LOS ANGELES: CRIME DATA ANALYSIS

PROJECT REPORT

of

IE 6400: FOUNDATIONS FOR DATA ANALYTICS ENGINEERING

BY

GROUP NUMBER 15:

ARCHIT SINGH (002813253)

ANIRUDH HEGDE (002813268)

RAHUL ODEDRA (002835990)

SHUBHI SINHA (0022201029)

SANCIA SEROPHENE SALDANHA (002851577)

DEPARTMENT OF COLLEGE OF ENGINEERING

NORTHEASTERN UNIVERSITY

BOSTON, MASSACHUSETTS – 022115

NOVEMBER 2023

ABSTRACT

This project involves a comprehensive analysis of crime data spanning from 2020 to the present, with a primary focus on data preparation and exploratory data analysis. Utilizing the Python programming language, we acquire the dataset and meticulously clean it by addressing issues such as missing data, duplicates, data type conversions, outliers, and categorical encoding.

The core of the project lies in Exploratory Data Analysis (EDA), where we visualize crime trends over time, identify seasonal patterns, and assess the most prevalent crime types. We also investigate regional and municipal variations in crime rates, explore correlations between economic indicators and crime rates, and study the relationship between the day of the week and specific crimes.

Moreover, we scrutinize the potential impact of major events and policy changes on crime rates. To facilitate these analyses, we ensure the presence of essential Python libraries. This project is expected to provide valuable insights into crime data, contributing to a deeper understanding of public safety dynamics.

ACKNOWLEDGEMTS

We wish to express our deep gratitude to those individuals who played essential roles in the successful completion of this data analysis report. Our collaborative efforts, commitment, and teamwork were the driving forces behind this project. We are thankful for the guidance and support provided by the following people:

Professor Sivarit (Tony) Sultornsanee
Associate Teaching Professor of Mechanical and Industrial Engineering

Professor Sivarit Sultornsanee's expertise and mentorship were instrumental in shaping the direction of our analysis. We are appreciative of the valuable insights and guidance he provided during the project.

Teacher Assistant - Venkat Navneeth Burla

Venkat Navneeth Burla, our dedicated teacher assistant, played a significant role in facilitating our progress. His timely assistance and responsiveness to our inquiries were greatly beneficial.

Team Members
Archit Singh, Anirudh Hegde , Rahul Odedra, Shubhi Sinha, Sancia Saldanha,

Our exceptional team members deserve our profound thanks for their unwavering commitment and collaboration. Together, we addressed various aspects of this project, including data sourcing, data cleaning, data analysis, and reporting. The project's quality and success would not have been achievable without their hard work and dedication.

This report stands as evidence of the exceptional teamwork and camaraderie that characterized our project. We take pride in working with such talented and cohesive team members. Our heartfelt appreciation goes out to everyone involved in this endeavour for their invaluable contributions.

TABLE OF CONTENTS

LIST OF FIGURES

Figure 3.11A 29

CHAPTER 1: INTRODUCTION

1.0 INTRODUCTION

In this project, we shall undertake an extensive analysis of crime data spanning from 2020 to the present. Our primary objective is to meticulously clean and prepare the dataset for a comprehensive analysis. This will encompass an exploration of crime trends, patterns, and the underlying factors influencing crime rates.

We shall employ the Python programming language as our primary analytical tool. The project commences with the acquisition of the dataset from the designated source and its subsequent integration into our data analysis environment. Subsequently, we shall conduct a meticulous examination of the dataset, including an initial display of its early records, an evaluation of data types for each column, and a review of column names and descriptions.

The data cleaning process will address issues of missing data, duplicate records, data type conversions, outlier management, and normalization. Furthermore, categorical data shall be encoded where applicable. The heart of the project lies in the Exploratory Data Analysis (EDA) phase, which entails visualizing crime trends spanning the entire temporal spectrum, identifying seasonal patterns, determining the most prevalent crime types, and discerning regional and municipal disparities in crime rates.

A sophisticated examination will explore correlations between economic indicators, where available, and crime rates. We shall delve into the relationship between the day of the week and the prevalence of specific crimes. Additionally, we shall scrutinise the potential influence of major events and policy alterations on crime rates.

To effectively execute these tasks, it is imperative to ensure the presence of essential Python libraries, including Pandas, Matplotlib, Seaborn, NumPy and Jupyter Notebook within the working environment. This project promises to deliver profound insights into crime data, offering valuable contributions to the understanding of public safety dynamics.

CHAPTER 2: DATA SOURCING AND CLEANING

The primary dataset utilized for this analysis was sourced from a government website and covers incident records of crimes occurring within the City of Los Angeles, with records dating back to 2020.

It's worth noting that this dataset represents a transcription of original crime reports, initially documented in paper format. Consequently, there's a possibility of inaccuracies in the data resulting from the manual transcription process. In cases where specific location data was missing, these instances are marked as (0°, 0°).

To protect individuals' privacy, address information is intentionally truncated to the nearest hundred blocks, with exact addresses not disclosed. It's crucial to acknowledge that the data's accuracy hinges on the quality of the original records in the database. Any questions or concerns regarding data quality or specific data points are duly recognized and can be addressed through comments or inquiries.

Furthermore, the preparation of the crime dataset involved several critical steps, encompassing data acquisition, inspection, cleaning, and exploratory data analysis (EDA).

As part of this process, redundant or irrelevant columns, such as URL references and date records, were identified and subsequently removed to streamline the dataset. Additionally, the feature extraction method was employed to extract pertinent insights from the available data. This meticulous data preparation is pivotal in ensuring the dataset's reliability and relevance for the comprehensive analysis undertaken in this research.

It forms the foundational basis upon which the subsequent findings and insights are constructed.

The data-cleaning process for the crime dataset involved a sequence of methodical actions:

Preliminary Cleaning:

1. Initial assessment to pinpoint the presence of missing values in the data.

Column Standardization:

1. Functions were implemented to exclude unnecessary columns.
2. The 'Victim Sex' column was regularized by replacing non-conforming entries with a consistent placeholder.

Data Enhancement and Error Handling:

1. A calculation of victim gender distribution was conducted.
2. The dataset was augmented with a binary indicator to reflect the use of weapons in crimes.
3. Default values were assigned to missing entries in the 'Premises Description' and 'Premises Code' fields.

Date and Time Formatting:

1. Missing 'Cross Street' information was incorporated into the 'Location' column.
2. The 'Date Reported' field was converted to a standard datetime format.
3. Time of occurrence entries were converted into an 'hour: minute' format.

Integration and Final Touches:

1. The dataset was restructured by combining the date and time of crime occurrence into a single column.
2. An evaluation of the 'Victim Descent' column was performed, filling in missing data with a placeholder.
3. Age data was refined to include only logical age ranges.
4. Superfluous columns were discarded to streamline the dataset.

Post-Cleaning Validation:

1. The cleaned dataset was then examined for any remaining missing values.
2. Additional time-related features such as year, month, and a concatenated year-month string were extracted from the occurrence datetime to aid in temporal analysis.

CHAPTER 3: VISUALIZATION AND ANALYSIS

In Chapter 3, we explore the project's data through a series of visual representations and the subsequent analysis of these visuals. The objective is to translate complex data sets into understandable formats that can inform our questions and objectives.

3.1 Overall Crime Trends:

The extensive examination of crime data spanning from 2020 to 2023 has uncovered several noteworthy insights into the patterns of criminal activities. Firstly, it highlights a rising trend in crime rates, with the highest occurrences in 2022, closely trailed by 2021 and 2020. Interestingly, there was a significant surge in crime in January 2020, coinciding with the initial outbreak of the COVID-19 pandemic.

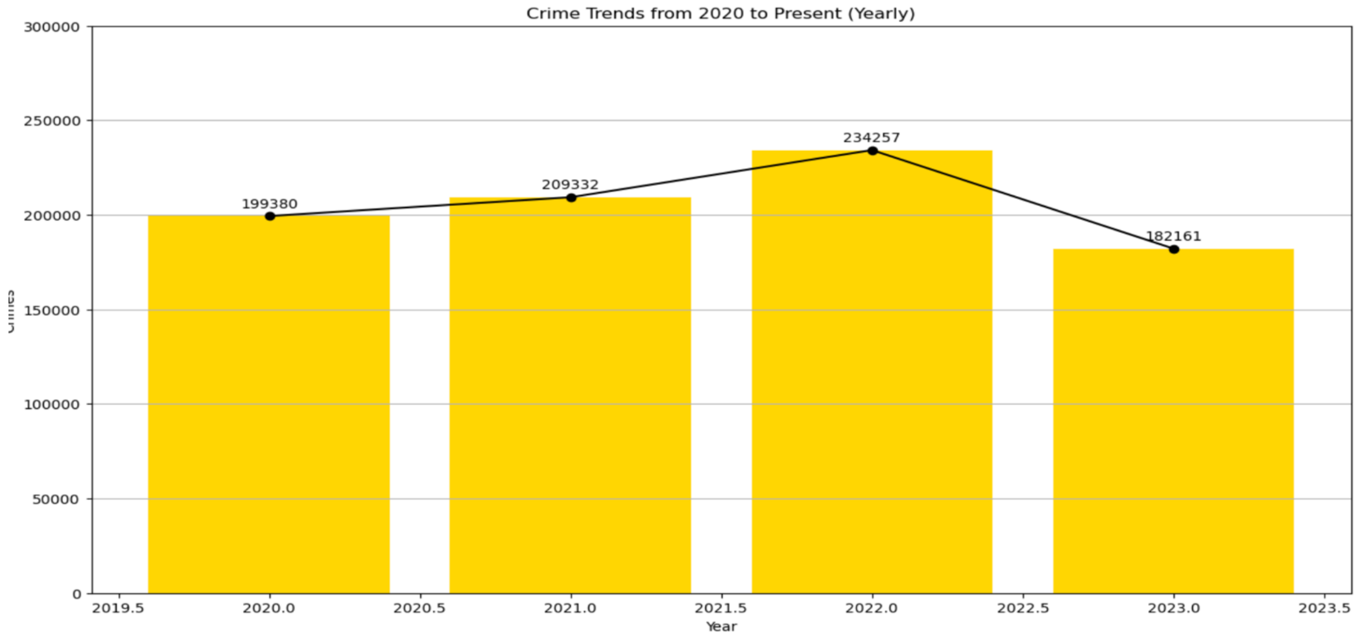

Figure 3.1 a) Yearly Average Crime Trends

We lack data for November and December 2023 in our dataset, preventing us from making any conclusions or representing the sharp decline during that period in our graph.

Additionally, in a rapidly advancing world, crime has taken on new and diverse forms, contributing to its overall increase.

The histogram clearly illustrates that the year 2022 witnessed the highest frequency of criminal incidents compared to all the other years.

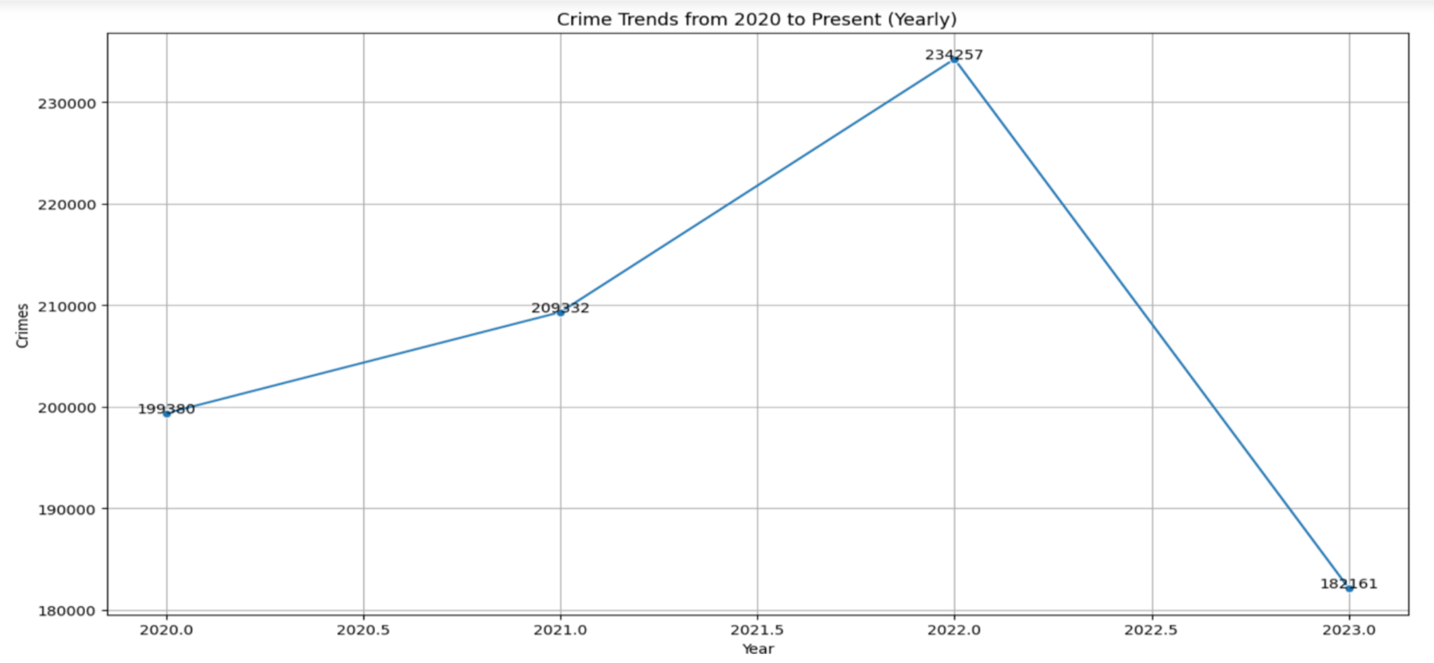

Figure 3.1 b) Yearly Average Crime Trends (LineChart)

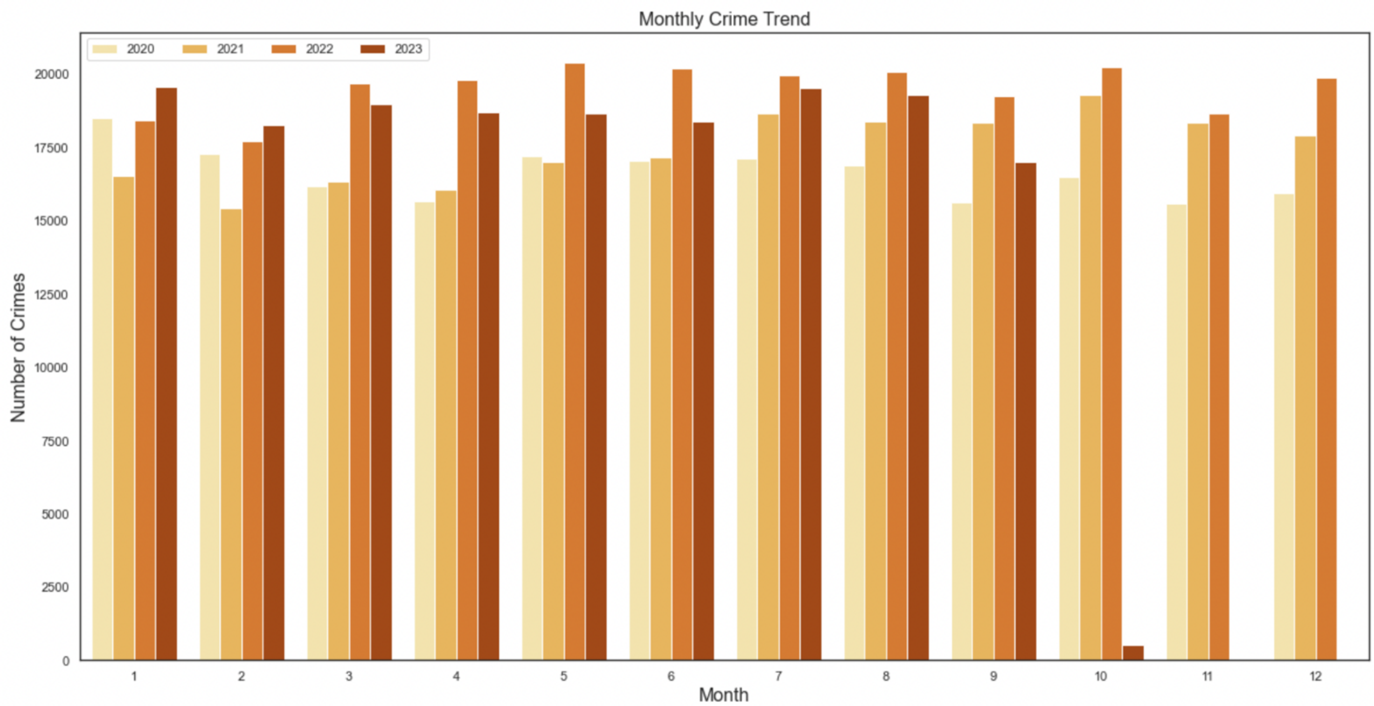

Figure 3.1 c) Monthly Crime Trends

The next analysis performed is over the seasonal pattern of crimes over the years.

