Clustering Results Report

By:- Sanya Arora

1. Introduction

This report provides an in-depth analysis of the customer segmentation performed using K-means clustering. The goal of this analysis is to identify distinct customer segments based on transactional behavior and demographic attributes, which can be leveraged for targeted marketing, personalized offers, and strategic business decisions.

2. Methodology

2.1 Data Sources

- **Customers Dataset:** Contains customer demographic information.
- **Products Dataset:** Product-related details.
- **Transactions Dataset:** Records of customer transactions.

2.2 Feature Engineering

The following features were derived for clustering:

- **Total Spending:** Sum of transaction values.
- **Frequency:** Number of transactions per customer.
- **Average Transaction Value:** Mean value of transactions.
- **Recency:** Number of days since the last purchase.

2.3 Data Preprocessing

- Date columns converted to datetime format.
- Aggregation of transactional data.
- Standardization of features using StandardScaler.

2.4 Clustering Algorithm

K-means clustering algorithm was applied to the scaled features to identify distinct customer segments.

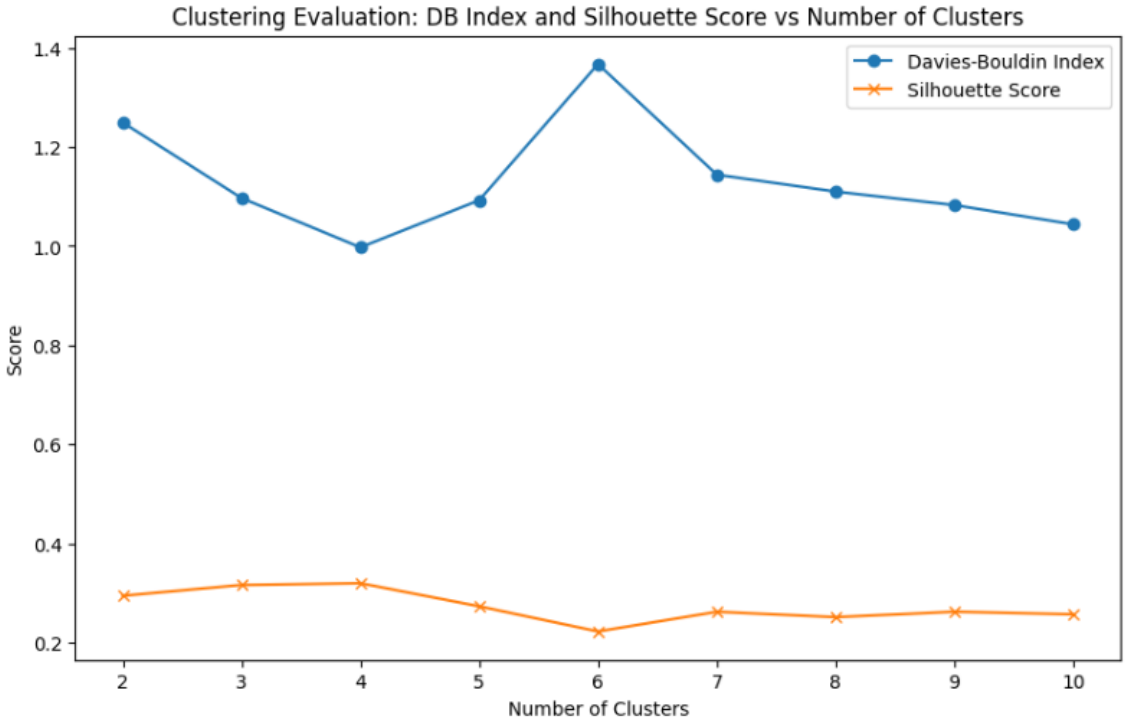
3. Clustering Evaluation Metrics

3.1 Optimal Number of Clusters

The optimal number of clusters was determined using the Davies-Bouldin (DB) Index and Silhouette Score.

- **Optimal Number of Clusters:** 4

Visualization: plot showing DB Index and Silhouette Score against different cluster numbers.

