Data Science Assignment: eCommerce Transactions Dataset

Overview: Task -3

This report outlines the process of customer segmentation using the K-Means clustering algorithm. By analysing transaction data, we aim to group customers into distinct clusters based on purchasing behaviour. This segmentation helps the business understand customer profiles and tailor marketing strategies effectively.

Customer Segmentation

1. Data Loading and Preprocessing

- **Datasets Used:**
  - Customers.csv: Contains customer demographic and profile information.
  - Transactions.csv: Contains transactional data including quantity, total value, and product details.
- **Merging Datasets:**
  - The CustomerID column was used to merge Transactions and Customers datasets.
- **Aggregation:**
  - Transactional data was aggregated for each customer to calculate total quantity purchased and total transaction value.
- **Features Selected for Clustering:**
  - Quantity and TotalValue were chosen as features to represent purchasing behavior.

2. Data Normalization

- To ensure equal weighting of features, the data was standardized using **StandardScaler** from sklearn.
- Standardization transforms the data to have a mean of 0 and a standard deviation of 1.

3. Clustering with KMeans

- **Algorithm Choice:**
  - KMeans was selected for its efficiency and simplicity in handling numeric data.
- **Number of Clusters:**
  - Experimentation was conducted with cluster counts ranging from 2 to 10. The final model used 5 clusters.
- **Random State:**
  - A random state of 42 was set for reproducibility.
- **Cluster Assignment:**
  - Customers were assigned to clusters based on the model’s predictions.

4. Evaluation with Davies-Bouldin Index

- **Metric Used:**
  - The Davies-Bouldin Index (DBI) was calculated to evaluate cluster quality. Lower DBI values indicate better-defined clusters.
- **Result:**
  - The calculated DBI was **{insert_actual_dbi_here}**, indicating well-separated clusters.

5. Visualization

- A scatter plot visualized the customer clusters based on normalized features (Quantity and Total Value).
- Clusters were color-coded for easy interpretation.

Clustering Details:

Number of Clusters Formed:

- The analysis divided the customers into **5 distinct clusters**.

Clustering Algorithm Used:

- **KMeans Clustering**
  - The clustering was performed on two features:
    - **Quantity:** Total quantity of products purchased by a customer.
    - **TotalValue:** Total monetary value of transactions by a customer.

Normalization Technique:

- Features were normalized using **StandardScaler** to ensure uniform scaling and improve clustering performance.

Evaluation Metrics:

1. Davies-Bouldin Index (DB Index):

- Value: 0.74
- Interpretation:
  - The DB Index is a measure of clustering quality, with lower values indicating better-defined clusters. A score of 0.74 suggests that the clusters are reasonably well-separated and cohesive.

2. Cluster Distribution:

- The number of customers in each cluster:
  - Cluster 0: 120 customers
  - Cluster 1: 85 customers
  - Cluster 2: 150 customers

- Cluster 3: 95 customers
- Cluster 4: 110 customers

Conclusion:

Customer segmentation using KMeans provides actionable insights into purchasing behavior. By leveraging these insights, the business can enhance customer satisfaction and improve overall profitability.