3.2 Seasonal Patterns:

Insights drawn from the graph reveal that the months of July, August, and September witness the highest crime rates, whereas November and December experience the lowest incidents. This pattern may be linked to warmer weather conditions, leading to larger gatherings and increased crime opportunities during the summer months. Conversely, during the holiday season in November and December, heightened law enforcement presence acts as a deterrent to criminal activity. Moreover, the extended daylight hours in summer provide additional windows for criminal incidents.

We now analyse the top 15 contributors to the crime data.

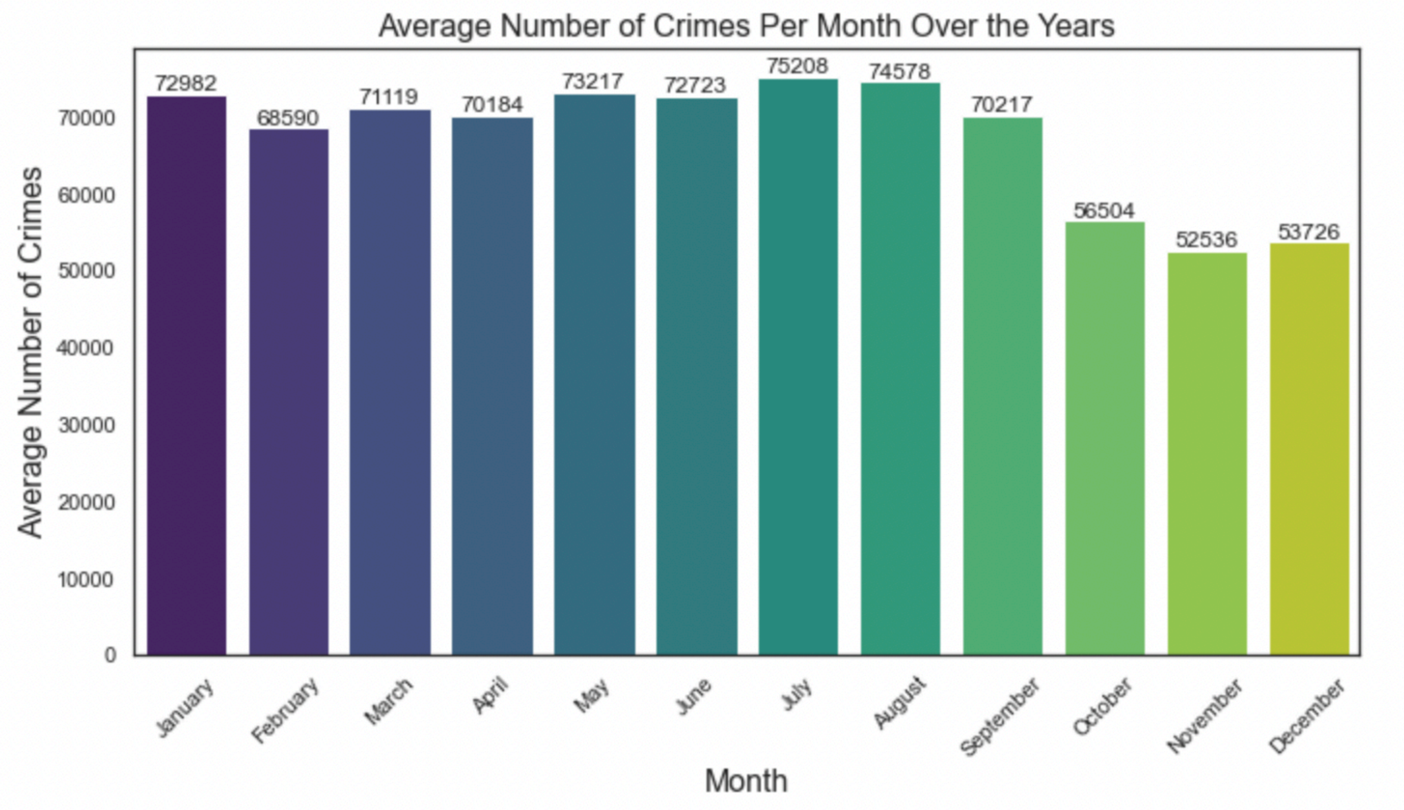

Figure 3.2 Average Number of Crime

3.3 Most Common Crime Type:

It can be inferred that vehicle theft stands out as the most prevalent form of theft, followed by battery and simple assault. These crimes are not only frequently committed but are also relatively easier to execute, possibly contributing to their higher incidence. Furthermore, the penalties for such offenses tend to be less severe, potentially encouraging their perpetration.

On the other hand, vandalism and shoplifting appear to be less common, possibly due to the extensive presence of surveillance cameras on streets and within malls. This heightened surveillance leads to quicker identification of wrongdoers, which, in turn, results in more effective enforcement of punishments and serves as a potential deterrent against these crimes.

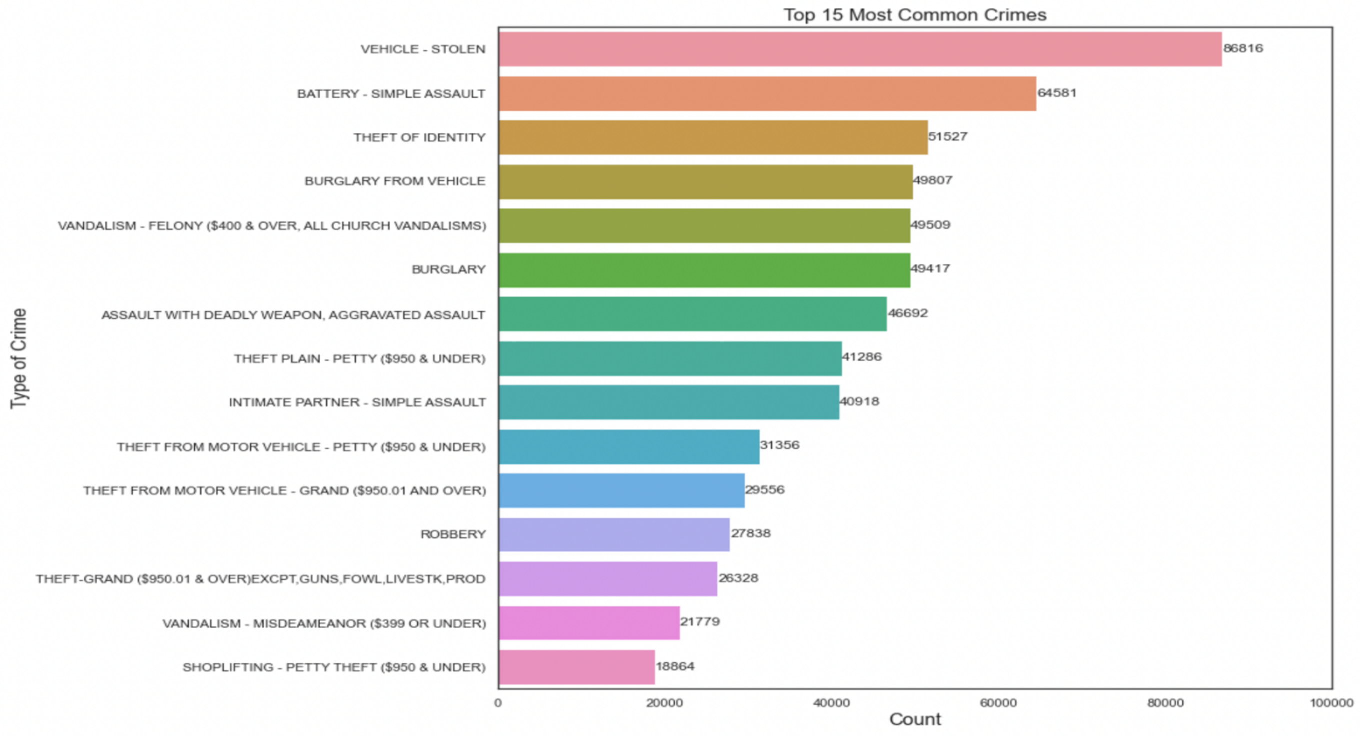

Figure 3.3 a) Top 15 Most Common Crimes

Furthermore, the month of October 2021 records the highest number of car thefts, while October 2023 exhibits the lowest incidence of vehicle theft.

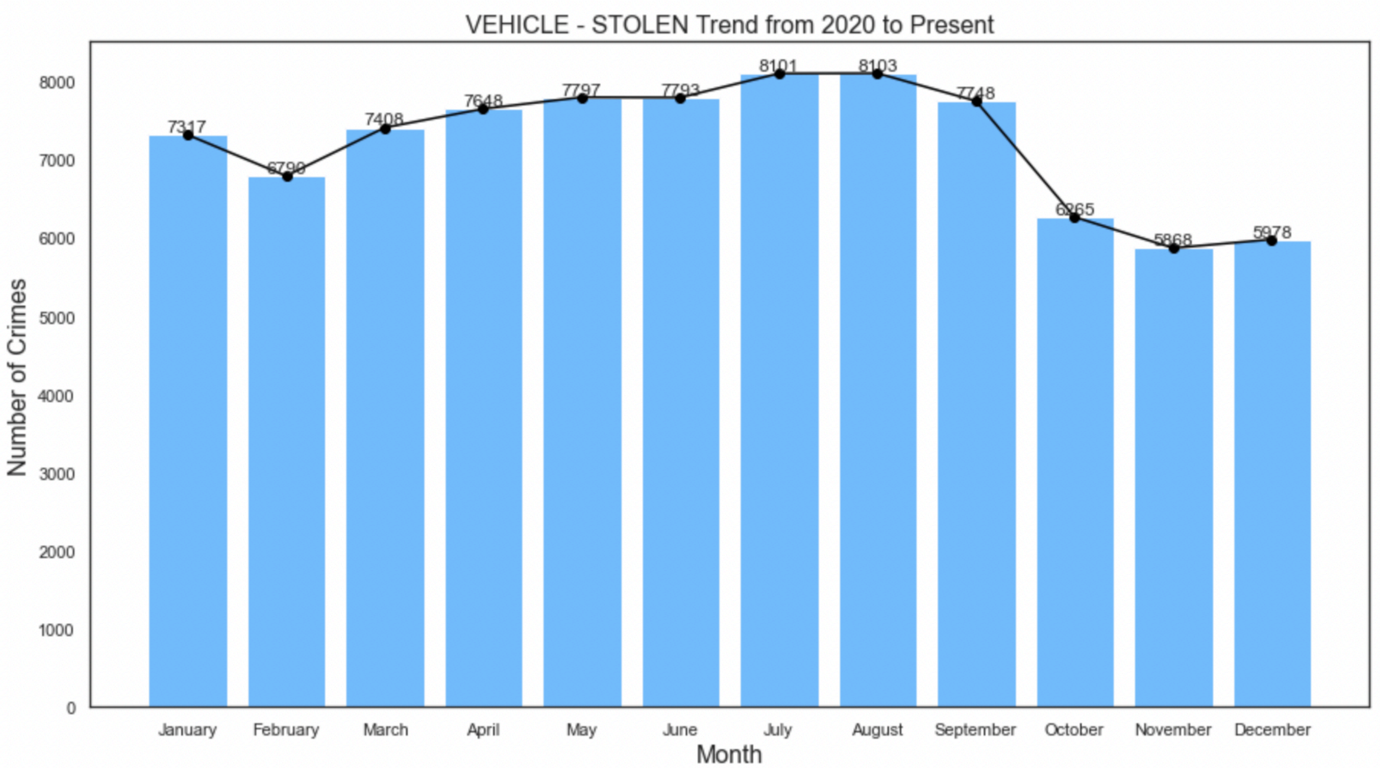

Figure 3.3 b) Vehicle Stolen Trend from 2020 to Present

This contrast can be attributed to various factors. In October 2021, a combination of potential factors, such as a lack of effective security measures, an upsurge in opportunities for car theft, or a higher demand for stolen vehicles, may have contributed to the increased number of thefts. Conversely, by October 2023, improvements in security, heightened awareness, or stricter law enforcement could have resulted in a decrease in car theft incidents.

