CRIME PATTERN ANALYSIS ON NYPD ARREST DATA

TEAM: SAI SUMANTH REDDY KACHI, SHASHANK PUPPALA, SUSHANTH REDDY MANDAPURAM, VAGBHAT DWIBHASHYAM

UMBC

INTRODUCTION

- Crime continues to be a critical concern in urban environments, particularly in densely populated cities like New York. Understanding the underlying patterns and trends in criminal activity is essential for developing effective prevention and intervention strategies.
- This project uses publicly available arrest data from the New York Police Department (NYPD) to analyze and find crime patterns across the city.
- Through a combination of data exploration, statistical analysis, and visualization techniques, the project examines key attributes such as type and description of offense, location of arrest, and demographic details of the perpetrator.
- By identifying recurring trends, crime hotspots and anomalies within the data, the project aims to generate meaningful insights that can support data-informed decision-making for public safety officials, urban planners, and community stakeholders.

OBJECTIVE:

To develop a machine learning model that can detect crime hotspots by analyzing geospatial arrest data from the NYPD. The model identifies abnormal concentrations of arrest locations (e.g., clusters of high activity across boroughs) and provide high crime regions called called crime hotspots. This enables a data-driven approach to understanding urban crime patterns and can support strategic decision-making in law enforcement and public safety planning.

DATASET DESCRIPTION:

The dataset contains 19 attributes and 260503 instances. Out of the 19 attributes 9 are numerical, 10 categorical. There are no redundant or duplicate rows. There are 1442 (<0.1%) missing values. All attributes: ARREST_KEY, ARREST_DATE, PD_CD, PD_DESC, KY_CD, OFNS_DESC, LAW_CODE, LAW_CAT_CD, ARREST_BORO, ARREST_PRECINCT, JURISDICTION_CODE, AGE_GROUP, PERP_SEX, PERP_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, New Georeferenced Column.

Dataset:
https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date/resource/c48f1a1a-5efb-4266-9572-769ed1c9b472

Important Attributes:
OFNS_DESC, LAW_CAT_CD, ARREST_PRECINCT
ARREST_BORO, AGE_GROUP, PERP_SEX, PERP_RACE,
Latitude, and Longitude.

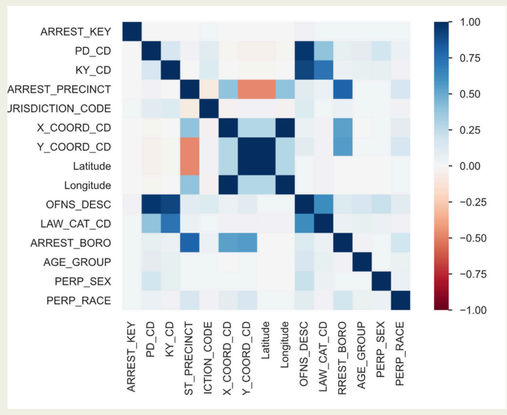
DATA PREPROCESSING:

- **Dimensionality Reduction:** A total of 10 attributes (ARREST_KEY, ARREST_DATE, PD_CD, PD_DESC, KY_CD, LAW_CODE, X_COORD_CD, Y_COORD_CD, JURISDICTION_CODE, and New Georeferenced Column) which are not useful to develop the model are dropped from the dataset.
- **Handling Missing Values:** The rows containing missing values after dropping unnecessary columns are dropped from the dataset.

EXPLORATORY DATA ANALYSIS

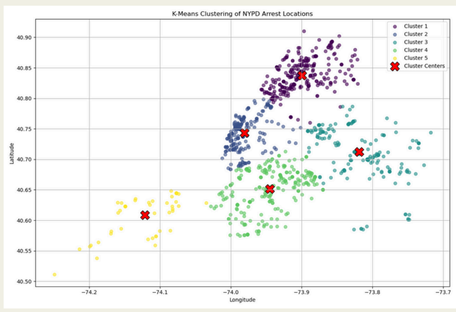
EDA played a crucial role in understanding the dataset and preparing it for modeling. Key techniques included:

- **Correlation Heatmap:** These helped identify relationships between features, highlighting attributes with strong positive or negative correlations to the target variable.
- **Distribution Plots:** Used to examine the spread of features, detect outliers, and understand patterns in data distributions.



