Customer Segmentation/Clustering

1. Objective

The primary objective is to identify the optimal number of clusters and group users based on their transaction patterns across different products. This involves performing customer segmentation using clustering techniques.

2. Approach

To determine the optimal clustering of users, unsupervised clustering algorithms are applied. In this case, the K-Means clustering algorithm and DBSCAN are utilized.

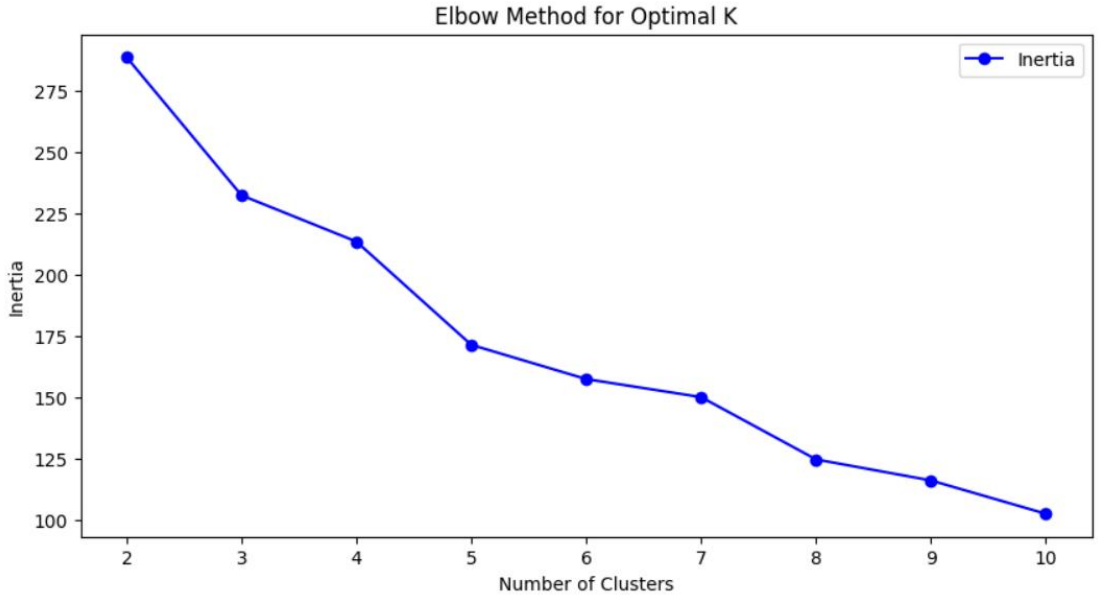
3. Techniques

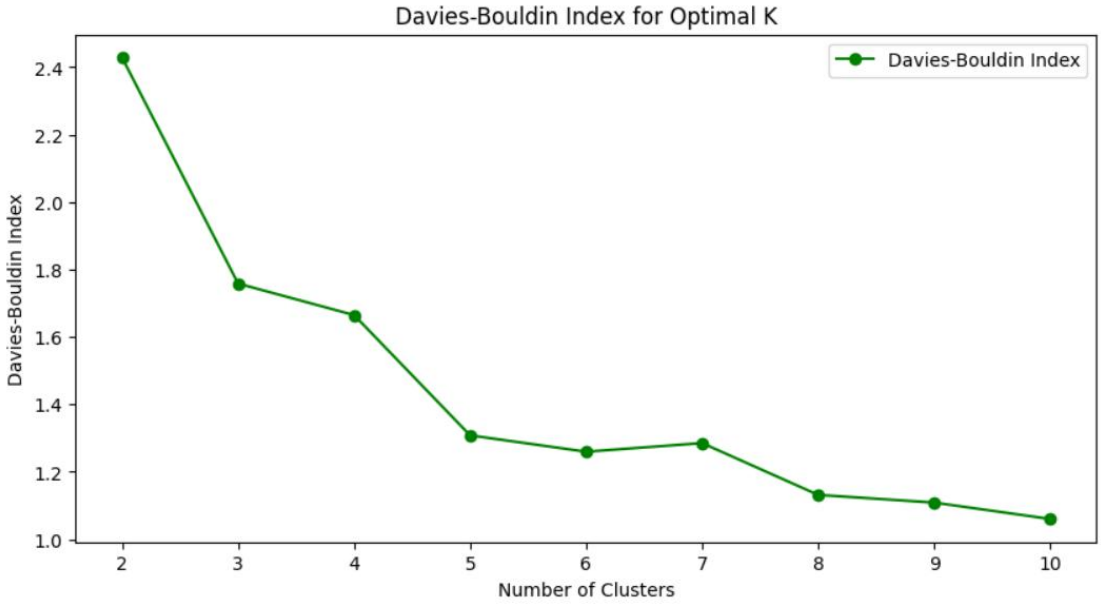
3.1 Data Preprocessing and Feature Engineering:
The dataset provided is clean, with no missing, duplicate, or outlier values. Therefore, additional preprocessing is unnecessary. However, for model training, data normalization is applied, and one-hot encoding is performed to ensure compatibility with the clustering algorithms.

