Data Science Intern

TASK 3 REPORT

By Aviral Soni

Task 3

Perform customer segmentation using clustering techniques. Use both profile information (from Customers.csv) and transaction information (from Transaction.csv)

- You have the flexibility to choose any clustering algorithm and any number of clusters between 2 and 10.
- Calculate clustering metrics, including the DB Index(Evaluation will be done on this).
- Visualize your clusters using relevant plots.

Ans:

Executive Summary:

This report is centered on a customer segmentation analysis based on the transaction behavior and profile data. The ambition was to identify distinctive customer groups and provide insight for targeted marketing as well as retention strategies.

Using K means clustering, 6 customer segments/clusters were identified.

Clustering Summary:

- Number of clusters formed: 6 clusters.
- Algorithm used: K-Means Clustering
- DB Index: 1.31(min) for 6 clusters.

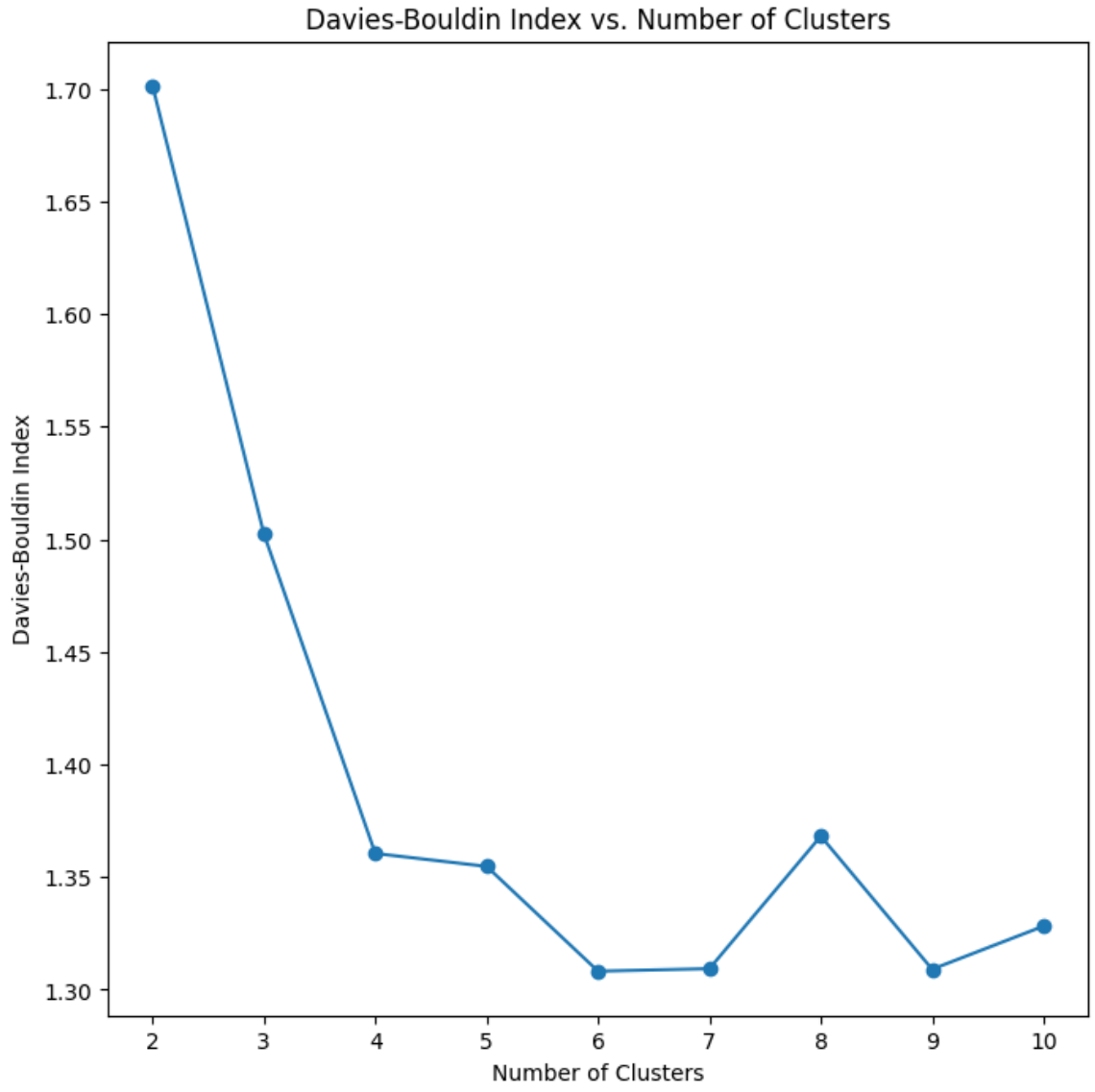
Data Preparation:

