CLEANING AND ANALYZING CRIME DATA

IE6400 – Foundation of Data Analytics Engineering

Final Report

Group Number 28

Batta Aditya Yadav (002874554)

Hrishik Bhaven Parmar (002815908)

Karan Dalal (002836524)

Aishwarya Belavakadi Subrahmanya (002820128)

1. Background and Project Context

1.1 Background and Project Overview

Crime investigation and forecasting are essential elements of law enforcement and public safety. The use of data analytics in understanding and addressing crime in an era of increasing data availability has become a valuable tool for law enforcement, policymakers and researchers The availability of comprehensive crime data allows for the adoption of a more data-driven and targeted approach to crime prevention, resource allocation, and policy development.

The "Cleaning and Analyzing Crime Data" project is a data analytics and engineering effort aimed at harnessing the power of data to provide insights into crime patterns, trends, and factors affecting crime rates. The project delivers real-world crime data from 2020 applied to the present, providing more information on types of crime, locations, and trends over time.

1.2 Project Objectives

The main objectives of this program are as follows:

1. **Data Acquisition:** To download and input crime data into our data analysis tool of choice
2. **Data Analysis:** Analyze and understand the structure of the data set, data types, and any available column descriptions.
3. **Data cleaning and preparation:** Ensure that the data set is ready for analysis by addressing issues such as missing data, duplication, data type changes, and standardization.
4. **Exploratory Data Analysis (EDA):** To identify valuable insights from various data sources, including overall crime trends, seasonal patterns, crime frequency, regional differences, relationships among economic factors, analysis of the week, and the impact of major events or policy changes.

1.3 Dataset Description and Source

About the Dataset

The dataset utilized in this project captures incidents of crime in the City of Los Angeles, dating back to 2020. It is important to note that this data is transcribed from original crime reports that were originally documented on paper. As a result, there may be some inaccuracies within the dataset due to human error or variations in transcription. For instance, some location fields may have missing data. Additionally, address fields are only provided up to the nearest hundred blocks to preserve the privacy and security of individuals. Despite these limitations, the dataset is as accurate as the data recorded in the original crime reports. Any questions or concerns about the data have been documented in the comments.

Source: Crime Data from 2020 to Present

1.4 The Significance of Data Analytics in Crime Analysis

For law enforcement, public safety, and policymaking, the incorporation of data analytics into crime analysis is crucial. In the following respects, it transforms the conventional wisdom on understanding, preventing, and reducing crime:

- **Data-Driven Decision Making:** In a criminal investigation, data analytics facilitates the shift from gut feelings to data-driven tactics. It enables analysts and law enforcement organizations to make well-informed judgments based on empirical evidence by utilizing large and diverse databases. This method improves policy development and resource allocation, making them more successful and efficient.
- **Crime Pattern Recognition:** With the use of data analytics, analysts may find complex patterns and trends in crime data that would be missed by more conventional analytical techniques. It makes it easier to pinpoint hotspots,

temporal patterns, and high-crime locations, which makes it possible to create focused crime prevention and intervention strategies.

- **Predictive Policing:** The creation of predictive police models that use previous data to forecast possible criminal activity is made possible by sophisticated data analytics. These models give law enforcement organizations important information on how to best allocate resources, including stepping up patrols in high-crime areas.
- **Resource Optimization:** Analytical data helps make wise decisions about how to use scarce resources. Law enforcement organizations may focus their efforts in high-crime areas by identifying them and doing so, which makes their operations more efficient.
- **Crime Reduction:** Through the identification of significant variables and patterns in criminal activity records, data analytics provides a practical understanding of the fundamental reasons behind crime. Customized policies and actions targeted at lowering crime rates and improving public safety can be informed by these data portions.
- **Transparency and Accountability:** The use of data analytics in law enforcement activities encourages openness. Thorough data analysis increases public trust and accountability by giving law enforcement's activities a transparent, objective foundation.
- **Continuous Improvement:** The use of data analytics enables the ongoing evaluation and modification of crime prevention and reduction tactics. Regular data analysis ensures that strategies remain successful and relevant, allowing law enforcement organizations to adjust to changing crime patterns.

2. Data Acquisition and Initial Inspection

2.1 Data Acquisition

After analyzing the data scraped from Equity Residential, it is concluded that there is, in fact, a relationship between apartment features and amenities and the price of the apartment. This was a conclusion drawn regardless of what city was being analyzed. Although price ranges for apartments varied by city, it was noted that all cities showed the same basic trends, such as the price of apartments increasing as the square footage increased. An additional conclusion drawn that was unexpected was that while pricing trends were generally consistent, the average prices themselves varied by zip code, perhaps due to outside factors such as proximity to public transportation, parks, etc.

We have used Tableau to bring out the insights from city-level data to highlight the apartments available in each city and all the amenities that are on offer for each of them. The various amenities and other features like price, square feet, and location play the major selection criteria for renting an apartment.

The data was cleaned and processed completely using Python's Pandas library, and the data was connected to Tableau. The project consists of two dashboards; the first dashboard is a general analysis of the apartments in the user-selected city, whereas the second dashboard tells us about individual cities and their respective amenities in comparison to another city that utilizes the same dataset to bring about different insights for the end user to choose and decide upon when moving into a new city.

2.2 Initial Inspection

We conducted an initial inspection of the dataset in this project part, exposing its size and structure. The dataset consists of 28 columns and a sizable 820,599 rows. We were able to close this gap even though column descriptions were initially missing by gathering metadata from outside sources. One notable column that became apparent was 'Part 1-2,' which classified

instances into Part 1 and Part 2 crimes or offenses. These two classifications provide important context for our dataset study by differentiating between major and less serious offenses.

This data inspection step lays the groundwork for the next stages of our investigation. It helps us understand the general structure of the dataset, identify any problems with the quality of the data, and create a schedule for preprocessing and data cleaning. With this crucial background knowledge, we may proceed with a more thorough analysis of the dataset, drawing insightful conclusions and useful patterns from the abundance of data it has.

3. Data Cleaning and Preparation

3.1 Dropping Unnecessary Columns

In this section, we initiated the data-cleaning process by examining and counting the missing values in each column. The results revealed that several columns had varying percentages of missing data. This assessment allowed us to gain a comprehensive understanding of the data quality issues in our dataset.

We decided to remove a few columns from the dataset that weren't relevant for our research to make it more manageable. 'DR_NO' (Division of Records Number), 'AREA' (identifier), 'Premis Cd' (code for location type), 'Weapon Used Cd' (code for weapon type used), 'Mocodes' (Modus Operandi codes), 'Crm Cd 1,' 'Crm Cd 2,' 'Crm Cd 3,' and 'Crm Cd 4' (codes for different types of crimes), 'Status' (case status), 'Status Desc' (status description), 'Cross Street' (distance information), and 'LOCATION' (street address) are among the columns that were removed. The primary goal of removing these columns was to streamline the dataset and make it more pertinent to our analytical goals. These columns either included unnecessary information or were unrelated to the analysis we intended to carry out.

3.2 Data Cleaning and Transformation - Date, Time, and Location

During the initial phase of the data transformation and cleaning procedure, we concentrated on the dataset's temporal features. First, we took the date out of the 'Date Rptd' column and changed its name to 'Date Reported'. We were able to improve the temporal representation as a result. To construct a consistent dataset free from the effect of missing data from the present month of October 2023, we then selected a specified period from January 1, 2020, to September 30, 2023. This guarantees that past crime trends are covered by our study up to this month.

Next, we processed the 'DATE OCC' column similarly, extracting the date, renaming it as 'Date Occurred,' and aligning the date format with 'Date Reported.' Additionally, we formatted the 'TIME OCC' column uniformly to show the time of day when occurrences happened. This

required changing the format to 24 hours, splitting the hours from the minutes, and renaming the column "Time Occurred." For clarity, we have changed the columns 'LAT' and 'LON' to 'Latitude' and 'Longitude'.

3.3 Data Cleaning and Transformation - Victim Information

We dealt with concerns connected to the victim's information at this point. First, by identifying and addressing problematic items, we examined and enhanced the 'Vict Age' column. We replaced the inaccurate information with more relevant numbers by filtering out negative ages and ages over 100. A more accurate age distribution is an aftereffect of this technique.

In addition, we included a new column called "Age Group" to organize victims according to age into categories like "Child," "Teen," "Adult," and "Old." Our capacity to analyze the dataset is improved by this segmentation, which sheds light on the relative contributions of different age groups to reported crimes.

The 'Vict Sex' column was the last area we worked on to enhance the data's uniformity and readability. We classified values ending in "H" and "-" as "unknown" after mapping them to complete gender names. This data transformation ensured that the column accurately represented the sex of the victim.

