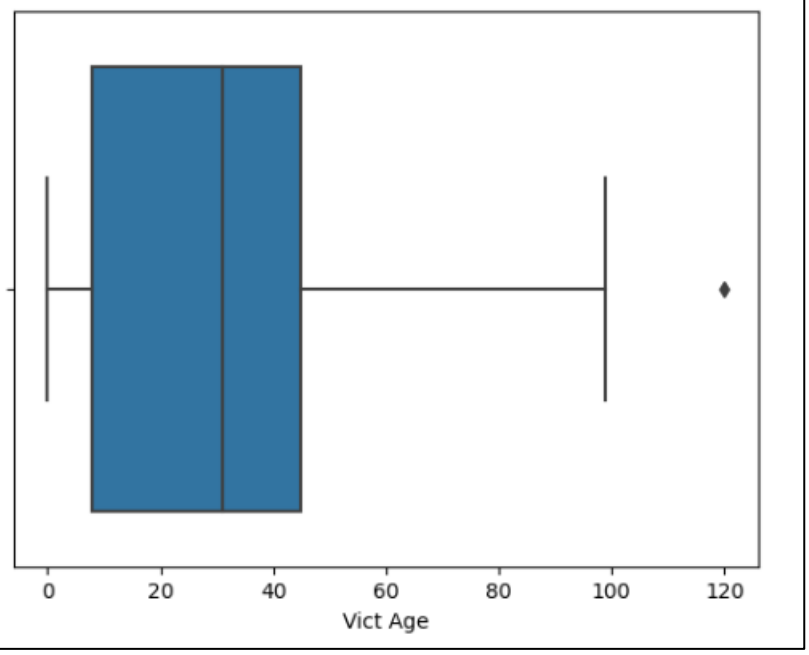


The preceding analysis of crime incidence in the context of specific policy changes reveals noteworthy patterns. During the initial implementation of COVID-19 restrictions and the emergence of Black Lives Matter (BLM) protests in 2020, it is evident that the monthly crime count remained consistently below the typical average crime rate. This trend persisted even in the face of temporary spikes in certain periods.

Conversely, the implementation of the second set of COVID-19 restrictions in 2020 also resulted in monthly crime counts that remained below the historical average. However, in the latter half of 2021, a discernible upswing in crime incidence became apparent, with monthly counts surpassing the established average crime rate. This surge in crime incidence persisted into the subsequent years of 2022 and 2023, consistently exceeding the historical average crime rate.

```
sns.boxplot(data=crime_df, x='Vict Age')
```

```
<Axes: xlabel='Vict Age'>
```

