Introduction:

Understanding crime data is essential for ensuring the safety and security of a city's residents. Many individuals choose to move to a new city with safety as their top priority. In this analysis, we aim to investigate how various factors influence crime rates in the city of Los Angeles from 2020 to the current date.

Our objective is to explore questions such as how demographic characteristics, economic factors, and Major Events impact crime rates in different regions. By delving into the intricate interplay between these factors and crime rates, we seek to provide valuable insights which may be used for enhancing public safety measures and promoting informed decision-making for individuals considering relocation to Los Angeles.

In this project, we perform Exploratory Data Analysis (EDA), which is akin to taking an initial glimpse at the data. EDA serves to acquaint us with a new dataset and unveil intriguing patterns or trends that lie within. Our primary goal is to gain familiarity with the dataset's features, identify any missing values, and compute essential statistical metrics. Through this preliminary exploration, we aim to uncover any initial indicators or associations that might guide our subsequent analyses and contribute to a comprehensive understanding of crime dynamics in Los Angeles.

Here we have used Python libraries like pandas, Seaborn, Matplotlib, Numpy, Scipy, Sklearn for cleaning the datasets and used the same for visualization of the data to draw conclusions from. Jupyter Notebook is used as a platform to conduct EDA. We performed several visualizations by analyzing the data for several questions to see their correlation and if there are any dependencies to identify several factors influencing crime rates.

There are three steps for doing this project, those are: Data Collecting, Data Processing, Statistical Analysis & Exploratory data analysis through visualization and correlation analysis.

```
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   DR_NO           807377 non-null  int64
 1   Date Rptd       807377 non-null  object
 2   DATE OCC        807377 non-null  object
 3   TIME OCC        807377 non-null  int64
 4   AREA            807377 non-null  int64
 5   AREA NAME       807377 non-null  object
 6   Rpt Dist No     807377 non-null  int64
 7   Part 1-2        807377 non-null  int64
 8   Crm Cd          807377 non-null  int64
 9   Crm Cd Desc     807377 non-null  object
 10  Mocodes         696010 non-null  object
 11  Vict Age        807377 non-null  int64
 12  Vict Sex        701468 non-null  object
 13  Vict Descent    701460 non-null  object
 14  Premis Cd       807368 non-null  float64
 15  Premis Desc     806901 non-null  object
 16  Weapon Used Cd  281174 non-null  float64
 17  Weapon Desc     281174 non-null  object
 18  Status          807377 non-null  object
 19  Status Desc     807377 non-null  object
 20  Crm Cd 1        807367 non-null  float64
 21  Crm Cd 2        59483 non-null   float64
 22  Crm Cd 3        1987 non-null    float64
 23  Crm Cd 4        58 non-null      float64
 24  LOCATION        807377 non-null  object
 25  Cross Street    129232 non-null  object
 26  LAT             807377 non-null  float64
 27  LON             807377 non-null  float64
dtypes: float64(8), int64(7), object(13)
```

Data Acquisition and processing:

Data is obtained from the crime dataset provided in the official website of the United States Government DATA.GOV and this data is downloaded into Jupyter Notebook which is the platform used for Python. The data required for getting correlation with crime rates and Economic factors was generated hypnotically to perform data analysis.

Data Inspection: The data which is loaded has a total of 807377 rows and 28 columns with the data types of float, int and objects. Here by getting a look at the dataset we can see that there were many null values in some cells which are to be addressed for maintain the data quality.

Data Cleaning: This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy. So, for maintain the data quality some columns are dropped which are not relevant for the analysis and has more than 80% of the data missing. As for the dealing with null values we have filled the null values of object columns with their mode and numeric columns with their mean to perform near to accurate analysis.

After this we have eliminated the duplicate rows to get manageable data. Another task in data processing is to identify inconsistent data, so for this data there can be inconsistence in the date and time writing styles as they were initially recorded on a paper and then uploaded into the database. Therefore, inconsistent data were replaced into a common format using python.

DR_NO	int64
Date Rptd	datetime64[ns]
DATE OCC	datetime64[ns]
TIME OCC	object
AREA	int64
AREA NAME	object
Rpt Dist No	int64
Part 1-2	int64
Crm Cd	int64
Crm Cd Desc	object
Mocodes	object
Vict Age	int64
Vict Sex	object
Vict Descent	object
Premis Cd	int64
Premis Desc	object
Weapon Used Cd	int64
Weapon Desc	object
Status	object
Status Desc	object
Crm Cd 1	int64
LOCATION	object
LAT	float64
LON	float64

Normalizing Numerical data: To ensure uniformity and comparability in the data, a technique known as min-max normalization was employed. This approach facilitates the scaling of values in a specific column to fit within the range of 0 to 1. The 'norm' function defined in the code accomplishes this by first calculating the range of values in the column (maximum minus minimum) and then applying a transformation that shifts the minimum value to 0 and the maximum value to 1. Before the normalization, the Shapiro-Wilk normality test was conducted on each column to check the distribution's normality. If the p-value from this test was less than the standard significance level of 0.05, indicating non-normality, the 'norm' function was applied to the column, ensuring that the data conforms to a standardized scale for further analysis and interpretation

Encoding Data: To enable the utilization of categorical data for analysis, a process known as label encoding was implemented. This technique involves converting categorical values into numerical representations. Each unique category within these columns was assigned a unique numerical label, allowing for easier computation and analysis. The Label Encoder function facilitated this process, transforming categorical data into a format that can be readily interpreted and processed by various machine learning models and statistical algorithms.

