DATA SCIENCE ASSESSMENT : ECOMMERCE TRANSACTIONS DATASET

TASK : 3 Clustering Analysis Report

INTRODUCTION :

This clustering analysis aims to identify groups of customers based on transactional data. The KMeans clustering algorithm was used to partition the data into distinct clusters, and several metrics were employed to assess the quality and coherence of the clustering results.

1. Number of Clusters Formed:

- The analysis determined that the optimal number of clusters to represent the customer base is **8**. This was identified after evaluating multiple cluster numbers and selecting the one that minimized the Davies-Bouldin Index and provided a reasonable Silhouette Score.

2. Davies-Bouldin (DB) Index:

- The **Davies-Bouldin Index (DBI)** for the selected number of clusters (8) is **0.7271**. This index is a key indicator of cluster separation. Lower values indicate that clusters are well-separated and have compact shapes. A DBI of 0.7271 suggests that the clusters are relatively well-separated, but there is still room for improvement, as some clusters might be overlapping or not as distinct as desired.

3. Silhouette Score:

- The **Silhouette Score** for the selected clustering solution is **0.4877**. This metric assesses how similar each data point is to its own cluster (cohesion) compared to other clusters (separation). A score closer to +1 indicates that the points are well-clustered, while a score near 0 suggests overlapping clusters, and negative values indicate misclassified points. With a score of 0.4877, this clustering solution

reflects a moderate degree of cohesion and separation, indicating that the clusters are reasonably well-defined but not perfectly distinct.

4. **Cluster Sizes**:

- The customers were grouped into 8 clusters with the following distribution:
    - **Cluster 2:** 199 customers
    - **Cluster 0:** 148 customers
    - **Cluster 5:** 138 customers
    - **Cluster 1:** 130 customers
    - **Cluster 4:** 125 customers
    - **Cluster 3:** 109 customers
    - **Cluster 6:** 89 customers
    - **Cluster 7:** 62 customers
- The distribution shows some clusters with significantly more customers than others. Clusters 2 and 0 are notably larger than clusters 7 and 6, which might suggest that certain customer behaviors or characteristics are more prevalent than others.

5. **Visualizing the Clusters**:

- **PCA Visualization**: To better understand the structure of the clusters, the dimensionality of the data was reduced using Principal Component Analysis (PCA) to two components. The clusters were visualized on a 2D scatter plot where each point represents a customer. The plot shows how well the clusters are separated along the two principal components.
    - The clusters appear to have distinct groupings, with some overlap indicating that further refinements could be made to improve cluster separability.
- **Pairplot**: A pairplot of the data, color-coded by the clusters, provides insights into the relationships between the features within each cluster.

It shows that some clusters have clear distinctions across different feature combinations, while others may have more variability or overlap.

6. General Insights:

- **Cohesion and Separation:** The clustering results indicate that there is a reasonable degree of cohesion within the clusters and some level of separation between them. However, the moderate Silhouette Score suggests that there may still be room for improvement in terms of distinguishing the clusters more clearly.
- **Customer Segmentation:** With the customer base segmented into 8 distinct clusters, businesses can now analyze these clusters for targeted marketing, customer behavior analysis, and personalized recommendations.
- **Cluster Size Variability:** The uneven distribution of customers across clusters could indicate varying degrees of customer importance or behavior patterns. Larger clusters may represent common customer profiles, while smaller clusters might represent niche or specialized behaviors.

7. Conclusion:

- The KMeans clustering results with 8 clusters offer a reasonably good segmentation of the customers, with decent separation and cohesion. This segmentation provides a foundation for understanding customer behaviors and can be leveraged for marketing strategies, product recommendations, and further customer analysis. While the results are promising, there is potential for refining the clustering solution for even better insights and customer differentiation.