EXPLORATORY DATA ANALYSIS

IE6400 – Foundations Data Analytics Engineering
Final Report

Group Number: 04

Aadarsh Praveen Selvaraj Ajithakumari (002832667)
Aravind Swamy(002847684)
Moheesh Kavitha Arumugam (002296201)
Hashwanth Moorthy(002830971)
Uma Maheshwari Deivasigamani (002847743)

INTRODUCTION:

Los Angeles

In Southern California, Los Angeles—also referred to as L.A.—is a sizable city with a wide range of cultures. Notable for its magnificent beaches, year-round moderate temperatures, and legendary entertainment industry, this city is the second largest in the United States. A multicultural enclave with a wide range of lifestyles, Los Angeles is a major center for Hollywood, music, film, and television worldwide. It is a quite well-known metropolitan center because of its many neighborhoods, attractions for the arts, and importance to the economy.

Police Department of Los Angeles (LAPD):

The criminal landscape in Los Angeles is varied, including violent crimes, property crimes, and neighborhood-specific gang activity. High levels of violence and crime are a result of the existence of active gangs in certain communities. Law enforcement, community initiatives, and technology are all used in ongoing attempts to address and lower the number of criminal incidents. The town is dedicated to improving community safety and putting anti-crime policies into place in an effort to make the area safer for its citizens.

In the city, the Los Angeles Police Department (LAPD) is leading the way in preventing crimes. They participate in community service, quickly respond to emergencies, and actively patrol neighborhoods. Policing to uphold community standards and foster trust. The department utilizes technology for surveillance and predictive policing, hires crime analysts to spot trends, and performs inquiries in order to resolve cases. The community is actively involved in efforts to prevent crime as part of the LAPD's multifaceted approach to improving public safety.

Aim & Approach:

Exploratory Data Analysis (EDA) for the dataset on criminal events in Los Angeles entails a systematic method to eliciting useful insights. Understanding the structure of the dataset and cleaning it are critical first steps in correcting missing or erroneous data entries. Basic summary statistics, temporal analysis, and geographical exploration demonstrate crime patterns over time and space. Analysis of crime categories and demographics might provide useful context. Data visualization assists in the presentation of data in an intelligible manner, while outlier identification and correlation analysis expand our comprehension. Finally, EDA provides a foundation for making educated decisions and developing successful crime prevention tactics.

TABLE OF CONTENTS:

DATA SET SUMMARY:

We have about 829778 Rows and about 28 Columns and deeper into the column description is below:

DATASET'S COLUMN DESCRIPTION:

- **DR_NO**: Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits.
- **API Field Name**: MM/DD/YYYY.
- **DATE OCC**: MM/DD/YYYY.
- **TIME OCC**: In 24 hour military time.
- **AREA**: The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21.
- **AREA NAME**: The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.
- **Rpt Dist No**: A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons.
- **Crm Cd**: Indicates the crime committed. (Same as Crime Code 1)
- **Crm Cd Desc**: Defines the Crime Code provided.
- **Mocodes**: Modus Operandi: Activities associated with the suspect in commission of the crime.
- **Vict Age**: Two character numeric.
- **Vict Sex**: F - Female M - Male X - Unknown.
- **Vict Descent**: Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian.
- **Premis Cd**: The type of structure, vehicle, or location where the crime took place.
- **Premis Desc**: Defines the Premise Code provided.
- **Weapon Used Cd**: The type of weapon used in the crime.
- **Weapon Desc**: Defines the Weapon Used Code provided.
- **Status**: Status of the case. (IC is the default).
- **Status DEsc**: Defines the Status Code provided.
- **Crm Cd 1**: Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
- **Crm Cd 2**: May contain a code for an additional crime, less serious than Crime Code 1.
- **Crm Cd 3**: May contain a code for an additional crime, less serious than Crime Code 1.
- **Crm Cd 4**: May contain a code for an additional crime, less serious than Crime Code 1.
- **LOCATION**: Street address of crime incident rounded to the nearest hundred block to maintain anonymity.

- **Cross Street**: Cross Street of rounded Address.
- **LAT**: Latitude.
- **LON**: Longitude.

All these columns are of different data types such as integer,float and object.Each column with corresponding data type with its null count is depicted below ,we can find this using the **info** function.We found about 8 float data types ,7 integer data type and about 13 object data types there in this dataset which has to be changed to appropriate types for better understanding and description.

DATA SOURCE:

The provided link directs you to a dataset available on the Data.gov platform, offering crime data spanning from the year 2020 to the present day. This dataset contains information about various criminal incidents in a particular region, likely Los Angeles, and may include details such as the type of crime, dates of occurrence, locations (though anonymized for privacy), and other related information. It serves as a valuable resource for data analysis, research, and insights into crime trends and patterns in the specified region.

Link for the dataset :

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

For this project the above dataset is already downloaded as a CSV file.

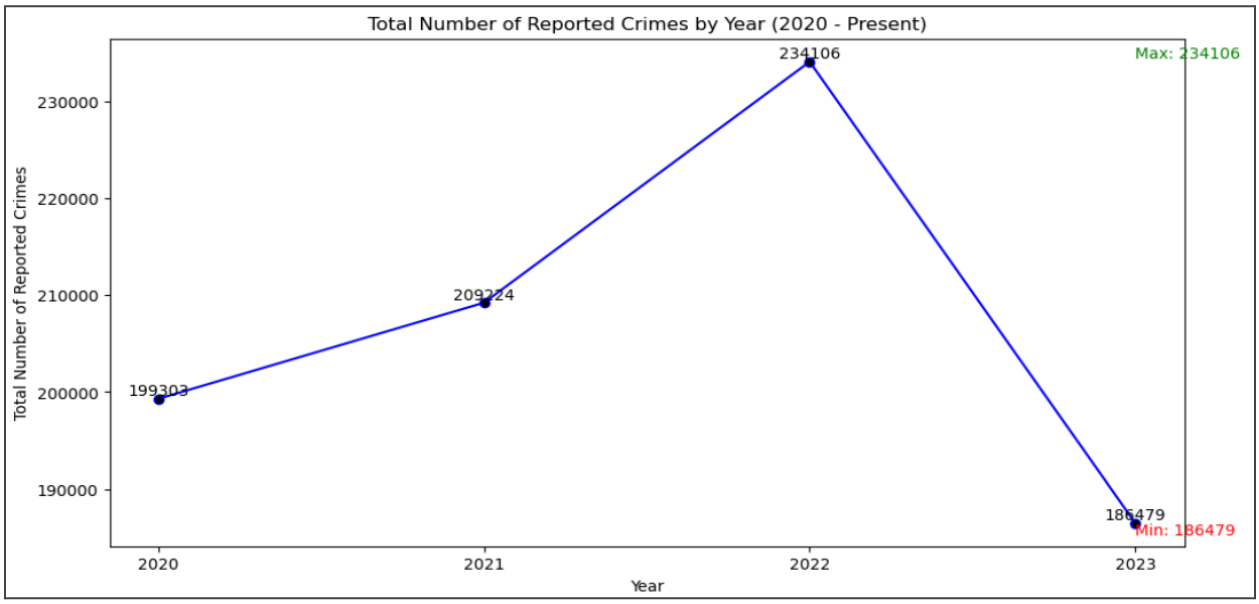
DATA CLEANING:

In order to have a better data quality ,reliability and analysis of the dataset we need to clean the crime_data_from_2020_to_present as its found with many missing data leading to bias and inaccuracies.The NaN are filled with appropriate data types, the integers columns are filled with 0 and others are filled with matching categorical values in the given dataset like **Vict Sex** is filled with **X** ,**Weapon Desc** as **No weapon** ,**Vict Descent** with **X** and **Cross Stree**t as **Unknown** which reduces the ambiguity and fake information storage. Since there is no duplicate value the scope for redundant data is ruled out.