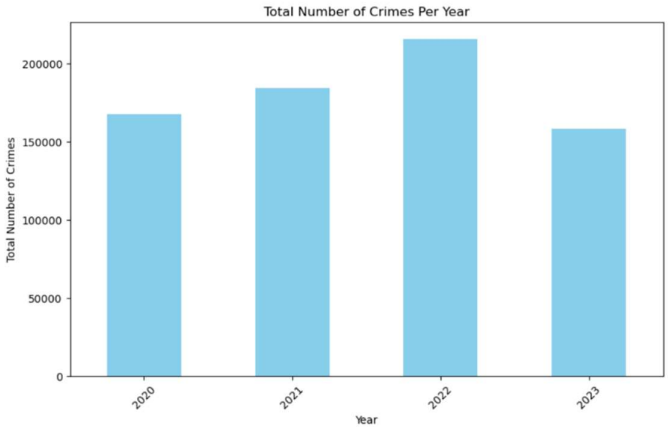
Exploratory Data Analysis Through Data Visualization:

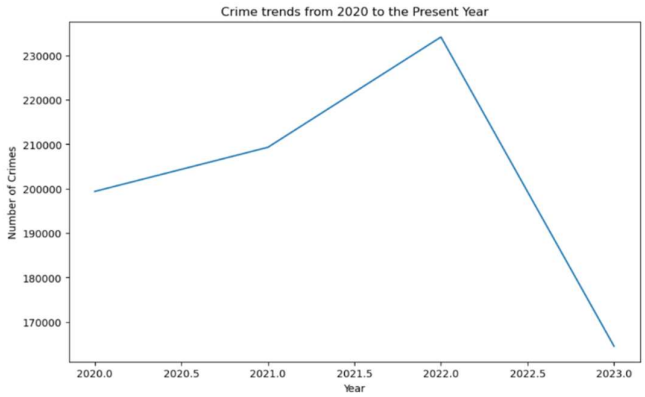
For the project of Analyzing the Crime Data we have calculated and plotted several trends to find any correlations and to find what factors affect the crime in the Los Angeles, like what are the common types of crimes, effect of major events and demographic factors.

Here we have employed time series forecasting by using prophet method to predict future crime trends based on historical data.

1. Overall Crime Trends:

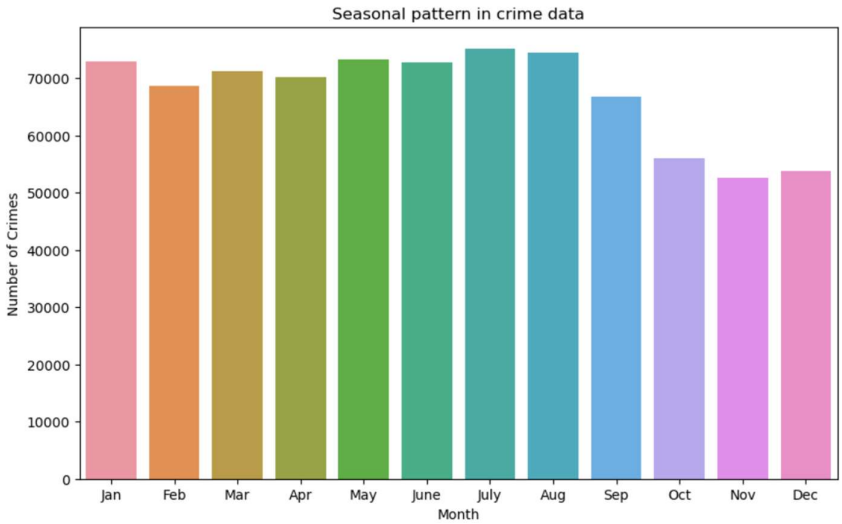
From the observation, 2022 experienced the highest reported crimes, with the count exceeding 200,000. This was followed closely by 2022. On the other hand, 2020 and 2023 witnessed a decrease in the crime rates, with 2023 having the lowest count of about 160,000. The crime rates have consistently increased from 2020 to 2022. There might be various reasons like pandemic, economic downturns, job losses, financial hardships or changes in law/policy.





