UCLA
STATISTICS & DATA SCIENCE

STATS 140XP December 10, 2023

# Smarter Crime Intervention: Predicting Crime Type Using Victim Sex, Location and Time

by

Rita Chau Bonnie Gu Yuki Kitamura Kaylin Lee
William Pan Ellen Wei

**Abstract:**

The LA Crime dataset under investigation is sourced from the Los Angeles Police Department (LAPD). Covering incidents from 2020 to the present, the dataset offers detailed information on crime types, severity, date, time, and location within the city. Our research question focuses on predicting crime types and severity based on temporal and geographical parameters, as well as identifying primary victims across different crime categories. The analysis is conducted through chi-squared hypothesis testing and a multinomial logistic regression model with a 24% classification accuracy. The test demonstrates a statistically significant association between crime type with area, time of occurrence, and victim sex. Future goals of this study could include an in-depth examination of temporal patterns, such as seasonality and special events, incorporation of external data sources, and alternate spatial clustering techniques. Furthermore, this dataset has potential in sociological research regarding victim demographics and vulnerable populations.

**UCLA – 2023**

# 1 Introduction

The data we analyzed came from the Los Angeles Crime dataset, obtained from the Los Angeles Police Department (LAPD). It constitutes a comprehensive repository of information encapsulating diverse criminal incidents from 2020 to the present. We aim to dissect the complexities of crime patterns in Los Angeles using details including crime types, severity, temporal nuances, and geographical specifics. As we investigate, we will employ various statistical methods to find patterns and get insights that not only enhance our understanding of crime in Los Angeles but also contribute to the development of effective crime prevention strategies. We address the following research question: Can we predict the type of crime based on when it occurred, the area it occurred in, and to whom it occurred?

# 2 Variables Of Study

To explore our research question, we selected 3 specific features to investigate. We used the crime code to extract the crime committed. There exist multiple crime code columns in the dataset but we chose to focus on the main crime for each instance as many events did not have more than one associated crime. The LAPD has 21 Community Police Stations referred to as geographic areas within the department and these were the areas used to determine location. These geographic areas or patrol divisions were given name designations that reference a landmark or the surrounding community that it is responsible for. For example, 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles. We chose to use area over other location information as it localizes communities. Furthermore, using more detailed information such as street, latitude, and longitude would result in too many categories, rendering any following models difficult to draw meaningful conclusions from.

We used 2 columns in correspondence to time: incident date (month, day, year) and incident time (in 24-hour military time). We also included victim sex as a categorical feature in our analysis.

The full dataset has 28 columns and 853 thousand rows where each row is a crime incident. The columns can be divided into 6 categories: location, crime type, time, demographic, crime details, and documentation.

**Location:** Location (specific address), cross street, latitude, longitude, area, the corresponding area name, and a district number which is a four-digit code that represents a sub-area within a geographic area.

**Crime type:** Crime code, the corresponding crime code description, crime codes 2,3,4 which contain codes for additional crimes less serious than the first listed crime, and part 1-2 (classification of Part 1 offense or Part 2 offense). Part 1 offense classifications include criminal homicide, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, arson, human trafficking (commercial sex acts), and human trafficking (involuntary servitude). Part 2 offenses encompass all other reportable classifications outside those defined as Part 1. These include other assaults, forgery and counterfeiting, fraud, embezzlement, stolen property, vandalism, weapons (carry, possessing, etc.), prostitution, sex offenses, drug abuse violations, gambling, driving under the influence, liquor laws, disorderly conduct, vagrancy, all other offenses.

