Exploratory Data Analysis

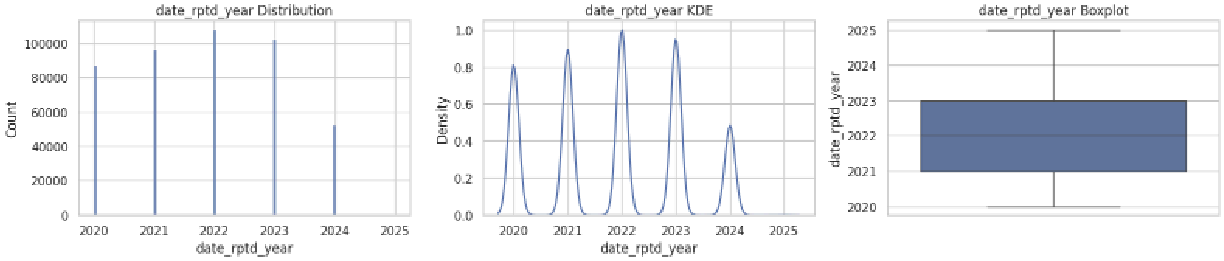
For our project, we obtained the data from the Los Angeles Police Department public records along with the U.S. Government Open Data Catalog. Our data consists entirely of structured data and has about 1,004,991 reported crime incidents dates ranging from January 2020 through June 2025. Each observation represents a single crime report and includes 34 variables related to time, location, victim demographics, and crime characteristics.

Before beginning the analysis, the dataset was cleaned to ensure accurate measurement of reporting delays and spatial patterns. We removed any records that contained missing or invalid occurrence/report dates, as the values are required to compute reporting delay. Also, additional filtering was applied to remove records with missing area identifiers for geographic analysis. After cleaning, the dataset decreased to about half its size (444007, 34) but still provided broad coverage across all LAPD areas.

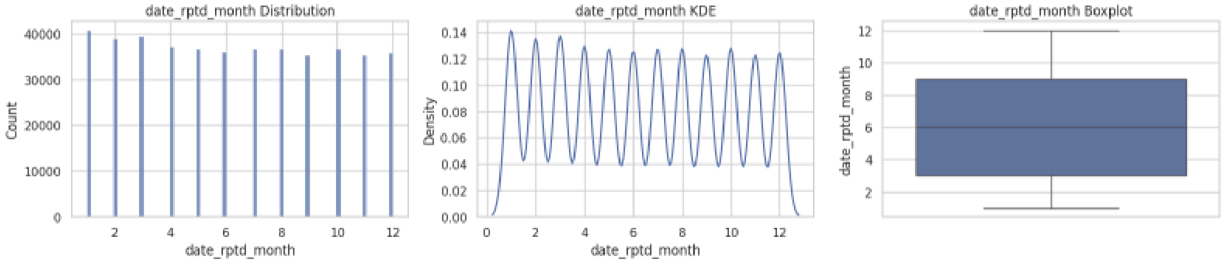
The reporting delay variable was calculated as the difference between the date and time a crime occurred and when it was reported. Multiple time based variables were created such as day, month, and year. We grouped the crimes into violent and non-violent based on the crime code descriptions, and the victim age was categorized into groups to support demographic analysis.

The initial exploration focused on the overall reporting behavior. The figure below shows the distributions of crime occurrence and report dates by year, month, and day. Crime counts show to peak around 2023, with a significant decline in 2024. When analyzing the monthly and daily patterns, they seem to be relatively stable, suggesting that crime reporting does not follow a strong seasonal trend at a broad level.