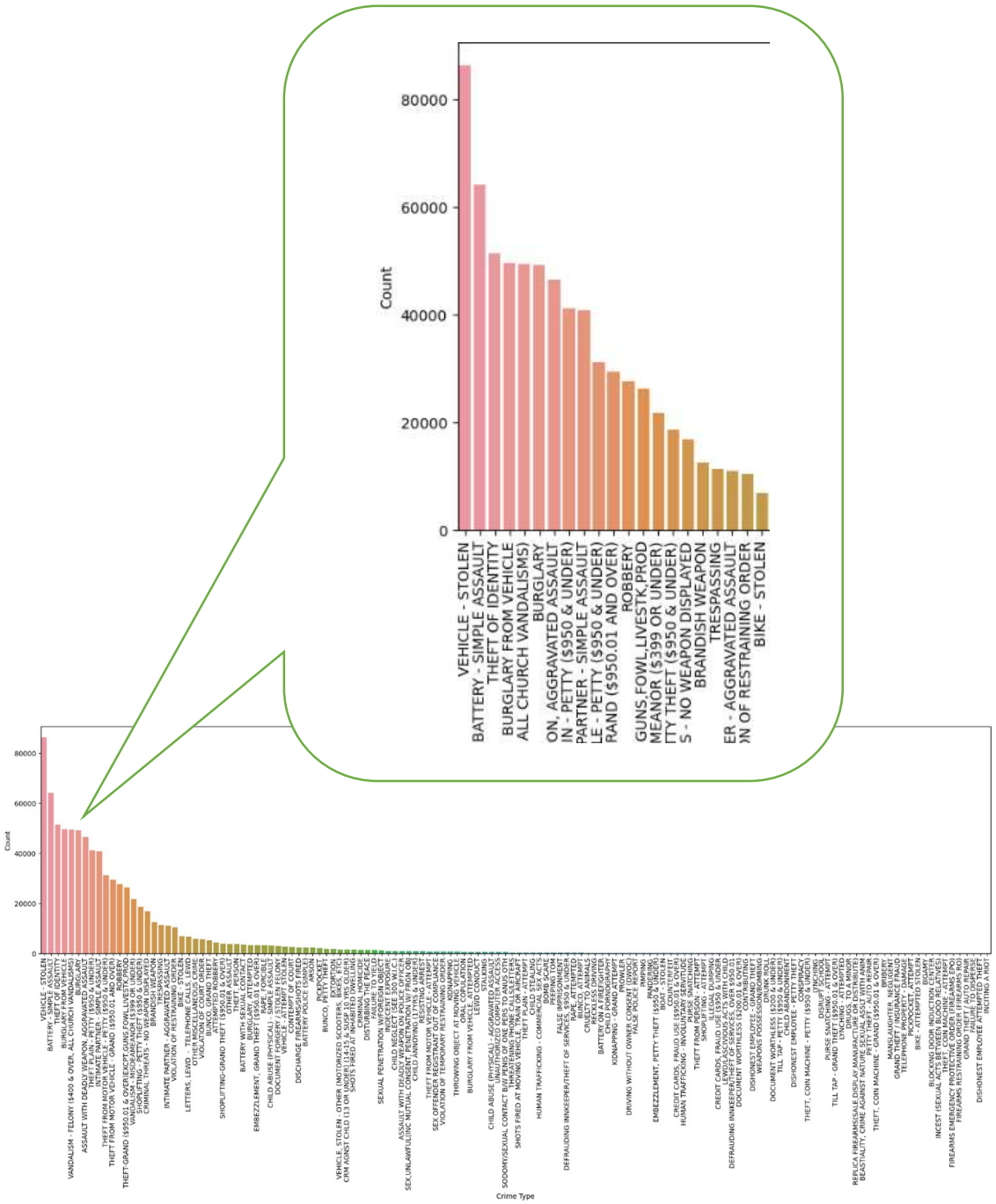
2. Seasonal Patterns:

Crime rates appear relatively consistent throughout the year, hovering between 50,000 and 70,000 incidents. July witnesses the highest crime rate above 70,000, followed by a slight decrease in August. January also saw a high crime rates. The months from March to August maintain a stable count, each above 60,000 crimes. A noticeable drop occurs in September, with rates continue to decrease till November. The rates slightly increase in December.



