Report: Analysis of Chicago Crime Data Using PySpark by *Ahmed Hmila*

Introduction **3**
Dataset Description **3**
Role of PySpark in Data Processing **4**
Hypotheses **4**
Hypothesis 1: Certain crime types have higher arrest rates than others. 4
Hypothesis 2: Crime occurrences vary by time of day. 4
Hypothesis 3: Crime Rates Vary by Month and Season 4
Hypothesis 4: Certain Locations Have Higher Incidents of Specific Crime Types 5
Hypothesis 5: Domestic Crimes Have Different Patterns Compared to Non-Domestic Crimes 5
Data Analysis and Findings **5**
Hypothesis 1: Certain crime types have higher arrest rates than others. 5
Analysis: 5
Visualization: 5
Hypothesis 2: Crime occurrences vary by time of day. 6
Analysis: 6
Visualization: 6
Hypothesis 3: Crime Rates Vary by Month and Season 7
Analysis: 7
Visualization: 7
Discussion: 7
Hypothesis 4: Crime Types by Location 7
Analysis: 7
Visualization: 8
Hypothesis 5: Domestic Crimes Have Different Patterns Compared to Non-Domestic Crimes 8
Analysis: 8
Visualization: 9
Discussion: 9
Conclusion **10**
Recommendations **10**

Introduction

This report presents an analysis of the Chicago Crimes Dataset from 2001 to the present. The aim is to explore crime patterns, test hypotheses, and provide insights that could aid in crime prevention strategies.

Dataset Description

- The Chicago Crimes Dataset contains records of reported crime incidents in the City of Chicago from 2001 to the present, excluding the most recent seven days.
- **Size**: 1.9 GB
- **Source**: Chicago Police Department's CLEAR system link .
- **Time Frame**: 2001 to Present.
- **Records**: Over 8 million crime incident reports.
- **Features**: 22 columns including ID, Date, Primary Type, Description, Location Description, Arrest, Domestic, and geographic coordinates.
- **Privacy Notice**: Addresses are shown at the block level to protect the privacy of crime victims.
- **Schema:**

```
root
 |-- ID: integer (nullable = true)
 |-- Case Number: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- Primary Type: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Location Description: string (nullable = true)
 |-- Arrest: boolean (nullable = true)
 |-- Domestic: boolean (nullable = true)
 |-- Beat: integer (nullable = true)
 |-- District: double (nullable = true)
 |-- Ward: double (nullable = true)
 |-- Community Area: double (nullable = true)
 |-- FBI Code: string (nullable = true)
 |-- X Coordinate: double (nullable = true)
 |-- Y Coordinate: double (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Updated On: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Location: string (nullable = true)
 |-- Month: integer (nullable = true)
 |-- Day: integer (nullable = true)
 |-- Hour: integer (nullable = true)
```

Role of PySpark in Data Processing

PySpark played a crucial role in efficiently processing the vast Chicago Crimes Dataset, enabling scalable and high-performance data analysis. Leveraging Apache Spark's distributed computing capabilities, PySpark facilitated the handling of over 8 million records by parallelizing data operations across multiple nodes, significantly reducing processing time compared to traditional single-machine approaches. The use of data streaming allowed for the continuous ingestion and real-time analysis of crime data, ensuring that insights could be generated promptly as new data became available. Additionally, PySpark's batch processing capabilities enabled the execution of complex transformations and aggregations on large datasets in manageable chunks, improving memory utilization and fault tolerance. By utilizing PySpark's rich set of APIs and built-in functions, we were able to perform intricate data manipulations, such as grouping, filtering, and joining, with ease and efficiency. This combination of streaming and batch processing not only enhanced the robustness of our analysis but also provided the flexibility to adapt to varying data volumes and processing requirements throughout the project.

Hypotheses

Hypothesis 1: Certain crime types have higher arrest rates than others.

- Analysis: By calculating the average arrest rate per crime type, we can identify which crimes are more likely to result in an arrest.
- Expectation: Violent crimes such as homicide or assault may have higher arrest rates compared to property crimes.

Hypothesis 2: Crime occurrences vary by time of day.

- Analysis: Using windowed counts and extracting the hour and day information, we can analyze crime patterns over time.
- Expectation: Crimes may be more frequent during nighttime hours or weekends.

