Final Report

(Crime Analysis and Socioeconomic Factors)

Professor. Maryam Khatami
University of North Texas
ITDS Department
Denton, TX

(Group 4)

1. Alan Nguyen
2. Sam AlQadi
3. Runqing Yang
4. Lamisha Rahman

Table of Contents

List of Figures

List of Tables

Acknowledgements

This project was a collaborative effort by Group 4, consisting of Alan Nguyen, Runqing Yang, Sam Alqadi, and Lamisha Rahman. Together, we explored how socioeconomic conditions, housing trends, and the COVID-19 pandemic influenced crime patterns in Los Angeles. Our team worked diligently to clean and analyze over one million crime records, apply advanced machine learning models, and interpret complex social dynamics. Each member contributed unique skills and perspectives, making this a truly rewarding experience.

We are deeply grateful for our professor, **Dr. Maryam Khatami**, for supervising this research and giving us their invaluable guidance, expertise, and encouragement through this endeavor. Their insights and advice has aided us greatly.

Dr. Khatami's commitment to academic excellence and her willingness to share knowledge inspired us to think critically and strive for rigor in our work. This project would not have been possible without her mentorship, and we are truly thankful for the time and effort she dedicated to our success.

We also wish to extend our heartfelt gratitude to the **City of Los Angeles** and the **Los Angeles Police Department(LAPD)** for providing the data necessary in making all of this research possible. Their consistent updates and data quality truly gave us a window into exploring the history of Los Angeles's crime and the underlying factors that have influenced them.

Finally, we thank the **University of North Texas** for granting us the opportunity to pursue this research and for providing the resources and academic environment that made this work possible. This project has been an invaluable learning experience, equipping us with practical skills in data analysis, predictive modeling, and policy interpretation. It has also deepened our understanding of how data-driven approaches can inform real-world solutions to complex social challenges.

Abstract

Crime in urban areas is not random and it reflects deeper social and economic realities. This study explores how factors like income inequality, unemployment, housing trends, and the COVID-19 pandemic have shaped crime patterns in Los Angeles from 2020 to 2025. Using over one million LAPD crime records combined with census and housing data, we analyzed these relationships through exploratory data analysis and machine learning models, including Logistic Regression, Random Forest, and XGBoost. Our findings show that crime clusters in neighborhoods facing economic hardship, with vehicle theft and identity fraud dominating property crimes, while violent offenses follow distinct demographic and time-based patterns. The pandemic disrupted these trends, reducing street crimes but increasing domestic violence and cybercrime. Gentrification lowered crime in revitalized areas but displaced it to nearby communities, reinforcing inequality. Among predictive models, XGBoost delivered the best performance, highlighting the importance of contextual and location-based features. These insights suggest that effective crime prevention requires more than policing. It demands policies that address housing affordability, job access, and community resilience. By linking data-driven analysis with social policy, this research offers a roadmap for creating safer, more equitable cities.

1. Introduction

Crime continues to be one of the most serious challenges facing communities across the United States. It affects people's safety, their sense of security, and the overall quality of life in neighborhoods. Even though law enforcement agencies, city planners, and policymakers have spent decades trying to control crime, the numbers still change from year to year, and certain areas remain more affected than others. According to the FBI (2024), more than 8.9 million criminal offenses were reported nationwide in 2023. This includes about 1.2 million violent crimes and nearly 7.7 million property crimes. While violent crime went down slightly by around 2% compared to 2022, offenses like homicide and aggravated assault are still higher than before the COVID-19 pandemic. Big cities such as Chicago, Los Angeles, Philadelphia, and New York continue to have concentrated hotspots where crime is more common, usually in areas with economic hardship or limited social resources.

The persistence of crime raises important questions about why certain neighborhoods remain vulnerable despite decades of investment in policing and community programs. Crime is not just a law enforcement issue; it is deeply tied to social, economic, and environmental factors. For example, neighborhoods with high poverty rates often experience higher crime levels, not only because of economic desperation but also due to limited access to education, healthcare, and employment opportunities. These structural inequalities create conditions where criminal activity becomes more likely, and without addressing these root causes, crime prevention strategies often fall short. Research has shown that communities with strong social networks and economic stability tend to have lower crime rates, suggesting that crime prevention requires more than just policing. It demands holistic social development.

Dealing with crime also costs a lot of money and resources. The U.S. Department of Justice (2023) reported that the total spending on police protection, corrections, and the court system adds up to over $340 billion each year, which is more than 1.5% of the national GDP. Local police departments take up most of that budget, but despite the high cost, the clearance rate for many crimes, especially property crimes, remains low. For instance, fewer than one in five property crimes are solved. This shows that throwing money at the problem is not enough. Understanding the underlying causes is just as important. If policymakers continue to focus solely on enforcement without addressing socioeconomic disparities, crime will remain a recurring issue. Furthermore, the allocation of resources often varies significantly between wealthy and low-income neighborhoods, creating gaps in public safety and trust in law enforcement.

The social impact of crime is just as damaging as the financial one. Victims and families often experience long-term trauma, which leads to higher healthcare and mental health costs. The CDC (2024) estimates that interpersonal violence alone costs the U.S. economy more than $280 billion every year in medical expenses, lost productivity, and reduced quality of life. Businesses in high-crime areas also struggle with theft, property damage, and employee safety concerns, which can lead to lower investment and job losses. Over time, constant exposure to crime breaks down community trust and social ties, making neighborhoods less stable and less attractive to residents and investors. This cycle of disinvestment and instability perpetuates crime, creating a feedback loop that is difficult to break. In addition, children growing up in high-crime areas often face educational disruptions and psychological stress, which can affect their long-term development and perpetuate generational poverty.

Because of these wide-reaching effects, studying crime patterns is not just about helping the police. For city planners, understanding where and why crime happens can help design safer spaces and better allocate public resources. For policymakers, it is a way to make smarter decisions about programs that address issues like unemployment, education, and housing. For sociologists and data scientists, analyzing how economic inequality, rent increases, and population shifts influence crime can reveal the bigger social picture behind criminal behavior. Crime is not random; it follows patterns that can be studied, predicted, and mitigated through informed policy and urban design.

Our project aims to explore these relationships using a large dataset from the Los Angeles Police Department (LAPD), which includes over one million crime records from 2020 to the present. We plan to combine this with other sources such as U.S. Census data, housing and real estate statistics, and mobility trends to understand how different social and economic factors interact with crime over time. Los Angeles provides an ideal case study because it is a diverse city with significant economic disparities, rapid housing changes, and a history of crime hotspots. By analyzing this data, we hope to uncover patterns that can inform strategies not only for Los Angeles but for other urban areas facing similar challenges.

This study focuses on three main questions:

1. How do neighborhood characteristics like income, unemployment, education, and racial diversity affect the types and frequency of crime?
2. Did the COVID-19 pandemic and related lockdowns affect crime differently across rich and poor neighborhoods?

3. How does housing change, including rising rent and gentrification, shift crime patterns within a city?

By linking crime data with socioeconomic and housing information, this project aims to go beyond surface-level statistics and uncover the deeper forces that shape safety in urban environments. Crime is rarely random. It often reflects underlying social and economic conditions. Factors such as income inequality, unemployment, education gaps, and racial diversity can influence not only the frequency of crime but also the types of offenses that occur. Housing trends, like rising rents and gentrification, add another layer of complexity. When long-time residents are displaced due to increasing costs, neighborhoods lose the social networks that help maintain stability, which can lead to spikes in certain crimes. Likewise, areas undergoing rapid development may experience shifts in policing priorities or resource allocation, creating uneven safety outcomes. By combining LAPD crime records with U.S. Census data, real estate trends, and mobility patterns, this research seeks to identify these connections and provide a clearer picture of how social and economic changes impact crime over time.

