Crime Data Analysis

IE6400 - Foundations of Data Analytics Engineering

Project Report

Group: 1

Kruthika Srinivas Vasisht (002798505)

Ruthvika Reddy Tangirala (002293262)

Sabarish Subramaniam A V (002243373)

Sarvesh Selvam (002874621)

Sneha Manjunath Chakrabhavi (002836841)

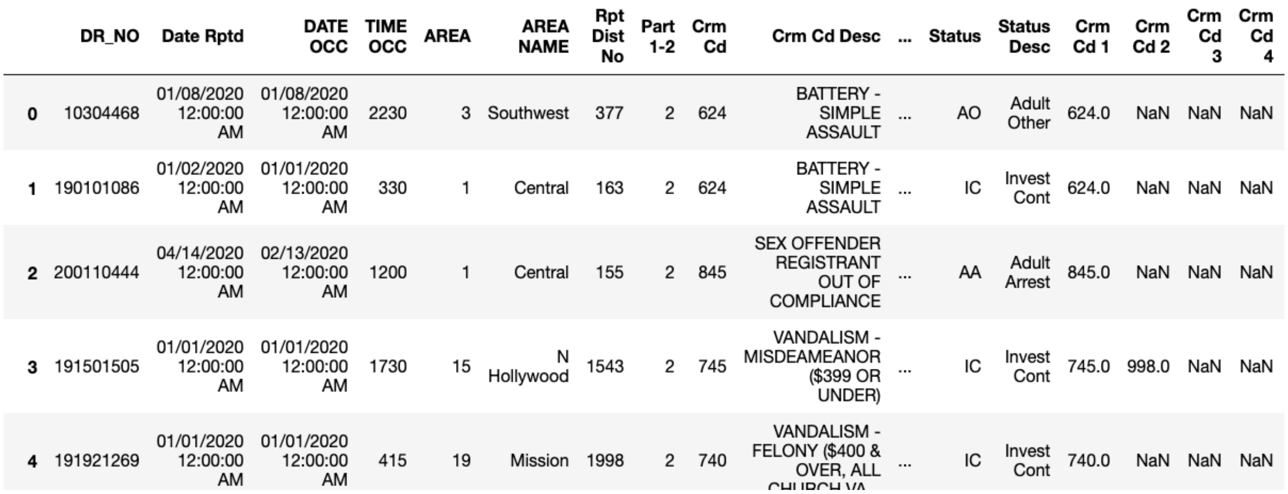
INTRODUCTION

Maintaining public safety and creating efficient law enforcement methods require an understanding of crime statistics. In this project report, we use Jupyter Notebook and Python to explore the complex field of crime data analysis. Our main source of information is the official U.S. government data archive, which offers thorough crime data from 2020 to the present. We employ time series forecasting techniques to predict future trends and find seasonal patterns in crime rates as well as variances across different geographic locations.

This analysis is important because it can provide information about the dynamics of crime in the US to law enforcement, legislators, and the general public. We aim to shed light on the variables driving crime rates and pinpoint locations with distinctive crime trends by examining the temporal patterns and geographic distribution of the data. In addition, by employing time series forecasting, we will be able to offer predictions about future crime trends, which will facilitate the creation of proactive policies to successfully combat crime.

We have manually cleansed and prepared the dataset from the official U.S. government data repository in order to do this study. We have developed a reliable and repeatable procedure for data analysis using Python and Jupyter Notebook. The project's results will be a vital resource for those seeking to comprehend and tackle the dynamic problems facing law enforcement and crime prevention in the US

DATA SOURCE



	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4
0	10304468	01/08/2020 12:00:00 AM	01/08/2020 12:00:00 AM	2230	3	Southwest	377	2	624	BATTERY - SIMPLE ASSAULT	...	AO	Adult Other	624.0	NaN	NaN	NaN
1	190101086	01/02/2020 12:00:00 AM	01/01/2020 12:00:00 AM	330	1	Central	163	2	624	BATTERY - SIMPLE ASSAULT	...	IC	Invest Cont	624.0	NaN	NaN	NaN
2	200110444	04/14/2020 12:00:00 AM	02/13/2020 12:00:00 AM	1200	1	Central	155	2	845	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	...	AA	Adult Arrest	845.0	NaN	NaN	NaN
3	191501505	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	1730	15	N Hollywood	1543	2	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	...	IC	Invest Cont	745.0	998.0	NaN	NaN
4	191921269	01/01/2020 12:00:00 AM	01/01/2020 12:00:00 AM	415	19	Mission	1998	2	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA	...	IC	Invest Cont	740.0	NaN	NaN	NaN

The crime dataset used in this analysis was obtained from the US government's official data repository, Data.gov. The dataset spans from the year 2020 to the present, encompassing a comprehensive record of criminal activities across various regions. To ensure data quality and integrity, the initial phase of the project involved meticulous data preparation procedures. The crime dataset preparation involved:

- checking and converting data types,
- checking for and removing null values,
- dropping unnecessary columns
- feature extraction.

The dataset presents an extensive record of crime incidents in Los Angeles since 2020, offering valuable insights into patterns of criminal activity. With various columns capturing crucial details such as dates, geographical information, crime codes, victim characteristics, and weapons used, the dataset facilitates comprehensive analysis for law enforcement, policymakers, and researchers.

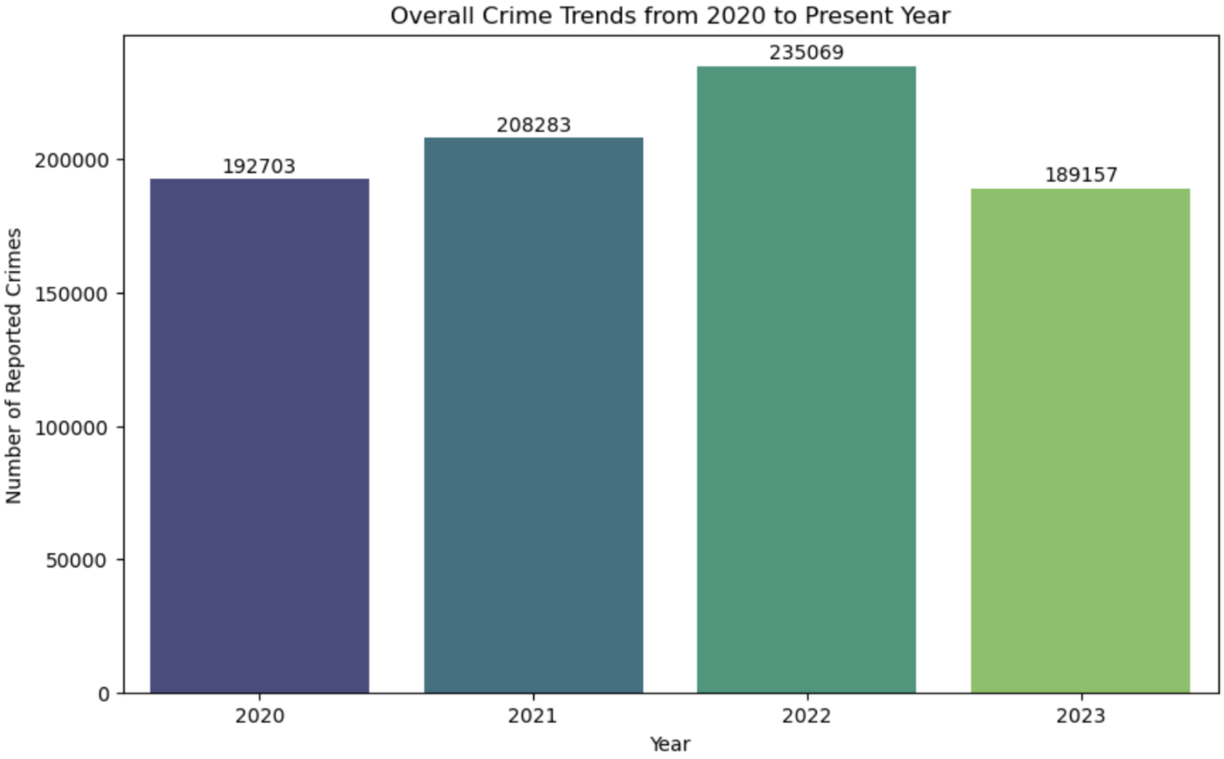
Key dataset attributes include the Division of Records Number (DR_NO) for unique identification, crime type (Crm Cd), victim details (Vict Age, Vict Sex, Vict Descent), and incident location information (LOCATION, LAT, LON). Additionally, the dataset includes codes and descriptions for the type of premises and weapon used, enabling in-depth contextual analysis. By undertaking these crucial steps, we aimed to streamline the dataset for effective analysis and exploration of crime patterns and trends.

