**Final Project**

**Development a Predictive Model for Identifying Future Crime Hotspots Using Historical and Geospatial Crime Data**

Data Understanding Report

Supervisor: Yoav Ziv

Ori Fogel 315729293

Iris Birman 208111377

31/1/2025

The project focuses on developing a predictive model aimed at identifying potential locations for future crime hotspots based on historical crime patterns. Our model will enable predictions of both the locations of future crime incidents and the expected types of crime, using geographic, temporal, and other relevant data.

# Data Collection

## Data Sources

Our data is sourced from an existing publicly available dataset found at the open data platform Data.gov. This dataset provides comprehensive crime data from 2020 to the present, that occurred in Los Angeles and is provided by the LAPD. The dataset’s reliability and scope make it an ideal foundation for our work. It is updated regularly, ensuring that the data remains current and reflective of recent developments. The inclusion of geospatial and temporal details also enables advanced analysis, such as identifying hotspots of criminal activity and evaluating time-based trends. The dataset includes various attributes that are essential for our analysis. Given its detailed nature and relevance, this source is sufficient to meet our project needs without requiring the purchase of supplemental data or the collection of additional data through surveys or tracking. However, if during the model development process we encounter significant data gaps or limitations, we will evaluate the possibility of incorporating additional data sources to enhance and complete the analysis.

**1.2 Initial Data Inspection**

The following attributes appear to be the most promising for the model:

* **LAT, LON, AREA, AREA NAME**: These geographic attributes help identify the precise locations of crimes, which are critical for predicting future crime hotspots.
* **DATE OCC, TIME OCC**: Temporal attributes enable the identification of patterns based on days, times, or seasons.
* **Crm Cd, Crm Cd Desc**: These represent the types of crimes, serving as the target variable.
* **Premis Cd, Premis Desc**: Information about the types of premises where crimes occurred, shedding light on the relationship between location and crime.
* **Weapon Used Cd, Weapon Desc**: Data on weapon use, which can provide valuable insights into understanding crime types and severity.

The following attributes may not provide significant predictive value and can potentially be excluded. These attributes either do not contribute meaningfully to the goals of the project or are redundant given the presence of other columns::

* **DR\_NO:** Appears to be a unique identifier for each report and does not offer predictive insights.
* **Cross Street**: Since this information is incomplete and contains many null values, it may be necessary to remove it.
* **Status, Status Desc:** Since the event's status (e.g., whether it is resolved or not) is not relevant to the objectives of the project, these attributes can be excluded.
* **Crm Cd 1**: This field contains information identical to that in the Crm Cd field, essentially duplicating the data, so it should be excluded.
* **Crm Cd 2, Crm Cd 3, Crm Cd 4**: These columns likely represent several additional crime codes associated with a single incident and are expected to have limited impact on the analysis compared to the primary crime code columns.

The dataset, with over 1 million records and spanning from 2020 to the present, provides ample data to make accurate predictions and draw generalized conclusions. Its comprehensive range of attributes, including geographic, temporal, and demographic information, allows for in-depth analysis of crime patterns. However, proper preprocessing is necessary to address issues like missing data and class imbalances to ensure reliable results.

The dataset contains many attributes, but not all are necessary for the chosen modeling methods, such as DBSCAN, KNN, and Decision Trees. Some attributes, like **Crm Cd 1-4** and **Cross Street** have limited predictive value and can be excluded. Redundant columns, such as **AREA** and **AREA NAME**, should also be simplified. To optimize the dataset and improve model performance, techniques such as correlation analysis, feature importance evaluation, and dimensionality reduction will be applied.

Currently, we are not merging different data sources, as the LAPD dataset provides sufficient information for our analysis. However, as the project progresses, we may incorporate additional sources, such as socioeconomic or weather data, to enhance the model’s accuracy. If merging becomes necessary, we will address challenges like data consistency and alignment to ensure accurate integration.

At this stage, we have not fully addressed how missing values will be handled but are aware it is a critical step in data preprocessing. As we move forward, we will carefully examine the dataset to determine the extent of missing data and its potential impact. Depending on the patterns observed, we will decide whether to apply imputation techniques, exclude specific attributes, or explore external data to fill gaps if necessary. This process will ensure the data is reliable and suitable for building a predictive model.