**Time:** Date reported, date occurred, time occurred (in 24-hour military time)

**Demographic:** Victim age, sex, descent

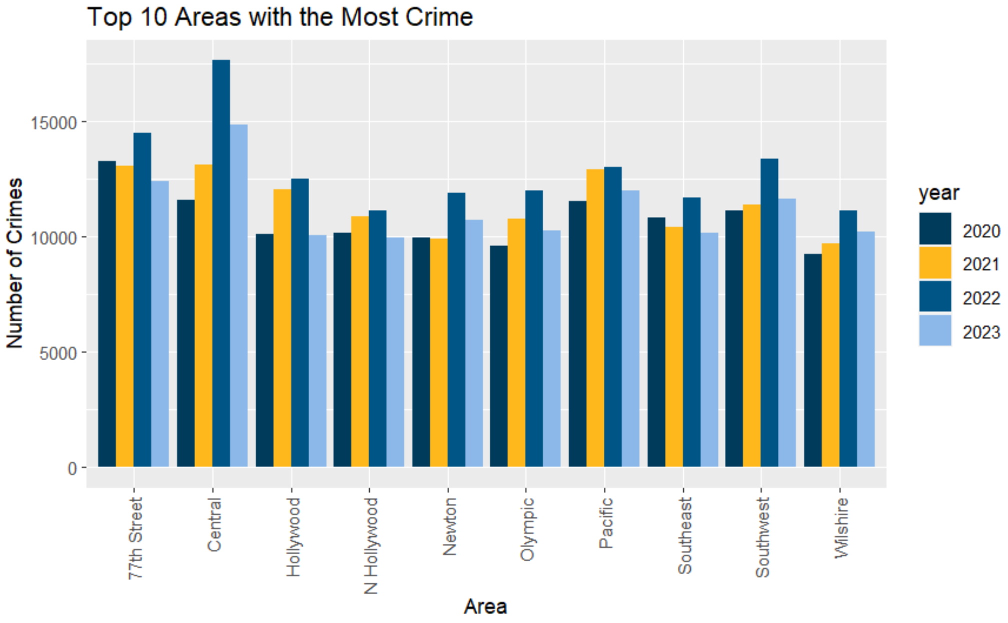
**Crime details:** Modus operandi (activities associated with the suspect in the commission of the crime), premise code (the type of structure, vehicle, or location where the crime took place), the corresponding premise definition, weapon used code, the corresponding weapon description

**Documentation:** Division of records number (official file number), the status of the case, the corresponding status description

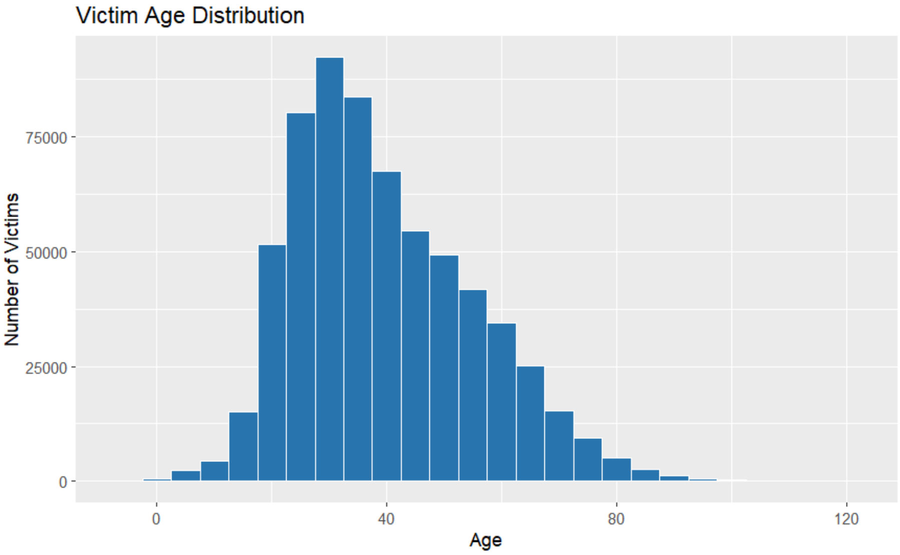
# 3 Descriptive Statistics

Number of unique values for each column:

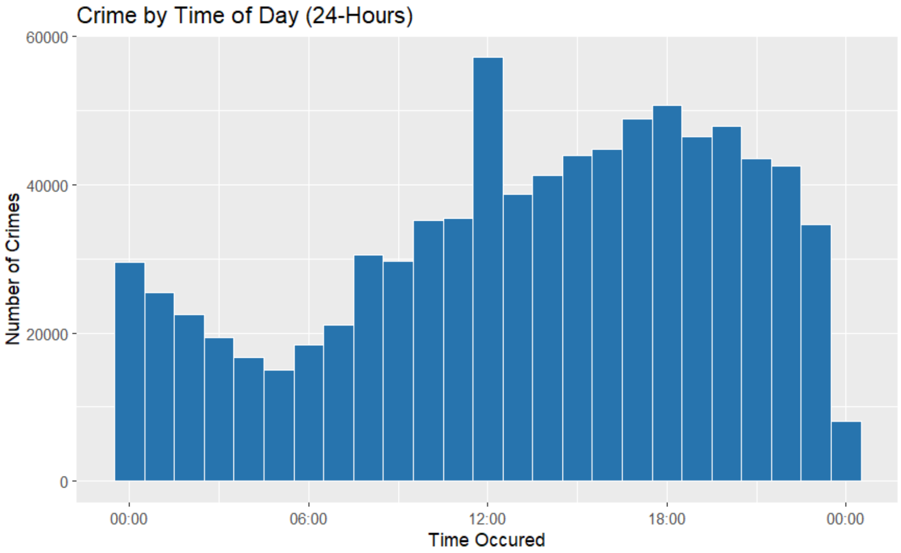
| Variable | Unique Values |
|---|---:|
| date_rptd | 1427 |
| date_occ | 1427 |
| time_occ | 1439 |
| area_name | 21 |
| rpt_dist_no | 1206 |
| crm_cd | 138 |
| crm_cd_desc | 138 |
| mocodes | 281423 |
| vict_age | 103 |
| vict_sex | 6 |
| vict_descent | 21 |
| premis_desc | 307 |
| weapon_desc | 80 |
| status | 6 |
| crm_cd_1 | 141 |
| crm_cd_2 | 121 |
| crm_cd_3 | 38 |
| crm_cd_4 | 7 |
| location | 64227 |
| cross_street | 9817 |
| lat | 5407 |
| lon | 4971 |



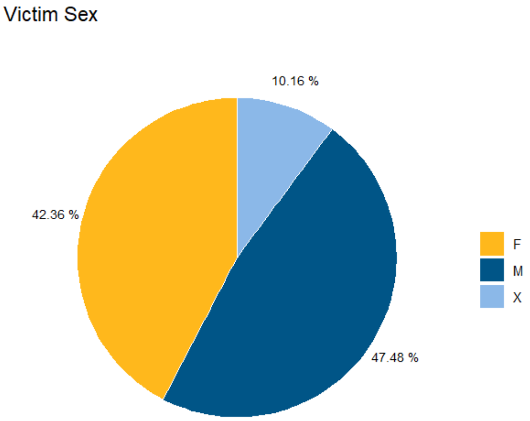
This is a comparative bar chart that illustrates the top 10 areas with the highest crime rates from 2020 to 2023. Notably, Central consistently records the highest number of crimes throughout these years, indicating a stable pattern in crime distribution. 2022 stands out as the year with the highest incidence of crimes across all areas.



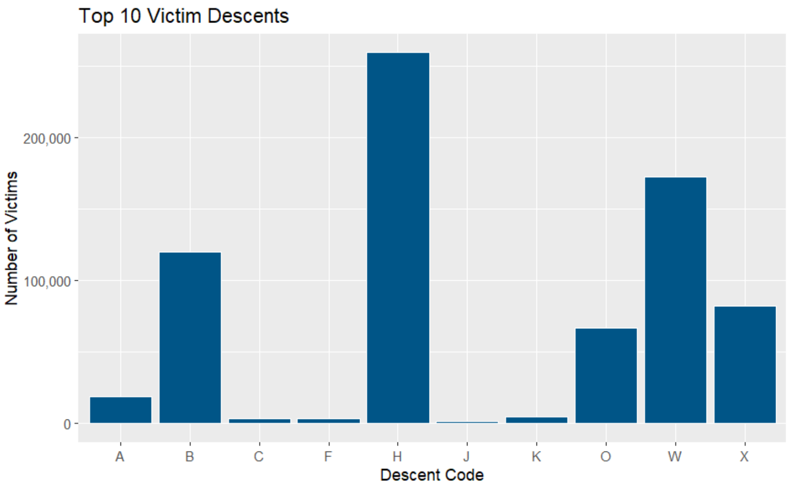
This bar chart illustrates the distribution of victims' ages ranging from 0 to 99. The distribution exhibits a slight right skew, suggesting that victims are mainly young adults, with the most common age range falling between 20 and 50 years.



