Clustering Report

Data Preparation:

merged profile and transaction data for better overview of data.

Then calculated total spend, average spend per transaction, total quantity and region of a specific customer.

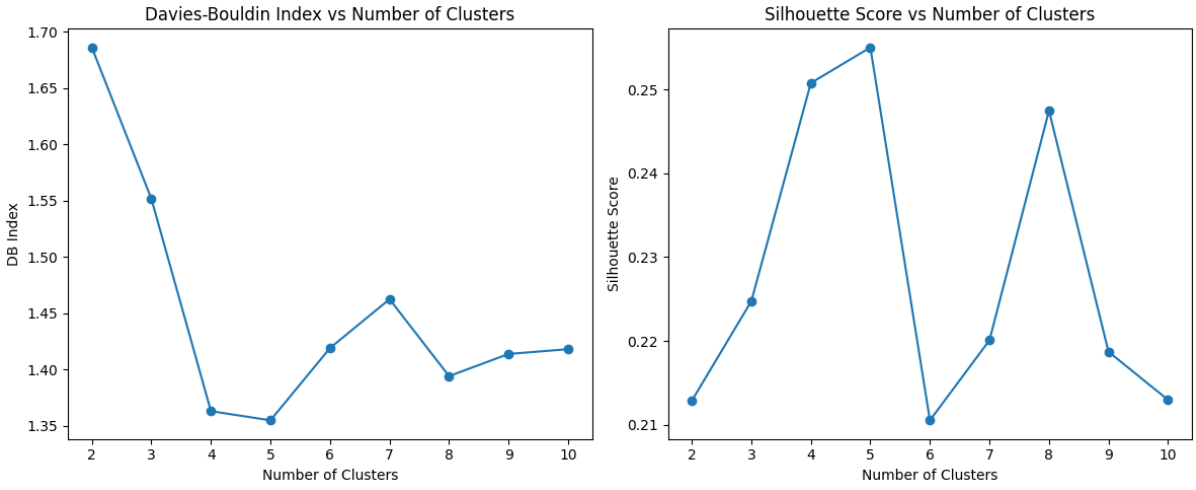
Used one-hot encoding for categorical feature like region and numerical features ware scaled accordingly.

Clustering Algorithm:

Used K-Means clustering from 2 to 10 and tested for best cluster using DB Index and Silhouette Score.

