Data Science Assignment: eCommerce Transactions Dataset

Task 3: Customer Segmentation / Clustering

- Perform customer segmentation using clustering techniques. Use both profile information
- (from Customers.csv) and transaction information (from Transactions.csv).
- You have the flexibility to choose any clustering algorithm and any number of clusters in
- between(2 and 10)
- Calculate clustering metrics, including the DB Index(Evaluation will be done on this).
- Visualise your clusters using relevant plots.

Deliverables:

- A report on your clustering results, including:
    - The number of clusters formed.
    - DB Index value.
    - Other relevant clustering metrics.
- A Jupyter Notebook/Python script containing your clustering code.

Evaluation Criteria:

- Clustering logic and metrics.
- Visual representation of clusters.

Objective

The goal of this project is to perform customer segmentation using clustering techniques based on both profile information (e.g., region) and transaction details (e.g., total spend, number of transactions, and average transaction value). This enables better understanding of customer behaviour and aids in targeted marketing strategies.

Steps Involved in Building the Solution

1. **Data Collection:**
    - Loaded customer profile data from Customers.csv.
    - Merged it with transaction data from Transactions.csv using CustomerID.
2. **Feature Engineering:** Created derived features:
        - Total Spent: Sum of transaction values for each customer.
        - Number of Transactions: Count of transactions per customer.
        - Average Transaction Value: Average value of transactions per customer.
    - Encoded the Region column as numerical data for clustering.
3. **Outlier Removal:** Applied the Interquartile Range (IQR) method to remove outliers
4. **Data Scaling:** Scaled features (total_spent, num_transactions, avg_transaction_value, Region) using StandardScaler to normalize data.
5. **Dimensionality Reduction:** Applied Principal Component Analysis (PCA) to reduce the feature set while retaining most of the variance.
6. **Clustering:**
    - Performed K-Means Clustering with the number of clusters ranging from 2 to 10.
    - Evaluated clustering performance using Davies-Bouldin Index and Silhouette Score.

7. **Optimal Cluster Selection:** Identified the optimal number of clusters (4) based on the lowest Davies-Bouldin Index value.
8. **RFM Analysis:** Conducted Recency, Frequency, Monetary (RFM) Analysis to categorize customers further:
   - Recency: Time since the last transaction.
   - Frequency: Number of transactions.
   - Monetary Value: Total amount spent.

9. **Visualization:**

   - Used pair plots, a 3D scatter plot, and PCA plots to visualize cluster distributions.
   - Generated a feature importance plot using a Decision Tree Classifier.