- The dataset merged customer profile information (Customers.csv) and transaction details (Transactions.csv).
- Aggregated transactional features per customer:
    - Total spending
    - Total transactions
    - Average transaction value
    - Total quantity purchased
    - First Transaction
    - Last Transaction
- Both datasets were merged on CustomerID (inner join).
- Categorical variables were encoded
- Columns which do not contribute were removed. Eg: CustomerID, CustomerName, SignupDate, FirstTransaction etc.
- Standardization was applied to normalize feature scales

Approach to find optimal number of clusters:

```
for k in cluster_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    cluster_labels = kmeans.fit_predict(scaled_data)
    db_score = davies_bouldin_score(scaled_data, cluster_labels)
    db_scores.append(db_score)
```

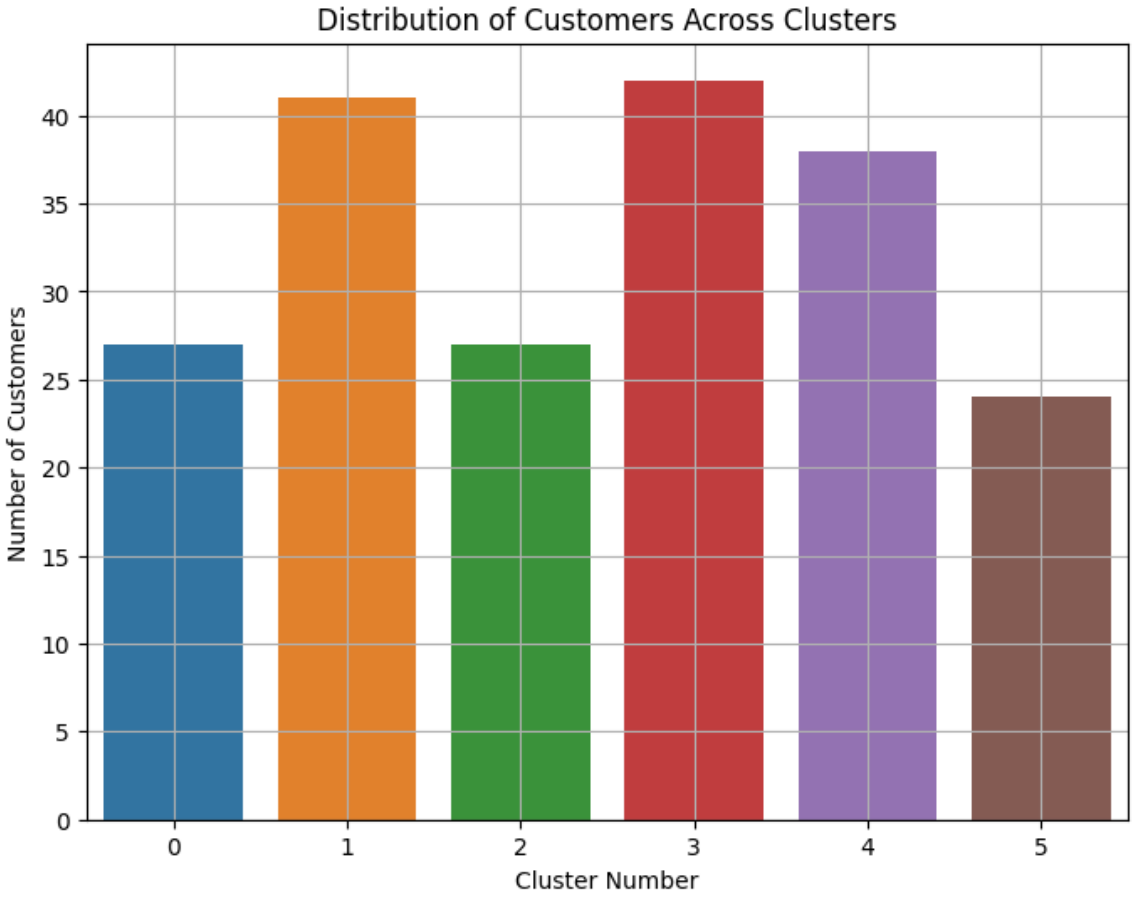
- This function was used to find the optimal number of clusters.
- The minimum DB score corresponds to the desired number of clusters.
- In this case the value was 6 clusters at DB score of 1.31.



Total Spending vs Total Transactions:

- Some clusters represent **high-value customers** with high spending and frequent transactions. (cluster 2)
- Other clusters consist of **low-value customers** with low transaction counts and spending. (cluster 3)

Distribution of Customers Across Clusters



Total Quantity vs Total Spending

- Some customers who purchase high total quantity are found in high spending clusters. This indicates bulk buying behavior.

- Some clusters have moderate spending but high quantity, which may represent cost-conscious customers buying lower-priced products in volume.
- Low total quantity and low spending clusters suggest minimal engagement and could be potential churn risks.

RECOMMENDATIONS:

1) Rehabilitate high value customers with loyalty programs.
2) Engage new customer base with high average transaction values.
3) Reward price sensitive to spend more but buy the same volume.