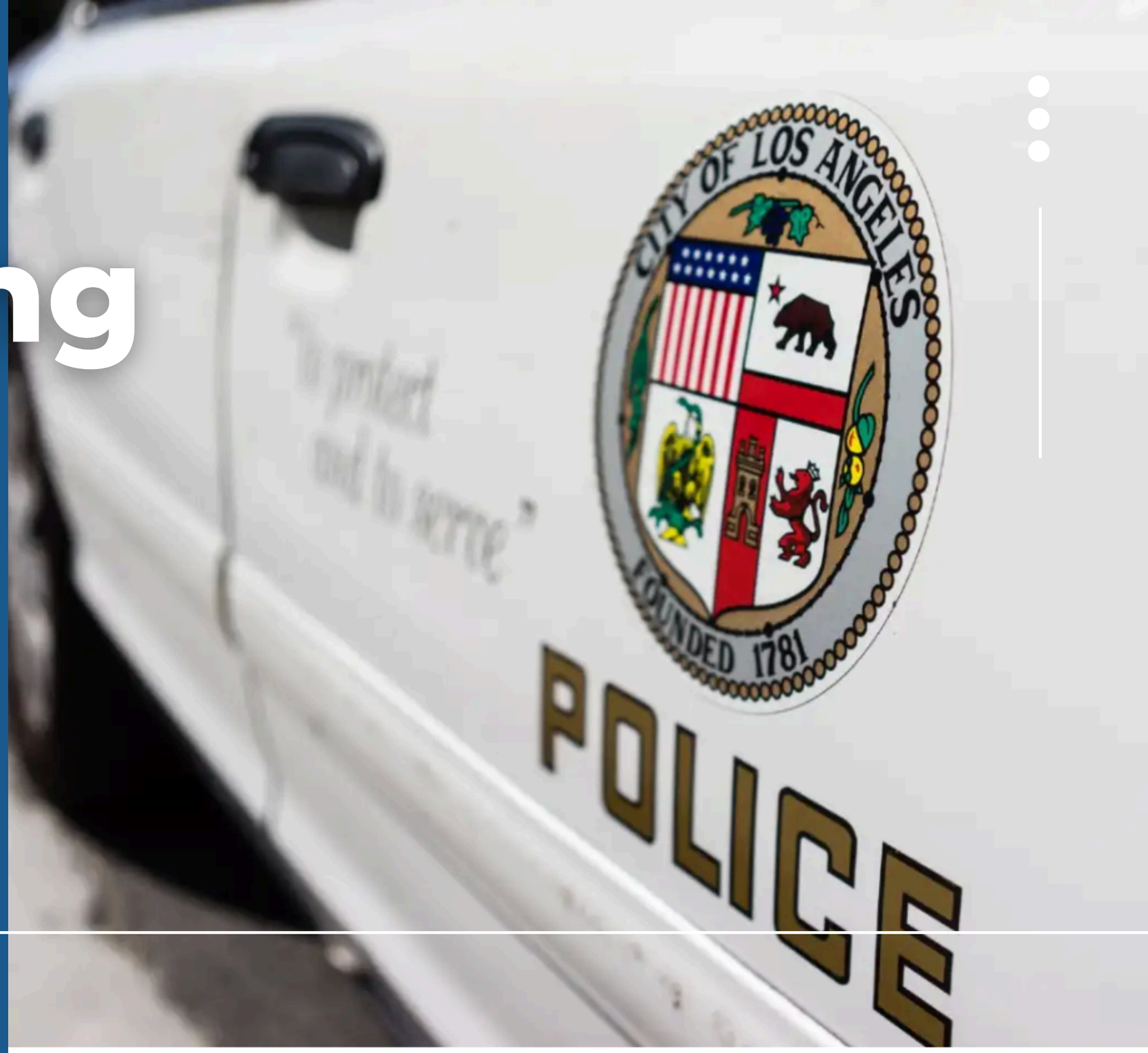
Predicting Crime

A datadriven analysis on the impact of machine learning models for crime prevention in Los Angeles

By Oliver Bohler



Problem Statement

The LAPD faces a critical shortage of officers, resulting in delayed 911 responses and decreased public safety. With the advance of new technology the fight against crime can be supported by machine learning models

This data driven approach aims to predict crime patterns and hotspots to help allocate resources more efficiently and reduce crime rates, benefiting both law enforcement and the community.

Approach

The goal is to analyze historical crime data and identify patterns in time, location, and victim demographics. Using predictive modeling, I aim to forecast future crime trends, enabling a more optimized resource allocation and proactive crime prevention strategy.

Goal 1

Develop machine learning models to predict crime outcomes, focusing on arrest likelihood and identifying high-risk areas and time periods.

Goal 2

Leverage data insights to help the LAPD efficiently allocate limited resources to zones and crime types with the highest impact on public safety.

Goal 3

Examine historical crime data to identify trends and patterns over the years, providing insights into how crime evolves post-pandemic and beyond.

