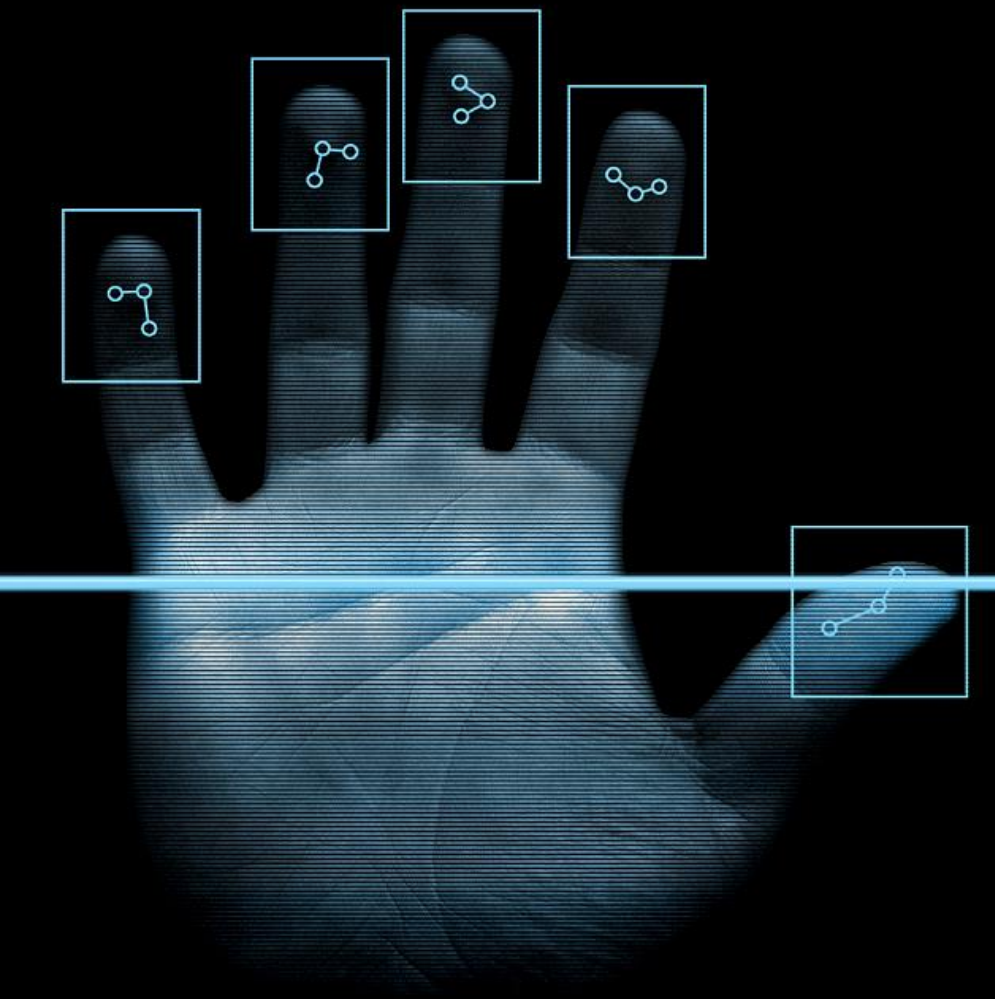


# Crime Data Cleaning and Analysis Report


| NUID | Name |
| --- | --- |
| 002242934 | Anjali Ingle |
| 002772623 | Avani Kala |
| 002291838 | Ameya Deshmukh |
| 002734199 | Wenzhe Zhang |
| 002922799 | Yangwenyu Peng |

# Table of Contents

## Executive Summary

This report delves into the analysis conducted on crime data of Los Angeles spanning from Jan 1, 2020, to Oct 16, 2023. The primary focus was on data cleaning, exploratory data analysis (EDA), and extracting meaningful insights regarding crime trends, patterns, and factors influencing crime rates. The report outlines the methodology, key findings, and conclusions derived from the analysis.

## Methodology

### 1. Data Cleaning and Preprocessing

The crime dataset in the specified CSV file was firstly loaded into a Pandas DataFrame, a structured data representation, making it amenable for subsequent processing phases and in-depth analytical examination.
Following the data inspection, we gained a preliminary understanding that this dataset consists of 815,882 rows and 28 columns, encompassing attributes with information on crime occurrences, victim demographics, crime types, weapons used, and geographic details.

Missing values were identified within a total of twelve columns within the dataset, namely: 'Mocodes,' 'Vict Sex,' 'Vict Descent,' 'Premis Cd,' 'Premis Desc,' 'Weapon Used Cd,' 'Weapon Desc,' 'Crm Cd 2,' 'Crm Cd 3,' 'Crm Cd 4,' and 'Cross Street.' As a response, a strategy was employed where we populated missing data points within pertinent columns with a designation of 'unknown' or 'undefined,' preserving the contextual relevance of the data.

To enhance the dataset's suitability for subsequent analytical procedures, a deliberate decision was also made to remove rows that exhibited missing values in the 'Premis Cd' and 'Crm Cd 1' columns, and to drop extraneous columns, namely 'Mocodes,' 'Crm Cd 2,' 'Crm Cd 3,' 'Crm Cd 4,' 'Cross Street,' and 'Weapon Used Cd,' which were deemed as not contributing significantly to the analytical objectives.

To ensure data integrity, the count of duplicate rows was checked and found none in the dataset.

In order to prepare data for future time-series analysis, the data in 'Date Rptd' and 'DATE OCC' columns was converted to datetime data type that contains

datetime values.

An outlier was detected while applying a box plot analysis on the variable 'Vict Age.' To address this issue, we constrained the range of 'Vict Age' within the limits defined by its minimum and maximum values.

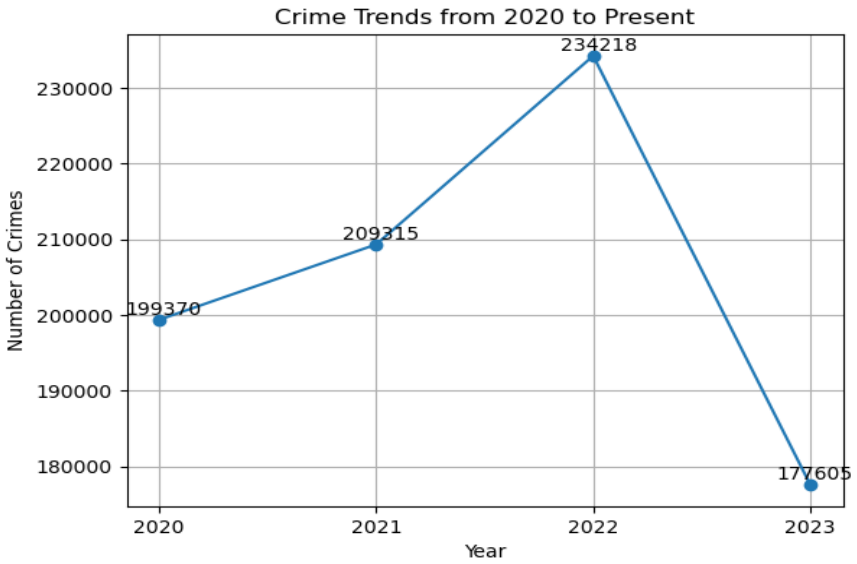
By applying LabelEncoder tool, text-based values in the categorical columns ‘'Vict Sex', 'Vict Descent', 'Status', 'Status Desc' were converted into numerical representations. This transformation serves to prepare the data for further analytical tasks that require numerical input, thus enhancing the data's compatibility with a wide range of algorithms and statistical methods.