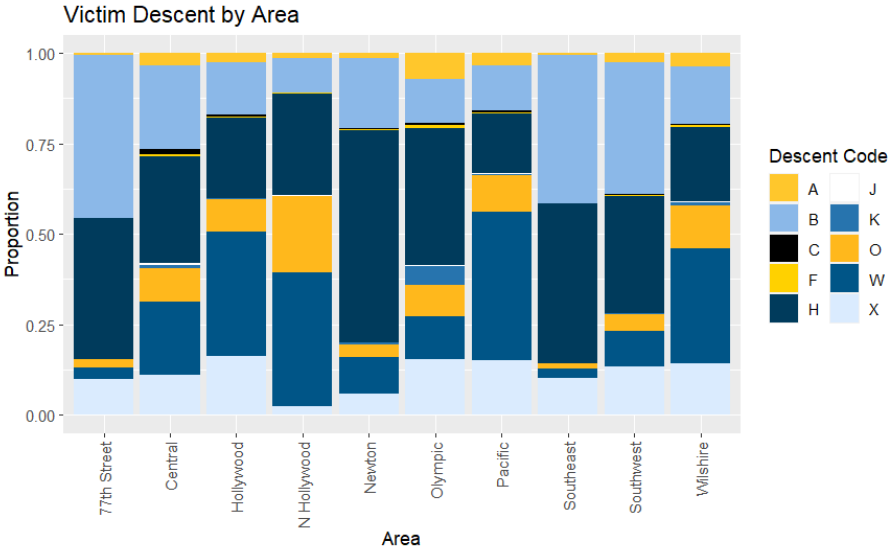
This bar chart illustrates the timing of crimes committed throughout the day. Crime rates peak at noon (12 - 1 PM) and in the evening. Conversely, there seems to be a notable decrease in the number of crimes recorded in the late night (11 PM - 12 AM) and early morning (4 - 7 AM).



The pie chart above shows the proportion of victims' sex, with F, M, and X representing female, male, and unknown, respectively. Male victims represent a plurality of victims, but the union of both non-male categories represents a majority of the victims.



The plot shows the top ten frequency of victim descents. The number of Hispanic victims is higher than any other race. The frequency is followed by White, Black, Unknown, and Other Pacific Islanders. Other victim descents have a fairly similar number of victims, which is significantly lower than Hispanic, White, and Black.



This plot shows the proportion of victim descent in each area. The sum of Black and Hispanic tends to have a higher proportion more than 50% in most of the area. Similarly, White seems to be the second highest proportion of victim descent. When the proportion of Black and Hispanic is less than 50%, the proportion of White is higher than 30%, which is more than double of general proportions of White in each area.

# 4 Experiment / Modeling

## 4.1 Hypothesis Testing

Our research question aims to explore the factors that can be utilized to predict the type of crime. Initially, we hypothesized a correlation between the time a crime occurs, the specific area in which it takes place, and the sex of the victim. In commencing our research, our primary objective was to investigate whether these three factors are strongly associated with the type of crime observed.

Type of Crime vs. Area It Occurs In

$H_0$: There is no significant association between the type of crime and the area it occurs in

$H_a$: There is a significant association between the type of crime and the area it occurs in

We performed the Chi-Square test and obtained X-squared = 190592, df = 2740, p-value $< 2.2e-16$. The results indicate that the p-value is less than 0.05. Therefore, we reject the null hypothesis. This leads us to the conclusion that there exists a statistically significant association between the type of crime and the area in which it occurs. This suggests that the specific geographic location is a relevant factor in predicting the type of crime.

Type of Crime vs. Time It Occurs $H_0$: There is no significant association between the type of crime and the hour at which it occurs

$H_a$: There is a significant association between the type of crime and the hour at which it occurs

We performed the Chi-Square test and obtained X-squared = 237446, df = 3151, p-value $< 2.2e - 16$. The analysis shows that the p-value is below the 0.05 threshold. Therefore, we reject the null hypothesis. This leads us to the conclusion that a statistically significant association exists between the type of crime and the hour at which it transpires. This suggests that the time of day is a meaningful factor in predicting the type of crime.

