Customer Segmentation: Unsupervised Learning Report

---

1. Description of the Data

The dataset used for this analysis contains detailed information about customer demographics, purchasing behavior, and interactions with a company. The key columns include:

- **Year_Birth**: The year of birth of the customer.
- **Education**: Education level (e.g., Graduation, PhD, etc.).
- **Marital_Status**: Marital status of the customer.
- **Income**: Annual income of the customer.
- **Kidhome & Teenhome**: Number of children and teenagers in the household.
- **Spending Columns**: Customer spending on different products (e.g., wine, meat, fruits, etc.).
- **NumWebPurchases & NumStorePurchases**: Number of purchases made via the web and stores.
- **Recency**: Number of days since the last purchase.
- **Response**: Whether the customer responded to a campaign.

The dataset originally consisted of **2,240 rows** and **29 columns**. Preprocessing included handling missing values (24 missing income values) and feature engineering (adding Family_Size and Total_Spent while removing redundant columns).

---

2. Main Objective

The primary objective of this analysis is to segment customers into distinct groups based on their purchasing behavior and demographic characteristics using clustering algorithms. This segmentation allows the company to tailor marketing strategies, improve customer retention, and target high-value customer groups effectively.

---

3. Model Variations and Selection

To achieve the objective, we applied multiple clustering algorithms:

K-Means Clustering

- **Optimal k**: The Elbow Method suggested that the optimal number of clusters lies at **k=4**.
- **Silhouette Scores**:
  - k=4: 0.512

    - k=3: 0.424

- K-Means performed well, with k=4 showing the highest silhouette score, indicating better-defined clusters.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- Explored varying eps values ([4.5, 4.0, 3.5, 3.0]) and min_samples ([2, 3, 5, 7]).
- **Best Configuration**:
    - **eps = 3**, **min_samples = 3**: Achieved **3 clusters** with a **silhouette score of 0.209**.
- DBSCAN performed moderately well, identifying dense clusters and noise effectively.

PCA for Shape Visualization

- Principal Component Analysis (PCA) reduced the data to two dimensions for visualization.
- The data exhibited a spherical distribution, favoring K-Means over DBSCAN for clustering.

---

4. Key Findings

1. **Cluster Characteristics**:
    - Each cluster revealed distinct spending habits and demographic traits.
    - For instance, one cluster contained high-income families with high spending on wines, while another comprised lower-income families with modest purchases.
2. **Model Comparisons**:
    - K-Means performed best with k=4 clusters, yielding higher silhouette scores and well-separated groups.
    - DBSCAN struggled to define clusters with significant noise due to the dataset's shape and density.

---

5. Suggestions for Next Steps

1. **Refine Features**:
    - Include more features like customer loyalty score or geographic region for better segmentation.
    - Test additional feature scaling methods like MinMaxScaler for DBSCAN.
2. **Test Other Algorithms**:
    - Experiment with **Hierarchical Clustering** for comparison.

   - Apply **Gaussian Mixture Models (GMM)** for probabilistic clustering.

3. **Address Noise**:
   - Revisit noise points identified by DBSCAN and analyze them as a potential "special" customer segment.

4. **Cluster Validation**:
   - Use metrics like Davies-Bouldin Index and Dunn Index to validate clustering performance further.

5. **Business Integration**:
   - Leverage these clusters for targeted marketing campaigns and personalization strategies.

This report showcases the segmentation process, highlights insights gained, and proposes actionable steps to enhance clustering results and business value.