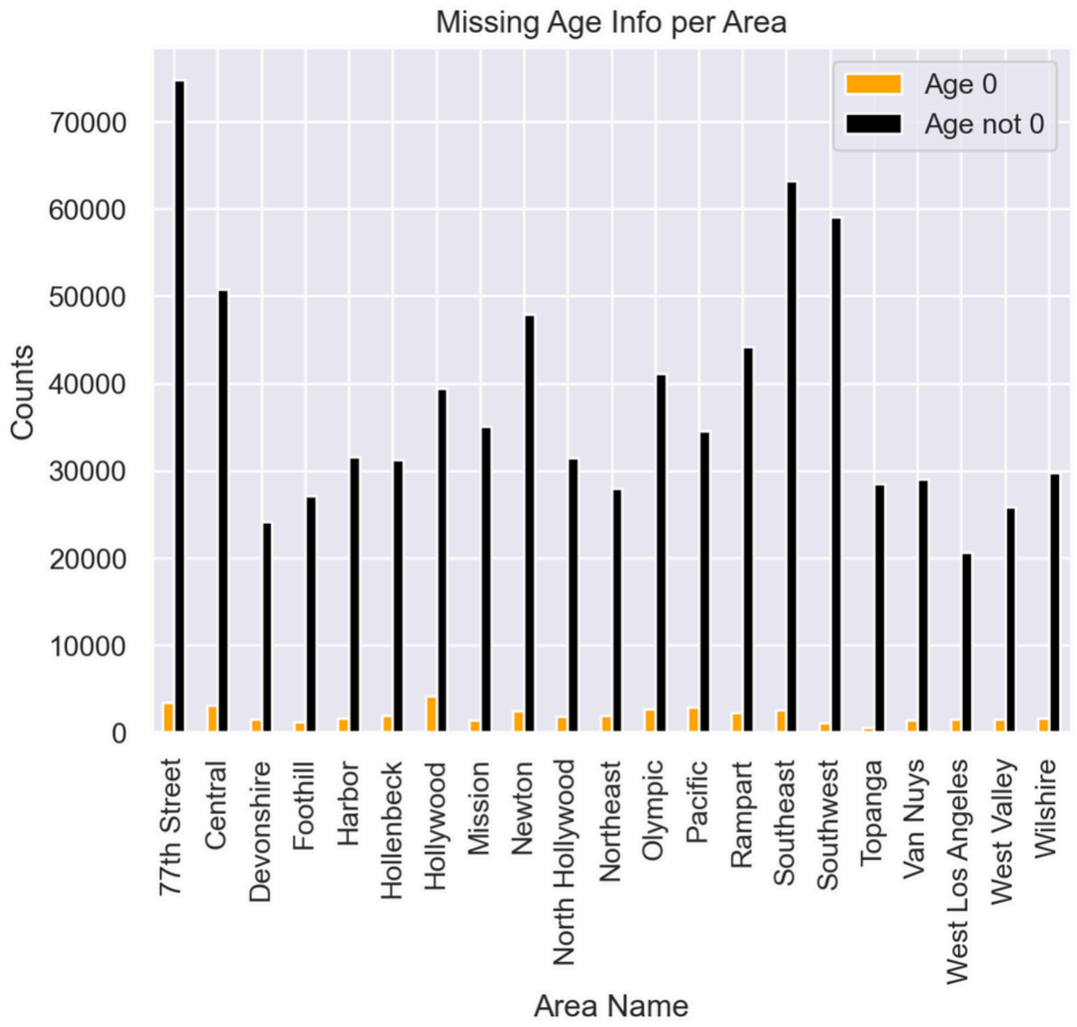
Goal 4

Uncover similarities and factors that lead to arrests, while analyzing the likelihood of gun involvement during certain hours to better protect police officers.

Data Cleaning

The data was sourced from data.gov and provided by the Los Angeles Police Department.

- Initial Dataset: Over 3 million crime entries from 2010–2023 across 17 jurisdiction districts of Los Angeles.
- The data included many missing values and errors so data wrangling steps were taken



Cleaning Process

Inconsistent/ Missing Data

Problem: Handwritten records later digitized led to typos, missing values
and inconsistencies.
Solution: Implemented thorough data validation, dropping irrelevant columns and imputing missing values systematically (e.g., mean imputation for Victim Age).

High Variability in Crime

Problem: The dataset included numerous crime codes with varying levels of detail, making analysis complex and noisy.
Solution: Grouped crimes into broader categories (e.g., Robbery & Burglary, Assault & Battery) to simplify and focus the analysis.

Irregular Temporal Trends

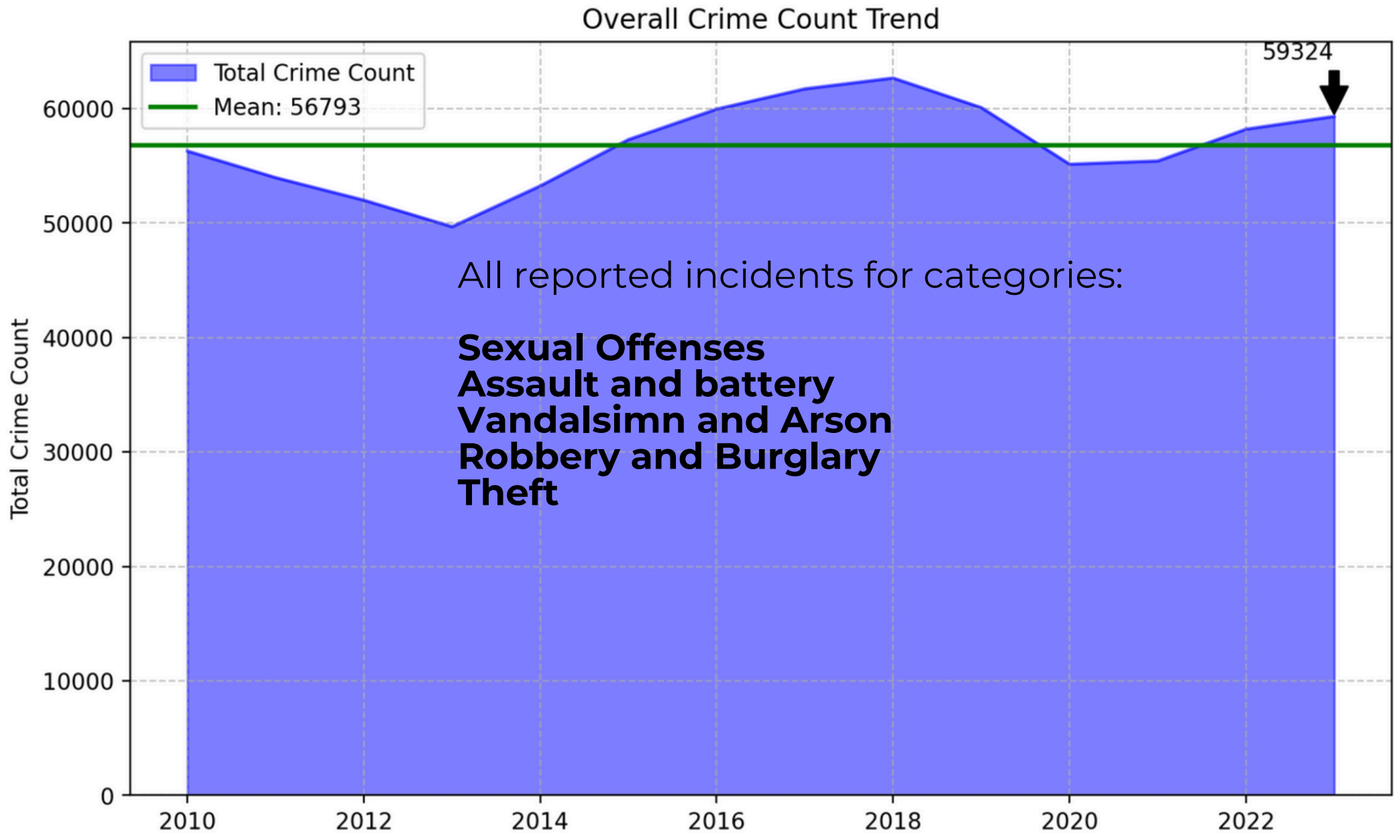
Problem: Crime data had irregular patterns and inconsistencies across years due to reporting differences and missing timestamps.
Solution: Standardized time formatting, adjusted for missing values, and created temporal features like year and hour to uncover actionable trends.

Complex Geographics

Problem: Original dataset included 17 jurisdictions that were overly detailed for analysis.
Solution: Consolidated jurisdictions into 4 broader zones using socio-economic and geographic similarities to provide actionable insights.

Exploratory Data Analysis

The main goal of EDA is to analyze the LAPD crime dataset to understand its structure, quality, and completeness for actionable insights. It focuses on identifying patterns and trends in crime, such as hotspots, victim demographics, and temporal variations. Finally, EDA prepares the data for predictive modeling, enabling resource optimization and improved crime prevention strategies.

