# **Final Report: NYC Shootings Cluster Analysis**

## **1. Project Overview**

The goal of this project was to analyze patterns in shooting incidents across New York City using unsupervised machine learning techniques, specifically clustering. By examining spatial, temporal, and demographic attributes, we aimed to identify meaningful groupings of incidents that can support data-driven public safety strategies.

## **2. Dataset Description**

The dataset includes historical records of shootings in NYC with attributes such as:

* **Date & Time** of occurrence
* **Geographic Coordinates** (latitude, longitude)
* **Victim and Perpetrator Demographics** (age, race, gender)
* **Jurisdiction Code**
* **Other contextual data** about each incident

Missing data in some demographic features was imputed using median values. Date and time fields were transformed to extract hour of day and convert dates into datetime format.

## **3. Methodology**

### **3.1 Data Preprocessing**

* Removed rows with missing location or timestamp data.
* Converted dates and times into usable numeric features (e.g., hour of day).
* Scaled features using StandardScaler to normalize the input data.

### **3.2 Feature Selection**

The features selected for clustering included:

* latitude, longitude (location)
* hour (time of day)
* victim\_age, perp\_age (demographics)
* jurisdiction\_code (contextual/legal relevance)

### **3.3 Clustering Technique**

* Applied **K-Means Clustering** with k=5 (based on elbow method).
* Each incident was assigned a cluster label.
* Dimensionality reduction via **PCA** was used to visualize the clusters in 2D.

### **3.4 Visualization**

* **PCA Scatter Plot**: Visualized clustering in reduced feature space.
* **Interactive Map (Plotly)**: Visualized geographic spread and hotspots of clusters.

## **4. Results**

### **4.1 Cluster Analysis**

* **Cluster 0**: Concentrated in high-density downtown Manhattan, mostly late-night incidents.
* **Cluster 1**: Spread across multiple boroughs, moderate age range for both victim and perpetrator.
* **Cluster 2**: Clustered in Brooklyn and Bronx with younger perpetrators and afternoon timing.
* **Cluster 3**: Early morning shootings, more frequent in residential outer-borough neighborhoods.
* **Cluster 4**: Jurisdiction-heavy cluster, indicating specialized law enforcement zones.

### **4.2 PCA Results**

The PCA plot showed clear separation between most clusters, suggesting that the clustering captured distinct patterns in the data.

### **4.3 Geospatial Insights**

* Multiple **hotspots** identified in Bronx and Brooklyn.
* Certain areas were repeatedly involved across time, showing persistent risk zones.

## **5. Recommendations**

* **Targeted Patrols**: Direct law enforcement resources to the identified hotspots, especially clusters with repeat incidents.
* **Time-Based Allocation**: Increase patrols during late-night and early-morning hours in identified zones.
* **Community Engagement**: Initiate outreach programs in high-risk clusters to address underlying social issues.
* **Further Analysis**: Incorporate additional layers (e.g., socio-economic data, gang activity reports) to enrich understanding.

## **6. Success Criteria Evaluation**

| **Criterion** | **Outcome** |
| --- | --- |
| Accuracy | Clusters showed distinct temporal, spatial, and demographic patterns. |
| Relevance | Clusters were interpretable and actionable for law enforcement. |
| Scalability | Model can be easily applied to new or expanded datasets. |
| Usability | Visualizations are intuitive and designed for non-technical stakeholders. |

## **7. Ethical Considerations**

* Data was handled responsibly with a focus on **privacy** and **bias reduction**.
* No personally identifiable information (PII) was included or processed.
* Demographic variables were used solely for pattern recognition and not for profiling.

## **8. Limitations**

* The dataset may contain underreporting or bias in law enforcement documentation.
* Temporal granularity (only using hour of day) may overlook day-of-week or seasonal patterns.
* Clustering is unsupervised; further validation through expert feedback is recommended.

## 

## **9. Conclusion**

The clustering analysis of NYC shootings successfully revealed actionable patterns across time, space, and demographics. These insights can support strategic decision-making and contribute to a more proactive approach to public safety.

Future work could include more advanced clustering algorithms like **DBSCAN** or **HDBSCAN** and integration with real-time data sources for dynamic policing models.