Project 1

Cleaning and Analyzing Crime Data

Final Report

Course - IE6400 Foundations Data Analytics Engineering
Section - SEC 01 Fall 2023 [BOS-2-TR]
Group Number - 23
Group Members -

Names	NUID	NU Email
Krishna Priya Gitalaxmi	002699970	gitalaxmi.k@northeastern.edu
Aarathi Saranya Pandravada	002810439	pandravada.a@northestern.edu
Pratham Pawar	002876059	pawar.pratha@northeastern.edu
Nirmit Ajay Sachde	002837262	sachde.n@northeastern.edu
Sri Sai Tarun Vemu	002840565	vemu.s@northeastern.edu

Objective

In this project, we worked with a real-world dataset containing crime data from 2020 to the present (2023) for Los Angeles city. Our goal was to clean and prepare the dataset for analysis, perform exploratory data analysis, and answer specific questions related to crime trends, patterns, and factors influencing crime rates.

Dataset:
We used the crime dataset available at Crime Data from 2020 to Present.

Data Sources

1. We used the crime dataset available at Crime Data from 2020 to Present. This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data.
    a. This dataset includes columns such as:
        i. Report number, date and time of occurrence,
        ii. area where the crime happened,
        iii. type of crime, weapon used if any,
        iv. Victim details,
        v. and the Location.

2. The data for major events was extracted from https://www.latimes.com/ featuring the major events that happened in LA City from 2020 to Present.

3. Population data for calculating crime rates was collected from: https://www.census.gov/quickfacts/fact/table/losangelescountycalifornia,CA/PST045222

4. The employment and income data for the unemployment rate was collected from: https://la.myneighborhooddata.org/
    - Used for correlating economic factors (Unemployment rate and Median household income) with the crime rate

5. The data for LAPD regions, neighborhoods each precinct covers , was extracted from:
https://www.lapdonline.org/

Summary of results:

Analysis of crime data in Los Angeles from 2020 to the present reveals valuable insights into crime patterns, shedding light on the various factors influencing crime rates and the types of crimes that are more likely to occur in the region. Notably, the year 2022 experienced the highest number of recorded crimes, a phenomenon attributed to the gradual return to normalcy after the COVID-19 pandemic.

Crime rates were calculated by considering the number of crimes per day, normalized based on the population per year, with a scaling factor of 1000. An unusual spike in crime rates is observed every New Year, likely due to increased socializing and alcohol consumption during the celebration. The most prevalent crime in Los Angeles is vehicle theft, which consistently ranks as the top reported offense

In a regional analysis of crime in Los Angeles, a visual representation using the LA map clearly distinguishes between crime count and crime rate. When assessing crime count alone, many regions appear to be less safe. However, when factoring in the crime rate, which considers population data, the city Central emerges as an outlier due to its comparatively lower population.

Economic factors emerged as a notable influence on crime rates within the region, with a discernible correlation between higher median household income and lower crime rates. Conversely, regions experiencing an increase in unemployment rates tended to witness an uptick in crime. These crucial insights have been effectively illustrated through visualizations.

A comprehensive day-of-the-week analysis revealed that Friday consistently stood out as the day when a significant number of crimes occurred. Furthermore, our analysis pinpointed the afternoon hours as the most common times for criminal activities to transpire. A visualization capturing the frequency of crimes at specific times on any given day provides a clear depiction of these patterns.

Notably, our data showed a gender disparity in crime commission, with a higher incidence of crimes committed by men compared to women, as indicated by a count plot. However, the crime category "Theft of Identity", exhibited a distinct trend, with a majority of cases involving women as perpetrators.

Regarding age groups, individuals in their 20s were found to be the most common age bracket associated with criminal activities.

A time series forecasting utilizing the Prophet model was performed to predict future crime trends for the upcoming year. The red dots represent these forecasts, and reassuringly, no unusual spikes or alarming patterns were detected in the predictions, suggesting a stable outlook for future crime rates.

Data Cleaning:

We started by cleaning the dataset, which involved the following steps:

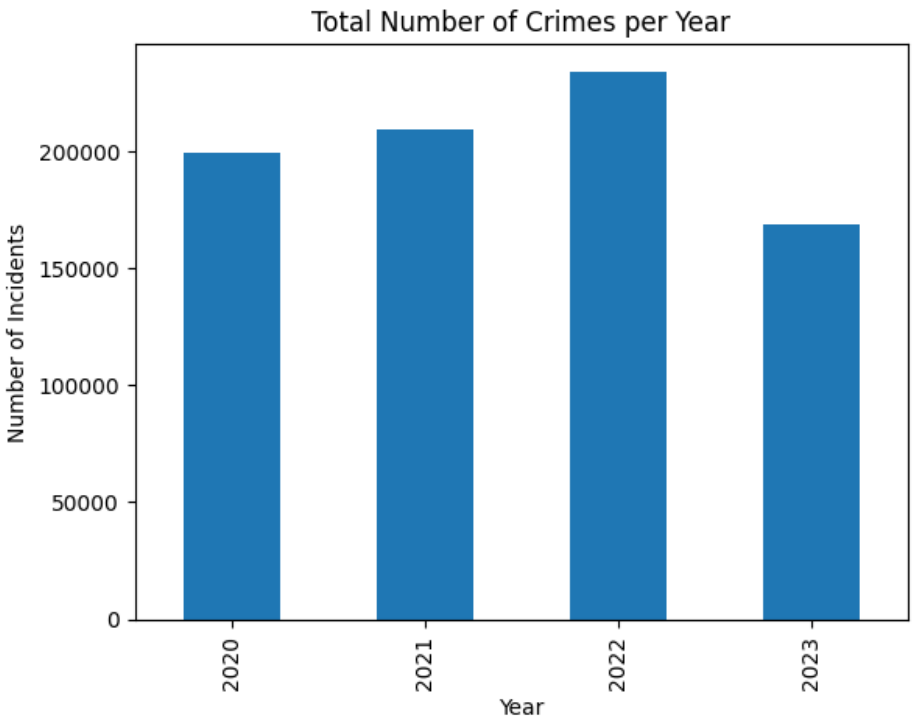
- Formatted date and time columns.
    - We have observed that both the "Date Reported" and "Date Occurred" columns contain time information in their formats. However, it is noteworthy that all time values in these columns are consistently represented as "12:00:00 AM". As a result of this uniformity, we have reached the conclusion that the time component can be safely removed from the date columns.
    - The column labeled "TIME OCC," which records the time at which the crime occurred, initially contained time information in an integer format representing a 24-hour clock time. We have rectified this format by converting it to the more conventional "hour:minute:second" format.

      AMAZON REVIEW DATASET
- Handled missing data.
    - For the “Vict Sex” column, the missing and the unwanted data was replaced by the accepted values i.e., “M”, “F” and “X” in the ratio they were present in the dataset.
    - For the “Vict Age” column, there were more than a lakh “0” and negative instances which were initially handled by filling them with the mean value as per standards, later on it was observed that the mean age stood out and gave a skewed picture of the actual scenario,Hence we decided on filtering out the inconsistent values for the purpose of analysis and visualization.

- Removed duplicate rows.
    - Removed all the duplicate rows from the dataset
- Standardized data types.
- Dealt with outliers.
- Encoded categorical data.

Findings

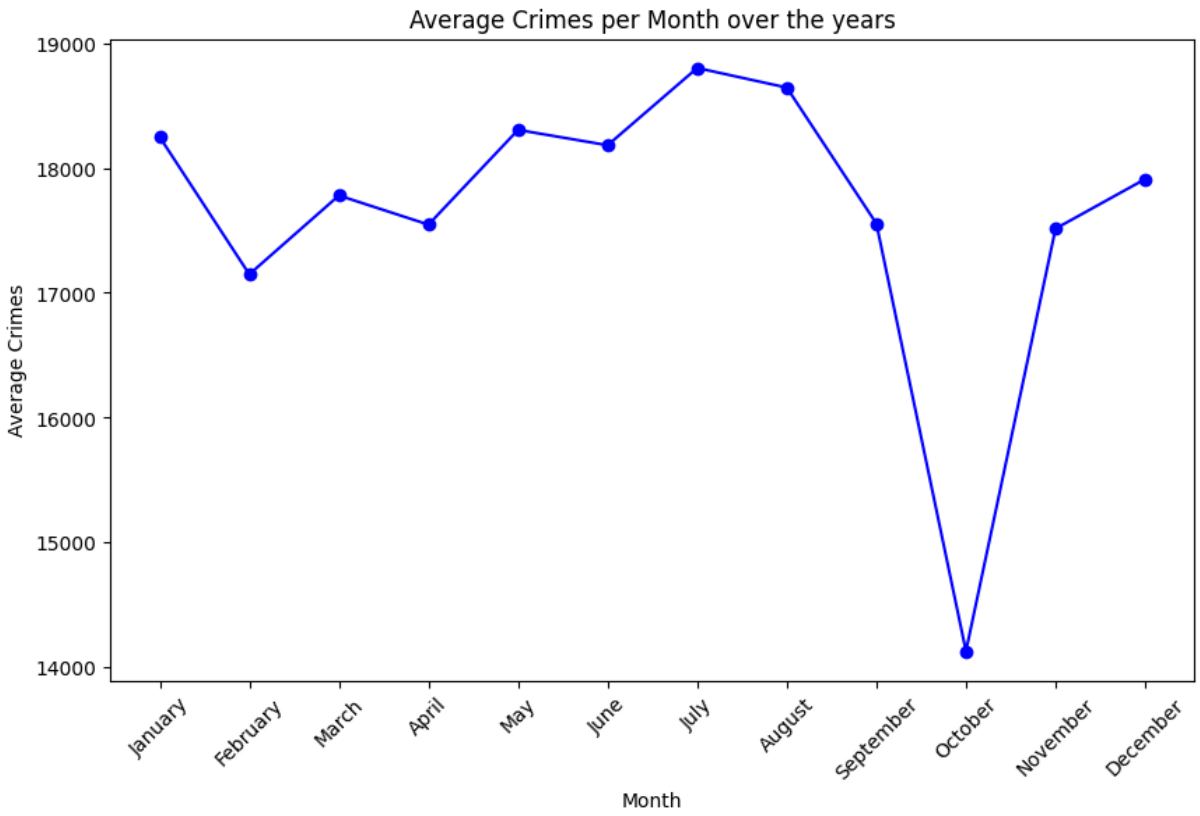
1. **Overall Crime Trends:**
    - We calculated and plotted the total number of crimes per year to visualize the trends.