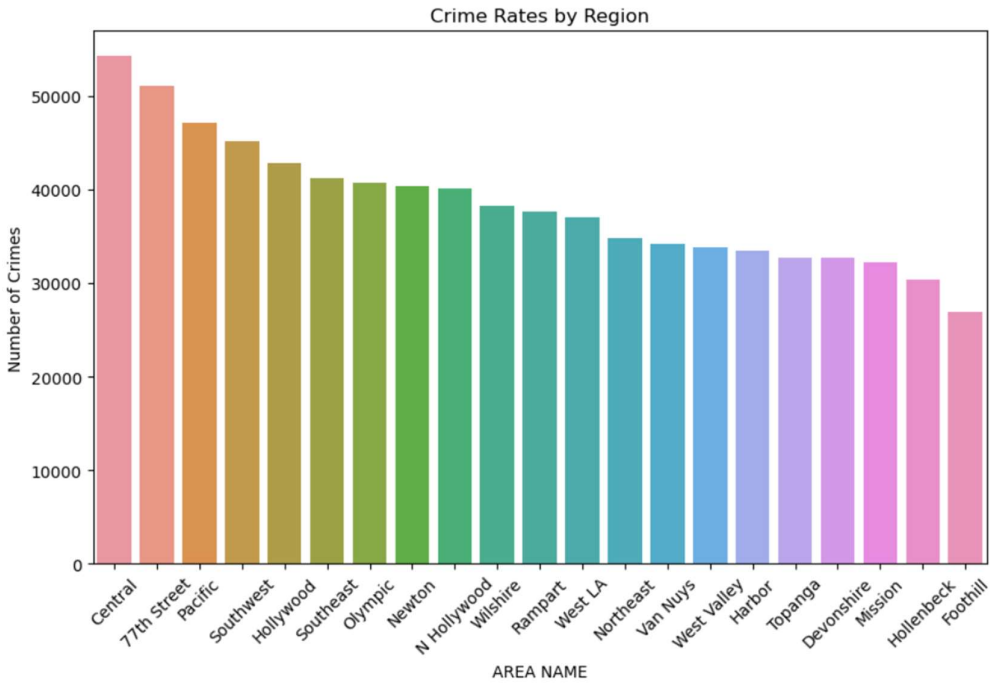
3. Most Common Crime Type:

The most prominent crimes based on the count appear to be related to category of theft, battery, and assault, with these having the highest incidences. “Vehicle Stolen” topped the crime list with 86,335 count. As we move towards the right side of the graph, the frequency of crimes decreases, showcasing a wide range of offenses from vandalism, burglary, and drug-related crimes to lesser-reported incidents such as embezzlement, fraud, and specific types of assault. The distribution gives a clear insight into the prevalence of specific crimes, with theft-related offenses dominating the statistics. The crime with least count was “Inciting a Riot”.



4. Regional Differences:

The bar graph showcases crime rates across different regions. The "Central" region experiences the highest crime rates, surpassing 50,000 incidents. Following Central, "77th Street" and "Pacific" regions also demonstrate significantly high crime rates, but none reach the levels of Central. As we transition to "Southwest," "Hollywood," and so on, there is a gradual decrease in the number of crimes, though they all remain in the upper range. Regions like "Van Nuys," "West Valley," and "Topanga" show comparatively lower numbers. However, by the time we reach the "Hollenbeck" and "Foothill" regions, the crime rates decrease notably, with "Foothill" having the lowest reported incidents.



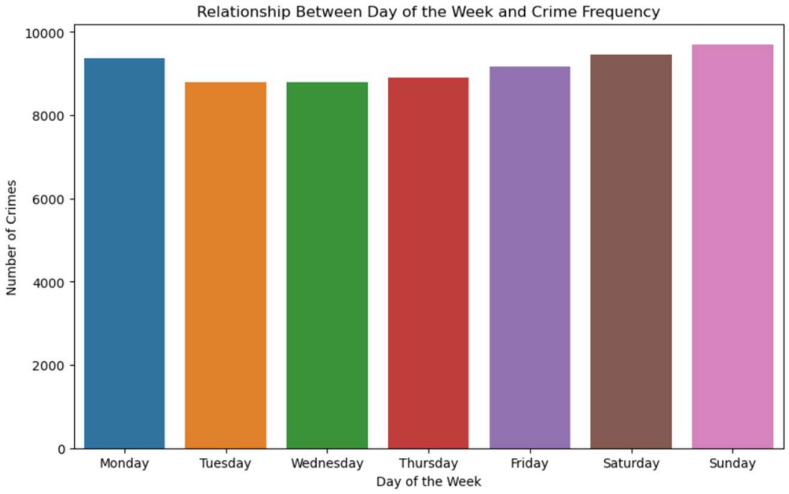
5. Correlation with Economic Factors:

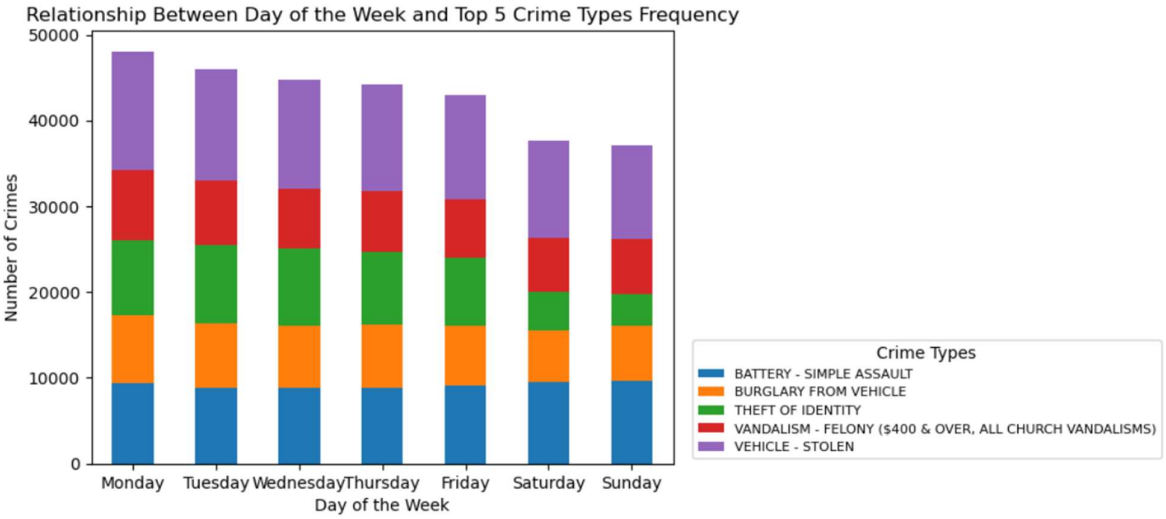
The heatmap illustrates the correlation between various economic factors and crime rate. Notably, there's a strong correlation between unemployment rate and crime count, indicating that as unemployment rises, so does the number of crimes. Conversely, GDP shows a moderately negative correlation with crime count, suggesting that as a region's GDP increases, crime tends to decrease. Income level and crime count appear to have a very weak or negligible relationship. Lastly, the poverty rate exhibits a relatively strong negative correlation with crime, implying that lower poverty levels are associated with low criminal activities.

