Los Angeles Police Department

Crime Reporting Delays Analysis

Rough Draft

By

Forest Flux Team

Victoire Migashane, and Sarah Mohammed

Arizona State University

DAT 490 - Spring A, 2026

Laurence Schneider

Feb 23, 2026

Table of Contents

Abstract/Executive Summary

Timely crime reporting is essential for effective policing, resource allocation, and public safety, though reporting behavior varies significantly across crime types, locations and populations. This project analyzes public crime data from the LAPD between 2020-2025 to help us investigate three questions: what factors influence crime reporting delays, what conditions predict crime hotspot intensity, and how victim age patterns differ across crime characteristics. After cleaning over a million records, reporting delay was calculated as the difference between the occurrence date and report date. The problem was modeled as a classification task distinguishing timely reports from delayed reports using decision tree, random forest, logistic regression, and XGBoost models. Spatial crime intensity was examined using count–based regression, and victim age patterns were analyzed using regression and exploratory visualizations techniques.

Results show consistent behavioral patterns where violent crimes are reported substantially faster than non-violent crimes, suggesting that perceived urgency strongly influences reporting decisions. Crimes that involve weapons are also reported quicker compared to financial and identity related offenses (mainly because the discovery is typically much later than when the crime occurred). Geographic crime concentration does not necessarily correspond to slower reporting, indicating crime volume and reporting behavior reflect different underlying processes. Age patterns further demonstrate that younger individuals are more frequently victims of violent offenses, while older individuals are more affected by fraud related crimes.

Among the predictive models, XGBoost produced the strongest performance with high accuracy and balanced recall across all reporting categories.

Our goal for this analysis was to have interpretive results rather than predictive. The models reveal behavioral tendencies in how residents interact with law enforcement rather than simply forecasting incidents.

Our findings suggest that delayed reporting is less a product of police response capacity and more a function of trust, awareness, and the perceived severity of harm. Policy implications include targeted outreach for financial crimes, multilingual reporting resources, and community based communication strategies rather than solely increasing enforcement in high crime areas. Predictive analytics can help with decision making, but ethical implementation requires acknowledging the historical date reflecting social inequalities and should guide support efforts rather than punitive deployment.

Project Plan

Organization Description

The Los Angeles Police Department (LAPD) serves as the municipal law enforcement agency for the City of Los Angeles, California. As the third-largest police department in the United States, the LAPD is responsible for a population of approximately 4 million residents across 502 square miles. The department operates 21 community police stations, referred to as "Areas," throughout the city and employs over 9,000 sworn officers and approximately 3,000 civilian personnel. The LAPD is structured into four main bureaus: the Office of Operations (patrol and specialized units), the Office of Special Operations (metro, K-9, SWAT), the Detective Bureau (investigative services), and the Office of Administrative Services (support functions). Oversight is provided by the Los Angeles Board of Police Commissioners, and the department reports directly to the Mayor of Los Angeles.

The mission of the LAPD is to safeguard the lives and property of residents, reduce the incidence and fear of crime, and enhance public safety by collaborating with diverse communities to improve quality of life. The department adheres to six core values: Service to Our Communities, Reverence for the Law, Commitment to Leadership, Integrity in All We Say and Do, Respect for People, and Quality Through Continuous Improvement. These values underscore a commitment to public safety, adherence to constitutional principles, exemplary leadership, the highest ethical standards, respect for all individuals, and a focus on innovation and ongoing improvement.

The current Chief of Police, Jim McDonnell, was sworn in as the 59th Chief on November 8, 2024. Chief McDonnell is a 29-year LAPD veteran and the first person to hold senior executive leadership roles at the three largest policing agencies in Los Angeles County: LAPD, the Los Angeles County Sheriff's Department (LASD), and the Long Beach Police Department (LBPD). He succeeded Dominic H. Choi, who served as Interim Chief from March to November 2024 and was the first Asian American to lead the LAPD. Michel Moore previously served as Chief from 2018 to 2024.

The LAPD Strategic Plan 2021-2023 identifies six primary goals to guide departmental operations and improvements. These goals include protecting Los Angeles through crime reduction and community partnerships, building public trust through engagement initiatives, enhancing organizational accountability by reducing uses of force and promoting fairness, modernizing technology to increase field efficiency, enriching training with cultural perspectives, and maximizing workforce potential through diverse recruitment and employee retention. The plan prioritizes constitutional policing, transparency, and continuous improvement in service delivery throughout all neighborhoods.

Headquarters: Los Angeles California (100 West 1st Street, Los Angeles, CA 90012)

Contacts: 911 for emergencies and (877) 275-5273 for non-emergencies

Website: https://www.lapdonline.org/

Learn more: https://www.lapdonline.org/lapd-organization/

History of Organization

The Los Angeles Police Department (LAPD) was established in 1869 when the city appointed six officers under City Marshal William C. Warren. Prior to this, Los Angeles depended on volunteer groups such as the Los Angeles Rangers, founded in 1853, to maintain order in a frontier town characterized by gambling, vice, and what was reportedly the highest murder rate in the United States during the Gold Rush era. The LAPD expanded gradually, and in 1889, Chief John Glass divided the city into four police districts to enhance supervision. The department achieved significant milestones by appointing Alice Stebbins Wells in 1910 as the first female police officer with full arrest powers in the United States, and Georgia Ann Robinson in 1916 as the first African American female officer. Despite these advancements, the department experienced persistent corruption in the early 1900s until Mayor Fletcher Bowron implemented substantial reforms in 1938, resulting in the removal of numerous corrupt officers and commissioners.

Between 1950 and 1966, Chief William H. Parker restructured the LAPD, establishing it as one of the most professional police departments in the United States. He introduced modern management systems and adopted the motto "To Protect and To Serve" in 1955. Despite these advancements, the department encountered significant challenges, including the 1965 Watts Riots and the 1991 Rodney King incident, both of which highlighted issues of police brutality and strained community relations. These incidents prompted comprehensive reforms through the Christopher Commission, which promoted greater diversity in hiring and enhanced accountability. In 2001, the U.S. Department of Justice imposed a consent decree on the LAPD to address civil rights violations; following extensive reforms, the decree was lifted in 2013. Currently, the LAPD emphasizes constitutional policing, community engagement, and transparency.

Learn more: https://www.lapdonline.org/history-of-the-lapd/

Community Background

Los Angeles is among the most diverse cities in the United States, with residents from over 140 countries and speakers of 224 languages distributed across approximately 214 neighborhoods. The city encompasses affluent areas such as Bel Air and Pacific Palisades, as well as working-class communities in South Los Angeles and the San Fernando Valley. This demographic diversity presents complex policing challenges that necessitate cultural competence and community-specific strategies. The Los Angeles Police Department (LAPD) operates 21 community police stations citywide, staffed by more than 9,000 sworn officers and 3,000 civilian personnel.

In response to community needs and to rebuild trust following historical tensions, the LAPD has implemented a Community Policing model that emphasizes partnership between law enforcement and residents. Each of the 21 geographic Areas is supported by a Community-Police Advisory Board (CPAB), which facilitates community input into local policing decisions. The department employs the SARA (Scanning, Analysis, Response, and Assessment) problem-solving framework to collaboratively identify and address crime issues. Additionally, the LAPD has enhanced transparency by making crime data publicly accessible. Nevertheless, Los Angeles continues to confront persistent challenges, including homelessness, property crime, gang activity, and the pursuit of equitable policing across neighborhoods of varying socioeconomic and racial composition.

Lean more: https://www.lapdonline.org/community-policing-unit/

Competitors and Finances

The Los Angeles Police Department (LAPD) operates with a budget of $2.14 billion for fiscal year 2025-26, reflecting an 8.1% increase from the previous year. When combined with federal,

state, and other funding sources, the total budget reaches approximately $3.3 billion. Key budget priorities include salary adjustments, recruitment initiatives, vehicle and helicopter replacements, and technology upgrades such as the Real-Time Crime Center and enhanced cybersecurity operations. The city aims to increase the number of officers to 9,084, with a long-term objective of hiring 9,500 officers by 2028 or earlier. As of 2024-2025, the LAPD employs over 9,000 sworn officers and approximately 3,000 civilian personnel distributed across 21 community police stations. These stations serve 4 million residents within 502 square miles, positioning the LAPD as the third-largest municipal police department in the United States.

The LAPD's primary peer agencies are the New York City Police Department (NYPD) and the Chicago Police Department (CPD). The NYPD, the largest municipal police department in the United States, has a fiscal 2025 budget of $5.8 billion and supports 48,844 full-time positions, including 35,001 uniformed officers. The department plans to expand to 40,000 officers by 2029. The Chicago Police Department operates with a fiscal year 2024 budget of nearly $2.0 billion but faces significant staffing challenges, functioning at only 87% capacity with 12,329 active employees as of January 2024, a decrease of 1,400 officers since 2019. All three departments face similar challenges, including recruitment difficulties, overtime costs, the need to modernize technology, and the need to balance effective crime reduction with accountability reforms. The LAPD distinguishes itself through its community policing model, which includes 21 Area-based Community-Police Advisory Boards and a strategic focus on diversity, equity, and inclusion.

Learn more:

https://nenc-la.org/2024/11/lapd-budget-approved-2-14-billion-spending-plan-for-2025-26/

Research Questions

Research Question 1: Predicting Crime Reporting Delays in Los Angeles

This research will investigate the factors contributing to the delay between the occurrence of a crime and its reporting to the Los Angeles Police Department. The study aims to identify which variables, including crime type, victim demographics, location, and timing, influence whether a crime is reported immediately or after a significant delay. Additionally, the analysis will examine whether these patterns differ between violent and property crimes. The decision to report a crime is not random; it may be influenced by factors such as fear, perceived severity, trust in law enforcement, or practical barriers such as lack of knowledge about reporting procedures or time constraints. By applying classification analysis to LAPD crime data from 2020 to 2025 (approximately 1,000,000 records and 28 variables), the research will determine which factors most strongly predict reporting delays. Insights from this analysis will assist the LAPD in identifying communities or crime types that encounter barriers to timely reporting, informing strategies to build trust, allocate resources effectively, and promote equitable access to police services across all neighborhoods.

Research Question 2: Predicting Crime Hotspot Intensity Across Los Angeles

This research will examine the factors that predict crime concentration and intensity across Los Angeles neighborhoods. Specifically, we aim to identify how temporal patterns (such as rush hour versus late night, weekdays versus weekends, and seasonal variations), the distribution of crime types, crime locations (including streets, parking lots, and residences), and historical crime data contribute to the emergence of crime hotspots or, conversely, to neighborhood safety. This

inquiry is significant because disparities in perceived and actual safety persist across Los Angeles communities. Utilizing regression analysis on LAPD crime data from 2020 to 2025, comprising approximately 1,000,000 records and 28 variables, we will aggregate data by area and time period. This approach will enable us to determine the specific conditions that increase vulnerability to crime in certain locations and times. By elucidating these patterns, our findings can inform LAPD resource allocation, support community-based interventions to address environmental risk factors, and promote equitable safety outcomes for all residents.

Research Question 3: Predicting Victim Age Patterns in Los Angeles Crime

This research will analyze the demographic patterns of crime victims in Los Angeles, with a focus on identifying factors that predict whether victims are children, young adults, middle-aged individuals, or seniors. We will assess how variables such as crime type, location, timing, and weapon involvement influence the age distribution of victims, and whether certain age groups are disproportionately targeted or exposed to specific risks. This analysis is important because different age groups experience distinct vulnerabilities and require tailored prevention strategies. For example, elderly individuals may be more susceptible to scams or physical attacks, while young people may be at greater risk in certain neighborhoods or at certain times. By applying regression analysis to LAPD crime data from 2020 to 2025, which includes approximately 1,000,000 records and 28 variables (including victim age), we aim to identify which demographic groups are most at risk for specific crimes and the underlying reasons. The results can inform targeted interventions, such as specialized outreach to seniors, safety education for young adults, and measures to ensure safe environments for children.

Project Hypotheses

Research Hypothesis 1

Violent crimes, such as assault with deadly weapons, are expected to exhibit significantly shorter reporting delays compared to non-violent crimes, such as theft. Specifically, violent crimes are anticipated to be reported within 24 to 48 hours of occurrence, whereas theft and fraud-related crimes are expected to demonstrate reporting delays exceeding several days and even months.

Research Hypothesis 2

Crime concentration in specific LAPD areas is expected to be significantly associated with nighttime hours (10 PM to 4 AM) and with commercial premises such as stores, parking lots, and gas stations. Commercial districts are expected to experience the highest crime intensity during late-night weekend hours, with incident rates 3 to 5 times higher than those in residential areas during daytime weekday hours. Furthermore, areas with historically high crime rates are projected to maintain elevated crime intensity regardless of temporal factors, indicating persistent geographic vulnerability.

Research Hypothesis 3

Violent crimes involving weapons are expected to disproportionately affect younger victims (ages 18 to 35), with over 60% of weapon-related assaults targeting this demographic. In contrast, property crimes such as theft and fraud are anticipated to exhibit a bimodal distribution, affecting both young adults (18 to 35) and seniors (60 and older). Crimes occurring during late-night hours (10 PM to 4 AM) are projected to predominantly victimize younger individuals,

whereas crimes during daytime business hours (9 AM to 5 PM) are expected to more frequently target middle-aged and senior victims. These patterns reflect differing daily activities and vulnerabilities across age groups.

About the Data

The dataset utilized in this project originates from the Los Angeles Police Department (LAPD) and is publicly accessible via the U.S. Government Open Data Catalog at data.gov. It is provided as a single CSV file containing both structured and unstructured variables that document crime records for the greater Los Angeles metropolitan area. The dataset comprises 1,004,991 reported crime records from January 1, 2020, to June 4, 2025, and includes 28 variables. The data are organized into five categories: Incident Identification, Temporal Information, Geographic Information, Victim Demographics, and Crime Characteristics. These categories collectively facilitate the investigation of crime reporting delays, which constitute the primary focus of this project.

Incident Identification

Each of the 1,004,991 crime records is uniquely identified by a set of core identification and classification variables. The division record number (DR_NO) functions as the unique identifier for each incident. The reporting district number (Rpt Dist No), LAPD area code, and area name specify the precinct responsible for the case. Each record may include up to four crime code fields (Crm Cd, Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4) and their corresponding descriptions (Crm Cd Desc), enabling both primary and secondary crime classifications for incidents involving multiple crime types. The method of operation (mocodes) provides further detail

regarding how the crime was committed. The status code (Status) and its description (Status Desc) track each case's current status.

Temporal Information

Temporal variables are central to this project. The dataset includes both the date of occurrence (Date OCC) and the date reported (Date Rptd), each with associated year, month, and day fields (Date OCC Year, Date OCC Month, Date OCC Day, Date Rptd Year, Date Rptd Month, Date Rptd Day). The time of occurrence (Time OCC) is also recorded, providing hour-level detail. Reporting delay is calculated at the day level by subtracting the date of occurrence from the date reported. This delay serves as the primary classification target for the predictive model.

Geographic Information

Geographic variables provide spatial context for each crime record. The LAPD area code (Area) and area name (Area Name) identify the police district where the crime occurred. More precise location information is available from the full address (Location), the nearest cross street (Cross Street), and the geographic coordinates (latitude [lat] and longitude [lon]). These variables enable geographic analysis, particularly when investigating crime hotspot intensity across Los Angeles neighborhoods.

Victim Demographics

The dataset contains three victim demographic variables: victim age (Vict Age), victim sex (Vict Sex), and victim descent (Vict Descent). Victim demographics such as age and descent may influence the likelihood or speed of crime reporting and may serve as the primary variable to classify which age groups are most affected by various crime types. Collectively, these

demographic variables provide essential context for understanding the human factors that influence crime reporting behavior.

Crime Characteristics

Crime characteristics variables describe the nature and context of each incident. The premise code (Premis Cd) and premise description (Premis Desc) specify the location type, such as residence, street, or commercial establishment. The weapon used code (Weapon Used Cd) and weapon description (Weapon Desc) indicate whether a weapon was involved and, if so, its type. The crime code description (Crm Cd Desc) is an unstructured text field that requires processing to create consistent crime-type categorizations, such as distinguishing between violent and non-violent crimes. These classifications are essential for testing the hypothesis that violent crimes will have significantly shorter reporting delays than non-violent crimes such as theft or fraud.

Measurements

A clear understanding of measurement is essential for any analytical project, particularly those involving predictive analytics. This project focuses on classification accuracy and patterns of reporting delay. The primary outcome variable in this research is reporting delay, defined as the time between when a crime occurred and when it was reported to the LAPD. This delay is calculated at the day levels to capture broad and detailed patterns. The delay is then transformed into a classification target, categorizing incidents as reported quickly (within 24 hours) or with a significant delay (exceeding 3 days). This classification serves as the predictive model's output variable.

The LAPD records all relevant information at the time a crime is entered into the system, including timestamps, location details, victim information, and crime characteristics. This results in a comprehensive dataset that directly supports the project's measurement objectives. Key predictors for the reporting delay classification model include crime type (violent versus non-violent), victim demographics (age, sex, and descent), geographic location (LAPD area), time of occurrence, and premise type. Each variable represents a potential factor influencing reporting speed, and collectively, they provide the predictive model with the necessary information to identify patterns in crime reporting behavior across Los Angeles.

Methodology

This project will employ statistical and predictive modeling to analyze crime patterns in Los Angeles using public crime data from 2020 to 2025. The primary objective is to examine how crime reporting behavior, crime concentration, and victim age patterns vary over time, across locations, and by crime characteristics.

Prior to analysis, the dataset will be cleaned and prepared by removing records with invalid or missing dates, addressing missing demographic values, and creating new variables as necessary. The key variable, 'crime reporting delay,' will be calculated as the difference between the date and time a crime occurred and the date it was reported to the Los Angeles Police Department.