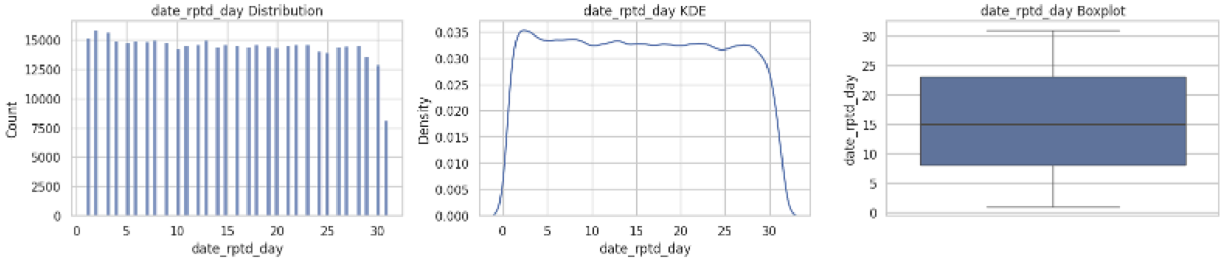
DATE_RPTD_YEAR



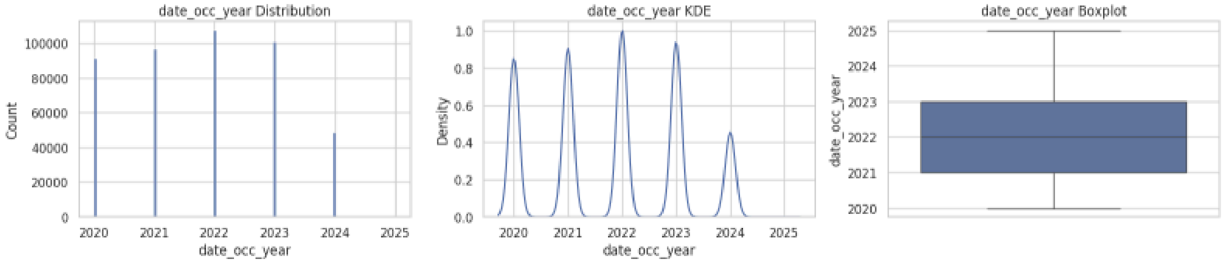
DATE_RPTD_MONTH



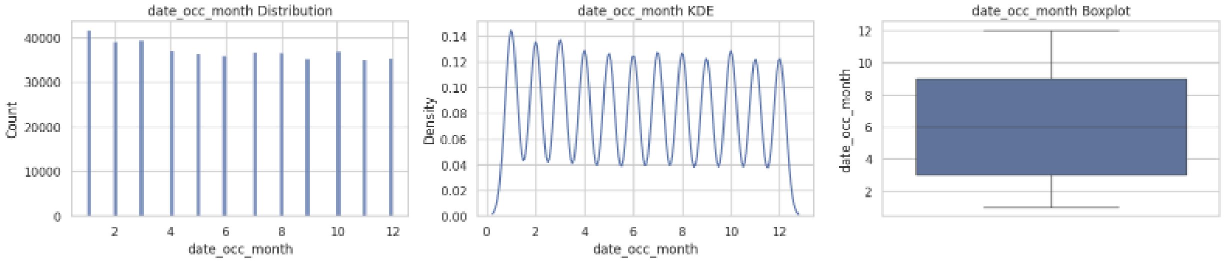
DATE_RPTD_DAY



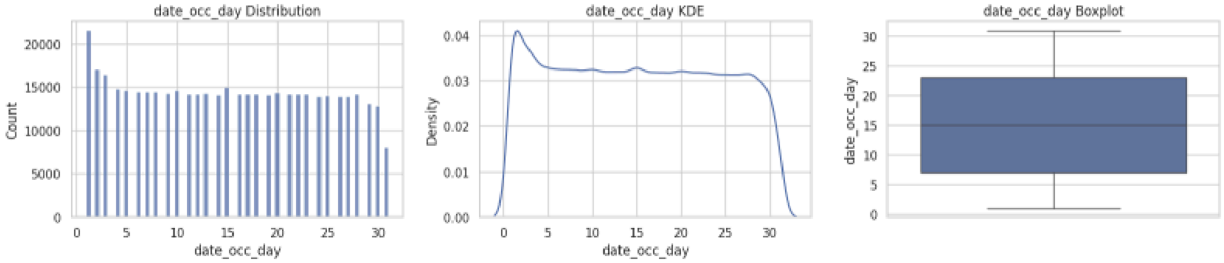
DATE_OCC_YEAR



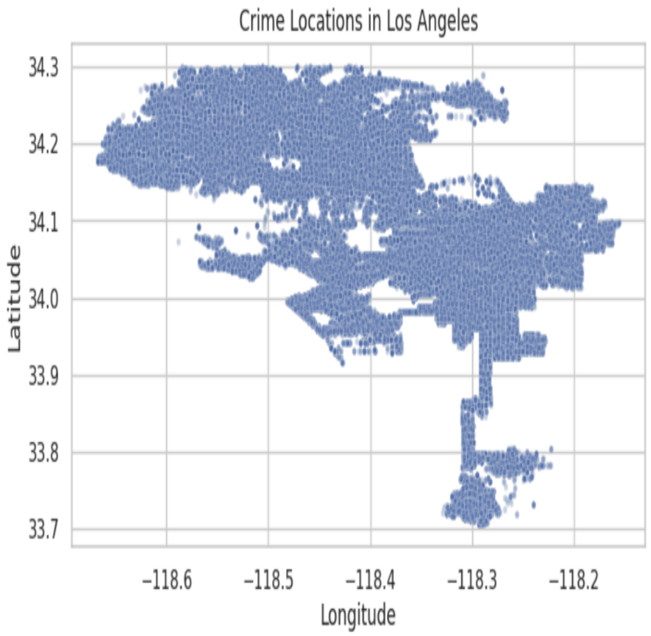
DATE_OCC_MONTH

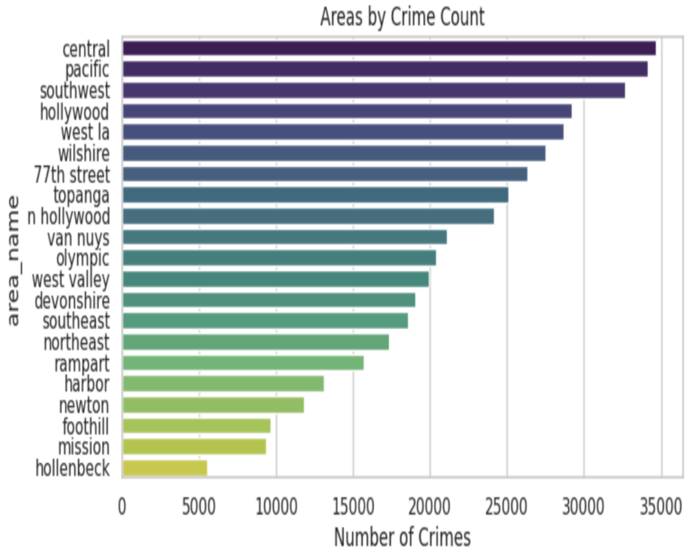


DATE_OCC_DAY

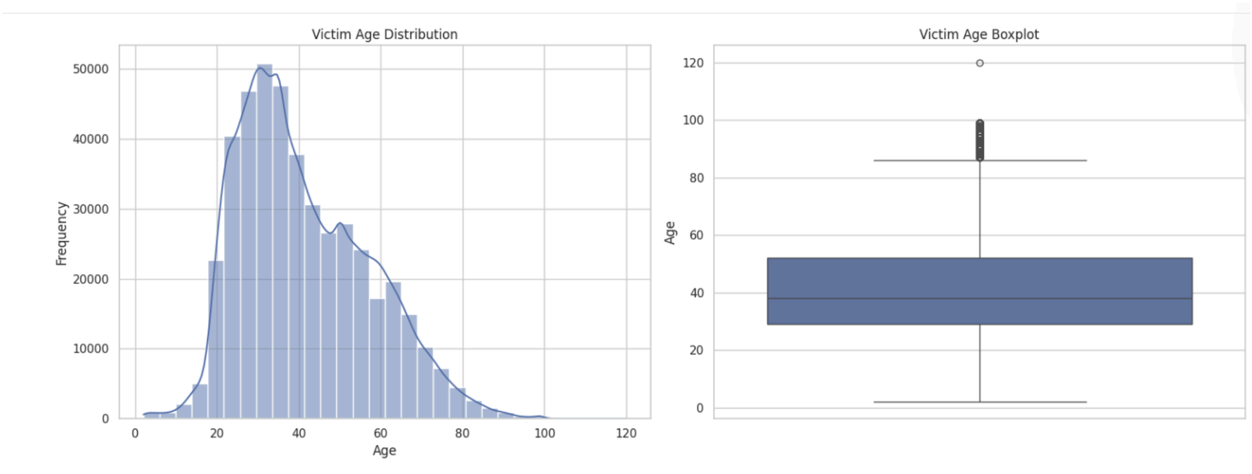


We then explored the geographic distribution of crimes. The figures contain plots of crime locations using latitude and longitude and show that crime incidents are highly concentrated within central Los Angeles rather than evenly distributed across the city. The second graph displays the pattern of total crime counts by LAPD area a little clearer. Areas like Central, Hollywood, and Pacific experience substantially higher crime volumes, whereas areas like Mission and Foothill report fewer incidents. Instead of indicating random fluctuation, these differences point to ongoing spatial vulnerability.