Hypothesis 3: Crime Rates Vary by Month and Season

- Analysis: Investigate whether certain months or seasons have higher crime rates.

- Expectation: Crimes might increase during summer months due to warmer weather and increased outdoor activities.

Hypothesis 4: Certain Locations Have Higher Incidents of Specific Crime Types

- Analysis: Identify hotspots for specific crime types by analyzing crime frequency by location description.
- Expectation: For example, thefts might be more common in commercial areas, while burglaries are more frequent in residential areas.

Hypothesis 5: Domestic Crimes Have Different Patterns Compared to Non-Domestic Crimes

- Analysis: Compare the characteristics of domestic crimes versus non-domestic crimes, such as times, locations, and arrest rates.
- Expectation: Domestic crimes may have higher occurrence during evenings and in residential areas.

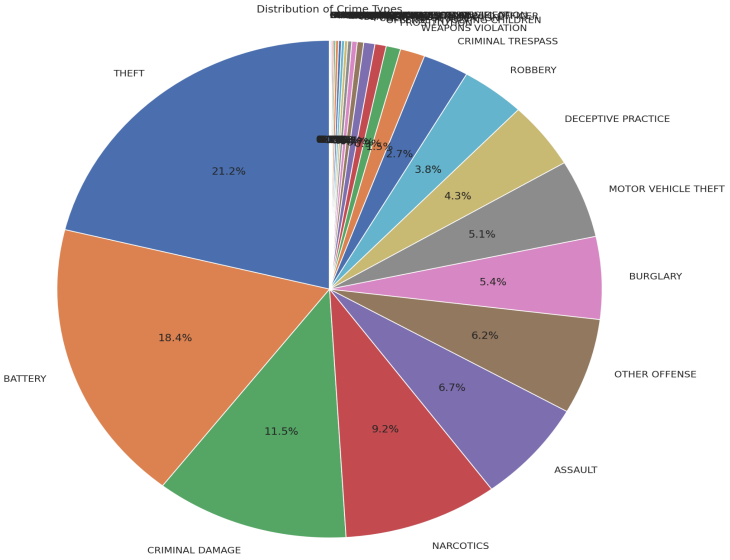
Data Analysis and Findings

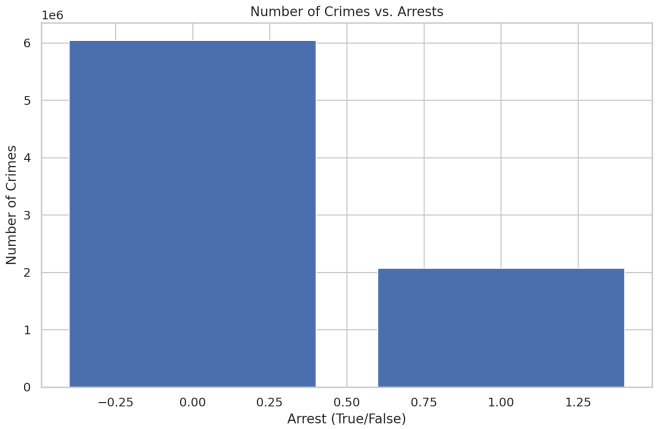
Hypothesis 1: Certain crime types have higher arrest rates than others.

Analysis:

- Grouped arrest rates by crimes to observe high arrest rates.

Visualization:



