Cleaning and Analyzing Crime Data

Foundations of Data Analytics - 6400

Project Report

Group - 2

Akshat Jain

Vamsi Krishna Jilla

Manoghn Kandiraju

Dhiksha Mathanagopal

Ananya Mahesh Shetty

1 INTRODUCTION AND RESEARCH QUESTION

Crime is a pervasive societal issue that has far-reaching consequences for public safety, community well-being, and law enforcement efforts. Analyzing crime data provides valuable insights into the trends, patterns, and factors that influence crime rates. This project aims to harness the power of data analysis to better understand and address these issues by examining a real-world dataset containing crime data from 2020 to the present.

In recent years, advancements in data collection and technology have made it possible to access and analyze vast amounts of crime-related information. The analysis of this data can help law enforcement agencies, policymakers, and communities make more informed decisions to enhance public safety and reduce crime rates.

"What are the key crime trends and patterns in the dataset, and what factors are contributing to variations in crime rates over time and across different locations?"

In this extensive data analysis project, our goal is to address a range of critical questions related to crime data spanning from 2020 to the present. We'll start by investigating temporal trends, with a focus on uncovering recurring patterns, seasonality, and long-term shifts in crime rates over time. We'll then turn our attention to spatial patterns, aiming to identify crime hotspots, analyze regional variations, and identify neighborhoods with consistently high or low crime rates. Further exploration will involve dissecting crime types to identify the most prevalent categories and track how they've changed over time. We'll also delve into demographic factors like age and gender to understand their connection with specific crimes and regional crime rates. Economic, environmental, and urban factors will be assessed for potential correlations with crime rate fluctuations. Additionally, we'll analyze the influence of policy changes and law enforcement strategies on crime rates. For those interested in advanced analysis, we'll explore the option of predictive modeling to forecast future crime rates using historical data and identified factors. Our objective is to provide a comprehensive understanding of crime dynamics and their determinants, aiding informed decision-making, enhancing public safety, and guiding policy recommendations.

2 SUMMARY OF THE RESULTS

The results presented in this analysis encompass several key stages of data processing, exploration, and visualization. The initial phase involved data acquisition, where a file was downloaded from Google Drive, saved as 'crime_data.csv,' and loaded into a Pandas DataFrame. The dataset's dimensions, in terms of the number of rows and columns, were calculated and displayed, providing essential information about the data structure.

Subsequently, data inspection and cleaning were conducted to enhance dataset usability. This phase involved displaying initial rows, capturing data types, and summarizing the dataset. Renaming of columns for clarity and the creation of a well-documented dataset were achieved. Furthermore, data cleaning tasks included handling missing values, converting date formats, and eliminating duplicates, resulting in a refined and informative dataset suitable for in-depth analysis.

The analysis also delved into the temporal aspects of the crime dataset, providing insights into crime trends over the years and monthly variations. Visualizations, such as line graphs and bar plots, were used to illustrate these trends. Notable events and their impact on crime trends, such as the COVID-19 lockdown, were highlighted using annotations. Additionally, seasonal crime patterns were explored, offering a clear representation of how crime occurrences change throughout the year.

The analysis extended to examining crime data by geographic area, revealing comparative crime rates across various regions. The top three crime types for each area were identified and presented in a bar plot, enhancing the understanding of prevalent crime types in different locations. The most common crime type in the dataset was also determined, followed by a temporal analysis to unveil its long-term trends. Moreover, crime distribution by the day of the week and the distribution of top 10 crime types across different days were visualized, highlighting the influence of weekdays on crime rates.

Furthermore, the analysis explored the relationships between specific crime types and demographic factors, such as gender, age, and descent of victims. Visualizations, including bar plots and box plots, shed light on these patterns and correlations, offering insights into the impact of these factors on crime occurrences.

A critical aspect of the analysis involved correlating crime data with economic indicators. The integration of economic data and subsequent correlation analysis allowed for a better understanding of potential relationships between economic factors and crime rates. A correlation matrix and heatmap were generated to visualize these associations.

In the advanced analysis section, the analysis presented sophisticated techniques such as time series forecasting with SARIMAX to predict monthly crime counts and crime location mapping to visualize the spatial distribution of crimes in Los Angeles. K-Means clustering was used to group crime locations based on crime counts and coordinates, with the Elbow Method and PCA for visualization. These advanced analyses provided deeper insights into crime patterns, predictions, and geographical aspects.

In summary, this comprehensive analysis covered data acquisition, cleaning, and exploration, revealing temporal, spatial, and demographic patterns within the crime dataset and offering insights into the potential influence of economic factors on crime rates. Advanced techniques were also employed to provide more sophisticated insights into crime trends and locations

3 RESULTS AND METHODS

3.1 Data Acquisition

Downloads a file from Google Drive with the given file ID, saves it as 'crime_data.csv', and loads it into a Pandas DataFrame. It then calculates and prints the number of rows and columns in the DataFrame. The output will be in the form of: "No of rows are <number_of_rows> and No of columns are <number_of_columns>."

3.2 Data Inspection

Conducts data inspection and column renaming for a dataset. It begins by displaying the initial rows of the dataset to provide a quick overview of the data. It then captures the data types of each column and prints a concise summary of the dataset, which includes the number of non-null entries and their data types. A dictionary is utilized to associate meaningful descriptions with each column, aiding in understanding their significance. The code subsequently renames specific columns according to a predefined mapping for clarity. The final output is a well-documented dataset with renamed columns, which enhances its usability and comprehensibility, making it more accessible for subsequent data analysis and exploration.