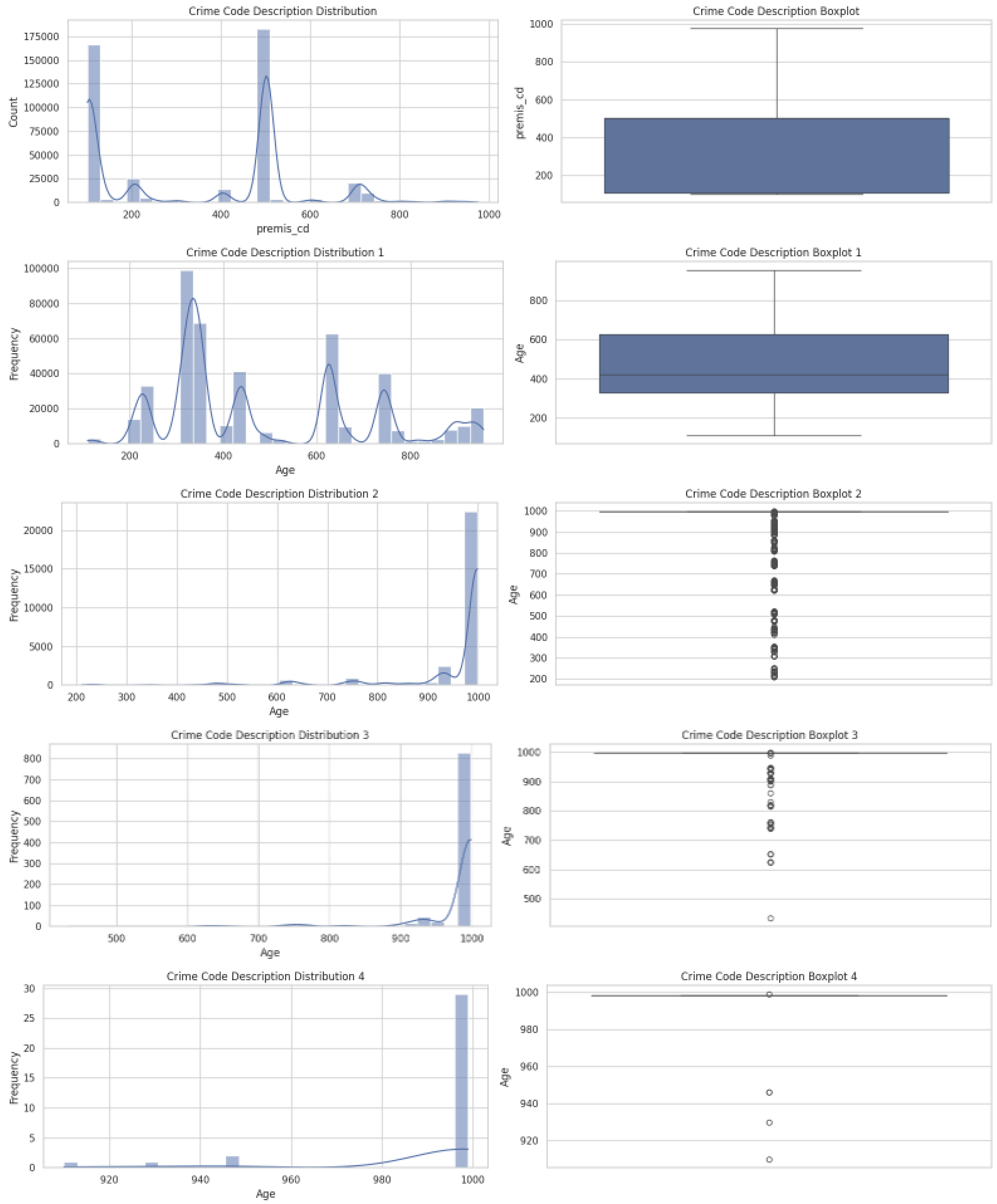




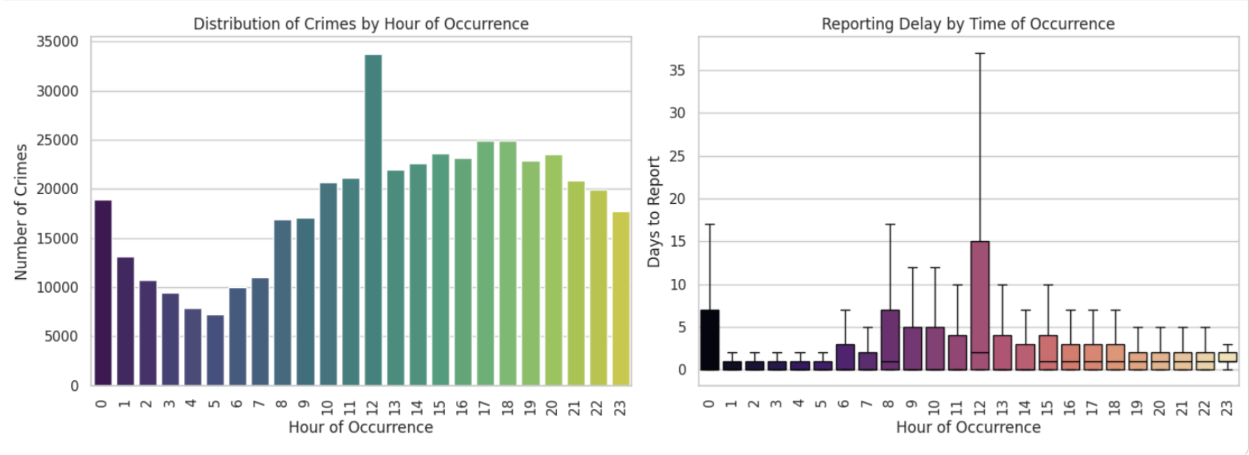
We then examined victim demographics to better understand who is most affected by crime. The histogram and boxplot displays the distribution of victim age and shows a unimodal pattern centered around young and middle aged adults. Crimes involving very young or elderly victims are less common but still present.



