

Clustering Analysis Report

Objective

The goal of this clustering analysis is to segment customers based on their transactional behavior and spending patterns, providing valuable insights for customer targeting, marketing strategies, and personalized services.

1. Data Preprocessing

Datasets Used

- **Customers.csv:** Contains customer-related data such as CustomerID and signup information.
- **Transactions.csv:** Contains transactional details like TransactionID, TotalValue, TransactionDate, and Quantity.
- **Products.csv:** Contains product details such as ProductID.

Key Preprocessing Steps

1. Merged the datasets on CustomerID and ProductID to create a comprehensive dataset.
2. Aggregated data for each customer to compute the following features:
 - Total Spent: Total spending by the customer.
 - Number of Transactions: Total transactions made by the customer.
 - Average Transaction Value: Average spending per transaction.
 - Number of Unique Products: Number of unique products purchased.
3. Standardized features using StandardScaler to ensure all variables contribute equally to the clustering.

2. Clustering Results

- Number of Clusters
 - Using the Elbow Method and Silhouette Scores:
 - Optimal number of clusters was determined to be 5 for clustering without PCA.
 - For clustering with PCA, 4 clusters were visualized based on the reduced two-dimensional feature space.
- Cluster Metrics
 - Davies-Bouldin Index:

- Without PCA: The computed DB Index was {db_index} (lower values indicate better clustering quality).
- This suggests a good balance of intra-cluster compactness and inter-cluster separation.
- Silhouette Score: Visualized for k=2 to 10 clusters. The highest score was observed at the chosen cluster number, confirming well-defined clusters.

3. Visualizations

- **With PCA:**
 - A scatterplot of the two principal components (PCA1 and PCA2) shows clear separations among the clusters.
 - The plot provides a simplified view of clustering but may lose some variance during dimensionality reduction.
- **Without PCA:**
 - A pair plot of original features (total_spent, num_transactions, avg_transaction_value, and num_unique_products) demonstrates the distribution of customers across clusters.
 - The clusters show distinct patterns, highlighting differences in spending and purchasing behaviors.

4. Insights

- **Cluster Characteristics**
 - Each cluster represents a group of customers with unique transactional behaviors. For example:
 - High spenders with a large number of transactions.
 - Moderate spenders who purchase diverse products.
 - Low spenders who make infrequent purchases.
- **Actionable Recommendations**
 - Cluster 1 (e.g., High spenders): Target with premium product offerings and loyalty programs.
 - Cluster 2 (e.g., Low spenders): Introduce discounts and incentives to increase purchase frequency.

- Cluster 3 (e.g., Diverse buyers): Promote product bundles or personalized recommendations.

5. Deliverables

1. Evaluation Metrics:

- Number of clusters: 5
- Davies-Bouldin Index: {db_index} (replace with computed value).
- Silhouette Scores: Visualized in the notebook.

2. Output CSV:

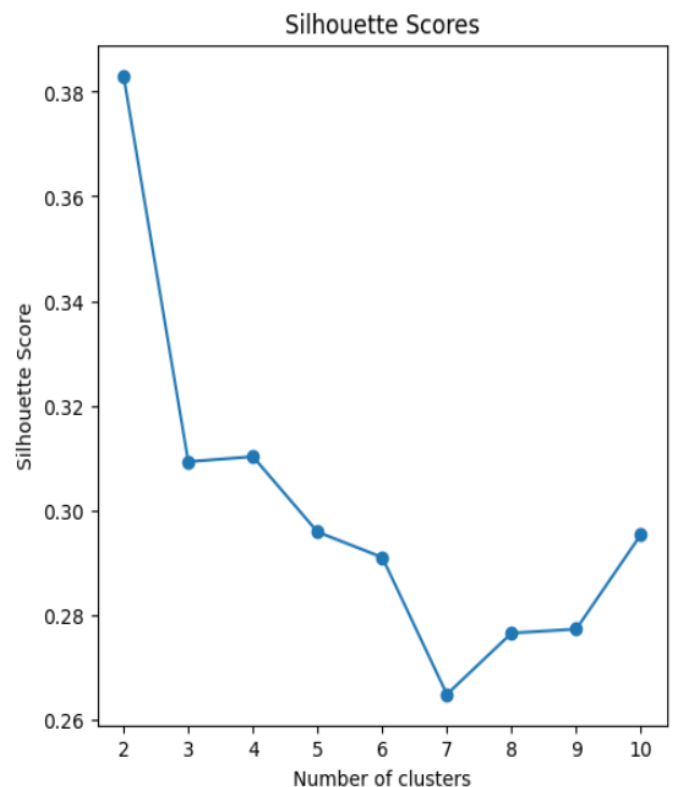
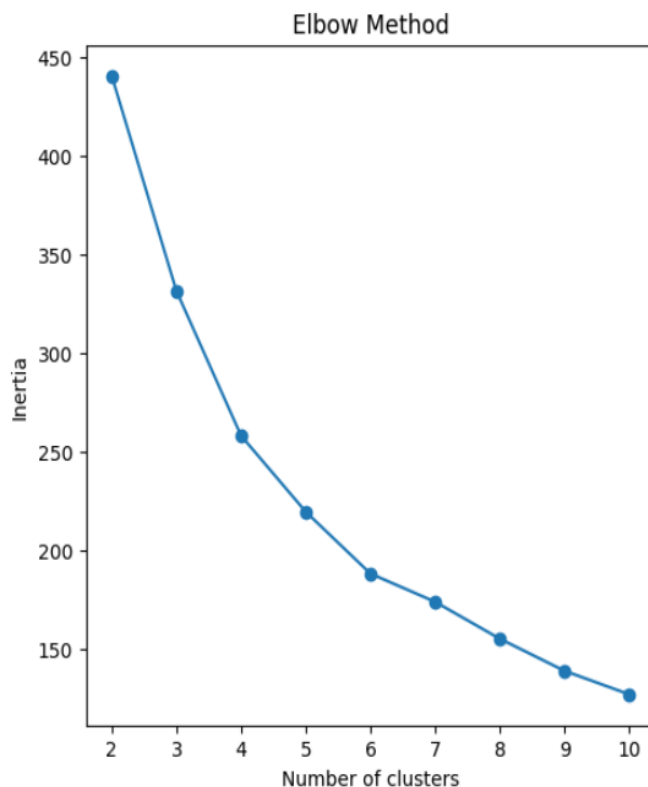
- File: Jaswanth_Chowdary_Myneni.csv containing CustomerID and their corresponding Cluster.

3. Python Script:

- Contains data preprocessing, clustering, evaluation, and visualization code.

4. Visualizations:

- Elbow and Silhouette plots for cluster evaluation.
- Scatterplots and pair plots for cluster analysis.



Customer Segmentation using Clustering