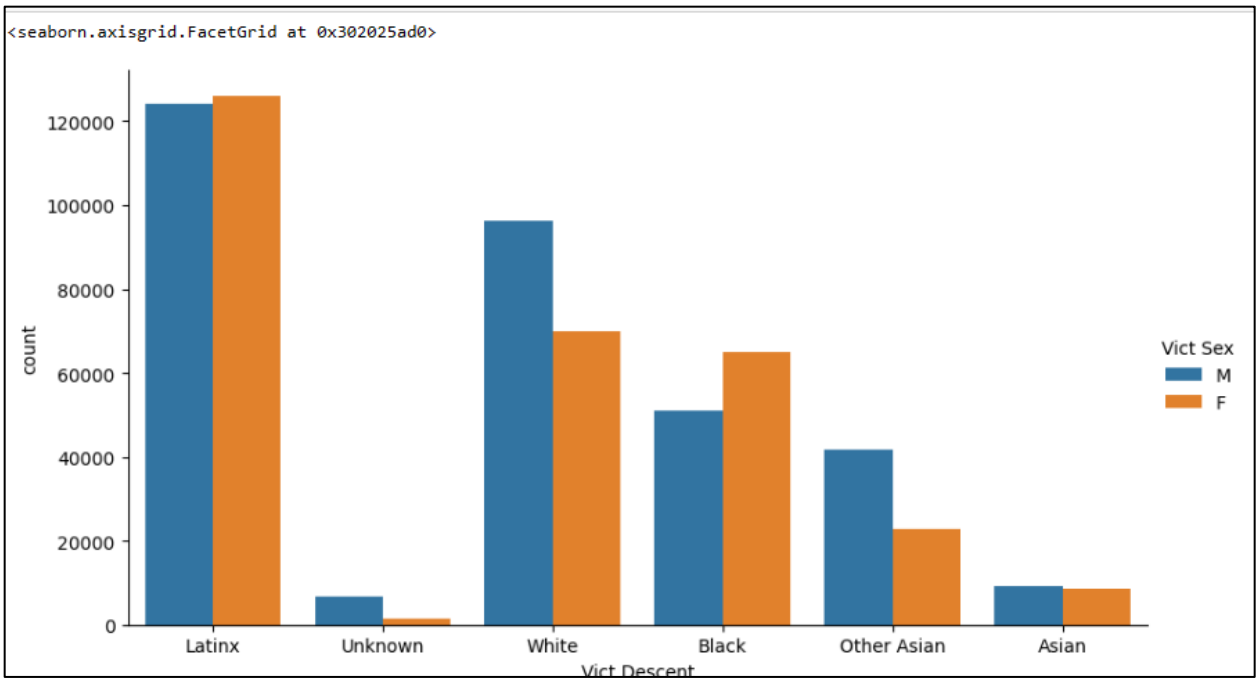


Key Observations:

Age Range: The boxplot provides a clear representation of the age distribution of crime victims. The majority of victims fall within the age group of 10 to 50 years, as indicated by the interquartile range (the box in the boxplot).

Central Tendencies: The boxplot shows the median age as the line within the box, which represents the 50th percentile of victim ages. Additionally, the box provides information about the first and third quartiles, helping to understand the central tendencies of the age distribution.

Outliers: The boxplot also highlights any potential outliers, which are data points that fall significantly outside the main body of the distribution. Outliers may represent unusual or extreme cases. They are typically depicted as individual data points beyond the "whiskers" of the boxplot.



we delve into the analysis of victim demographics, specifically focusing on the victim's descent (ethnicity or race) and gender. The catplot is used to visualize the count of victims based on their descent and gender, with specific emphasis on the six most prevalent victim descents.

Key Observations:

Latinx Victims: Among the analyzed victim descents, Latinx victims are the most frequent, with both male and female victims exceeding 120,000. This suggests that Latinx individuals are disproportionately represented among the victims in the dataset.

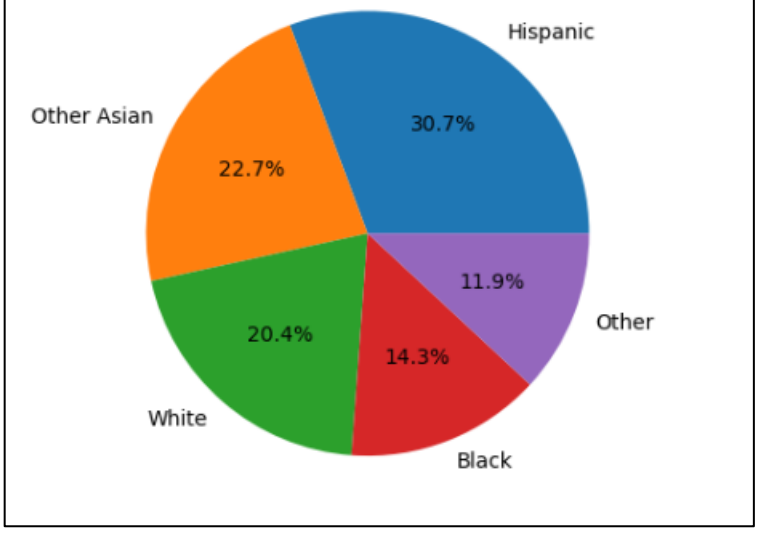
Unknown Descent: Victims with an "Unknown" descent have a relatively low count, with fewer than 20,000 cases. This may be due to limited or missing information about their descent.

White Victims: White victims are among the top three most common descents, with counts similar to those of Latinx victims. This indicates a significant presence of White victims in the dataset.