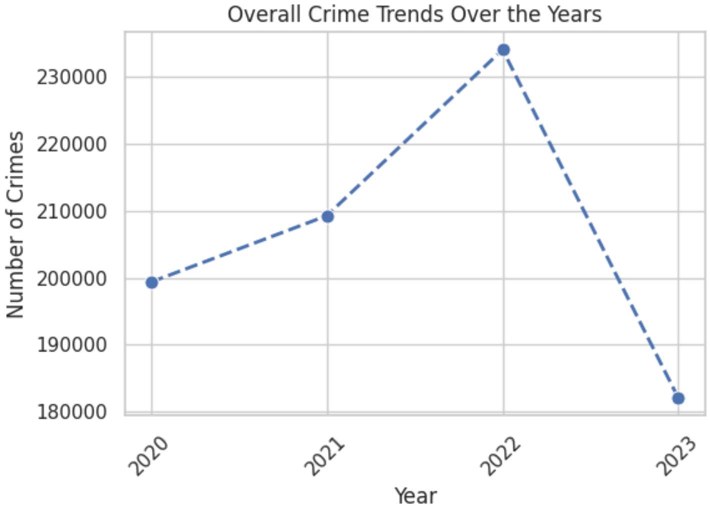
3.3 Data Cleaning & EDA

Conducts a series of data cleaning and exploratory data analysis (EDA) tasks on the dataset. It initially explores the dataset by counting the occurrences of unique values in specific columns, such as geographic areas and victim genders, while also identifying missing data. Column conversion and cleaning operations are performed, including filling missing values in the 'Crm_Cd1' column with zeros and converting it to integers. The 'DATE_OCC' column is converted to a datetime format for better date handling. Duplicates based on the 'DR_NO' column are removed to ensure data integrity. Further examination of missing data is conducted after these cleaning steps. Finally, a box plot is generated to visualize the distribution of victim ages, with quartiles and potential outliers highlighted. The end result is a more refined and informative dataset, facilitating subsequent data analysis and interpretation, along with valuable insights into crime areas and victim ages.

3.3.1 OVERALL CRIME PER YEAR

The first part of the code visualizes the overall crime trends per year. It calculates the number of crimes for each year using the 'DATE_OCC' column, sorts them chronologically, and then plots a line graph using Matplotlib and Seaborn. The graph displays the evolution of crime rates, with markers for individual years, allowing for a clear understanding of any long-term trends.

.

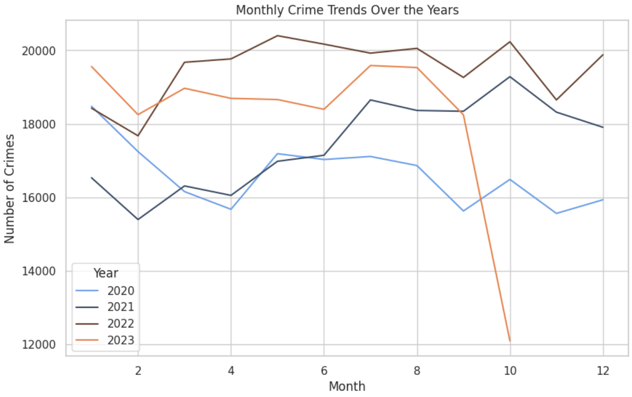


- From the Yearly Crime Trends, we can see that although there has been a steady rise in crime from 2020 to 2022, with a peak in 2022, there has been a sharp decline in the number of crimes in 2023

3.3.2 MONTHLY CRIME OVER YEARS

The second part of the code explores monthly crime trends across multiple years. It first converts the 'DATE_OCC' column to a datetime format and then extracts the year and month information. It aggregates the data to count monthly crimes, creating a new DataFrame. The code proceeds to generate a line plot with Seaborn, showing how monthly crime counts have varied over the years. Different colors represent different years, offering insights into any seasonality or recurring patterns in crime occurrences.

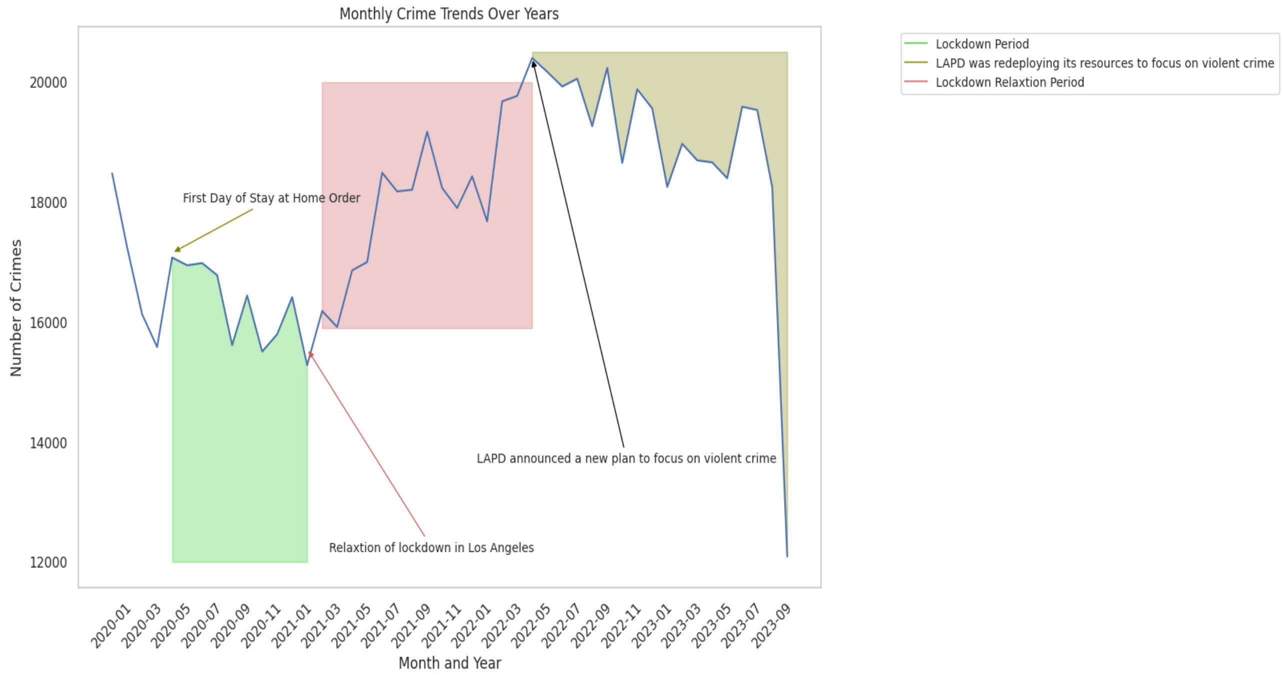
Both sections utilize data visualization techniques to provide valuable insights into the dataset's temporal crime trends, from yearly variations to monthly patterns.