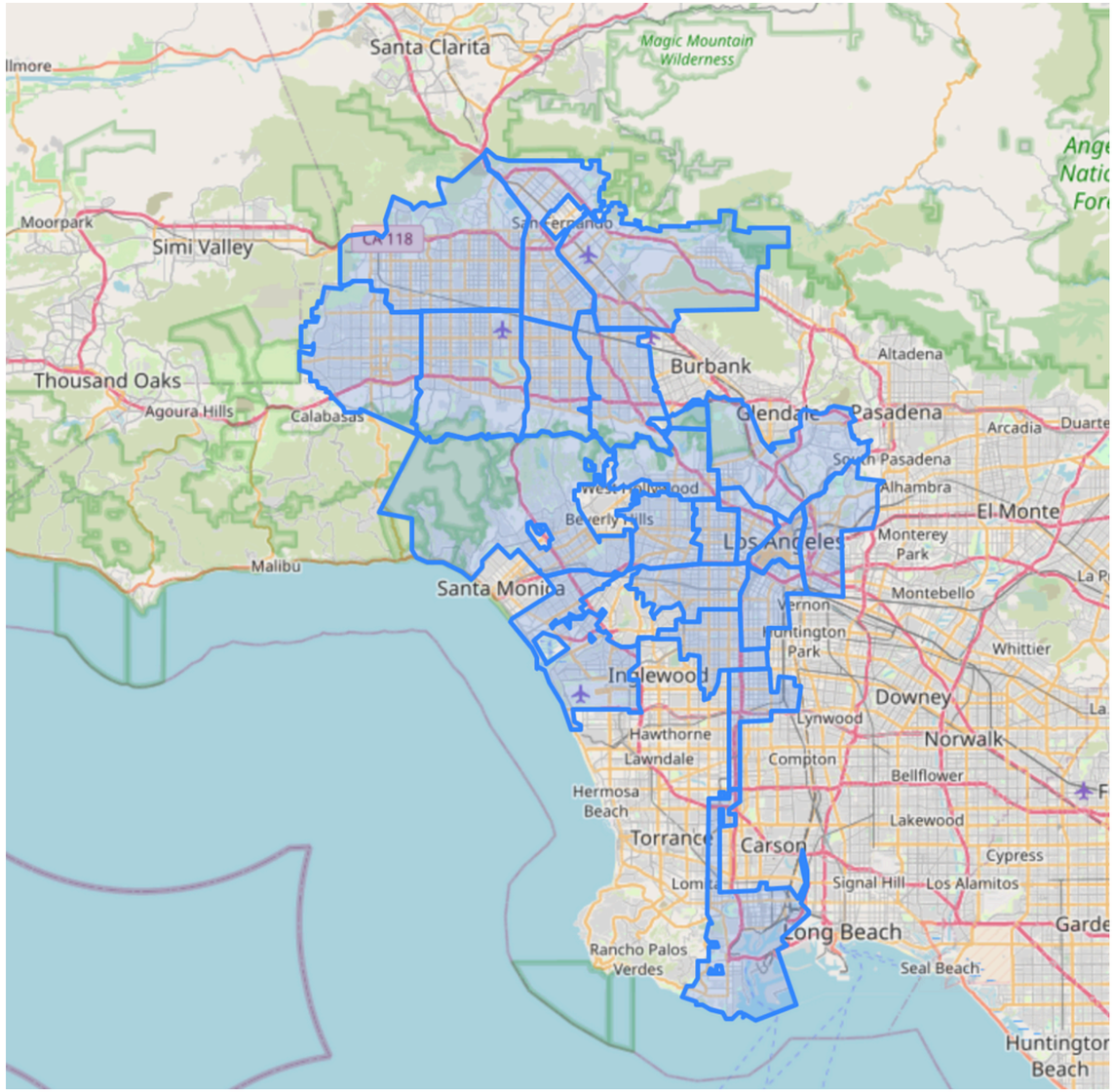


Mapping Zones

To visualize the geographic distribution of crime, the 17 LAPD jurisdictions were grouped into four main zones and plotted on a map:

1. **Hollywood/ Downtown**: High-density urban areas, including Downtown LA.
2. **Westside and Wilshire**: Coastal and affluent neighborhoods.
3. **San Fernando Valley**: Suburban areas with mixed residential and commercial spaces.
4. **South LA**: Predominantly residential and industrial areas.

The map provided a clear and simplified view of these zones and an html link is provided in the project folder



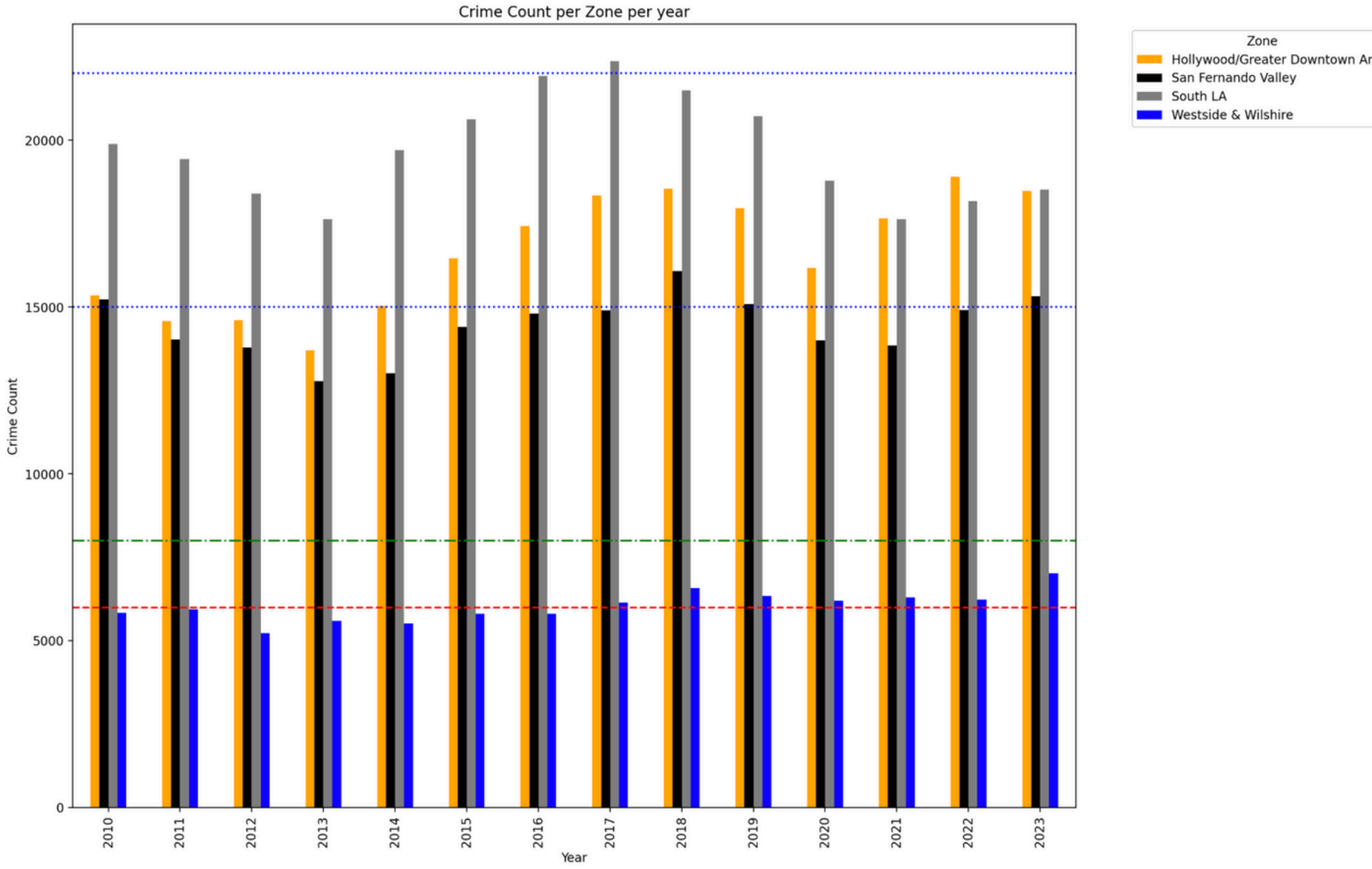
Data at a Glance

This bar chart visualizes the yearly crime count across four main zones in Los Angeles from 2010 to 2023.