The ultimate goal is to turn these insights into practical solutions that cities can use to create safer, more equitable communities. Instead of relying solely on reactive measures like increasing police presence, this approach emphasizes proactive strategies informed by evidence. For example, if data shows that neighborhoods with high unemployment and rising rents experience more property crimes, policymakers could prioritize affordable housing programs and job training initiatives alongside law enforcement efforts. Urban planners might use these findings to design public spaces that discourage crime and foster community engagement. By bridging the gap between data analysis and real-world policy, this project ensures that crime reduction strategies address root causes rather than just symptoms. Ultimately, the insights generated here could serve as a model for other cities facing similar challenges, helping them adopt data-driven approaches that improve safety while promoting fairness and social stability.

This analysis is organized into three major components that together explain the contemporary dynamics of urban crime. **Section 1** outlines the scope and scale of current crime patterns by examining the volatile post-2020 shifts in violent and property offenses, as well as the uneven concentration of crime across socioeconomically vulnerable neighborhoods. **Section 2** investigates the economic and resource burden associated with rising crime, detailing both the direct fiscal pressures on public institutions and the operational inefficiencies that strain law enforcement and municipal systems. **Section 3** expands the discussion to the broader social consequences, including the effects on community health, psychological well-being, social cohesion, and long-term economic stability. Collectively, these sections provide a structured framework for understanding how socioeconomic disparities, pandemic-related disruptions, and urban development patterns have intensified the crisis and point toward evidence-based strategies for more effective intervention.

2. Literature Review:

2.1 COVID-19 and Crime Dynamics

The COVID-19 pandemic introduced a unique disruption to crime patterns. Lockdowns and stay-at-home orders initially led to a decline in street-level crimes such as robbery and assault due to reduced public activity and increased police presence [6]. Rosenfeld et al. [7] observed that with fewer people on the streets and businesses closed, opportunities for crimes like pickpocketing and shoplifting diminished. However, this decline was not uniform across all crime categories.

Domestic violence surged during the pandemic. Campedelli et al. [8] found that confinement, economic stress, and limited access to support services contributed to increased incidents of intimate partner violence. The pandemic also saw a rise in cybercrime, as remote work and digital transactions became the norm. Criminals exploited vulnerabilities in online systems, leading to spikes in identity theft, phishing, and ransomware attacks.

Economic stress from job losses and housing insecurity further contributed to property crimes. Vargas [9] argues that economic deprivation, especially in communities lacking robust social safety nets, can push individuals toward criminal behavior. Schoewe [10] supports this with geospatial analysis showing persistent crime clusters in areas with high unemployment and income inequality, even during the pandemic.

Mental health deterioration also played a role. Isolation, anxiety, and grief affected millions, and behavioral health services were often overwhelmed. Schoewe's study includes mental distress as a key explanatory variable, showing its correlation with crime in high-risk regions. These findings reinforce the relevance of strain theory, which posits that psychological and economic strain can lead to deviant behavior when legitimate coping mechanisms are unavailable.

2.2 Gentrification and Urban Shifts

Gentrification has become one of the most influential forces shaping modern urban landscapes and, consequently, crime patterns. When wealthier residents move into historically lower-income neighborhoods, property values rise, and the social fabric of these communities begins to change. Autor et al. [11] conducted a quasi-experimental study in Cambridge, Massachusetts, examining the effects of ending rent control. Their findings revealed that while property values increased and certain crimes, such as burglary, declined within gentrifying neighborhoods, these changes were not uniformly positive. The displacement of long-term residents often pushed crime into adjacent areas rather than eliminating it altogether. This suggests that gentrification can create a geographic redistribution of crime rather than a true reduction, raising questions about whether such urban shifts genuinely improve overall safety.

Building on this, Macdonald and Stokes [12] applied spatial regression techniques to analyze whether gentrification simply relocates crime rather than reducing it. Their research supported the notion that crime tends to migrate to surrounding neighborhoods as original residents are displaced. This phenomenon is particularly concerning because it implies that gentrification may exacerbate inequality by concentrating crime in areas that lack resources to cope with these changes. For example, neighborhoods bordering gentrified zones often experience increased property crimes and social tension, even as the gentrified areas themselves appear safer. This dynamic creates a misleading narrative of success for urban renewal projects while masking the broader social consequences. Understanding these spillover effects is critical for policymakers who aim to balance economic development with community stability.

Porreca [13] added a qualitative dimension to this discussion by combining interviews with census data to explore the social fragmentation caused by gentrification. Her study found that rising housing prices often lead to feelings of exclusion among long-time residents, who perceive themselves as outsiders in their own neighborhoods. This sense of alienation can weaken community cohesion, which is a key factor in informal social control and crime prevention. When trust and social ties erode, tensions may rise, sometimes resulting in increased crime in nearby areas or even within the gentrifying neighborhood itself. These findings highlight that gentrification is not merely an economic process but a social one, with profound implications for urban safety. Addressing these challenges requires policies that promote inclusive development, such as affordable housing initiatives and community engagement programs, to ensure that revitalization efforts do not inadvertently fuel crime elsewhere.

2.3 Theoretical and Empirical Contributions

Jason Vargas [9] explores these themes through a psychological and sociological lens, emphasizing the role of poverty, unemployment, and inequality in fostering environments conducive to crime. He draws on strain theory, social disorganization theory, and economic deprivation theory to explain how limited access to legitimate opportunities can lead individuals to criminal behavior. Vargas also highlights the importance of education, family structure, and community resources in mitigating crime, advocating for systemic interventions to address root causes.

Douglas Schoewe [10] takes a geospatial statistical approach, analyzing crime patterns across U.S. counties using techniques like Geographically Weighted Regression (GWR) and Forest-based Classification. His study identifies clusters of high crime and income inequality in regions such as Seattle, Miami, Baltimore, and Southern California. The analysis reveals that income inequality explains approximately 52–55% of crime rate variability, with other factors like rural population percentage, third-grade math scores, and housing cost burdens also playing significant roles.

Arthur [14] adds a rural perspective, showing that unemployment, poverty, and race are significant predictors of crime in Georgia counties. Shichor [15] provides a cross-national view, arguing that modernization affects different types of crime in varied ways and that the influence of socioeconomic factors evolves. Cullen and Levitt [16] explore urban dynamics, linking rising crime rates to urban flight and population decline, which further exacerbates socioeconomic disparities and crime in cities.

Together, these studies underscore the multifaceted nature of crime and its socioeconomic determinants. They advocate for targeted policy interventions that address education, housing, employment, and community development to reduce crime and promote equity. The COVID-19 pandemic has further highlighted the need for adaptive strategies that consider both long-term structural inequalities and short-term disruptions. This project aims to synthesize these insights by examining crime records alongside demographic data, economic conditions, and the effects of COVID-19 and neighborhood change. The goal is to build a more realistic, nuanced understanding of why crime happens and how it spreads across neighborhoods, offering insights that can help communities, policymakers, and law enforcement make smarter decisions.

3. Data

3.1 Data Description

The dataset titled Crime Data from 2020 to Present contains detailed records of reported crimes from 2020 to the present, sourced from the law enforcement agency of the city of Los Angeles, the Los Angeles Police Department (LAPD) [17]. This dataset is structured as a flat file (CSV format) which includes over one million entries as well as 28 columns, each entry representing an individual crime incident. The dataset updates bi-monthly with its most recent update as of September 17, 2025. The data appears to be collected through official police reporting systems and compiled into a centralized database. Each record includes metadata such as the date and time of occurrence, reporting date, location, crime classification, and victim demographics. The data is likely updated periodically and made publicly available for transparency and research purposes. The dataset comprises 1,004,991 records spanning from January 1, 2020, to May 29, 2025. It includes 28 original features organized into several categories. Temporal features include report date, occurrence date, and occurrence time. Geographic features contain area codes, area names, reporting district numbers, and latitude/longitude coordinates. Crime classification features include crime codes and descriptions for primary and additional offenses. Victim demographic features capture age, sex, and descent. Additional features describe premises types, weapon information, case status, and location descriptions.