Research Question 1: Predicting Crime Reporting Delays in Los Angeles

The research will be transformed into a classification problem. We will have the crimes labeled as either reported within 48 hours or reported with a significant delay (3+ days). Decision Tree, random forest, logistic regression, and XGBoost, will be used to determine factors such as crime type, victim demographics, location, and time to help us predict delayed reporting.

Research Question 2: Predicting Crime Hotspot Intensity Across Los Angeles

The analysis of crime hotspot intensity will involve grouping data by areas within the Los Angeles Police Department and by intensity levels. Regression models for count data, such as negative binomial regression, will be used to determine how temporal patterns, premise types, and historical crime levels predict areas of highest crime concentration in the city.

Research Question 3: Predicting Victim Age Patterns in Los Angeles Crime

The analysis of victim age patterns will categorize victims as children, young adults, middle-aged adults, or seniors. Multinomial logistic regression will be used to examine how crime type, weapon involvement, time of occurrence, and location predict the likelihood of victimization for each age group.

Computation method and output

All data processing and analysis will be conducted using Python and R. Regression models will be developed with statistical modeling tools, and standard data analysis libraries will be applied for data cleaning and feature engineering.

For classification models addressing research questions 1 and 3, performance will be evaluated using accuracy, precision, recall, F1 score, and ROC-AUC. Confusion matrices will also be employed to analyze classification errors, particularly when delayed reporting cases are incorrectly predicted as timely.

The hotspot analysis for research question 2 will assess model fit and predictive strength using goodness-of-fit measures and comparisons between observed and predicted crime counts. Spatial

visualizations, such as heat maps and area-level crime plots, will be utilized to effectively communicate hotspot patterns.

The outputs for this project will include cleaned and processed datasets, regression results and coefficient tables, model performance metrics, and visualizations of reporting delays, crime hotspots, and victim age distributions. These outputs are intended to be interpretable while providing robust analytical evidence.

Output Summary

Research Question 1: Predicting Crime Reporting Delays in Los Angeles

The analysis will indicate which crimes, locations, and victim characteristics are most strongly associated with delayed reporting. The results will highlight differences between violent and non-violent crimes and identify communities where barriers to timely reporting are present.

Research Question 2: Predicting Crime Hotspot Intensity Across Los Angeles

The analysis will reveal which areas within Los Angeles experience the highest crime intensity and the periods when hotspots are most active. Maps and time-based summaries will demonstrate how crime concentration changes across neighborhoods, providing context for why certain areas are perceived as less safe.

Research Question 3: Predicting Victim Age Patterns in Los Angeles Crime

The analysis will identify which age groups are most vulnerable to specific crime types and conditions. The results will demonstrate how age patterns vary by crime type, weapon involvement, and time of day, providing insight into age-specific risks across Los Angeles.

Campaign Implementation

Los Angeles is one of the largest and most diverse cities in the United States, with over four million residents distributed across neighborhoods that vary significantly in safety, resources, and trust in law enforcement. Due to these disparities in crime experiences, the Los Angeles Police Department (LAPD) must strategically allocate resources, communicate with residents, and implement targeted prevention efforts.

Research Question 1: Predicting Crime Reporting Delays in Los Angeles

Addressing the first research question, identifying patterns in delayed crime reporting enables the LAPD to focus outreach efforts on crimes and communities where reporting barriers are most significant. If non-violent crimes such as theft or fraud are consistently reported late, the department could implement targeted public awareness campaigns to explain the importance of timely reporting, even for seemingly minor offenses. Initiatives including online reporting tools, multilingual reporting guides, and community workshops in affected neighborhoods may help reduce these barriers and improve data accuracy for the LAPD.

Research Question 2: Predicting Crime Hotspot Intensity Across Los Angeles

Regarding the second research question, understanding crime hotspot intensity allows the LAPD to deploy officers and resources more efficiently. The department could prioritize staffing during high-risk hours and in consistently vulnerable areas identified through research, rather than distributing officers evenly across the city. This data-driven approach may reduce crime rates while minimizing unnecessary policing in low-risk neighborhoods.

Research Question 3: Predicting Victim Age Patterns in Los Angeles Crime

For the third research question, identifying age patterns enables the LAPD to tailor prevention strategies to specific populations. Depending on the demographic affected, initiatives such as social media campaigns, support centers, and targeted programs can be implemented to increase awareness and provide relevant information.

Exploratory Data Analysis

Data Preparation

Data were collected from the Los Angeles Police Department public records and the U.S. Government Open Data Catalog. The dataset includes both structured and unstructured data, comprising approximately 1,004,991 reported crime incidents from January 2020 to June 2025. Each record represents a single crime report and contains 28 variables, including time, location, victim demographics, and crime characteristics.

Before analysis, all column names were standardized to lowercase with underscores, and all string values were converted to lowercase to ensure consistency. Date fields (date_rptd and date_occ) were parsed into a datetime format, and six new columns were extracted to capture the year, month, and day for each date. String representations of null values were replaced with appropriate null types, and missing value counts were assessed for all columns.

Records with missing values in premis_cd and crm_cd_1 were removed, as these fields are essential for crime classification and geographic analysis. Duplicate records were eliminated, retaining only the first occurrence. Further filtering excluded records with invalid entries in vict_sex (coded as ‘h’ or ‘-’), vict_descent (coded as ‘h’ or ‘-’), premis_desc, and victim ages

outside the range of 1 to 100. Following all cleaning steps, the dataset was reduced from 1,004,991 records to 443,803 records across 43 variables.

Several new features were engineered to facilitate analysis. The cross_street field was standardized into a new cross_street_refactored column using regex-based normalization to consolidate directional prefixes and street suffix variants (for example, “N Broadway St” and “Broadway Blvd” were both mapped to “broadway”). Victim descent was grouped into five broader categories (White, Black, Asian, Unknown, and Other) in a new vict_descent_grouped column. The time_occ field was zero-padded and parsed into a time object, with a separate hour column extracted. Crime classification was derived from the part_1-2 field, mapping values of 1 and 2 to “Violent” and “Non-Violent,” respectively. Victim age was binned into six groups: 0–17, 18–29, 30–44, 45–59, 60–74, and 75+ years. Incident hour was mapped to a time-of-day period: Morning (5am–12pm), Afternoon (12pm–5pm), Evening (5pm–9pm), and Night (all other hours).

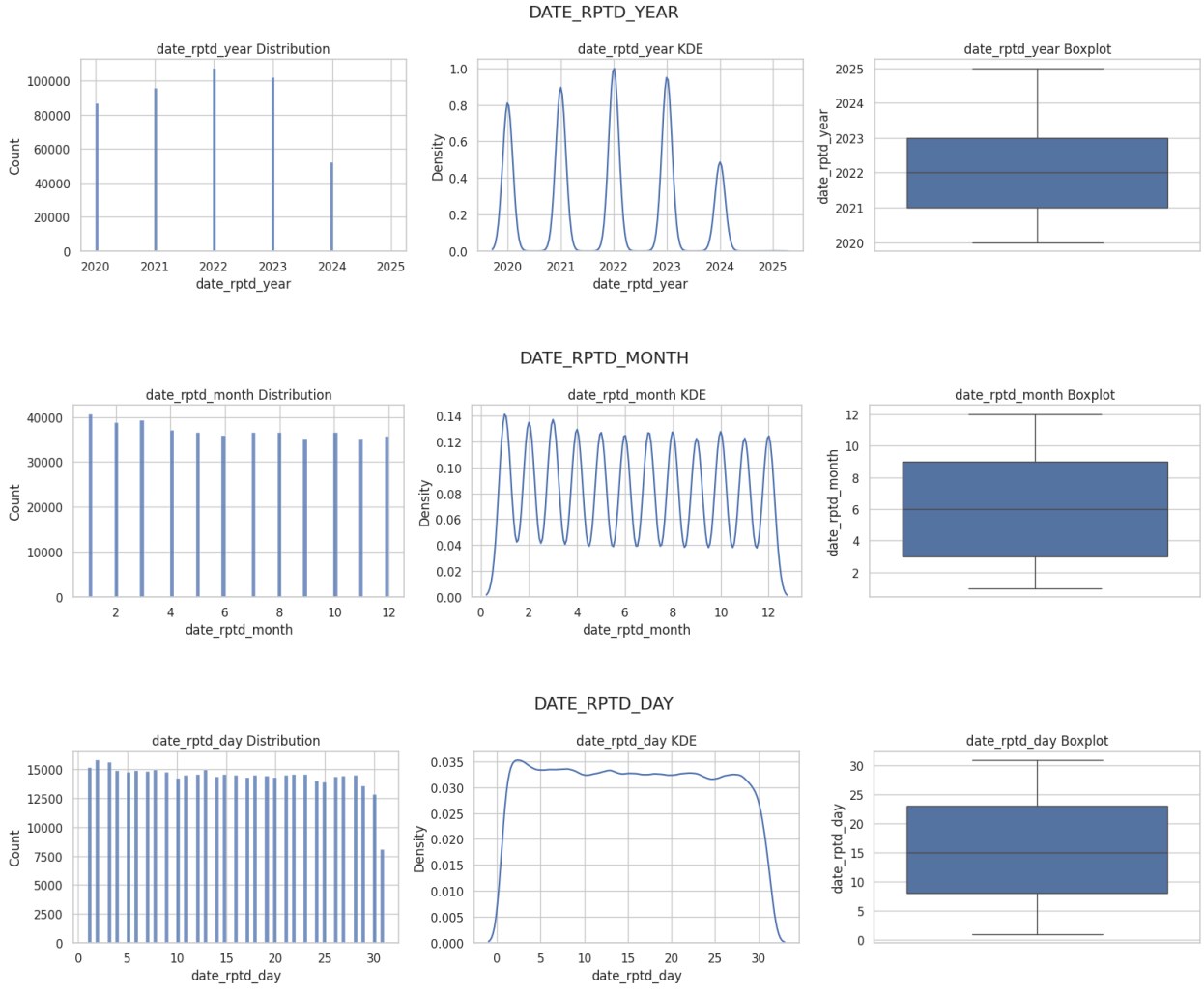
The reporting delay was calculated as the difference in days between date_occ and date_rptd and stored as days_to_report. A binary rpt_duration variable was created to indicate whether a crime was reported within 0–2 days (0) or 3 or more days (1). Records with null reporting delay values were removed after this step.

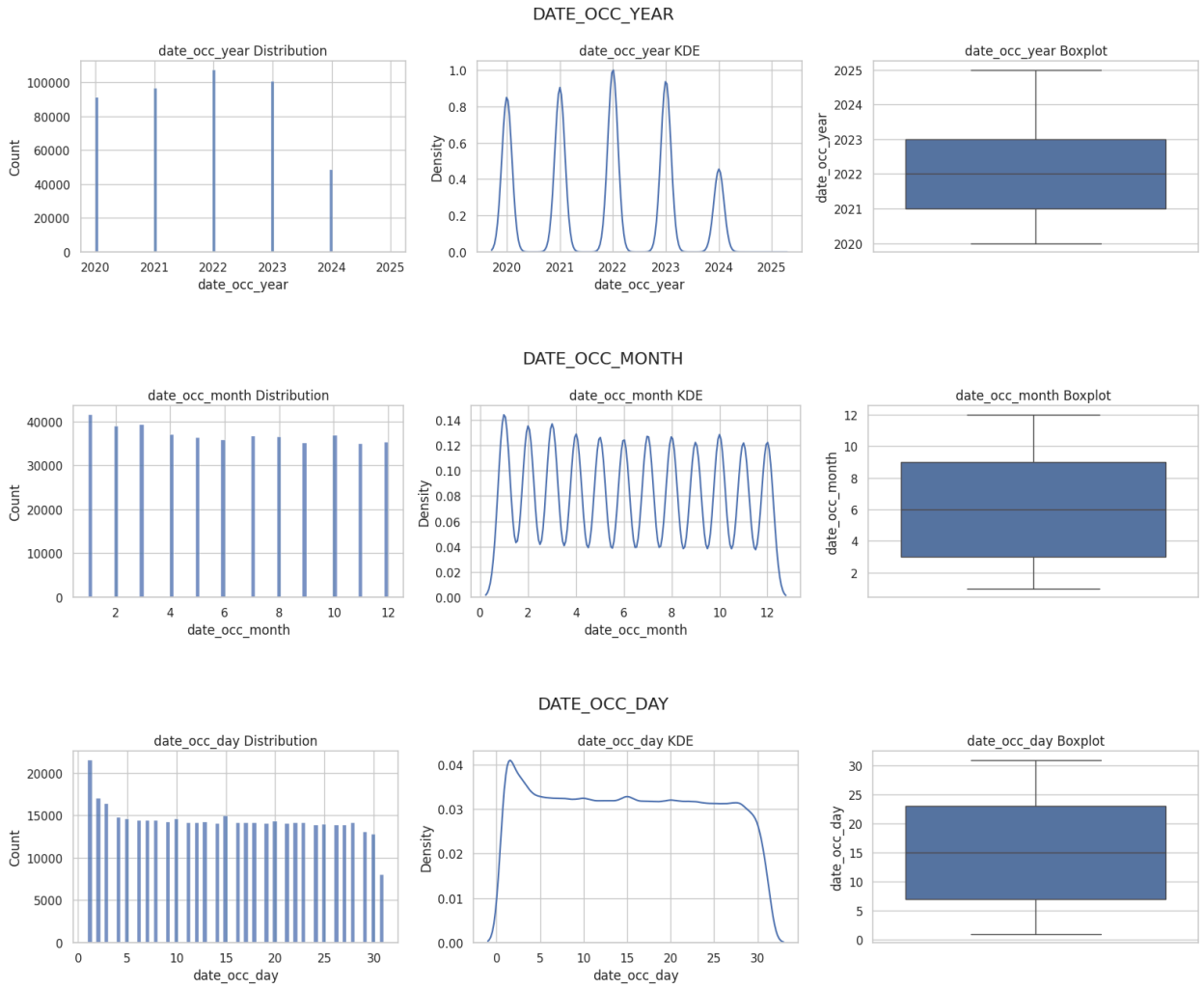
To address research question 2, crime counts were aggregated by area name, year of occurrence, and time period. This process produced a second dataframe with 436 rows and 5 columns: area name, year of occurrence, time period, crime count, and average reporting delay.

Univariate Analysis

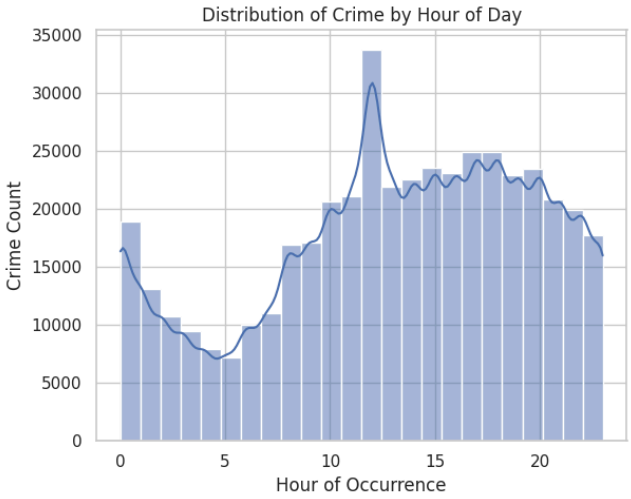
Temporal Analysis

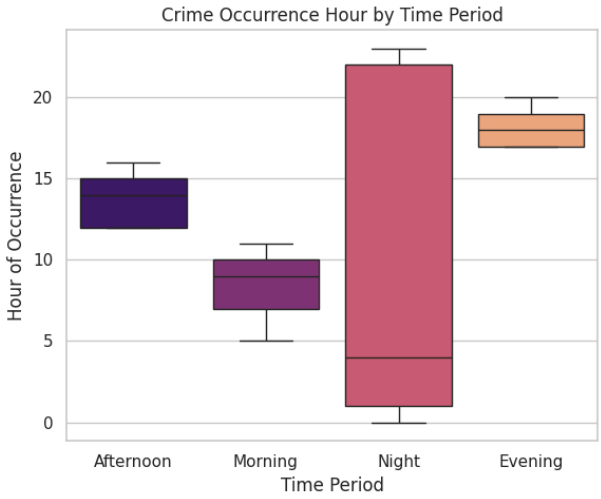
The initial exploration focused on overall reporting behavior. The figure below presents the distributions of crime occurrence and report dates by year, month, and day. Crime counts peaked around 2023, followed by a significant decline in 2024. Analysis of monthly and daily patterns indicates relative stability, suggesting that crime reporting does not exhibit a strong seasonal trend at a broad level.





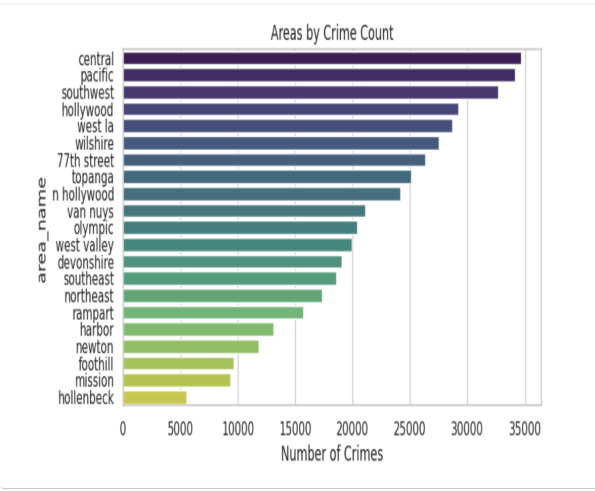
To examine the distribution of crime throughout the day, the hour of occurrence was visualized using both a histogram and a boxplot grouped by time period. The histogram indicates that crime activity reaches its lowest point in the early morning hours between 5 and 7 am, increases sharply to a midday peak around noon, and remains elevated through the afternoon and evening before declining after 10 pm. The boxplot supports the defined time period boundaries, demonstrating that the Night category spans the widest range of hours, reflecting its inclusion of both late-night and early-morning incidents, whereas the Morning and Afternoon periods are more narrowly concentrated.