3.4 Data Cleaning and Transformation - Victim Descent, Location, and Weapon

We addressed the 'Vict Descent' column to improve the dataset's informational value and ease of use. By employing a predetermined mapping to offer descriptive labels for the coded values, we improved the readability of the data. Values with a minus sign (-) were likewise labeled as "unknown." Using the 'Premis Desc' column, we took a similar tack by renaming it 'Crime Location' to better represent the kind of place where each crime was committed and substituting 'Unknown' for null values.

Finally, we improved the 'Weapon Desc' column by calling it 'Weapon Used Description' to make it clearer. We made sure that in this column, null values were shown as "unknown." After these changes, the dataset was better organized, useful, and prepared for more research.

4. Exploratory Data Analysis (EDA)

4.1 Visualizing Overall Crime Trends (Yearly)

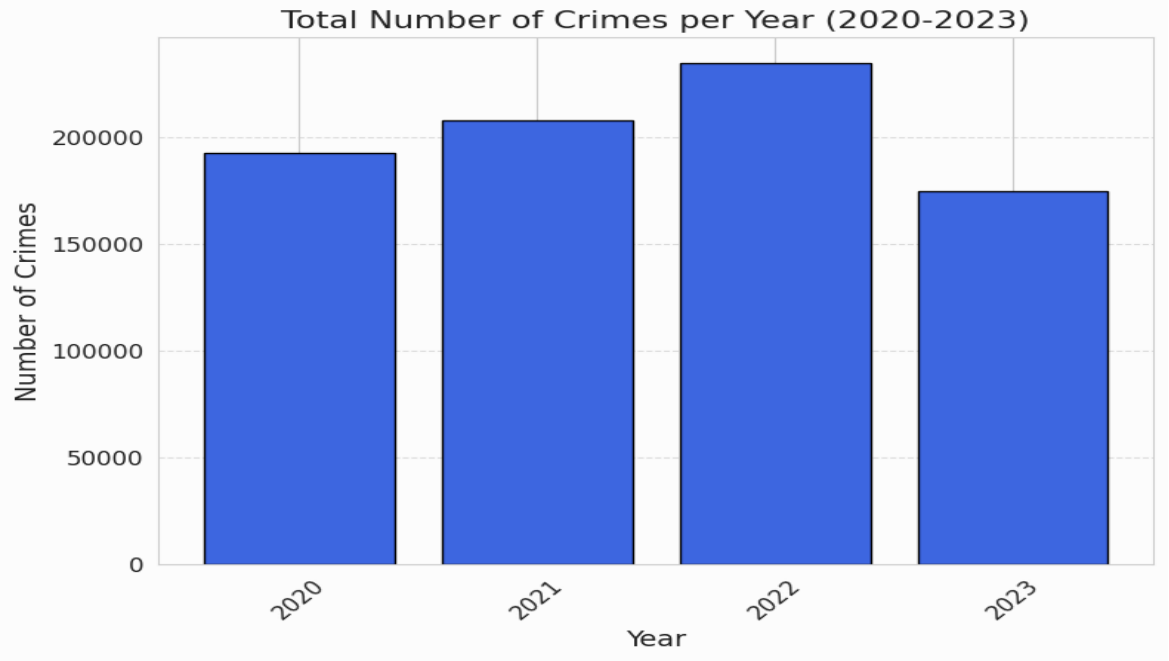


fig4.1.1

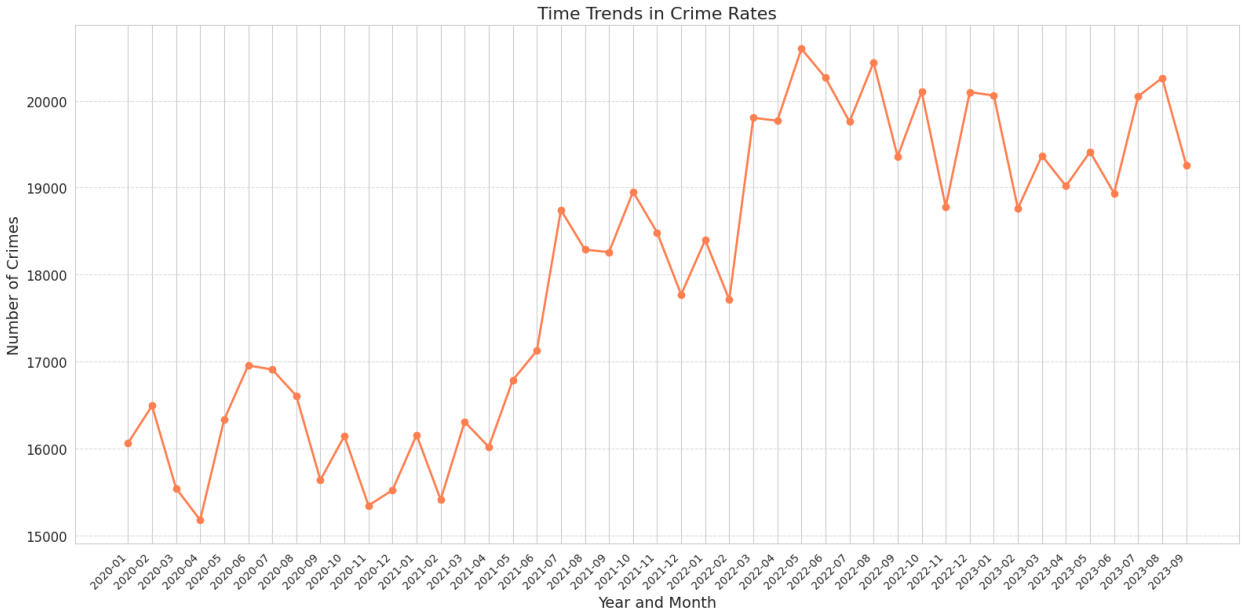


fig4.1.2

Figures **fig4.1.1 and fig4.1.2** show how many crimes are reported each year/month. There are several key points to take away from this visualization.

The overall trend shows that the number of crimes is on the rise between 2020 and 2023. The visualization also shows that there are some year-on-year variations in crime rates. The peaks in the chart are associated with the highest crime rates during that particular year. Although the figures show that crime rates are rising from 2020 to 2022, a fall in 2023 may have been due to data irregularities, and additional analysis needs to be made before an overall picture of criminal trends can be established. Moreover, when analyzing the trends in crime, it is appropriate to take into account other factors such as changes in law enforcement practice, seasonal patterns, demographics, and societal factors. This data is critical for law enforcement and policymakers because it allows them to adapt policies to the changing nature of crime throughout the year.

4.2 Analyzing Seasonal Patterns in Crime Data

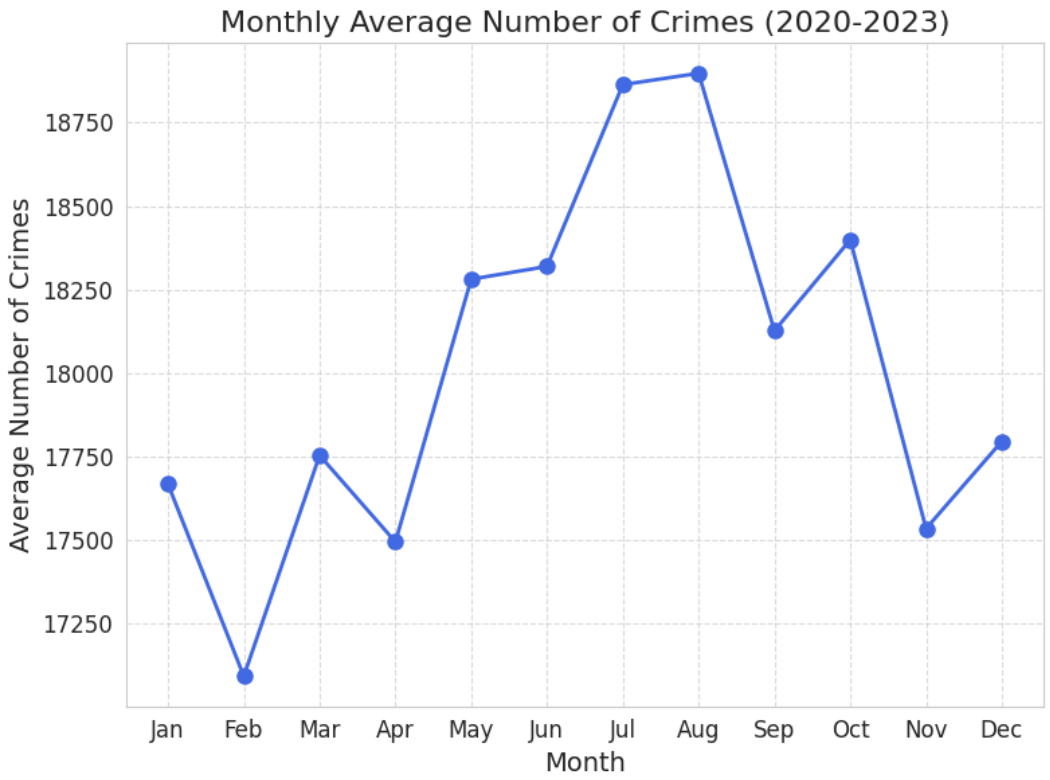


fig4.2.1

Figure **fig4.2.1** illustrates the monthly average crime rates over the selected period. Notably, the highest averages occur in July and August during the summer months, when warmer weather often leads to increased criminal activity. Conversely, crime rates are lower from November to February during the winter, likely due to reduced outdoor activity. December shows a slight increase, possibly due to holiday-related activities, while November has the lowest average crime rate, coinciding with reduced outdoor activities and the holiday season.