## 2. Exploratory Data Analysis (EDA)

In the EDA phase, we conducted a comprehensive analysis of the dataset, addressing key questions and visualizing trends and patterns:

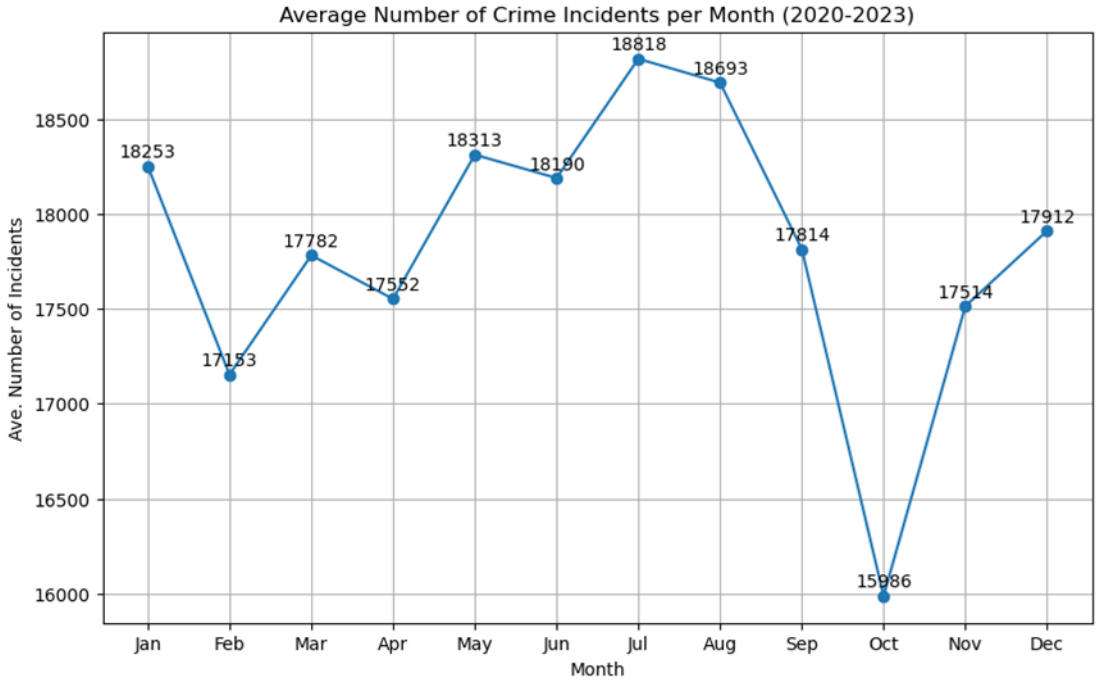
### 1) Overall crime trend

When we examine the overarching crime trends through visualization, it becomes evident that crime incidents saw an escalation starting in 2020, reaching a peak in 2022 with 234,218 cases, followed by a notable decline to 177,605 in 2023.

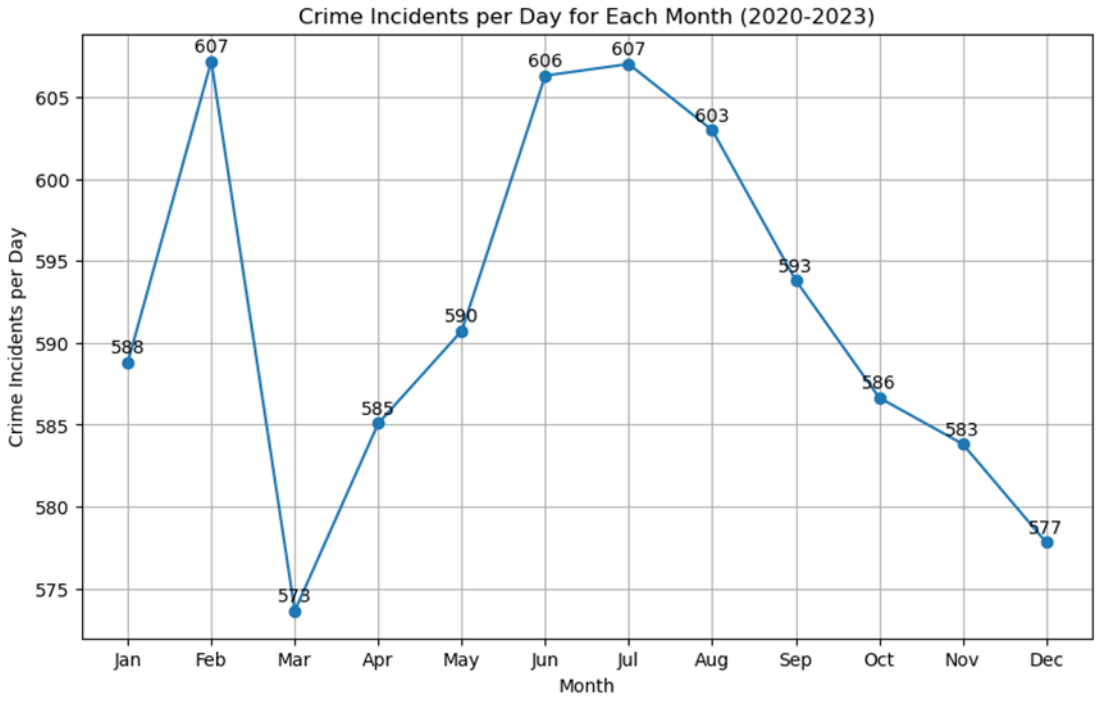


## 2) Seasonal patterns:

Upon analyzing the average number of crimes per month, spanning from 2020 to October 16, 2023, as illustrated in the line plot below, a discernible pattern emerges. Crime incidents peak during July, maintaining heightened levels throughout the summer months. Nonetheless, it is crucial to acknowledge the variations in the number of days within each month, as longer durations might potentially correlate with increased incidents.



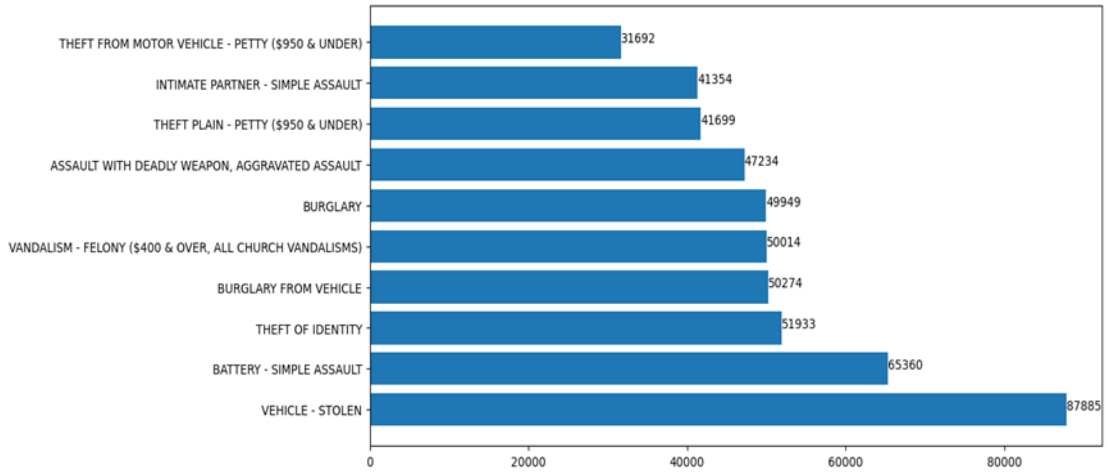

To appropriately adjust for these temporal differences, we computed the 'Crime Incidents per Day for Each Month' by dividing the sum of crime incidents per month by the respective days in each month. Special attention was given to 1) February 2020, characterized as a leap year with 29 days in that February; 2) October 2023, where 15 days of data are missing; 3) Missing data in November and December 2023.

