Predicting Crime Hotspots for Public Safety Using Big Data Analytics

Big Data Analytics

By

Ibrahim Mehmood
(2021748)

Zain Nofal
(2021723)

FCSE
GHULAM ISHAQ KHAN INSTITUTE

Abstract

Crime continues to be a major issue in both urban and rural settings, affecting public safety, economic development, and overall quality of life. Existing crime prevention approaches are mostly reactive, typically addressing incidents after they occur instead of actively reducing potential threats. This project intends to use Big Data technologies, predictive analytics, and spatial analysis to identify potential crime hotspots and offer practical insights for law enforcement and policy makers.

Dataset Overview

The dataset, sourced from the Chicago Police Department, contains detailed records of reported crimes. Key attributes include:

- **Crime Details:** Primary Type, Description, Arrest, and Domestic.
- **Temporal Data:** Date, Year, and Month.
- **Geographical Data:** Community Area, Location Description, Latitude, and Longitude.

```
 |-- ID: integer (nullable = true)
 |-- Case Number: string (nullable = true)
 |-- Date: string (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: integer (nullable = true)
 |-- Ward: integer (nullable = true)
 |-- Community Area: integer (nullable = true)
 |-- FBI Code: string (nullable = true)
 |-- X Coordinate: integer (nullable = true)
 |-- Y Coordinate: integer (nullable = true)
```

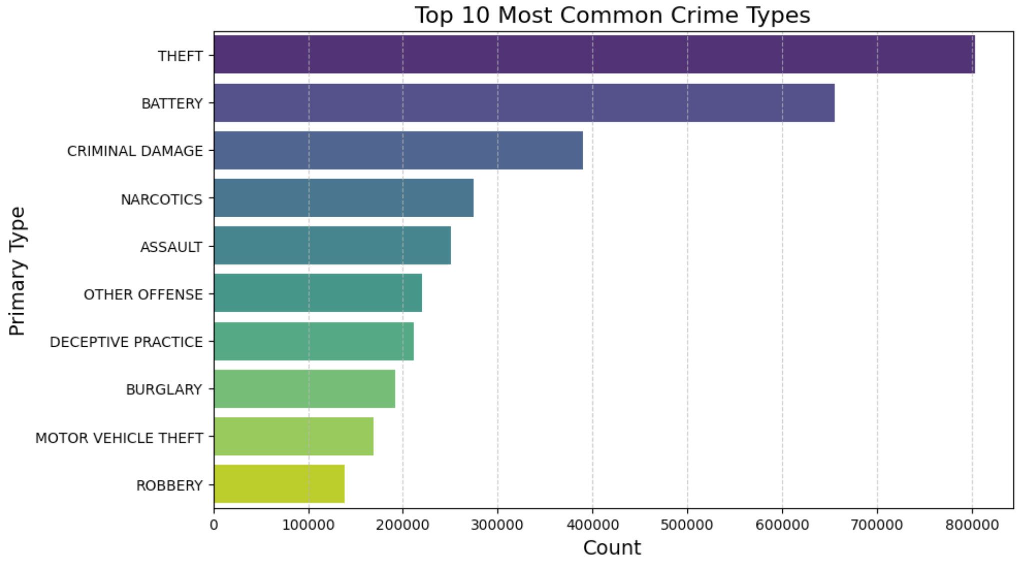
```
|-- Year: integer (nullable = true)
|-- Updated On: string (nullable = true)
|-- Latitude: double (nullable = true)
|-- Longitude: double (nullable = true)
|-- Location: string (nullable = true)
```