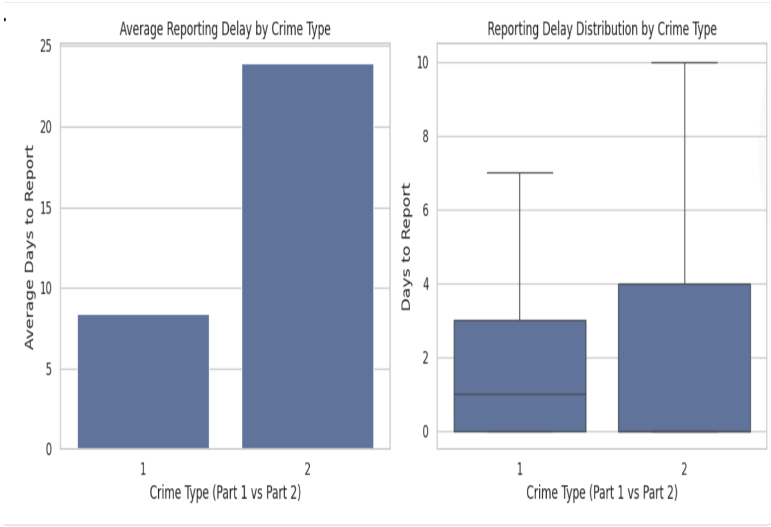
We then created a histogram that displays victim age but broken down by crime type. Clear differences emerge showing that violent crimes tend to involve younger victims, while non-violent crimes, typically fraud related offenses, disproportionately affect older adults.

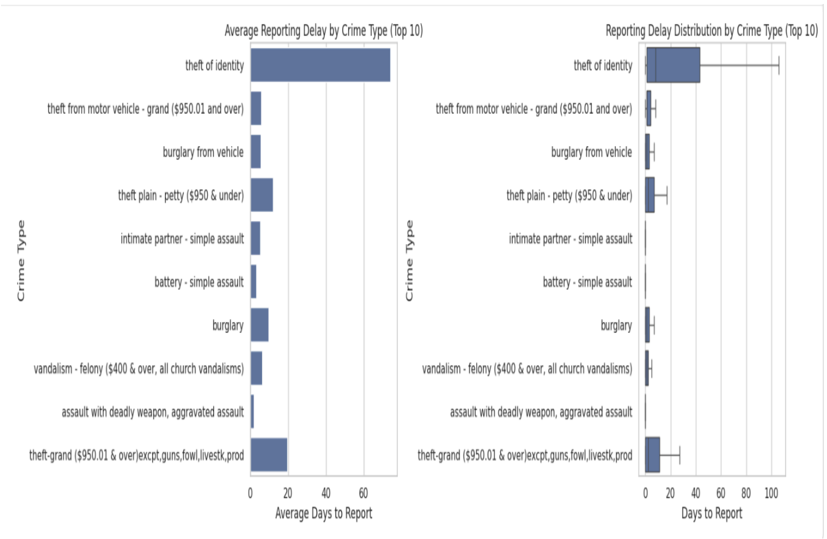


Temporal patterns were also explored. We examined the crime frequency by the hour of occurrence and discovered that incidents peak during evening and late night hours. The table also suggests that time of occurrence may influence how quickly a crime is reported, potentially due to accessibility or perceived urgency.

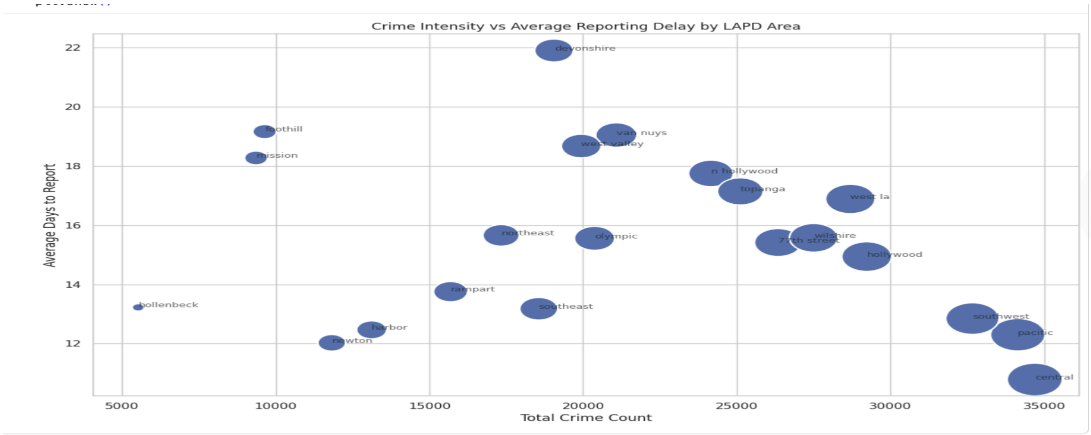


We then examined reporting delay more directly by creating a graph that shows the average reporting delays by LAPD area and reveals meaningful variation between districts. Some areas consistently experience longer reporting delays than others, which may reflect differences in access to reporting resources, community trust or socioeconomic factors. Additionally examined was the reporting delay between violent and non-violent crimes. We found that violent crimes are reported significantly faster and with less variability, while non-violent crimes exhibit longer and more dispersed delays. This pattern provides strong preliminary support for hypothesis 1.





Lastly, the relationship between crime concentration and reporting behavior was explored. The scatterplot displays the relationship between crime concentration and reporting behavior by comparing total crime volume with average reporting delay by LAPD area. While some high crime areas also show longer reporting delays, this relationship is not consistent across all areas. This suggested that crime intensity and reporting behavior are related but distinct processes that should be modeled separately.



From this information, a conclusion can be drawn about relationships between crime type, time of occurrence, geographic location, victim demographics, and reporting behavior. These findings guided the decisions used later in the project and confirm that the dataset contains meaningful structure relevant to all 3 research questions.