Lab Assignment: Clustering-Based Customer Segmentation from Online Retail Data

Objective

- Perform data preprocessing and feature engineering for customer profiling.
- Apply and compare different clustering algorithms (excluding link-based methods).
- Evaluate the clustering results.
- Demonstrate theoretical understanding of clustering techniques.

Dataset

Online Retail Dataset (UCI Repository)
URL: https://archive.ics.uci.edu/dataset/352/online+retail
Contains transactional data for a UK-based online retail store, including:

`InvoiceNo`, `StockCode`, `Description`, `Quantity`, `InvoiceDate`, `UnitPrice`, `CustomerID`, `Country`

Part 1: Data Preparation and Feature Engineering

Resource: RFM Feature Engineering

1. **Clean the data**
   - Remove records with missing `CustomerID`.
   - Filter for a single country (e.g., `United Kingdom`) to reduce noise.
   - Remove cancelled orders (where `InvoiceNo` starts with "C").
2. **Generate customer-level features**
   For each `CustomerID`, compute:
   - **Recency**: Days since last purchase (relative to max date in dataset)
   - **Frequency**: Number of purchases (distinct invoices)

    - **Monetary**: Total value of purchases (`Quantity × UnitPrice`)
      This forms your **RFM feature vector**.

3. **Standardize** the RFM data before clustering.

4. **Visualize** the RFM distribution (e.g., pairplot, histograms).

Part 2: Apply and Analyze Clustering Methods

Apply the following clustering methods on the RFM vectors and visualize the resulting clusters (e.g., via PCA):

2.1 K-Means Clustering

- Use the **Elbow Method** to find the optimal number of clusters.

- Report centroids and interpret customer segments.

2.2 Hierarchical Clustering (AGNES)

- Try **single**, **complete**, and **average** linkage.

- Plot dendrograms and discuss how to decide the number of clusters.

2.3 DBSCAN

- Use distance plots to choose `eps`, and choose a reasonable `min_samples` value.

- Visualize clusters, noise points, and discuss advantages over K-means.

Part 3: Clustering Evaluation

Use at least the following evaluation metrics for each clustering result:

- **Silhouette Score**

- **Inter-cluster vs intra-cluster distances**

- Brief **interpretation** of each cluster (e.g., high-value vs low-value customers)

Part 4: Theoretical Understanding

Include answers to the following in your report:

1. Explain the strengths and limitations of each clustering method you applied.
2. What assumptions does each algorithm make about data structure?
3. Why might DBSCAN detect “noise” while K-means cannot?
4. Discuss why scaling the features before clustering was necessary.
5. What would be the implications of using different distance metrics?

Submission Requirements

- Well-commented **Python code** (preferably in a Jupyter notebook).
- A short **report in IEEE Conference Format (max 4 pages)** covering:
  - Feature engineering
  - Clustering approach and visualization
  - Evaluation metrics
  - Theoretical answers
  - Interpretation of clusters and business insights