# Data Description

**2.1. Amount of Data**

The dataset contains 1,004,847 rows and 28 columns, providing a substantial amount of data for analysis. With many observations, it allows for a comprehensive exploration of crime patterns, while the diverse set of attributes enables a multi-faceted analysis of various factors influencing crime.

**2.2. Value Types**

The dataset includes a variety of attribute types, both numeric and categorical. Numeric attributes include int64 (e.g., DR\_NO, AREA, TIME OCC) and float64 (e.g., LAT, LON, Premis Cd), which represent numerical values such as identifiers, geographical coordinates, and certain crime-related codes. Categorical attributes are represented as object type, such as Date Rptd, Crm Cd Desc, Vict Sex, and Status, which contain string values that describe dates, crime descriptions, victim demographics, and event statuses. The dataset also includes both int64 and float64 types for crime codes (Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4), which represent additional crime categories.

A detailed breakdown of the attribute types, as determined by the dtypes command in Pandas, is provided in Appendix A.1 (Table 1).

A description of the attribute meanings and their respective types, as documented in the official dataset, is available in Appendix A.2 (Table 2).

**2.3. Coding Schemes**

The dataset contains several categorical attributes, each with its own coding scheme:

1. **Vict Sex:** This attribute represents the gender of the victim, using the following coding: M for Male, F for Female, X for Unknown.
2. **Status:** Represents the status of the case, using codes such as:

'AA' for Adult Arrest

'IC' for Invest Cont.

'AO' for Adult Other

'JA' for Juv. Arrest

'JO' for Juv. Other

'CC' with a status description of UNK

1. **Vict Descent:** This attribute describes the descent or ethnicity of the victim, with various codes representing different ethnic groups.

Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian

1. **Mocodes:** (Modus Operandi: Activities associated with the suspect in commission of the crime) is also a categorical attribute that contains codes from a list. The complete list of MO codes used in this dataset can be found in Appendix A.3.

# Data Exploration

In this phase, we will perform an initial analysis of the data to identify trends, patterns, and outliers that will help validate or refute the hypotheses established during the business understanding phase. We will utilize various visualization tools such as histograms, bar charts and others to gain a deeper understanding of the relationships between variables and uncover meaningful insights.

In this data exploration phase, we have formulated several hypotheses to investigate the relationships between different variables in the dataset.

These hypotheses are as follows:

**Hypothesis 1:** Certain types of crimes (indicated by Crm Cd) are more prevalent in specific areas (AREA).

Is there a relationship between the area and the type of crime?

**Hypothesis 2:** Crimes involving weapons (indicated by Weapon Used Cd) are more likely to occur at certain times (TIME OCC).

**Hypothesis 3:** There may be correlations between the age of the victim (Vict Age) and the type of crime (Crm Cd).

**Hypothesis 4:**There are gender and descent-based differences in the distribution of victims

In our initial exploration of the dataset, several attributes appear promising for further analysis. Specifically, **Crm Cd (Crime Code)** is a key attribute that will allow us to explore the distribution and prevalence of different crime types. The **AREA (Area of occurrence)** attribute is also significant, as it can help us identify crime patterns and hotspots, potentially linking certain crimes to specific geographical areas. Additionally, **Vict Age (Victim's age)** stands out as an important factor to examine in relation to crime types, as it may reveal age-related trends in victimization. The **Weapon Used Cd (Type of weapon used)** attribute offers insight into the types of weapons involved in various crimes, and examining this in combination with other factors like the time of occurrence could provide further understanding. **TIME OCC (Time of occurrence)** is a crucial variable to analyze, as it could help identify any time-based patterns or correlations in crime activity, such as whether certain crimes are more likely to occur at specific hours of the day or night.

Lastly, **Crime Location** (LAT, LON) is an important geographic attribute that can help visualize crime hotspots and facilitate spatial analysis of crime distribution.

Our explorations have indeed revealed several new characteristics about the data that were not initially apparent. One notable observation is the significant variation in crime distribution across different areas, with certain regions showing much higher crime rates than others. Additionally, we discovered that most reported crimes do not involve weapons, and most crimes occur during the afternoon hours, suggesting that time-based patterns might exist, but they are not strongly associated with crimes involving weapons. Furthermore, we observed that the Victim Age attribute displays varying trends depending on the type of crime, with younger victims being more frequently targeted in certain crime categories. This could suggest a correlation between victim age and crime type that requires deeper investigation. Finally, during our analysis, we examined the 10 most reported types of crimes and explored their distribution across different geographic areas. This allowed us to identify patterns and trends regarding how certain crime types are more prevalent in specific locations, suggesting that the location may have a significant impact on the occurrence of crimes.

