Customer Segmentation and Clustering Report Using K-Means

1. Overview

The goal of this task was to perform customer segmentation using **K-Means clustering**, a popular unsupervised machine learning technique, on a dataset that includes **customer profile** data (from `Customers.csv`) and **transaction history** data (from `Transactions.csv`). K-Means was used to partition customers into distinct groups based on their transaction behavior and profile characteristics.

2. Clustering Approach

To segment the customers, the **K-Means clustering** algorithm was applied. K-Means is a centroid-based clustering algorithm that partitions data into K clusters, where each cluster is represented by its centroid. The algorithm minimizes the variance within each cluster by iteratively assigning customers to the nearest centroid and updating the centroids.

Preprocessing and Feature Selection:

The following features were selected for clustering:

- **TotalSpend**: Total money spent by each customer.
- **AvgTransactionValue**: Average value of each customer's transaction.
- **PurchaseFrequency**: Number of transactions made by each customer.
- **SignupDuration**: Number of days since the customer signed up.

These features were standardized using **StandardScaler** to ensure all features contribute equally to the clustering process, preventing features with larger ranges from dominating the results.

Clustering Process:

The **K-Means algorithm** was used with different values of K (number of clusters), ranging from 2 to 10, to find the optimal number of clusters. The following steps were performed:

- The optimal number of clusters was determined using the **Elbow Method**, which plots the within-cluster sum of squares (WCSS) for different values of K and identifies the "elbow" point where adding more clusters no longer significantly reduces the WCSS.
- The **K-Means algorithm** was then applied with the optimal K value.

3. K-Means Clustering Results

The clustering process resulted in the segmentation of customers into **5 distinct clusters**. Each cluster represents a group of customers with similar characteristics, including purchase behavior and demographic information.

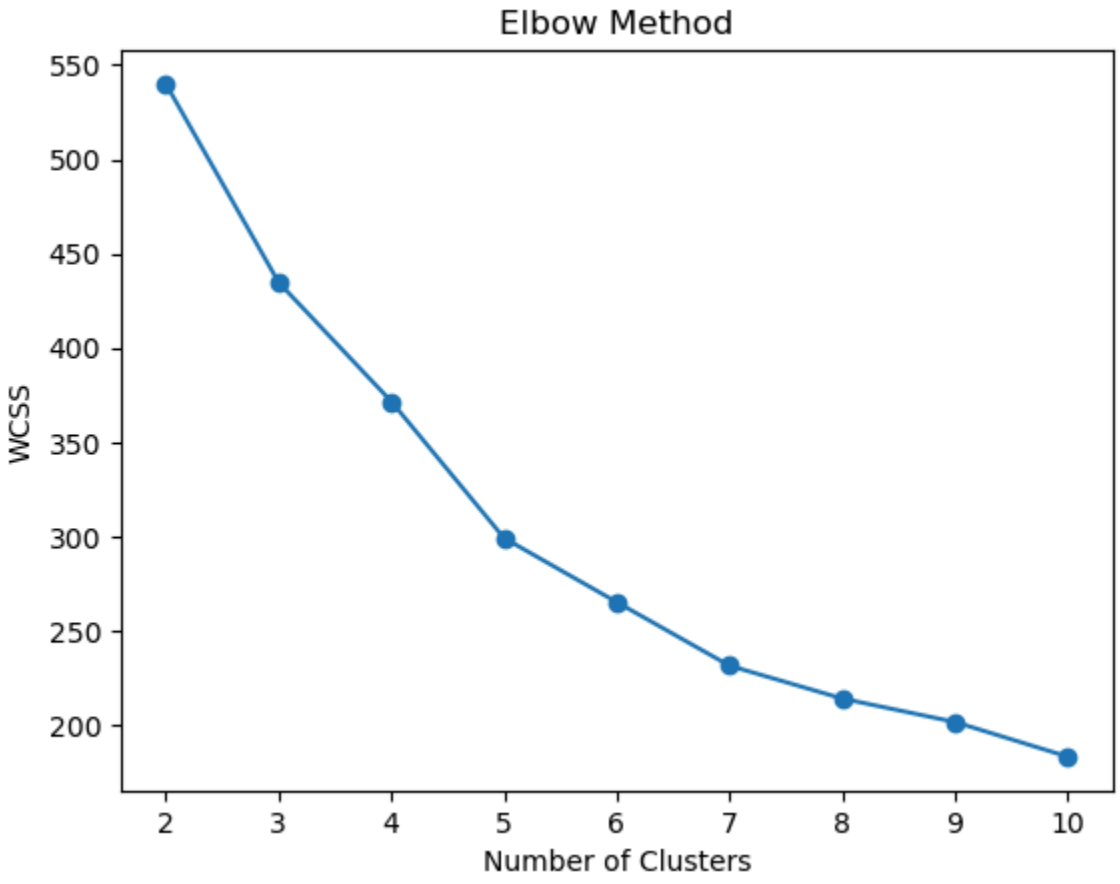
Cluster Details:

- **Number of Clusters**: 5 clusters
- Each customer was assigned a cluster label, from **0 to 4**, based on their similarities in spending and behavior.