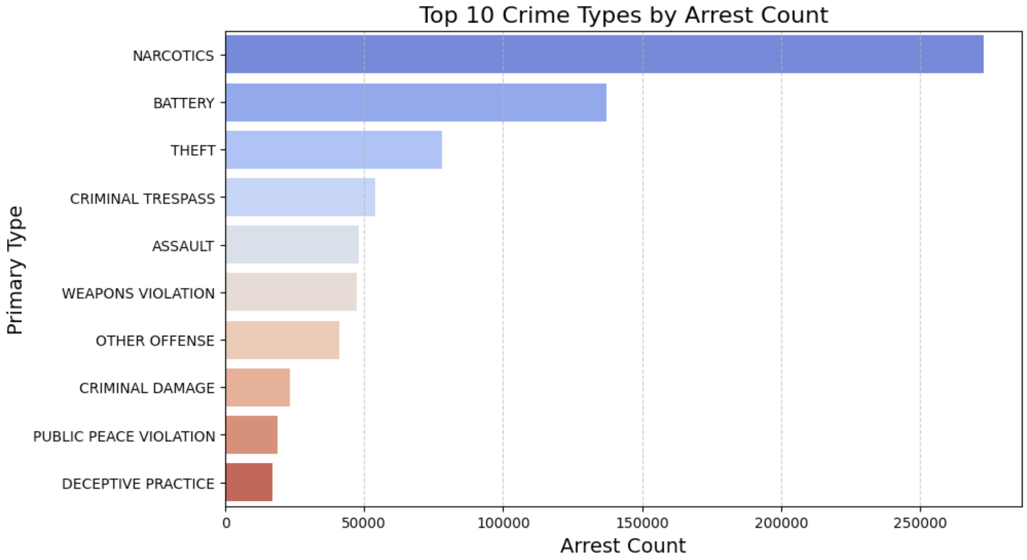
Data Preprocessing

1. **Data Loading:**
   - Combined multiple CSV files into a single Spark DataFrame.
   - The Total Rows: 7,870,291, Total Columns: 22.
2. **Missing Value Analysis:**
   - **Case Number:** 2 missing values, probably because of incomplete recordkeeping or data entry errors.
   - **Location Description:** 9,638 missing values, probably because that location details were not recorded for these crimes.
   - **Geographical Data(X Coordinate, Y Coordinate, Latitude, Longitude, Location):** Each of these columns has 35,442 missing values, probably because the geographical details are not available for these records.
   - **Administrative Columns (District, Ward, Community Area):** Missing values in District (1), Ward (96), and Community Area (397) because records lack complete administrative details.
   - **Columns with No Missing Values:** Columns like ID, Date, Primary Type, Arrest, Domestic, and Year are fully populated which is useful.
3. **Data Cleaning:**
   - Filtered rows with null or invalid entries in critical columns (e.g., Date, Arrest).
   - Converted Date column to proper timestamp format.
4. **Derived Features:**
   - Extracted Year, Month, and Day of the Week from Date.

