Task 3 – Customer Segmentation / Clustering

Overview:

This report summarizes the customer segmentation analysis using clustering techniques based on the customer and transaction information from the customer.csv and transaction.csv datasets.

Number of clusters Formed:

The analysis identified 5 clusters as the optimal customer segmentation.

Clustering Metrics:

1. Calculated the Davies-Bouldin Index 1.37 for the clustering. Lower values indicate better clustering, but it shows a higher similarity ratio of each cluster with the cluster most similar to it.
2. Initializes a PCA object to reduce the data dimensions to 2 components. Fits the PCA model to the scaled customer data and transforms it into a 2-dimensional space. The resulting array customer_data_2d contains the 2 principal components for each customer.

Additional Insights:

Cluster Separation:
The data points are color-coded according to different clusters, with distinct groups visible in the plot. This indicates that the clustering algorithm has effectively segmented the customers into distinct groups based on their characteristics.
Cluster Distribution:
The distribution of points within each cluster varies. Some clusters are more densely packed, indicating homogeneity within those groups, while others are more spread out, suggesting diverse customer profiles within those clusters.
Cluster Overlap:
There may be some overlap between clusters, where data points from different clusters are close to each other. This overlap could indicate similarities between certain customer segments or transitional behavior between clusters.
Outliers:
Any data points that are far from the main cluster centers can be considered outliers. These customers may exhibit unique behavior that sets them apart from the majority of the customer base.