Los Angeles Crime Data Analysis

100565371

College of Science and Engineering, University of Derby

[100565371@unimail.derby.ac.uk](mailto:100565371@unimail.derby.ac.uk)

**Abstract.** The report examines crime trends in Los Angeles, focusing on identity theft, using a dataset from 2020 to the present. It reveals demographic disparities, with Hispanic, White, and Black populations being most victimized. High crime rates were found in Central, 77th Street, and Pacific. Identity theft incidents are reported later than other crimes, with most occurring in single-family and multi-unit dwellings. Machine learning models were applied to predict daily identity theft incidents.

1. Introduction

Like many big cities, Los Angeles has recently struggled with different crime-related issues influenced by a range of factors including geographical variations, demographic shifts, and the nature of crime itself. Understanding this dynamic is essential in formulating effective measures to promote public and community well-being.

The Los Angeles Police Department is responsible for ensuring these measures are in place to protect people and their properties. The department faces various challenges in ensuring its goal is fulfilled, as it finds it difficult to answer several questions that arise while trying to solve this problem.

The study (Rosenfeld & Austin, 2023) uses statistical models to forecast annual changes in property and violent crime rates from the 1960s until 2021. The models project a rise in property and violent crime rates in Los Angeles in 2022, with a gradual increase in violent crime. Previous studies concluded that several elements affect the probability of criminal activity in Los Angeles, such as the time of day, location, gender, and ethnicity.

This report examines the correlation between crime victims' demographics and location, aiming to improve Los Angeles residents' safety of living. Python programming is used to analyze crime data, identify patterns, identify crime hotspots, and predict crime numbers using regression algorithms. The analysis aids law enforcement and public safety.

1. Dataset Description

The dataset used for this analysis is a compilation of crime data(2020-Present) of about 21 areas in the city of Los Angeles done by the Los Angeles Police Department (LAPD). The dataset contains 28 columns including the time and crime location, geographic unit, demographic information of the victims, and the description of offences with about 879106 observations.

The crime report includes various details such as the Division of Number, Date reported, Date and Time of occurence, Area ,Crime description, Mocodes , Victim Age, Victim Sex, Victim Descendant, Premise, Weapon used, Status of crime, Additional codes related to the crime, Location, Latitude and Longitude. These details help in understanding the crime and its perpetrator.

* 1. Dataset Overview and Data Cleaning

The analysis of Los Angeles crime trends requires a thorough understanding of the dataset. The dataset was inspected for missing values in variables such as Mocodes, victim sex, victim descendant, premise code, weapon description, weapon used code, and cross street. These values may have been lost due to law enforcement not recording demographic details, poor documentation, or unknown weapon identification. Missing values were replaced with appropriate placeholders to maintain consistency. Outliers were identified using the Inter-Quartile Range (IQR) based outlier method, and the data was visualized using boxplot as seen in the code file of this report. The latitude and longitude values should be between -90 and 90, -180 and 180 respectively but the value in the dataset includes invalid entries like 0. To fix this inconsistency, rather than removing the rows entirely the invalid entries were only excluded while dealing with these variables to avoid bias in the result. Informed permission, confidentiality, honesty, and allegiance are among the ethical considerations that must be balanced with the right to privacy for crime victims and their families. Transparency, accuracy, and a broad audience are all necessary for an effective presentation.

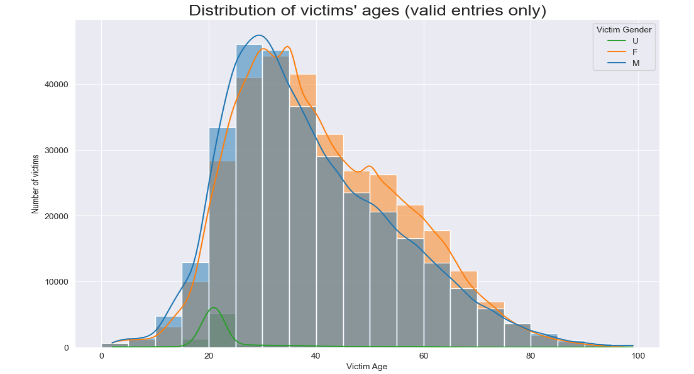
* 1. Victims Analysis

The pie chart in the appendix reflects the diversity of crime victims by ethnic group and indicates which groups are more frequently victimized relative to the crime data. The figures were represented in the context of the overall ethnic group in the dataset.

The heatmap in appendix shows that Hispanics, Whites, Blacks, and Hawaiians are more vulnerable to crime incidents, possibly due to their population size or presence in areas with higher crime. The Central area has the highest incidences, while West LA, Olympic, Northeast, and Harbour have moderate to lower incidences. Asians have less incidences in certain neighborhoods. The Hispanic group has a larger frequency of events, while the Black group has a higher number.

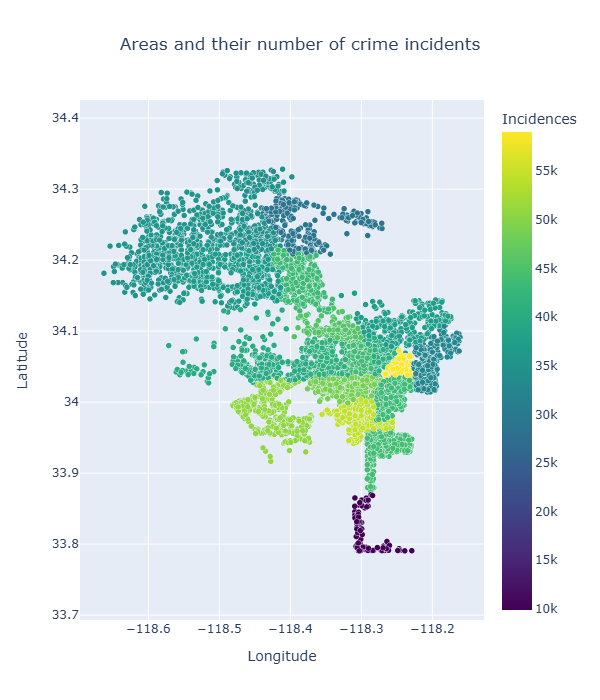
Strategies for community support and crime prevention may need to receive more concentrated attention in areas where there is a persistently high incidence rate. The disparity in event counts between ethnic groups may indicate the presence of underlying environmental, social, or economic factors that impact reporting or crime rates.

Based to the data, both male and female younger individuals, especially those between the ages of 20 and 30, are the most frequently reported victims of crime. The victim counts decline with age, indicating that younger people are either more likely to report crimes or more likely to be victims themselves. It is significant to highlight that because of reporting rates, the type of crimes committed, and other sociocultural factors that may affect victimization or the likelihood of reporting a crime, the actual number of male and female victims may not be equal to the population. However, the representation indicates that there are more men among the victims of crime. But while men are more likely to become victims of crime overall, women are more likely to become victims of grand theft, property theft, kidnapping, rape, and other. Because the gendered victims are unknown, both men and women are undervalued. This could be because some victims choose not to disclose their identity because of embarrassment, cultural norms, trauma, or a desire to escape the charade on social media.

**Fig. 1.** 

* 1. Geographical Analysis

From the result of the analysis done on the area of the crime location, it happened that the Central region of the Los Angeles has the highest crime rate, followed by the 77th Street and Pacific and in descending order as shown in the figure below, please refer to the code for a more interactive view of the plot.

