IT5006 Project: Predictive Policing - Milestone 2

Group 5

Team Members:

Li LingYu	A0304263N
Liu Yichao	A0304386A
Wu Jingyan	A0305022B
Yang Qingshan	A0162526H
Ye Fangda	A0310135Y

March 27, 2025

1 Exploratory Data Analysis

1.1 Introduction

This section presents an exploratory analysis of the preprocessed Chicago crime dataset using visualizations created in Tableau. The goal is to identify key patterns related to the temporal distribution of crimes, their spatial concentration, and correlations between crime attributes. The analysis primarily draws upon composite visualizations summarizing temporal and spatial patterns, supplemented by insights into correlations.

1.2 Temporal Patterns in Crime Incidents

Understanding when crimes occur is crucial. Figure 1 provides a composite overview of temporal patterns, aggregating crime cases by month, weekday, and hour, also indicating the proportion of cases resulting in an arrest (the orange portion representing the proportion of cases resulting in an arrest, and the blue portion representing those without an arrest).

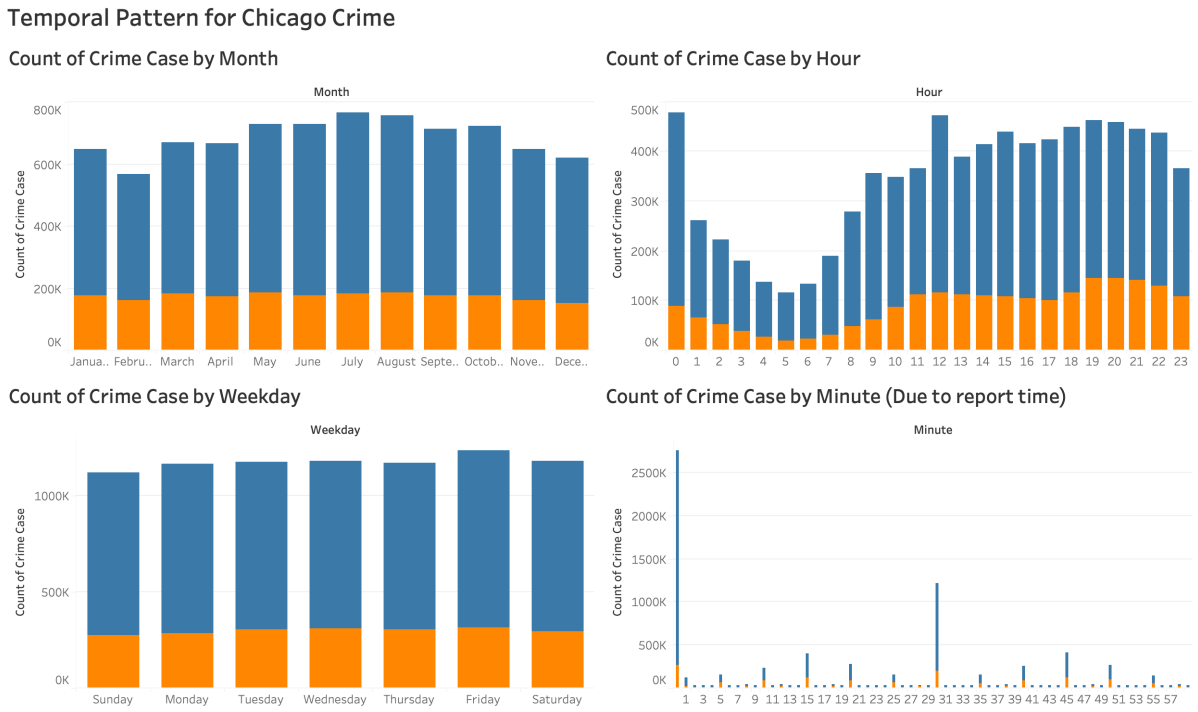

Figure 1: Composite View of Temporal Crime Patterns.

