1 Dataset

In this data analysis project, we will use the open dataset from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting system, including crime data from 2001 to present (excluding the most recent 7 days) in Chicago. The dataset contains 22 columns, providing information about temporal aspects (e.g., date, year, updated_on), spatial aspects (e.g., block, location, ward, community area, beat), categorical features (e.g., FBI code, domestic, primary type), case handling (e.g., arrest, domestic), and detailed descriptions of the incidents.

To ensure the training data reflects recent crime trends, we focus only on the data after 2023, which is around 500,000 rows. Within the dataset, some columns exhibit varying degrees of missing data: location_description (0.44%), ward (0.0004%), community_area (0.0004%), location (0.279%, the same to coordinates). Also, some data require type conversion for further analysis.

Thus, data preprocessing for the crime data includes missing data handling, feature extraction, data cleaning, and normalization.

1.1 Data Cleaning

In data cleaning, we address redundancies and inconsistencies in the data format.

Some of the features are not necessary in our future training, either containing redundant inofrmation or are irrelevant for model training, thus we drop the features below:

- Block and location. Block is too granular for prediction and can be replaced by community area and district. Location can be replaced by longitude and latitude as they contain equivalent information.
- Primary_type and description. Almost all primary_type can be inferred from first two digits (including leading zeros) of IUCR code, and description for the last two digits. Although there are exceptions such as case started with 13 can be "criminal trespass" or "criminal damage", but they are trivial and will not significantly affect the model. Therefore, we drop these two columns and replace them with the related IUCR code.
- X-coordinate, Y-coordinate, updated_on.These features are not used in future modeling.

With 138 unique values for location_description, we need to simplify the future training by limiting the number of entries for this column. We group the descriptions with occurrences under 60 to "Others".

There are three kinds of format inconsistencies in Chicago crime data, and should be handled differently.

First, the date column is not in proper datetime format, which we transformed

to standard datetime format for better time-based analysis. Longitude and latitude also are transformed to float for future missing value handling.
Also, we standardize the location description by stripping the whitespace and special characters and converting them to lowercase. The up-mentioned IUCR alternative to description data also include string, we delete those rows to avoid confusing future machine learning.
Lastly, as arrest and domestic columns are originally checkbox data, we convert them to integers (1 for True and 0 for False) for facilitate modeling.
Because there are no duplicates detected in this dataset, we did not perform duplicate removal to clean the data.

1.2 Missing Data Handling

As missing data occurs primarily in spatial and descriptive columns, different strategies are applied: spatial data are inferred by other spatial columns without loss and descriptive column are filled with specific values.
For location coordinates, we use the median coordinates of the cases that happened in the same block or district to impute the missing data. Similarly, we use the most frequent wards and community area data from the same district to suggest the missing ward and community.
For location description, as the description may vary case by case, we choose to fill in "unknown" to handle this missing value.

1.3 Temporal and Spatial Aggregation

Time and space should be two of the most crucial data in the Chicago crime dataset. However, with limited date and location data, the dataset could be too specific and less informal. Thus, we did temporal and spatial aggregation to extract information and take deeper look in the data.
To get more clear information about temporal occurrences, we extract the year, month, day, and hour from the date column. Additionally, a new column is created to distinguish weekdays from weekends, providing insights into weekly crime patterns.
We also aggregate the crime counts per block and district to identify crime density in different areas.

1.4 Feature Standardization

To prepare categorical features for machine learning, we apply one-hot encoding to the location_description column, turning it into numerical features suitable for training.