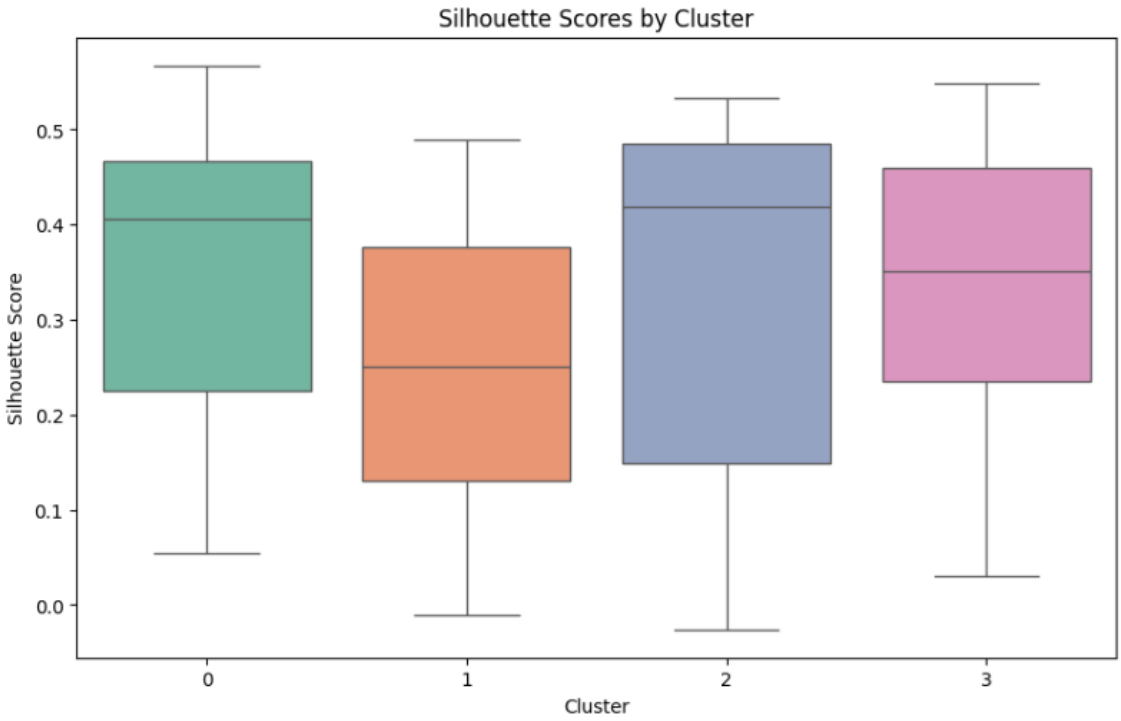



```
Optimal number of clusters: 4
DB Index: 0.9975879026509441
Silhouette Score: 0.32029123091701694
```

3.2 Evaluation Metrics

- **Davies-Bouldin Index:** 0.9976 (lower is better, indicates compact and well-separated clusters).
- **Silhouette Score:** 0.3203 (higher indicates better-defined clusters).
- **Calinski-Harabasz Index:** 101.4487 (higher is better, indicates dense and well-separated clusters).

Visualization: Boxplot of Silhouette Scores by Cluster.



:

