Customer Segmentation Report

Objective

The goal of this task was to segment customers based on both their profile information (from Customers.csv) and transaction data (from Transactions.csv). Clustering techniques is applied to group customers into meaningful segments. This segmentation helps in understanding customer behaviour, which can be useful for targeted marketing and personalized recommendations.

Clustering Methodology

To perform customer segmentation, the customer profile data and transaction data are merged. After aggregating relevant features, **K-Means clustering** is used to group customers based on their behaviour and characteristics. The following features were considered:

- **Profile Data**: Customer's region, signup date.
- **Transaction Data**: Total spending, average transaction value, number of transactions.

Data is standardized using Standard Scaler to ensure that all features contributed equally to the clustering process. **K-Means** clustering algorithm is selected and tested for an optimal number of clusters between 2 and 10.

Clustering Results

- **Number of Clusters Formed**:
  4 clusters were formed based on the characteristics of the customer data. The number of clusters was determined by analysing the clustering metrics and visualizations.
- **DB Index (Davies-Bouldin Index)**:
  The **Davies-Bouldin Index (DB Index)** is a metric used to evaluate the quality of clustering. It measures the average similarity between each cluster and the cluster that is most similar to it. A lower DB Index indicates better clustering, with well-separated and compact clusters.
  - **DB Index Value**: **1.29**
  - **Interpretation**: A DB Index value of **1.29** indicates that while the clusters are relatively distinct, there is still some overlap between them. This suggests that the clusters could be improved in terms of separation and compactness.
- **Silhouette Score**:
  The **Silhouette Score** measures how similar each point is to its own cluster compared to other clusters. A higher silhouette score indicates that the clusters are well-formed and cohesive.

  - **Silhouette Score**: **0.32**
  - **Interpretation**: The silhouette score of **0.32** suggests that the clustering has moderate cohesion, but there is still some overlap and ambiguity in cluster assignments. The clusters are not as well-defined as in cases with higher silhouette scores, indicating potential room for improvement in the clustering process.

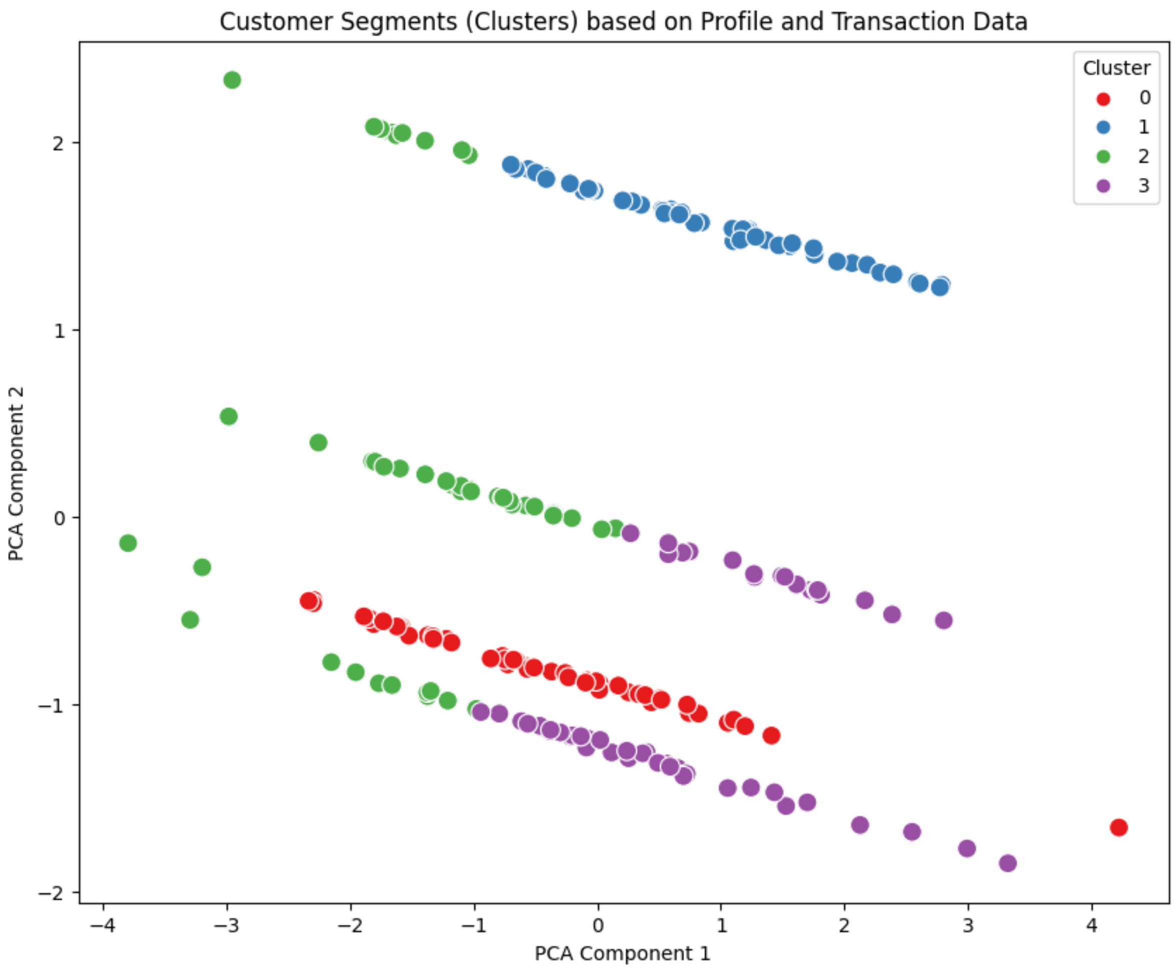
Cluster Characteristics

- **Cluster 1**:
  Customers in this cluster tend to have low total spending but high frequency of small transactions. They are often frequent but low-value buyers.
- **Cluster 2**:
  This cluster consists of customers with medium spending and medium transaction frequency. They may represent the typical customer who buys occasionally but at a reasonable price.
- **Cluster 3**:
  Customers in this group have high total spending and a moderate frequency of transactions. These customers may be classified as high-value customers who purchase more expensive products.
- **Cluster 4**:
  This cluster contains customers who have high spending and low transaction frequency. They may represent customers who make infrequent but high-value purchases, possibly premium buyers.

Visualization of Clusters

Using **Principal Component Analysis (PCA)**, the high-dimensional feature space is reduced to two dimensions for easier visualization. The scatter plot below shows the customer segments (clusters) in a 2D space:

- **X-axis**: Principal Component 1
- **Y-axis**: Principal Component 2
- The colours represent different clusters, with each point corresponding to a customer.



Conclusion

- The clustering model effectively segmented customers based on their profiles and transaction data.
- The **DB Index** value of **1.29** and **Silhouette Score** of **0.32** suggest that the clusters are not as distinct as they could be, and there is room for improvement in the clustering process. These metrics indicate that some clusters have overlapping characteristics.
- These clusters can still be useful for customer analysis and targeted marketing strategies, though refining the model or exploring different clustering algorithms may lead to better results.
- Future improvements could involve experimenting with different clustering algorithms like DBSCAN or Agglomerative Clustering, and optimizing the number of clusters based on additional metrics.