RESULTS AND METHODS

Overall crime trends

To comprehend the overall trends, we initially computed and visualized the total reported crimes for each year, providing valuable insights into the evolving crime rates over time. The data reveals a distinct pattern in the reported crime numbers. In 2020, there was a significant decrease, likely due to unique circumstances. Subsequently, there was an upward trend, peaking in 2022, followed by a gradual decline.

The analysis demonstrates varying annual crime counts, with the highest in 2022, closely followed by 2021.

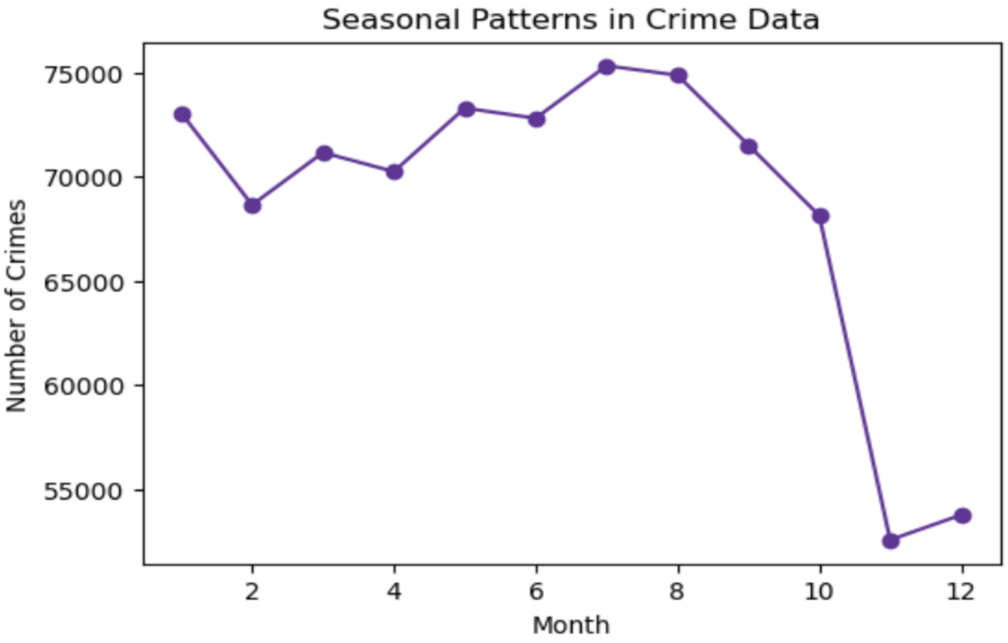


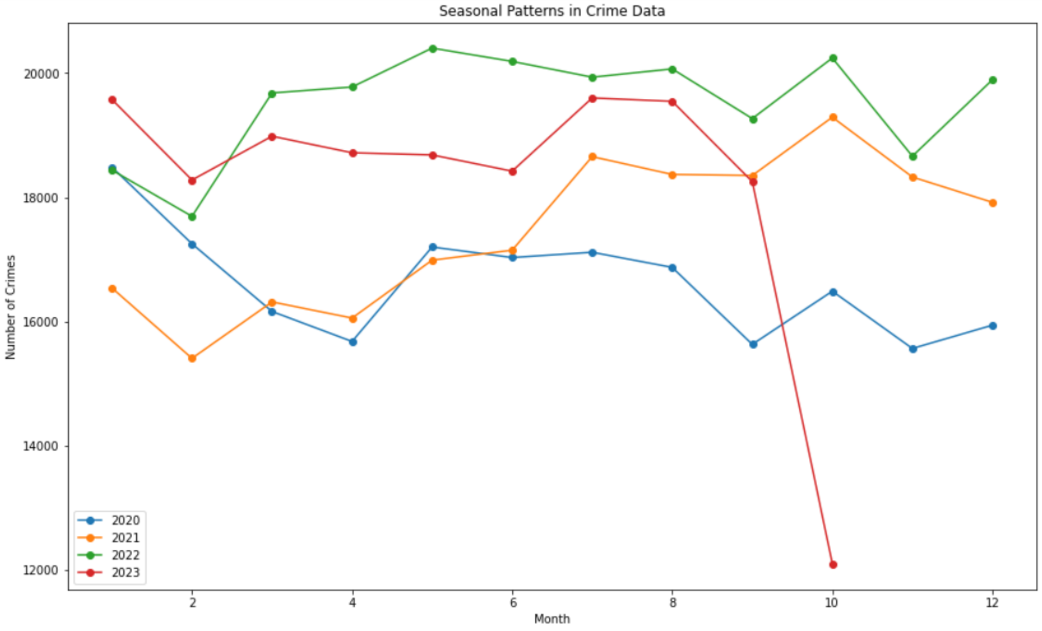
The year with the highest crime rate is 2022.

The year with the lowest crime rate is 2023.

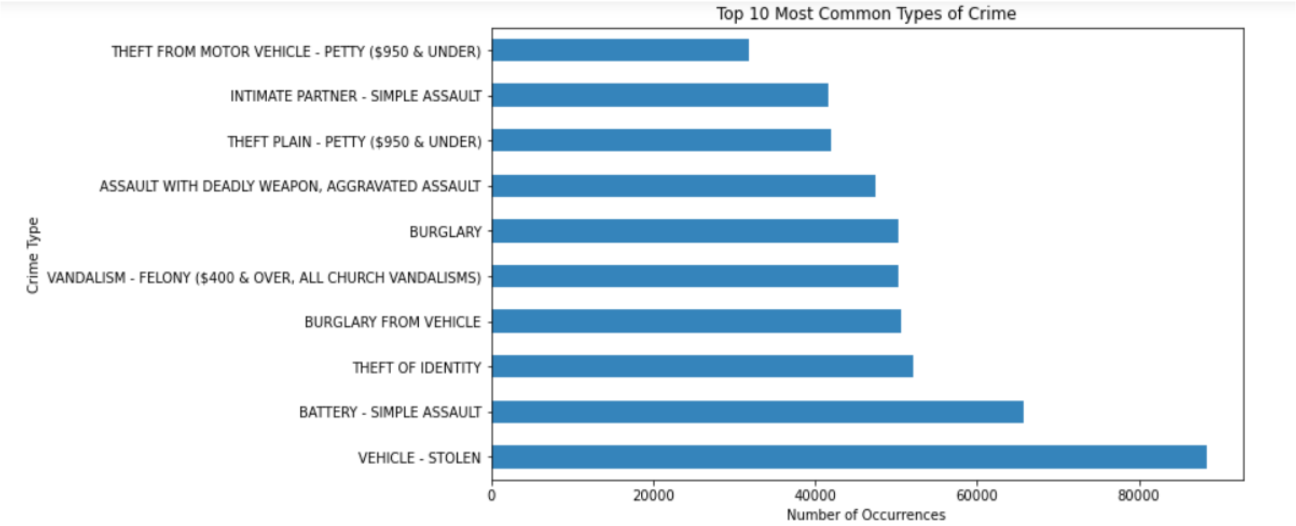
Seasonal crime patterns

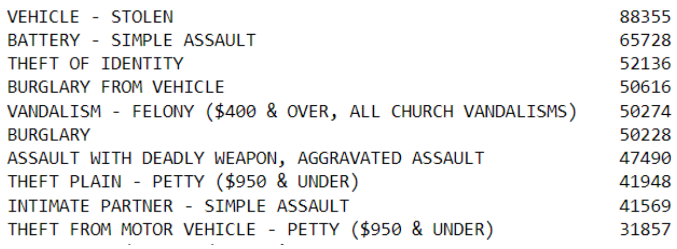
The months of June through August saw the highest crime rate. An extensive analysis of the seasonal crime patterns, especially from June to August, reveals a notable increase in criminal activity. This observation points to a possible seasonal concentration of certain kinds of crimes. By identifying particular crime trends and hotspots, more investigation into these patterns can help law enforcement authorities better allocate resources and carry out focused intervention measures. Seasonal variation analysis can also shed light on the temporal dynamics of criminal behavior, which makes it easier to build effective crime prevention strategies in a timely manner.