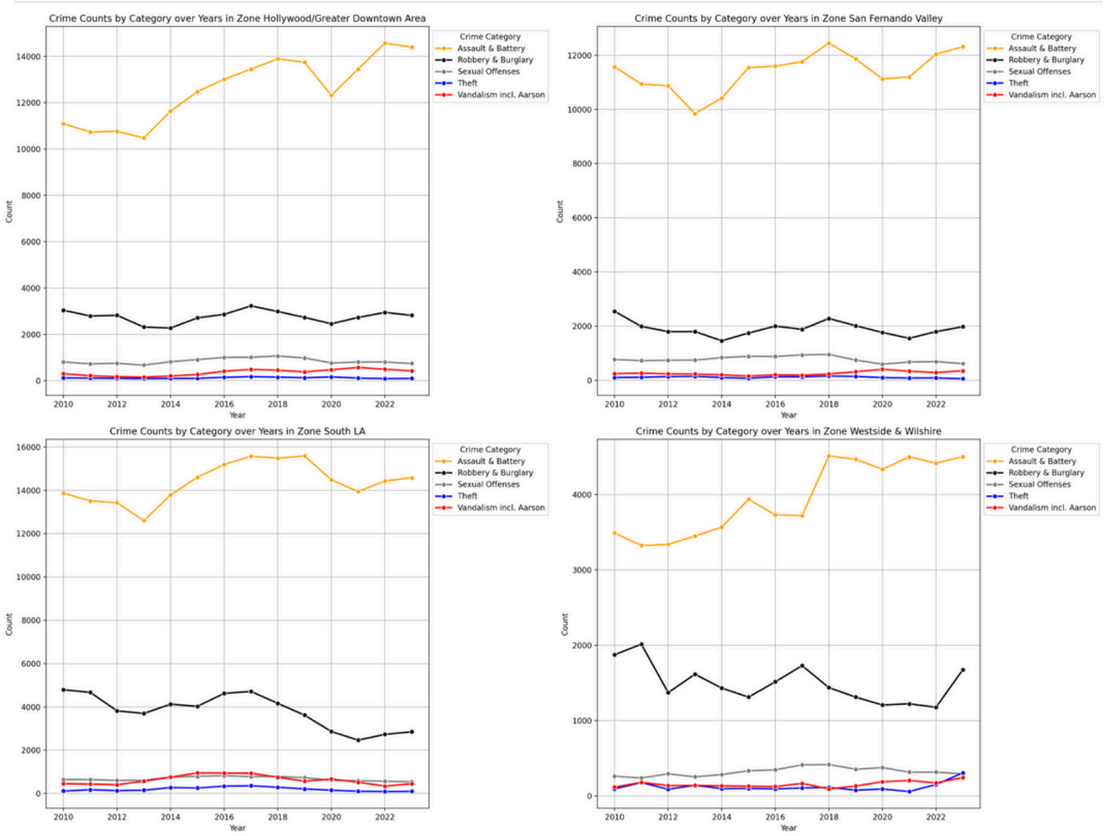
Key Observations:

a. Hollywood/Greater Downtown Area consistently reports high crime rates, followed by San Fernando Valley.
b. South LA shows moderate crime levels across all years.
c. Westside & Wilshire has the lowest crime rates, with significant stability over time.

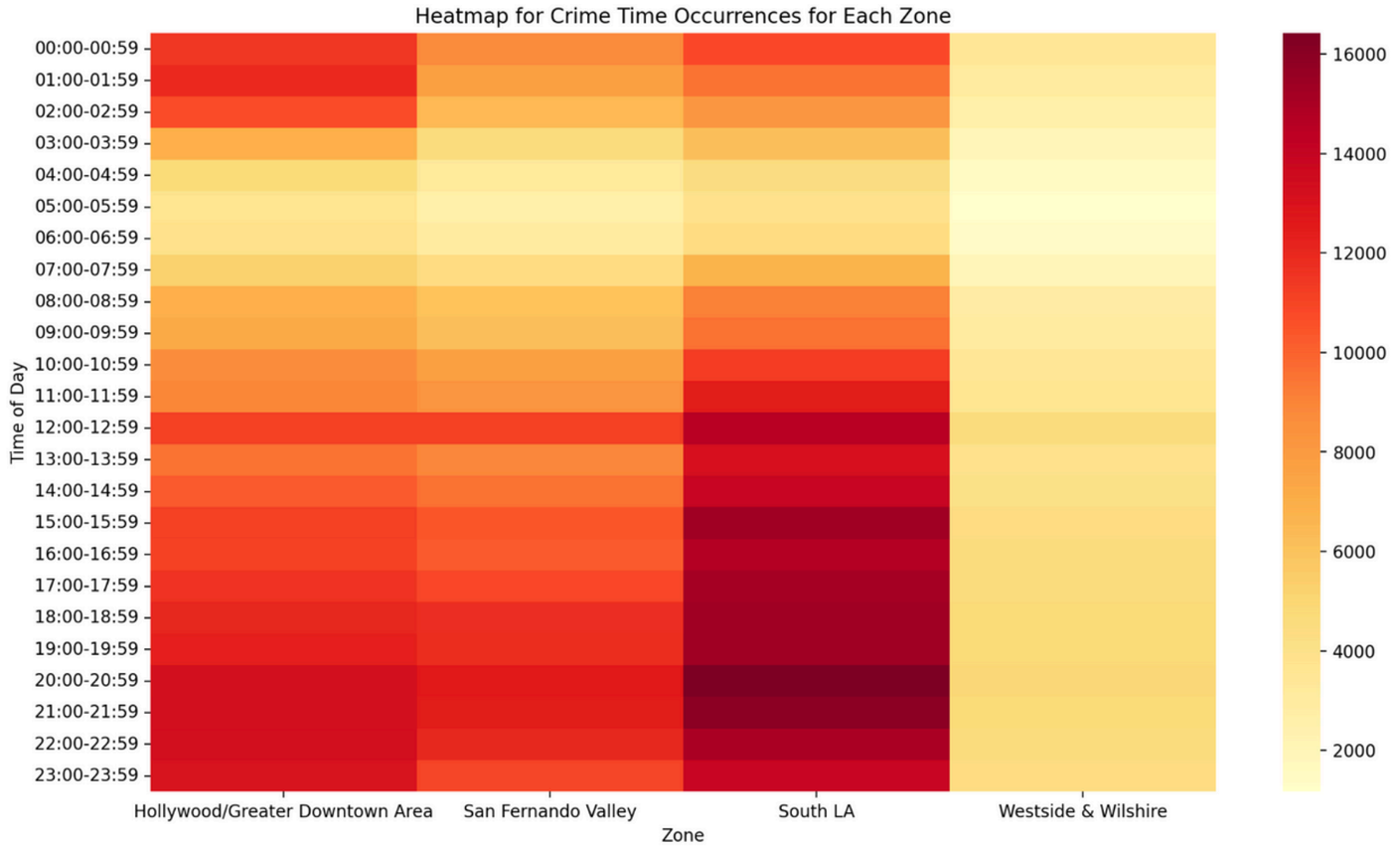
Insights: The distinct variations in crime count by zone emphasize the need for zone-specific resource allocation and targeted crime prevention strategies.