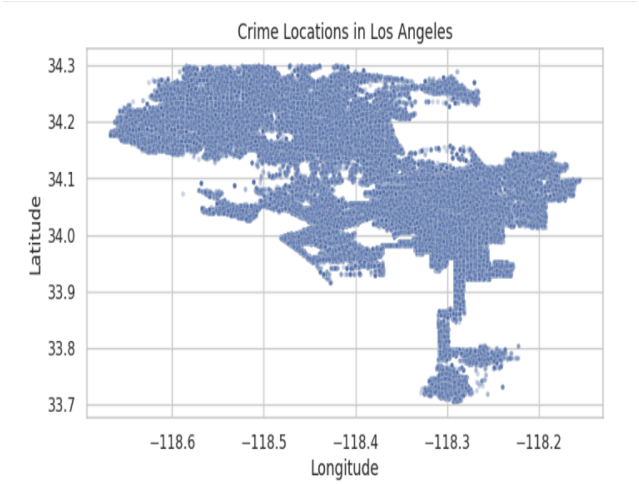




Geographic Ama;ysis

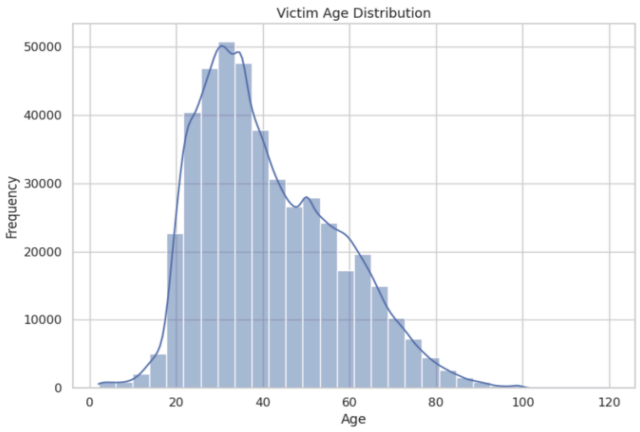
The geographic distribution of crimes was subsequently examined. The figures present plots of crime locations by latitude and longitude, demonstrating that incidents are highly concentrated in central Los Angeles rather than evenly distributed across the city. The second graph more clearly illustrates the pattern of total crime counts by the Los Angeles Police Department (LAPD) area. Central, Hollywood, and Pacific areas experience substantially higher crime volumes, whereas Mission and Foothill report fewer incidents. These differences suggest persistent spatial vulnerability rather than random fluctuation.

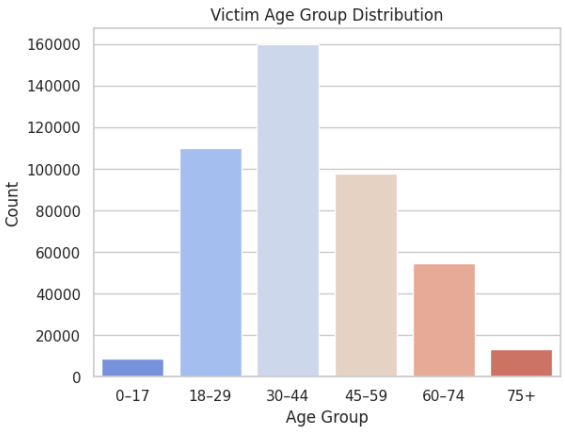




Victim Demographics Analysis

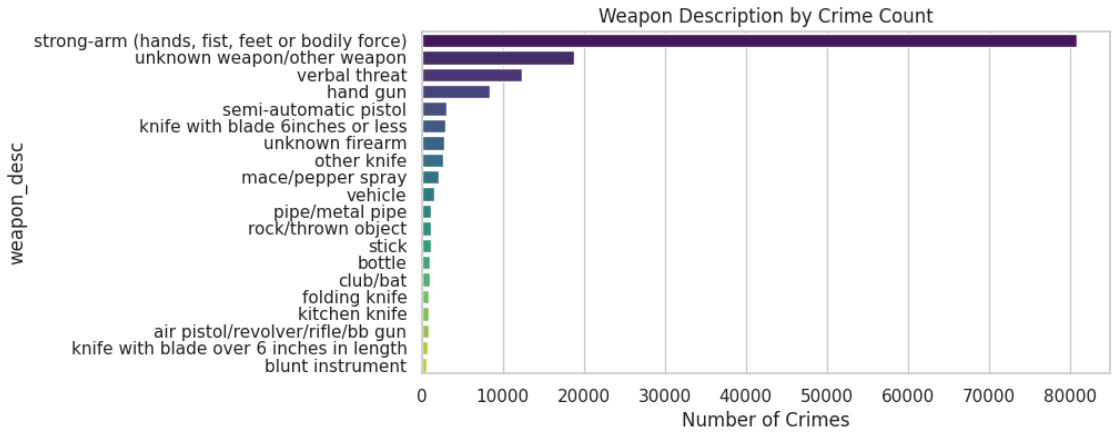
Victim demographics were examined to identify which groups are most affected by crime. The histogram and count plot displays the distribution and count of crime by victim age, revealing a unimodal pattern centered on young and middle-aged adults. Although crimes involving very young or elderly victims are less common.



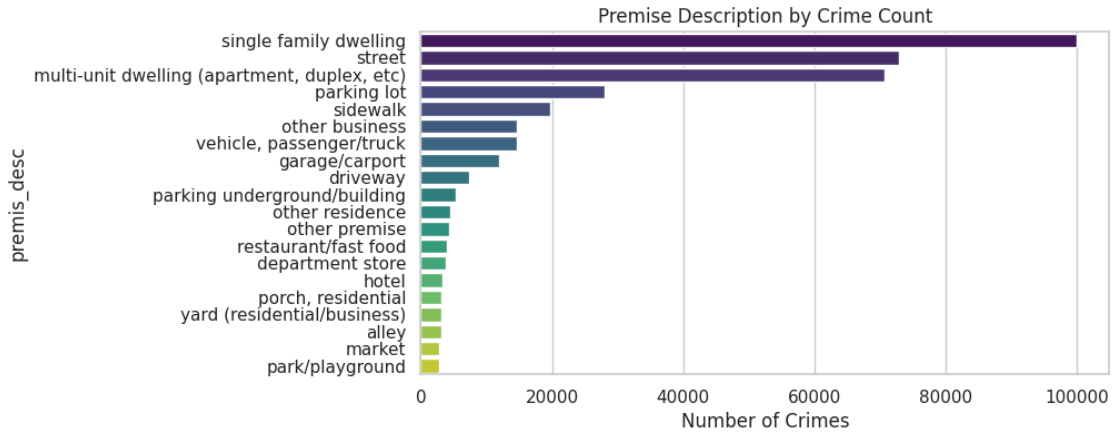


Crime Characteristics Analysis

An analysis of the top 20 weapon types used in crimes indicates that physical force, including hands, fists, or feet, was the most prevalent, appearing in over 80,000 incidents. This figure is more than four times higher than any other weapon type. Unknown weapons ranked second, followed by verbal threats and handguns. Knives and semi-automatic pistols were also frequently reported, while blunt objects and thrown items comprised the remainder of the list.



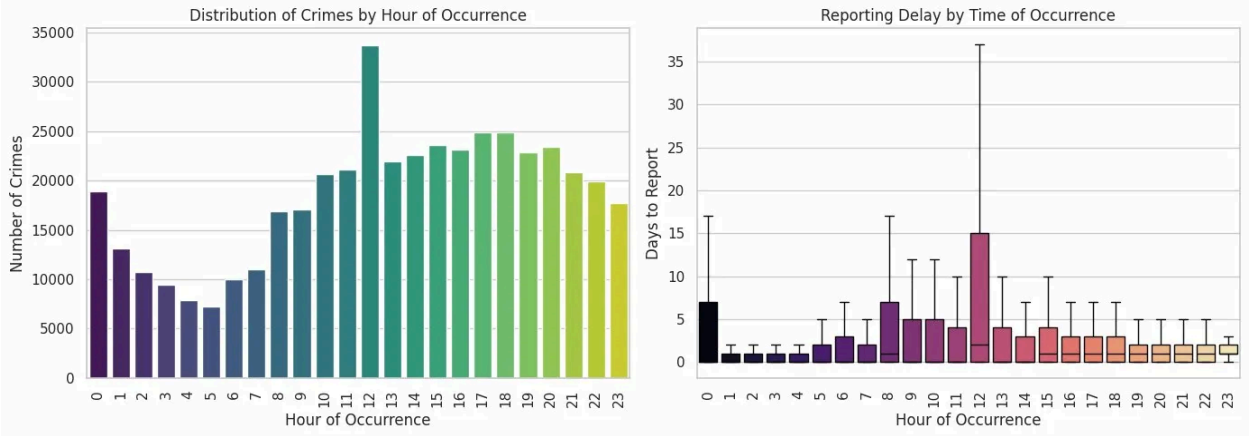
The analysis also examined the most common crime locations. Single-family homes accounted for the highest number of incidents, with approximately 100,000 cases, followed by streets and apartment buildings. Parking lots and sidewalks were the next most frequent locations, indicating that a significant proportion of crimes occur in public spaces. Businesses, vehicles, and garages were also common sites, whereas restaurants, stores, hotels, and parks were less frequently reported.



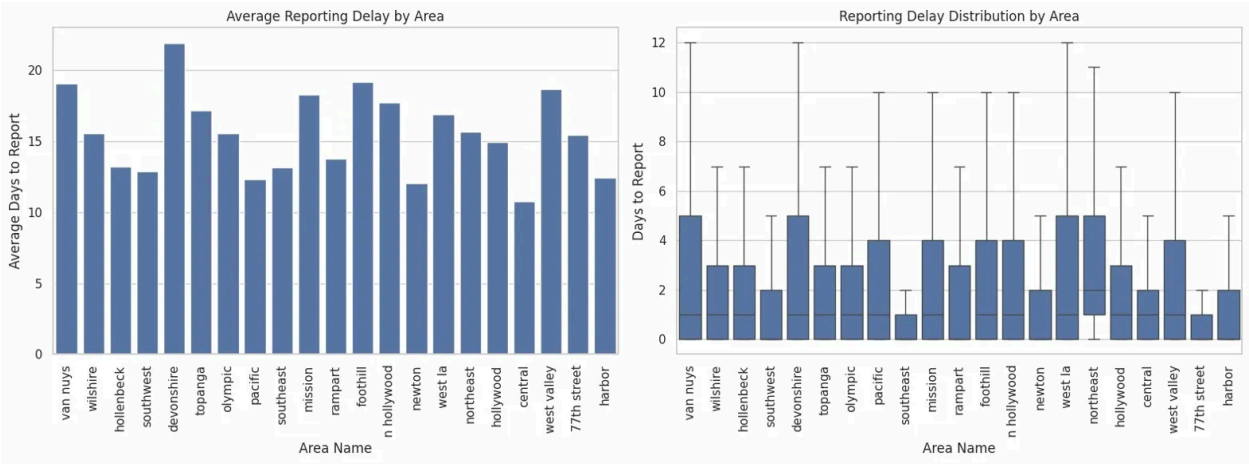
Bivariate Analysis

The analysis first examined whether crimes reported at certain hours of the day are reported faster or slower. The count plot indicates that crime volume reaches its lowest point in the early

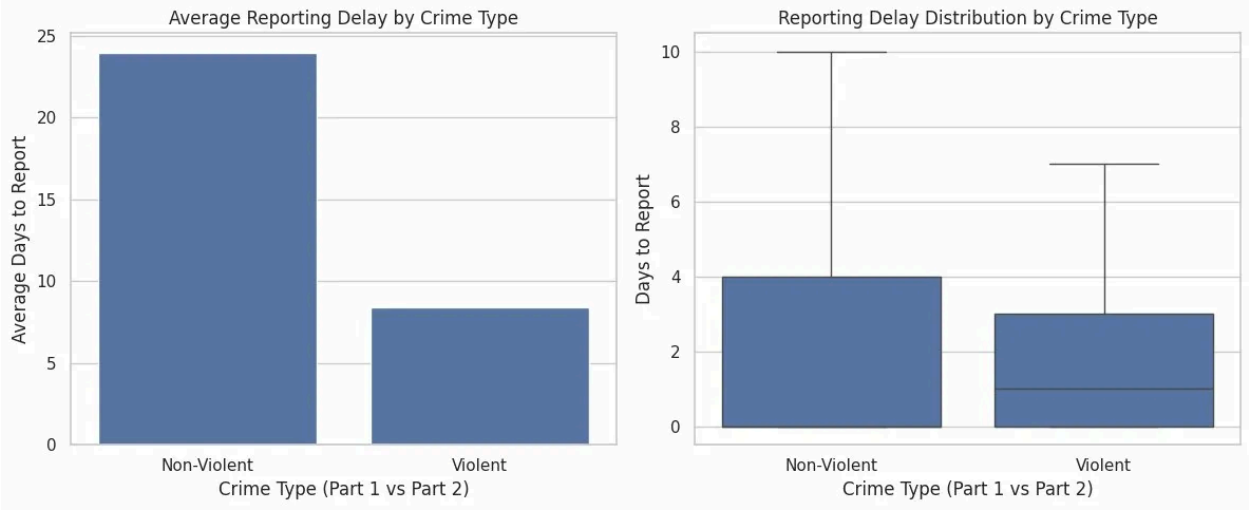
morning hours, peaks sharply around noon, and remains elevated through the afternoon and evening. The boxplot demonstrates that reporting delays are relatively consistent across all hours, although crimes occurring around midnight and noon exhibit slightly higher median delays. These findings suggest that the time of a crime has a limited impact on how quickly it is reported.



The analysis then examined how reporting delays vary across LAPD areas. Average delays ranged from approximately 11 to 22 days, depending on the area, with Devonshire exhibiting the longest average and Central among the shortest. The boxplot confirms that most areas have similar median delays of around 1 to 5 days, indicating that higher averages are likely driven by a small number of cases with very long delays rather than a widespread pattern of slow reporting in those areas.

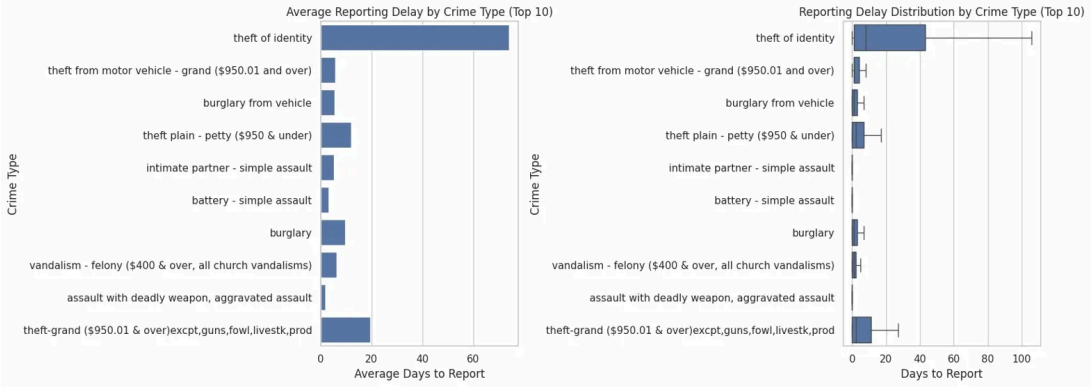


Reporting delays were compared between violent and non-violent crimes. Non-violent crimes averaged approximately 24 days to report, nearly three times the average for violent crimes at around 8 days. The boxplot indicates that violent crimes have a tighter distribution, suggesting they are reported more consistently and quickly. This pattern aligns with the expectation that violent incidents prompt immediate action from victims or witnesses.

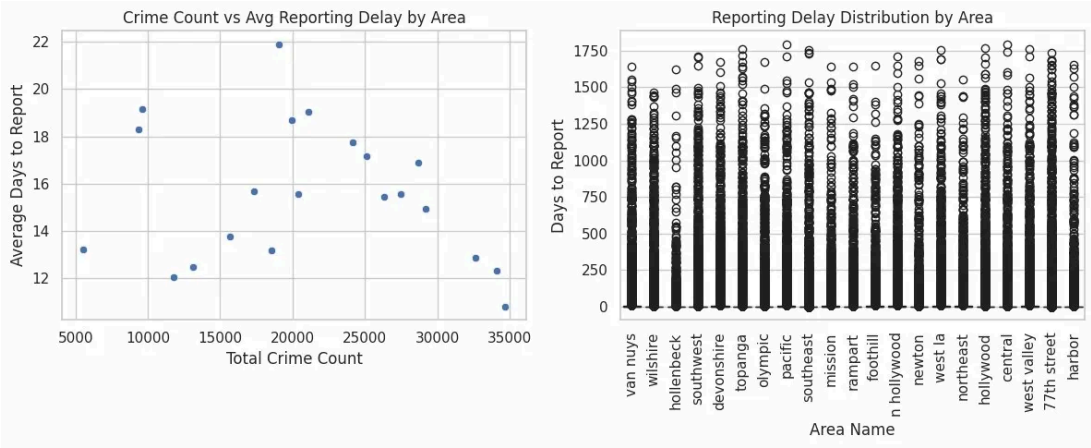


Among the top 10 most common specific crime types, identity theft had an average reporting delay of approximately 70 days, substantially higher than for any other crime type. This result is expected, as victims often do not discover identity theft until weeks or months after it occurs. All

other crime types in the top 10, including vehicle burglary, assault, and petty theft, had much shorter average delays of under 15 days.

