Mohammad Bagher Soroush - 400130273 - ML 6 Report

1 .Explanation of Process Steps

The clustering analysis in the provided notebook involves the following steps:

Data Preprocessing:

- Importing Data: The dataset is read into a Pandas DataFrame.
- Categorical Encoding: Categorical variables (Channel, Region) are transformed into numerical features using one-hot encoding.
- Normalization: Numerical features are standardized using StandardScaler to ensure all variables contribute equally to the clustering process.

Clustering Implementation:

- K-means Algorithm:
- The number of clusters is initially determined using the Elbow Method.
- Clustering is performed on the normalized dataset.

Dimensionality Reduction:

- PCA (Principal Component Analysis) is applied to reduce dimensions for visualization purposes.

Evaluation:

- Silhouette Score: Measures the compactness and separation of clusters.
- Elbow Method: Uses the Within-Cluster Sum of Squares (WCSS) to find the optimal number of clusters.

2. Silhouette Scoring vs. Elbow Method

Silhouette Score:

- This metric measures how similar an object is to its own cluster compared to other clusters.
- Values range from -1 (incorrect clustering) to +1 (dense and well-separated clusters).
- It evaluates the "tightness" of clusters and is useful for validating the quality of clustering.

Elbow Method:

- Plots WCSS against the number of clusters.
- The "elbow" point indicates the optimal number of clusters, where adding more clusters minimally improves the WCSS.
- This method focuses on the trade-off between cluster count and variance explained.

Comparison:

- The Silhouette Score provides a direct evaluation of cluster quality, while the Elbow Method helps decide the cluster count.

- Both methods complement each other: Elbow determines the number of clusters, and Silhouette evaluates their effectiveness.

3. Cluster Analysis

Overview:

The clustering results indicate groupings based on customer spending habits. For instance:

- Cluster 1: High spenders across all product categories, potentially representing wholesale buyers.
- Cluster 2: Customers with moderate spending, likely smaller businesses or individual consumers.
- Cluster 3: Low spenders or niche buyers focusing on specific categories.

Behavior Analysis:

- By analyzing cluster centroids, we can deduce spending patterns and identify which features drive each cluster.
- Example insights:
- Customers in Cluster 1 may require premium services.
- Cluster 3 might benefit from targeted marketing campaigns.

4. Hyperparameters in K-means

(a) init

- Specifies the initialization method for cluster centroids.
- Options: k-means++ (default, improves convergence), random (random centroid initialization).
- Effect: Better initialization (e.g., k-means++) reduces the risk of poor convergence.

(b) n_init

- Number of times the algorithm runs with different centroid seeds.
- Default: 10.
- Effect: Higher values increase reliability but require more computation time.

(c) max_iter

- Maximum number of iterations for a single run of K-means.
- Default: 300.
- Effect: Ensures convergence but may increase runtime if set too high without significant gains.