**Fig. 2.** 

* 1. Temporal Analysis

The crime rate in Los Angeles has increased over time. On average, from January 2020 to January 2024 there are 597 crimes each day.

On a weekly basis, as represented in appendix it happens that more crimes tend to happen over the weekend starting from Friday till Monday implying that crime activities increase towards and during the weekend. Suggesting that extra measures be put in place like law enforcement officials positioning in spots where crimes are most likely to occur.

On a monthly basis as represented in appendix, the crime rate is even for each month, but more crime happens in January, this could be because of the holidays in December towards January. There is usually more pressure to spend more during this period and people end up with less funds in January.

The chart in appendix shows the number of crime occurrences over the years, 2024 seems to have the lowest number of crimes but this is due to the fact data for 2024 ends in January therefore the impact data collected during this period is insignificant. Crime has increased gradually since 2020 with an insignificant decrease in 2023 indicating that crime everyone has a lot of work to do pertaining to the reduction of crimes in Los Angeles, not just, the Law enforcement individuals need to be at alert.

**Fig. 3.** A colorful squares with white text

Description automatically generated

The heatmap clearly shows that the lunchtime and early afternoon hours are when crime is most active, with a notable spike in activity from 8:00 to 18:00. Since there is the least criminal activity between 00:00 and 06:00, early mornings and late nights are often safer. The little monthly differences highlight how important the time of day is in determining the frequency of crimes. Businesses can improve security during vulnerable hours, law enforcement organisations can optimise patrol plans with this information, and the public should remain cautious during peak hours.

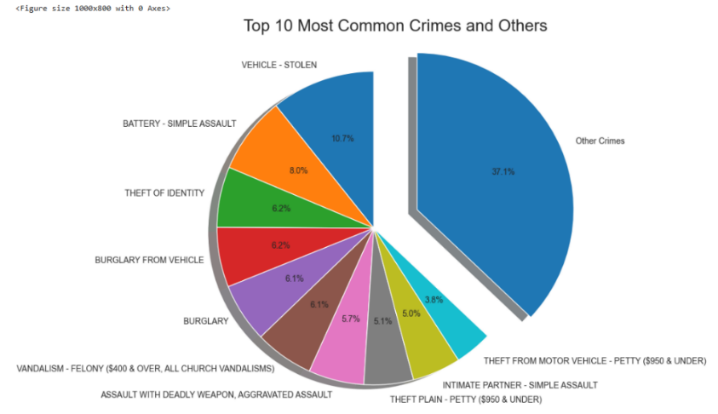
* 1. Crime Analysis

The crimes were categorized into property crimes and violent crimes. Property crimes involve the loss of personal belongings which include theft, burglary, arson, vandalism, etc. Violent crimes involve the use of weapons or hostile actions like the use of fists including assault, sex offenses, homicide, robbery, etc. Looking at a broader categorization of these crimes it appears that theft is more prevalent as seen in the appendix below.

The crime description attribute has about 139 unique values which indicates 139 unique crime descriptions. The top ten crimes make up about two-thirds of the population.

The analysis reveals that property crimes, especially crimes related to theft, are more common in Los Angeles, as represented in the figure below, with car thefts being the most common type of crime. These crimes should be the focus of law enforcement personnel and community preventive initiatives. Simple Assaults and Identity thefts also stand out explaining the need for public awareness campaigns and preventive actions.

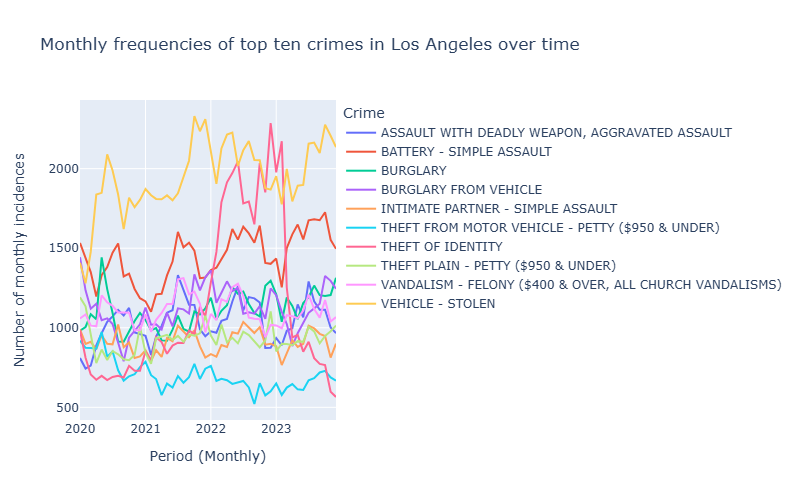
The analysis also shows that although the most prevalent crimes are widely known and noticeable, there is a wide range of less common but nonetheless important crimes that are in the "Others" category as shown in the figure below, which should not be ignored. To achieve a complete strategy to lower crime rates overall, this data-driven method could direct the distribution of resources, police patrols, and community assistance programs to the locations and crime categories most at danger. measures against these crimes.

**Fig. 4.** 

I focused on the top ten crimes and went into more detail about offences that pose a greater risk to public safety. I went further with the Crime analysis by checking the trends of the top ten criminal activities over the years and they are represented in the appendix below.

Representing these crimes on a weekly, monthly, and yearly trend, the graph in appendix displays a variety of crimes, including vehicle theft, identity theft, vandalism, theft from motor vehicles, simple assault, and assault with a lethal weapon. Week-to-week variations are notable, which is to be expected given the potential for short-term swings in crime rates. Certain crimes have discernible peaks at various times; for instance, Identity theft has multiple notable peaks that are noticeably greater than the rates of other crimes.

A clearer picture of the general patterns in monthly crime rates is given by the graph in the appendix, which smoothens out some of the fluctuation observed in the weekly data. We can observe patterns or trends with monthly representation that are not as obvious with weekly data, including seasonal variations or the results of policy changes or events. Just like the weekly trend, the identification trend of the monthly and yearly trend of common crimes is likewise indicating a notable uptrend of identity theft towards 2022–2023.

**Fig. 5.** 

Car theft continues to be the most prevalent crime in history. Nonetheless, identity theft has increased dramatically over the past year and has shown an erratic pattern (it is possible that the peaks coincide with the payment week). This indicates that citizens of Los Angeles should be issued further cautions about safeguarding their personal information, particularly because identity thieves have become increasingly adept at using various data harvesting techniques. Individuals are more susceptible to deception because of the growing necessity for data collection, which is understandable given how data analytics and machine learning have advanced modern technology.

Most crimes are reported early enough, as shown in the appendix, like aggravated assault, simple assault, burglary, and vandalism. Theft is usually not reported on time, especially Identity theft. Looking at Identity theft, it takes about a few days even a month for people to report these crimes, it is rare to see this crime being reported the same day. This indicates that identity theft takes time to be discovered, the victims may not know their identity has been stolen until they start noticing some issues especially with their finances. Victims might end up paying for goods or bills they did not buy or use. Victims even find out that their details have been used to take loans or even as guarantors for loans.

#### Identity Theft Analysis

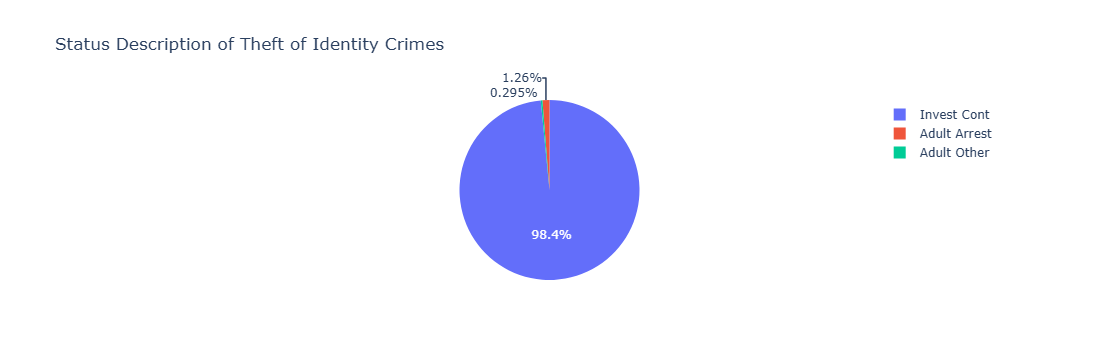
Identity theft is a crime that violates the right to privacy and confidentiality because it involves using someone’s information without consent. US economy faces significant threat from identity theft, resulting in losses of $442 million in 1995, $450 million in 1996, and $745 million in 1997, with predicted cost exceeding $2.3 billion. (Allison et al., 2005). Financial crimes are rising steadily because of identity theft.

Identity theft happens everywhere, including family dwellings, multi-unit buildings, finance companies, cyberspace, cash points (ATM), on the street shopping malls and stores. Still, it is so unfortunate that according to the chart in the appendix, it happens more in single-family dwellings and multi-unit dwellings. Single-family homeowners could feel more secure and private in their homes, which makes them less careful about protecting their personal data. Their lack of awareness can render individuals more susceptible to identity theft methods, including social engineering and phishing frauds. Shared amenities and services, like mailrooms, and laundry rooms, are common in multi-unit housing and may provide several people with access to personal information. Identity thieves can intercept mail or gain access to private documents because of this common environment. To reduce the risk of identity theft, it is important to establish security measures, raise awareness among inhabitants, and practise vigilance, regardless of the kind of home.

The chart in the appendix below shows that Identity theft usually does not involve the use of weapons which explains the reason the happens mostly where people live. Identity theft is a kind of crime that leaves you unaware as it involves non-violent means such as hacking and phishing which informs the need for cybersecurity.

The pie chart in the appendix displays the status description of identity theft crimes. The chart is divided into three categories: Invest Cont, Adult Arrest, and Adult Other.

The largest portion of the pie chart, taking up 98.5%, is labeled "Invest Cont." which stands for "Investigation Continued." The second-largest portion, at 1.24%, is labeled "Adult Arrest," suggesting that a small percentage of these crimes have resulted in the arrest of an adult suspect. The smallest portion, at 0.29%, is labeled "Adult Other," which could represent other outcomes for adult suspects, such as charges being dropped, or cases being closed for distinct reasons This indicates that most of the identity theft cases are still under active investigation and have not been resolved, with only a small fraction leading to arrests or convictions.

**Fig. 6.** 

1. Machine Learning
   1. Regression

In this section, I will explore different regression algorithms including Random Forest, Linear Regression, Gradient Boosting regression, K-Nearest Neighbours (KNN) to predict the number of daily incidents of identity theft per area and compare their accuracy using regression evaluation metrics.

### Data Preparation:

* Grouping Data: The data is initially grouped by date and area name to obtain the daily count of identity theft incidences per area. This step aggregates the data to provide a clearer picture of the frequency of identity theft occurrences in different areas over time.
* Feature Engineering: One-hot encoding converts categorical data into numerical format for machine learning algorithms, ensuring each unique area is represented as a binary vector with separate features.
* Combine Features: The one-hot encoded features are combined with the daily incident counts. This step integrates the engineered features with the target variable, creating a dataset suitable for machine learning model training.

### Machine Learning Preparation:

* Prepare Data for Machine Learning: The features (X) are separated from the target variable (y). The 'DATE OCC' column is dropped from the features as it is not used for prediction. Similarly, the 'DailyIncidents' column, representing the target variable, is separated from the features.
* Split Data: The data is split into training and testing sets using the train\_test\_split function from scikit-learn. This step is crucial for evaluating the performance of machine learning models. The training set trains the models, while the testing set assesses their predictive accuracy.

#### Principal Component Analysis

The dataset's dimensions were decreased, and its redundancy was decreased using PCA. A portion of the modified dataset obtained by the PCA technique is shown in Table 1.

**Table 1.** Subset of X\_train\_pca

|  |  |  |
| --- | --- | --- |
| pci | pc2 | pc3 |
| -0.693151 | 0.029779 | 2.400188 |
| 0.159296 | -0.517305 | -0.606642 |
| -0.140221 | -0.147874 | -0.241105 |
| -0.270243 | -0.297711 | 1.237249 |
| -0.172689 | 0.077526 | -0.195858 |

**Table 2.** Subset of X\_test\_pca:

|  |  |  |
| --- | --- | --- |
| pci | pc2 | pc3 |
| -0.171317 | 0.076300 | -0.193263 |
| -0.948753 | 2.514969 | -1.612665 |
| 0.156291 | -0.513711 | -0.604171 |
| 0.156291 | -0.513711 | -0.604171 |
| 0.196073 | -0.199640 | -0.100728 |

**Table 3.** Evaluation Metrics

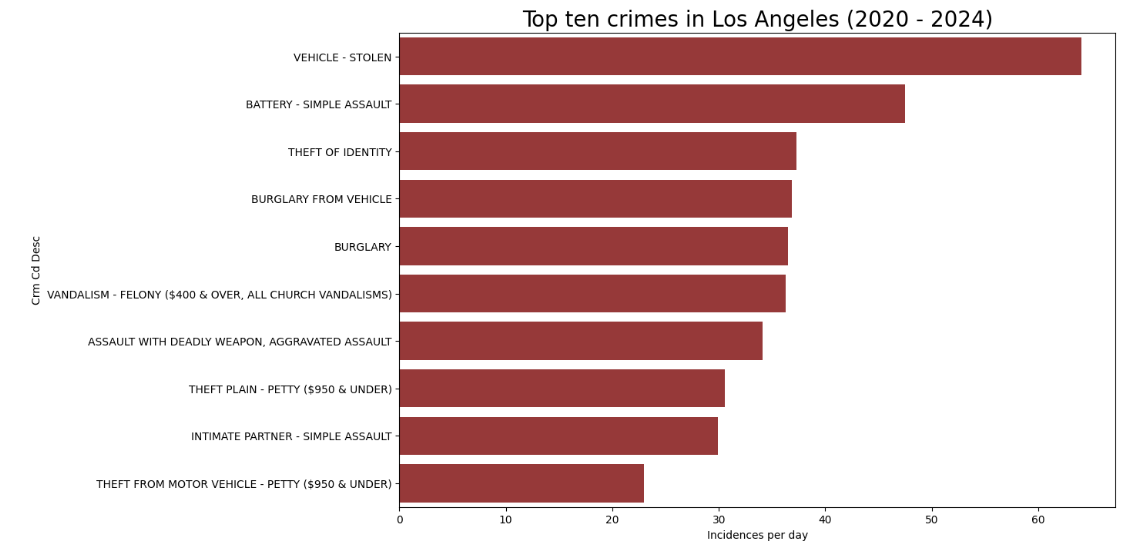
|  |  |  |
| --- | --- | --- |
| model | Mse | R2 score |
| Random Forest | 16.317467 | 0.000009 |
| Linear Regression | 16.130550 | 0.011464 |
| Gradient Boosting (GBR) | 16.129132 | 0.011551 |
| K-Nearest Neighbours (KNN) | 17.749319 | -0.087740 |

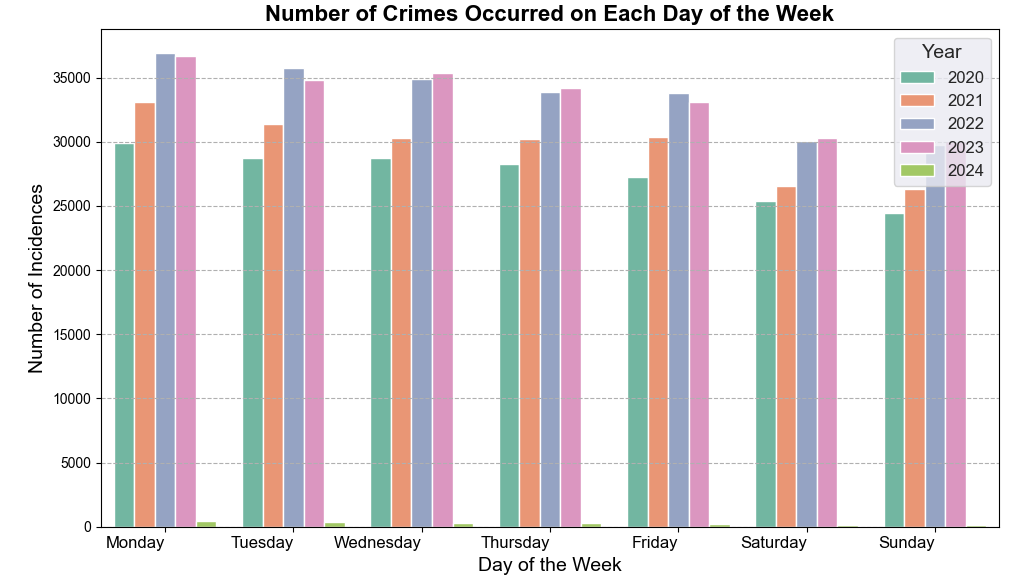
The result of the analysis concludes that the models, Random Forest and KNN do not perform well in predicting daily incidences of identity theft in Los Angeles. The high MSE values and low (or negative) R² scores suggest that the models fail to capture the complex patterns in the data. The Gradient Boosting Regressor and Linear Regression models showed the lowest MSE and R2 scores, respectively, indicating better performance in variance explanation. However, all models had R2 scores close to zero, indicating limited explanatory power of the features.

1. Conclusion

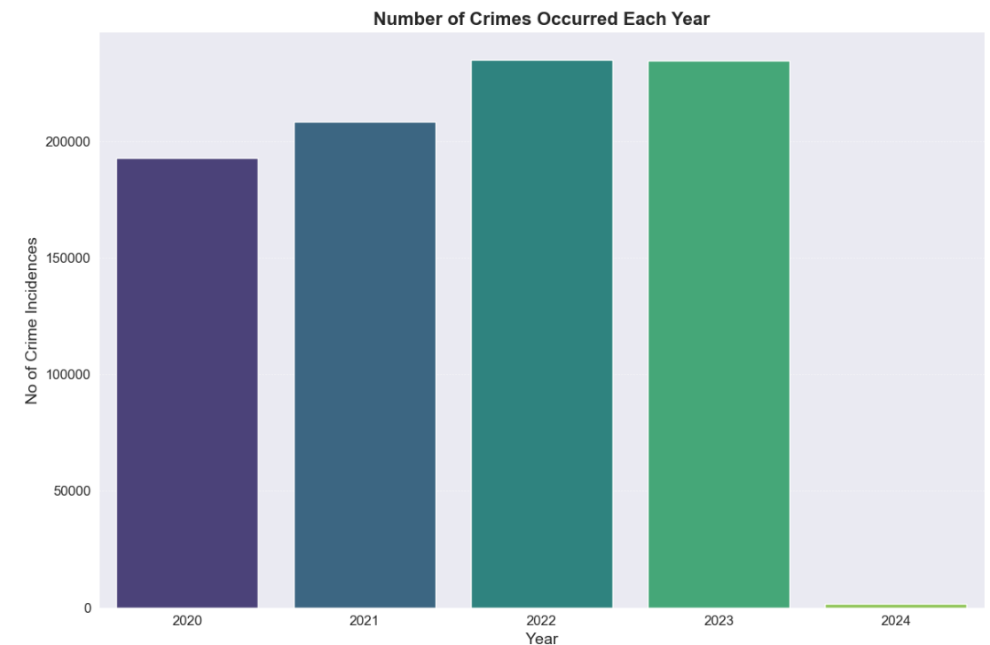
With R2 values close to zero, the models developed to forecast daily identity theft incidents demonstrated minute predictive ability. With an R2 score of just 0.011551, the Gradient Boosting Regressor model had the best performance. This suggests that accurate prediction is not possible with the characteristics and preprocessing methods in place. Other relevant factors, such as socioeconomic variables, weather conditions, or different types of crimes, should be included to increase the model's efficacy. Acquire additional datasets or more detailed data that may offer greater forecasting capacity. To better capture the pattern of incidents over time, temporal dependencies using lag features or time series forecasting models should be considered.

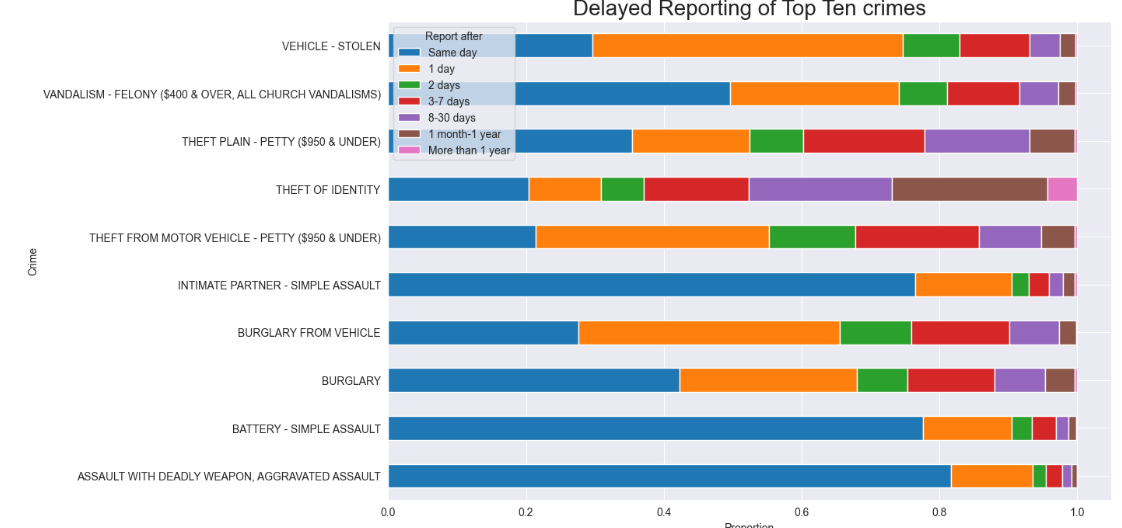
1. References
2. https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about\_data
3. Allison, S. F., Schuck, A. M., & Lersch, K. M. (2005, January). Exploring the crime of identity theft: Prevalence, clearance rates, and victim/offender characteristics. Journal of Criminal Justice, 33(1), 19–29. https://doi.org/10.1016/j.jcrimjus.2004.10.007
4. Rosenfeld, R., & Austin, J. (2023). The Future of Crime in Los Angeles and the Impact of Reducing the Prison Population on Crime Rates. *CrimRxiv*. https://doi.org/10.21428/cb6ab371.b2179eae
5. Appendices

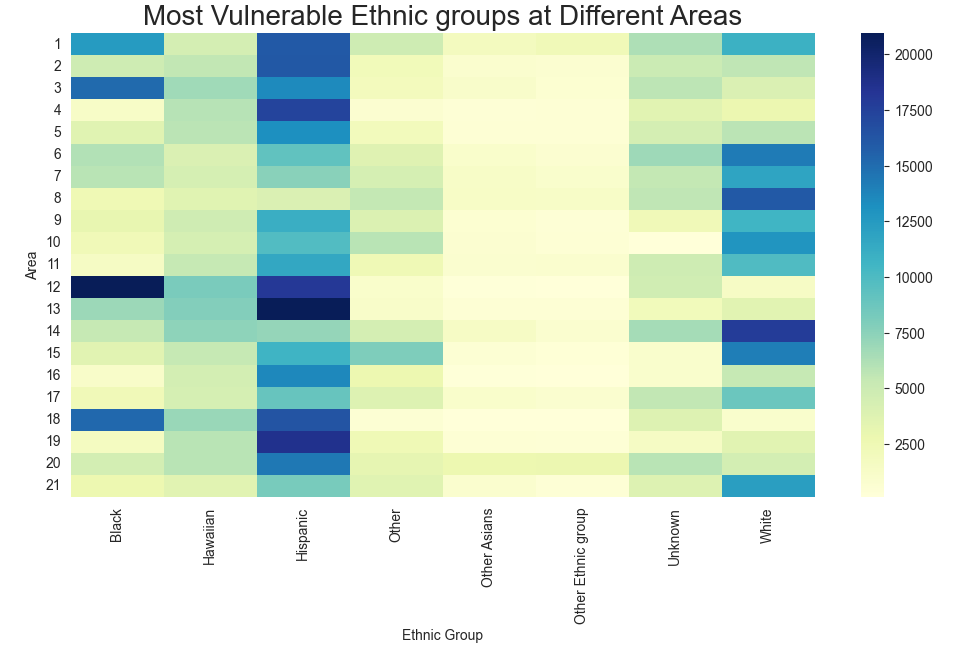
**Fig. 7.** 

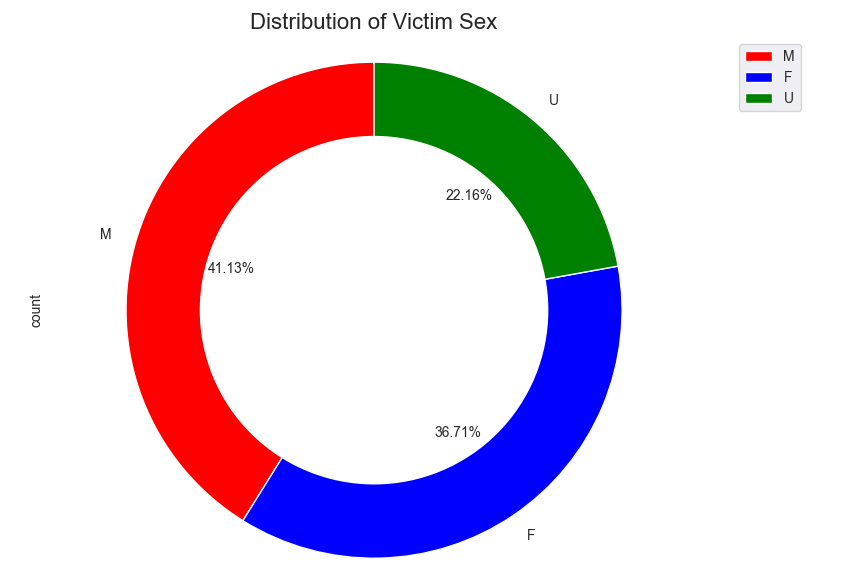
**Fig. 8.** 

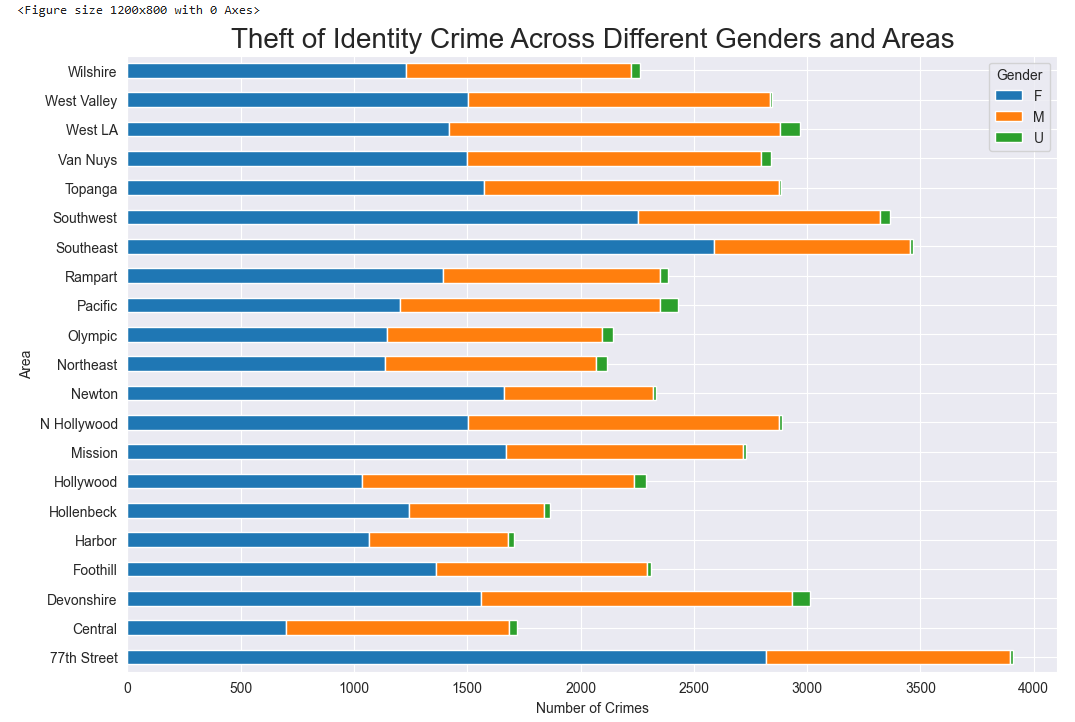
**Fig. 9.** 

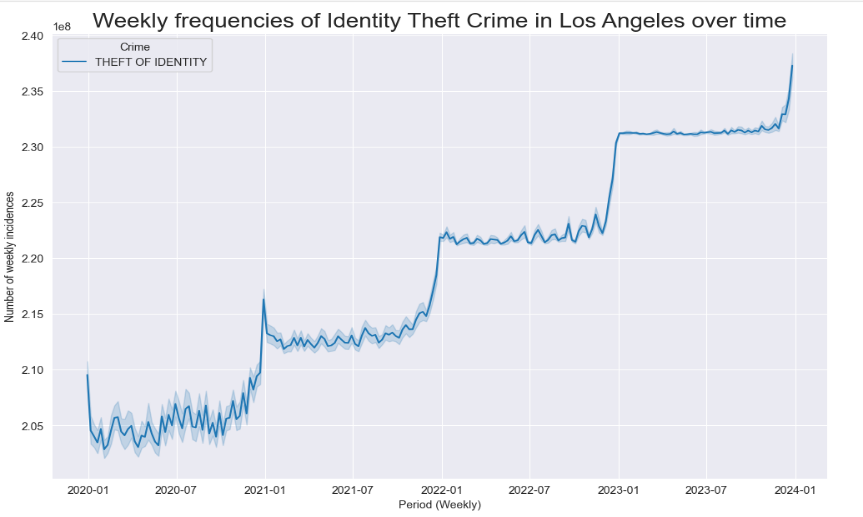
**Fig. 10.** 

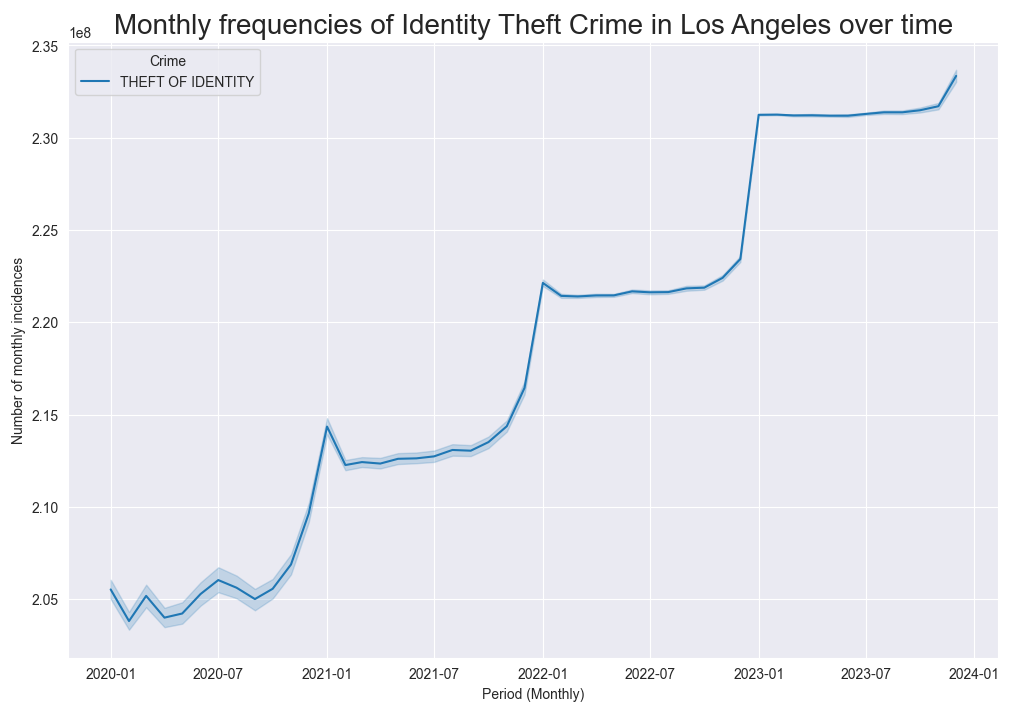
