Clustering and Dimensionality Reduction Report

1. Introduction

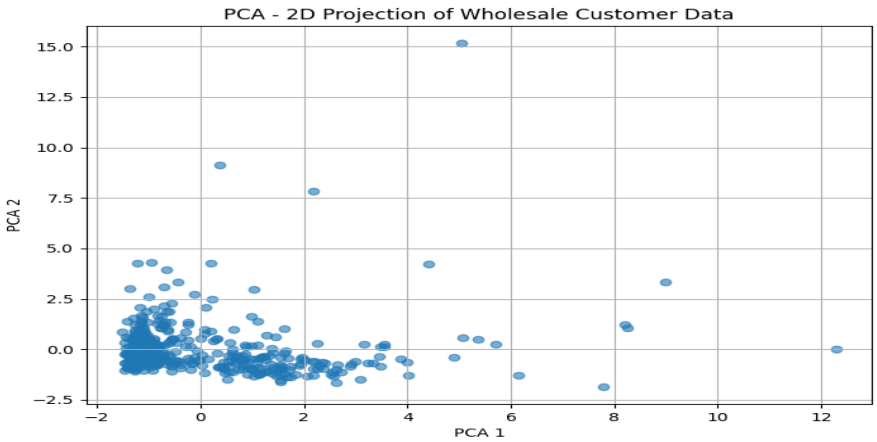
This report presents the implementation of clustering and dimensionality reduction techniques on the Wholesale Customer dataset as part of the Digital Empowerment Network - Week 02 assignment. The objective is to apply unsupervised learning using PCA for dimensionality reduction and multiple clustering methods (K-Means and DBSCAN) for customer segmentation.

2. Data Preparation & Preprocessing

The dataset 'Wholesale customers data.csv' was loaded and initially explored. Missing values and duplicate records were checked. Duplicate rows were removed. Non-numeric columns such as 'Region' and 'Channel' were excluded as clustering was performed only on numerical data. The dataset was then normalized using StandardScaler to ensure equal weighting of features.

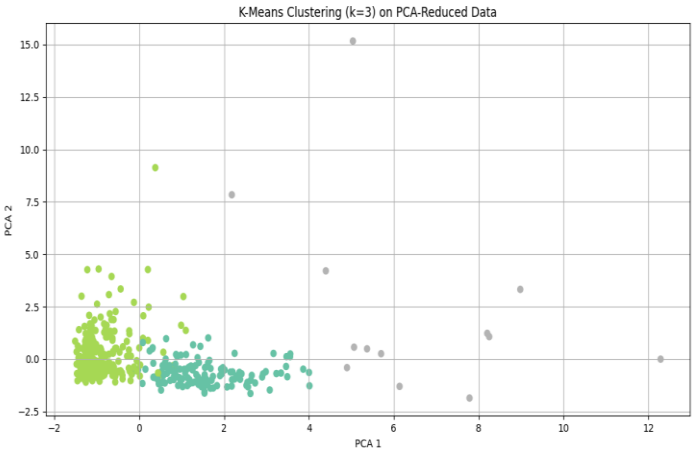
3. Dimensionality Reduction using PCA

Principal Component Analysis (PCA) was applied to reduce the data to two dimensions for better visualization. The explained variance ratio from the two components was noted to understand how much information was retained.



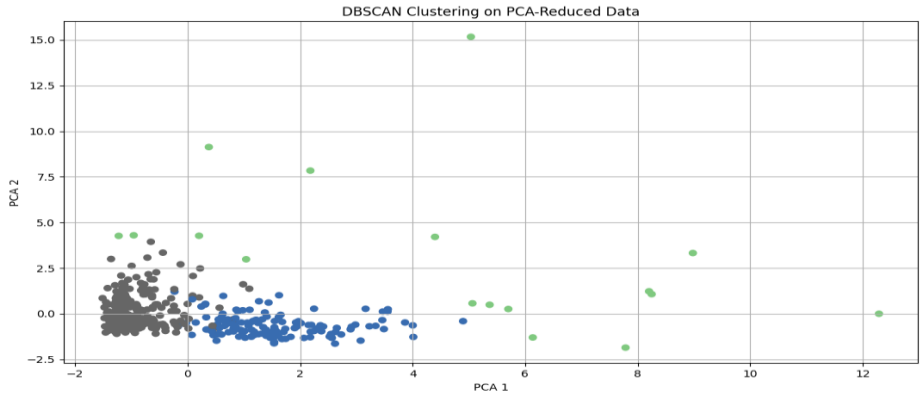
4. K-Means Clustering

The Elbow Method was used to determine the optimal number of clusters by plotting the inertia values against different values of k. Based on the plot, k=3 was chosen. The K-Means algorithm was applied, and clusters were visualized in the reduced PCA space.