During the data exploration phase, we uncovered several new insights that were not initially apparent, and these findings have influenced our original hypotheses. For instance, in Hypothesis 1, where we speculated a connection between crime types and their areas of occurrence, we observed significant variations in crime distribution across different regions. Some areas showed much higher crime rates than others, which not only supports our initial hypothesis but also calls for further investigation into the factors that contribute to these regional differences. In Hypothesis 2, which suggested that crimes involving weapons are more likely to occur at certain times, we found that the majority of crimes did not involve weapons, and most crimes occurred in the afternoon hours. This discovery led us to revise our hypothesis, as while time-based patterns were observed, they were not strongly linked to crimes involving weapons. Regarding Hypothesis 3, which explored the correlation between the victim's age and the type of crime, we observed trends indicating that younger victims are more frequently targeted in certain types of crimes. This suggests a potential correlation that requires further investigation to fully understand the relationship. Finally, in Hypothesis 4, we examined gender and descent-based differences in the distribution of victims. Our analysis revealed significant patterns in the gender and descent distribution of crime victims, which confirmed our initial hypothesis and further highlighted the importance of considering these factors in crime analysis. These explorations have led to a deeper understanding of the dataset, refining and sometimes challenging our original hypotheses, and they have opened new avenues for further analysis.

Upon reviewing our data exploration, we have found that our initial data science goals remain unchanged. The primary objective of the project continues to be the development of a predictive model that identifies potential locations for future crime hotspots, based on historical crime data. Our model will still focus on predicting both the locations and the types of crimes likely to occur, leveraging geographic, temporal, and other relevant variables. While the exploration has provided deeper insights into the data, these findings have reaffirmed our original goals without altering their direction.

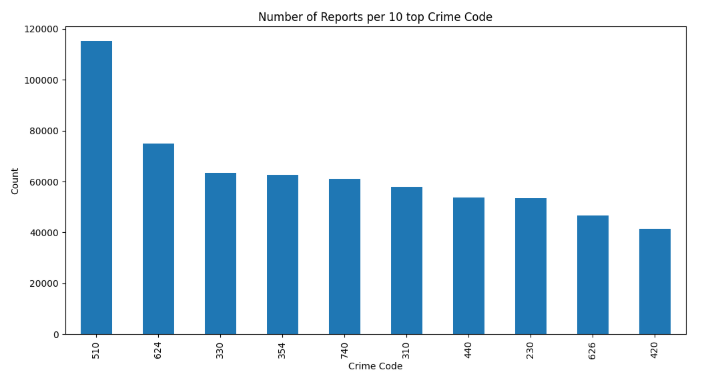
תמונה שמכילה טקסט, תרשים, צילום מסך, עיגול

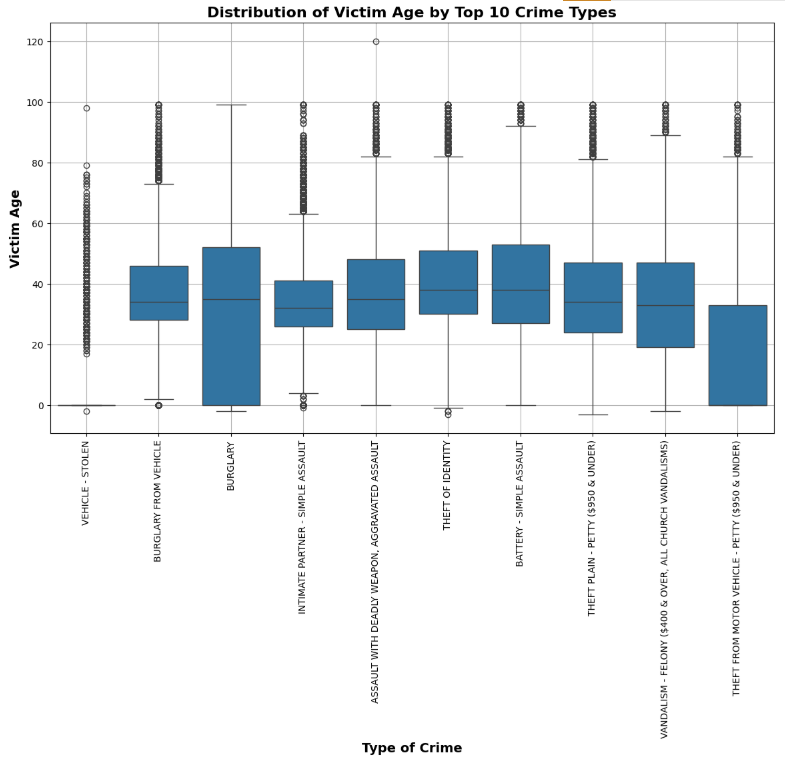
התיאור נוצר באופן אוטומטי**תמונה שמכילה טקסט, צילום מסך, קו, עלילה

