IE 6400: Foundations of Data Analytics

Project 1

Cleaning and Analyzing Crime Data

Group 18

Member Details:

Manogna Devalla : devalla.m@northeastern.edu
Amogha Gadde : gadde.am@northeastern.edu
Divya Babulal Shah : shah.divyab@northeastern.edu
Kumari Simran : simran.k@northeastern.edu
Abhishek Kumar Sinha : sinha.ab@northeastern.edu

Contributions:

Percentage of Effort Contributed by Manogna: 20%
Percentage of Effort Contributed by Amogha: 20%
Percentage of Effort Contributed by Divya: 20%
Percentage of Effort Contributed by K.Simran: 20%
Percentage of Effort Contributed by Abhishek: 20%

Submission Date: Nov-03-2023

Acknowledgements

Foremost, we would like to express our sincere gratitude to our Professor Sivarit Sultornsanee for giving us the opportunity to make this project. We would also like to thank our TAs for the continuous support of our study and research while making this project, for their patience, motivation and enthusiasm and immense knowledge.

We are thankful that we got an amazing team and for all the stimulating discussions and for all the experience and knowledge that we have gained by working together. We would also like to thank our parents and all our friends that have motivated and for never doubting our ability to complete this project.

Contents

Introduction

This project delves into the realm of real-world data, offering a fascinating journey into the intricate world of criminal activities and the underlying factors that influence them. In this project, we focus on a dataset encompassing crime data spanning from the year 2020 to the present day. Our primary objective is to uncover the rich insights hidden within this data by employing data cleaning, exploratory data analysis, and advanced analytical techniques.

Crime data provides a unique window into the socio-economic and demographic landscape of an area. By unraveling the patterns and trends in this data, we aim to gain a comprehensive understanding of the dynamics that govern criminal activities. This knowledge is invaluable for law enforcement, policymakers, and community stakeholders in their efforts to create safer and more secure environments.

Our mission begins with the crucial task of data cleaning and preparation. We shall navigate through the dataset with a meticulous eye, addressing issues such as missing data, duplicates, and data type inconsistencies. Ensuring the data's integrity is paramount as it forms the foundation for our subsequent analysis.

The heart of this project is the Exploratory Data Analysis (EDA). Here, we delve into the depths of the dataset, employing various visualizations, statistical analyses, and hypothesis testing. The aim is to paint a vivid picture of crime trends over time, seasonal patterns, the prevalence of specific crime types, regional variations, and much more. With these insights, we can better understand the ever-evolving world of crime and its intersection with societal and economic factors.

We also venture into advanced analyses, including predictive modeling techniques such as time series forecasting. These models provide us with the power to foresee future crime trends, which can be instrumental in planning and policymaking. Moreover, we explore correlations between demographic factors, economic conditions, and crime rates to discover any intriguing relationships.

We aim to provide a holistic view of the complex and dynamic world of crime data analysis. This report will shed light on the patterns, anomalies, and trends that shape our understanding of criminal activities that was discovered after performing exploratory data analysis on the crime data provided.

Objective

In this project, we have worked with a real-world dataset containing crime data from 2020 to the present. 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. The dataset contained the crime data of LA from Jan 2020 – Oct 2023.

Data Inspection

- The data set contains 18333614 rows and 26 columns.
- Out of the 26 columns, 15 columns are of numeric type, 2 are date type and 9 are string type.
- The ‘DATE OCC’ and ‘Date Rptd’ are columns with date values in the range 01/01/2020 – 10/02/2023.
- The ‘Vict Sex’ column contains categorical data with values ['F' : Female, 'M' : Male, 'X' : NonBinary, 'H' : Transgender, '-' : Unknown][1]
- The ‘Vict Descent’ column contains categorical data with values ['B' - Black or African American, 'H' - Hispanic or Latino, 'X' – Other, 'W' – White, 'A' – Asian, 'O' - Middle Eastern or Arab, 'C' - American Indian or Alaska Native, 'F' – Filipino, 'K' – Korean, 'I' – Indian, 'V' – Vietnamese, 'Z' – Chinese, 'J' – Japanese, 'P' - Pacific Islander, 'Unknown' - Data Missing or Not Specified, 'G' - Guamanian or Chamorro, 'U' – Hawaiian, 'D' – Samoan, 'S' – Samoan, 'L' – Laotian, '-' - Unknown or Not Specified][2]

Data Cleaning

- The date columns were present as string. We converted them to date type.
- There were Missing Values in the data. They were handled in the following ways:
  - The ‘Mocodes’ and ‘Cross Street’ columns were dropped.
  - The ‘Vict Sex’ and ‘Premis Cd’ columns had all their null values dropped.
  - The ‘Vict Descent’ and ‘Premis Desc’ had all their null values imputed with the values ‘Unknown’ and ‘NA’ respectively.
- Duplicate values in the data were dropped.
- Columns with categorical data such as ‘Vict Sex’ and ‘Vict Descent’ were encoded.
- We have also ensure that there is only 1-1 mapping between columns like ‘Crm Cd’ and ‘Crm Cd Desc’, and ‘Area’ and ‘Area Name’.

Exploratory Data Analysis

Overall Crime Trends:

Calculate and plot the total number of crimes per year to visualize the trends.

- The total number of crimes was calculated from 2020 to 2023.
- It was observed that the year 2022 had the highest number of crimes. The minimum number of crimes in 2022 was in February with 15352 crimes and maximum in June with 17766 counts.
- The second highest number of crimes took place in 2021 where the minimum count was in February with 13203 crimes and maximum in October with 16520 crimes.
- The crime rate was minimum in February in 2021 and 2022.
- In 2020, the maximum number of crimes took place in January and minimum in April.
- The total crimes decreased rapidly from 2022 to 2023 from 204150 to 146586.

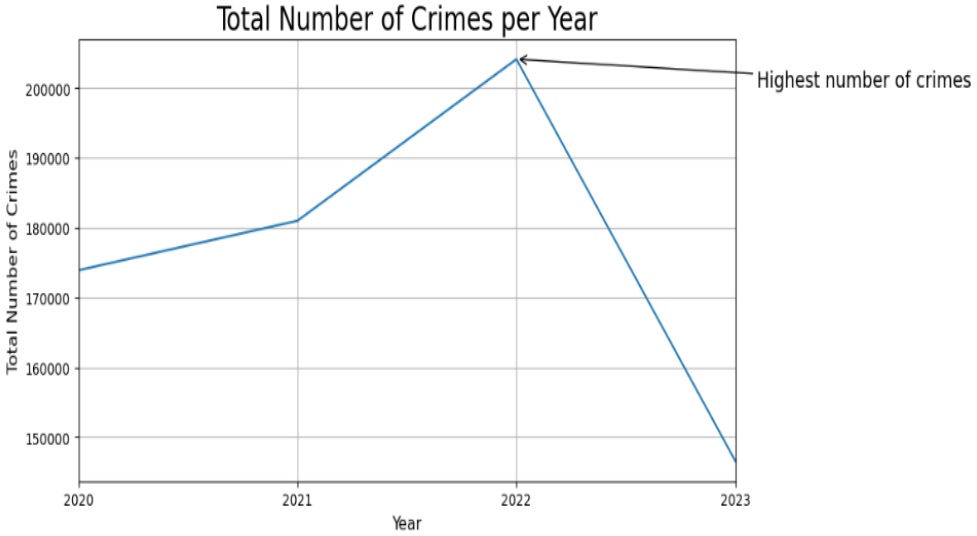

Fig 1: Line Chart for Total Number of Crimes per Year

Seasonal Patterns:

Group the data by month and analyze the average number of crimes per month over the years.

- 2020: The highest number of crimes was in January. The average number of crimes decreased from January to April then increased to a range of 15000 – 14000 from May to August. It reaches to lowest in November.
- 2021: The number of crimes was lowest in February. It gradually increased throughout the year. It reaches its highest count in October.
- 2022: The number of crimes have the highest rate compared to the other years. The lowest rate was in February within the range of 15000-16000. The highest crime count was in May and June.

- 2023: The crime rate decreased from January to June and then slowly increased till August. The count rapidly declined from August to September.

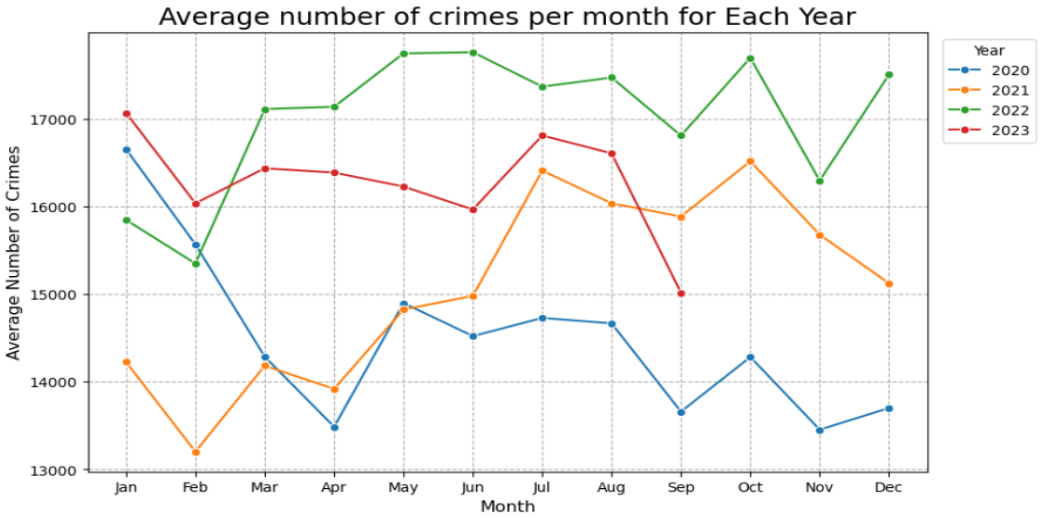

Fig 2: Line Chart for Average Number of Crimes per Month for each Year

Most Common Crime Type:

Count the occurrences of each crime type and identify the one with the highest frequency.

- The most prevalent crime type in the dataset is "BATTERY - SIMPLE ASSAULT," with a frequency of 64,559 reported cases. This indicates that simple assault-related incidents are the most reported crimes in the area, highlighting the significance of addressing and preventing such incidents to enhance public safety and security.

Analysis of the top 10 Crime Types:

- The graph below displays the number of occurrences for the top 10 types of crimes.
- Theft of Identity is the second most common crime type, with 51,506 occurrences. Identity theft is a significant issue and has a high frequency in reported incidents.
- Burglary from Vehicle and Vandalism - Felony ($400 & Over, All Church Vandalisms) have similar frequencies, with approximately 49,783 and 49,483 occurrences, respectively. This suggests a notable number of vehicle burglaries and felony vandalism cases.
- Burglary is another common crime type, with 49,351 occurrences. Burglaries, in general, are reported frequently.

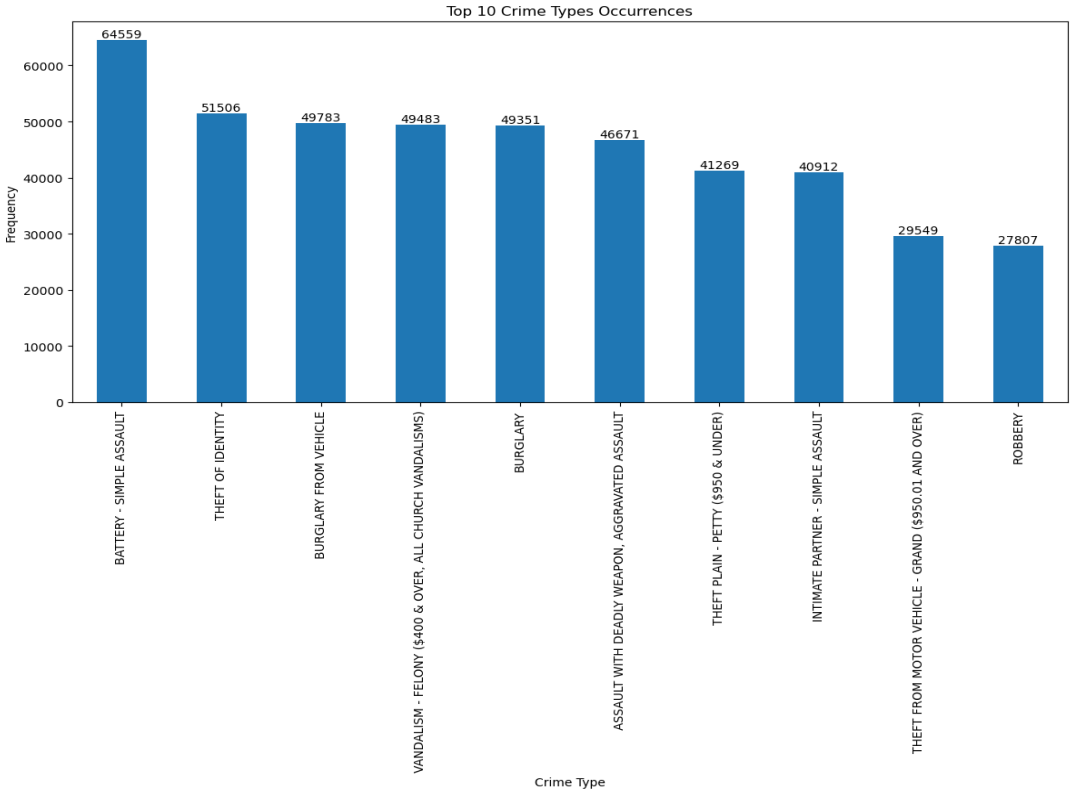

Fig 3: Bar Chart for Top 10 Crime Types Occurrences

Regional Differences:

Group the data by region or city and compare crime rates between them using descriptive statistics or visualizations.

Analysis of Total Number of Crimes by Region:

- The area, Foothill, has the lowest crime count of 22889.
- The area, Central, has the highest crime count of 50470.
- Central, 77th Street, and Pacific are the areas with the highest reported crime counts, with 50,470, 43,788, and 40,758 cases, respectively. This suggests that these areas may have higher crime rates compared to others.
- Foothill has the lowest reported crime count among all the listed areas, with only 22,889 cases. This area appears to have a relatively lower crime rate compared to the others.
- Hollywood, Southwest, and Olympic have similar crime counts, ranging from 39,216 to 39,264 cases. These areas have moderately high reported crime rates.
- Harbor, Mission, and Hollenbeck have relatively lower reported crime counts, ranging from 24,894 to 28,215 cases, indicating that these areas may have lower crime rates compared to the top-ranking areas.

