Los Angeles Crime Analysis

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

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 day. Utilizing Logistic and Poisson regression models, our research aims to unearth the pivotal factors influencing the seriousness of crime commited and predict the number crime occurrences in a given area and time respectively. Key findings from the first research objective indicate a relationship between factors like victim demorgraphic and time of day on the seriousness of crime committed. Findings from the second research objective 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(LAPD), encompassing a vast array of reported crimes recorded from 2020 to the present day and updated by the LAPD on a `weekly` basis. The dataset offers a rich tapestry of information, offering a comprehensive view of the crime landscape in Los Angeles.

As of **13-Oct-2023**, the Dataset has **815,882** observations of **28** variables and the information provided can broadly be classified into the following categories:

1. `Location`: Latitude, Longitude, Area, Street, District
2. `Victim Demographic`: Age, Gender, Ethnicity
3. `Crime Description`: Type of Crime, Investigation Outcomes, Weapon Usage
4. `Date and Time`: Date Reported, Date Occurred, Time Occurred
5. `Identifier/Classifier`: Crime Record Identifier, Mocodes

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 is dynamic in nature and updated on a weekly basis, a fixed snapshot of the data as of **13-Oct-2023** was used throughout the research process to prevent any errors or fluctuations of the outcomes and results.

3.1 Data Cleaning & EDA

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

1. Distance to precinct: using the Latitude (LAT) and Longitude (LON) columns from the dataset and publicly available co-ordinates of the precincts in Los Angeles we calculated the distance between the spot of crime occurrence and the precinct it was reported in, to understand if there was a relationship between the distance and the seriousness of crime.
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: Empty values (systemic) 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 for analysis and modelling
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, 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.

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

3.2 Research Question 1

Question: **"What are the strongest indicating factors that influence the seriousness of crime committed"**

The **binary** outcome variable "crime type" was created by converting the "Part 1 2" column from the original dataset which indicates the seriousness of crime committed, the following mapping was performed 1-"serious", 2-"non-serious" and the result was saved as as a factor.

Logistic Regression was chosen for this analysis since it is well-suited for binary outcomes such as 'serious' and 'non-serious'. Logistic regression is also particularly effective for inference since the output is easily interpretable, 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 combined with the "X"-unknown category from the dataset. for Victim sex, the type "H" was excluded from the analysis since it had very few (88) observations comparatively.

In our a priori selection, factors like time of occurrence, distance to the precinct, demographic information (gender, age, descent), and weapon usage were included, hypothesizing that these significantly influence crime seriousness. We also explored the interaction term with sex and descent, based on the hypothesis that specific genders and descents might be interrelated. However, the model with the interaction term did not yield statistically significant results, leading to its removal from the final model. Additionally, several assessment methods were considered for fitting the model to the dataset.

Firstly, regarding collinearity, we chose to use the Variance Inflation Factor (VIF) to assess collinearity within the model. We also employed deviance and the Akaike Information Criterion (AIC) to find the most fitting model by comparing two alternative models, excluding victim sex and descent. Based on the relationship between crime seriousness and 1, we evaluated the model's fit using McFadden's pseudo R^2 value. Additionally, we assessed the model using the ROC curve and confusion matrix through prediction methods. These approaches helped us determine the extent to which the model fits the dataset and its effectiveness in addressing the research question.

3.3 Research Question 2

Question: **"What is the predicted number of crimes for a given area and time period?"**

Each row of our dataset is an individual crime incident, to get the count of crimes which is required for answering this research question, we aggregated the data at a Location(Area)-Month-Weekday level and got the count of crimes at this level.

Poisson regression was chosen for this analysis as it is an optimal approach for modeling count data and the output can be interpreted easily. This model allows us to integrate time, date, and location variables, capturing the essence of crime occurrences over various periods and areas.