Seasonal Variations: The chart can aid in the identification of potential seasonal variations in crime. The following visualization provides a clear and simple representation of the average number of crimes recorded each month. It is a fundamental instrument for monitoring monthly crime trends, facilitating more in-depth analysis, and assisting decision-making in the fields of crime prevention and law enforcement.

4.3 Identifying the Most Frequent Crime Type

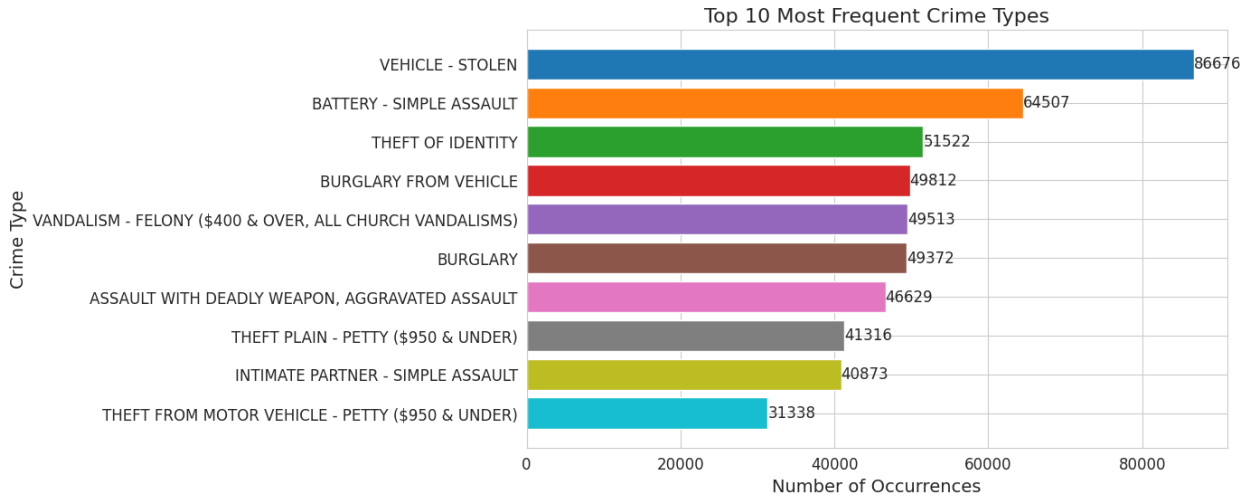


fig4.3.1

It was found according to **fig4.3.1** that 'VEHICLE - STOLEN' was the most common crime type, accounting for 86,680 instances. Most criminals steal vehicles because it is easy to manipulate

car keys if it is not properly locked. Furthermore, we identified the top ten most common crime categories, which range from 'BATTERY - SIMPLE ASSAULT' to 'VANDALISM - MISDEMEANOUR ($399 OR UNDER).'

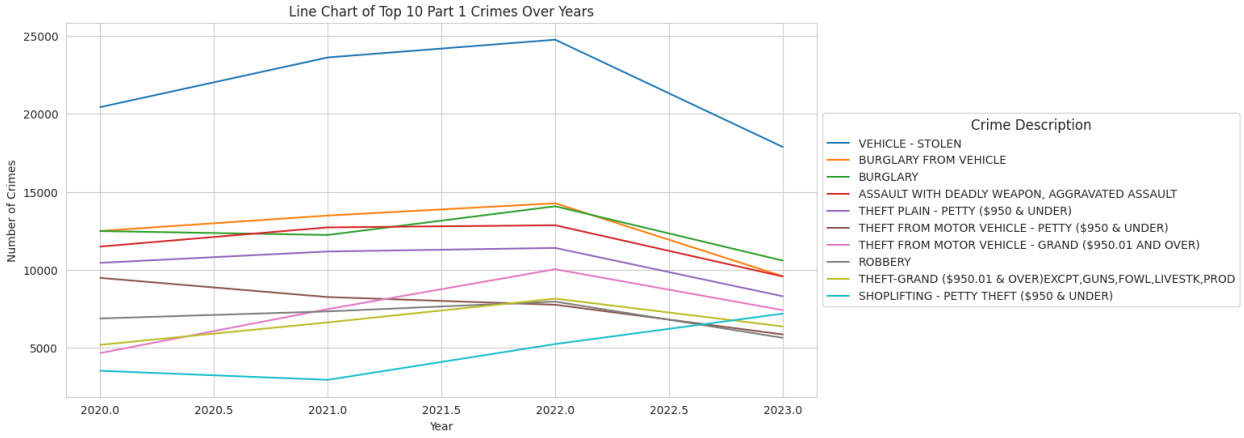


fig4.3.2

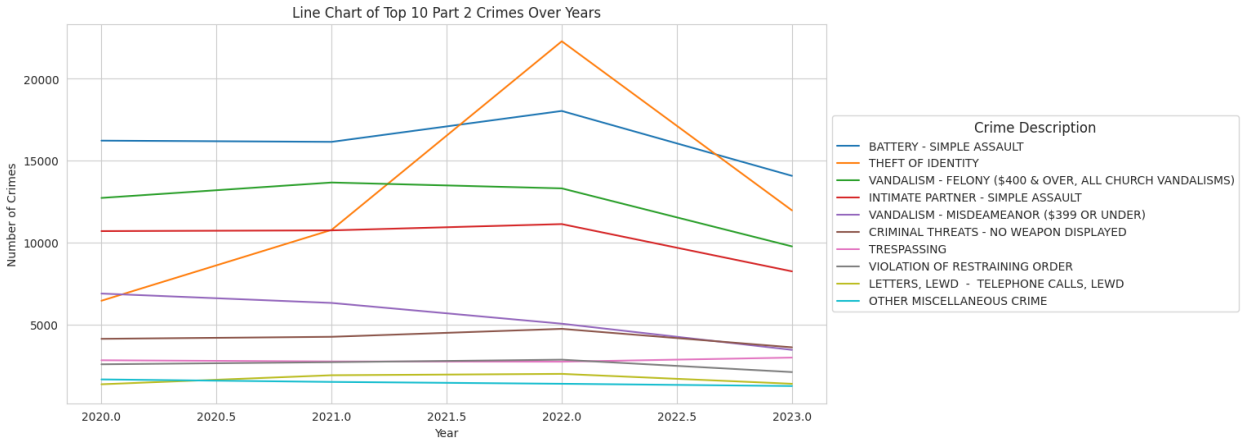
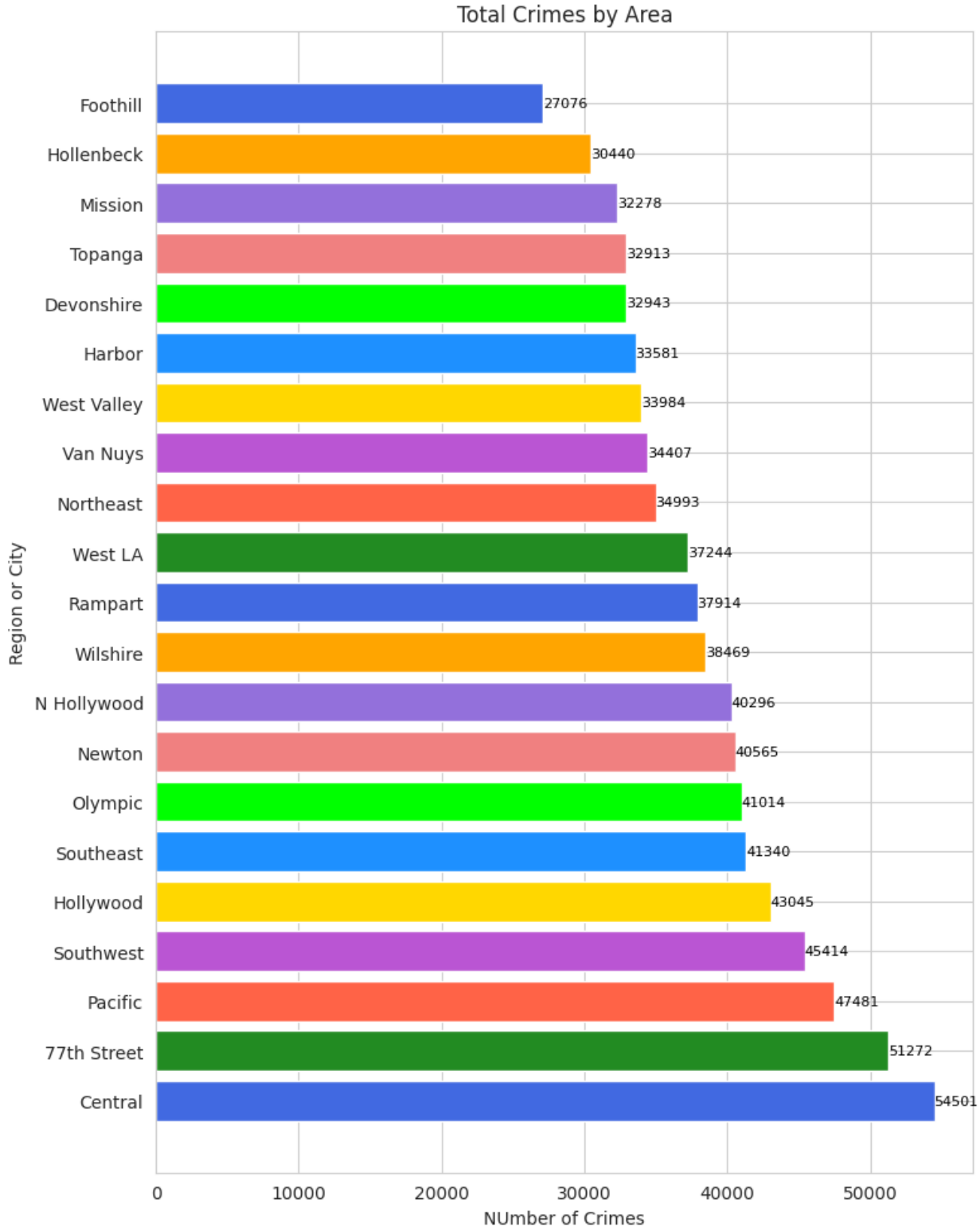