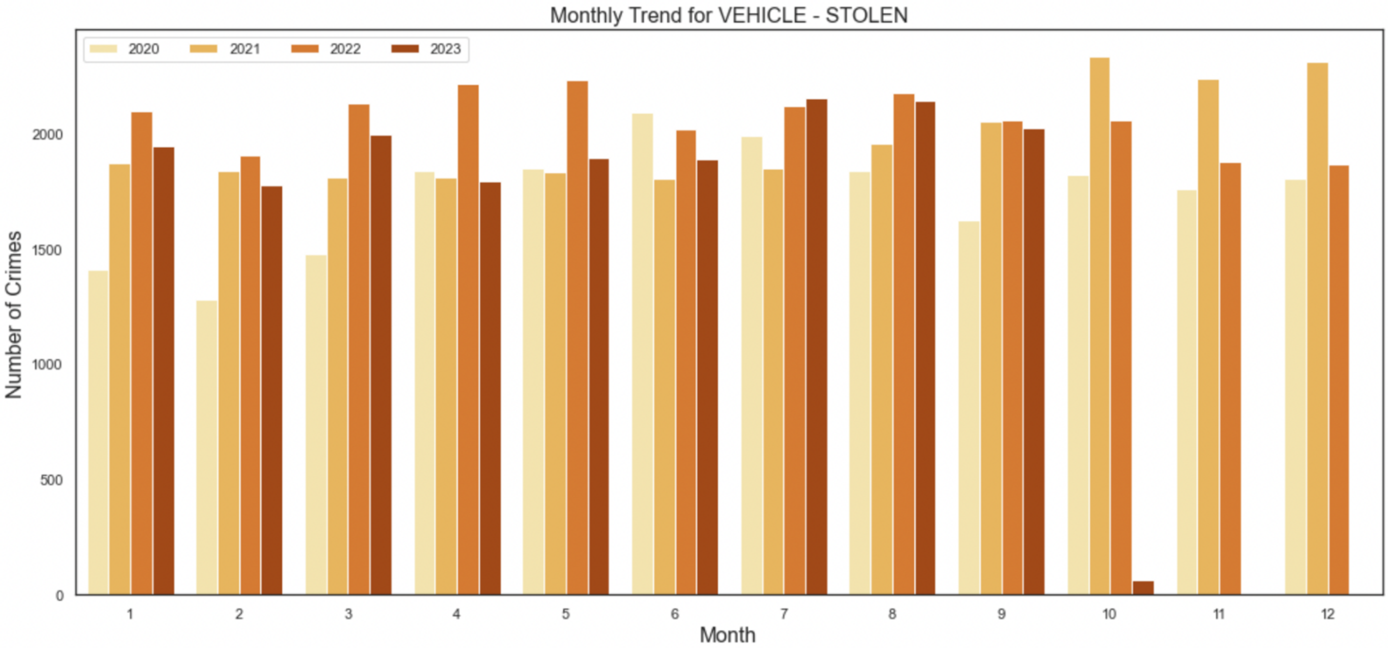

Figure 3.3 c) Monthly Trend for Vehicle Stolen

3.4 Regional Difference:

Now we investigate if there are any notable differences in crimes between differences The highest crime rates are observed in Central LA, with 77th Street and Pacific areas following closely behind. Conversely, the lowest incidence of crime is reported in Hollenback and Foothill regions.

This discrepancy can be explained by various factors. In Central LA, a higher population density, increased economic disparities, and urban environments may contribute to elevated crime rates. 77th Street and Pacific areas might face similar urban challenges, leading to their ranking right after Central LA .Conversely, in Hollenback and Foothill regions, factors such as a lower population density, stronger community bonds, and potentially more effective law enforcement efforts may lead to decreased criminal activity

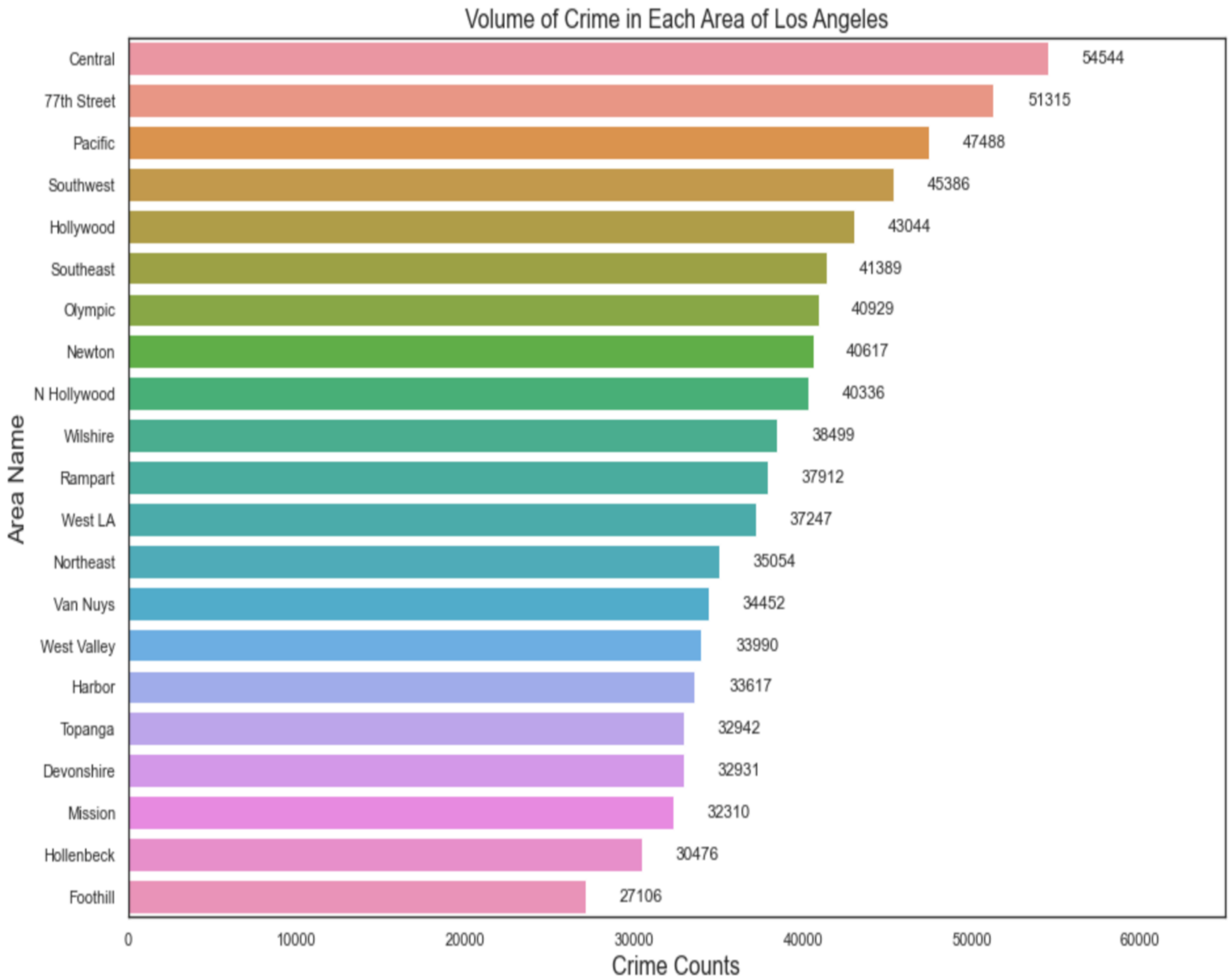

Figure 3.4 Volume of Crime in Each Area of Los Angeles

3.5 Correlation with Economic Factors:

Now we will explore the correlation between the economic factors and crime rates. The insights derived from the line graph depicting the correlation between economic factors and crime rates reveal a complex interplay of social and economic dynamics. One key observation is that areas marked by economic deprivation, including high poverty rates and limited job opportunities, exhibit higher crime rates. Individuals facing economic hardship may resort to criminal activities as a means of survival or to enhance their living conditions. Moreover, regions with pronounced income inequality often experience elevated crime rates, driven by frustration among those in lower-income brackets.

This feeling of resentment may fuel criminal behaviour as a form of protest or as a means to bridge the wealth gap. Insufficient access to quality education, healthcare, and social services, often tied to economic factors, contributes to a higher likelihood of engaging in criminal activities. Additionally, the illicit drug trade, prevalent in economically disadvantaged areas, significantly impacts crime rates, particularly related to violent and property crimes. Understanding the intricate relationship between economic conditions and crime is crucial for policymakers to develop comprehensive strategies aimed at reducing crime rates while improving the overall well-being of communities.