Given the prediction nature, we've narrowed down our model to the following variables:

- Predictor Variables: Month, Weekday, Area
- Outcome Variable: Count of crimes at the predictor variable level

To refine and optimize the model, **cross-validation** and **stepwise forward** selection was used

To assess model fit, the data was split into Train and Test and the model was used to predict the number of crimes on Test dataset, metrics such as RMSE and R^2 were studied to evaluate the performance of the model

4. Results

4.1 Exploratory Data Analysis

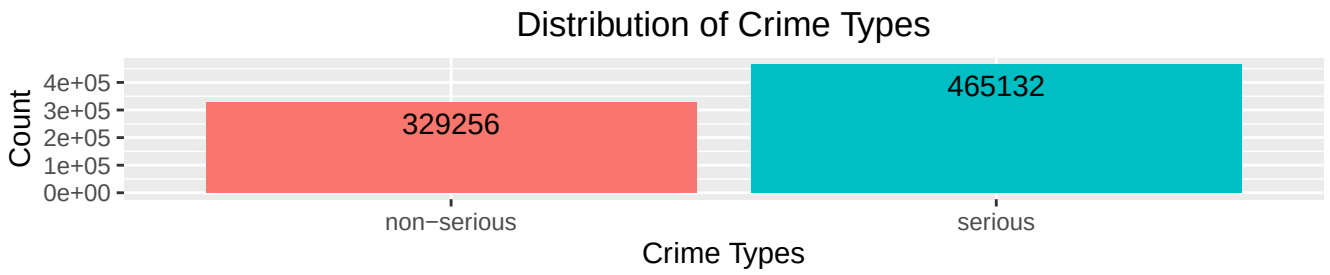
The following observations were made during the EDA process.

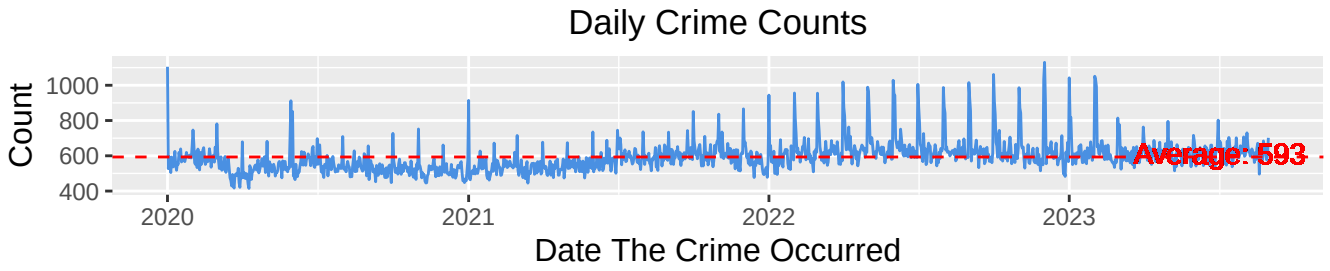
About **60%** of the crimes committed are serious crimes, and on and average **593** crimes occur in a day with with a few days having very high number of crimes.

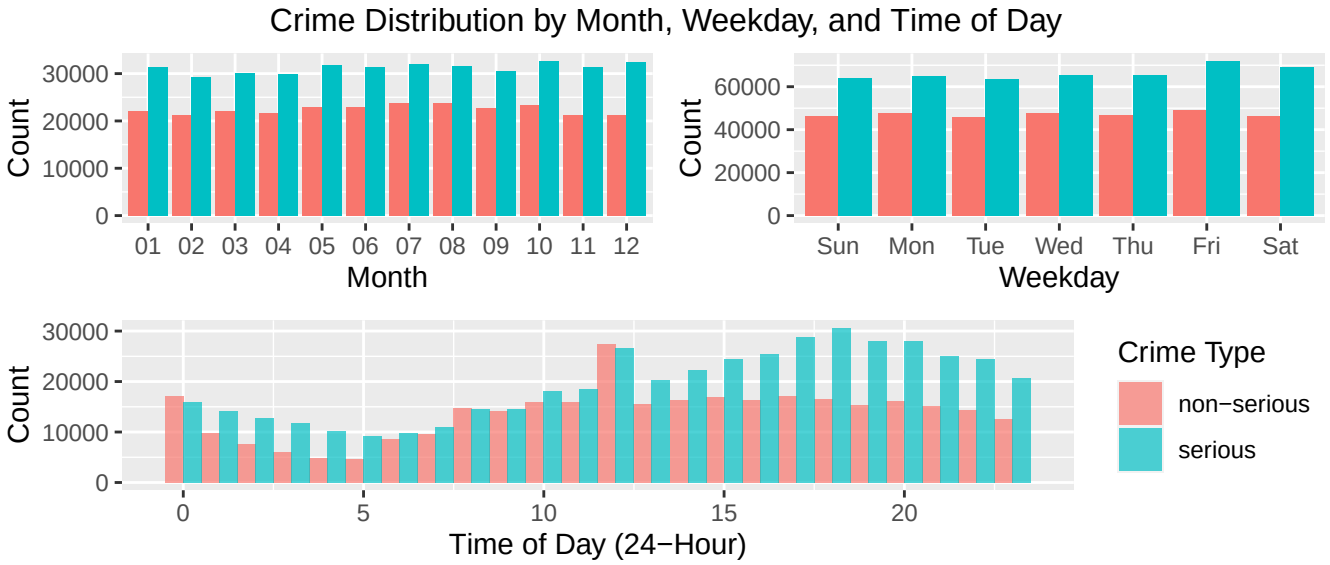
There is a minor variation in the proportion and count of crimes when compared on a monthly and weekday level. The variation is more prominent at the 'Time of Day' level.

