Customer Segmentation Report (Task 3)

1. Overview

In this task, we performed **customer segmentation** using clustering techniques based on both **profile information** from the `Customers.csv` and **transaction information** from the `Transactions.csv`. We applied the **KMeans clustering algorithm** to group customers into segments and evaluated the clustering performance using **Davies-Bouldin (DB) Index** and **Silhouette Score**. The goal was to segment customers into meaningful groups to provide insights for further analysis or targeted strategies.

2. Clustering Approach

We used the following approach for customer segmentation:

- **Algorithm**: KMeans clustering, chosen for its simplicity and efficiency in partitioning data into clusters.
- **Number of Clusters**: We selected **4 clusters**, as it provided a reasonable balance between model complexity and interpretability. The number of clusters could be adjusted based on further analysis (e.g., through methods like the Elbow Method, but for simplicity, we used 4 here).

  The chosen clusters were:

  - Cluster 0: High transaction value customers
  - Cluster 1: Low transaction frequency customers
  - Cluster 2: Mid-range transaction value customers
  - Cluster 3: High frequency, low value customers
- **Features Used for Clustering**:
  - **Total Transaction Value**: The sum of the monetary value of transactions by each customer.
  - **Quantity**: The total quantity of items purchased by each customer.
  - **Region**: The geographic location of the customer (one-hot encoded).

3. Clustering Evaluation

We evaluated the clustering results using the following metrics:

- **Davies-Bouldin (DB) Index**: This index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB index indicates better clustering performance (i.e., clusters are well-separated and compact).
  - **DB Index**: **0.808**
    The relatively low DB index indicates that the clusters formed are fairly distinct from one another, suggesting that the clustering algorithm worked well in separating different customer segments.

- **Silhouette Score**: The silhouette score measures how similar each point is to its own cluster compared to other clusters. It ranges from -1 to 1, with a higher score indicating better-defined clusters.
    - **Silhouette Score**: **0.527**
      This score suggests that the clusters are moderately well-defined. A higher silhouette score would indicate stronger clustering results.

4. Visual Representation of Clusters

We visualized the customer segments in two dimensions using **Principal Component Analysis (PCA)**, which reduces the dimensionality of the data for easier visualization while preserving as much variance as possible.

- **Cluster Visualization**: The plot below shows the segmentation of customers in the 2D PCA space. Each point represents a customer, and colors represent different clusters. The plot provides an intuitive sense of how customers are grouped based on their transaction behavior and profile.
    - The clusters appear to be well-separated, further confirming the effectiveness of the KMeans algorithm in creating meaningful segments.

5. Conclusion

- **Number of Clusters**: 4 customer segments were created, based on transaction value, quantity, and region.
- **Clustering Evaluation Metrics**:
    - **DB Index**: 0.808 (indicates reasonably good cluster separation).
    - **Silhouette Score**: 0.527 (indicates moderately well-defined clusters).
- **Visual Representation**: The PCA plot shows distinct clusters, suggesting that the KMeans algorithm successfully segmented customers into meaningful groups.

6. Future Considerations

- **Fine-Tuning**: The clustering result can be improved by experimenting with different numbers of clusters (using methods like the Elbow method) and using additional features such as customer demographics.
- **Actionable Insights**: The segmentation can guide targeted marketing strategies, personalized offers, or customer retention efforts based on the characteristics of each segment.

7. Output

- The clustering results are saved in the file `Customer_Segmentation.csv`, which includes the **CustomerID** and the assigned **Cluster** for each customer.