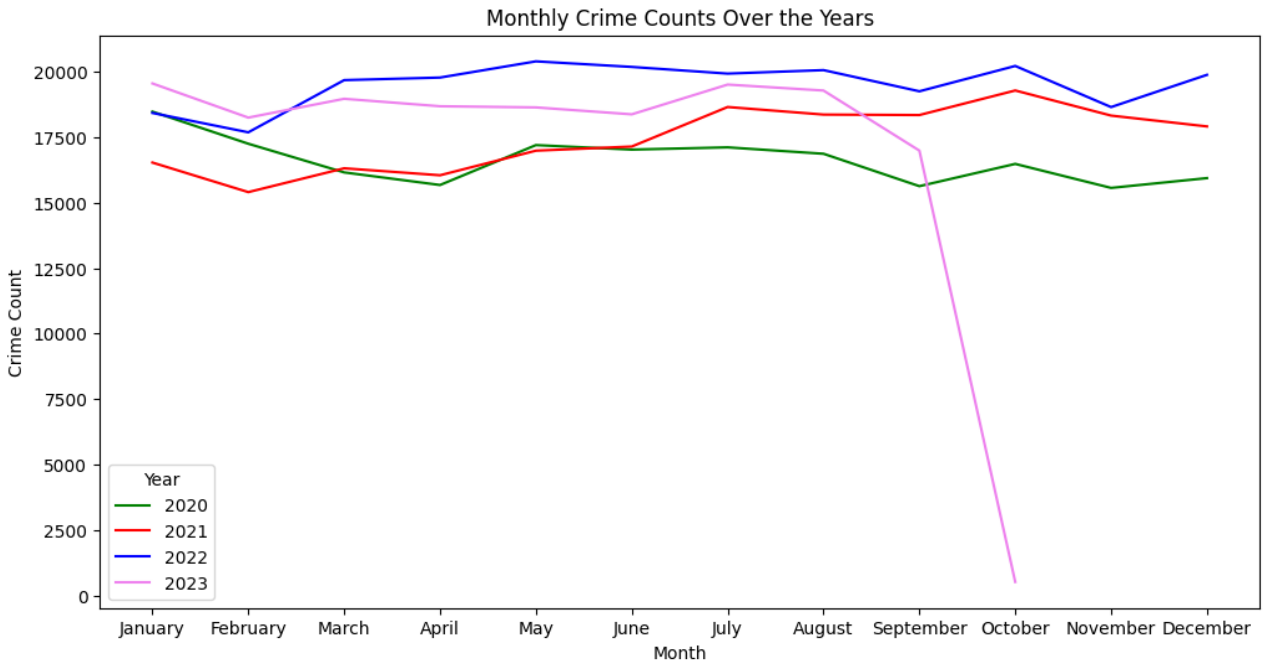
- The bar graph provides a picture of reported crime counts over a span of four years, implying a significant shift in the number pre covid and post covid. Notably, a substantial increase in crime occurrences transpired after 2021, with the number of reported incidents witnessing a peak.
- A closer examination reveals that the year 2022 crossed the threshold of 200,000 reported crimes. By the midway point of 2023, we find that the number of incidents remained high, hovering at nearly 150,000, which signals an ongoing challenge in maintaining public security. In contrast, the years 2020 and 2021 displayed a rather similar pattern in the total number of crimes reported annually, suggesting a stable trend during this period.

2. **Seasonal Patterns:**

- Seasonal patterns were analyzed by grouping the data by month and calculating the average number of crimes per month over the years.
- A line plot was created to visualize the average crimes per month over the years.

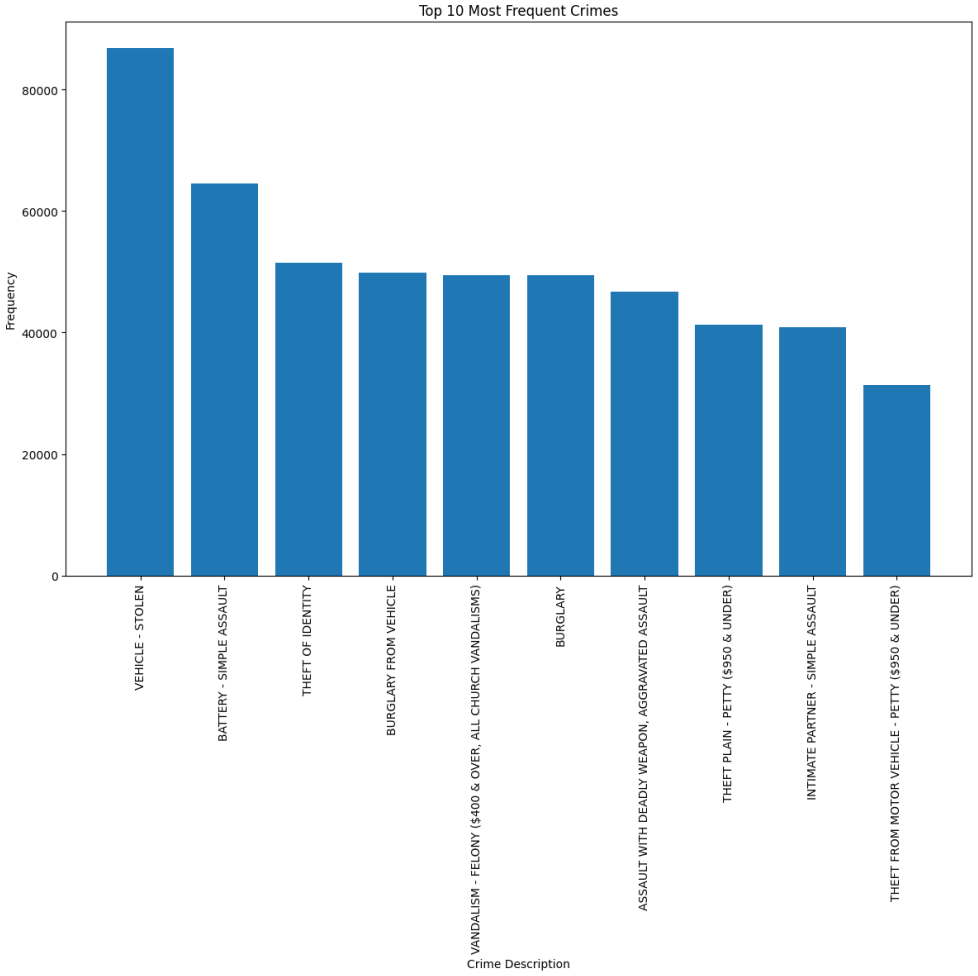


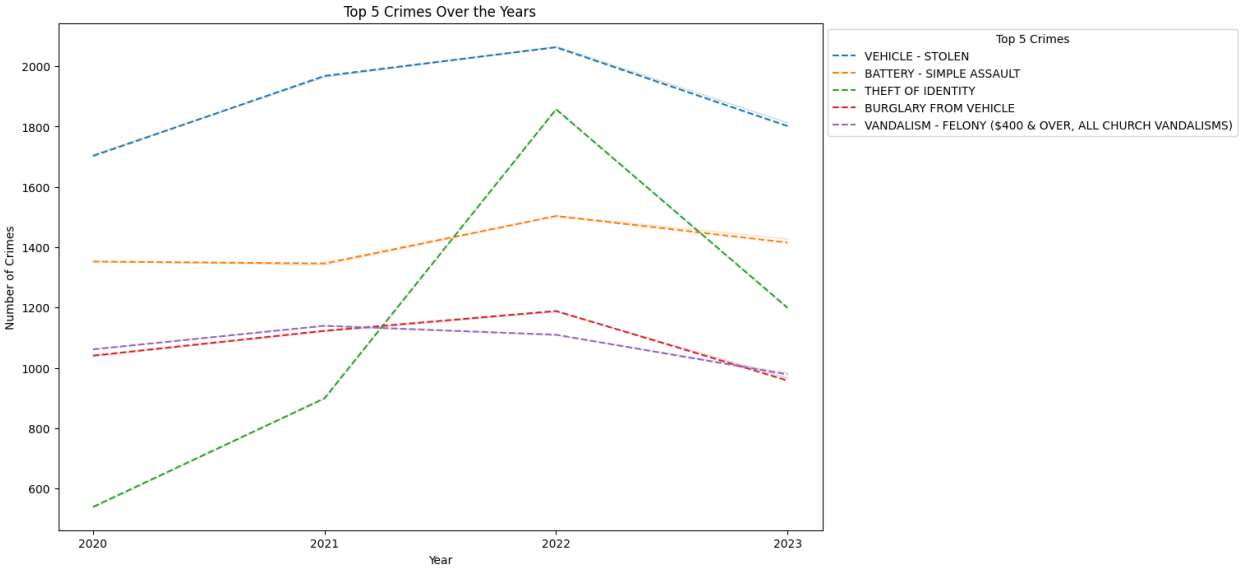
- Seasonal crime patterns can be effectively illustrated by grouping the data on a monthly basis and analyzing the average number of crimes occurring each month over multiple years. A line plot is a suitable visualization method for tracking changes in crime rates from one month to the next.
- Upon examination of the line plot, it is evident that the crime rate peaks in July, which corresponds to the warmest month of the year. Consequently, the summer season experiences the highest incidence of criminal activities. The noticeable steep decline in the line graph during October is not indicative of a decrease in crime rate, but rather due to the fact that the data for October 2023 is incomplete, as the month is still in progress.
- To gain a more comprehensive understanding of seasonal crime patterns, we created a line plot that depicts crime counts for each year. This visualization clearly reveals that the crime rate tends to be higher during the summer and autumn months, while it is comparatively lower during the winter and spring months.



3. **Most Common Crime Type:**

- We counted the occurrences of each crime type and identified the one with the highest frequency.
    - The most common crime is with Crime Code: **510**
    - Description: **VEHICLE - STOLEN**
    - Frequency: **86816**
- We also plotted a bar graph for top 10 crimes occured, they were
    - VEHICLE - STOLEN
    - BATTERY - SIMPLE ASSAULT
    - THEFT OF IDENTITY
    - BURGLARY FROM VEHICLE
    - VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...
    - BURGLARY
    - ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
    - THEFT PLAIN - PETTY ($950 & UNDER)
    - INTIMATE PARTNER - SIMPLE ASSAULT
    - THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)


Top 10 Most Frequent Crimes
Frequency
0
20000
40000
60000
80000
VEHICLE - STOLEN
BATTERY - SIMPLE ASSAULT
THEFT OF IDENTITY
BURGLARY FROM VEHICLE
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
BURGLARY
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
THEFT PLAIN - PETTY ($950 & UNDER)
INTIMATE PARTNER - SIMPLE ASSAULT
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
Crime Description