Crime Count per Zone



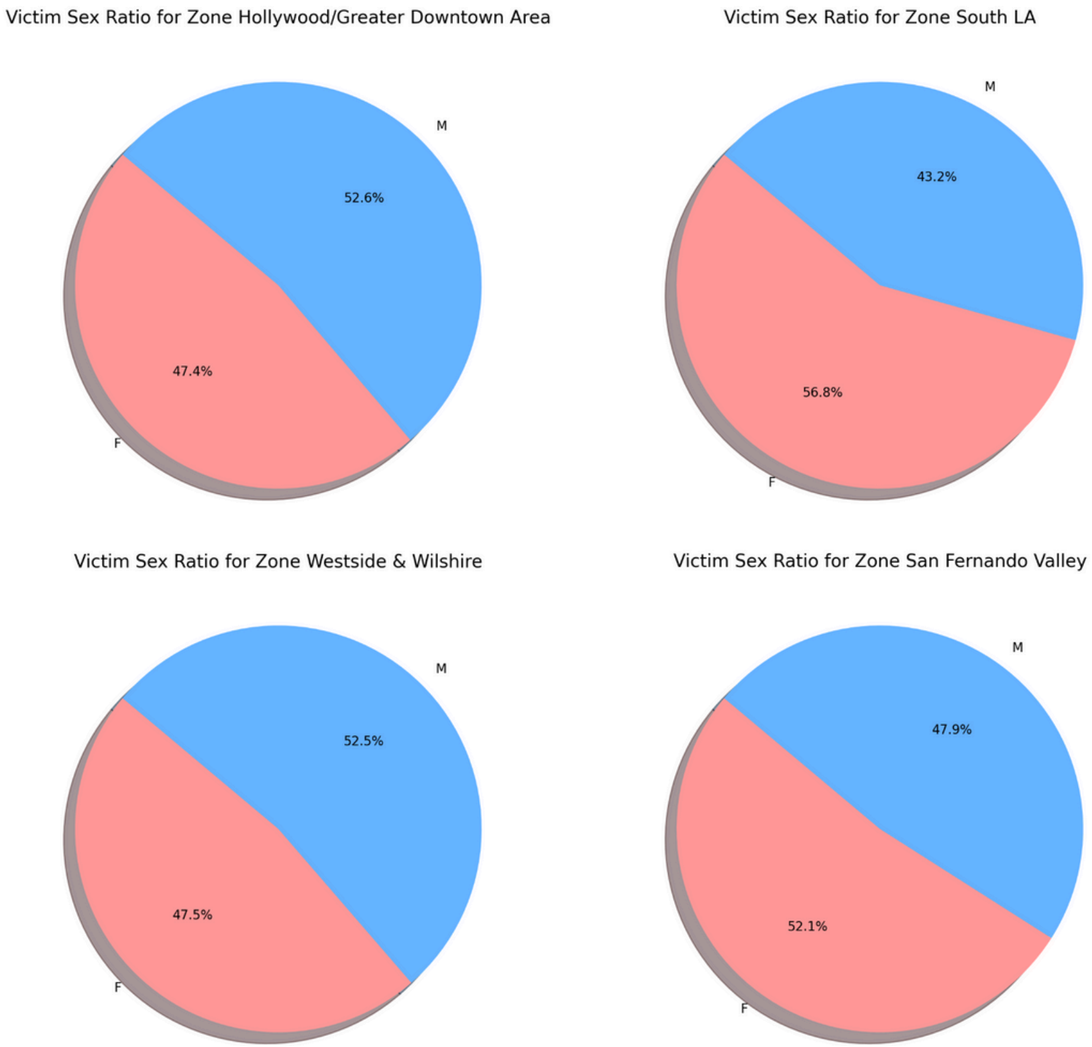
Heatmap Los Angeles



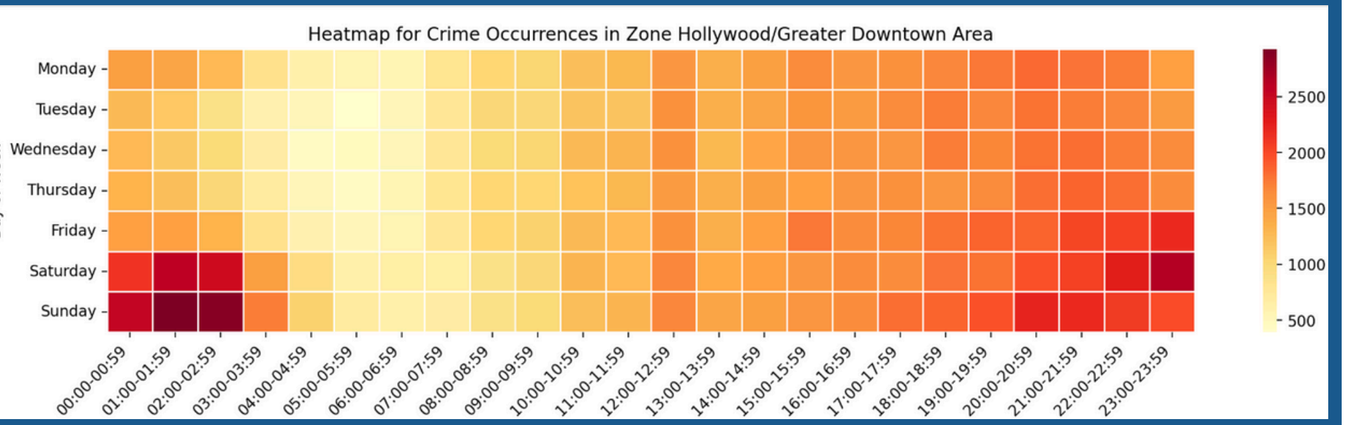
Zones at a Glance

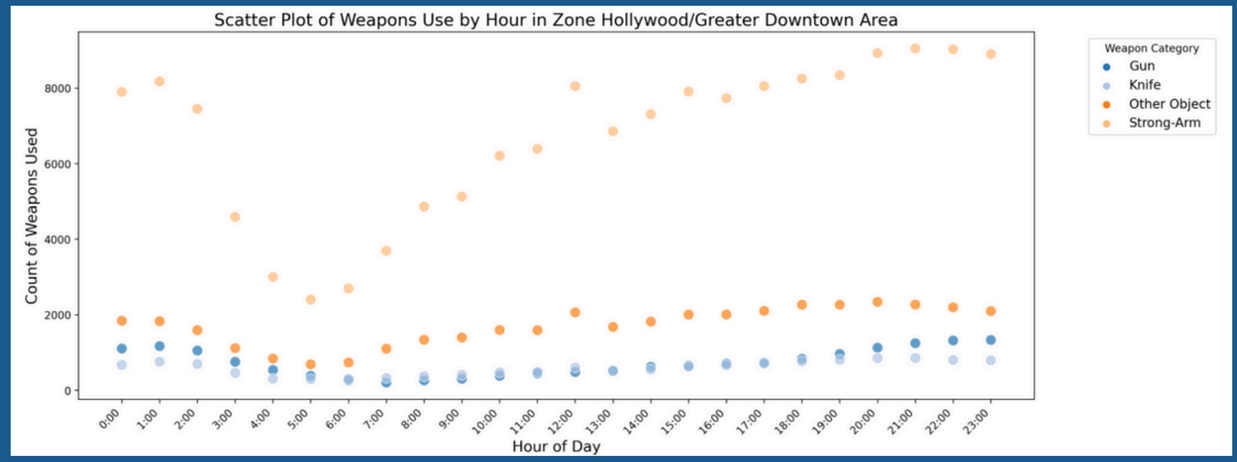
The following slides show the distribution of crimes in each zone as well as specifics about weapon usage and time of inciddents.

