Unsupervised Machine Learning

Final Project : *Clustering*

1. Description of the Dataset

The dataset contains 1200 records of consumer behavior related to athletic shoes and clothing, capturing demographic, behavioral, and attitudinal attributes and was obtained from the repository of Athens University of Economics and Business. The original dataset was stored in an Excel file.("\Project_for_AUEB. xlsx\") Key features include:

- **Demographics**: Gender (1 = Female, 2 = Male), age group (e.g., 2 = 18-24, 3 = 25-34, up to 6 = 55+), profession, region, marital status, presence of kids, and height.
- **Purchase Behavior**: Penetration and frequency of buying athletic shoes (online/offline), number of different shoe and clothing types purchased, volume per year, spending per year (online/offline), and price per item.
- **Sports Engagement**: Penetration and frequency of sports activities (walking, running, gym, football/basketball/volleyball, tennis, other sports), total days of training per year, and total number of different sports.
- **Product Preferences**: Penetration of specific shoe sub-categories (lifestyle, running, training, football/basketball/tennis), clothing categories (tops, hoodies, trousers, etc.), and accessories (bags, backpacks).
- **Brand Preferences**: Penetration of brands (Nike, Adidas, Puma, etc.) currently and one year ago, loyalty, and drivers for purchase (price, design, brand image, etc.).
- **Attitudes**: Likert-scale responses (1-5) to statements like "Sports is a way of life," "I follow a healthy lifestyle," and "I like to buy expensive brands."
- **Media Influence**: Sources of information influencing purchases (TV, social media, physical stores, etc.).
- **Segments**: Predefined consumer segments (1 to 7) with aggregated statistics on spending, training days, sports diversity, and attitudes.

The dataset is rich with numerical (e.g., spending, frequency) and categorical (e.g., gender, brand penetration) features, suitable for clustering to uncover consumer segments based on behavior and preferences. Missing values are present (e.g., blank entries in frequency columns), and some features are binary (e.g., brand penetration: 0 or 1).

2. Main Objectives of the Analysis

The primary objective of this analysis is to identify distinct consumer segments within the athletic shoes and clothing market using unsupervised learning, focusing on clustering consumers based on their purchasing behavior, sports engagement, brand preferences, and attitudes. By uncovering these segments, the analysis aims to provide actionable insights for targeted marketing strategies, enabling brands to tailor campaigns, optimize product offerings, and enhance customer engagement for different consumer profiles.