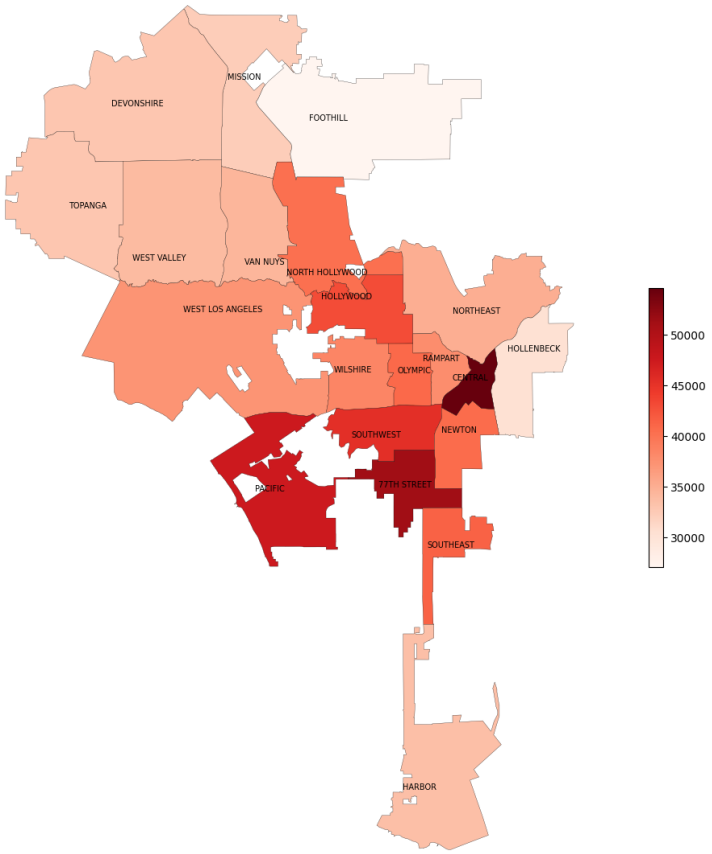

Top 5 Crimes Over the Years
Number of Crimes
600
800
1000
1200
1400
1600
1800
2000
2020
2021
2022
2023
Year
Top 5 Crimes
VEHICLE - STOLEN
BATTERY - SIMPLE ASSAULT
THEFT OF IDENTITY
BURGLARY FROM VEHICLE
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)

The graph helps to identify and visualize the temporal patterns of the top 5 crimes over the years. It allows for the comparison of how the frequency of these crimes has changed annually.
VEHICLE - STOLEN has consistently been the highest occuring crime over 3 years followed by BATTERY- SIMPLE ASSUALT at the second position and BURGLARY FROM VEHICLE and VANDALISM going side by side at the third position.
But the crime THEFT OF IDENTITY took a huge jump, from being the lowest occurred crime in 2020 to being the highest occured crime in 2022 and then again dipping down in 2023.

4. **Regional Differences:**

- **Region Data:**
    - The code merges the LAPD division shapefile data with the crime count data by the area name ('APREC' column).
    - It then creates a geospatial visualization of the crime count using matplotlib, with different areas shaded in colors based on crime counts.

Region Wise Distribution of Crime Count across Los Angeles



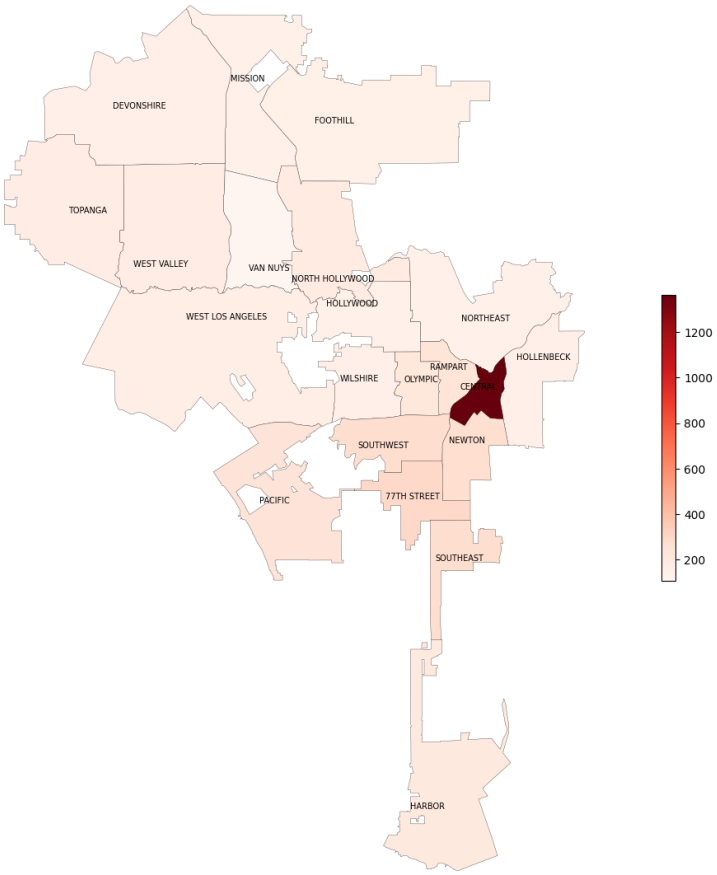
- Central has the largest count of crimes (54556 in the year 2021) followed by 77th street and Pacific regions.
- But central appears to be an outlier when you consider the population of each region into account and calculate the crime rate.

- **Population Data:**
  - The population data is merged with the existing shapefile data based on the 'APREC' column, ensuring that population figures are associated with each area.

- **Crime Rate Calculation:**
  - Two new columns are added to the shapefile DataFrame: 'crime rate per thousand' and 'crime rate.' These columns represent crime rates per thousand people and crime rates (as a fraction of the population), respectively, for each area.

- **Crime Rate Visualization:**

Region Wise Distribution of Crimes Rates(per thousand people) across Los Angeles

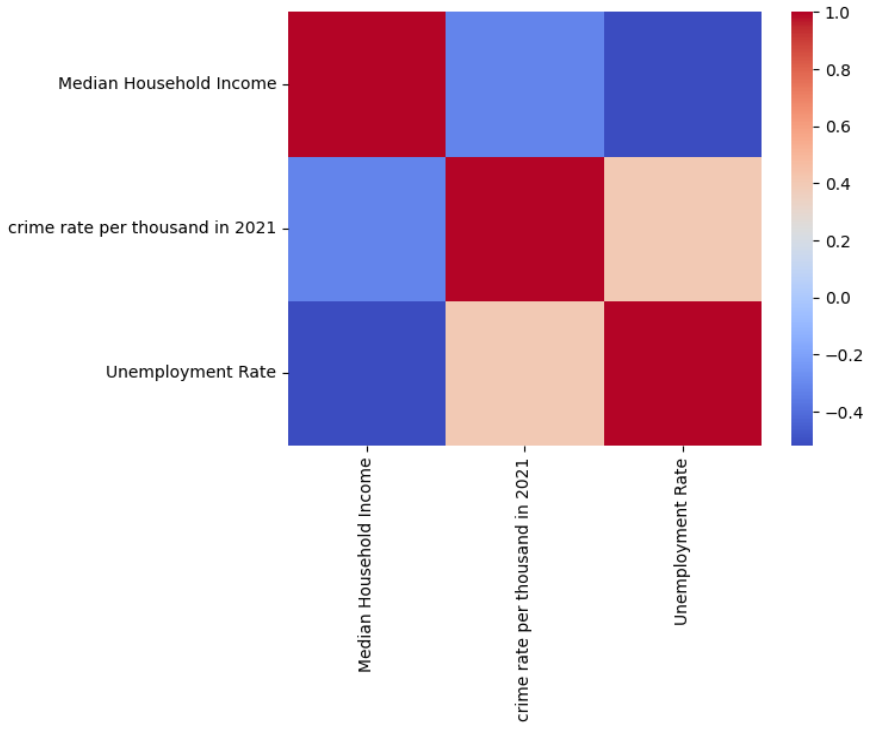


- A second geospatial visualization is created, this time showing crime rates per thousand people in different areas of Los Angeles.
- Crime rate=(crime count in the area / population of the area) * scaling factor
- The scaling factor we have considered is 1000, which gives us the crime rate per thousand people
- The central region is seen to have an extremely high crime rate since it is a less populous neighborhood.

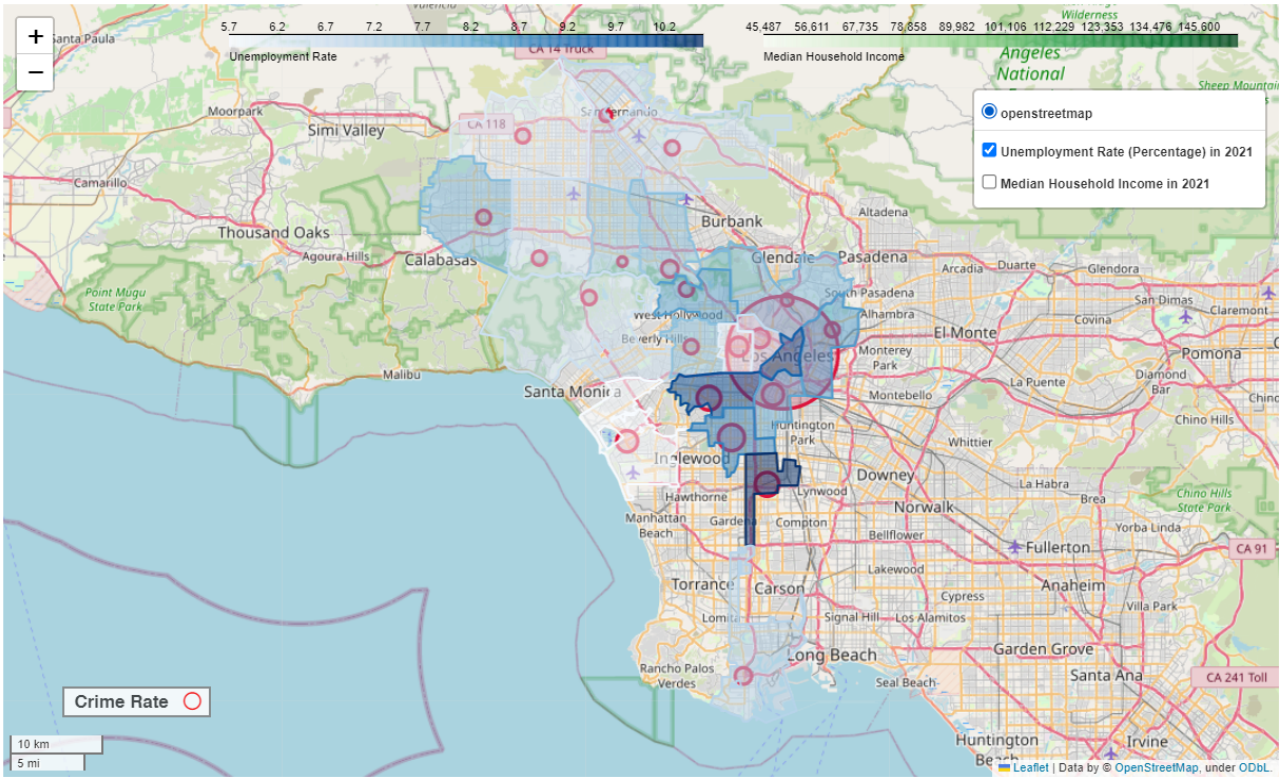
5. Correlation with Economic Factors:

- We collected economic data for the same time frame and assessed the relationship between economic factors and crime rates through correlation analysis.
- The two economic factors we have considered are
  - 1. Median House Income
  - 2. Unemployment Rate

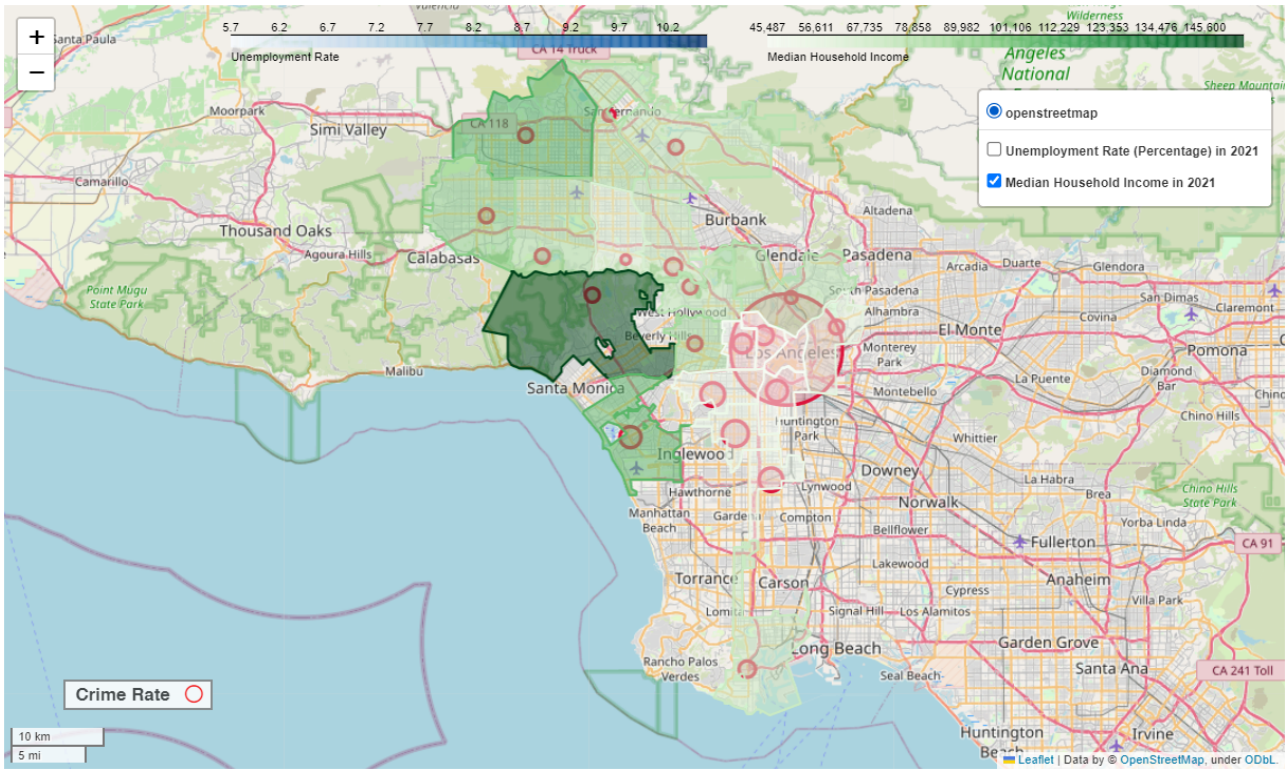
	Median Household Income	crime rate per thousand in 2021	Unemployment Rate
Median Household Income	1.000000	-0.318089	-0.519101
crime rate per thousand in 2021	-0.318089	1.000000	0.403328
Unemployment Rate	-0.519101	0.403328	1.000000



- Crime rate vs Unemployment Rate: correlation value= 0.403328
  It shows a weak positive correlation, which means as Unemployment rates increase across regions the Crime rate is seen to increase as well.
- Crime rate vs Median Household Income: correlation value= -0.519101
  It shows a strong negative correlation, which means in our context the areas with higher Median Household Income show lower crime rates.
- Unemployment Rate vs Median Household Income: correlation value: -0.318089
  There is a weak negative correlation between Unemployment Rates and the Median Household Income.



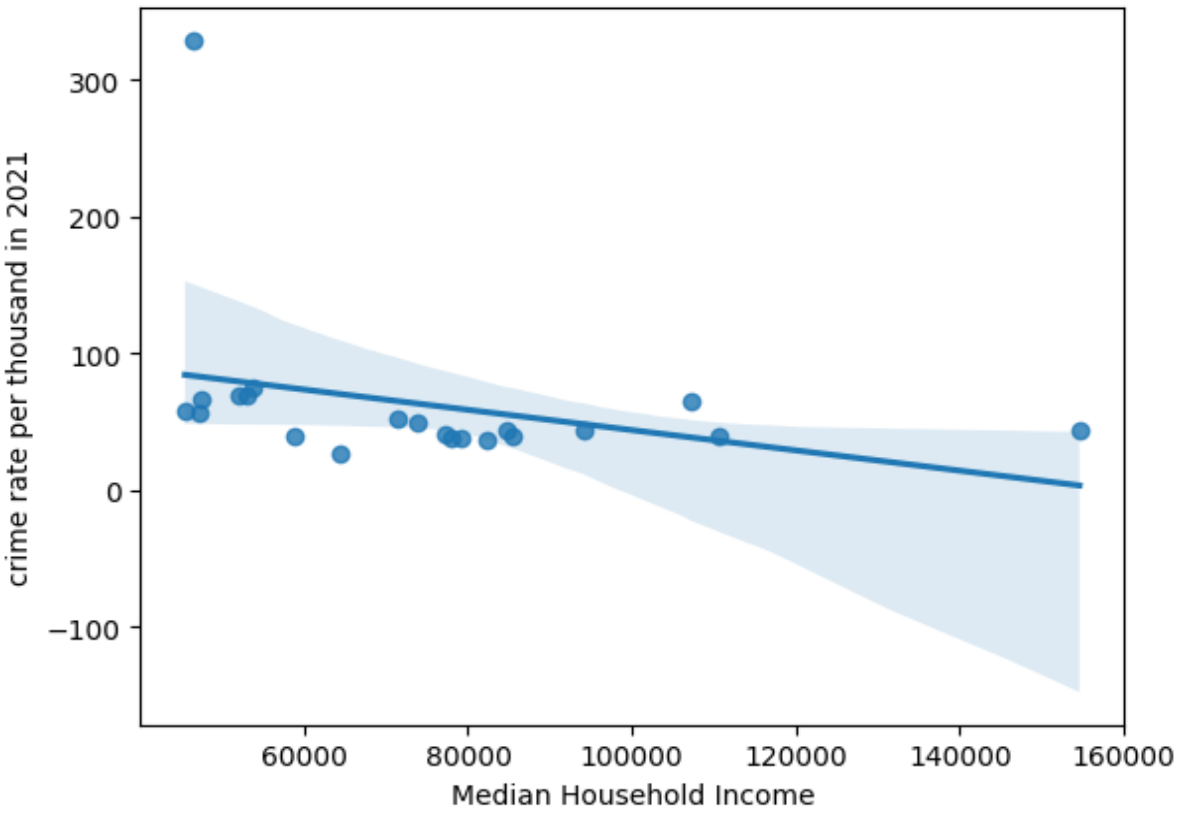
Choropleth Map Showing Unemployment Rate vs Crime Rate

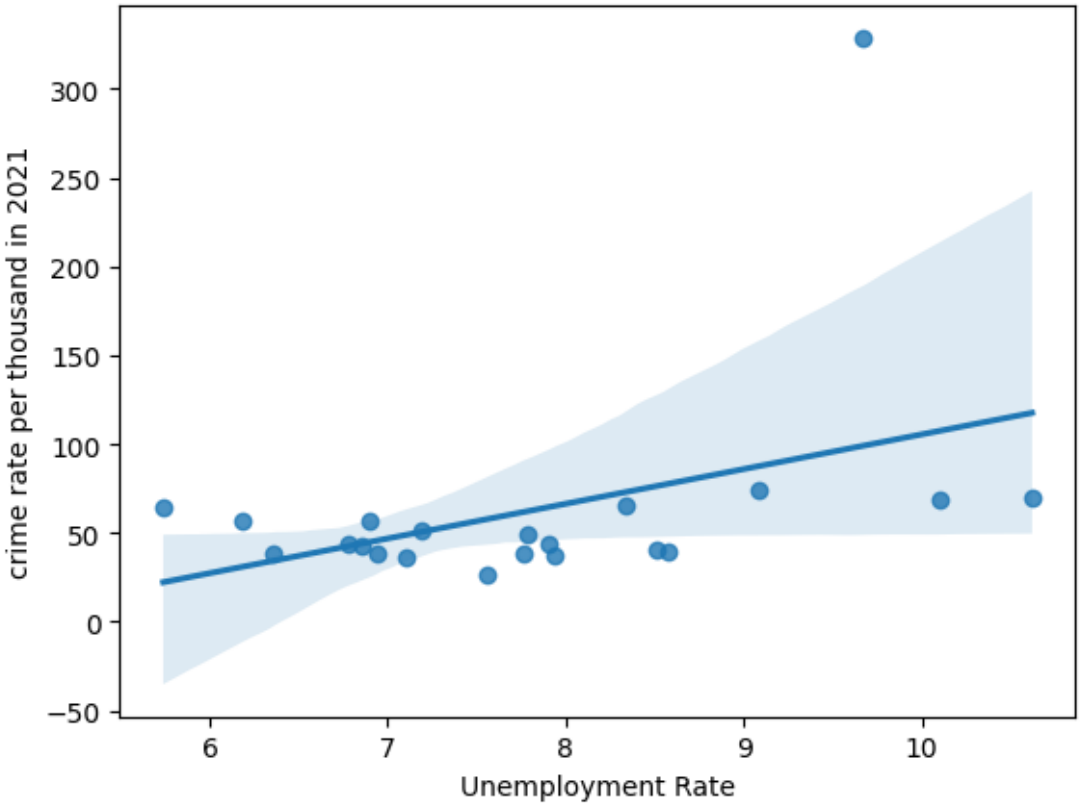


Choropleth Map Showing Median Household Income vs Crime Rate

Region Wise Distribution of Crime count and Crime rate per thousand people:

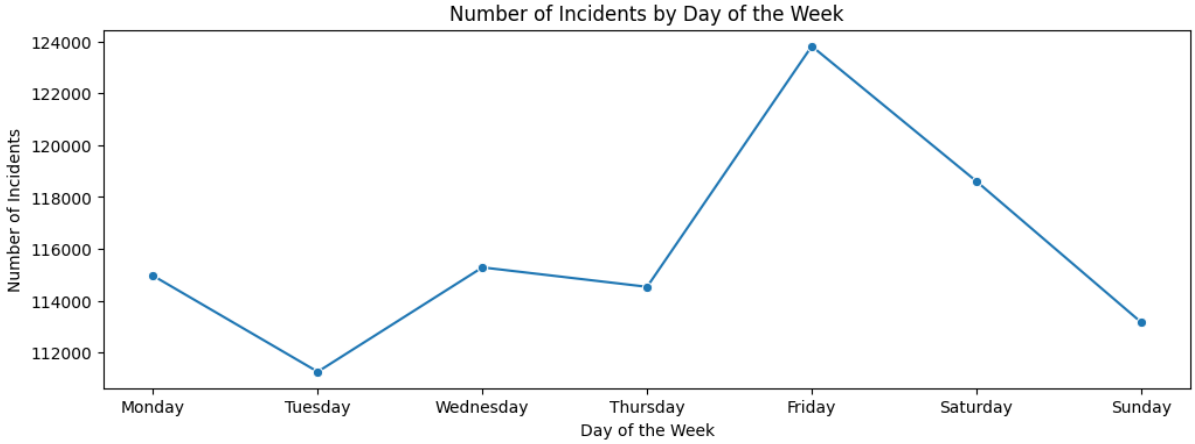
Central registered the highest count of crimes in the year 2021, totalling 54,556 incidents, followed by the 77th Street and Pacific regions. However, when adjusting for the population of each region and calculating the crime rate using the formula Crime Rate = (Crime Count in the Area / Population of the Area) * Scaling Factor, where the chosen scaling factor is 1000 (resulting in the crime rate per thousand people), Central appears as an outlier due to its relatively smaller population. Consequently, Central exhibits an exceptionally high crime rate, which is primarily attributed to its lower population in comparison to other neighborhoods.


crime rate per thousand in 2021
300
200
100
0
−100
60000
80000
100000
120000
140000
160000
Median Household Income

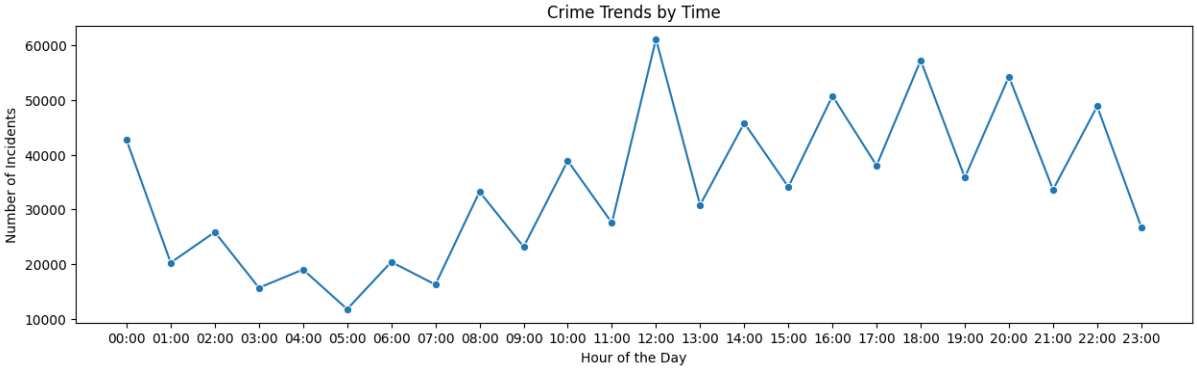

crime rate per thousand in 2021
300
250
200
150
100
50
0
−50
6
7
8
9
10
Unemployment Rate

6. **Day of the Week Analysis:**

- We grouped the data by the day of the week and analyzed crime frequencies for each day.

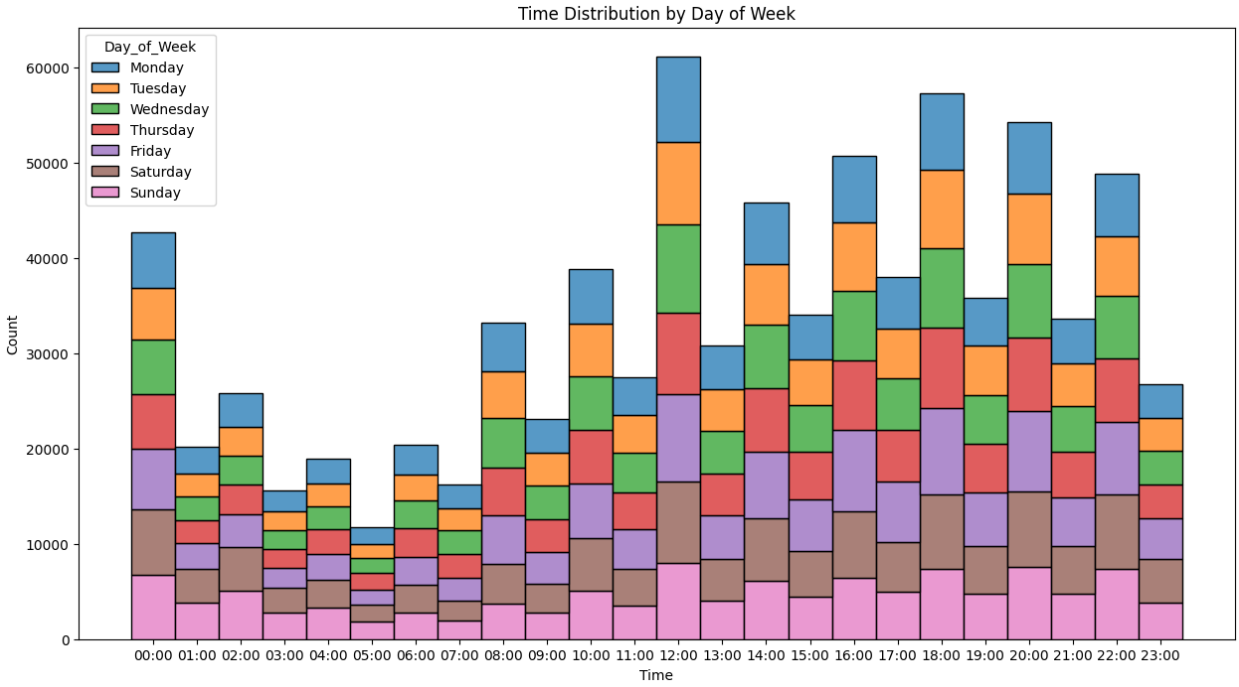


- This line graph shows the crime frequencies for each day (Monday through Sunday). It allows us to observe patterns and trends during the typical day of the week. It is observed that over the week highest number of crimes are comitted on Fridays and the least on Tuesday.



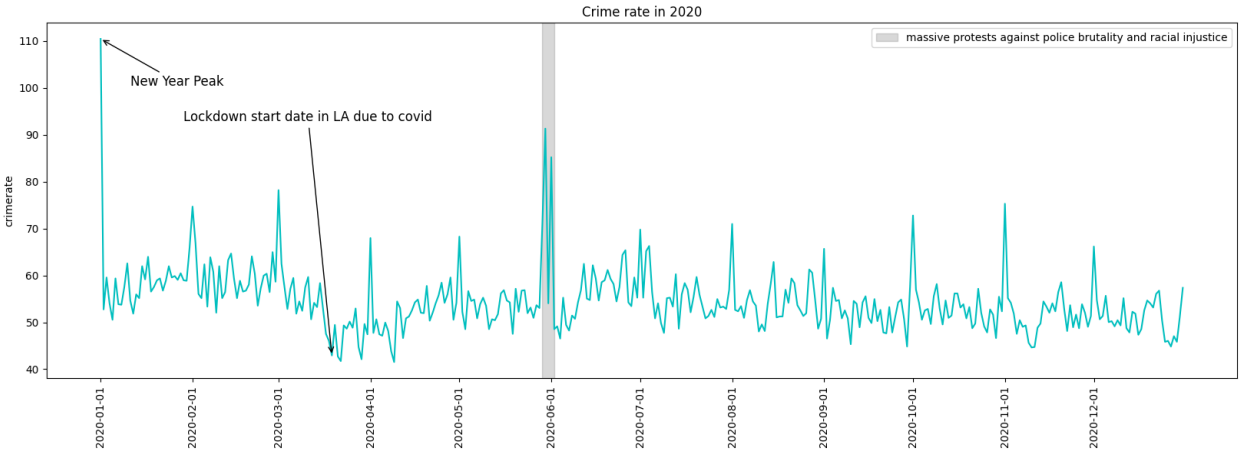
- It has been noted that the peak frequency of crimes occurs around noon during a day.
- The overall incidence of crimes is lower during the daytime (from 05:00 to 12:00), reaching its maximum at noon, and then rising during the afternoon, evening, and nighttime (until 01:00).
- Crime occurrences decrease during the late night hours.

The stacked bar chart in the below shows you the crime distribution by day of week with time. The visualization gives an idea about the frequency of crimes on a given day at a given time.



7. **Impact of Major Events:**
   - We identified major events or policy changes during the dataset period and analyzed crime rate changes before and after these events.

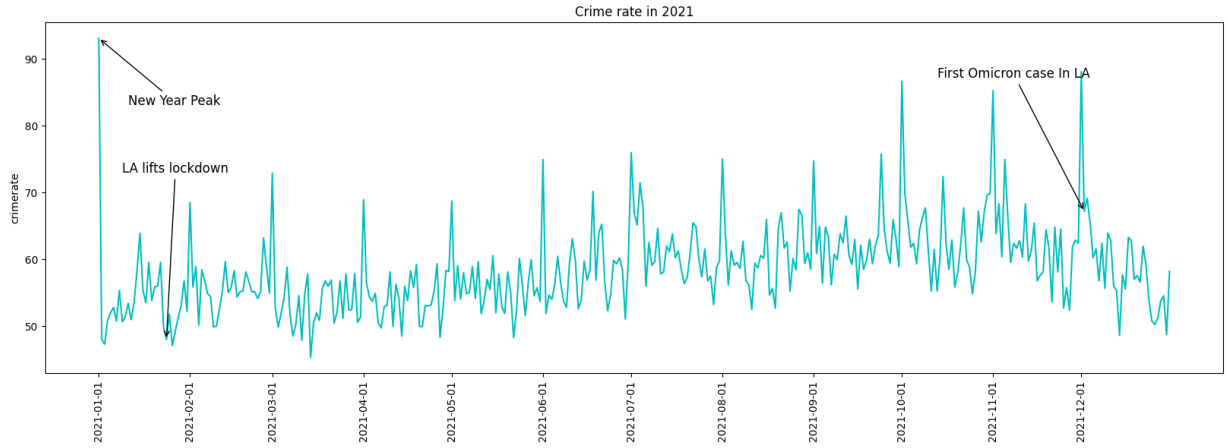
   - In 2020



   - A distinctive and recurring peak in the crime rate is observed every New Year's Day. This peak, which is close to 100, quickly returns to normal levels shortly after the holiday.