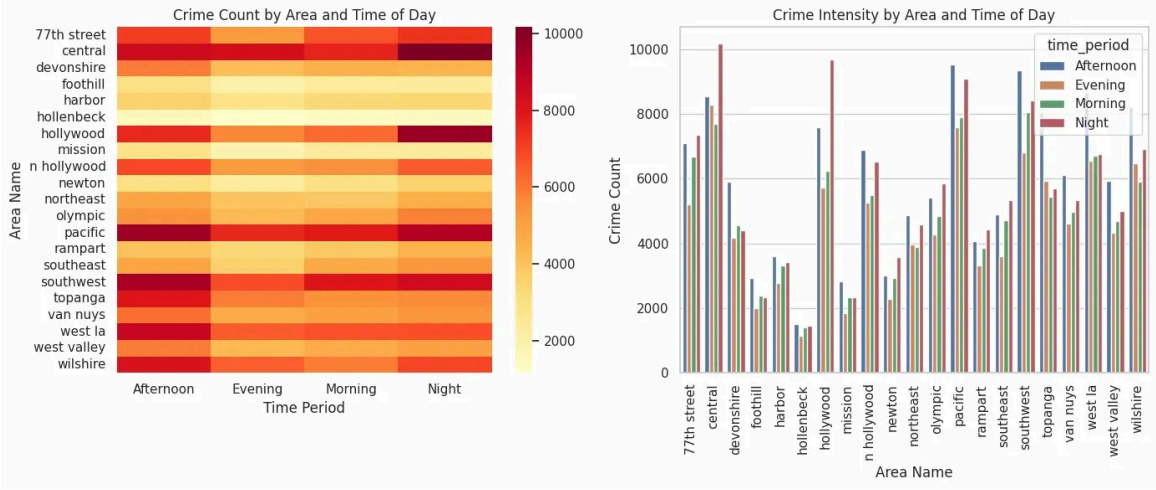


The analysis examined whether areas with higher crime counts also tend to have longer reporting delays. The scatter plot shows no clear relationship between total crime counts and average reporting delays across areas. High-crime areas do not consistently exhibit longer or shorter delays, suggesting that crime volume does not significantly affect how quickly residents report incidents.

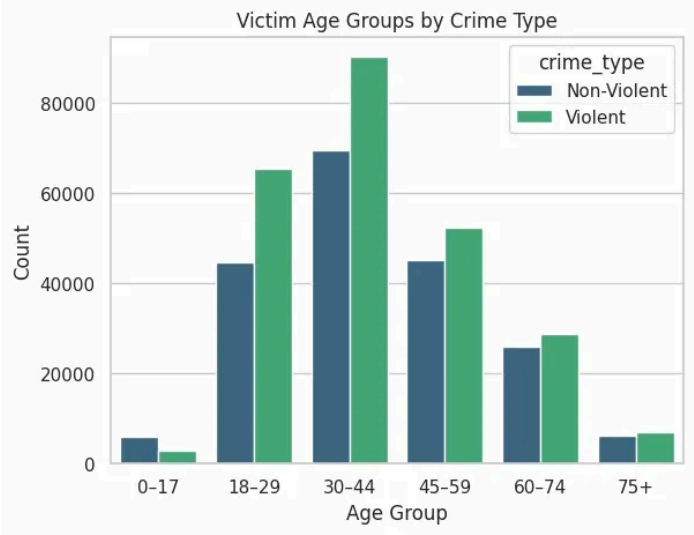


Finally, the distribution of crime activity across times of day within each LAPD area was analyzed. The heatmap indicates that night and afternoon are the busiest periods in nearly all areas, with Hollywood, Central, and Pacific recording the highest overall crime counts. Morning

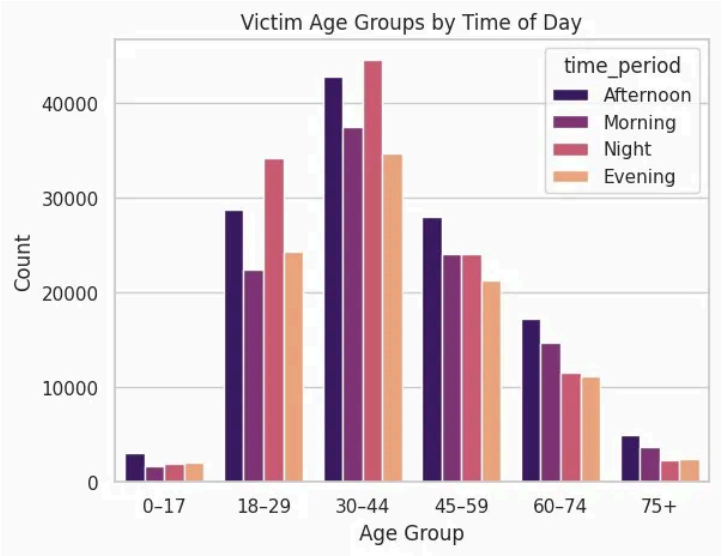
consistently exhibits the lowest activity across all areas. The grouped bar chart corroborates this pattern, with night crimes leading in most areas and morning crimes consistently trailing.



An analysis of victim age groups by crime type indicates that adults aged 30–44 are the most frequently victimized for both violent and non-violent crimes, followed by the 18–29 and 45–59 age groups. Violent crimes consistently outnumber non-violent crimes across all age groups, with the largest disparity observed in the 30–44 range. Victims under 18 and those aged 75 and above exhibit the lowest overall counts.

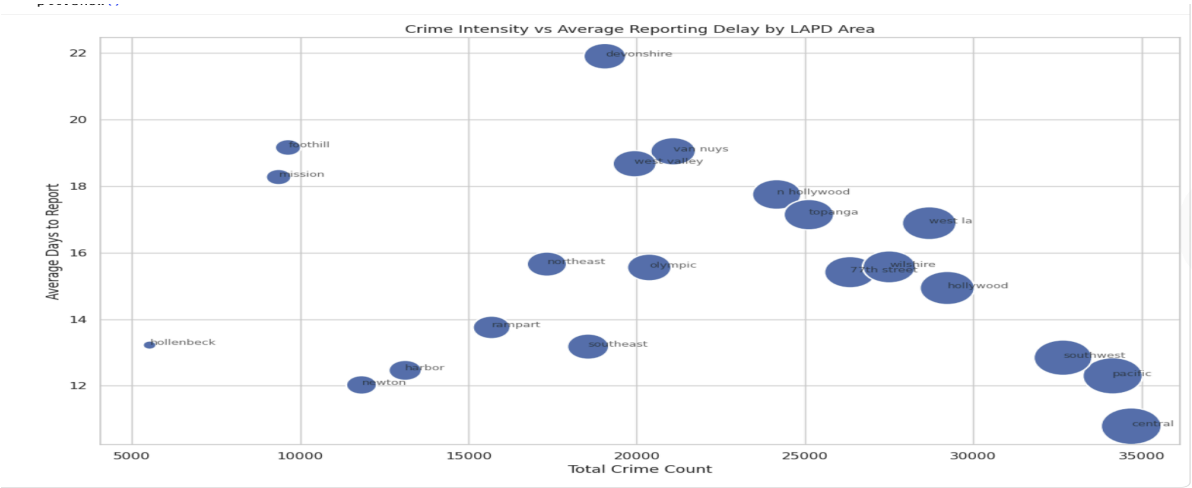


The distribution of victim age groups across different times of day reveals that the 30–44 age group consistently experiences the highest crime counts in all time periods. Night and afternoon are the most active periods for nearly every age group, while mornings show the lowest counts. The overall age distribution remains consistent across time periods, indicating that the demographic profile of victims is relatively stable regardless of when the crime occurs.

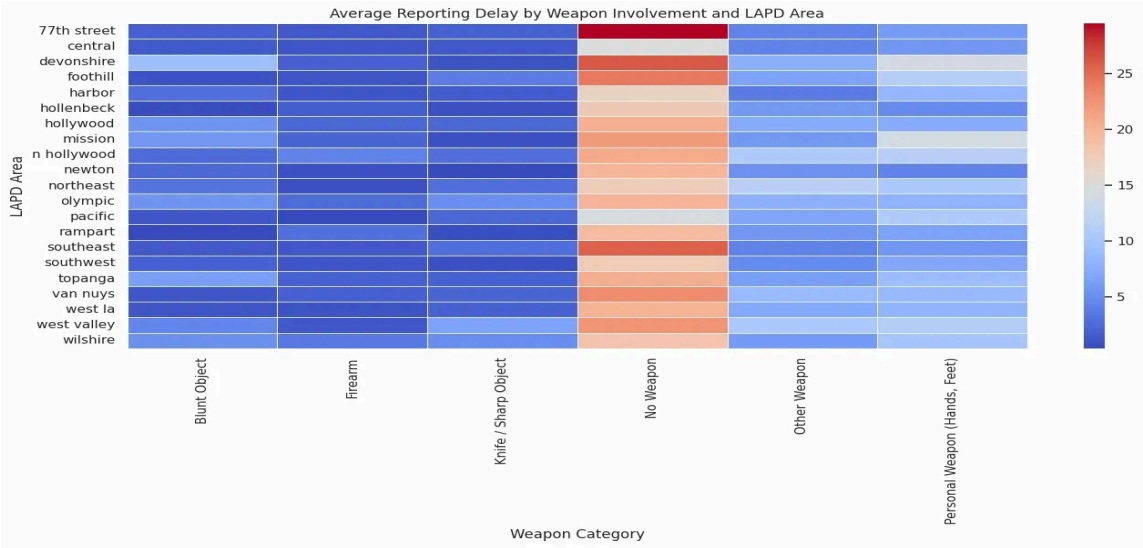


Multivariate Analysis

The relationship between crime concentration and reporting behavior was examined. A scatterplot illustrates this relationship by comparing total crime volume with average reporting delay across LAPD areas. Although certain high-crime areas exhibit longer reporting delays, this pattern is not consistent throughout all areas. These findings indicate that crime intensity and reporting behavior are related yet distinct processes that warrant separate modeling.



The analysis also considered whether crimes involving different weapon types are reported at varying speeds and whether this varies by area. The heatmap indicates that crimes without a weapon involved consistently experience the longest reporting delays across nearly all LAPD areas, with particularly high delays in Southeast and 77th Street. In contrast, crimes involving firearms, knives, and blunt objects are reported much faster, as indicated by the darker blue tones in those columns. This pattern is consistent across areas, suggesting that the presence of a physical weapon is a stronger determinant of faster reporting than geographic location.



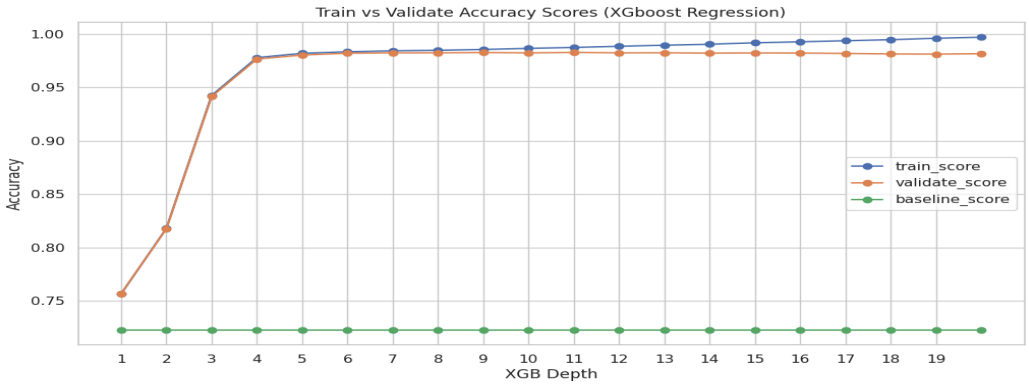
Modeling, Visualizations and Evaluation

Pre-processing

Prior to modeling, data preparation involved encoding, splitting, and saving the data as a CSV file. Categorical variables, including area name, victim descent, victim sex, weapon category, crime type, and premise description, were one-hot encoded using dummy variables, with the first category omitted to prevent multicollinearity. The encoded columns were combined with numeric features such as hour of occurrence, report number, occurrence, and report day, crime codes, and the binary reporting duration target variable. The dataset was then split into training, validation, and test sets at 60/20/20, with stratification on the target variable to maintain class balance. This process produced three subsets for model development and evaluation. The same 60/20/20 split was applied to the aggregated crime counts dataframe for count-based regression models, but without stratification because the target is continuous. Both the cleaned full dataset and all six split files were saved to disk for subsequent modeling.

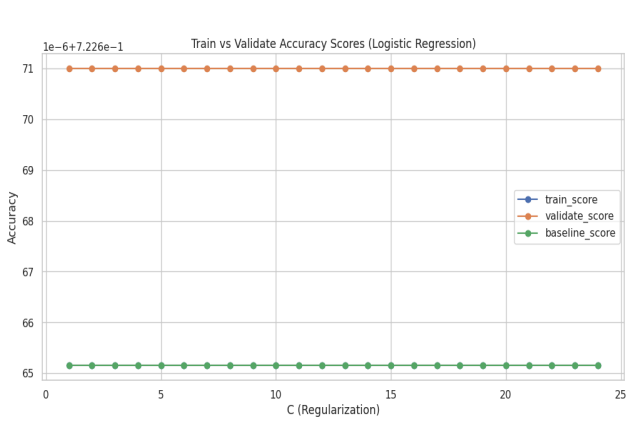
XGboost

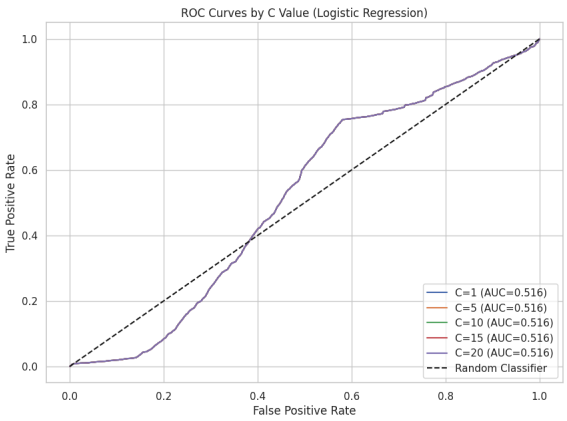
XGBoost was trained with a binary logistic objective, 200 estimators, a learning rate of 0.1, and 80% row and feature subsampling per tree. Validation accuracy increased sharply between depths 1 and 4, stabilizing at approximately 0.982 compared to a baseline of 0.723. The minimal gap between training and validation accuracy across all depths indicates strong generalization and limited overfitting.

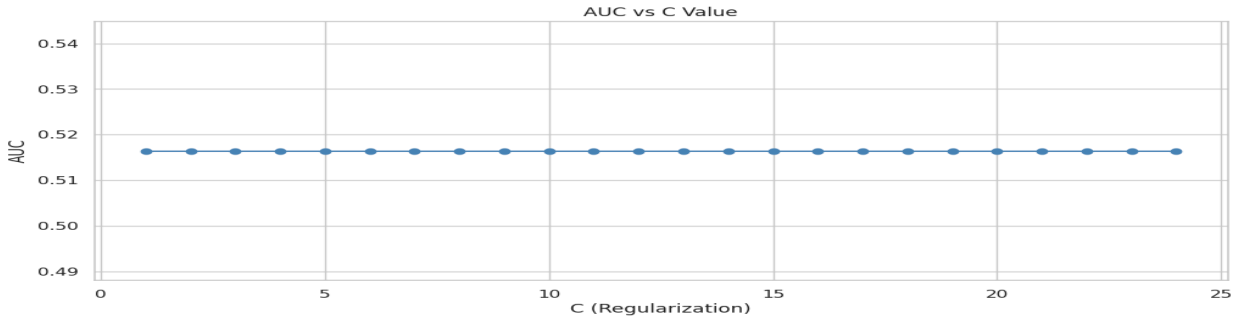


Logistic Regression

Logistic Regression was trained across C values ranging from 1 to 24. Both training and validation accuracy remained flat at approximately 0.71 across all regularization strengths, and the AUC held steady at 0.516, which is only marginally above the 0.5 threshold of a random classifier. The consistently flat ROC curves indicate that the model is unable to meaningfully separate the two classes, regardless of regularization. These results demonstrate that a linear decision boundary is inadequate for capturing the nonlinear relationships in the data, rendering logistic regression unsuitable for this classification task.