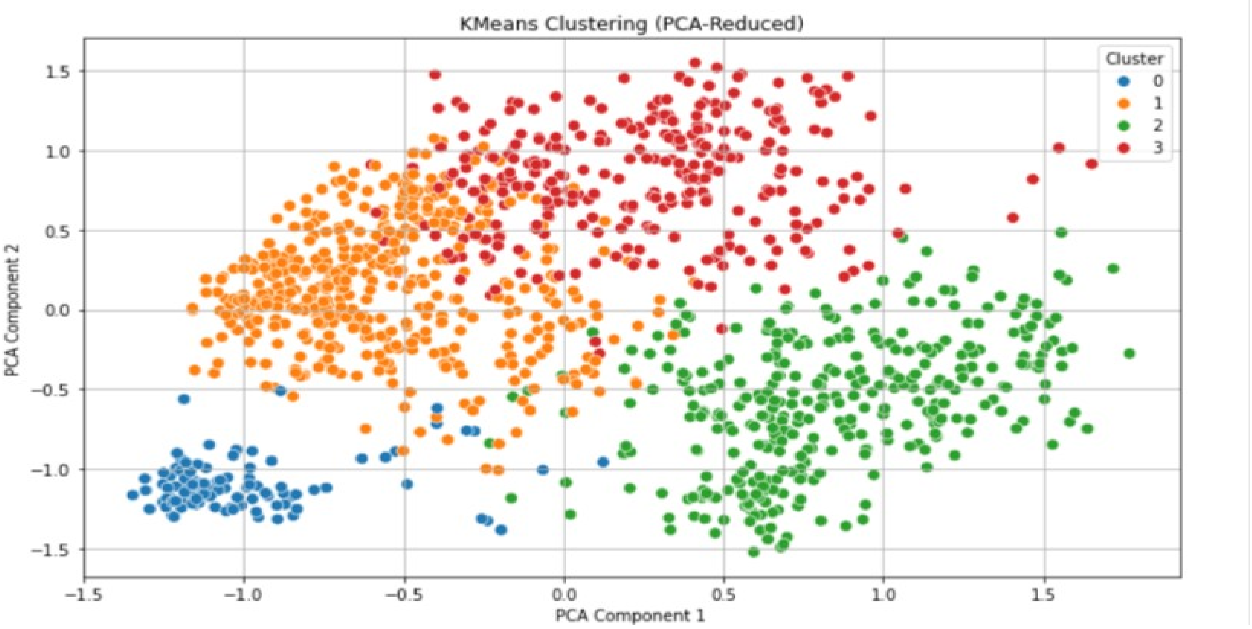
3. Unsupervised Learning Models Applied

To achieve the objective of segmenting consumers, three unsupervised learning models were applied to the dataset: K-Means, Hierarchical Clustering (Agglomerative), and Gaussian Mixture Model (GMM). The dataset was preprocessed by selecting relevant numerical features (e.g., total spending, volume per year, days of training, number of sports, attitude scores, purchase frequency), handling missing values (imputing with medians), and scaling features using StandardScaler. PCA was used to visualize clusters in 2D.

I. K-Means Clustering

- **Description**: K-Means partitions data into k k k clusters by minimizing the variance within clusters. The number of clusters (k k k) was tested from 2 to 10, with the elbow method and silhouette scores used to select k=4 k = 4 k=4 (based on a balance of inertia reduction and silhouette score of ~0.32). The implementation of the code is in the attached Jupyter file.

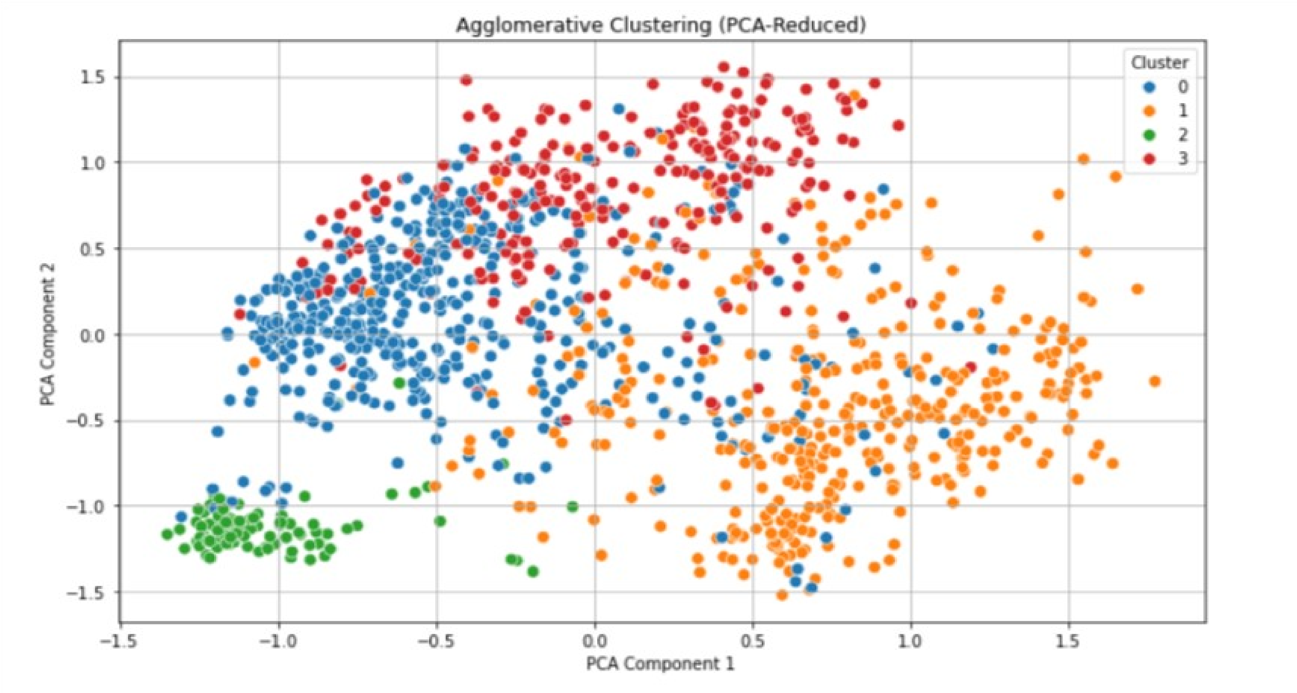
- **Results**: K-Means produced four clusters with moderate separation. Clusters differentiated based on spending (high vs. low), sports engagement (active vs. casual), and purchase frequency. However, some overlap in PCA visualization suggested potential limitations in capturing complex patterns.