fig4.3.3

We focused on determining the most common in this Part 2 analysis (Fig. 4.3.3). With 64,509 reported instances, the study found that 'BATTERY - SIMPLE ASSAULT' was the most often occurring Part 2 offense. 'OTHER MISCELLANEOUS CRIME', on the other hand, was the least common Part 2 crime, with only 5,864 reported incidences.

4.4 Regional Differences in Crime Rates



(Fig. 4.4.1)

The total number of crimes in 21 distinct areas or cities is shown visually in this bar chart. Foothill has the lowest crime rate at 3.34%, while Central has the highest overall crime rate, making up around 6.82% of all recorded offenses. These areas report 38,627 offenses on average**. (Fig4.4.1)**

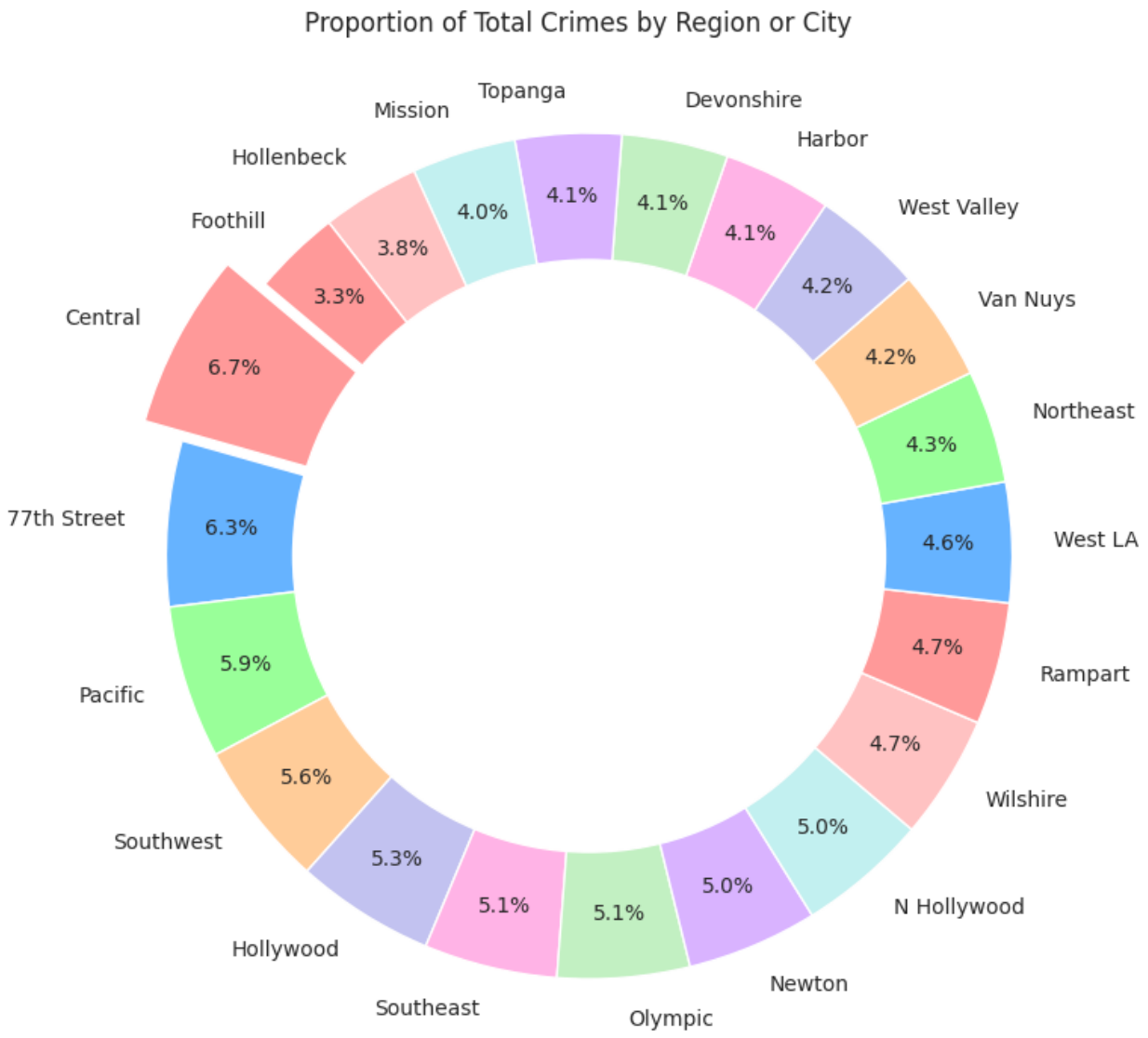


Fig4.4.2

The diagram (fig4.4.2) depicts a pie chart that shows the percentage of total offenses in various cities or areas. It determines the proportion of all crimes in every area and illustrates the differences in the distribution of crimes. According to the graph, "Central" has the greatest proportion of all offenses at 6.72%, followed by "77th Street" at 6.32%. An explosion effect has been created to highlight the "Central" region. The resultant pie chart gives a fast and simple summary of the regions with the greatest and lowest crime rates by visualizing the distribution of crimes across those regions.