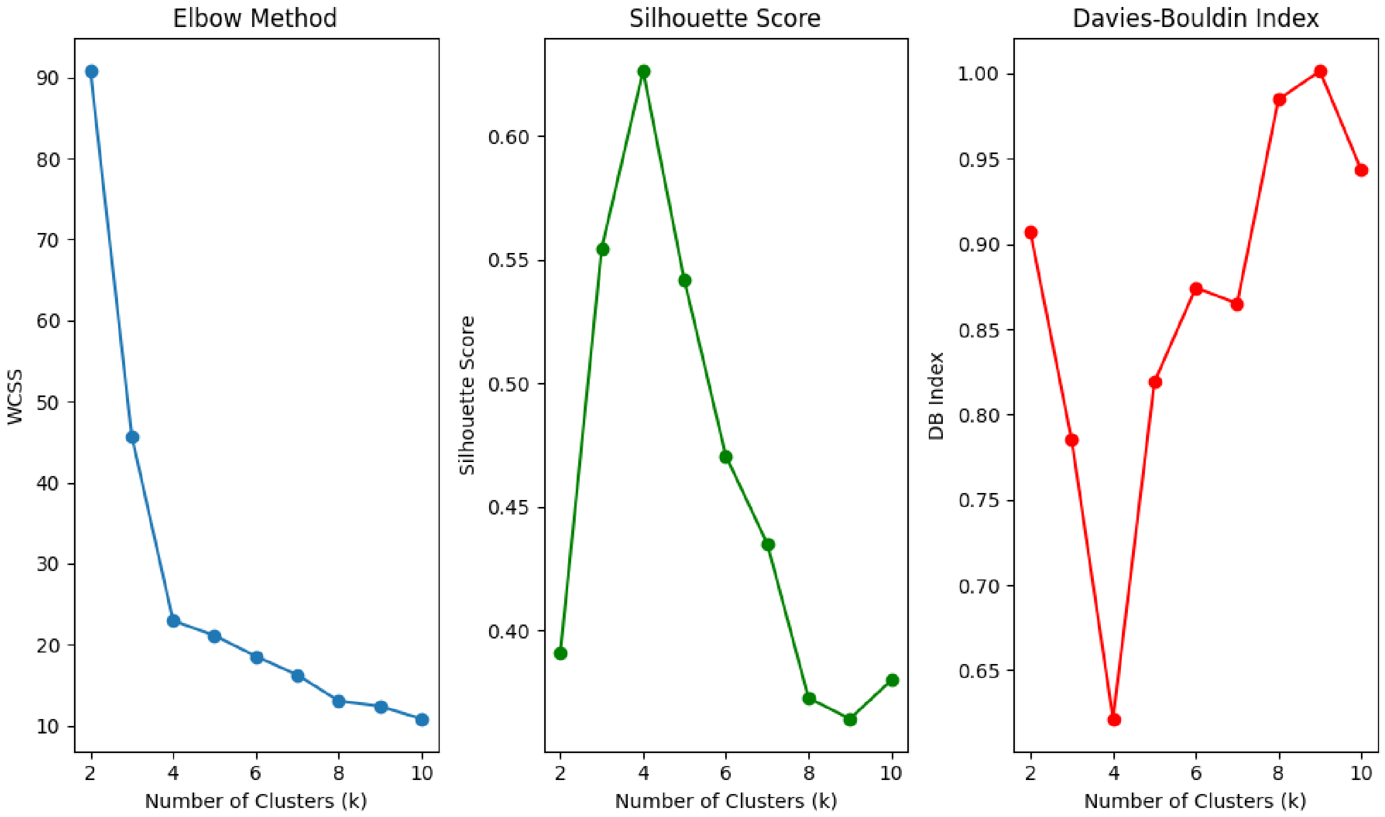
Optimal Results

- The optimal number of clusters was determined by minimizing the Davies-Bouldin Index.
- **Optimal Number of Clusters:** 4
- **Davies-Bouldin Index:** 0.6215427171683032
- **Silhouette Score:** 0.6263578055834661

Conclusion

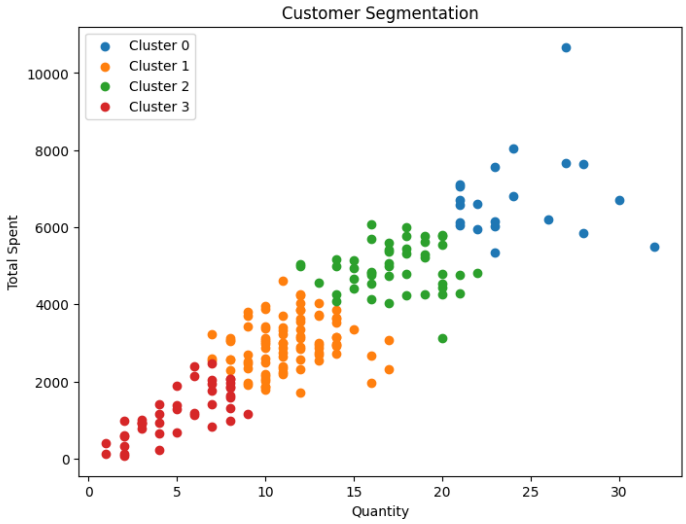
- The clustering analysis successfully segmented customers into distinct groups based on transactional and profile data.
- The optimal number of clusters (4) was chosen based on the Davies-Bouldin Index.
- Insights from clustering can help in designing personalized marketing strategies and improving customer engagement.