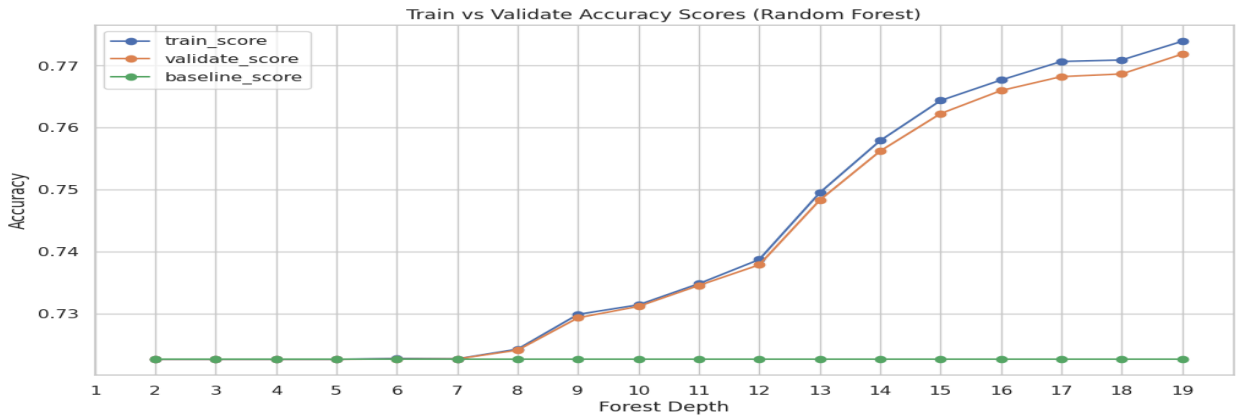






Random Forest

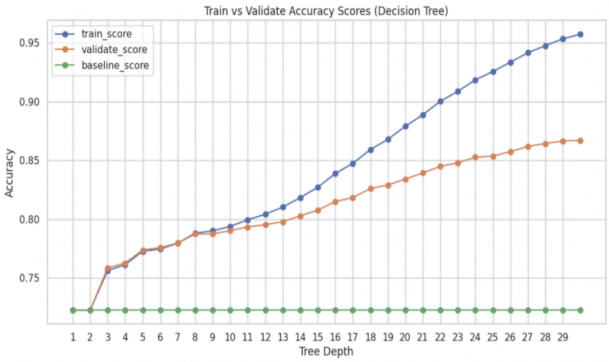
Random Forest was trained with 100 estimators, with both maximum depth and minimum samples per leaf set to the tested depth value. Validation accuracy improved gradually from approximately 0.72 at shallow depths to about 0.77 at deeper levels, consistently outperforming the baseline. The small gap between training and validation scores throughout indicates reasonable generalization, though the modest accuracy ceiling suggests the ensemble does not fully capture the data's underlying complexity.

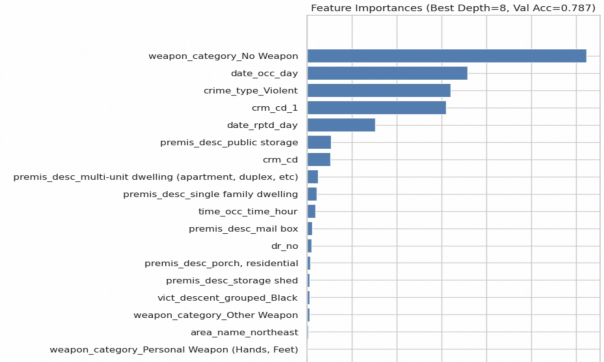


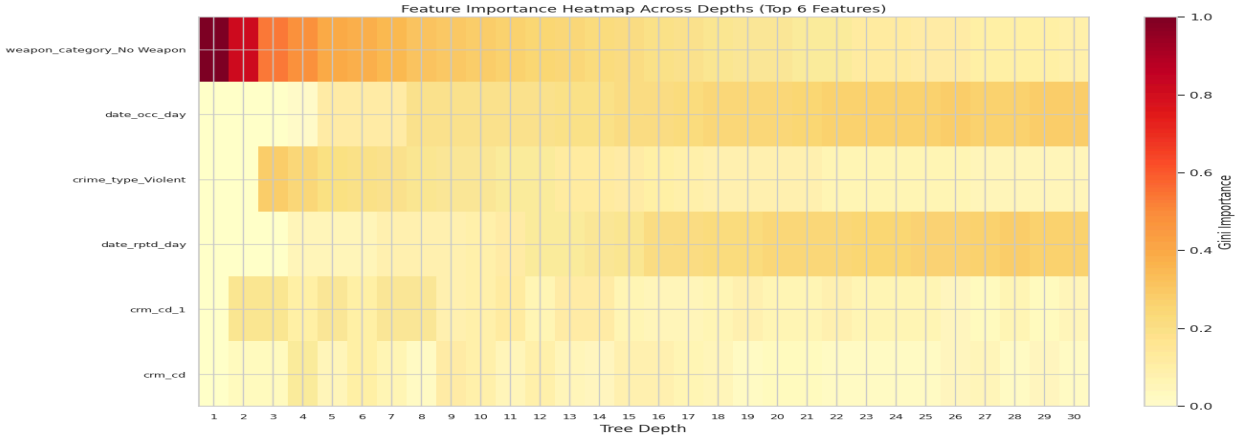
Decision Tree

The Decision Tree was trained across depths from 1 to 30. Validation accuracy improved steadily, reaching approximately 0.86-0.87 at deeper depths, well above the 0.723 baseline. However, training accuracy increased more rapidly, exceeding 0.95 at deeper levels, and the

widening gap between training and validation accuracy indicates increasing overfitting with greater depth. At the optimal validation depth of 8, the most important feature, by a significant margin, was whether a weapon was involved, followed by day of occurrence, crime type, crime code, report day, and premise type. The feature importance heatmap demonstrates that the weapon category was dominant at shallow depths, but its relative importance decreased as additional features became relevant at greater depths.







Poisson Regression

Prior to model fitting, the mean and variance of crime counts were compared. The mean was approximately 1,018, while the variance was approximately 298,992, nearly 300 times larger, confirming severe overdispersion. Although Poisson regression is technically appropriate when

variance exceeds the mean, a dispersion statistic of 87.6 indicates that the Poisson model is a poor fit for these data. In practice, this results in severe underestimation of standard errors, leading to coefficient p-values that appear falsely significant. The primary issue is that the Poisson model assumes equal mean and variance, yet crime counts vary substantially across LAPD areas. For example, Central and Hollywood experience far more incidents than Hollenbeck or Mission, violating this assumption.

Generalized Linear Model Regression Results

Dep. Variable:	crime_count	No. Observations:	261
Model:	GLM	Df Residuals:	235
Model Family:	Poisson	Df Model:	25
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-13906.
Date:	Fri, 20 Feb 2026	Deviance:	25619.
Time:	23:50:09	Pearson chi2:	2.06e+04
No. Iterations:	5	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	283.5253	4.846	58.510	0.000	274.028	293.023
C(area_name)[T.central]	0.3887	0.011	35.480	0.000	0.367	0.410
C(area_name)[T.devonshire]	-0.2466	0.013	-19.064	0.000	-0.272	-0.221
C(area_name)[T.foothill]	-1.0262	0.014	-72.361	0.000	-1.054	-0.998
C(area_name)[T.harbor]	-0.6571	0.016	-41.014	0.000	-0.689	-0.626
C(area_name)[T.hollenbeck]	-1.6086	0.019	-84.927	0.000	-1.646	-1.572
C(area_name)[T.hollywood]	0.1476	0.011	13.919	0.000	0.127	0.168
C(area_name)[T.mission]	-0.9440	0.016	-58.230	0.000	-0.976	-0.912
C(area_name)[T.n hollywood]	-0.1232	0.012	-10.674	0.000	-0.146	-0.101
C(area_name)[T.newton]	-0.7792	0.015	-53.656	0.000	-0.808	-0.751
C(area_name)[T.northeast]	-0.5311	0.012	-43.562	0.000	-0.555	-0.507
C(area_name)[T.olympic]	-0.2779	0.012	-23.260	0.000	-0.301	-0.254
C(area_name)[T.pacific]	0.2764	0.011	25.887	0.000	0.255	0.297
C(area_name)[T.rampart]	-0.4340	0.014	-29.937	0.000	-0.462	-0.406
C(area_name)[T.southeast]	-0.3785	0.014	-27.907	0.000	-0.405	-0.352
C(area_name)[T.southwest]	0.1908	0.011	17.266	0.000	0.169	0.212
C(area_name)[T.topanga]	-0.0034	0.011	-0.310	0.757	-0.025	0.018
C(area_name)[T.van nuys]	-0.1939	0.012	-16.602	0.000	-0.217	-0.171
C(area_name)[T.west la]	0.1038	0.012	8.711	0.000	0.080	0.127
C(area_name)[T.west valley]	-0.2540	0.011	-22.139	0.000	-0.277	-0.232
C(area_name)[T.wilshire]	0.0377	0.012	3.194	0.001	0.015	0.061
C(time_period)[T.Evening]	-0.3036	0.007	-41.023	0.000	-0.318	-0.289
C(time_period)[T.Morning]	-0.2314	0.006	-40.106	0.000	-0.243	-0.220
C(time_period)[T.Night]	-0.0874	0.007	-13.065	0.000	-0.101	-0.074
date_occ_year	-0.1366	0.002	-57.084	0.000	-0.141	-0.132
avg_reporting_delay	-0.0032	0.000	-7.558	0.000	-0.004	-0.002

Negative Binomial Regression

To address overdispersion, a Negative Binomial Generalized Linear Model (GLM) was fit using the same predictors: area name, time period, year of occurrence, and average reporting delay. The resulting dispersion parameter of 0.13 indicates a much tighter fit to the data compared to the Poisson model. Most area and time period coefficients were statistically significant, and the negative year coefficient aligns with the observed decline in crime counts after 2023. Incident Rate Ratios indicated that Central, Hollywood, Pacific, and Southwest had higher crime rates than the reference area, while Hollenbeck, Foothill, and Mission exhibited substantially lower rates. Morning and evening periods had lower counts than the afternoon. On the validation set,

the model achieved a mean absolute error (MAE) of 253 and a root mean squared error (RMSE) of 315, with test performance close behind at an MAE of 271 and RMSE of 340. The validation dispersion statistic of 135 suggests some residual overdispersion in the held-out data, indicating that the model captures broad patterns but struggles with the most extreme counts.

Generalized Linear Model Regression Results

Dep. Variable:	crime_count	No. Observations:	436
Model:	GLM	Df Residuals:	410
Model Family:	NegativeBinomial	Df Model:	25
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3398.5
Date:	Sat, 21 Feb 2026	Deviance:	180.19
Time:	00:09:16	Pearson chi2:	52.7
No. Iterations:	9	Pseudo R-squ. (CS):	0.2309
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	341.4175	117.014	2.918	0.004	112.075	570.760
C(area_name)[T.central]	0.3186	0.318	1.001	0.317	-0.305	0.942
C(area_name)[T.devonshire]	-0.4149	0.315	-1.318	0.187	-1.032	0.202
C(area_name)[T.foothill]	-1.0678	0.314	-3.404	0.001	-1.683	-0.453
C(area_name)[T.harbor]	-0.6797	0.318	-2.141	0.032	-1.302	-0.057
C(area_name)[T.hollenbeck]	-1.5863	0.313	-5.064	0.000	-2.200	-0.972
C(area_name)[T.hollywood]	0.1116	0.316	0.353	0.724	-0.509	0.732
C(area_name)[T.mission]	-1.0375	0.318	-3.264	0.001	-1.660	-0.415
C(area_name)[T.n hollywood]	-0.0797	0.318	-0.251	0.802	-0.702	0.543
C(area_name)[T.newton]	-0.7771	0.317	-2.448	0.014	-1.399	-0.155
C(area_name)[T.northeast]	-0.5844	0.304	-1.924	0.054	-1.180	0.011
C(area_name)[T.olympic]	-0.2378	0.316	-0.751	0.452	-0.858	0.382
C(area_name)[T.pacific]	0.2889	0.317	0.910	0.363	-0.333	0.911
C(area_name)[T.rampart]	-0.5472	0.313	-1.749	0.080	-1.160	0.066
C(area_name)[T.southeast]	-0.3970	0.313	-1.268	0.205	-1.010	0.217
C(area_name)[T.southwest]	0.2571	0.317	0.812	0.417	-0.364	0.878
C(area_name)[T.topanga]	-0.1757	0.307	-0.573	0.567	-0.777	0.425
C(area_name)[T.van nuys]	-0.3154	0.311	-1.014	0.311	-0.925	0.294
C(area_name)[T.west la]	0.0799	0.317	0.252	0.801	-0.541	0.701
C(area_name)[T.west valley]	-0.3155	0.315	-1.003	0.316	-0.932	0.301
C(area_name)[T.wilshire]	0.0559	0.316	0.177	0.860	-0.564	0.676
C(time_period)[T.Evening]	-0.1832	0.175	-1.045	0.296	-0.527	0.160
C(time_period)[T.Morning]	-0.1421	0.138	-1.030	0.303	-0.413	0.128
C(time_period)[T.Night]	0.0051	0.161	0.032	0.975	-0.310	0.321
date_occ_year	-0.1653	0.058	-2.861	0.004	-0.279	-0.052
avg_reporting_delay	0.0033	0.011	0.307	0.759	-0.018	0.024

Incident Rate Ratios (IRR):

Intercept	1.886884e+148
C(area_name)[T.central]	1.375191e+00
C(area_name)[T.devonshire]	6.604222e-01
C(area_name)[T.foothill]	3.437796e-01
C(area_name)[T.harbor]	5.067632e-01
C(area_name)[T.hollenbeck]	2.046866e-01
C(area_name)[T.hollywood]	1.118049e+00
C(area_name)[T.mission]	3.543475e-01
C(area_name)[T.n hollywood]	9.234354e-01
C(area_name)[T.newton]	4.597584e-01
C(area_name)[T.northeast]	5.574338e-01
C(area_name)[T.olympic]	7.883602e-01
C(area_name)[T.pacific]	1.334986e+00
C(area_name)[T.rampart]	5.785922e-01
C(area_name)[T.southeast]	6.723593e-01
C(area_name)[T.southwest]	1.293185e+00
C(area_name)[T.topanga]	8.388661e-01
C(area_name)[T.van nuys]	7.295279e-01
C(area_name)[T.west la]	1.083149e+00
C(area_name)[T.west valley]	7.294302e-01
C(area_name)[T.wilshire]	1.057489e+00
C(time_period)[T.Evening]	8.325991e-01
C(time_period)[T.Morning]	8.675185e-01
C(time_period)[T.Night]	1.005097e+00
date_occ_year	8.476431e-01
avg_reporting_delay	1.003310e+00

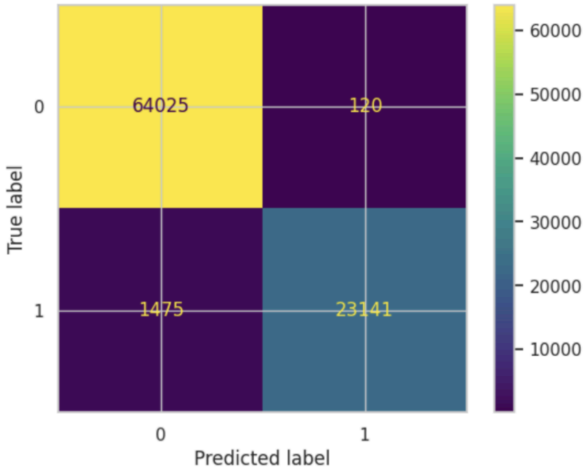
dtype: float64

Post Analysis

The XGBoost model performed very well, reaching an overall accuracy of 98% on the test set. This is a 35.5% improvement over the baseline accuracy of 72.27%. The classification report shows that XGBoost worked well for both classes. For low reporting delay (class 0), recall was perfect at 100%, and precision was high at 98%. For high reporting delay (class 1), precision was 99%, and recall was 94%. Both macro average and weighted average F1-scores were 98%, indicating the model effectively managed the class imbalance.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	64145
1	0.99	0.94	0.97	24616
accuracy			0.98	88761
macro avg	0.99	0.97	0.98	88761
weighted avg	0.98	0.98	0.98	88761

The confusion matrix supports these findings, with 64,025 true negatives and 23,141 true positives, and very few misclassifications. The model had only 120 false positives (0.19% error rate for low reporting delay) and 1,475 false negatives (6% error rate for high reporting delay). The low false-positive rate and strong true-positive results indicate that XGBoost learned to distinguish between the classes effectively. This makes it a strong choice for deployment, as it generalizes well and avoids overfitting.



Methodology

Our attempts at determining the crime reporting delays relied heavily on the data provided by the LAPD and public government data sets. Several preprocessing steps were taken before modeling, including the cleaning of data that did not contain timestamps making it hard to determine reporting delay as well as removing victim ages that were recorded as 0 or implausible values.

Reporting delay was calculated as the difference in days between the date a crime occurred and the date it was reported. Exploratory plots show that reporting delay is right skewed, with most incidents reported within a few days but a long tail extending beyond 60 days and because of this skewness, reporting delay was later converted into a classification variable rather than modeled purely as a continuous outcome.

Crime types were also grouped into violent (part 1) and non-violent (part 2) categories based on LAPD crime classifications. Temporal variables including house of occurrence, day of month, and month were extracted from the original date fields and spatial variables were constructed using LAPD area identifiers.

Research Question 1: Predicting Crime Reporting Delays in Los Angeles

Reporting delay was framed as a binary classification problem. Crimes reported within 48 hours were labeled as timely, while crimes reported after 3 days or more were labeled as delayed. Incidents reported between these thresholds were excluded to create clearer and separation between classes. Bar charts and boxplots comparing violent and non-violent crimes show that violent crimes have an average reporting delay of approximately 8 days, while non-violent crimes average over 23 days. The boxplots also show that nonviolent crimes exhibit longer upper tails, proving that delay behavior differs by crime severity.

The time of day plots show that times that occur over night, (between 12am - 5am), tend to have longer median reporting delays than daytime crimes, which supports the inclusion of hour of occurrence as a predictor. Day of month plots show relatively stable averages across most days, with the exception of day1, which exhibits a high mean due to outliers. And because the

distributions by day of month are relatively flat, that variable was included but did not dominate model performance.

All results were assessed using a combination of visual diagnostics, numerical metrics, and statistical testing for each research question. Our modeling focused on predicting crime reporting delays in Los Angeles. We ran four models suited for binary classification, including decision tree, random forest, logistic regression, and XGBoost. Model performance was evaluated using accuracy, with the best performing model being XGBoost at 98% accuracy in both training and testing, achieving a 26% improvement over the baseline prediction of 72% accuracy.