- On the right: Zonal distribution of Victims Male and Female



Hollywood & Downtown

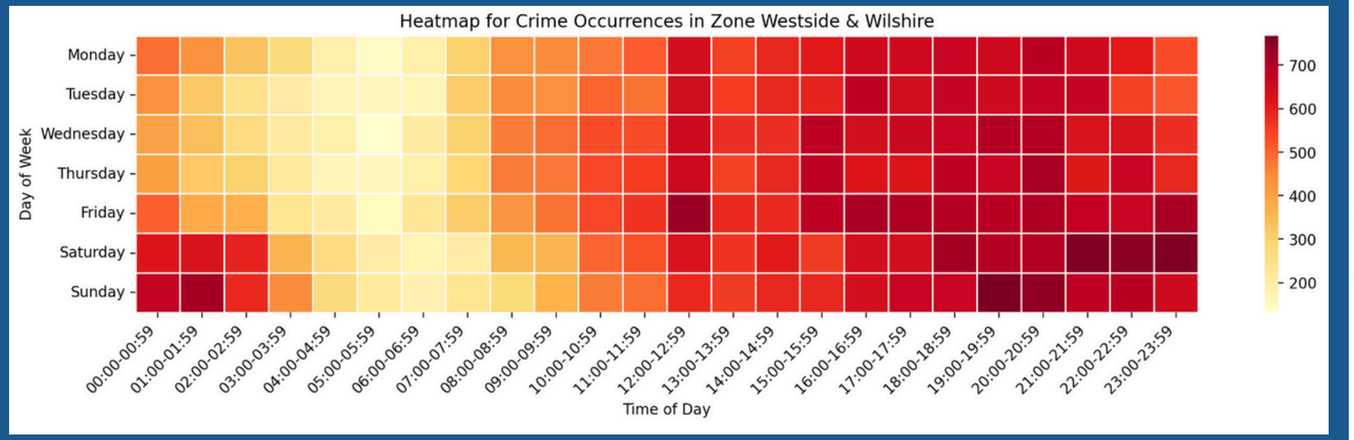


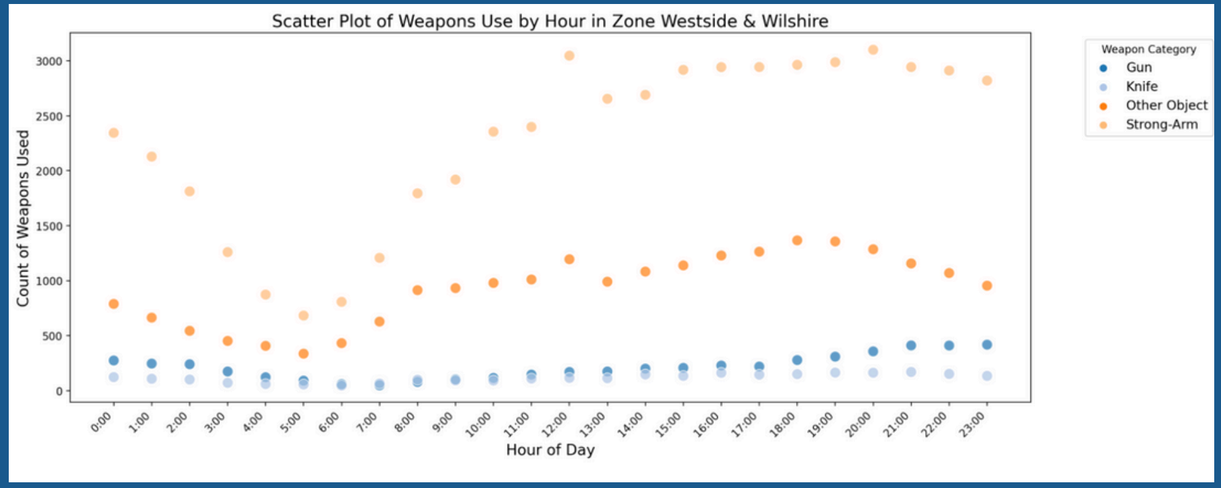


Crime Category

Total Arrest Rate for Zone Hollywood/Greater Downtown Area: 43.16%
- Crime Category: Assault & Battery, Arrests: 84818, Investigations: 91155, Arrest Rate: 48.20%
- Crime Category: Robbery & Burglary, Arrests: 9204, Investigations: 29484, Arrest Rate: 23.79%
- Crime Category: Sexual Offenses, Arrests: 4662, Investigations: 7197, Arrest Rate: 39.31%
- Crime Category: Theft, Arrests: 413, Investigations: 1255, Arrest Rate: 24.76%
- Crime Category: Vandalism incl. Aarson, Arrests: 1538, Investigations: 3452, Arrest Rate: 30.82%