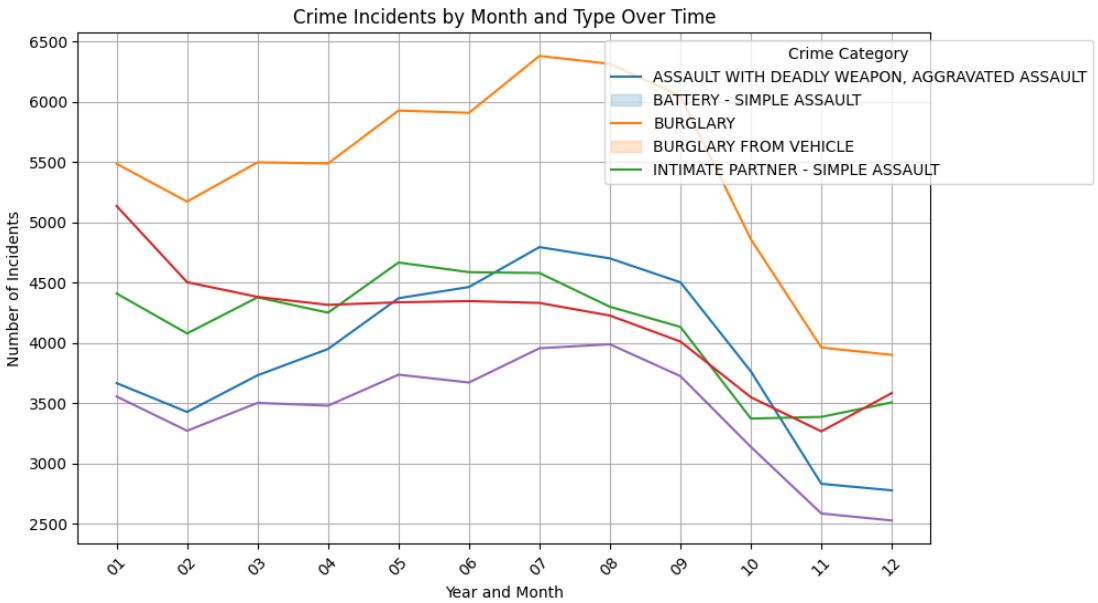


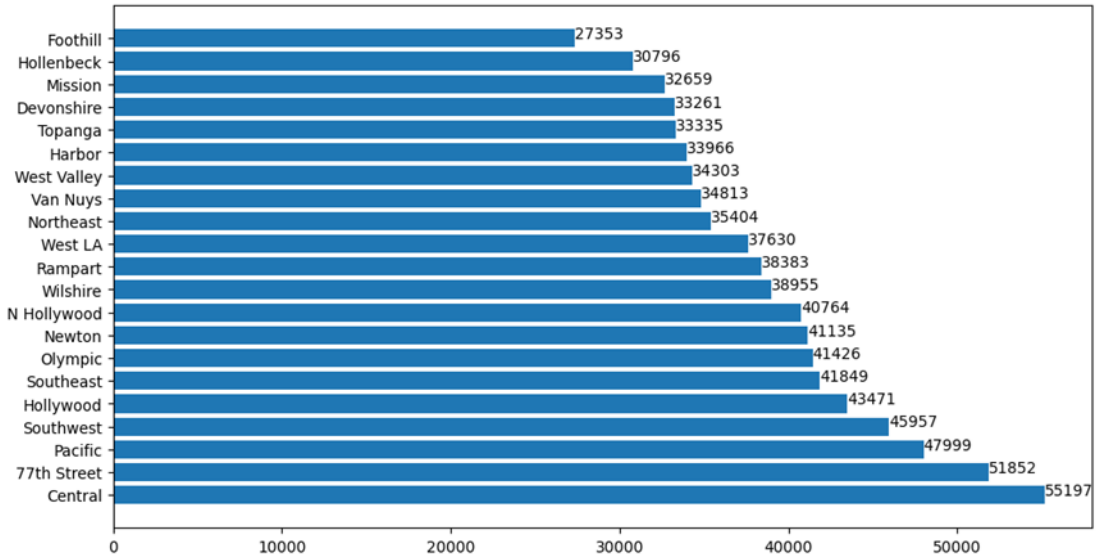

Consequently, the presented visualization above distinctly highlights February as the month with the most pronounced daily incident frequency, succeeded by a significant decline in March, marking the lowest point. Subsequently, the pattern reveals an increase in crime rates as temperatures rise from March to July. Conversely, with the decrease in temperatures and the onset of colder weather, there is a corresponding decline in crime rates from July to December.

## 3) Top 10 most prevalent crime types

When the data is categorized by different types of crimes, a distinct pattern becomes evident. The most prevalent type of crime is 'Vehicle Stolen,' consistently maintaining the highest frequency throughout the year, closely followed by 'Battery-Simple Assault' and 'Theft of Identity.' Remarkably, these crime categories exhibit similar seasonality patterns across various months.





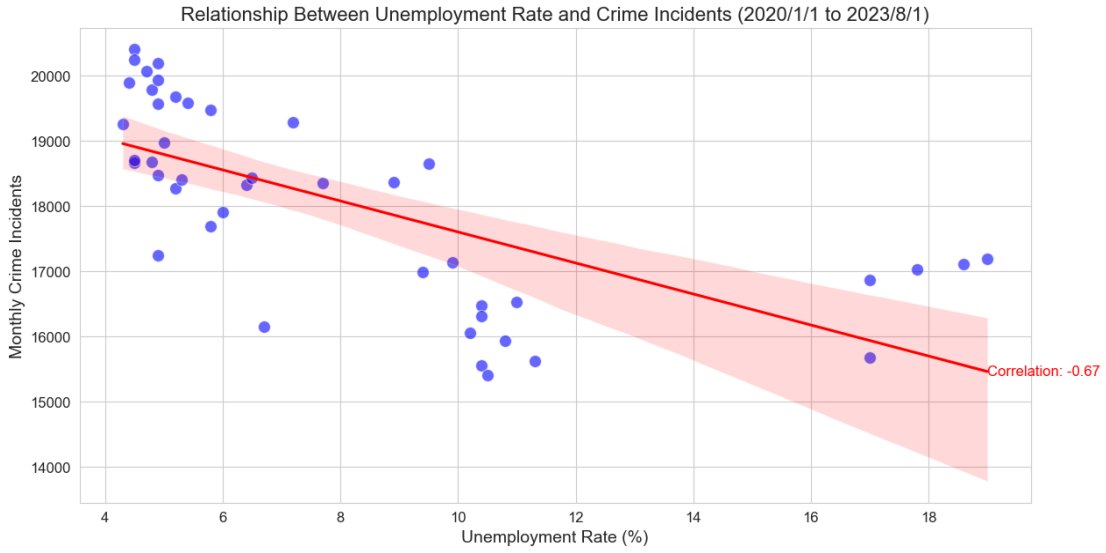

As seen in the graph above, the frequency of crime incidents is noticeably linked to specific geographic areas, with the 'Central' part experiencing a higher concentration of criminal activity, while 'Foothill' stands out as a relatively secure location, with the fewest reported incidents.