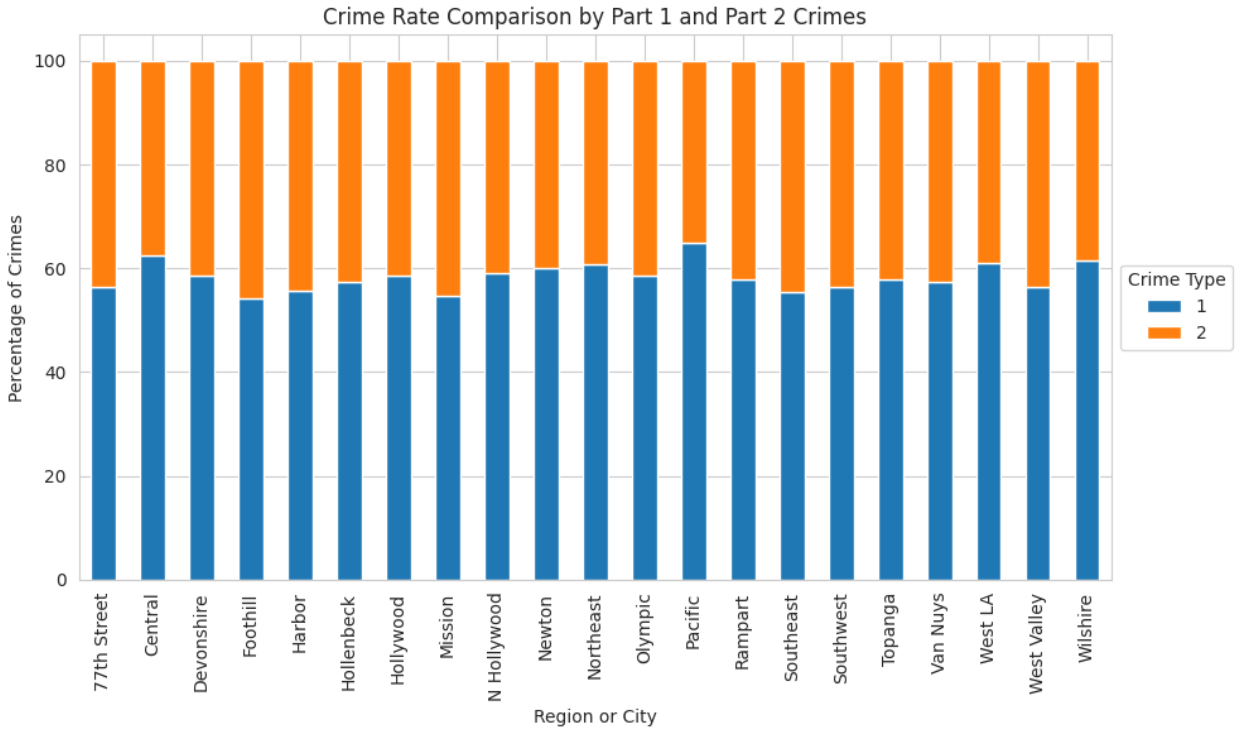


Fig4.4.4

This (fig4.4.4) contrasts how Part 1 and Part 2 offenses are distributed across various geographic areas. It demonstrates that with Part 2 crimes coming in second, Part 1 crimes make up the majority of criminal offenses in most locations, accounting for over 50% of all offenses. Areas like the Pacific and Olympic have a very high percentage of Part 2 criminal offenses.

4.5 Correlation Analysis with Economic Factor

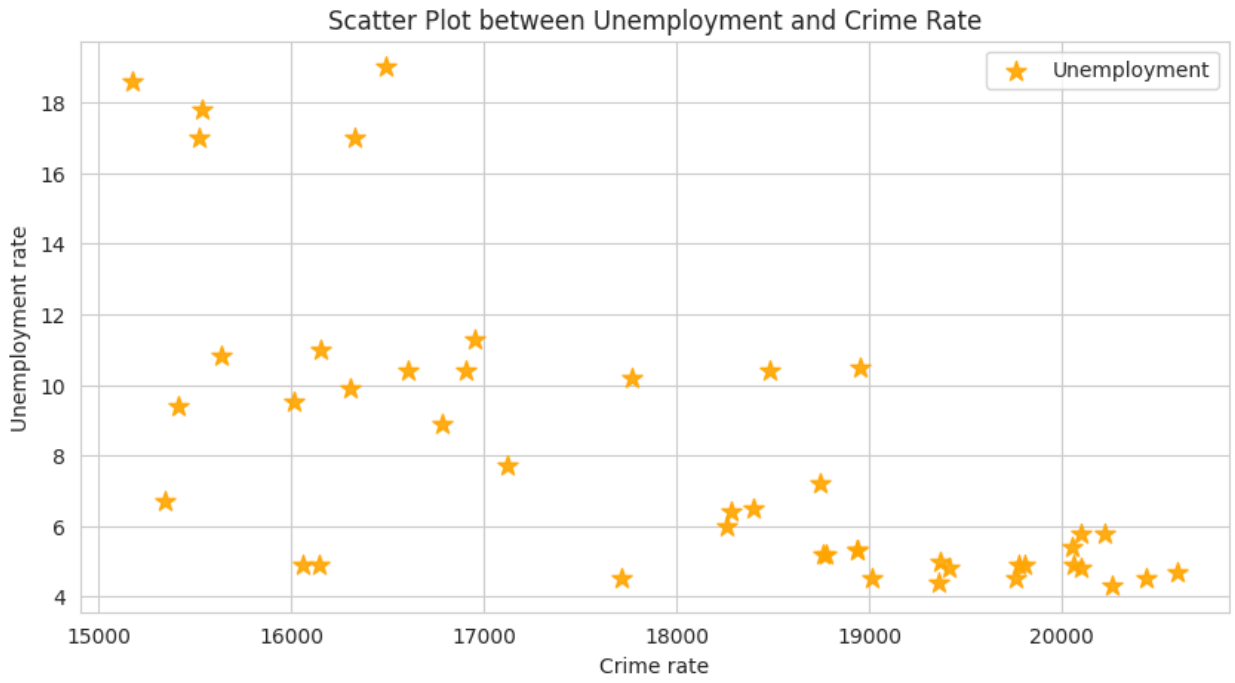


Fig4.5.1

This indicates that while the "Crime Rate" rises, showing an increase in the number of recorded crimes, the "Unemployment Rate" tends to fall, suggesting that fewer individuals are out of work. In other words, these two variables move oppositely.

This inverse association may appear to be contradictory, but it could be due to a variety of variables. When crime rates rise, for example, authorities and communities may be motivated to focus on improving safety and creating jobs, which can lead to a fall in unemployment. As a result, the negative relationship between crime and unemployment can be a complicated and dynamic interplay driven by a variety of circumstances

4.6 Day of the Week Analysis

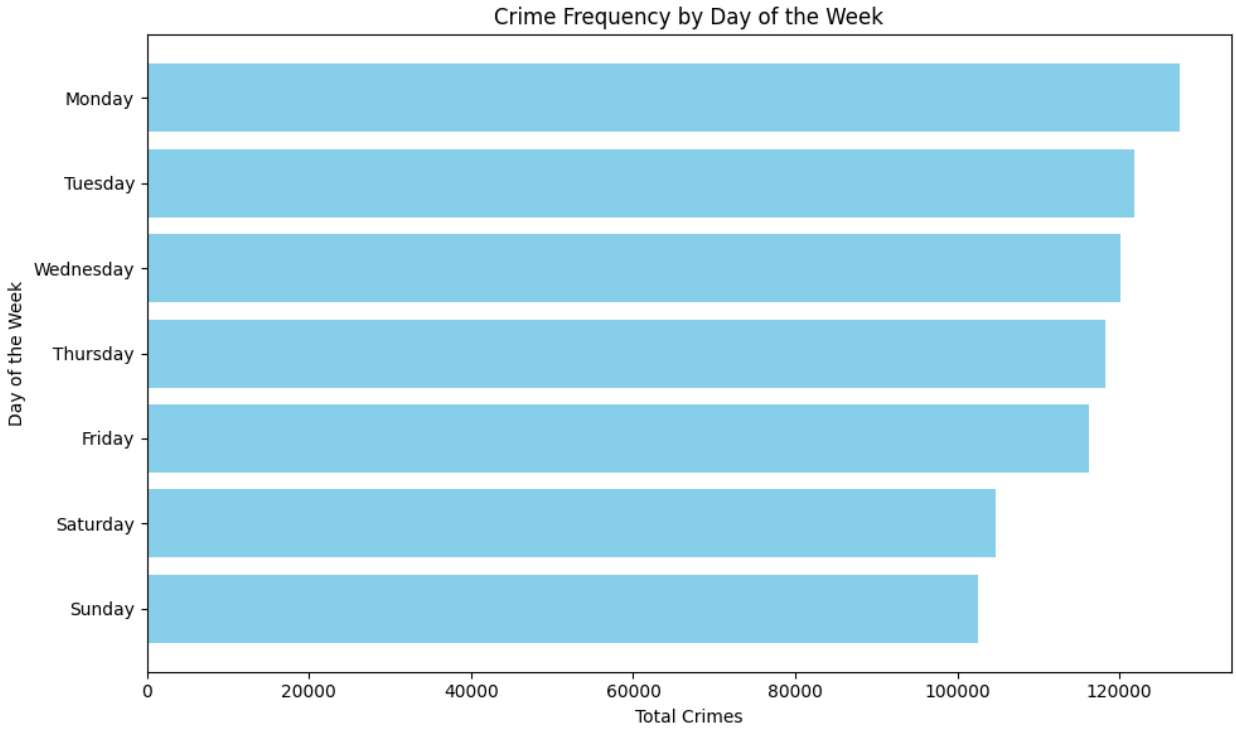


Fig4.6.1

The research on crime frequency by weekday yielded intriguing insights into criminal activity patterns. Mondays had the most recorded crimes, with 127,397, while Sundays had the fewest, with 102,511 incidents. This trend suggests a possible link between the start of the workweek and an increase in criminal activity, which might be attributed to the return to routine and business activities. Weekends, on the other hand, had a lower crime rate, indicating that people were likely engaged in leisure and recreational activities. The findings highlight the importance of temporal aspects in understanding crime trends and can help law enforcement agencies allocate resources and develop targeted crime prevention tactics. (fig4.6.1)

