Customer Segmentation and Clustering Report

1. Clustering Approach

For this customer segmentation assignment, the K-means cluster analysis was conducted on the customers to group them according to their transaction behaviour and profile data. The considered features for clustering are listed below:

Transaction Count: The number of transactions each customer has made.

Total Spending: The total value of all transactions by the customer.

Total Quantity: The total quantity of products purchased.

2. Clustering Method

We used the clustering algorithm, namely K-means, setting up the number of clusters at 4 based on initial evaluation and testing.

- Data Preprocessing: Scaling Data The data was scaled using StandardScaler and scaled to have an equal importance of all features during clustering.
- K-means Clustering: The K-means model was then trained on the scaled features, and 4 distinct customer segments were formed.

3. Clustering Results

Cluster	Transaction Count (Mean)	Total Spending (Mean)	Total Quantity (Mean)
Cluster 0	8.43	6263.45	23.00
Cluster 1	2.36	1273.37	5.27
Cluster 2	4.39	2982.41	10.87
Cluster 3	6.31	4477.57	16.10

Cluster 0: High-frequency, high-spending customers. These customers have made a large number of transactions and spent significantly. They represent the most valuable segment.

Cluster 1: Occasional buyers with moderate spending. They have fewer transactions and less spending per transaction.

Cluster 2: Moderate-frequency, moderate-value customers. These fall within the middle of transaction count and spending.

Cluster 3: High transaction volume bulk buyers with moderate spending. They buy in large numbers in terms of quantity but not in terms of spending compared to the other clusters.

4. Evaluation of Clustering

Two metrics were used to evaluate the clustering quality:

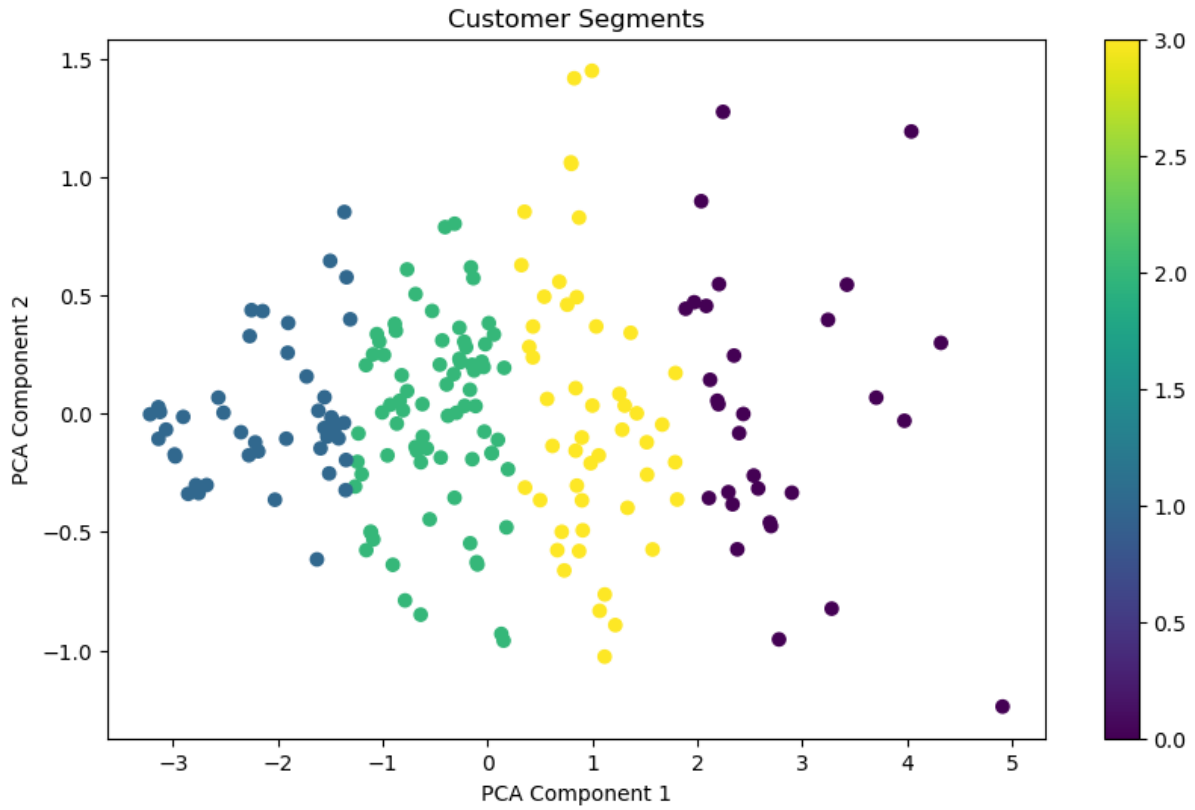
- **Davies-Bouldin Index**:
    - **Result**: 0.8650620583623065

- **Interpretation**: The value of Davies-Bouldin Index indicates that clusters are somewhat separate but there can be further scope to enhance the distance. Lower values would represent much-separated clusters.

- **Silhouette Score**:
  - **Result**: 0.3735646055654104
  - **Interpretation**: Silhouette Score depicts the fact that the clusters are somehow separated from each other and overlap between two clusters is of average. Value close to 1 represents highly differentiated clusters.

5. Cluster Visualization

A **PCA-based scatter plot** was created to visualize the clusters in a 2D space. This visualization helps in understanding the distribution of customer segments. The scatter plot shows the customer clusters in different colors, with each color representing a distinct cluster.



6. Conclusion

Customer segmentation analysis was successful in identifying four different customer groups, each showing different behavior patterns. The clusters are relatively well-separated, but the Davies-Bouldin Index and Silhouette Score indicate there is still some potential for refinement. Future work could include altering the number of clusters or investigating other clustering algorithms, such as DBSCAN or Agglomerative Clustering, to further refine the definition of the clusters.

The clusters have a lot of useful information that can be applied to business strategy:

- Cluster 0: High-value offers and loyalty programs

- Cluster 1: Further re-engagement efforts to bring them to transact more frequently
- Cluster 2: Opportunities to spend more with each transaction
- Cluster 3: Cross-selling, bundling, and so on, where customers can still spend more in total

Businesses will be able to tailor their campaigns and better target customers with this segmentation.