5. DBSCAN Clustering

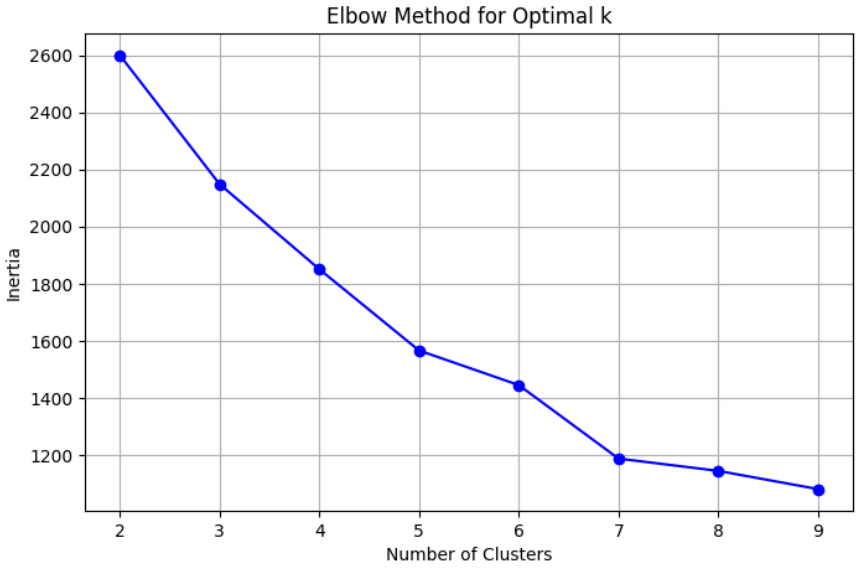
DBSCAN was applied using eps=2 and min_samples=5. The resulting clusters were visualized using PCA-reduced data. DBSCAN does not require pre-specifying the number of clusters and is capable of identifying noise points.



6. Clustering Evaluation

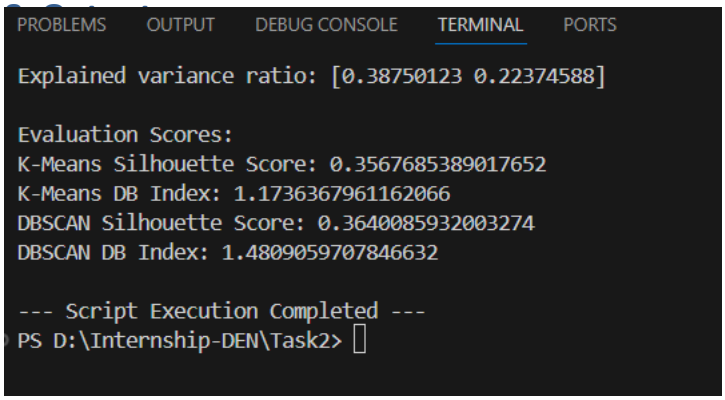
To evaluate clustering performance, Silhouette Score and Davies-Bouldin Index were calculated. Higher silhouette scores and lower DB index values indicate better clustering

quality. K-Means achieved meaningful clusters with acceptable scores, while DBSCAN's performance varied depending on parameters.



7. Conclusion

This project demonstrates the application of PCA and clustering algorithms for customer segmentation. K-Means provided well-separated clusters, while DBSCAN revealed density-based groupings. The experiment highlights the effectiveness of dimensionality reduction and the importance of parameter tuning in clustering.


PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
Explained variance ratio: [0.38750123 0.22374588]
Evaluation Scores:
K-Means Silhouette Score: 0.3567685389017652
K-Means DB Index: 1.1736367961162066
DBSCAN Silhouette Score: 0.3640085932003274
DBSCAN DB Index: 1.4809059707846632
--- Script Execution Completed ---
PS D:\Internship-DEN\Task2>