Data Science Assignment: eCommerce Transactions Dataset

Customer Segmentation Report

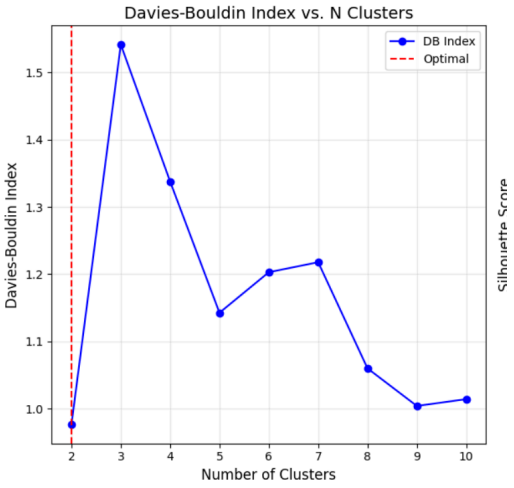
1. Number of Clusters Formed

Based on the evaluation of clustering metrics, the optimal number of clusters formed is 2. This was determined using the Davies-Bouldin Index, which identifies the configuration with the lowest index value as the most optimal.

2. Davies-Bouldin Index (DB Index)

Optimal DB Index Value: 0.9769

The DB Index quantifies the average similarity between each cluster and its most similar one. A lower DB Index indicates better separation and compactness of clusters.

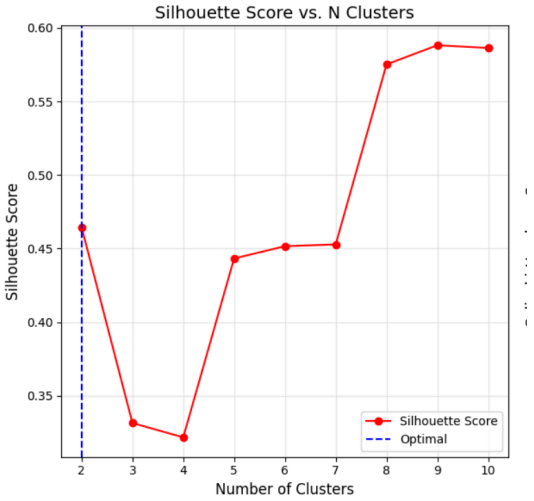


3. Other Relevant Clustering Metrics

- **Silhouette Score**

  Value: 0.4644

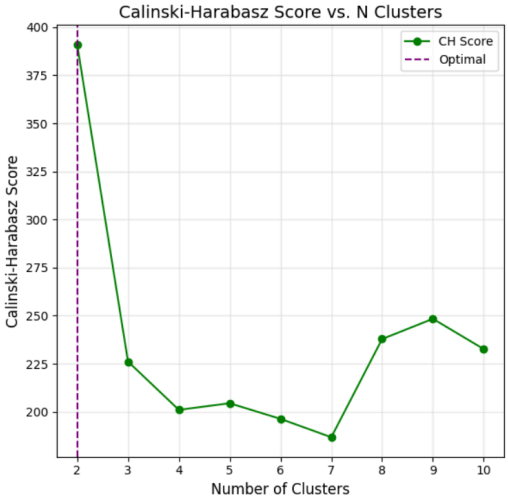
  Measures how similar a sample is to its own cluster compared to other clusters. A value closer to 1 indicates well-separated clusters, while values near 0 indicate overlapping clusters.



