IS 507: Data, Statistical Models, & Information

Professor Yang Wang

Statistical Analysis of Chicago Crime (2019-2024)

Researchers:

Arundhati Raj (raj9)

Tejas Honmode (honmode2)

Riddhi Bhapkar (bhapkar2)

Pranav Charakondala (prc4)

Michael Khalaf (mkhala6)

Revised December 15th, 2024

Abstract
The study initiates with the selection of a dataset specific to the topic of Chicago crime for a course centered around statistical analysis. Appropriate methodologies that demonstrate mastery of those discussed within the collegiate course scope. Measures were taken to reduce the compilation time of the dataset, with variables extracted to form subsets relevant to the segmented research questions. Findings include key insights on Chicago crime statistics and provide leeway for recommendations. Lastly, the researchers open the discussion for limitations, flaws, and future work associated with the subject matter.

Introduction
The City of Chicago's residents, tourists, commuters, businesses, and stakeholders all share a space so common and rich in opportunity that they also share the same risks. It's important to know these risks. But Chicagoans wonder what kind of risks, so as this information will allow them to make better informed decisions on their travels, whereabouts, awareness, and conduct as they go about their lives.

Problem Statement
Chicago crime is a trivial element to awareness for any of the relevant stakeholders of the city itself. Statistical analysis can reveal trends and patterns in this

Main Dataset

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data

Audience & Relevance
The report introduces the scope of the issue surrounded on Chicago's diverse crime statistics and chooses to address an audience of residents, travelers, tourists, executives, businesses, and public stakeholders of Chicago assets.

Research Objective
What can be revealed through statistical analysis of crime occurring in Chicago from the years 2019 to 2024, encompassing a period before lockdown & after?

Methods & Analysis

- **Data Collection**: The primary dataset was sourced from the Chicago Data Portal, supplemented with datasets on arrests, police sentiment scores, school progress reports, and police station locations.
- **Data Cleaning**: Missing and null values were removed or transformed to maintain consistency. Outliers were handled to ensure reliability.

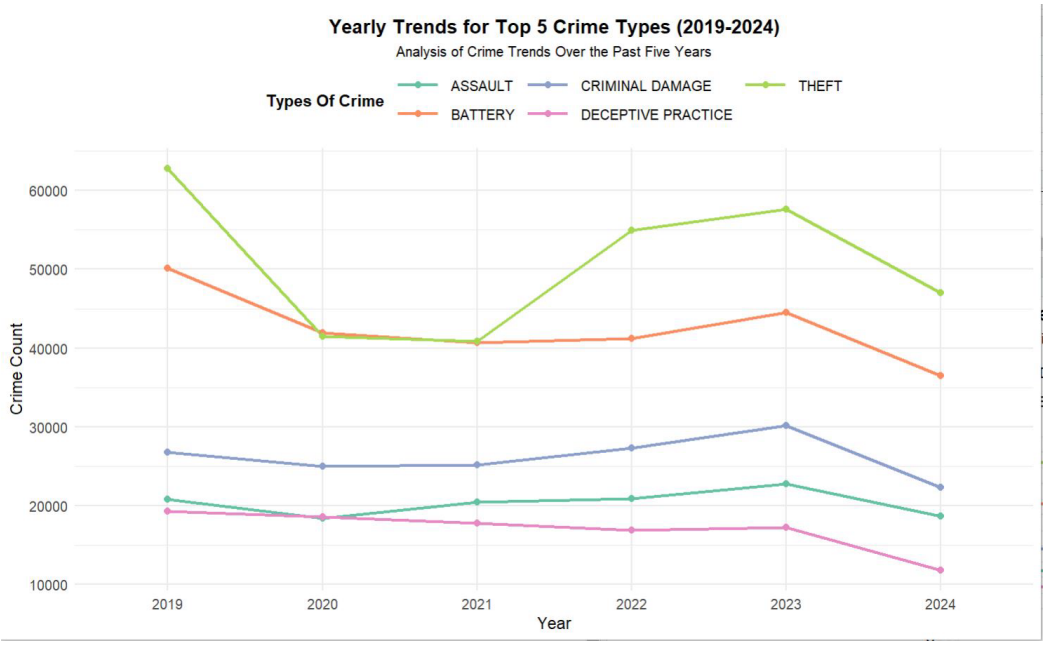
- **Exploratory Data Analysis (EDA)**: Patterns in crime types, seasonal trends, and locational influences were explored using visualizations like line graphs, heatmaps, and summary tables.
- **Time Series Analysis**: Trends in crime frequency over time were analyzed.
- **Geospatial Analysis**: Heatmaps were used to identify spatial concentrations of crimes.
- **Clustering and Correlation**: To determine relationships between socioeconomic status, location type, and crime rates