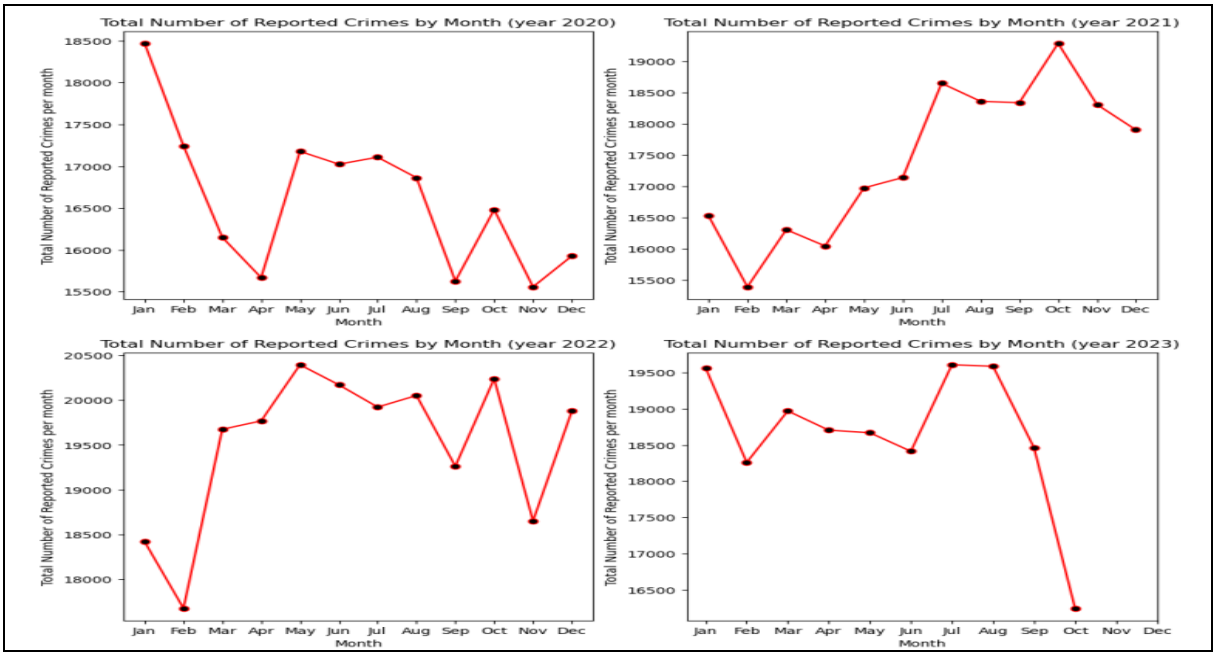
EDA:

In this section , we will do multiple tasks like exploring the crime dataset and visualizing the overall crime trend over the years and find important factors like the most common crime and seasonal crime trends. In addition we will also explore the differences in the crime rates across various regions in Los Angeles and investigate if there is a direct

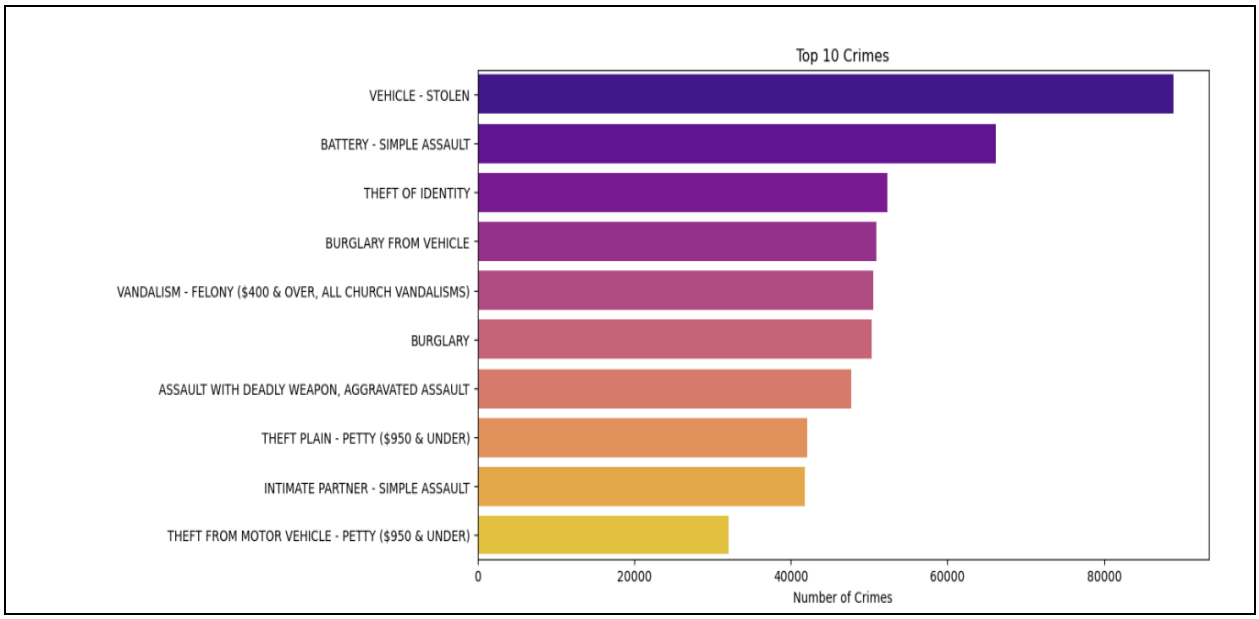
correlation between the economic factors and crime rates. Finally we will also analyze and find if there are any influences on crime rates based on days. We also use suitable graphs to present the gained insights,



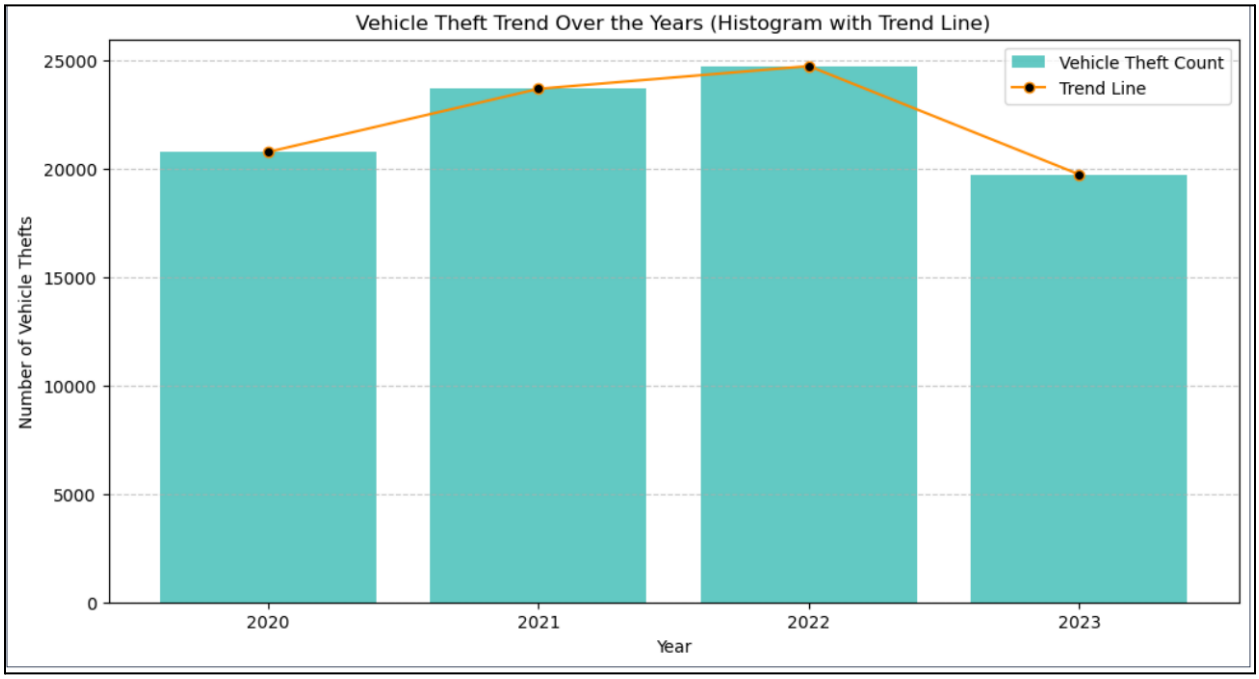
The graph above displays the trends in crime from 2020 to the present. It is evident from the graph that there was a noticeable increase in crimes reported between 2020 and 2023. There was a modest increase in the crimes from 2020 to 2021 and there was a peak in events reported in 2022, which was followed by a sharp decrease in 2023. This pattern suggests that since the peak year, recorded crimes have generally decreased. However, there are still yearly fluctuations in the quantity of cases reported, which are probably caused by several factors. According to recent data, there has been a decline in reported incidents and events in 2023 as compared to prior years. A deeper exploration of the underlying forces is necessary to comprehend the intricacies of these trends.
After this we will also analyze and visualize the seasonal influence on crime patterns,



