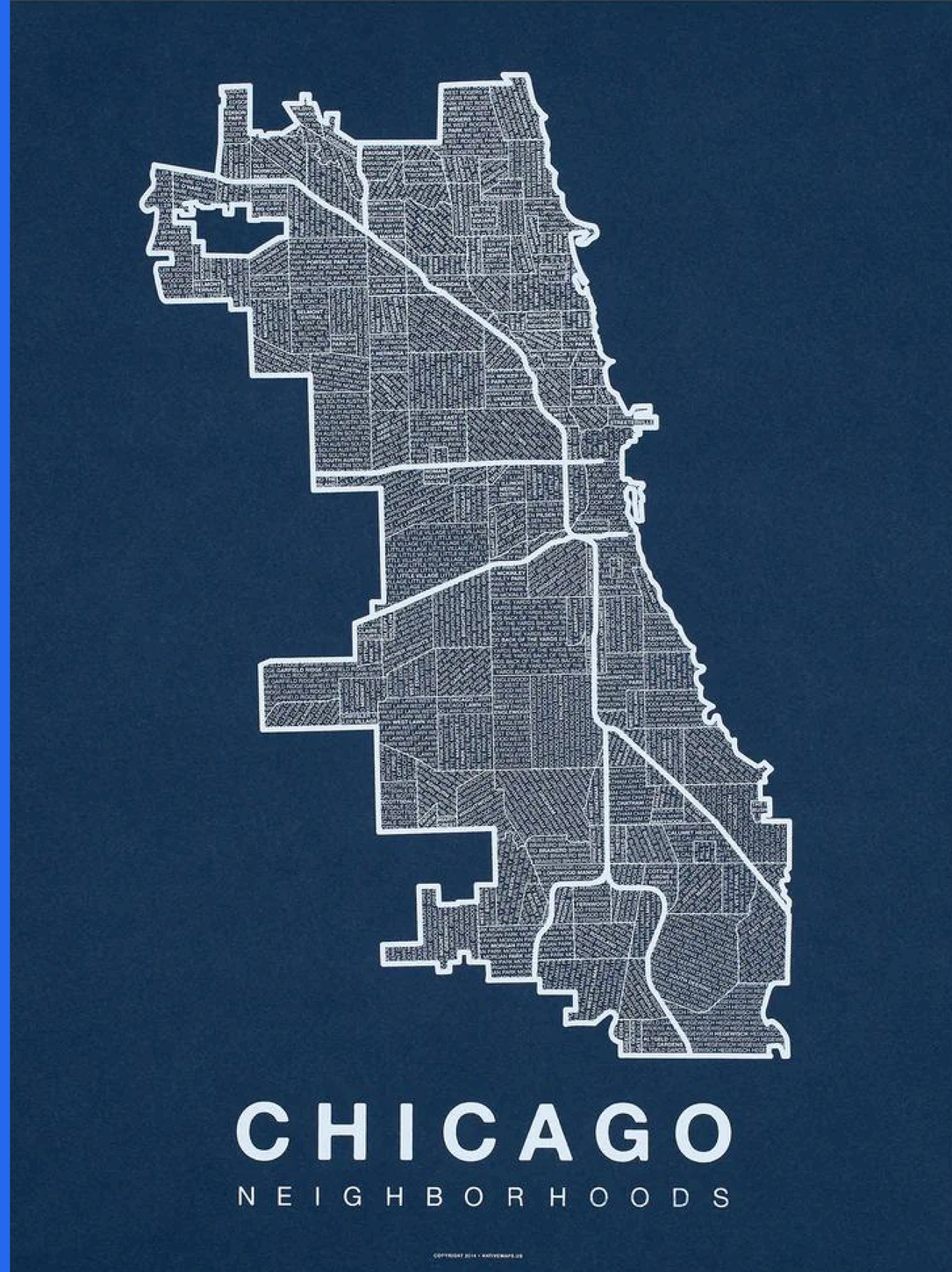
Criminal Chicago Big Data Project

team16

May, 2025



Agenda

1 Business Problem

2 Dataset Description

3 Data Insights

4 Analysis of Results

5 Stages Results

6 Challenges

7 Demo

Meet the Team

Amine Trabelsi

Data Exploratory Analysis and storytelling

Abdurahmon Abdukhamidov

Model Training and evaluation

Elena Tesmeeva

Dashboards analysis and Training Monitoring

Nikita Rashkin

Data Processing and Storing

Business Problem

Estimating the time when the next crime happens can be crucial for many federal organizations as well as ordinary people who would like to get insights about their safety in their city and district of living