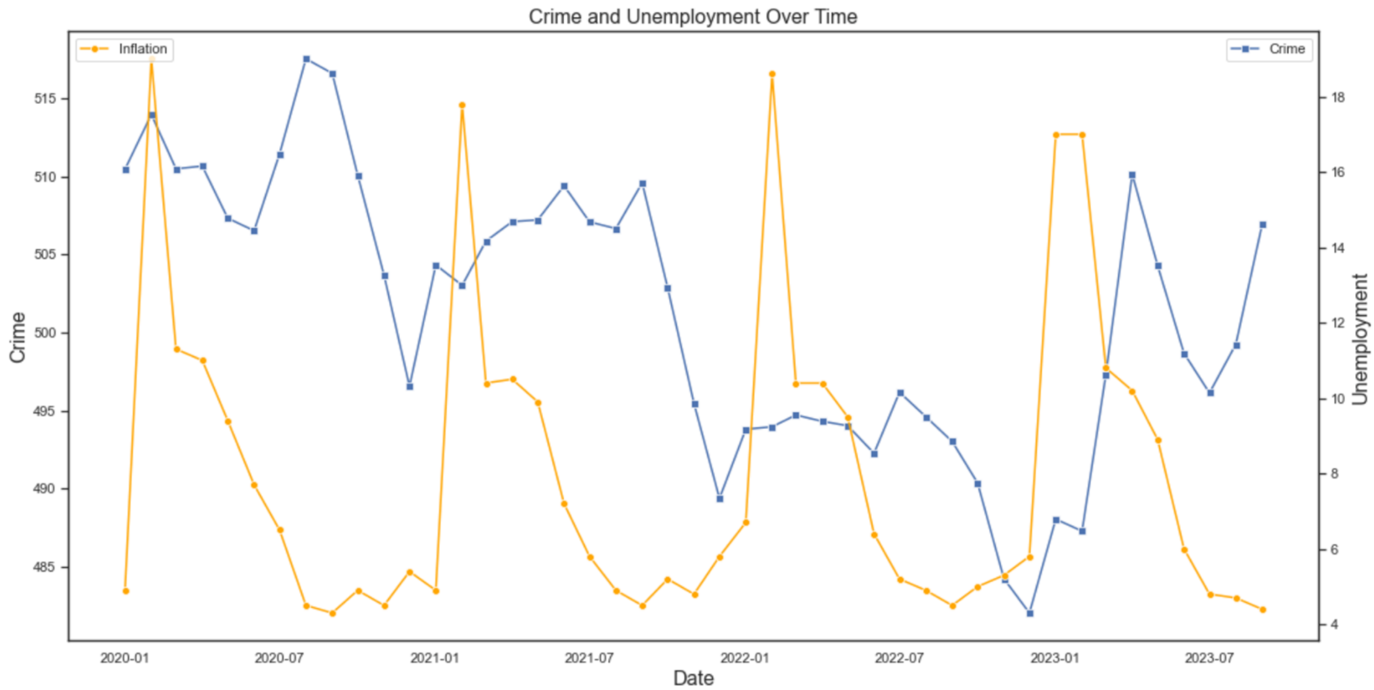

Figure 3.5 a) Crime and Unemployment Over Time

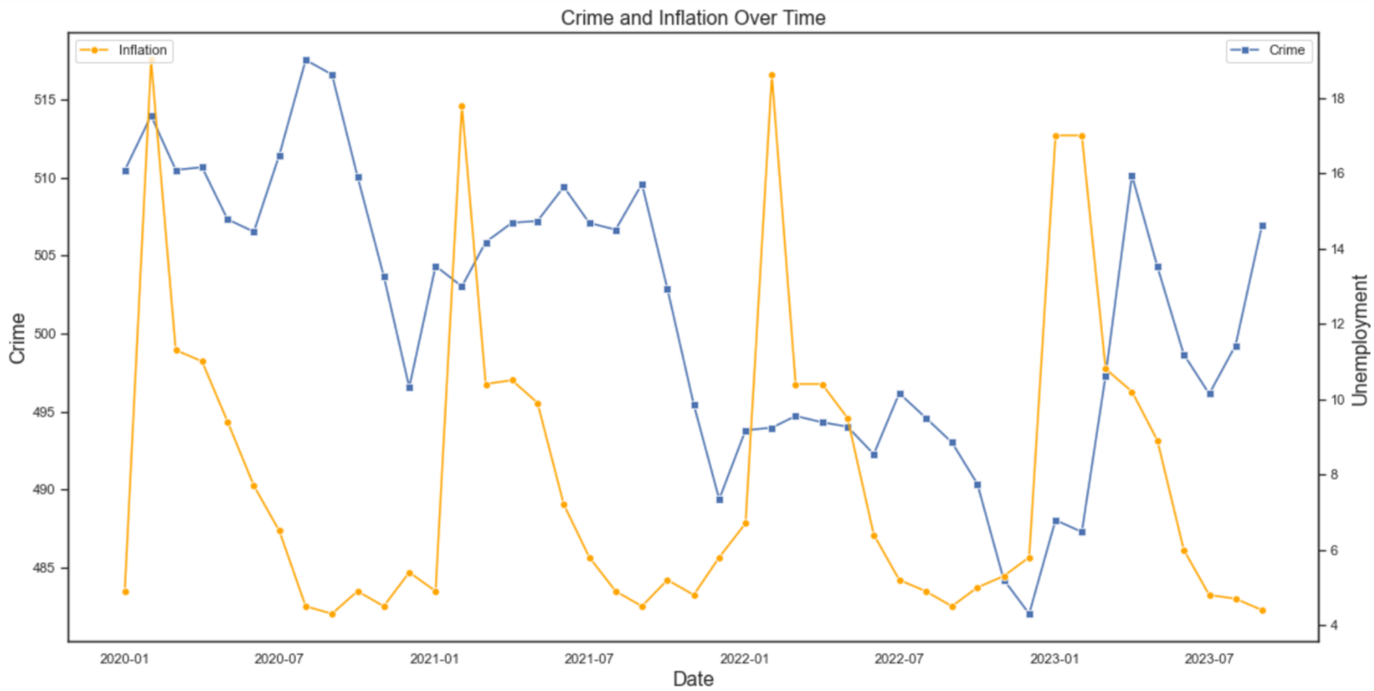

Figure 3.5 b) Crime and Inflation Over Time

	Crime	Unemployment
Crime	1.000000	-0.095023
Unemployment	-0.095023	1.000000

Figure 3.5 c) Correlation of Crime and Unemployment

	Crime	Inflation
Crime	1.000000	0.261255
Inflation	0.261255	1.000000

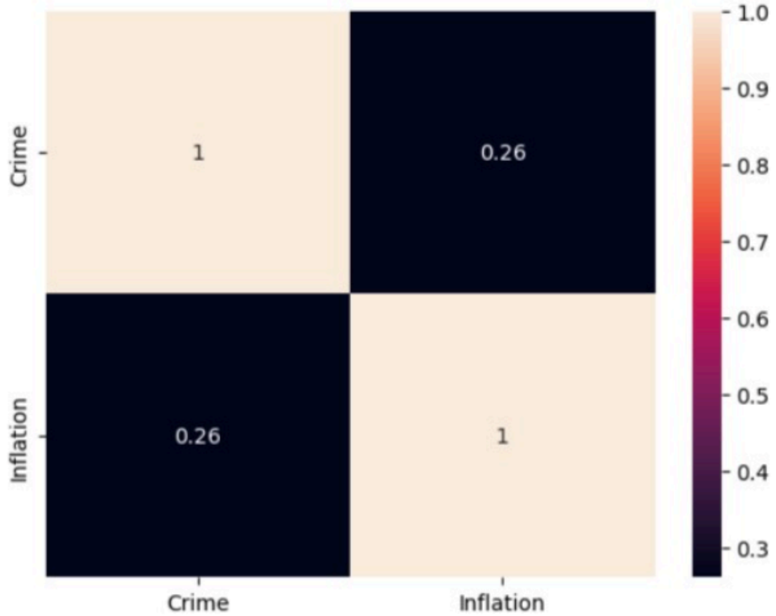

Figure 3.5 d) Correlation of Crime and Inflation

3.6 Day of the Week Analysis:

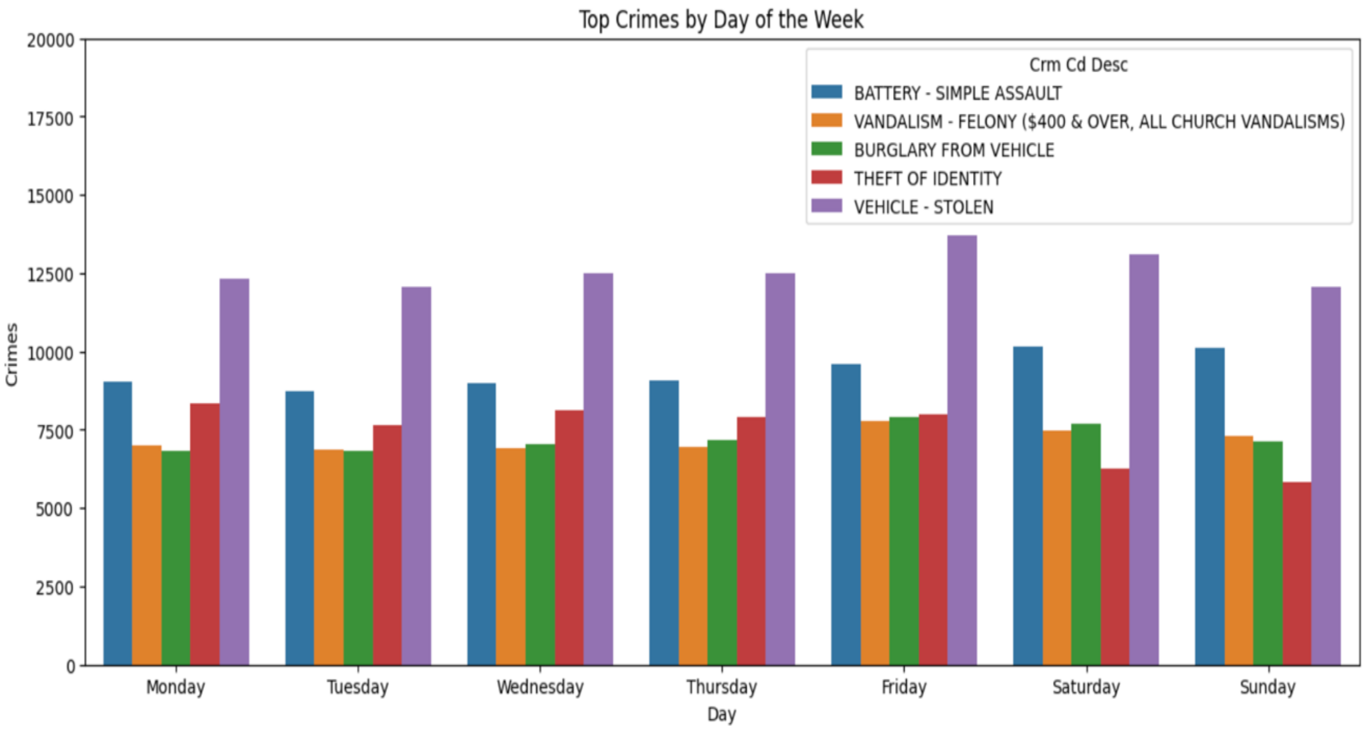

Figure 3.6 Top Crimes by Day of the Week

In the above analysis, it can be observed that during the weekend (Saturday, Sunday) theft of identity is reduced compared to other days of the week. Another pattern observed during all the days of the week is that the highest crime is Vehicle stolen, followed by Battery- simple assault, and then Vandalism. These three trends are constant throughout the week.

3.7 Impact of Major Events:

a) George Floyd Protest
b) US Elections 2020
c) Minimum Wage for selected immigrant workers
d) Covid-19