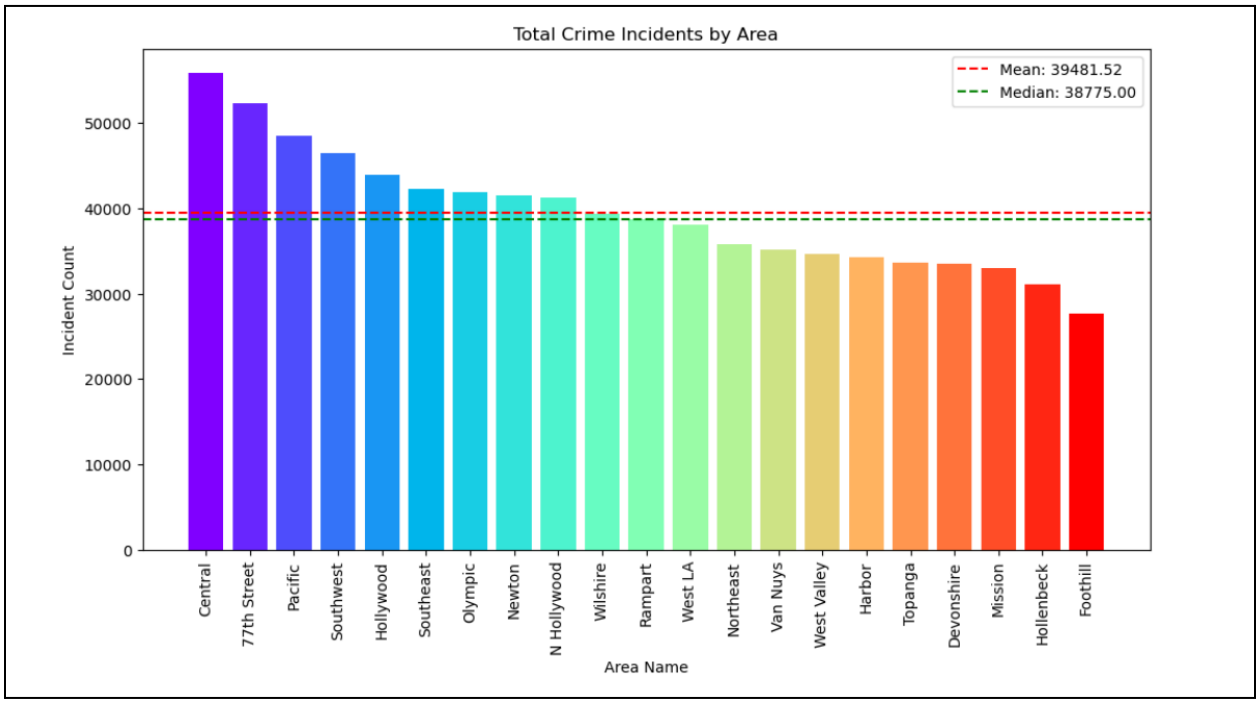
The graphs above show that there was a considerable variation in the total number of reported crimes in 2020. The data shows that January (18,469 incidents) and May (17,178 cases) had the highest number of crimes committed. With 15,559 cases reported, November had the fewest crimes overall. This illustrates the monthly variations in events that are reported; crime rates may be influenced by situational or seasonal factors. In 2021, there were large variations in the monthly total of recorded offenses. October (19,285 cases) was the month with the most crimes reported, followed by July (18,647 cases). February, on the other hand, had the fewest offenses (15,393 total). These monthly variations in crime reports raise the possibility that seasonal or contextual factors may have an impact on crime rates.In 2022, there were discernible variations in the quantity of offenses reported for each month. May (20,396 cases) was the month with the highest number of crimes reported, followed by October (20,240 cases). March had the fewest offenses reported, with 19,675 cases. These monthly variations suggest that crime rates could be affected year-round by external factors or seasonal trends. According to the data available, there is a less noticeable trend in reported crimes in 2023 compared to previous years. The most reported offenses occurred in the month of July, while figures for the following months are fairly stable. October has the fewest incidents according to the data, which indicates a less variable rate of crime; however, it is important to consider the pattern over the course of the entire year once the data is available for more thorough analysis.

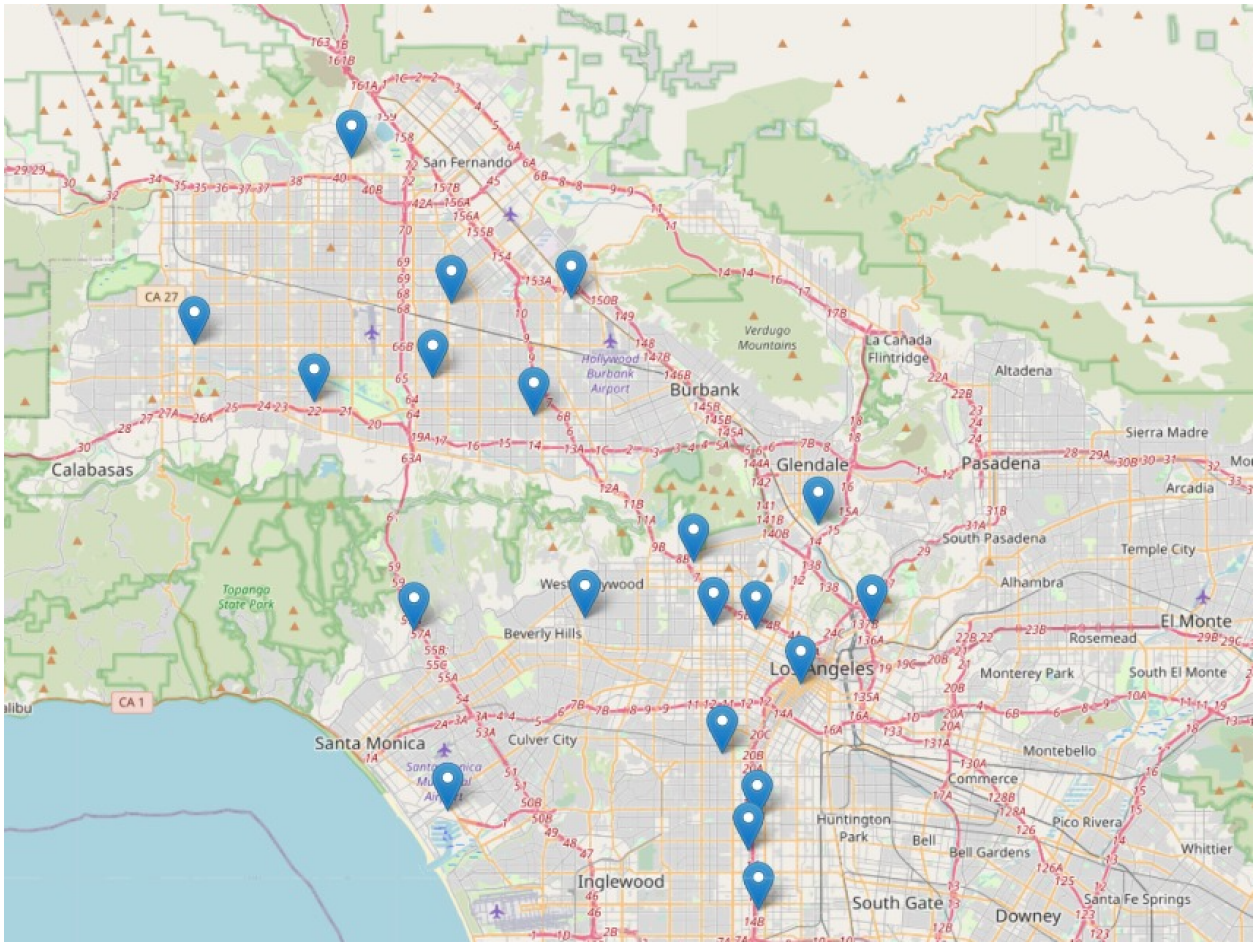


Next ,in order to find the crimes over the year by each month ,we would use a bar graph with Crm Cd Desc .The list of the top ten reported crimes offers valuable insights into prevalent societal concerns. Topping the list is "Vehicle - Stolen," underscoring a pressing issue of car theft. Following closely are "Battery - Simple Assault" and "Theft of Identity," indicating heightened worries regarding personal safety and identity theft. Additionally, burglary and vandalism feature prominently, underscoring the urgency to address property-related offenses and enhance overall safety.



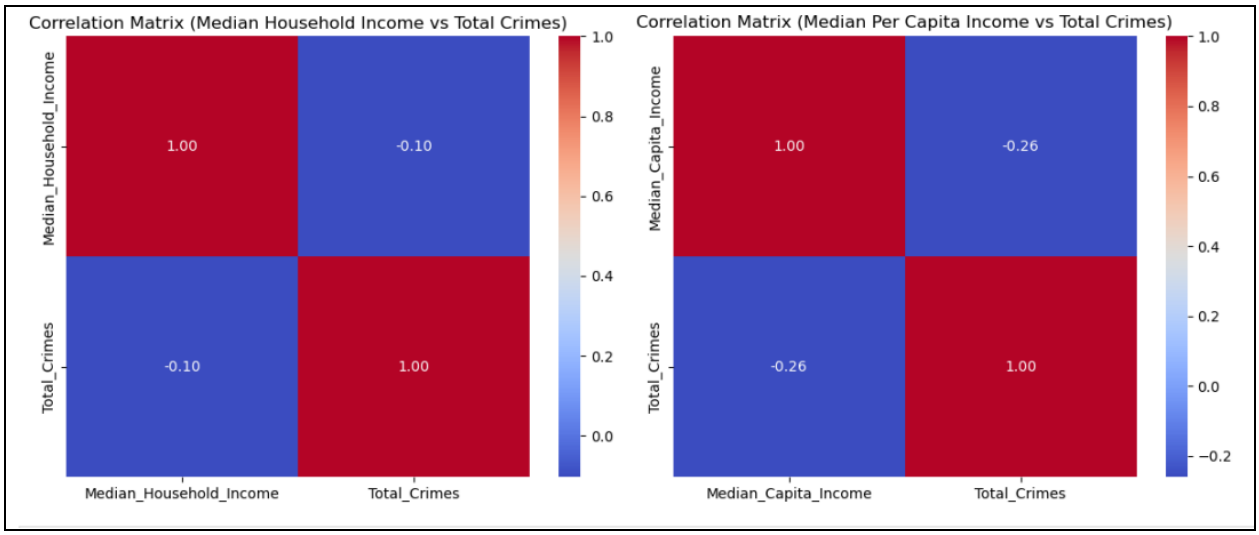
Next, in order to find the total vehicle theft over years we have used a histogram plot with a trend line. Vehicle theft number changes from year to year, but a clear pattern shows it increases every year. The data indicates these thefts go up each year. The number of car thefts climbed from 20,765 in 2020 to 24,723 in 2022, showing a steady increase. In 2023 however there was a substantial decrease with only 19,727 reported cases. The shift in the trend can be caused due to variety of factors, demanding a thorough examination of the underlying causes of these variances.





The bar chart exhibits the number of reported crimes in various regions, arranged in ascending order. It allows for a quick and easy comparison of crime rates across different areas, where taller bars represent higher crime counts and shorter bars denote lower reported incidents.

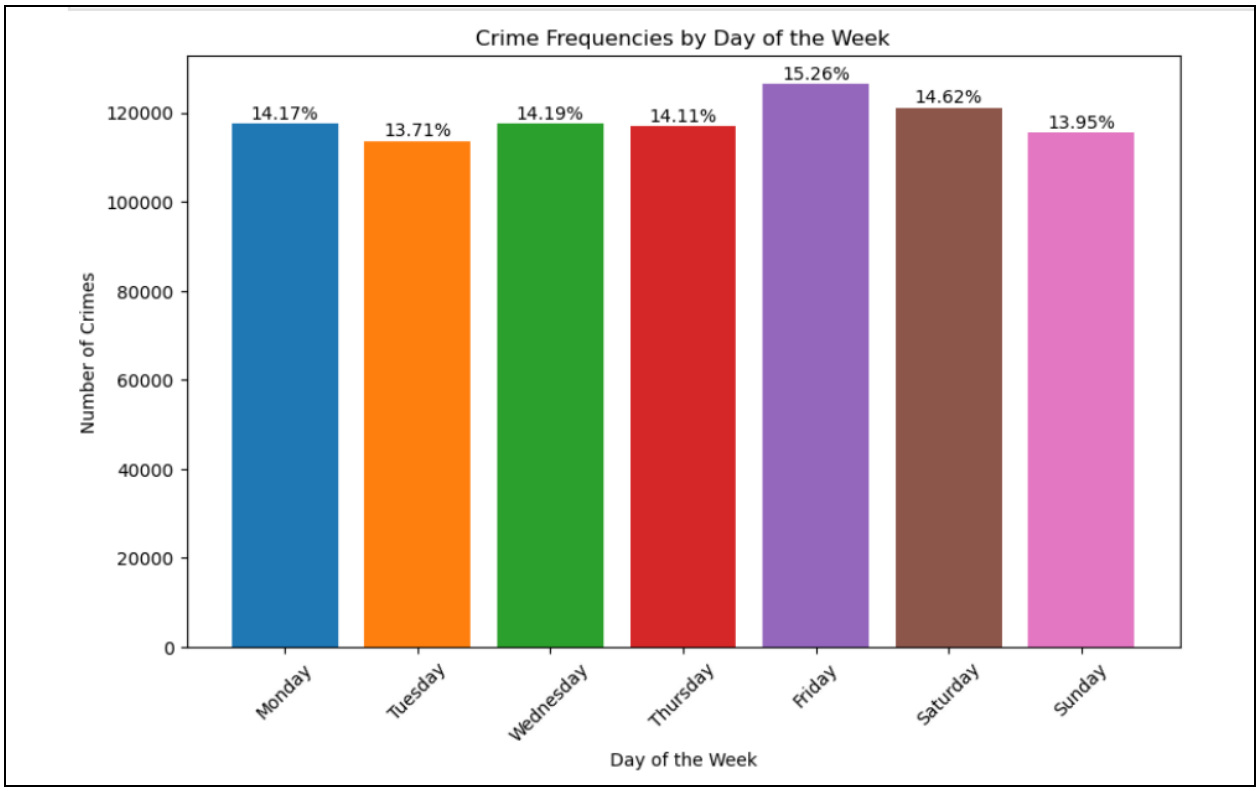
Our investigation delves into understanding the potential relationship between crime rates and diverse economic factors such as income levels, emigration rates, homelessness, and unemployment rates. The primary aim is to explore whether any correlation exists between these economic variables and crime rates observed over time. The report focuses on the number of occurrences in different locations, highlighting Central as the area with the highest count of incidents at 55,890. With 77th Street and Pacific following closely, reporting 52,309 and 48,542 instances, respectively, indicating significant law enforcement involvement. Additionally, the crime rates in the top 10 cities exceed the average of the total crime rates.