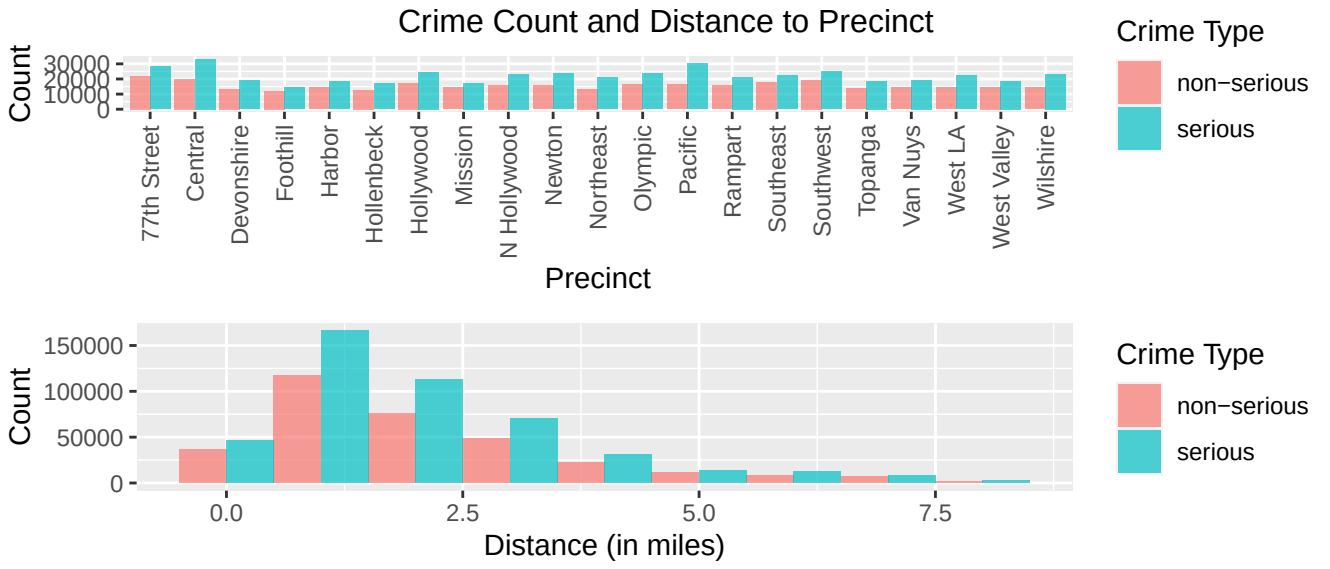
There is a minor variation in the proportion and count of crimes at different precincts.