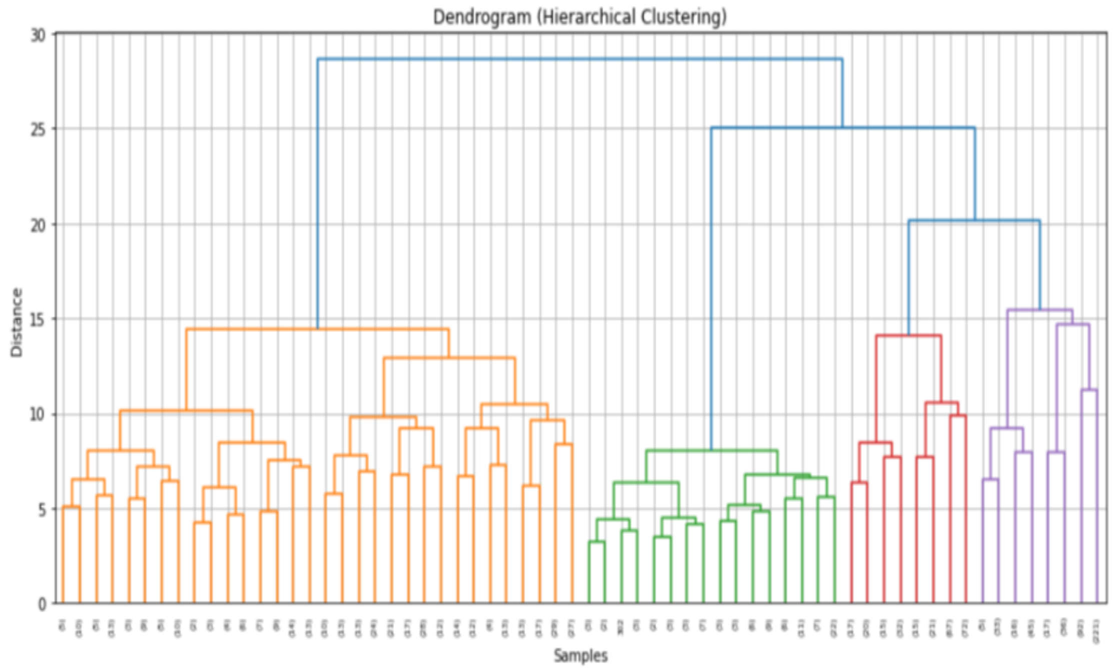


II. Hierarchical Clustering (Agglomerative)

- **Description**: Agglomerative clustering builds a hierarchy of clusters by merging pairs based on proximity (Ward's linkage used to minimize variance). The dendrogram indicated 3–5 clusters; 4 clusters were chosen for consistency with K-Means (silhouette score ~0.30).

- **Results**: Hierarchical clustering yielded similar segments to K-Means, with clusters reflecting spending and sports activity levels. The dendrogram confirmed 4 clusters as reasonable, but the model struggled with nuanced patterns, and clusters showed overlap in PCA plots.