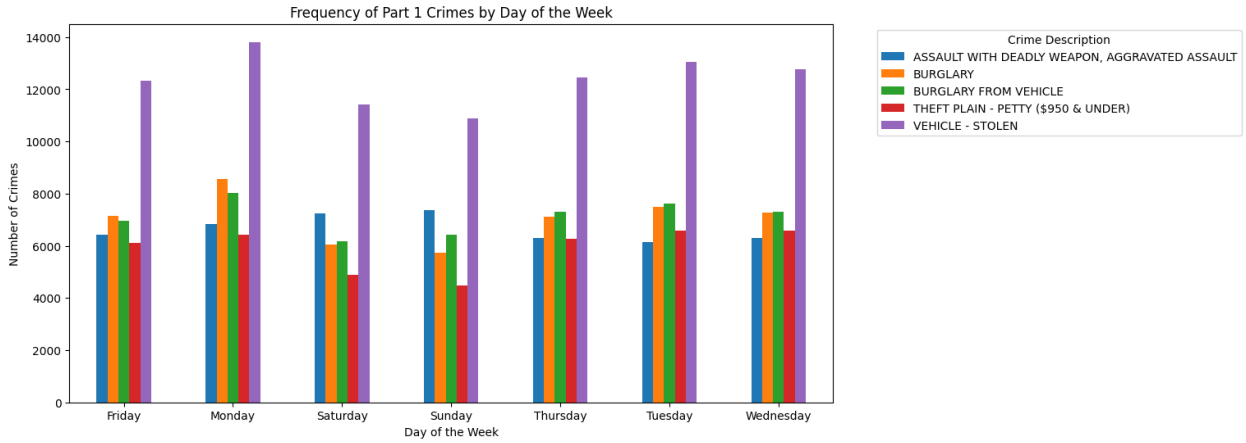


Fig4.6.2

This (fig4.6.2) focuses on the analysis of the top 5 Part 1 offenses and their corresponding frequencies on various days of the week. A bar chart is used to visually represent the data, highlighting different trends in these crimes on different days. For example, the most common day for Assault with a Deadly Weapon (7370 instances) is Sunday, whereas the most common day for Burglary (8568 cases) is Monday. Tuesdays saw the highest number of vehicle burglaries (7614 occurrences), while Fridays saw the highest number of petty thefts (6109 incidents). In contrast, wit12,31616 complaints, the frequency of vehicle theft is highest on Fridays. This research offers insightful information on the weekly fluctuations in these Part 1 crime incidences.

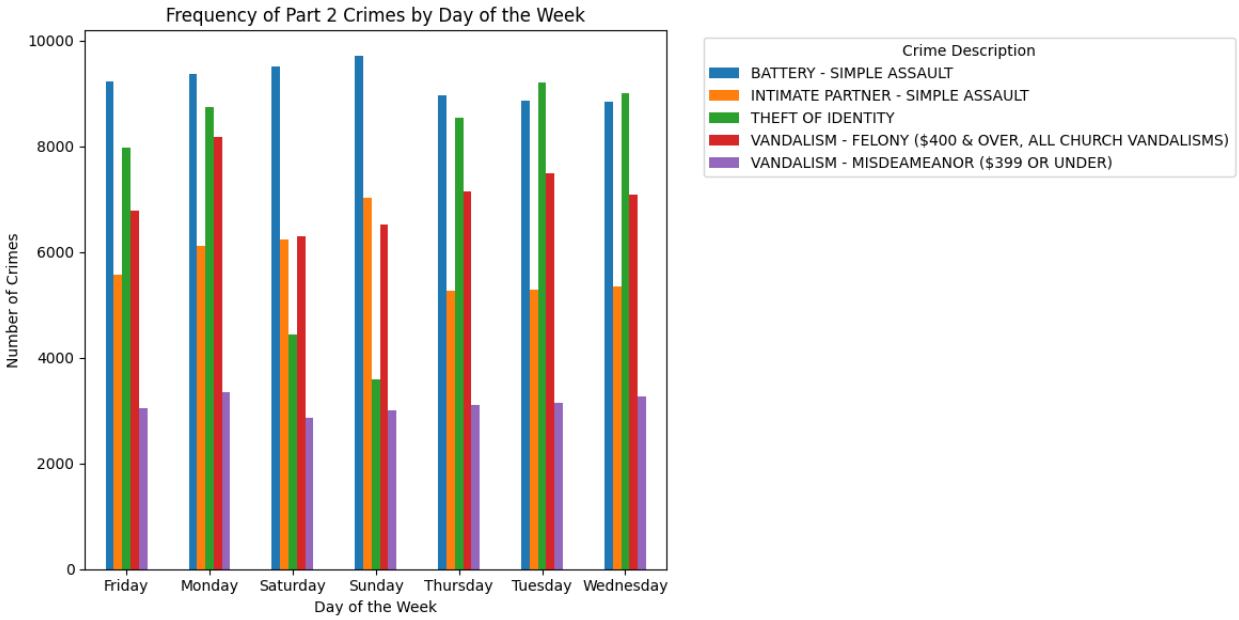


Fig4.6.3

The pattern in (fig4.6.3) is fairly similar for Part 2 offenses, with greater rates of crime throughout the week. There are fewer cases of "THEFT OF IDENTITY" and "BATTERY - SIMPLE ASSAULT" on Saturdays compared to weekdays. Throughout the week, "VANDALISM" exhibits its own distinct patterns in both Part Offenses and Part 2 offenses.

4.7 Impact of Major Events or Policy Changes

The major events and policy changes mentioned can have various effects on crime trends. Here's how these events and policy changes may influence crime patterns:

- COVID-19 Pandemic (2020 - Mid 2021): The pandemic had complex effects. Lockdowns and social distancing measures led to reduced economic activity and employment, potentially resulting in increased property crimes such as burglaries and thefts. At the same time, domestic violence incidents and online scams surged. However, some types of crimes, like those occurring in public spaces, decreased due to restrictions. We can clearly observe the crime trend declining in mid of 2020 and at the start of 2021.

- Vaccination Campaign (2021): The vaccine rollout allowed for a gradual return to normalcy, which could have affected crime patterns as businesses reopened and people resumed normal activities. We can clearly see the increase in crime trends after mid-2021.
- Infrastructure Investment and Jobs Act (Nov 2021): Infrastructure projects can lead to changes in local economies, affecting the job market and potentially influencing property crime rates in construction-heavy areas.
- (Late 2021 to Late 2022): The economic downturn marked by a declining stock market, widespread layoffs, increasing unemployment rates, and rising inflation has left its mark on crime statistics.
- Property crimes have shown an observable uptick, as individuals facing financial stress may resort to theft, burglary, and vandalism in an attempt to address economic hardships.
- Simultaneously, financial crimes, including fraud and scams, have surged, taking advantage of the heightened vulnerability of financially distressed individuals. As the labor market became constricted and layoffs occurred, white-collar crimes also saw an increase. The strain of economic instability has the potential to escalate into interpersonal conflicts, contributing to an elevated incidence of violent crimes, particularly domestic violence.
- Furthermore, prolonged economic uncertainty has occasionally culminated in civil unrest and protests, resulting in crimes associated with public disorder. Substance abuse issues, often exacerbated by economic stress, have led to drug-related crimes and public intoxication. The consequences of these economic challenges have prompted fluctuations in law enforcement budgets, which have been redistributed to address specific economic-related crimes.
- Mental health-related crimes have also been affected, with more disturbances and crises necessitating police intervention. It is essential to acknowledge the intricate relationship between economic conditions and crime, but these statistics

underscore the profound influence of economic difficulties on the crime landscape during this period.