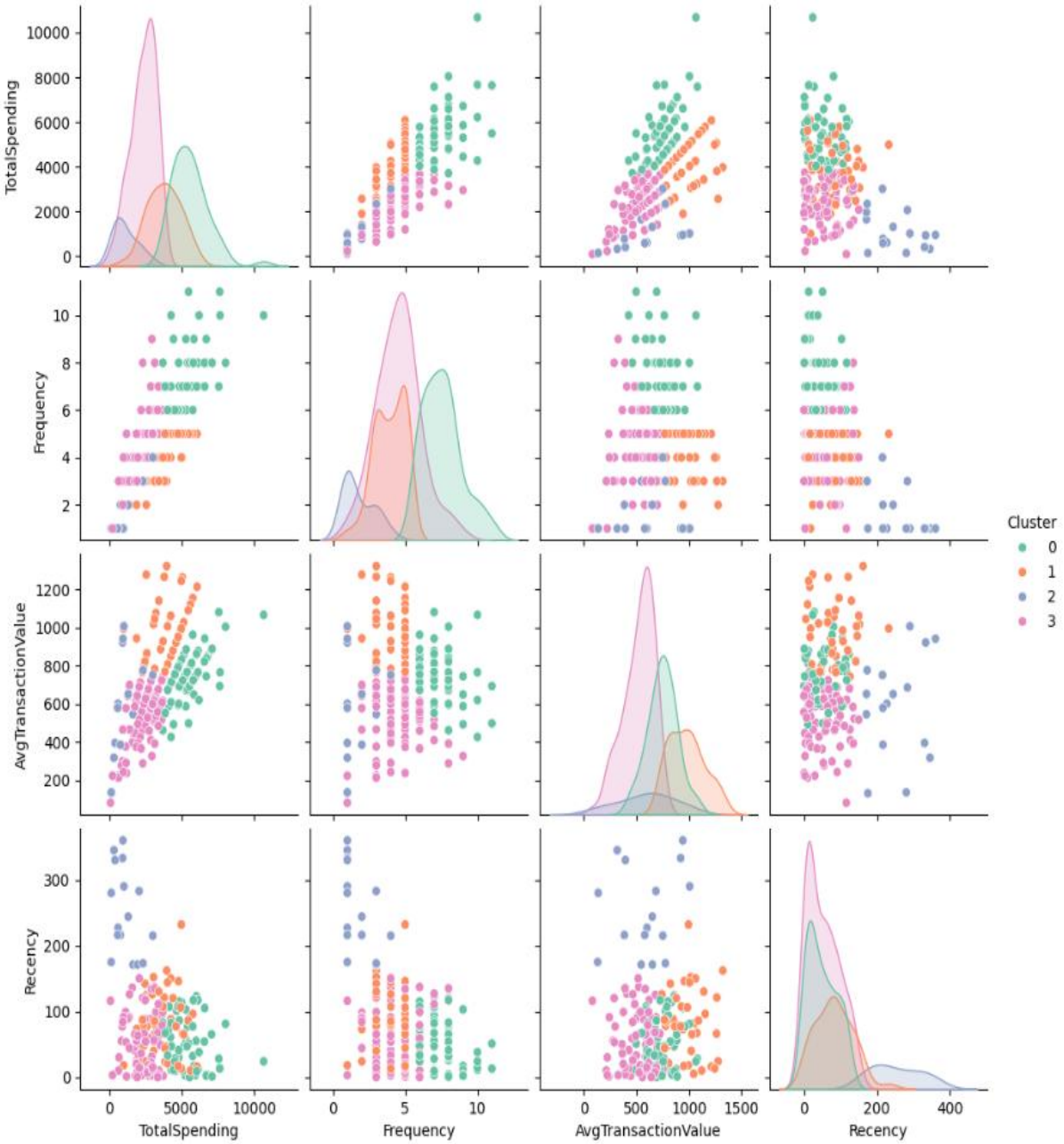
4. Clustering Results

4.1 Cluster Centers

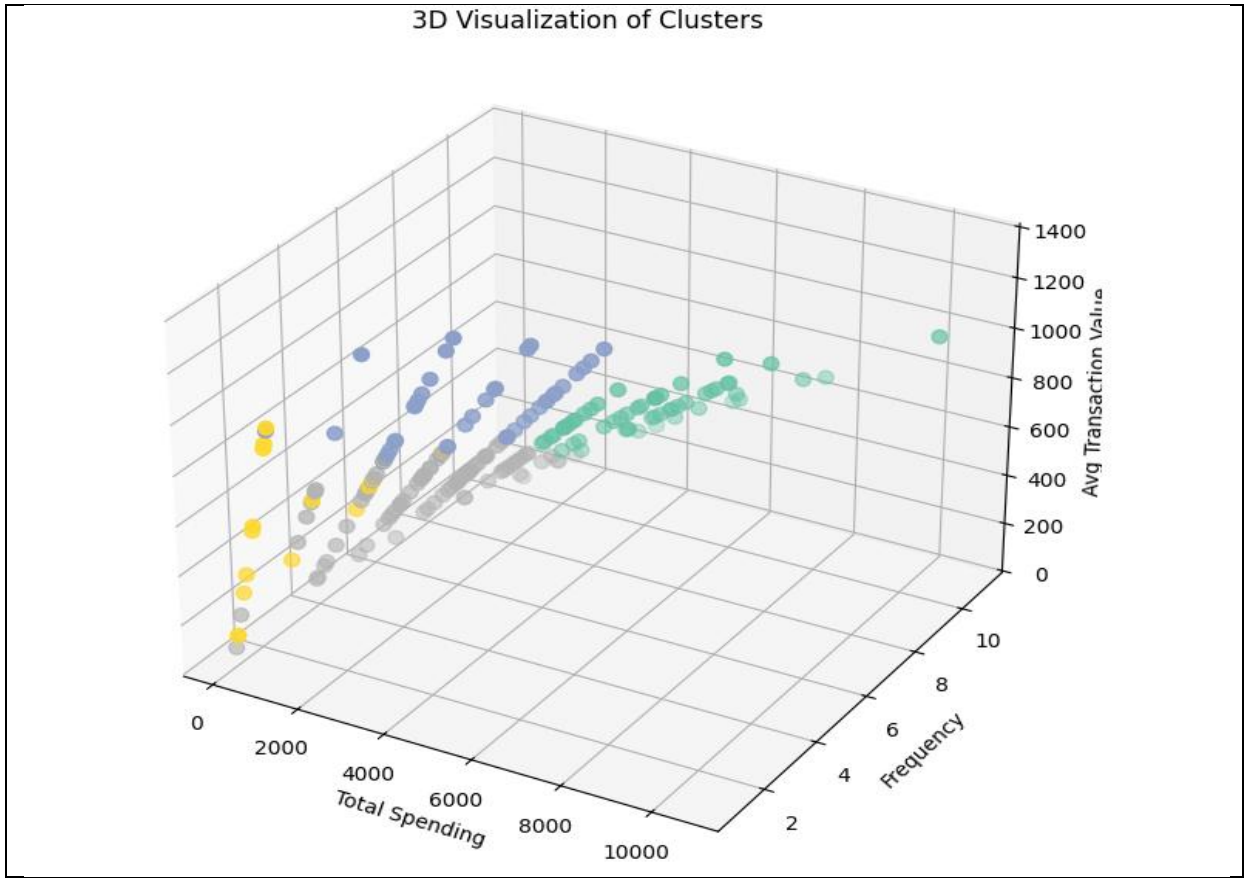
The following table shows the mean values of features for each cluster:

Cluster	Total Spending	Frequency	Avg Transaction Value	Recency
0	1.159981	1.153062	0.251799	-0.349450
1	0.177620	-0.510926	1.221203	0.125380
2	-1.277299	-1.465073	-0.396918	2.514288
3	-0.611555	-0.231428	-0.694595	-0.305004

Visualization: Pairplot of clusters across features.



Visualization: 3D scatter plot for Total Spending, Frequency, and Avg Transaction Value.