Most common types of crimes

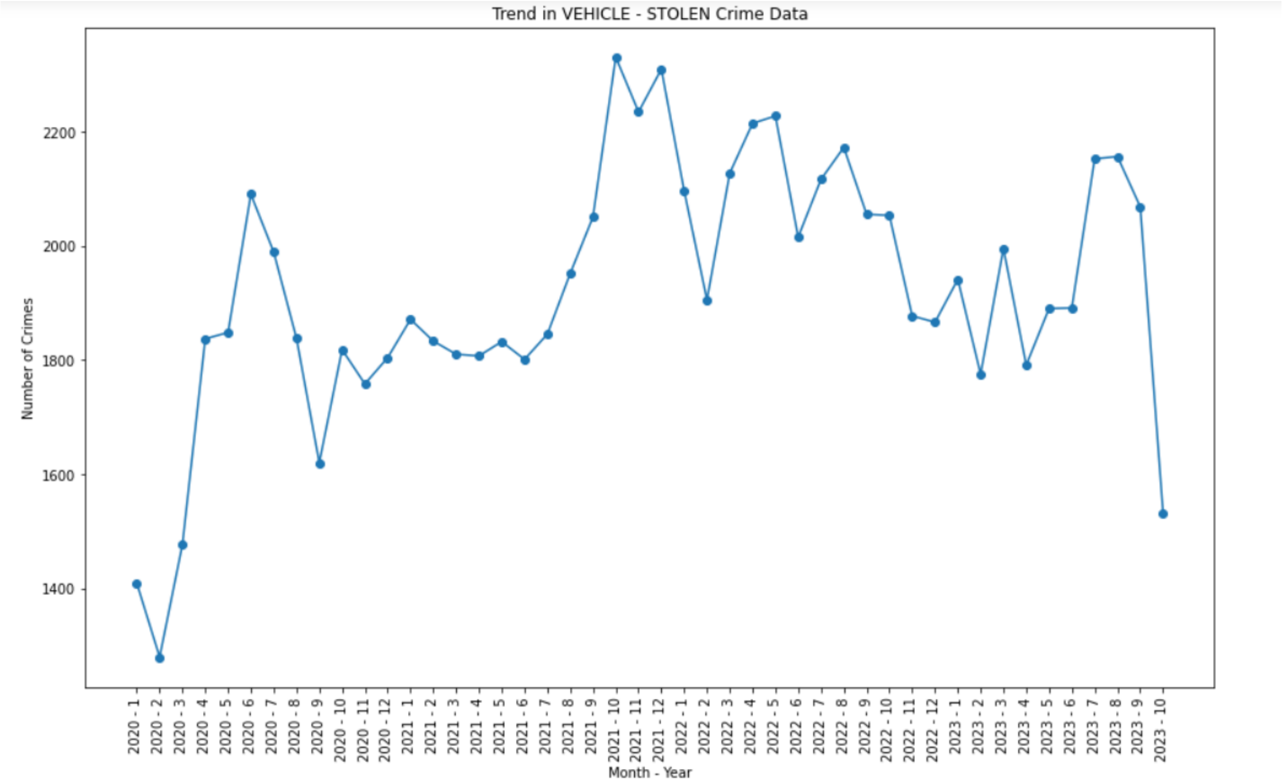




```
VEHICLE - STOLEN                                          88355
BATTERY - SIMPLE ASSAULT                                  65728
THEFT OF IDENTITY                                         52136
BURGLARY FROM VEHICLE                                     50616
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)   50274
BURGLARY                                                  50228
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT            47490
THEFT PLAIN - PETTY ($950 & UNDER)                        41948
INTIMATE PARTNER - SIMPLE ASSAULT                         41569
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)           31857
```

The results clearly show that, when compared to the other stated crime types, "VEHICLE - STOLEN" is the most common sort of crime in Los Angeles. This indicates that vehicle theft is a significant problem in the city and calls for concentrated efforts to solve the problem and enhance automotive security measures. Furthermore, the high rate of "VANDALISM - FELONY" and "BURGLARY FROM VEHICLE" highlights the significance of putting targeted policies into place to counteract crimes involving property.

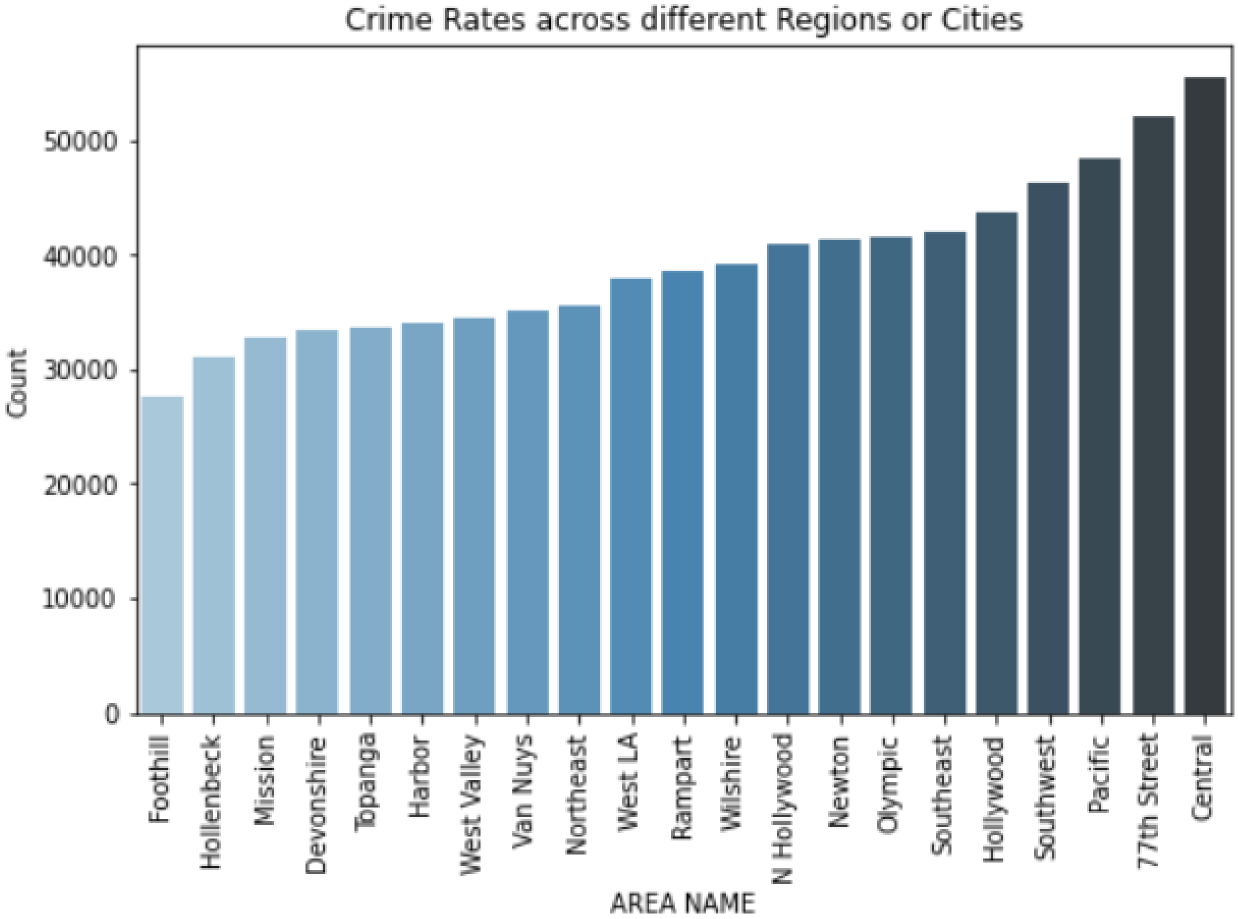
The fact that the terms "BATTERY - SIMPLE ASSAULT" and "ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT" are on the list further emphasizes the frequency of violent crimes and the necessity of strong law enforcement and community safety initiatives in order to deter and deal with them. The incidence of "THEFT OF IDENTITY" emphasizes even more how important it is for the city to improve identity protection and cybersecurity.