6. Day of the Week Analysis:

Sunday sees the highest number of reported incidents, approaching 10,000 cases. From Tuesday to Friday, there's a noticeable decrease in criminal occurrences, with numbers around 8,000 incidents. Despite minor fluctuations throughout the week, the data suggests that the beginning of the week, specifically Monday, has a heightened crime rate compared to the rest of the week.

The second graph shows the relationship between the days of the week and the frequency of the top 5 crime types. The crimes listed include "Battery - Simple Assault", "Burglary from Vehicle", "Theft of Identity", "Vandalism - Felony ($400 & over, All Church Vandalisms)" and "Vehicle - Stolen". It shows consistent trend across the week for these crime types with only minor fluctuations in frequency from Monday to Sunday. Notably, "Vehicle- Stolen " appears to be the most committed crime every day.





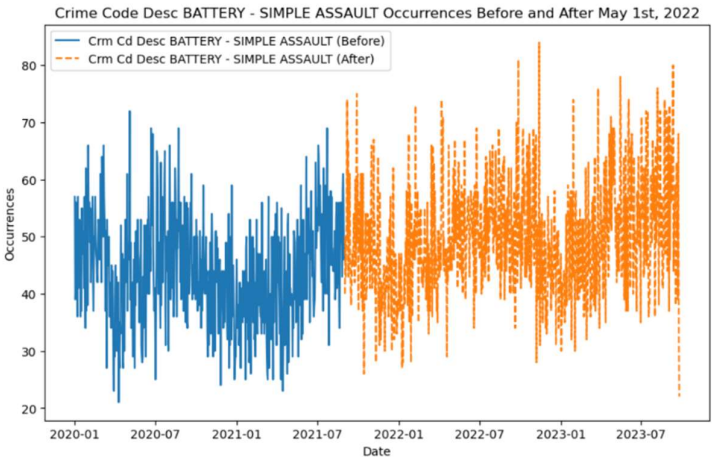
7. Impact of Major Events:

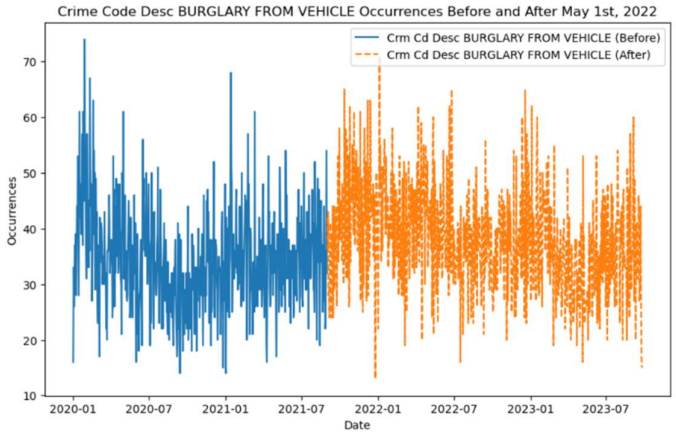
This analysis is based post second wave of COVID-19 pandemic, Major event.