3.2 Model Training and Evaluation:
Two clustering algorithms—K-Means and DBSCAN—are employed, and their performance is evaluated with varying cluster counts and hyperparameter values.

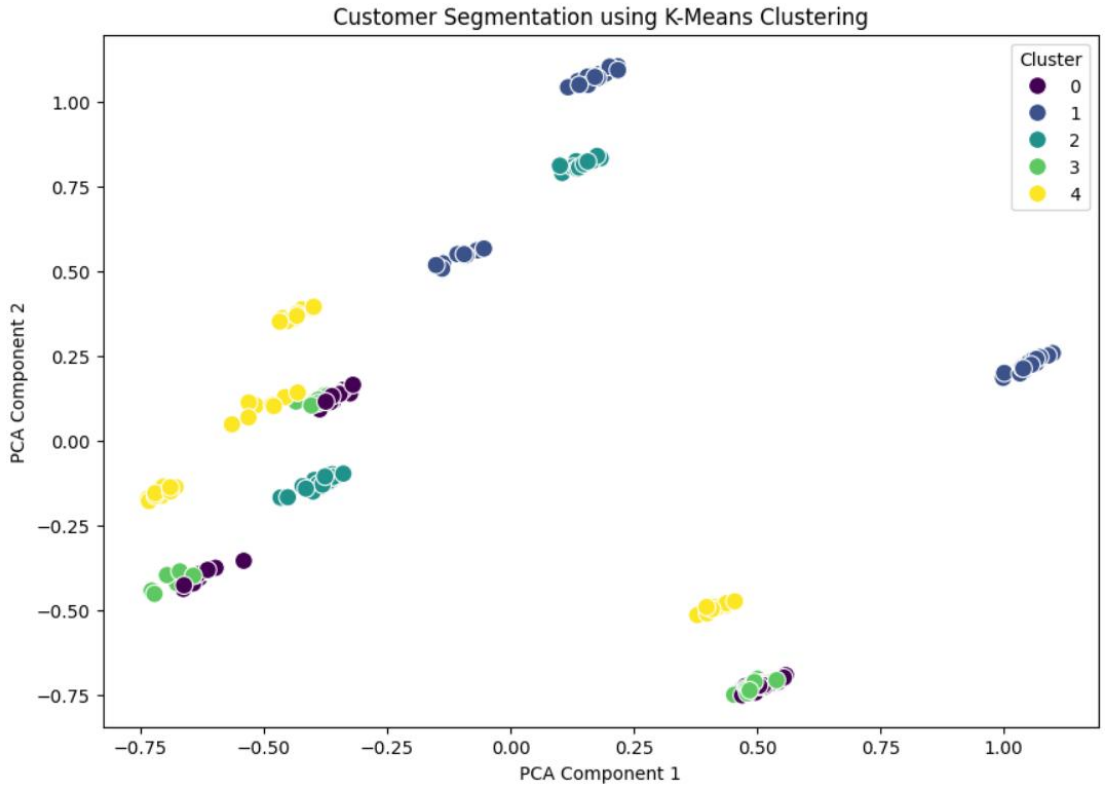
Selection of Optimal Clusters

K-Means Results:
The Elbow Method is used to determine the optimal number of clusters for the K-Means algorithm. The elbow is observed at k=5, though it is not sharply defined. Evaluation results indicate a Davies-Bouldin Index (DBI) of 1.308 for k=5 and 0.956 for k=10. Based on the Elbow Method, k=5 is chosen to balance interpretability and avoid overfitting.