Analyzing Figure 1:

- The top-left chart displays monthly counts, revealing distinct seasonality. Crime incidents peak during warmer summer months (May-August) and decrease during colder months (lowest in February), suggesting an influence of weather or outdoor activity.
- The top-right chart illustrates hourly crime counts, demonstrating a clear diurnal cycle. Crime reports are minimal in the early morning (approximately 4-6 AM), increase throughout the day, with pronounced peaks at 12 PM (noon) and 24 PM (midnight).
- The bottom-left chart illustrates weekly counts. Crime levels appear relatively consistent across weekdays, potentially with a slight increase on Fridays.
- The arrest overlay, visible across these charts, suggests that arrests might constitute a slightly higher proportion during peak crime hours, although the overall number of reported cases vastly outweighs arrests for most periods.
- The bottom-right chart shows counts by minute, likely reflecting reporting artifacts (e.g., reports logged on the hour) rather than actionable crime patterns.

1.3 Spatial Distribution of Crime

Figure 2 illustrates where crimes are concentrated, examining both specific location types and broader geographical areas within Chicago.

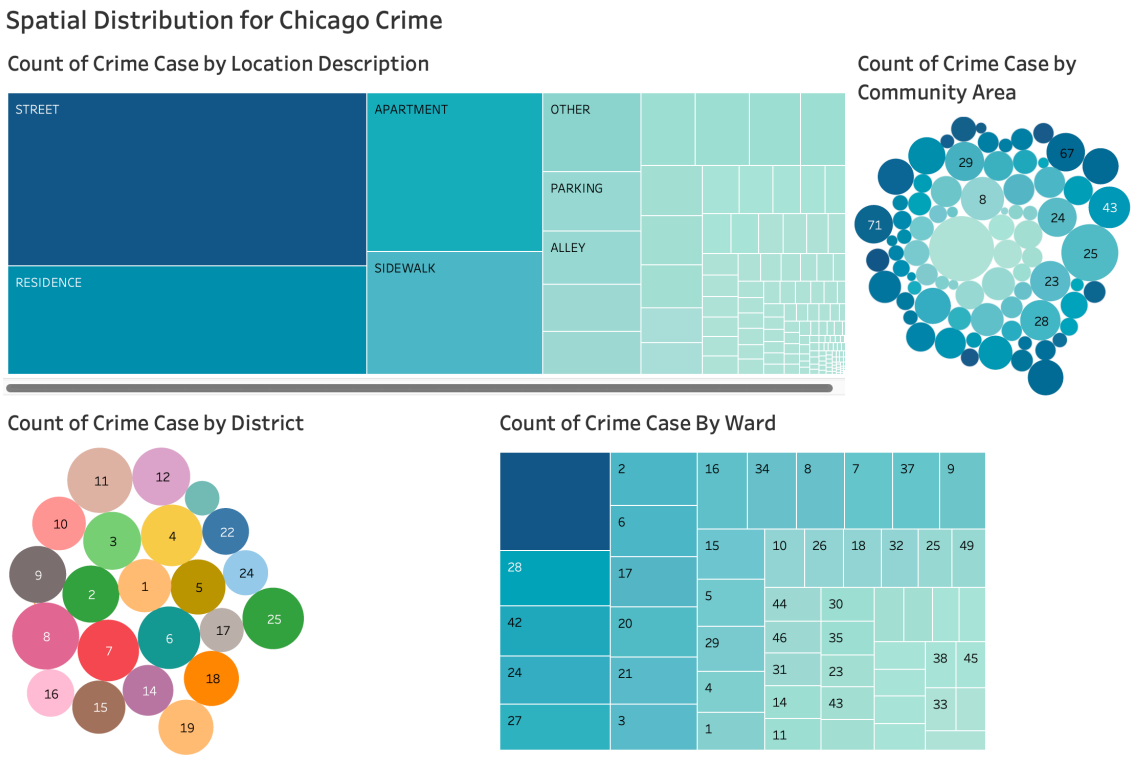

Figure 2: Composite View of Spatial Crime Patterns.

Analyzing Figure 2:

- A treemap (top-left) details common Location Descriptions. It shows crimes most frequently occur in public or accessible areas, predominantly on STREETs, followed by residential settings (RESIDENCE, APARTMENT) and SIDEWALKs.
- Geographical visualizations, including a bubble chart for Community Area (top-right), a bubble chart for District (bottom-left), and a treemap for Ward (bottom-right), demonstrate significant spatial concentration. Crime is not uniformly distributed; certain community areas, police districts, and wards bear a much higher burden of reported incidents than others, clearly indicating geographical hotspots."