The data analysis pertaining to "VEHICLE - STOLEN" crimes reveals noticeable variations in their frequency over time as well as within certain months. Interestingly, a consistent pattern indicates higher incidences in the latter months of the year, though there are minor differences from year to year. In addition, the data shows that from the start of 2020 to the end of 2022, "VEHICLE - STOLEN" offenses generally increased. But by the end of 2023, there's a noticeable decrease in these occurrences.

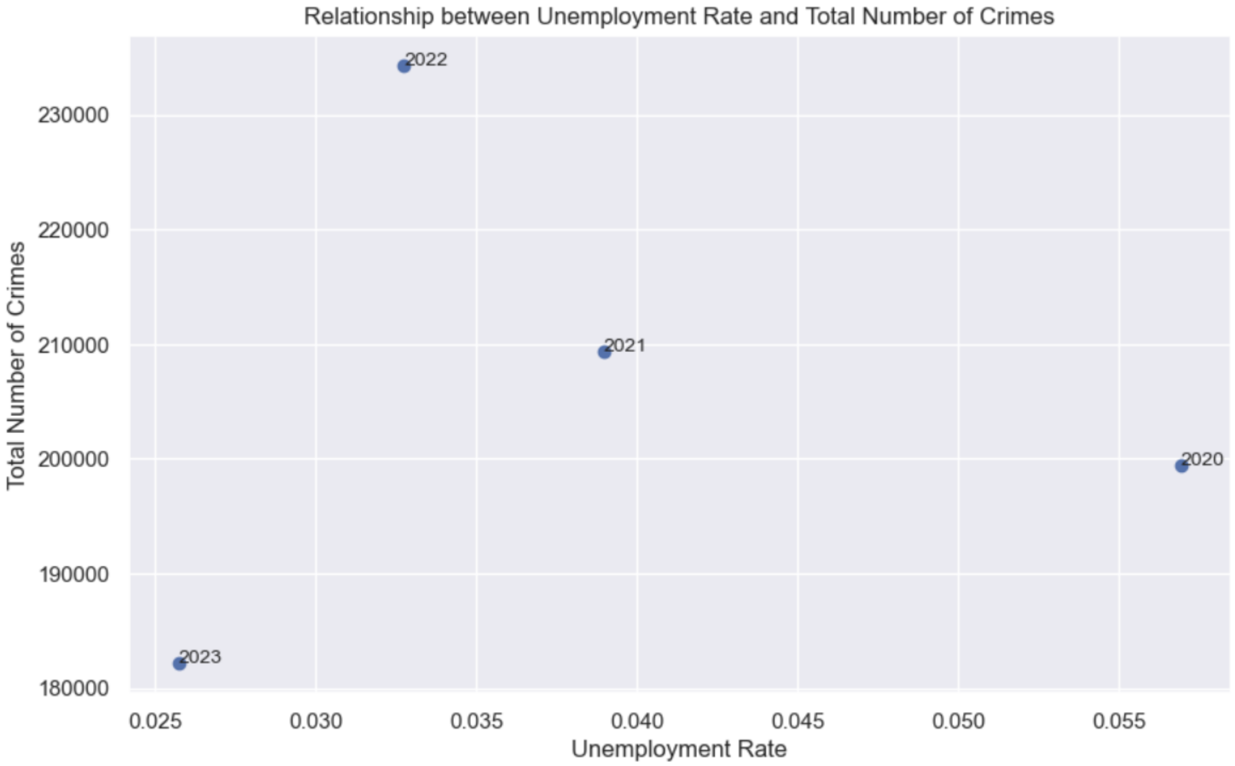
These results suggest that there may be seasonal patterns and temporal variations in the number of vehicle theft incidents in the city. Knowing these trends can help law enforcement organizations create focused plans to address and reduce the frequency of this particular kind of crime.

Regional difference



Among the listed regions, the Central area records the highest count of reported crimes, with 55,567 incidents, closely followed by the 77th Street region with 52,087 incidents. In contrast, the Foothill and Hollenbeck regions have the lowest counts, with 27,497 and 30,980 incidents, respectively. These statistics highlight the varying levels of reported criminal activities across different regions in Los Angeles, underscoring the need for tailored law enforcement strategies to address the specific challenges faced by each area.

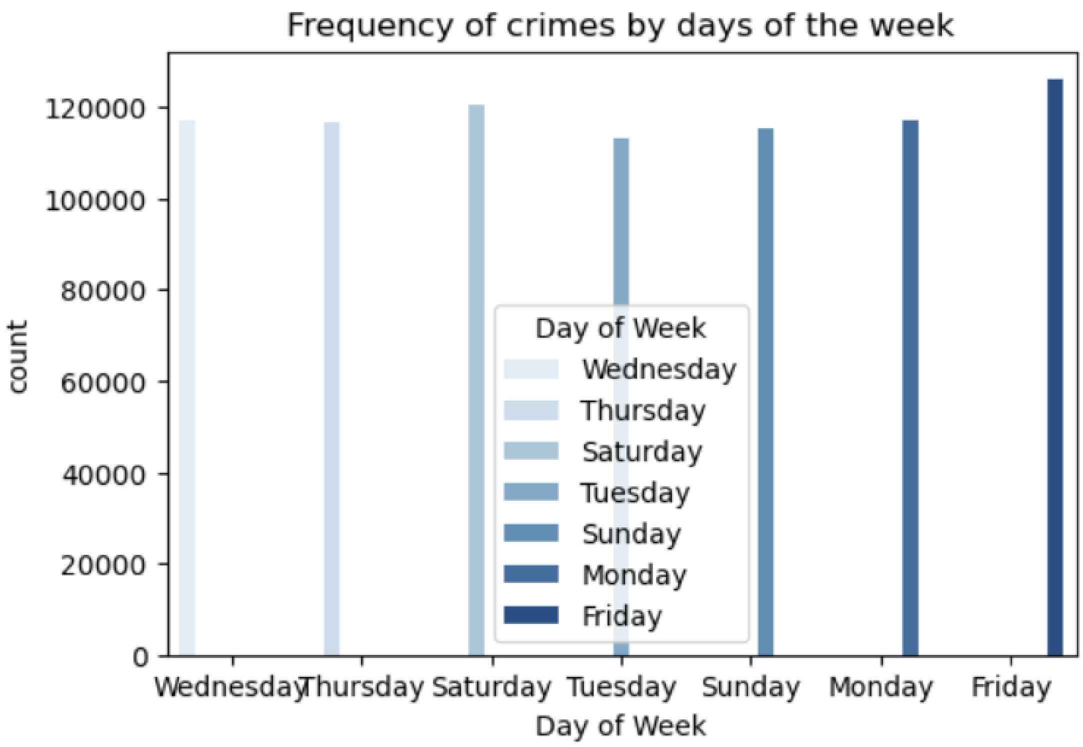
Correlation with Economic Factors

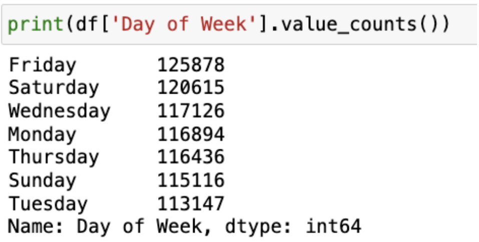


Pearson Correlation: 0.02
P-value: 0.98

The overall number of crimes and the unemployment rate appear to have a very weak positive link, as indicated by the high p-value of 0.98 and the Pearson correlation coefficient of 0.02. This suggests that there may be a little trend for the overall crime rate to go up in parallel with an increase in the unemployment rate. Nonetheless, the fact that the correlation is nearly zero suggests that the relationship is essentially insignificant. Therefore, it is apparent from the correlation study that the unemployment rate has little to no effect on the total number of crimes in the dataset, indicating that other factors might have a greater impact on the likelihood of crimes in the particular scenario.