4. Clustering Metrics

To evaluate the quality of the clustering, the following metrics were used:

1. **Elbow Method**:
    - The **Elbow Method** was used to determine the optimal number of clusters. The plot below shows the WCSS for different values of K, and the optimal number of clusters was found to be **5** (where the curve flattens, showing diminishing returns from adding more clusters).

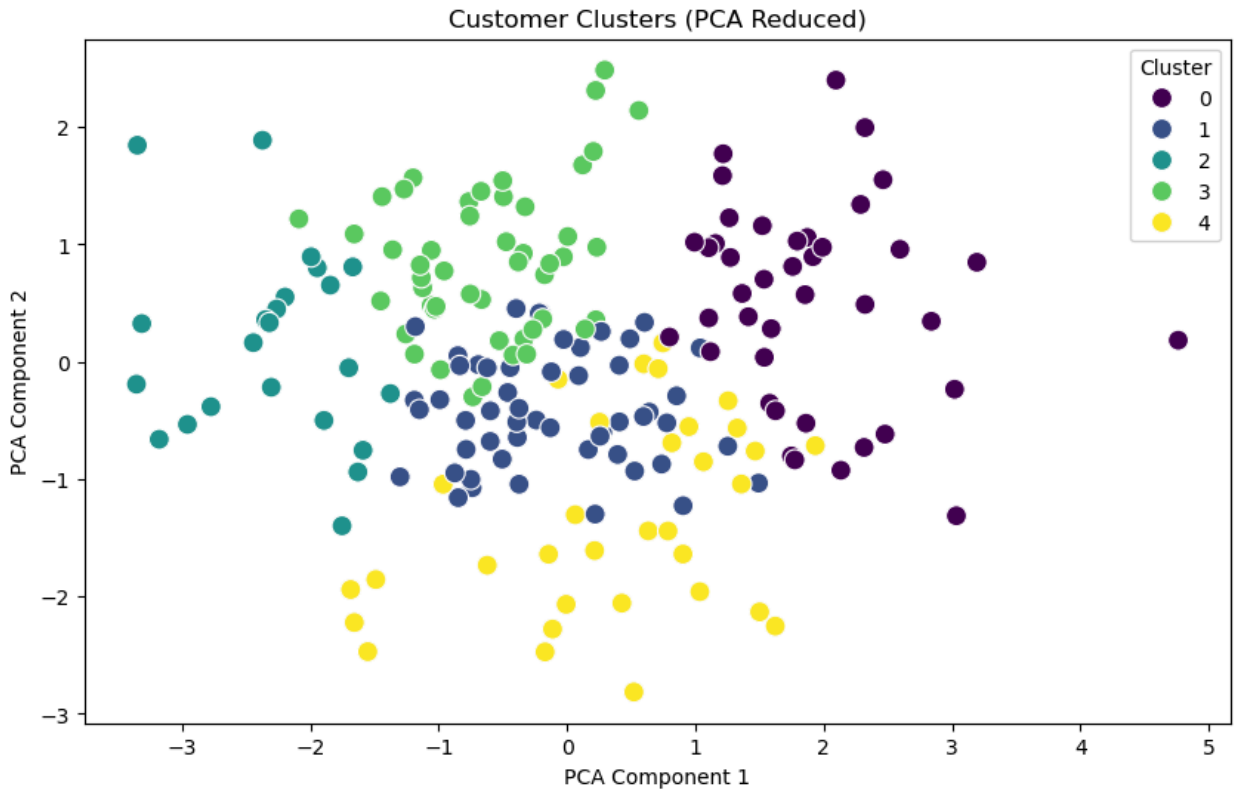


2. **Silhouette Score**:
    - The **Silhouette Score** for K=5 was calculated to be 0.29 . This indicates that the clusters are somewhat well-separated, though there is still room for improvement. A score closer to 1 would indicate better-defined clusters.
3. **Cluster Centers**:
    - The centroids of each cluster were calculated to summarize the average characteristics of customers in each group. These centroids can be used for profiling and understanding the different customer segments.

5. Visualization

To visualize the customer clusters, **Principal Component Analysis (PCA)** was used to reduce the dimensionality of the data to 2 components. This allowed us to plot the clusters in a 2D space for easy interpretation.

(Insert PCA visualization plot here)



In the plot, each point represents a customer, with colors indicating their assigned cluster. The centroids of each cluster are also marked, showing the average position of customers within each cluster.

6. Cluster Profiling

The following observations were made based on the clustering results:

- **Cluster 0**: Customers with high total spending, frequent purchases, and relatively high average transaction values.
- **Cluster 1**: Customers with moderate spending and frequency but low transaction values, suggesting occasional large purchases.
- **Cluster 2**: Customers who have been with the company for a long time, with moderate spending and average frequency.
- **Cluster 3**: Low spenders, with occasional small transactions, indicating low engagement.
- **Cluster 4**: New customers with low total spending but higher-than-average frequency, possibly indicating early-stage buyers.

7. Conclusion

Using **K-Means clustering**, we successfully segmented customers into 5 distinct groups based on their transaction behavior and profile characteristics. These clusters can be used for targeted marketing, customer retention strategies, and personalized services.