For finding the correlation between the economic factors and the crime rates, we plot a correlation matrix. For this purpose we create a dataframe with the median household income and median per capita income over the years 2020-2023 in Los Angeles. We also add a column counting the total number of crimes each year in Los Angeles,

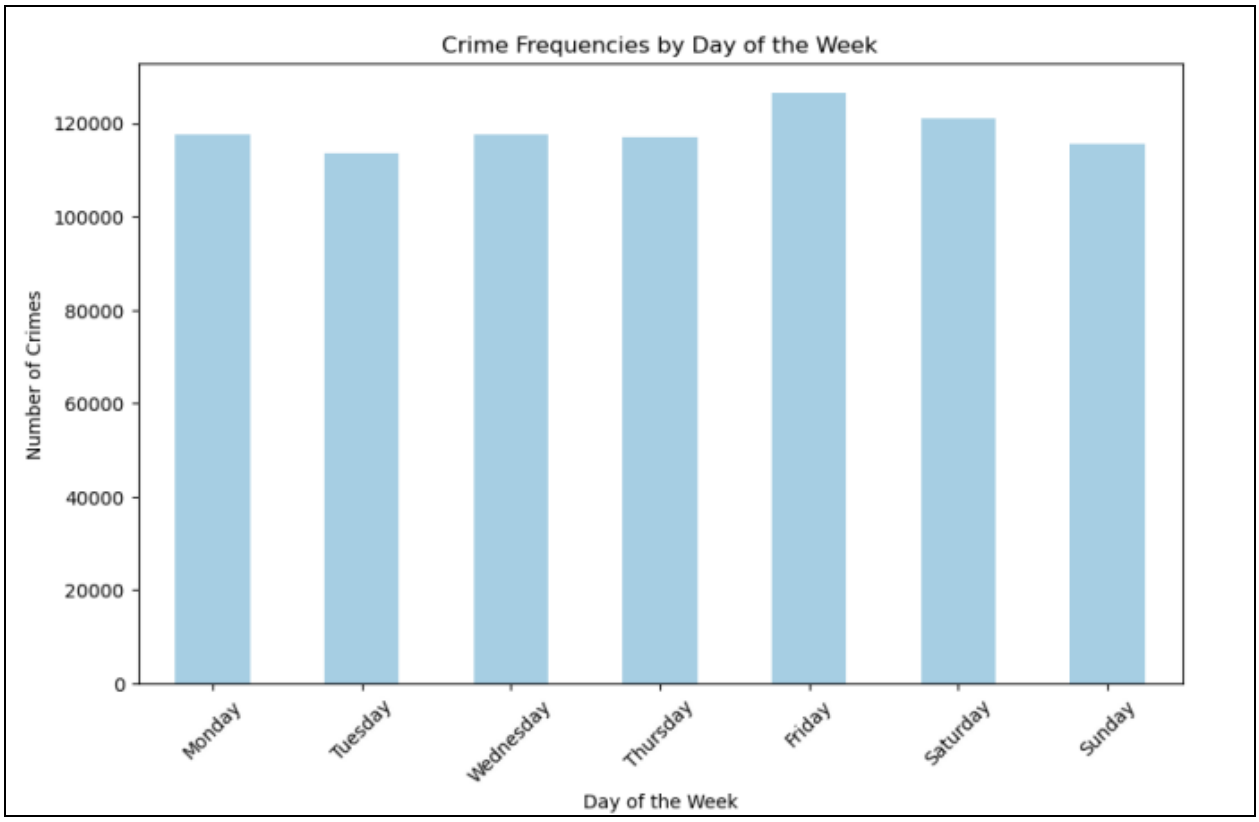
We create a correlation matrix to find the correlation between the crime rates and economic factors.The correlation matrix shows a weak negative correlation of approximately -0.10 between median household income and total reported crimes. This suggests that areas with slightly higher income levels tend to have slightly lower crime rates, but the relationship is not strong. It's important to note that multiple factors influence crime rates, and income is just one piece of the puzzle. Further analysis and consideration of additional variables are needed to fully understand crime trends

The correlation matrix shows a moderate negative correlation of approximately -0.26 between median per capita income and total reported crimes. This suggests that areas with higher per capita income tend to have lower crime rates. While the relationship is more substantial than with median household income, it's essential to recognize that crime rates are influenced by various factors, and per capita income is just one element in this complex interplay. Further analysis and consideration of additional variables are needed for a comprehensive understanding of crime trends

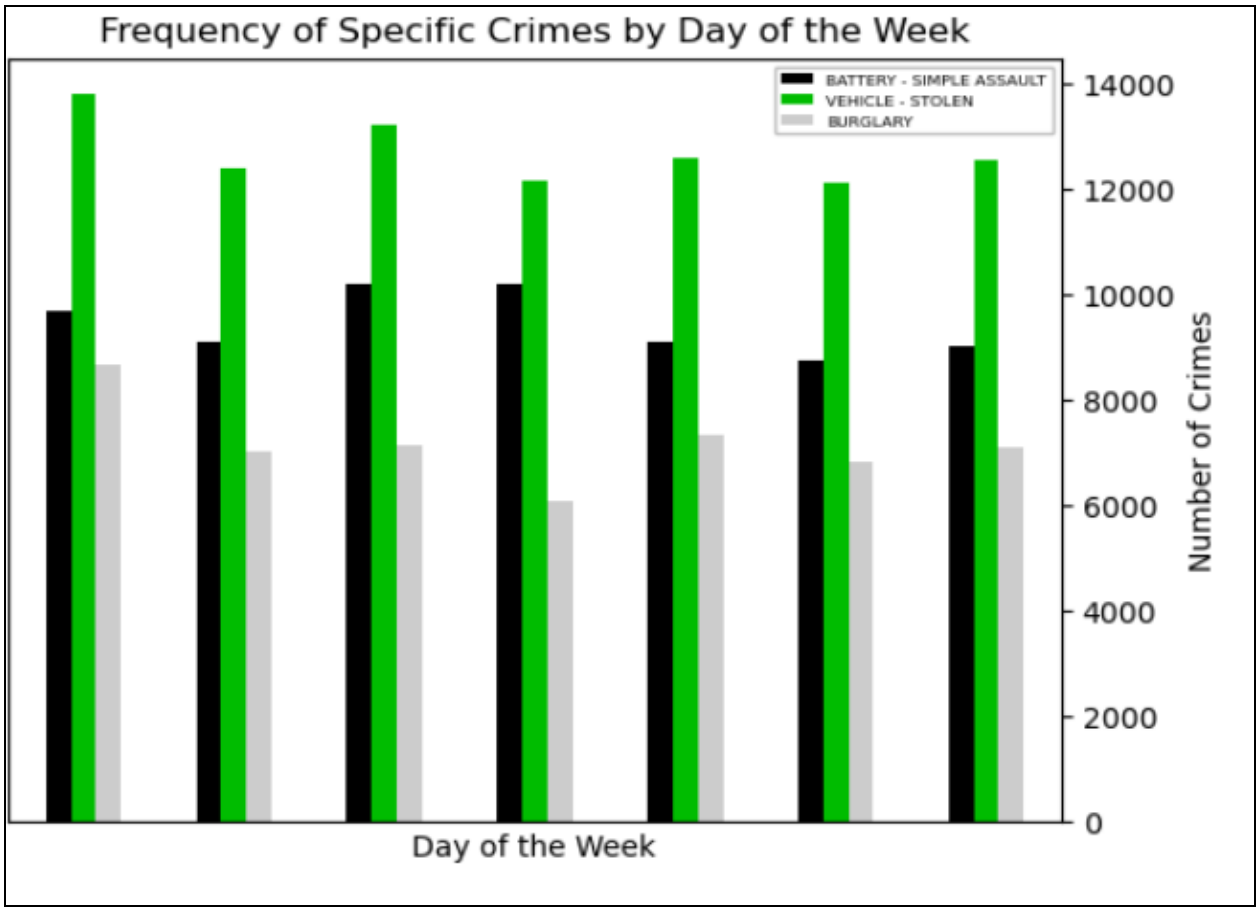


The bar chart gives a graphical representation of the number of crimes that have happened on all the days of the week. From the chart we can infer that the maximum number of crimes which is 15.26% took place on friday which is the last day of the weekday. The least number of crimes took place on Tuesday, which is 13.71% . Overall the crimes took place in all the days of the week. The data indicates that there are fluctuations in reported crimes depending on the day of the week. Fridays record the highest number of incidents, closely followed by Saturdays and Wednesdays, pointing to increased criminal activity during weekends and mid-week periods. Notably, Mondays and Thursdays also report significant crime rates, while Tuesdays stand out as having the lowest number of reported crimes among the weekdays. These observations provide valuable information for law enforcement and community efforts, allowing them to allocate resources effectively and tailor crime prevention strategies to account for these temporal patterns in criminal incidents.