4.2 Cluster Sizes

Cluster Number of Customers

0 56

1 42

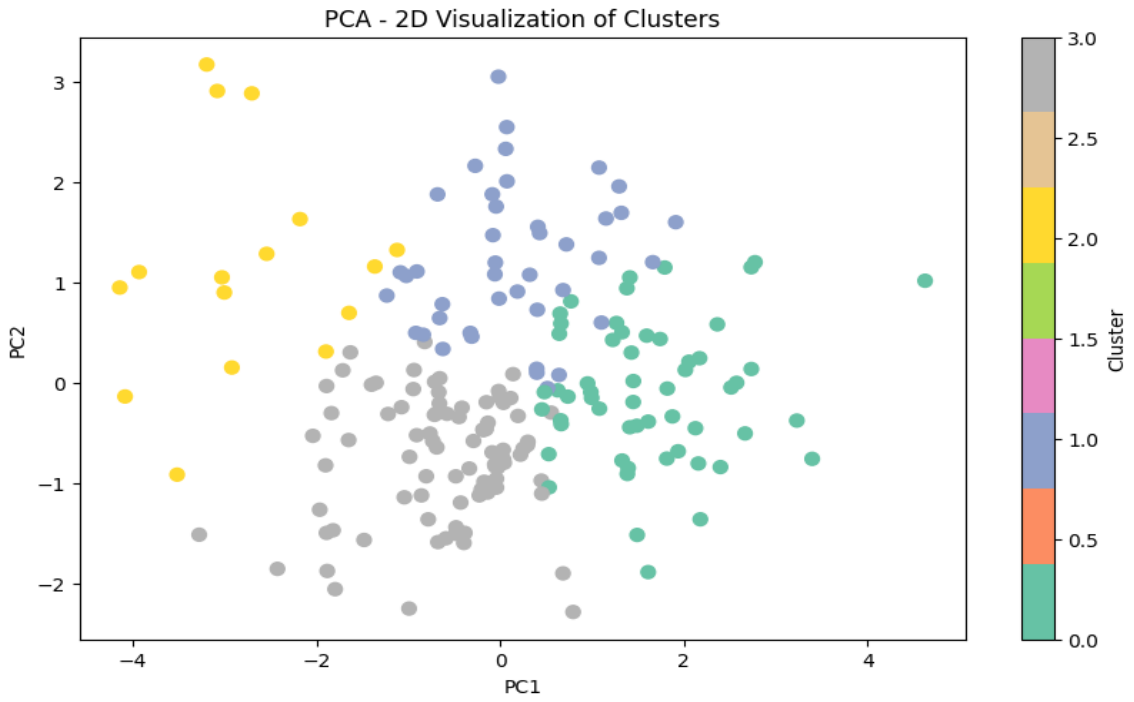
2 16

3 85

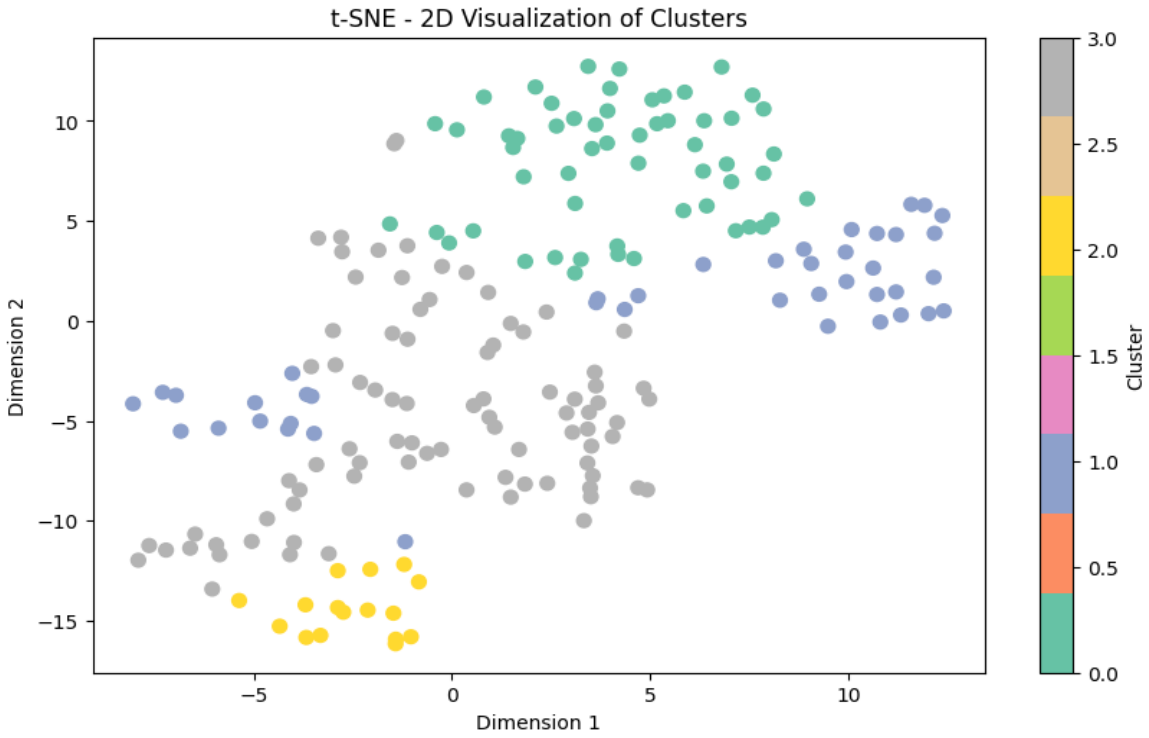
4.3 Cluster Profiling

Cluster	Total Spending	Frequency	Avg Transaction Value	Recency
0	5587.84	7.55	747.34	49.16
1	3792.02	3.90	977.42	82.76
2	1132.33	1.81	593.38	251.81
3	2349.35	4.52	522.73	52.31

Visualizations : PCA 2D Visualization with cluster centers highlighted



- t-SNE Visualization for cluster separation.