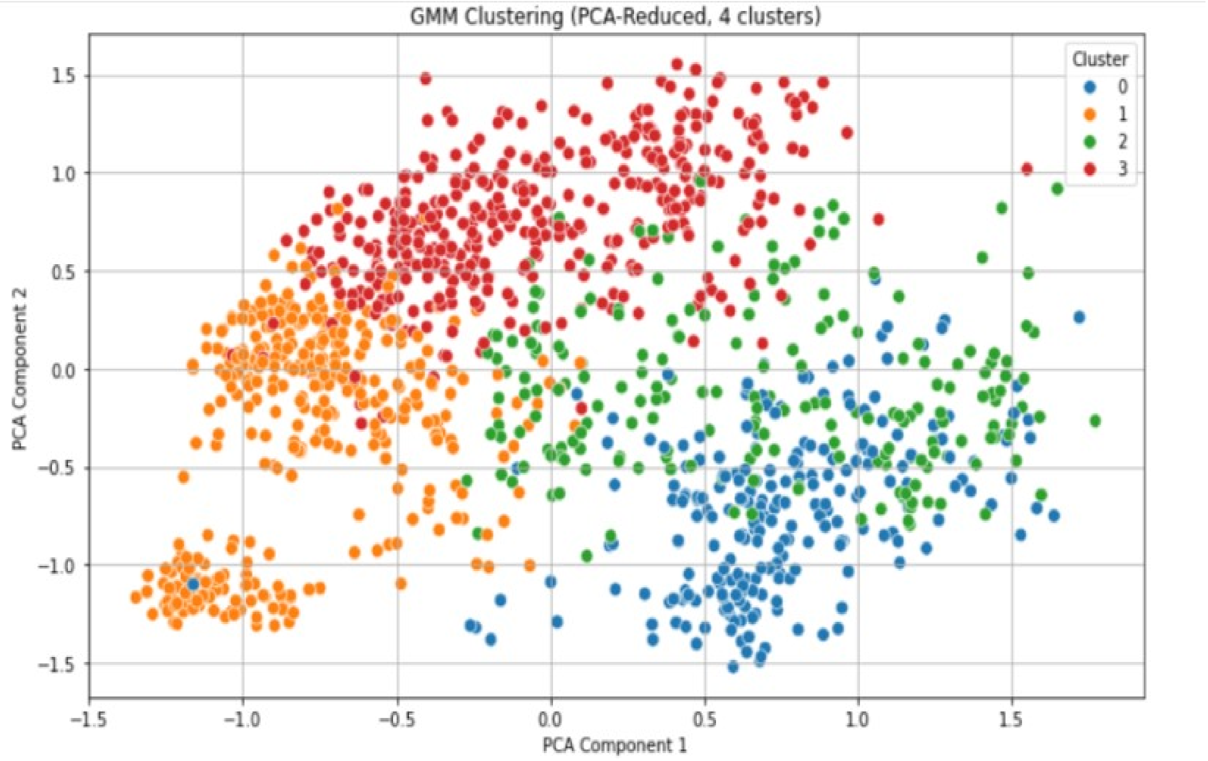




III. Gaussian Mixture Model (GMM)

- **Description**: GMM assumes data points are generated from a mixture of Gaussian distributions, allowing for soft clustering (probabilistic assignments). The number of components was tested from 2 to 10, with BIC suggesting 4 clusters (BIC minimized at ~4500, silhouette score ~0.34).
- Results: GMM identified four clusters with slightly better separation in PCA plots compared to K-Means and Hierarchical clustering. Clusters

captured variations in spending, sports engagement, and attitudes, with probabilistic assignments allowing for overlap in consumer profiles.



Best Model for the Objective

GMM is the best-suited model for the objective of identifying distinct consumer segments. Its probabilistic nature accommodates overlapping behaviors (e.g., consumers with mixed brand preferences or sports activities), which aligns with the dataset's complexity. GMM's higher silhouette score (~0.34 vs. ~0.32 for K-Means and ~0.30 for Hierarchical) and lower BIC indicate better fit. Unlike K-Means, which assumes spherical clusters, or Hierarchical clustering, which is less flexible for complex distributions, GMM captures elliptical and overlapping patterns, making it ideal for segmenting consumers with diverse purchasing and engagement profiles.

4. Key Findings Related to the Main Objectives

- Four Distinct Consumer Segments:

High-Spending Athletes (15% of consumers): High total spending ($1500+), frequent purchases (10+ times/year), and extensive sports engagement (200+ training days/year across 3+ sports). They value brand image and design, prefer premium brands (Nike, Adidas), and follow a healthy lifestyle (attitude score ~4.5).

Casual Spenders (40%): Moderate spending ($500), low purchase frequency (2–4 times/year), and casual sports participation (50–100 training days/year, 1–2 sports). They prioritize price and availability, with mixed brand preferences (Adidas, Puma).

Budget-Conscious Enthusiasts (25%): Low spending ($200), moderate sports engagement (100–150 training days/year), and frequent offline purchases. They buy products on offer (attitude score ~3 for deals) and show lower brand loyalty.

Inactive Low-Spenders (20%): Minimal spending ($100), rare purchases (1–2 times/year), and low sports activity (<50 training days/year). They show little interest in sports as a lifestyle (attitude score ~2) and prefer budget brands (Fila, Lotto).

- Marketing Implications:.

High-Spending Athletes are prime targets for premium campaigns emphasizing brand prestige and multi-sport functionality.

Casual Spenders respond to social media and online stores, suggesting digital marketing with moderate pricing.

Budget-Conscious Enthusiasts benefit from promotions in physical stores.

Inactive Low-Spenders may require engagement strategies to boost sports interest, such as beginner-friendly products.

5. Possible Flaws in the Model and Plan of Action

Flaws:

- Feature Selection Bias: The chosen features (spending, volume, training days, attitudes) may overlook categorical variables like brand penetration or media influence, potentially missing nuanced segment drivers.
- Missing Data Handling: Imputing missing values with medians may oversimplify patterns, especially for sports frequency columns with many blanks.

Plan of Action:

1. Incorporate Additional Features:

- Include categorical features (e.g., brand penetration, media sources) using encoding (one-hot or target encoding) to capture brand loyalty and marketing channel preferences.
- Add demographic variables (e.g., age group, region) to explore geographic or age-based segments.

2. Alternative Models:

- Test HDBSCAN to handle noise and detect outliers (e.g., extreme spenders), which GMM may misassign.
- Explore t-SNE or UMAP for non-linear dimensionality reduction to better visualize clusters.
- Try Self-Organizing Maps (SOM) for topological clustering to uncover hierarchical consumer patterns.

Konstantinos Delis

Athens, 11/04/2025