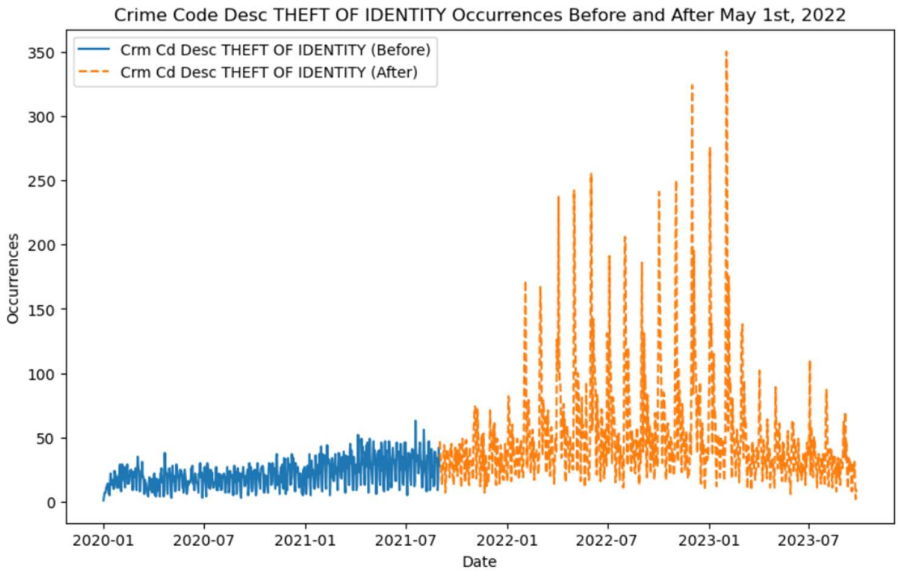
The occurrence of "**Simple Assault**" crimes before and after May 1st, 2022. It seems that the occurrences were fluctuating between 2020 and mid-2022 but then there's a notable increase after May 1st, 2022.

The second plot, the occurrence of "**Burglary from Vehicle**" crimes before and after May 1st, 2022. This plot shows a similar trend, occurrences fluctuated throughout.

The occurrences of "**Theft of Identity**" crimes remained relatively stable and low, oscillating between approximately 0 to 50 occurrences per period. No significant spikes or patterns are apparent before May 1st, 2022 . But, the frequency of crimes rises sharply, with multiple spikes reaching up to around 350 occurrences after May 1st, 2022. After reaching its peak, there is a gradual decline but the occurrences are still significantly higher than the period before May 1st, 2022.

