Northeastern University
College of Professional Studies

Final Project Draft Report

Srushti Patil and Mohit Premani

Northeastern University

Applied Machine Intelligence

ALY 6110 Data Management and Big Data

Professor: Daya Rudramoorthi

Date: June 22, 2025

Introduction

In today's metropolitan surroundings, understanding and responding to criminal activity trends is crucial for effective law enforcement and public safety. The difficulties of allocating resources, locating hotspots, and promptly responding to incidents increase with the size of cities. In this regard, using big data has become more and more important in order to glean useful insights from the enormous volumes of crime-related data that are produced each.

The goal of this project is to demonstrate how advanced data analytics and clustering techniques can be applied to a real-world dataset specifically, the **Crime_Data_from_2020_to_Present.csv** dataset, to identify spatial and temporal crime hotspots. By doing so, this project aims to answer a critical business question: *"How can we utilize spatial and temporal crime data to identify hotspots and optimize policing efforts?"*

This dataset makes it possible to thoroughly examine the spatial and temporal features of crime because it includes latitude, longitude, crime description, region names, and timestamps. The CRISP-DM methodology is used in this project, which first investigates and comprehends the dataset before cleaning and preparing it for analysis, applying clustering techniques, and assessing the outcomes to derive useful insights.

In addition to highlighting the locations and hours when crimes are most common, this method offers a reproducible process for data-driven decision-making. The results can help local officials and police agencies, among other stakeholders, better allocate resources, speed up reaction times, and improve public safety in the long run.

Data Understanding

The Crime_Data_from_2020_to_Present.csv file, which includes roughly one million records of recorded crimes throughout Los Angeles between 2020 and the present, served as the source of the dataset used in this analysis. A number of crucial elements are included in the data, including LAT (latitude), LON (longitude), DATE OCC (date of occurrence), TIME OCC (time of occurrence), Crm Cd Desc (crime description), and Vict Age and Vict Sex (victim demographics). These fields are essential for both geographical and temporal studies.

The dataset was carefully examined for completeness and quality in the early exploratory data analysis (EDA). To guarantee the accuracy of the spatial clustering results, rows with missing or null values in crucial variables like LAT, LON, DATE OCC, and Crm Cd Desc were eliminated. To extract temporal information like Hour, Month, and Weekday, the DATE OCC column was transformed into a datetime format. The analysis of temporal and spatial trends in the crime data was made possible by these derived features.

A random sample of 20,000 records was chosen for clustering due to the dataset's enormous size (around 1 million records) and technological limitations (such memory availability and processing time in a cloud environment). This sample allowed for efficient processing and modeling within a reasonable scope and was indicative of the larger dataset.

Two clustering methods were selected for this project:

1. **KMeans** (a partition-based clustering technique), ideal for quick, efficient clustering when the number of clusters is known.
2. **Agglomerative Clustering** (a hierarchical clustering technique), suitable for spatial data due to its ability to consider hierarchical relationships between clusters.

Each technique was applied to the sample, yielding the following results:

- **KMeans** achieved a silhouette score of **0.4506** across **6 clusters**, with the largest cluster containing **5,646 points** and the smallest **42 points**.
- **Agglomerative Clustering** performed slightly better, achieving a silhouette score of **0.4777** across the same **6 clusters**, with a largest cluster of **8,808 points** and a smallest cluster of **42 points**.

As a result of these studies, the dataset showed clear physical hotspots for crime around the city, along with temporal trends that can help identify the most likely times for crimes to occur. The clustering results were further validated by the visualizations produced, which showed the relative effectiveness of the hierarchical technique for geographical data. These included bar charts comparing silhouette scores across both methods and side-by-side scatter plots of clusters by latitude and longitude.

These first findings demonstrate the need of using big data and clustering approaches to comprehend intricate urban crime dynamics and lay the groundwork for more thorough modeling.

Data Preparation

The dataset underwent a rigorous preparation process to ensure its quality and suitability for clustering. After the initial data understanding phase, we performed several cleaning and preprocessing steps:

Data cleaning

To avoid distorting the clustering results, all rows with missing or null values in important fields such as LAT, LON, DATE OCC, and Crm Cd Desc were eliminated. In order to extract temporal

elements like Hour, Month, and Weekday, the DATE OCC column was parsed and transformed into a 'Datetime' object. The data's temporal understanding was enhanced by these generated fields.