Research Question 2: Crime Hotspot Intensity Across Los Angeles

To analyze crime concentration, incidents were aggregated by LAPD area and time period, producing crime counts per area. Exploratory bar charts show that crime is uneven across areas with Central, Pacific, Southwest, and Hollywood, exceeding over 30,000 incidents each, whereas Hollenbeck and Mission had less than 10,000 incidents each.

Scatterplots comparing total crime count and average reporting delay show no strong linear relationship, indicating that crime intensity and reporting behavior should be treated as related but distinct processes.

Crime counts were modeled using count based regression methods. Poisson regression was considered initially but exploratory variance checks suggested overdispersion. As a result, negative binomial regression was used better to account for excess variance in crime counts.

The temporal heat maps also show that crime intensity peaks during evening and late night houses, specifically between 6pm and 12 am, and that high crime areas remain consistently high over the years.

Research Question 3: Victim Age Patterns in Los Angeles Crime

Victim age was modeled both as continuous variable and as grouped categories such as children, young adults, middle aged, and seniors). Histograms show a unimodal age distribution centered around the late 30's with a median victim age near 38-40 years. Box plots reveal a right tail extending into older ages with seniors appearing as less frequent but meaningful outliers.
Plots comparing victim age by crime type show that violent crimes skew younger while non violent crimes such as fraud and identity theft skew older. Time of day plots further show that late night crimes disproportionately involve younger victims, while daytime crimes more frequently affect middle aged and older individuals.

Challenges

We encountered several challenges throughout this project. The first major challenge involved data completeness and reliability. The original data set selected contained over a million records, but a large portion of the data had missing or implausible values for key variables such as victim demographics, timestamps, and premise descriptions. Any of the reports that were missing that data was removed, as it prevented us from accurately calculating the report delay of crimes. After cleaning the dataset, the report was reduced to almost less than half of its original size. By removing the reports with missing data, we reduced the potential of including potential bias within our report.
The second challenge was defining reporting delay in a meaningful way. Reporting delay is highly skewed; most crimes are reported quickly, while a handful take weeks or months to report. Modeling delay as a continuous variable caused instability and made patterns difficult to interpret. To address this, we converted the variable into a binary classification problem →

Timely vs. Delayed Reporting. By doing this conversion, it improved model performance, interpretability, as well as simplified a complex behavioral process.

A third challenge involved class imbalance and model evaluation. After assessing the data, we found that the majority of the data fell into the timely reporting category, which risked the models appearing accurate simply by predicting the majority class. To confirm that the model was learning meaningful patterns rather than exploiting imbalance, multiple evaluation metrics such as precision, recall, F-1 score and confusion matrices were used. Despite the strong accuracy, the model did not guarantee real world reliability because human reporting behavior is influenced by factors not present within the dataset such as trust in law enforcement, fear of retaliation, or knowledge of reporting procedures.

A fourth challenge we stumbled upon was interpreting spatial patterns. High crime areas did not always correspond to longer reporting delays, suggesting that crime volume and reporting behavior are distinct processes. This made it difficult to separate whether observed differences were caused by environment risk factors, population density, or social factors that were not included nor measured within the our dataset.

Ethical Recommendations

Our review of LAPD crime data shows significant differences in how quickly crimes are reported, depending on the area, the people involved, and the type of crime. Even though the XGBoost model predicts these patterns with 98% accuracy, strong results do not mean it is always ethical to use. Before using these models to decide where to send patrols, we need to address three important concerns.

First, sending more police to areas flagged as hotspots can lead to over-policing in communities that already have frequent law enforcement contact. Central, Hollywood, and Pacific, for example, have both high crime rates and longer delays in reporting, but just adding more officers does not solve the reasons people wait to report crimes. Issues like language barriers, distrust of police, fear of retaliation, and not knowing how to report are likely causes. These communities might benefit more from tools in multiple languages, community liaison programs, and working with local organizations that residents trust.

Second, losing half the data because of missing information makes us question if our models truly represent all parts of Los Angeles. Missing data is not random; it often stems from problems in data collection, which likely affect marginalized neighborhoods the most. Before using the model, we need to ensure it works well across every LAPD area and group, not just where the data are complete. People should also know how their communities are labeled as hotspots and what that means for them.

Third, we must be careful not to let patterns in the data lead to stereotypes or make us pay less attention to certain cases. Our results show that violent crimes happen more to younger people, fraud affects older adults, and reporting times vary by victim sex. Still, these trends do not tell us about each person's experience. Just because weapon-related crimes are reported faster does not mean we should doubt victims who take longer to report, since trauma affects everyone differently (Lanthier et al., 2018). Prevention programs should be shaped with community input to avoid messages that might stigmatize anyone.

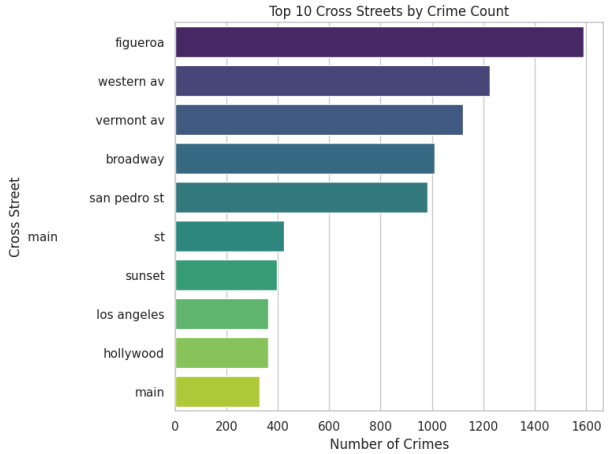
In the end, crime is often a sign of bigger social issues like poverty, inequality, and past trauma (Patterson, 1991). Predictive models can help guide decisions, but using them responsibly means also investing in social support and solutions.

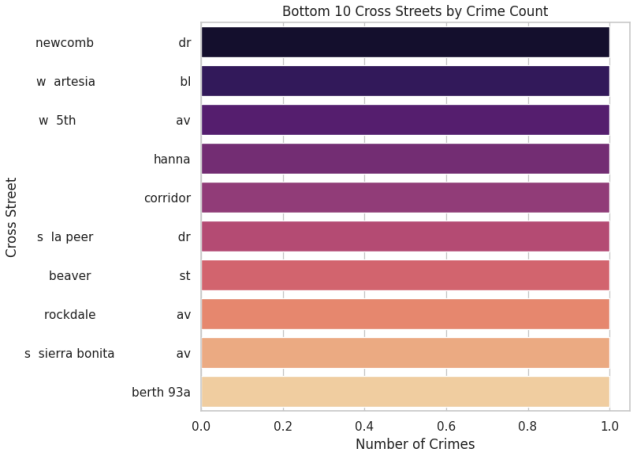
References

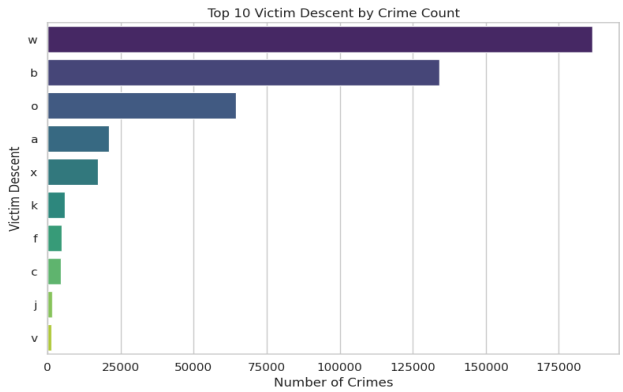
1. Los Angeles Police Department. *LAPD organization*. https://www.lapdonline.org/lapd-organization/
2. Los Angeles Police Department. *History of the LAPD*. https://www.lapdonline.org/history-of-the-lapd/
3. Los Angeles Police Department. *Community policing unit*. https://www.lapdonline.org/community-policing-unit/
4. North East Neighborhood Council. (2024, November). *LAPD budget approved: $2.14 billion spending plan for 2025–26*. https://nenc-la.org/2024/11/lapd-budget-approved-2-14-billion-spending-plan-for-2025-26/
5. Lanthier, S., Du Mont, J., & Mason, R. (2018). Responding to delayed disclosure of sexual assault in health settings: A systematic review. *Trauma, Violence, & Abuse*, 19(2), 251-265. https://doi.org/10.1177/1524838016659484
6. Patterson, E. B. (1991). Poverty, income inequality, and community crime rates. *Criminology*, 29(4), 755-776. https://doi.org/10.1111/j.1745-9125.1991.tb01087.x

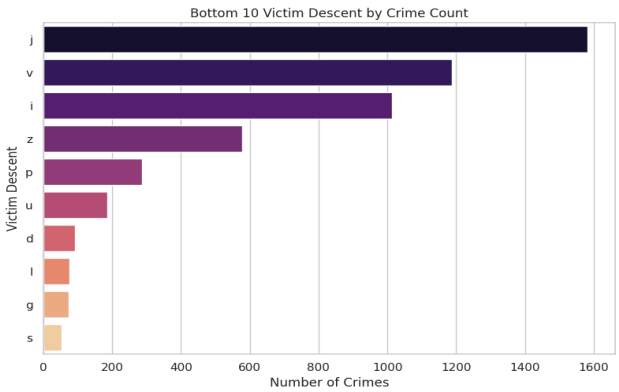
Appendix

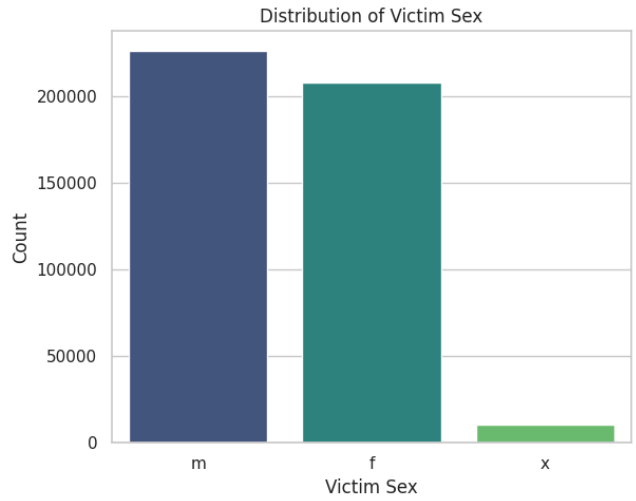
EDA

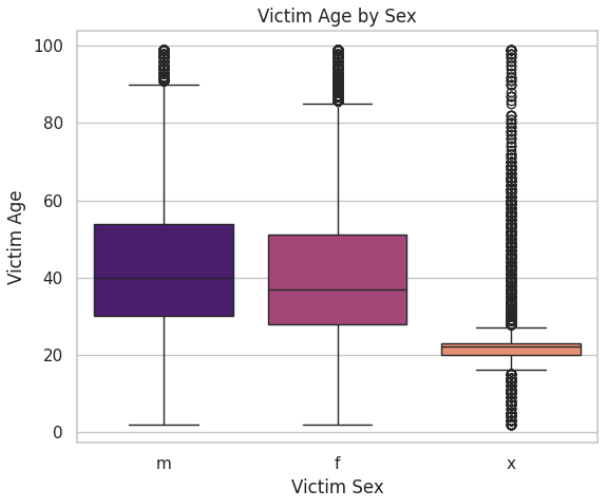


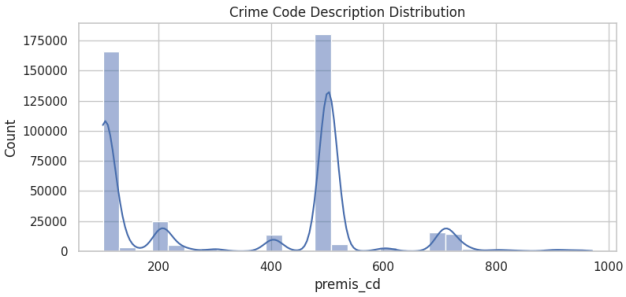


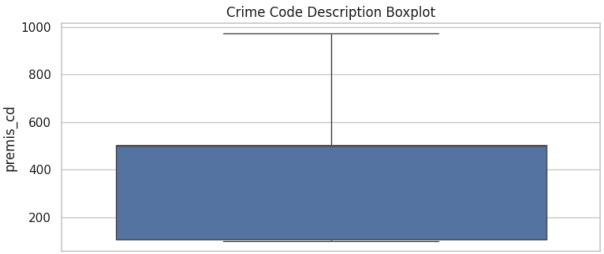


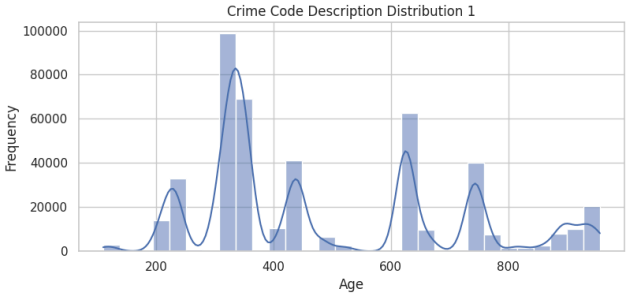


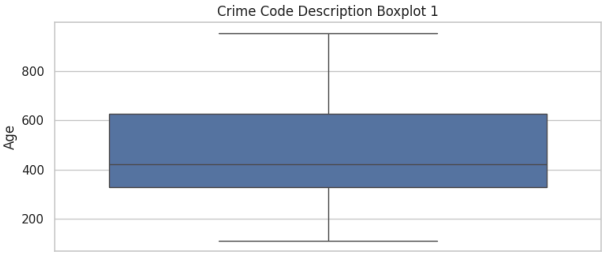





Crime Code Description Distribution
Count
premis_cd
0
25000
50000
75000
100000
125000
150000
175000
200
400
600
800
1000

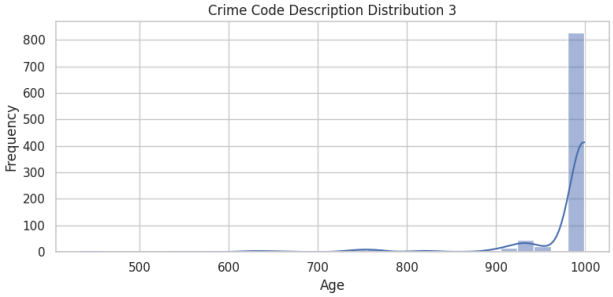

Crime Code Description Boxplot
premis_cd
200
400
600
800
1000


Crime Code Description Distribution 1
Frequency
Age
0
20000
40000
60000
80000
100000
200
400
600
800


Crime Code Description Boxplot 1
Age
200
400
600
800


Crime Code Description Distribution 2
Frequency
Age
0
5000
10000
15000
20000
200
300
400
500
600
700
800
900
1000

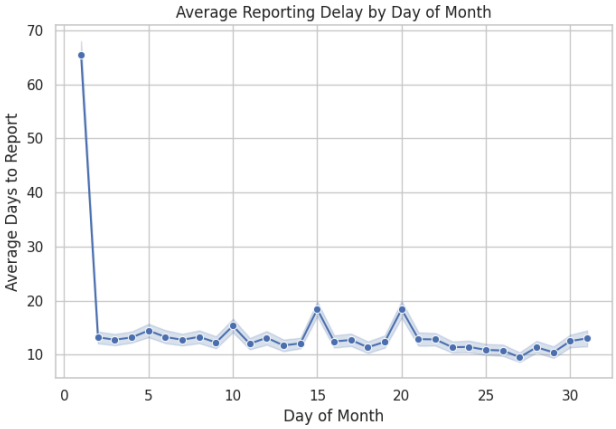

Crime Code Description Boxplot 2
Age
200
300
400
500
600
700
800
900
1000

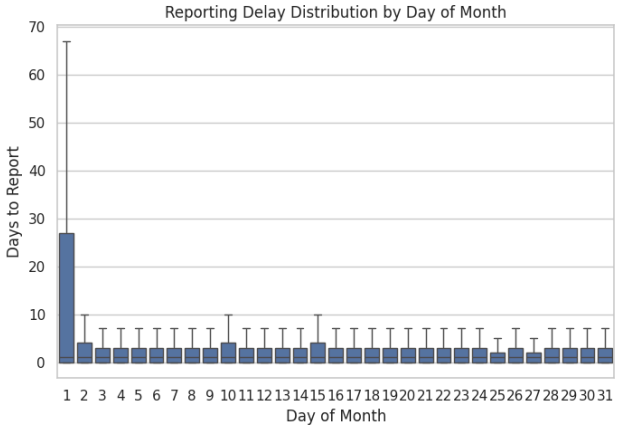

Crime Code Description Distribution 3
Frequency
Age
0
100
200
300
400
500
600
700
800
500
600
700
800
900
1000

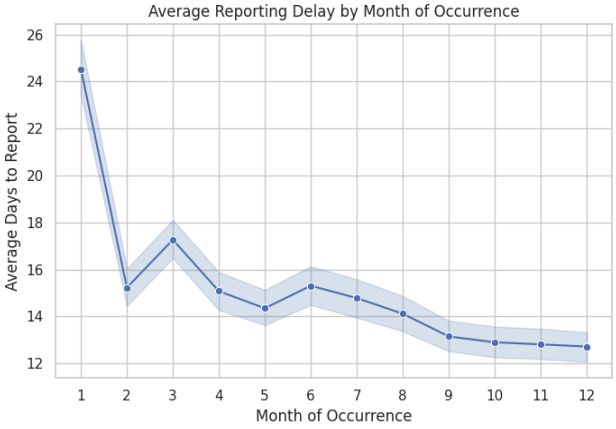

Crime Code Description Boxplot 3
Age
500
600
700
800
900
1000

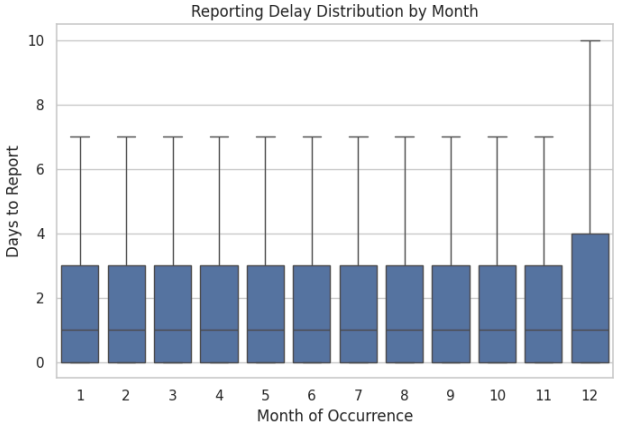

Crime Code Description Distribution 4
Frequency
Age
0
5
10
15
20
25
30
920
940
960
980
1000