Black Victims: Black victims, both male and female, also have substantial counts, with males slightly exceeding 50,000 and females reaching nearly 60,000.

Other Asian and Asian Victims: Both "Other Asian" and "Asian" victims have relatively low counts, with both males and females having fewer than 20,000 cases. This suggests that crimes involving individuals from these descents are less frequent in the dataset.

```
plt.pie(finalised_categories,labels=lab, autopct='%1.1f%%')
plt.show()
```



Key Observations:

The pie chart reveals the following distribution of victim descents:

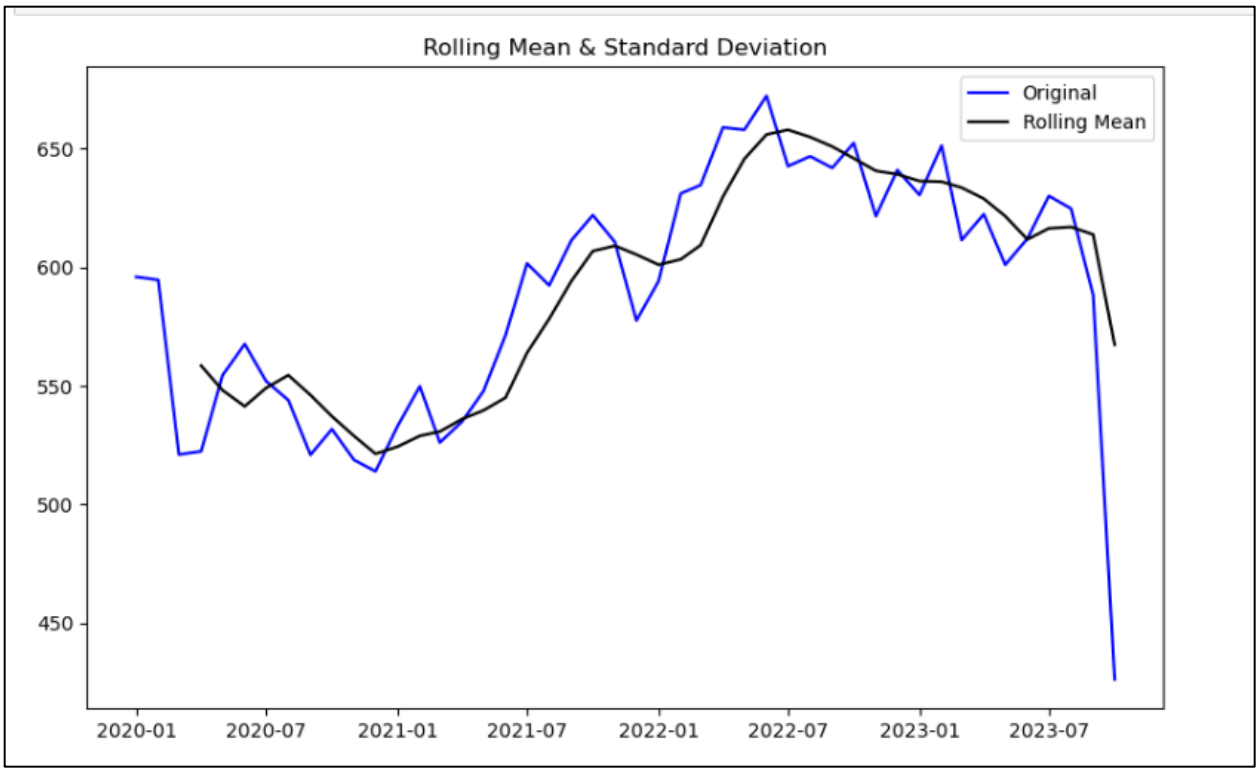
Hispanic: Hispanic victims account for the largest portion, making up approximately 30.7% of the total victim count. This suggests a significant presence of Hispanic victims in the dataset.

White: White victims represent approximately 20.4% of the total victim count. While their proportion is substantial, it is notably lower than that of Hispanic victims.

Black: Black victims make up around 14.3% of the total victim count, indicating their presence in the dataset.

