Report 3

CLUSTER ANALYSIS OF
NEW YORK CITY'S STOP-QUESTION-FRISK (SQF) DATA

By

PRAJWAL NAGARAJ

Abstract:
This report presents the results of cluster analysis performed on the Stop, Question, and Frisk (SQF) dataset for the year 2012, focusing on crime location clustering in Manhattan for felony incidents and clustering stopped individuals by reasons for stop. Additionally, it explores the potential applications of cluster analysis in the dataset and evaluates the clustering results using internal validation measures.

Data Preparation:

1. **Definition and Preparation of Class Variables:**
    - Class variables were defined as attributes relevant to the analysis, including crime location (xcoord and ycoord) and reason for stop (crimsusp).
    - Irrelevant columns such as personal identifiers, date and time of stop, and detailed police encounter information were dropped from the dataset.
2. **Final Dataset Description:**
    - The final dataset used for classification consists of crime location coordinates (xcoord and ycoord) and the reason for stop (crimsusp).
    - The scale and range for the new combined variables (xcoord and ycoord) represent geographical coordinates within the Manhattan area.

Modelling:

1. **Cluster Analysis:**
    - Crime location clustering for felony incidents in Manhattan was performed using the K-means algorithm.
    - Stopped individuals were clustered based on the reasons for stop, allowing for the identification of distinct groups within the dataset.
    - Cluster analysis was also explored for other potential applications, such as clustering individuals based on demographic attributes or police encounter outcomes.
2. **Determining Suitable Number of Clusters:**
    - The optimal number of clusters for each method was determined using the Elbow method, which identifies the point of inflection (elbow) in the plot of Within-Cluster Sum of Squares (WCSS) against the number of clusters.
3. **Internal Validation Measures:**
    - Internal validation measures, including the Elbow method plot and visualization of clusters with centroids, were used to describe and compare the clustering results.

- The Elbow method plot demonstrated the optimal number of clusters for felony incidents in Manhattan, while the cluster visualization provided insights into the spatial distribution of crime locations.

Evaluation:

The results of the cluster analysis provide valuable insights into the spatial distribution of felony incidents in Manhattan and the grouping of stopped individuals based on the reasons for stop. The most interesting findings include:

- The identification of distinct geographical clusters for felony incidents, which can inform law enforcement agencies and policymakers about high-crime areas.
- The clustering of stopped individuals based on the reasons for stop, revealing patterns in police encounters and potential disparities in law enforcement practices.
- The versatility of cluster analysis in exploring various aspects of the dataset, such as demographic characteristics, police encounter outcomes, and spatial-temporal patterns of crime.

These findings can be used to guide resource allocation, strategic planning, and targeted interventions aimed at reducing crime rates, improving community relations, and promoting equitable policing practices.

1. **Elbow Method for Optimal Number of Clusters for Felony in Manhattan Dataset**:
    - It plots the Within-Cluster Sum of Squares (WCSS) on the y-axis against the Number of Clusters on the x-axis.
    - The WCSS is a measure used to evaluate the performance of a clustering algorithm, where lower values generally indicate better clustering by minimizing intra-cluster variance.
    - The graph shows a sharp decline in WCSS as the number of clusters increases from 1 to around 5 or 6, after which the decrease in WCSS becomes more gradual. This point, where the rate of decrease sharply changes, is known as the "elbow," suggesting that the optimal number of clusters for this dataset might be around 5 or 6.
2. **Clustering for Felony in Manhattan**:
    - It is a scatter plot showing the geographical distribution of felony incidents in Manhattan.
    - The x-axis represents the x-coordinate (likely longitude), and the y-axis represents the y-coordinate (likely latitude).
    - Data points are colored differently based on the cluster they belong to, indicating that the clustering algorithm has grouped the felony incidents into distinct geographical areas.

- Each cluster is represented by a different color, and red dots within each cluster may represent centroids or specific points of interest.

These visualizations are useful for understanding the spatial distribution of felony incidents in Manhattan and determining the number of clusters that best capture the inherent grouping in the data. This can aid in resource allocation, strategic planning, and targeted interventions for crime prevention.