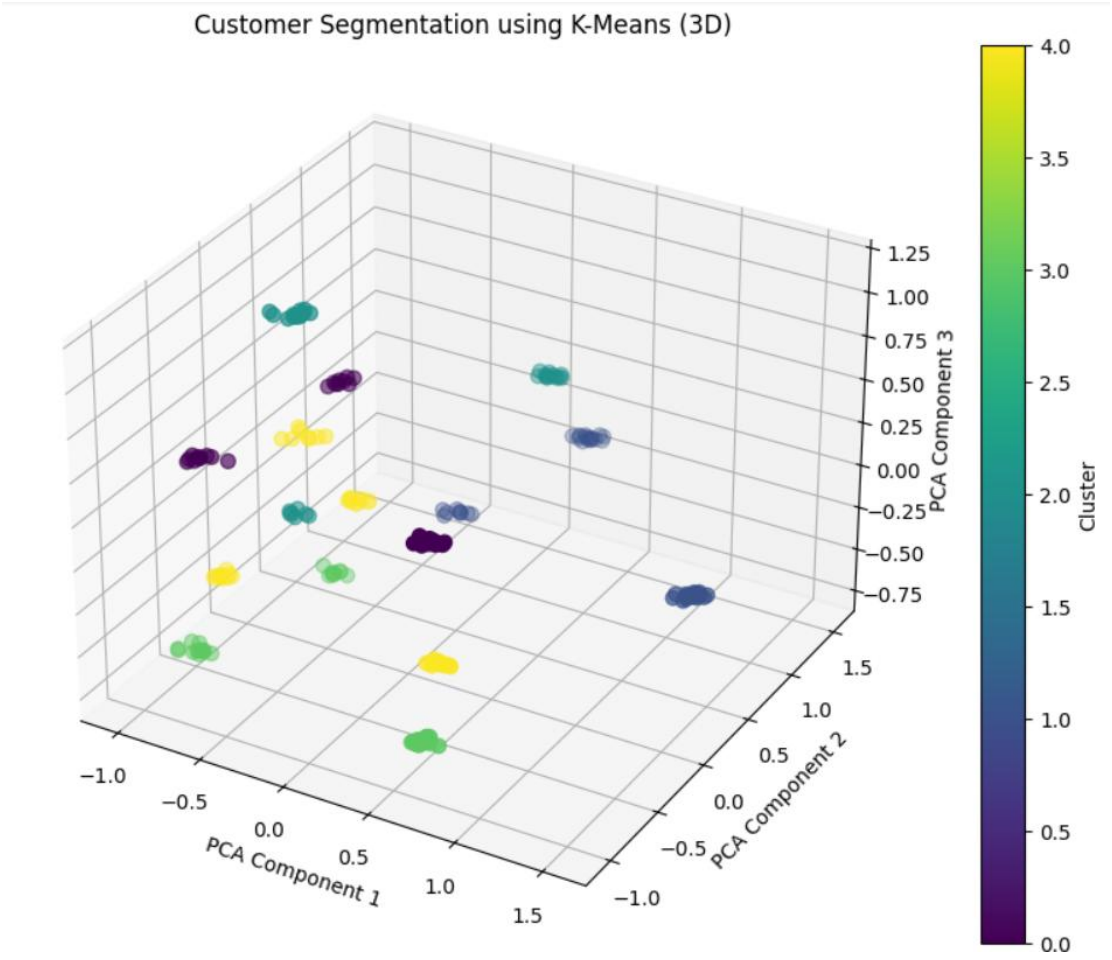




Visualization using PCA:



The 3-D segregation variant shows a much clearer description of cluster segregation with k=5 leading to pretty segregated groups of customers.

