

Deciphering Crime Dynamics in Los Angeles

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 committed and predict the number of crime occurrences in a given area and time respectively. Key findings from the first research objective indicate a relationship between factors like victim demographic 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](#) (Los Angeles Police Department, 2023) 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 modeling:

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 modeling
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, Week-day etc. were calculated from the columns present in the dataset.

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

We additionally conducted checks for missing values, particularly in the columns of interest recognizing that missing data could introduce bias or affect the models' robustness. Wherever the missing data was systematic or substantial, they were handled appropriately.

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 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.

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. We evaluated the model's fit using McFadden's pseudo R² 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 R² were studied to evaluate the performance of the model.

4. Results

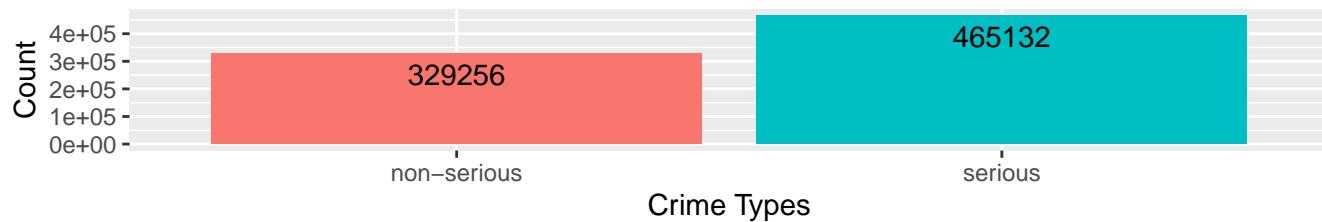
4.1 Data Cleaning & EDA

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

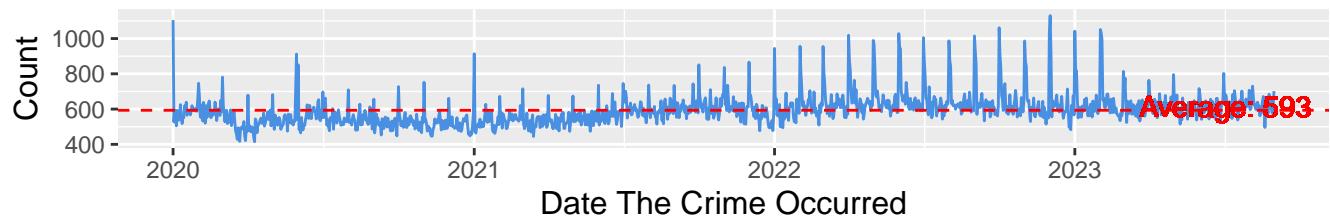
The following observations were made during the EDA process.