התיאור נוצר באופן אוטומטי**Visual graphs that were made during the data exploration:

תמונה שמכילה תרשים, צילום מסך, עלילה, טקסט

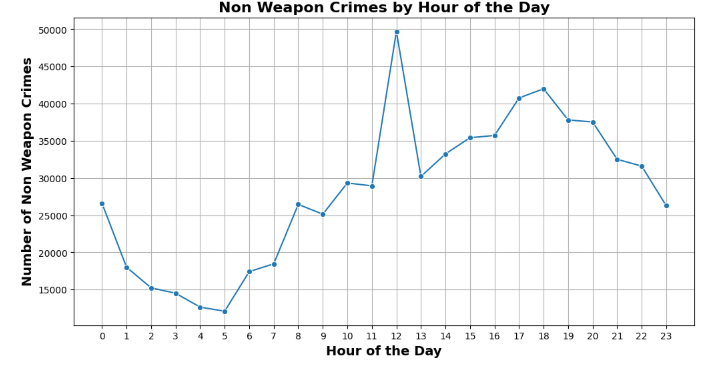
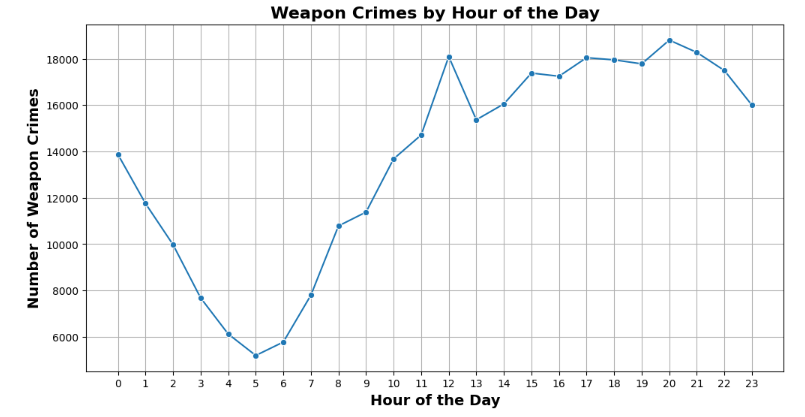
התיאור נוצר באופן אוטומטיתמונה שמכילה צילום מסך, טקסט, עלילה, צבעוני

התיאור נוצר באופן אוטומטי

****

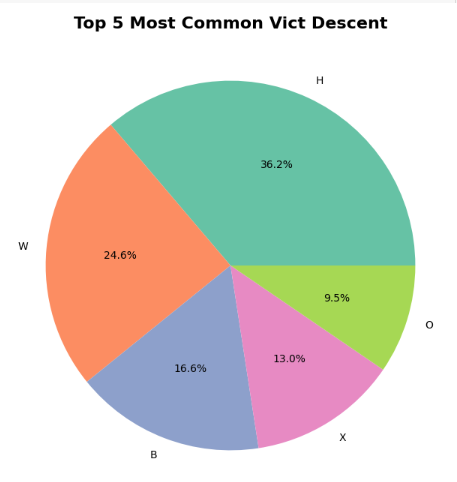
תמונה שמכילה טקסט, תרשים, עלילה, צילום מסך

התיאור נוצר באופן אוטומטי



תמונה שמכילה טקסט, צילום מסך, תרשים, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, תרשים, צילום מסך, עיגול

התיאור נוצר באופן אוטומטי

תמונה שמכילה צילום מסך, עלילה, תרשים, טקסט

התיאור נוצר באופן אוטומטי

# Data Quality:

In our data exploration, we identified several issues related to missing or incorrect values. Specifically, we found that columns such as "Mocodes" and "Vict Sex" contain missing values (null), with a total of 13 attributes having missing data. A summary table detailing all columns with missing values and their respective counts is provided in Appendix B.1 (Table 1). However, some of the missing (null) values indicate the absence of the corresponding attribute. For example, in the features related to weapon usage and its description, missing values suggest that no weapon was used.