- From the monthly crime trends over the years, we can see that the number of crimes in each month of 2023 have been higher than 2020 and 2021 but are often less than 2022.

3.3.3 INVESTIGATING ANY IMPACT BECAUSE OF GLOBAL EVENTS

The code first converts the 'DATE_OCC' column to a datetime format, allowing for date-based operations. It then calculates monthly crime counts by grouping the data based on the year and month of the 'DATE_OCC' column, resulting in a count of crimes per month over the years. Matplotlib is employed to generate a line plot illustrating these monthly crime counts, with the x-axis representing the month and year and the y-axis showing the number of crimes. Custom x-axis labels are added, displaying the month and year for every other data point while rotating for improved readability. The code also utilizes annotations with arrows and labels to emphasize notable events, such as the commencement of the stay-at-home order in May 2020 and the relaxation of the lockdown in February 2021, providing contextual information. The resulting plot displays the monthly crime trends over the years, with clear variations in the number of crimes over time.



- From the results, we can clearly see that the 1st day of lockdown saw decline in crime rates, with occasional spikes.
- After the relaxation of the lockdown, there is also a clear rise in the number of crimes.
- The number of crimes sees a decline after the LAPD announced a new plan to focus on violent crime.

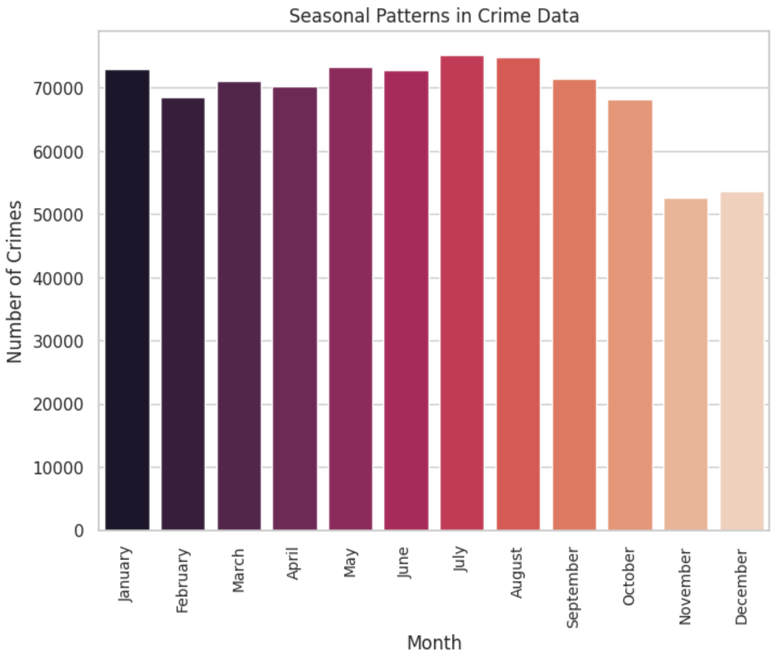
3.3.4 SEASONAL CRIME DATA

The code first creates a dictionary called month_names to map numeric month values (1 to 12) to their corresponding month names. It then adds a new 'Month' column to the dataset by extracting the month component from the 'DATE_OCC' column and mapping it to the month names using the created dictionary.

Next, the code groups the data by the 'Month' column and counts the number of crimes reported in each month. To ensure the bar plot displays the months in chronological order, the results are reindexed based on the predefined month names.

A bar plot is created using Seaborn's barplot function, where the x-axis represents the months, the y-axis shows the number of crimes, and the color palette is set to 'rocket' for visual appeal. The resulting plot provides a clear visual representation of seasonal patterns in crime data, allowing viewers to observe trends in crime occurrences throughout the year, all while maintaining readability and interpretability.

Seasonal patterns show that the November and December period sees a decline in the number of crimes

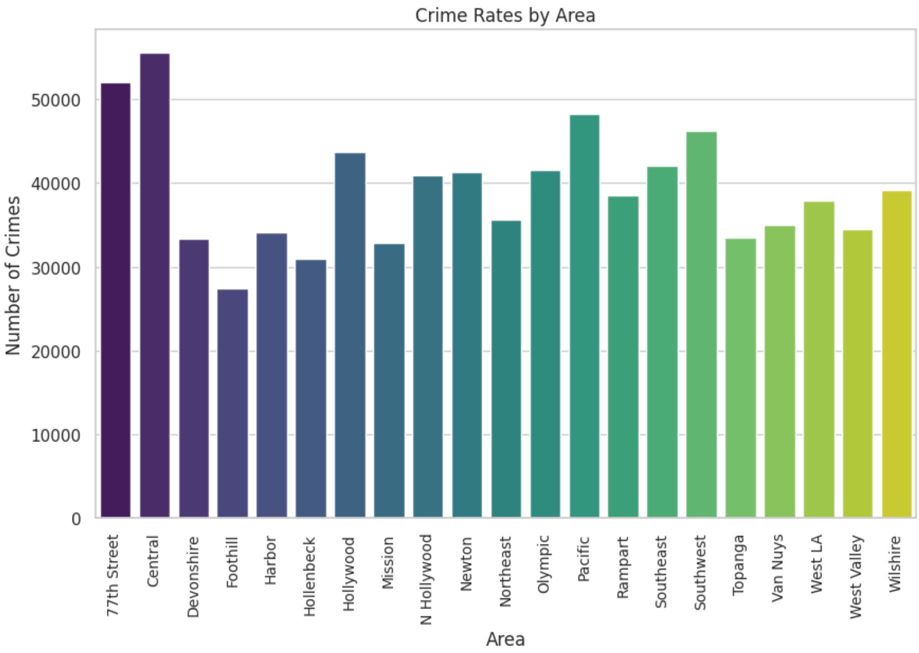


- Crime usually peaks in July-August period and drops to lowest in November-December period over this period of 3 years.