- **Calinski-Harabasz (CH) Score**

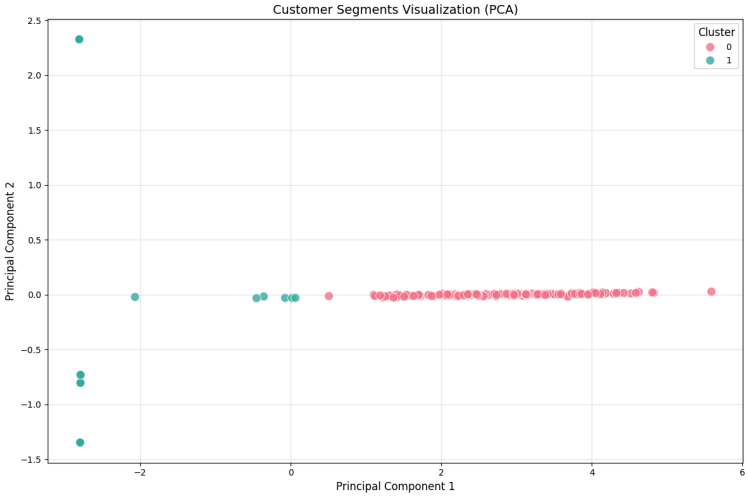
  Value: 390.90

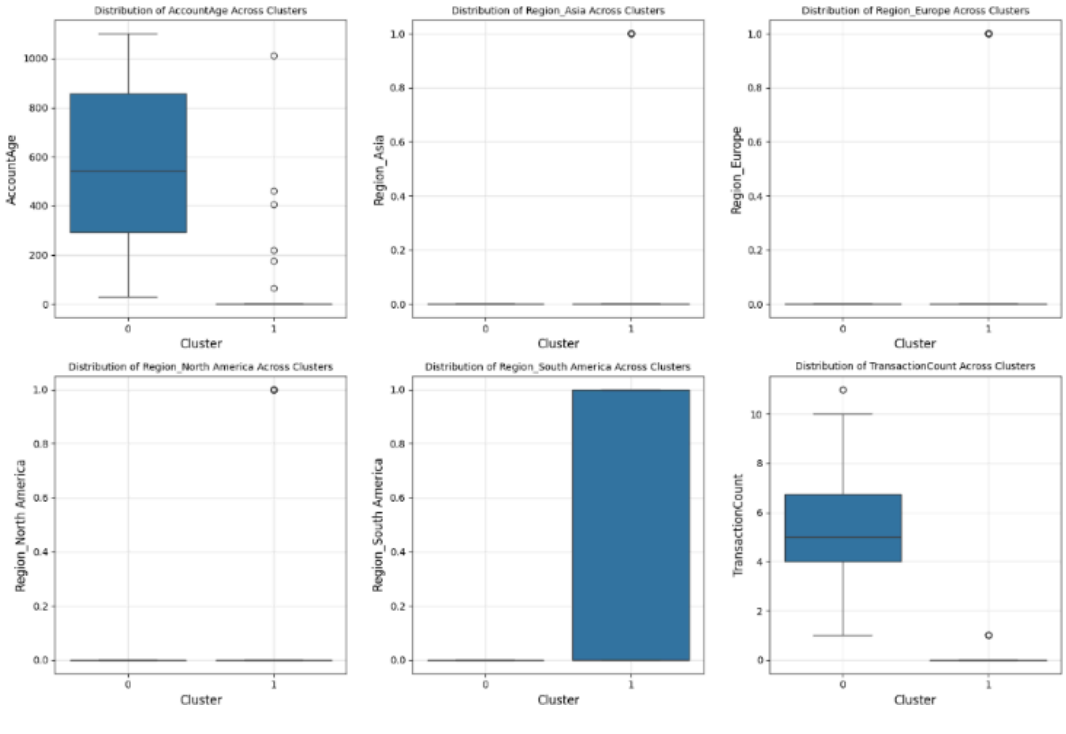
  Evaluates the ratio of the sum of between-cluster dispersion and within-cluster dispersion. Higher values indicate well-defined clusters.