IMPLEMENTED MODELS WITH RESULTS:

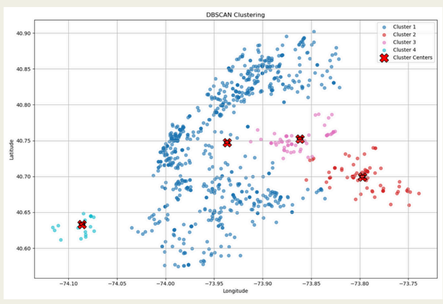
K-Means Clustering




```
Cluster Centers (Latitude, Longitude):
Cluster 1: (40.8374, -73.8996)
Cluster 2: (40.7431, -73.9809)
Cluster 3: (40.7122, -73.8190)
Cluster 4: (40.6518, -73.9449)
Cluster 5: (40.6090, -74.1215)
```

- The K-Means clustering algorithm was applied to geospatial arrest data to uncover crime hotspots across New York City. The model identified five cluster centers based on the latitude and longitude of arrests, effectively highlighting areas with concentrated criminal activity
- These correspond to well-known high-crime zones in **Manhattan, the Bronx, Brooklyn, Queens, and Staten Island**, respectively. While the model successfully segmented the city into meaningful crime zones, it assumes clusters are spherical and evenly sized—limitations that may not fully capture the complexity of urban crime patterns.

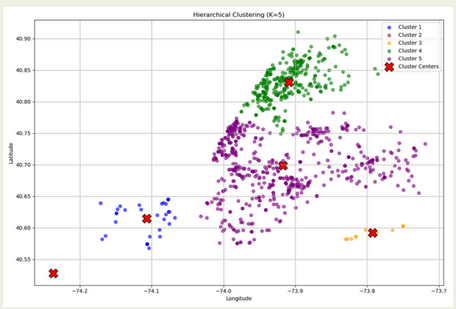
DBSCAN Clustering




```
Cluster Centers (Latitude, Longitude):

Cluster 1 center: (40.7472, -73.9367)

Cluster 2 center: (40.6993, -73.7970)

Cluster 3 center: (40.7525, -73.8616)

Cluster 4 center: (40.6332, -74.0861)
```

The DBSCAN algorithm was applied to arrest location data to uncover natural crime clusters without predefining the number of regions. Unlike K-Means or standard hierarchical methods, DBSCAN is density-based, allowing it to detect clusters of varying shapes, sizes, and densities while effectively identifying sparse regions as noise. The model successfully revealed nuanced, irregular crime zones in areas like downtown **Manhattan and eastern Brooklyn—regions** often missed by simpler clustering techniques. Its ability to manage noise and non-linear boundaries makes DBSCAN particularly well-suited for analyzing complex urban crime patterns.

Hierarchical Clustering



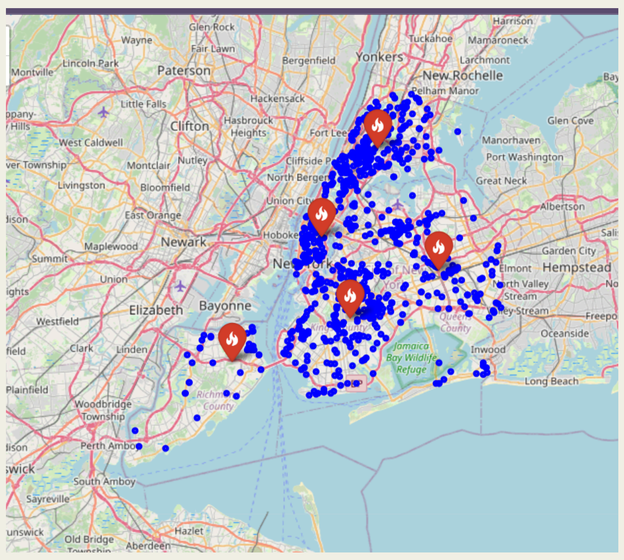

```
Cluster Centers (Latitude, Longitude):
Cluster 1: (40.6146, -74.1070)
Cluster 2: (40.5281, -74.2371)
Cluster 3: (40.5921, -73.7923)
Cluster 4: (40.8305, -73.9091)
Cluster 5: (40.6985, -73.9170)
```

Hierarchical clustering was employed on geospatial arrest data to identify layered patterns of criminal activity across New York City. Using average linkage and the Haversine distance metric, the model grouped arrest locations into five spatial clusters, revealing distinct hotspots that align with known high-crime boroughs such as Manhattan, Brooklyn, and the Bronx. This method's strength lies in its flexibility—it does not require a predefined number of clusters and can reveal nested structures in the data. However, its computational complexity and sensitivity to noise may affect scalability and robustness in larger datasets.