Day of the Week Analysis

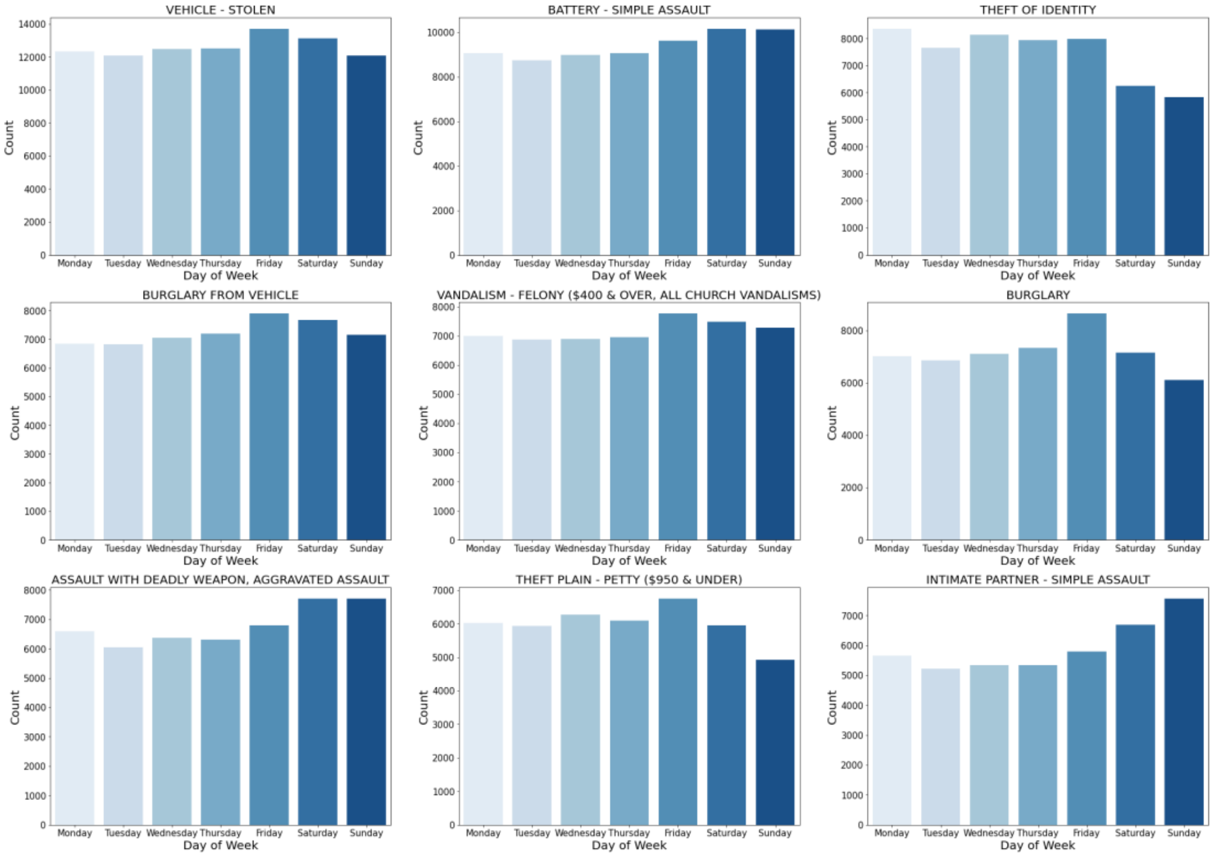




```
print(df['Day of Week'].value_counts())

Friday       125878
Saturday     120615
Wednesday    117126
Monday       116894
Thursday     116436
Sunday       115116
Tuesday      113147
Name: Day of Week, dtype: int64
```

Analyzing the numerical data reveals that the highest number of reported crimes occurs on Fridays, reaching 125,878 occurrences, closely followed by Saturdays with 120,615 reported crimes. Wednesdays, Mondays, Thursdays, Sundays, and Tuesdays demonstrate slightly lower counts, with 117,126, 116,894, 116,436, 115,116, and 113,147 reported crimes, respectively. These findings suggest a consistent level of criminal activities during weekdays, with a notable increase on Fridays and Saturdays, indicating potential trends in criminal behavior according to the day of the week. Understanding these patterns can aid law enforcement agencies in optimizing resource allocation and scheduling patrols to effectively manage and address the varying levels of criminal activity throughout the week.