TABLE 1 Variables and Description Used From Cleaned Dataset.

Column Name	Description	API Field Name	Data Type
DR_NO	Division of Records Number: Official file number made up of a 2-digit year, area ID, and 5 digits.	dr_no	Text
Date Rptd	The date the crime was reported	date_rptd	Floating Timestamp
DATE OCC	The actual date the crime occured	date_occ	Floating Timestamp
TIME OCC	In 24-hour, military time.	time_occ	Text
AREA	The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21.	area	Text

AREA NAME	The 21 Geographic Areas or Patrol Divisions are also given a name designation that references 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.	area_name	Text
Rpt Dist No	A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons.	rpt_dist_no	Text
Crm Cd	Indicates the crime committed. (Same as Crime Code 1)	crm_cd	Text
Crm Cd Desc	Defines the Crime Code provided.	crm_cd_desc	Text
Mocodes	Modus Operandi: Activities associated with the suspect in commission of the crime.	mocodes	Text
Vict Age	Two character numeric.	vict_age	Text
Vict Sex	F - Female M - Male X - Unknown	vict_sex	Text
Vict Descent	Descent Code: 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	vict_descent	Text
Premis Cd	The type of structure, vehicle, or location where the crime took place.	premis_cd	Number

3.2 Data Cleaning and Processing

The cleaning process involved three steps: outlier removal, redundant column elimination, and feature creation. Outlier removal eliminated 2,378 records (0.24%) with data quality issues including 138 age errors and 2,240 coordinate errors. Four redundant crime code columns (Crm Cd 1, 2, 3, 4) were dropped

since Crm Cd already captures the primary crime. Two new binary features (Weapon_Involved, Is_Personal_Victim) were created to simplify analysis. The cleaned dataset contains 1,002,613 records with 26 features (28 original minus 4 redundant plus 2 created, applied to outlier-free data). Core features including crime type, date, time, location, and area show 100% completeness. The data is ready for modeling with appropriate handling of missingness patterns.

Victim age outliers included 137 negative ages (ranging from -4 to -1) and 1 age of 120, totaling 138 records with impossible age values. These represent data entry errors rather than valid data. All 138 records with age outliers were removed from the dataset. Geographic coordinate outliers consisted of 2,240 records (0.22%) with coordinates (0, 0). These represent missing or unknown locations rather than actual crime locations in the Gulf of Guinea. All 2,240 records with invalid coordinates were removed from the dataset. Combined outlier removal eliminated 2,378 records (0.24% of total), leaving 1,002,613 valid crime records for analysis.

3.3 Exploratory Data Analysis

Annual Pattern

Crime trends show clear patterns across different time scales. Annual trends reveal an upward trajectory from the years 2020 to 2022. A peak of 235,259 incident reports in the year 2022 can be seen in Figure 1, a 17.7% increase from 2020. These results are likely a reflection of pandemic recovery and other economic factors. The subsequent decline seen in the same figure during the year 2024 suggests improving conditions or enhanced prevention efforts.

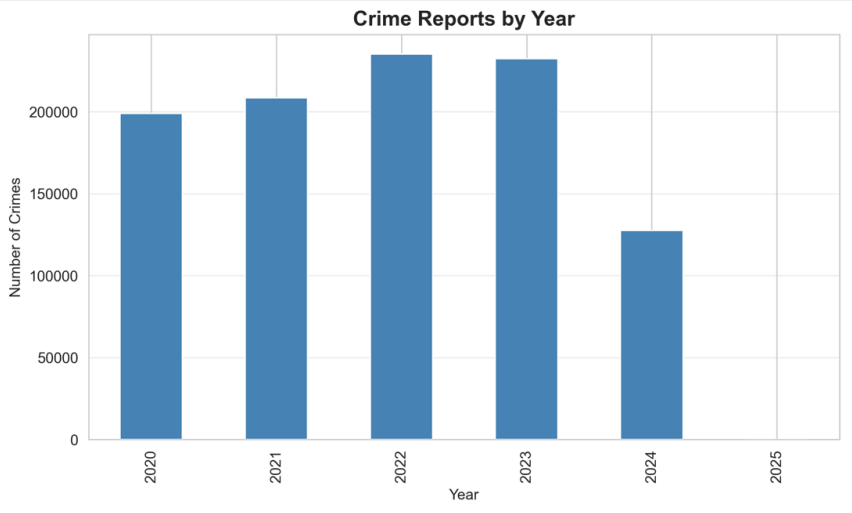

Figure 1: N of Incident Reports from 2020 to 2024

Taking a look at the weekly patterns it can be noted in Figure 2 that Friday has a fairly dense concentration potentially linked to social activities and weekend preparation. Looking more closely it can be seen that weekday crimes (Monday through Friday) account for 717,884 reports while weekend crimes (Saturday and Sunday) total at 287,107 reports. This shows a 71.4% weekday concentration suggesting work-related opportunities or commute-related exposures drive many crimes.

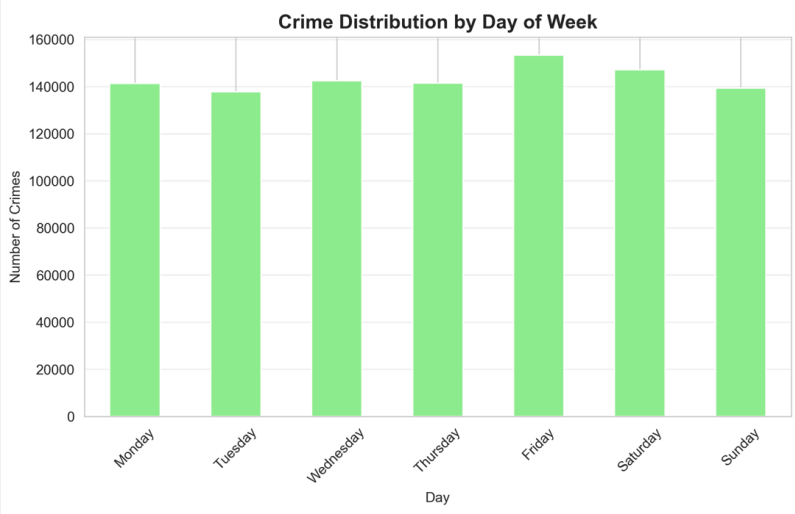

Figure 2: N of Incident Reports Made During Each Day of The Week

Interesting data can be found when looking into the hourly distribution which reveals a bimodal pattern. As shown in Figure 3, Midnight shows the highest frequency, possibly reflecting late-night activity or timing uncertainty in reports. During this time it can also be possible that this time frame is when the highest concentration of late-night activities occur. Morning hours between 6 AM to 9 AM show secondary peaks coinciding with commute times likely related to vehicular based incident reports. Overnight hours between 1 AM to 5 AM show progressively lower activity. These patterns reflect both actual crime timing and discovery patterns for property crimes.