Data Analysis

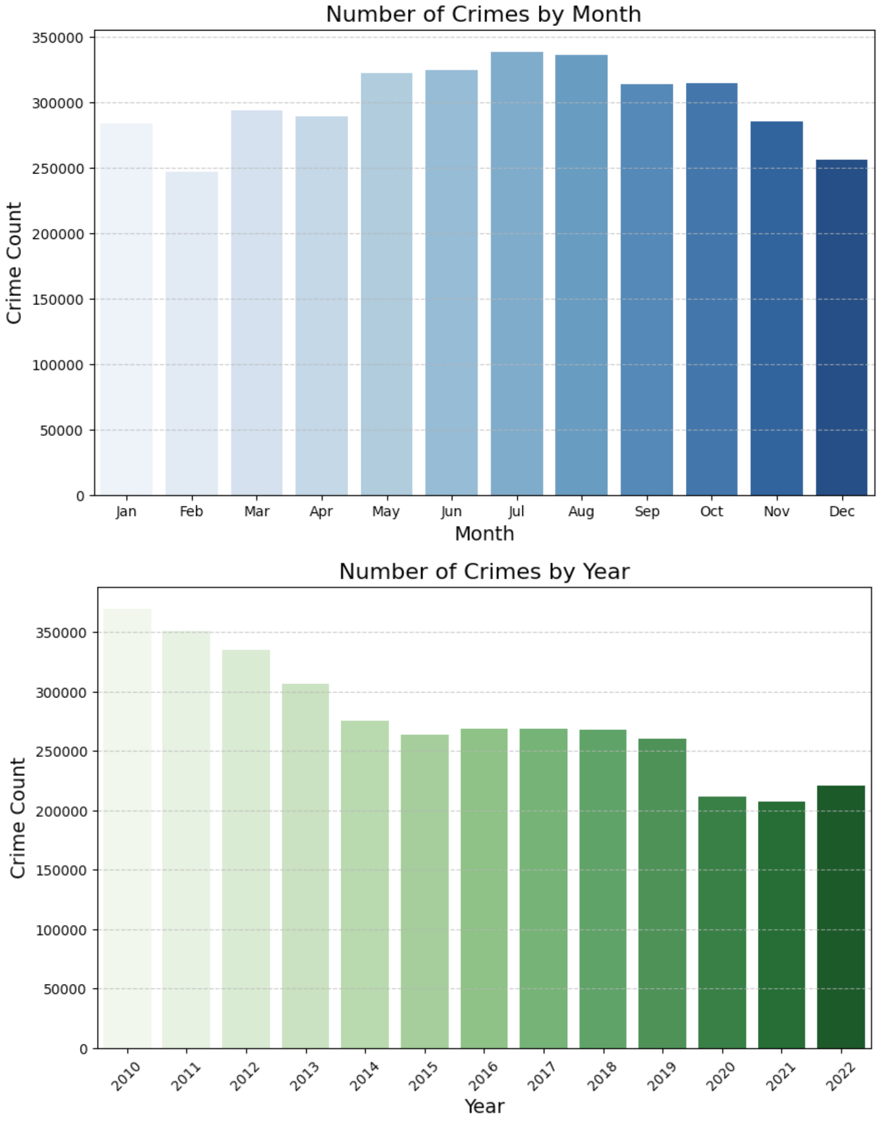
- **Crime Type Distribution:** The dataset is dominated by crimes like Theft (803,057 incidents) and Battery (655,296 incidents), as well as Criminal Damage, Narcotics, and Assault, highlighting these as the most common and prevalent crime categories.
- **Arrest Counts:** Arrest rates vary widely by crime type. For instance:
  - Public Peace Violations have a relatively high number of arrests (18,726).
  - Offenses Involving Children (4,667) and Stalking (340) show lower absolute arrests, probably because of different law enforcement approaches or reporting practices.





Temporal Analysis:

Number of Crimes by Time:



Peak Crime Months:

- July (338,908) and August (336,115) have the highest crime counts. This could be due to higher outdoor activities and gatherings outside.
- June (324,940) and May (322,384) also have high crime levels.

Low Crime Months:

- February (247,174) has the lowest crime count, possibly due to lower outdoor activity (because of cold weather) and fewer days in the month.

- December (256,573) also has lower crime counts, possibly due to holiday seasons and colder weather.