5. Cluster Stability Analysis

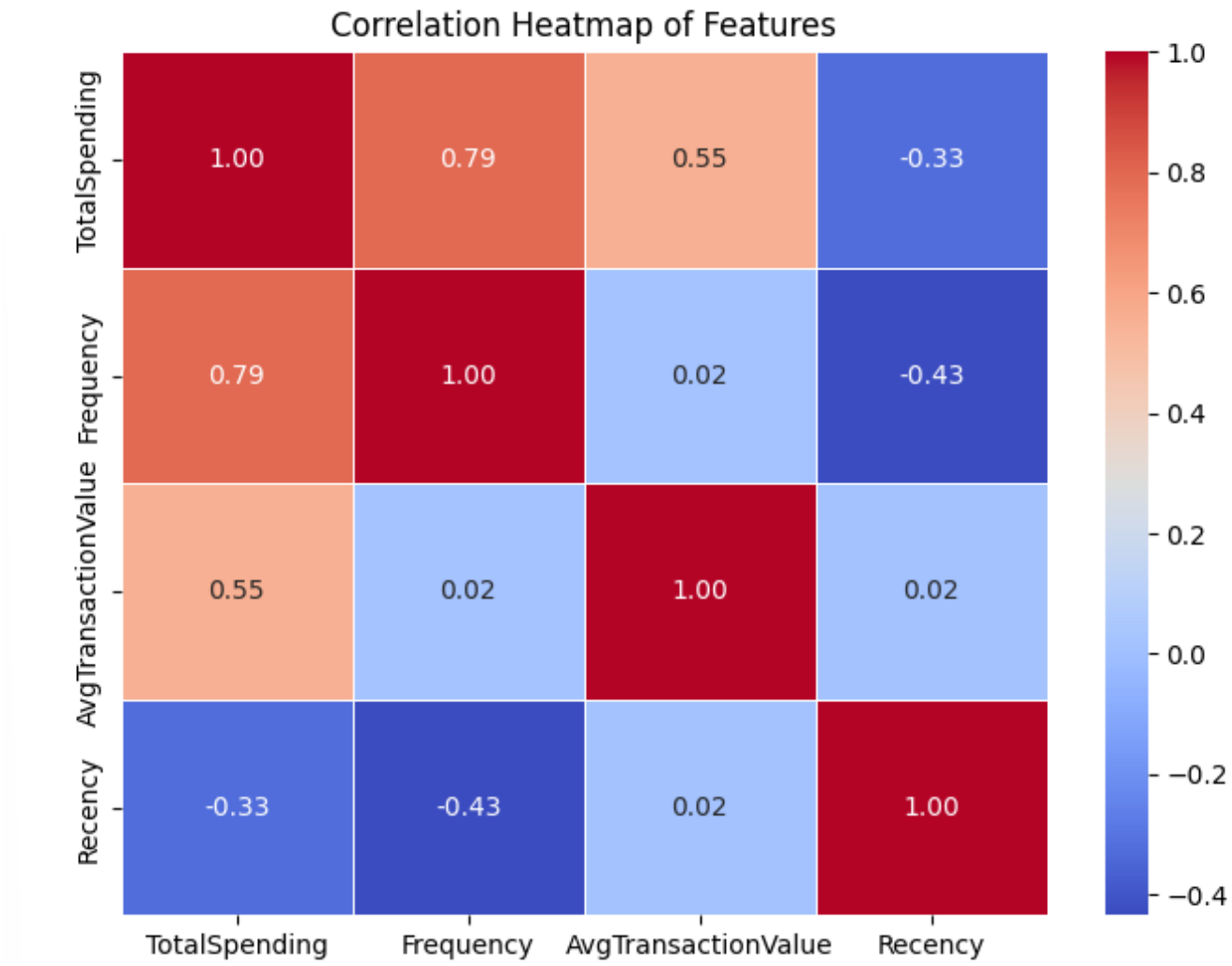
5.1 Bootstrap Sampling Consistency

- **Cluster Consistency across bootstrap samples:** 29.15%

6. Correlation Analysis

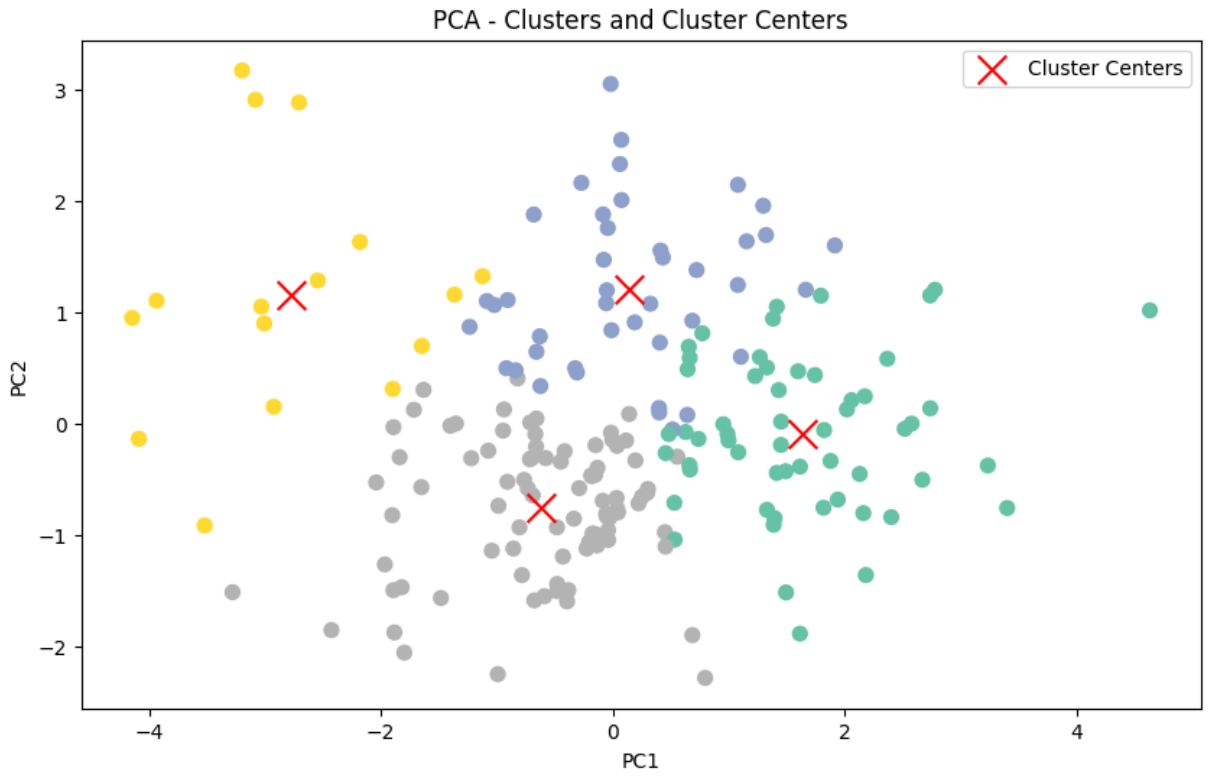
6.1 Feature Correlation Heatmap

A heatmap was generated to visualize the correlation among features:



7. Conclusion

- Four distinct customer segments were identified.
- Cluster 0 represents high spenders with frequent transactions.
- Cluster 2 has infrequent buyers with long recency periods.
- The clustering model shows moderate silhouette scores and stable performance metrics.
- Recommendations include targeted marketing strategies based on customer segmentation.



- **Cluster 0 (High Spenders):** Loyalty programs and premium offers.
- **Cluster 1 (Moderate Spenders):** Upsell campaigns to boost frequency.
- **Cluster 2 (Inactive Customers):** Reactivation campaigns.
- **Cluster 3 (Average Customers):** Retention strategies.

```
Cluster Consistency across bootstrap samples: 29.15%
Optimal Number of Clusters: 4
Davies-Bouldin Index: 0.9976
Calinski-Harabasz Index: 101.4487
```