Dataset Description (part 1)

- ID: Unique identifier for the record.
- Case Number: Chicago Police internal reference number.
- Date: Date and time of the incident (sometimes estimated).
- Block: Partially redacted street-level address.
- IUCR: Illinois Uniform Crime Reporting code, tied to Primary Type.
- Primary Type: General crime category (e.g., THEFT, BATTERY).
- Description: Subcategory of the Primary Type.
- Location Description: Description of where the crime occurred (e.g., STREET, RESIDENCE).
- Arrest: Boolean indicating whether an arrest was made.

Dataset Description (part 2)

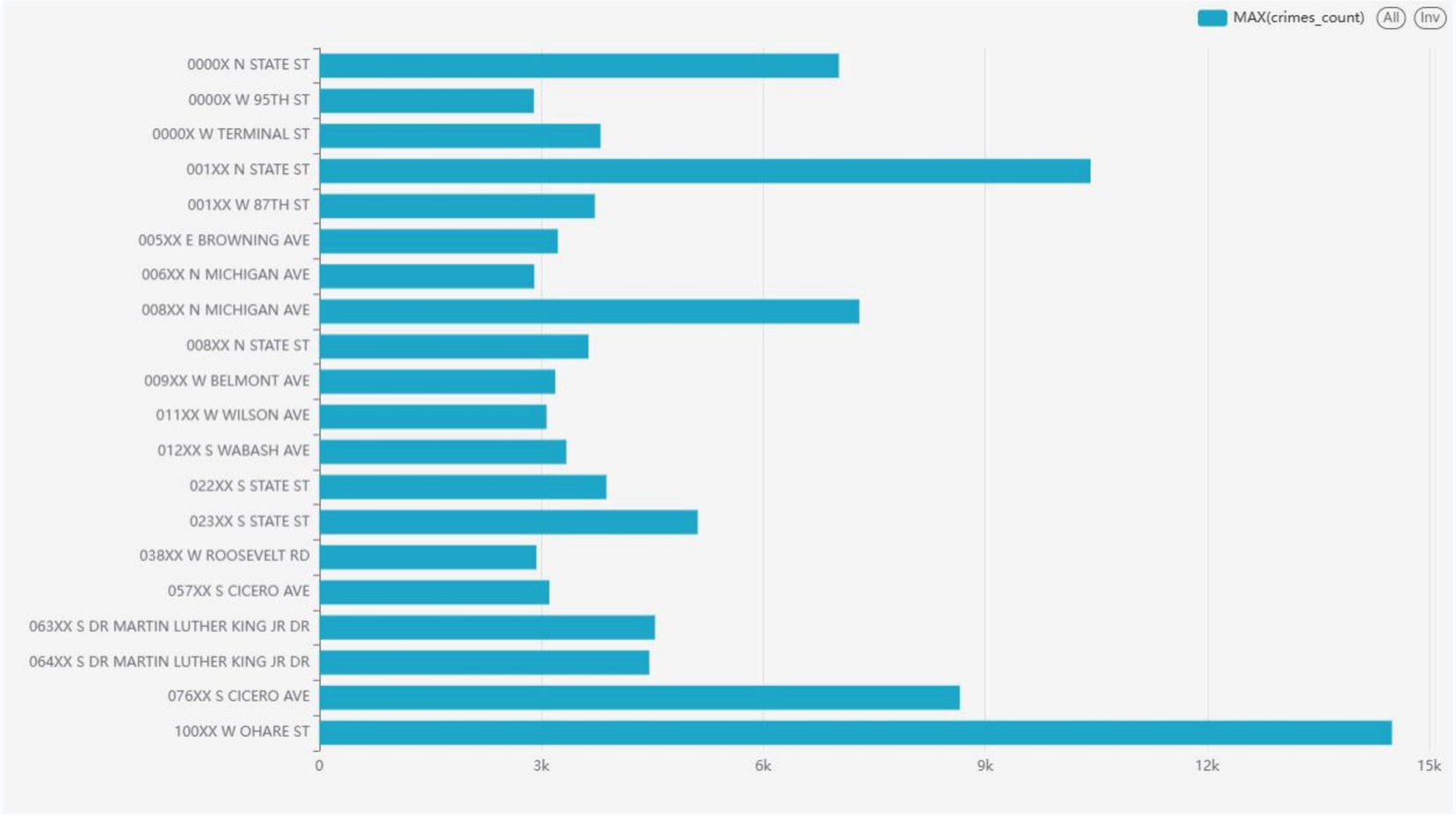
- Domestic: Boolean indicating whether the crime was domestic-related.
- Beat: The smallest police patrol area.
- District: Police district code.
- Ward: City Council district where the crime occurred.
- Community Area: One of Chicago's 77 community areas.
- FBI Code: Crime classification code used by the FBI.
- X Coordinate / Y Coordinate: UTM coordinates (spatial location, partially redacted).
- Year: Year the incident occurred.
- Updated On: Last update timestamp of the record.
- Latitude / Longitude: Geographic coordinates (partially obfuscated).
- Location: Combined geospatial field