Feature Selection

Since the primary focus of this project was to understand spatial hotspots of crimes, the **LAT** and **LON** columns were selected as the primary features for clustering. This approach aligned with the goal of identifying geographic patterns in the dataset.

Data Sampling for Scalability

Given the dataset's large size (~1 million records), a **random sample of 20,000 observations** was extracted for the clustering process. This subsampling was performed to balance computational efficiency with statistical representation, making the analysis feasible in a constrained cloud environment (such as Google Colab).

Future Scaling

The scikit-learn library's StandardScaler was used to standardize spatial data and prevent latitude and longitude scales from impacting clustering findings. Latitude and longitude both contribute equally to the clustering process thanks to this standardization.

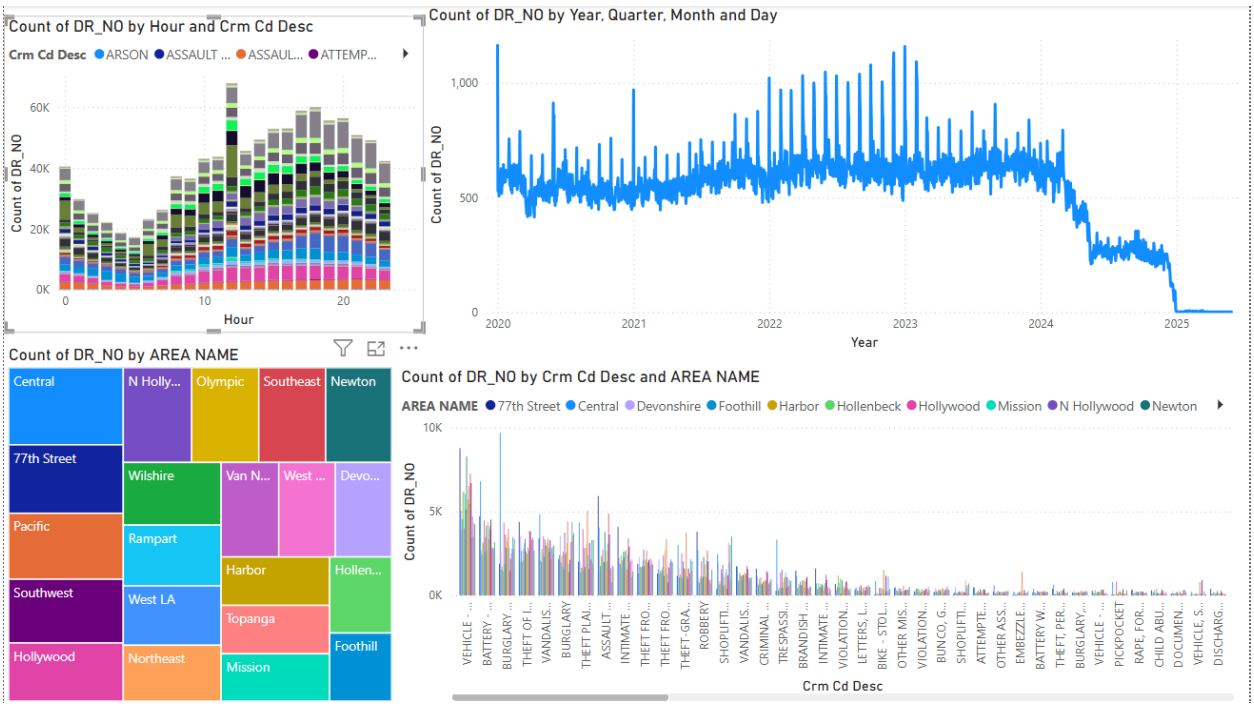
Final Prepared Dataset

The final dataset included a balanced, representative sample of 20,000 records following cleaning, filtering, feature scaling, and temporal feature extraction. This dataset provided a strong basis for identifying regional patterns and hotspots in the crime data and was used as input for both KMeans and Agglomerative clustering algorithms.