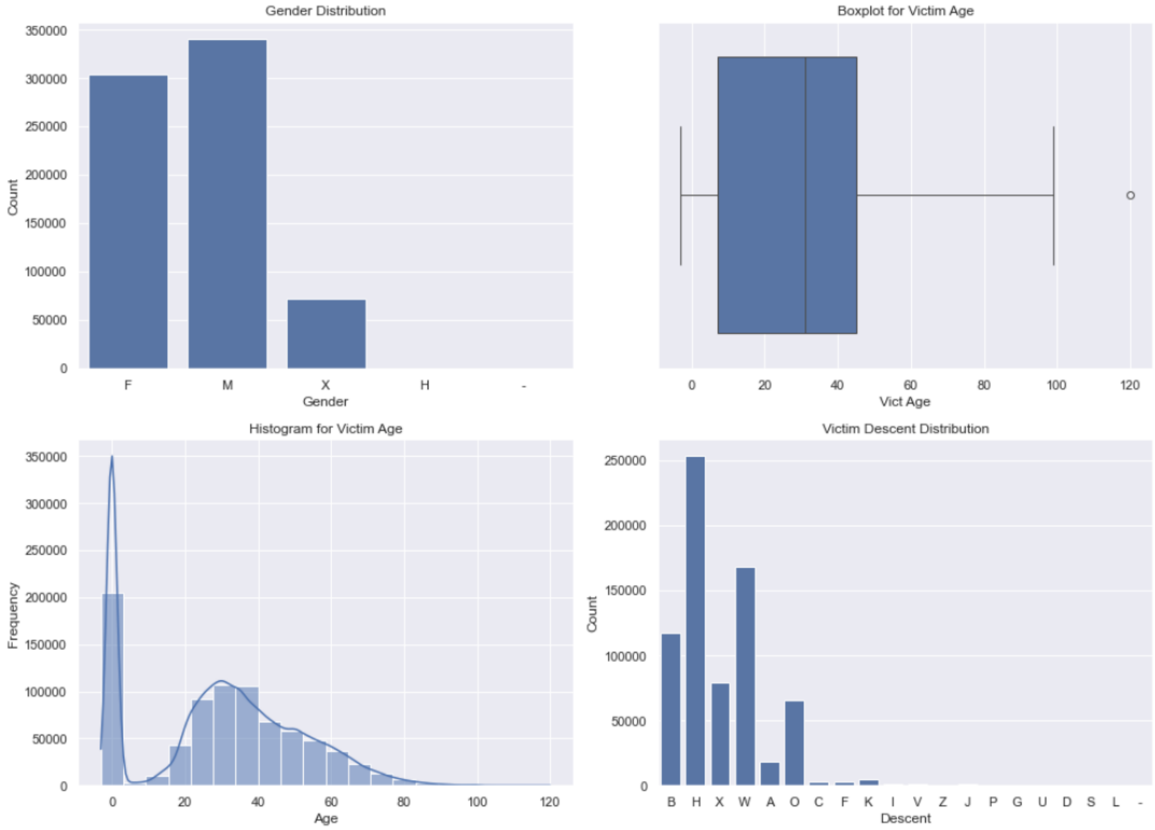


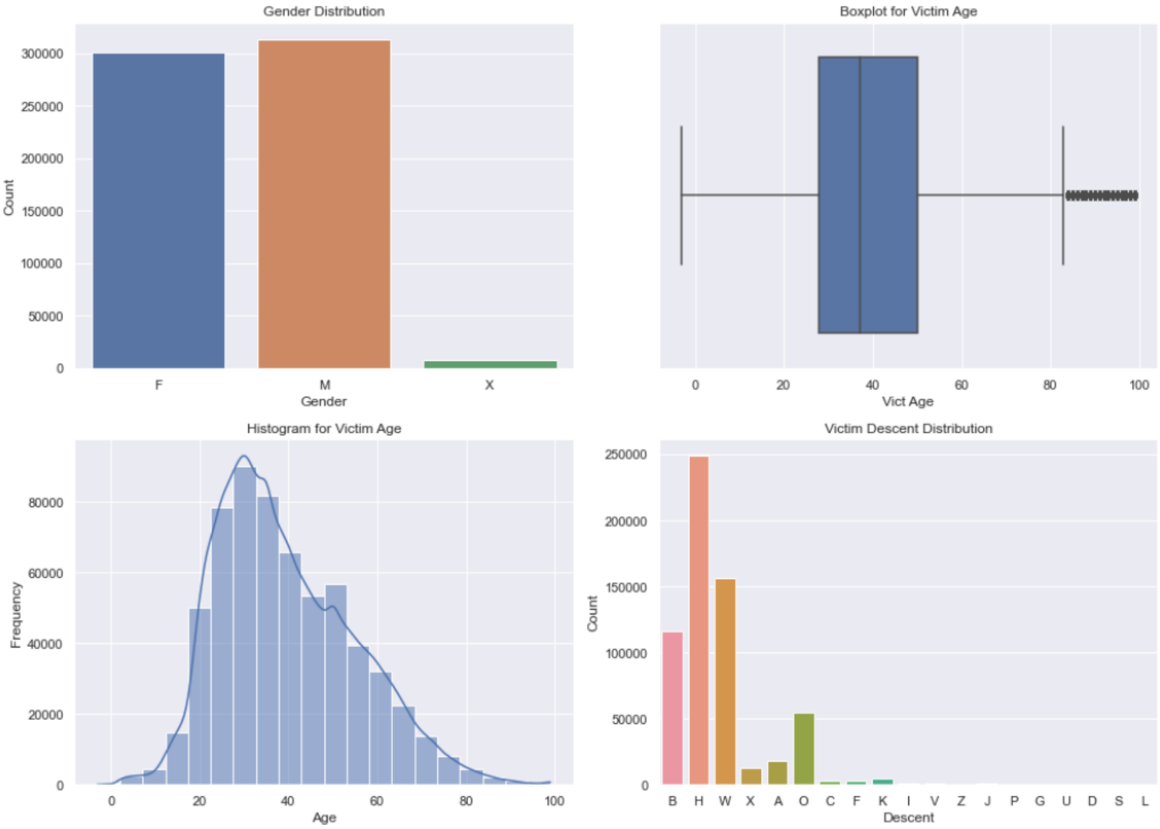
Frequency of crimes by days of the week for top 9 crimes

Throughout the week, the counts for various crime types show consistency, with minor fluctuations each day. Saturdays, in particular, tend to have slightly higher reported crime counts for several top crimes, including "Vehicle - Stolen," "Battery - Simple Assault," "Theft Plain - Petty ($950 & Under)," "Burglary From Vehicle," "Burglary," "Vandalism - Felony," "Robbery," "Assault with Deadly Weapon, Aggravated Assault," and "Intimate Partner - Simple Assault."

Certain offenses like "Battery - Simple Assault" and "Intimate Partner - Simple Assault" exhibit slightly higher counts during the weekend, likely due to increased social activities. Similarly, crimes such as "Vandalism - Felony" and "Robbery" also show a slight increase on weekends, possibly influenced by reduced surveillance or increased opportunities. Understanding these trends can assist law enforcement in addressing specific challenges related to weekend crime patterns and ensuring public safety during these times.