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

Distribution of Crime Types

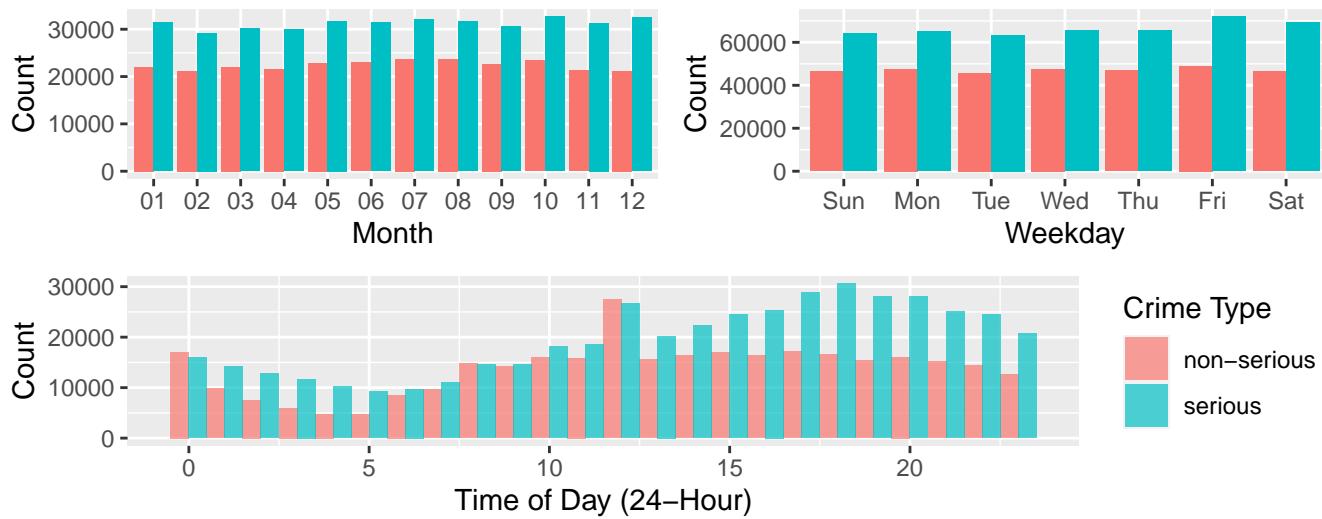


Daily Crime Counts

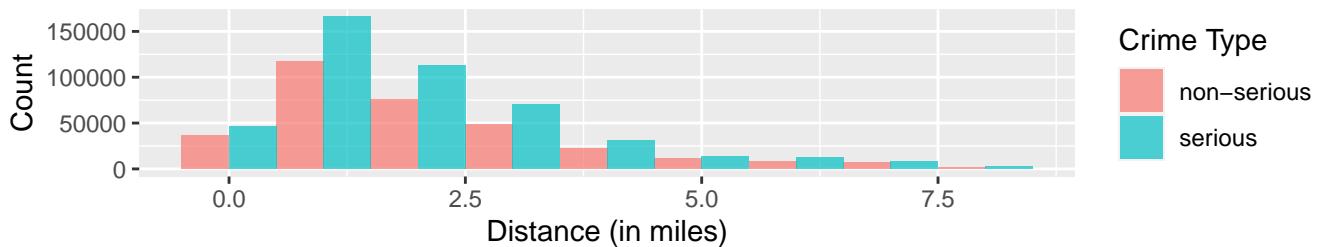
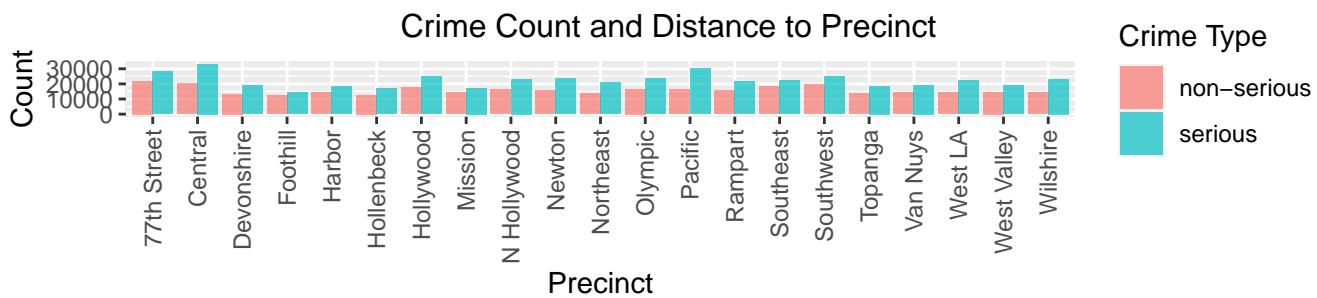


There is a minor variation in the proportion and count of crimes when compared on a monthly and weekday level. The variation is more prominent at the 'Time of Day' level.

