Customer Segmentation / Clustering

Task: 3

Objective:

The goal was to segment customers using clustering techniques based on both **profile information** and **transaction behaviour**.

Results

Using KMeans Clustering

1. Number of Clusters Formed

- **4 Clusters** were formed using the **KMeans** clustering algorithm, based on customer features such as total spend, average quantity purchased, product diversity, and transaction frequency.

2. Davies-Bouldin Index (DB Index)

- **DB Index Value**: **1.0157**
- A lower DB Index indicates better clustering performance. This value suggests a good balance between intra-cluster compactness and inter-cluster separation.

3. Other Relevant Clustering Metrics

- **Compactness**: Data points within each cluster are closely packed, indicating meaningful groupings of customers based on similar purchasing behaviours.
- **Separation**: Clusters are distinct from each other, showing clear differences in customer groups.
- **Cluster Behaviour**:
- **Cluster 1**: High-spending, frequent buyers who represent loyal, high-value customers.
- **Cluster 2**: Moderate spending but high product diversity, indicating customers with varied purchase patterns.
- **Cluster 3**: Low-spending, infrequent buyers who may be less engaged.
- **Cluster 4**: Balanced spenders with moderate frequency, representing mid-tier customers.

4. Insights from Clustering

The clustering results provide actionable insights into customer segmentation, allowing for targeted strategies:

- Focus on retaining high-value customers (Cluster 1) through loyalty programs.
- Engage low-spending customers (Cluster 3) with personalized offers to increase spending.
- Cater to diverse-purchasing customers (Cluster 2) by offering variety-focused marketing campaigns.

This clustering analysis, supported by the DB Index and visualizations, provides a clear understanding of customer segments for informed decision-making.

Using DBSCAN

The **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** algorithm was applied with the following parameters:

- **eps**: 0.5 (radius of neighbourhood)
- **min_samples**: 5 (minimum number of points to form a cluster)

This algorithm does not require pre-specifying the number of clusters and can automatically identify noise points (outliers).

4. Clustering Results:

- **Clusters Formed:** The algorithm identified multiple clusters, with some points being labeled as noise (-1 label).
- The **number of clusters** excluding noise was calculated using the value counts of the Cluster column.

Clusters Distribution:

- Total number of clusters formed: 5 (including the noise points labeled as -1)
- Noise points (cluster = -1): 45 customers (outliers)

5. Clustering Metrics:

- **Davies-Bouldin Index (DBI):**
  - **DB Index Value**: 1.58 (lower values indicate better clustering performance; ideal is 0)
- **Silhouette Score:**
  - **Silhouette Score**: 0.27 (ranges from -1 to 1; higher values indicate better-defined clusters)

6. Visualizations:

1. **Scatter Plot**: A scatter plot was used to visualize how customers are distributed based on two features: **Total Spend** vs **Average Quantity Purchased**. The points were coloured based on their cluster label.
2. **Pair Plot**: A pair plot of features (Total Spend, Avg Quantity Purchased, Product Diversity, Frequency) was generated, with clusters represented by different colours. This helped in visually analysing the relationships between features within each cluster.

7. Interpretation & Insights:

Cluster Distribution:

- **Cluster 0**: Contains the majority of customers.
- **Cluster 1**: Small cluster, characterized by customers with higher product diversity but lower spend.
- **Cluster 2**: Contains frequent shoppers with relatively high average purchase quantities.
- **Cluster 3**: Includes a group of customers with very low spend and product diversity.