Research Questions

RQ1: *What are the trends of top crimes committed from before COVID-19 Lockdown to today (ie.2019 to 2024)?*

Methodology

This analysis examines crime trends in Chicago from 2019 to 2024, focusing on the top five crime types. Using an R script, the data was cleaned and prepared by filtering records within the specified time range (2019–2024) and converting the `date` column into a proper format for temporal analysis. A `year` column was extracted to enable annual aggregation, while crime categories were grouped by `primary.type` and summarized to identify the top five crime types. A subset focusing on these crimes was created, and trends were visualized using **ggplot2**, with a line graph emphasizing year-over-year variations through clear legends and a "Set2" color palette.



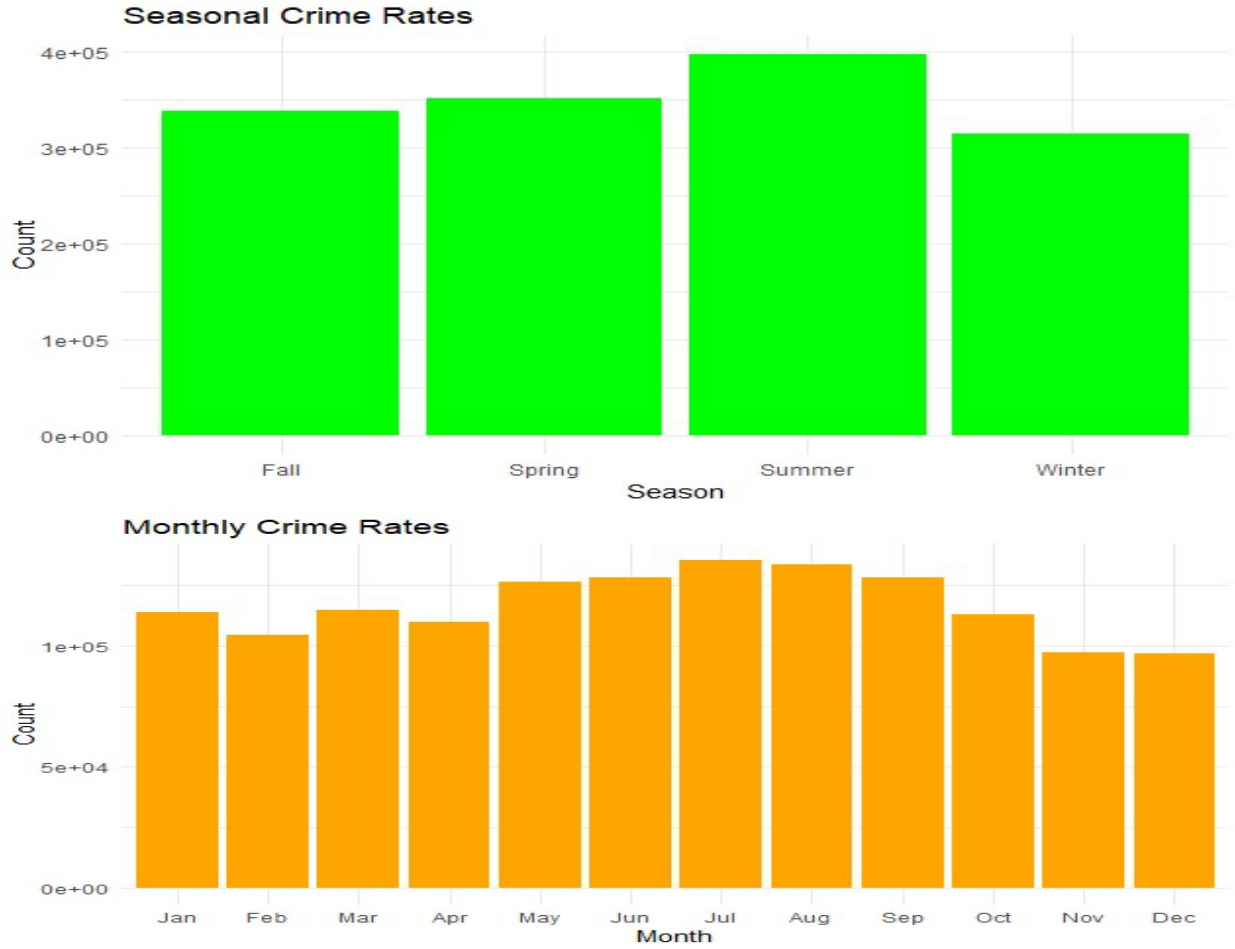
Findings

The results revealed distinct trends. **Theft** was the most prevalent crime, showing a sharp decline during the COVID-19 lockdown in 2020 but rebounding post-lockdown, peaking in 2023. **Battery** followed a similar pattern, stabilizing after an initial decline. **Assault**, **Criminal**