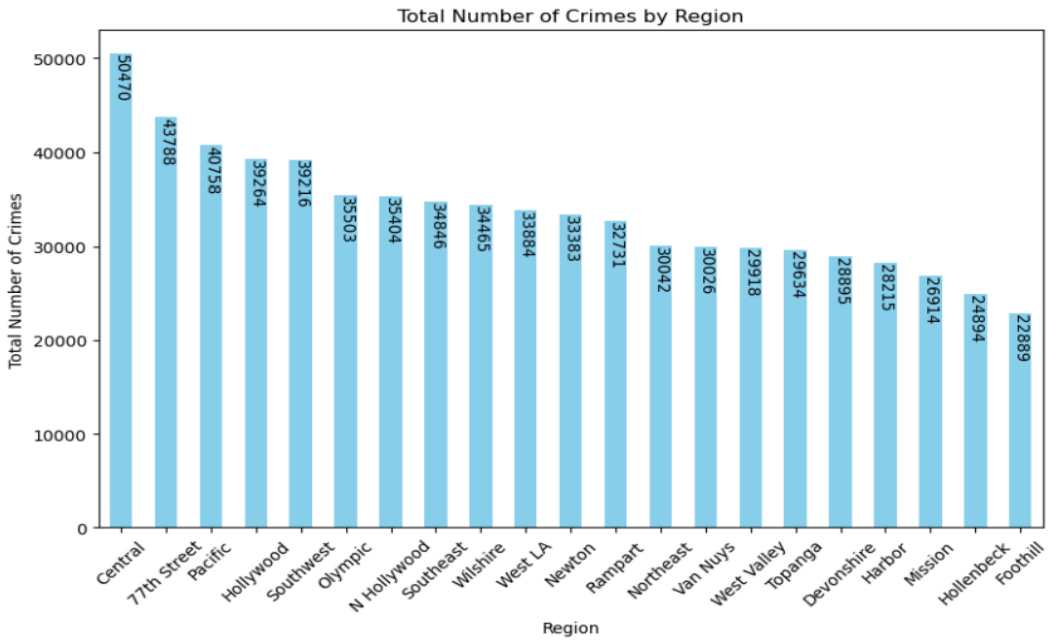

Fig 4: Bar Chart for Total Number of Crimes by Region

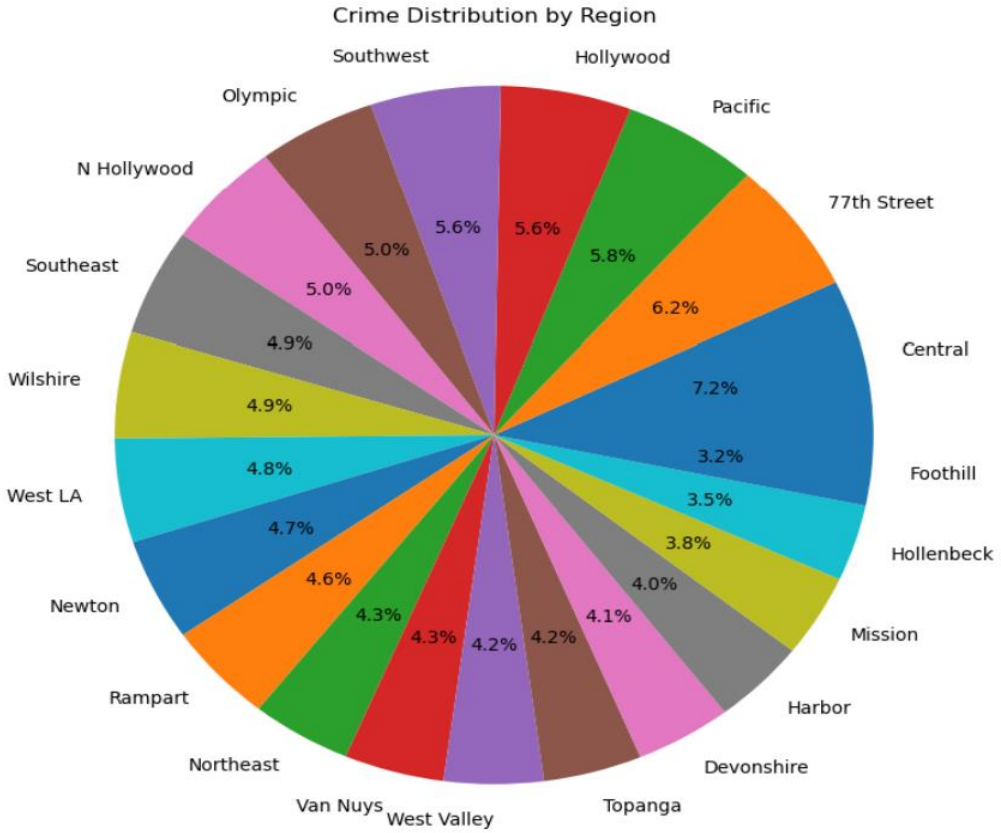

Fig 5: Pie Chart for Crime Distribution by Region

- The area, Foothill, has the lowest percentage of crime count of 3.2%.
- The area, Central, has the highest percentage of crime count of 7.2%.

Analysis of Top 5 Crime Type Distribution by Region:

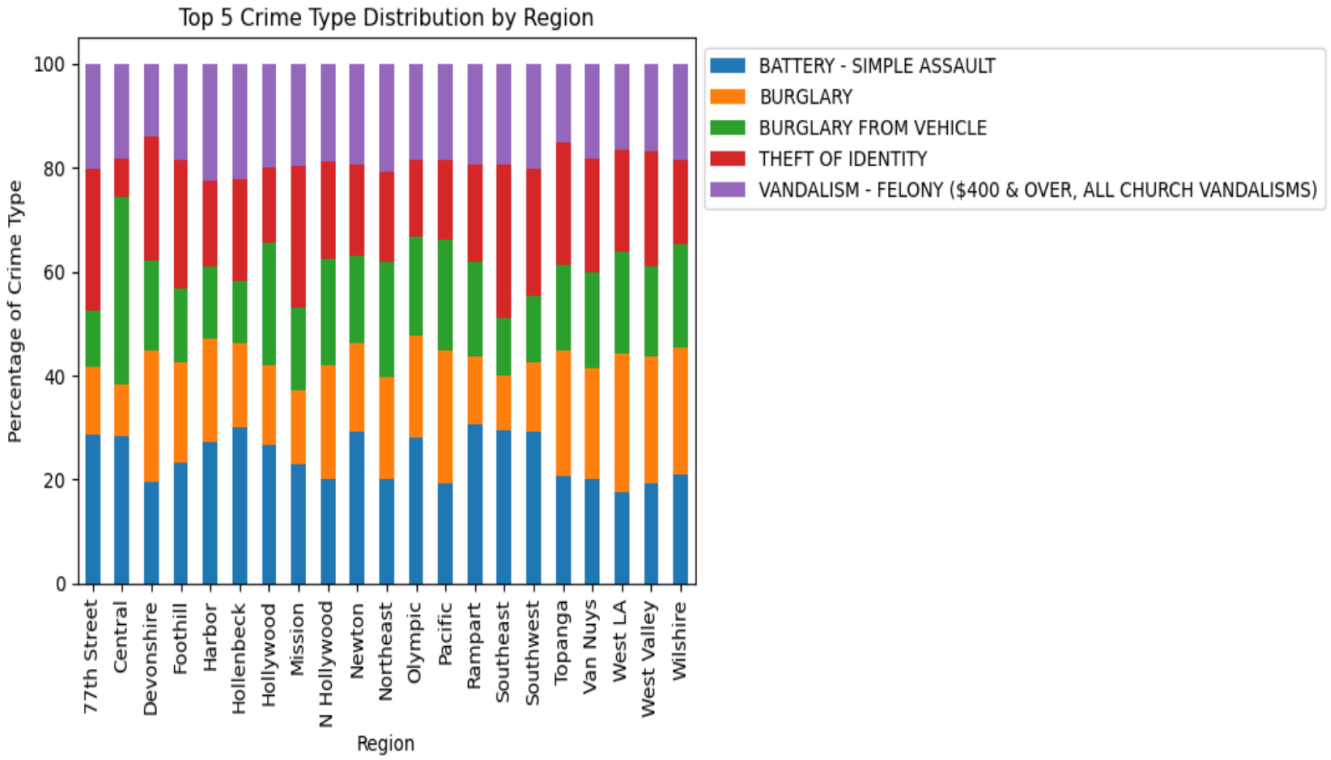

Fig 6: Stacked Bar Chart for Top 5 Crime Type Distribution by Region

- Variation in Crime Rates: There is a significant variation in the crime rates for different types of crimes across different areas. For example, "BATTERY - SIMPLE ASSAULT" is highest in the 77th Street area, while "BURGLARY FROM VEHICLE" is highest in Central area.
- Regional Differences: Different areas have different crime profiles. For instance, "VANDALISM - FELONY" is relatively high in Harbor and Hollenbeck areas, while "THEFT OF IDENTITY" is higher in areas like Southeast.
- Identity Theft: "THEFT OF IDENTITY" seems to be relatively high in Mission and 77th Street areas, indicating potential issues related to identity theft in those regions.
- Vandalism Rates: "VANDALISM - FELONY" rates vary, with Harbor and Hollenbeck having higher rates, while Topanga has the lowest rate.

Analysis of Top 5 Crime Type Distribution by Region and Gender:

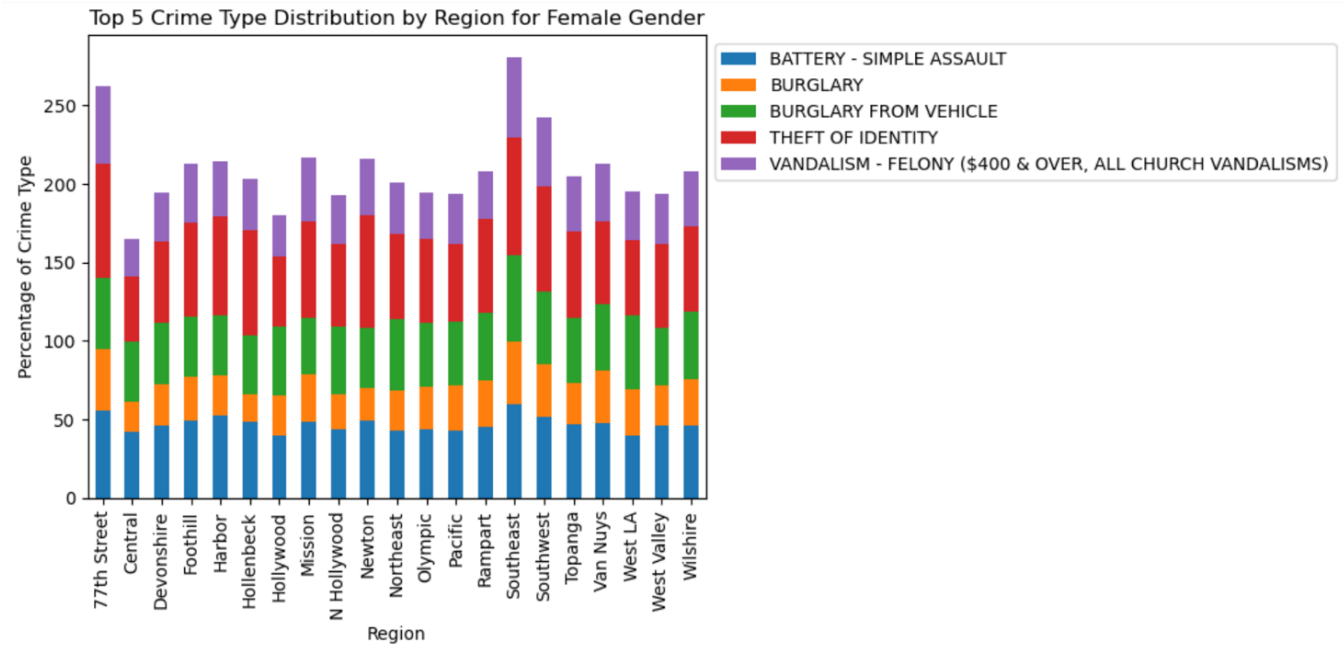

Fig 7: Stacked Bar Chart for Top 5 Crime Type Distribution by Region for Female Gender

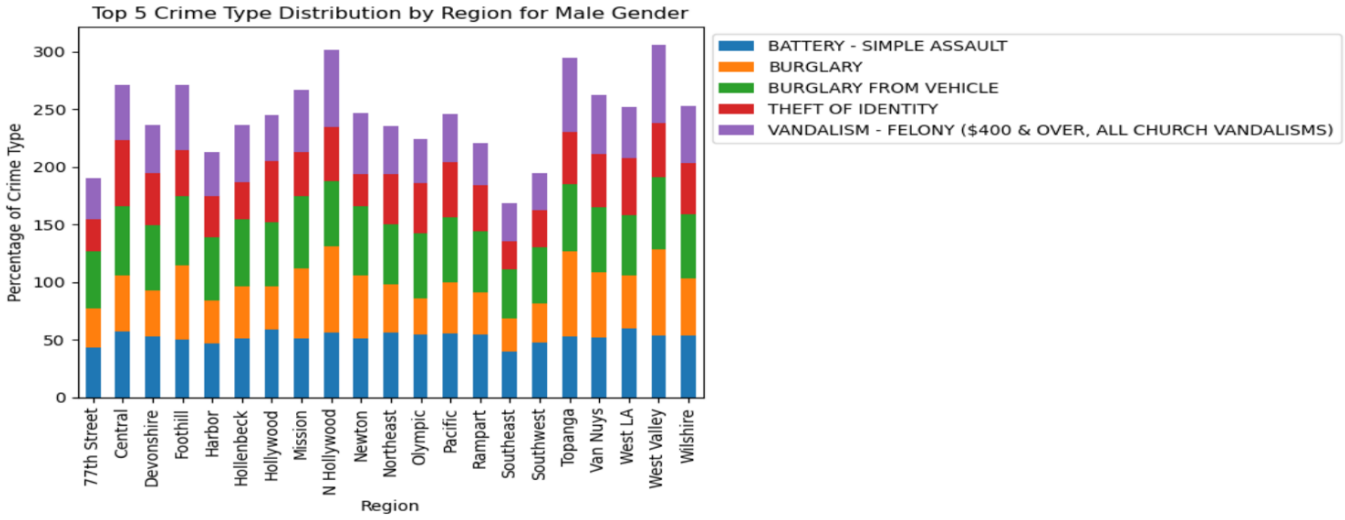

Fig 8: Stacked Bar Chart for Top 5 Crime Type Distribution by Region for Male Gender