1.4 Correlation Analysis: Crime Type, Time, Location, and Arrests

Beyond the general temporal and spatial trends, further analysis reveals correlations involving crime types. Visualizations breaking down crime counts by primary type over months and hours show that 'THEFT' and 'BATTERY' are dominant categories, largely driving the overall temporal patterns observed in Figure 1. These detailed views can also reveal if specific crime types have unique peak times or seasonality.

Crime types are also strongly associated with specific locations. For instance, 'THEFT' is particularly prevalent on 'STREET's and in retail environments, while 'BURGLARY' is more common in 'RESIDENCE's, aligning with the locations identified in Figure 2.

Finally, the relationship between crime type and arrest outcome is critical. Figure 3 illustrates arrest ratios across different primary crime types. It highlights that while 'THEFT' and 'BATTERY' have the highest volumes, their arrest rates are relatively low compared to types like 'NARCOTICS', 'HOMICIDE', or 'WEAPONS VIOLATION'. This disparity likely reflects differences in evidence availability, solvability, or policing priorities.

1.5 Summary of EDA Findings

The exploratory analysis highlights several core characteristics of the reported Chicago crime data:

- Crime follows distinct temporal patterns, peaking seasonally in summer and diurnally in the afternoon/evening, with additional peaks observed at 12 PM (noon) and 24 PM (midnight) in the hourly distribution.

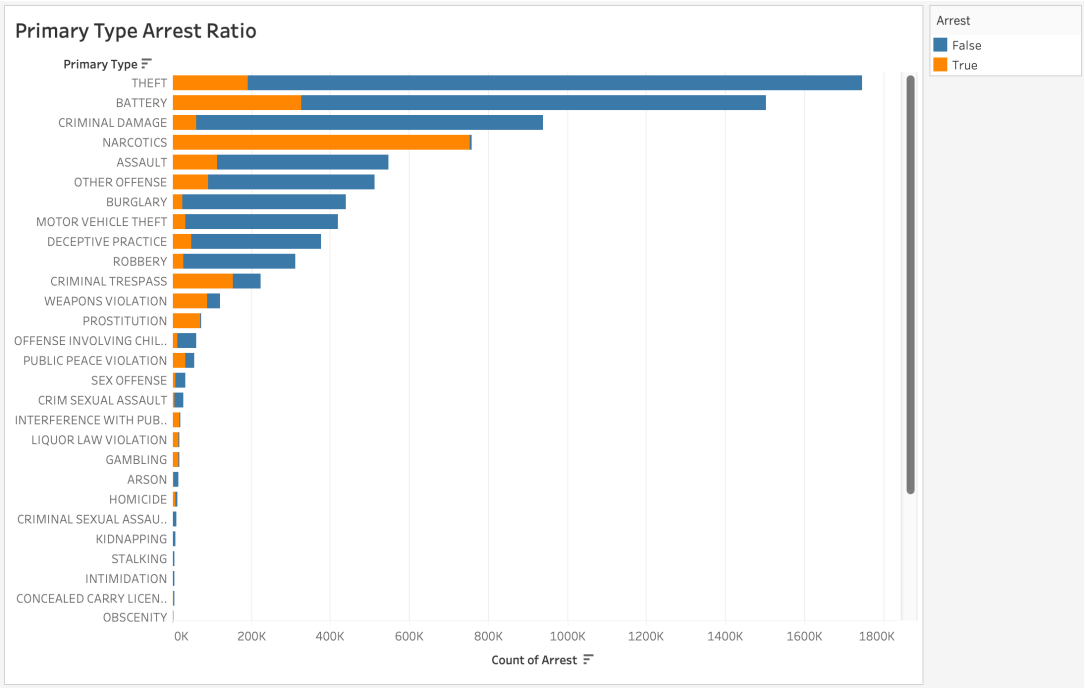

Figure 3: Arrest Ratio by Primary Crime Type (Illustrative, based on Figure 6 in source PDF).