Damage, and **Deceptive Practices** exhibited consistent patterns, with minimal fluctuations. Notably, deceptive practices, such as fraud, did not escalate post-COVID, suggesting the shift to digital platforms did not significantly increase such crimes. These findings underscore the varying impacts of external events, such as the pandemic, on different crime categories.

RQ2: *Are there specific days, months, or seasons where certain types of crimes are more likely to occur?*
This analysis shows the seasonal/monthly variations in crime in Chicago. Descriptive data analysis and visualization techniques have been used to answer this research question.

Analysis 1:



Findings:

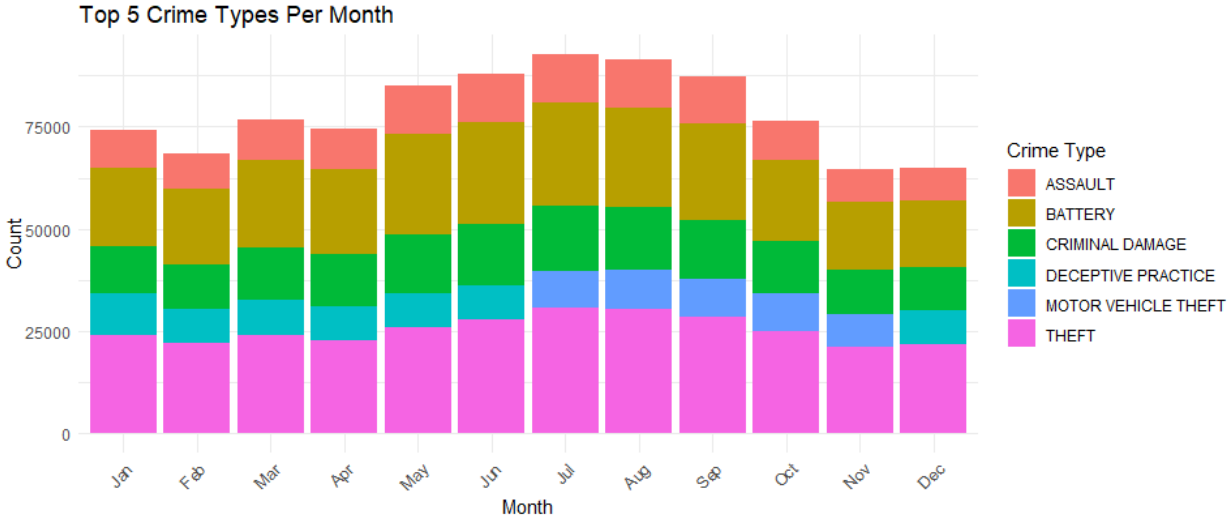
- Average number of incidents increase by about 15%-20% between the months of May and October.
- Average number of incidents increase by about 18% in summer.