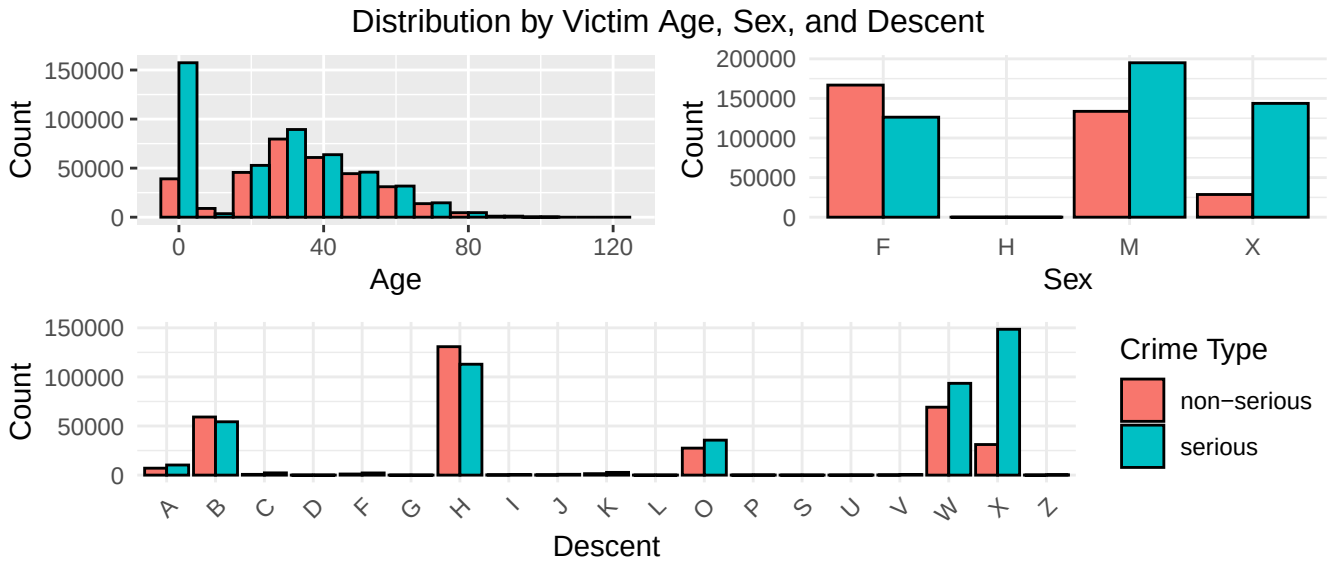
The victim demographic data - 'Age', 'Sex' and 'Descent' have a large number of missing values(systematic). There is a clear observable variation of the number and proportion of the type of crimes with respect to these variables. (Refer to Appendix for Meaning of Letter Codes)


Distribution of Crime Types
Count
4e+05
3e+05
2e+05
1e+05
0e+00
329256
465132
non−serious
serious
Crime Types


Daily Crime Counts
Count
1000
800
600
400
Average: 593
2020
2021
2022
2023
Date The Crime Occurred


Crime Distribution by Month, Weekday, and Time of Day
Count
30000
20000
10000
0
01 02 03 04 05 06 07 08 09 10 11 12
Month
60000
40000
20000
0
Sun Mon Tue Wed Thu Fri Sat
Weekday
0 5 10 15 20
Time of Day (24−Hour)
Crime Type
non−serious
serious


Crime Count and Distance to Precinct
Count
30000
20000
10000
0
77th Street
Central
Devonshire
Foothill
Harbor
Hollenbeck
Hollywood
Mission
N Hollywood
Newton
Northeast
Olympic
Pacific
Rampart
Southeast
Southwest
Topanga
Van Nuys
West LA
West Valley
Wilshire
Precinct
Crime Type
non−serious
serious
150000
100000
50000
0
0.0
2.5
5.0
7.5
Distance (in miles)
Crime Type
non−serious
serious



4.2 Logistic Regression

The Logistic Regression model was fit on the selected columns as mentioned earlier and it gave the following (selected) `exponentiated` parameters as the output:

	Coefficient Estimates	Standard Errors	p-values	Confidence Interval (2.5%, 97.5%)
(Intercept)	1.1932	1.798e-02	< 2e-16 ***	(1.1519, 1.2361)
Time hours occurred	1.0159	3.704e-04	< 2e-16 ***	(1.0151, 1.0166)
Victim Age	0.9965	1.513e-04	< 2e-16 ***	(0.9962, 0.9968)
Victim Sex: Male	1.8827	5.271e-03	< 2e-16 ***	(1.8634, 1.9023)
Victim Descent: Japanese	1.7731	7.185e-02	1.56e-15 ***	(1.5423, 2.0443)
Victim Descent: White	0.8616	1.659e-02	< 2e-16 ***	(0.8340, 0.8900)
Distance to precinct	0.9998	5.826e-06	< 2e-16 ***	(0.9998, 0.9999)
Weapoon: Used	0.5995	5.269e-03	< 2e-16 ***	(0.5933, 0.6057)

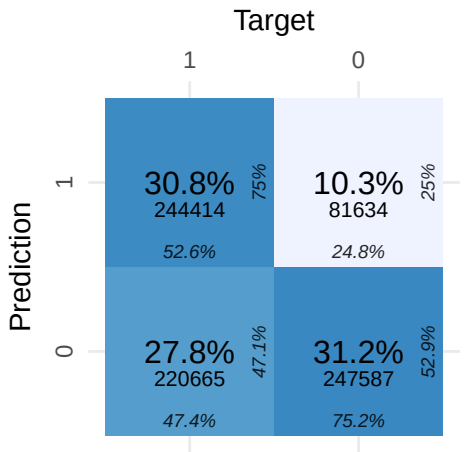
The coefficients above represent the odds of particular category or unit increase in value being a victim of serious crime as compared to the reference level for that category. E.g. A Male is 1.88 times more likely to be a victim of a serious crime as compared to a female.