- Spatial distribution is highly uneven, concentrated in specific districts and common location types like streets and residences.
- 'THEFT' and 'BATTERY' are the most frequent crime types overall.
- Arrest rates vary significantly by crime type, not directly correlating with report volume.

These observations provide valuable context for understanding crime dynamics and guide further analysis and modeling efforts.

2 Preprocessing

2.1 Introduction

This section presents the steps of data preprocessing of the Chicago crime dataset. The goal of data processing is to clean the dataset and prepare the data for future use in the model and prediction.

2.2 Data Cleaning and Preparation

We performed a cleaning and preparation in the following steps:

S1: Read the dataset and use the data visualization methods to check the structure and attributes of the dataset.

S2: Identified missing values in critical columns such as Location Description, Latitude, and Longitude. For missing values in data cleaning, usually, we have three different ways to deal with this.

- Drop missing value.
- Fill in the missing value with our own assumption.
- Train the model with existing value to predict missing value.

As shown in Figure 4, the missing value ratio is low, and we chose to drop the missing value. Some features have very few missing values so that we cannot see them in the plot, so we use pandas to find exactly which are missing, then we drop these missing values in District, Location Description, X Coordinate, Y Coordinate, Latitude, Longitude, Location, Community Area, Ward.

S3: For data analysis in the future, we converted the date-time columns into a standardized format (DateTime) and transformed categorical variables into appropriate formats.

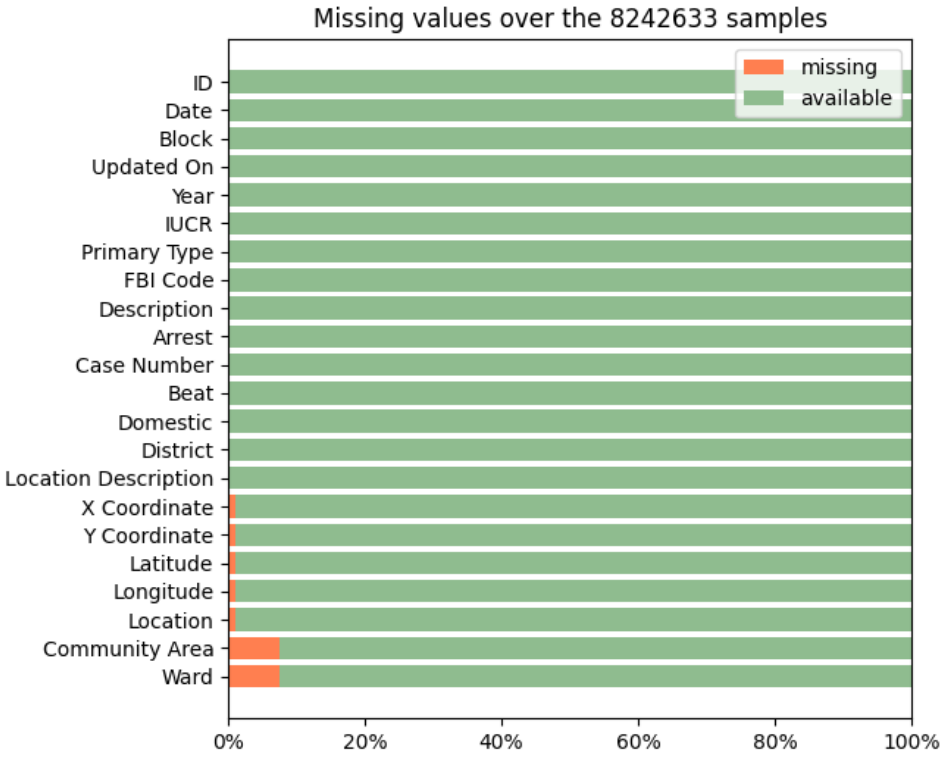

Figure 4: Missing values in the original dataset.

S4: Duplicate records were removed to maintain data integrity. Filtered out irrelevant or inconsistent data points.

Updated On, Year, Location, X Coordinate, and Y Coordinate were dropped as they either contained duplicate information or were not relevant to spatial analysis.

IUCR, FBI Code, and Description were removed since Primary Type already sufficiently represents crime categories.