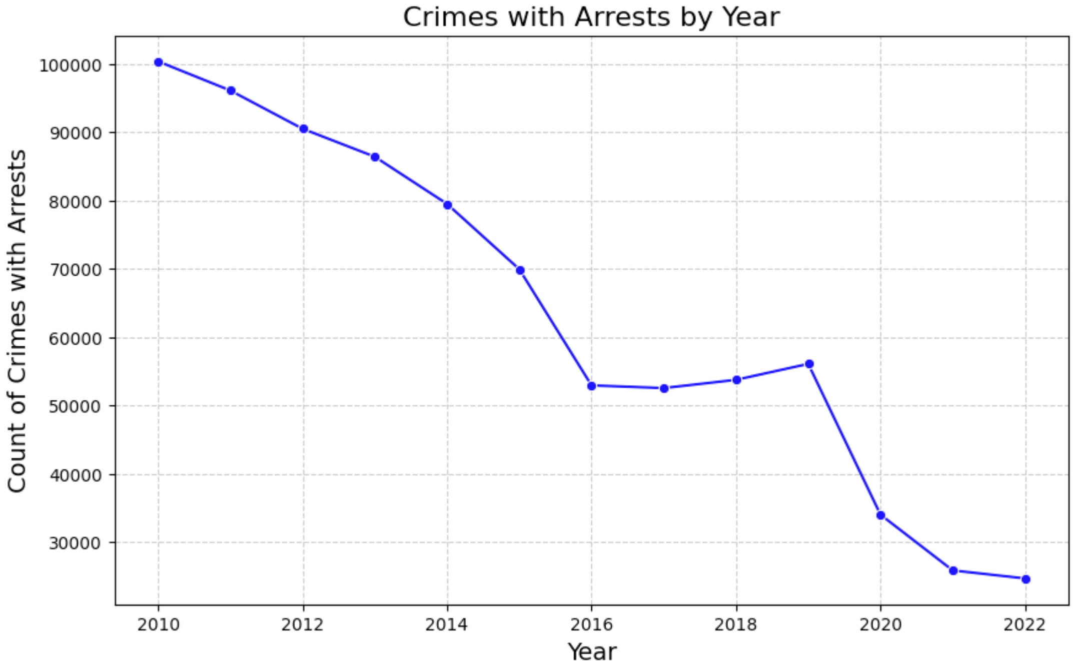
Seasonality:

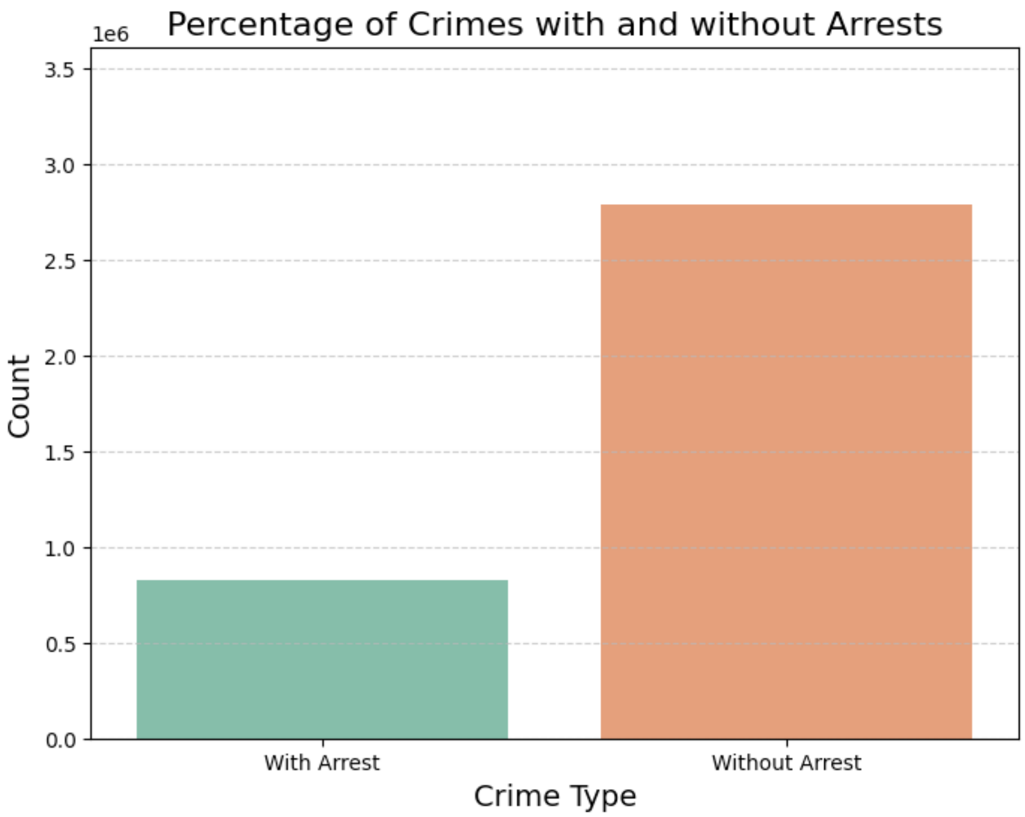
- There is a clear seasonal pattern with crime rates peaking in summer (June–August) and dropping in winter (November–February).