4.8 Detecting Outliers and Unusual Patterns

In order to enhance the quality and reliability of the dataset, several critical data preprocessing steps were applied. First and foremost, a careful examination of the 'Victim Age' column revealed the presence of outliers and anomalies, including negative and extreme age values. To address this issue, a systematic approach was adopted. Rows with 'Victim Age' less than or equal to 0 were identified and processed. Mean imputation was performed for non-zero, non-negative age values, yielding a set of fixed age values around the calculated mean. This approach was employed to replace erroneous or missing age data effectively. Additionally, an outlier with an extraordinarily high age of 120 was removed from the dataset.

Subsequently, 'Vict Sex' and 'Vict Descent' columns underwent a transformation process. 'Vict Sex' values were replaced with more descriptive labels, such as 'Male,' 'Female,' and 'Unknown.' Null values in this column were filled with 'Unknown' to ensure consistency. Similarly, 'Vict Descent' values were mapped to meaningful categories, and missing data was also labeled as 'Unknown.' These measures were taken to enhance the interpretability of the dataset and maintain consistent treatment of unknown information. Overall, these data preprocessing steps were instrumental in refining the quality and utility of the dataset for subsequent analysis and insights.

To address the issue of missing values in the 'Weapon Description' column, I systematically filled these gaps with the label 'Unknown.' This approach ensures data completeness, maintains consistency, and allows for a comprehensive analysis of crime-related information. It also emphasizes transparency in acknowledging the absence of weapon details in the dataset.

4.9 Analyzing Demographic Factors

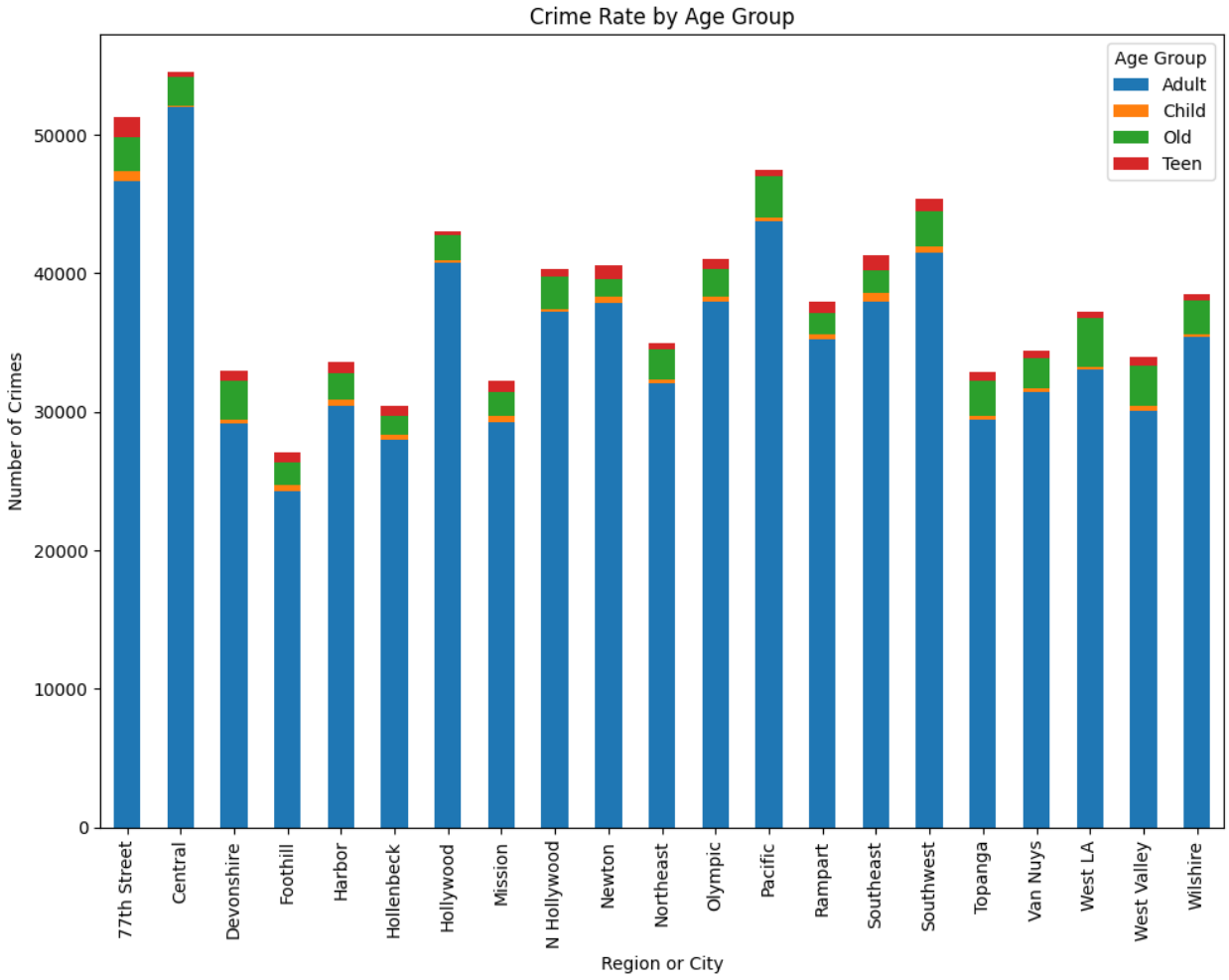


Fig4.9.1

The analysis delves into age group demographics, showing that the "Adult" age group consistently contributes the most to crime rates across all regions, comprising between 88% and 95% of all incidents. On the other hand, the smallest part of the crime rate is usually the "child" age group, with a percentage of less than 2%.

4.10 Predictive Analysis

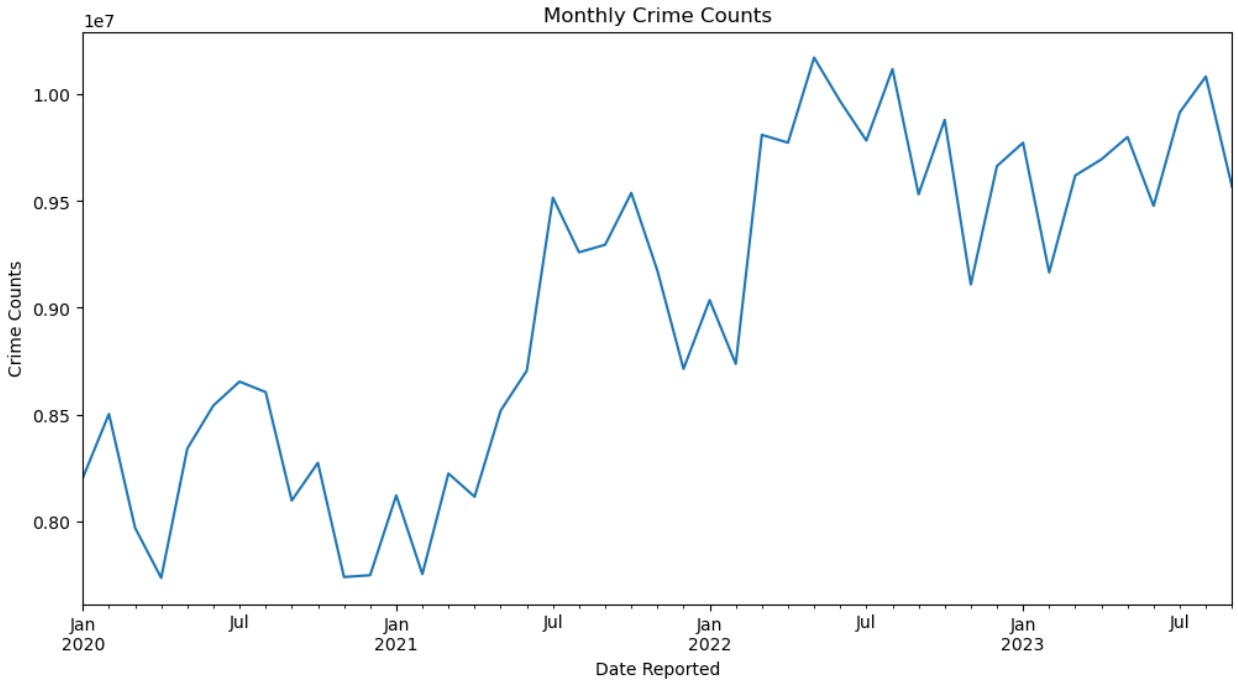


Fig4.10.1


Crime Code
1e7
1.0
0.8
1e6
Trend
9.5
9.0
8.5
Seasonal
0
−500000
Resid
250000
0
−250000
2020-05
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09