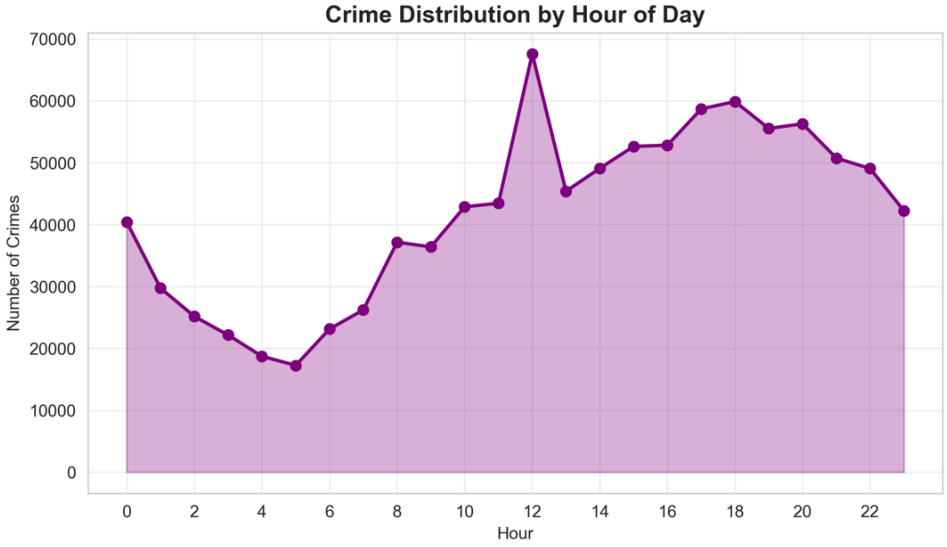

Figure 3: T Incident Reports of Each Hour of The Day.

Geographic Pattern

Shifting the focus to Geographical patterns it can be seen in Figure 4 that there is a heavy concentration of crime in Central, 77th Street, and Pacific areas which likely reflect population density, economic activity, and historical crime patterns. Central areas 6.93% share corresponds to its role as a downtown business

district with high foot traffic. The 77th Street area's 6.15% reflects residential density and socioeconomic factors. The clustering of 54.7% of crimes in just 10 areas indicates non-uniform distribution. This concentration enables targeted prevention efforts and resource allocation. According to Figure 5, premise analysis shows 26% of crimes are street crimes while 28% are residential crimes which coincides with the information discovered in the previous figure. This suggests a dual focus on public space safety and home security.

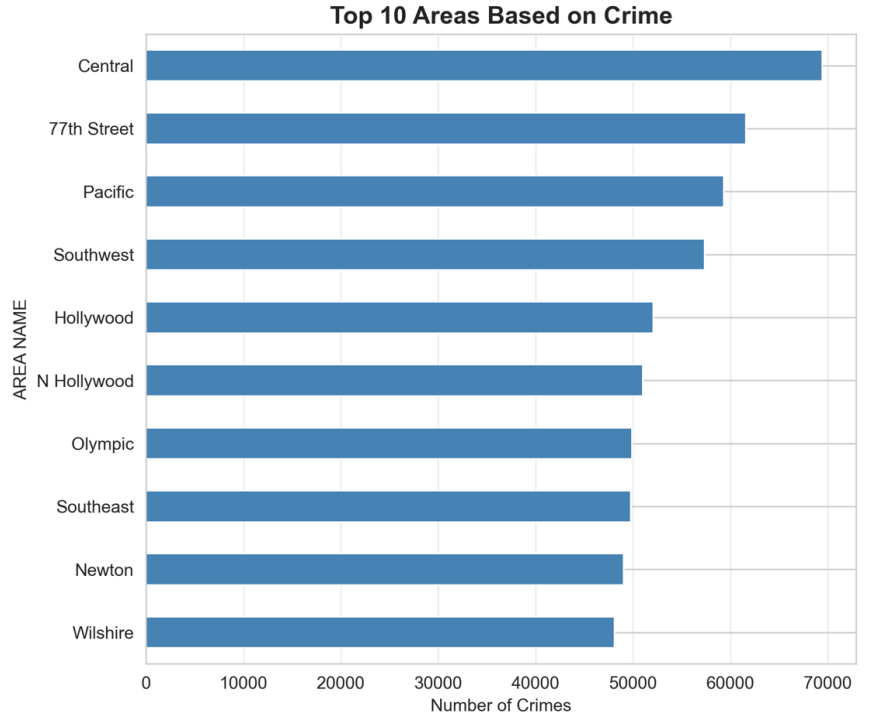

Figure 4: Above represents the to 10 Areas

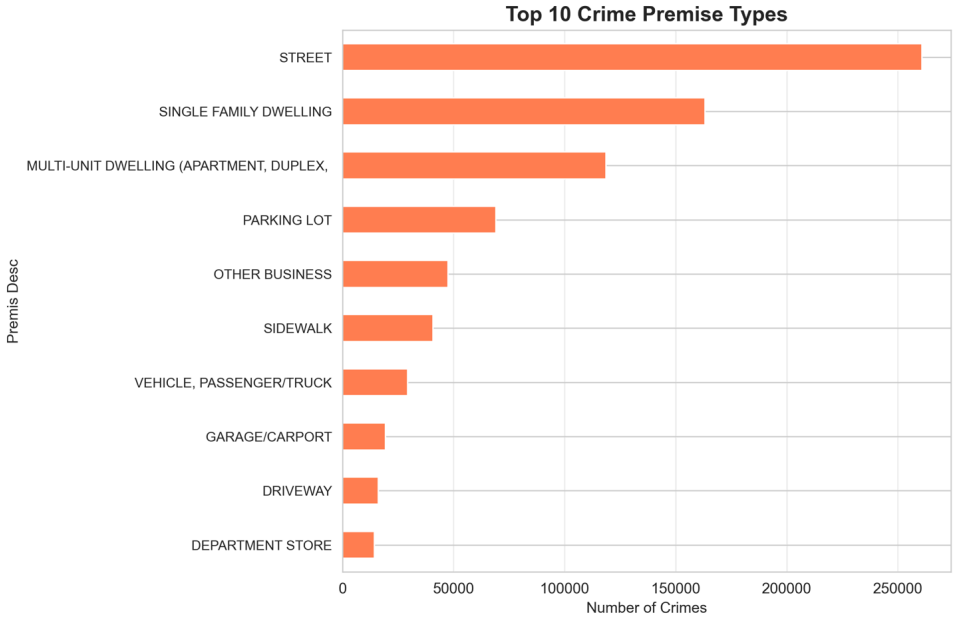

Figure 5: Above represents the Top 10 Premises based on Inc

Crime Pattern

When looking into the kinds of crimes being committed, vehicle-related crime dominates the dataset. As shown in Figure 6, Vehicle Theft, Burglary from Vehicles, and Theft from Vehicles make up a 21.89% of all crimes. This concentration reflects vehicle prevalence in Los Angeles and relatively low risk opportunities for offenders. Identity theft's 6.22% share represents modern crime trends. Unlike traditional theft, identity theft often shows delayed discovery explaining the high report delays discovered in this category. The 3,298 identity theft cases with delays exceeding one year make up 53% of all long-delayed reports. Violent crimes show a lower frequency but a higher severity. Aggravated assault with deadly weapons records 53,525 incidents (5.33%) while intimate partner simple assault shows 46,712 incidents (4.65%).