## 4) Explore correlations between economic factors and crime counts

### a) Unemployment rate

Unemployment is a condition in which individuals capable of participating in the workforce actively seek employment but are unable to secure jobs. This phenomenon carries significant economic importance as it serves as a critical indicator reflecting an economy's performance and holds substantial implications for societal well-being. Prior to our analysis, we anticipated a positive correlation between the unemployment rate and crime figures, driven by the idea that individuals facing unemployment or financial hardship may turn to criminal activities to alleviate their economic difficulties.

The dataset for the unemployment rate in Los Angeles County was sourced from the U.S. Bureau of Labor Statistics, spanning from January 1900 to August 2023, with data recorded monthly. To align the time frame of two datasets, we selected the unemployment rate dataset covering January 2020 to August 2023 and aggregated daily crime incidents into monthly records within this period.




An observed correlation coefficient of -0.67 indicates a moderate negative relationship between the unemployment rate and the frequency of monthly crime incidents. With p-value of 5.222700124903396e-07 determined by the Pearson test, we do not reject the null hypothesis and conclude that the correlation is statistically significant. In essence, as the unemployment rate increases, there is a corresponding reduction in the number of monthly crime incidents. This negative correlation may appear counterintuitive and challenges our initial assumptions. However, this outcome could be influenced by various other factors such as increased community surveillance, changes in routine and opportunities for crime, and other social factors that might reduce crime even as unemployment rises.

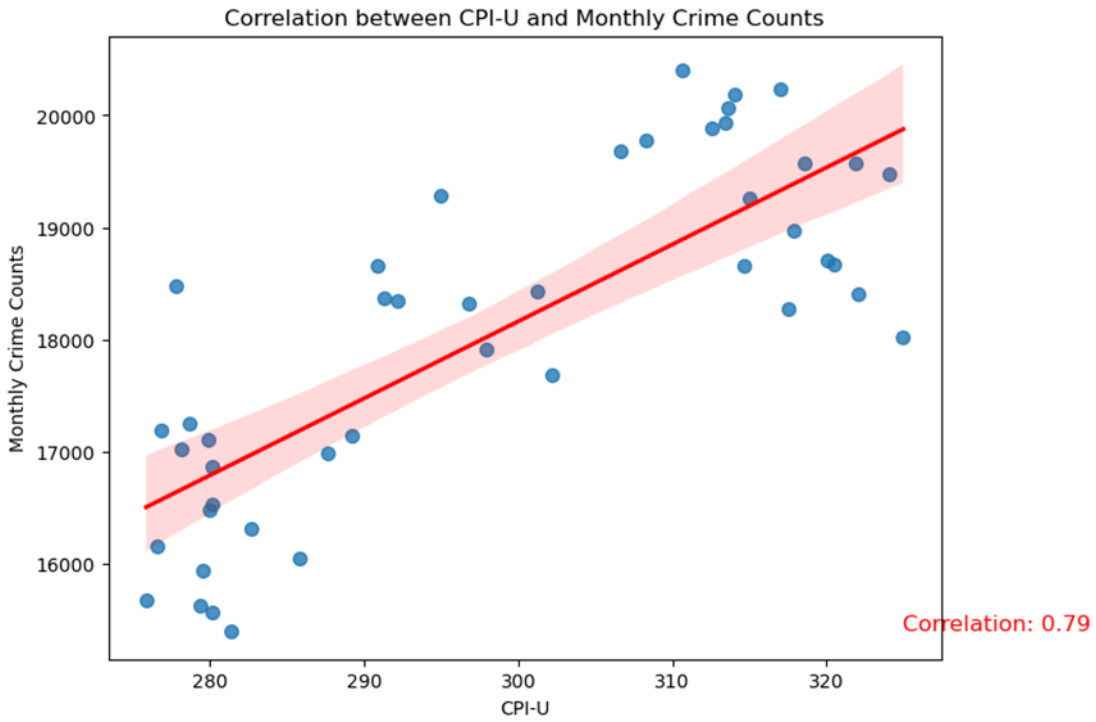
### b) CPI-U

The Consumer Price Index (CPI) serves as a gauge of the average fluctuation in the prices paid by urban consumers for a representative assortment of goods and services. It specifically measures the inflation experienced by consumers in their everyday expenses.

We accessed a dataset concerning the Consumer Price Index for All Urban Consumers (CPI-U) in the Los Angeles-Long Beach-Anaheim, CA area, available on the U.S. Bureau of Labor Statistics website. Opting for CPI-U over CPI-W (Consumer Price Index for Urban Wage Earners and Clerical Workers) was deliberate, as it encompasses approximately 93 percent of the total population, according to the U.S. Bureau of Labor Statistics. This choice was made to reflect the prevalent impact on the lives of people in LA.

Since the CPI-U dataset covers the period from January 2020 to September 2023 on a monthly basis, we aligned the crime dataset to end on September 30, 2023, matching it with the CPI-U dataset. To obtain the monthly crime incidents, we organized the data by year and month, aggregating the total count of crime incidents for each month within every year.

We proceeded to employ a scatter plot as a visual means to investigate the correlation between the two variables:




From the graph, a noticeable positive correlation is observed, and we further conducted Pearson test. The correlation coefficient of 0.79 suggests a robust positive linear relationship between the variables. The P-value, which is significantly lower than the commonly accepted threshold of 0.05, is approximately 7.048587960250187e-11. This supports the existence of a strong and significant correlation.

### c) Inflation Rate

We extended our investigation to examine the inflation rate, which is derived from the CPI. The monthly inflation rate is determined by examining the percentage change in the Consumer Price Index (CPI) over a single month. It provides insights into whether the cost of goods and services increases or decreases during that time frame. Given the similar nature to the CPI, we expected a positive correlation to exist between the inflation rate and the monthly crime numbers.

To compute the monthly inflation rate, we employed the following formula:

$$Inflation\ Rate = \frac{(\text{CPI of the current month} - \text{CPI of the previous month})}{\text{CPI of the previous month}} \times 100$$

To address the missing value in January 2020, we utilized CPI-U data in December 2019 sourced from the US Bureau of Labor Statistics.


Upon conducting a correlation analysis between the monthly inflation rate and crime counts, we observed a weak positive correlation of 0.27. However, it's crucial to note that the P-value (0.068) in the Pearson Test exceeds the significance level of 0.05. Consequently, we lack substantial evidence to support the presence of a statistically significant correlation between monthly crime counts and the inflation rate within the provided dataset.