EDA

Data Insights

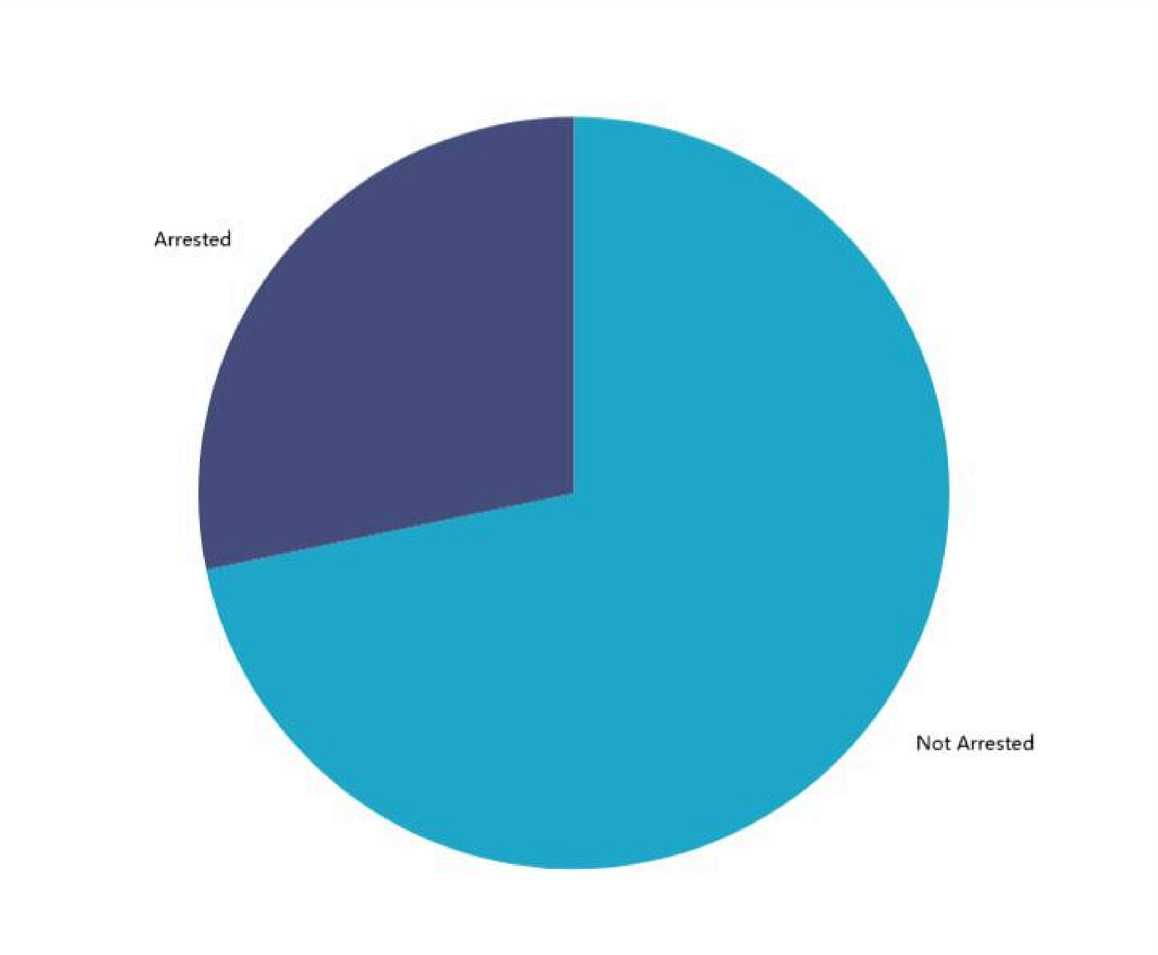
By analyzing crimes by location, type, arrest rate, and category, we can identify critical hotspots, evaluate policing effectiveness, and understand the social context behind specific offenses especially domestic-related crimes.

