Clustering Analysis Report

Introduction

In this analysis, we applied KMeans clustering to customer transaction data with the aim of segmenting customers based on their purchasing and demographic qualities. The dataset contained features such as Quantity, TotalValue, and Price, which were scaled using a StandardScaler to ensure that each feature contributed equally to the clustering process. The objective was to determine the optimal number of clusters that best segmented the data.

Clustering Approach

KMeans clustering is an unsupervised learning algorithm that groups data into clusters by minimizing the sum of squared distances between points and their corresponding centroids. To determine the optimal number of clusters, we performed a grid search testing values from 2 to 10 clusters, evaluating each configuration using the Davies-Bouldin Index (DBI).

Evaluation Metrics

The Davies-Bouldin Index (DBI) was chosen as the evaluation metric due to its ability to balance cluster cohesion and separation. A lower DBI indicates better clustering, where the clusters are more compact and well-separated.

Results:

The optimal number of clusters was determined to be 2, based on the lowest Davies-Bouldin Index of 0.7596. This suggests that two clusters provide the best balance between compactness and separation in the data. Below is the table of DBI values for each tested number of clusters:

	n_clusters	davies_bouldin_index
0	2	0.759606
1	3	1.072367
2	4	1.991391
3	5	5.127449
4	6	4.631310
5	7	4.414384
6	8	2.103944
7	9	2.004123
8	10	1.936859

Conclusion

The clustering analysis revealed that 2 clusters best segmented the customer data, with distinct groups based on transaction frequency and value. Future work could explore the use of other clustering algorithms like DBSCAN or hierarchical clustering or incorporate additional customer features for more granular segmentation.