Additionally, we discovered that the "Vict Age" column contains erroneous values, where some entries display negative numbers. There are 135 such invalid entries, which will need to be addressed in the data cleaning phase.

No anomalies were found regarding coding inconsistencies, such as nonstandard units of measurement or value inconsistencies, nor were any issues detected related to bad metadata, such as mismatches between a field's apparent meaning and its defined name or description.

**Appendices:**

**Appendix A:**

**Appendix A.3: Attribute Types Based on dtypes Output**

תמונה שמכילה טקסט, גופן, צילום מסך, מסמך

התיאור נוצר באופן אוטומטיTable 1 presents a detailed breakdown of the attribute types as determined by the output of the dtypes command in Pandas.

**Appendix A.2: Attribute Meanings and Types**

Table 2 provides an overview of attribute meanings and their respective types, as documented in the official dataset description. The full reference can be found at the following link:[**Crime Data Attribute Definitions**](https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data)**.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Col #** | **Column Name** | **Description** | **API Field Name** | **Data Type** |
| 1 | DR\_NO | Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits | dr\_no | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 2 | Date Rptd | MM/DD/YYYY | date\_rptd | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| 3 | DATE OCC | MM/DD/YYYY | date\_occ | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| 4 | TIME OCC | In 24 hour military time. | time\_occ | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 5 | AREA | The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21. | area | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 6 | AREA NAME | The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles. | area\_name | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 7 | Rpt Dist No | A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons. Find LAPD Reporting Districts on the LA City GeoHub at | rpt\_dist\_no | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 8 | Part 1-2 |  | part\_1\_2 | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| 9 | Crm Cd | Indicates the crime committed. (Same as Crime Code 1) | crm\_cd | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 10 | Crm Cd Desc | Defines the Crime Code provided. | crm\_cd\_desc | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 11 | Mocodes | Modus Operandi: Activities associated with the suspect in commission of the crime. | mocodes | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 12 | Vict Age | Two character numeric | vict\_age | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 13 | Vict Sex | F - Female M - Male X - Unknown | vict\_sex | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 14 | Vict Descent | Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian | vict\_descent | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 15 | Premis Cd | The type of structure, vehicle, or location where the crime took place. | premis\_cd | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| 16 | Premis Desc | Defines the Premise Code provided. | premis\_desc | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 17 | Weapon Used Cd | The type of weapon used in the crime. | weapon\_used\_cd | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 18 | Weapon Desc | Defines the Weapon Used Code provided. | weapon\_desc | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 19 | Status | Status of the case. (IC is the default) | status | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 20 | Status Desc | Defines the Status Code provided. | status\_desc | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 21 | Crm Cd 1 | Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious. | crm\_cd\_1 | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 22 | Crm Cd 2 | May contain a code for an additional crime, less serious than Crime Code 1. | crm\_cd\_2 | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 23 | Crm Cd 3 | May contain a code for an additional crime, less serious than Crime Code 1. | crm\_cd\_3 | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 24 | Crm Cd 4 | May contain a code for an additional crime, less serious than Crime Code 1. | crm\_cd\_4 | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 25 | LOCATION | Street address of crime incident rounded to the nearest hundred block to maintain anonymity. | location | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 26 | Cross Street | Cross Street of rounded Address | cross\_street | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| 27 | LAT | Latitude | lat | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| 28 | LON | Longitude | lon | [Number](https://dev.socrata.com/docs/datatypes/number.html) |

**Appendix A.3: Link to MO Codes List**  
The full list of MO codes can be accessed at the following link:  
[MO Codes Numerical List (Los Angeles Data)](https://data.lacity.org/api/views/2nrs-mtv8/files/4591b6bf-5846-4de0-9fb0-8780a77a036c?download=true&filename=MO_CODES_Numerical_20191119.pdf)

**Appendix B:**

**Appendix B.1: Summary of Columns with Missing Values**

Table 1 presents a summary of all dataset columns containing missing values, along with their respective counts.

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי