Report On Chicago Crime

data (2001-2023) for Data Visualization

DS612 : Introduction To Data Visualization	
Team : T_05	
202418044	Kashish Patel
202418017	Sujal Dhrangdhariya
202418030	Manish
202418016	Devang Choudhary

Dataset link :

https://drive.google.com/drive/folders/1e4PbcKjKJrJFjIXzD1nbcObAj2blkDYU?usp=drive_link

Index

Overview

This project analyzes Chicago crime data from **2001 to 2023**, leveraging data visualization to uncover crime patterns, temporal trends, arrest rates, and the nature of domestic incidents. The primary aim was to provide an interactive and insightful view of how crimes have evolved over time and what categories dominate the criminal landscape.

The solution uses dashboards to present layered insights including time-of-day trends, crime category breakdowns, domestic violence proportions, and arrest effectiveness. Through this approach, the project not only enhances understanding of the city's safety concerns but also enables data-driven storytelling for stakeholders in policy, law enforcement, and urban development.

This report highlights key takeaways on public safety, policing effectiveness, and crime seasonality for better decision-making.

1 Data Observation

1.1 Introduction

The data for this project comes from the **Chicago Crime Dataset (2001–Present)**, which is published by the **City of Chicago** and collected by the **Chicago Police Department**. It is updated every day and contains detailed information about each reported crime, such as the type of crime, where and when it happened, and whether an arrest was made. The police collect this data from official reports and then clean and organize it before sharing it with the public. Since the data is shared by an official government source and updated regularly, it is considered to be **trustworthy and reliable**. However, it only includes crimes that were actually reported to the police, so some incidents may be missing. Overall, this dataset is a great source for studying crime trends, patterns, and locations in Chicago.

Dataset Link :- https://www.kaggle.com/datasets/nathaniellybrand/chicago-crime-dataset-2001-present

1.2 Columns & Rows Introduction

Columns : There are 22 columns Available in dataset
Rows : There are around 77.82 lakhs rows Available in dataset

Attribute	**Description**
ID	A unique numeric identifier automatically assigned to each crime record. It helps to uniquely distinguish each incident in the dataset.
Case Number	The official case reference used by the Chicago Police Department. It may

	appear on official documents and helps in tracking investigations.
Date	The specific date and time when the crime was reported to have occurred. Useful for time-based analysis such as trends by hour, day, or year.
Block	The anonymized street-level address (e.g., “002XX N STATE ST”), rounded to preserve privacy. Indicates general location without revealing exact address.
IUCR	Illinois Uniform Crime Reporting code, a numeric code that classifies crimes consistently across the state. Useful for standardized crime classification.
Primary Type	The major crime category, such as “THEFT”, “ASSAULT”, or “BURGLARY”. This is often used as a grouping variable in analysis and visualization.
Description	A more specific breakdown of the crime, under the Primary Type. For example, “SIMPLE” under “BATTERY” or “OVER $500” under “THEFT”.
Location Description	The type of place where the incident occurred, such as “STREET”, “RESIDENCE”, or “PARKING LOT”. Useful for understanding context and hotspot locations.
Arrest	A boolean (TRUE/FALSE) field indicating whether an arrest was made during or shortly after the incident.

	Helps evaluate police response or enforcement.
Domestic	Indicates if the incident was related to domestic violence, such as disputes between family or household members.
Beat	The smallest police geographic unit in Chicago, representing a patrol area. Important for identifying crime patterns at a very local level.
District	A larger administrative unit comprising multiple beats. Chicago is divided into 22 police districts, used for regional policing strategy.
Ward	Political division of the city, used for city council representation. Useful in connecting crime patterns with political or policy boundaries.
Community Area	One of 77 standardized neighborhoods in Chicago used for planning and public services. Frequently used in spatial visualizations and reporting.
FBI Code	A federal-level crime classification code assigned by the FBI. Enables cross-jurisdictional crime comparisons and federal reporting.
X Coordinate	The horizontal (X) coordinate in the Illinois State Plane coordinate system. Used in GIS mapping and spatial joins.
Y Coordinate	The vertical (Y) coordinate in the Illinois State Plane system. Works with X Coordinate to plot the precise point on a map.
Year	The calendar year when the crime occurred. Useful for tracking yearly

	trends and evaluating long-term policy impact.
Updated On	The timestamp showing the last time the crime record was updated in the database — often due to new case information or corrections.
Latitude	The north–south geographic coordinate of the crime location. Used in geospatial mapping to identify the crime's position on the globe.
Longitude	The east–west geographic coordinate paired with latitude to pinpoint the crime location.
Location	A combined (latitude, longitude) field that represents geographic coordinates as a point — ready for direct use in mapping tools like Tableau or GIS.

1.3 Problems In Dataset

Working with the Chicago Crime Dataset (2001–2023) (approx. 77.82 Lakhs rows) posed several challenges during loading and preprocessing. Key issues included:

1. Large Dataset Size

- ➢ Full dataset caused memory overload and slow performance in tools like Excel or Tableau.
- ➢ Basic operations became time-consuming or caused crashes.

2. Irrelevant Columns

- ➢ Columns like IUCR, FBI Code, and Case Number were not useful and added unnecessary bulk.

3. Missing & Inconsistent Data

- Fields like Location Description had null values.
- Category names had inconsistencies (e.g., "STREET" vs. "Street") impacting grouping.

4. Date & Type Conversion

- Date field required conversion from string to datetime.
- Boolean fields (Arrest, Domestic) were stored as text and needed formatting.

1. Stratified Sampling Setup

- 10% stratified sampling was used to ensure fair representation.
- Rare categories needed manual checks to retain at least one record.

6. Imbalanced Data Distribution

- Some crime types or years had very few records, making balanced sampling tricky.
- Direct random sampling would have missed low-frequency but important categories.

7. Processing Time

- Preprocessing the entire dataset (cleaning, type conversion, sampling) took significant time.
- Had to shift initial processing to Python due to tool limitations.

8. Tool Compatibility

- Full dataset couldn’t be handled smoothly by Tableau or Excel.
- Final 10% sample was used for analysis after cleanup in Python.

2 Sampling Method, Statistical Testing & Data Validation

To create a manageable yet representative subset of the Chicago Crime Dataset, Stratified Sampling was used. A new column named Strata was created by combining the Year and Primary Type fields, ensuring that each subgroup had enough entries to represent the overall distribution. A 10% sample was selected using train_test_split with the stratify parameter to maintain the distribution of crime types across years.

2.1 Chi-Square Test of Independence (For Arrest Column)

Why Used: This test is suitable for comparing two categorical distributions — in this case, the distribution of Arrest values (True/False) between the full dataset and the sample.
Formula: $\chi^2 = \sum((O_i - E_i)^2 / E_i)$
Where:
O_i = Observed frequency
E_i = Expected frequency under the assumption of independence
Test Result: P-value > 0.05
This indicates there is no significant difference in arrest distributions between the sample and full dataset.

2.2 Kolmogorov–Smirnov (KS) Test (for Hour column)

Why Used: The KS test compares the distributions of two continuous (numerical) variables — here, the distribution of crime hours in the sample vs. full dataset.
Formula: $D = \sup_x |F_1(x) - F_2(x)|$

Where:
$F_1(x)$, $F_2(x)$ are the empirical cumulative distribution functions (ECDF) of the two datasets.
D is the maximum distance between the ECDFs.
Test Result: P-value > 0.05
This means the hour-wise distribution of crimes is not significantly different between the full data and the sample.

2.3 Statistical Test Results Summary

Feature	Test Used	Test Statistic	P-Value & Interpretation
Arrest	Chi-Square Test	0.2484	0.6182 → No significant difference. Sample is valid.
Domestic	Chi-Square Test	0.3295	0.5660 → Proportionally similar. Sample is valid.
Location Description	Chi-Square Test	213.4843	0.5356 → No significant difference. Sample is valid.
Community Area	Chi-Square Test	59.0559	0.9360 → Distribution consistent. Sample is good.
Hour	KS Test	0.0006	0.9538 → Distribution preserved. Sample is representative.

2.4 Handling Null Values

After sampling, some records in the dataset had missing or incomplete values. To ensure data quality, a combination of deletion and imputation techniques was used. Records with missing Location Description were removed, as this information is essential for understanding the context of crimes. For entries where the Community Area was missing but latitude and longitude were available, values were imputed using a reference file from the Chicago Data Portal, which links geographic coordinates to community areas. Similarly, when latitude or longitude was missing, the average location of other records within the same community area was used to fill in the missing values. These preprocessing steps helped maintain the integrity of the sampled dataset for accurate analysis and visualization. In total, only 0.2% of the sampled data was dropped from location description

2.5 Removal Attributes from data with Reason

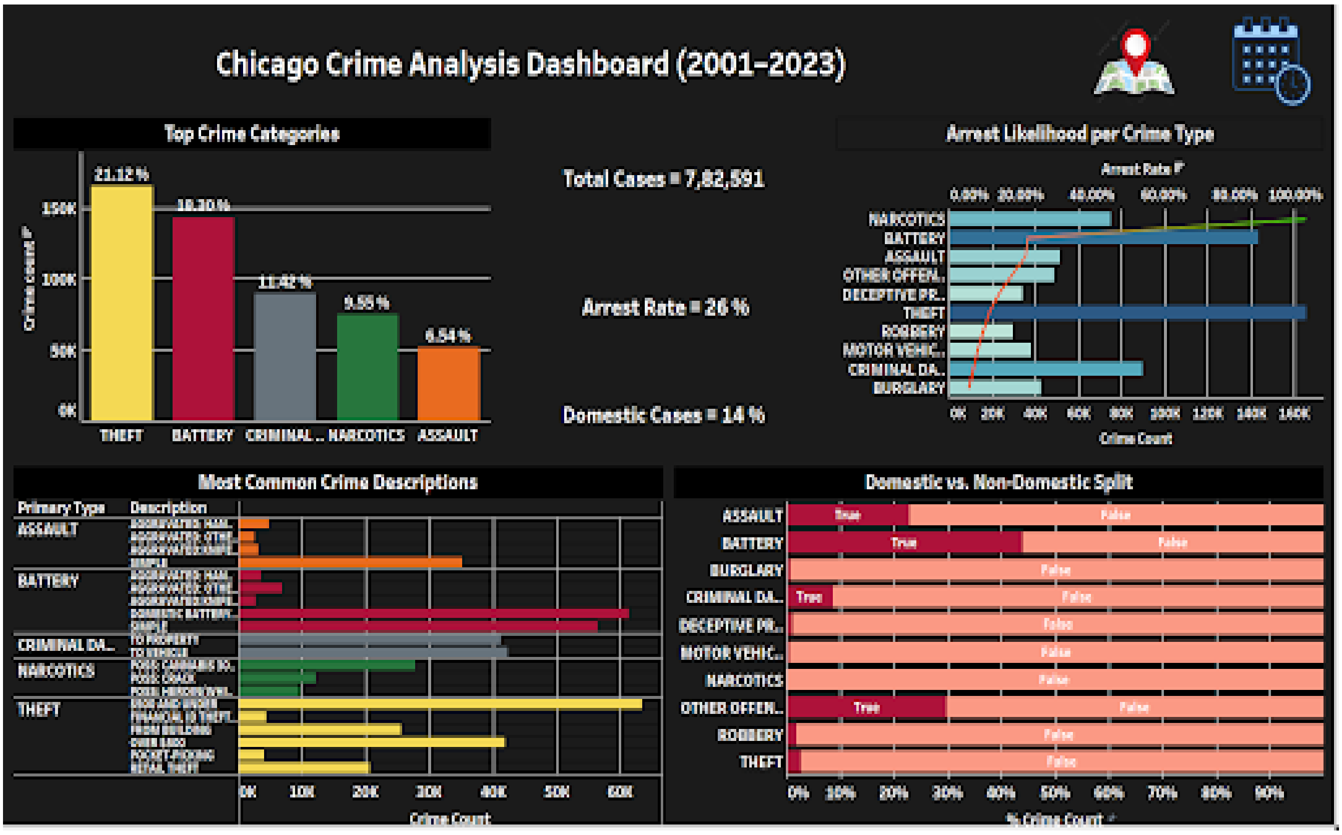
Attribute	Reason for Removal
Case Number	This is just an identifier used by the police; it holds no analytical value and does not contribute to trends, patterns, or insights.
Block	The exact block-level address isn't necessary for high-level analysis and may raise privacy concerns; community area or latitude/longitude provides sufficient spatial granularity.
IUCR	This is a numeric code for crime classification, but it's redundant since Primary Type and Description already offer more readable and useful categorizations.

Beat	A very small policing unit; for broader insights, using District or Community Area is more practical and interpretable for non-specialist audiences.
District	Often overlaps in purpose with Community Area, which is more publicly understood and standardized in most reports and maps. May be dropped to avoid redundancy.
Ward	A political boundary not directly relevant to crime pattern analysis; unless you're analyzing policy impact or council-level differences, it's not useful.
FBI Code	A generic federal classification — it's often too broad or redundant when Primary Type and Description already provide granular crime types.
X Coordinate	These are in State Plane projection (not geographic coordinates), which are less intuitive and harder to visualize than Latitude/Longitude.
Y Coordinate	These are in State Plane projection (not geographic coordinates), which are less intuitive and harder to visualize than Latitude/Longitude.
Updated On	This shows when the database was last modified, not when the crime occurred; it doesn’t contribute to the actual crime analysis.

Location	This is just a merged version of Latitude and Longitude, which are already available separately for mapping — keeping both is redundant.

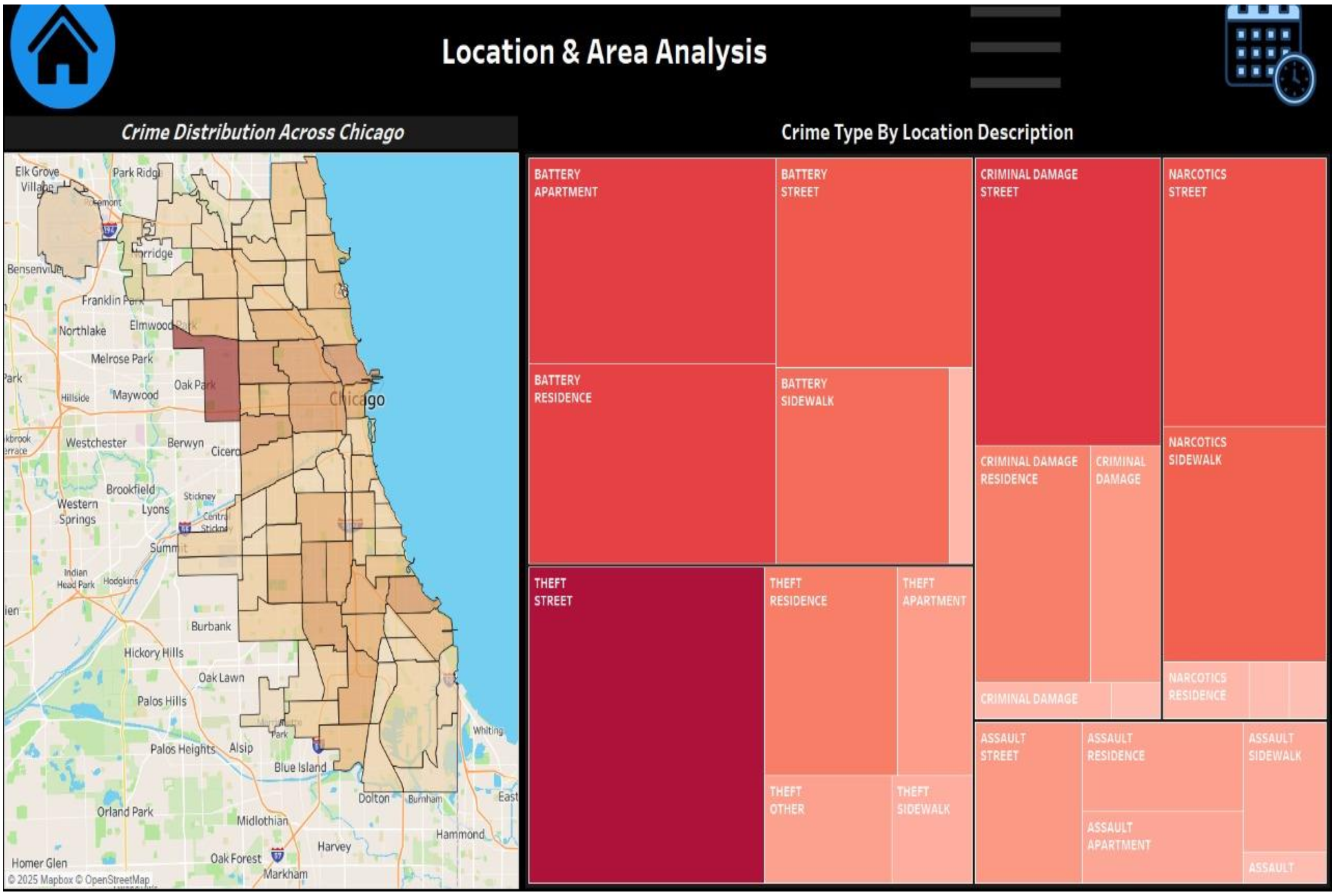
3 Dashboard & Chart Visualization

3.1 Dashboard Views



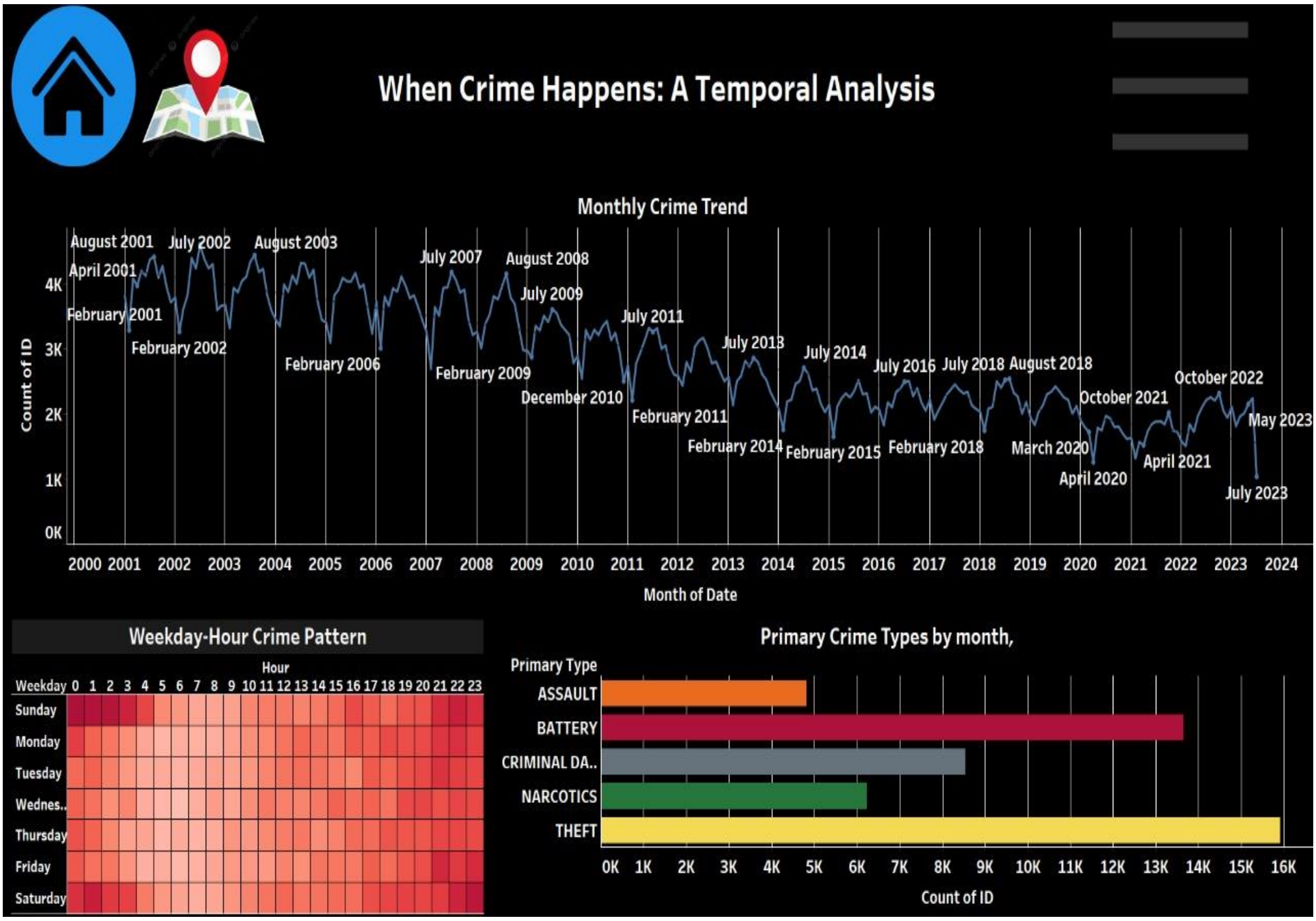
Hypothesis 1 Proof

Certain Locations Have Higher Incidents of Specific Crime Types



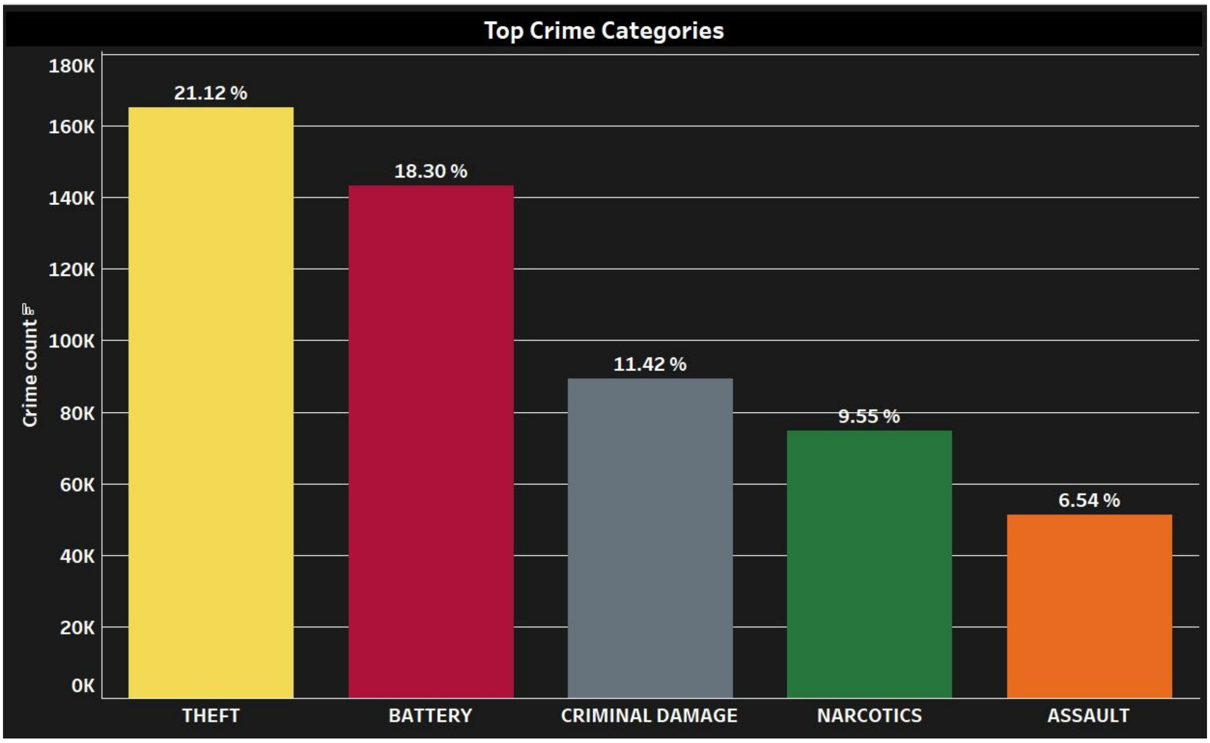
Hypothesis 2 Proof

Crime Rates Vary by Month and Season



3.2 Chart Visualization

3.2.1 Top Crime Categories



Attributes Used: Primary Type (Categorical), Count of Records (Quantitative)
Marks: 1D
Channels: Position (Vertical), Color (Hue)

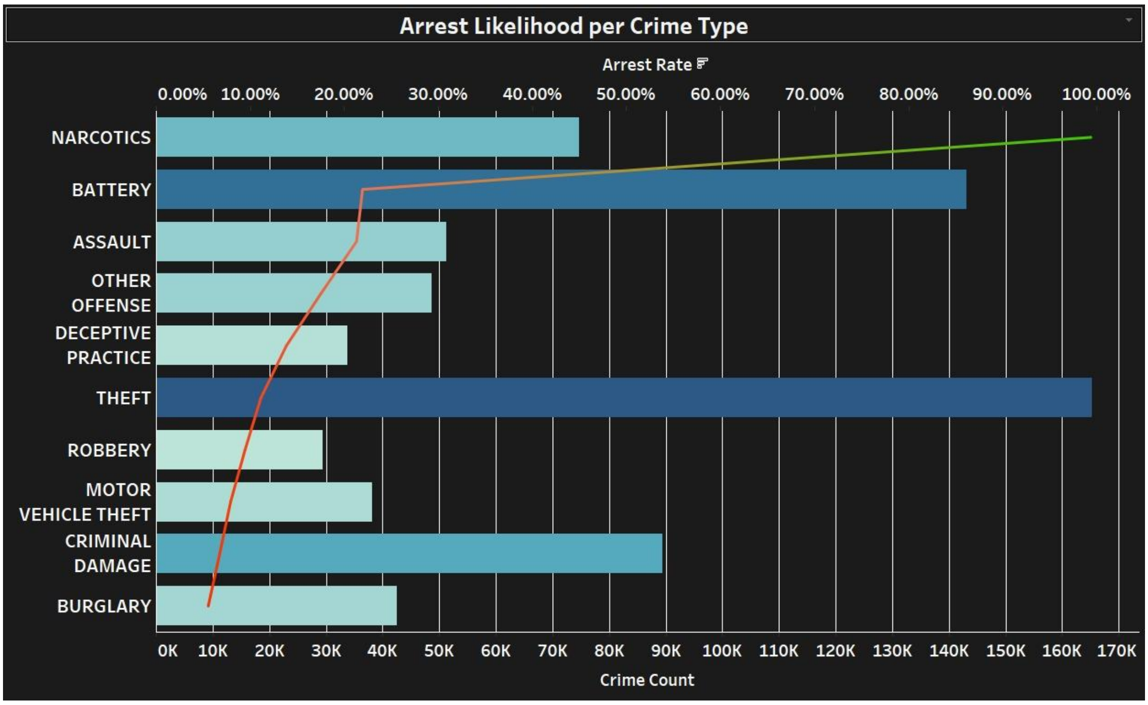
Why this Visualization Scheme?

• **Purpose**: Compare frequency of top crime types
• **Key Insight**:

Theft is highest (185K+), followed by **Battery** (143K+)

Assault is lowest among top 5 (51K+)
• **Scalability**: Can add filters like year, location, or arrest
• **Task Abstraction**: Ranking, Comparison

3.2.2 Arrest Likelihood per Crime Type



Attributes Used: Primary Type (Categorical),Count of ID(Quantitative),Arrest Rate (Quantitative - %)

Marks: 1D(bar), 0D(line)
Channels: Position (Horizontal), Length (Bar), Color (Arrest Rate)

Why this Visualization Scheme?

• **Purpose:** Show how arrest likelihood varies across crime types
• **Key Insight:**

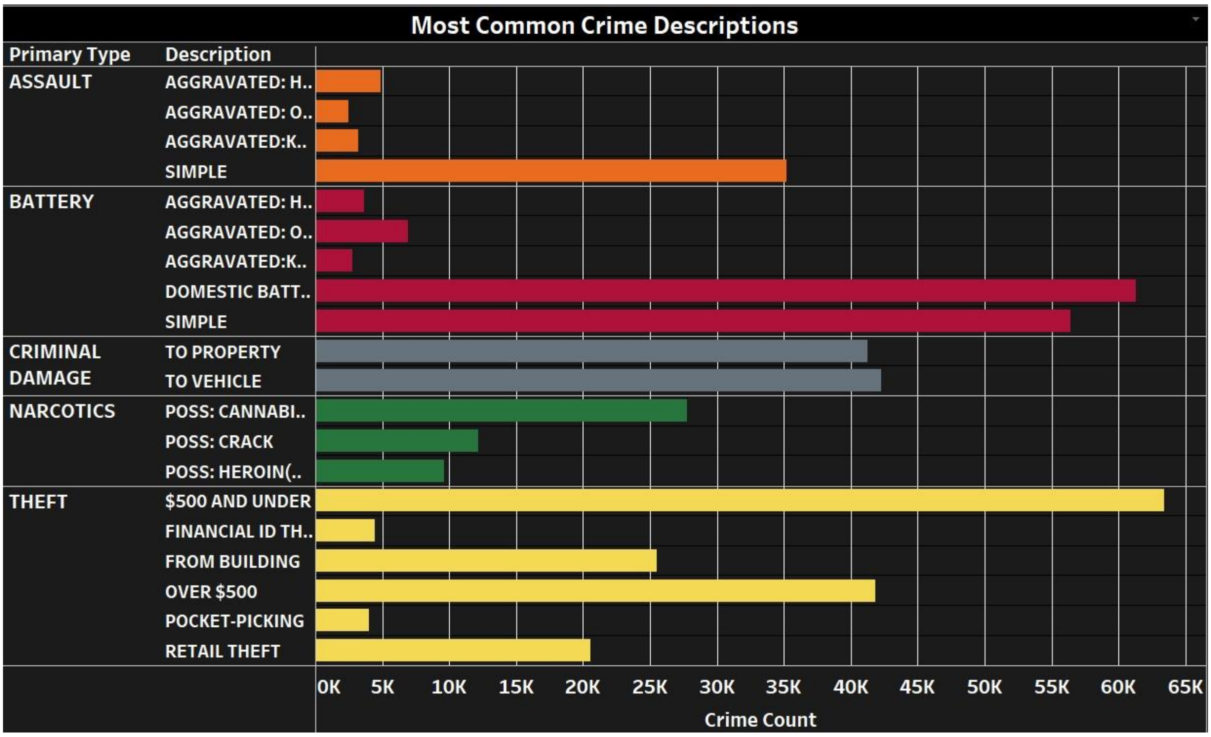
Narcotics **has highest arrest rate (~90%)**

Theft **is most common crime but has one of the lowest arrest rates (~5.5%)**

Strong disparity between crime volume and arrest likelihood
• **Scalability:** Easily extendable with filters (year, location, etc.)

• **Task Abstraction:** Association & Pareto Analysis (Volume vs. Effectiveness)

3.2.3 Most Common Crime Descriptions



Attributes Used: Primary Type, Description, Count of ID
Marks: 1D
Channels: Length (Vertical), Color (Primary Type), Position (Horizontal)

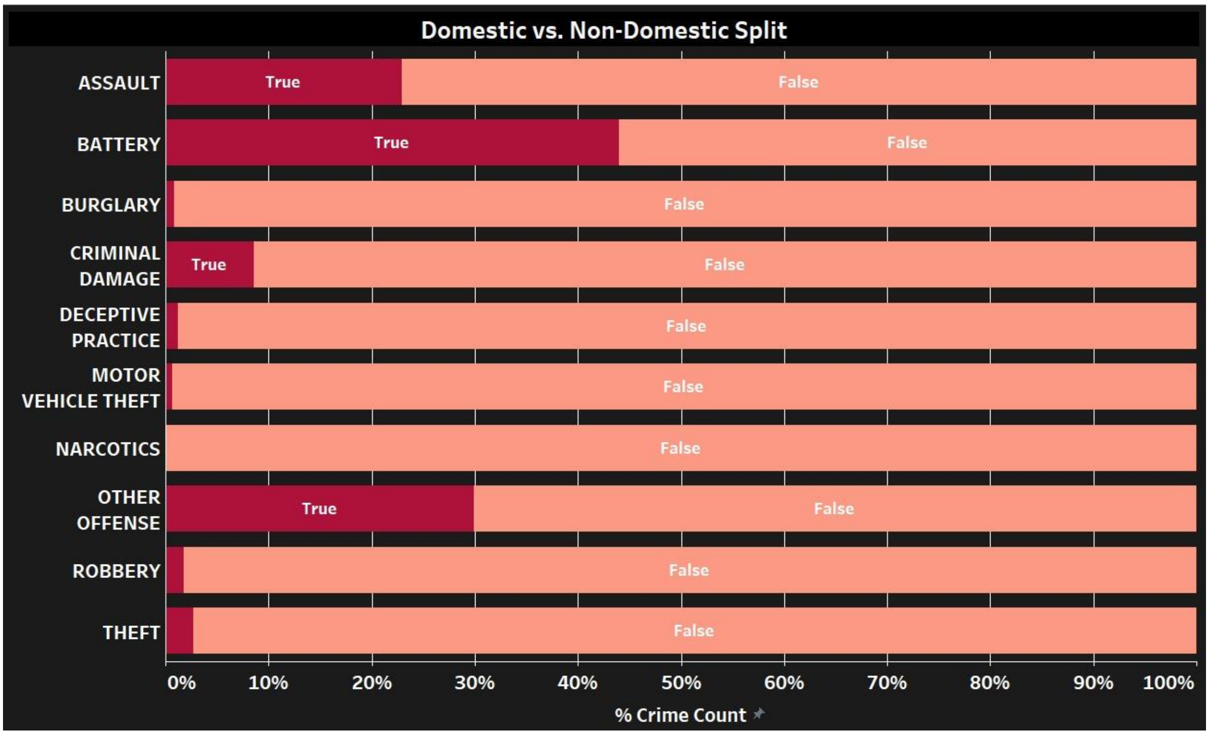
Why this Visualization Scheme?

- **Purpose**: Identify most frequent crime descriptions per category
- **Key Insight:**

Simple and *Domestic Battery* dominate in Battery

<$500 Theft most common in Theft
- **Scalability:** Easily filterable by description, type, or location
- **Task Abstraction:** Categorization + Frequency comparison

3.2.4 Domestic vs. Non-Domestic Split



Attributes Used: Primary Type (Categorical), Domestic (Boolean), Count of ID (Quantitative)
Marks: 1D (Stacked Bar**)**
Channels: Position (Horizontal), Color (True/False), Category (Vertical)

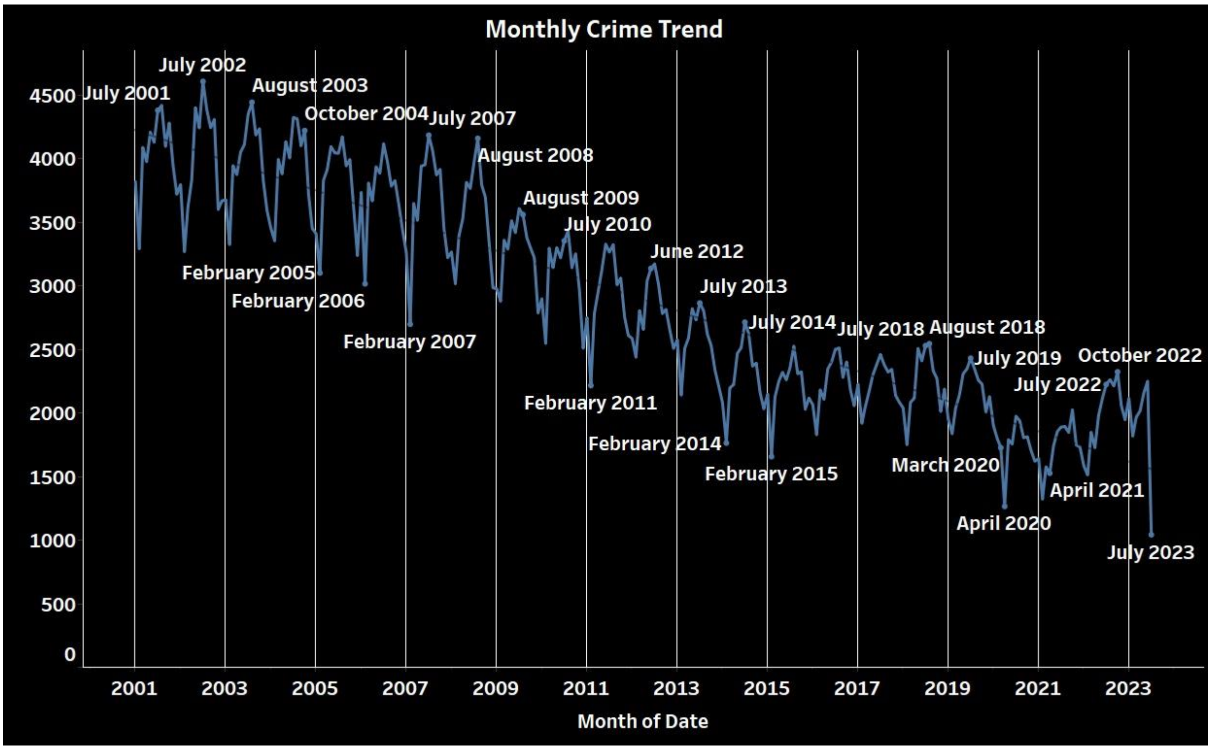
Why this Visualization Scheme?

• **Purpose:** Show proportion of domestic vs. non-domestic cases per crime type
• **Key Insight:**

Battery **has highest share of *domestic* crimes**

Most other types are majorly *non-domestic*
• **Scalability:** Can add filters like year, district, arrest status
• **Task Abstraction:** Proportion, Comparison

3.2.5 Monthly crime trend



Attributes Used: Month of Date (Temporal), Count of ID (Quantitative)
Marks: 0D (Line + Trend Line)
Channels: Position (Vertical for count), Position (Horizontal for time), Line (Pattern)

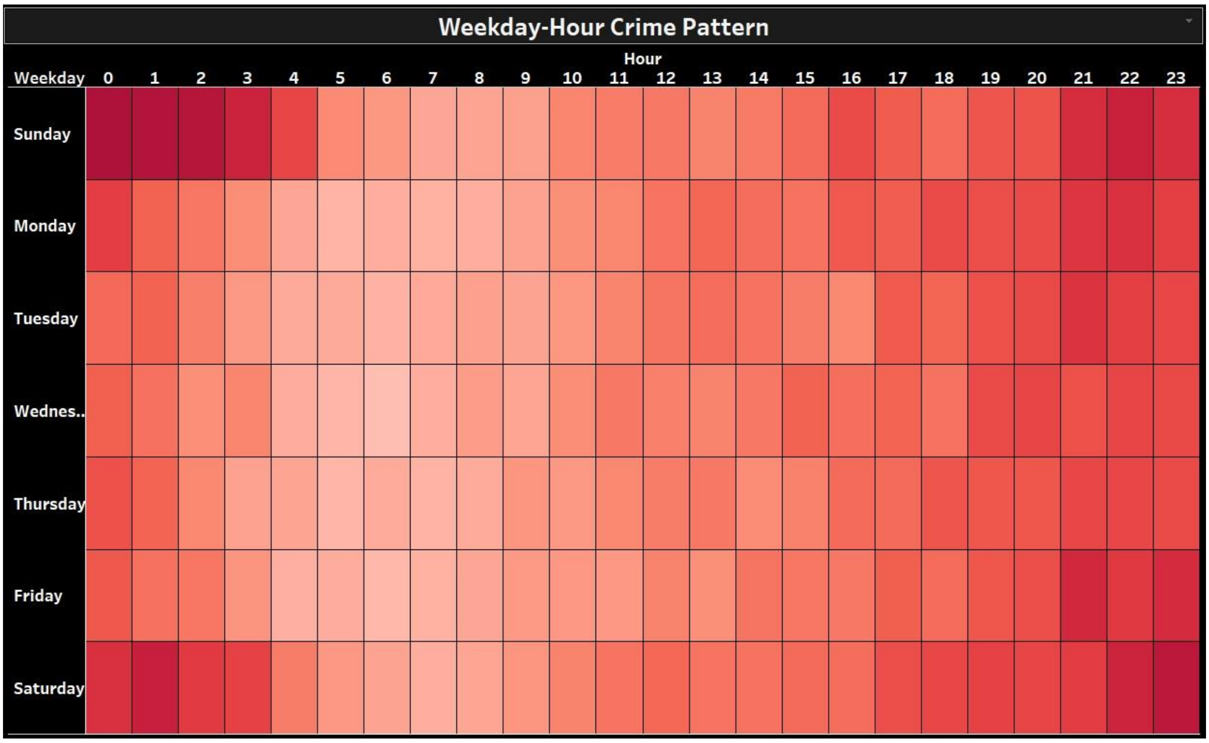
Why this Visualization Scheme?

- **Purpose:** Show trend of crimes over time
- **Key Insight:**

Crime peaked around 2001–2008**, then** gradual decline

Lowest point observed around April 2020 **(COVID effect)**
- **Scalability:** Filters like crime type, location, or arrest can be added
- **Task Abstraction:** Trend, Change Over Time, Extremes

3.2.6 Weekday-Hour Crime Pattern



Attributes Used: Weekday (Categorical), Hour (Quantitative), Count of ID (Quantitative)
Marks: 0D
Channels: Position (Horizontal & Vertical), Color (Intensity = Count)

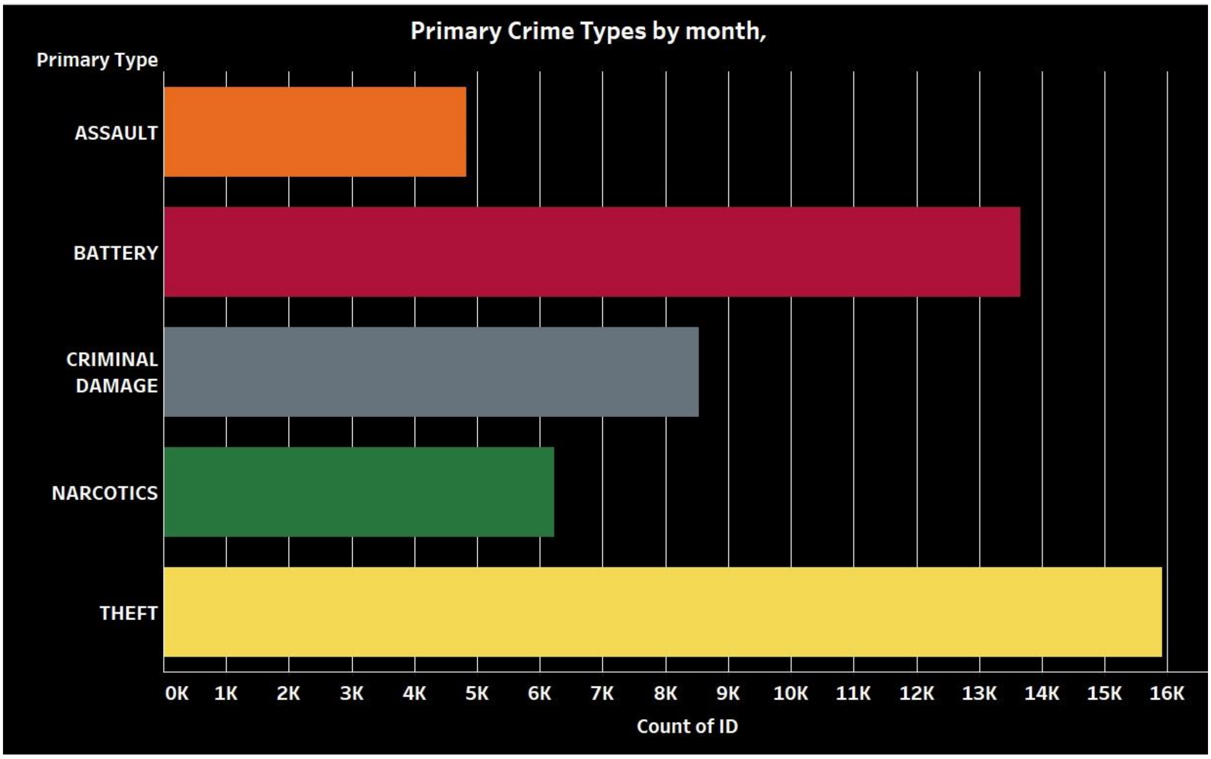
Why this Visualization Scheme?

- **Purpose:** Identify crime patterns by day and hour
- **Key Insight:**

Crimes peak between noon and midnight

Weekends (esp. Saturday night) **show higher activity**
- **Scalability:** Can filter by year, location, crime type
- **Task Abstraction:** Pattern detection, Frequency comparison

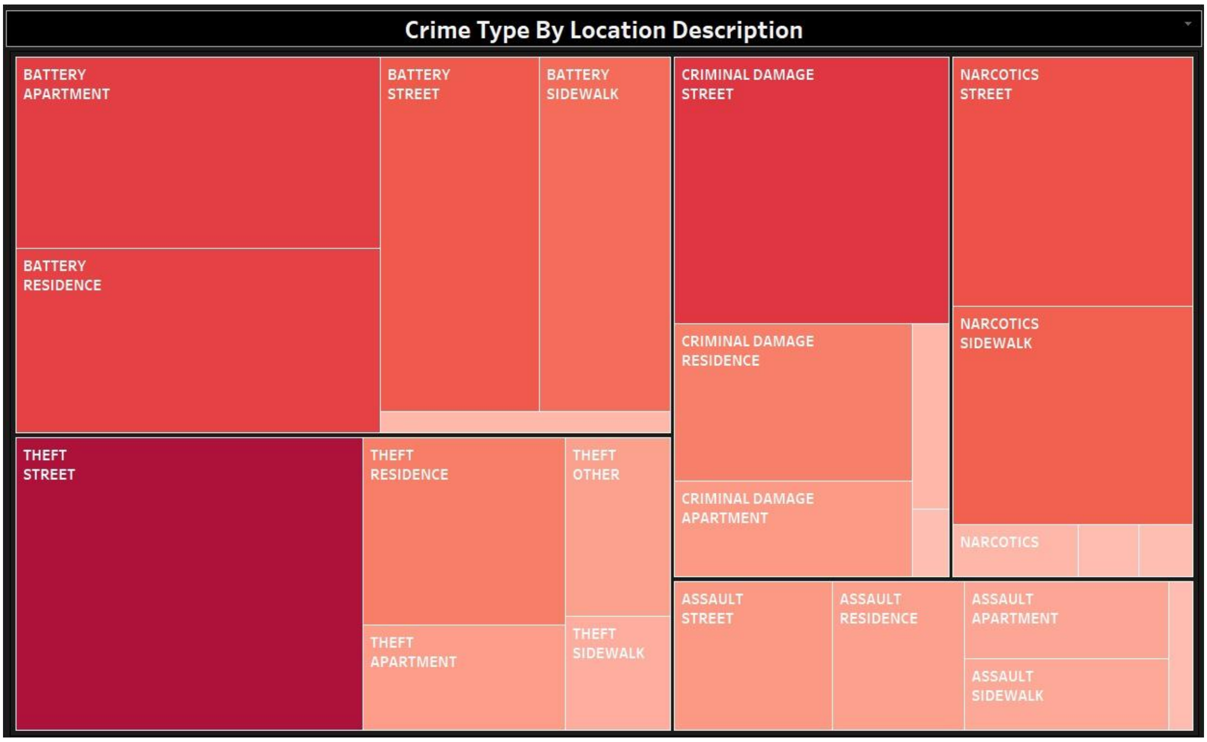
3.2.7 Primary Crime Types by Month, Day



Attributes Used: Weekday (Categorical), Hour (Quantitative), Count of ID (Quantitative)
Marks: 1D
Channels: Position (Horizontal & Vertical), Color (Intensity = Count**)**

Why this Visualization Scheme?
• **Purpose:** Identify crime patterns by day and hour
• **Key Insight:**
Crimes peak between noon and midnight
Weekends (esp. Saturday night) show higher activity
• **Scalability:** Can filter by year, location, crime type
• **Task Abstraction:** Pattern detection, Frequency comparison

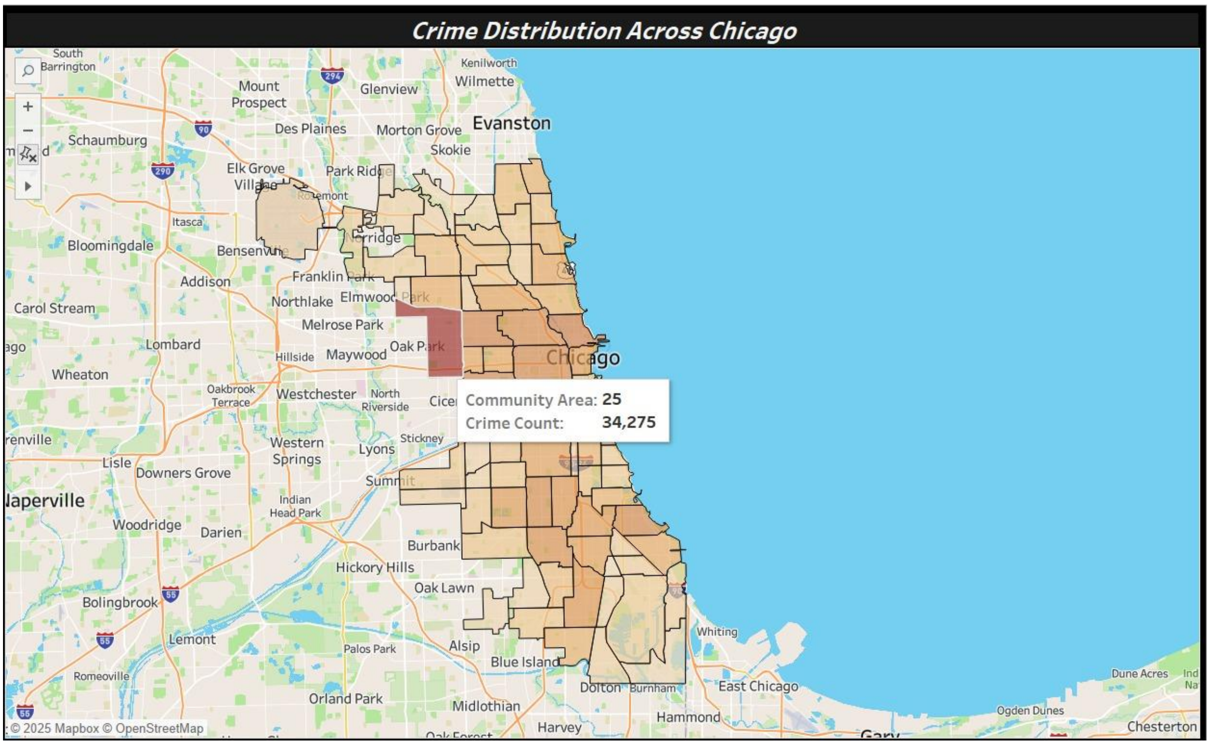
3.2.8 Crime Type by Location Description



Attributes Used: Primary Type (Categorical), Location Description (Categorical), Number of Crimes (Quantitative)
Marks: 2D (Rectangle Area)
Channels: Area (Size = Number of Crimes), Color (Intensity = Number of Crimes)

Why this Visualization Scheme?
• **Purpose:** Understand distribution of crime types across various locations
• **Key Insight:**
Battery and Theft are most frequent, especially in Apartments and on Streets
Narcotics and Criminal Damage show significant presence on Streets and Sidewalks
• **Scalability**: Can filter by year, community area, crime type
• **Task Abstraction:** Part-to-whole relationship, Frequency comparison

3.2.9 Crime Distribution Across Chicago



Attributes Used: Community Area (Categorical), Count of ID (Quantitative), Geographic Coordinates (Spatial)
Marks: 0D (Geographical area fill)
Channels: Color (Intensity = Crime Count), Position (Latitude/Longitude)

Why this Visualization Scheme?
• **Purpose:** Visualize spatial distribution of crimes across Chicago
• **Key Insight:**
Certain community areas (e.g., Area 25) have significantly higher crime counts
High crime zones cluster near city center and certain western regions
• **Scalability:** Can filter by primary crime type, year, or specific community area
• **Task Abstraction:** Geospatial pattern identification, Outlier detection

4 Conclusion

Key findings

1. **Theft is the most common type of crime** reported in the data.
2. **Even when some crimes happen a lot, arrests don’t always match**, showing a gap.
3. **Crime numbers changed over the years**, with a few years showing sudden rises or drops.
4. **Some weekdays show more crimes than others**, pointing to weekly patterns.

Solutions

1. **Made a dashboard** to show crime data clearly and in an easy-to-read format.
2. **Compare crime numbers with arrests** to check how often police take action.
3. **Helped find problem areas** in the city where police might need to focus more.
4. Designed the visuals in a way that makes it easy for police or officials to use them.

Impact

2. By showing where and when crimes happen most, this project helps police focus on high-risk areas and keep people safer.
3. Domestic crime data can guide support groups and helplines to help families and individuals in need.
4. Students, researchers, and analysts can study real-world crime data and learn how to use visualization tools.
5. Trends in past crimes can help predict and prevent future crimes with early action.