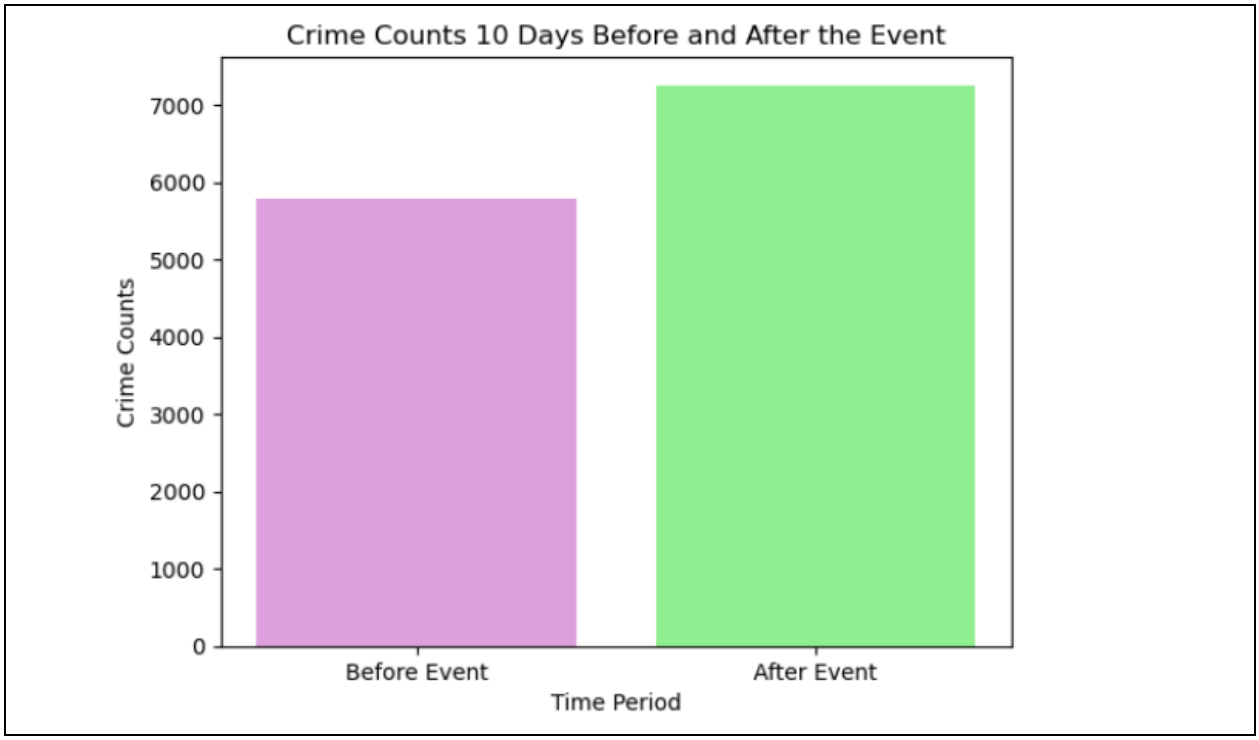
The crime frequency graph by day of the week shows that crime is most common on Mondays, followed by Wednesdays and Fridays. Crime is least common on Sundays. This pattern is likely due to a number of factors, including:People are more likely to be away from their homes and businesses on weekdays. This makes it easier for criminals to commit crimes such as burglary and car theft.



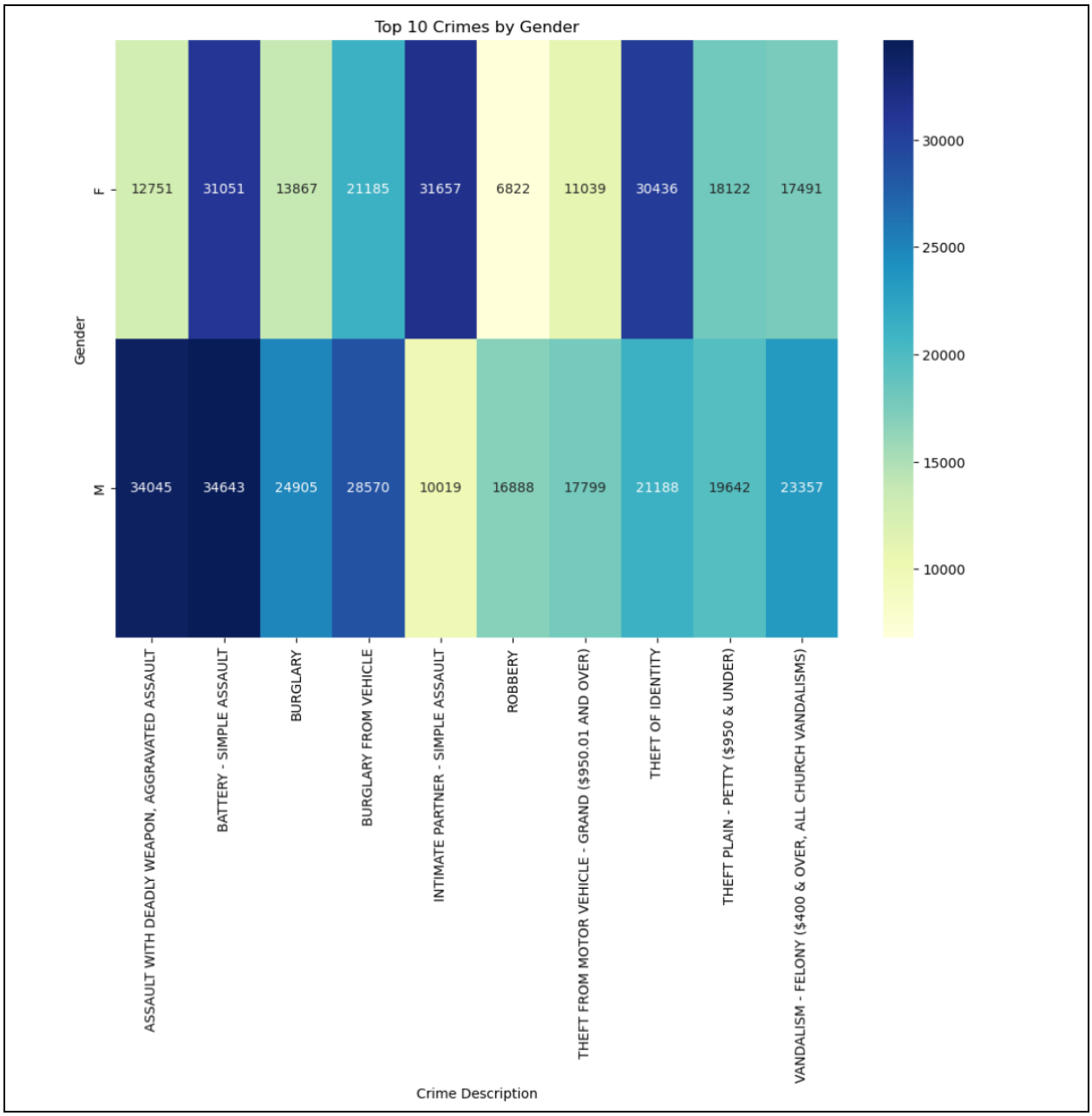
There are more people out and about on weekends. This increases the opportunities for crimes such as assault and robbery.There may be different law enforcement which could have caused the fluctuation in pattern over the years.



The frequency of specific crimes by day of the week is generally higher on weekdays than on weekends, with the highest crime rates on Mondays and Wednesdays. This is likely due to a number of factors, including people being more likely to be away from their homes and businesses on weekdays, and more people being out and about on weekends.Crime frequency is highest on Mondays and Wednesdays, and lowest on Sundays.