3.3.5 CRIME BY AREA

The plot effectively presents a comparative view of crime rates across various areas, making it easy to identify regions with higher or lower crime activity. Generates a bar plot that identifies the top three crime code descriptions for each geographic area. It first groups the data by 'AREA_NAME' and 'Crm_Cd_Desc', counting the number of crimes in each category. Then, it sorts the results by area and count in descending order. The final plot displays the top three crime code descriptions for each area, with bars representing the counts, allowing for a comparative view of the most prevalent crime types in different areas. The 'viridis' color palette is applied, enhancing readability and distinguishing between areas.

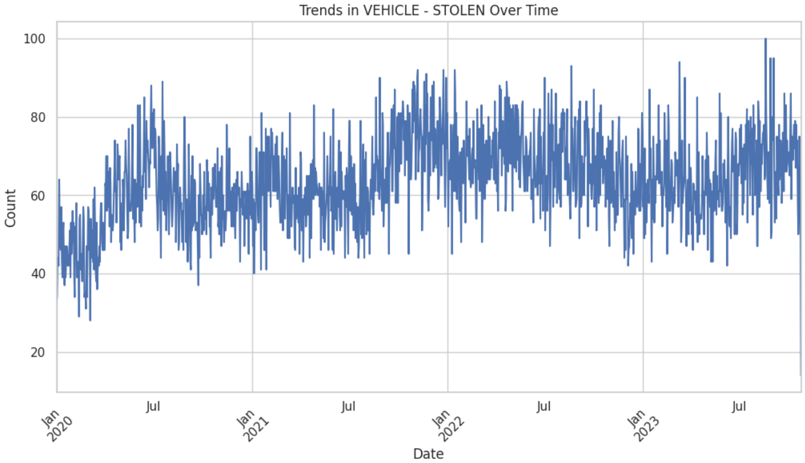
- We can clearly see that Central and 77th Street are the 2 areas with the highest frequency of crimes, while Foothill is the area with the lowest crime rate.

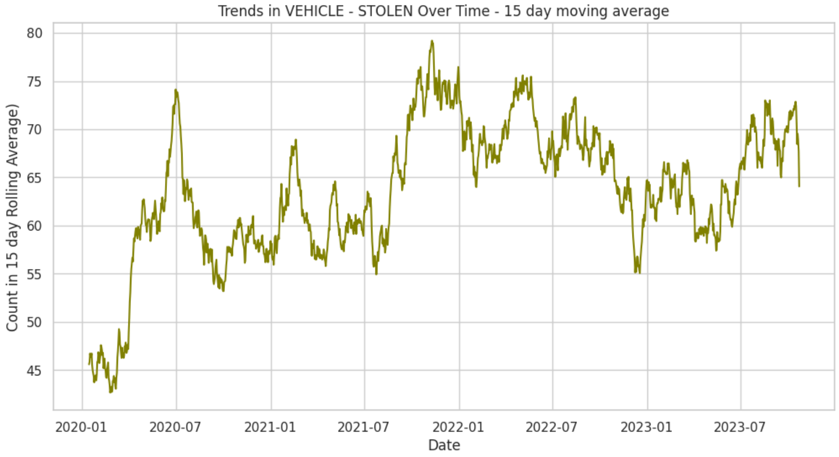


3.3.6 MOST COMMON CRIME

This code calculates and prints the most common crime type in the dataset. It first counts the occurrences of each crime type using the 'Crm_Cd_Desc' column and then identifies the crime type with the highest occurrence. The result is printed as follows: "The most common crime type is 'crime_type' with 'occurrence_count' occurrences," providing valuable information about the most prevalent crime in the dataset. The code initially identifies the most common crime type in the dataset and reports its prevalence. It then delves into the temporal analysis of this primary crime type, generating two plots. The first plot illustrates the count of this crime type over time, providing insights into its trends and fluctuations. The second plot employs a 15-day rolling average to smooth out short-term variations, offering a clearer view of broader trends in the crime's occurrences. This combined analysis yields valuable insights into the most prevalent crime type's temporal patterns and helps in identifying long-term trends while minimizing short-term noise.

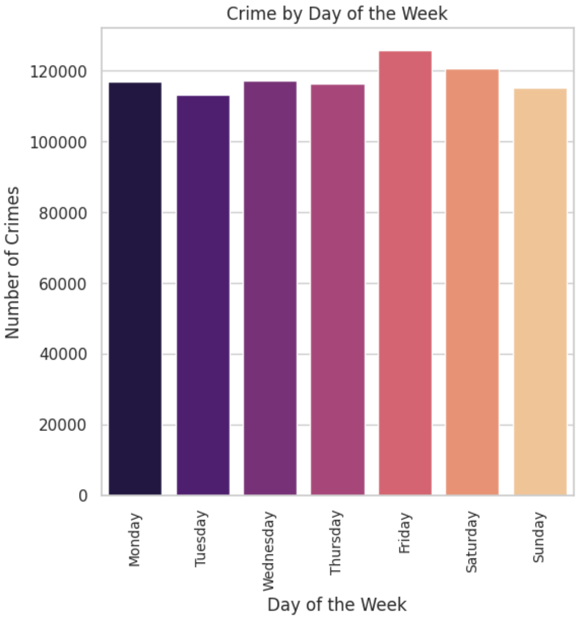
- Most common crime was found to be Stolen Vehicles.
- The graphs demonstrate the trends of the Stolen Vehicle crime.





