Los Angeles Crime Analysis

Group 13: Titus(tra29), Revanth(rg361), Suim(sp699), Meixiang(md480)

1. Abstract

This analysis delves into the intricate dynamics of crime in Los Angeles, employing an extensive dataset from the Los Angeles Police Department (LAPD), covering incidents from 2020 to present. Utilizing Logistic and Poisson regression models, our research aims to unearth the pivotal factors influencing the seriousness and frequency of crimes. Key findings indicate significant temporal and spatial variations in crime rates, offering valuable insights for law enforcement and public safety strategies. This study not only enhances the understanding of crime patterns but also aids in resource allocation and preventive measures.

2. Introduction

Crime, a complex and multifaceted social issue, demands meticulous investigation to inform and enhance public safety measures. This report capitalizes on a detailed dataset from the Los Angeles Police Department, encompassing a vast array of **815,882** crime reports. Recorded from 2020 to the present day, the dataset offers a rich tapestry of information, ranging from location specifics and victim demographics to crime types and investigative outcomes (Los Angeles Police Department, 2023). Each entry in the dataset is a unique amalgamation of several variables, including latitude, longitude, area, demographic details of the victim, and more, offering a comprehensive view of the crime landscape in Los Angeles.Central to our analysis are two pivotal research questions: First, we explore the determinants impacting the severity of crimes, a query crucial for preventive strategies and public awareness. Second, we aim to predict crime frequencies in specific areas and timeframes, a task vital for strategic law enforcement deployment and community safety initiatives. By employing logistic regression, we seek to understand the factors that significantly influence the seriousness of crimes. Concurrently, through Poisson regression modeling, we aim to forecast crime occurrences, accounting for a variety of predictors such as time, date, and geographic factors.Our investigation is not merely an academic exercise but a crucial endeavor to discern patterns and predictors within the urban landscape of Los Angeles. The insights gleaned from this analysis we're hoping can be instrumental in helping law enforcement agencies, policymakers, and the community at large, enabling a data-driven approach to crime prevention and optimized policing efforts. Furthermore, the study contributes to a broader understanding of criminal behavior, aiding in the development of effective community outreach programs and mitigating risk factors associated with crime.

3. Methodology

Since the Dataset used for this project is dynamic in nature, a fixed snapshot of the data as of **13-Oct-2023** was used to prevent any errors or fluctuations of the results.

Exploratory Data Analysis (EDA) was performed to understand the trends and relationship between the crime occurrences and the following factors:

1. Temporal: Time of occurrence, weekday, month
2. Precinct: Number of crimes reported, distance to precinct
3. Victim Demographic: Age, Ethnicity, Gender

4. General: proportion of crime types, daily rate

The following cleaning and processing steps were performed on the data before it was used for the analysis and modelling:

1. Distance to precinct: we calculated using the Latitude (LAT) and Longitude (LON) columns from the dataset and publicly available co-ordinates of the precincts in Los Angeles
2. Data Type Conversion: Columns were converted to their relevant Datatypes e.g: Conversion of Date related columns to Datetime, Victim Sex, Gender as Factors etc.
3. Bucketing: Columns like Time of crime occurrence were bucketed in order to decrease the number of unique levels and avoid over-fitting
4. Imputation: Blank and empty values in certain columns of interest like Victim Gender and Ethnicity were filled or combined with place holder values (e.g. "X") so that they can be used in the modelling phase
5. Time Restriction: Since the data is dynamic in nature and the most recent crimes may not be reported yet, the time-frame for analysis was restricted and we considered only data up to 31-Aug-2023 so that there is no skewing of the results
6. Helper Columns: New columns such as number of days between crime occurrence and reporting were calculated, Month, Weekday etc. were calculated from the columns present in the dataset.

Post the restrcition of the Time-frame for analysis, The number of observations reduced to **794,388**, a reduction of **21,494** (~2.5%) observations.

Research Question 1

The first research question is addressed using logistic regression, which is well-suited for binary outcomes such as 'serious' and 'non-serious'. Logistic regression is particularly effective for inference, allowing us to identify factors that influence the seriousness of crimes committed. Prior to model construction, it was essential to eliminate rows with blank values or "-" for two specific variables: victim sex and victim descent. These missing values were impractical to impute and were replaced with 'unknown' during model configuration, due to the dataset's large size. Subsequently, we removed all 'unknown' values from each variable.

Based on a priori selection, we included factors like time of occurrence, distance to the precinct, demographic information (gender, age, descent), and weapon usage, hypothesizing that these significantly impact crime seriousness. Regarding collinearity, the Variance Inflation Factor (VIF) values for each variable were around 1, indicating stability in the model.

Research Question 2

The focus of our statistical investigation addresses pivotal questions concerning the predictors of crime severity and the forecast of crime occurrences.Our methodology was designed to provide insights into crime patterns and to predict the number of crimes in a given area and time period.

Utilizing **Poisson regression**, an optimal approach for modeling count data, we aim to predict the frequency of crimes in Los Angeles with precision and statistical accuracy, leveraging the model's suitability for such discrete outcome variables. This model allows us to integrate time, date, and location variables, capturing the essence of crime occurrences over various periods and areas. By employing a Poisson regression model, we aim to dissect the frequency of crimes within the multifaceted urban setup of Los Angeles, identifying temporal and spatial hotspots of criminal activity.

Given the prediction nature, we've narrowed down our model to the following variables, and undergone **stepwise forward** selection and **cross-validation** to refine and optimize the predictive performance.

List of Predictor Variables

- Independent Variables

  - Time and Date: Time Occurred, Date Occurred

  - Geographic Factor: Area

- Outcome Variable

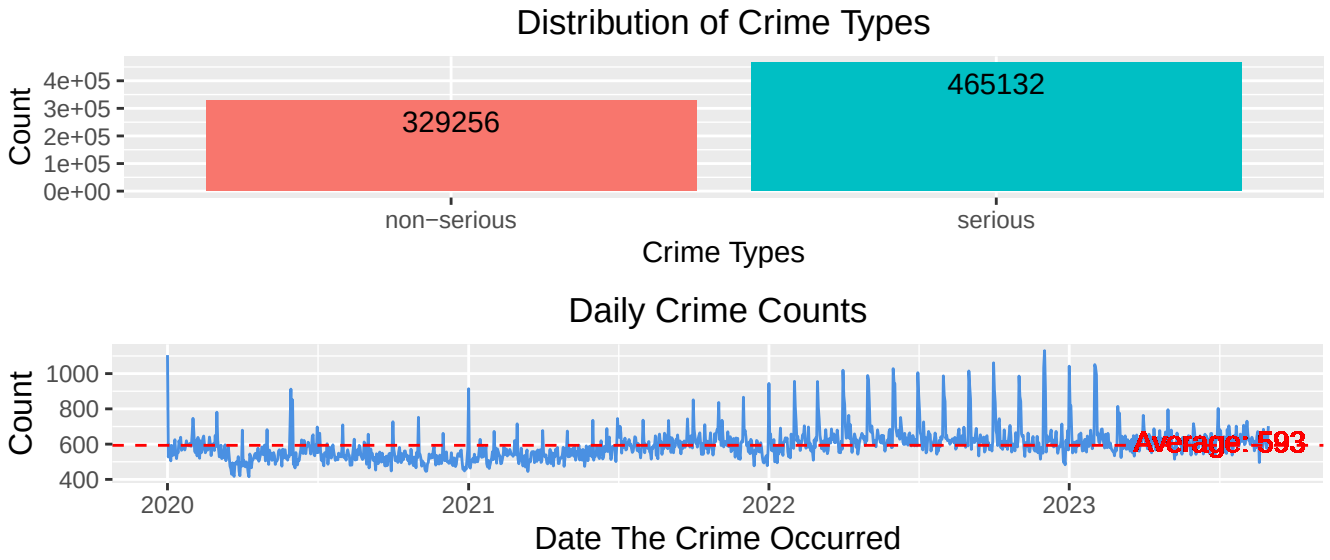
  - Count of Crimes (Generated by grouping the data set by the selected predictor variables)

4. Results

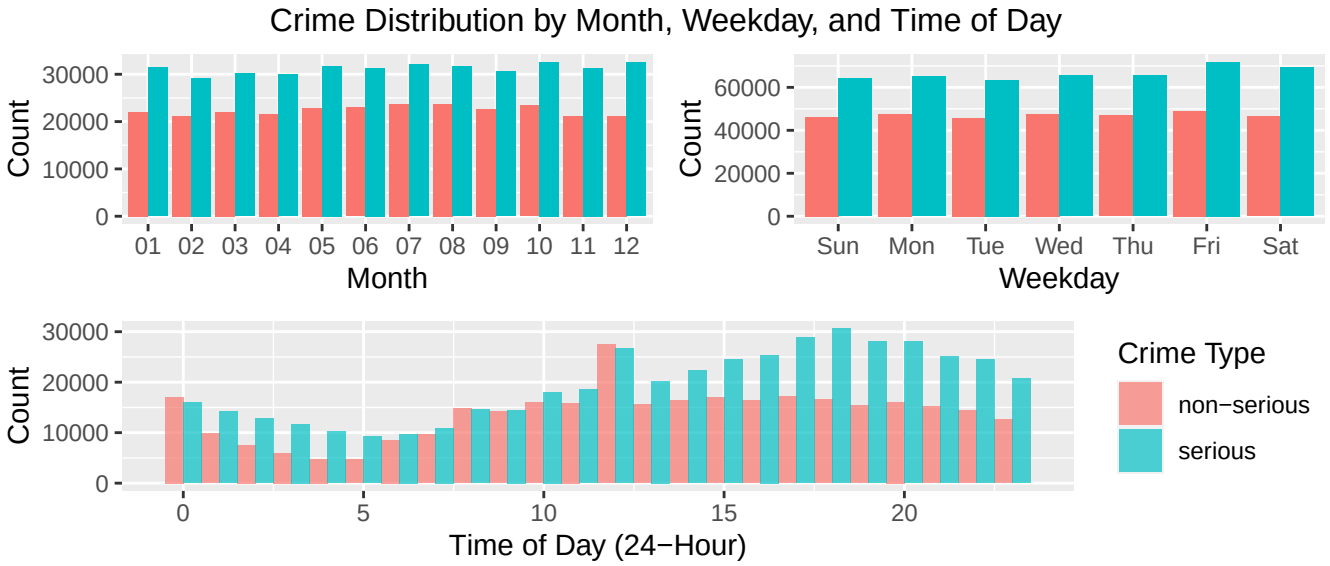
4.1 Exploratory Data Analysis

The following observations were made during the EDA process.

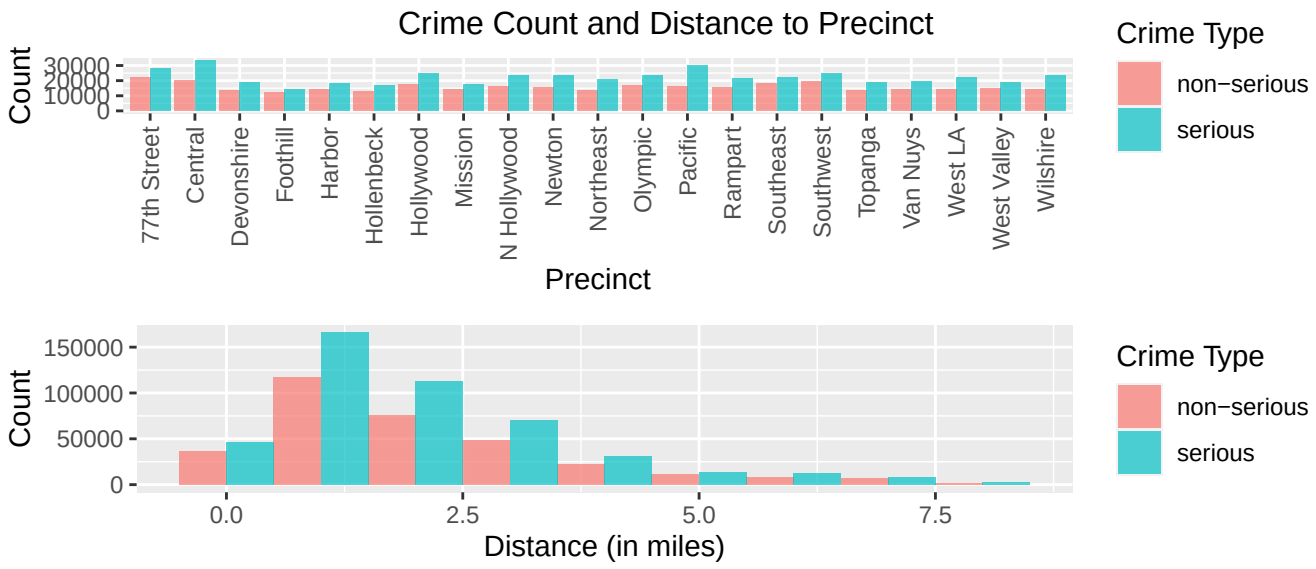
About **60%** of the crimes committed are serious crimes, and on and average **593** crimes occur in a day with with a few days having very high number of crimes.



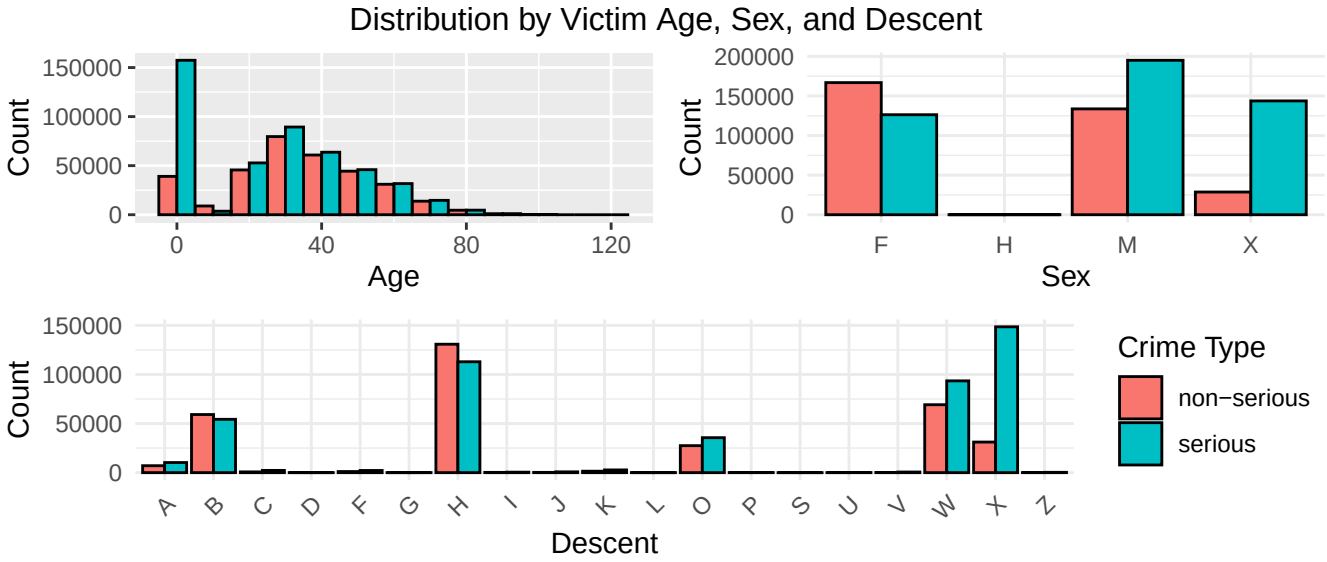
There is a minor variation in the proportion and count of crimes when compared on a monthly and weekday level. The variation is more prominent at the 'Time of Day' level.



There is a minor variation in the proportion and count of crimes at different precincts.



The victim demographic data - 'Age', 'Sex' and 'Descent' have a large number of missing values(systematic). There is a clear observable variation of the number and proportion of the type of crimes with respect to these variables. (Refer to Appendix for Meaning of Letter Codes)



4.2 : Logistic Regression

The model was refined using forward stepwise selection and cross-validation. The goodness of fit was indicated by a null deviance of 829808, leading to a significant model improvement with an AIC of 829856. However, the McFadden's pseudo R^2 value of 0.0367, though low, suggests limited explanatory power.

The model's accuracy was approximately 60.07%, not sufficiently high to be definitive. A Kappa value of 0.1951, while slightly better than chance, also indicates modest performance. Sensitivity and specificity are crucial in evaluating the model. Sensitivity (true positive rate) effectively identifies actual serious cases, while specificity (true negative rate) accurately identifies non-serious cases. These metrics collectively offer a comprehensive view of the model's ability to distinguish between serious and non-serious cases.

The model's insights include:

- Time hours occurred: Night or early morning crimes were more likely to be serious.

- Distance to precinct: Closer proximity to certain precincts correlated with serious crimes.
- Demographic Information: Men and younger individuals were more often victims of serious crimes, with certain racial groups being more susceptible. Among the various descents, individuals of Cambodian, Filipino, Japanese, Korean, Vietnamese, and Asian Indian descent were more likely to be victims of serious crimes compared to other racial groups.
- Weapon Used: The presence of a weapon did not significantly increase the likelihood of a crime being serious. Instead, crimes committed without a weapon had a higher probability of occurring.

Model 2 : Poisson

With forward step wise selection and cross validation methods to improve model, the deviance measures goodness of fit, with a null deviance of 35064 and a residual deviance of 5204, indicating a substantial improvement in model fit, resulting in an AIC value of 15356 with 4 Fisher Scoring iterations.

Our preliminary evaluation of the model's predictive accuracy, gauged through the RMSE, yielded a value of 42.59 to estimate the average difference between the predicted and observed crime counts. Additionally, the R-squared value of 0.82 reflected a strong linear relationship between the observed and predicted crime counts, explaining a substantial proportion of the variance in the crime data.

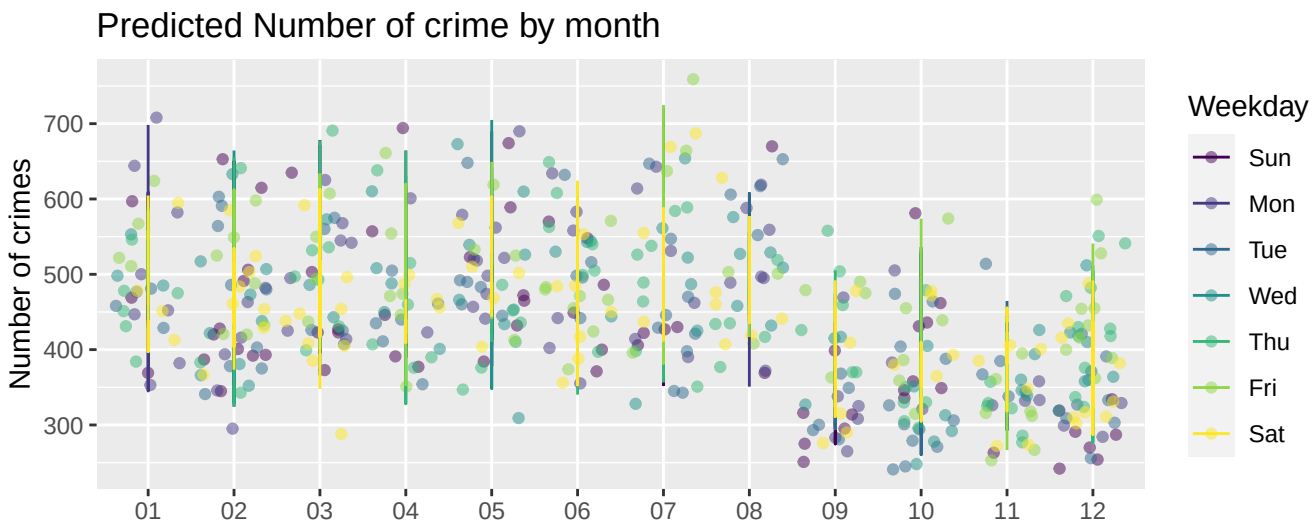
Preliminary examination of the model outputs suggests distinct spatial and temporal crime patterns.

- **Day of the Week Matters:**Different weekdays impact crime differently. For example, crimes are more likely on weekends compared to others weekdays, and these differences are confirmed by the p-values associated with these coefficient.
- **Monthly Trends in Crime:**Crime rates vary by month. Some months, like February and September, see fewer crimes, while others, like July and August, experience more. These patterns are statistically significant.
- **Geographical Differences:**Crime rates differ by police area. Some areas have higher crime rates (e.g., "Southwest"), while others have lower rates (e.g., "Hollywood"). These variations are statistically significant, providing insights into local crime trends.

However, with a dispersion value of 4.07, it indicates potential problems with the model assumptions, likely due to factors such as Population Heterogeneity and Model Misspecification. Zero-Inflation and Correlation Among Observations have been ruled out. To address these issues, consider switching to a negative binomial model for better handling over-dispersion and troubleshoot Population Heterogeneity and Model Misspecification to improve the model.

Plot the predictions alongside observed counts, depicts the pattern of the number of crime varies from month to month weekday and weekends.

We can skip the below part since we have a similar plot in the EDA section, moght help save space


Predicted Number of crime by month
Number of crimes
700
600
500
400
300
01
02
03
04
05
06
07
08
09
10
11
12
Weekday
Sun
Mon
Tue
Wed
Thu
Fri
Sat