Clustering Report

Overview:

The clustering analysis was performed on a dataset containing customer transaction details. The goal was to segment customers based on their transactional behavior using KMeans clustering. Feature selection and preprocessing steps were carried out before applying the clustering algorithm, and the optimal number of clusters was determined using the Davies-Bouldin Index.

1. Number of Clusters Formed:

- The optimal number of clusters (`k`) was determined to be **9** based on the Davies-Bouldin Index and a visual analysis of its values for `k` ranging from 2 to 10.

2. Davies-Bouldin Index (DB Index):

- The **DB Index** for the selected `k=9` is **0.9925**.
  - A lower DB Index indicates better cluster compactness and separation.
  - This value suggests that the clusters are reasonably well-separated and compact, but there is room for improvement in terms of inter-cluster separation.

3. Silhouette Score:

- The **Silhouette Score** for `k=9` is **0.2885**.
  - A Silhouette Score close to 1 indicates well-separated clusters, while a score near 0 suggests overlapping clusters.
  - The relatively low score indicates that while the clusters exist, some overlap or noise may reduce their distinctiveness.

4. Cluster Characteristics:

Each cluster represents a group of customers with distinct behavioral patterns. The average values of key features for each cluster are as follows:

Cluster	TotalValue	Quantity	Price	DaysSinceSignup
0	833.43	3.66	231.54	698.06
1	377.39	3.44	111.06	102.46
2	503.76	1.37	368.16	45.13
3	198.42	2.28	94.93	689.06
4	984.51	2.48	402.66	426.73
5	1535.92	3.67	420.10	650.73
6	1278.00	3.58	358.93	-6.57
7	500.56	1.43	354.27	691.15
8	185.69	1.46	127.67	74.84

- **Key Observations:**
  - Cluster 5 represents customers with the **highest average transaction values** (`TotalValue = 1535.92`) and frequent high-priced purchases.
  - Cluster 8 has the **lowest total value and quantity**, indicating low engagement and spending.
  - Cluster 6 shows **negative DaysSinceSignup**, suggesting possible data anomalies requiring further investigation.

5. Key Visualizations:

- **Davies-Bouldin Index Plot:** A visual decline in the DB Index supports selecting `k=9`.
- **Pairplot Visualization:** Displays clear distinctions between clusters for `TotalValue`, `Quantity`, and `Price`, though some overlap is observed.
- **Feature Distributions by Cluster:** Histograms reveal how features like `TotalValue` and `DaysSinceSignup` vary across clusters.