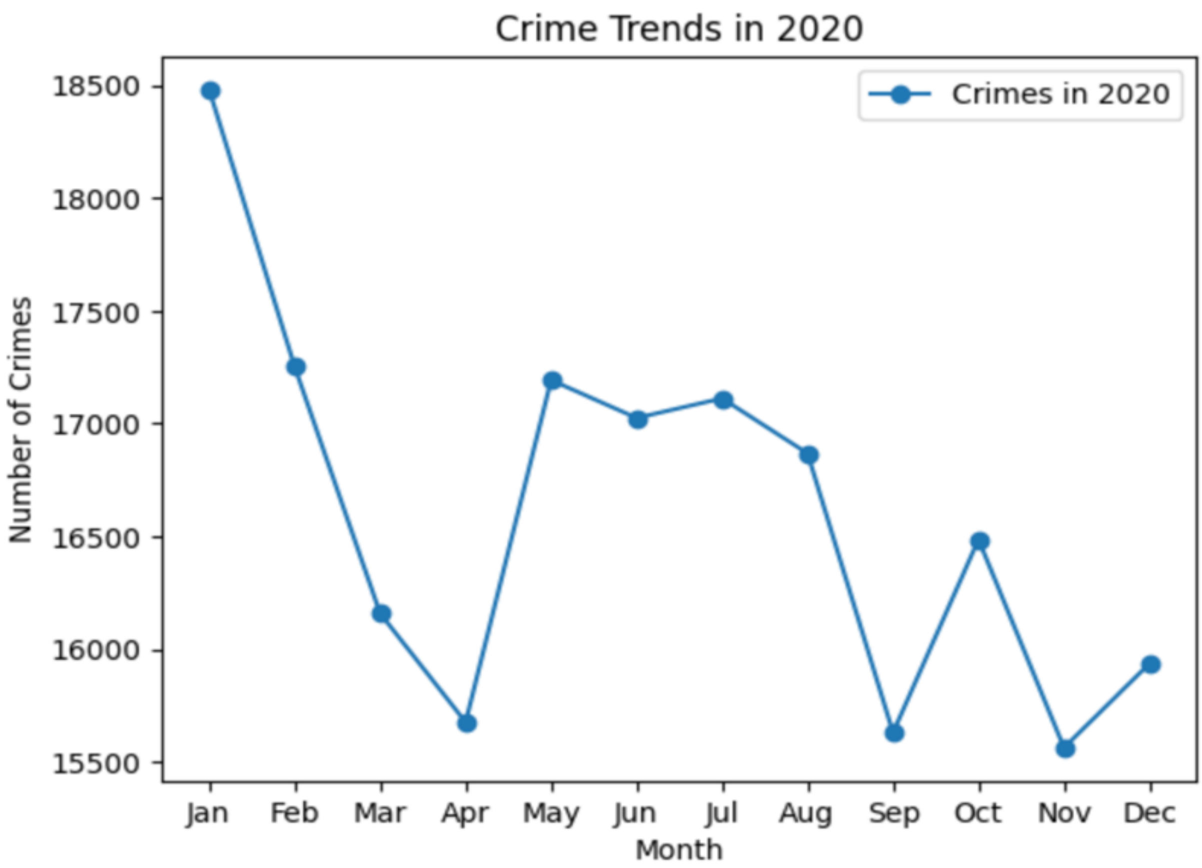

Figure 3.7 a) Crime Trends in 2020

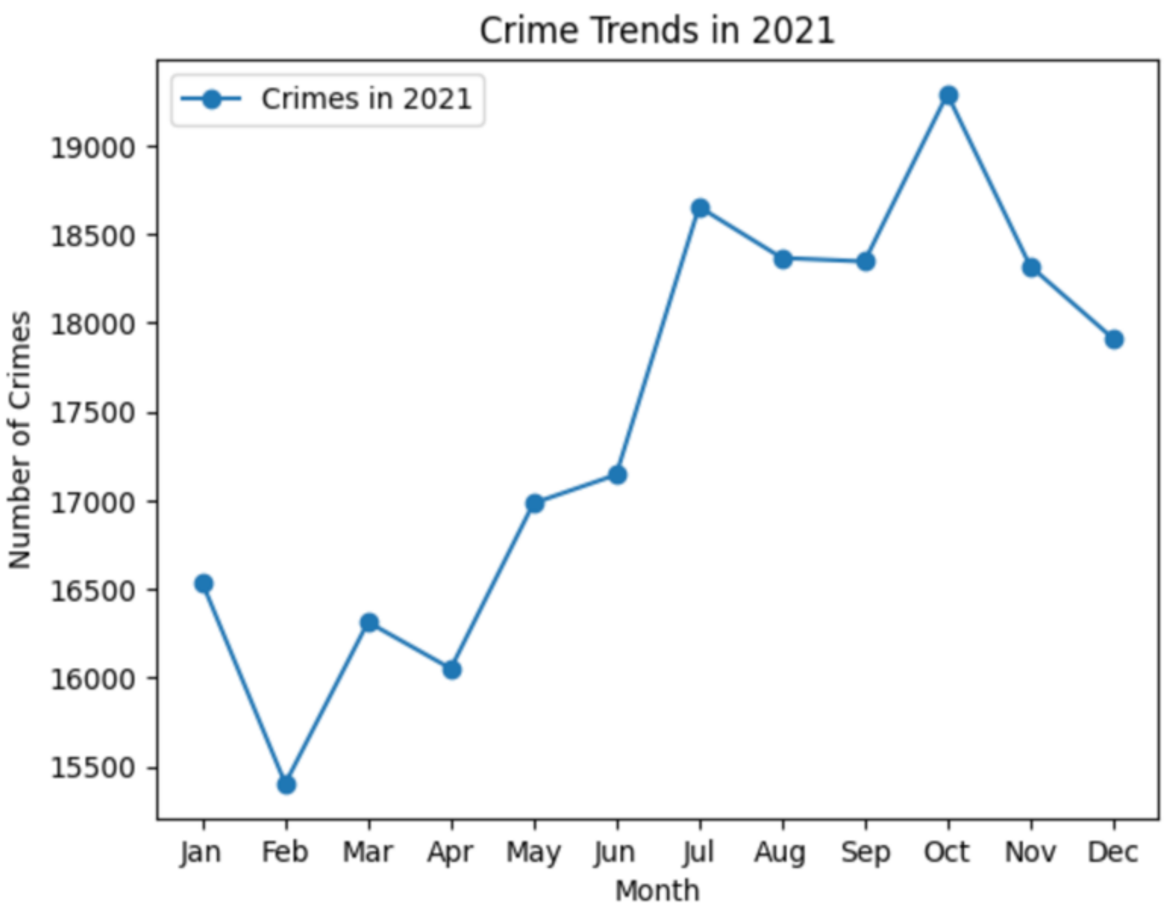

Figure 3.7 b) Crime Trends in 2021

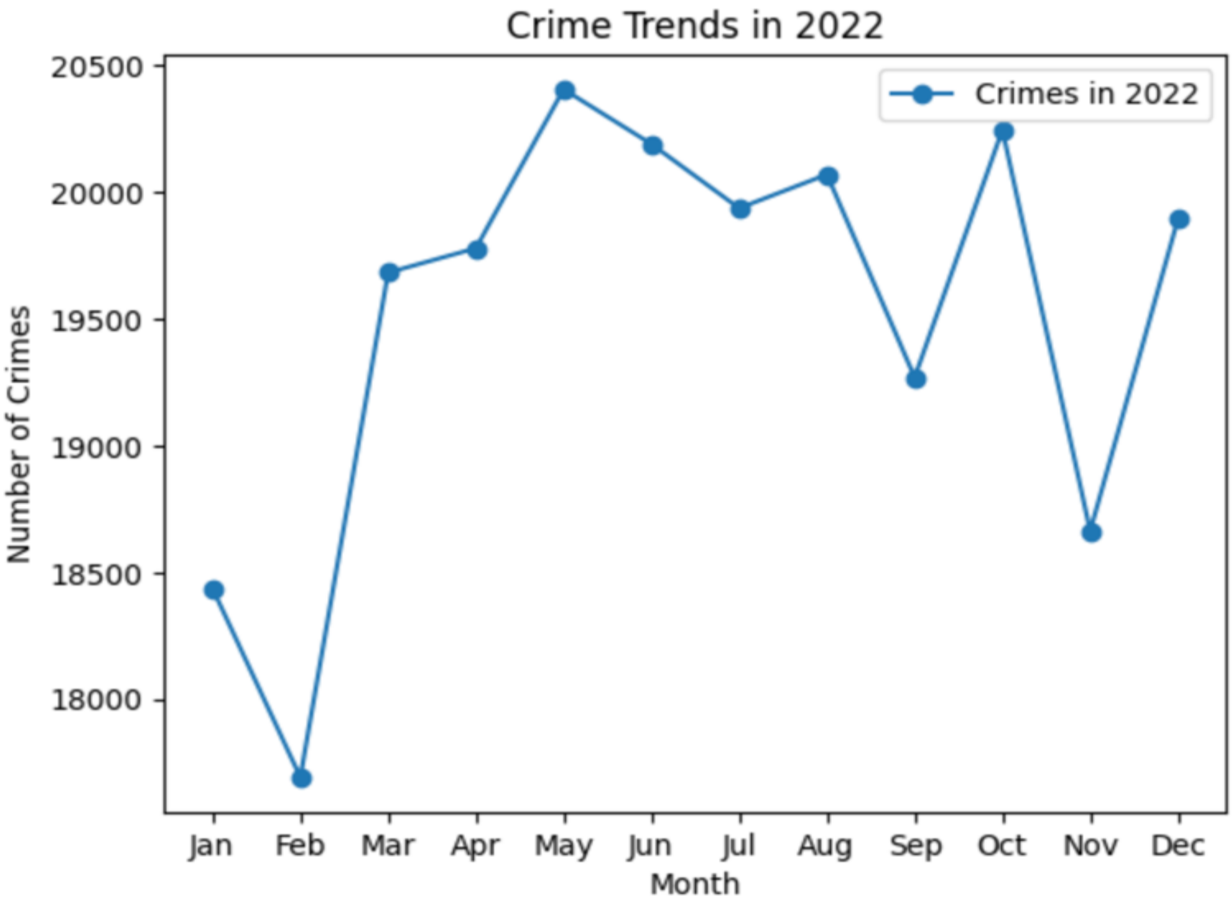

Figure 3.7 c) Crime Trends in 2022

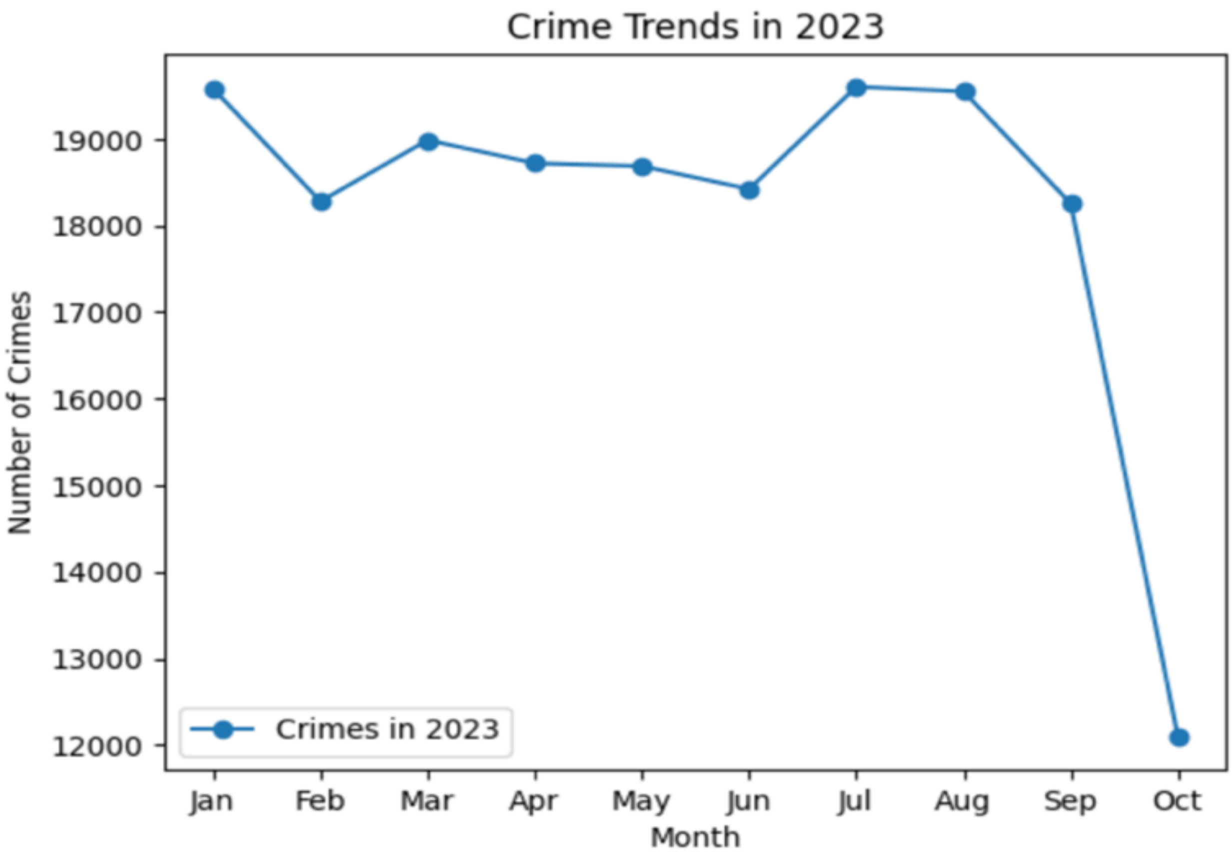

Figure 3.7 d) Crime Trends in 2023

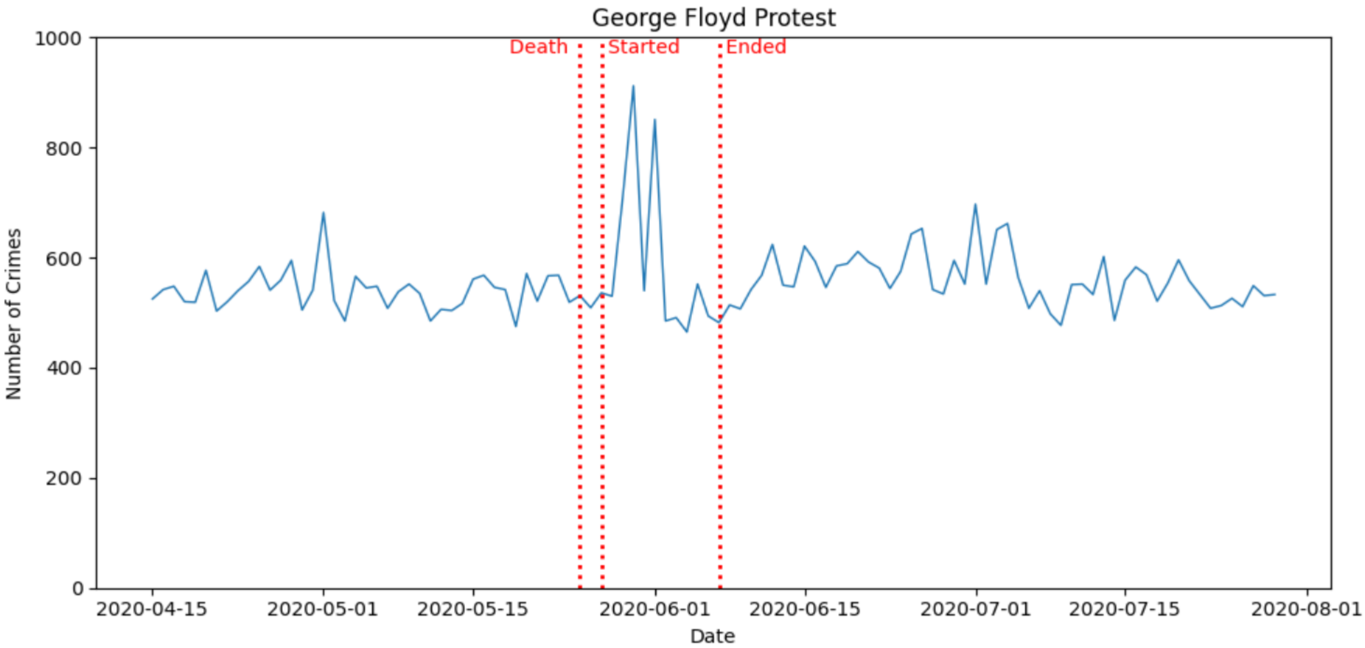

Figure 3.7 e) George Floyd Protest

In May and June for the year 2020, there was a spike in the crime rate due to the George Floyd protests.

In Los Angeles, crime rates in October 2021 may have been influenced by a range of local and seasonal factors. The autumn season's arrival, with its cooler temperatures and longer nights, could have created conditions more conducive to certain types of criminal activity.

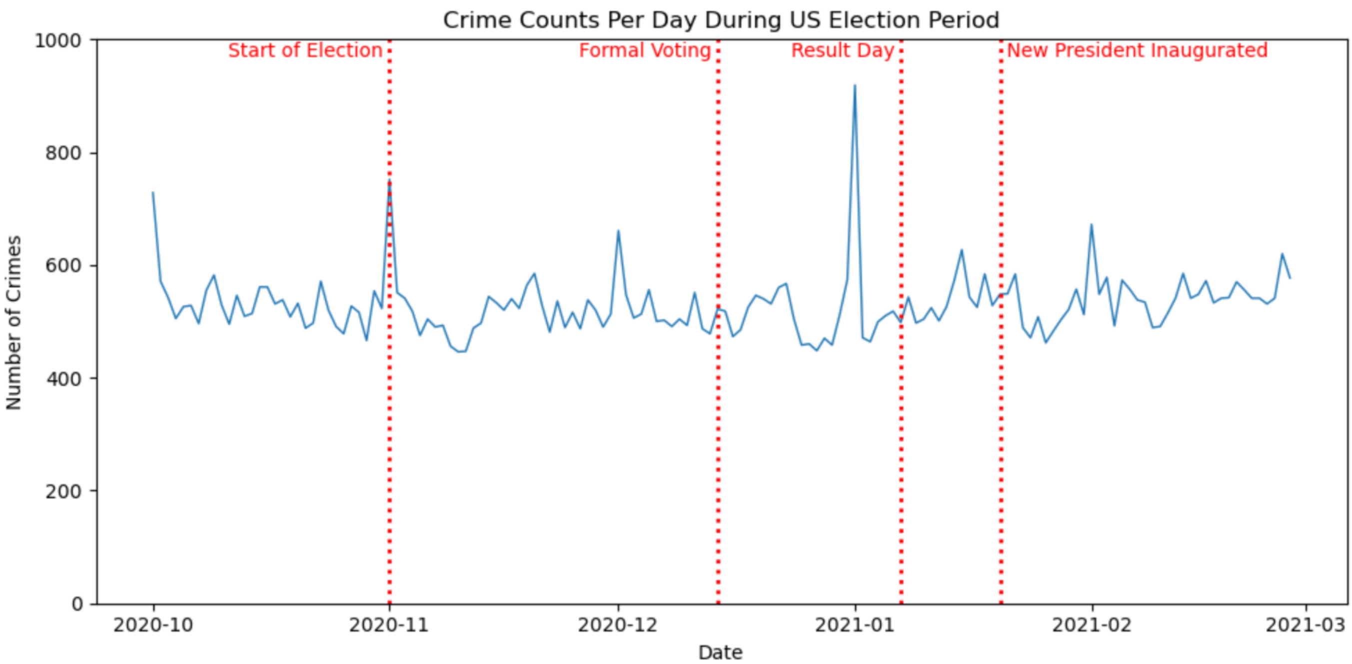

Figure 3.7 f) US Elections 2020

Additionally, the celebration of Halloween, a widely observed holiday, can sometimes lead to increased incidents of vandalism and property damage.

Social and cultural events in the area, as well as economic pressures brought about by challenging times, may have contributed to fluctuations in crime rates.

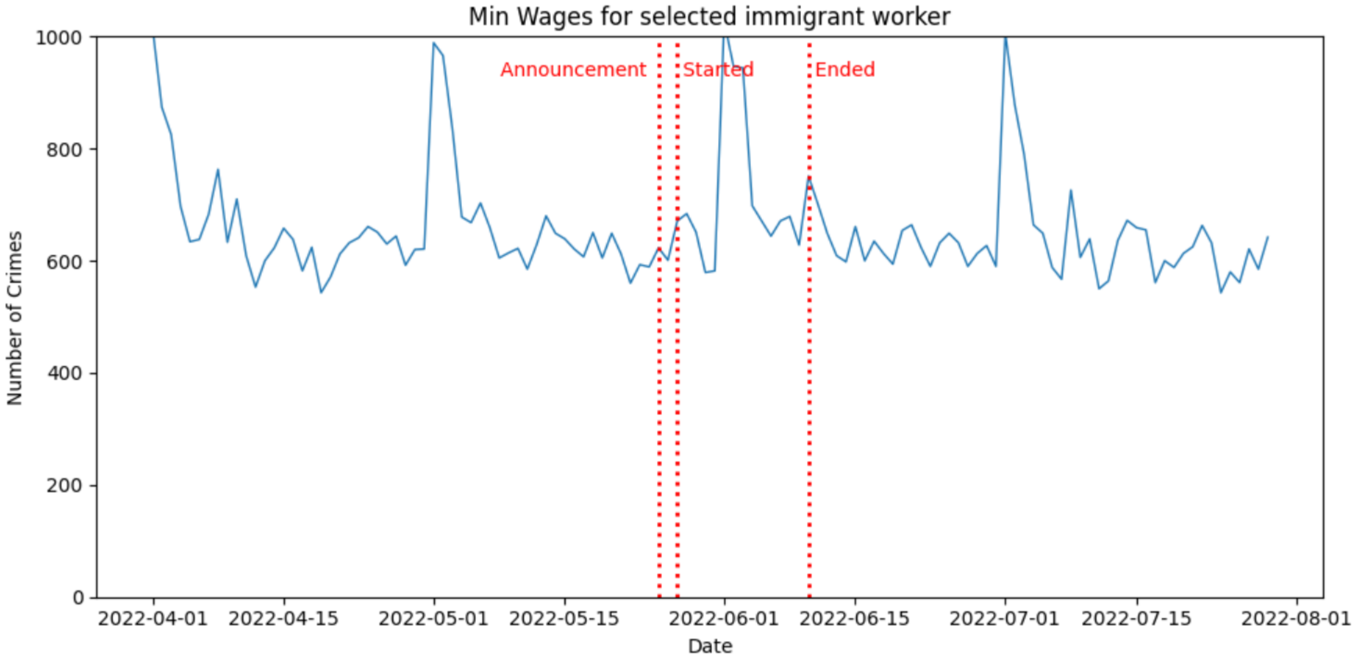

Figure 3.7 g) Minimum Wages for Selected Immigrant Worker

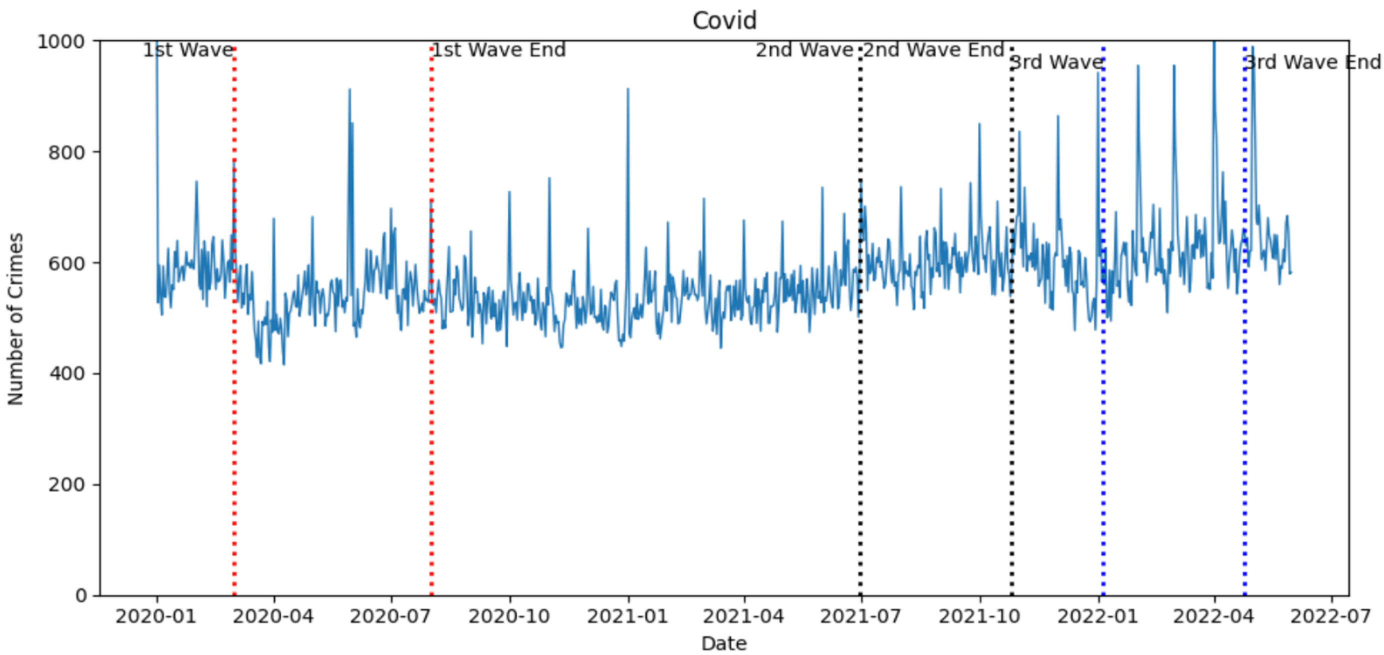

Figure 3.7 h) COVID - 19

Moreover, the ongoing impact of the COVID-19 pandemic, with its disruptions to employment, education, and social services, could have played a role in altering crime patterns. Changes in law enforcement strategies, community dynamics, and local events may have also influenced the elevated crime rates in Los Angeles during that particular month.

