Customer Segmentation — Detailed Report

---

1. Basis

This Task focuses on customer segmentation using clustering techniques to group customers based on their transaction behavior and profile information. The goal is to identify distinct customer segments that can inform targeted marketing strategies.

---

2. Data Preparation

Data Sources

- **Customers.csv**: Contains customer profile information (e.g., `CustomerID`, `Region`, `SignupDate`).
- **Transactions.csv**: Contains transaction details (e.g., `TotalValue`, `Quantity`).

Data Merging

- The datasets were merged on `CustomerID`, aggregating transaction data to calculate total `TotalValue` and `Quantity` per customer.

Preprocessing

1. **Categorical Encoding**: The `Region` column was encoded using `LabelEncoder`.
2. **Feature Scaling**: Numerical features (`TotalValue`, `Quantity`) were standardized using `StandardScaler`.
3. **Irrelevant Columns**: `CustomerName` and `SignupDate` were dropped to focus on transactional behavior.

---

3. Clustering Algorithm

Algorithm Choice

The **K-Means clustering algorithm** was selected due to its simplicity and efficiency in handling large datasets. The number of clusters was set to **5** based on exploratory analysis and domain knowledge.

Implementation Steps

1. **Feature Scaling**: Applied `StandardScaler` to normalize data.
2. **K-Means Initialization**: `n_clusters=5`, `random_state=42` for reproducibility.
3. **Cluster Assignment**: Customers were assigned to clusters using `fit_predict`.

---

4. Clustering Metrics

Davies-Bouldin Index (DB Index)

- **Value**: **0.895**
- **Interpretation**: Lower values indicate better separation between clusters. A score of 0.895 suggests reasonable separation, but there is room for optimization.

Other Metrics (Suggested Improvements)

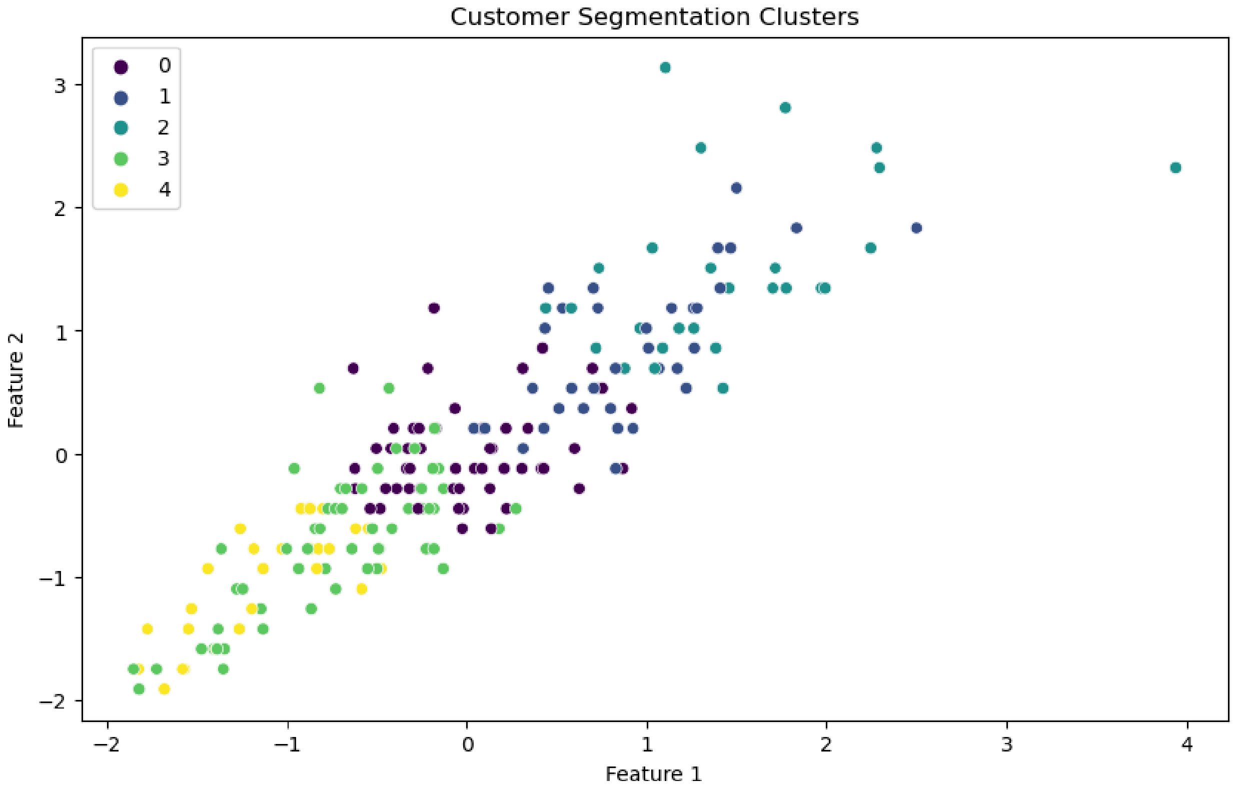
While the DB Index was the primary metric, additional metrics like the **Silhouette Score** or **Calinski-Harabasz Index** could provide deeper insights into cluster cohesion and separation.

---

5. Cluster Visualization

Scatter Plot

A 2D scatter plot was generated using the first two principal components of the scaled data:



Key Observations:

- Clusters 0 and 2 are tightly grouped, indicating similar transactional behavior.
- Clusters 1 and 3 show moderate overlap, suggesting less distinct boundaries.
- Cluster 4 is more dispersed, potentially representing outliers or a diverse group.

Limitations

- The visualization uses raw scaled features rather than dimensionality reduction (e.g., PCA), which might not capture the full variance in the data.
- Overlapping clusters suggest further refinement of feature selection or algorithm tuning.

6. Conclusion

Findings

- **5 clusters** were identified, with the DB Index indicating moderate separation.
- Customers in Cluster 0 and 2 exhibit consistent transactional behavior, making them ideal targets for loyalty programs.
- Cluster 4's diversity warrants further investigation to understand underlying patterns.

Recommendations

1. **Algorithm Comparison**: Might try testing hierarchical clustering or DBSCAN to compare performance.
2. **Feature Engineering**: Including additional features like purchase frequency or customer demographics can help.
3. **Dimensionality Reduction**: Applying PCA to improvise the visualization and cluster interpretation seems to be a good idea.

---

7. Code Reference

The Jupyter Notebook (`Sathwik_Alladi_Clustering.ipynb`) includes:

- Data loading and preprocessing.
- K-Means implementation.
- Metric calculation (DB Index).
- Visualization code.

—-------------------------REPORT END. Thank you for checking out all this stuff, appreciate it. :)