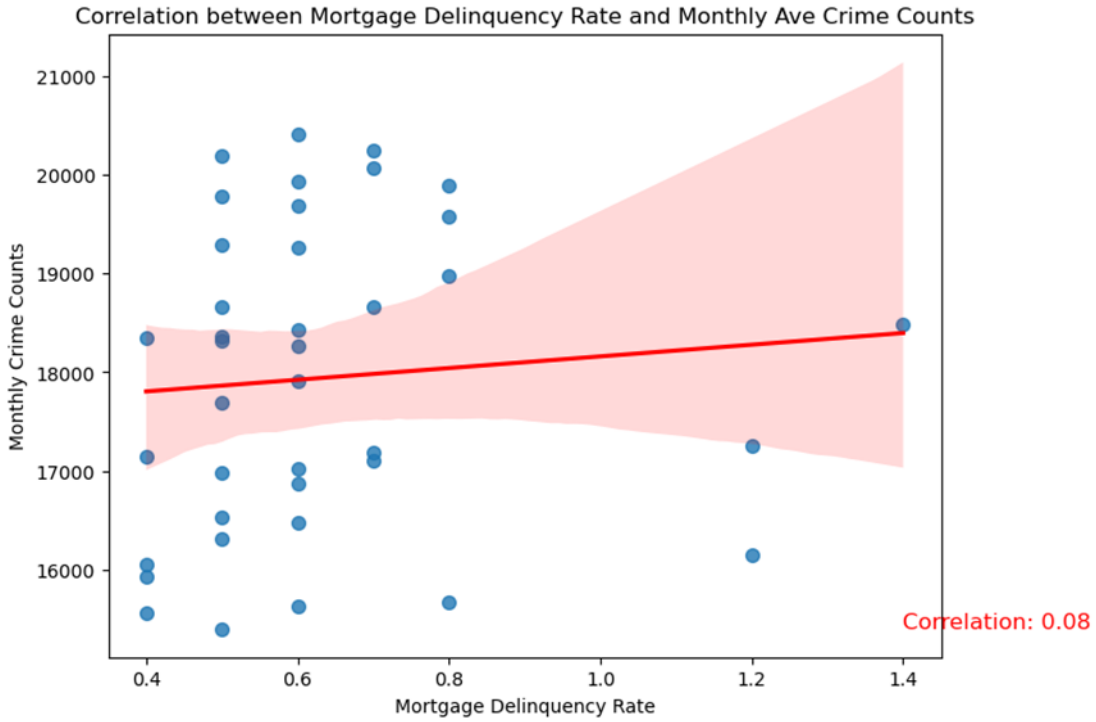
### d) Mortgage Delinquency Rate

We retrieved the dataset of 30-89 mortgage delinquency rates in LA from the Consumer Financial Protection Bureau. This rate serves as a metric for early-stage delinquencies and holds significance within the housing and financial sectors. It offers insights into potential payment issues among borrowers in the early stages of mortgage delinquency. Prior to correlation analysis, we assumed that higher mortgage delinquency rates could indicate financial stress within a community.

Economic difficulties, such as housing payment issues, might lead to higher stress levels among individuals, potentially contributing to an increase in certain types of crimes.

The data for the mortgage delinquency rate spans from January 2020 to March 2023. To ensure a consistent time frame, we adjusted the crime dataset to conclude on March 31, 2023.

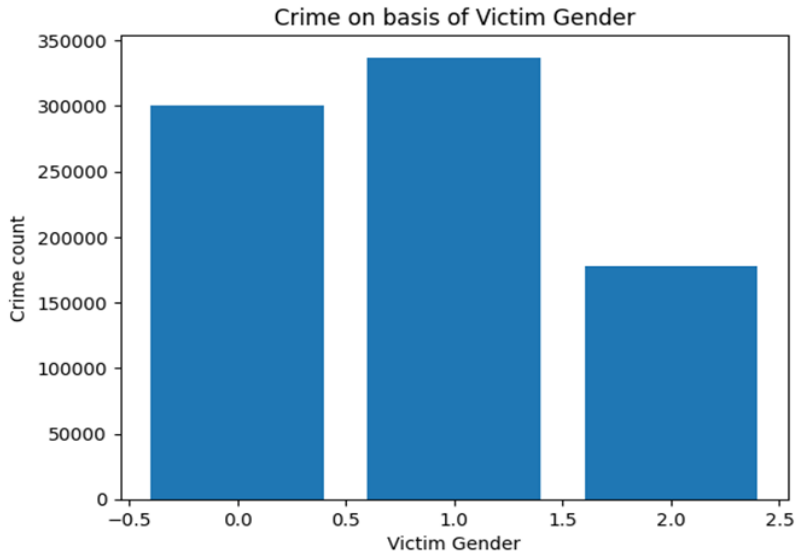
Subsequently, a correlation analysis was carried out between the monthly average crime counts and the mortgage delinquency rate in LA.




The nearly horizontal regression line in the scatter plot indicates an exceedingly weak or negligible linear relationship, which is supported by the correlation coefficient of 0.08 obtained from the Pearson test. Furthermore, the associated p-value was computed to be 0.609, surpassing the conventional significance level of 0.05. This indicates a failure to reject the null hypothesis. In essence, the data does not substantiate a statistically significant correlation between the variables under examination.

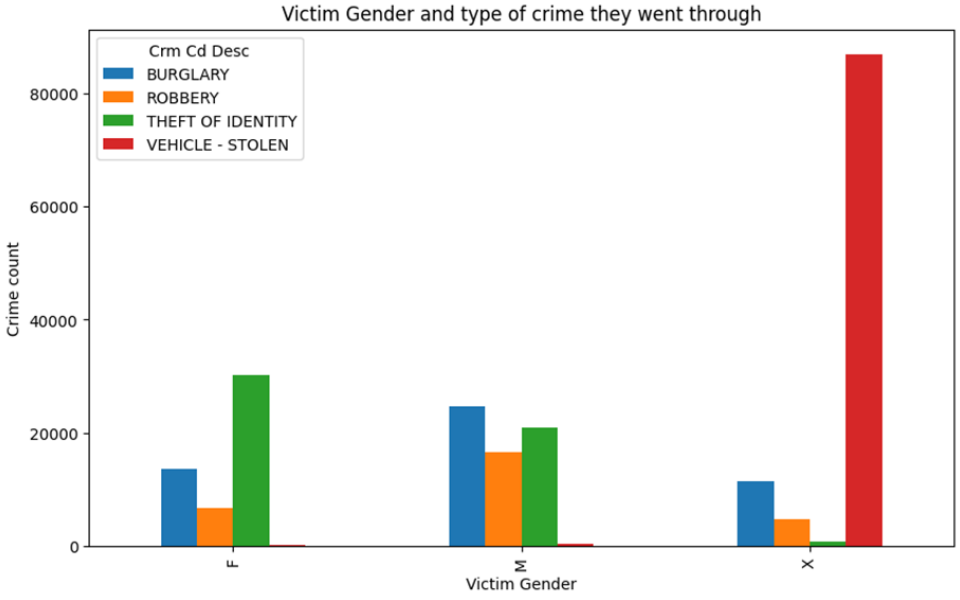
## 5) Demographic factors with specific types of crimes

### a) Crime on basis of Gender




The above chart shows that the Male Victims are slightly more prone to crime than the female. Nevertheless, the presence of approximately 20% of cases categorized as having an unknown gender (X) represents a notable proportion. This substantial percentage has the potential to introduce a degree of distortion in the analysis of the data.