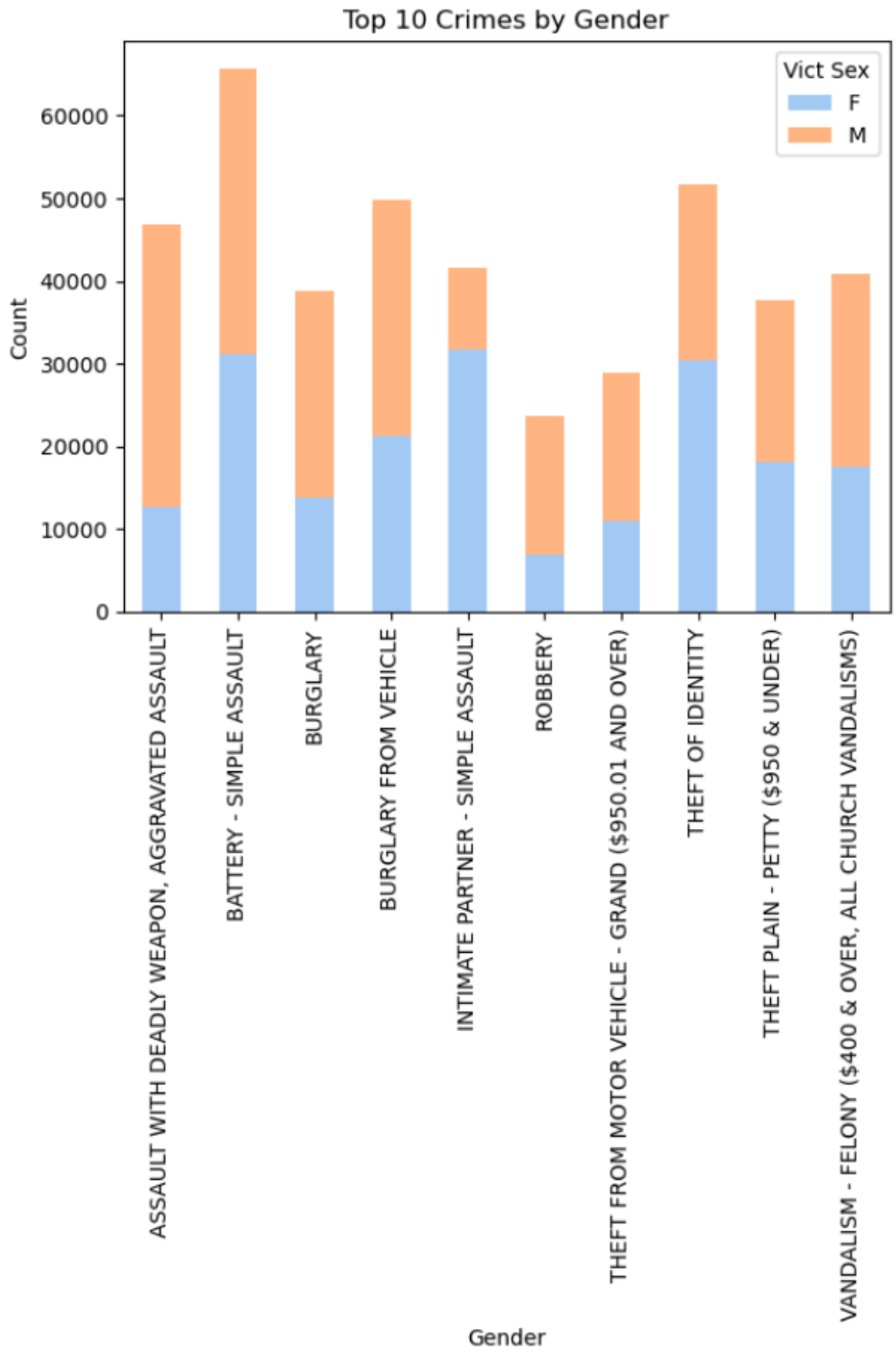


The above bar chart gives the number of crime counts that have happened before and after an event. On November 28 2022 a carnival took place in Los Angeles, the name of the event was “Camp Flog Gnaw Carnival” , the number of crime events increased rapidly once the event got over. The graph tells us that the crime rate has increased due to the result of the event.

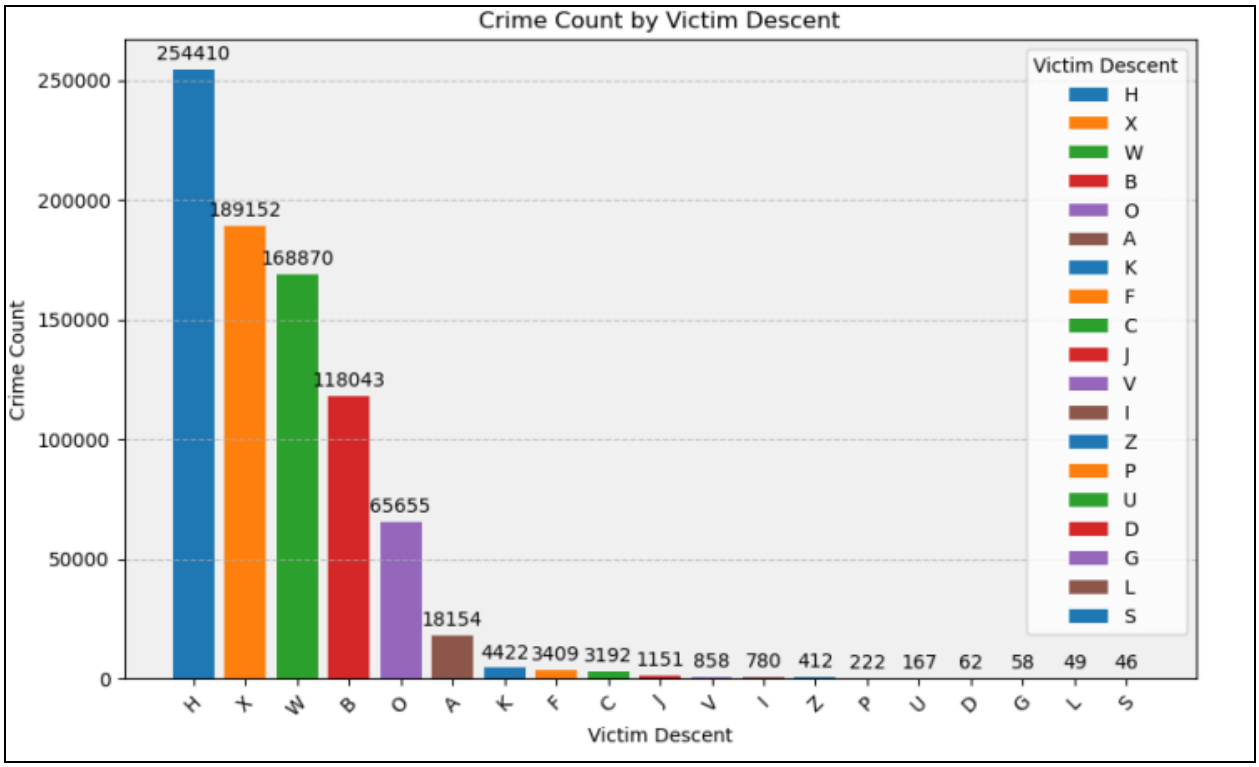


Male Victims: The graphical representation (heat map) reveals that the highest count of victims in reported crimes comprises males, indicating a greater frequency of males as victims in the dataset.

Female Victims: Following male victims, the second most prevalent group consists of female victims. Although the count of female victims is lower than that of males, it still constitutes a noteworthy portion of the reported crimes.

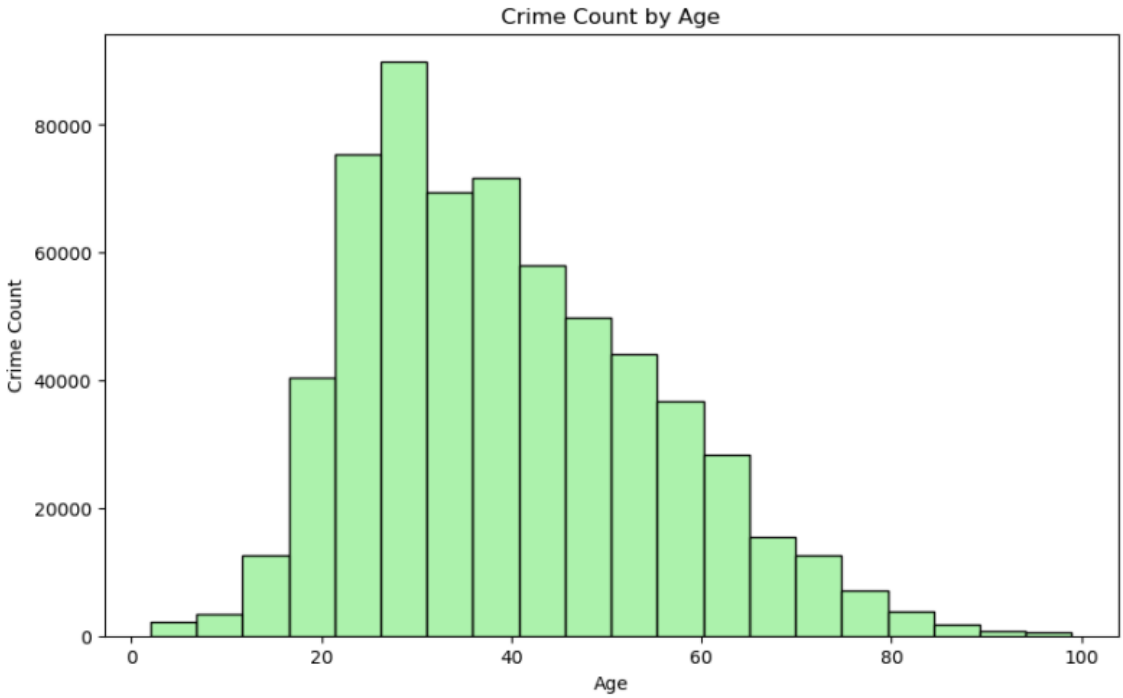


The stacked bar chart gives us information about the number of crimes performed by each gender. From the chart we can infer that the maximum number of crimes has been committed by male and the crimes done by females are relatively less compared to male.



