Clustering Logic and Metrics

Clustering logic and metrics refer to the methodology and evaluation criteria used to group similar data points into clusters in a dataset and assess the quality of those clusters.

Clustering Logic:

Clustering involves grouping data points into clusters based on their similarity or distance in the feature space. The following steps outline the logic:

1. Data Preparation:
   - Combine relevant features from the dataset (e.g., customer profile and transaction details).
   - Normalize the data to ensure all features contribute equally to the clustering.

2. Choosing a Clustering Algorithm:
   - K-Means Clustering: Divides data into k clusters, minimizing intra-cluster variance.
   - DBSCAN: Groups data points based on density, identifying noise or outliers.
   - Hierarchical Clustering: Builds a hierarchy of clusters by iteratively merging or splitting them.

3. Determine the Optimal Number of Clusters:
   - Elbow Method: Evaluate the sum of squared distances (inertia) and identify the "elbow" point where adding more clusters results in diminishing returns.
   - Silhouette Score: Measure how similar data points in a cluster are to points in other clusters.

4. Assign Clusters:
   - Train the model on the dataset and assign cluster labels to each data point.

5. Visualization:

- Use techniques like PCA or t-SNE to reduce dimensionality and visualize clusters in 2D or 3D space.

Clustering Metrics:

Metrics are used to evaluate the performance and quality of the clusters. Key metrics include:

1. Davies-Bouldin Index (DB Index):

- Measures the average similarity ratio of clusters to their sizes.

- Lower values indicate better-defined clusters.

Formula:

DB = (1/k) * SUM(max((sigma_i + sigma_j) / d_ij))

where:

- sigma: Cluster dispersion.

- d_ij: Distance between cluster centroids i and j.

2. Silhouette Score:

- Measures how similar a data point is to its own cluster compared to other clusters.

- Range: -1 to 1 (higher is better).

3. Inertia (Within-Cluster Sum of Squares):

- Sum of squared distances between each point and its cluster centroid.

- Lower inertia indicates tighter clusters.

4. Purity (for labeled datasets):

- Measures the extent to which clusters contain a single class of data points.

5. Dunn Index:

- Ratio of the minimum inter-cluster distance to the maximum intra-cluster distance.

- Higher values indicate better clustering.

Example Application:

For the given eCommerce dataset:

1. Combine features like Region, SignupDate, TotalValue, and TransactionCount.

2. Use K-Means with normalized data to create 5 clusters.

3. Evaluate clusters using the DB Index.

4. Visualize clusters using PCA to understand their separability.

By employing these metrics and steps, you can ensure meaningful and actionable clustering results.