**Project Documentation: Unsupervised Learning on Retail Data**

\*\*1. Introduction:\*\*

- \*\*Objective:\*\*

The project aims to apply unsupervised learning techniques on a retail dataset to explore patterns and relationships within the data.

- \*\*Tools and Libraries:\*\*

The code is written in Python and utilizes popular libraries such as pandas, numpy, matplotlib, seaborn, scikit-learn, and scipy.

\*\*2. Data Exploration and Preprocessing:\*\*

- \*\*Data Loading:\*\*

The dataset is loaded from an Excel file ('Online Retail.xlsx') using the pandas library.

- \*\*Data Cleaning:\*\*

Duplicate records are removed, and irrelevant columns such as 'StockCode,' 'Description,' 'CustomerID,' and 'InvoiceDate' are dropped.

- \*\*Outlier Removal:\*\*

Outliers in the 'Quantity' and 'UnitPrice' columns are identified and removed using z-scores.

- \*\*Data Visualization:\*\*

Various visualizations are created using seaborn and matplotlib to understand the distribution and relationships within the cleaned dataset.

Build models

\*\*3. K-Means Clustering:\*\*

- \*\*Feature Selection and Standardization:\*\*

The 'Quantity' and 'UnitPrice' features are selected for clustering and standardized using StandardScaler.

- \*\*Determination of Optimal Clusters:\*\*

The Elbow Method is employed to identify the optimal number of clusters (k) for K-Means.

- \*\*K-Means Clustering:\*\*

The K-Means algorithm is applied with the optimal number of clusters, and the resulting clusters are visualized.

- \*\*Silhouette Score:\*\*

The silhouette score is calculated to evaluate the quality of the clustering.

\*\*4. K-Nearest Neighbors (KNN) Classification:\*\*

- \*\*Feature Scaling:\*\*

The features are scaled using StandardScaler.

- \*\*Train-Test Split:\*\*

The dataset is split into training and testing sets.

- \*\*KNN Model Training:\*\*

A KNN classifier is trained on the training set.

- \*\*Model Evaluation:\*\*

The accuracy, confusion matrix, and classification report are calculated to evaluate the KNN model's performance.

\*\*5. Hierarchical Clustering:\*\*

- \*\*Hierarchical Clustering:\*\*

A hierarchical clustering dendrogram is created using the features 'Quantity' and 'UnitPrice.'

\*\*6. Isolation Forest for Anomaly Detection:\*\*

- \*\*Isolation Forest Model:\*\*

An Isolation Forest model is trained to detect anomalies (outliers) in the dataset.

- \*\*Anomaly Visualization:\*\*

Anomalies are visualized on a scatter plot of 'Quantity' and 'UnitPrice.'

- \*\*Summary:\*\*

The project involves exploratory data analysis, unsupervised learning, and anomaly detection on a retail dataset.