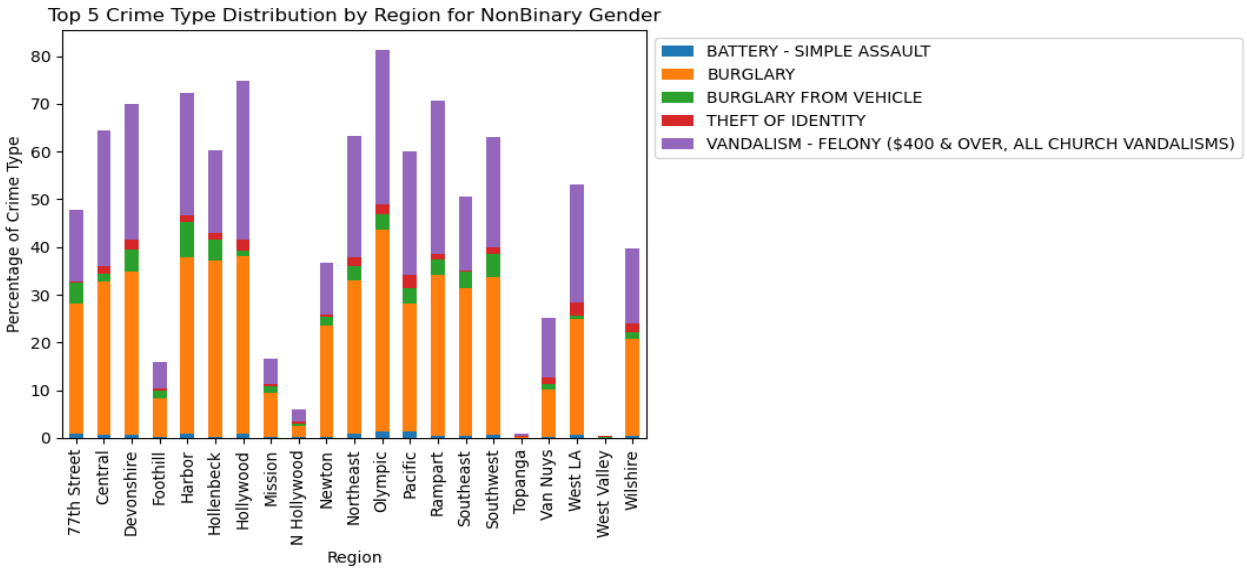

Fig 9: Stacked Bar Chart for Top 5 Crime Type Distribution by Region for NonBinary Gender

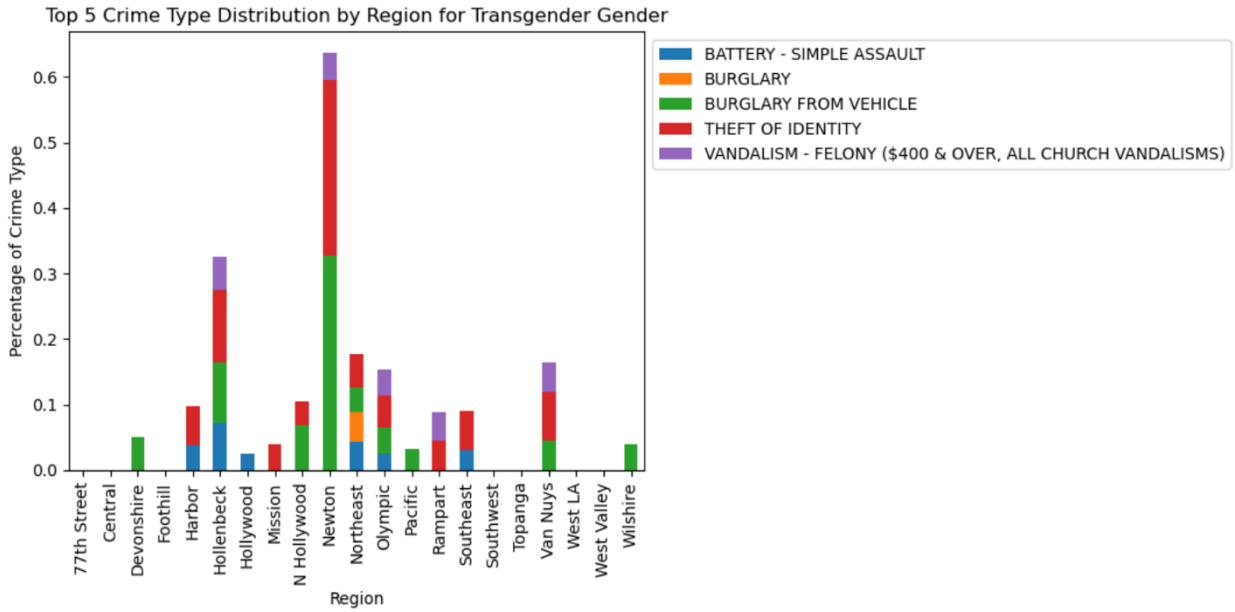

Fig 10: Stacked Bar Chart for Top 5 Crime Type Distribution by Region for Transgender Gender

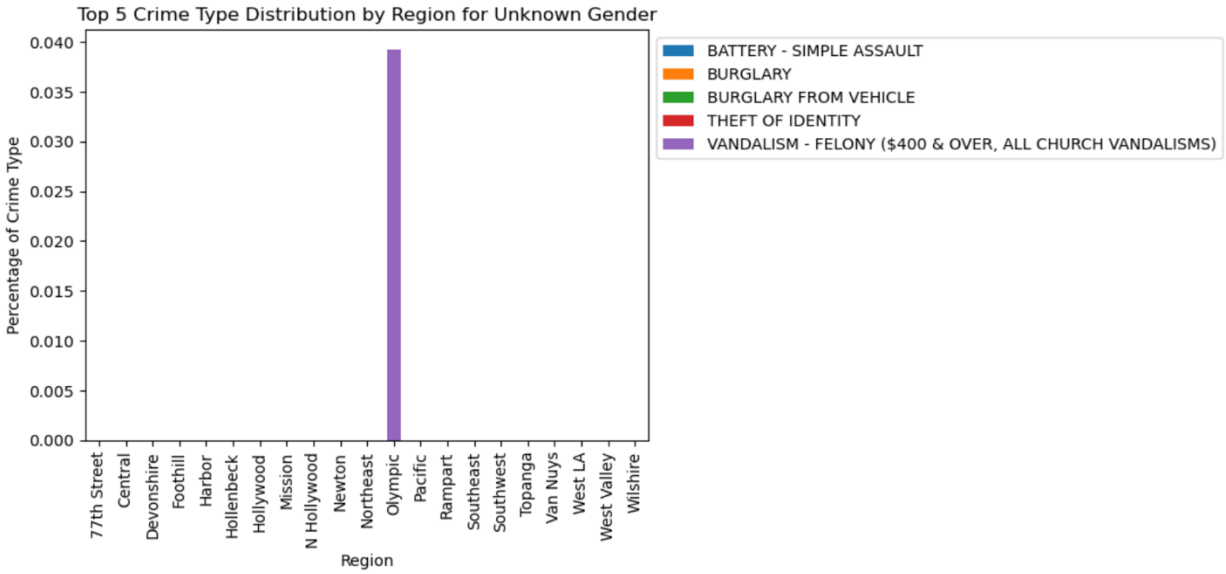

Fig 11: Stacked Bar Chart for Top 5 Crime Type Distribution by Region for Unknown Gender

- Gender Disparities in Crime: Across different areas, there are variations in the gender distribution of individuals involved in different types of crimes. For example, in the "77th Street" area, there is a higher percentage of "Female" involvement in "BATTERY - SIMPLE ASSAULT" compared to "Male" involvement, while the opposite is observed for "BURGLARY FROM VEHICLE."
- Crime Type Variation: The data shows that different types of crimes have distinct gender distributions. For example, "BURGLARY" typically has a higher percentage of "Male" involvement, while "THEFT OF IDENTITY" tends to have a higher percentage of "Female" involvement.
- Nonbinary Category: The "NonBinary" category, which represents individuals who do not identify strictly as "Male" or "Female," is present in some crime categories but is significantly less represented in most cases.
- Gender and Vandalism: "VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)" generally has a more balanced distribution between "Male" and "Female" involvement, with smaller percentages for "Nonbinary" and "Transgender".
- Significant Unknown Data: The "Unknown" category is prevalent in most areas and crime categories. This indicates that for a large portion of reported crimes, the gender identity of the individuals involved is not specified or known.

Analysis of Distribution of Victim Ages in Different Areas/ Regions:

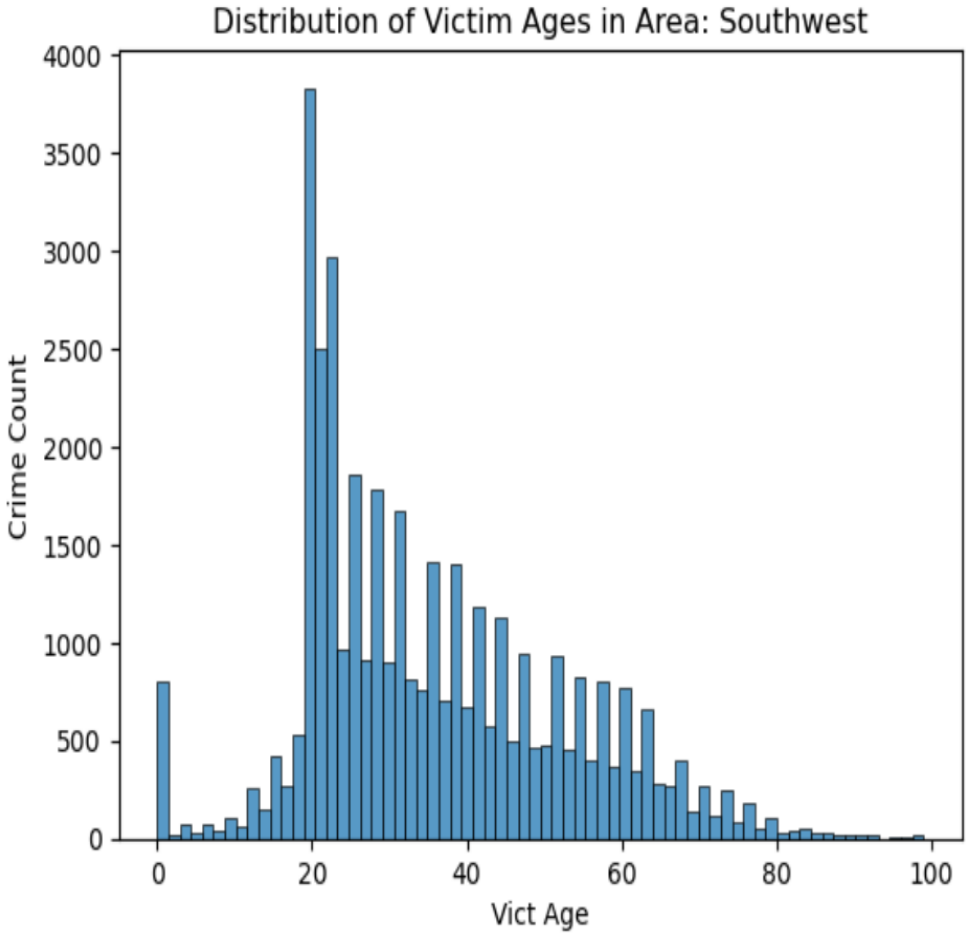

Fig 12: Histogram for Distribution of Victim Ages in Area : Southwest

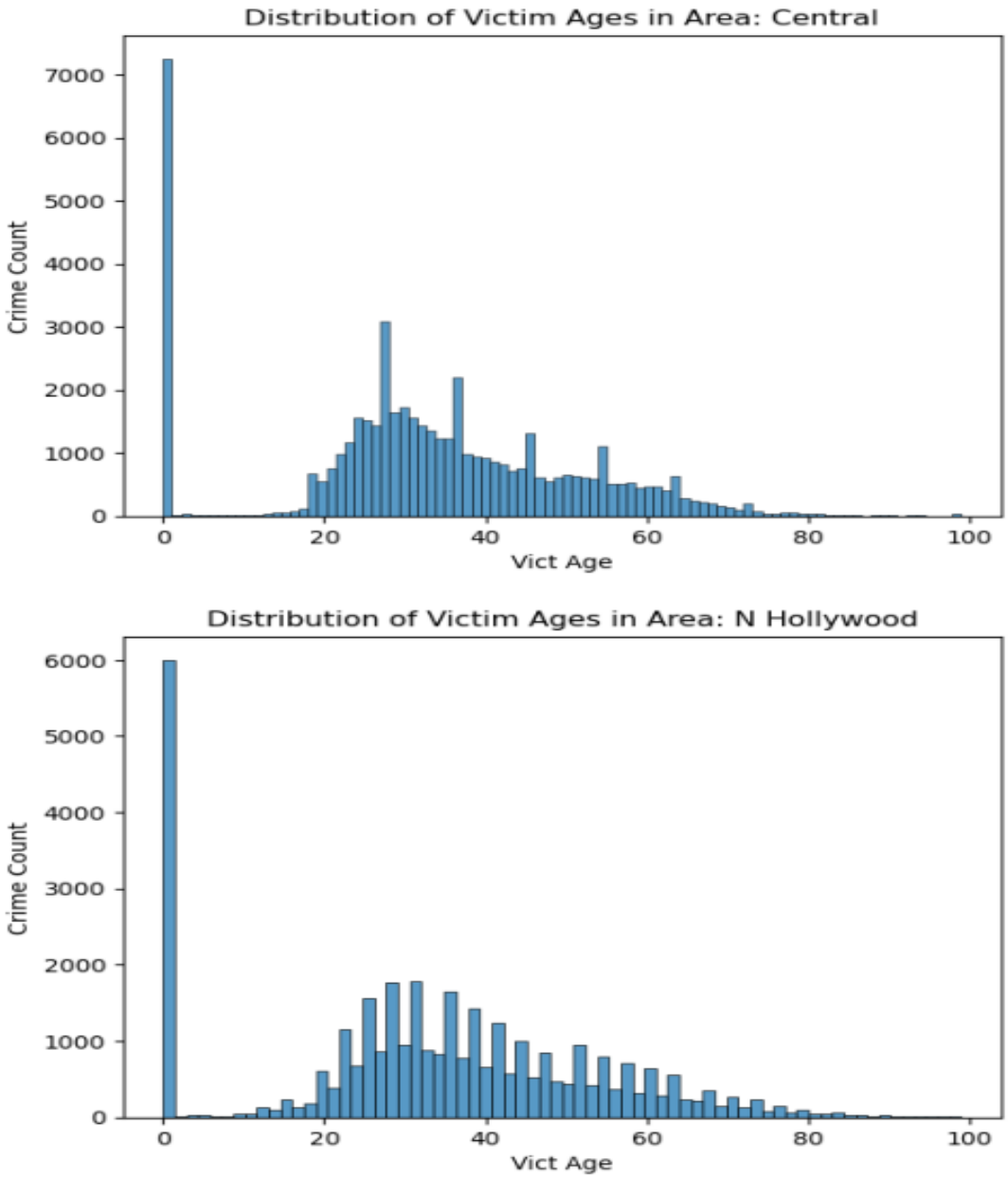

Fig 13: Histogram for Distribution of Victim Ages in Areas : Central , N Hollywood