Other Asian: Victims with an "Other Asian" descent account for approximately 22.7% of the total victim count. This suggests a notable representation of individuals from this group.

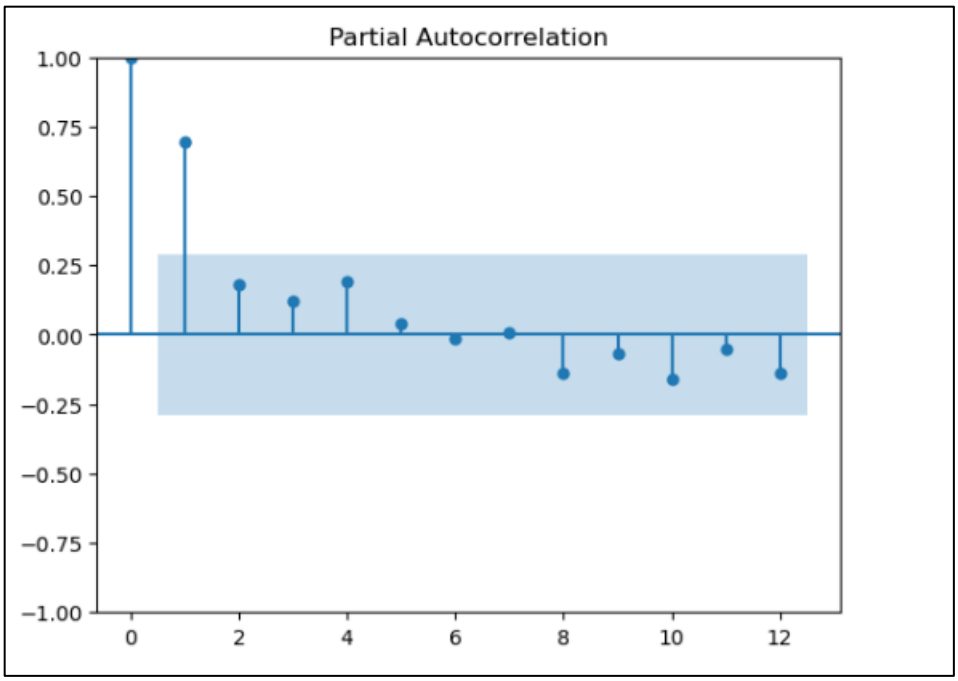
Other: The "Other" category, which likely includes various descent groups not explicitly mentioned, represents approximately 11.9% of the total victim count.



Key Observations:

Original Data (Blue Line): The blue line represents the original monthly crime counts, which exhibit fluctuations over time. These fluctuations may be influenced by various factors, including seasonal variations or changes in crime patterns.

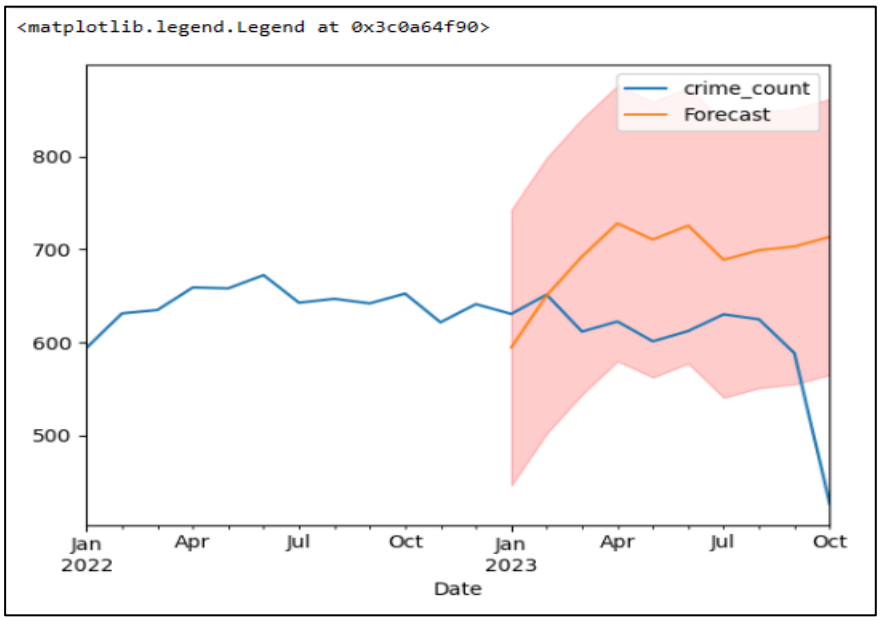
Rolling Mean (Black Line): The black line represents the rolling mean, which is calculated with a window of four months. The rolling mean smooths out the data and helps identify long-term trends. From the plot, it's evident that the rolling mean is not constant, and it fluctuates over time.