Crimes per neighborhood



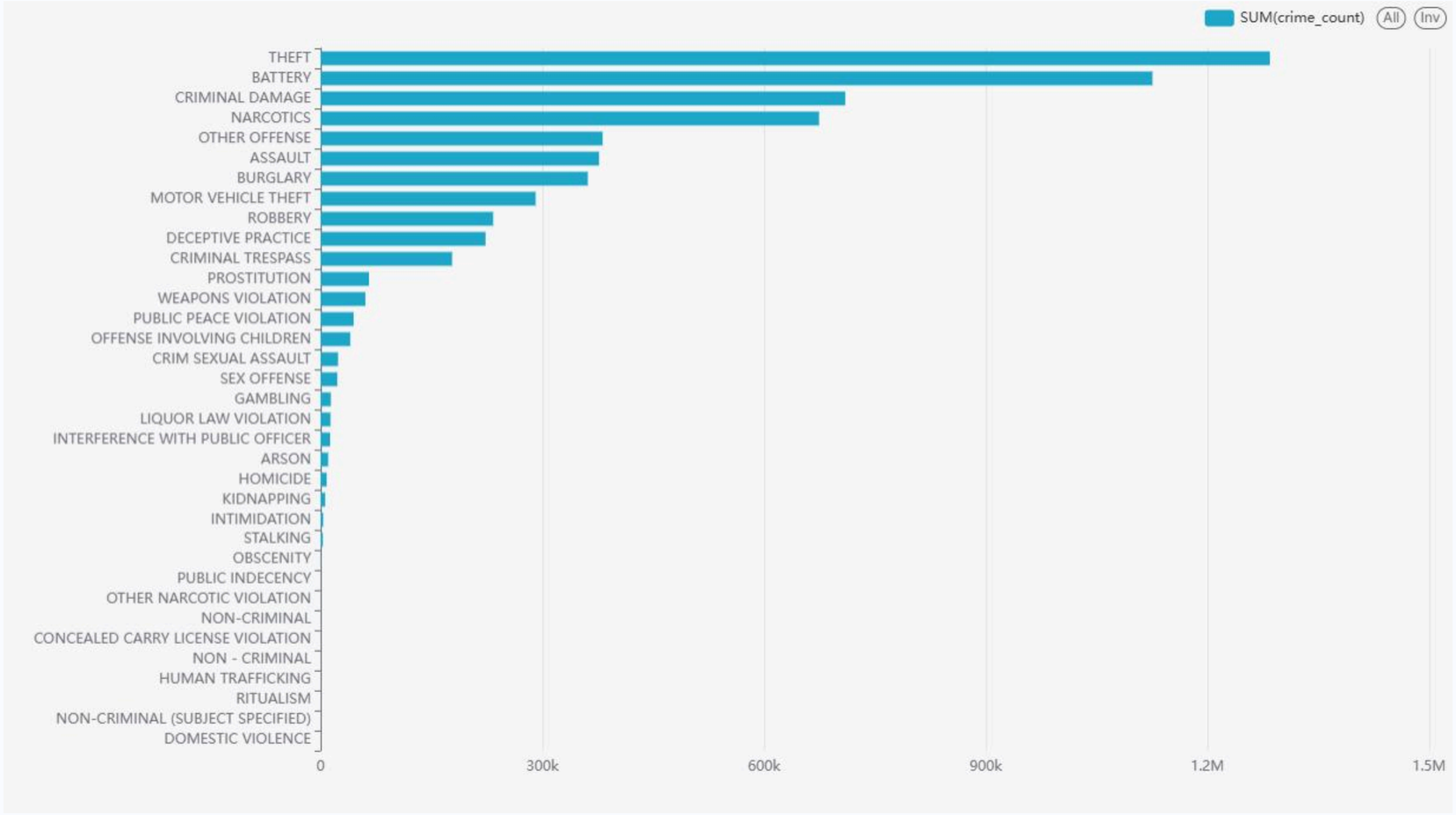
100XX W OHARE ST is the most crime-heavy location, with 14,499 reported incidents, making it 4,073 cases higher than the next location, 076XX S CICERO AVE (8,659).

Arrest rate for crimes



Arrest rates are low: out of approximately 6.17 million reported crimes, only 1.74 million resulted in arrests meaning over 71% of crimes did not lead to an arrest.

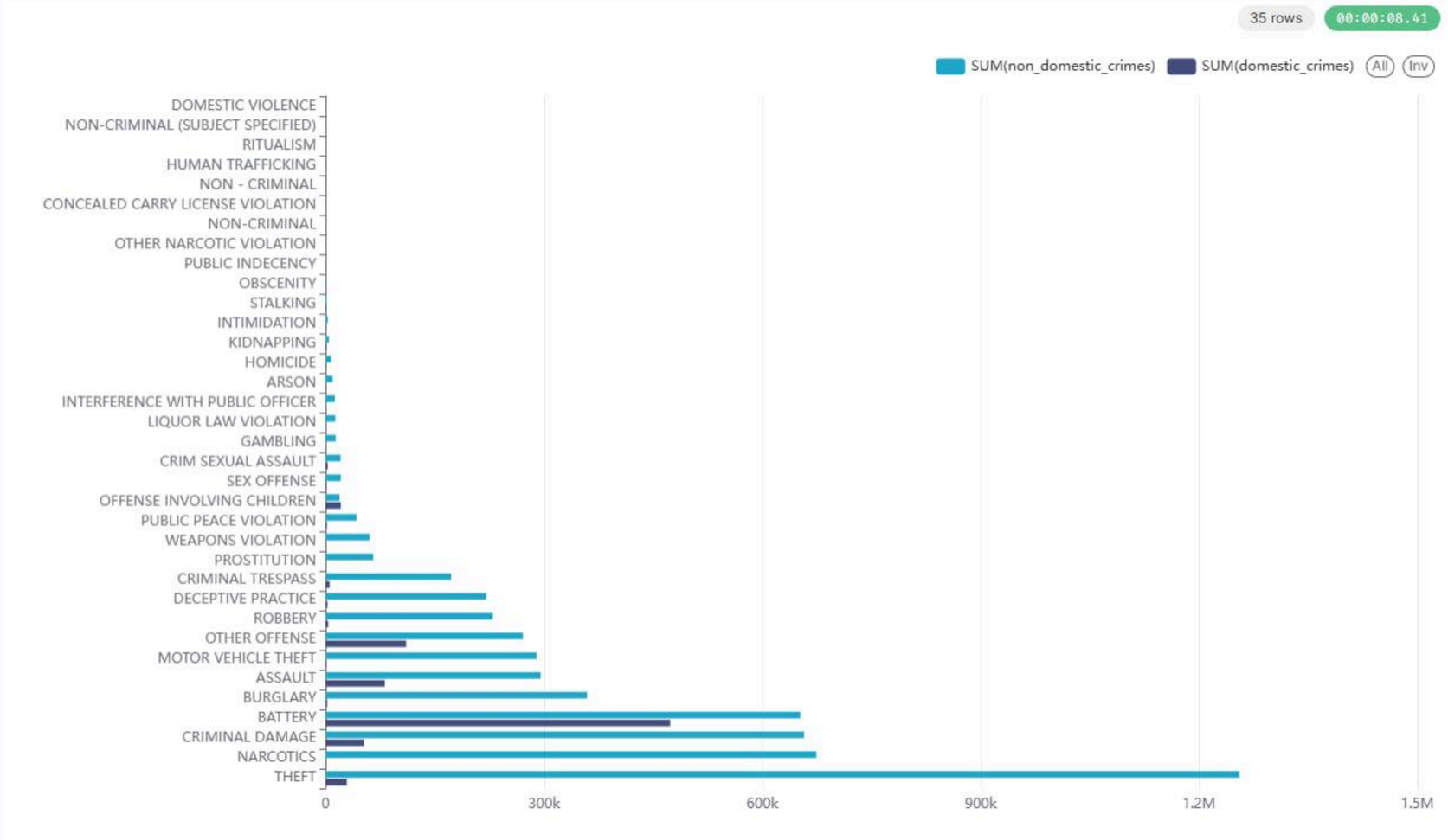
Categories of Crimes



The top five crime types by volume are:

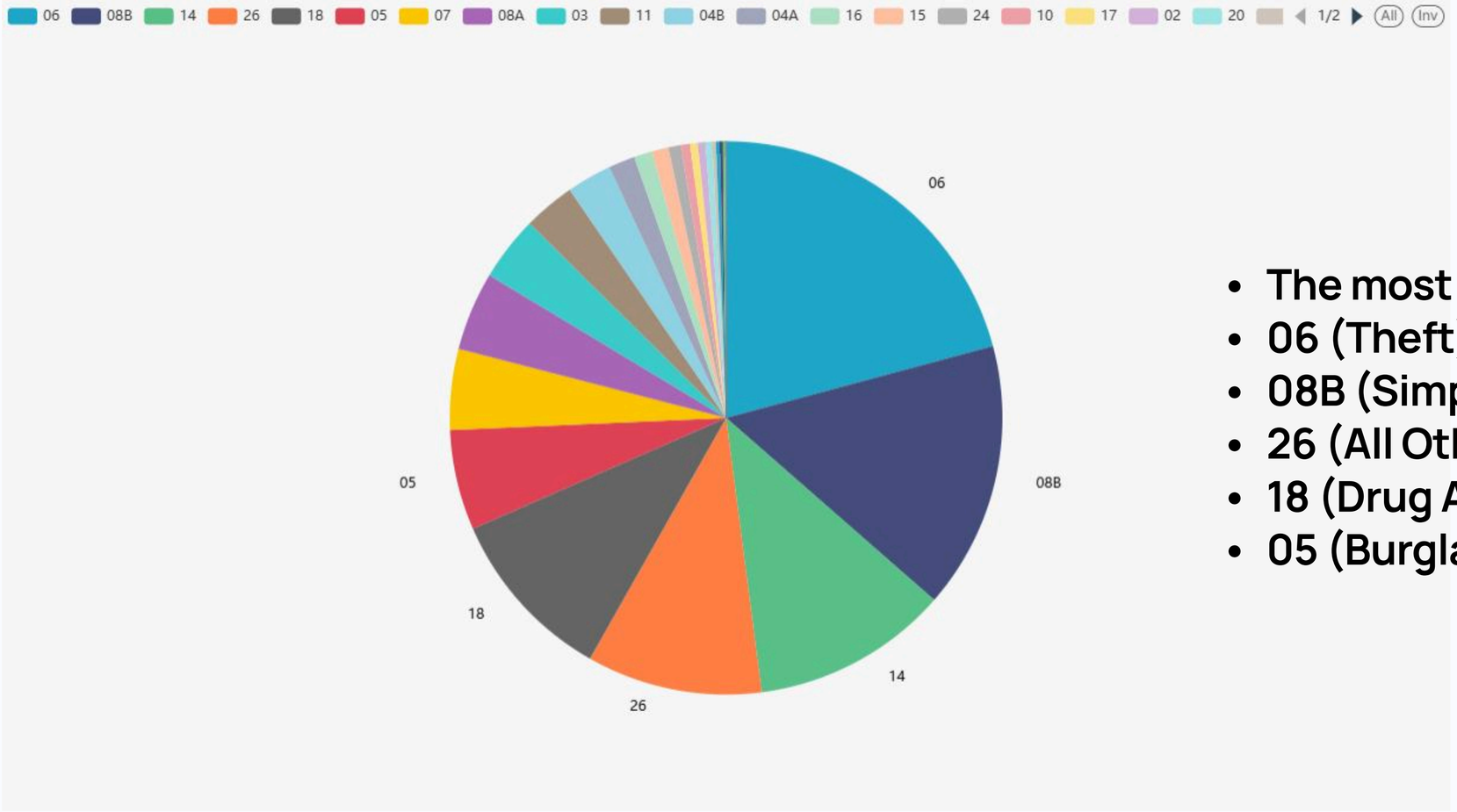
- Theft: 1,284,608 cases
- Battery: 1,125,725 cases
- Criminal Damage: 710,082 cases
- Narcotics: 674,556 cases
- Other Offense: 381,714 cases

Domestic vs non-Domestic



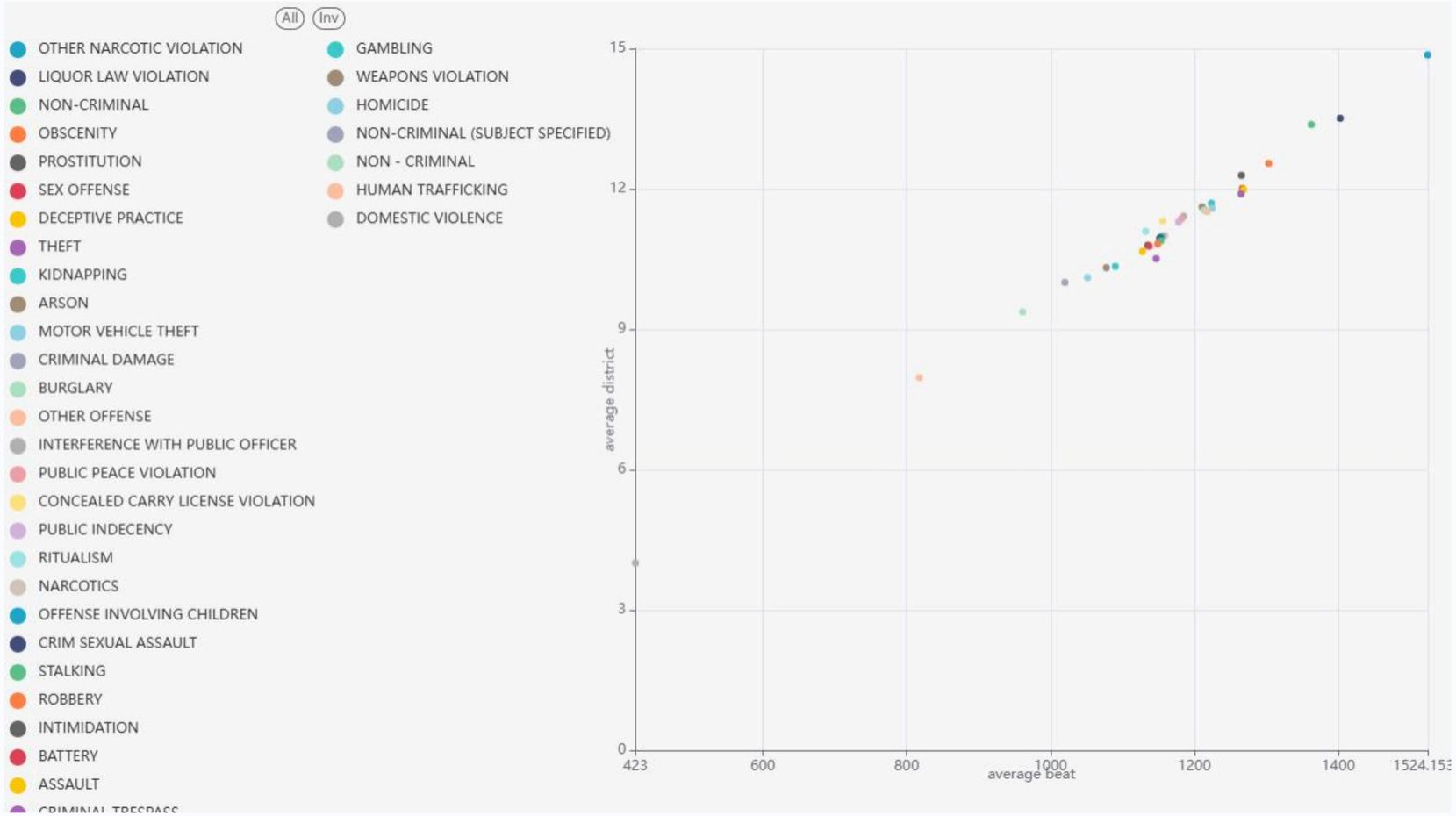
- Battery has 473,499 domestic-related incidents, representing 42% of its total.
- Assault follows with 81,371 domestic cases, about 21.6% of the category.
- Criminal Damage and Other Offense also have significant domestic shares (7.4% and 29%, respectively).

Codes used for crimes



- The most common FBI crime codes are:
- 06 (Theft): 1,284,608 cases
- 08B (Simple Assault): 963,362 cases
- 26 (All Other Offenses): 633,337 cases
- 18 (Drug Abuse Violations): 631,977 cases
- 05 (Burglary): 361,624 cases

Location analysis



Crime distribution by district and beat shows that offenses like Other Narcotic Violations and Liquor Law Violations are concentrated in higher-numbered beats and districts (e.g., average beat >1400)

Stage Results

We achieved an error rate of less than 0.5 hours in predicting the time of the next crime

Analysis of Results

LinearRegression

Train R^2: 0.6821
Train MAE: 1921.4928 seconds
Test R^2: 0.0577
Test MAE: 2352.1970 seconds

SVR

Train R^2: -0.013025762904651295
Test R^2: -0.017938046323558998
Train MAE: 36990.68875229361
Test MAE: 35227.85355072754

LGBMRegressor

Train R^2: 0.3214487328422152
Test R^2: 0.003460908525462103
MAE: 1351.1564149316278
RMSE: 9672.096064682535
R^2: 0.3214487328422152

Challenges

Data Preparation:

- Duplicates in dataset
- Missing numbers
- Feature selection

EDA:

- Data files were not automatically deleted once the table is dropped and would result in duplicate columns.
- Manually syncing the database columns in superset after change.

DEMO

Thank You

Team 16