**Fig. 11.** 

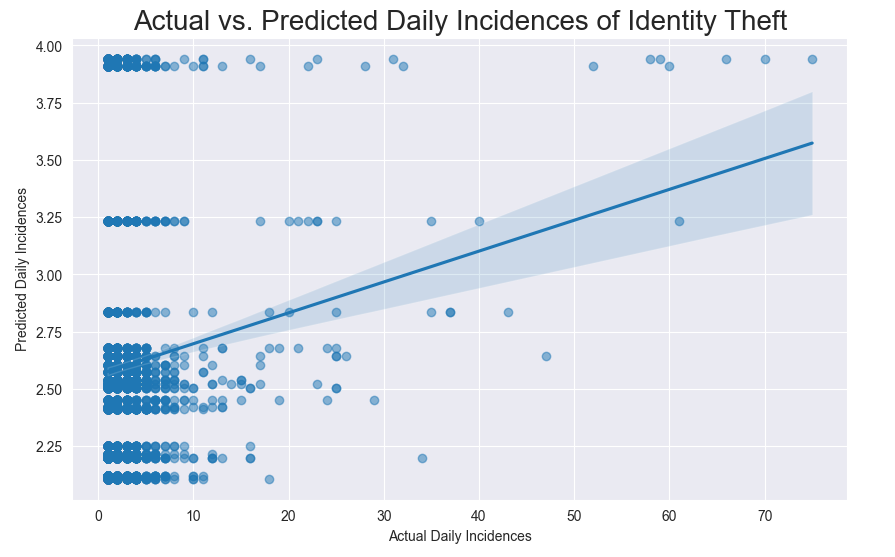
**Fig. 12.** 

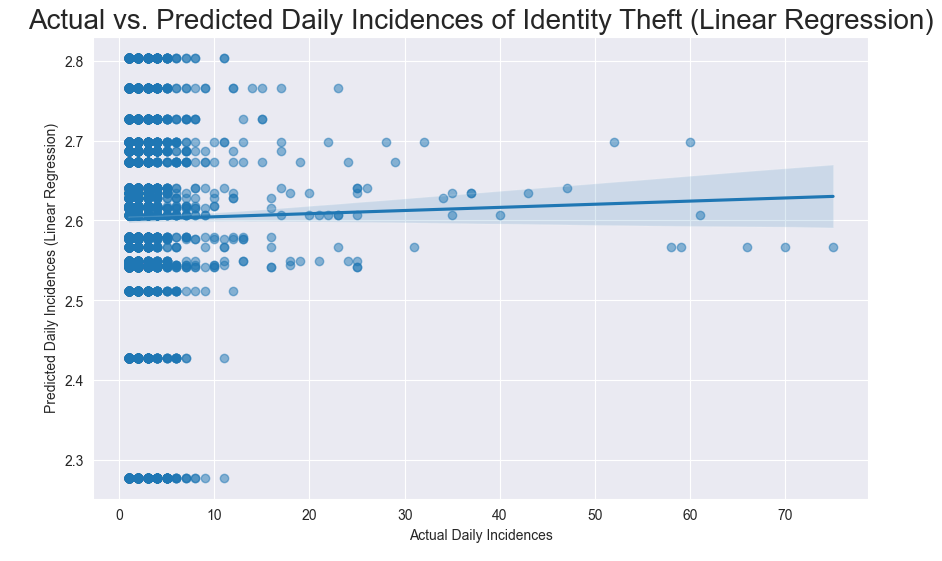
**Fig. 13.** 

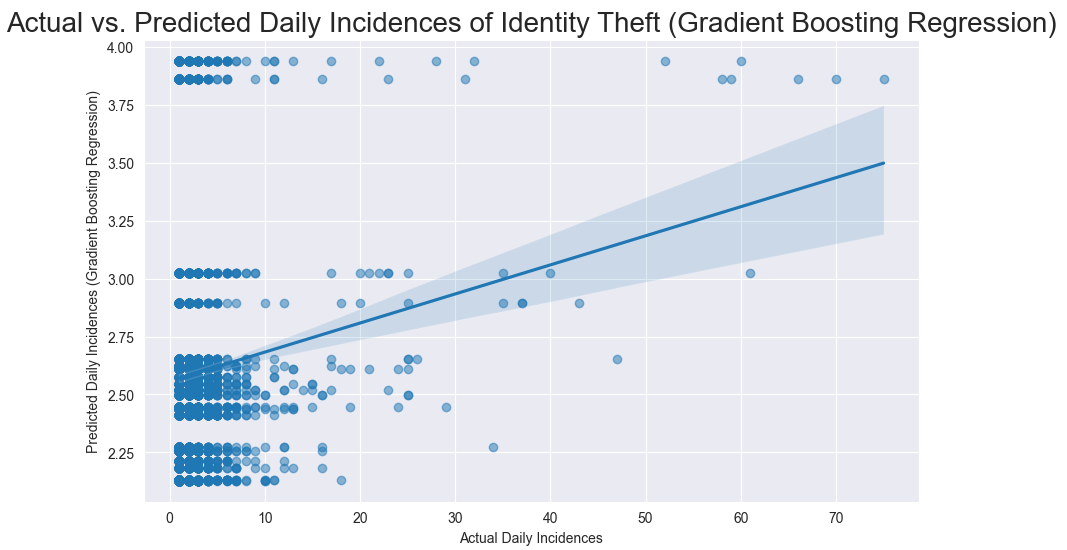
**Fig. 14.** 

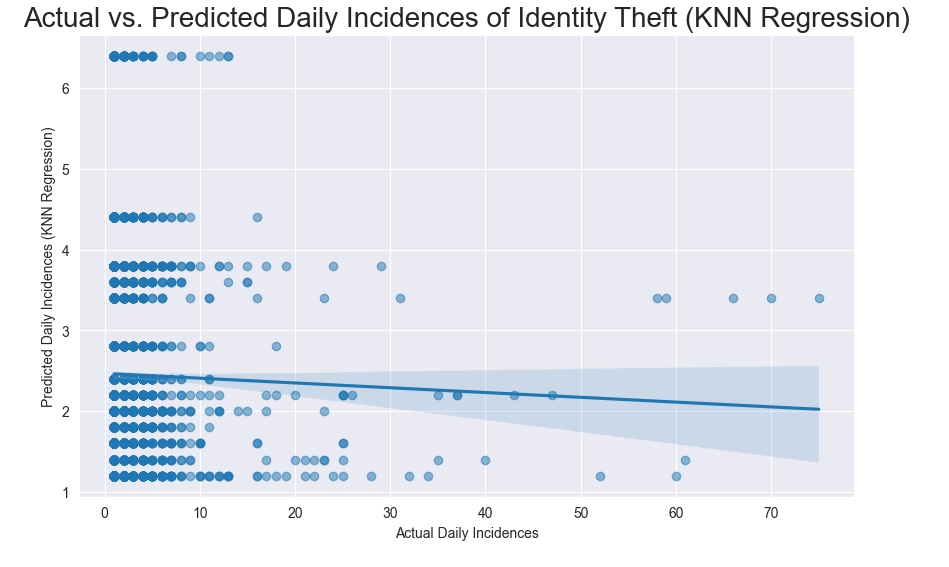