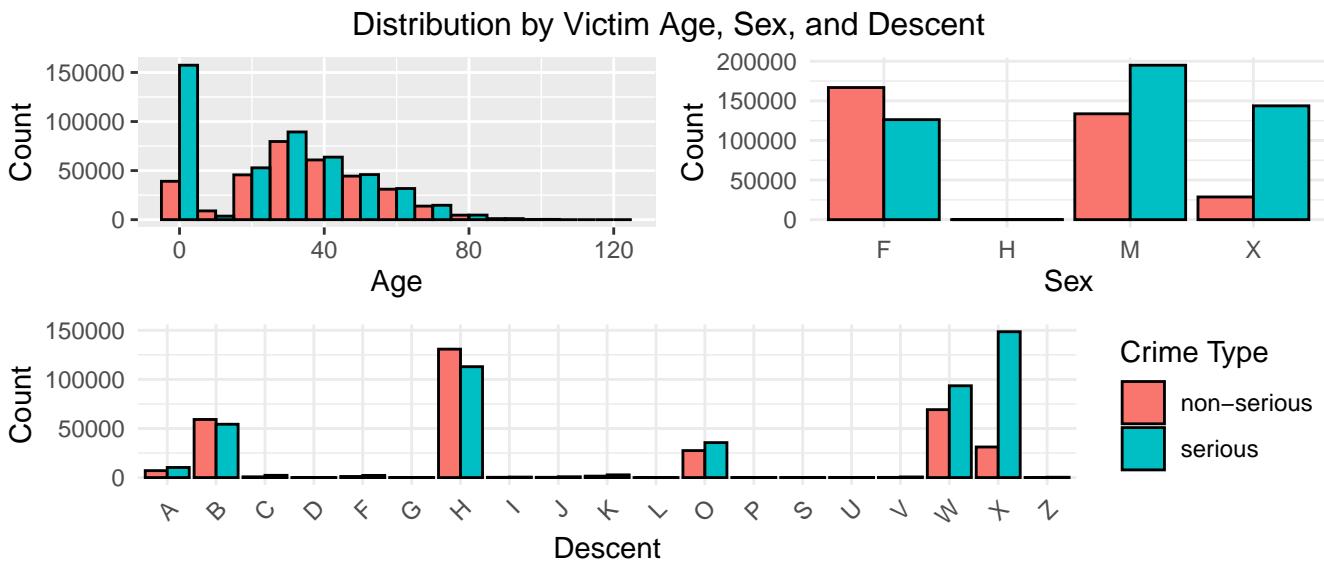
Crime Distribution by Month, Weekday, and Time of Day



There is a minor variation in the proportion and count of crimes at different precincts, the crime rate initially increases with the increase in distance from the precinct and then decreases.



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)



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 Logistic Regression analysis since it had very few (88) observations comparatively.

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)

	Coefficient Estimates	Standard Errors	p-values	Confidence Interval (2.5%, 97.5%)
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)
Weapon: Used	0.5995	5.269e-03	< 2e-16 ***	(0.5933, 0.6057)

Note: Full Model Coefficients are present in Appendix

The coefficients indicate how the likelihood of being a victim of a serious crime changes with each category or unit increase in a given variable, relative to a baseline category. For instance, being male increases the odds of being a victim of a serious crime by 1.88 times compared to being female.

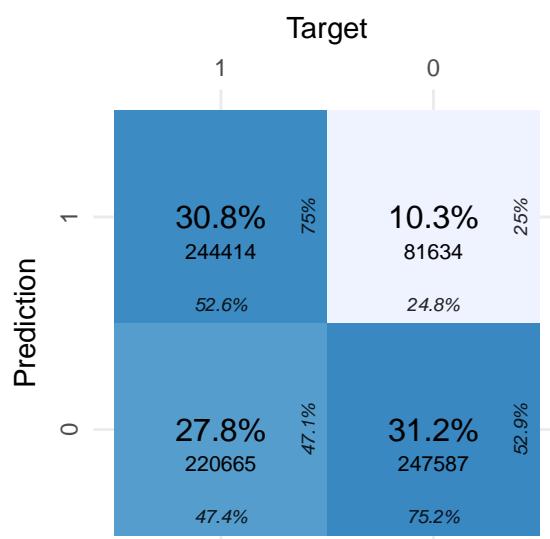
The model was compared with other models and null-model to study interactions and assess the model.

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 R²** value of 0.09004, though low, suggests limited explanatory power.

The **VIF** score of most of the parameters was around 1 and only the Victim Descent and Sex was above 10, however we decided to keep both of them in the model since they describe different demographic characteristics.

The model has a **AUC** value of 0.7 which indicates that the model has a decent ability to distinguish between classes, performing better than random chance. This value suggests that the model is moderately effective in predicting outcomes accurately.