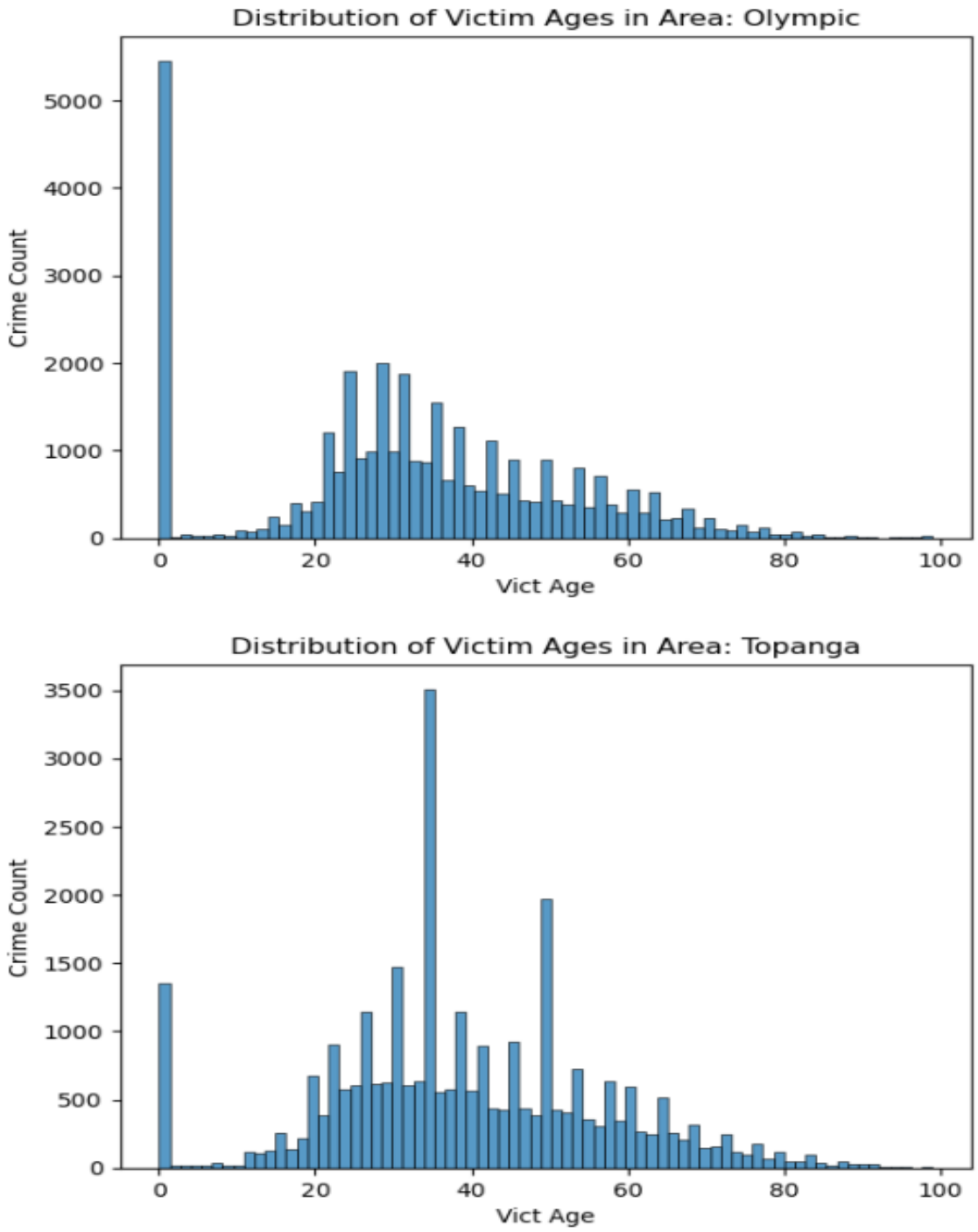

Fig 14: Histogram for Distribution of Victim Ages in Areas : Olympic, Topanga

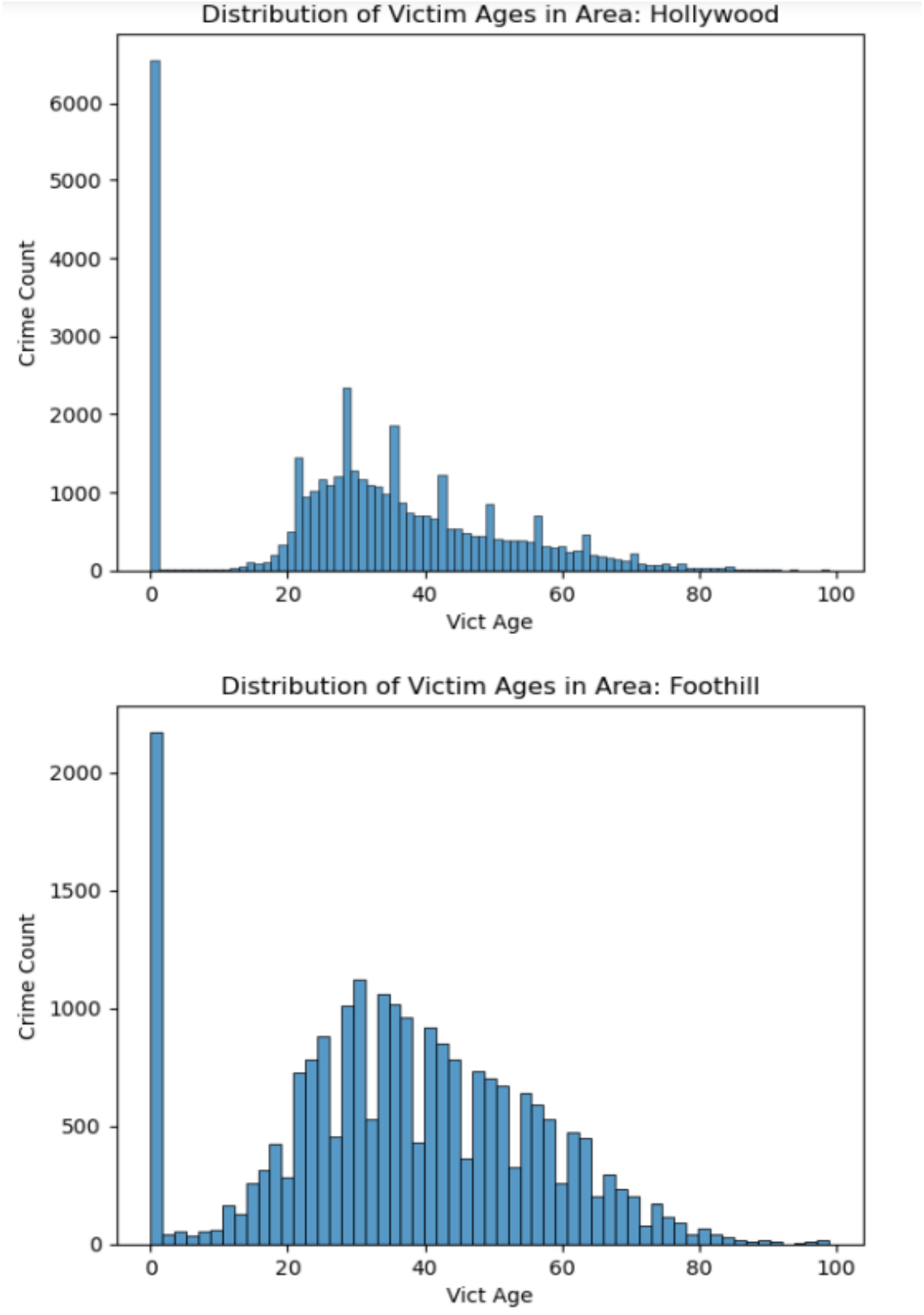

Fig 15: Histogram for Distribution of Victim Ages in Areas : Hollywood, Foothill

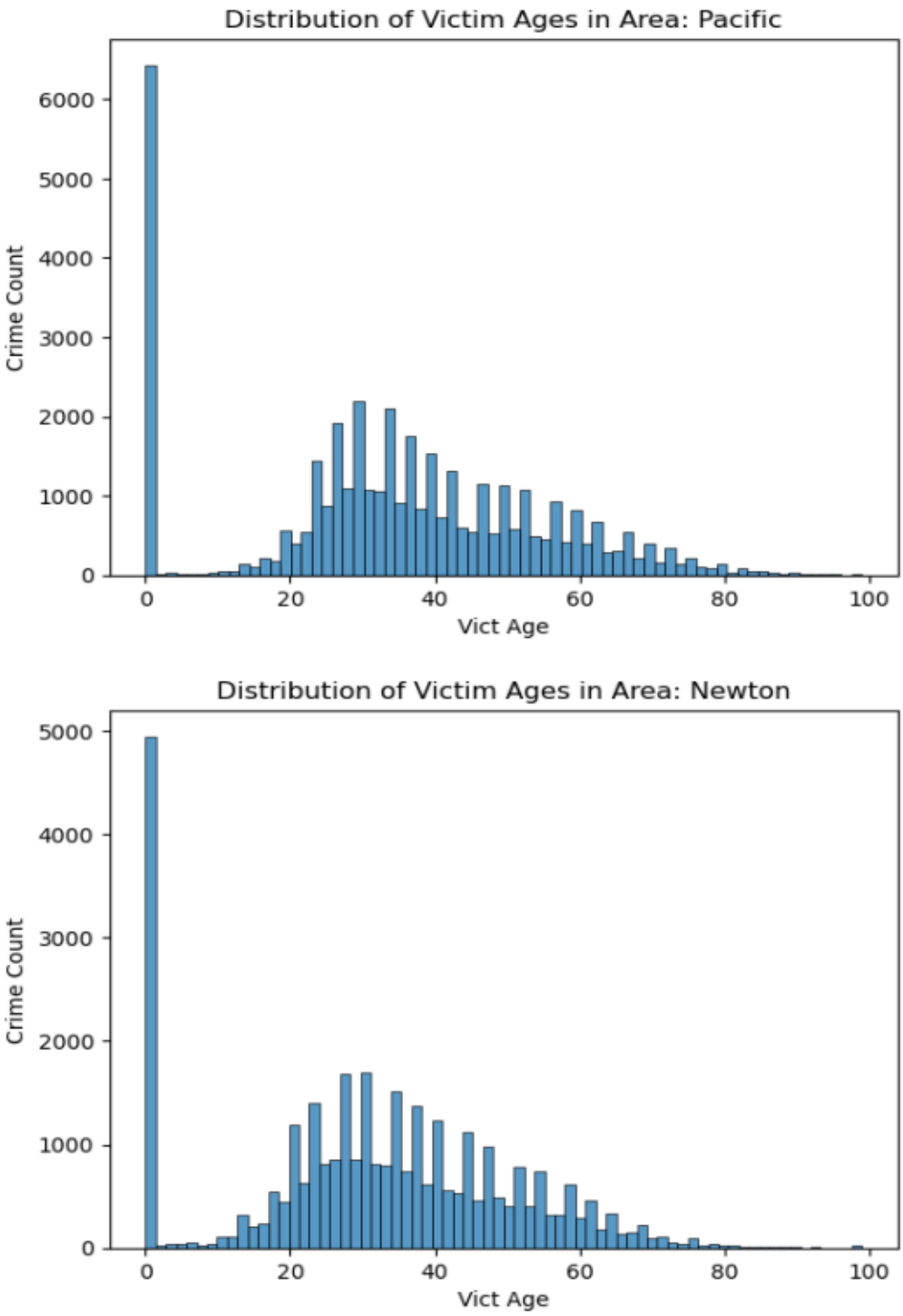

Fig 16: Histogram for Distribution of Victim Ages in Areas : Pacific, Newton

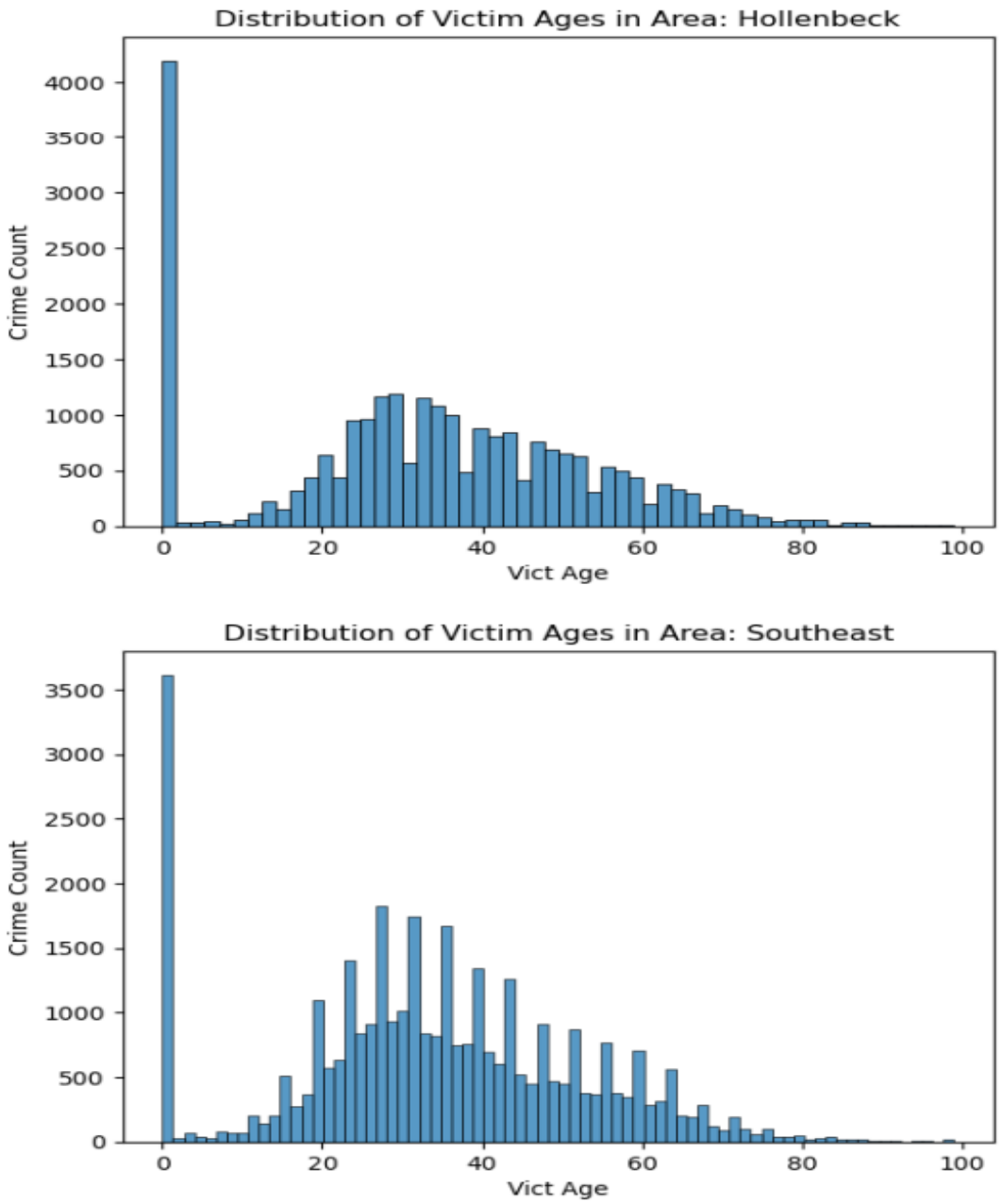

Fig 17: Histogram for Distribution of Victim Ages in Areas : Hollenbeck, Southeast