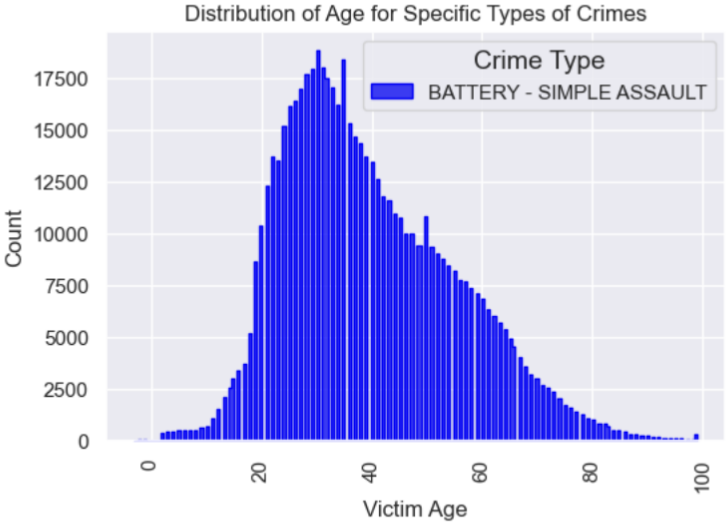
Outliers and Anomalies

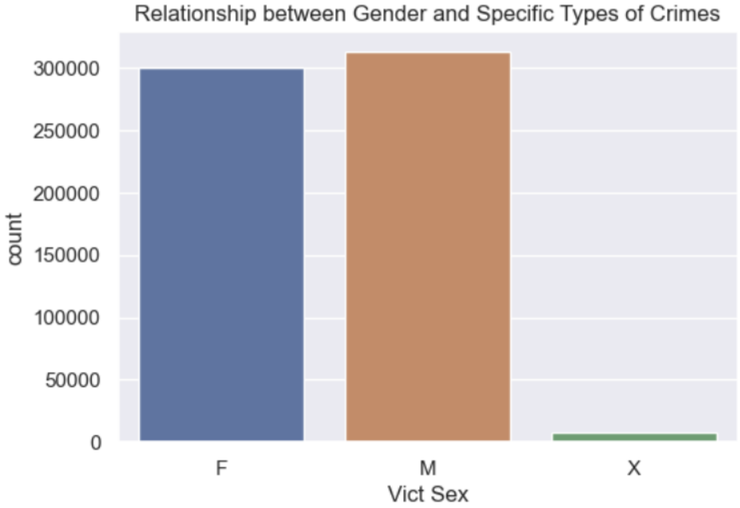




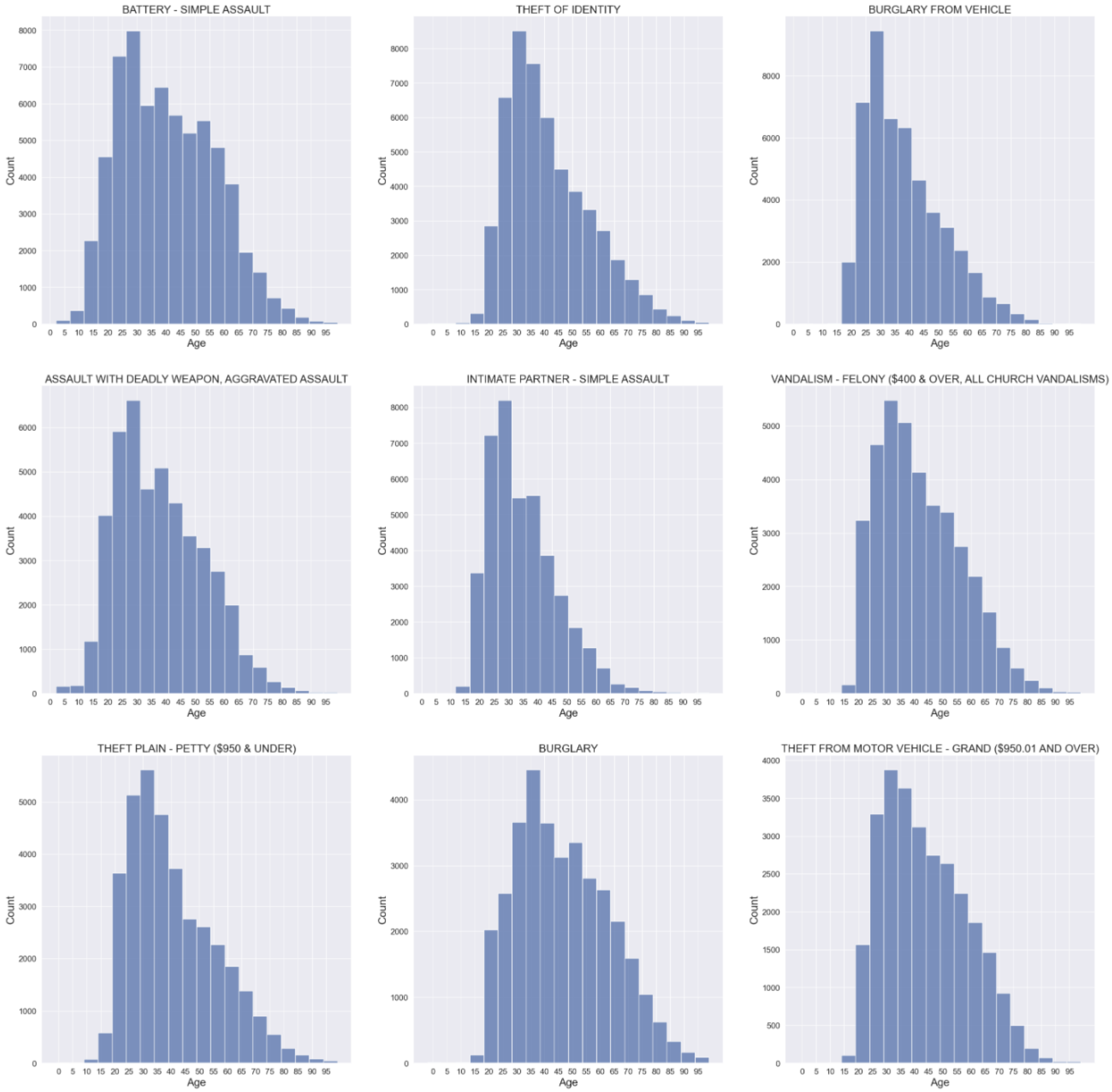
- The value 120 in the 'Vict Age' column appears to be an outlier based on the calculated quartiles and interquartile range. An age of 120 is unusually high and uncommon in typical demographic distributions. This could be due to data entry errors, missing values, or potentially exceptional cases that need to be investigated further.
- From the histogram for age, it can be seen that a lot of crimes has an age of 0 which is not reasonable and possible hence removing zero
- From the box plot it can be observed that an age greater than 100 is an outlier
- From the column description in the origin website, age is only M, F, X, and H and - are inconsistent data
- From the column description in the origin website, descent doesn't suppose to have "-"

Demographic Factors





From the above graphs, it is inferred that most number of criminals are between the ages of 20 and 40. And similarly, it is inferred that men are involved in the most number of crimes.



Distribution of Victim Age for Specific Types of Crimes

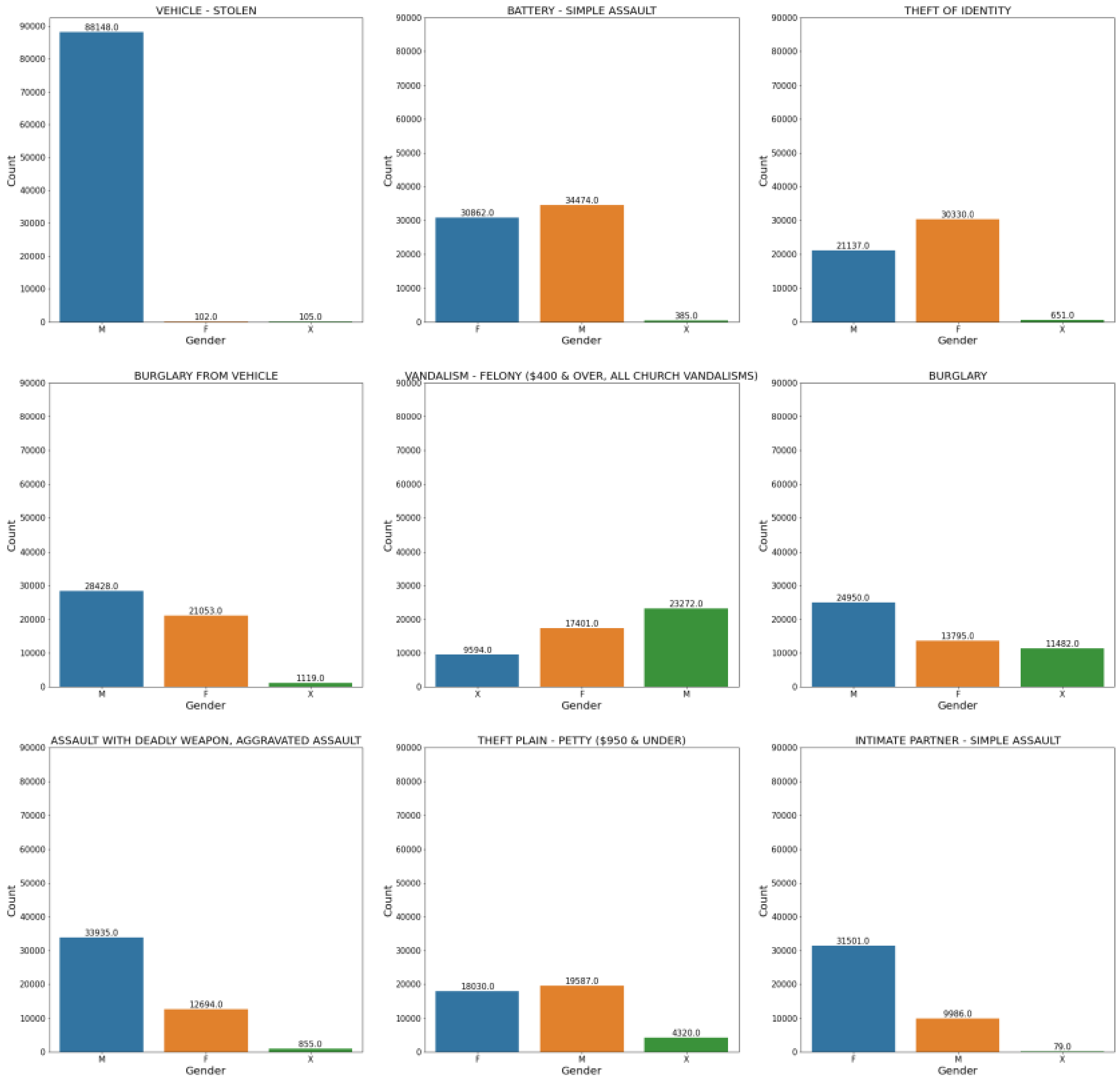
The data analysis for various crime types reveals distinct age trends within different categories of criminal activities.

- In the case of "BATTERY - SIMPLE ASSAULT," the data demonstrates a more spread-out distribution, with the peak count appearing at the age of 30.
- "THEFT OF IDENTITY" shows a peak count at the age of 30, indicating that individuals around this age might be more commonly affected by identity theft.
- The distribution for "BURGLARY FROM VEHICLE" is relatively uniform, with a slight peak count at the age of 30.
- "ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT" showcases a diverse

distribution, with the peak count at the age of between 25 to 30.

- "INTIMATE PARTNER - SIMPLE ASSAULT" indicates a slightly more dispersed age distribution, with the peak count at between 25 to 30
- "VANDALISM - FELONY exhibits a more diverse age distribution, with the peak count at 30 years.
- "THEFT PLAIN - PETTY ($950 & UNDER)" demonstrates a relatively uniform age distribution, with the peak count at the age between 25 to 30.
- In the case of "BURGLARY," the distribution displays a more uniform pattern, with the peak count at the age of 30.
- In the case of "THEFT FROM MOTOR VEHICLE - GRAND," the distribution displays a more uniform pattern, with the peak count at the age of 30.

Finally, "INTIMATE PARTNER - SIMPLE ASSAULT" indicates a slightly dispersed age distribution, with the peak count between the ages of 25 to 30, highlighting the involvement of individuals within this age range in cases of simple assault within intimate relationships. Understanding these age-specific patterns can assist in targeted crime prevention strategies and interventions aimed at specific age groups within the community.