The model was compared with other models and null-model to study interactions and asses the model.

The goodness of fit was indicated by a null deviance of 980739.3, leading to a significant model improvement with an **AIC** of 980789.3. However, the McFadden's pseudo R^2 value of 0.09004, though low, suggests limited explanatory power.

The VIF score of most of the parameters was less than 3 and only the Victim Descent was above 10.

A confusion matrix was created and the following metrics were studied to evaluate the model:



The following table briefly describes the results of the confusion matrix:

Metric	**Value**
Accuracy	0.6187
95% Confidence Intervals	(0.6177, 0.6198)
Kappa	0.2611
Sensitivity	0.5217
Specificity	0.7558

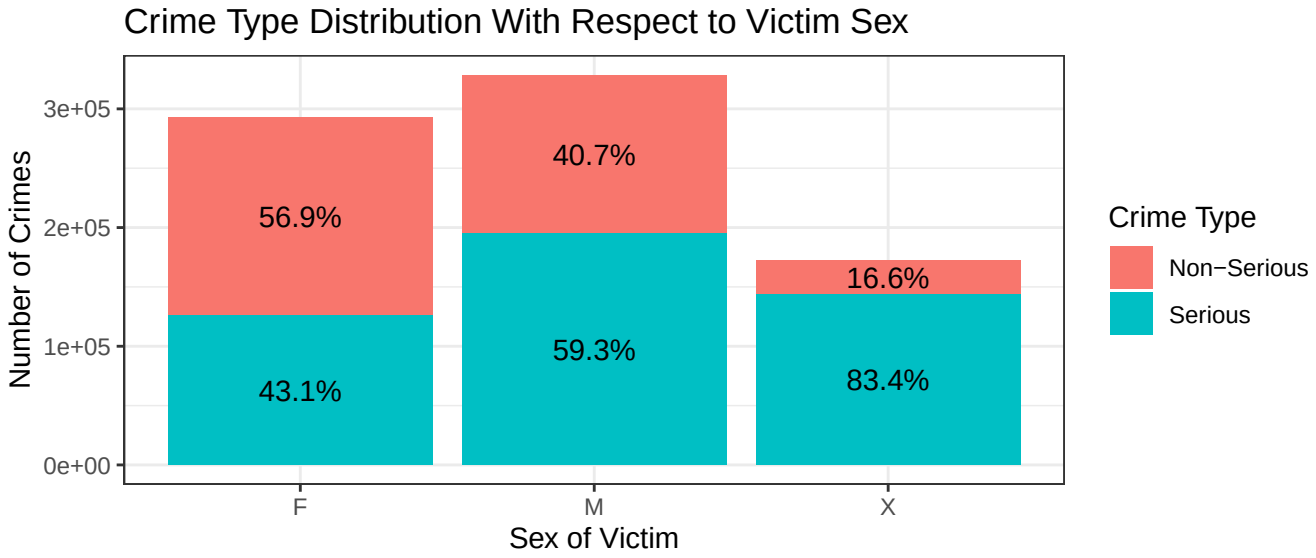
The model's accuracy was approximately 61.87%, not sufficiently high to be definitive. A Kappa value of 0.2611, while slightly better than chance, also indicates modest performance. Sensitivity and specificity are crucial in evaluating the model. Sensitivity (true positive rate) effectively identifies 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.

These results provide insights into the factors that influence the occurrence of serious crimes. While many variables show statistical significance, the study does have its limitations. To enhance the model, incorporating additional variables or exploring new datasets could be beneficial. A major challenge lies in managing the 'Unknown' categories within Demographic Information, which is essential for improving the study's accuracy and relevance. In our modeling configuration, we have removed all such values from the dataset.

The below the graph establishes one of the outputs of the model that Males are more likely to be a viticm of violent crime as compared to females:



4.3 Poisson Model Regression

The Logistic Regression model was fit on the selected columns as mentioned earlier and it gave the following (selected) `exponentiated` parameters as the output:

	Coefficient Estimates	2.50% Confidence Interval	97.50% Confidence Interval	P Values
(Intercept)	663.6024	655.4907	671.7956	0.00E+00
Weekday: Saturday	0.9705	0.9638	0.9773	0.00E+00
Month: 02	0.9399	0.9283	0.9517	0.00E+00
Month: 08	1.0163	1.0046	1.0281	6.20E-03
Area: Hollywood	0.8362	0.8236	0.8491	0.00E+00
Area: Southwest	0.8898	0.8769	0.9029	0.00E+00

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.

The model was made to predict the count of crimes on the Test dataset and these values were compared against the original count of crimes and the following metrics were used to evaluate the model:

Metric	**Value**
RMSE	42.59
R-Squared	0.82

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.

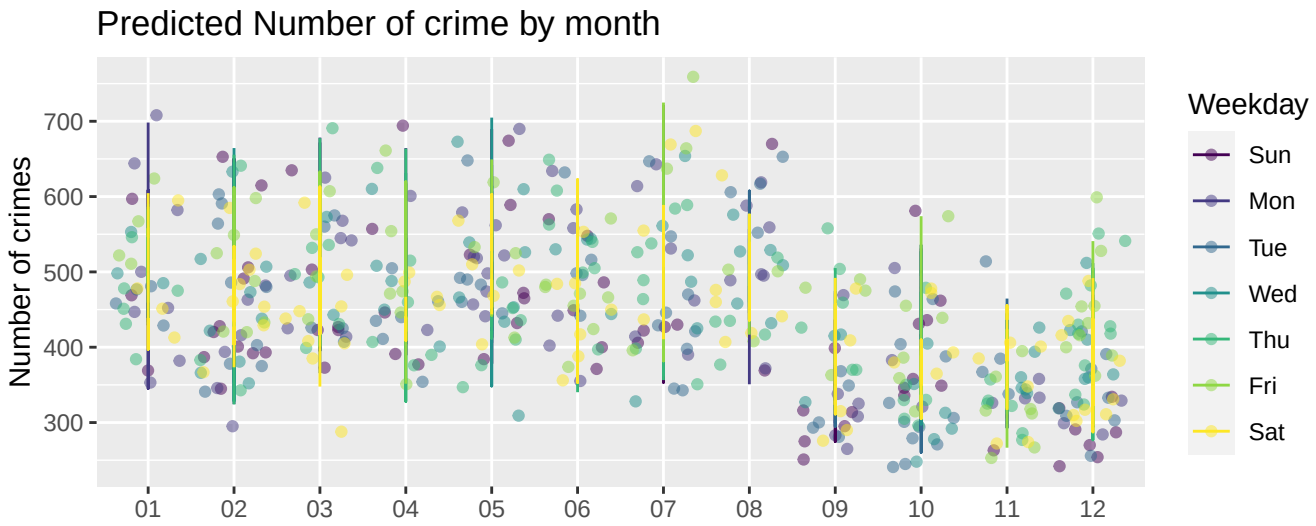
Preliminary examination of the model outputs suggests distinct spatial and temporal crime patterns.

- **Day of the Week:** 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.
- **Month:** 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.
- **Area:** 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.

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



5. Conclusion

Our analysis of Los Angeles crime data using Logistic and Poisson regression models reveals significant insights while also highlighting key limitations and future research directions. The Logistic model uncovers that crimes during night or early morning hours, proximity to certain precincts, and victim demographics are critical factors in determining crime severity. Surprisingly, weapon usage did not significantly influence the seriousness of a crime. Despite these insights, the model's limited explanatory power, as reflected in its accuracy and McFadden's pseudo R2 value, suggests the need for further refinement.

The Poisson model, on the other hand, highlights temporal and spatial variations in crime patterns. Weekdays and certain months exhibit distinct crime trends, and geographical disparities in crime rates across police areas offer valuable insights for targeted law enforcement efforts. However, issues with model assumptions and over-dispersion necessitate consideration of alternative models like negative binomial regression in future studies.

These findings, while insightful, must be viewed in the context of the study's limitations. The challenge of managing 'Unknown' categories in demographic data, potential model misspecification, and the need for more comprehensive variables are areas for future improvement. This study's endeavor to understand and predict crime patterns in Los Angeles aligns with our initial motivation to aid public safety strategies and crime prevention. Moving forward, enhancing model accuracy and incorporating broader datasets could provide deeper insights into the dynamics of crime, ultimately contributing to more effective public safety planning and community welfare.