A confusion matrix was created and the following metrics were studied to evaluate the model, “1” indicates the positive class of serious crimes:



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

Metric	Value
Accuracy	0.6194
95% Confidence Intervals	(0.6183, 0.6205)
Kappa	0.2614
Sensitivity	0.5255
Specificity	0.7520

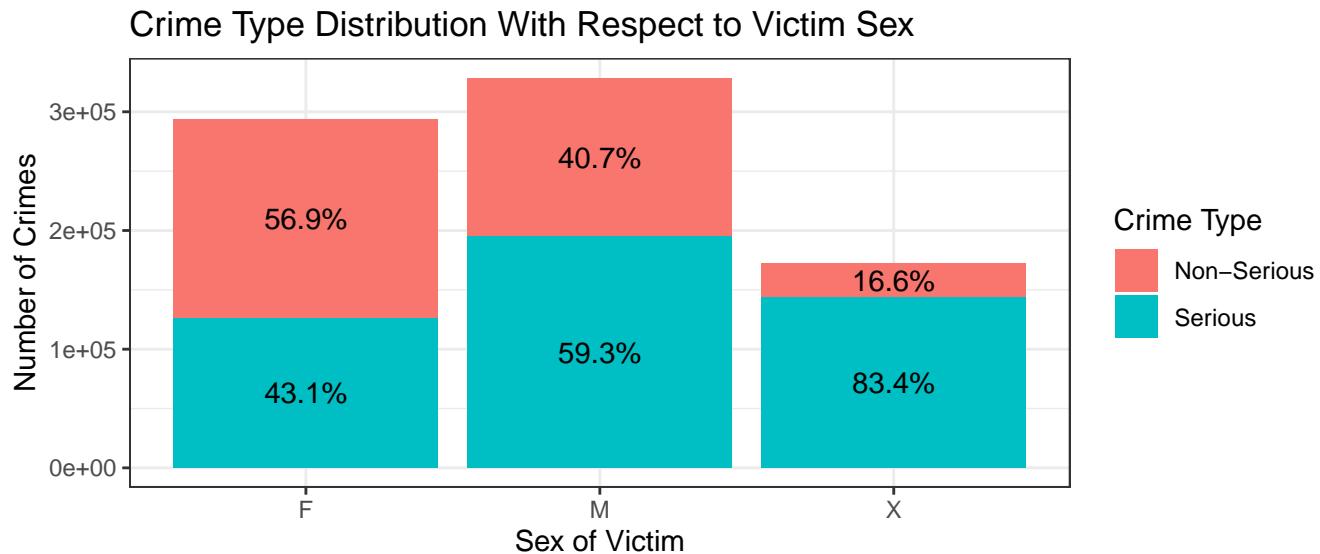
The model's accuracy was approximately 61.9%, not sufficiently high to be definitive. A Kappa value of 0.2614, while slightly better than chance, also indicates modest performance. Sensitivity and specificity are crucial in evaluating the model. Sensitivity (true positive rate) effectively identifies actual serious cases, while specificity (true negative rate) accurately identifies non-serious cases. Our model is better at correctly identifying non-serious crimes than serious ones, as it accurately identifies non-serious crimes in about 76% of cases vs ~52% of correctly identifying the 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 Usage: 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 victim of violent crime as compared to females:



Note Model Diagnostic Plots are present in Appendix

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:	0.9705	0.9638	0.9773	0.00E+00
Saturday				
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

Note: Full Model Coefficients are present in Appendix

The coefficient estimates represent the proportion in the number of crimes as compared to the reference level keeping all other variables constant. E.g. Controlling for other factors like Weekday and Area, the expected number of crimes increases by ~1% when the month is August as compared to January.

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	43.88
R-Squared	0.82

Our preliminary evaluation of the model's predictive accuracy, gauged through the RMSE, yielded a value of 43.88 to estimate the average difference between the predicted and observed crime counts, in comparison to the average crime count ~450 in the Test dataset indicates that the model is effective in predicting the number of crimes. Additionally, the R-squared value of 0.82 reflects a strong linear relationship between the observed and predicted crime counts, explaining a substantial proportion of the variance in the crime data.