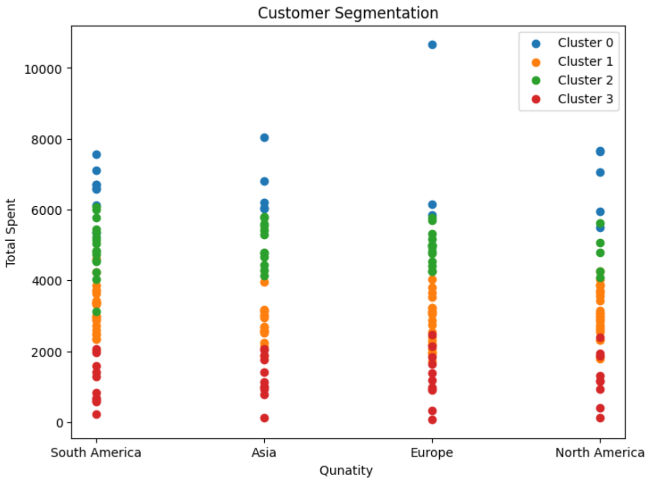
Result Images



DB Index: 0.72127971818163

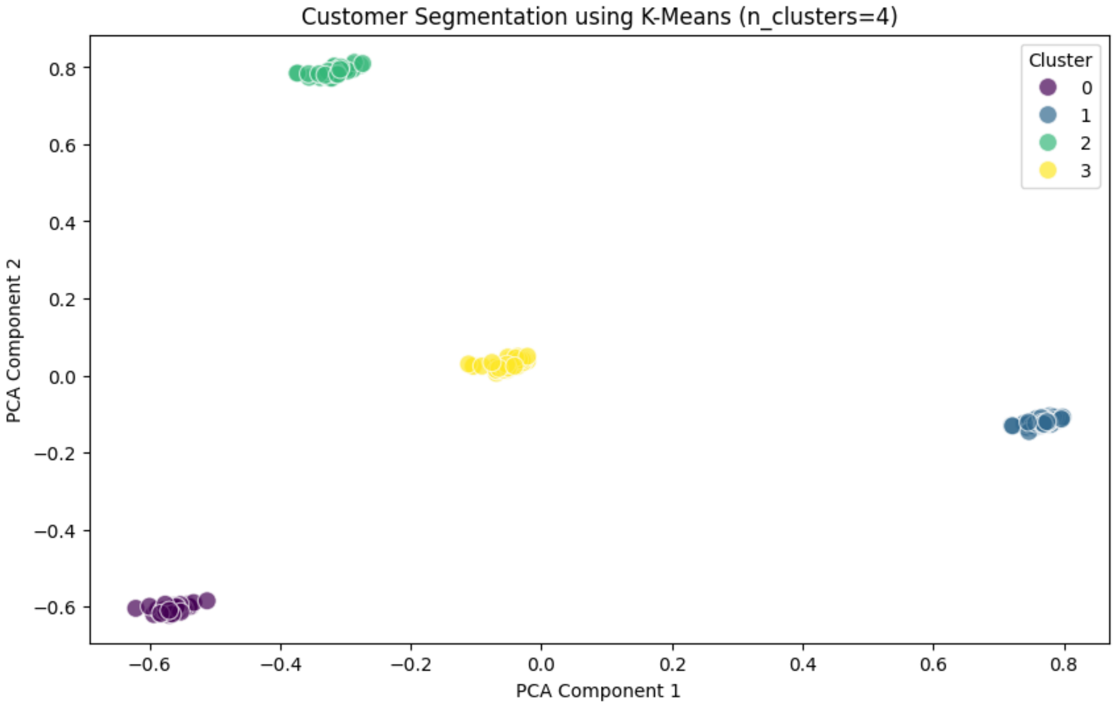
Silhouette Score: 0.4497501059000638





Best Silhouette Score: 0.6263578055834661

Davies-Bouldin Index: 0.6215427171683032



Srikal Kakula

Srikalkakula@gmail.com