Other Observations:

- Crimes have shown a consistent decline over the years, from 369,664 in 2010 to 207,430 in 2021. This reflects improvement in law enforcement and societal changes.
- 2020 (211,426) has a significantly lower count, likely due to the COVID-19 pandemic, which reduced public activity and opportunities for crime.

Arrest Trends over Time:





22.79% of all reported crimes resulted in arrests, suggesting that only about 1 in 5 crimes led to an arrest. This indicates significant room for improvement in resolving reported incidents.

- In 2010, arrests were highest at 100,370, matching with the overall higher crime rates in earlier years.
- Arrest counts declined over the years.:
    - From 2015 (69,926) to 2021 (25,810), arrests dropped by 63%.
- Lowest Arrests:
    - In 2022, arrests hit a new low of 24,639, indicating a decline post-pandemic.

Spatial Analysis:

Most Crime-Prone Areas:

- Community Area 25 (Austin) tops the list with 222,537 crimes, indicating it as a significant hotspot.

- Other highly crime-prone areas include 8 (Near North Side), 43 (South Shore), 29 (West Town), and 28 (Englewood). These neighborhoods probably have a high population density or face socioeconomic challenges.

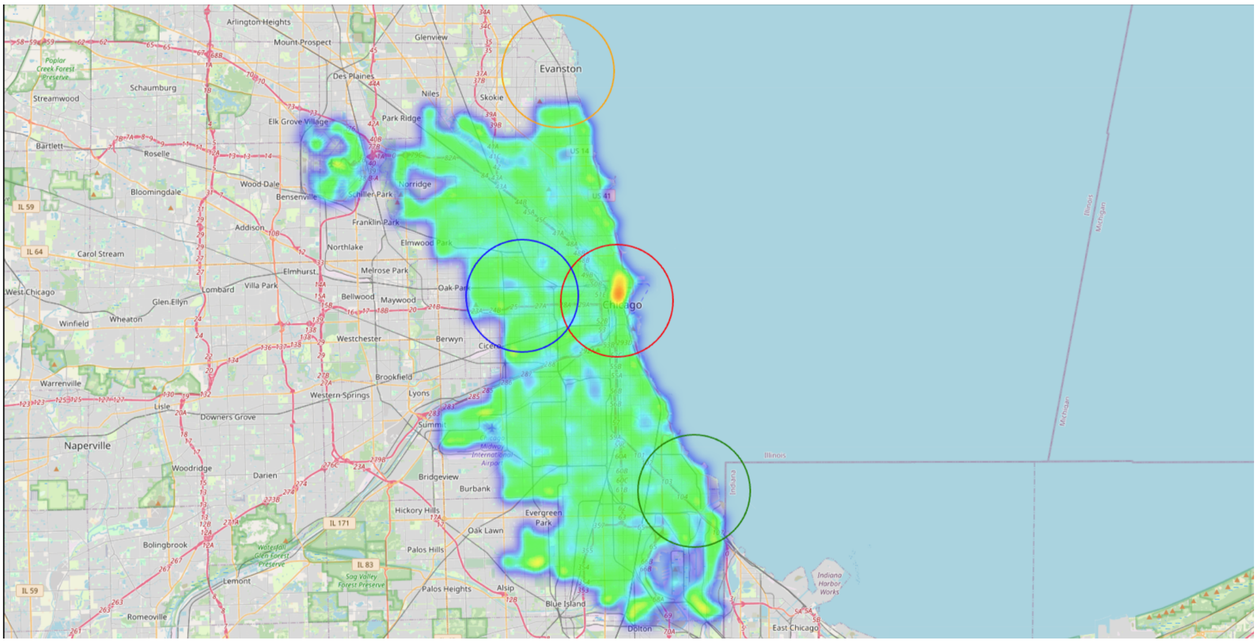
Geographical Insights:

