Data Science Assignment: eCommerce Transactions Dataset Analysis

Report on Customer Segmentation

1. Number of Clusters Formed: After performing the clustering analysis, I determined the optimal number of clusters to be **5**. This was based on the **Elbow Method**, where I observed a clear reduction in the rate of decrease as the number of clusters increased. Beyond 5 clusters, adding more did not significantly improve the model's quality.

2. DB Index Value: The **Davies-Bouldin Index (DB Index)** for the clustering solution is **0.8967**. This index measures the separation and compactness of the clusters. A lower DB Index indicates better-defined clusters with minimal overlap. The value of 0.8967 suggests that the clustering solution is reasonably good, with distinct clusters and minimal overlap between them.

3. Other Relevant Clustering Metrics:

- **Silhouette Score**: The **Silhouette Score** for the clustering is **0.3606**. This metric quantifies how similar each sample is to its own cluster compared to other clusters. A higher score indicates better-defined clusters. While the score is not very high, it still suggests that the clustering is somewhat effective, with some overlap between the clusters.

4. Cluster Visualization: To visualize the customer segments, I used Principal Component Analysis (PCA) to reduce the dimensionality of the data to two dimensions. The scatter plot of the resulting PCA components clearly shows the separation between the five clusters, with each cluster occupying a distinct area in the plot. This visualization confirms the clustering results and provides insights into how customers are distributed across the segments.

The clusters formed provide useful insights into the different types of customers and the metrics used (DB Index) suggest that the clustering model is effective.