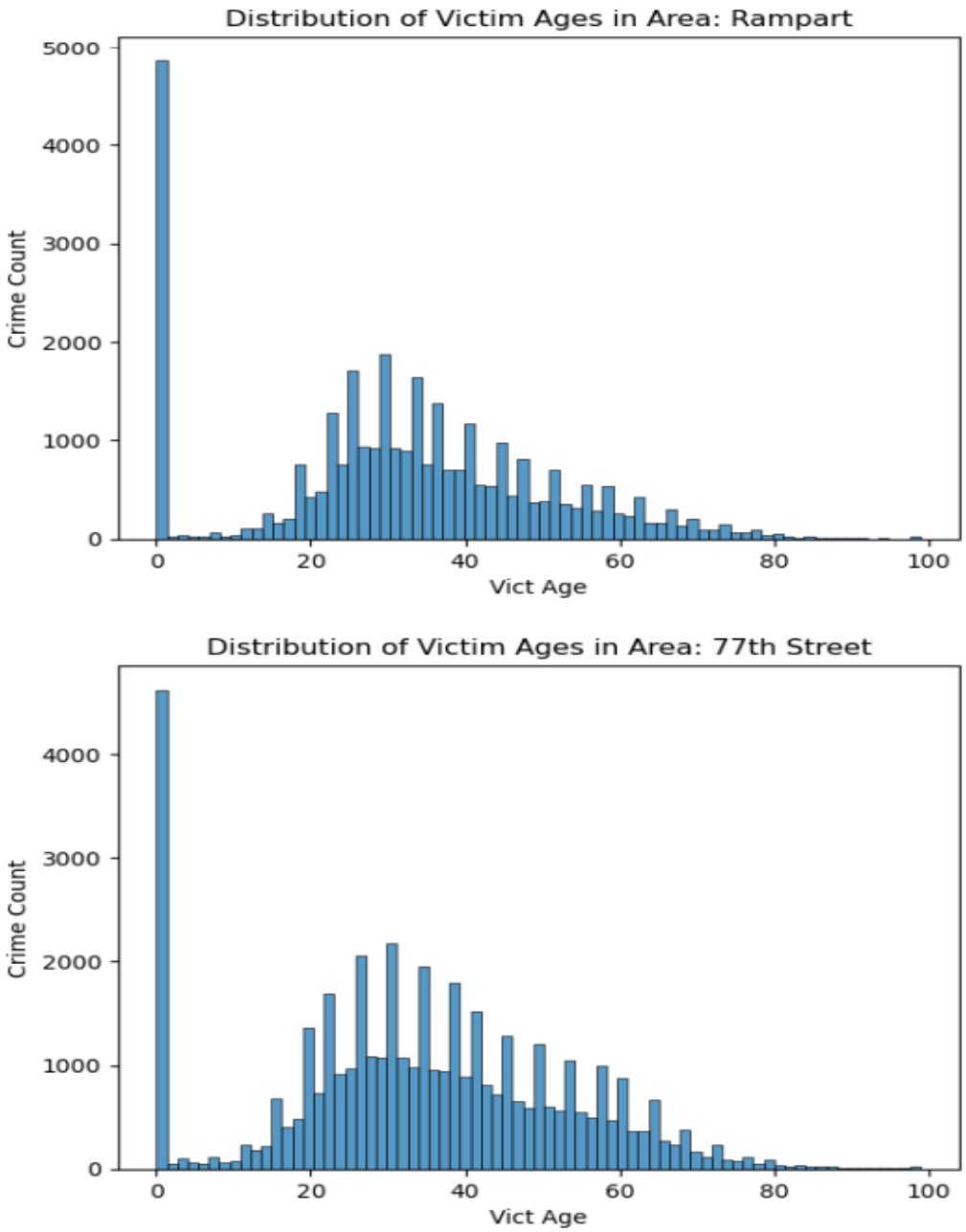
Distribution of Victim Ages in Area: Rampart
Crime Count
Vict Age
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
Distribution of Victim Ages in Area: 77th Street
Crime Count
Vict Age
0
1000
2000
3000
4000
0
20
40
60
80
100

Fig 18:

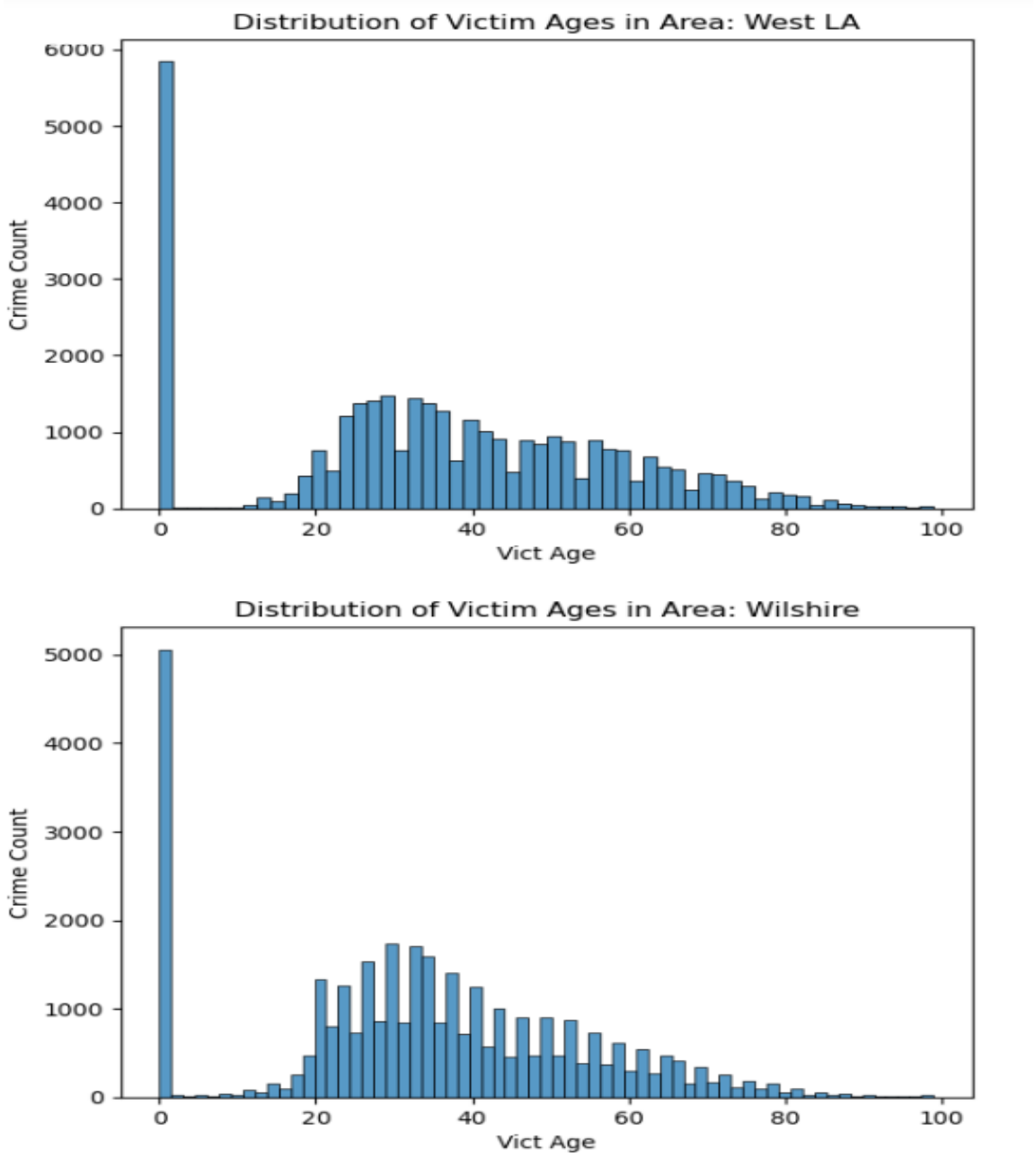

Fig 19: Histogram for Distribution of Victim Ages in Areas : West LA, Wilshire

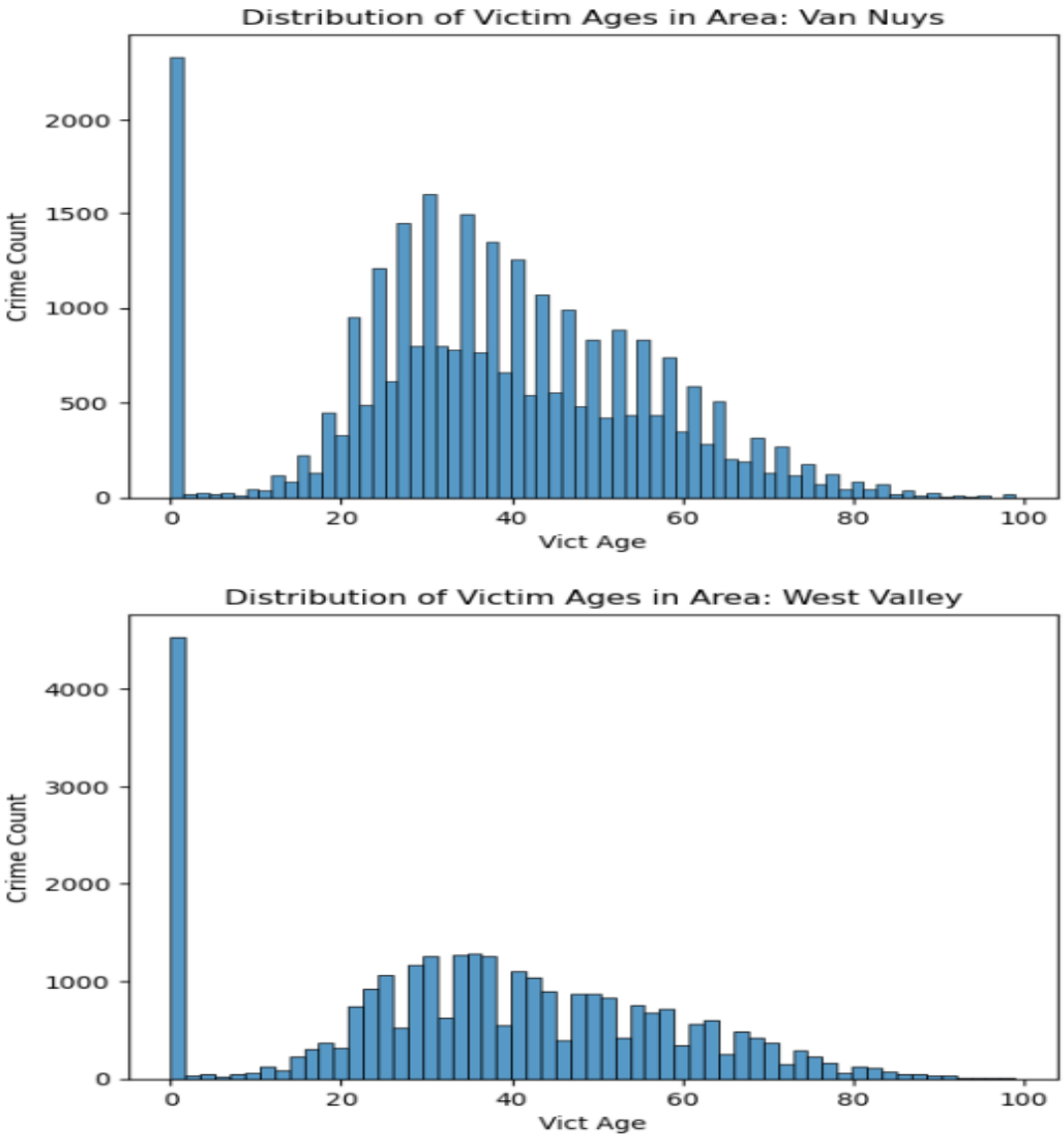

Fig 20: Histogram for Distribution of Victim Ages in Areas : Van Nuys, West Valley