DB index = 1.3547088760955934

Silhouette Score = 0.2549935635490153

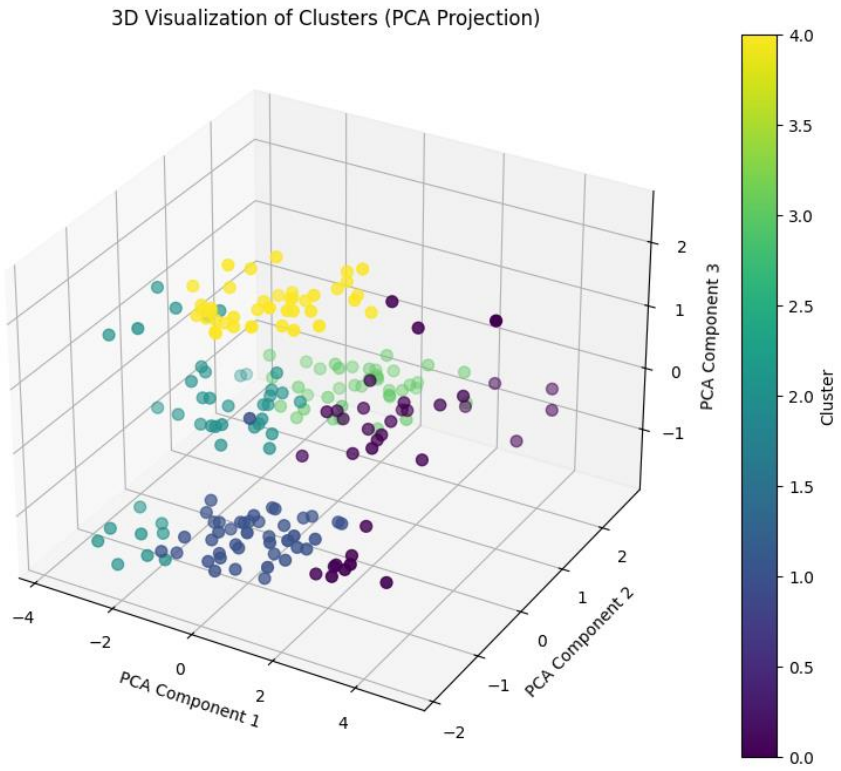


The graph shows the optimal clusters to be 5

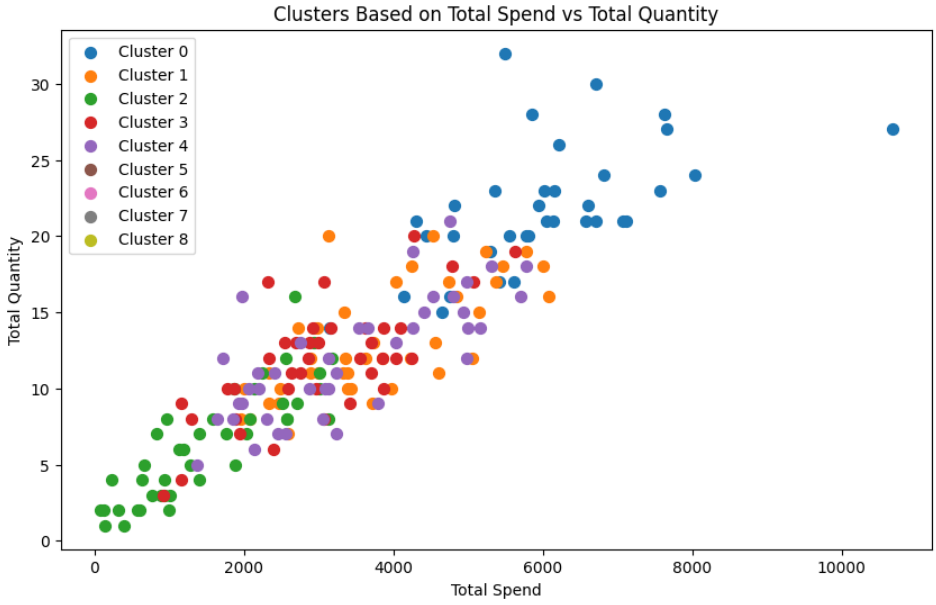
Optimal Clusters:

The optimal number of clusters was found to be 5 based on lowest DB Index (1.35)

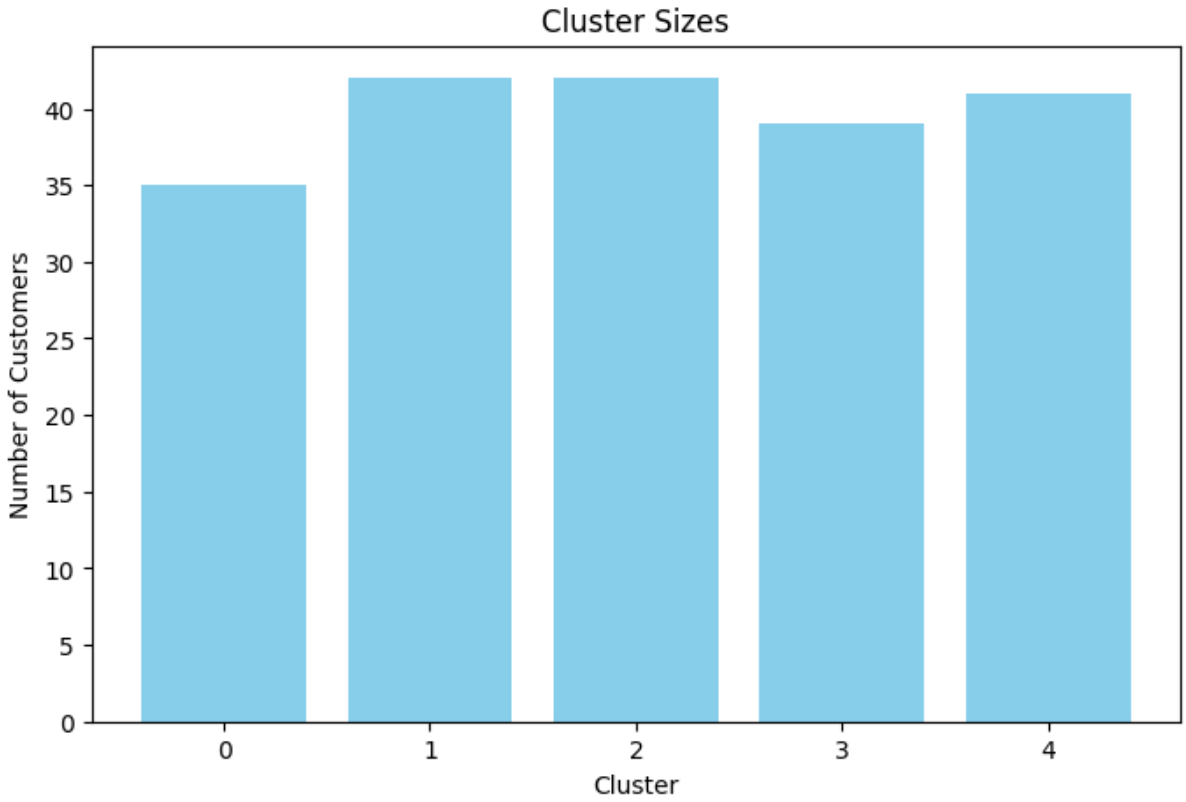
Visualization:



Here I reduced the feature to 3 using PCA and drew the graph for clusters.



Scatter plots of Total Spend vs. Total Quantity show clear distinctions between clusters.



A bar chart shows the size of each cluster which has almost equal no of customers.