**Fig. 15.** 

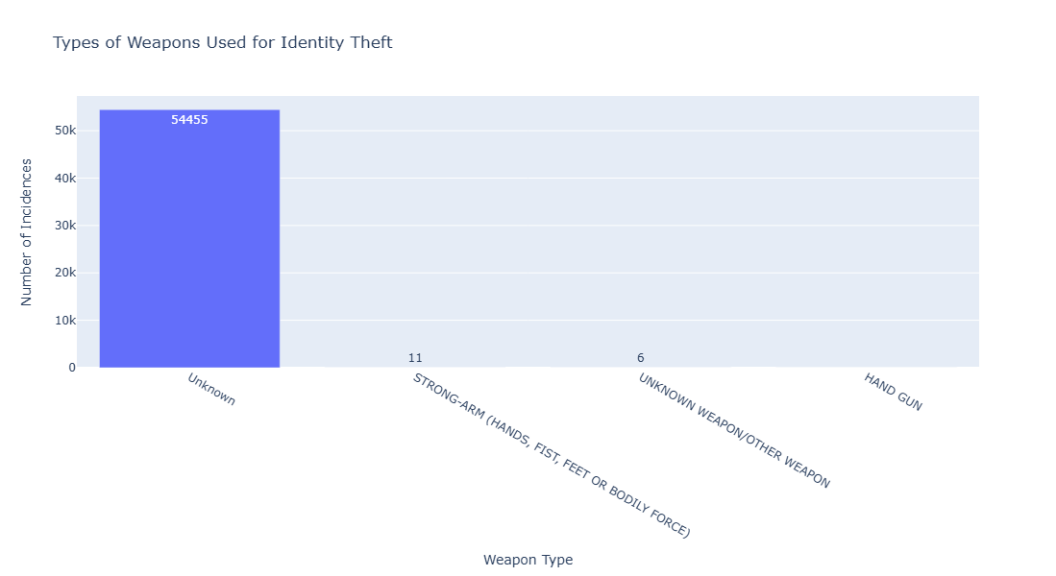
**Fig. 16.** 

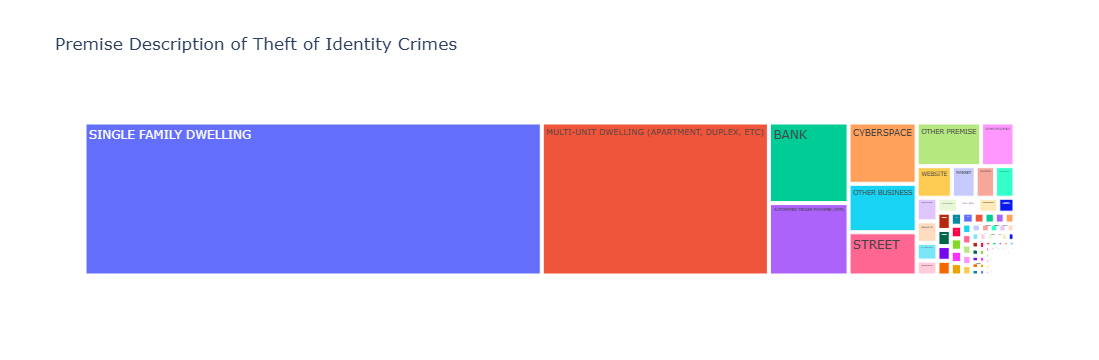
**Fig. 17.** 

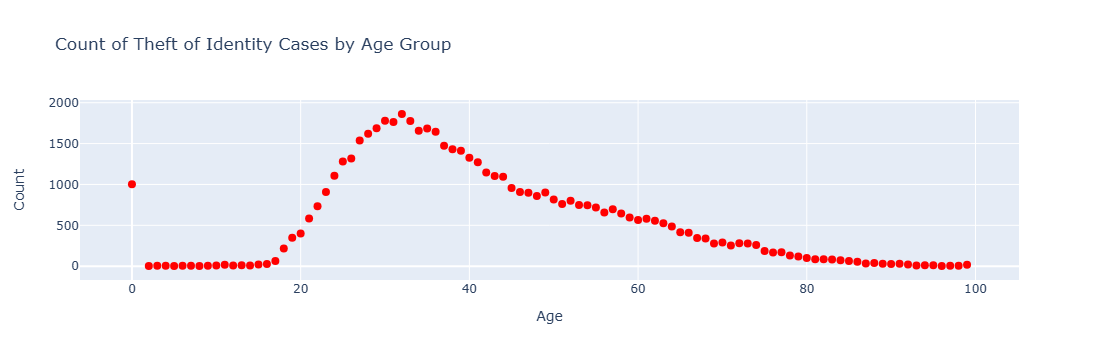
**Fig. 18.** 

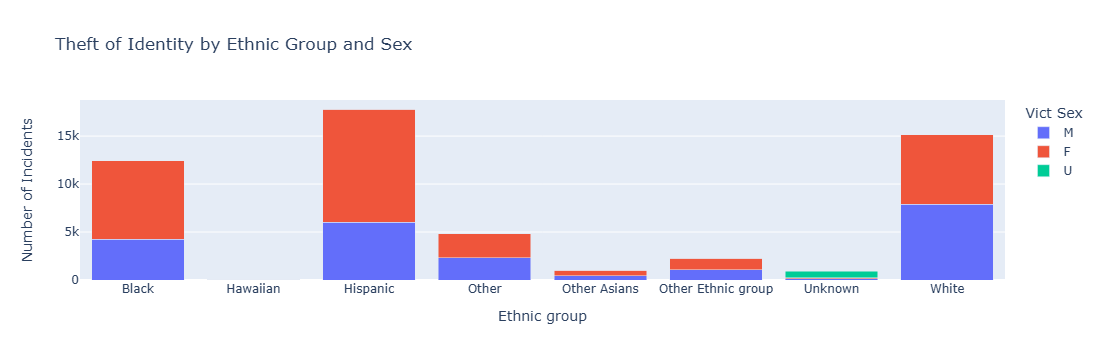