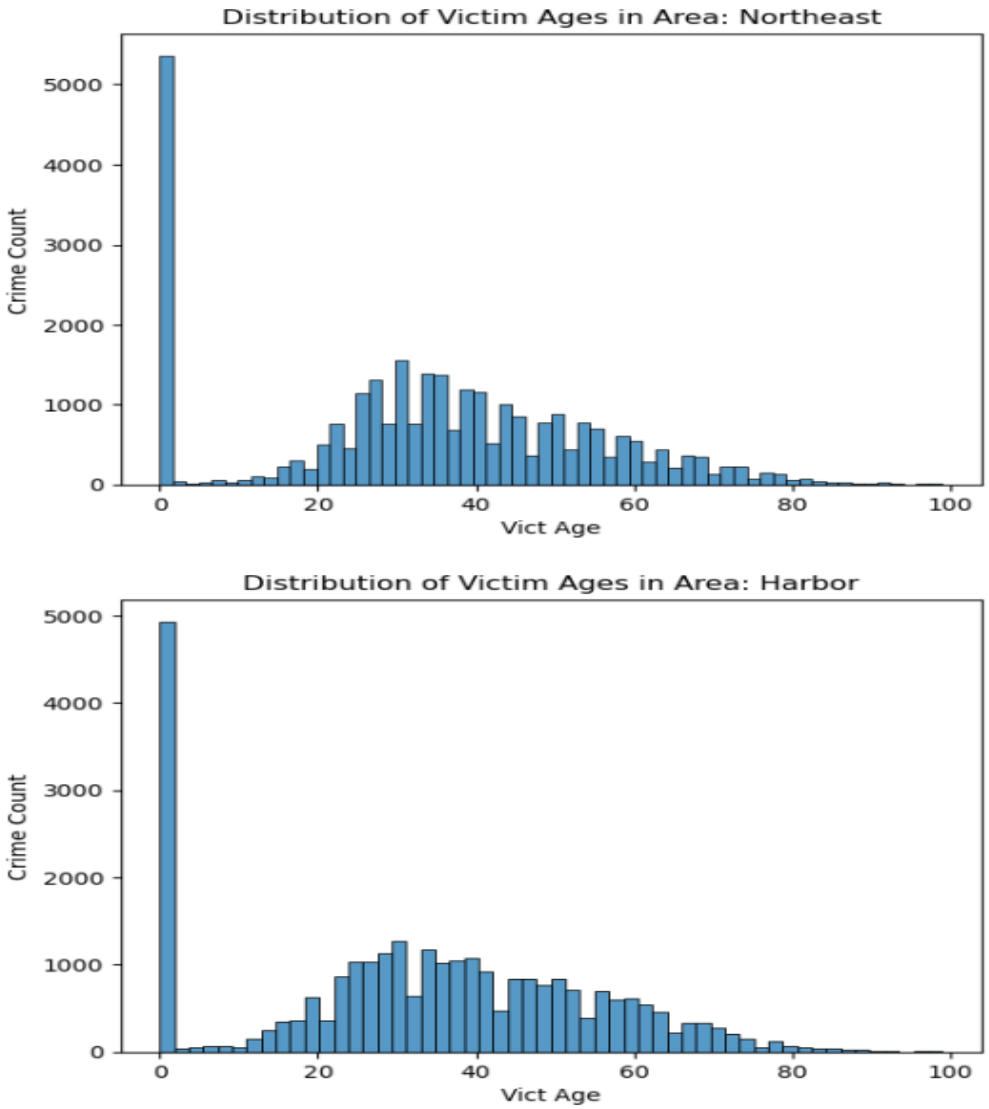

Fig 21: Histogram for Distribution of Victim Ages in Areas : Northeast, Harbor

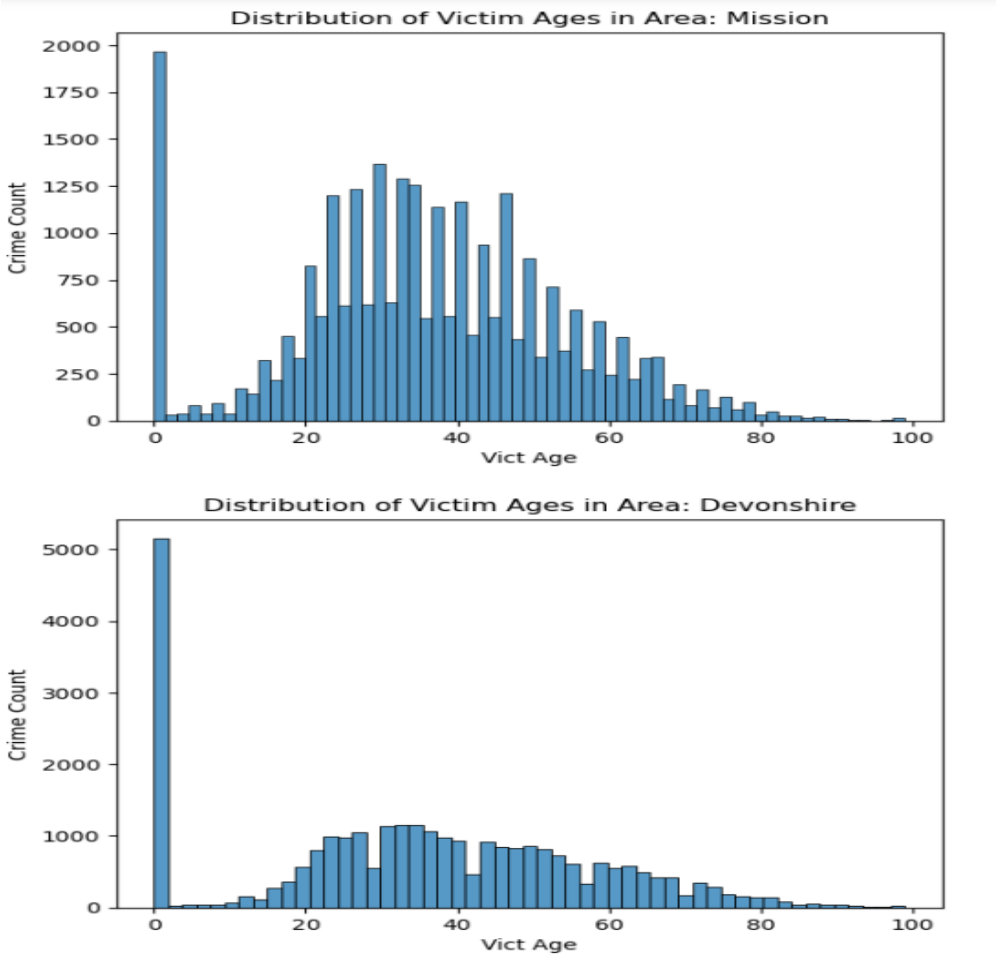

Fig 22: Histogram for Distribution of Victim Ages in Areas : Mission, Devonshire

- In most of the areas, the age group with the highest count is "0," which may represent an unspecified or unknown age group. This suggests that for many incidents, the age of the victim is not recorded or available.
- However, in "Southwest" and "Topanga," the age groups with the highest counts are "21" and "35," respectively. This indicates that in these areas, there are significant numbers of incidents or crimes involving individuals in these specific age groups.
- Some areas have particularly high counts of incidents, such as "Central," "Hollywood," "Pacific," and "West LA." These areas have counts in the range of 5,000 to 7,000 incidents.
- "Mission" has the lowest count among the listed areas, with only 1,967 incidents recorded for the unspecified age group.

Correlation with Economic Factors:

Collect economic data for the same time frame and use statistical methods like correlation analysis to assess the relationship between economic factors and crime rates.

We are going to test the following economic factors with crime rates in the city:

- Unemployment Rate: (monthly data between 2020-2023)
- Poverty Rate: (yearly data between 2020-2023)
- Gross Domestic Product(GDP): (yearly data between 2020-2023)
- Per Capita Personal Income: (yearly data between 2020-2023)
- Median Household Income: (yearly data between 2020-2023)
- Income Inequality: (yearly data between 2020-2023)
- Poverty Gap: (yearly data between 2020-2023)
- Consumer Price Index: (monthly data between 2020-2023)
- Inflation Rate: (monthly data between 2020-2023)
- Housing Price Index: (yearly data between 2020-2023)
- Housing Affordability Index: (yearly data between 2020-2023)
- Housing Inventory: (yearly data between 2020-2023)
- Federal Funds Rates: (monthly data between 2020-2023)

Correlation of economic factors with crime rates in the city after statistical analysis:

- **Unemployment Rate:** The Correlation coefficient was -0.63 and the p value was 0.00000001. Since the p value was less than 0.05, we rejected the null hypothesis. Hence, we can say that there is a statistically significant correlation of the Unemployment Rate and Crime Count in the city.
- **Poverty Rate:** The Correlation coefficient was -1.00 and the p value was 1.0000. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Poverty Rate and Crime Count in the city.
- **Gross Domestic Product (GDP):** The Correlation coefficient was 0.74 and the p value was 0.4721. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the GDP and Crime Count in the city.
- **Per Capita Personal Income:** The Correlation coefficient was 1.00 and the p value was 1.0000. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Per Capita Personal Income and Crime Count in the city.
- **Median Household Income:** The Correlation coefficient was 1.00 and the p value was 1.00. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Median Household Income and Crime Count in the city.
- **Income Inequality:** The Correlation coefficient was -1.00 and the p value was 1.000. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Income Inequality and Crime Count in the city.
- **Poverty Gap:** The Correlation coefficient was 1.00 and the p value was 1.000. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say

that there is no statistically significant correlation of the Poverty Gap and Crime Count in the city.

- **Consumer Price Index:** The Correlation coefficient was 0.64 and the p value was 0.00000001. Since the p value was less than 0.05, we rejected the null hypothesis. Hence, we can say that there is a statistically significant correlation of the Consumer Price Index and Crime Count in the city.
- **Inflation Rate:** The Correlation coefficient was 0.69 and the p value was 0.00000001. Since the p value was less than 0.05, we rejected the null hypothesis. Hence, we can say that there is a statistically significant correlation of the Inflation Rate and Crime Count in the city.
- **Housing Price Index:** The Correlation coefficient was 0.98 and the p value was 0.1194. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Housing Price Index and Crime Count in the city.
- **Housing Affordability Index:** The Correlation coefficient was 0.98 and the p value was 0.1194. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Housing Affordability Index and Crime Count in the city.
- **Housing Inventory:** The Correlation coefficient was -0.18 and the p value was 0.2368. Since the p value was not less than 0.05, we rejected the null hypothesis. Hence, we can say that there is no statistically significant correlation of the Housing Inventory and Crime Count in the city.
- **Federal Funds Rates:** The Correlation coefficient was 0.38 and the p value was 0.0109. Since the p value was less than 0.05, we rejected the null hypothesis. Hence, we can say that there is a statistically significant correlation of the Interest Rate and Crime Count in the city.

Day of the Week Analysis:

Group the data by day of the week and analyze crime frequencies for each day.

Analysis of Day of Occurrence of Crime:

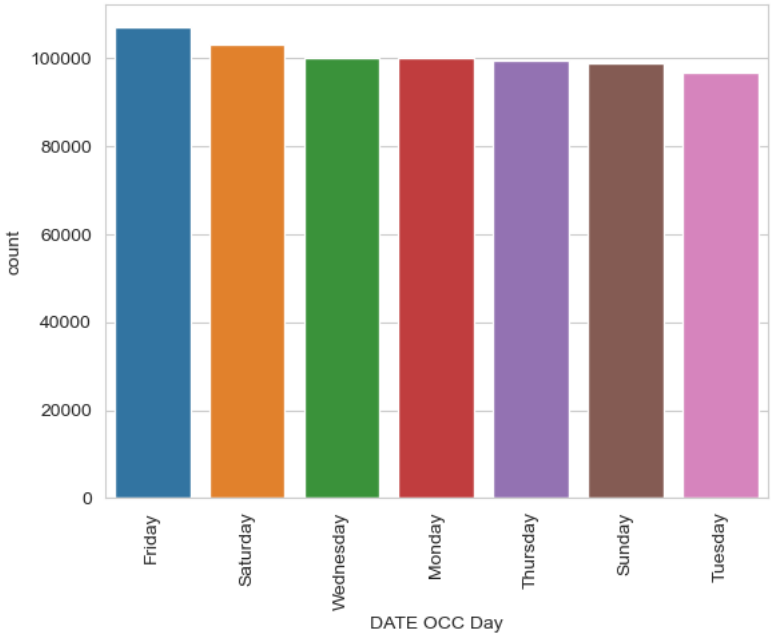

Fig 23: Bar Chart for Day of Occurrence of Crime

From the above chart, we can tell that most crimes happen on Friday and Saturdays. This suggests that it is more likely for crimes to happen on the beginning of weekends when people go out for gatherings.

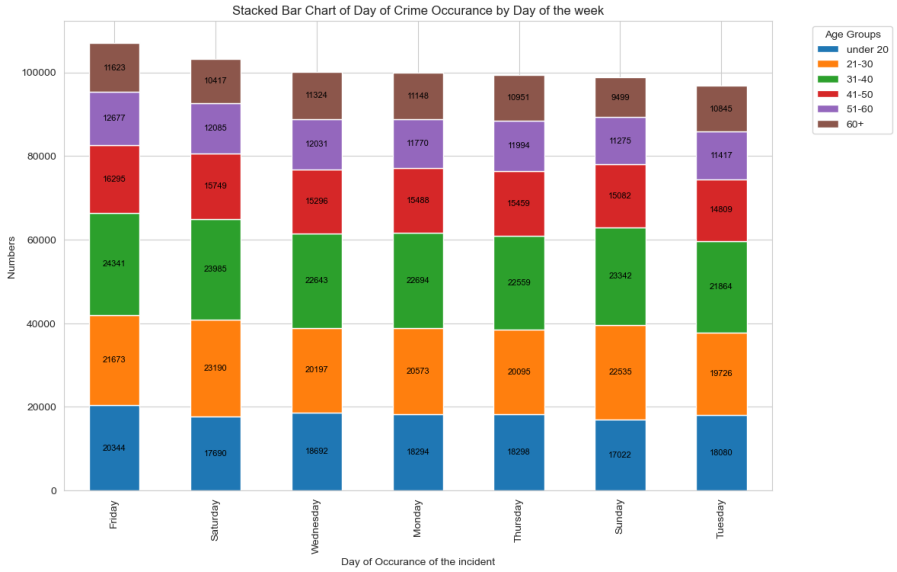

Fig 24: Stacked Bar Chart for Day of Occurrence of Crime by Age Group

The above chart shows that most of the crimes are committed to the people in the age group under 20, 20-30 and 30-40.

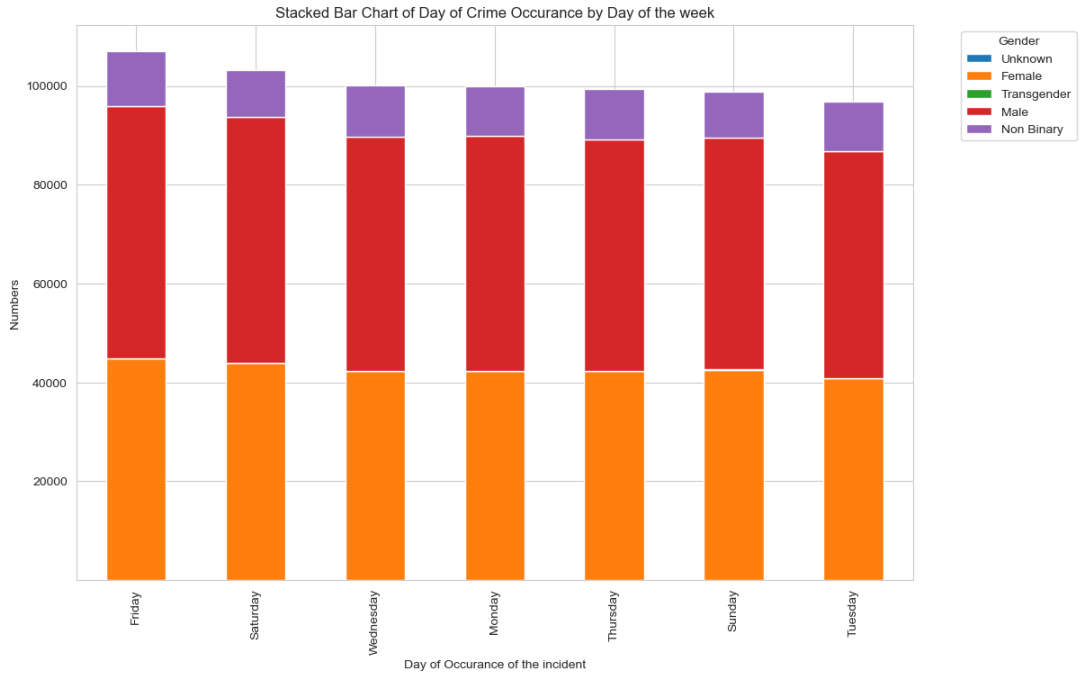

Fig 25: Stacked Bar Chart for Day of Occurrence of Crime by Day of the Week

The above chart shows that the crimes are mainly committed against Female, Male and Non-Binary people.