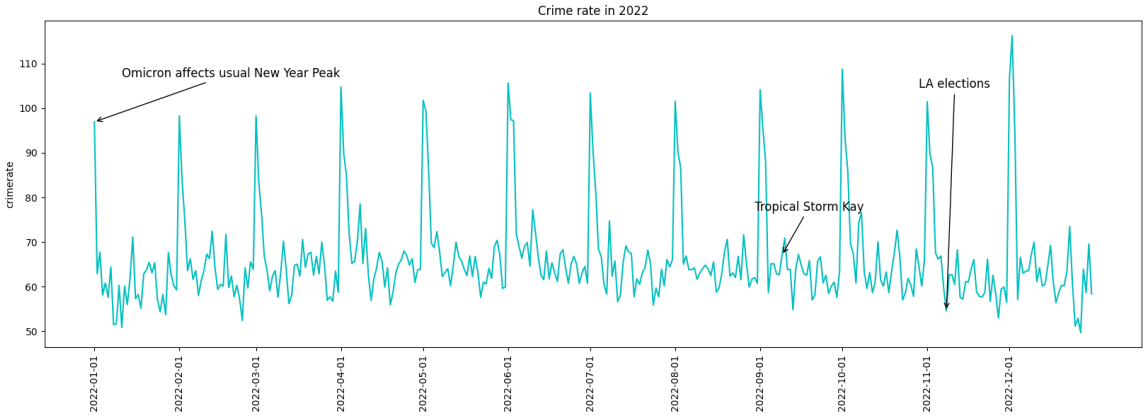
- The onset of the COVID-19 lockdown in Los Angeles on March 19, 2020, had a significant impact on the crime rate. Prior to the lockdown, the average crime rate exceeded 60, but it subsequently decreased to the 40s and remained at this lower level.
- In June 2020, Los Angeles witnessed massive protests against police brutality and racial injustice, following the death of George Floyd in Minneapolis. These protests began on May 29, 2020, and escalated to the point where Mayor Eric Garcetti declared a curfew for the city on June 2, 2020. This period saw notable fluctuations in the crime rate, likely influenced by the unrest and curfew imposed during the protests.

- In 2021



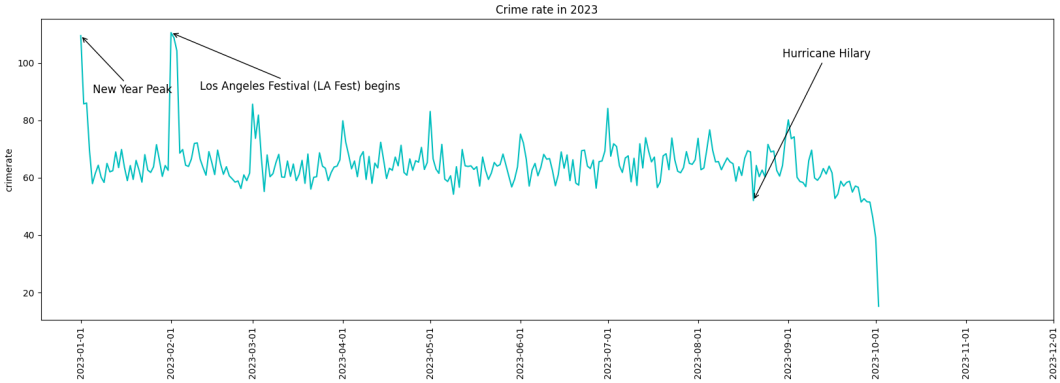
- In the year 2021, a notable peak in the crime rate, reaching close to 100, is clearly visible on New Year's Day. This spike in crime rate on this holiday is a recurring pattern.
- An evident increase in the crime rate was observed immediately after the removal of the lockdown in Los Angeles on January 24, 2021. This increase suggests a correlation between the relaxation of lockdown measures and an uptick in criminal activities.
- Following the first reported case of the Omicron variant in Los Angeles on December 2, 2021, a significant decline in the crime rate is noticeable. This suggests a potential relationship between the emergence of the Omicron variant and a decrease in the crime rate, possibly due to heightened public health measures and restrictions in response to the variant's spread.

- In 2022



- In the year 2022, the recurring pattern of a temporary spike in the crime rate reaffirms the influence of the New Year's holiday on the crime rate.
- The occurrence of Tropical Storm Kay on September 9, 2022, had a notable impact on the crime rate, causing it to reach a downward peak. Severe weather events can disrupt normal activities and affect crime rates.
- Throughout the year 2022, there are multiple peaks in the crime rate. These peaks are likely associated with various events such as games, award functions, and shows that resumed in full swing after the easing of COVID-19 restrictions. Such large-scale events can contribute to fluctuations in the crime rate.
- In the Los Angeles elections of 2022, which took place on November 8, 2022, a new mayor, Mayor Black, was elected. Mayor Black holds the distinction of being the first African woman to assume the mayoral office in Los Angeles. Her introduction of new policies aimed at addressing homelessness appears to have had a significant impact on the crime rate. Subsequent months show fewer peaks, suggesting that her policies may be contributing to a reduction in certain types of criminal activities in the city.

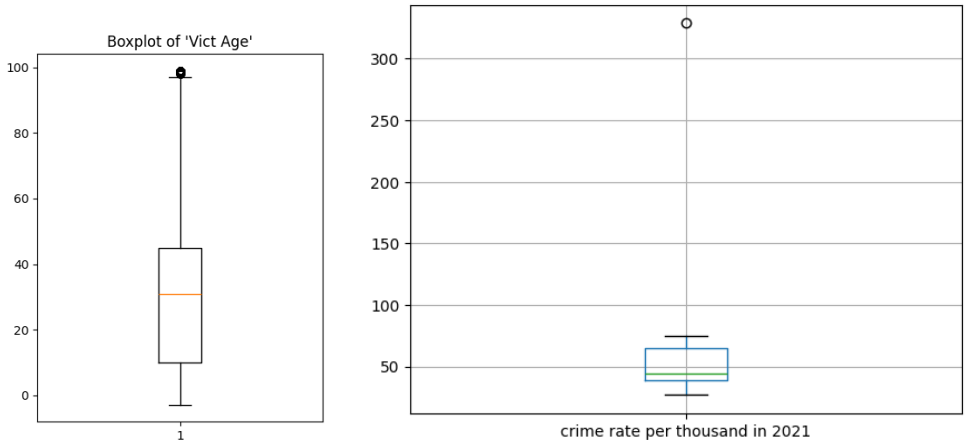
- In 2023



- In the year 2023, the New Year peak in the crime rate is exceptionally high, surpassing 100. This is a notable and recurring pattern observed during the New Year holiday.
- There is a discernible spike in the crime rate on the same day that the LA Fest began on February 1, 2023. The commencement of such events can sometimes coincide with an increase in certain types of criminal activities.
- The occurrence of Hurricane Hilary leads to a significant downward peak in the crime rate. Severe weather events like hurricanes can disrupt normal activities, including criminal activities, and result in a temporary decrease in the crime rate.

8. Outliers and Anomalies:

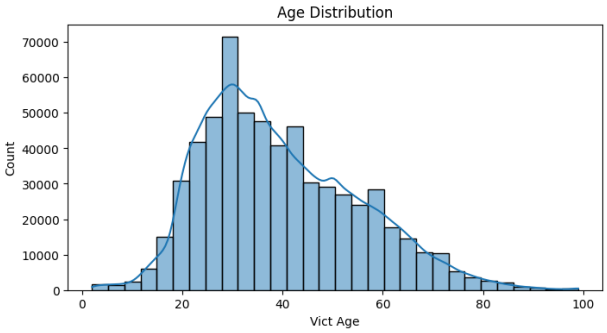
- We used statistical methods and data visualization techniques to identify dataset outliers and investigate unusual patterns.

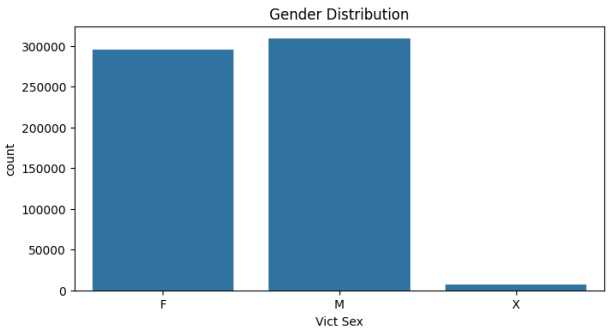


- In identifying outliers in the provided crime dataset, "victim age" column has notable outliers, alongside the crime rates column, which was generated during the processing of data.
- Upon close examination, it becomes evident that individuals aged above 20 up to 40 years constitute the most frequent age group among crime victims. Those under the age of 20 make up the second-highest group of crime victims. Notably, a very small portion of individuals above the age of 90 falls into the outlier category.
- In the crime rate per thousand in 2021, the outlier is the "central region" where the crime rate is abnormally high because of the extremely smaller population compared to other regions.

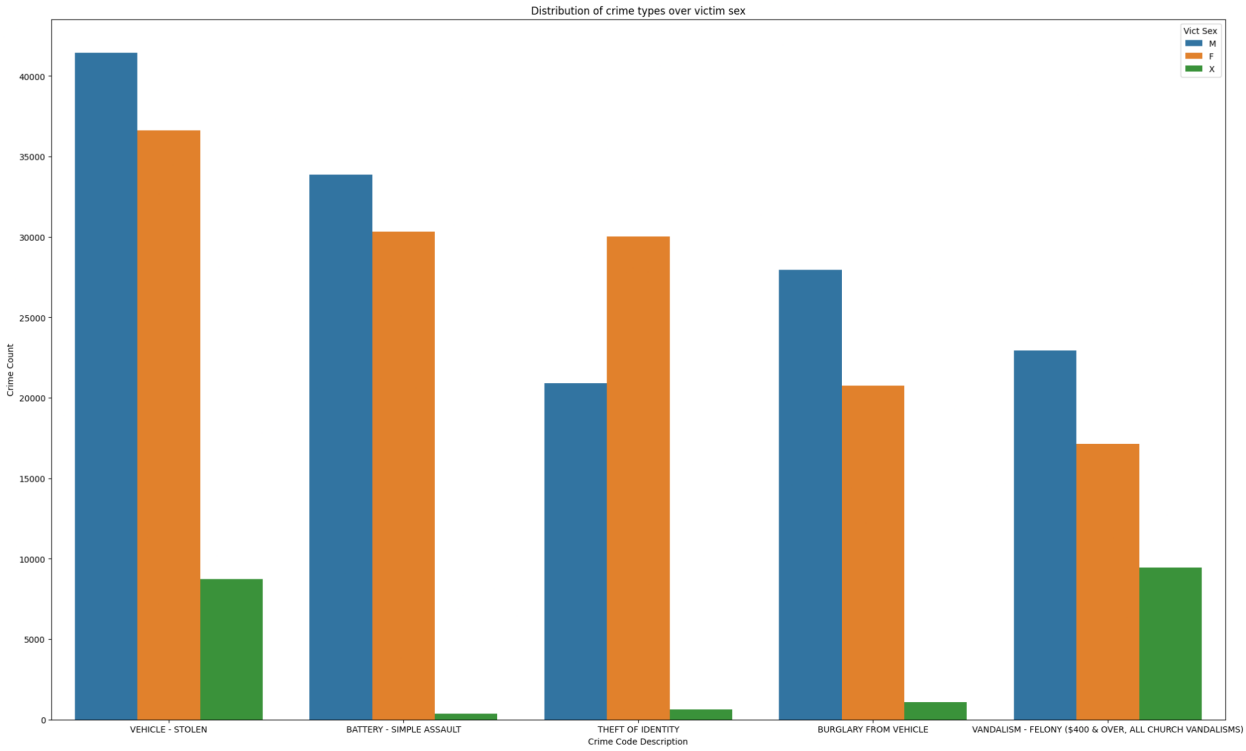
9. **Demographic Factors:**

- We analyzed the dataset to identify any patterns or correlations between demographic factors (e.g., age, gender) and specific types of crimes.