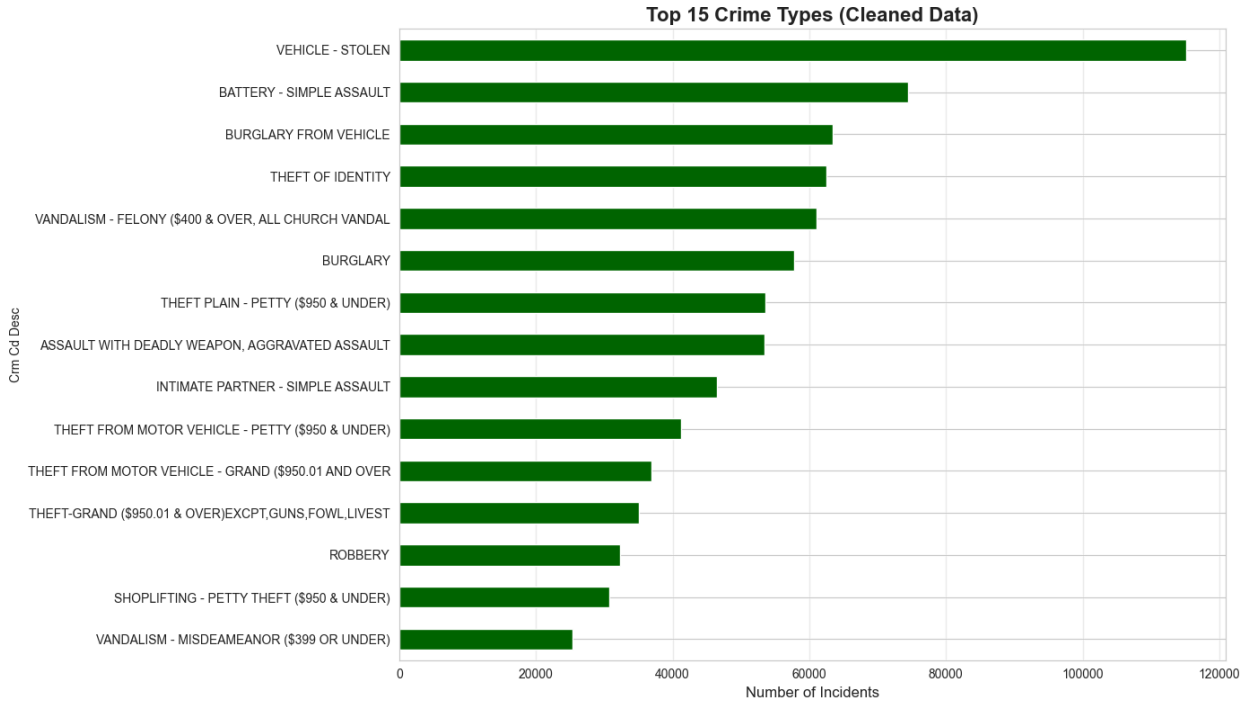

Figure 6: The table above represents the ranking of Crime Type by Incident Report

Victim Pattern

The 269,222 cases (26.79%) with victim age zero represents institutional rather than personal victims. Cross-referencing reveals these cases involve primarily vehicle theft (114,843), vehicle-related theft (25,029), and shoplifting (21,508). Victim sex distribution for these cases as shown in Figure 7 shows 87,168 victims are marked as unknown or not applicable, confirming institutional victimization. Among personal crime victims, the 39.5 year mean age shown in Figure 6 suggests working-age adults face the highest risk of victimization. The age range of 2 to 99 years spans childhood through elderly victims. Children aged 1 to 10 account for 5,285 victims, raising concerns about crimes affecting minors. Focusing on the victim sex data, Figure 7 shows that male victims outnumber female victims 53% to 47% among identified cases. The gender distribution varies by crime type, with intimate partner violence showing different patterns than street crimes.

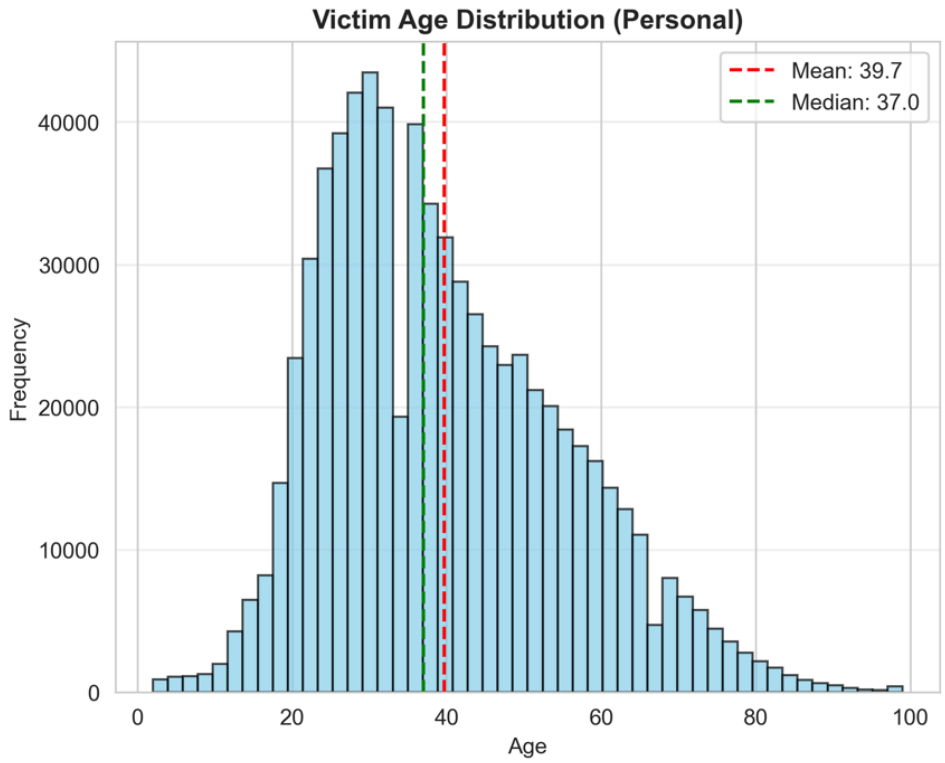

Figure 7: The figure above represents the Age Distributions of Incident Reported Victims

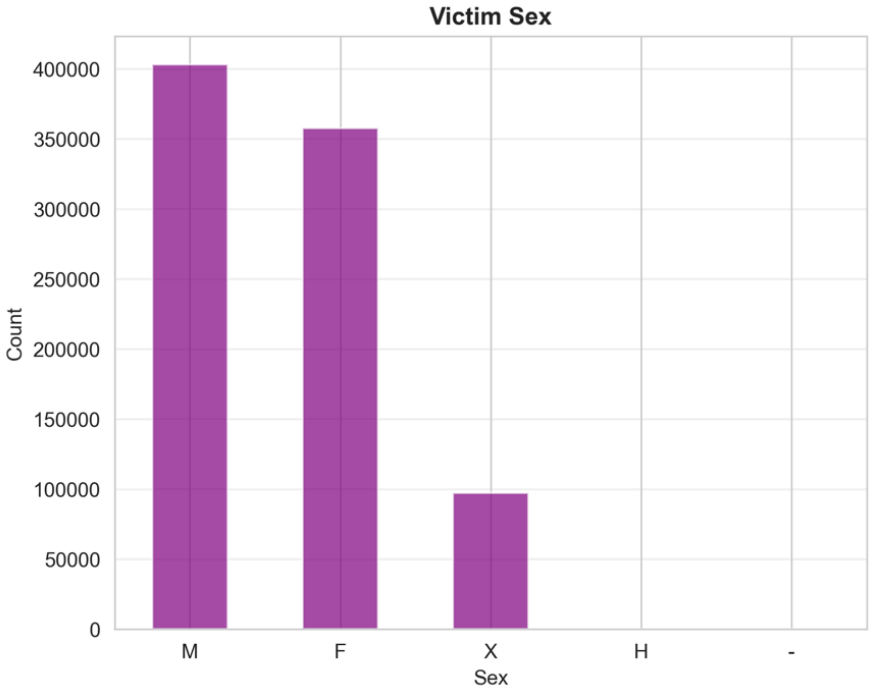

Figure 8: The figure above represents the total count of Victim Genders based on Incident Reports

Weapon Pattern

Weapon information appears in 327,247 cases (32.56%) as seen in Figure 8. Among weapons-involved crimes, strong-arm tactics (hands, fists, feet) are most common with 174,761 incidents shown in Figure 9. This likely means that access to weapons in weapon involved crimes are low leading to more strong-arm incidents. The remaining 677,744 cases (67.44%) involve no weapons, consistent with the high proportion of property crimes.