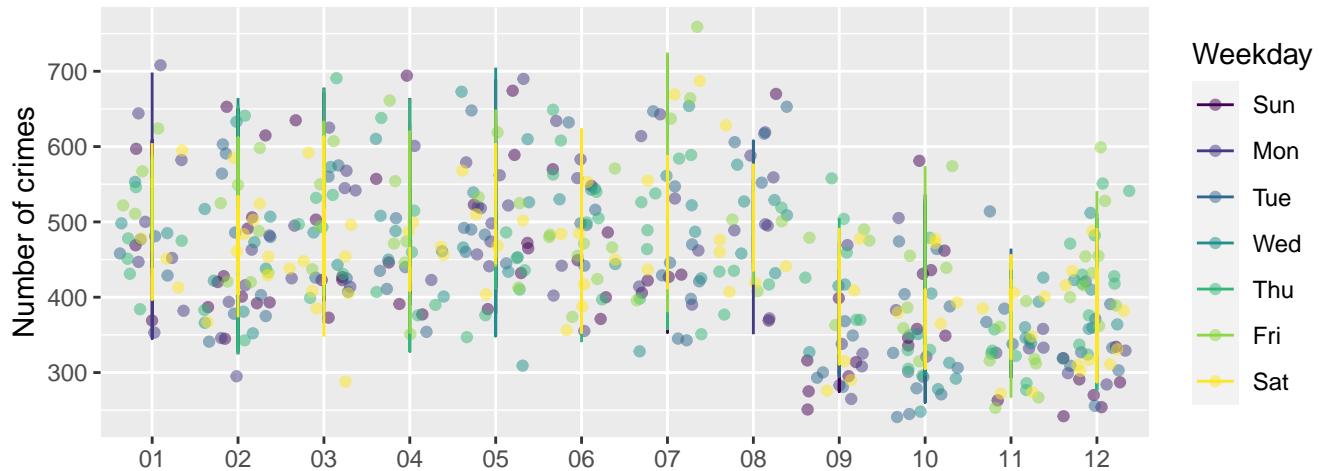
Preliminary examination of the model outputs suggests distinct spatial and temporal crime patterns.

- **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 4.10, 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.

The plot below shows the relationship between the Crime counts and Temporal factors:

Predicted Number of crime by month



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.

References

- Los Angeles Police Department. (2023). Crime data from 2020 to present. City of Los Angeles. Retrieved from <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>

Appendix

Mapping for Victim Descent Code to Victim Descent:

Victim Descent Code	Victim Descent
A	Other Asian
B	Black
C	Chinese
D	Cambodian
F	Filipino
G	Guamanian
H	Hispanic/Latin/Mexican
I	American Indian/Alaskan Native
J	Japanese
K	Korean
L	Laotian
O	Other
P	Pacific Islander
S	Samoan
U	Hawaiian
V	Vietnamese
W	White
X	Unknown
Z	Asian Indian

Mapping for Victim Sex Code to Victim Sex:

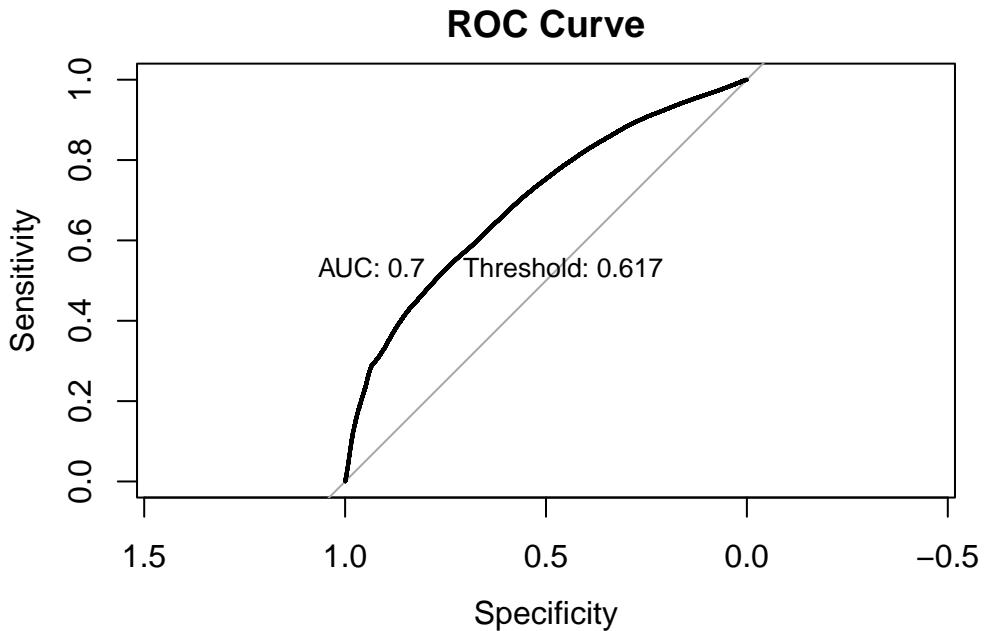
Victim Descent Code	Victim Descent
M	Male
F	Female
X	Unknown/Other

