ZEOTAP DS ASSIGNMENT

Name: Ponnekanti Pranathi

TASK-3

Customer Segmentation/ Clustering

I have first merged the Customers.csv, Transactions.csv and Products.csv file for my first task itself into a ‘merged_transactions_updated.csv’ file. I have used the same file for this task as well. I have used k-means clustering for this with a predefined number of clusters **4** since the business aims to identify customer segments with different purchasing behaviors. I have considered the following features as metrics to cluster the data.

1. Transaction Data:

- **Total Spent** (CalculatedTotalValue): The total amount a customer has spent across all transactions. This gives an idea of how much money a customer has invested in the products.
- **Transaction Count**: The total number of transactions made by the customer. This helps in identifying customers who make frequent purchases versus those who purchase infrequently.
- **Average Transaction Value** : The average value spent per transaction by the customer. This helps to identify if customers tend to purchase high-value products or low-value products.

  Clustering based on total spending, transaction frequency, and average spend helps new customers based on their purchasing habits and financial behavior. The number of transactions is useful to determine customer loyalty or engagement with the business. Customers with higher transaction counts might be considered more engaged.(as already mentioned in task-1, reward points help us to analyze the shopping habits of the customers).

2. Customer Profile Data:

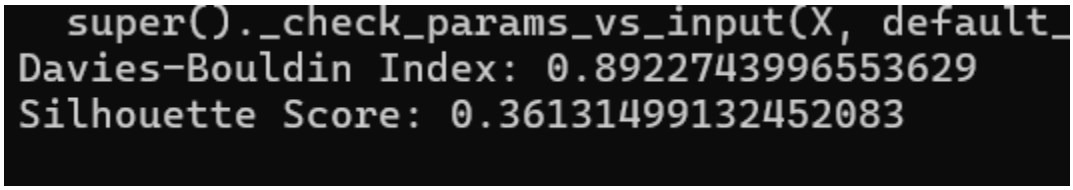
- **Region**: The geographical region of the customer is used as a metric.

Time-Based Analysis: Features such as SignupDate and TransactionDate could also be used if we want to cluster based on when customers became active and how their purchasing patterns evolve over time.

All other clustering techniques based on different metrics I have already used in task-1 with clear plots.

In summary, the clustering was based primarily on financial behavior (total spending, transaction count, and average transaction value), which directly reflects customer engagement and purchasing habits and region(region based recommendations we get based on the regional products also).

I have used silhouette score and DB index to evaluate. The results can be seen below.



We see that the Silhouette score is better than Davies-Bouldin Index for this particular model. After this I applied PCA after scaling the data. The PCA model reduces the original data from 3 features to 2 principal components, which capture the most important information.