1. Introduction

This report outlines the creation of an Interactive Clustering Analysis Dashboard utilizing Dash (Plotly). The primary function of this dashboard is to enable users to examine a dataset, visualize their results, and dynamically adjust clustering parameters. Its objective is to facilitate users in investigating and comprehending the data structure through the application of clustering algorithms such as KMeans and DBSCAN.

2. Dataset

The Wholesale Customers dataset used in this project contains annual spending amounts of wholesale customers on different product categories. Numerical features include the following:

- **Fresh**: Annual spending on fresh products.
- **Milk**: Annual spending on milk products.
- **Grocery**: Annual spending on grocery products.
- **Frozen**: Annual spending on frozen products.
- **Detergents_Paper**: Annual spending on detergents and paper products.
- **Delicassen**: Annual spending on delicatessen products.

3. Clustering Algorithms

The dashboard implements two clustering algorithms:

3.1 KMeans Clustering

KMeans is a centroid-based clustering algorithm that partitions the data into **k** clusters, where each data point belongs to the cluster with the nearest centroid. The algorithm requires the user to specify the number of clusters (**k**) in advance. In the dashboard, the user can adjust the number of clusters using a slider.

3.2 DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups together points that are closely packed, while marking points that are far away as outliers. DBSCAN requires two parameters:

- **eps (epsilon)**: The maximum distance between two samples for them to be considered as in the same neighborhood.
- **min_samples**: The minimum number of points required to form a dense region.

In the dashboard, the user can adjust both **eps** and **min_samples** using sliders.

4. How the Dashboard Works

1. **Data Loading and Preprocessing**:
   - The dataset is loaded and preprocessed by standardizing the numerical features using **StandardScaler**. This ensures that all features have the same scale, which is important for clustering algorithms.
2. **Clustering**:
   - The user selects a clustering algorithm (KMeans or DBSCAN) and adjusts the parameters using the sliders.
   - The clustering algorithm is applied to the standardized data, and the cluster labels are assigned to each data point.
3. **Visualization**:
   - The scatter plot is updated to show the clustering results, with each cluster represented by a different color.
   - The summary table is updated to show the mean values of each feature for each cluster.
4. **Interactivity**:
   - The dashboard is interactive, meaning that any change in the algorithm or parameters will immediately update the scatter plot and summary table.
5. **Data Download**:
   - The user can download the clustered data as a CSV file for further analysis.

5. Conclusion

The **Interactive Clustering Analysis Dashboard** provides a user-friendly interface for exploring clustering algorithms and visualizing their results. The dashboard allows users to interactively adjust clustering parameters, visualize the results, and download the clustered data for further analysis. The implementation of both **KMeans** and **DBSCAN** algorithms provides flexibility in exploring different clustering approaches.

This project demonstrates the power of **Dash (Plotly)** for building interactive web-based dashboards for data analysis and visualization. The modular structure of the code makes it easy to extend the dashboard with additional clustering algorithms or features in the future.

Additional Resource:

Dataset link: https://archive.ics.uci.edu/dataset/292/wholesale+customers
Colab link: DW_Assignment2.ipynb