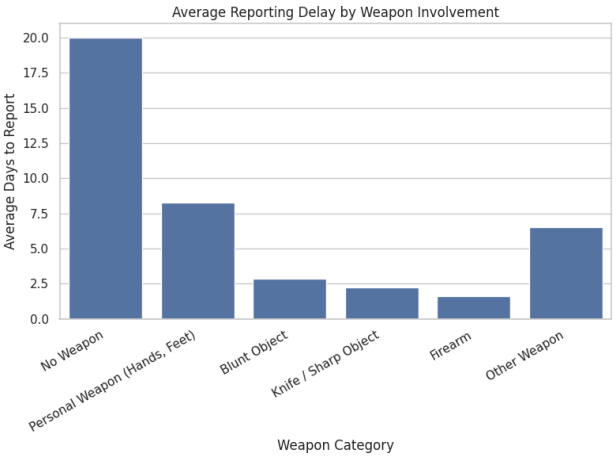

Crime Code Description Boxplot 4
Age
920
940
960
980
1000

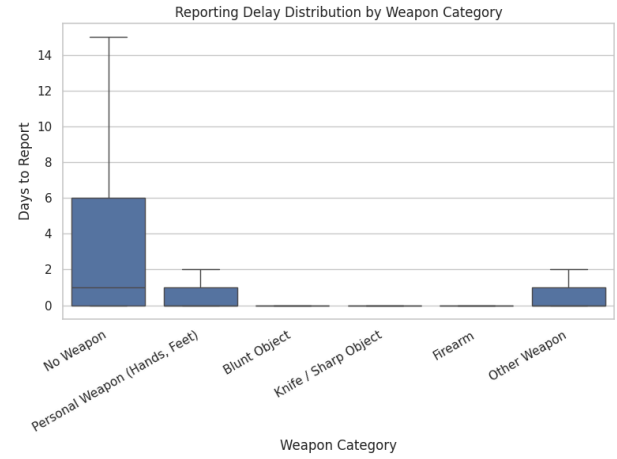

Average Reporting Delay by Day of Month
Average Days to Report
Day of Month

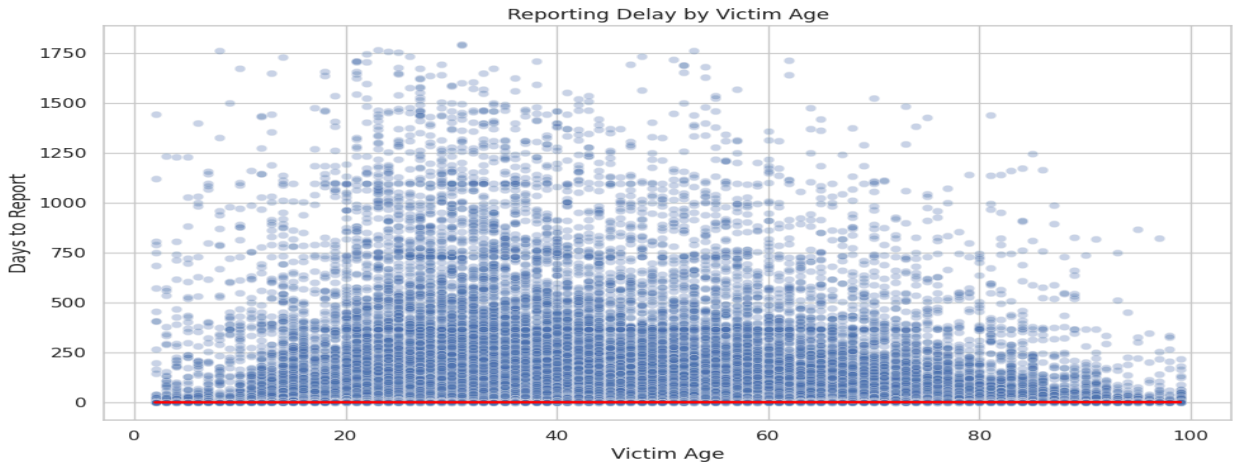

Reporting Delay Distribution by Day of Month
Days to Report
Day of Month

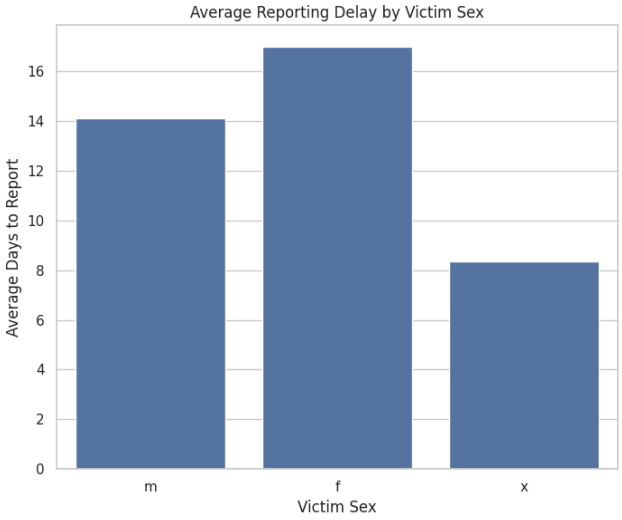

Average Reporting Delay by Month of Occurrence
Average Days to Report
Month of Occurrence

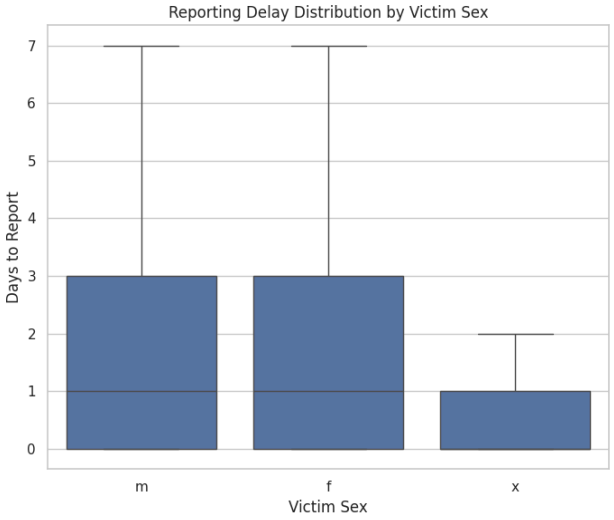

Reporting Delay Distribution by Month
Days to Report
Month of Occurrence

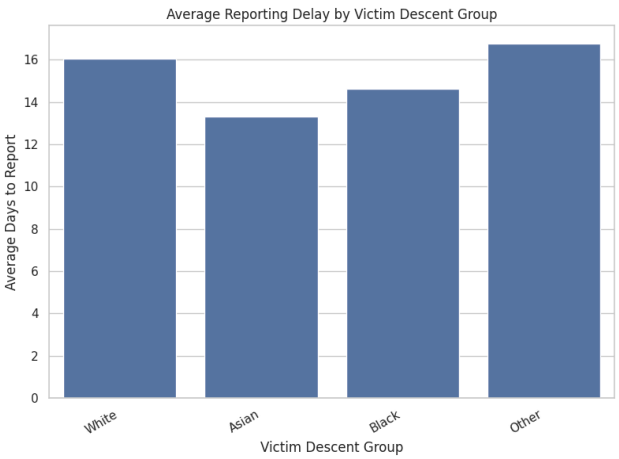

Average Reporting Delay by Weapon Involvement
Average Days to Report
No Weapon
Personal Weapon (Hands, Feet)
Blunt Object
Knife / Sharp Object
Firearm
Other Weapon
Weapon Category

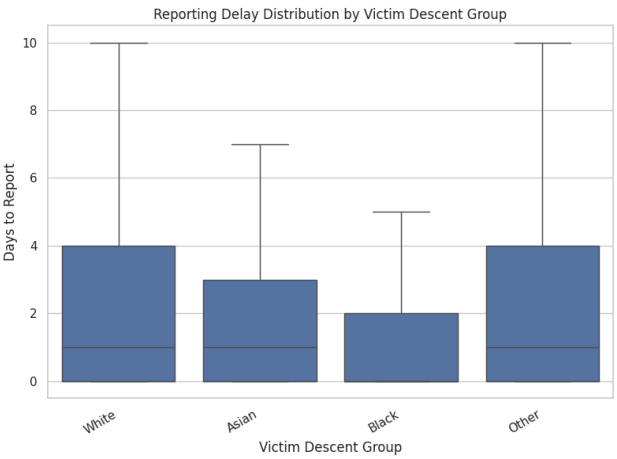

Reporting Delay Distribution by Weapon Category
Days to Report
No Weapon
Personal Weapon (Hands, Feet)
Blunt Object
Knife / Sharp Object
Firearm
Other Weapon
Weapon Category

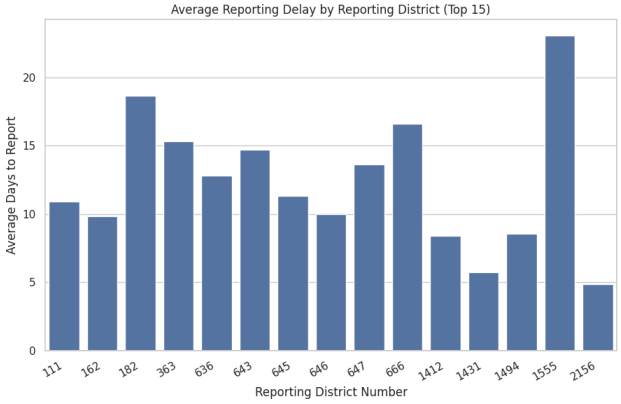

Reporting Delay by Victim Age
Days to Report
Victim Age
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100

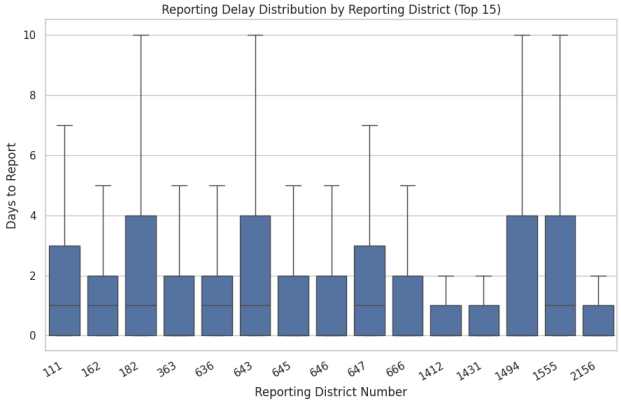

Average Reporting Delay by Victim Sex
Average Days to Report
Victim Sex
0
2
4
6
8
10
12
14
16
m
f
x

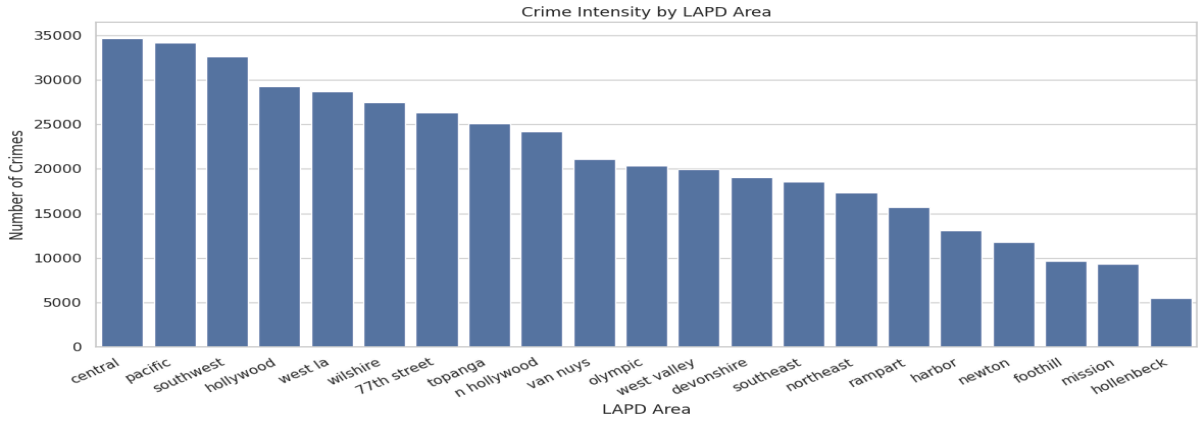

Reporting Delay Distribution by Victim Sex
Days to Report
Victim Sex
0
1
2
3
4
5
6
7
m
f
x

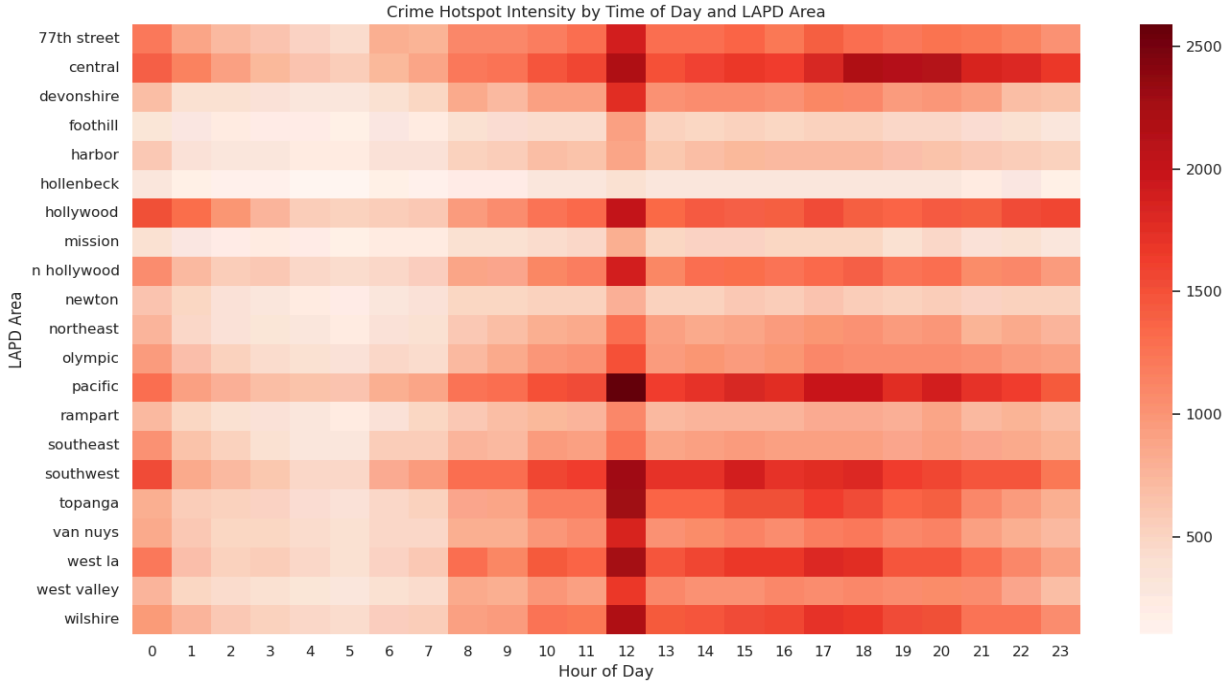

Average Reporting Delay by Victim Descent Group
Average Days to Report
Victim Descent Group
0
2
4
6
8
10
12
14
16
White
Asian
Black
Other

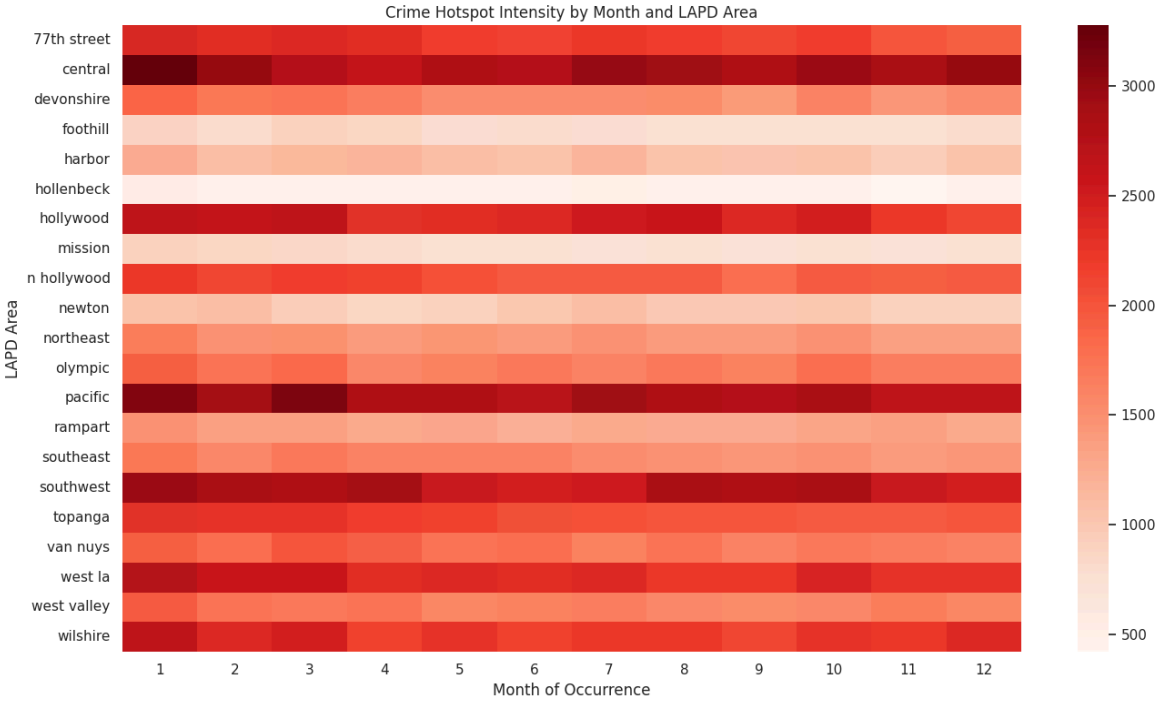

Reporting Delay Distribution by Victim Descent Group
Days to Report
Victim Descent Group
0
2
4
6
8
10
White
Asian
Black
Other

