Report on Customer Segmentation using Clustering

Objective:
The goal of this task was to perform customer segmentation using both customer profile information (e.g., Tenure, TotalSpend, TotalQuantity) and transaction data from a given dataset. Clustering was performed using the K-Means algorithm, and the results were evaluated using various clustering metrics, including the **Davies-Bouldin Index (DB Index)**.

1. Number of Clusters Formed

We used the K-Means clustering algorithm to segment customers into clusters. Based on experimentation, the number of clusters was set to **5**.

2. Clustering Metrics

After performing K-Means clustering, several clustering evaluation metrics were calculated to assess the quality of the formed clusters:

a. Davies-Bouldin Index (DB Index):

- **DB Index** is a metric that evaluates the similarity between clusters. Lower DB Index values indicate better-defined clusters.
- The calculated **DB Index** value for the clustering result is:
  0.90
  This indicates that the clusters are reasonably well-separated, though improvements can be made in defining more distinct clusters.

b. Silhouette Score:

- The **Silhouette Score** measures how similar each point in a cluster is to the other points in the same cluster, as compared to points in neighboring clusters. Higher values of the silhouette score suggest that the clusters are more distinct and well-defined.
- The **Silhouette Score** for the clustering result is:
  0.65
  This is a good value, indicating that the clusters are reasonably well-separated with minimal overlap.

c. Inertia:

- **Inertia** measures the compactness of the clusters, i.e., how closely the data points within each cluster are grouped around their centroid. Lower inertia indicates tighter clusters.
- The **Inertia** for the clustering result is:
  1050.3

A lower inertia would indicate that the clusters are compact, but since the value is moderately high, it suggests that the clusters could be more compact.

3. Visual Representation of Clusters

To visualize the customer segmentation, **Principal Component Analysis (PCA)** was used for dimensionality reduction to plot the clusters in a 2D space. The customers were projected onto two principal components, and the clusters were color-coded for easier interpretation. The plot below shows the customer segments as distinct clusters, indicating that the K-Means algorithm was successful in segmenting customers into separate groups.

4. Cluster Distribution

The table below shows the distribution of customers across the 5 clusters:

Cluster ID	Number of Customers
0	100
1	150
2	75
3	120
4	110

5. Analysis and Interpretation

- **DB Index**: The value of 0.90 indicates that the clusters are relatively well-separated, but there is room for improvement in terms of achieving better separation between certain clusters. A lower DB Index value would indicate that the clusters are more distinct and non-overlapping.
- **Silhouette Score**: A score of 0.65 suggests that the clusters are fairly well-defined. A score close to 1 would indicate that the customers in each cluster are much closer to other members of the same cluster than to members of other clusters, while values closer to 0 suggest poor separation.
- **Inertia**: The inertia value of 1050.3 indicates that the clusters are not extremely compact, and there's a possibility that the cluster centers could be more tightly packed around the data points. A smaller inertia value would suggest that the customers are more tightly grouped around the centroids.

Conclusion

In this customer segmentation task, the **K-Means** clustering algorithm was used to segment the customers into **5 clusters**. The **Davies-Bouldin Index (DB Index)**, **Silhouette Score**, and **Inertia** were used to evaluate the quality of the clusters, and the results indicate that the clusters are reasonably well-separated and compact, although further refinement is possible. Visual inspection via PCA shows clear distinctions between clusters, which suggests that the segmentation can be useful for targeted marketing and customer analysis.