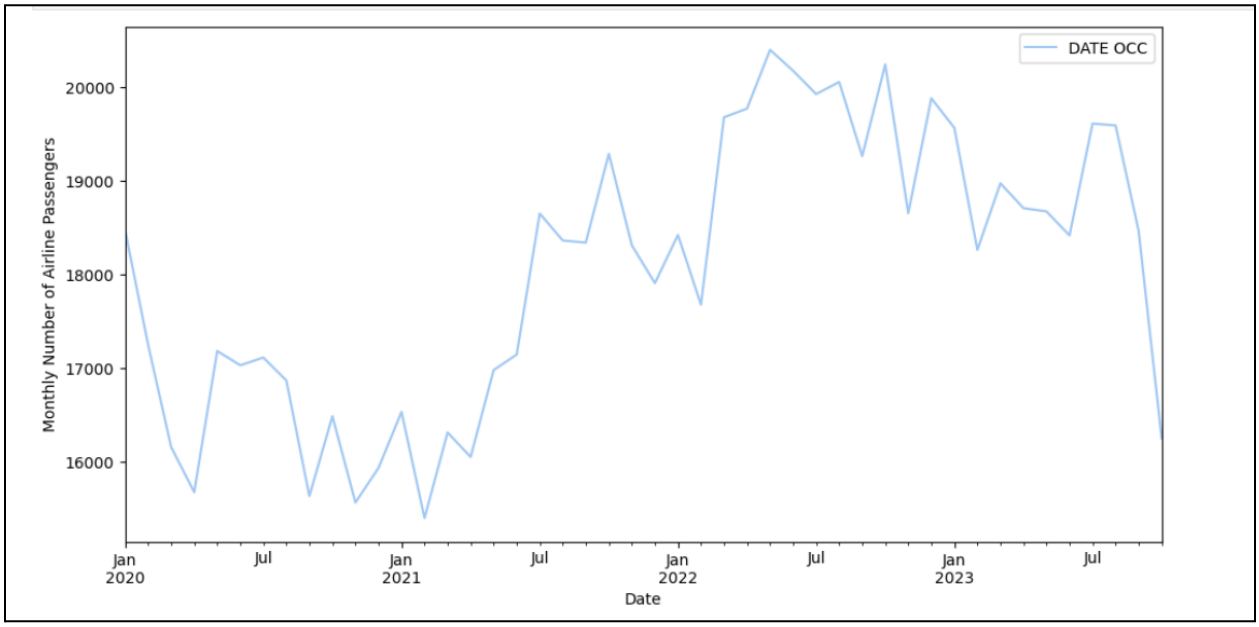
The graph shows that the most common victim descent is Hispanic/Latino/Mexican, followed by White, Black, and Other Asian.

The graph also shows that there has been a steady increase in the number of crimes committed against all victim descents over the past five years. However, the rate of increase has been highest for Hispanic/Latino/Mexican victims.

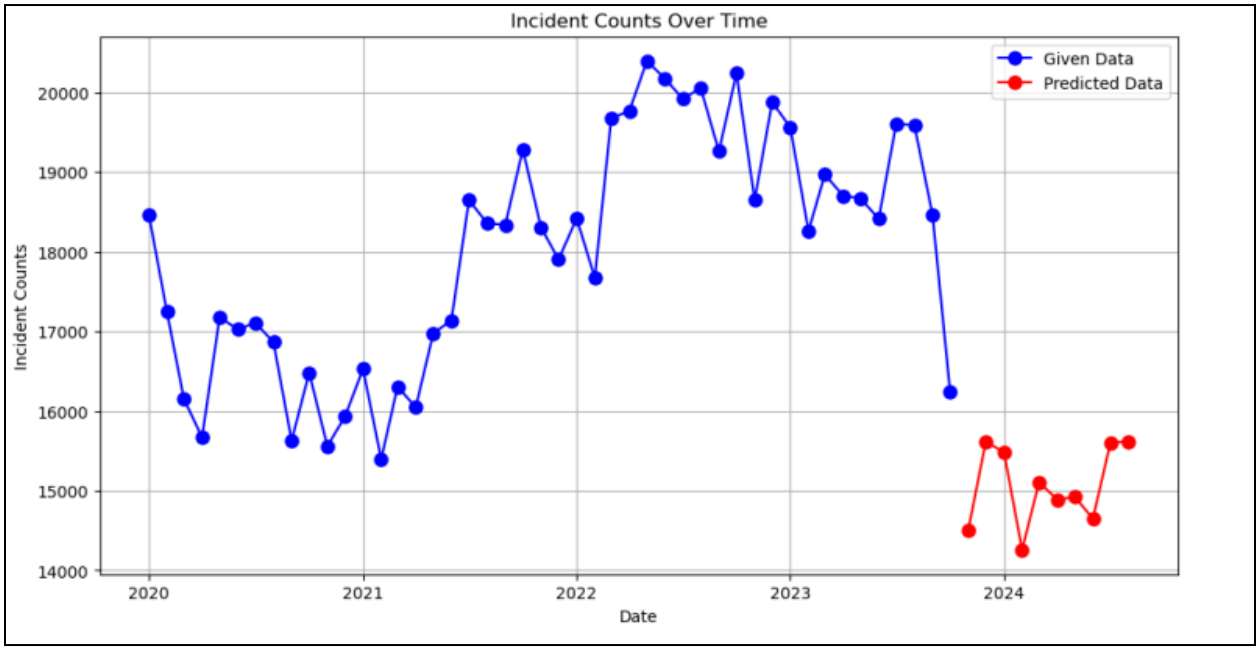


The above graph gives information about the crime rates by age. As it is evident from the graph the crime rate is higher among the people between the ages 20-40, with its peak around the age of 28.

Advanced Analysis:



The above graphical representation tells us about the number of crimes that have happened in Los Angeles city over the past few years. There is a steady rise in the line plot over the years from 2020 to 2022 but there is a decline in the number of crimes in the year 2023.



The predicted line is denoted with red color and blue line is used for representing with given data .The graph shows that the number of reported incidents has been decreasing steadily over time, while the predicted incidents have remained relatively flat. This suggests that the measures that have been taken to reduce incidents are having a positive effect. However, it is important to note that the predicted incidents are still relatively high, so there is still room for improvement.

–

Conclusion:

In conclusion, this project analyzes the crime dataset of the Los Angeles city over the years 2020 to 2023. The dataset was cleaned using necessary steps like identifying missing data and duplicate rows after which we performed steps like data type conversions, handling outliers, and normalized numerical data wherever necessary.

And in the EDA part, we gained a lot of comprehensive insights by analyzing and visualizing the crime datasets , we visualized the crime rates across years and analyzed seasonal impacts on the crime rates, we also found the most common crimes and how the economic factors can influence crime rates. We also identified the temporal influence on crime rates based on days of the week. Analysis of the regional differences on crime rates would help to strategize localized plans. In addition to all this ,we used time series forecasting (predictive modeling technique) to predict the future crime trends. All these valuable insights can be used by the LAPD to better understand and control the crimes occurring in Los Angeles.