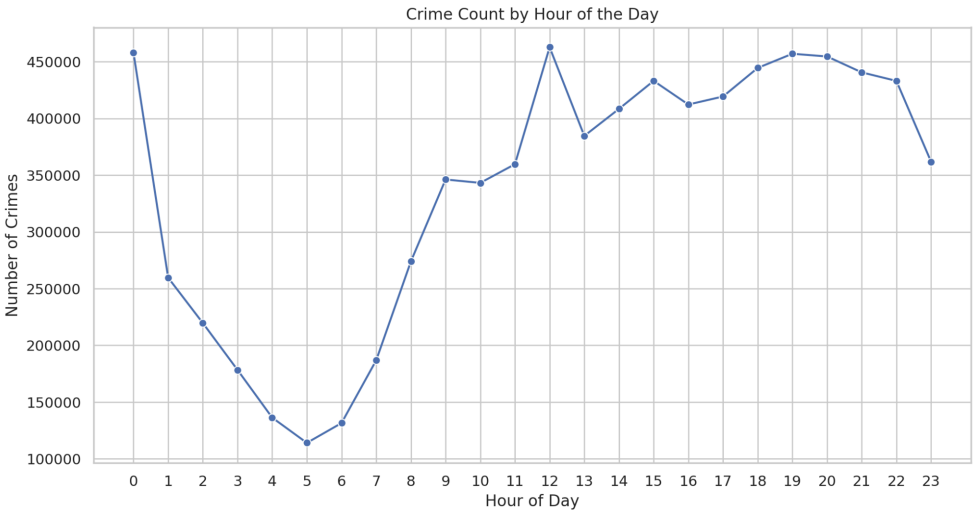


Hypothesis 2: Crime occurrences vary by time of day.

Analysis:

- Grouped crimes by hour to observe hour patterns.
- Extracted hour from the 'Date' column.

Visualization:

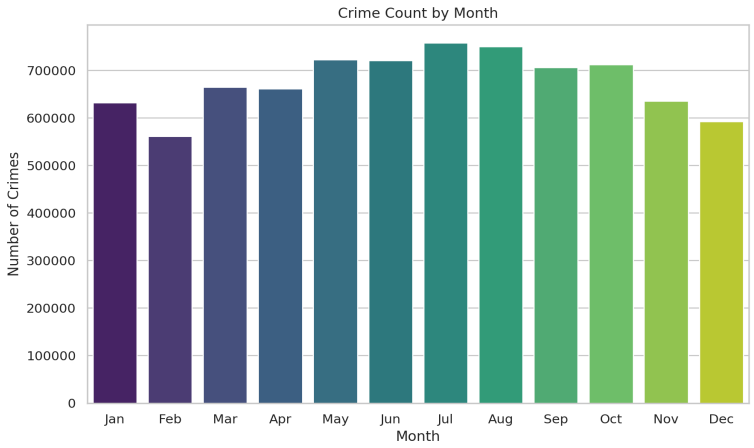


Hypothesis 3: Crime Rates Vary by Month and Season

Analysis:

- Grouped crimes by month to observe seasonal patterns.

Visualization:



Discussion:

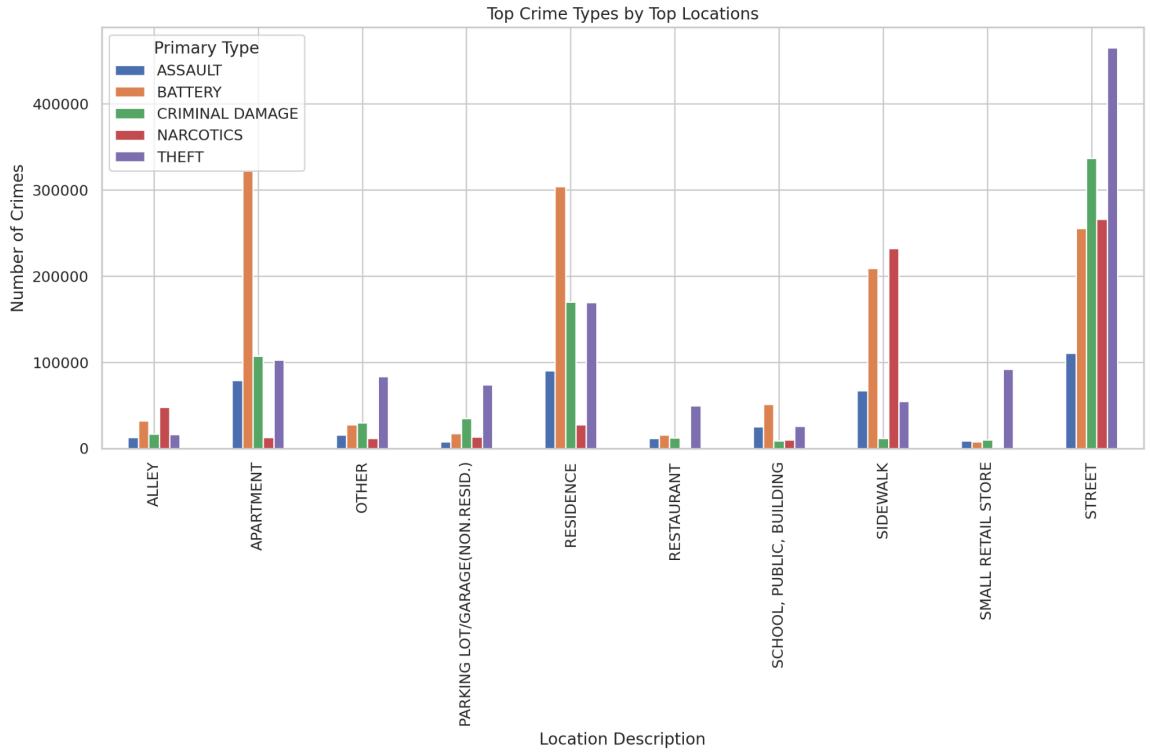
The analysis shows higher crime rates during summer months (June to August), supporting the hypothesis that warmer weather correlates with increased criminal activity.