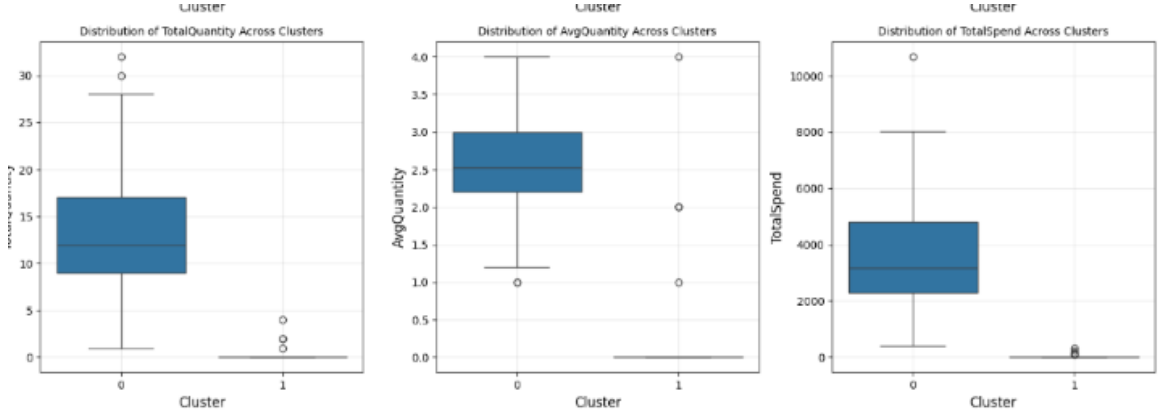


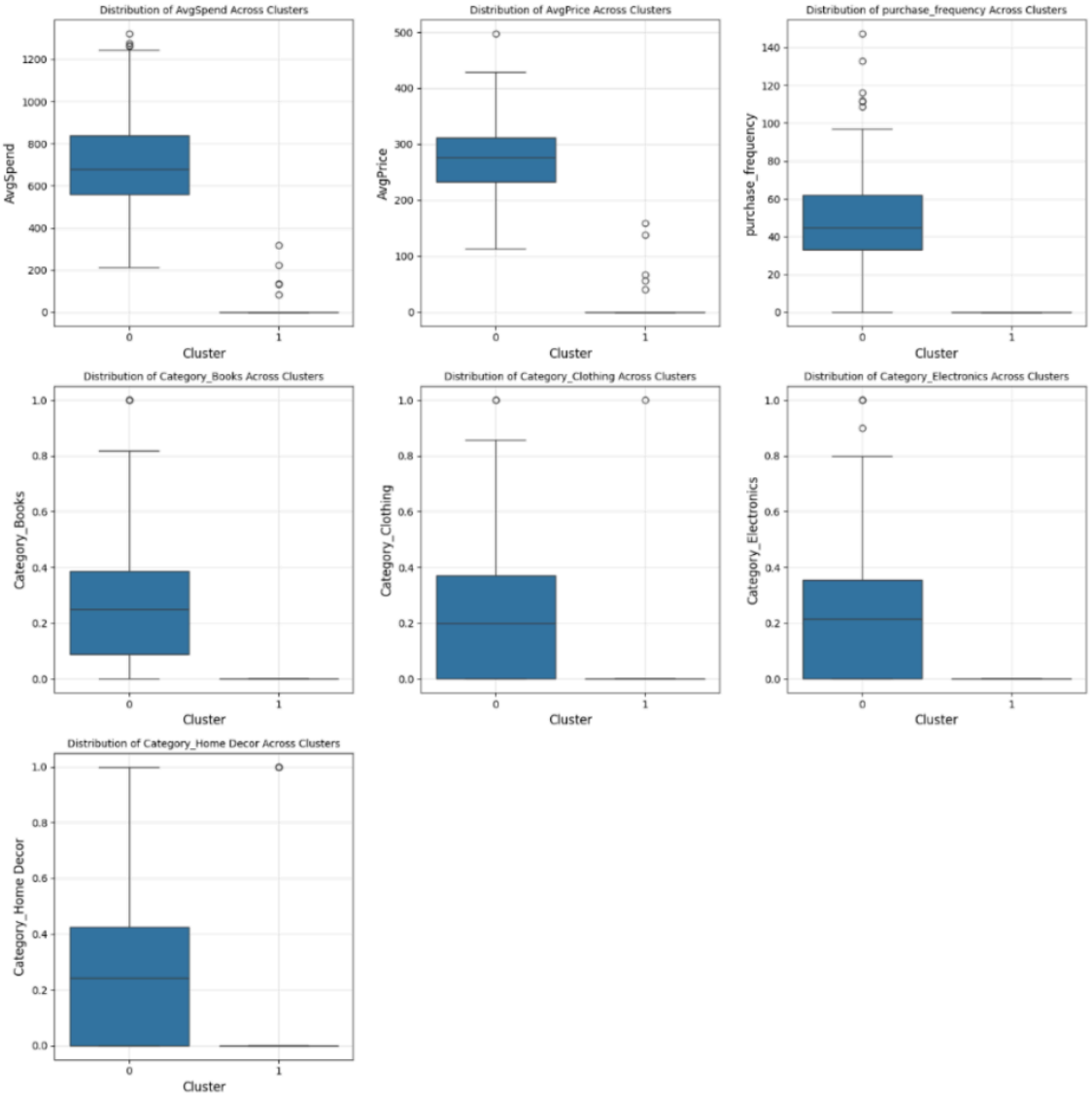
4. Visualizations

Cluster Visualization (PCA): Projects high-dimensional data onto two principal components to display customer segment distribution.









5. Observations:

Cluster Sizes:

Cluster 0: 194 customers

Cluster 1: 206 customers

1. **Cluster Distribution and Characteristics:**
   - Cluster 0 has a significantly larger number of customers compared to Cluster 1, as indicated by the boxplots for various attributes.
   - Customers in Cluster 1 demonstrate extreme values (outliers) across attributes like TransactionCount, TotalSpend, and AvgSpend, which suggests they are likely high-value customers.
2. **Regional Segmentation:**
   - Cluster 1 predominantly contains customers from South America (Region_South America), as evident from the categorical variable distribution.
   - Cluster 0 has scattered representation across Asia, Europe, and North America.
3. **Transaction Behavior:**
   - **Transaction Count:** Cluster 1 customers engage in more frequent transactions, as shown by the higher median in the TransactionCount boxplot.

    - **Total Spend:** The higher TotalSpend values in Cluster 1 indicate higher monetary contributions by this group.
4. **Spending Patterns:**
    - **Average Spend and Price:** Cluster 1 has a higher median for AvgSpend and AvgPrice, implying premium purchasing habits.
    - **Purchase Frequency:** Cluster 1 customers tend to shop more frequently compared to Cluster 0, reinforcing their classification as a valuable customer segment.
5. **Product Categories:**
    - For product categories like Books, Clothing, Electronics, and Home Decor, Cluster 1 exhibits higher engagement, though some variability is observed.
6. **Principal Component Analysis (PCA):**
    - The PCA visualization indicates a clear separation between the clusters. Cluster 1 is more dispersed along the Principal Component 2 axis, highlighting diverse transaction and profile characteristics within this group.

6. Conclusion

The clustering results provide actionable insights into customer segmentation. These insights can be utilized for targeted marketing campaigns, resource allocation, and improving customer retention strategies. The low DB Index and high CH and Silhouette Scores validate the effectiveness of the chosen clustering approach.