Project Title: Clustering Wholesale Customers Based on Item Purchases

Project Overview:

For this project, the use case is to find the segment group of wholesale customers based on the item purchased. Dataset used will be wholesale customer dataset from the public domain.

This dataset is chosen as clustering customer segments is one of the common topics seen. The clustering of variables/ features using unsupervised learning often allows discovery of inter class high similarity and intra class high dis-similarity. This clustering is easily reproducible in other contexts beside the customer segment. Other domains such as automobile insurance, credit card users, telco customer churn and ecommerce shoppers.

As these modeling are based on historical data with certain assumptions. This project did not take into account other external factors such as mode of purchase (credit card/ cash), direct delivery or self collect and time of purchase transaction.

Data Source

The primary data source for this project is a wholesale customer dataset from the public domain. The dataset should include the following information:

Features related to item purchases (e.g., product categories, quantities, spending)
Any additional customer-related information (if available)

Data Dictionary:

- FRESH: annual spending (m.u.) on fresh products (Continuous)
- MILK: annual spending (m.u.) on milk products (Continuous)
- GROCERY: annual spending (m.u.)on grocery products (Continuous)
- FROZEN: annual spending (m.u.)on frozen products (Continuous)
- DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
- DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous)
- CHANNEL: customers Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel (Nominal)
- REGION: customers Region - Lisnon, Oporto or Other (Nominal)

Note: Monetary units (m.u.)

Data Source: https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

Scope

The project scope includes:

- Data preprocessing, including handling missing values and feature selection.
- Applying unsupervised learning algorithms for clustering.
- Evaluation of clustering results using appropriate metrics.
- Interpretation of customer segments and their characteristics.
- Documentation of the analysis process for reproducibility.
- No consideration of external factors such as payment method, delivery method, and time of purchase transaction.

Key Deliverables

- Exploratory Data Analysis (EDA) report, including data summary and initial insights.
- Clustering model that segments wholesale customers.
- Evaluation report with clustering performance metrics.
- Customer segment profiles, detailing characteristics of each group.
- Codebase and documentation.
- Final project report summarizing the process and findings.

Project Implementation

1. Exploratory Data Analysis (EDA)

Perform EDA to understand the data, visualize patterns, and gain initial insights into customer behavior.

2. Clustering Model Development

Apply clustering algorithms to segment wholesale customers based on item purchases.

3. Model Evaluation

Evaluate the clustering results using appropriate metrics using brelative metrics

- Elbow method
- Silhouette method

4. Interpretation of Results

Interpret customer segments and provide detailed profiles of each group.