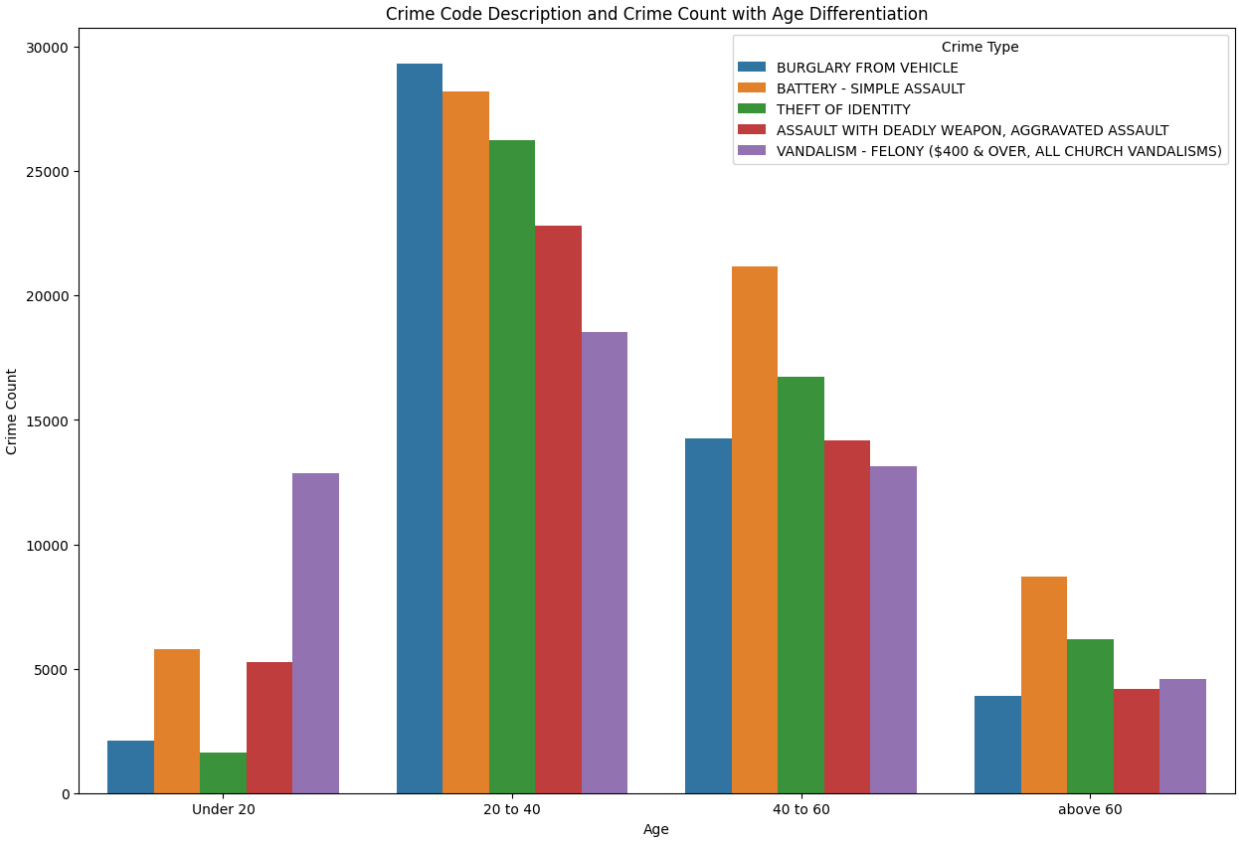




- The age distribution with crime count represents the frequency of criminal incidents across different age groups. This analysis provides valuable insights into the correlation between age and criminal activity within a given dataset. The age distribution is typically visualized using a histogram or a bar chart, with age groups on the x-axis and corresponding crime counts on the y-axis. It is seen that there are a lot of 20 to 40 aged group people as victims.
- The gender distribution with crime count provides a comprehensive overview of the prevalence of criminal incidents across different genders within a given dataset. This analysis is crucial for understanding the relationship between gender and criminal activity, shedding light on potential patterns, trends, and disparities in the commission of crimes. Typically visualized through a bar chart or pie chart, with genders represented on the x-axis and corresponding crime counts on the y-axis, this analysis allows for a clear comparison of criminal involvement between different gender groups. We found that there are almost same but a few more in male victims.



- The analysis of the distribution of the top 5 crime types over victim sex provides valuable insights into the patterns and variations of criminal incidents based on the gender of the victims. This examination helps to discern the prevalence of specific crime types among different genders, shedding light on potential gender-specific vulnerabilities and crime dynamics.
- Typically visualized through stacked bar charts or grouped bar charts, with crime types on the x-axis, the count of crimes on the y-axis, and separate bars or color-coding for each victim sex, this analysis enables a comparative exploration of how the top 5 crime types manifest across male, female, and potentially other gender categories.

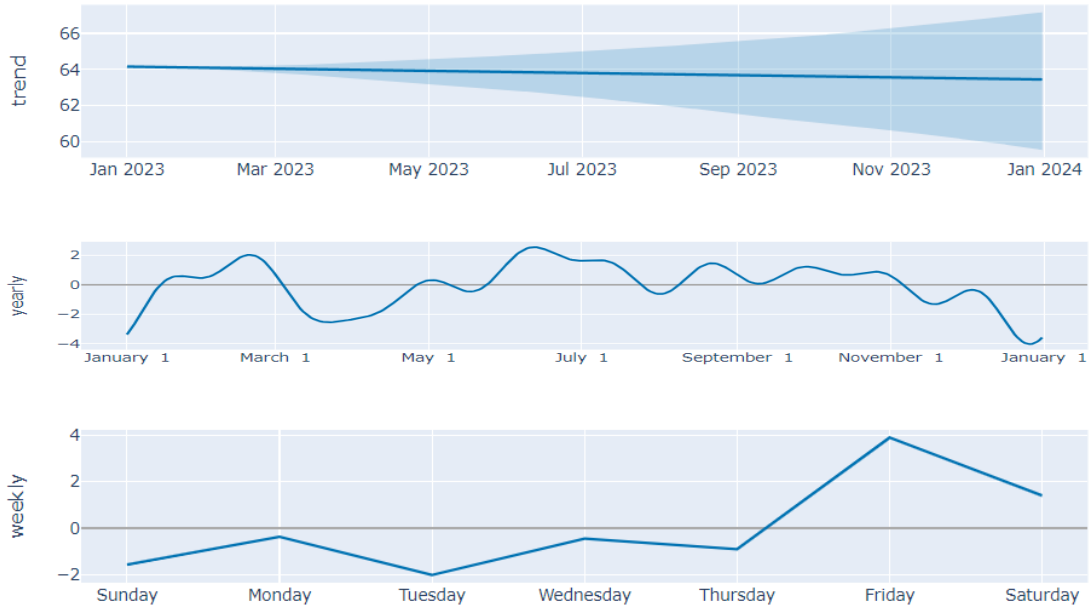


- Analyzing the distribution of the top 5 crime types over victim age provides a comprehensive overview of how specific criminal incidents are distributed across different age groups. This examination offers insights into the dynamics of criminal victimization, revealing potential trends and variations in the prevalence of certain crime types among various age demographics.
- Visualized through histograms or grouped bar charts, with age groups on the x-axis, crime types on the y-axis, and separate bars or color-coding for each crime type, this analysis facilitates a comparative exploration of how the top 5 crime types manifest across different age categories.

10. Predicting Future Trends:

- We employed time series forecasting methods, such as Prophet, to predict future crime trends based on historical data.

