1. Clustering Analysis

Introduction

Clustering is an unsupervised machine learning technique used to group similar data points based on inherent patterns. This project applies clustering methodologies to analyze a dataset, aiming to identify meaningful patterns and segmentations within the data. The primary objective is to uncover hidden structures that can be utilized for targeted business strategies, customer segmentation, or anomaly detection.

Data Preprocessing

Before applying clustering techniques, data preprocessing was performed, which included:

- **Handling Missing Values**: Missing values were imputed using mean, median, or mode based on the data type.
- **Feature Scaling**: Since clustering algorithms are distance-based, standardization using Min-Max Scaling or StandardScaler (Z-score normalization) was applied.
- **Feature Selection**: Unimportant features were removed to enhance model performance and prevent noise in clustering results.

Clustering Methods Used

1. **K-Means Clustering**: This technique partitions data into k clusters based on the centroid proximity. The Elbow Method was used to determine the optimal number of clusters by plotting the Within-Cluster-Sum-of-Squares (WCSS) against k values.

2. **Hierarchical Clustering**: This method was explored to visualize cluster relationships using dendrograms.

3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: A density-based approach was considered to handle noise and identify arbitrarily shaped clusters.

Results and Insights

- The K-Means model effectively grouped data into distinct clusters, providing meaningful segmentations.
- Hierarchical clustering confirmed the optimal cluster count identified through K-Means.
- DBSCAN helped identify outliers, making it useful for anomaly detection.
- Visualizations such as scatter plots and cluster heatmaps were used to interpret the cluster characteristics.