Westside & Wilshire

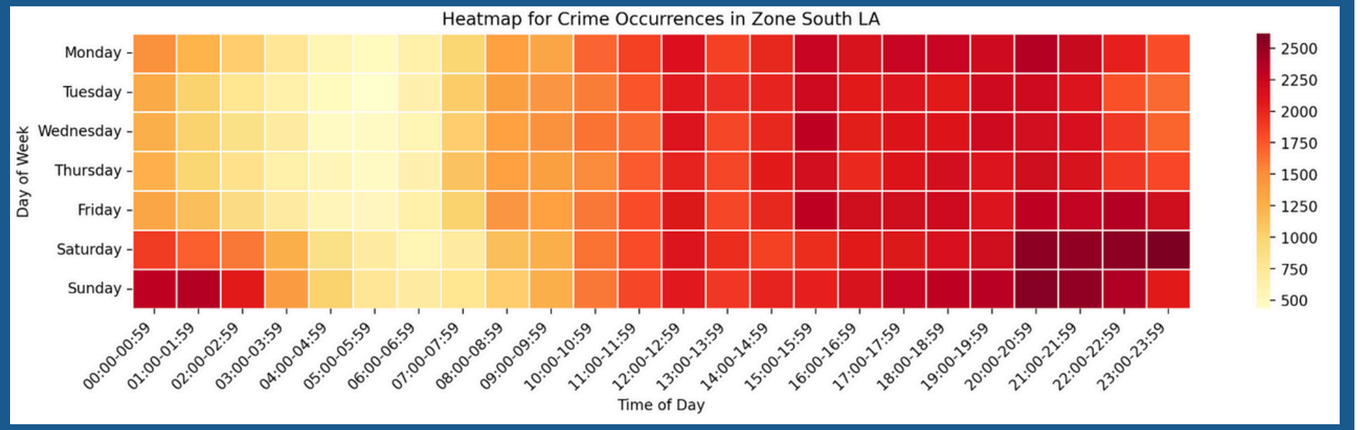


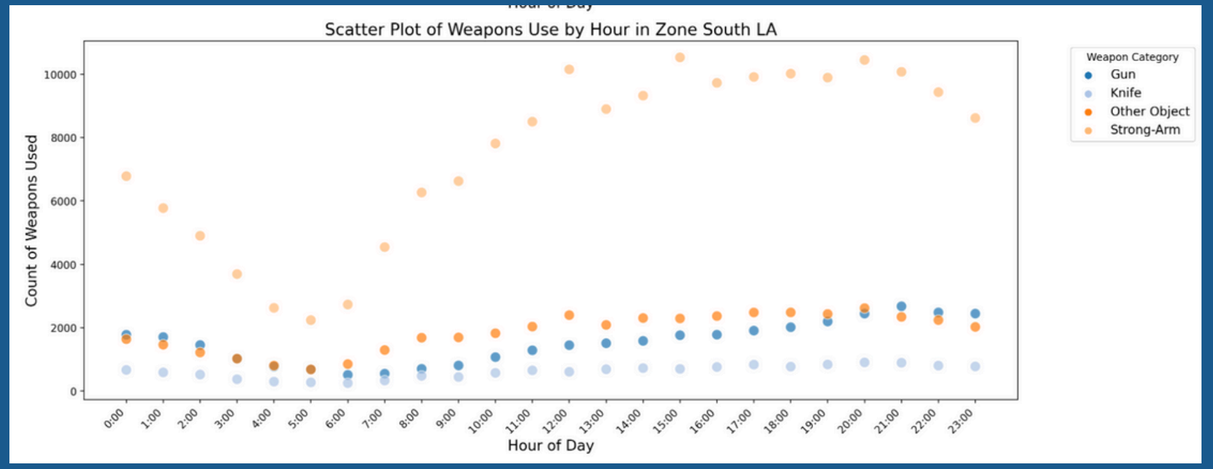


Crime Category

Total Arrest Rate for Zone Westside & Wilshire: 37.43%
- Crime Category: Assault & Battery, Arrests: 25619, Investigations: 29702, Arrest Rate: 46.31%
- Crime Category: Robbery & Burglary, Arrests: 3417, Investigations: 17480, Arrest Rate: 16.35%
- Crime Category: Sexual Offenses, Arrests: 1902, Investigations: 2587, Arrest Rate: 42.37%
- Crime Category: Theft, Arrests: 151, Investigations: 1538, Arrest Rate: 8.94%
- Crime Category: Vandalism incl. Aarson, Arrests: 563, Investigations: 1595, Arrest Rate: 26.09%

South LA

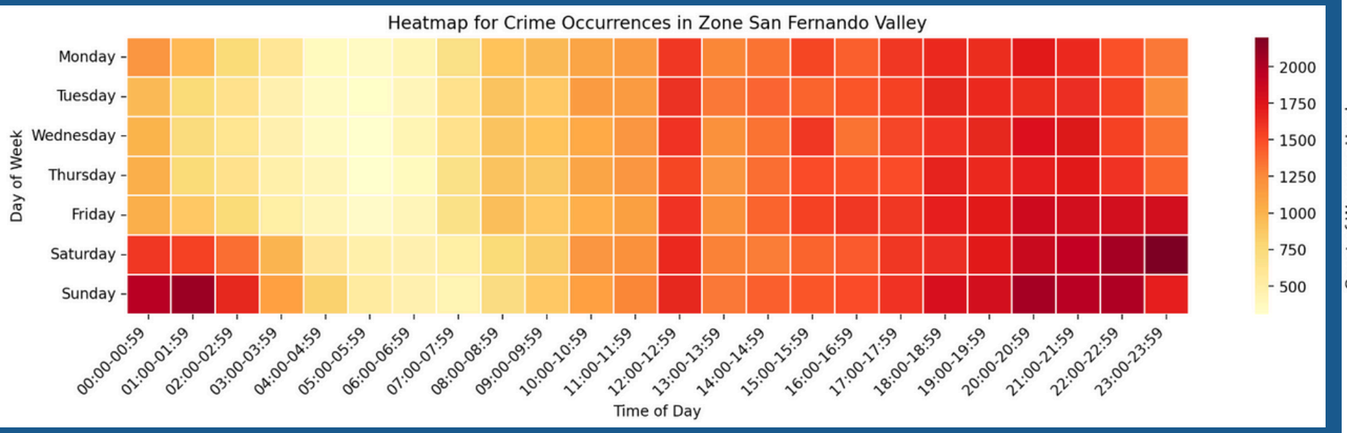


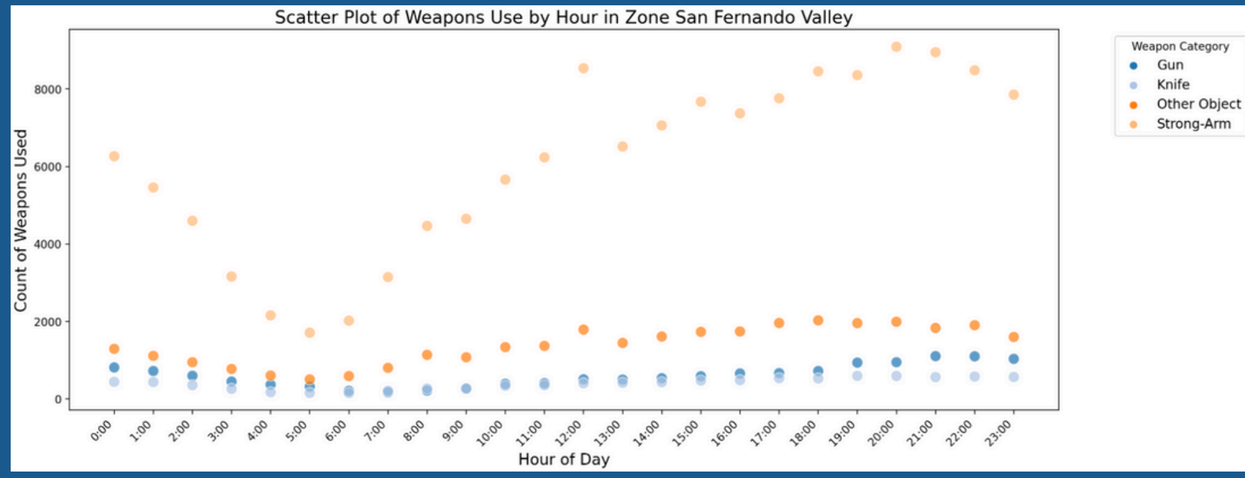


Crime Category

Total Arrest Rate for Zone South LA: 36.93%
- Crime Category: Assault & Battery, Arrests: 82930, Investigations: 118127, Arrest Rate: 41.25%
- Crime Category: Robbery & Burglary, Arrests: 11701, Investigations: 41446, Arrest Rate: 22.02%
- Crime Category: Sexual Offenses, Arrests: 3996, Investigations: 5505, Arrest Rate: 42.06%
- Crime Category: Theft, Arrests: 574, Investigations: 2184, Arrest Rate: 20.81%
- Crime Category: Vandalism incl. Aarson, Arrests: 2434, Investigations: 6324, Arrest Rate: 27.79%

San Fernando Valley





Crime Category

Total Arrest Rate for Zone South LA: 36.93%
- Crime Category: Assault & Battery, Arrests: 82930, Investigations: 118127, Arrest Rate: 41.25%
- Crime Category: Robbery & Burglary, Arrests: 11701, Investigations: 41446, Arrest Rate: 22.02%
- Crime Category: Sexual Offenses, Arrests: 3996, Investigations: 5505, Arrest Rate: 42.06%
- Crime Category: Theft, Arrests: 574, Investigations: 2184, Arrest Rate: 20.81%
- Crime Category: Vandalism incl. Aarson, Arrests: 2434, Investigations: 6324, Arrest Rate: 27.79%

Feature Engineering

one-hot encoding

bool values for target variable
1 = arrest,
0 = no arrest
Categorical features were encoded using one-hot encoding

Feature Selection

Premise
Crime Category
Time
Zone
dataset was split into training and test sets

Imbalanced Classes

Class weights were computed to penalize misclassifications of the minority class. Weights ensure the model does not overly favor the majority class (0: No Arrest).