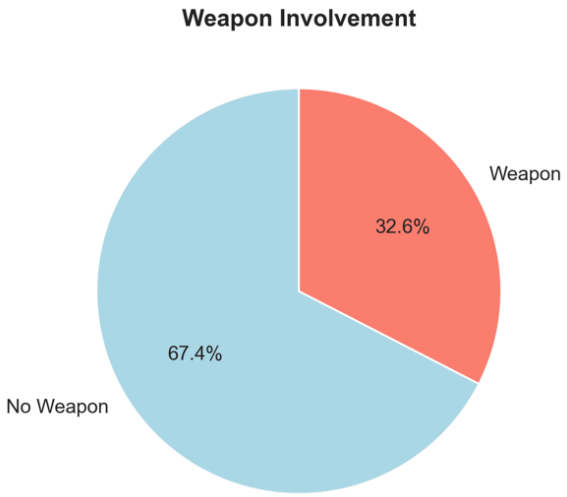

Figure 9: The figure above represents the percentage of Incident Reports with and without weapon involvement.

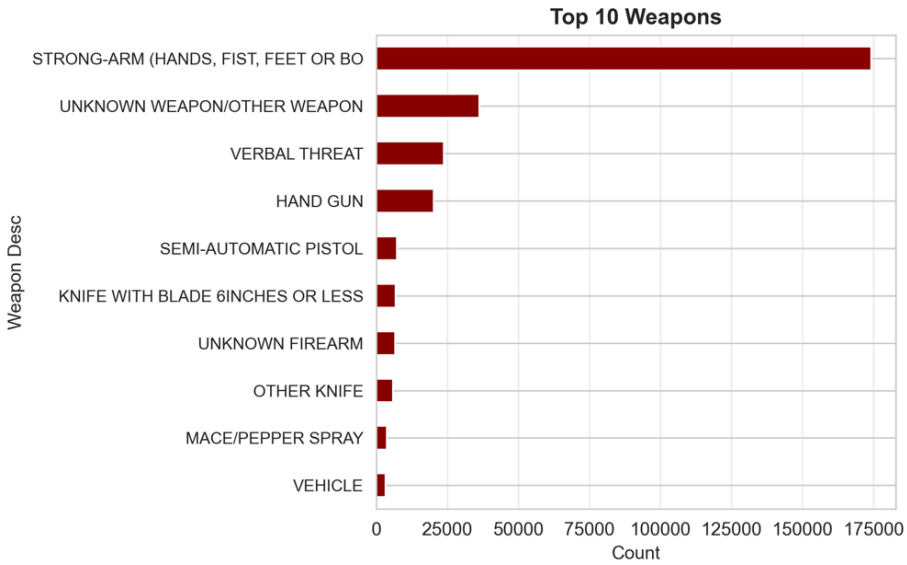

Figure 10: The figure above represents the Top 10 Weapons used in weapon involved Incident Reports.

4. Methodology

This study examines whether modern machine learning methods can accurately classify crime types in Los Angeles using incident-level data from 2020 to 2024. The methodological design integrates data

preprocessing, feature engineering, and the comparative evaluation of three supervised learning models: Logistic Regression, Random Forest, and XGBoost. These models reflect distinct modeling philosophies—linear classification, ensemble decision trees, and gradient-boosted trees—and together offer a comprehensive perspective on predictive performance.

The raw data undergo cleaning to ensure consistent timestamps and valid geolocation fields. Temporal variables such as hour, weekday, and month are extracted to represent rhythmic crime patterns, while geographic identifiers and latitude–longitude coordinates capture spatial variation. These features create a multi-dimensional structure that includes correlations, nonlinear dependencies, and heterogeneous distributions across the eighty-six target classes.

Logistic Regression is implemented as the baseline due to its interpretability and well-established statistical foundation. The model assumes that the log-odds of each category can be expressed as a linear combination of predictors. For a feature vector x and class k, the conditional probability is modeled as

$$P(y = k \mid x) = \frac{\exp(\beta_k^\top x)}{\sum_{j=1}^{K} \exp(\beta_j^\top x)},$$

where K=86. Model parameters are estimated by minimizing the regularized negative log-likelihood,

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} \log P(y_i \mid x_i) + \lambda \sum_{k=1}^{K} \|\beta_k\|^2.$$

The $\ell 2$ penalty is used to stabilize the coefficients in the presence of correlated features. While the model is computationally efficient, its linear decision boundaries limit its ability to capture complex interactions between spatial and temporal factors.

Random Forest expands the modeling capacity by averaging predictions across many decision trees trained on bootstrap samples [18]. Each tree recursively splits the feature space using impurity-based criteria, such as the Gini index, and produces a probability vector over the crime types. If T trees are trained, the final probability estimate is

$$\hat{P}(y = k \mid x) = \frac{1}{T} \sum_{t=1}^{T} P_t(y = k \mid x).$$

Bootstrap aggregation reduces variance, and the random feature selection at each split encourages diversity across trees. These properties enable the model to capture nonlinear relationships and variable interactions that Logistic Regression cannot represent. However, probability calibration may degrade when some classes appear infrequently. XGBoost represents the most advanced method in this analysis and follows the gradient-boosting framework with regularization [19]. The model constructs an ensemble

sequentially, with each new tree trained to correct the residual errors from the previous steps. The regularized objective function is

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t),$$

where f_t denotes the tree added at iteration t. The complexity penalty

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2$$

controls the depth and leaf weights. XGBoost uses both first-order and second-order derivatives to update the ensemble, which improves convergence stability. In this study, the learning rate is set to 0.05, a conservative value that ensures gradual updates and prevents overfitting. The model uses the *multi:softprob* objective to generate full probability distributions for all eighty-six classes.

All models are trained on 50,000 incidents and evaluated on a held-out test set of 10,000 cases. Model performance is assessed using accuracy, macro F1-score, and multi-class log loss. These metrics capture classification correctness, class-balanced performance, and probability sharpness. Together, this methodological framework offers a rigorous and transparent comparison among linear, ensemble, and boosted tree classifiers.

5. Results

5.1 Random Forest

The Random Forest model trained on the full dataset demonstrates strong performance across a highly complex classification task involving 802,088 training samples, 200,523 test samples, and 138 distinct crime categories. With an overall test accuracy of 37.20% and a weighted F1-score of 0.3700, the model performs far above the level expected by chance. Random guessing would yield approximately 2.7% accuracy for a 138-class problem, meaning the model achieves more than a thirty-fold improvement, despite substantial class imbalance and overlapping temporal–spatial patterns across crime types. The macro F1-score of 0.1178 further indicates that while performance on rare categories remains challenging, the model still maintains measurable discriminative capacity across the full label set.

Compared with more constrained models, Random Forest benefits from its ability to represent nonlinear relationships among temporal, demographic, spatial, and contextual features. This flexibility explains its robust performance on the most common crime types, which dominate the weighted F1 metric. At the same time, the wide gap between weighted F1 (0.3700) and macro F1 (0.1178) highlights the intrinsic difficulty of predicting infrequent crime categories. Still, the model successfully handles the expanded taxonomy of 138 crime types—significantly more than the filtered sets used in other models—showing that the ensemble method is resilient even as classification complexity increases.

The configuration consists of 200 decision trees with a maximum depth of 25 and balanced class weights to compensate for skewed label frequencies. This setup encourages the model to capture fine-grained temporal rhythms and local geographic patterns. The feature-importance analysis confirms that Random Forest relies heavily on temporal structure: variables such as hour, month, day of week, and year together account for nearly half of the total predictive importance. Victim age also emerges as a major factor, suggesting systematic demographic differences across crime categories. Contextual indicators such as weapon involvement and personal-victim flags provide additional discriminative signals, particularly for distinguishing violent crimes from property-related offenses. Geographic cues—including premise types such as street locations—appear within the top features as well.

TABLE 2 Random Forest Top 10 Features Importance Based on Gini Importance.