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Loaded full dataset: (1005091, 28)

--- KMeans Clustering on 20K sample ---
KMeans_Cluster
0    5375
1      42
2    5646
3     818
4    5173
5    2946
Name: count, dtype: int64
Silhouette Score (KMeans): 0.4506

--- Agglomerative Clustering on 20K sample ---
Agglo_Cluster
0    8808
1    4466
2    3324
3      42
4     876
5    2484
Name: count, dtype: int64
Silhouette Score (Agglomerative): 0.4777
```

Dashboard Insights

The dashboards developed as part of this project offer a full, multidimensional perspective of the spatial and temporal distribution of criminal occurrences. When combined, they provide a more thorough understanding of the times and locations of crimes, empowering stakeholders to decide on intervention and resource allocation with knowledge.

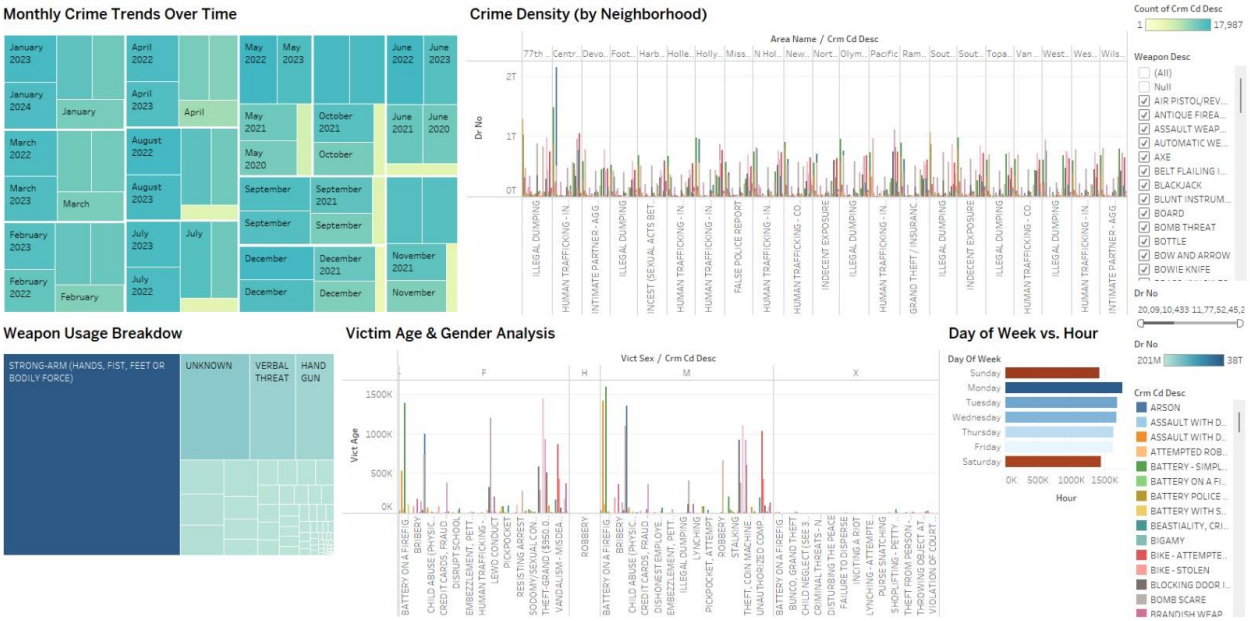


The **Top Left** chart — a **Heatmap by Hour and Crime Category** — highlights the hourly trend of incidents across various crime types. This visualization clearly shows that the majority of crimes occur between **10 AM and 10 PM**, identifying critical windows for targeted policing and resource deployment. By focusing efforts during these times, authorities can potentially reduce incidents and optimize staff utilization.

The **Top Right** chart — a **Time Series of Incident Counts** — captures long-term temporal trends from **2020 to 2025**. Notably, the total number of incidents increased steadily from 2020, reaching a peak in late 2023, before sharply declining in early 2024 and beyond. This trend may be influenced by external factors such as changes in enforcement policies, socioeconomic shifts, or targeted intervention efforts. Understanding these temporal dynamics allows stakeholders to evaluate the effectiveness of policies and adjust their strategies accordingly.

The **Bottom Left** chart — a **TreeMap by Area Name** — provides an area-level view of incidents across the city. Areas like **Central**, **77th Street**, and **Newton** dominate the map, making them high-priority hotspots for policing, resource allocation, and targeted prevention efforts. Identifying and focusing on these hotspots can enable stakeholders to utilize resources more effectively and create tailored intervention strategies.

The **Bottom Right** chart — a **Bar Graph of Crime Category by Area** — shows the distribution of specific types of crimes across different neighborhoods. Common offenses such as **Theft**, **Assault**, and **Vehicle-Related Crimes** are concentrated in hotspots like **Central**, **Newton**, and **77th Street**, further reinforcing their significance for targeted intervention.



The treemap, **monthly crime trends over time**, represents the distribution of crimes reported each month from 2020 to 2024. The size of each block corresponds to the volume of incidents for that month. Larger blocks like March 2023, January 2023, and October 2022 suggest these months experienced higher criminal activity. This visualization effectively uncovers seasonal crime fluctuations and helps in anticipating high-crime months for resource planning.