3.3.7 CRIME BY DAY OF THE WEEK

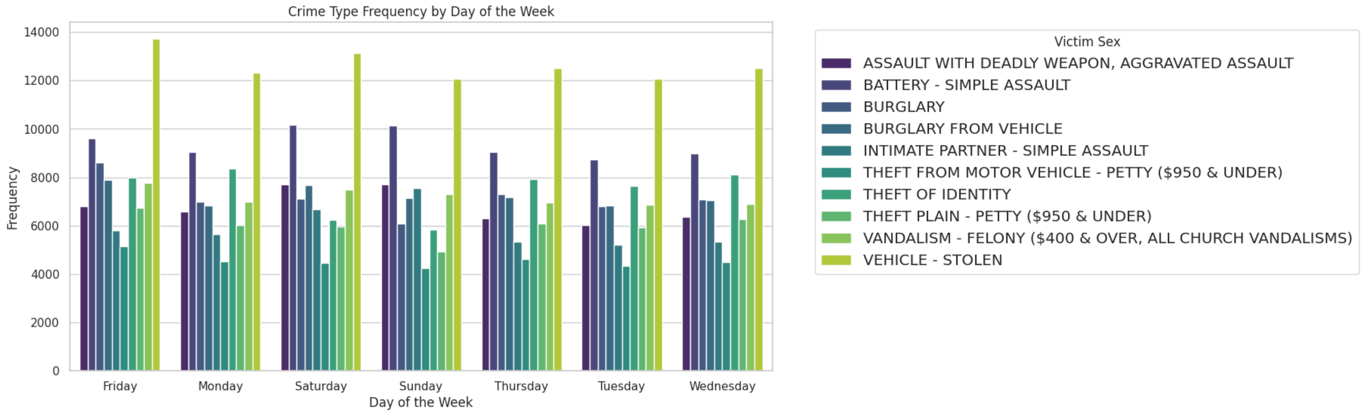
The code converts date data to the day of the week, then visualizes the distribution of crimes, demonstrating that weekends (Saturday and Sunday) exhibit higher crime rates, using a bar plot with the 'magma' color palette.



- The graph clearly shows that Friday has the highest number of crime counts.

3.3.8 CRIME COUNT DISTRIBUTION FOR TOP 10 CRIMES ON EACH DAY OF THE WEEK

The code first converts date data to the day of the week and then focuses on the top 10 crime types by counting their occurrences. It creates a bar plot using Seaborn, where each bar represents the frequency of a specific crime type on different days of the week. The color palette 'viridis' is applied for distinction. The result illustrates the distribution of the top 10 crime types across the days of the week, offering insights into when these crimes are more prevalent. It provides a detailed view of crime patterns by day and crime type, helping to identify any potential correlations or trends in their occurrence based on the day of the week.

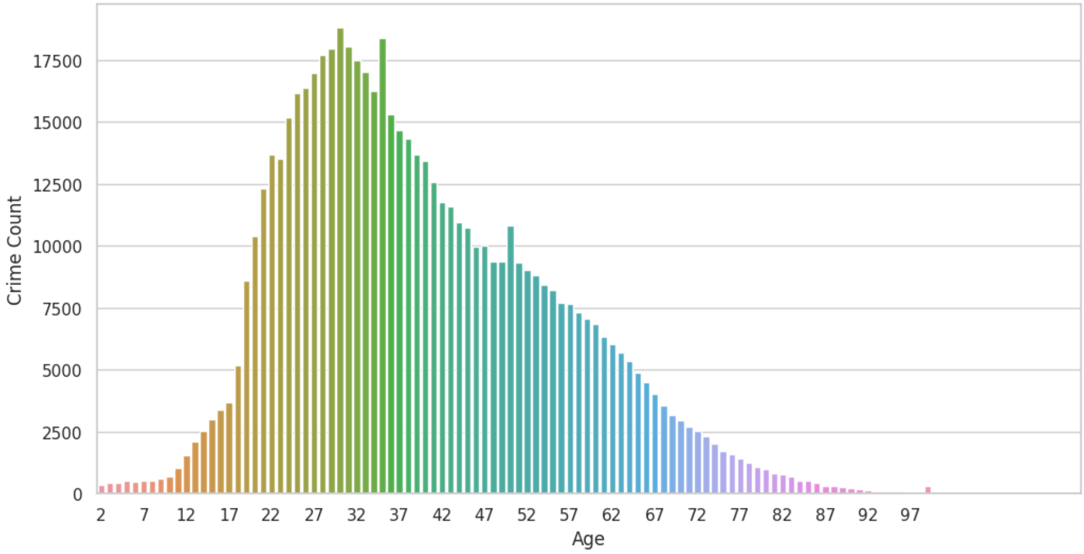


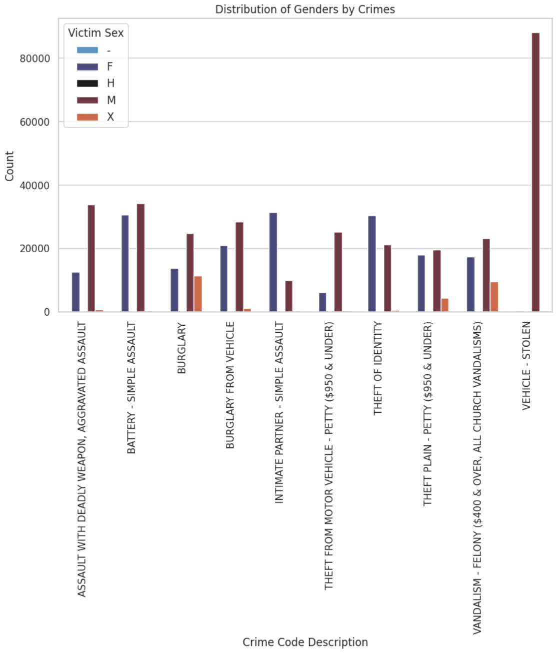
- Vehicle-Stolen crime occupies the top spot on all days of the week Battery simple assault being the second highest on each day of the week.
- The order of crime counts doesn’t change drastically on any day of the week.
- Distribution of crimes on each day is more or less the same with minute differences.