Logistic Regression Model Coefficients Full:

	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 Sex: Unknown	3.4659	2.507e-02	< 2e-16 ***	(3.2995, 3.6402)
Victim Descent: Black	0.7144	1.693e-02	< 2e-16 ***	(0.6910, 0.7385)
Victim Descent: Chinese	1.6801	4.516e-02	< 2e-16 ***	(1.5385, 1.8365)
Victim Descent: Cambodian	1.4699	3.002e-01	0.1994	(0.8321, 2.7216)
Victim Descent: Filipino	1.1828	4.074e-02	3.78e-05 ***	(1.0922, 1.2814)
Victim Descent: Guamanian	1.0696	2.884e-01	0.8156	(0.6120, 1.9082)
Victim Descent: Hispanic/Latin/Mexican	0.6354	1.635e-02	< 2e-16 ***	(0.6154, 0.6561)
Victim Descent: American Indian/Alaskan Native	1.0689	7.907e-02	0.3992	(0.9164, 1.2495)

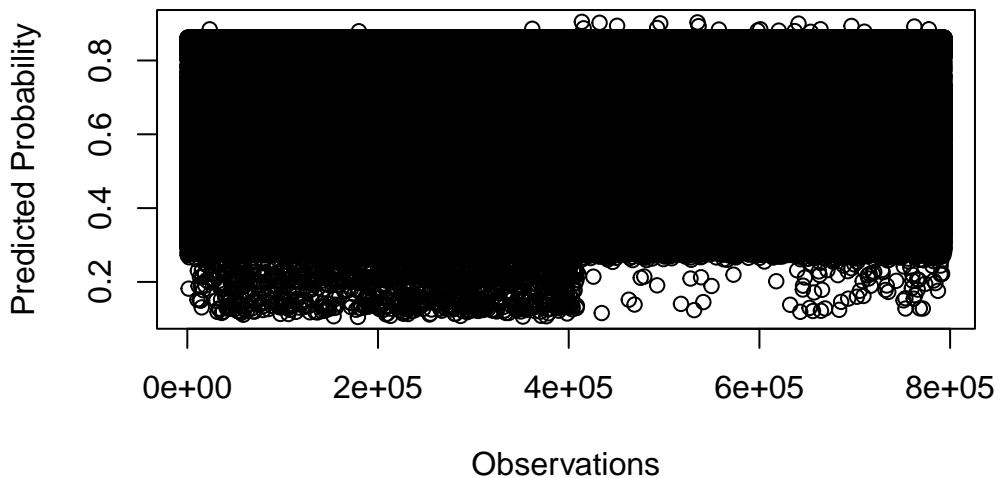
	Coefficient Estimates	Standard Errors	p-values	Confidence Interval (2.5%, 97.5%)
Victim Descent: Japanese	1.7731	7.185e-02	1.56e-15 ***	(1.5423, 2.0443)
Victim Descent: Korean	1.1302	3.635e-02	0.0008 **	(1.0526, 1.2138)
Victim Descent: Laotian	0.5307	2.953e-01	0.0319 *	(0.2944, 0.9438)
Victim Descent: Other	0.7804	1.782e-02	< 2e-16 ***	(0.7536, 0.8082)
Victim Descent: Pacific Islander	1.0433	1.446e-01	0.7694	(0.7880, 1.3899)
Victim Descent: Samoan	0.6829	3.072e-01	0.2145	(0.3733, 1.2552)
Victim Descent: Hawaiian	1.4184	1.759e-01	0.0469 *	(1.0120, 2.0197)
Victim Descent: Vietnamese	1.9475	8.509e-02	4.75e-15 ***	(1.6518, 2.3062)
Victim Descent: White	0.8616	1.659e-02	< 2e-16 ***	(0.8340, 0.8900)
Victim Descent: Unknown	1.0460	2.853e-02	0.1153	(0.9891, 1.1062)
Victim Descent: Asian Indian	1.8502	1.209e-01	3.61e-07 ***	(1.4663, 2.3569)
Distance to precinct	0.9998	5.826e-06	< 2e-16 ***	(0.9998, 0.9999)
Weapon: Used	0.5995	5.269e-03	< 2e-16 ***	(0.5933, 0.6057)