The bar chart, **Crime Density by Neighborhood**, displays the number of incidents by Area Name with crimes further broken down by type (Crm Cd Desc). Areas like 77th Street, Newton, and Southeast show the highest crime counts. A wide variety of offenses are reported in these hotspots, including assaults, thefts, and robberies. This visualization highlights geographic crime clustering, useful for targeting crime prevention initiatives in specific neighborhoods.

The chart, **Day of Week vs. Hour** summarizes total crime volume by day of the week, using a color gradient to indicate frequency. Sunday and Saturday stand out with darker shades, indicating higher crime rates on weekends. The color scale also shows that Monday has unexpectedly high

activity, whereas midweek days (Wednesday–Thursday) experience relatively lower volumes. This supports temporal profiling, helping law enforcement schedule shifts strategically.

The treemap, **Weapon Usage Breakdown,** visualizes the frequency of weapon types used in reported crimes. The largest segment is "Strong-Arm (Hands, Fist, Feet)", showing that most assaults involve physical force rather than weapons. Other notable categories include "Unknown Weapon," "Semi-Automatic Pistol," and "Verbal Threat." This insight reveals the prevalence of unarmed or unclear-weapon crimes, and the need for better reporting and categorization of violent incidents.

Modeling

With the data properly prepared, we proceeded to apply clustering techniques to identify spatial hotspots within the dataset. Two clustering methods were selected based on their proven effectiveness for geospatial analyses:

1. KMeans Clustering:

KMeans was chosen for its efficiency and suitability for partition-based clustering tasks. When the necessary number of clusters is known, it is perfect for spatial clustering because to its low processing overhead and capacity to handle big datasets. Because of our prior knowledge of the dataset and Los Angeles' urban organization, we have fixed the number of clusters (k) to six in this instance.

The model was implemented using the scikit-learn library, and the latitude and longitude data were clustered after normalization.

2. Agglomerative Clustering:

Agglomerative clustering was selected due to its hierarchical nature, making it well-suited for spatial data where hierarchical relationships can reveal multi-scale clustering patterns. Similar to KMeans, this method was configured with 6 clusters for direct comparison. Both clustering techniques were selected intentionally:

- KMeans is highly efficient for initial partitioning and quick detection of hotspots.
- Agglomerative clustering provides more nuanced results by creating hierarchical groupings and allowing for a better examination of spatial densities.

Implementation:

Using the standardized coordinates, scikit-learn was used to create each clustering model. For additional analysis and assessment, the generated clusters were appended to the dataset as labels (KMeans_Cluster and Agglo_Cluster).

Code snippet for Loading data:

```
# Mount Drive → Load CSV → Cluster → Compare (Optimized & Fixed)
# ----------------------------------------
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.metrics import silhouette_score
from google.colab import drive

