**Final Project**

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

Data Preparation Report

Supervisor: Yoav Ziv

Ori Fogel 315729293

Iris Birman 208111377

29/3/2025

# Selecting Data:

Based on our initial data understanding and analysis, we carefully selected the most relevant features for our data science goals while excluding redundant or less useful ones.

The following attributes were removed from our dataset:

* DR\_NO – A unique identifier that does not contribute to predictive analysis.
* DATE\_RPTD – Represents the date the crime was reported, which is less relevant since we already have the actual occurrence time.
* AREA NAME, CRM CD 1– Duplicate columns providing less relevant information compared to the AREA NUMBER and CRM CD attributes.
* CRM Cd Desc – This attribute provides a verbal description for CRM CD.
* RPT DIST NO, PART 1-2, STATUS, STATUS DESC, Premis Cd and Premis Desc – Attributes that are not significant for our analysis.
* CRM CD 2-4, Cross Street, Mocodes – These columns have limited contribution to our model and high percentage of missing values.
* LOCATION – column that provides a verbal description of the crime location. Given that the models we will be working with cannot handle string-based features, we will use the 'LON' and 'LAT' columns, which represent the location in a precise numeric format.

Some of the removed columns were supplementary textual descriptions of existing attributes, others contained excessive missing values, and some were deemed irrelevant for our predictive model. Our decisions ensure that the dataset remains focused, reducing redundancy while retaining critical information for accurate analysis.

# Cleaning Data

As part of the data cleaning process, we took several steps to address issues identified in the selected data for analysis:

In the Vict Age column, we decided to replace outliers (such as values of 0 and negative values) with the mean value of the column. This was done to prevent invalid values from affecting the model and to maintain consistency in the data.

In the Vict Sex and Vict Descent columns, we decided to fill the missing values with "X," which represents "unknown" This ensures that we handle missing data in a meaningful way, preserving the integrity of the dataset for analysis.

In the Weapon Use column, we decided to treat the missing values as an indication that no weapon was used in the incident, based on the assumption that missing data in this context implies the absence of weapon involvement.

# Constructing New Data

To improve our model, we constructed new data by deriving additional attributes. We added a Day column, which represents the day of the week on which the crime occurred, and an Hour column, represents the time of day the crime took place. These columns were introduced to convert the date and time into numerical formats that can be more effectively utilized by the model. Additionally, we added a Category column, which categorizes the types of crimes into five categories: 'Property Offenses', 'Sexual Assault', 'Assault', 'Theft', and 'Other'. This was done to reduce the variety of crime types, thereby enhancing the model's accuracy and ensuring more precise predictions.

Furthermore, we introduced a Weapon Used binary column, based on the Weapon Used Cd column, to indicate whether a weapon was used in the crime. We also created a Crime Severity column, derived from the CRM CD 2-4 columns, to assess the severity of the crime. These new attributes were designed to provide more meaningful input for the model, helping to improve its predictive capabilities.

# Integrating Data

In our analysis, no additional data integration was performed as part of the project. We did not merge or append any external datasets to the existing data, as the current dataset provided sufficient information for addressing our data science goals. Therefore, no merging or appending actions were required, and the focus remained on cleaning, transforming, and deriving new attributes from the available data.

# Formatting Data

For our model building, we plan to use the Random Forest and XGBoost algorithms. Both models do not require the data to be in a specific order but do require that the data be formatted properly. Specifically, we ensure that all categorical variables are either one-hot encoded or label encoded, and all numerical variables are in a numeric format suitable for the models. Additionally, we have taken care to handle any missing data appropriately before fitting the models. We have also prioritized columns in numeric format over those in string format that provide the same information. No further data manipulations or specific ordering were necessary beyond these standard preprocessing steps.

# Exploratory Data Analysis (EDA)

This section presents the analysis of the distribution of the target variables, including geographical locations (longitude and latitude) and crime categories. Statistical summaries and visualizations will be used to explore their relationships.

1. תמונה שמכילה טקסט, צילום מסך, גופן, כחול חשמלי

   תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.Target variable- category types:
2. תמונה שמכילה טקסט, צילום מסך, גופן, מספר

   תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.Target variable- LON:

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

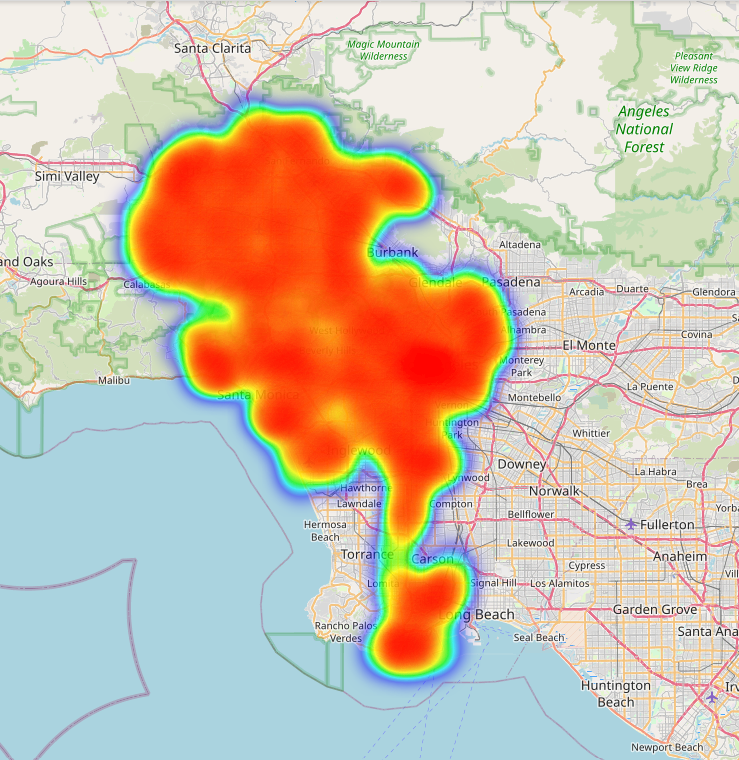
1. תמונה שמכילה טקסט, צילום מסך, גופן, מספר

   תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.Target variable- LAT:

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

Heatmap for the geographical locations (longitude and latitude):



Correlations between variables:

תמונה שמכילה טקסט, צילום מסך, תרשים, ריבוע

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

1. Important variables:

תמונה שמכילה תרשים, עלילה, צילום מסך, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה טקסט, צילום מסך, גופן, מספר

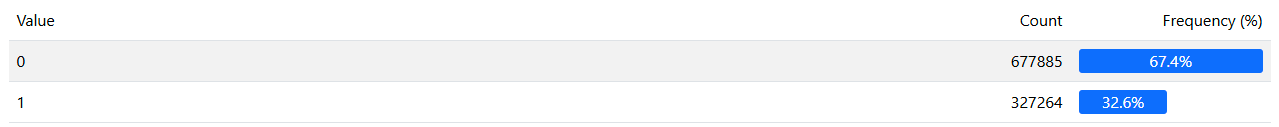
תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.Vict age:

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.Hour:

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

Weapon use:

1. תמונה שמכילה תרשים, צילום מסך, מלבן, טקסט

   תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.Relationships between variables:

* Chi-square test between category and weapon use: chi2=676271.93, p-value=0.0000. The p-value is less than 0.05, the relationship is statistically significant.