# ***Exploratory Data Analysis on Crime Analysis in Los* Angeles *(2020-2024)***

## Introduction

This project presents an in-depth Exploratory Data Analysis (EDA) of crime incidents reported in **Los Angeles** between **2020 and 2024**, using data sourced from the **Los Angeles Police Department (LAPD)**. The primary objective of this analysis is to uncover meaningful patterns and insights related to crime types, temporal trends, victim profiles, geographic hotspots, and law enforcement response.

Using Python and libraries such as Pandas, Matplotlib, and Seaborn, the dataset was thoroughly cleaned, preprocessed and enriched through feature engineering techniques. Key variables such as crime code description, area name, victim sex, weapon description, and status description were used to visualize and interpret the underlying patterns.

The analysis reveals that **theft-related crimes** dominate the overall crime landscape, with areas like **77th Street**, **Southwest**, and **Central LA** reporting consistently high crime volumes. **Crimes peak during the evening hours and weekends**, with noticeable seasonal fluctuations, including a dip during the early COVID-19 phase. While males constitute the majority of victims, crime impacts all demographics. Notably, a significant proportion of cases result in reports taken without arrests, especially in non-violent categories.

This EDA lays a strong foundation for future work in predictive modeling, crime forecasting, and spatial analysis. The insights derived can support **evidence-based policing strategies, resource allocation**, and **public awareness initiatives** in the city of Los Angeles.

## Purpose and Problem Statement

Understanding and mitigating crime is a central concern for urban safety and governance. In a sprawling metropolis like **Los Angeles**, crime patterns vary not only by location and time but also by type, motive, and response. With growing urban populations, changing socio-economic dynamics, and the evolving nature of crime (especially post-2020 pandemic), data-driven insights are essential for effective law enforcement and policy decisions.

This project aims to conduct a **comprehensive Exploratory Data Analysis (EDA)** on Los Angeles crime data from **2020 to 2024**, focusing on extracting patterns across time, location, victim demographics, crime categories, and police responses. Through this, the project seeks to:

* Identify **high-frequency crime types and areas** ("hotspots")
* Explore **temporal patterns** in crime (hourly, weekly, monthly, and yearly)
* Examine **victim profiles** and their association with crime categories
* **police status outcomes** (arrest made, report taken, etc.)
* Lay the foundation for **predictive modeling** or future geospatial analysis

The ultimate goal is to provide **actionable insights** for stakeholders including law enforcement agencies, policymakers, community safety groups, and data scientists interested in crime prediction and prevention.

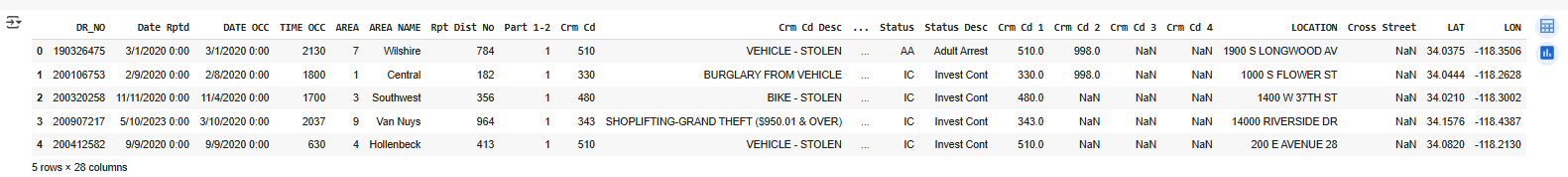
## Dataset Overview & Column Descriptions

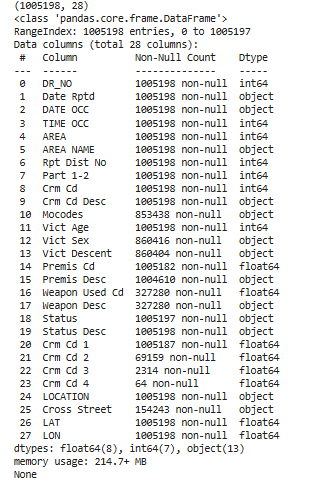
The dataset used in this project was sourced from the **Los Angeles Police Department (LAPD)** and contains detailed records of reported crimes in Los Angeles from **January 2020 to 2024**. Each record corresponds to a unique crime incident and includes a wide range of features describing the nature, location, time, and victim characteristics of the crime.

**Dataset Summary:**

* **Time span**: 2020 – 2024
* **Unit of observation**: Individual crime report
* **Initial shape**: ~960,000 rows × 28 columns (before cleaning)
* **Final shape**: ~850,000 rows × 19 columns (after cleaning and dropping columns with >30% null values)

The dataset used for this analysis is the "Crime Data from 2020 to Present," which contains records of reported crimes. The dataset has 1,005,198 rows and 28 columns.





*Figure 1: original dataset*

## Data Cleaning & Preprocessing

Raw crime datasets often contain missing values, inconsistent formats, irrelevant fields, and duplicate entries, all of which can lead to misleading analyses if not addressed. A systematic cleaning process was applied to the Los Angeles crime dataset to ensure it was consistent, complete, and analysis-ready.

**Steps Performed**

**1. Column Renaming**

To improve readability and prevent syntax issues in Python, all column names were standardized by:

* Converting to **lowercase**
* Replacing spaces with **underscores**

**2. Dropping High-Null Columns**

Columns with more than 30% missing values were dropped to reduce noise and ensure reliability.

Dropped columns included: crime\_code\_3, crime\_code\_4, cross\_street, and others not critical to the analysis.

**3. Imputing Remaining Missing Values**

For columns with minor missing values:

* **Mode imputation** was applied (filling nulls with the most frequent value in each column).  
  This preserved distributional integrity without introducing outliers.

**4. Datetime and Time Processing**

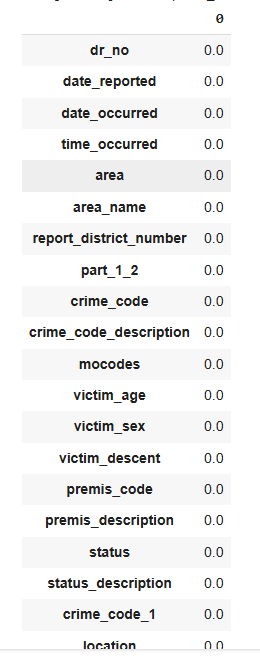
* Date occurred and date reported were converted to **datetime objects**.
* Time occurred was converted to 4-digit strings and padded using zfill(4).
* From date occurred, the following fields were extracted:
  + year, month, day, hour, day name (weekday)

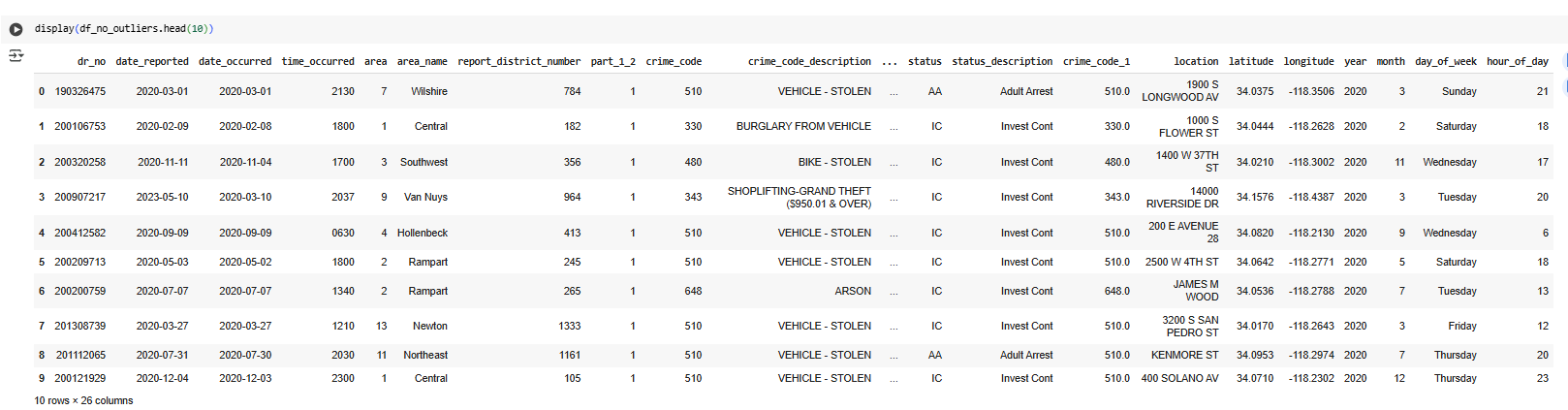
**5. Data Type Conversion**

* Columns like victim\_age and time\_occurred were converted to appropriate numeric/string types.
* Categorical features (like victim\_sex, area\_name) were treated accordingly for grouping and plotting.

**6. Duplicate Removal**

* Any fully duplicated rows were checked and removed to prevent double-counting in frequency-based visualizations.





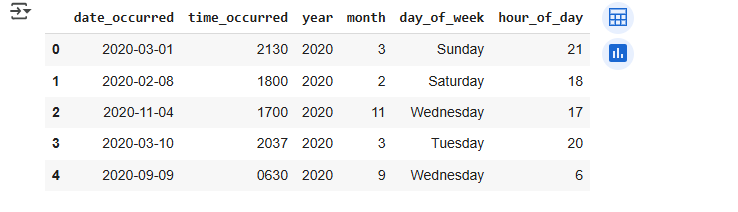
*Figure 2: Cleaned dataset*

## Section 6: Feature Engineering

To support more insightful visualizations and deeper pattern recognition, several **derived features** were engineered from the original dataset. These additions played a vital role in temporal, categorical, and spatial analyses throughout the EDA process.

**1. Temporal Features**

To examine seasonal and hourly crime patterns, the following features were extracted from date occurred and time occurred:



*Figure 3: temporal features*

**2. Spatial Feature Preparation**

* latitude and longitude were preserved and validated for **geo-scatter plots**.
* These were used to visualize **crime density and geographic clustering** within LA.

## Exploratory Data Analysis (EDA)

## Univariate Analysis

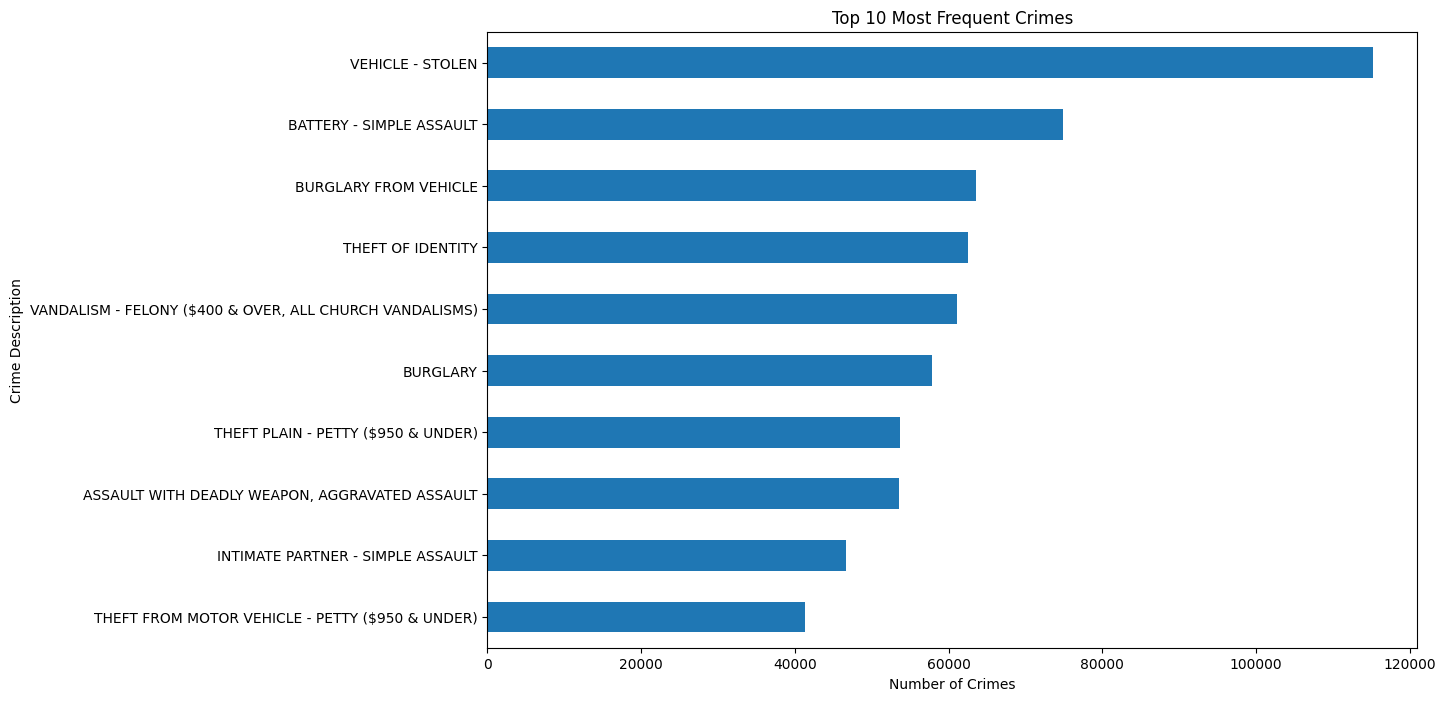
This section explores the **distribution of individual features**, focusing on identifying the most frequent crime types, high-crime areas, demographic trends, used.

**Most Common Crime Types**

A bar chart was plotted to visualize the **Top 10 most frequent crime types** reported in Los Angeles from 2020 to 2024.

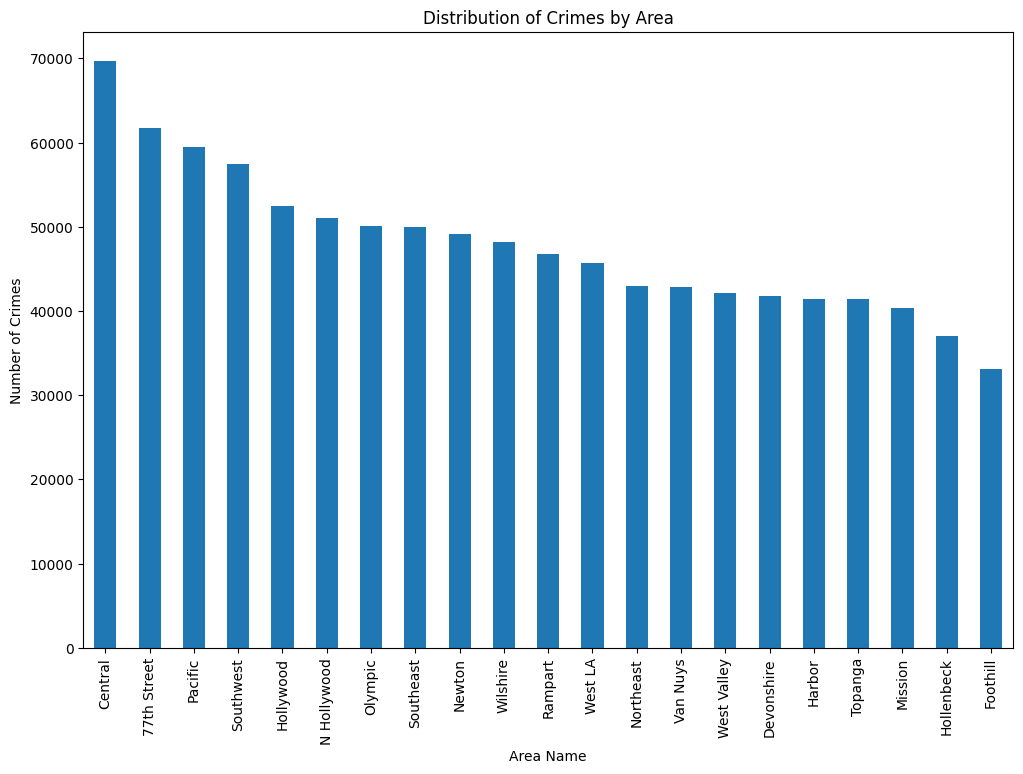
I**nsight:**

* *Theft-related crimes*, including **Theft from Vehicle**, **Battery**, and **Burglary**, were the most commonly reported.
* These top 10 crimes accounted for a substantial portion of the total crime volume.

  
 *Figure 4: Bar Plot — Top 10 Most Frequent Crime Types*

**Distribution of Crimes by Area**

To visualize the distribution of crimes across different geographical areas to identify which areas have the highest crime rates.



*Figure4: Distribution of Crimes by Area*

**Insight:**

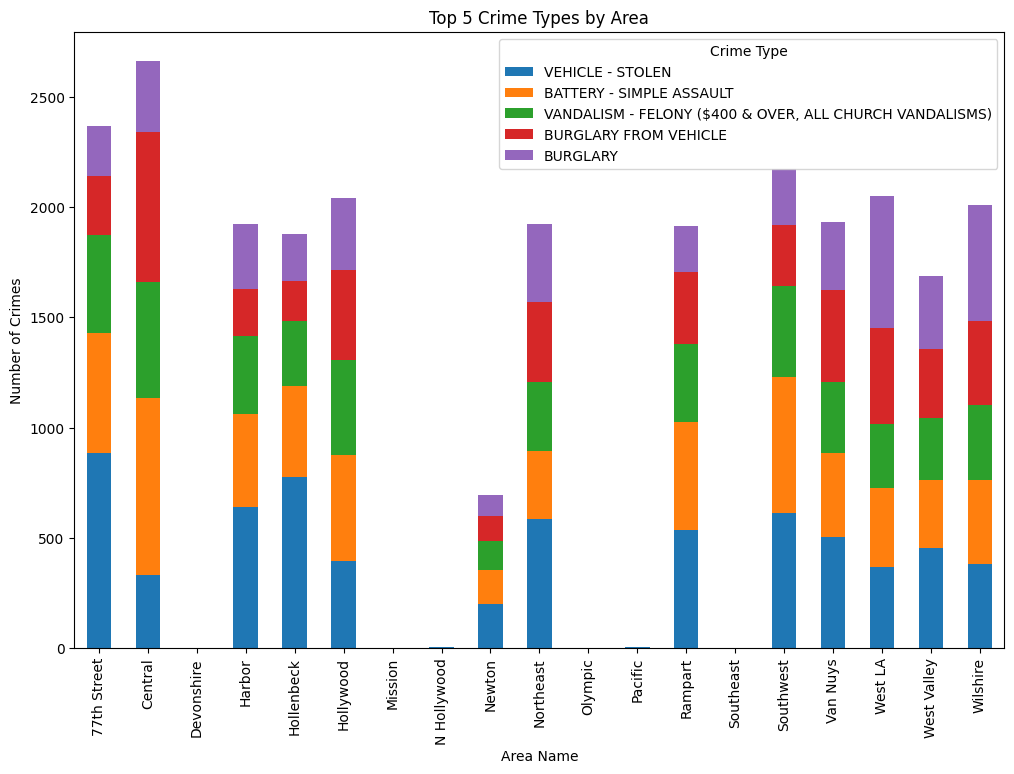
* The "Central" area has the highest number of crimes, which is expected given its high population density and commercial activity.
* "Southeast" and "77th Street" also have a high number of crimes, indicating that these are also high-crime areas.
* Areas like "West LA" and "Devonshire" have relatively fewer crimes.
* This information can be used to allocate law enforcement resources more effectively, with a greater focus on the high-crime areas. It also suggests that a more granular, area-specific approach to crime prevention is needed.

### Bivariate

This section explores how **two variables interact**, focusing on crime type distribution across geographic areas and how various combinations influence crime dynamics.

**1. Crime Type by Area (Grouped Stacked Bar Chart)**

A stacked bar chart was created to show the **top 5 crime types** across different LAPD areas. This reveals how specific types of crime are **concentrated geographically**.



*Figure 5: Top 5 Crime Types by LAPD Area (Grouped Stacked Bar Chart)*

**Insight:**

* **77th Street, Central, and Southwest** areas have the highest crime volume overall.
* **Vehicle thefts** dominate in nearly every area.
* Certain regions, such as **Rampart and Hollenbeck**, show relatively balanced crime distribution across multiple categories.
* This analysis helps identify which divisions might need **category-specific interventions**.

**2. Monthly Trends of Crime Types (Line Plot)**

A line chart was plotted to show **monthly fluctuations** in the top 5 crime types. Each line represents a unique crime category, helping reveal **seasonal or event-driven variations**.

  
 *Figure 6: Monthly Trends of Top 5 Crime Types (Line Chart)*

**Insight:**

* **Vehicle theft** remains consistently high throughout the year, although it slightly dips in later months.
* **Battery and burglary** show fluctuating behavior, with peaks around **April to June**.
* **Vandalism** incidents increase toward mid-year and remain volatile.
* This suggests **month-wise resource allocation** could help mitigate high-incidence periods.