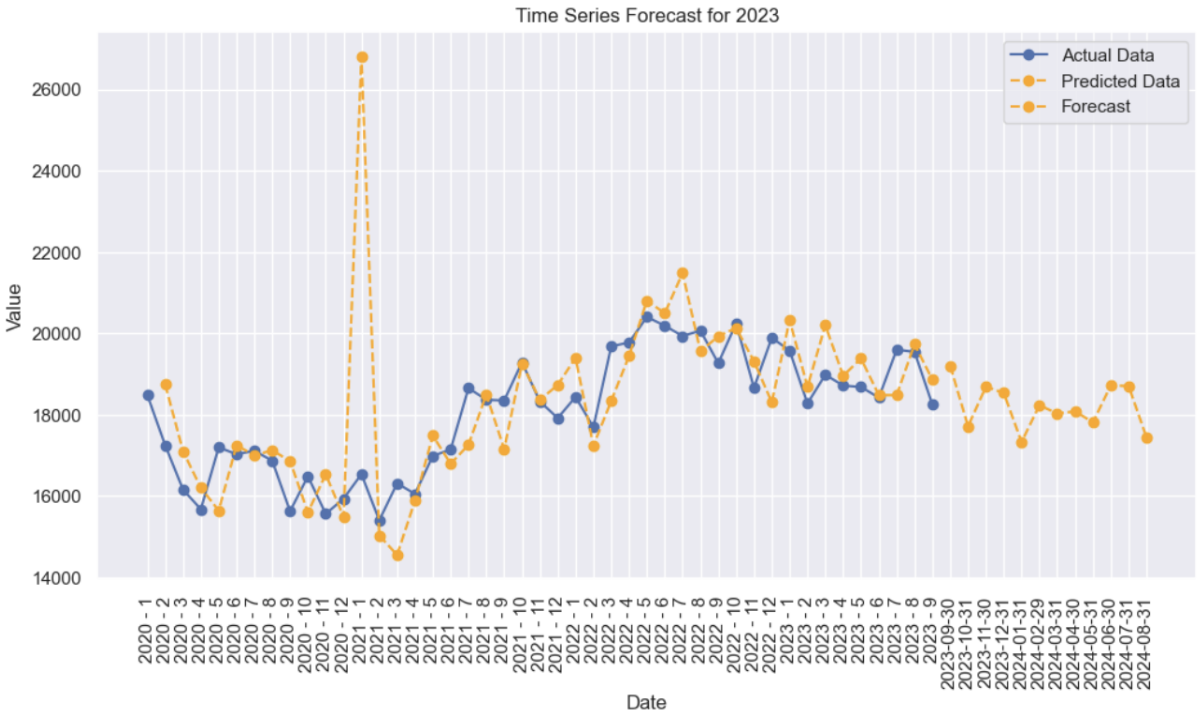


- The analysis of victim demographics for different crime types reveals nuanced patterns regarding gender distribution in various criminal incidents.
- Concerning "VEHICLE - STOLEN," the data indicates a notably higher count of male victims, suggesting that males are more commonly targeted in cases of vehicle theft.
- In the case of "BATTERY - SIMPLE ASSAULT," the data portrays a relatively balanced distribution between male and female victims, with a slightly higher count of female victims compared to males, indicating a more even victimization pattern in incidents of simple assault.
- For "THEFT OF IDENTITY," the data suggests a higher count of female victims, indicating a higher likelihood of females being targeted for identity theft compared to males.
- Similarly, "BURGLARY FROM VEHICLE" demonstrates a more balanced gender

distribution, with a slightly higher count of male victims, indicating a relatively even victimization pattern in this type of crime.

- The data for "VANDALISM - FELONY" reflects a substantial discrepancy, with a higher count of male victims compared to female victims, suggesting that males are more commonly targeted in felony vandalism cases.
- Furthermore, "BURGLARY" indicates a higher count of male victims compared to females, suggesting that males are more commonly targeted in burglary incidents.
- "ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT" displays a significantly higher count of male victims compared to females, indicating a higher likelihood of males being targeted for this type of assault.
- On the other hand, "THEFT PLAIN - PETTY" demonstrates a relatively balanced gender distribution, with a slightly higher count of female victims, suggesting a relatively even victimization pattern in petty theft cases.
- Finally, "INTIMATE PARTNER - SIMPLE ASSAULT" reveals a notably higher count of female victims compared to males, suggesting that females are more commonly targeted in cases of intimate partner violence. Understanding these gender-specific trends can aid in the development of targeted interventions and support services tailored to the specific needs and vulnerabilities of different gender groups affected by these crimes.

Predicting Future Trends



- The message "CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH" means that the reduction in the objective function value is less than the specified threshold (likely the default machine precision).
- The total number of iterations performed is 49, which includes function evaluations and other steps involved in the optimization process.
- The algorithm made 57 function evaluations during the optimization process, suggesting that it probed the objective function at different points in the parameter space to find the optimal solution.
- The projected gradient at the final iteration is very small (1.438D-04), indicating that the gradient is close to zero, and the algorithm is close to the optimal solution.
- The final function value achieved by the SARIMAX model is 5.7306000291302803, which likely corresponds to the minimized error or loss function.

SUMMARY AND RESULT

The comprehensive analysis of the crime data for the City of Los Angeles uncovers crucial insights into the shifting crime trends over the years. Notably, 2020 saw a significant decline in reported crimes, likely due to unique circumstances, followed by an upward trajectory peaking in 2022 and a subsequent gradual decline. Seasonal patterns revealed heightened criminal activity during the summer months, emphasizing the importance of understanding temporal dynamics for effective law enforcement strategies.

Furthermore, the prominence of "Vehicle - Stolen" and "Vandalism - Felony" emphasizes the need for targeted policies to combat property-related crimes, while the occurrences of violent crimes like "Battery - Simple Assault" and "Assault with Deadly Weapon" underscore the necessity of robust law enforcement and community safety initiatives. The analysis of demographic data indicates a notable concentration of criminals between the ages of 20 and 40, with men being more frequently involved in various types of crimes.

Additionally, the weak correlation between the overall number of crimes and the unemployment rate suggests that other factors might significantly influence crime rates. Understanding the varying levels of criminal activities throughout the week, with heightened incidents on Fridays and Saturdays, aids law enforcement in resource allocation and scheduling patrols effectively.

In conclusion, this analysis highlights the significance of a data-driven approach in understanding crime dynamics, aiding policymakers and law enforcement agencies in implementing targeted interventions and proactive measures to ensure public safety and well-being in the City of Los Angeles.

LIMITATIONS AND FUTURE WORK

One major obstacle we faced in our criminal data analysis work was the amount of data that was available. Although the official US government source provided us with our data, the dataset was restricted to the years 2020–present. Due to this restriction, we were unable to conduct long-term historical analysis or look into crime trends that occurred after this comparatively little time frame. The dataset should be broadened to incorporate historical data in future iterations of this study so that we can better anticipate future trends in crime and recognize long-term patterns. Furthermore, a more thorough knowledge of the variables impacting crime patterns would be possible with the inclusion of additional demographic and socioeconomic data.

The data's granularity is another significant drawback. Aggregated crime statistics are included in the dataset, which is frequently arranged by county or city. Access to more detailed data, perhaps down to the neighborhood or block level, would be helpful for a more thorough review. With the use of this finely detailed data, we would be able to pinpoint certain crime hotspots, comprehend regional trends, and modify our crime prevention tactics as necessary. Looking ahead, obtaining such precise data through cooperation with local law enforcement organizations would significantly improve the project's analytical capabilities.

It is imperative to investigate more sophisticated predictive modeling methods in future research. Although time series forecasting was used in our project to estimate future patterns in crime, more advanced methods, including machine learning models, might be used to increase the accuracy of these forecasts. Law enforcement agencies and legislators would also receive timely insights and more potent tools to combat and prevent crime if real-time data sources were integrated and our Jupyter Notebook-based research was made more interactive. These improvements are essential to improving the relevance and usefulness of our study for the good of communities and public safety.