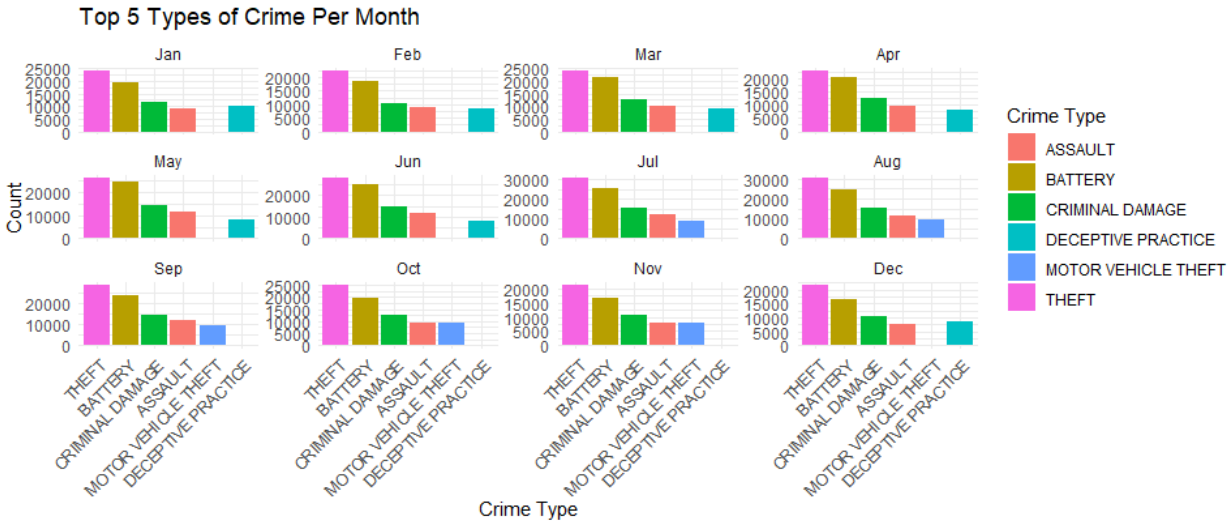
Possible Causes:

- **Weather Changes –** As weather becomes warmer, outdoor activity increases, leading to more opportunities for crimes.

- **Tourism –** Tourism increases in summer leading to more crowds, which can attract opportunistic crimes like pickpocketing and theft
- **School Breaks –** May to September/October encompasses summer vacation for schools. Younger individuals who lack supervision may participate in criminal activities.

Analysis 2:





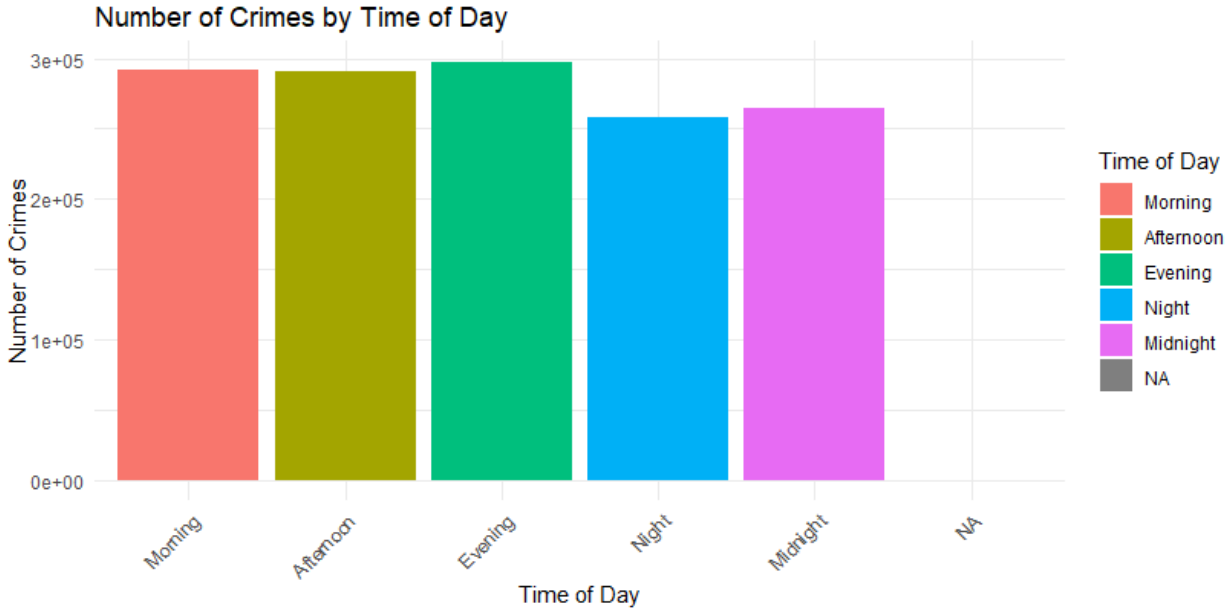
Findings:

- The top 4 crimes reported each month consistently are theft, battery, criminal damage and assault.
- Between the months of July to November, vehicle theft takes the 5th spot.