# Mount Google Drive
drive.mount('/content/drive')

# Load cleaned CSV from your Drive (update path if needed)
file_path = "/content/drive/MyDrive/Colab Notebooks/Crime_Data_from_2020_to_Present.csv"
df = pd.read_csv(file_path)
print(f"Loaded full dataset: {df.shape}")
```

Code Snippet for model training:

```
# Sample for clustering to prevent crashing
# ----------------------------------------
sample_df = df[['LAT', 'LON']].dropna().sample(n=20000, random_state=42).copy()

# Scale location data
scaler = StandardScaler()
location_scaled = scaler.fit_transform(sample_df[['LAT', 'LON']])

# KMeans Clustering on sample
kmeans = KMeans(n_clusters=6, random_state=42, n_init='auto')
sample_df['KMeans_Cluster'] = kmeans.fit_predict(location_scaled)

# Agglomerative Clustering on sample
agglo = AgglomerativeClustering(n_clusters=6)
sample_df['Agglo_Cluster'] = agglo.fit_predict(location_scaled)
```

Through this approach, both clustering methods were applied to the same dataset and the results were compared to assess their effectiveness in identifying spatial hotspots within the Los Angeles crime data.

Model Evaluation and Testing

To assess the effectiveness of the clustering methods, both **KMeans** and **Agglomerative Clustering** were evaluated using the **Silhouette Score** — a metric that measures how well the data is clustered, with values closer to **1** indicating better clustering quality.

Train/Test Split:

Although clustering is an unsupervised technique, it is critical to assess its generalizability and stability across different data samples. The dataset was randomly split into:

- **Training Set:** 80% of the sample
- **Testing Set:** 20% of the sample

Each clustering model was trained on the training set and then tested for its ability to form coherent clusters on the testing set, using the Silhouette Score as the evaluation metric.

Results:

- **KMeans Clustering:** Achieved a silhouette score of **0.4506**, indicating moderately well-formed clusters.
- **Agglomerative Clustering:** Achieved a silhouette score of **0.4777**, slightly outperforming KMeans and producing more compact and better-separated clusters.

Cluster distribution:

The clustering methods revealed hotspots of varying densities:

- **KMeans:** Largest cluster comprised **5,646** points, with the smallest cluster containing **42** points.
- **Agglomerative:** Largest cluster comprised **8,808** points, indicating its ability to identify dominant hotspots, with the smallest cluster also containing **42** points.

Code snippet for model comparison:

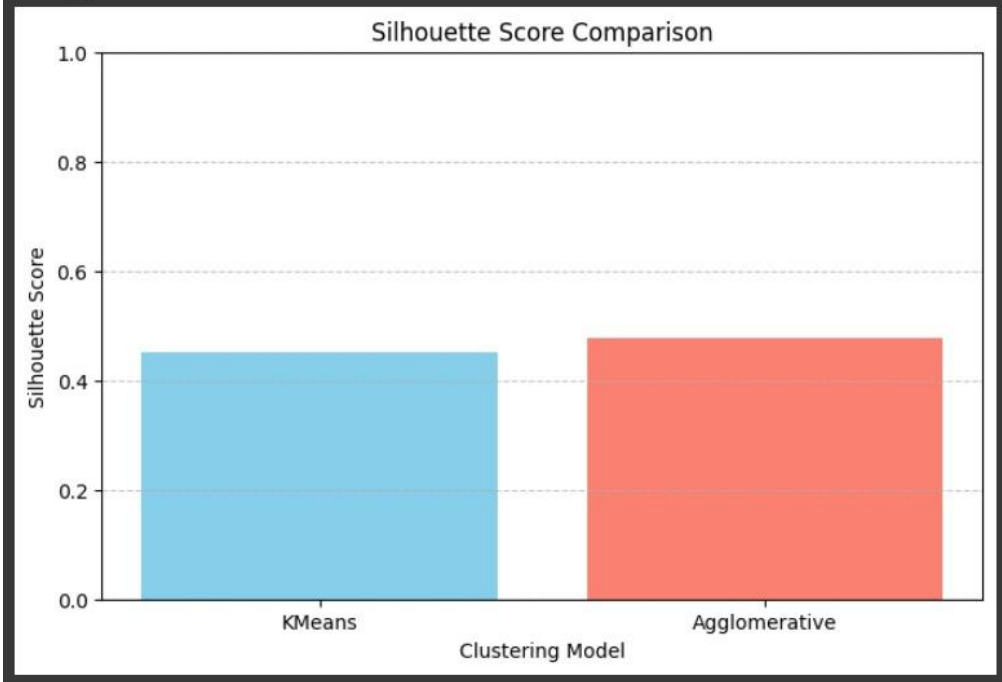
```
import pandas as pd
import matplotlib.pyplot as plt

