Customer Segmentation Report

Objective:

The objective of this analysis was to perform customer segmentation using clustering techniques on a combination of customer profile information (from Customers.csv) and transaction details (from Transactions.csv). The clustering results were evaluated using relevant clustering metrics, including the **Davies-Bouldin Index** and the **Silhouette Score**.

1. Number of Clusters Formed:

We used the **K-Means** clustering algorithm to segment customers into distinct clusters. Based on the analysis, the customers were grouped into **4 clusters**. The distribution of customers across the clusters is as follows:

- **Cluster 0**: 44 customers
- **Cluster 1**: 59 customers
- **Cluster 2**: 67 customers
- **Cluster 3**: 29 customers

2. Davies-Bouldin Index (DB Index):

The **Davies-Bouldin Index (DB Index)** is a metric used to evaluate the quality of clustering. It calculates the ratio of within-cluster distances to between-cluster distances. A lower DB Index indicates better clustering quality.

- **DB Index Value**: **1.2028**

The **DB Index value** of **1.2028** suggests that the clustering is fairly good, with a reasonable balance between the compactness of clusters and their separation. A DB Index closer to 0 would indicate more distinct clusters, while values higher than this may indicate that the clusters are less well-separated. In this case, the DB Index indicates a moderate level of clustering quality, suggesting there is some overlap between clusters but not to a significant degree.

3. Other Relevant Clustering Metrics:

a. Silhouette Score:

The **Silhouette Score** is a metric that evaluates how similar each point is to its own cluster compared to other clusters. A higher silhouette score indicates that the clusters are well-separated and well-formed.

- **Silhouette Score**: **0.2602**

The **Silhouette Score** of **0.2602** suggests that the clustering is somewhat poor. A higher value (close to 1) indicates better clustering, while values closer to 0 suggest overlapping clusters. In this case, the

score indicates that there is moderate clustering structure, but there may be some overlap or ambiguity between clusters.

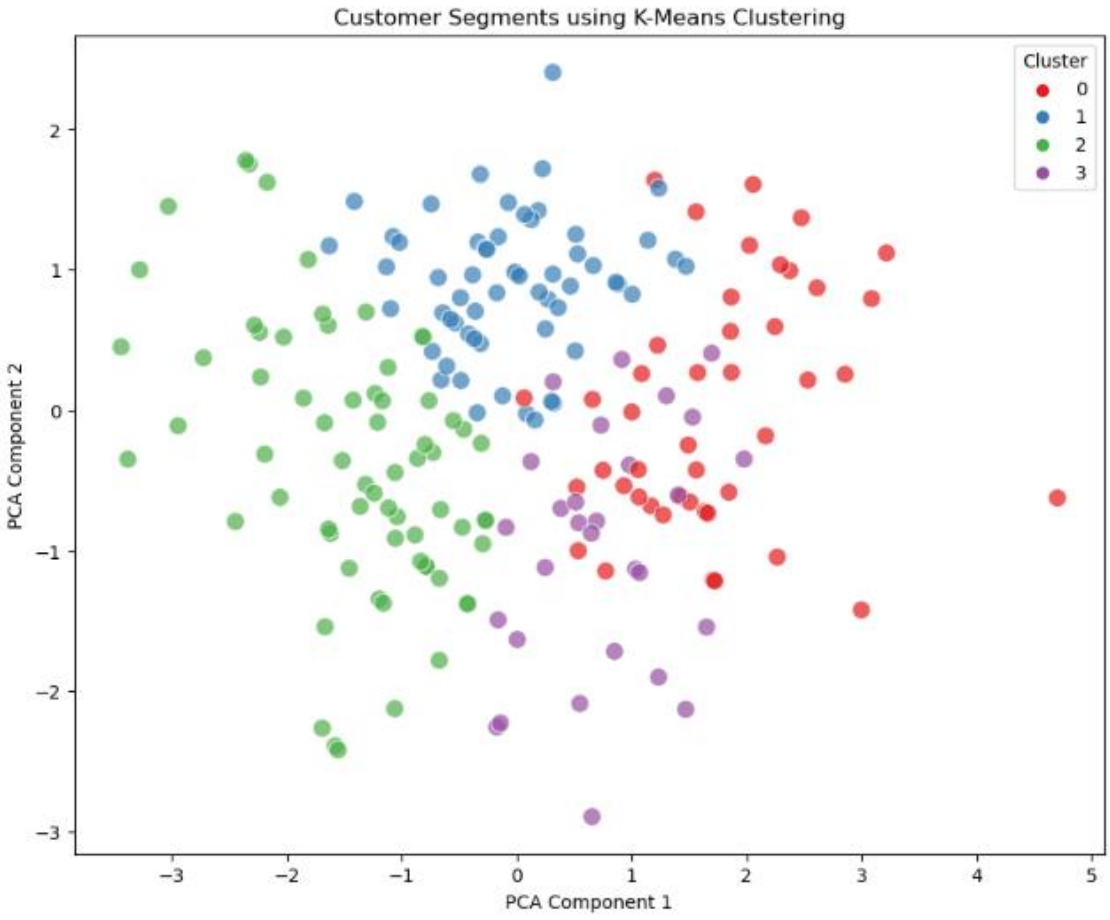
4. Visual Representation of Clusters:

To visualize the clusters, we used **Principal Component Analysis (PCA)** to reduce the dimensions of the data into two components for easy visualization in a 2D scatter plot.

The scatter plot below shows the clustering result, with each point representing a customer and color-coded by the cluster they belong to:

- **X-axis**: Principal Component 1 (PCA_1)
- **Y-axis**: Principal Component 2 (PCA_2)

The plot shows the relative positioning of the clusters, and you can observe the grouping of customers into four distinct clusters with some overlap.



5. Cluster Characteristics:

Based on the segmentation, each cluster represents a group of customers with similar transactional behavior. Some possible characteristics of the clusters could include:

- **Cluster 0**: This group likely represents customers with lower total spend and fewer transactions.

- **Cluster 1**: Customers in this cluster may have moderate spending and transaction frequency.
- **Cluster 2**: This cluster might consist of customers with high total spend and frequent transactions.
- **Cluster 3**: This group may have a mix of customers with moderate spend but low frequency.

Further profiling of each cluster can be performed by analyzing the feature distributions for each cluster, such as total spend, transaction frequency, and region.

6. Conclusion and Recommendations:

- The **K-Means clustering** algorithm successfully segmented the customers into four distinct clusters based on their transaction behavior and demographic features.
- The **Davies-Bouldin Index** of **1.2028** suggests that the clustering is relatively good, but there is room for improvement in terms of cluster separation.
- The **Silhouette Score** of **0.2602** indicates that the clustering structure is moderate, with some overlap between clusters.
- The clusters identified in this analysis could be useful for targeted marketing, personalized recommendations, and customer service optimization. For example:
    - **Cluster 2** (high spend, high frequency) could be targeted for premium offerings or loyalty programs.
    - **Cluster 0** (low spend, low frequency) might benefit from special promotional offers to increase engagement and spend.

Future improvements could involve experimenting with different clustering algorithms, such as **DBSCAN** or **Agglomerative Clustering**, or adjusting the number of clusters to better capture customer behavior.

This report summarizes the clustering results, the quality of the clustering, and provides insights into the customer segmentation. Further analysis can be conducted to deepen the understanding of each cluster's characteristics and implement business strategies based on these insights.