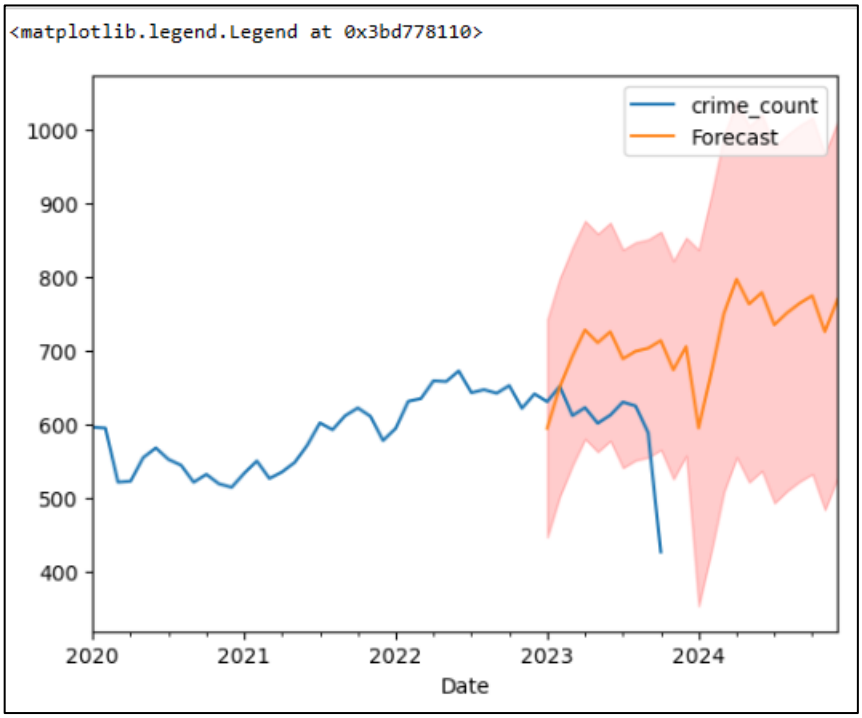
Key Observations:

Confidence Region (Blue Shaded Area): The blue shaded region represents the 95% confidence region. This region helps you assess the statistical significance of the partial auto-correlation values.

Interpretation: If most of the lag values fall within the confidence region, it suggests that the partial auto-correlation values do not have a strong statistical significance. In other words, there may not be a strong linear relationship between the current observation and its past observations at those lags.



The RMSE is approximately 121.85, indicating that the model's predictions deviate from the actual values by this average amount. The higher RMSE suggests that the model didn't perform as well as desired, and it can be attributed to the underlying autocorrelation in the crime data. Autocorrelation measures the relationship between a data point and past data points in a time series.



In the forecast graph, we observe a specific pattern that includes a dip followed by a sudden spike, which can be attributed to seasonal variations within the crime data. This pattern occurs periodically, and the SARIMAX model attempts to capture and reproduce these recurring fluctuations. The dip represents a decrease in the crime rate, suggesting a relatively calmer period, while the sudden spike indicates a rapid increase in the crime rate, signifying a period of heightened criminal activity. The SARIMAX model incorporates a seasonal component that is essential for replicating these variations, as it adjusts for recurring patterns, such as weekly, monthly, or yearly cycles. The sudden spikes and dips are often associated with specific events or trends that tend to repeat at certain intervals, like weekends, holidays, or other seasonal factors. Therefore, the model aims to predict these fluctuations based on historical data.