Fig4.10.2

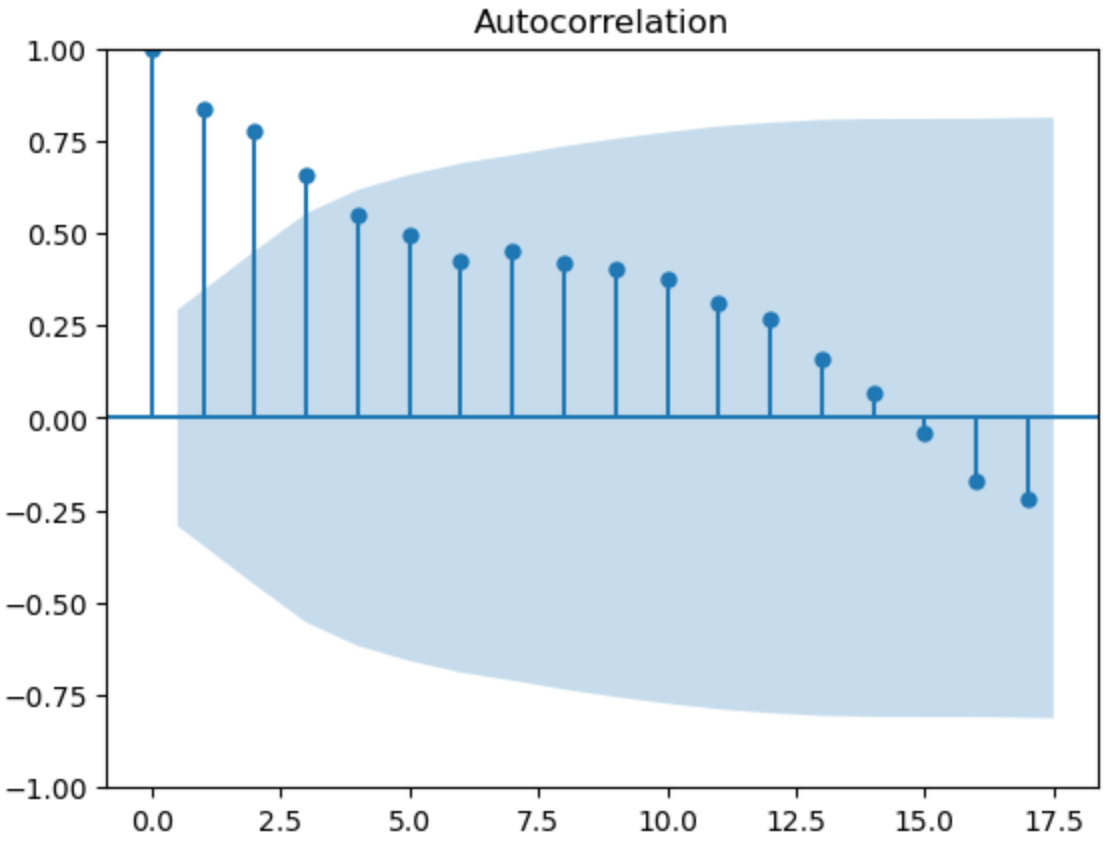

Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Fig4.10.3

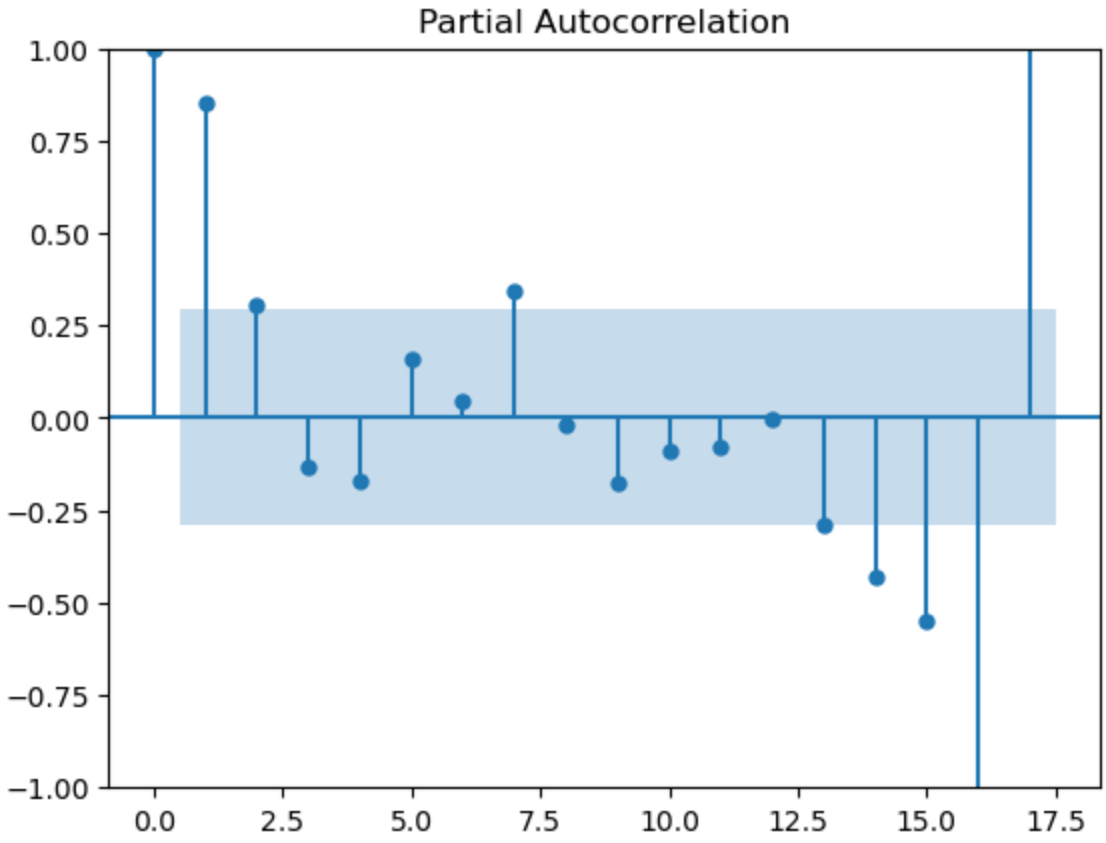


Fig4.10.4

Monthly crime count time series data are used in this research. The first step involves visualizing the time series and resampling the data to a monthly frequency. To identify patterns, seasonality, and residuals, the data is broken down. Applying differencing makes the data steady. Future patterns in criminal activity are predicted by fitting an ARIMA model. Calculating standard errors and confidence intervals helps evaluate how reliable the forecast is. Planning related to law enforcement and public safety can benefit from this analysis's foundation for forecasting and visualizing future crime trends.

5. Conclusion

5.1 Summary of Key Findings

- Between 2020 and 2023, overall crime patterns indicated a rise, with significant year-to-year variability.
- There are seasonal trends in crime, with summertime seeing increased crime rates.
- Among the most frequent crime categories were "Battery - Simple Assault" and "Vehicle - Stolen".
- There were clear regional disparities in crime rates, with "Central" and other places having greater rates.
- The study looked at the link between crime rates and economic variables and found that the unemployment rate had an inverse association with crime rates.
- An examination of reported crimes by day of the week showed that Sundays had the fewest crimes and Mondays had the most.
- Crime statistics were notably impacted by significant events and policy changes, including the COVID-19 pandemic, vaccine efforts, and economic swings.
- Data preparation techniques improved the quality of the dataset by addressing outliers and missing values.
- Demographic elements, such as age groups, were analyzed to identify their contributions to crime rates.

5.2 Implications and Recommendations

These findings have ramifications for scholars, legislators, and law enforcement. A useful tool for data-driven decision-making in resource allocation and crime prevention is data analytics. Targeted intervention tactics can be aided by identifying seasonal patterns, comprehending the most prevalent crime types, and addressing geographical variations. The research also emphasizes how important demographic data, significant events, and economic variables are in determining crime patterns. It is advised that law enforcement organizations keep spending

money on data analytics to support evidence-based decision-making, track variations in crime on different days, and take socioeconomic factors into account when creating anti-crime measures.

5.3 Limitations and Future Research Directions

The research recognizes its limitations, including the possibility of errors in the transcribed crime data and missing data in the dataset. Future lines of inquiry include:

- Investigating in further detail how significant occurrences and legislative changes affect crime rates
- Evaluating the efficacy of particular crime prevention and reduction techniques.
- Examining the connection between various economic variables and the prevalence of crime
- Identifying micro-level patterns by doing more detailed investigations at the neighborhood level.
- Adding more cities or regions to the dataset will give a more comprehensive view.
- The basis for improving crime analysis and creating well-informed plans to advance public safety and law enforcement effectiveness is provided by this initiative.