CLUSTERING REPORT

Objective of the assignment

The goal of this clustering exercise was to segment customers based on their profile and transaction data to uncover patterns and create actionable groups. This involved:

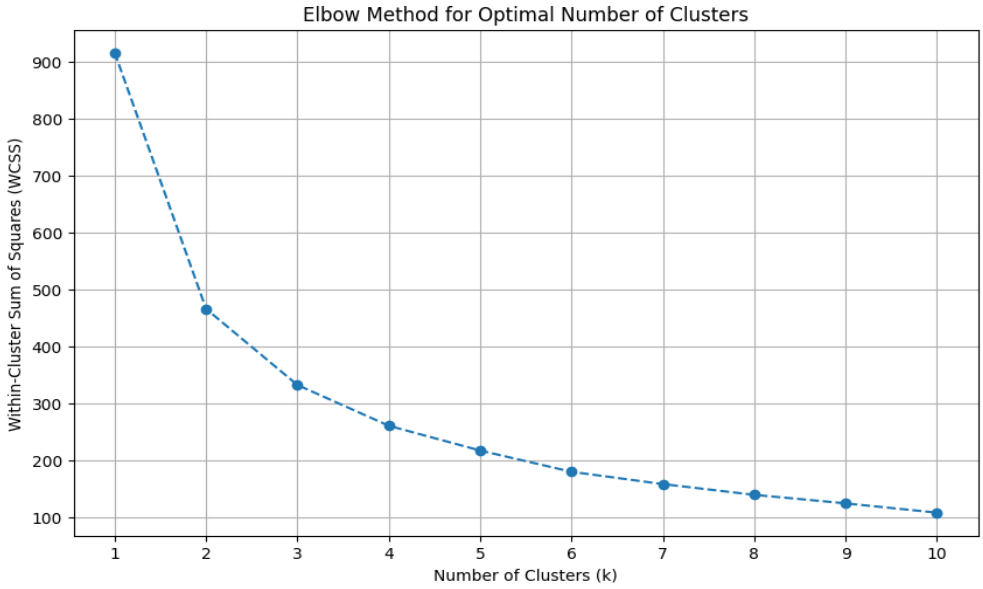
1. Determine the optimal number of clusters to differentiate customers.
2. **DB Index** and **Silhouette Score** are the evaluation criteria used in this assignment.
3. Create visualizations to get a better understanding of the customers clusters.

Data Preparation

- **CSV Files used are**:
    - Customers.csv
    - Transactions.csv
- **Features of the customers used**:
        - Total spending done by the customer.
        - Average transaction value.
        - Total quantity purchased during the transaction.
        - Number of unique products.
        - Number of transactions done by the customer.
    - These features were normalized using **StandardScaler** to ensure the accuracy and compatibility of the clustering model.

Methodology

1. **Clustering Algorithm**:
    - K-Means clustering was used for its efficiency and effectiveness on numerical dataset.
2. **Elbow Method**:
    - The Elbow Method was used to determine the optimal clusters count. The optimal cluster range was exists between **4 to 6** clusters.



3. **Metrics used for Evaluations**:
    - **Davies-Bouldin Index (DB Index)**: Lower is the value of DB Index, better the clusters formed.
    - **Silhouette Score**: Higher value signifies better cluster formation.

Results

Number of Clusters Formed

On applying the metrics and visualizing the elbow method graphs, the **range of 4 - 6 comes out to be the optimal range** for cluster formation.

Evaluation Metrics

- **Davies-Bouldin Index**: **0.898**
    - Indicates well-separated clusters with minimal overlap.
- **Silhouette Score**: **0.348**
    - Suggests moderately compact and well-separated clusters.