Model Testing

Linear Regreassion

Benefit: Establishes a baseline model for predicting crime patterns and trends by simplifying the relationship between features.
Evaluation: Performed well as an initial model, but its simplicity limited its ability to capture non-linear patterns. Improvement could involve incorporating more sophisticated features to better model crime complexity.

knn-Neighbors

Benefit: Provided localized crime pattern insights, making it effective for community-level analyses.
Evaluation: Achieved the lowest ROC AUC score (0.6613) due to sensitivity to noise and imbalanced data. Improvements could focus on optimizing hyperparameters and addressing class imbalance to enhance performance.

XGBoost

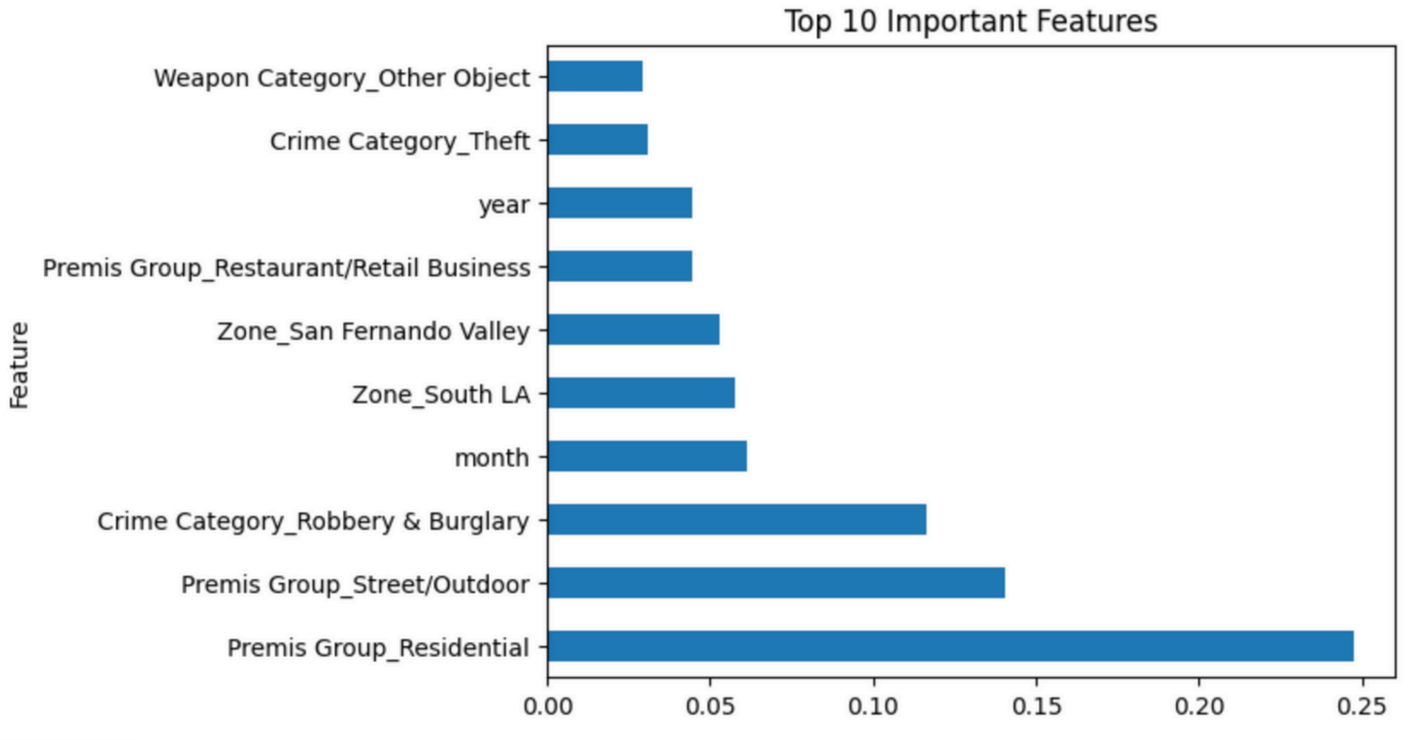
Benefit: Delivered robust predictions with a strong focus on complex relationships, achieving competitive scores across most metrics.
Evaluation: While its ROC AUC score (0.7341) and precision were strong, recall for Class 1 was relatively lower (59%). Further optimization of hyperparameters could improve recall while maintaining precision.

Random Forest

Benefit: Demonstrated balanced performance across metrics, valuable insights with feature importance.
Evaluation: Achieved the highest ROC AUC score (0.7367) and strong precision for Class 0 (72%), but recall for Class 1 remained moderate (63%). Improvements could focus on fine-tuning for minority class prediction without compromising overall accuracy.

Most Important Features

The feature importance analysis reveals that location-based features (e.g., Premis Group_Residential and Zones) and crime-specific attributes (e.g., Crime Category_Robbery & Burglary) are the most influential factors in predicting arrests. Temporal trends also play a significant role, suggesting that a deeper exploration of seasonal and geographic patterns could yield actionable insights for law enforcement strategies.



Final Model Comparison

Results Summary

Metric	Random Forest	XGBoost	Optimized XGBoost	CatBoost	Logistic Regression	Gradient Boosting	k-NN
ROC AUC Score	0.7367	0.7341	**0.7375**	0.7309	0.7057	0.7049	0.6613
Accuracy	68%	68%	68%	68%	66%	66%	62%
Class 0 Precision	**72%**	71%	71%	70%	71%	71%	69%
Class 0 Recall	71%	75%	**76%**	75%	66%	67%	63%
Class 1 Precision	62%	63%	**64%**	**64%**	59%	59%	55%
Class 1 Recall	**63%**	59%	57%	58%	**65%**	64%	61%
Class 1 F1-Score	**63%**	61%	61%	61%	62%	61%	58%

Future Improvements

Class imbalance persisted in the analysis due to the significantly higher number of non-arrest cases compared to arrests. However, the models demonstrated how valuable data collection and predictive analytics can be in identifying crime patterns and improving resource allocation.

Arrest-Related Context

Include additional datasets that capture factors influencing arrests, such as police response times, officer deployment statistics, or resource availability, to understand why certain cases result in arrests while others do not.

Expand Temporal Data

Incorporate detailed temporal features like seasonality, specific holidays, or major public events, as these may influence arrest likelihood and crime patterns, helping balance the dataset across different contexts.

Social/Behavioral Data

Datasets reflecting social dynamics, such as public complaints, community watch reports, or social media activity, to identify unreported or minor crimes that might not lead to arrests but contribute to understanding broader patterns.

Victim/Offender Profiles

Collect more granular data on victim and suspect characteristics (e.g., repeat offenses, socioeconomic background, or relationship to the crime) to better predict outcomes, particularly for cases with lower arrest probabilities.