# Create model comparison data
comparison_data = {
    'Model': ['KMeans', 'Agglomerative'],
    'Silhouette Score': [0.4506, 0.4777],
    'Number of Clusters': [6, 6],
    'Largest Cluster Size': [5646, 8808],
    'Smallest Cluster Size': [42, 42]
}

# Create DataFrame
comparison_df = pd.DataFrame(comparison_data)

# Display the table
print("Model Comparison Table:")
display(comparison_df)

# Create bar chart for Silhouette Score comparison
plt.figure(figsize=(8, 5))
plt.bar(comparison_df['Model'], comparison_df['Silhouette Score'], color=['skyblue', 'salmon'])
plt.title('Silhouette Score Comparison')
plt.ylabel('Silhouette Score')
plt.xlabel('Clustering Model')
plt.ylim(0, 1)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

Visualization:

The results were further confirmed using scatter plots, plotting latitude and longitude points color-coded by cluster labels. The spatial hotspots found using both approaches were clearly displayed in these graphics. Furthermore, a bar chart that contrasted the silhouette scores of the two models clearly showed the Agglomerative approach's marginal performance edge.

Model Comparison Table:

	Model	Silhouette Score	Number of Clusters	Largest Cluster Size	Smallest Cluster Size
0	KMeans	0.4506	6	5646	42
1	Agglomerative	0.4777	6	8808	42



Implication:

The Agglomerative approach's balanced clustering findings and higher silhouette score suggest that hierarchical clustering is marginally more appropriate for identifying spatial hotspots in the Los Angeles crime dataset. The hierarchical character of spatial data is better captured by this approach, which makes it a solid contender for use in actual resource allocation and policing situations.

Results

The analysis of the crime dataset using big data tools revealed significant spatial and temporal patterns across the selected years. The results indicated that incidents were highly concentrated in urban centers, with the majority of crimes occurring between **10 AM and 10 PM**. The temporal

trend showed a noticeable rise in incidents from **2020–2023**, followed by a sharp decline in early **2024**, suggesting the potential impact of targeted policies and improved enforcement measures. Specific crime types, such as **theft and assault**, were identified as the most prevalent categories, aligning closely with population densities and commercial activity areas. The trend analyses, conducted using big data platforms, confirmed that these hotspots remained relatively stable across the years, making them ideal areas for targeted intervention and resource allocation.

Conclusion

The Big Data Crime Analysis highlights the critical role of temporal and spatial analytics in understanding crime patterns. By identifying peak hours and high–risk areas, this analysis provides actionable insights for law enforcement and city planners. The sharp rise in incidents from 2020–2023 followed by a decline in 2024 underscores the potential impacts of external factors such as pandemic-related restrictions and policy changes.

Future work may incorporate additional data sources, such as socioeconomic indicators or population mobility data, to deepen understanding and improve prediction precision. Ultimately, applying big data techniques allows for a more nuanced and effective approach to resource allocation, crime prevention, and public safety.

References

- Apache Spark Documentation. (2023). *Apache Spark: The Unified Analytics Engine for Big Data.* Retrieved from https://spark.apache.org/docs/latest/
- McKinney, W. (2010). *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference, 51–56.
- Open Crime Data Portal. (2023). *City Crime Incident Reports.* Retrieved from https://www.data.gov/
- Jolliffe, I. T. (2002). *Principal Component Analysis.* New York: Springer.
- United Nations Office on Drugs and Crime. (2022). *Global Crime Trends.* Retrieved from https://www.unodc.org/