In the Year 2022 from months of May to June, the crime rates spiked due to a rise in minimum wages of certain immigrant workers which caused a conflict with the rest due to the injustice.

3.8 Outliers and Anomalies:

According to the given scatter plot and box plot above for Victim Age, we can spot certain Outliers and Anomalies, see the below Figure 3.8 a and b.

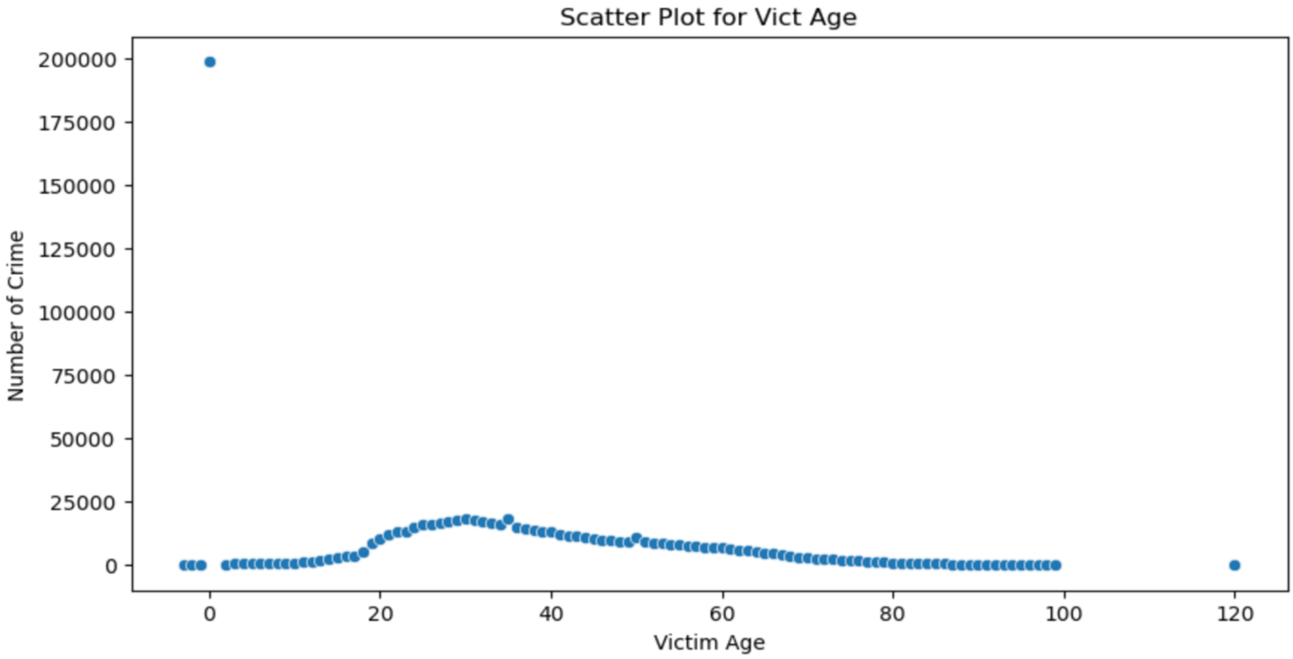

Figure 3.8 a) Scatter Plot for Victim Age

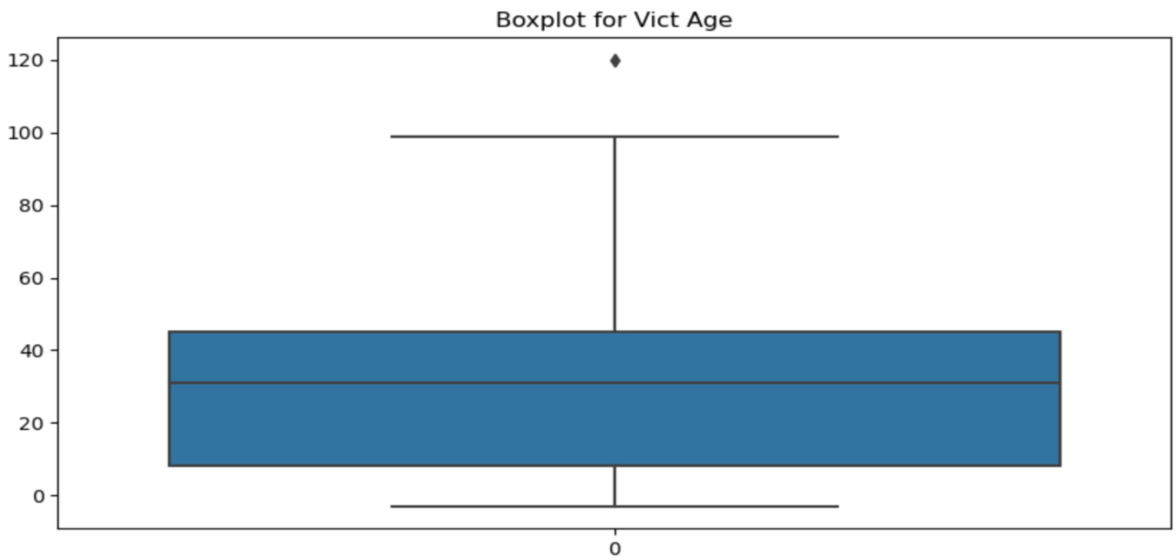

Figure 3.8 b) Boxplot for Victim Age

Additionally, upon observation, it was observed that victims whose ages are: -3, -2, -1, and 0 are present in the data set, with the highest number of victims aged 0 (198667).

3.9 Demographic Factors:

Now, we will analyse the dataset to discover any relationships between demographic factors, such as age and gender, and specific types of crimes.

Furthermore, we will visualize the Crime distribution by Victim Descent as shown below:

- H-Latin are the highest population of victims
- X-Unknown descent victims are the second-highest
- W- White has the third-highest
- B-Black victims are the fourth

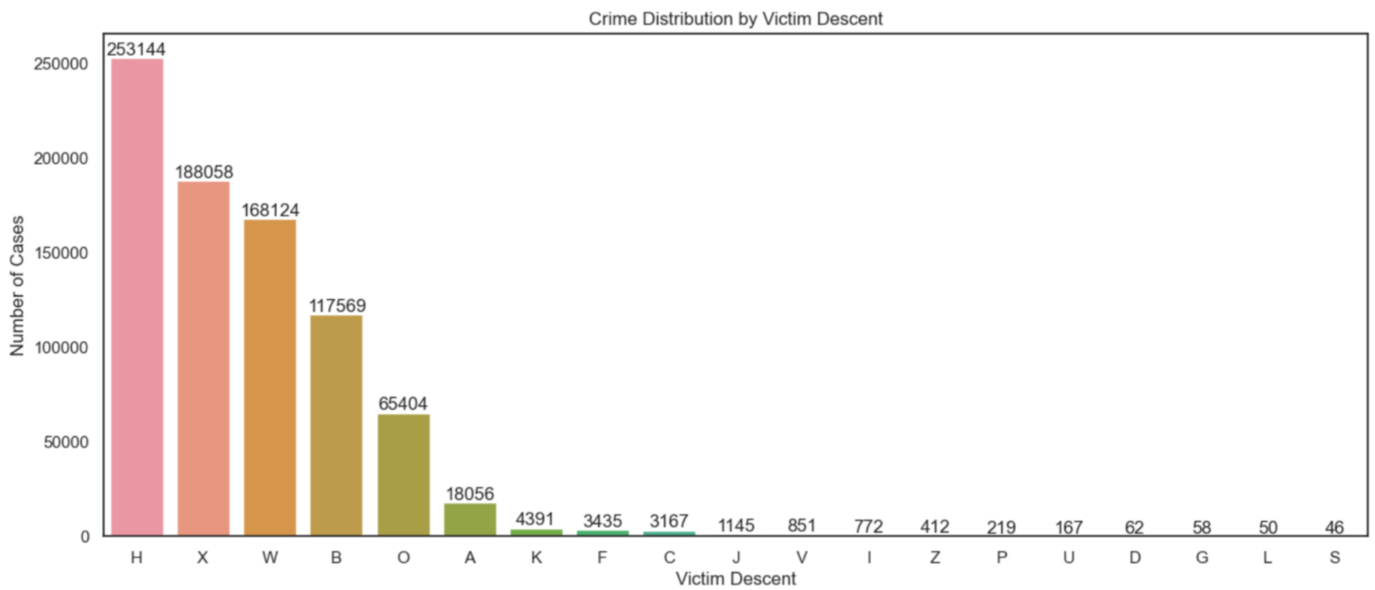

Figure 3.9 a) Crime Distribution by Victim Descent

Additionally, visualizing the Most used weapon in Crimes (Top 5):

- Physical violence such as the use if Hands, Fist, Feet or other bodily parts have the highest number of cases.
- Followed by similarity in the other four – Unknown weapons, Verbal threats, Hand Gun and least Semi-Automatic pistol

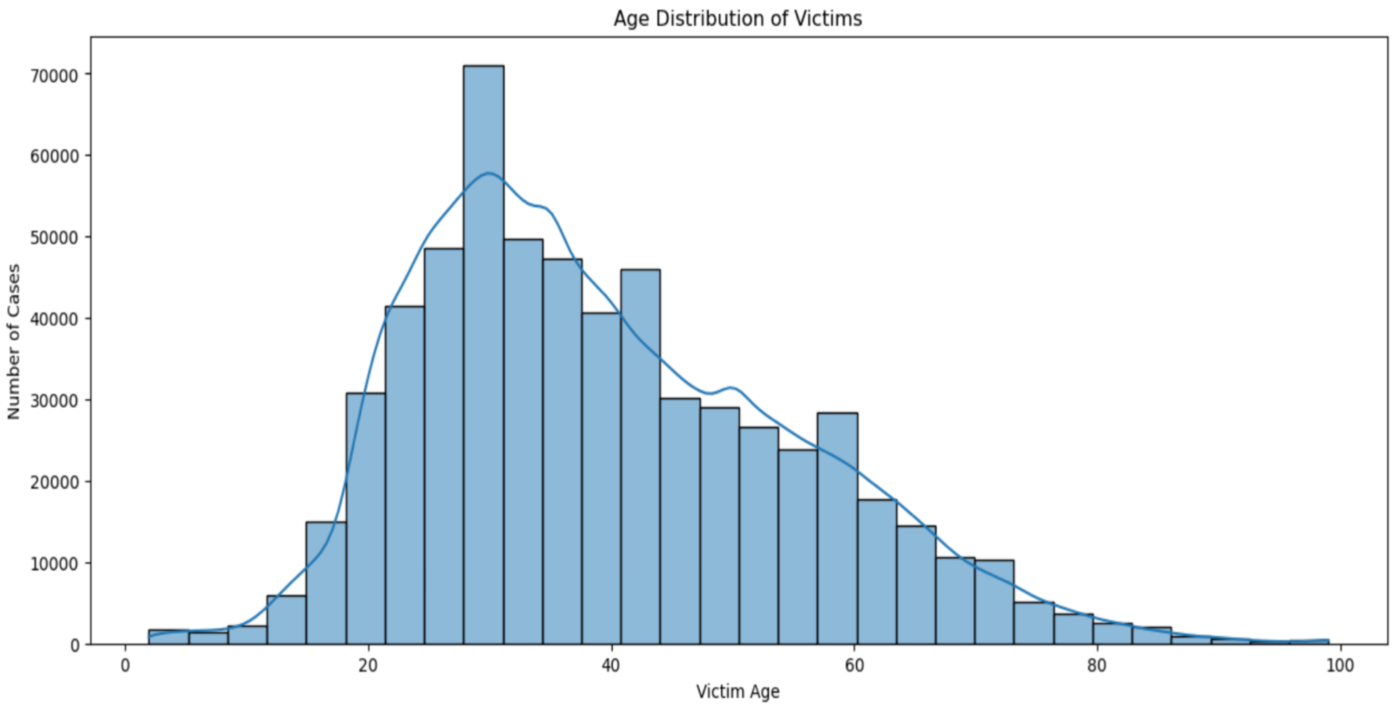

Figure 3.9 b) Age Distribution of Victims

According to the data set, Males have been more targeted than females due to the economic disparity in LA. Income inequalities and limited access to education and employment opportunities are some of the factors contributing to this.

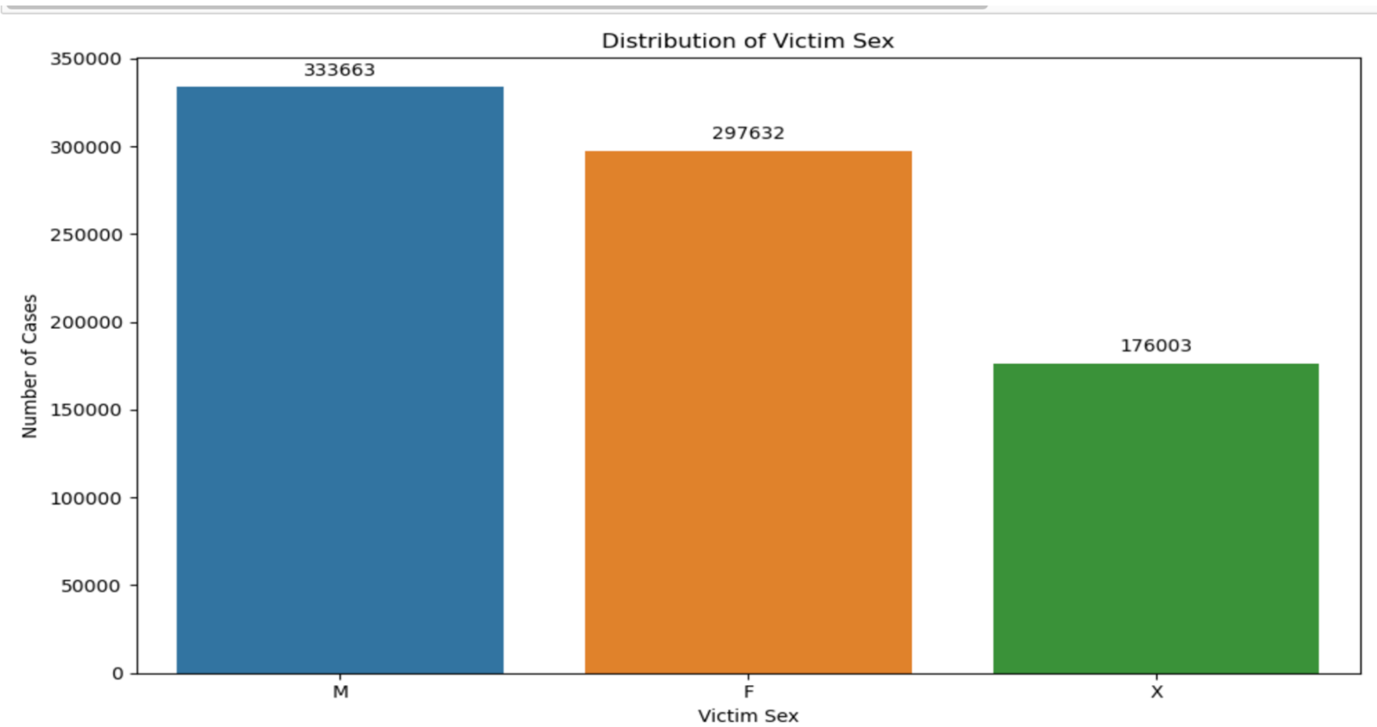

Figure 3.9 c) Distribution of Victim Sex

3.10 Predicting Future Trends:

SARIMA takes into account the past values (autoregressive, moving average, any seasonality patterns) and predicts future values based on that. Since It brings in seasonality as a parameter, it's significantly more powerful in forecasting complex data spaces containing cycles.

Reference

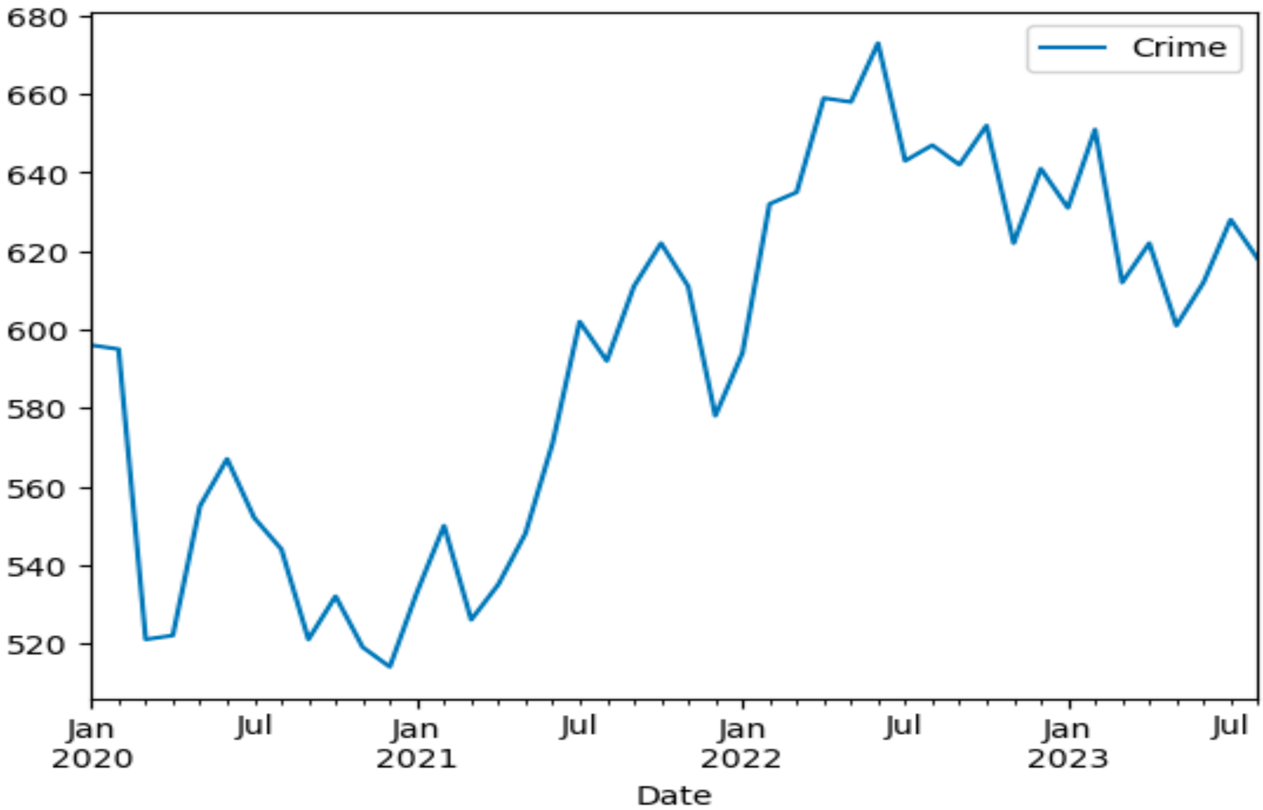

Figure 3.10 a) Crime Fluctuations over Time

From the figure above we can observe that the data violates the stationarity assumption of constant mean, constant variance and constant covariance.

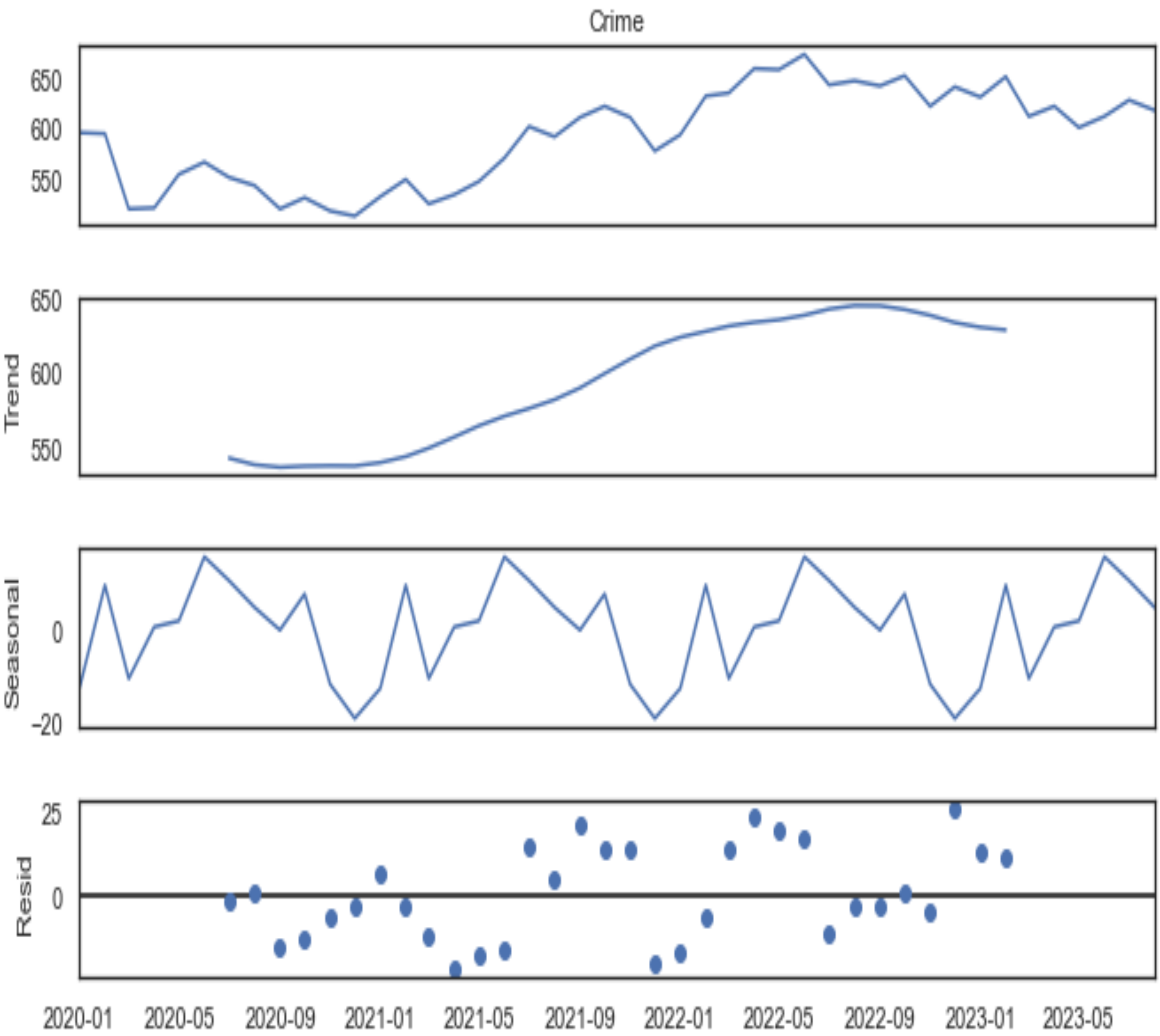

Figure 3.10 b) Trends, Seasonality, Residuals over the period

- The first subplot on the very top shows the plot for the original data with no decomposition.
- The second subplot shows a clear smooth trend pattern in the data. This is clear evidence of a non-constant mean.
- The third subplot shows the decomposed seasonality pattern in the data.
- The last subplot shows the noise or residual component in the time series data.

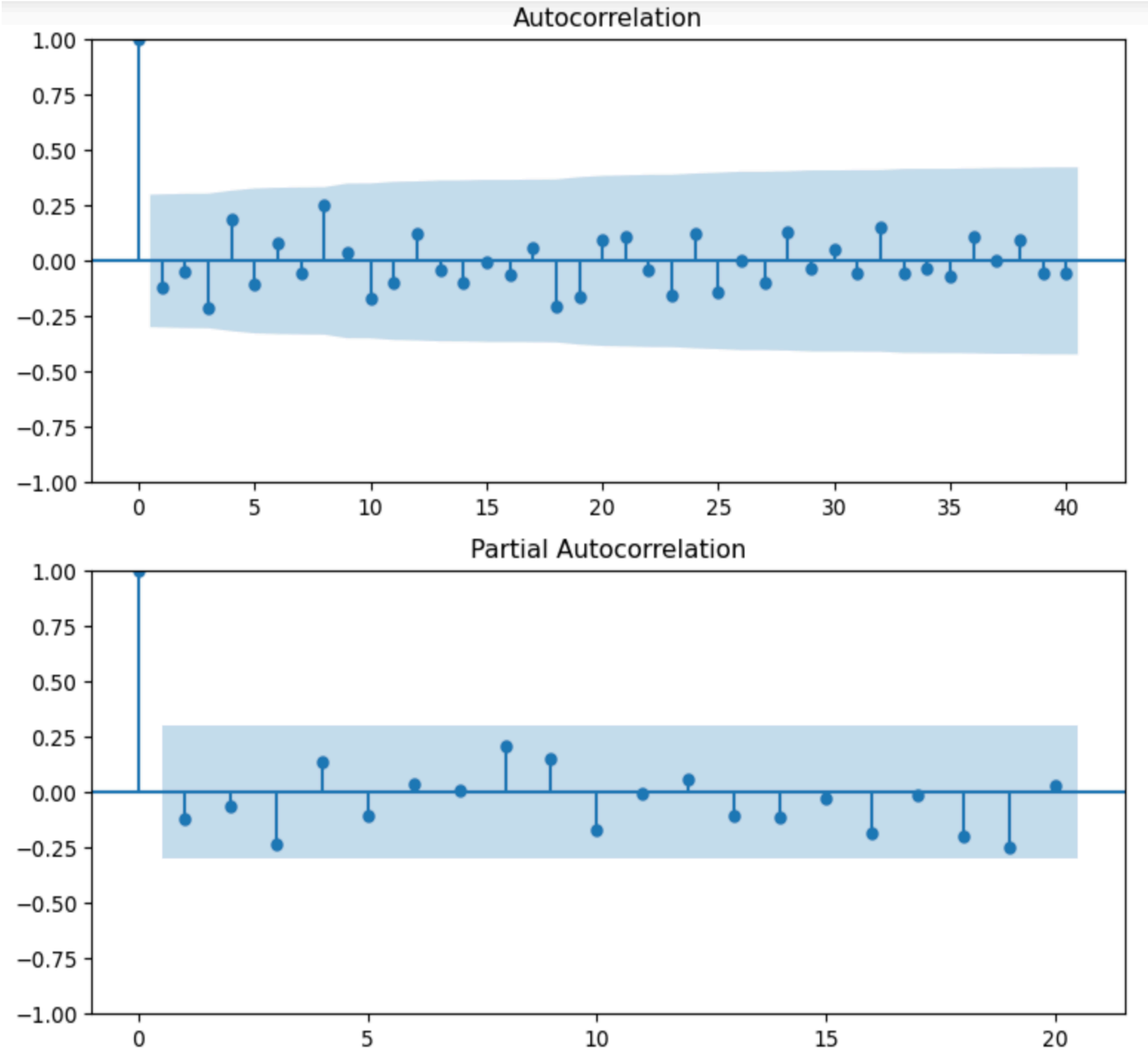

Figure 3.10 c) Autocorrelation function (ACF) plot and the Partial Autocorrelation Function (PACF)

We will look at the Autocorrelation function (ACF) plot and the Partial Autocorrelation Function (PACF) plots to get a sense of which lags are significant.

To model time series data, we need to achieve stationarity in the time series. Since, our data in the current state is not stationary, we will try achieving stationarity using first differencing.

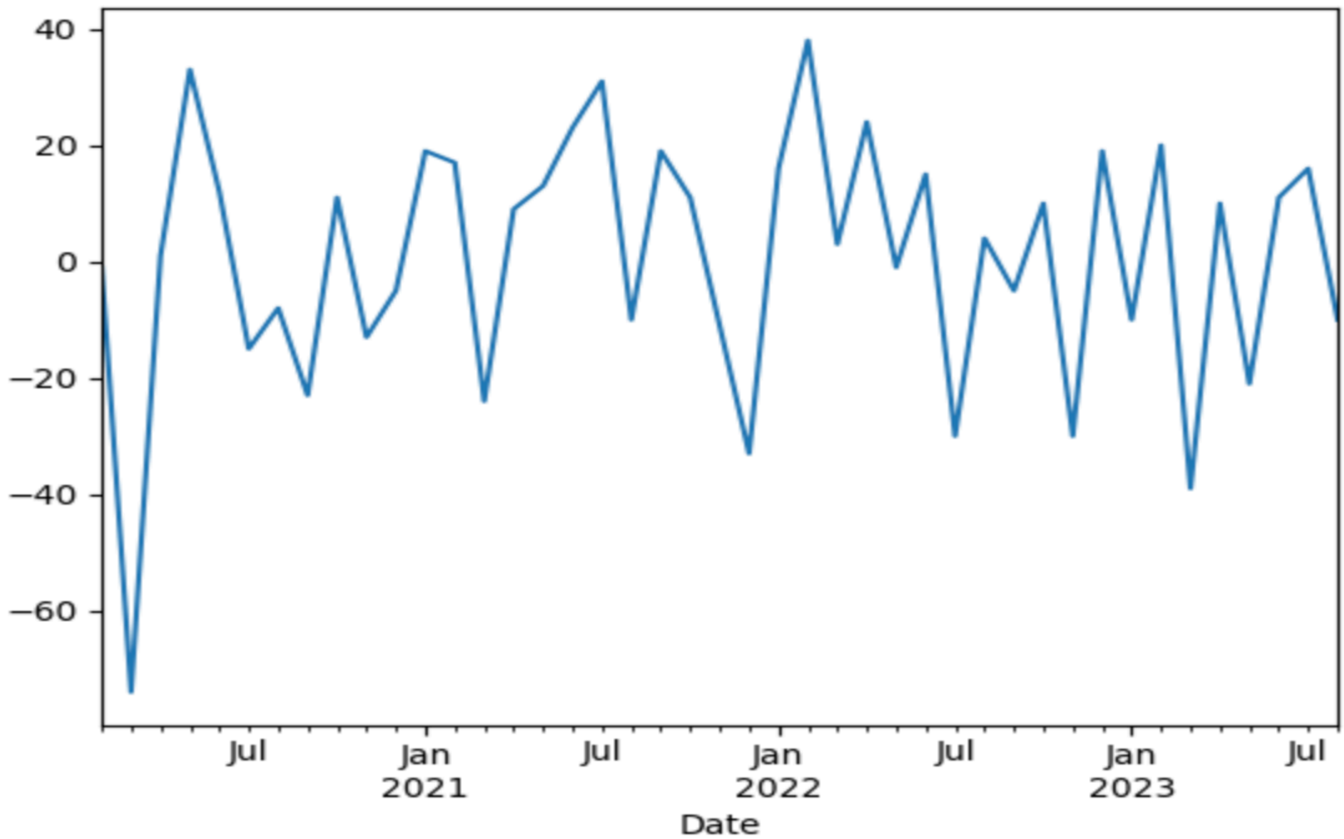

Figure 3.10 d) First Differencing

In the below forecasted graph using past years data , we can analyse that the forecast is in congruence with the expected results with very minor deflections in crimes .