Block and Location Description were removed as they provide overly specific location details, which are not useful for general crime prediction.

To enhance model generalizability, beat (the smallest police patrol area), District (comprising multiple beats), Ward (city council district), and Community Area (one of 77 Chicago community areas) were excluded. Instead, geographic coordinates were retained for spatial analysis using grid-based segmentation.

Domestic was removed as it is only relevant if predicting domestic crimes specifically.

The Primary Type was initially dropped because its utility in the modeling stage was uncertain. It can be reintroduced later if needed.

By removing these columns, the dataset becomes more structured, reducing noise and improving the efficiency of further analysis and predictive modeling.

2.3 Remove Spatial Outliers

To better predict results, we conduct a shrinking tail to remove spatial outliers. Figure 5 illustrates the final results after this step.

2.4 Temporal and Spatial Aggregation

To simplify the dataset, and make it more structured and suitable for crime prediction models, we effectively incorporate spatial and temporal information.

- Spatial Aggregation: Crime data is grouped into a 10×10 grid based on geographic coordinates, and the original latitude and longitude values are removed.
- Temporal Aggregation: The timestamp is standardized to hourly intervals to facilitate time-series analysis.
- Splitting Data by Arrest Status: Crime incidents are categorized into ”arrest” and ”not arrest” groups. A time-series dataset is constructed by counting crimes per grid and hour,

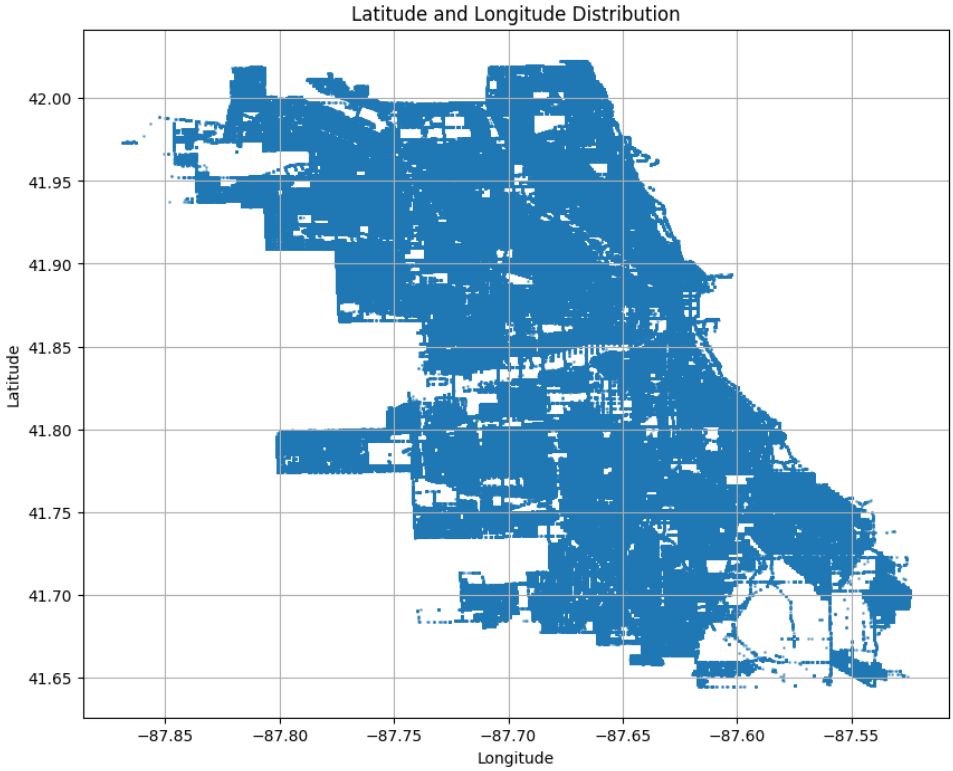

Figure 5: Latitude and Longitude Distribution.

ensuring that missing timestamps are filled with zero values.

2.5 Summary of Data Preprocessing

These preprocessing steps significantly improved data quality, making it more suitable for analysis. The cleaned dataset can now be effectively used for crime pattern detection, trend analysis, and predictive modeling.