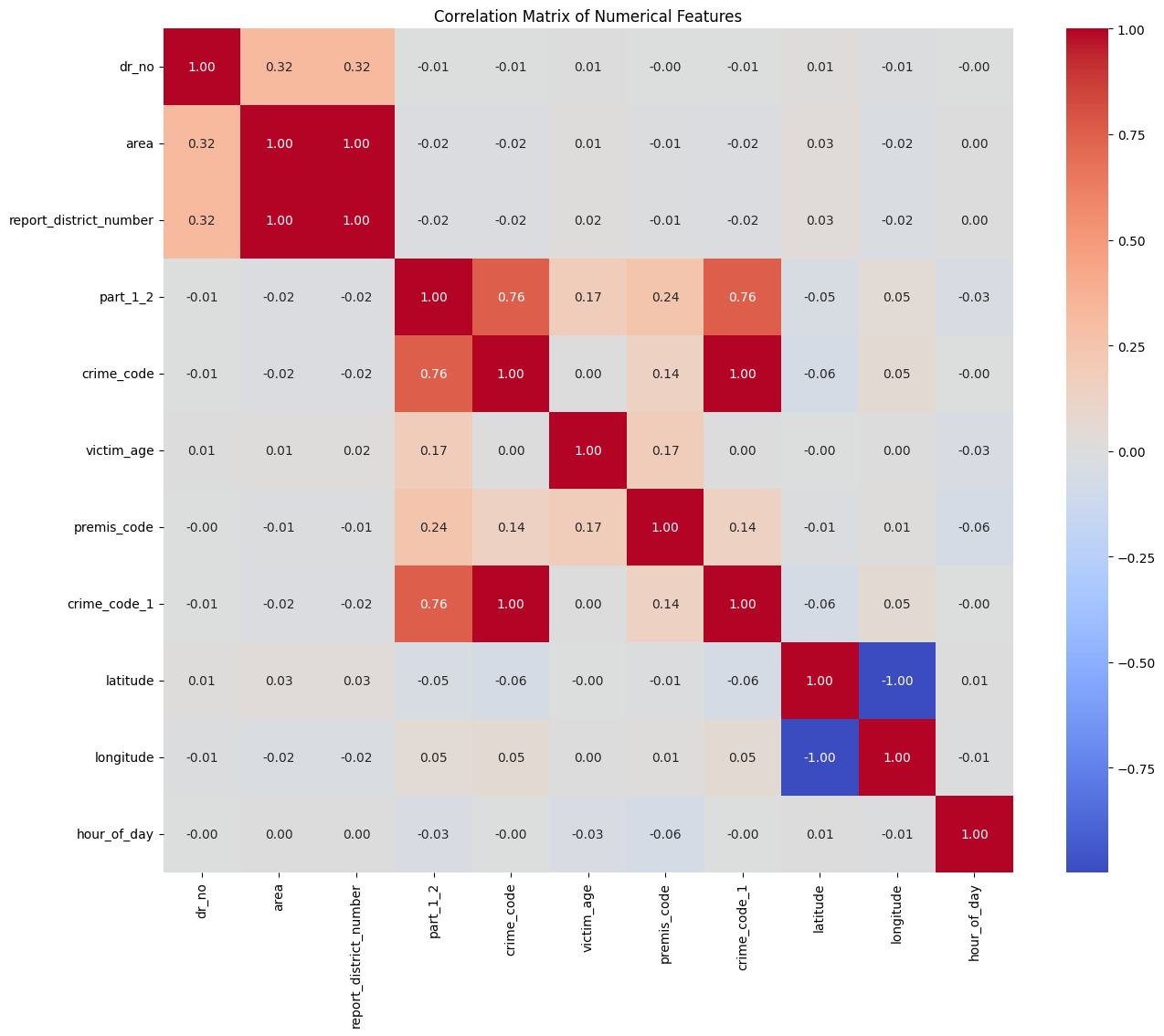
**Summary of Bivariate Insights:**

* Crime type is not evenly distributed across the city — some divisions are hotspots for specific offenses.

## Multivariate analysis

Multivariate analysis helps examine the **interactions between multiple variables simultaneously**, uncovering deeper relationships not visible through univariate or bivariate plots alone. In this project, both **correlation analysis** and **dimensionality reduction (PCA)** were performed to understand these complex interactions.

**1. Correlation Heatmap of Numerical Features**

A heatmap was generated to visualize the **pairwise correlation** among numerical columns such as victim\_age, hour\_of\_day, crime\_code, latitude, and longitude.

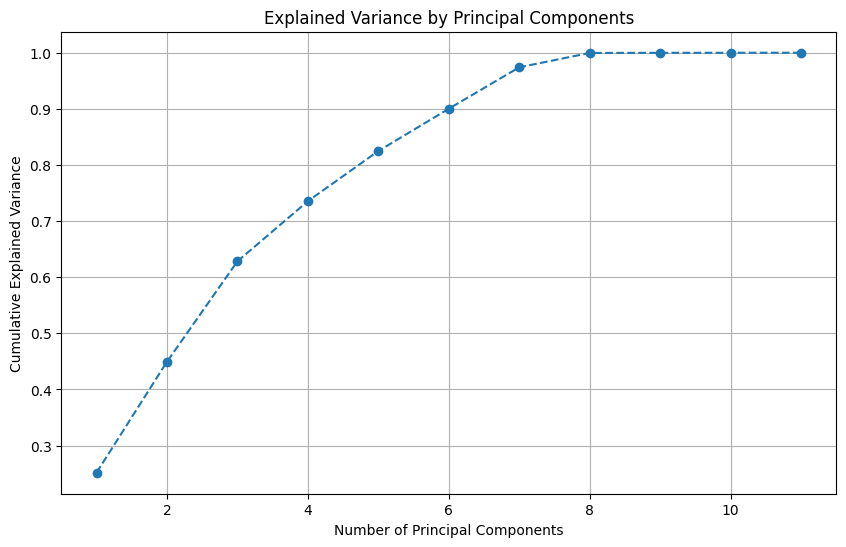
*Figure 7: Heatmap — Correlation Matrix of Numerical Features*

**Insight:**

* Most variables show **low correlation** with each other, indicating that the dataset is **rich in diverse, non-redundant features**.
* Notable weak positive correlations:
  + crime\_code with part\_1\_2 and weapon\_used\_code
  + victim\_age with premis\_code
* Weak negative correlation between latitude and longitude, which may reflect **spatial orientation** within LA.

**2. PCA — Principal Component Analysis**

Principal Component Analysis was applied to reduce the dimensionality of the numeric dataset. The **explained variance curve (scree plot)** shows how much information is retained as the number of principal components increases.

  
 *Figure 8: Line Chart — PCA Explained Variance by Components*

**Insight:**

* The **first 5–6 components** capture around **85–90% of the total variance**.
* This indicates that much of the meaningful variation in the data can be retained while reducing dimensionality — useful for future tasks like **clustering** or **ML classification**.

**Summary of Multivariate Insights:**

* Crime features are mostly independent, making the dataset suitable for **complex modeling**.
* PCA shows potential for efficient **dimensionality reduction** without major information loss.

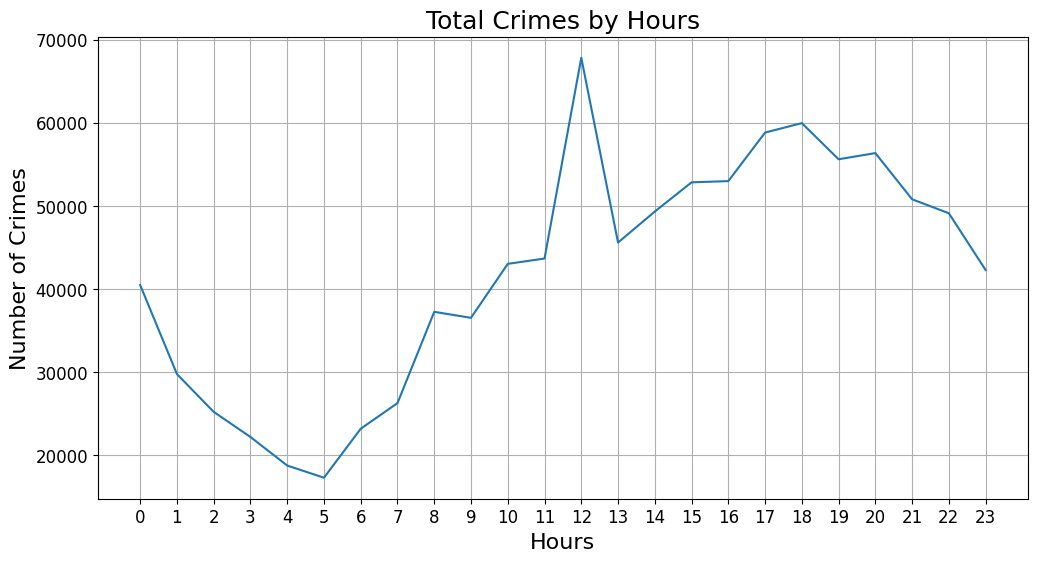
### Trends and Seasonality (Time Series)

Temporal analysis was conducted to examine how crime activity fluctuated over time in Los Angeles from 2020 onward. This involved analyzing **monthly trends**, **seasonality**, and **daily counts**, with a focus on identifying patterns and outliers.

**1. Total Crimes by Hours**

**Objective:**

To visualize the distribution of crimes throughout the day to identify peak and low crime hours.



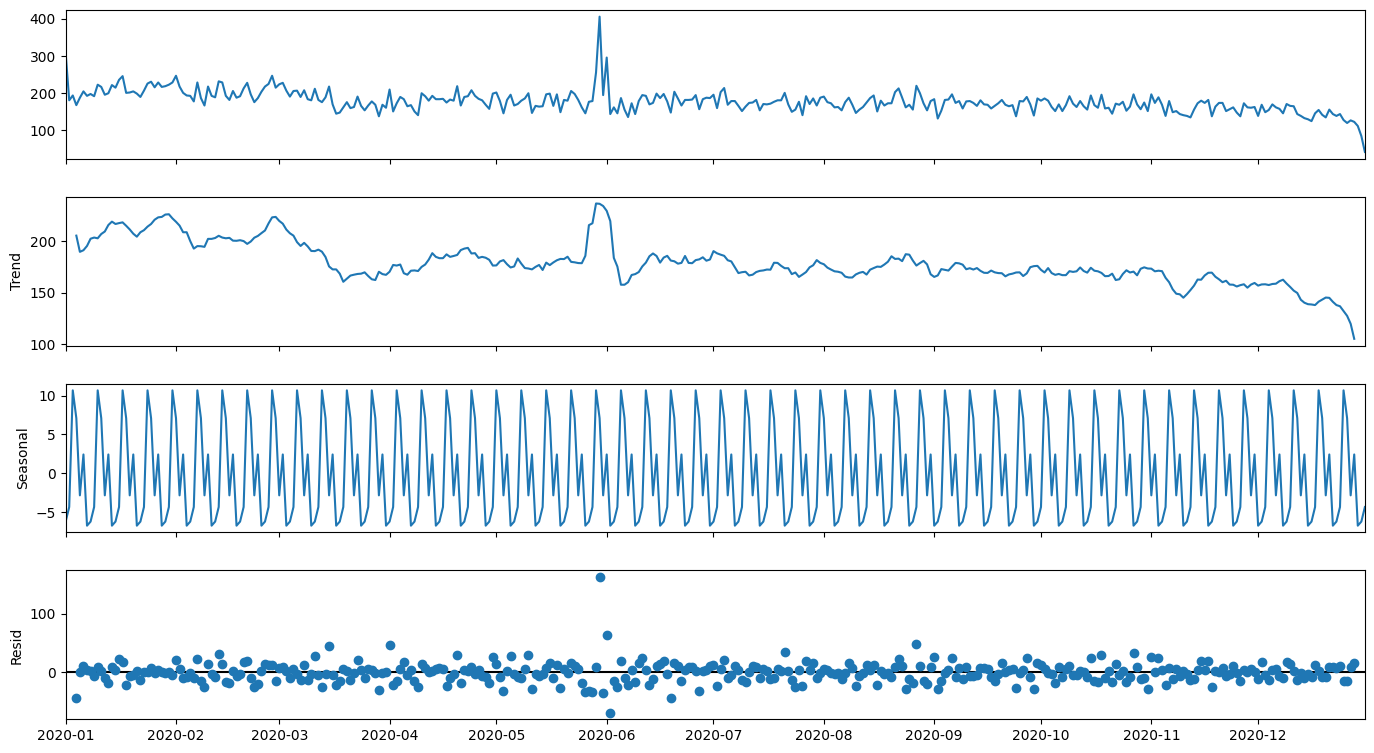
*Figure 9: Total Crimes by Hours*

**Insight:**

* The number of crimes starts to rise in the morning, with a significant increase from 8 AM onwards.
* The peak time for crimes is in the late afternoon and evening, specifically between 3 PM and 7 PM.
* There is a sharp decline in criminal activity during the late night and early morning hours, with the lowest point being around 4 AM.
* This pattern suggests that most crimes are committed when people are active and moving around, with a decrease during sleeping hours. This information can be used to optimize police patrols and resource allocation throughout the day.

**2. Time Series Decomposition**

A daily-level time series was constructed and decomposed using the **seasonal decomposition** technique to extract its components: **trend**, **seasonality**, and **residual noise**.

   
 *Figure 10: Seasonal Decomposition of Daily Crime Counts (2020)*

**Insight:**

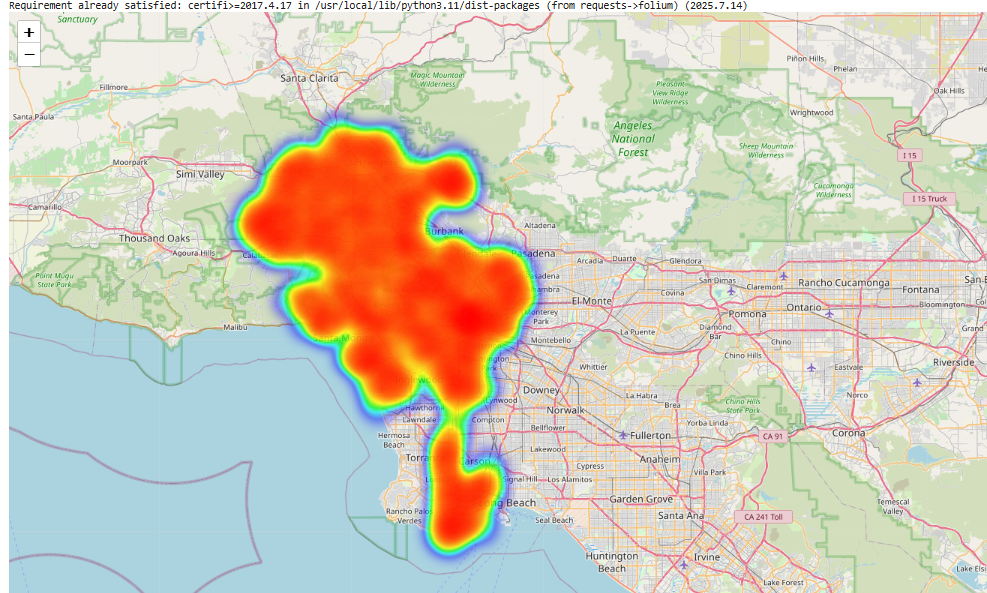
* The **trend component** reveals a noticeable dip during mid-2020, aligning with **COVID-19 lockdowns**.
* The **seasonal component** shows consistent weekly cyclicality — likely driven by **weekday/weekend effects**.
* **Residuals** remain mostly stable, with some unexpected spikes possibly tied to public events or disruptions.

### Spatial Analysis

Geo-spatial analysis provides a critical lens to understand how crime is distributed across the urban geography of Los Angeles. Leveraging the latitude and longitude data in the dataset, interactive maps were created using **Folium** to reveal patterns, clusters, and vulnerable zones.

**1. Crime Hotspot Heatmap**

A Folium-based **heatmap** was created using the folium.plugins.HeatMap class, centered around the **mean coordinates** of the crime dataset.



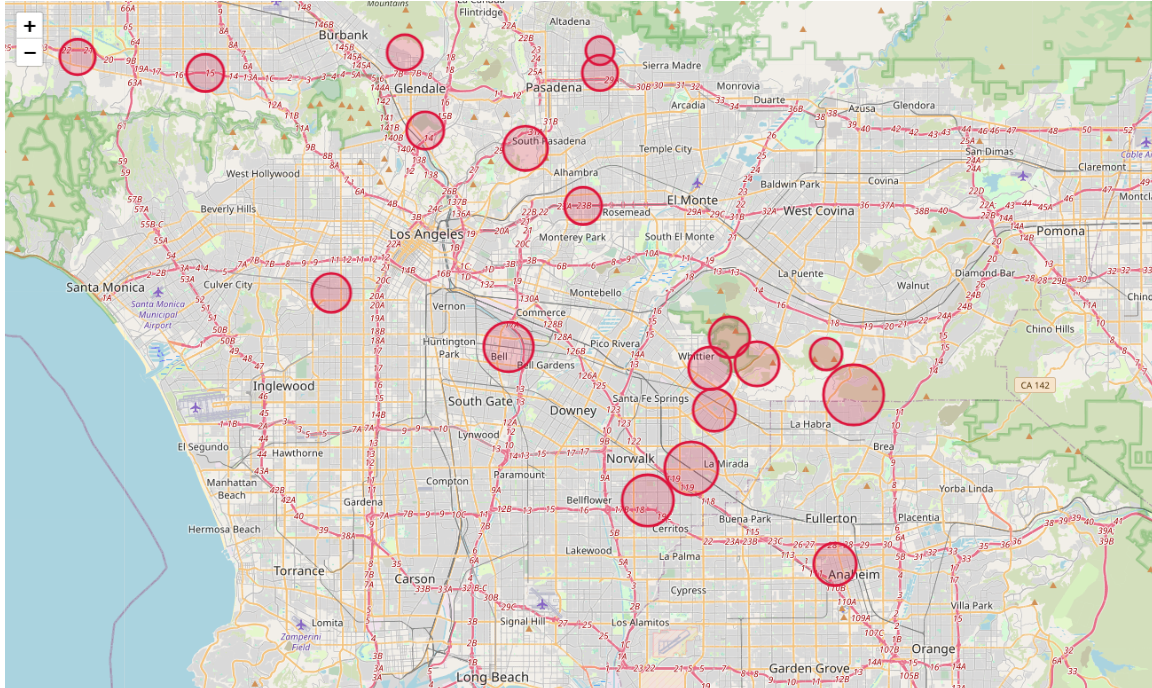
*Figure 11: Crime Density Heatmap of Los Angeles*

**Insight:**

* High crime density is observed in the **central and southern regions** of LA, particularly around:
  + **Downtown LA**
  + **77th Street**
  + **Southwest Division**
* These hotspots suggest concentrated areas of criminal activity, which may benefit from enhanced law enforcement presence.

**2. Choropleth-style Bubble Map by Area**

To simulate a **choropleth-style distribution**, the dataset was grouped by area\_name, and each area’s **mean coordinates and total crime count** were used to create **circle markers**.



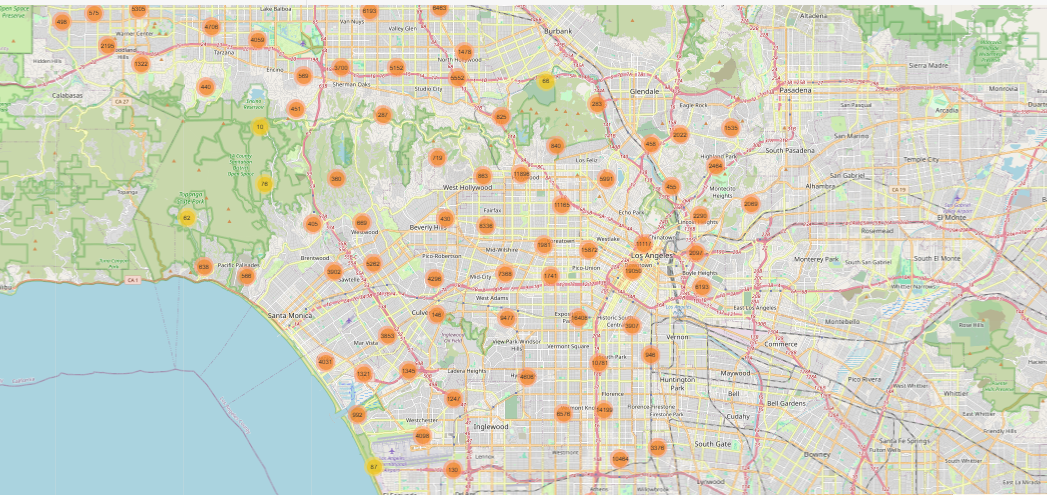
*Figure 12: Circle Marker Map — Crime Volume by Area*

**Insight:**

* The **size of each marker** reflects the total number of crimes in that LAPD division.
* The visual confirms that **West LA**, **Harbor**, and **Northeast** areas have comparatively lower crime rates.

**3. Female Victim Map**

A targeted map was generated by filtering crimes where victim\_sex == 'F' to visualize **crimes against women** across LA.



*Figure 13: Crime Map — Female Victim Incidents*

**Insight:**

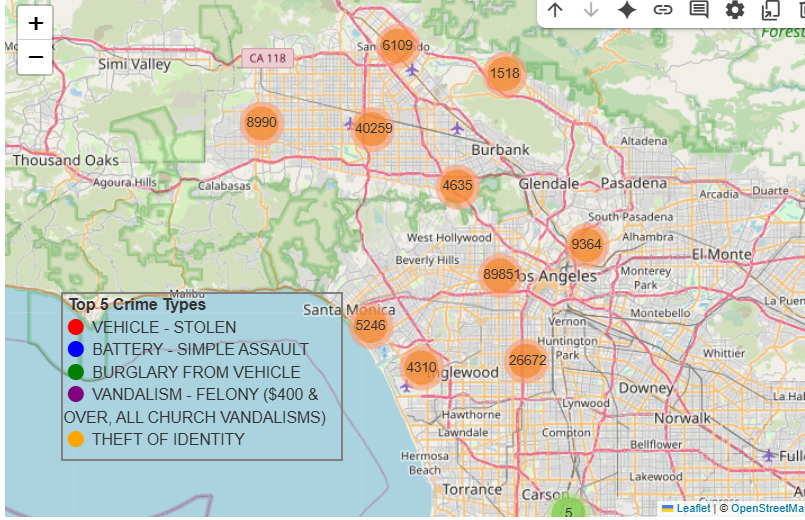
* Crime against women shows similar clustering to overall crime, but there are **additional pockets** in residential zones that are otherwise low-crime for male victims.
* This highlights the need for **gender-sensitive policing strategies**.

**Summary of Geo-Spatial Insights:**

* Crime is not evenly distributed — specific neighborhoods act as **repeat hotspots**.
* Visualization reveals not just **frequency**, but also **vulnerability zones** by gender and location.
* These maps can directly support **patrolling optimization, community alerts**, and **urban safety design**.

**4. Top 10 Most Frequent Crimes**

To identify the most common types of crimes reported in the dataset.



*Figure14: Top 10 Most Frequent Crimes*

**Insight:**

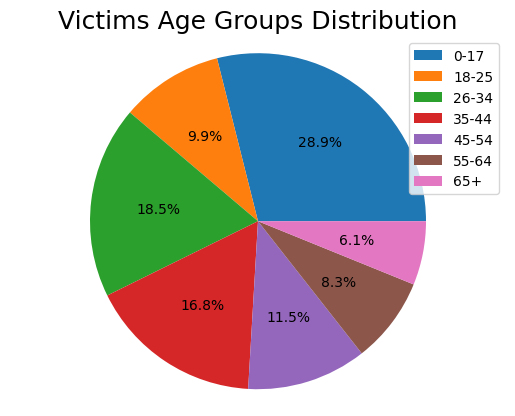
* The most frequent crime is "VEHICLE - STOLEN," indicating a significant problem with auto theft in the area.
* "BATTERY - SIMPLE ASSAULT" and "BURGLARY FROM VEHICLE" are the second and third most common crimes, respectively, highlighting the prevalence of both violent and property crimes.
* The list is dominated by property crimes, including various forms of theft and vandalism.
* This information can be used to prioritize crime prevention efforts and public awareness campaigns. For example, campaigns focused on vehicle security could be particularly effective.

## Victim Demographics Analysis

**1.Victims Age Groups Distribution**

Objective:

To visualize the distribution of victims across different age groups to understand which age groups are most affected by crime.



*Figure 15: Victims Age Groups Distribution*

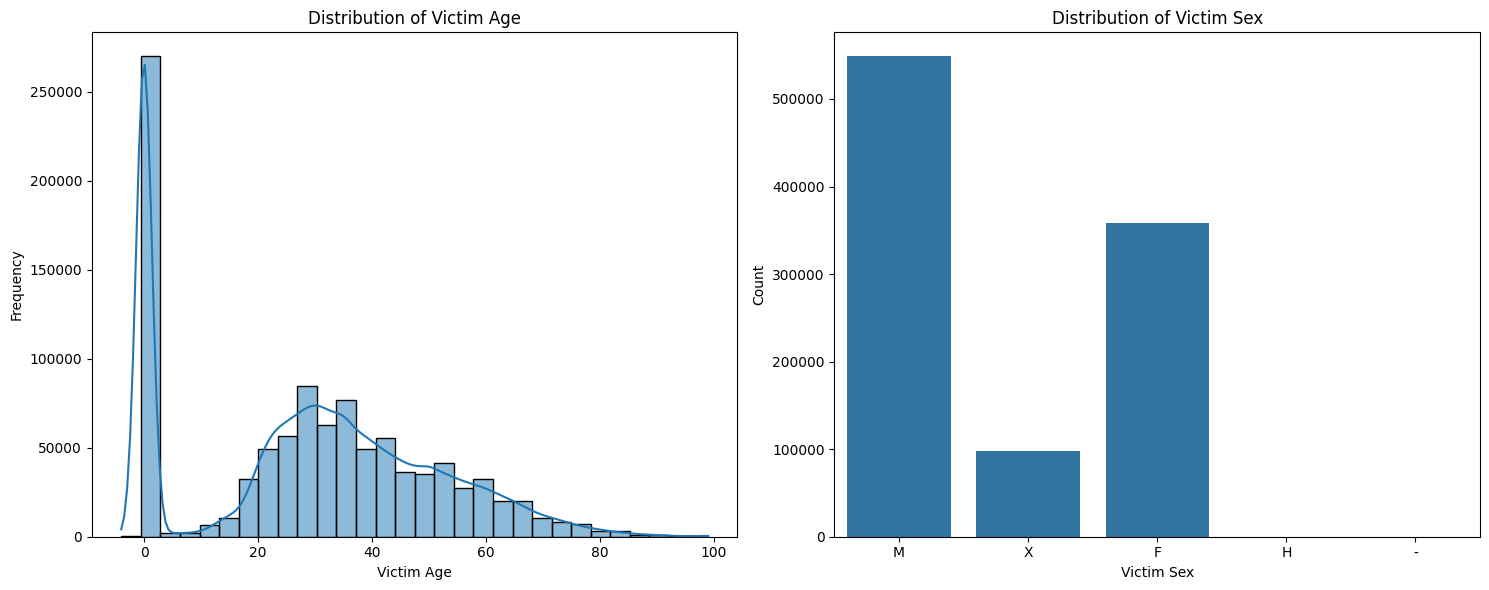
**Insight:**

* The pie chart shows that the "0-17" age group has the largest proportion of victims, which is a significant finding. This could be due to a variety of factors, including the nature of the crimes being committed (e.g., school-related incidents, juvenile delinquency).
* The "26-34" and "35-44" age groups also represent a substantial portion of the victims.
* The "65+" age group has the smallest proportion of victims.
* This information is crucial for developing age-specific crime prevention programs and victim support services. For example, programs aimed at youth safety and awareness could be particularly beneficial.

**2.Distribution of Victim Age and Sex**

**Objective:**

To visualize the distribution of victim ages and sexes to understand the demographic characteristics of crime victims.



*Figure 16: Distribution of Victim Age and Sex*

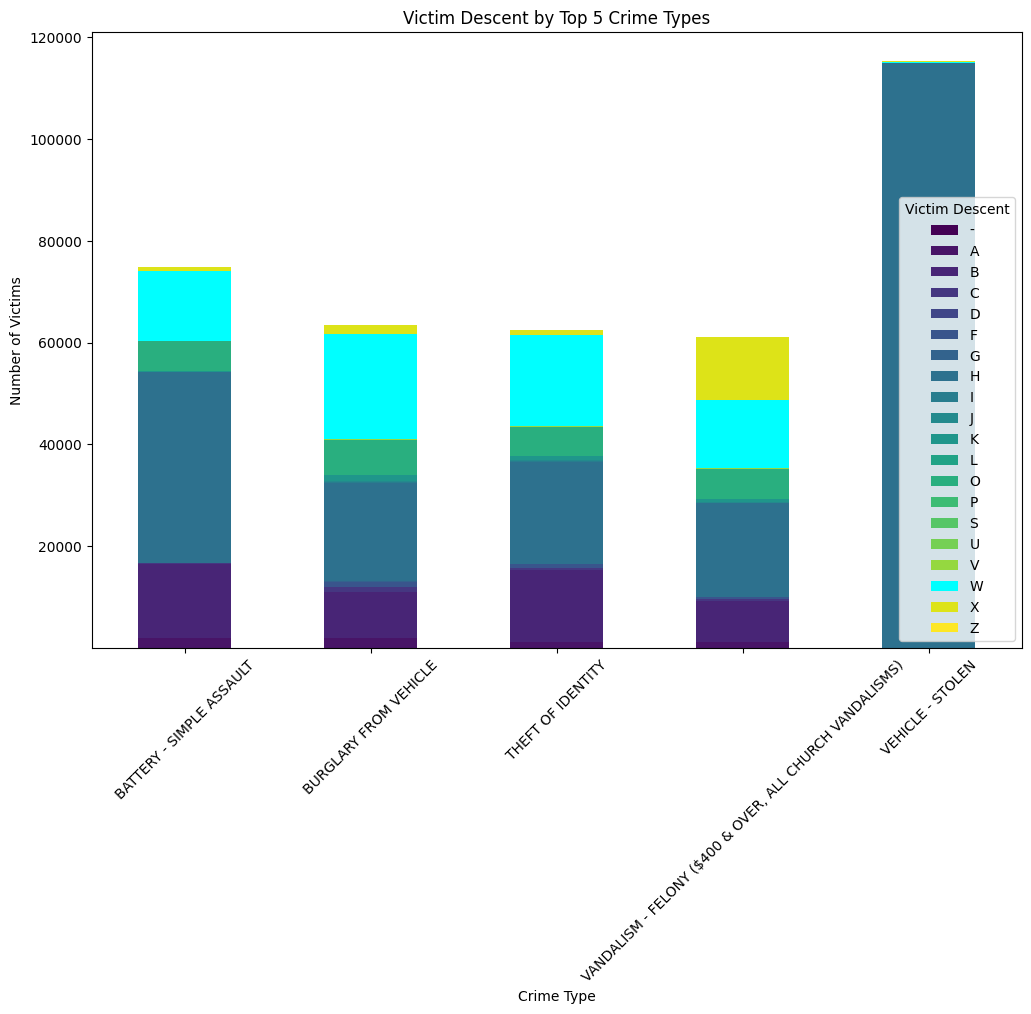
**Insight:**

* The histogram of victim age shows that younger individuals are more likely to be victims of crime, with a peak in the 25-34 age group.
* The count plot of victim sex shows that males are more frequently victims of crime than females.
* This information can be used to tailor victim support services and crime prevention programs to the specific needs of different demographic groups.

**2.Victim Descent by Top 5 Crime Types**

**Objective:**

To visualize the relationship between victim descent and the top 5 most frequent crime types.



*Figure 17: Victim Descent by Top 5 Crime Types*

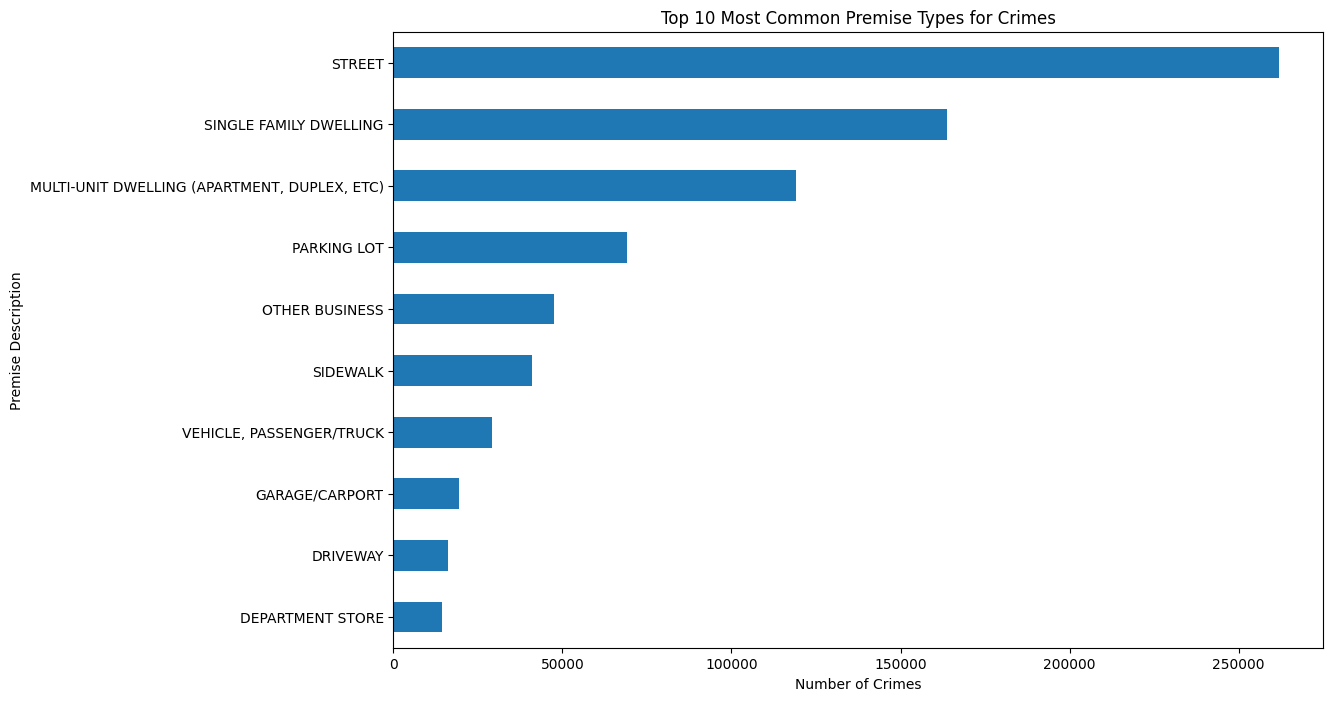
**Insight:**

* The stacked bar chart shows the distribution of victim descent for each of the top 5 crime types.
* It is evident that certain descent groups are disproportionately affected by specific types of crime. For example, in the case of "VEHICLE - STOLEN", the 'H' (Hispanic/Latin/Mexican) and 'W' (White) descents are the most frequent victims.
* This information can be used to identify and address potential biases in the criminal justice system and to develop culturally sensitive victim support services.

**Top 10 Most Common Premise Types for Crimes**

**Objective:**

To identify the most common locations where crimes occur.



*Figure 18: Top 10 Most Common Premise Types for Crimes*

**Insight:**

* The horizontal bar chart shows that "STREET" is the most common premise for crimes, followed by "SINGLE-FAMILY DWELLING" and "PARKING LOT".
* This indicates that a significant number of crimes occur in public spaces and residential areas.
* This information can be used to inform crime prevention strategies, such as increasing police presence in public areas and promoting home security measures.

## **Findings & Insights**

This section consolidates the major findings from the exploratory data analysis of the Los Angeles crime dataset (2020–2024), revealing meaningful patterns in **crime type**, **temporal distribution**, **demographics**, and **geographic concentration**.

**Crime Type Patterns**

* **Theft-related crimes** (e.g., vehicle theft, burglary) are by far the most common in LA.
* Other high-frequency crimes include **battery**, **vandalism**, and **assault**, indicating persistent personal and property security concerns.

**High-Risk Areas**

* The LAPD areas of **77th Street**, **Southwest**, and **Central** consistently report **the highest volume of crimes**.
* These divisions are also hotspots across multiple crime categories, as confirmed by heatmaps and bubble maps.

**Time-Based Trends**

* **Crimes are more frequent during evening hours** and **on weekends**, particularly Saturdays and Fridays.
* Monthly and daily time series analyses show seasonal effects, with visible dips during **COVID-19 lockdown periods (2020)**.
* Weekly seasonality is evident, suggesting the value of **predictive scheduling** for law enforcement.

**Victim Demographics**

* **Males are more commonly victims** overall, but a significant number of incidents involve female victims — particularly in residential zones.
* Most common victim ethnicities include **Hispanic**, **Black**, and **White**, reflecting the diverse urban demographic.

**Police Response & Arrest Rates**

* The majority of cases were logged as **“Report Taken”**, with relatively few resulting in **“Arrest Made”**.
* Arrest rates vary significantly by crime type, with violent crimes slightly more likely to lead to arrests.

**Spatial Clustering & Vulnerable Zones**

* Geographic plots reveal **dense clustering in inner-city areas** like Downtown LA, with some spread to peripheral zones.
* Specific heatmaps for **female victim crimes** show **additional vulnerable zones**, not flagged by general trends.

## Future Scope

* Develop **predictive models** to forecast:
  + Crime occurrence by area and time (using classification or regression)
  + Arrest likelihood based on crime type, location, and victim details
* Implement **time-series forecasting** (ARIMA, Prophet, or LSTM) for:
  + Monthly or daily crime volume trends
  + Seasonal crime patterns
* Build an **interactive dashboard** (using Power BI, Tableau, or Dash) to:
  + Visualize crime heatmaps and area-wise comparisons
  + Allow filtering by date, area, crime type, or victim attributes
  + Support real-time decision-making for law enforcement
* Expand **geo-spatial analysis** with:
  + Shapefiles or Geo JSON for proper district-level mapping
  + Advanced techniques like kernel density estimation (KDE)
* Propose **policy recommendations**:
  + Time-specific patrolling strategies based on hourly trends
  + Gender-focused interventions in female-targeted hotspots

## Conclusion

This project presented a detailed Exploratory Data Analysis (EDA) of crime data in Los Angeles from 2020 to 2024, aiming to uncover actionable insights and uncover deeper patterns in public safety and law enforcement response. Through data cleaning, feature engineering, and advanced visualization, the project transformed raw incident-level data into interpretable trends and evidence-based findings.

The analysis revealed that theft-related crimes dominate the city’s criminal landscape, with specific LAPD areas like 77th Street, Southwest, and Central reporting significantly higher incidents. Temporal patterns showed increased criminal activity during evening hours and weekends, with seasonal fluctuations aligning with real-world events like COVID-19 lockdowns.

Demographic analysis highlighted that while males form the majority of crime victims, females face distinct risks that warrant attention. The relatively low arrest rates across many crime types underscore the need for further investigation into the justice response pipeline. Geospatial analysis added another layer of depth, pinpointing crime hotspots and helping identify zones where public safety can be significantly enhanced through focused policing and urban interventions.

**Overall Observation:**

The EDA strongly suggests that crime in Los Angeles is **concentrated in specific areas and time windows**, influenced by both **social and spatial factors**. These findings are valuable for:

* Targeted policing
* Urban planning
* Community engagement
* Designing data-driven intervention programs

This report serves as a robust baseline for future modelling , predictive analytics, and policy formulation. Whether the goal is to build a real-time crime forecasting system, allocate law enforcement resources more efficiently, or design community-based interventions, the foundation laid by this EDA equips decision-makers and data scientists alike with the necessary understanding to act.

## **References**

* <https://www.kaggle.com/datasets/venkatsairo4899/los-angeles-crime-data-2020-2023>
* **License by -** ([U.S. Government Works](https://www.usa.gov/government-works/))