Possible Causes:

- **Increased Traveling –** People travel for holidays during summer and fall and may leave their vehicles unattended in unsecure places.
- **Juvenile Crime –** Studies have shown that individuals between the ages of 15-19 are statistically more likely to commit vehicle thefts. Summer vacations and return-to-school periods can increase the rates of such crimes.

Analysis 3:



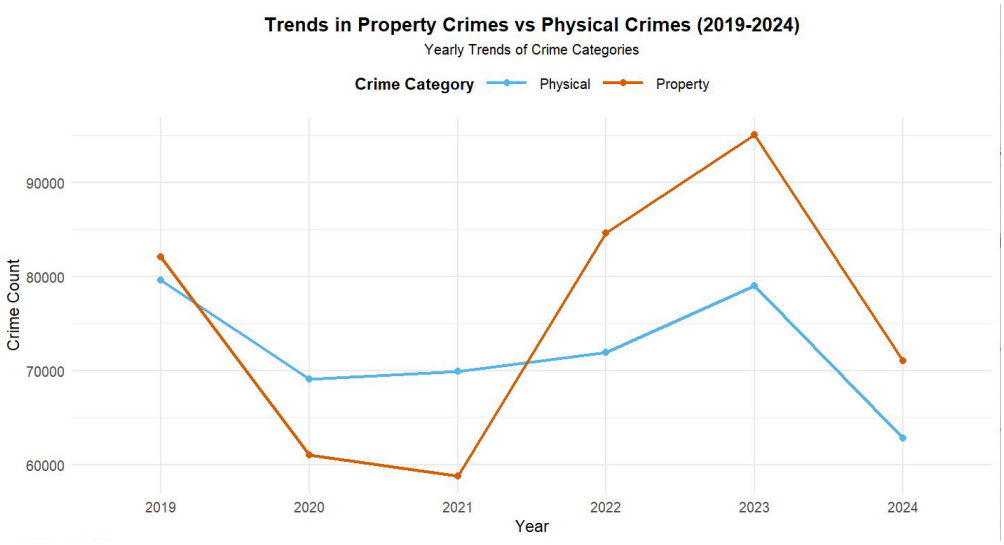
Findings:

- Almost 60% of the crimes are committed between dusk and dawn.

Possible Causes:

- **Cover of Darkness –** The lack of daylight allows criminals to hide better. The effectiveness of cameras also reduces which makes it easier to commit crimes.
- **Reduced Police Presence –** Law enforcement resources might be stretched thinner at night due to limited personnel which can lead to reduced patrolling and increase in response time to criminal incidents.

RQ3: *What are the trends in property crimes compared to trends in physical crimes?*