**Fig. 19.** 

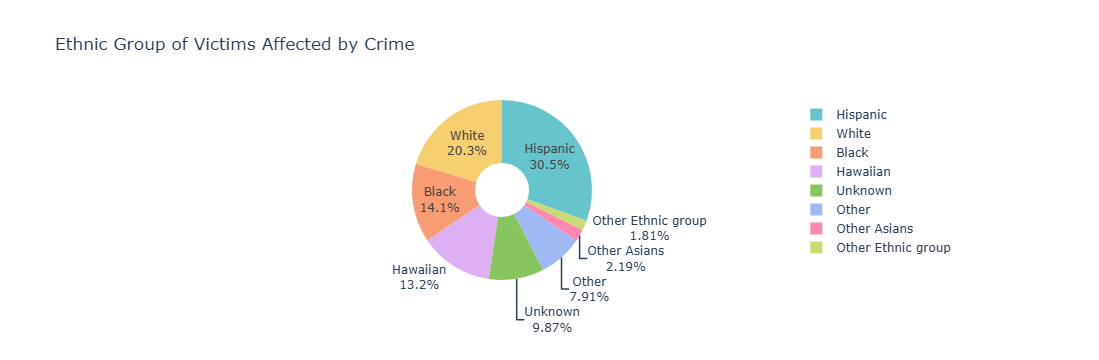
**Fig. 20.** 

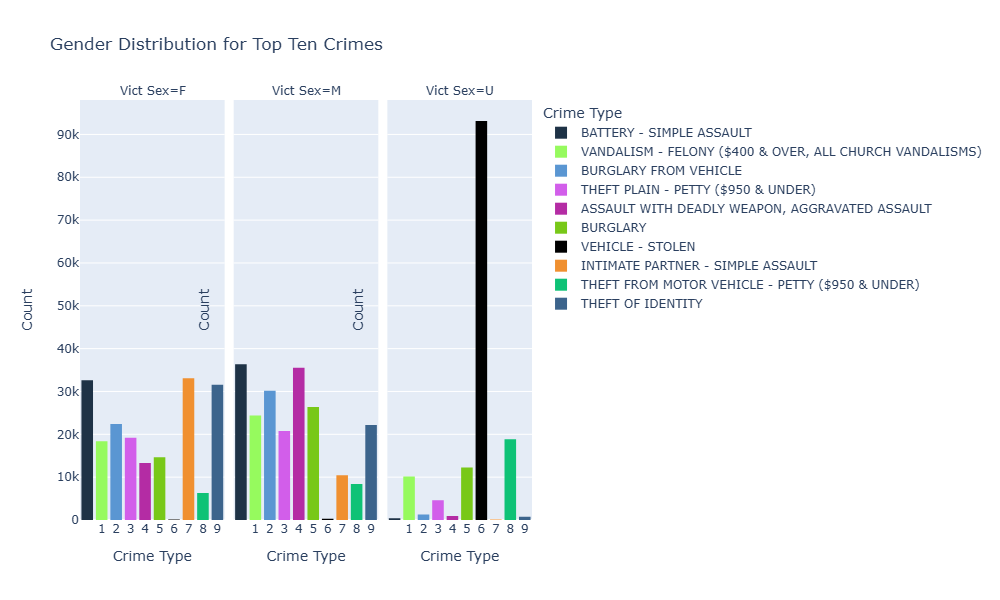
**Fig. 21.** 

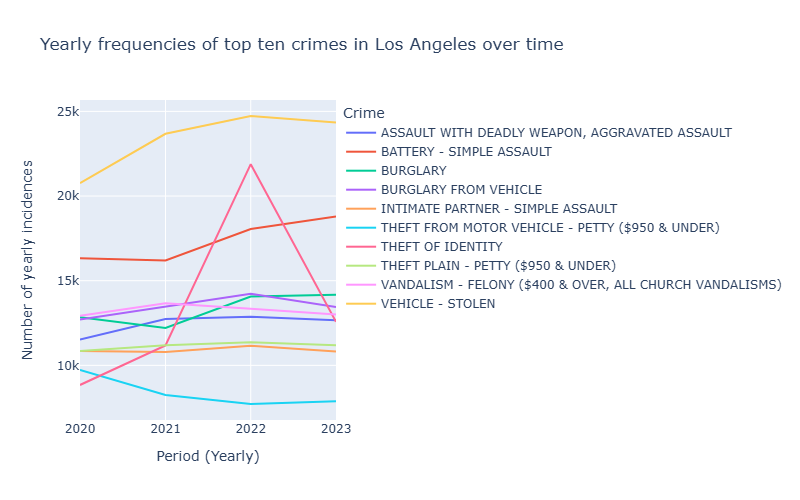
**Fig. 22.** 

**Fig. 23.** 

**Fig. 24.** 

**Fig. 25.** 

**Fig. 26.** 

**Fig. 27.** 

**Fig. 28.** 