Analysis of Day of Reporting of Crime:

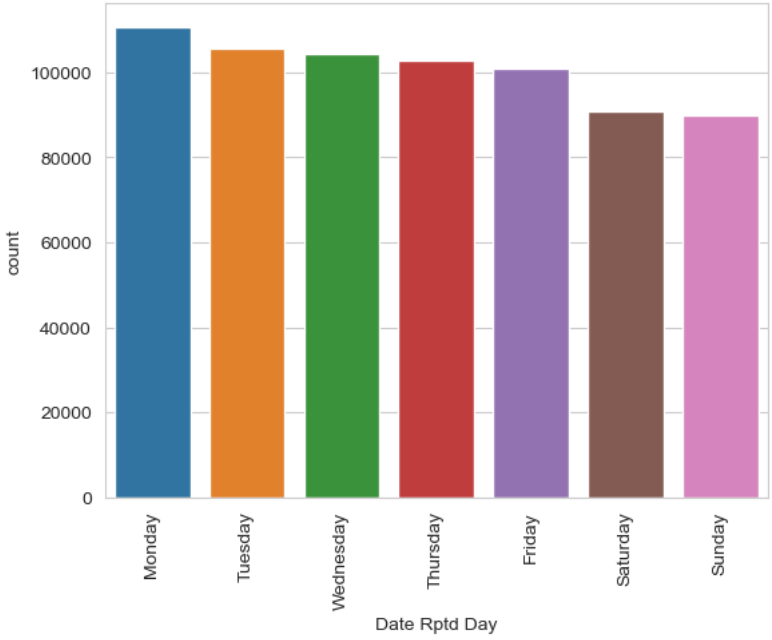

Fig 26: Bar Chart for Day of Reporting of Crime

From the above chart, we can tell that most crimes are reported on Mondays and Tuesdays. This suggests that it is more likely for people to report the crimes happened to them after the weekend is over.

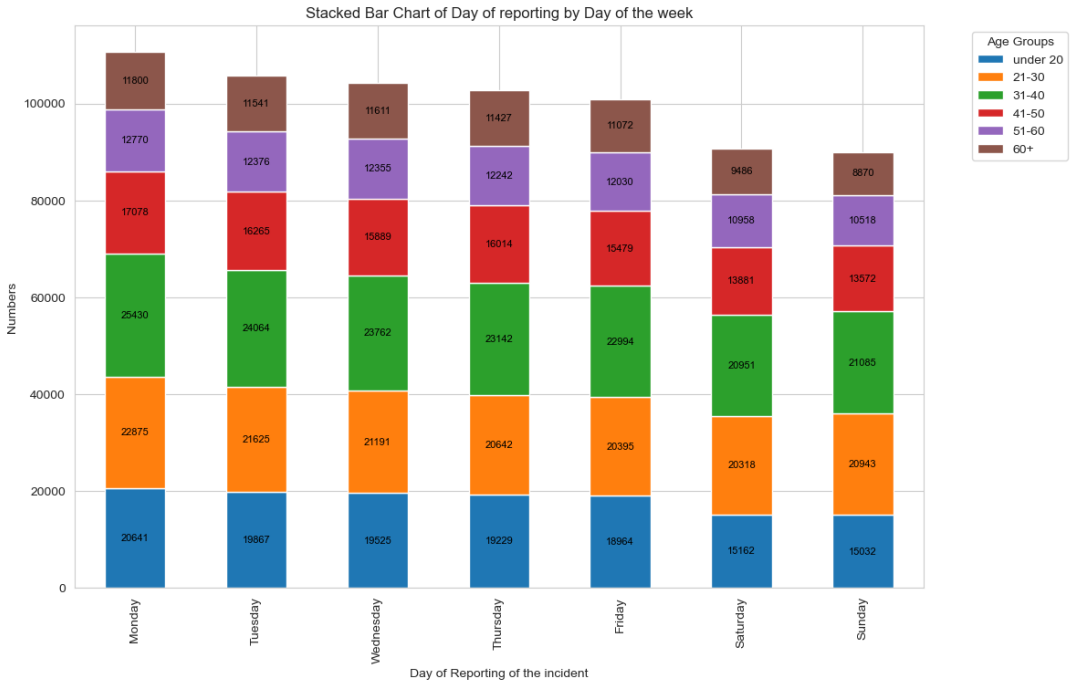

Fig 27: Stacked bar plot showing the day of reporting of incident by ages.

The above chart shows that most of the crimes are reported by the people in the age groups under 20, 20-30 and 30-40.

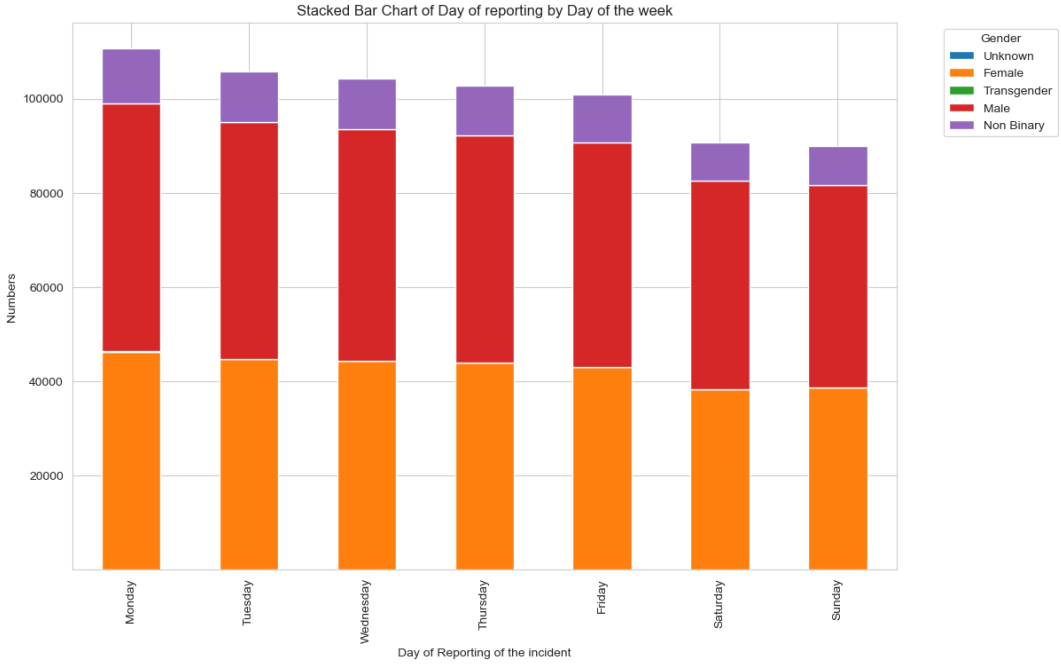

Fig 28: Stacked Bar chart showing day of reporting of incident by gender.
The above chart shows that the crimes are mainly reported by Female, Male and Non-Binary people.

Impact of Major Events:

Identify major events or policy changes during the dataset period and analyze crime rate changes before and after these events.

The time-period (2020-2023) under consideration includes several major events that may have had an impact on crime patterns. To analyze these potential impacts, we conducted a t-test, which is a statistical method used to compare means and assess the significance of differences between two groups. In the context of crime analysis, a t-test allows us to evaluate whether there are statistically significant changes in crime rates or patterns associated with these events.

a. COVID-19:

- The COVID-19 pandemic, which began in early 2020, has had widespread societal impacts, including on crime rates
- The result indicated that the analysis had detected a significant difference in crime rates before and after the COVID lockdown, as the null hypothesis had been rejected based on the t-test results.
- This suggests that the lockdown had a statistically significant impact on the number of crimes.

b. George Floyd Incident:

- The tragic event involving George Floyd in May 2020 led to widespread protests and discussions about racial justice and police reform.
- By conducting a t-test, the aim was to determine if there were any statistically significant changes in crime rates in the aftermath of the incident
- By the result of the t-test conducted, it can be concluded that there was significant difference in the number of crimes happened before and after the incident.

c. December Holiday Period (for three years 2020, 2021, & 2022):

- The December holiday period generally exhibits variations in crime rates.
- T-tests were conducted to evaluate whether significant differences in crime rates existed during the holiday periods across three consecutive years: 2020, 2021, and 2022.

i. *2020 Analysis*:

- For the December holiday period in 2020, t-tests were employed to scrutinize crime data.
- Results: No statistically significant change in crime rates was observed during this period in 2020 when compared to the non-holiday period.

ii. *2021 Analysis*:

- The analysis extended to include the December holiday period in 2021.
- Results: In contrast to 2020, a statistically significant change in crime rates was identified during the holiday season in 2021.

iii. *2022 Analysis*:

- T-tests were utilized to assess the December holiday period in 2022.

- Results: Similar to 2020, the data indicated no statistically significant change in crime rates during the 2022 holiday season.

- The findings imply that, while the holiday season did not significantly affect crime rates in 2020 and 2022, a distinct increase in crimes was evident in 2021.
- This emphasizes the need to consider annual variations and the specific dynamics of each holiday period when studying crime patterns and developing targeted strategies for crime prevention and law enforcement during the holiday season.

d. White House Attack (January 6, 2021):

- The attack on the U.S. Capitol on January 6, 2021, was a significant and unprecedented event in American history.
- T-tests were employed to investigate potential changes in crime rates and specific incident patterns in the aftermath of this event.
- The analysis did not reveal statistically significant alterations in crime rates or incident patterns, indicating that, from a statistical standpoint, the event did not immediately impact overall crime trends. Further localized research may be needed to explore specific effects.

e. US Presidential Election 2020 Results:

- Analysis of crime rates following the announcement of the U.S. Presidential election results in November 2020, using t-tests, revealed no statistically significant changes in criminal behavior.
- These findings suggest that the U.S. Presidential election outcome did not have an immediate substantial impact on crime rates in the examined area.

Outliers and Anomalies:

Use statistical methods or data visualization techniques to identify dataset outliers and investigate unusual patterns.

Using statistical methods and data visualization techniques, we identified the outliers and investigated the following unusual patterns in the victim age distribution.

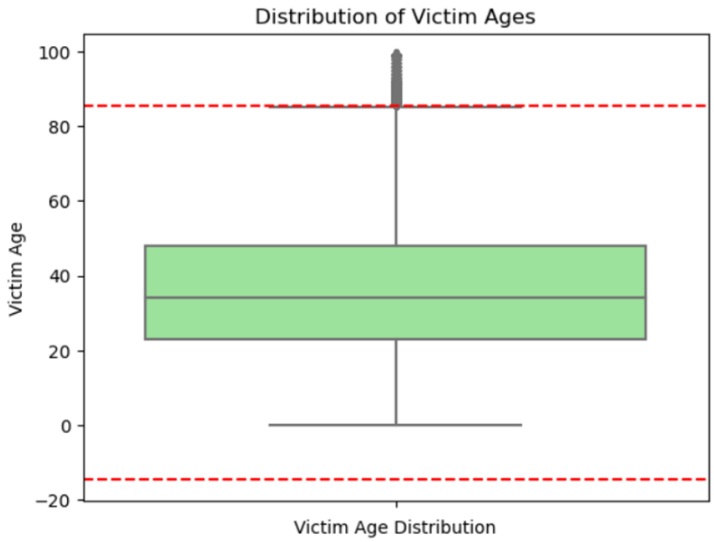

Fig 29: Boxplot showing the outliers in victim ages

The graph illustrates that a predominant portion of victims falls within the age range of 40 to 60, with a distinct peak at the age of 50. However, the graph also reveals the presence of outliers, situated both above and below the primary distribution.

- The median victim age is 50 years old, but there is a wide range of ages among victims. This suggests that crime can affect people of all ages, but older adults may be at an increased risk.
- There are outliers observed in the data, above the 3rd quartile.
- These outliers may be due to errors in data collection, unusual events, or targeted crimes against specific age groups.

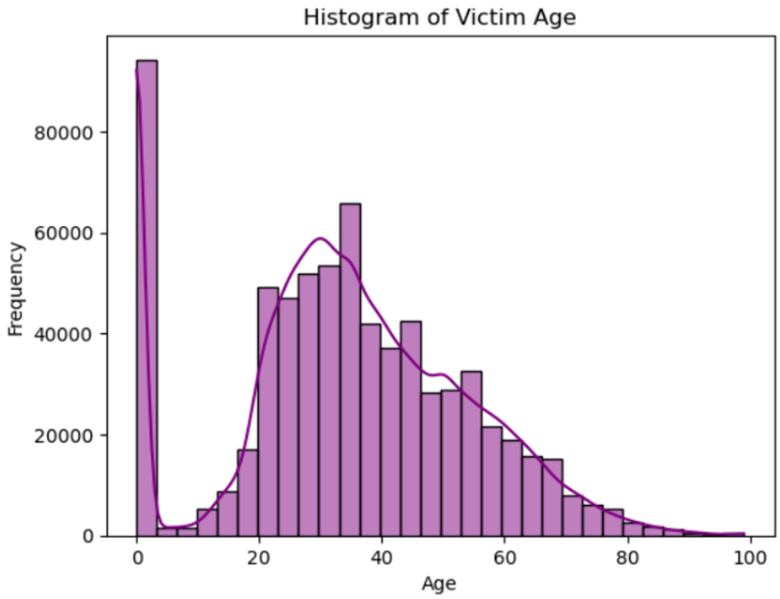

Fig 30:Histogram showing distribution of victim ages.

- The histogram plot determines that the median age of victims stands at 50 years. This indicates an even split where half of the victims are younger than 50, and the other half are older.
- Furthermore, the distribution of victim ages exhibits a right skew, implying that the data is characterized by a higher prevalence of younger victims in comparison to older individuals.