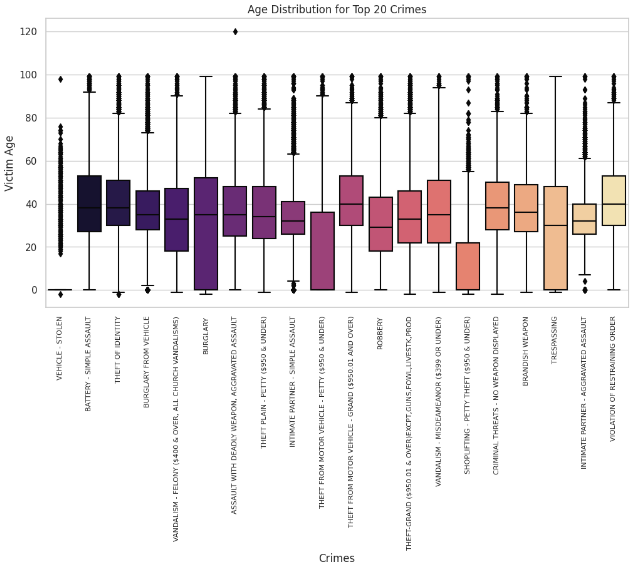
3.3.9 Analyzing to identify any patterns or correlations between demographic factors specific types of crimes

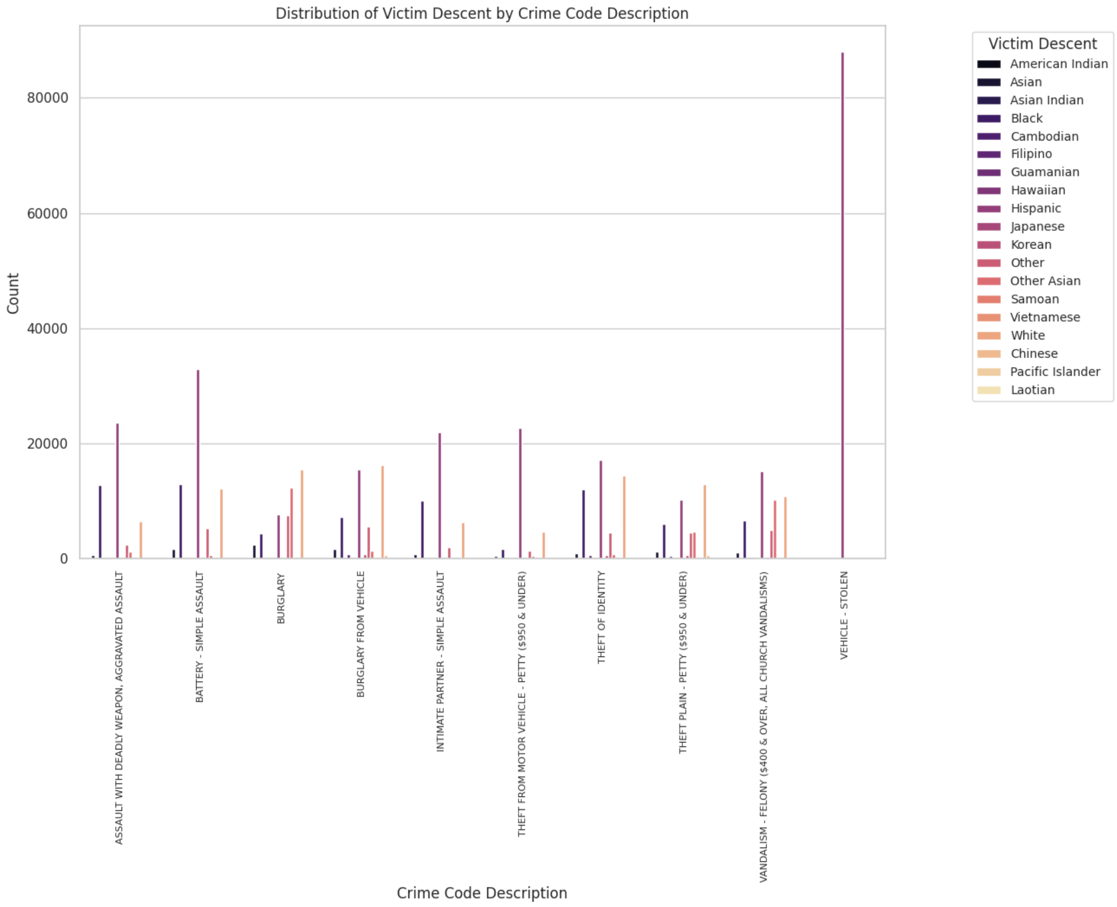
Here we analyzed by performing the distribution of Victim Sex for top 10 crimes, Age Distribution for top 20 Crimes, Crime Count distribution for all ages, Distribution of Victim Descent for top 10 crimes:

The code performs an extensive analysis of the crime dataset, examining various dimensions and relationships within the data. It starts by investigating the distribution of gender among crime victims for the top 10 crime types, using a bar plot to reveal gender-related patterns in crime occurrences. Next, it explores the age distribution of victims for the top 20 most common crimes through a box plot, providing insights into age-related trends and variations. The code also counts the number of crimes by victim age and presents the results as a bar plot, offering a clear overview of the age groups most affected by crimes. Lastly, it investigates the distribution of victim descent by crime type for the top 10 crime descriptions, using a bar plot to identify potential descent-related patterns in crime occurrences. These analyses collectively provide a comprehensive understanding of crime data and reveal valuable insights into the relationships between crime types, victim demographics, and victim characteristics.