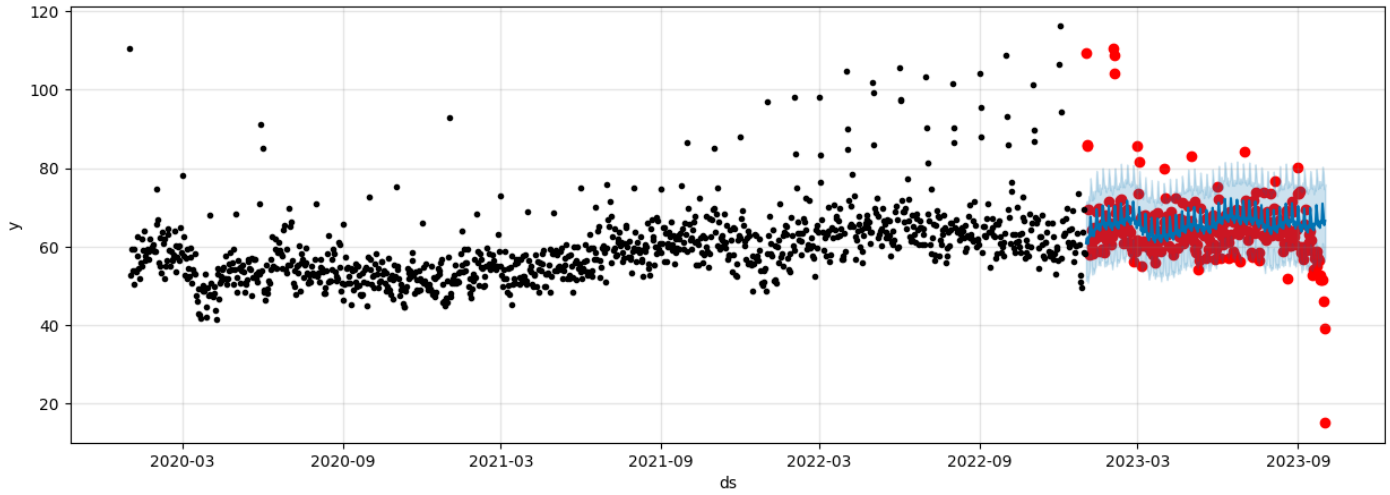
Trends yearly, monthly and weekly

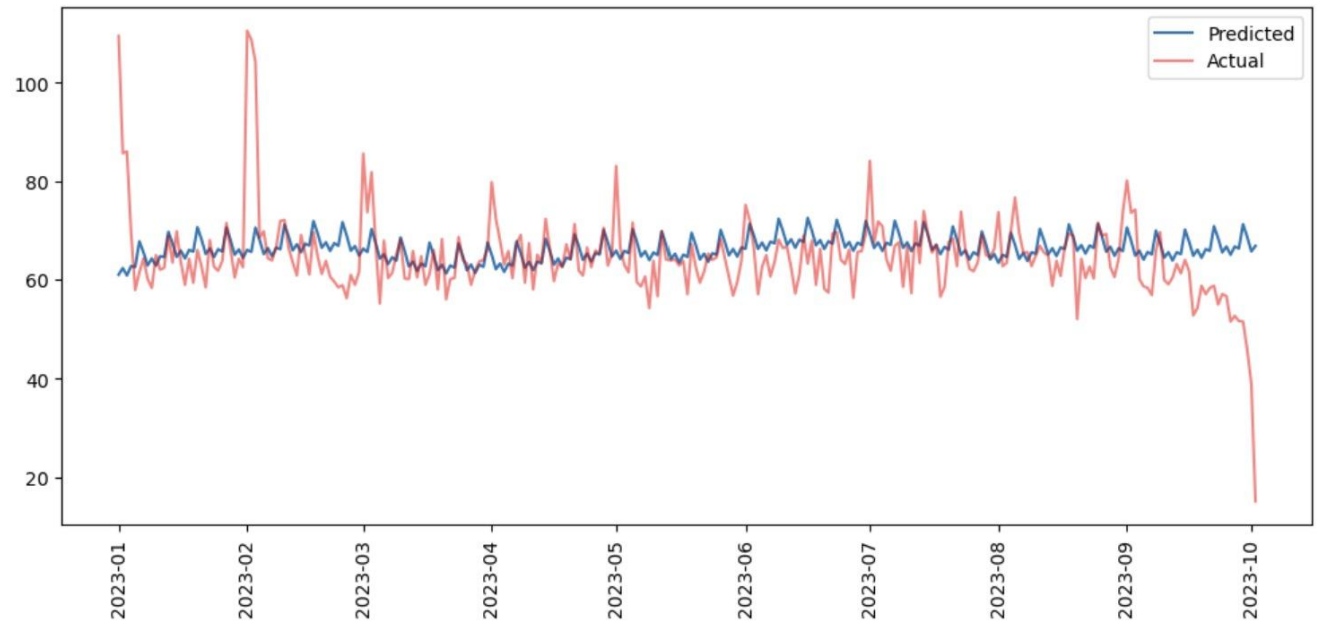


- The trends identified by the Prophet model closely align with the patterns observed in our previous weekly and yearly analyses. This consistency across different analytical approaches reinforces the reliability of our findings.

Our data was divided into two distinct sets to evaluate the prediction model's performance: a training set covering data up to January 2023 and a separate test set comprising data from January 2023 to October 2023. The model was trained on the training set, and its predictions were extended through October. This approach allowed us to assess the model's predictive accuracy by comparing its forecasted values with the actual data in our test set. The visualizations below provide a clear representation of this process and the model's performance:

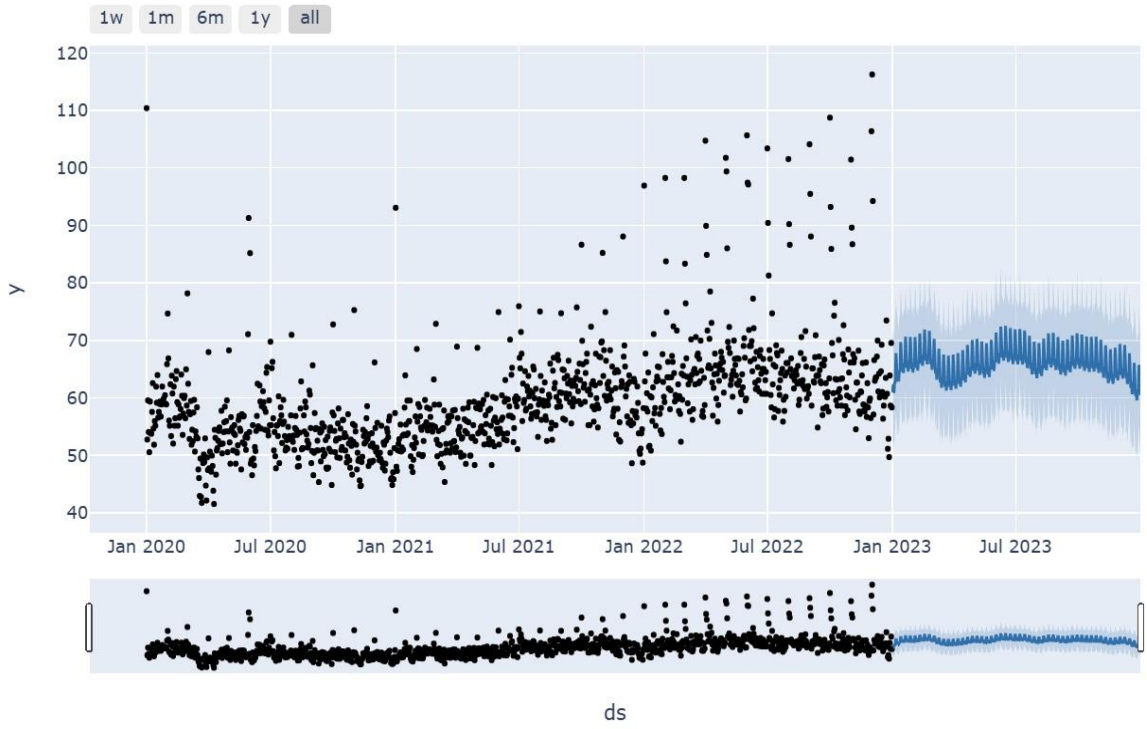
Actual values from the test set plotted as red dots in the forecast plot



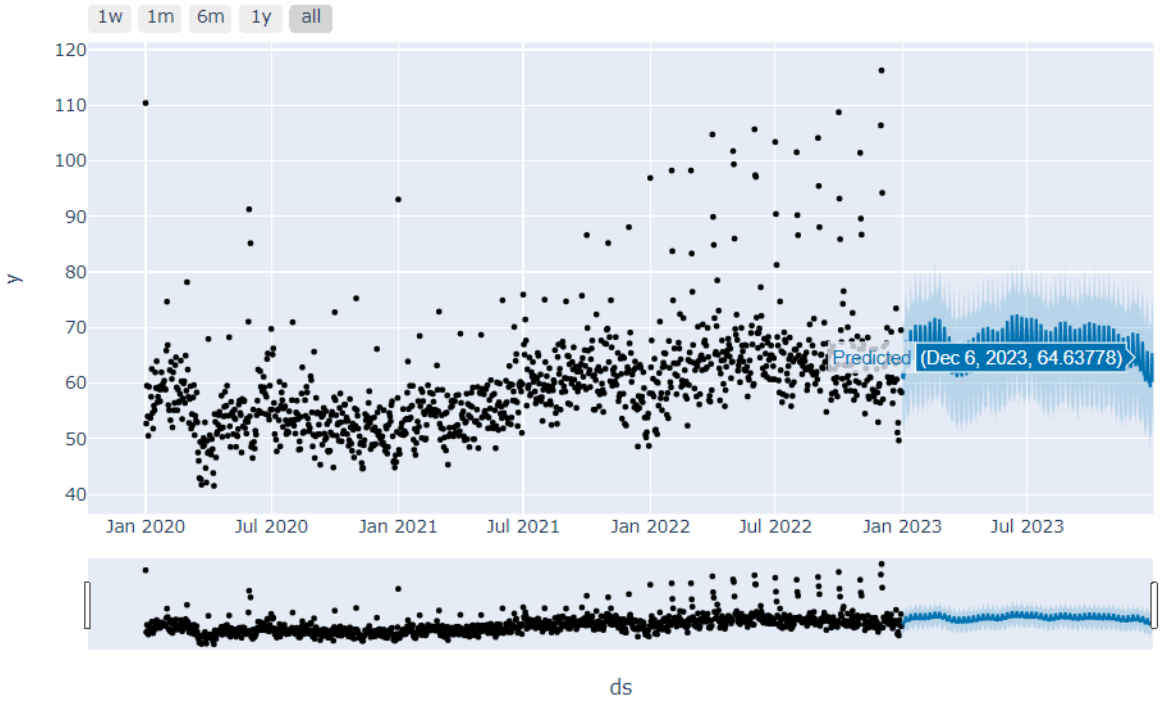


Once the model's accuracy was assessed on the test set, it was then applied to predict the entire crime rate for the year 2023. The visual representation provided below displays the predicted values as represented by the blue bars. The absence of significant spikes or irregularities in the predictions indicates a stable outlook for future crime rates. This predictive model offers valuable insights by providing estimates for future crime rates, which can be utilized for implementing strategies to reduce and manage crime effectively.

2023 Predicted Crime Rate



Hovering over the blue bars to visualize the predicted crime rate



Model Performance

MAE (Mean Absolute Error): MAE measures the average absolute difference between the observed (actual) values and the forecasted values. It provides a measure of the model's forecast accuracy. A lower MAE indicates better accuracy.

MAPE (Mean Absolute Percentage Error): MAPE calculates the average percentage difference between the observed values and the forecasted values. It's expressed as a percentage and measures the relative accuracy of the model. A lower MAPE suggests a more accurate model.

RMSE (Root Mean Squared Error): RMSE is a commonly used metric to measure the overall error of the model. It calculates the square root of the average squared differences between observed and forecasted values. RMSE penalizes larger errors more than MAE and is often used when larger errors are of greater concern.

In the context of our analysis, the values obtained for these metrics are as follows:
MAE - Prophet: 5.338357137784114
MAPE - Prophet: 0.09145407398730145
RMSE - Prophet: 76.40121553041475
These values provide insight into the performance of the time series forecasting model generated by Prophet. A lower MAE and MAPE are preferred as they indicate higher accuracy, and a lower RMSE suggests reduced error. The interpretation of these metrics should be aligned with the specific characteristics of the time series data and the nature of the forecasting problem under consideration.

Conclusion

In conclusion, this project provided valuable insights into crime data trends and patterns in the specified region. Our analysis highlighted the most common crimes, seasonal variations, and correlations with economic factors. Additionally, we investigated the impact of major events on crime rates and explored predictive modeling to forecast future trends.

The results of this analysis can be used for informed decision-making, resource allocation, and policy development. Further studies and analyses can build upon these findings to address specific questions related to crime and safety.

This report summarizes our data cleaning and analysis process, highlighting key insights and any interesting patterns or trends discovered. The accompanying Jupyter Notebook and presentation provide a more in-depth look into the project's details and findings. We believe this analysis provides valuable information for decision-makers and stakeholders interested in understanding and addressing crime-related issues in the specified region.