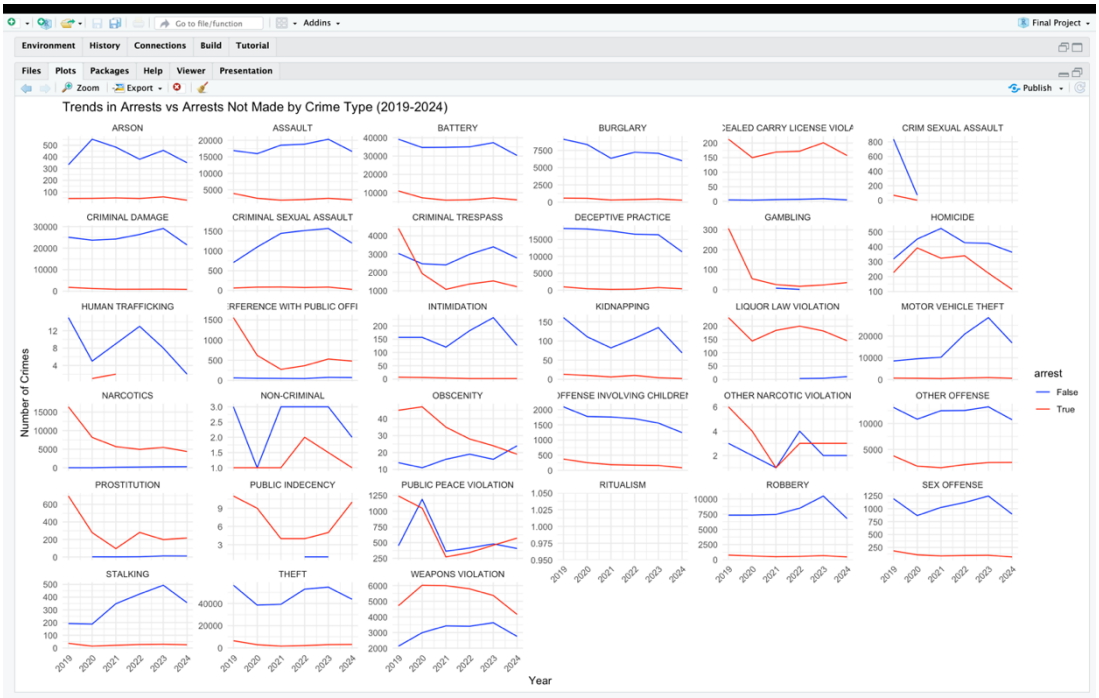


Findings
Line graph paired comparison introduces key observations: both categories progress from a steeper decline in 2019 through 2020 to an incline from 2021 to 2022. Both reach their highest in year 2023 with a decline through 2024. Property crimes may have skyrocketed after lockdown.

Methodology
Time series analysis allows for observations to be made about fluctuations in these tallied incidents within a specific window. [grep] commands were issued through a Kali Linux terminal to extract subset datasets from the master CSV list. Manual analysis and subjective grouping of these major crime categories into a group of physical crime ("BATTERY", "ASSAULT", "SEXUAL ASSAULT", "HOMICIDE", "ROBBERY") & property crime ("THEFT", "BURGLARY", "FORGERY", "ARSON", "MOTOR VEHICLE THEFT"). An R script handled dataset filtering and categorization. Dates were converted to datetime format for year-based aggregation, missing values replaced with zeroes using replace_na, and subsets merged using full_join. Visualizations were generated with ggplot2, producing a time-series line chart.

RQ 4: What are the trends in arrests made versus arrests not made across each crime category from 2019-2024?

Methodology
To analyze trends in arrests versus non-arrests for various crime types from 2019 to 2024, the dataset was loaded and cleaned. Column names were clarified, the `arrest` field was converted into a factor variable, and the `arrest_date` column was formatted as a proper date using `as.Date()`. The data was then filtered for the years 2019–2024, ensuring focus on recent trends. Using the **dplyr** package, the data was grouped by `primary_type` (crime type), year, and arrest status, with crime counts calculated using `summarise()`. Visualization was achieved using **ggplot2**, where line plots depicted trends over time, with arrests shown in blue and non-arrests in red. The `facet_wrap()` function created separate panels for each crime type, enabling clear comparisons.



Findings

The resulting graph reveals key patterns. **Theft**, the most prevalent crime, consistently shows a wide gap between arrests and non-arrests, highlighting enforcement challenges. Crimes like **Battery** and **Assault** exhibit stable trends, but non-arrests remain dominant. For crimes like **Narcotics** and **Public Peace Violations**, arrests have declined, possibly reflecting policy changes. Low-volume crimes such as **Stalking**, **Human Trafficking**, and **Sex Offenses** show inconsistent arrest rates.