- The majority of crimes occur on streets (844,413), consistent with public offenses like theft, assaults, and vandalism.
- Residences (586,435) and apartments (494,985) together indicate notable instances of domestic or localized crimes, including burglaries and domestic violence.

Outdoor vs. Indoor Crimes:

- Outdoor activity-related crimes are highlighted by locations such as sidewalks (344,524), parking lots (86,373), and alleys (77,288).
- Indoor or confined spaces like small retail shops (79,079) and restaurants (68,653) also add to crime statistics, indicating theft or other issues in commercial zones.

Hotspot Detection:



High Concentration Areas:

- **Red**: Downtown Chicago (Loop, Near North Side, West Loop): The heatmap indicates a significant concentration of crime in downtown Chicago, particularly in the Loop, Near North Side, and West Loop. This is anticipated because of the higher population density, presence of businesses, entertainment venues, and substantial pedestrian activity.
- **Blue**: Garfield Park and Surrounding Areas: West Garfield Park and East Garfield Park show significant levels of crime, probably associated with socioeconomic issues and longstanding crime patterns.
- **Green**: South Chicago and Southeast Regions: The heatmap also highlights hotspots extending to areas like Englewood and surrounding neighborhoods, where crime rates have been historically high.
- **Orange**: Northern Suburban Areas: Areas situated to the north of Jefferson Park and Skokie seem to have fewer crime incidents, indicating a more subdued gradient. This is consistent with the lower population densities and the residential nature of the suburban regions.
- **Crime Spread Along Major Roads and Parks:** Visible alignment of hotspots along major roads, highways, and public spaces like Garfield Park and Humboldt Park suggests that crime incidents tend to cluster along high-traffic corridors.

Advanced Analysis – Clustering Hotspots:

To identify geographical crime hotspots, K-Means clustering was applied to the dataset using the latitude and longitude attributes. This unsupervised learning technique groups geographical data points into clusters, allowing for the identification of areas with high crime densities.

Implementation Steps:

1. **Feature Preparation:**
   A VectorAssembler was used to combine Latitude and Longitude columns into a single feature vector required by the clustering algorithm.
2. **Applying K-Means Clustering:**
   The K-Means algorithm was configured with k=5 clusters (adjustable based on the analysis) and trained on the prepared feature set. The algorithm grouped data points into five geographical clusters.

3. **Cluster Centers:**

   The geographical centers of the clusters were calculated and retrieved for interpretation.

4. **Adding Cluster Predictions:**

   The cluster assignments were added as a new column in the DataFrame, enabling the visualization of crime density across different clusters.

```
+------------+-------------+----------+
|    Latitude|    Longitude|prediction|
+------------+-------------+----------+
|41.915306069|-87.686639247|         2|
|41.728192429|-87.600985433|         0|
| 41.81227369|-87.748176594|         3|
|41.940221932|-87.669039008|         2|
|41.762066981|-87.699077348|         3|
|41.763263853|-87.596998313|         0|
|41.940892158|-87.654138616|         2|
|41.691818181|-87.678096311|         3|
|41.917921699|-87.758418952|         1|
|41.865187198|-87.660234403|         4|
```

Conclusion:

This project highlights the potential of big data analytics in enhancing public safety. Predicting crime hotspots can help law enforcement agencies proactively address safety concerns and improve community trust.