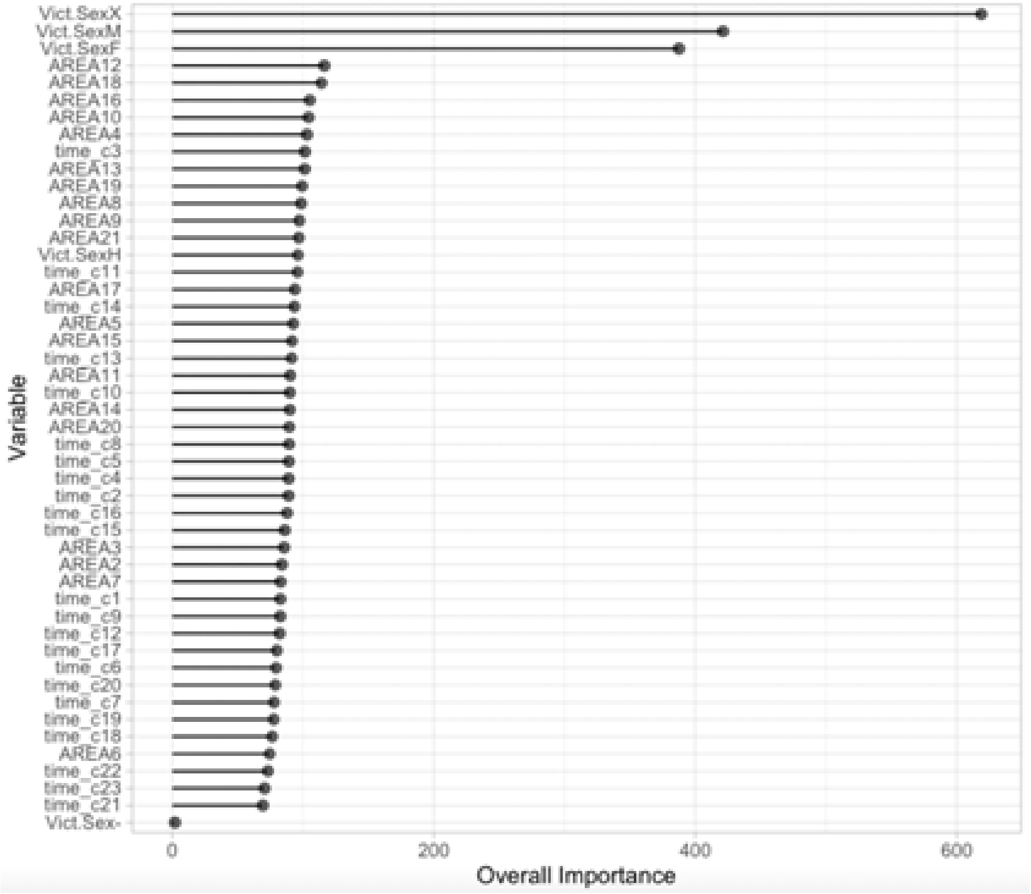
Type of Crime vs. Victim Sex $H_0$: There is no significant association between the type of crime and victim sex

$H_a$: There is a significant association between the type of crime and victim sex

We performed the Chi-Square test and obtained X-squared = 1267789, df = 685, p-value $< 2.2e - 16$. The result shows that the p-value is smaller than 0.05. Therefore, we reject the null hypothesis. This leads us to the conclusion that there is a statistically significant association between the type of crime and victim sex. This suggests that the sex of the victim is a useful factor in predicting the type of crime.

## 4.2 Classification Model

We constructed a classification model with types of crime as target and area, time, and victim sex as predictors. For the predictor time, we converted it into hours. We split the data randomly with 80% training and 20% testing. Then, we fitted a multinomial logistic regression model and the model achieved an accuracy of around 24% on both training and testing sets, which was considerably better than a naive baseline that would make classification based on the most frequent crime type (around 10%). The contribution of each predictor was also evaluated through the caret package in R, where the sex of the victim was shown to be the most important.