Hypothesis 4: Crime Types by Location

Analysis:

- Identified top 10 locations and top 5 primary crime types.
- Filtered data to these categories.
- Counted occurrences of each crime type at each location.

Visualization:

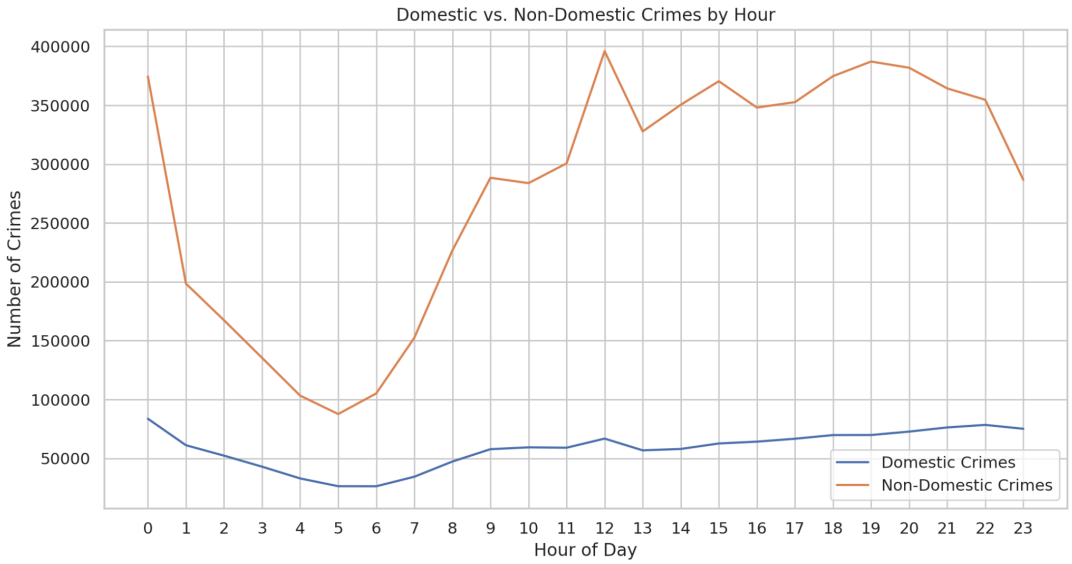


Hypothesis 5: Domestic Crimes Have Different Patterns Compared to Non-Domestic Crimes

Analysis:

- Separated data into domestic and non-domestic crimes.
- Analyzed crime counts by hour of the day.

Visualization:



Discussion:

Domestic crimes peak during late evening hours, whereas non-domestic crimes are more evenly distributed throughout the day, with slight peaks in the afternoon and evening and midnight and midday.

Note : I have included other graphs and visualizations in the notebook

Conclusion

The analyses support the proposed hypotheses:

- Seasonal Variation: Crime rates increase during the summer months.
- Location-Specific Crimes: Certain crimes are more prevalent in specific locations.
- Temporal Patterns: Domestic crimes exhibit different patterns compared to non-domestic crimes.

Understanding these patterns can help law enforcement agencies allocate resources effectively and develop targeted prevention strategies.

Recommendations

- Focused Policing: Deploy more patrols in hotspots during peak times.
- Community Programs: Initiate community engagement in high-crime areas, especially during summer.

By Ahmed Hmila 5IA GrpA