### b) the gender of the victim and the category of crime they experienced



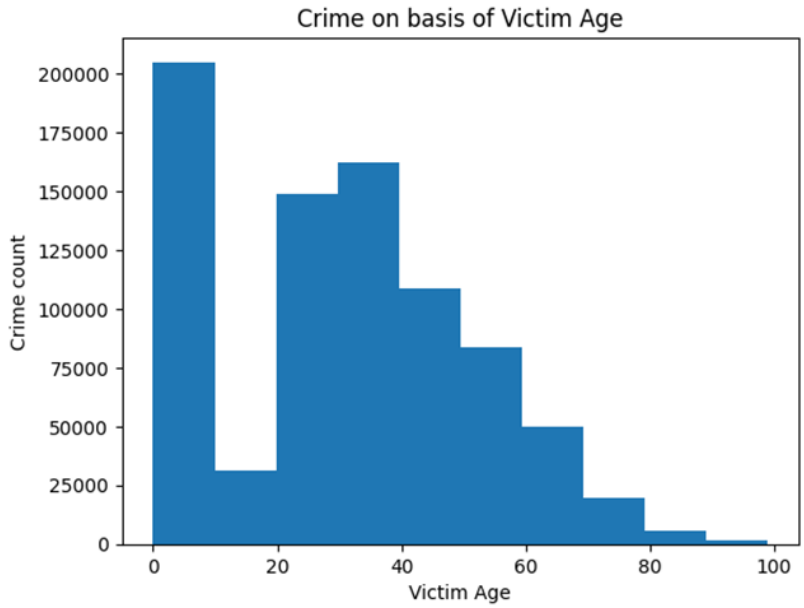
The primary observations derived from the above bar chart are as follows:

i. The data indicates that females exhibit a higher susceptibility to identity theft.

ii. Conversely, males demonstrate a greater vulnerability to burglary according to the provided information.

iii. The unidentified category predominantly comprises victims of vehicle theft, suggesting challenges in determining the gender of the owners of stolen vehicles.

### c) Crime on basis of Victim Age




There are a lot of victims whose age is unknown, which are filled with 0, while the predominant demographic of victims falls within the age range of 20 to 40 years.

### d) Victim Age range and type of crime they went through

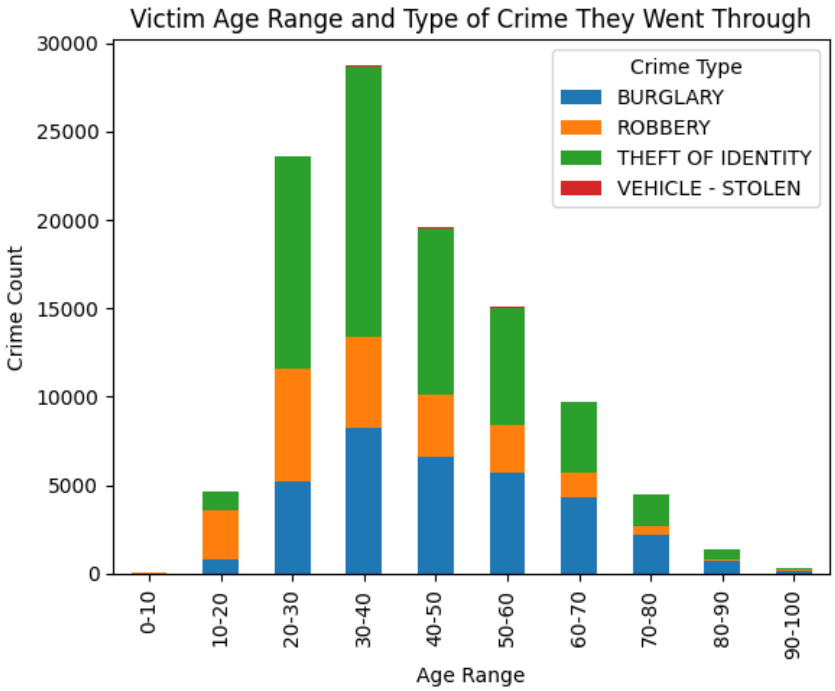
From the graph below, we observe that:

i. Victims in the age range of 30 to 40 years mainly went through "Theft of Identity" crimes.

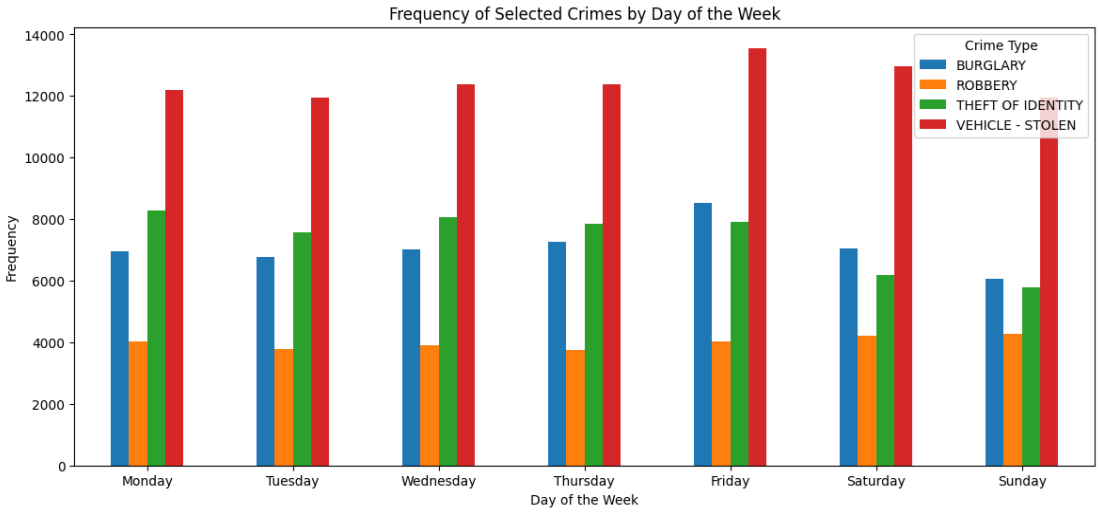
ii. Victims in the age range of 20 to 30 years suffered from "Robbery" crimes.

iii. "Vehicle Stolen" crime type is most prevalent among the 0-10 years age

group, which suggests that these are possibly cases of vehicle theft where the age of the victim is not known (marked as 0 in the dataset).



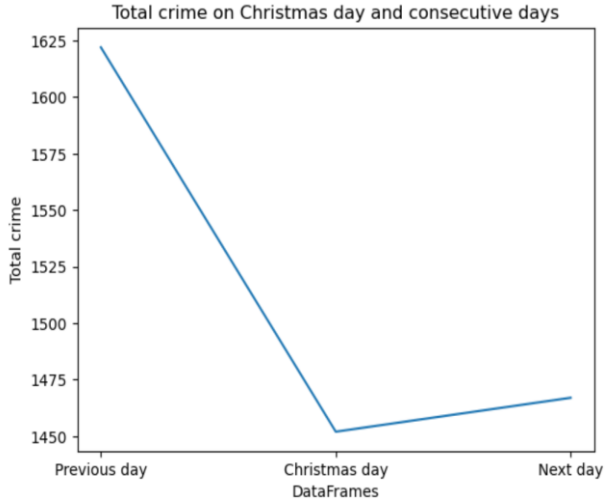

## 6) The relationship between the day of the week and the frequency of certain crime type



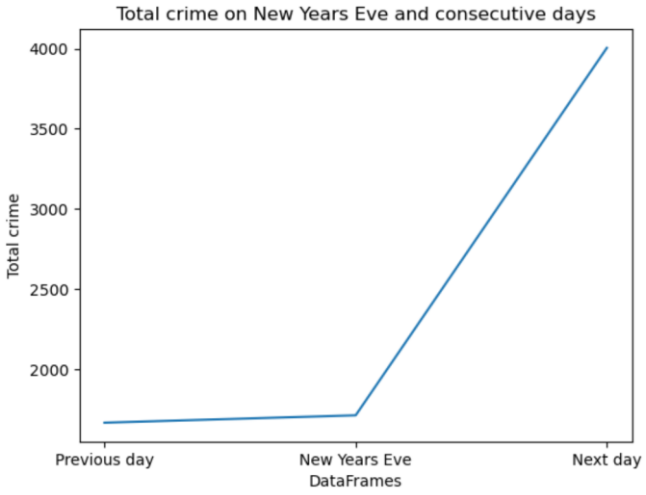

Irrespective of the day of the week, various crime types exhibit a consistent frequency, suggesting the absence of any discernible relationship between the day of the week and the type of crime.

## 7) Impact of major events or policy changes on crime rates

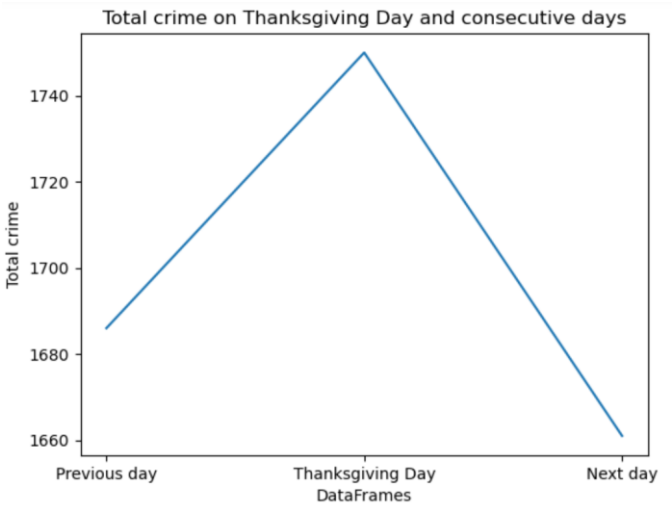
For Major Events like Christmas, we can observe the crime rates drop significantly on the day of Christmas every year.



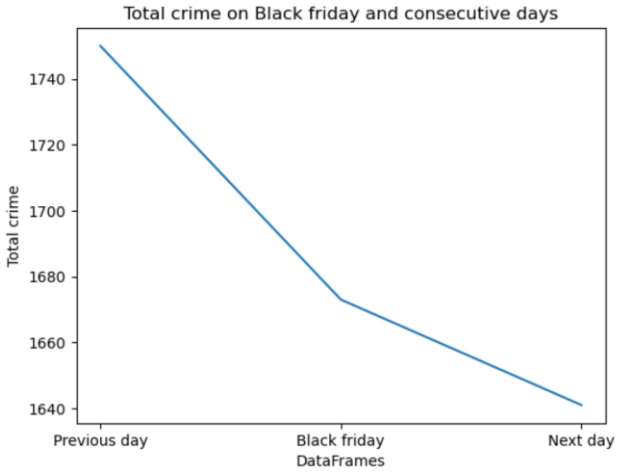

On the contrary , for New Years we can see a spike on the next day, indicating an increase in crime rate every start of the year.




Similarly on Thanksgiving Day, we observe a spike in the crime rates for that day. The crime rates got lower down for the next day.




For Black Friday, we can observe a noticeable descent in the crime rates for the next day as well.




From the above graphs, we can observe a spike in crime on the following days:

1) In the month of January, we can observe a descent in crimes after the 3rd of January. There's a slight spike on the 15th of January.

2) In February, we can observe a slight descent in crimes around the 4th of the month followed by another one around the 27th of the month. The crime rates are average mid-month.

3) In March, we can observe a significant descent in crimes around the 4th of the month followed by average crime rates throughout the month.

4) In April, we can observe a significant descent in crimes around the 6th of the

month followed by a spike on the 10th of April. The crime rates look average for the rest of the month.

5) In May, we can observe a huge descent in crimes around the 4th of the month followed by average crime rates mid-month. There's a spike observed on the 30th of May.

6) In June, we can observe a significant descent in crimes around the 4th of the month followed by multiple spikes from the 10th to the 27th of June.

7) In July, we can observe a significant descent in crimes around the 4th of the month followed by multiple spikes between the 6th to 24th of July.

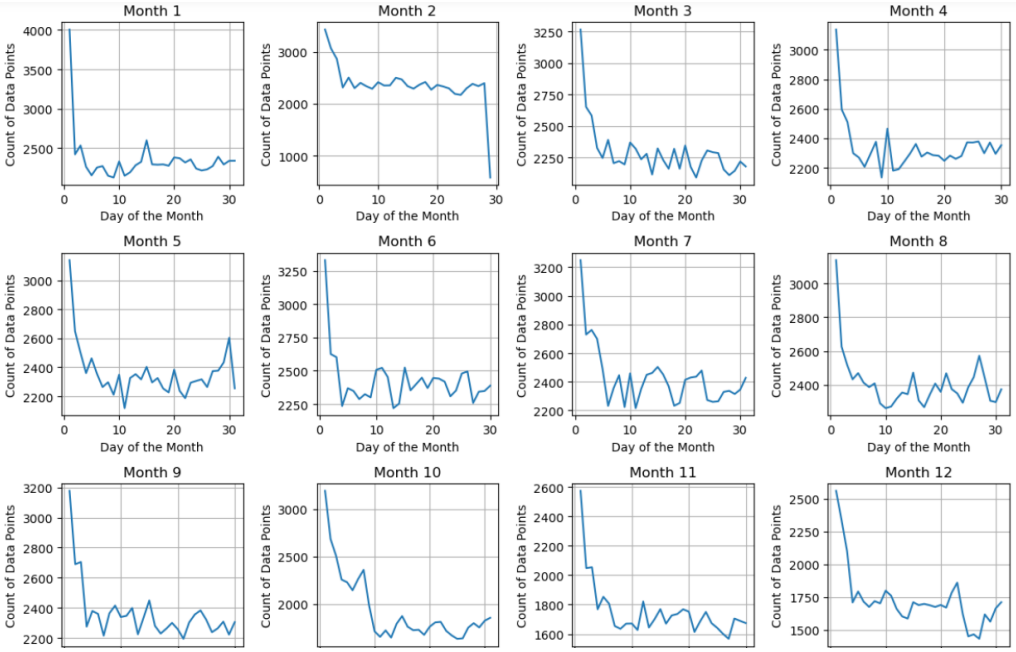
8) In August, we can observe a significant descent in crimes around the 10th of the month followed by average crime rates throughout the month and a spike around the 27th of July.

9) In September, we can observe a huge descent in crimes around the 4th of the month followed by average crime rates throughout the month and multiple spikes on the 7th, 15th, and 23rd of September.

10) In the month of October, we can observe a significant descent in crimes around the 6th of the month followed by a spike around the 8th of October. The crime rates are average for the rest of the month.