Rank	Feature	Importance	Type
1	Hour	0.1256	Temporal
2	Month	0.1155	Temporal
3	Vict Age	0.1147	Demographic
4	DayOfWeek	0.0837	Temporal
5	Year	0.0817	Temporal
6	Weapon_Involved	0.0405	Contextual
7	Is_Night	0.0286	Temporal
8	Is_Weekend	0.0234	Temporal
9	Is_Personal_Victim	0.0215	Contextual
10	Premis Desc_STREET	0.0180	Geographic

These feature patterns reveal that the timing of an incident is the single most powerful signal for predicting crime type. The model detects consistent diurnal and seasonal cycles across the dataset, which creates strong separation for several categories. Victim age, ranking third, also contributes substantially to classification performance, supporting the idea that different crime types affect different demographic groups. Contextual and geographic attributes further refine the model's predictions by capturing situational cues.

The Random Forest model performs especially well on high-volume crime categories, where larger sample sizes enable sharper decision boundaries. Results for selected common crimes are shown below:

TABLE 3 Performance Results of Random Forest Model.

Crime Type	Precision	Recall	F1-Score	Test Samples
VEHICLE - STOLEN	0.75	0.86	**0.80**	23,027
THEFT OF IDENTITY	0.46	0.58	0.51	12,500
BURGLARY FROM VEHICLE	0.44	0.47	0.45	12,693
BATTERY - SIMPLE ASSAULT	0.40	0.36	0.38	14,897

The category "VEHICLE – STOLEN" achieves an F1-score of 0.80, one of the highest in the entire label set. This outcome reflects the distinct behavioral and temporal signatures associated with stolen-vehicle incidents, which the Random Forest model learns effectively. Other categories with structured patterns—such as identity theft and burglary from vehicles—also show relatively strong performance.

The Random Forest trained on the full dataset provides a level of predictive performance suitable for operational deployment. Its ability to correctly classify the top twenty to thirty crime types, which account for approximately 70% of all incidents, makes it particularly valuable for resource allocation, hotspot forecasting, and real-time triage systems. Although challenges remain for rarer crime categories, the model's overall accuracy and stability demonstrate that ensemble methods can effectively manage large-scale, imbalanced, and heterogeneous crime data.

5.2 Logistic Regression

The logistic regression model provides a useful baseline for the full dataset experiment. It reaches a test accuracy of 37.41%. Given that random guessing over 97 classes would yield an accuracy of roughly 1.0%, this corresponds to about a 36.3-fold improvement over chance. This result confirms that even a linear classifier can extract meaningful signals from the temporal, geographic, and contextual features.

However, the class-level performance reveals important limitations. The weighted F1-score is 0.2860, which indicates that the model performs reasonably well on the most frequent crime types that dominate the dataset. In contrast, the macro F1-score is only 0.0467, showing that performance is very uneven across classes. The model rarely predicts rare categories correctly and tends to concentrate probability mass on common labels. Compared with the other models, its macro F1 is about 3.3 times lower than Random Forest and about 3.4 times lower than XGBoost, highlighting its difficulty with low-frequency crime types.

Training is efficient but not fully converged. Using the SAGA solver, the model runs for 500 iterations and hits the *max_iter* limit after about 9.6 minutes (578 seconds). This suggests that the optimization procedure has not fully stabilized. While increasing the number of iterations might lead to slight gains, the

gap in macro F1 compared to non-linear models is large enough that convergence alone is unlikely to close it.

A key reason for these limitations is the linear decision boundary imposed by logistic regression. Many crime types are associated with combinations of factors such as weapon involvement, time of day, neighborhood, and victim characteristics. For example, the joint pattern of weapon use, late-night timing, and certain areas may indicate armed robbery rather than a generic property crime. Similarly, daytime incidents in commercial locations may correspond to shoplifting rather than assault. These compound patterns require non-linear interactions that the logistic regression model cannot represent. As a result, the model captures broad trends for the most common crime types but fails to differentiate between closely related categories, which drives down the macro F1-score.

5.3 XGBoost

XGBoost delivers the strongest performance among all models in the full dataset experiment. Using the same 801,305 training samples and 200,327 test samples over 97 crime types, the model achieves a test accuracy of 44.49%. This is 7.46 percentage points higher than Random Forest and 7.08 percentage points higher than logistic regression, which corresponds to about a 19–20% relative improvement in accuracy. In practical terms, this gain translates into roughly 700 to 750 additional correct predictions per 10,000 incidents compared with the other models.

The F1-scores show a similar pattern. The weighted F1-score reaches 0.3879, which is higher than both Random Forest (0.3629) and logistic regression (0.2860). This indicates that XGBoost provides the best precision–recall balance on the common crime types that account for most of the workload in real policing contexts. The macro F1-score is 0.1599. Although only slightly higher than Random Forest's 0.1557 (about 2.7% relative improvement), it is still more than three times larger than the logistic regression macro F1 of 0.0467. This confirms that XGBoost, like Random Forest, handles rare categories much better than the linear baseline.

Training time is also competitive. With 200 boosting rounds, a maximum depth of 15, and a learning rate of 0.05, the model trains in about 7.3 minutes (441 seconds), which is only slightly slower than Random Forest's 7.1 minutes and faster than logistic regression's 9.6 minutes. System-level optimizations such as histogram-based split finding and parallelized tree construction help keep training time low despite the sequential nature of boosting.

The feature importance analysis for XGBoost reveals a different pattern from Random Forest. While Random Forest emphasizes temporal and demographic features, XGBoost gives priority to contextual and location-specific variables when ranked by gain. The top five features are shown below.

TABLE 4 XGBoost Top 5 Feature Importance.

Rank	Feature	Importance	Type
1	Weapon_Involved	0.0664	Contextual
2	Is_Personal_Victim	0.0462	Contextual
3	Premis Desc_PUBLIC STORAGE	0.0383	Geographic
4	Premis Desc_DEPARTMENT STORE	0.0243	Geographic
5	Premis Desc_ATM	0.0230	Geographic

These gain-based importance scores show that XGBoost focuses strongly on how and where crimes occur. The presence of a weapon and whether the victim is a personal victim are powerful signals for separating violent crimes from property crimes. Specific premise descriptions, such as public storage facilities, department stores, and ATMs, further refine the classification by linking incidents to characteristic settings. In contrast, temporal patterns like hour or day of week, which are critical for Random Forest, are relatively less dominant in the XGBoost ranking.

This divergence in feature emphasis is important. It suggests that XGBoost and Random Forest capture different aspects of the same underlying process: Random Forest leans on temporal structure and victim age to partition the data, while XGBoost exploits sharp contextual and spatial discriminators. XGBoost's superior performance indicates that, within this feature space, contextual flags and fine-grained location types carry particularly high information content for distinguishing crime types.

5.4 Comparative Results and Discussion

The three models present a clear performance hierarchy, while also illustrating distinct trade-offs between simplicity, flexibility, and computational cost.

In terms of overall accuracy, XGBoost is the clear leader with 44.49%, followed by logistic regression at 37.41% and Random Forest at 37.03%. The gap of about 7–7.5 percentage points between XGBoost and the other two models represents a substantial operational improvement, especially when scaled to large volumes of incidents. For every 10,000 test cases, XGBoost correctly labels several hundred more crimes than either logistic regression or Random Forest.

When weighted F1 is considered, which reflects performance weighted by class frequency, the same ordering holds. Logistic regression attains a weighted F1 of 0.2860, Random Forest improves that to 0.3629, and XGBoost reaches 0.3879. The jump from logistic regression to Random Forest indicates the value of non-linear decision boundaries and feature interactions for the common crime types. The further improvement from Random Forest to XGBoost shows the benefit of sequential error correction and more targeted split selection.