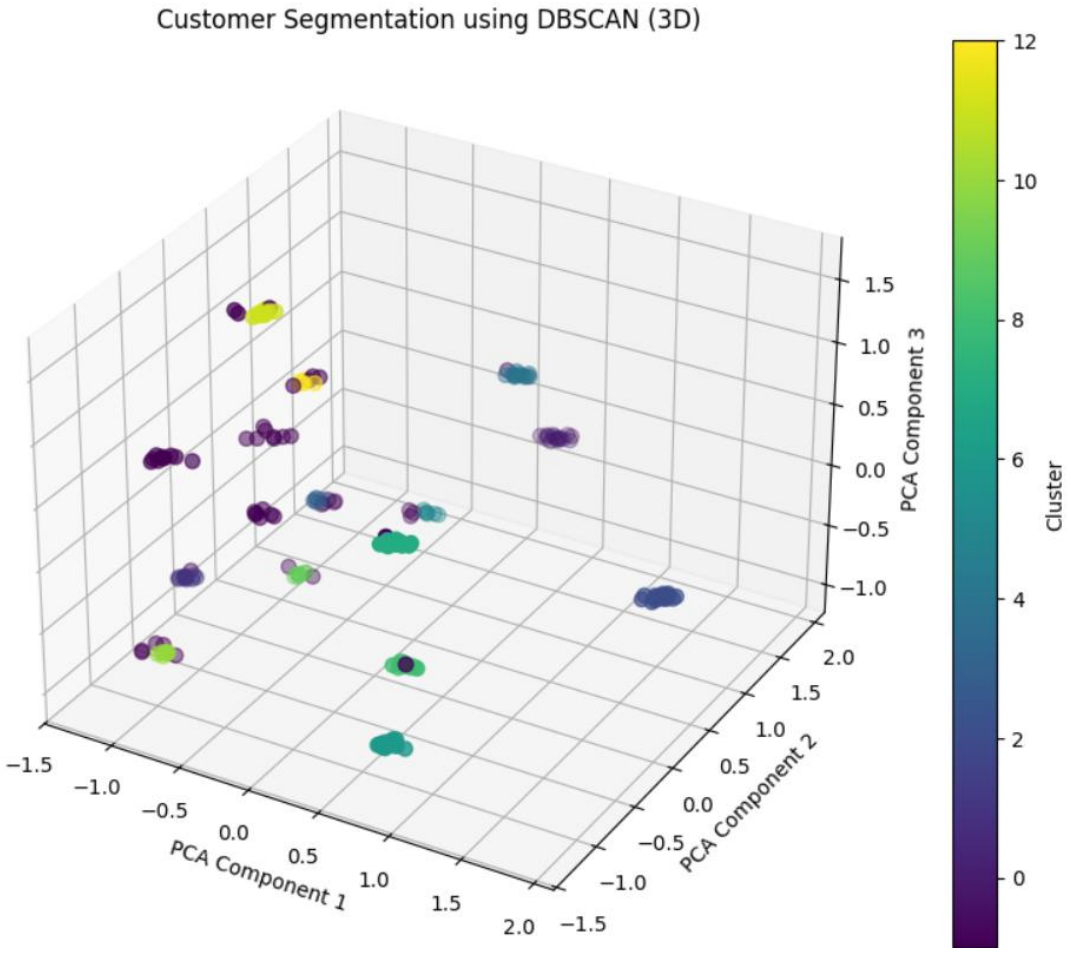


Evaluation Metrics:

```
Davies-Bouldin Index for 5 clusters: 1.3082932415373 78
Silhouette Score for 5 clusters: 0.30990063490812336
Calinski-Harabasz Index for 5 clusters: 46.48634126336333
```

DBSCAN Results:
For DBSCAN, varying epsilon (ε) values are tested. With ε=0.5, the model generates 14 clusters with a DBI of 1.307. Reducing ε to 0.3 results in 5 clusters with a DBI of 1.134 and a very high number of outlier points, making it unusable and inappropriate.

A 3-D visualization of these clusters describe a very vague segmentation as showed by 2 component PCA analysis.



Conclusion

The application of **K-Means** and DBSCAN clustering algorithms demonstrates that customer segmentation can be effectively achieved with thoughtful hyperparameter tuning. K-Means, with **k=5**, offers an optimal balance of interpretability and performance, as supported by the Elbow Method and DBI evaluation with a value of **1.308**. DBSCAN, on the other hand, provides flexible clustering based on density, with its performance varying significantly based on the epsilon parameter. Overall, these techniques provide valuable insights into user behavior and segmentation, enabling more targeted strategies for customer engagement and decision-making.