11) In November, we can observe a significant descent in crimes around the 3rd of the month followed by average crime rates throughout the month with a few spikes around 12th, 15th , 20th, and 23rd of November.

12) In December, we can observe a significant descent in crimes around the 4th of the month followed by average crime rate rates mid month. There's a noticeable spike around 23rd of December.

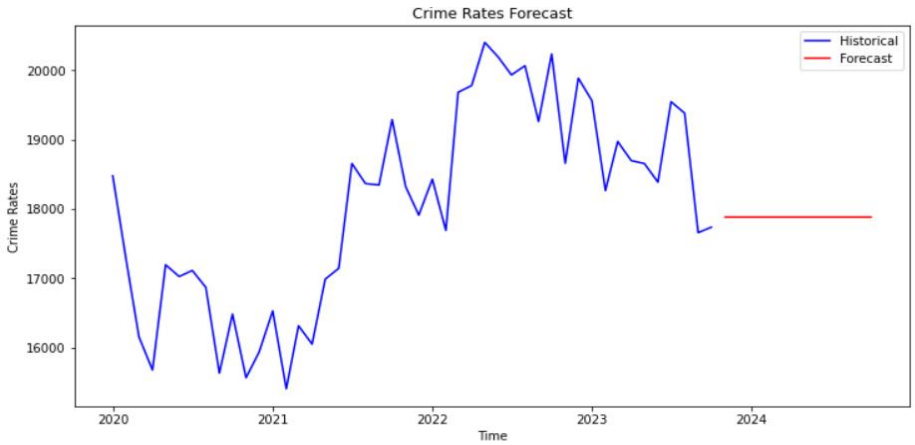

Month 1
Month 2
Month 3
Month 4
Month 5
Month 6
Month 7
Month 8
Month 9
Month 10
Month 11
Month 12
Count of Data Points
Day of the Month

## 3. Time Series Analysis and Prediction

Models:

1. LSTM Model: We initially considered using the Long Short-Term Memory (LSTM) model due to its recurrent neural network architecture, which excels in capturing long-term dependencies within data sequences. LSTM is well-suited for tasks involving sequence prediction. However, the LSTM model took a considerable amount of time to forecast future crime rates. It predicted the crime rate for the date immediately following the last available data point, which might not be practical for some forecasting needs.

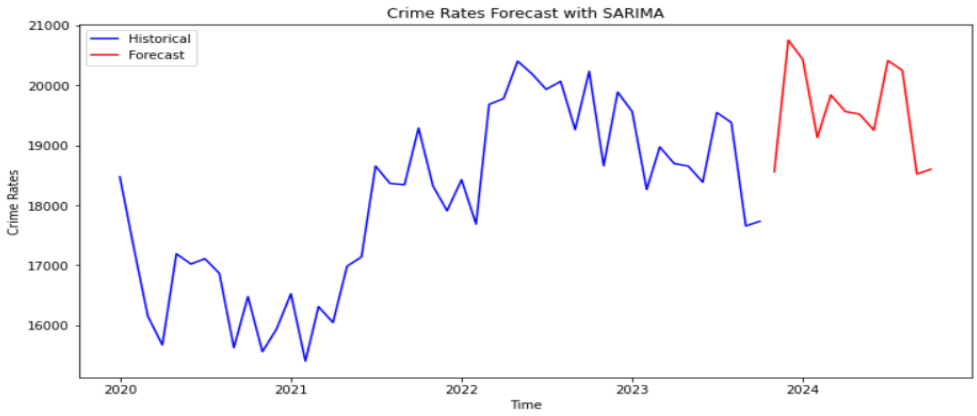
2. ARIMA Model: As an alternative, we explored the ARIMA model. ARIMA is a classical time series forecasting method known for its simplicity and effectiveness. It involves differencing to make the time series stationary and modeling autoregressive and moving average components. However, the ARIMA model did not yield favorable results, as indicated by higher AIC values. A lower AIC is preferred for a better-fitting model.




AIC: 732.328

3. SARIMA Model: Finally, we employed the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model. SARIMA extends the ARIMA model to handle seasonality in the data. The SARIMA model provided a forecast for crime data. The best SARIMA model identified through stepwise search minimized the AIC, indicating a good fit to the data. The specific model parameters are ARIMA(0,0,1)(0,1,0)[12] with an

intercept term.
AIC: 604.369




In summary, the SARIMA model (ARIMA with seasonality) provides the most favorable results, offering a practical forecast for crime data. While the ARIMA and LSTM models were considered, the SARIMA model demonstrated its effectiveness in capturing the seasonal patterns in the data, resulting in a lower AIC value and better model fit.

## Conclusion

The comprehensive analysis of crime data in Los Angeles from 2020 to October 2023 provided invaluable insights into the dynamics, trends, and influencing factors of criminal activities in the region. The methodology involved thorough data cleaning, exploratory data analysis (EDA), and the application of various statistical techniques to derive meaningful conclusions.

The data cleaning process ensured the dataset's integrity by addressing missing values and outliers, enhancing its suitability for analysis. Exploratory data analysis revealed intriguing trends such as the escalation in crime incidents from 2020 to a peak in 2022, followed by a decline in 2023.

Seasonal patterns unveiled fluctuations in crime rates corresponding to specific months, potentially linked to temporal variations and weather changes. Moreover, the analysis showcased distinct crime types prevalent in various geographical areas, shedding light on the concentrations and diversities of criminal activities.

Examining the correlation between economic factors and crime counts revealed unexpected relationships. While a negative correlation between unemployment rates and crime incidents challenged assumptions, a strong positive correlation was found between the Consumer Price Index and crime rates. However, the inflation rate and mortgage delinquency exhibited weak or negligible correlations with crime figures.

Demographic factors such as gender and age were explored in relation to specific crime categories. While males showed a higher susceptibility to certain crimes, the presence of unknown genders posed challenges in interpreting complete demographic patterns.

Further analysis explored the relationship between the day of the week and certain crime types, showcasing consistent frequencies across different days, suggesting a lack of discernible patterns in this regard.

Additionally, major events and policy changes displayed interesting correlations with crime rates, indicating fluctuations before, during, and after significant dates, highlighting the impact of specific occasions on criminal activities.

The Time Series Analysis and Prediction segment explored various models to forecast future crime rates. The SARIMA model emerged as the most effective, considering seasonal patterns, delivering a lower AIC value, and demonstrating a better fit to the data compared to other models such as ARIMA and LSTM.

In conclusion, the analysis of crime data in Los Angeles has provided substantial insights into the complex nature of criminal activities. The correlations discovered between economic indicators and crime rates, along with the influence of various demographic and temporal factors, underscore the multifaceted nature of crime dynamics. The findings and forecasting models established in this report can serve as a foundation for future policy-making, law enforcement strategies, and further in-depth studies aimed at crime prevention and societal well-being.

## Appendices

Federal Reserve Bank of St. Louis. (n.d.). Unemployment Rate in California [CALOSA7URN]. FRED Economic Data.

https://fred.stlouisfed.org/series/CALOSA7URN

U.S. Bureau of Labor Statistics. (n.d.). Consumer Price Index - Data. https://www.bls.gov/cpi/data.htm

Consumer Financial Protection Bureau. (n.d.). Mortgages 30-89 days delinquent. Consumer Financial Protection Bureau. https://www.consumerfinance.gov/data-research/mortgage-performance-trends/mortgages-30-89-days-delinquent/