# 5 Conclusion

The findings from the Chi-Square tests and the multinomial logistic regression model evaluated the relationship between various factors and the type of crime. The results of the Chi-Square tests indicate a statistically significant association between the type of crime and all the following variables: the geographical area, the time at which it occurs, and victim sex. This emphasizes the importance of considering the geographical context, time of crime occurrence, and victim sex when predicting and understanding criminal activities.

The results from fitting a multinomial logistic regression model present more insights into the important predictors for predicting crime types. Our multinormal regression model achieved both training and testing accuracy of around 24%, which is higher than usual crime prediction (10%). The analysis of statistically significant components of the model reveals the victim sex as the most important factor.

In summary, while the Chi-Square tests emphasize the significance of geographic location, time of day, and victim sex in predicting the type of crime with the same p-value, the multinomial logistic regression model highlighted victim sex as the most important variable. While all three factors are important for crime prediction, further exploration and refinement of models focusing on the victim's sex may be needed to enhance predictive accuracy.

# 6 Challenges of the Study

While this study can provide important insights about general predictors of crime in Los Angeles, the conclusions of this study are merely exploratory and should serve as a motivation for further research. We want to highlight some of the challenges of this study and the implications that this may have.

1. *Limited data availability*

   Despite the abundance of data present in this dataset, a significant challenge to the generalization of this study emerges due to the lack of information on certain key features containing essential information necessary for making more concrete and nuanced conclusions. For example, while we can conclude the association between crime type and area, there is not enough background information about the overall demographics and geography of each area to control for external variables that may also have an impact on the predominant crime occurrences of certain areas. This limitation significantly hindered our ability to conduct a comprehensive analysis and draw robust conclusions about the causes of certain crime patterns and critical aspects of the Los Angeles population.

2. *Time Constraints*

   One of the notable challenges encountered during the research was the limitation imposed by time constraints, particularly in seeking input from experts in the domain. The compressed timeline of the study restricted our ability to engage with sociologists, law enforcement organizations, and other primary sources who could provide valuable insights and guidance on the intricacies of crime patterns in Los Angeles. For example, we could refine methods of imputation and control for the unbalanced dataset to further refine our predictive model, but we sacrificed precision for generalization due to the lack of expert input on these refinement techniques. The importance of expert consultation in expanding domain knowledge and validating findings cannot be overstated.

# 7 Recommendations for the Future

As above, we have provided valuable insights into crime dynamics from 2020 to the present, exploring the LA crime dataset provided by the Los Angeles Police Department (LAPD). As we conclude our analysis of how to predict crime types and severity based on temporal and geographical parameters and identify key victims, we would like to briefly outline specific recommendations for some future research.

During our analysis, we found that an in-depth examination of temporal patterns, such as seasonal changes, holidays, and special events, related to crime could greatly increase our understanding of criminal dynamics. Creating a sub-model for distinct periods and highlighting temporal anomalies by incorporating external data sources would not only give law enforcement a better understanding of criminal dynamics but would also provide a more nuanced approach to predicting and responding to fluctuations in criminal activity.

Although the current research has focused on a broader geographic area, we would also recommend ways to explore the use of distance-specific analysis or spatial clustering techniques in specific areas. This is because these microscopic spatial resolutions can identify regional centers of crime and with further consultation, support law enforcement in more precise interventions.

Furthermore, exploring advanced machine learning techniques such as neural networks, or recurrent neural networks to further refine the predictive models, further from the analytical methods we used, will also be able to create more detailed predictive models, and optimize model performance, especially by incorporating characteristic engineering and hyper-parameter tuning, is expected to improve the predictive accuracy of crime types and severity.

In addition to predictive modeling, we would like to further highlight the potential of this dataset in sociological research, particularly concerning victim demographics and studying vulnerable populations. By cross-referencing data from external sources, there is great potential in intersectional analysis that reveals the overlapping impact of multiple demographic factors on crime and victimization.

Applying these specific recommendations to future research efforts will provide a more sophisticated, accurate, and ethical understanding of the criminal dynamics of Los Angeles. This will support the development of effective crime prevention strategies to create safer areas.