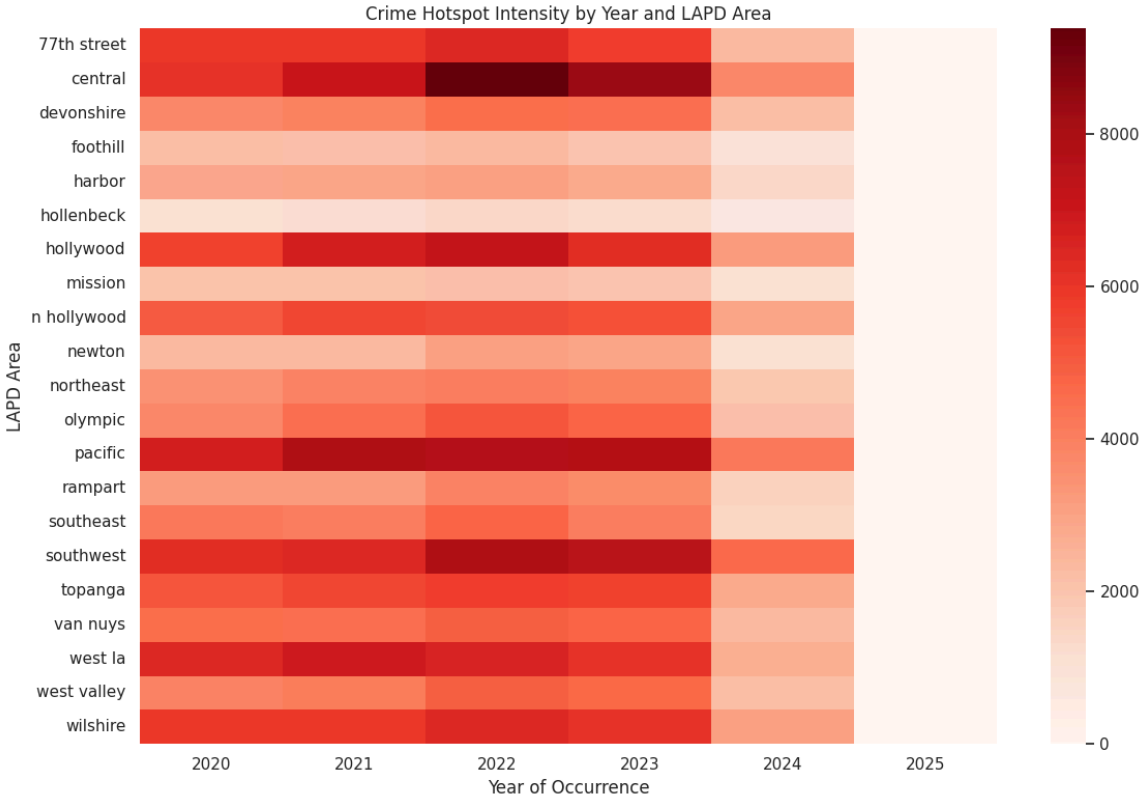

Average Reporting Delay by Reporting District (Top 15)
Average Days to Report
Reporting District Number
111
162
182
363
636
643
645
646
647
666
1412
1431
1494
1555
2156

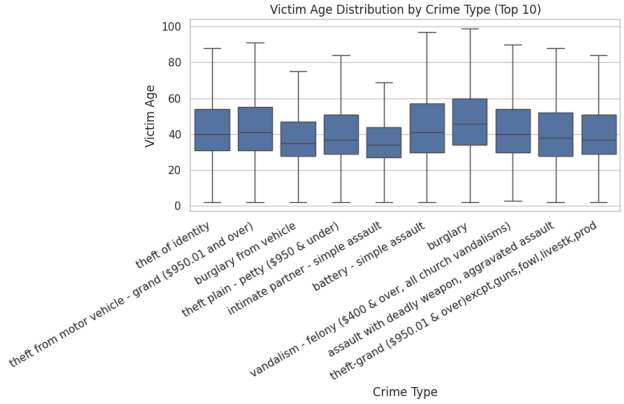

Reporting Delay Distribution by Reporting District (Top 15)
Days to Report
Reporting District Number
111
162
182
363
636
643
645
646
647
666
1412
1431
1494
1555
2156

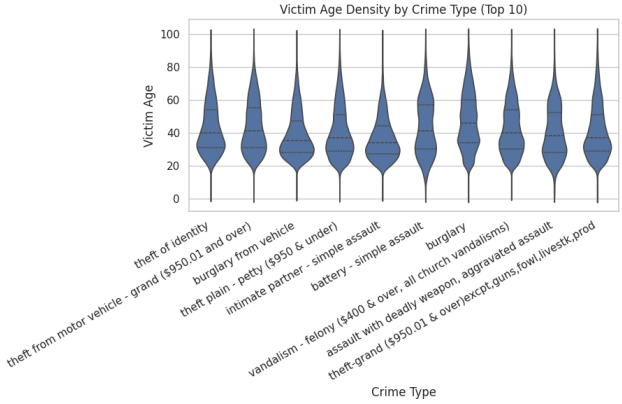

Crime Intensity by LAPD Area
Number of Crimes
LAPD Area
central
pacific
southwest
hollywood
west la
wilshire
77th street
topanga
n hollywood
van nuys
olympic
west valley
devonshire
southeast
northeast
rampart
harbor
newton
foothill
mission
hollenbeck

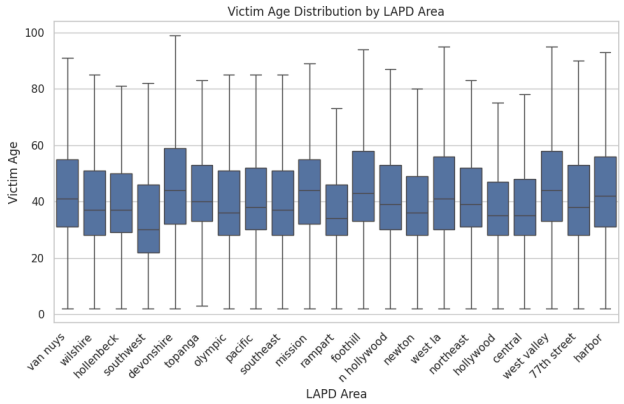

Crime Hotspot Intensity by Time of Day and LAPD Area
LAPD Area
77th street
central
devonshire
foothill
harbor
hollenbeck
hollywood
mission
n hollywood
newton
northeast
olympic
pacific
rampart
southeast
southwest
topanga
van nuys
west la
west valley
wilshire
Hour of Day

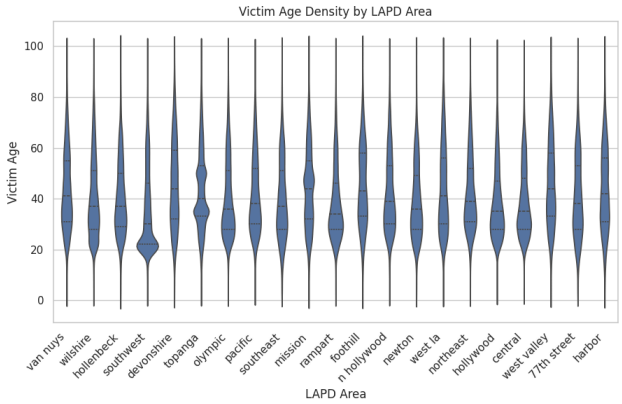

Crime Hotspot Intensity by Month and LAPD Area
77th street
central
devonshire
foothill
harbor
hollenbeck
hollywood
mission
n hollywood
newton
northeast
olympic
pacific
rampart
southeast
southwest
topanga
van nuys
west la
west valley
wilshire
LAPD Area
1
2
3
4
5
6
7
8
9
10
11
12
Month of Occurrence
3000
2500
2000
1500
1000
500

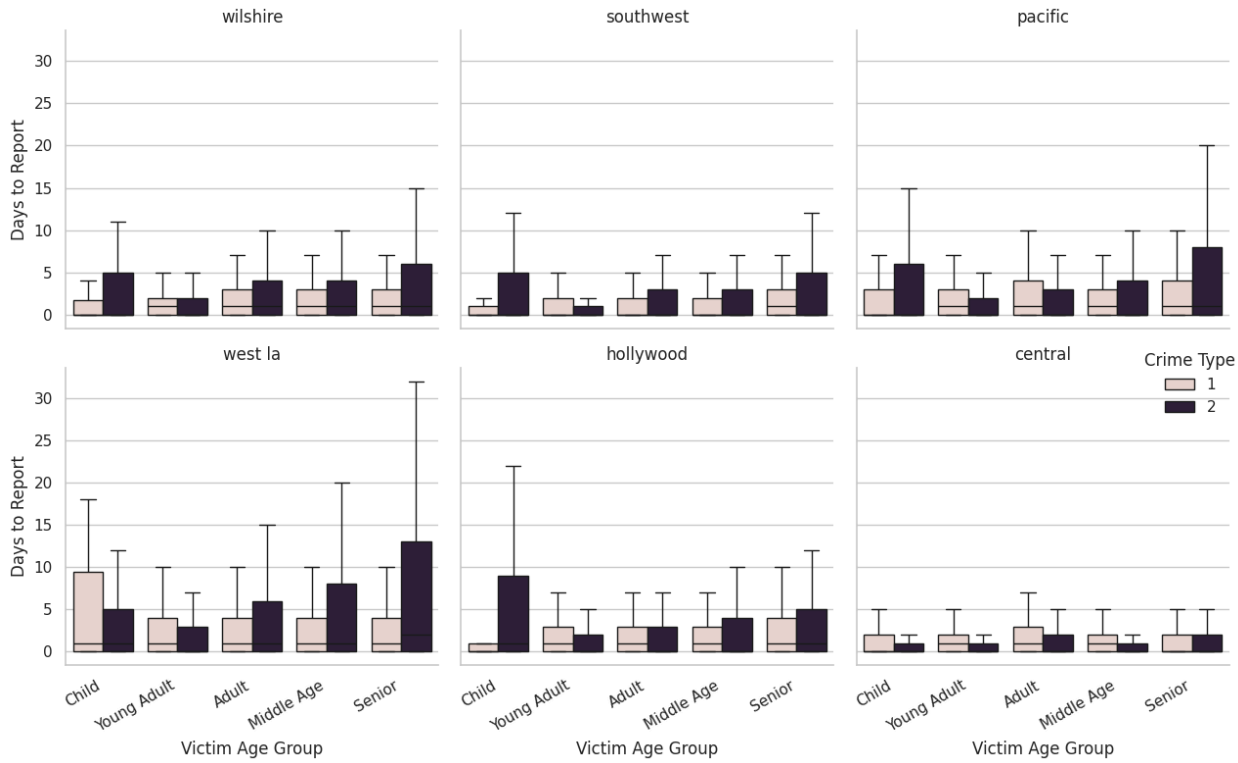

Crime Hotspot Intensity by Year and LAPD Area
77th street
central
devonshire
foothill
harbor
hollenbeck
hollywood
mission
n hollywood
newton
northeast
olympic
pacific
rampart
southeast
southwest
topanga
van nuys
west la
west valley
wilshire
LAPD Area
2020
2021
2022
2023
2024
2025
Year of Occurrence
8000
6000
4000
2000
0


Victim Age Distribution by Crime Type (Top 10)
Victim Age
Crime Type
theft of identity
theft from motor vehicle - grand ($950.01 and over)
burglary from vehicle
theft plain - petty ($950 & under)
intimate partner - simple assault
battery - simple assault
burglary
vandalism - felony ($400 & over, all church vandalisms)
assault with deadly weapon, aggravated assault
theft-grand ($950.01 & over)excpt,guns,fowl,livestk,prod


Victim Age Density by Crime Type (Top 10)
Victim Age
Crime Type
theft of identity
theft from motor vehicle - grand ($950.01 and over)
burglary from vehicle
theft plain - petty ($950 & under)
intimate partner - simple assault
battery - simple assault
burglary
vandalism - felony ($400 & over, all church vandalisms)
assault with deadly weapon, aggravated assault
theft-grand ($950.01 & over)excpt,guns,fowl,livestk,prod


Victim Age Distribution by LAPD Area
Victim Age
LAPD Area
van nuys
wilshire
hollenbeck
southwest
devonshire
topanga
olympic
pacific
southeast
mission
rampart
foothill
n hollywood
newton
west la
northeast
hollywood
central
west valley
77th street
harbor


Victim Age Density by LAPD Area
Victim Age
LAPD Area
van nuys
wilshire
hollenbeck
southwest
devonshire
topanga
olympic
pacific
southeast
mission
rampart
foothill
n hollywood
newton
west la
northeast
hollywood
central
west valley
77th street
harbor


wilshire
southwest
pacific
west la
hollywood
central
Days to Report
Victim Age Group
Crime Type
1
2
Child
Young Adult
Adult
Middle Age
Senior

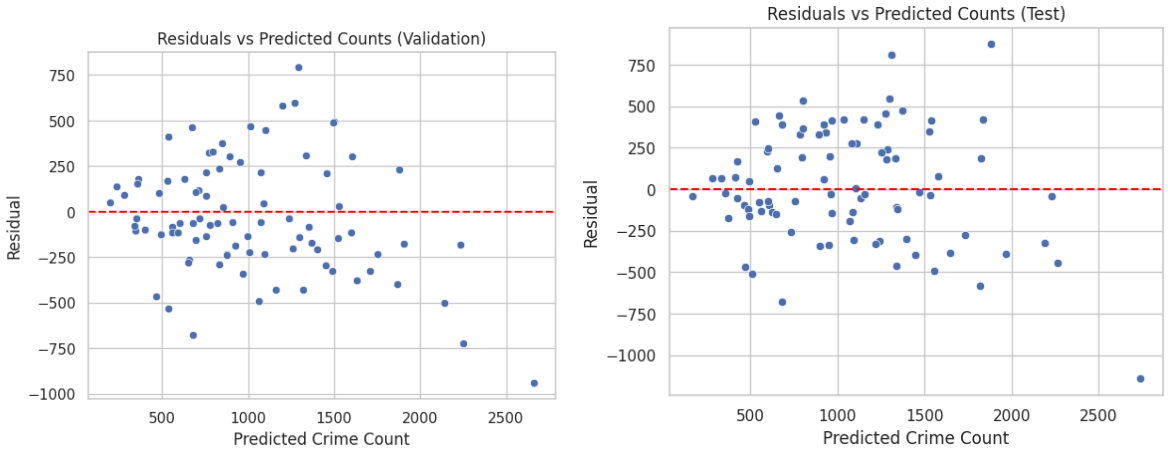
Stats testing

1. Does the speed of crime reporting vary by hour of the day?
    a. Kruskal-Wallis H Test | $H = 12{,}410.71$, $p < 0.05$. The null hypothesis is rejected, indicating that reporting delays differ significantly across hours of the day.
2. How do reporting delays vary across LAPD areas?
    a. Kruskal-Wallis H Test | $H = 13{,}710.34$, $p < 0.05$. The null hypothesis is rejected, demonstrating significant differences in reporting delays across at least one LAPD area.
3. How do reporting delays differ among specific crime types?
    a. One-Way ANOVA | $F = 2{,}834.01$, $p < 0.05$. The null hypothesis is rejected, indicating that at least one of the top 10 crime types has a mean reporting delay that differs significantly from the others.
4. Does weapon involvement influence the speed of crime reporting?
    a. Welch's Two-Sample t-Test | $T = -63.53$, $p < 0.05$. The null hypothesis is rejected. Crimes involving a weapon were reported significantly faster, with a mean of 6.4 days, compared to 20.0 days for crimes without a weapon.
5. How do reporting delays vary across victim age groups?
    a. Kruskal-Wallis H Test | $H = 1{,}297.02$, $p < 0.05$. The null hypothesis is rejected, indicating significant differences in reporting delays across at least one victim age group.
6. Do reporting delays differ by the sex of the victim?

a. Welch's Two-Sample t-Test | $T = -11.80$, $p < 0.05$. The null hypothesis is rejected, demonstrating a statistically significant difference in mean reporting delay between male and female victims.

7. Is there an association between high-crime areas and longer reporting delays?
    a. Spearman Rank Correlation | $\rho = -0.222$, $p = 0.333$. The null hypothesis cannot be rejected. There is no statistically significant relationship between crime volume and average reporting delay across LAPD areas.

Modeling



Code

All code requirements are available in the project's GitHub repository:

https://github.com/Forest-Flux/LAPD_crime_analysis