Customer Segmentation and Clustering Report

1. Number of Clusters Formed

After performing clustering on the combined dataset of customer profile information and transaction data, the optimal number of clusters was determined to be **5 clusters**. The selection of this number was based on both the elbow method and silhouette analysis, ensuring a balance between compact and well-separated clusters.

2. DB Index Value

The Davies-Bouldin Index (DB Index) for the clustering results was calculated to be **1.000518628183939**. A lower DB Index indicates that the clusters are compact and well-separated

3. Other Relevant Clustering Metrics

- **Cluster Sizes**:
- Cluster 0: 53 data points
- Cluster 1: 23 data points
- Cluster 2: 36 data points
- Cluster 3: 31 data points
- Cluster 4: 56 data points

4. Visual Representation of Clusters

The clustering results were visualized using:

- **Scatter Plot**: A 2D scatter plot based on PCA (Principal Component Analysis) was used to represent clusters in reduced dimensions. Each cluster was represented with a unique color to highlight the separation.
- **Heatmap**: A heatmap was created to show the relationship between customer attributes and cluster assignments.
- **Cluster Distribution**: A bar chart was used to represent the size of each cluster.

5. Summary and Insights

- Customers in **Cluster 4** tend to purchase higher-value products frequently, suggesting theymight be premium customers.
- **Cluster 0** contains customers who predominantly purchase low-value items but do soconsistently, identifying a loyal but budget-conscious segment.
- **Cluster 2** consists of customers who purchase sporadically, indicating an opportunity fortargeted marketing to improve engagement.
- (Additional insights derived from clustering metrics and visual analysis.)

6. Methodology and Tools

- **Data Preprocessing**: Missing values were handled, and all data was scaled usingStandardScaler.
- **Clustering Algorithm**: The clustering was performed using the K-Means algorithm with a rangeof cluster numbers (2 to 10).
- **Evaluation Metrics**: The DB Index, silhouette score, and inertia were used for evaluating thequality of clustering.
- **Tools Used**: Python libraries such as scikit-learn, pandas, matplotlib, and seaborn.