Best Model:

Based on the Silhouette Scores of the models, K-Means performed best with a score of 0.4776.

```
Silhouette Scores:
KMeans: 0.4776
Hierarchical: 0.4568
DBSCAN: 0.0541
```



Crime Profiles:

- Crime profiles help decision-makers identify and respond to crime more effectively by understanding who is committing the crimes.
- These insights help in finding motives of crime and help in esigning policies, targeting crime prevention efforts, and improving the effectiveness of law enforcement.
- Based on the number of crimes, the type of crime, and demographics of the perpetrator, crime profiles have been identified which occur more frequently.
- The top 10 crime profiles along with number of crimes are as follows:

```
1.BLACK Male aged 25–44 committed 7188 Misdemeanor crimes
  (ASSAULT 3 & RELATED OFFENSES)

2.WHITE HISPANIC Male aged 25–44 committed 4957 Misdemeanor crimes
  (ASSAULT 3 & RELATED OFFENSES)

3.BLACK Male aged 25–44 committed 4603 Felony crimes
  (FELONY ASSAULT)

4.BLACK Male aged 25–44 committed 4495 Misdemeanor crimes
  (OTHER OFFENSES RELATED TO THEFT)

5.BLACK Male aged 25–44 committed 4297 Misdemeanor crimes
  (PETIT LARCENY)

6.BLACK Male aged 25–44 committed 4110 Felony crimes
  (MISCELLANEOUS PENAL LAW)

7.BLACK Male aged 25–44 committed 3664 Misdemeanor crimes
  (VEHICLE AND TRAFFIC LAWS)

8.WHITE HISPANIC Male aged 25–44 committed 3402 Misdemeanor crimes
  (PETIT LARCENY)

9.BLACK Male aged 45–64 committed 3097 Misdemeanor crimes
  (PETIT LARCENY)

10.BLACK Female aged 25–44 committed 3038 Misdemeanor crimes
  (ASSAULT 3 & RELATED OFFENSES)
```

FUTURE WORK :

Future work involves expanding the analysis across multiple years to identify long-term trends and seasonal patterns in arrest data. Integrating socioeconomic data such as unemployment rates, education levels, and housing conditions could provide deeper context behind arrest distributions. A geospatial analysis using GIS tools can be conducted to visualize crime hotspots and their correlation with neighborhood characteristics. Machine learning models could be developed to predict potential spikes in arrests based on environmental or social indicators. Sentiment analysis on police reports or social media could also reveal public perception and real-time incident reporting trends.

CONCLUSION:

This project demonstrates the power of unsupervised learning techniques in uncovering meaningful patterns in urban crime data. By leveraging geospatial clustering algorithms—K-Means, Hierarchical Clustering, and DBSCAN—we identified distinct crime hotspots across New York City and explored the spatial distribution of arrests with greater granularity. Each model offered unique advantages: K-Means provided clear segmentation, Hierarchical Clustering revealed nested spatial hierarchies, and HDBSCAN captured irregular, noise-tolerant clusters. These insights, coupled with demographic and offense-level analysis, can support smarter, data-driven strategies for law enforcement and urban planning. The findings underscore the value of combining machine learning with public data to drive informed, actionable decisions in public safety.

QR CODE

REFERENCES:

1. Wang, F., & Brown, D. (2011). The spatio-temporal modeling of crime by using spatial and temporal kernel density estimation. Computers, Environment and Urban Systems, 35(5), 323–337. https://doi.org/10.1016/j.compenvurbsys.2011.07.001
2. Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal, 21(1), 4–28. https://doi.org/10.1057/palgrave.sj.8350066
3. Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546
4. Esri. (2020, June 10). Crime mapping using GIS. Esri Blog. https://www.esri.com/about/newsroom/blog/crime-mapping-using-gis/
5. Medina, J. (2019). Crime mapping and spatial data analysis using R. CRC Press.