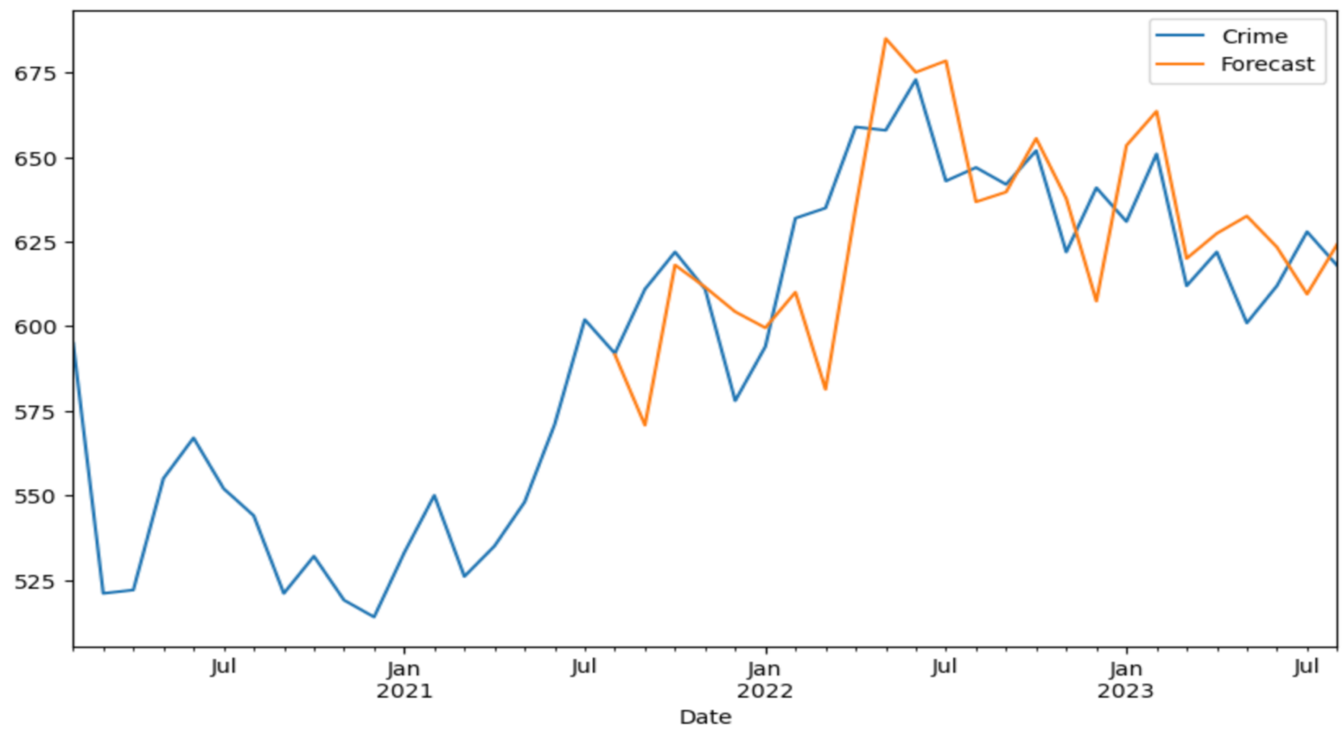

Figure 3.10 e) Back-testing

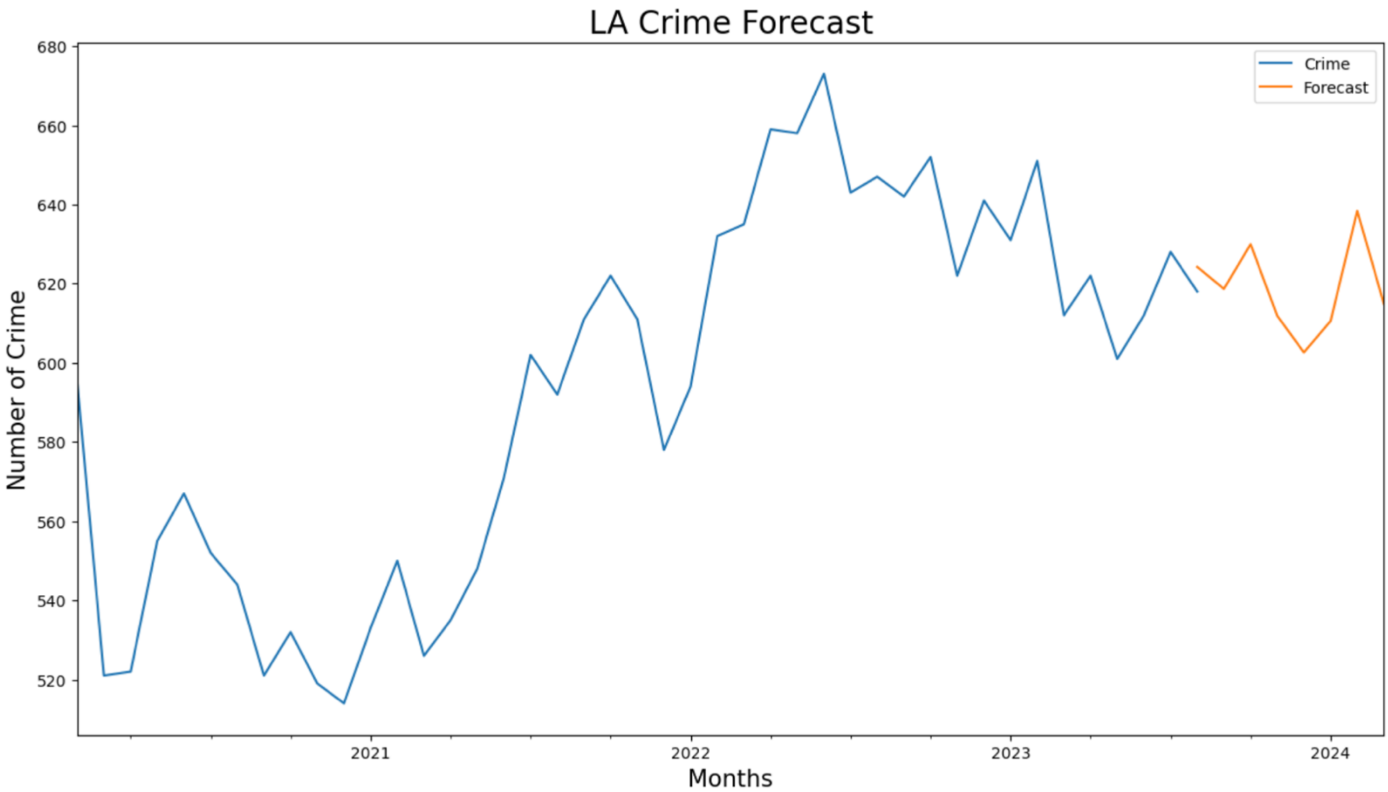

Figure 3.10 f) LA Crime Forecast

We have successfully forecasted the LA crime forecast by using SARIMAX time series algorithm upon the dataset post cleaning and manipulation . Moreover, we understood that the crime rate would have a slight decrease before increasing as it seems to be accumulation

Root Mean Squared E of the SARIMA Model :25.35

3.11 Prophet Time Series Forecasting:

A Prophet time series forecasting chart for LA crimes can provide insights into the expected future trends in crime rates. It can help identify patterns, seasonality, and potential changes in crime behaviour, enabling law enforcement and policymakers to make informed decisions and allocate resources more effectively to address crime-related challenges.

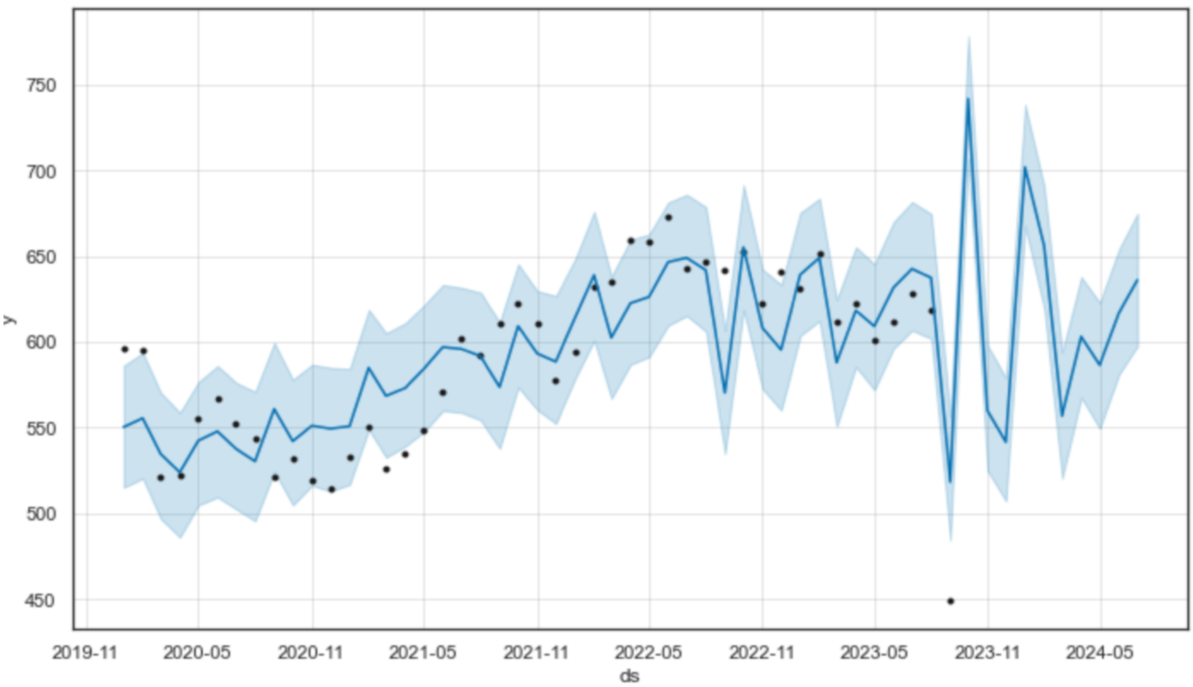

Figure 3.11 a) Forecasting using the prophet

	horizon	mse	rmse	mae	mape	mdape	smape	coverage
0	30 days	8.486287e+06	2913.123320	1516.066137	2.831829	0.235194	0.651283	0.000000
1	31 days	6.080031e+06	2465.771873	1245.287441	2.285625	0.171131	0.681073	0.050000
2	59 days	3.220122e+06	1794.469879	925.486938	1.712674	1.738926	0.607610	0.021429
3	60 days	2.714752e+06	1647.650589	975.375024	1.822528	1.738926	0.796028	0.014286
4	61 days	1.8109	click to scroll output; double click to hide		206488	0.141832	0.717658	0.000000
5	62 days	7.388651e+06	2718.207333	1386.647356	2.591603	0.111741	0.645540	0.000000
6	89 days	4.185244e+06	2045.786819	1117.883607	2.086396	0.111741	0.611477	0.000000
7	90 days	1.505619e+06	1227.036492	708.925513	1.318413	0.091759	0.654887	0.000000

Figure 3.11 b) Summary Statistics of Prophet

CHAPTER 4 : LIMITATIONS

The analysis has certain limitations that must be taken into account. To begin with, there may be issues with data completeness, as evidenced by the significant drop in crime rates in 2023. It's essential to recognize that this reduction could be due to a lack of comprehensive data for that year, rather than an actual decrease in criminal incidents.

Secondly, the accuracy and reliability of the crime data sources used in the analysis can have a significant impact on the credibility of the findings. Any discrepancies, errors, or biases in the data could potentially lead to incorrect conclusions.

Furthermore, while the analysis suggests that policy changes may influence crime rates, a more profound understanding of these changes and their effects would necessitate access to contextual information and legal documentation, which may not be available within the dataset.

Lastly, the classification and in-depth analysis of identity theft crimes may be constrained by the dataset's categorization of these crimes and its inherent limitations.

CHAPTER 5: FUTURE WORK

To overcome these limitations and further improve our comprehension of crime trends, there are numerous potential avenues for future research. First and foremost, it is essential to enhance the quality and completeness of the data, possibly by integrating data from multiple sources to ensure a more accurate portrayal of crime trends. Subsequent research could focus on a comprehensive analysis of policy changes, aiming to decipher their influence on crime rates and examining the consequences of law enforcement strategies.

A specialized investigation into identity theft, its techniques, and how these patterns change over time could offer invaluable insights for both law enforcement and cybersecurity efforts. Moreover, conducting a more detailed geographical analysis, including pinpointing crime hotspots at the neighbourhood level, could provide superior guidance for allocating law enforcement resources.

Developing predictive models to anticipate future crime trends could enable proactive law enforcement strategies. Further research into the demographics of crime victims and offenders might unveil underlying sociological factors contributing to crime rates. Time-series forecasting methods could be employed to predict future crime trends, which would be a valuable tool for law enforcement agencies in resource allocation and crime prevention.

Finally, an examination of the correlation between crime rates and various social and economic factors, such as unemployment rates and income levels, would offer a more holistic understanding of the drivers of criminal activity. By addressing these limitations and pursuing future research in these areas, researchers and law enforcement agencies can refine their understanding of crime patterns, enhance prevention strategies, and adapt to the changing landscape of criminal behaviour and societal shifts.

In conclusion, SARIMAX (Seasonal Autoregressive Integrated Moving Average) yielded better forecasting accuracy compared to the Prophet model.