Macro F1, which averages performance across all classes equally, tells a more nuanced story. Logistic regression has a macro F1 of 0.0467, confirming that it struggles severely with rare classes and concentrates on the majority categories. Random Forest raises macro F1 to 0.1557, more than a threefold increase, demonstrating much better coverage of low-frequency crimes. XGBoost further lifts macro F1 to 0.1599. The difference between Random Forest and XGBoost is modest in absolute terms but confirms that XGBoost does not sacrifice rare-class performance to gain accuracy; instead, it improves both.

Training time also matters for practical deployment. Logistic regression is the slowest in this setup at 9.6 minutes. Random Forest and XGBoost are closely matched at 7.1 and 7.3 minutes respectively, meaning that both non-linear models can be retrained regularly without prohibitive cost. Given that XGBoost achieves the best accuracy and F1-scores with only a small increase in training time over Random Forest, it offers the best overall trade-off for production use.

The differences in feature importance patterns highlight the complementary nature of Random Forest and XGBoost. Random Forest emphasizes victim age and temporal structure, using when crimes occur and who is affected as primary signals. XGBoost, in contrast, focuses on contextual flags and specific premises, using how and where crimes occur as main discriminators. This complementarity suggests that ensemble strategies combining both models could further improve performance, for example through model stacking or weighted averaging.

Overall, logistic regression serves as a useful but limited baseline, capturing only the simplest patterns. Random Forest offers a strong balance between performance and speed, with much better treatment of rare classes. XGBoost provides the best performance across all major metrics with only a minor training-time cost, and therefore stands out as the primary candidate for deployment in crime-type prediction systems.

5. Conclusion

This study set out to explore how different factors, including economic conditions, housing patterns, and the effects of COVID-19, shape crime in Los Angeles. By analyzing more than one million LAPD crime records together with demographic and neighborhood data, we found that crime does not occur randomly. Instead, it consistently clusters in communities experiencing financial strain, unstable housing, and limited social resources. Vehicle theft, residential burglary, and identity theft appeared most frequently, showing how long-term structural challenges and new opportunities for crime both influence daily life in the city. Our findings also show that crime patterns shift across time, location, and victim characteristics. Daily routines help explain the weekly and hourly spikes in crime, while neighborhood disadvantages clarify why certain areas remain high-risk. The COVID-19 pandemic added another layer of pressure. Domestic incidents increased, identity-related crimes became more common, and street crime responded to changes in mobility and economic stress. These results highlight that crime is shaped by broader social and economic forces and cannot be understood through policing patterns alone.

To answer our research questions, we examined three core issues. The first question asked how neighborhood characteristics, such as income, unemployment, education, and racial diversity, influence

the types and frequency of crime. The results show clear spatial clustering. Crime is more common in areas facing economic hardship and limited community resources. The second question focused on whether COVID-19 affected crime differently across affluent and disadvantaged neighborhoods. The evidence shows that pandemic disruptions worsened existing inequalities. Domestic violence and identity theft increased especially in vulnerable communities. The third question examined how housing changes, including rising rents and gentrification, shift crime patterns across the city. The results suggest that while crime may decrease within gentrifying neighborhoods, it often moves into nearby areas, which strengthens existing inequalities rather than eliminating crime.

While this study provides valuable insights, several limitations should be acknowledged. The analysis relies on publicly available LAPD data, which may include reporting delays, underreporting, or inconsistencies in victim or location information. Socioeconomic and housing data were aggregated at the neighborhood level, which limits our ability to analyze very localized patterns. The machine learning models used in this study can only reflect the features available in the dataset, meaning that important factors such as real-time mobility, informal economic activity, or community-level interventions were not included. The COVID-19 period also presents unique behavioral disruptions that make it difficult to separate temporary changes from long-term trends. Future work could address these limitations by integrating more detailed socioeconomic indicators, mobility information, and qualitative insights from the communities involved. Even with these constraints, the analysis demonstrates the value of combining exploratory data analysis and predictive modeling to understand urban crime. Machine learning tools highlight where risks are concentrated and which contextual factors, such as time of day, location type, and weapon involvement, carry the strongest predictive value. However, predictions alone cannot meaningfully reduce crime. Effective solutions require broader social policies, including expanded affordable housing, improved job access, strengthened community programs, and better access to education and mental health support.

Overall, this project shows that crime is closely connected to the wider social and economic conditions of a city. By integrating data-driven insights with equitable, community-oriented policies, cities like Los Angeles can become safer, more resilient, and more inclusive. Future research that incorporates richer and more dynamic data sources can deepen these insights and support long-term strategies for improving public safety and community well-being.

References

[1] G. S. Becker, “Crime and punishment: An economic approach,” J. Political Econ., vol. 76, no. 2, pp. 169–217, 1968.

[2] I. Ehrlich, “Participation in illegitimate activities: A theoretical and empirical investigation,” J. Political Econ., vol. 81, no. 3, pp. 521–565, 1973.

[3] S. Raphael and R. Winter-Ebmer, “Identifying the effect of unemployment on crime,” J. Law Econ., vol. 44, no. 1, pp. 259–283, 2001.

[4] P. Buonanno, “Crime and labour market opportunities in Italy (1993–2002),” Labour, vol. 17, no. 4, pp. 559–582, 2003.

[5] J. M. Brush, “Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties,” Econ. Lett., vol. 96, no. 2, pp. 264–268, 2007.

[6] D. Abrams, “COVID and crime: An early empirical look,” J. Public Econ., vol. 194, pp. 104344, 2021.

[7] R. Rosenfeld, E. Grigg, and A. I. Spivak, “COVID-19 and crime: A preliminary examination of the impact of the pandemic on crime in the United States,” Crime Justice, vol. 50, no. 1, pp. 1–20, 2020.

[8] G. M. Campedelli, F. Aziani, and L. Favarin, “Exploring the impact of COVID-19 lockdowns on domestic violence: Evidence from Italy,” J. Interpers. Violence, vol. 36, no. 23–24, pp. NP12562–NP12581, 2021.

[9] J. Vargas, “The impact of socioeconomic factors on crime rates,” Addict Criminol., vol. 6, no. 4, pp. 161, Aug. 2023.

10] D. Schoewe, “Crime and Socio-Economic Factors in the U.S.: A Quantitative Geospatial Statistical Analysis,” Jun. 2023. [Online]. Available: https://storymaps.arcgis.com/stories/ebc9f03e60074c9f9b4b36d22601b8aa

[11] D. Autor, C. Palmer, and P. A. Pathak, “Housing market spillovers: Evidence from the end of rent control in Cambridge, Massachusetts,” J. Political Econ., vol. 122, no. 3, pp. 661–717, 2014.

[12] Z. Macdonald and L. Stokes, “Gentrification and crime displacement: A spatial analysis,” Urban Stud., vol. 57, no. 2, pp. 345–362, 2020.

[13] A. Porreca, “Gentrification and social fragmentation: Crime trends in transitioning neighborhoods,” Soc. Sci. J., vol. 60, no. 1, pp. 45–59, 2023.

[14] J. A. Arthur, “Socioeconomic predictors of crime in rural Georgia,” Criminal Justice Rev., vol. 16, no. 1, pp. 29–41, 1991.
[15] D. Shichor, “Crime patterns and socioeconomic development: A cross-national analysis,” Criminal Justice Rev., vol. 15, no. 1, pp. 64–78, 1990.

[16] J. B. Cullen and S. D. Levitt, “Crime, urban flight, and the consequences for cities,” Rev. Econ. Stat., vol. 81, no. 2, pp. 159–169, 1999

[17] City of Los Angeles, Sept. 2025, “Crime Data from 2020 to Present” Los Angeles Police Department, [Online], Available: https://catalog.data.gov/dataset/crime-data-from-2020-to-present

[18] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[20] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. New York, NY, USA: Wiley, 2013.