Methods

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.