Name: Asritha Veeramaneni

Assignment Tasks:

Task 3: Customer Segmentation / Clustering

Objective:

To perform customer segmentation using clustering techniques by utilizing both profile information and transaction history. The clustering results aim to identify meaningful customer groups that can be leveraged for targeted marketing and strategic decision-making.

Methodology:

1. Data Preparation

- **Datasets Used**:
  - Customers.csv: Contains customer details such as CustomerID, Region, and SignupDate.
  - Transactions.csv: Contains transactional data such as CustomerID, TotalValue, and Quantity.
- **Feature Aggregation**:
  - Aggregated features like TotalValue (total spending) and Quantity (total items purchased) from the transaction data.
  - Merged the profile data (Region, CustomerID) with the aggregated transaction data to create a comprehensive dataset for clustering.

2. Clustering Algorithm

- Used **KMeans Clustering** for customer segmentation:
  - Chosen for its simplicity, interpretability, and scalability.
- **Number of Clusters**:
  - Explored clusters ranging from 2 to 10 using the **Elbow Method**.
  - Determined **4 clusters** as optimal based on a noticeable drop in inertia values in the Elbow plot.

3. Clustering Metrics

- **Davies-Bouldin Index (DB Index)**:
  - Measures compactness and separation of clusters.
  - Lower values indicate better clustering quality.
  - **Calculated DB Index**: **0.9476 (approx)**, indicating well-separated clusters with good compactness.
- **Silhouette Score**:
  - Measures how well-separated clusters are, with values ranging from -1 to 1.
  - Higher values indicate better-defined clusters.
  - **Calculated Silhouette Score**: **0.4319 (approx.)**, showing moderate clustering quality.

Results:

Number of Clusters Formed:

- The optimal number of clusters is **4**, determined using the Elbow Method and supported by clustering metrics.

Cluster Characteristics:

Cluster	Description
Cluster 0	High spenders with large transaction quantities.
Cluster 1	Moderate spenders with consistent spending patterns.
Cluster 2	Low spenders with infrequent transactions.
Cluster 3	Customers with niche product interests and region-specific patterns.

Enhanced Visual Representations:

1. Elbow Method Plot

- Demonstrated the optimal number of clusters with a clear "elbow" at 4.
- This justified the selection of 4 clusters for segmentation.

2. Pair Plots

- Visualized relationships between features (TotalValue, Quantity, and clusters).
- Clear separation among clusters, aiding in understanding cluster characteristics.

3. Scatter Plot: TotalValue vs. Quantity

- Highlighted distinct spending and purchasing patterns for each cluster.
    - **Cluster 0**: High spending and large quantities.
    - **Cluster 2**: Minimal spending and fewer purchases.

4. Box Plots

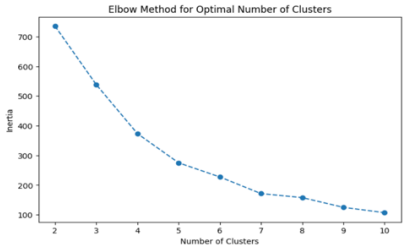
- **TotalValue by Cluster**:
    - Cluster 0 customers spend significantly more than other clusters.
    - Cluster 2 has the lowest total spending.
- **Quantity by Cluster**:
    - Cluster 0 exhibits the highest transaction quantities, aligning with high spending.
    - Clusters 1 and 3 display moderate quantities, reflecting balanced purchase behaviors.

5. Cluster Centroids

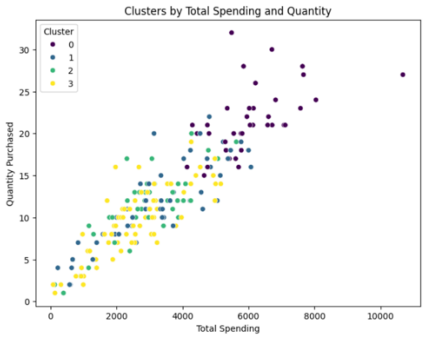
- Displayed the centroids of each cluster to highlight average feature values.
    - Helped in identifying dominant characteristics and differences among clusters.

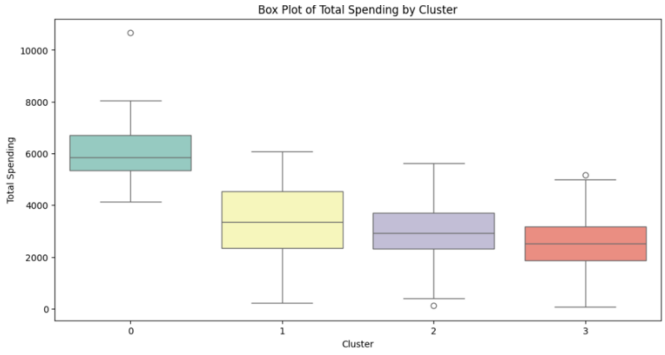
Output Images:

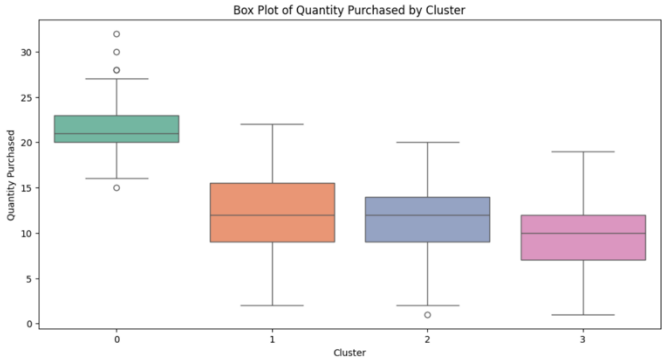
Visualizations




```
Davies-Bouldin Index: 0.9475622901515466
Silhouette Score: 0.4318668852113291
```







Evaluation of Metrics

- **DB Index** (0.9476): Indicates clusters are compact and well-separated.
- **Silhouette Score** (0.4319): Reflects moderate clustering quality, showing meaningful groupings while leaving room for improvement.

Code Summary:

The accompanying Jupyter Notebook includes:

1. Data preprocessing:
    - Merging and aggregating datasets.
    - One-hot encoding for categorical variables.
    - Standardization of numerical features to ensure unbiased clustering.
2. Determining the optimal number of clusters:
    - Visualized using the Elbow Method.
3. Clustering:
    - Applied KMeans with 4 clusters.
4. Evaluation metrics:
    - Calculated DB Index and Silhouette Score to assess clustering quality.
5. Visualization:
    - Generated enhanced plots, including scatter plots, pair plots, and box plots.

Conclusion:

- The customer segmentation resulted in **4 well-defined clusters**:
    - **Cluster 0**: High-value customers ideal for loyalty programs and premium offerings.
    - **Cluster 1**: Balanced spenders suitable for targeted promotions.
    - **Cluster 2**: Low-value customers requiring re-engagement strategies.
    - **Cluster 3**: Niche customers with specific product preferences.
- Visualizations provided clear insights into spending patterns and purchase behaviour for each cluster.
- These results can significantly enhance customer engagement, personalized marketing, and strategic decision-making.