Customer Segmentation / Clustering Report

Overview of the Clustering Approach:

For this task, I used **Agglomerative Clustering**, a popular hierarchical clustering method, to segment customers into distinct groups based on their transaction and profile data. After experimenting with different linkage methods, we decided on using the **ward** method, which aims to minimize the variance within each cluster.

Results:

1. **Number of Clusters**:

We ended up with **4 clusters**. This was determined after analyzing the results of the **elbow method**, which showed that 4 clusters was the optimal number for this dataset.

2. **Clustering Metrics**:

• **DB Index (Davies-Bouldin Index)**:

The DB Index measures the average similarity ratio of each cluster to the most similar one. A lower value means better separation of the clusters. Our DB Index came out to **1.0121**. While this isn't super low, it suggests that our clusters are reasonably separated and compact.

• **Silhouette Score**:

The Silhouette Score gives an idea of how well each point fits into its own cluster compared to other clusters. It ranges from -1 (wrong cluster) to +1 (very well-clustered). For our clustering, the score is **0.2950**, which is moderate. It tells us that while the clusters are not super tight or highly distinct, they are still meaningfully separated.

• **Homogeneity**:

This score tells us whether each cluster contains only similar members, and ours is perfect at **1.0000**. This means that every customer within a cluster is very similar to each other, which is great for customer segmentation.

• **Completeness**:

The Completeness Score shows how well all members of the same class are grouped together in a cluster. Our score is also **1.0000**, which means every customer that should be in the same cluster is, with no errors.

Conclusion:

Overall, the **Agglomerative Clustering** method with 4 clusters gives us fairly good results. The clusters are reasonably compact and separated (as shown by the DB Index and silhouette score), and the homogeneity and completeness scores indicate that the segmentation is both coherent and accurate.