Cluster Composition

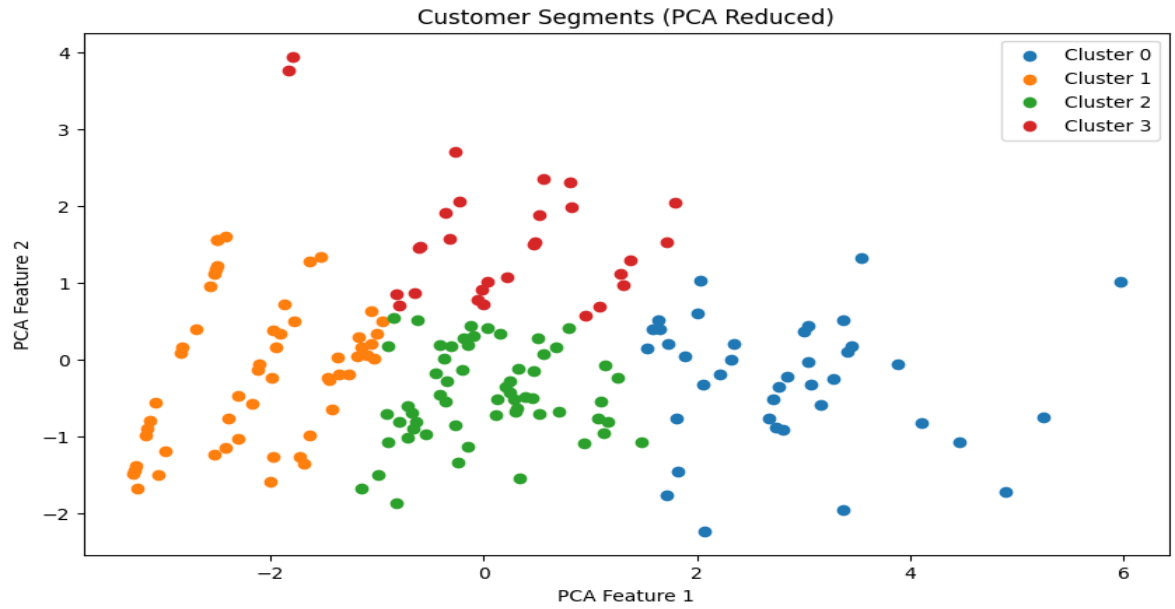
Each cluster exhibited distinct characteristics:

1. **Cluster 0:** High-value, frequent buyers with diverse product preferences.
2. **Cluster 1:** Low-value, infrequent buyers with minimal product diversity.
3. **Cluster 2:** Moderate spenders with steady transaction frequency and decent variety.
4. **Cluster 3:** High spenders with lower purchase frequency and limited product variety.

Visualization of the Clusters

The clusters were visualized using **scatter plots** for interpretability. Higher Dimensionality was reduced to 2-dimensional plane using Principal Component Analysis.

Each cluster was differentiable with others clusters showcasing the effectiveness of the clustering algorithm.



Insights from the clusters

Cluster	avg_total_spending	avg_transaction_value	avg_total_quantity	avg_num_products	avg_num_transactions
0	5550.087692	726.686323	20.102564	7.487179	7.769231
1	1117.736786	553.316012	4.446429	1.964286	2.000000
2	2698.781695	581.205119	11.152542	4.610169	4.677966
3	3860.410690	1069.830259	11.172414	3.620690	3.724138

1. **Cluster 0 (High-Value, Engaged Customers)**
    - Highest spending, most frequent transactions, generating most of the sales.
    - Should provide exclusive offers or promo codes for their loyalty.

2. **Cluster 1 (Low-Value, Less frequent transactions)**
    - Lowest spending, least transactions, not buying multiple products.
    - Target them with personalized coupons and offers for boosting sales.

3. **Cluster 2 (Moderate Spenders)**
    - Moderate spending, steady transaction frequency, buying an average variety of products.
    - Encourage higher spending via upselling or cross-selling strategies.

4. **Cluster 3 (High-Spenders with Lower Frequency of transactions)**
    - High transaction value but less frequent purchases and buying less diverse products.
    - Offer subscription models or discounts to encourage repeat purchases.