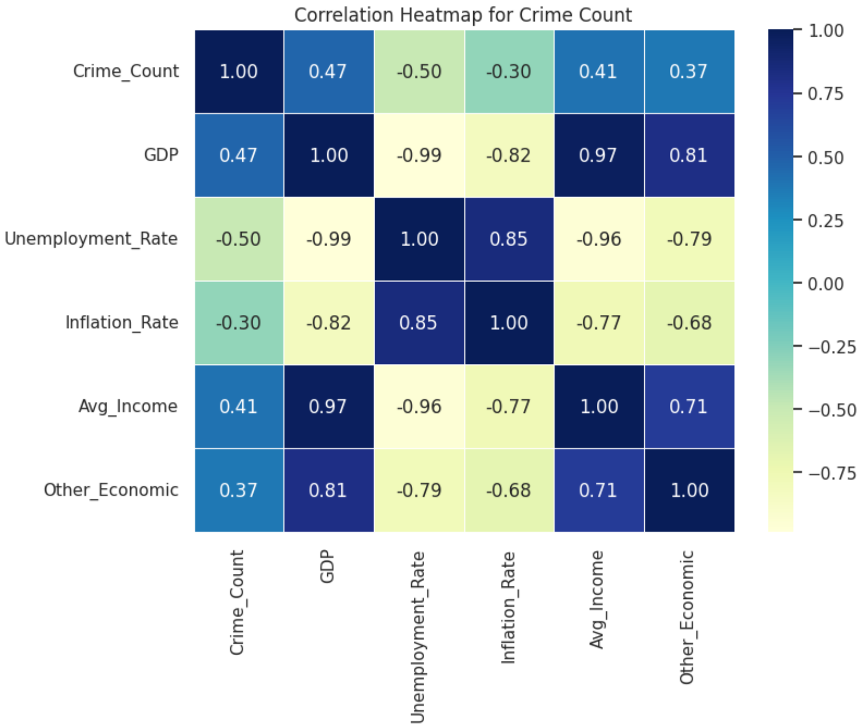


- In the Vehicle-Stolen crime majority of the victims are Men.
- Most of the Vehicle-Stolen Victims belong to one descent and this crime has relatively higher occurrences than any other crimes.

- The analysis reveals that Males of ages 20-35 are most likely to be victims of crimes.
- Age distribution for Shop-lifting and theft from motor vehicles is similar and mostly under 40 which is relatively young compared to other crimes.
- Gender distribution is abnormal in Vehicle-Stolen, Intimate Partner Assault crimes where first one is female dominated and the latter is male dominated.

3.3.10 Correlation with Economic Factors

Data Preparation: The code creates a DataFrame for the economic data, converts the date column to datetime, extracts the year and month, and then merges the economic data with the crime data based on the 'year_month' column. Correlation Analysis: The merged dataset is used to compute a correlation matrix, which measures the degree of linear relationship between economic indicators and crime counts.A heatmap is generated using Seaborn to visualize the correlation matrix, with annotations displaying correlation coefficients. This analysis provides insights into the potential relationships between economic factors and crime rates over time, offering a better understanding of how changes in economic indicators might be associated with variations in crime counts.

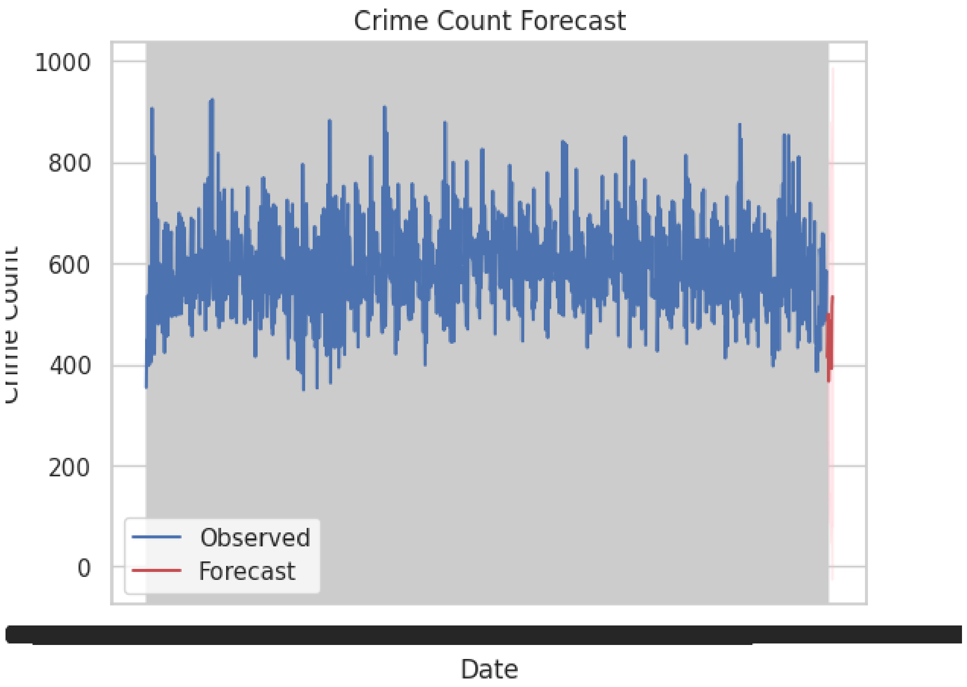


- From the correlation matrix, we can see that Number of Crimes is most correlated to the Unemployment Rate.
- Inflation Rate comes second in terms of correlation .
- Observe here negative correlation ideal for our context.

3.4 Advanced Analysis

3.4.1 Time Series Forecasting with SARIMAX

The code utilizes the Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) model to forecast monthly crime counts. It fits the model to the observed data and generates a forecast for future periods. The results are visualized with observed and forecasted crime counts, accompanied by confidence intervals.



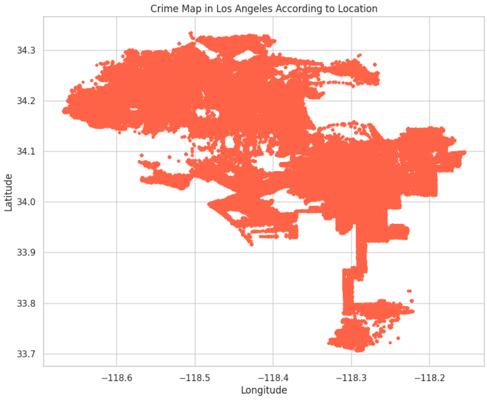
It preprocesses historical crime count data, forecasts future crime counts, and visualizes the results. The SARIMAX model, with order (1, 1, 0) and seasonal order (1, 1, 0, 12), is used to capture both non-seasonal and seasonal patterns. The line plot shows observed historical crime counts, while the red line represents the forecasted counts, with a pink shaded area denoting the confidence interval.