RQ 5: *Which kinds of locations or settings have crimes evidently occurred more commonly?*

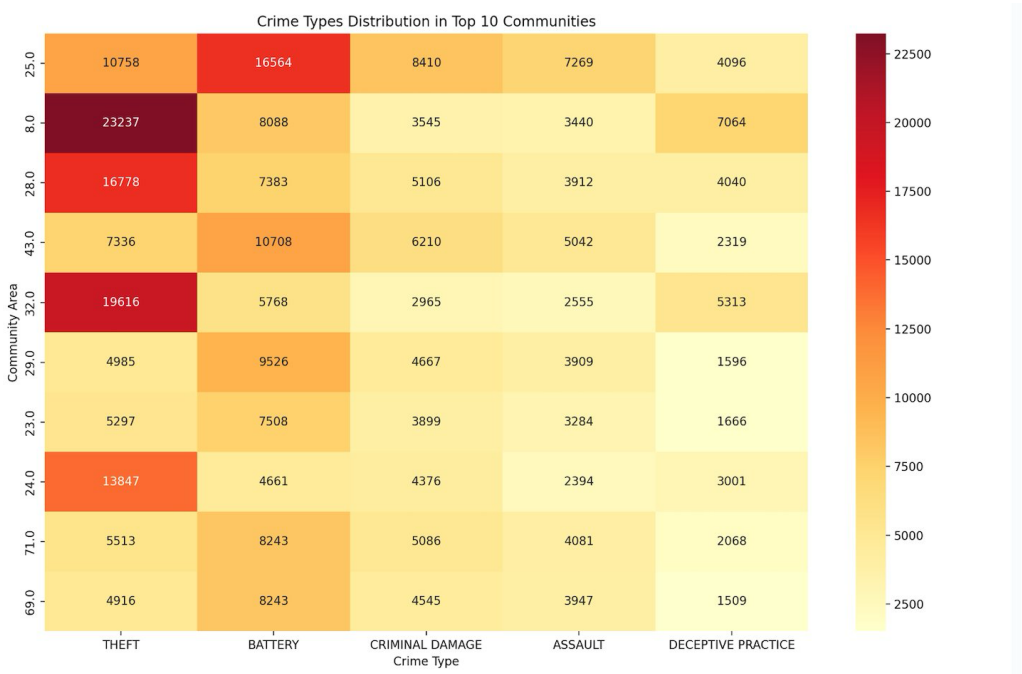
Methodology

To address this research question, we analyzed crime occurrences across different location types and identified patterns in spatial crime distributions. The dataset provided information on crime types, location descriptions, and community areas. The data preparation phase involved filtering relevant fields—Location.Description, Primary.Type, and Community.Area—and removing rows with missing or null values. Locations were grouped into meaningful categories such as "Residential Areas," "Streets," and "Public Spaces," while crime counts were aggregated using the dplyr library. A heatmap visualization method was chosen to highlight crime density patterns across the Top 10 Community Areas, with crime types on the x-axis and community areas on the y-axis. Labeled crime counts and a gradient color scale (yellow to dark red) allowed for quick identification of areas with elevated crime frequencies and clear comparisons across settings.

Findings

The findings revealed that Theft is the most frequently reported crime type, followed by Battery and Assault, particularly in public spaces and residential neighborhoods. Community Areas 8 and

32 reported the highest overall crime counts, highlighting persistent crime hotspots. Property crimes like Theft and Criminal Damage were prominent in public spaces and streets, while residential areas exhibited higher rates of domestic crimes such as Battery and Assault. These patterns suggest that location type significantly influences crime frequency and nature.



Future Scope

External factors like social and economic conditions likely influenced these crime trends. Future research could focus on ungrouped crime categories, geographic patterns, and detailed descriptions. Enhancing visualizations with additional R libraries and refining axis markers could improve clarity and reveal deeper insights.

These findings underscore disparities in arrest trends across crime types. Future work could examine correlations with socioeconomic factors, evaluate policy impacts, and incorporate predictive models to refine law enforcement strategies. Temporal analyses could help observe long-term trends, while integrating external datasets, such as socioeconomic indicators and police patrol data, could provide actionable insights to improve public safety.

Expanding on these insights, advanced time series models like ARIMA or Prophet could forecast future crime rates and guide proactive crime prevention. Geospatial tools, such as leaflet or sf, could enhance location-based analyses, helping policymakers and urban planners allocate resources more effectively. Additionally, segmenting data into pre-, during-, and post-pandemic periods could clarify the lasting effects of lockdowns and reforms. These methods would empower stakeholders to address urban safety challenges with precision and insight.