Logistic Regression ROC-AUC Curve:



Logistic Regression Predicted Probabilities for observations:

Predicted Probabilities



Poisson Regression Model Coefficients Full:

	Coefficient Estimates	2.50% Confidence Interval	97.50% Confidence Interval	P Values
(Intercept)	663.6024	655.4907	671.7956	0.00E+00
Weekday: Monday	1.0563	1.0491	1.0635	0.00E+00
Weekday: Tuesday	1.0155	1.0085	1.0225	0.00E+00
Weekday: Wednesday	0.9843	0.9776	0.9911	0.00E+00
Weekday: Thursday	0.9729	0.9663	0.9795	0.00E+00
Weekday: Friday	0.9846	0.9778	0.9915	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: 03	0.9715	0.9599	0.9832	0.00E+00
Month: 04	0.9515	0.9404	0.9628	0.00E+00
Month: 05	0.9971	0.9852	1.0092	6.40E-01
Month: 06	0.9868	0.9752	0.9987	2.94E-02
Month: 07	1.0335	1.0214	1.0457	0.00E+00
Month: 08	1.0163	1.0046	1.0281	6.20E-03
Month: 09	0.7242	0.715	0.7335	0.00E+00
Month: 10	0.768	0.7581	0.778	0.00E+00
Month: 11	0.7217	0.7125	0.7309	0.00E+00
Month: 12	0.7238	0.714	0.7336	0.00E+00
Area: Central	1.0655	1.0509	1.0803	0.00E+00
Area: Devonshire	0.6458	0.6357	0.6561	0.00E+00
Area: Foothill	0.5268	0.518	0.5356	0.00E+00
Area: Harbor	0.6578	0.6474	0.6683	0.00E+00
Area: Hollenbeck	0.5846	0.5749	0.5944	0.00E+00
Area: Hollywood	0.8362	0.8236	0.8491	0.00E+00
Area: Mission	0.6311	0.621	0.6413	0.00E+00
Area: N Hollywood	0.7796	0.7681	0.7914	0.00E+00
Area: Newton	0.7854	0.7736	0.7974	0.00E+00
Area: Northeast	0.6819	0.6711	0.6928	0.00E+00
Area: Olympic	0.7951	0.7828	0.8075	0.00E+00

Coefficient	Estimates	2.50% Confidence Interval	97.50% Confidence Interval	P Values
Area: Pacific	0.9285	0.9153	0.9419	0.00E+00
Area: Rampart	0.7423	0.731	0.7538	0.00E+00
Area: Southeast	0.8072	0.7954	0.8193	0.00E+00
Area: Southwest	0.8898	0.8769	0.9029	0.00E+00
Area: Topanga	0.6326	0.6224	0.6429	0.00E+00
Area: Van Nuys	0.6688	0.6583	0.6793	0.00E+00
Area: West LA	0.7223	0.7108	0.734	0.00E+00
Area: West Valley	0.6547	0.6443	0.6654	0.00E+00
Area: Wilshire	0.7505	0.7395	0.7617	0.00E+00