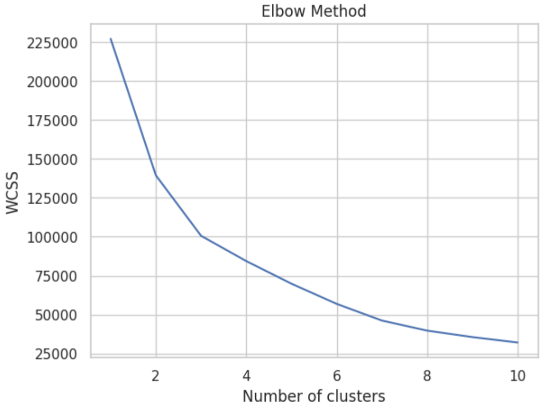
- The line plot visually compares the observed historical crime counts (in blue) with the forecasted values (in red).
- The pink shaded area represents the uncertainty or confidence interval around the forecasted values.
- The forecasted crime counts for the next 12 months are presented in the printed output.

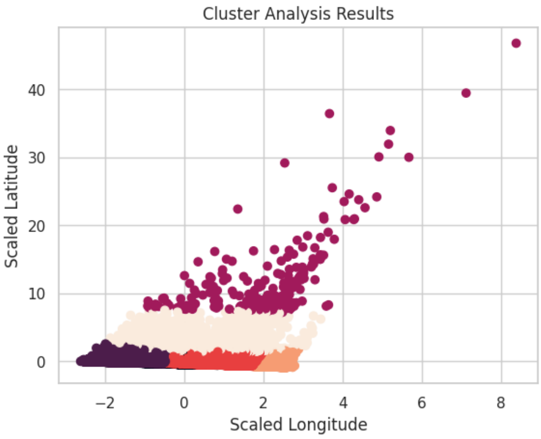
3.4.2 Crime Location Mapping

The code creates a scatter plot to map crime incidents in Los Angeles based on their geographical coordinates (latitude and longitude). This visualization provides insights into the distribution of crimes across the city.

K-Means Clustering for Crime Locations: The code uses K-Means clustering to group crime locations based on crime counts and geographical coordinates. The Elbow Method is applied to determine the optimal number of clusters. The Within-Cluster-Sum-of-Squares (WCSS) is plotted to find an elbow point, indicating the most appropriate cluster count. K-Means clustering is then executed with the chosen number of clusters. Principal Component Analysis (PCA) is applied to visualize the clustered data in a two-dimensional space.







- According to the elbow method, no. of clusters are chosen to be 6 and clustering analysis is performed to see six clusters. Observe that x and y coordinates are scaled not original.

4 Data Sources

Our main goal in this study is to examine an actual dataset that includes crime statistics from 2020 to the present. The main objective is to clean and prepare this dataset for detailed examination, carry out exploratory data analysis (EDA), and answer certain queries about patterns, trends, and factors affecting crime rates. This dataset includes crime episodes that have occurred in the City of Los Angeles since 2020.The basis for comprehending and analyzing crime trends, patterns, and variables impacting crime rates from 2020 to the present is provided by this dataset. data is essential for this project because data comes from legitimate law enforcement organizations, guaranteeing the information's dependability and accuracy. The dataset is expected to contain a wealth of information, including the type and location of crimes, dates and times of incidents, descriptions, and potentially demographic details about suspects and victims.

This dataset provides a valuable resource for examining and addressing criminal activities within the specified time frame. In addition to the core crime dataset, supplementary data sources may also be considered to enhance the analysis. These sources could include demographic data, economic indicators, and local events calendars.Demographic data can shed light on how crime rates correlate with population characteristics, such as age, gender, and socioeconomic factors. Economic indicators may help explore the impact of economic conditions on crime trends. Furthermore, local event data could be useful in understanding how significant events or gatherings influence crime rates in specific areas. Integrating these additional data sources would provide a holistic perspective on the factors contributing to crime patterns and trends, enabling a more thorough and insightful analysis of the topic.

Overall, the data sources for this analysis are diverse and multifaceted, with the primary crime dataset forming the core foundation. By combining official crime data with supplementary information, we aim to gain a comprehensive understanding of crime trends, patterns, and their underlying drivers, facilitating informed decision-making and policy recommendations for addressing and mitigating criminal activities in the studied period.

5 Limitations and Future Work

Despite the extensive analysis conducted in this study, there are several limitations to consider. First, the quality and accuracy of the crime dataset depend on the data sources and reporting procedures. Inaccuracies or underreporting of crimes may introduce biases into the analysis. It is crucial to acknowledge that the dataset represents reported crimes and may not fully capture the entirety of criminal activities.

Another limitation pertains to the temporal analysis of crime trends. While the analysis highlights long-term and monthly patterns, it does not account for the dynamic nature of crime. The influence of external factors on crime, such as changes in law enforcement practices, socio-economic conditions, or societal events, is not deeply explored. Future work could involve integrating additional contextual data to provide a more comprehensive understanding of these influences.To address these limitations and further enhance the analysis, there are several avenues for future research. Firstly, incorporating external datasets, such as socio-economic indicators, weather conditions, or demographic shifts, can provide a more holistic perspective on the factors influencing crime patterns. This expanded dataset could enable more robust predictive modeling and deeper insights into the drivers of crime.

Additionally, conducting spatial analysis at a more granular level, such as neighborhood or precinct-based analysis, can offer a more detailed view of geographic crime disparities. This could help law enforcement agencies allocate resources more effectively and implement targeted crime prevention strategies. The analysis could benefit from the implementation of more advanced machine learning and predictive modeling techniques, such as deep learning or time series forecasting with recurrent neural networks. These methods could potentially provide more accurate crime predictions and uncover hidden patterns within the data.

Moreover, investigating the impact of policy changes, legislative reforms, or socio-economic interventions on crime rates could be an essential aspect of future research. Understanding how changes in these factors affect crime trends can aid in evidence-based policy-making and crime prevention efforts.In summary, future work should focus on integrating additional data sources, conducting more detailed spatial analysis, employing advanced modeling techniques, and exploring the impact of external factors to enhance the depth and accuracy of crime analysis, ultimately contributing to more effective crime prevention and law enforcement strategies.