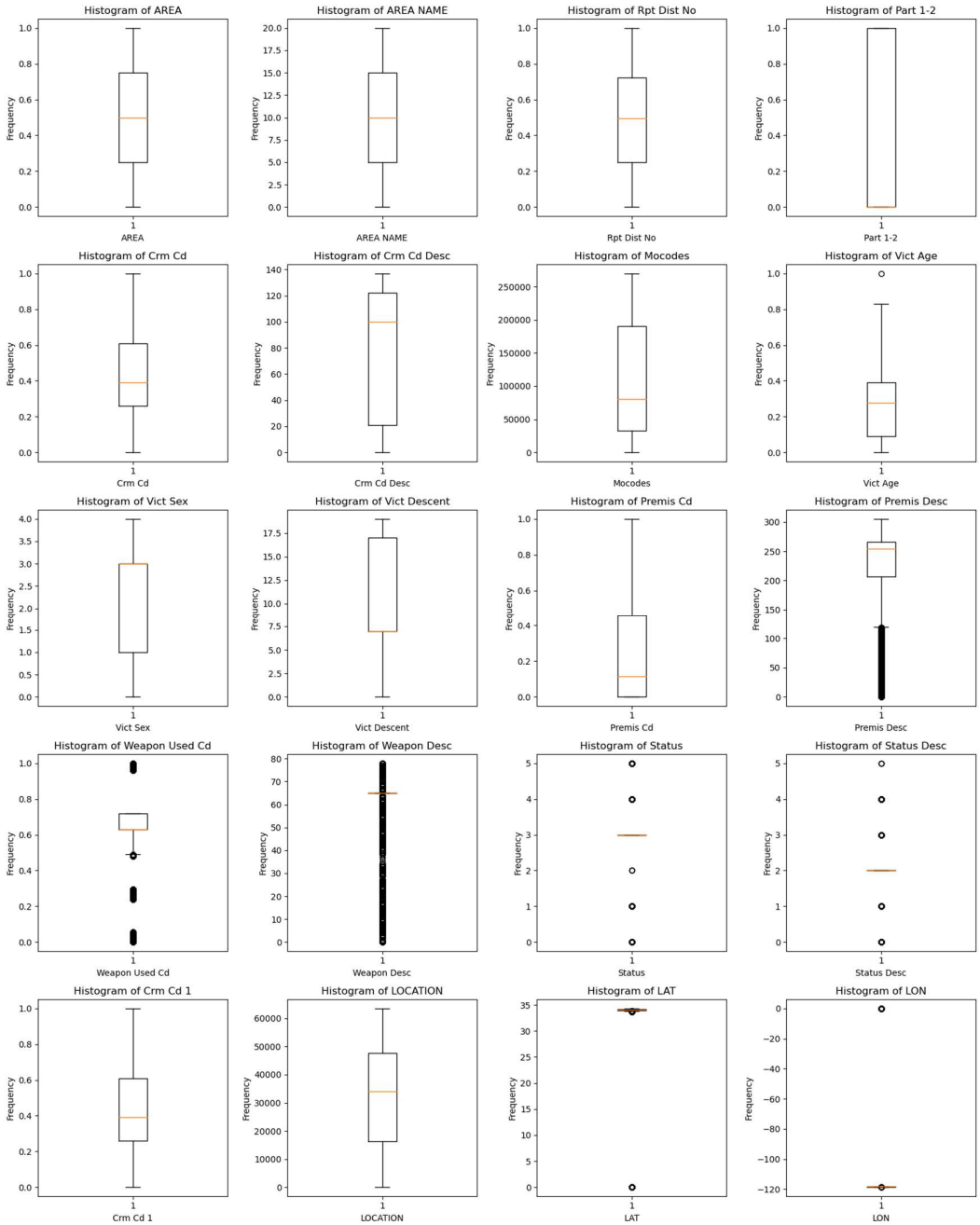






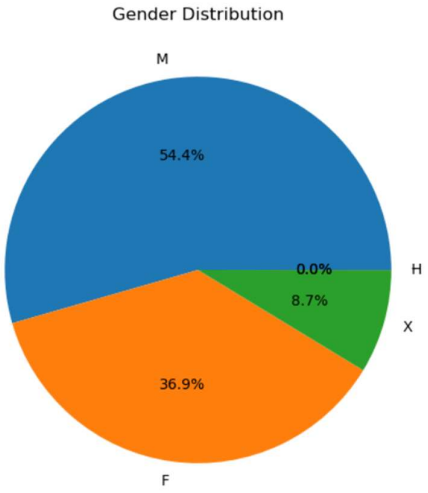
8. Outliers and Anomalies:

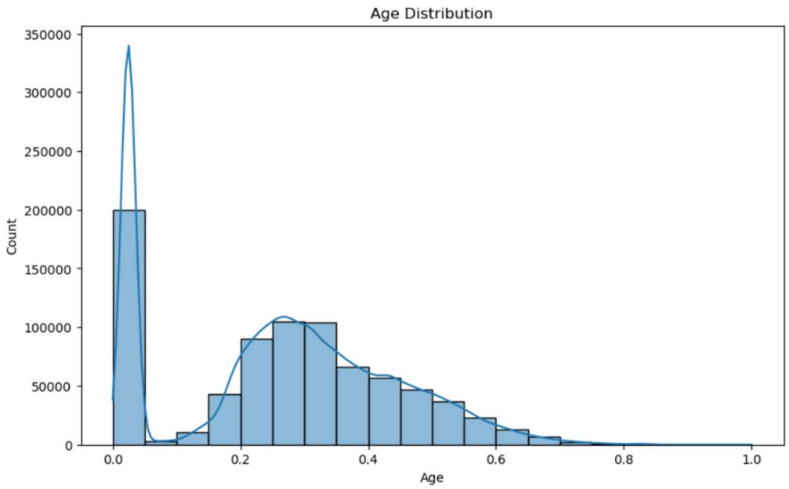
Most attributes exhibit a concentrated distribution, indicating consistency in the data. However, some notable exceptions include the "Weapon Desc", "Status", "Status Desc", “Premis Desc” and coordinate histograms "LAT" and "LON", which display outliers. Specifically, the "Weapon Desc" histogram has a few extreme values that stand out, suggesting irregularities or rare occurrences in the weapon descriptions. The "LAT" and "LON" histograms show some data points that deviate significantly, suggesting potential errors or anomalies in location data recording. The attributes like "AREA," "AREA NAME," and "Crm Cd" seem relatively well-distributed without apparent outliers, indicating uniformity in those aspects of the dataset.

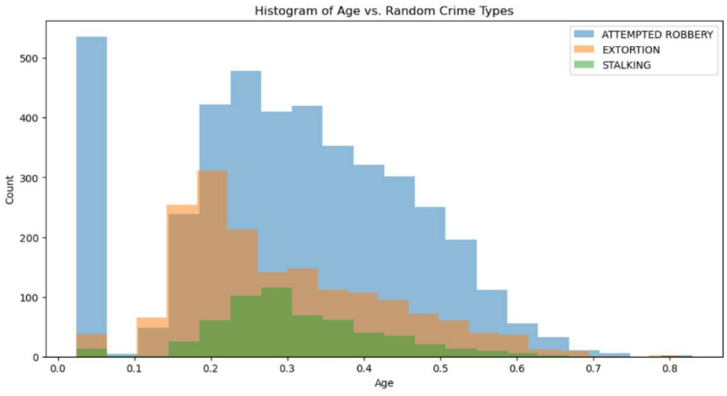


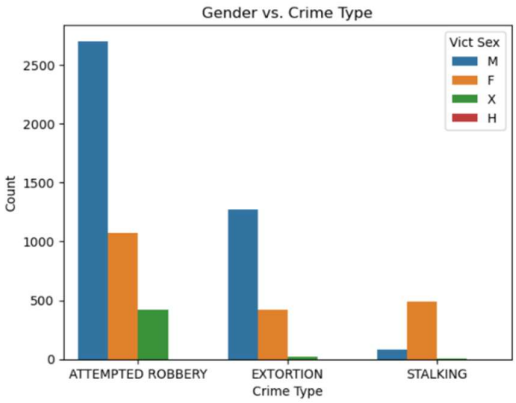
9. Demographic Factors:

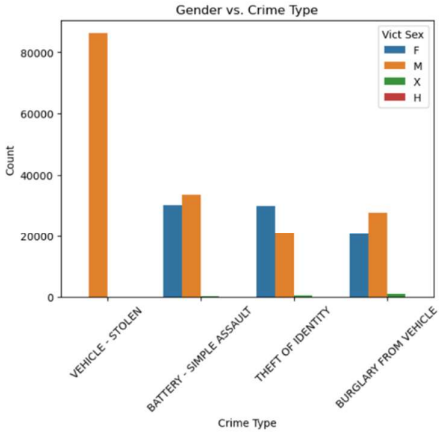
Young age groups, especially near the start of the age distribution, represent a significant portion of the population. If this data is about crime victims or perpetrators, it would be important to correlate age with specific crime types to see if younger age groups are more or less likely to be affected. Males are the predominant group in the provided population and also appear to be the predominant victims in the selected crime types. Robbery seems to be the favourite crimes. For crimes like 'Battery - Simple Assault' and 'Theft of Identity' both Male and Female show similar crime rates.





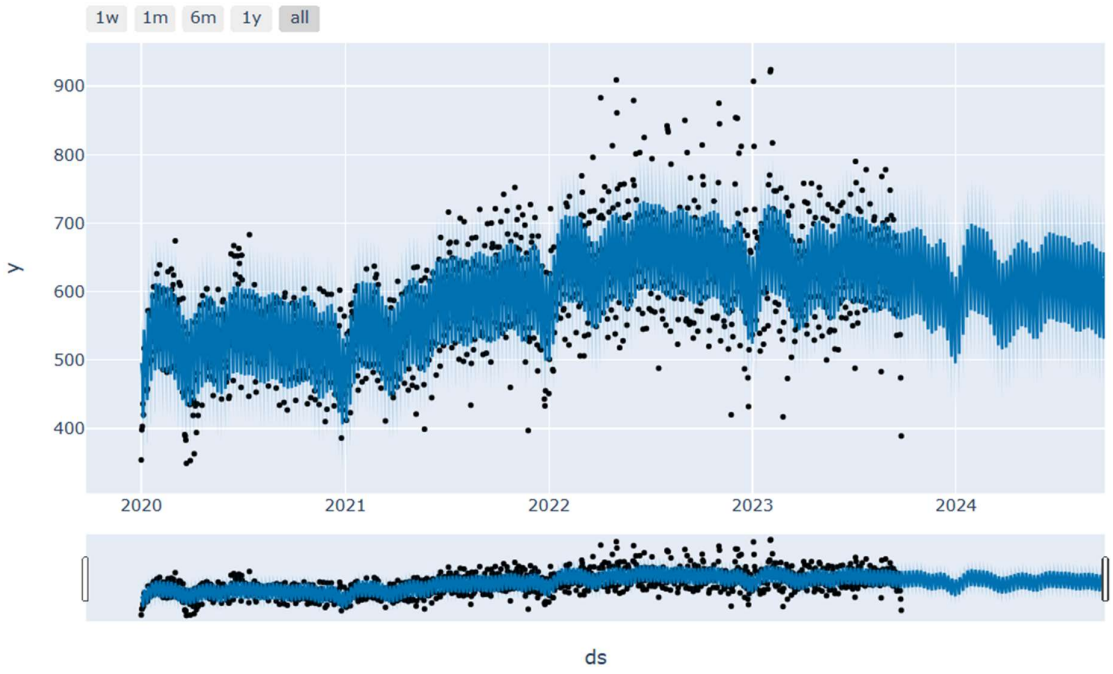






10. Predicting Future Trends:

The chart illustrates a crime forecast utilizing the Prophet model, spanning from 2020 to 2024. The observed crime rates, represented by the black dots, show fluctuating patterns over the past years. From mid-2023 onwards, the model predicts a general decline in crime rates, as indicated by the blue-shaded region. This shaded area also signifies the model's uncertainty interval, suggesting that while a decrease is anticipated, there are bounds to the prediction's confidence. The forecasted pattern implies a more stable and possibly safer environment in the upcoming year from mid-2023.



Limitations and Future Work:

While this analysis provides valuable insights into the factors influencing crime rates in the city for different regions, it has certain limitations. The data sources used has some limitations in terms of accuracy and timeliness.

Additionally, the analysis is based on historical data, and crime rates can change over time due to various factors. Future work in this project could involve updating the data with more accurate and recent information to provide with up-to-date insights. It would also be beneficial to expand the analysis to include a wider range of cities for a more comprehensive understanding of crime trends based on its neighboring areas.

In conclusion after analyzing the crime data, we have found several important insights. It is evident that there is a relationship between demographic characteristics, economic conditions, law enforcement efforts, and crime rates.

This also can be used as a powerful tool for individuals looking to make informed decisions when going to this city and law enforcement and use this to predict the crime trends. In total it offers a foundation for further research and enhancements to improve the accuracy and usability of the information provided.