Demographic Factors:

- These 16 crime type out of 137 contributes to 80% of crime
- In the majority of crime types, men have the highest crime counts

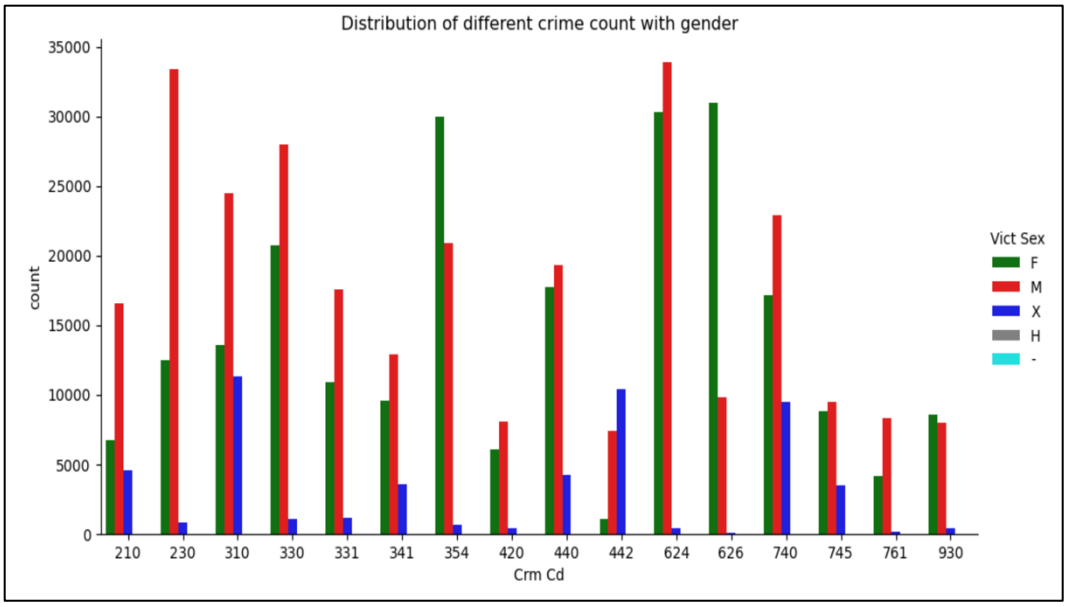

Fig 31: Bar plot showing the distribution of crime count by victim gender

- The age group between 20 and 40 experiences the highest incidence of crime type 624, specifically Battery – Simple Assault
- Individuals aged 30 to 40 are primarily impacted by crime type 354, which is Theft of Identity
- The age group ranging from 25 to 35 experiences the greatest influence of crime type 626, namely, Intimate Partner simple assault

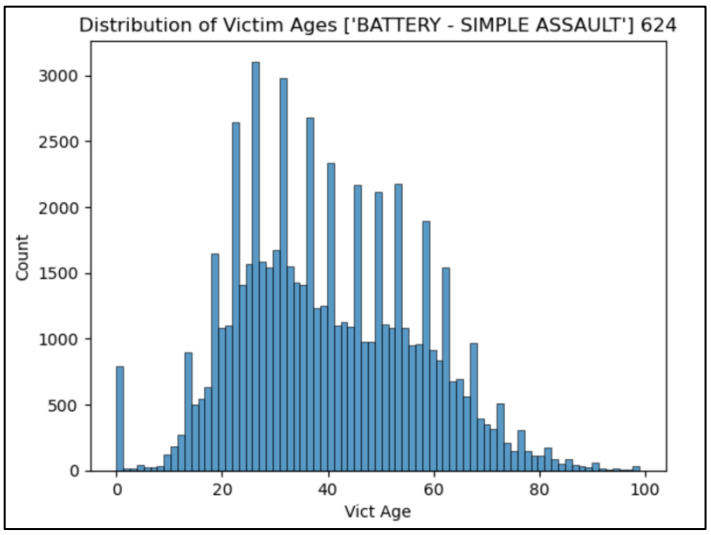

Fig 32:Histogram showing distribution of Victim age per reported incident

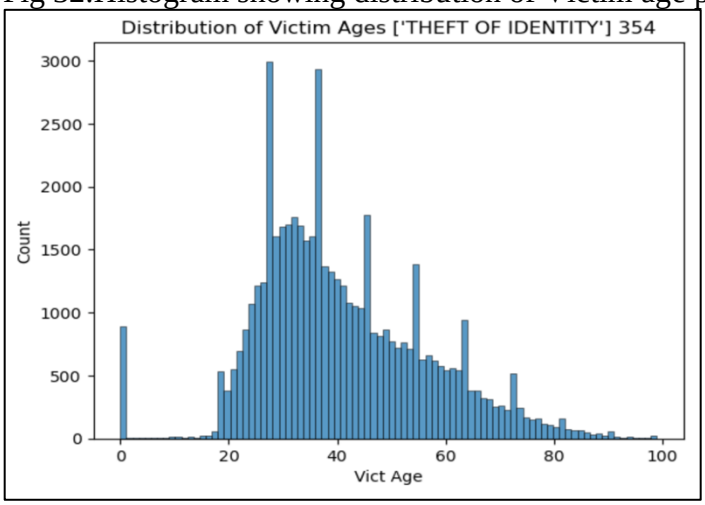

Fig 33: Histogram showing the distribution of victim age per reported incident

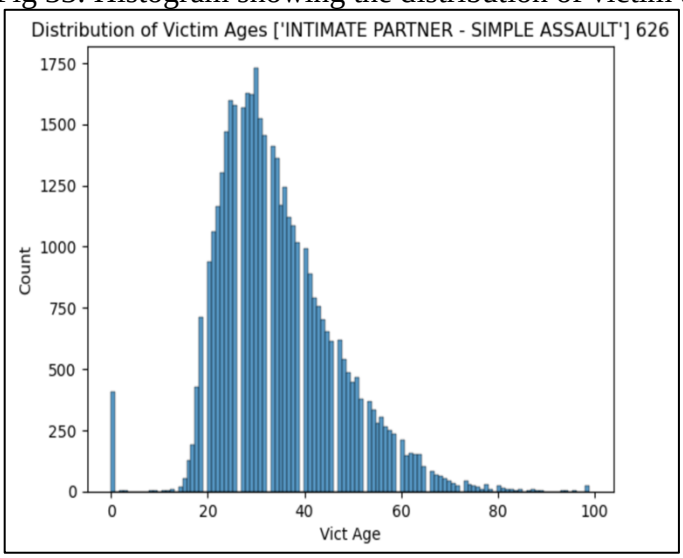

Fig 34: Histogram showing the distribution of victim ages per reported incident

Advanced Analysis

Time Series Forecasting Methods with ARIMA:

ARIMA (Autoregressive Integrated Moving Average) is a widely used time series forecasting model. It's a combination of autoregressive (AR) and moving average (MA) components with differencing (I). These components help capture and model different patterns and trends in time series data.

Predict Future Crime Trends:

The objective here is to make predictions about the daily future behavior of a crime rate at National level with and without regressors. It involves analyzing historical crime data to understand its patterns and using this understanding to forecast what is likely to happen in the future.

Baseline Forecast without Regressors:

The analysis starts by developing a basic forecast without considering external factors. This provides a starting point for understanding the inherent trends in the historical crime data and how well the ARIMA model captures those trends.

Inclusion of External Factors:
To improve the accuracy of crime rate predictions, the analysis considers external factors that might affect crime trends. For example:

- **Unemployment Rate**: High unemployment can sometimes lead to an increase in certain types of crimes.
- **Inflation Rate**: Inflation can impact the cost of living and economic conditions, which may influence crime rates.
- **Federal Funds Rates**: Interest rates can influence economic conditions, which, in turn, can impact crime.
- **Day, Month, Week, and Year**: Temporal factors like the day of the week, month, or year can reveal seasonality in crime patterns (e.g., more crimes during weekends or in certain months).

Forecast:

- Trend line of actual crime count vs predicted crime count at daily level with no regressor using ARIMA model is shown below.

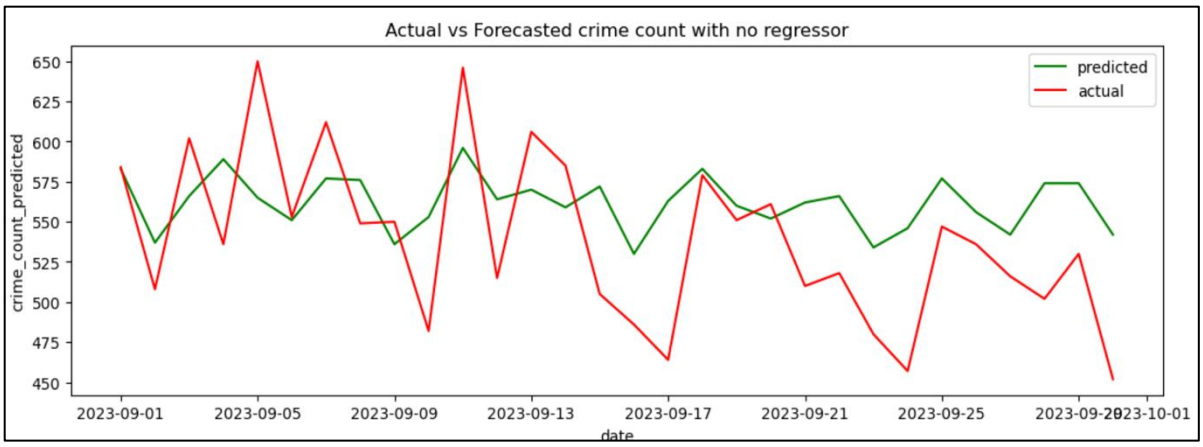

Fig 35: Line chart showing the actual and forecasted crime count with no regressor.

- Trend line of actual crime count vs predicted crime count at daily level with day as a regressor using ARIMA model is shown below.

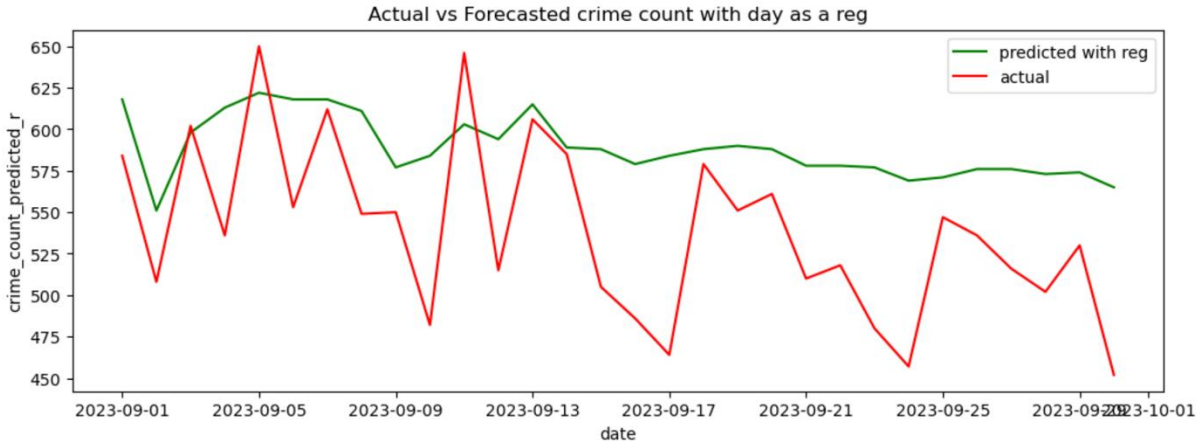

Fig 36: Line chart showing the actual and forecasted crime count with day as a regressor.

- Trend line of actual crime count vs predicted crime count at monthly level with Unemployment rate, Inflation rate, Federal Funds rate and consumer price index as a regressor using ARIMA model is shown below.

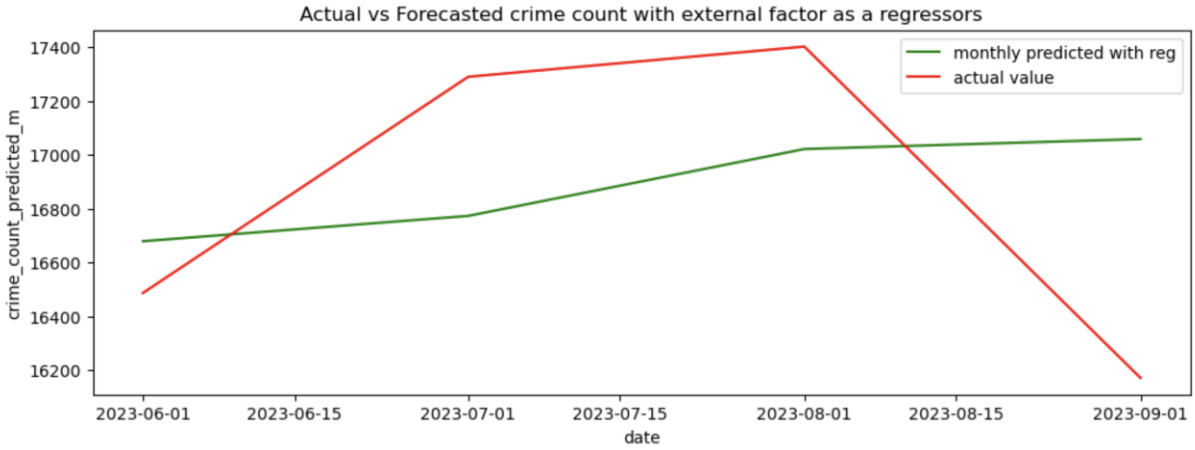

Fig 37: Line chart showing the predicted and actual crime when we use external regressor

Incorporating Regressors:

After creating the baseline forecast, the analysis introduces the external factors (regressors) into the model one at a time. This step allows for an assessment of how each external factor influences the accuracy of the crime trend predictions. For example, by including unemployment rates, the analysis can assess whether there's a significant correlation between unemployment and crime rates.

This approach combines statistical modeling (ARIMA) with external economic and temporal factors to create more accurate predictions of future crime trends. By first establishing a baseline forecast and then adding regressors, it provides insights into the factors that drive changes in crime rates, helping policymakers and law enforcement agencies make informed decisions.

References

Economic Data were collected from the following websites:

- Housing Price Index Data: https://fred.stlouisfed.org/series/ATNHPIUS06037A
- Housing Inventory Data: https://fred.stlouisfed.org/series/ACTLISCOU6037
- Unemployment Data: https://fred.stlouisfed.org/series/CALOSA7URN
- GDP LA Data: https://data.oecd.org/gdp/gross-domestic-product-gdp.htm
- Poverty Gap: https://data.oecd.org/inequality/poverty-gap.htm
- Poverty Rate: https://data.oecd.org/inequality/poverty-rate.htm#indicator-chart
- Income Inequality: https://data.oecd.org/inequality/income-inequality.htm#indicator-chart
- Per Capita Personal Income: https://fred.stlouisfed.org/series/PCPI06037
- Median Household income: https://fred.stlouisfed.org/series/MHICA06037A052NCEN
- Inflation Rate: https://www.usinflationcalculator.com/inflation/current-inflation-rates/(please note, we made a csv from the data mentioned on the website from 2020 –sep 2023)
- Consumer Price Index: https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/(please note, we made a csv from the data mentioned on the website from 2020 –sep 2023)
- Interest Rate: https://fred.stlouisfed.org/series/FEDFUNDS