**Customer Segmentation (CLUSTERING)**

**Objective**

This project is dedicated for customer segmentation using clustering methods in unsupervised machine learning for the purpose of identifying groups of customers based on their behaviors and attributes.

**Benefits from Analysis**

1. ***Better marketing plan***: Understand customer behaviors to make offers, promotions and gifts aways according to income and spending score.
2. ***Customer retention***: Facilitate better services and improve satisfaction for high-spending customers.
3. ***Resource allocation***: Plan marketing budgets more effectively by focusing on valuable segments based on behaviors of customers.
4. ***Inventory management***: Predict which customer group is likely to purchase what category of products and manage stock accordingly.
5. ***Customer lifecycle management***: Identify high-value customers, potential churners, and new leads based on demographic clusters.

**Methods**

Three main clustering methods were used in this project and selected the best one:

1. K-Means
2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and
3. Gaussian Mixture Model (GMM)

**Data Information**

The small shopping mall customer data set [from [Kaggle](https://www.kaggle.com/datasets/shwetabh123/mall-customers)] is used as a sample data, which contains **200** **records**, each row representing individual customers.

**Attributes:**

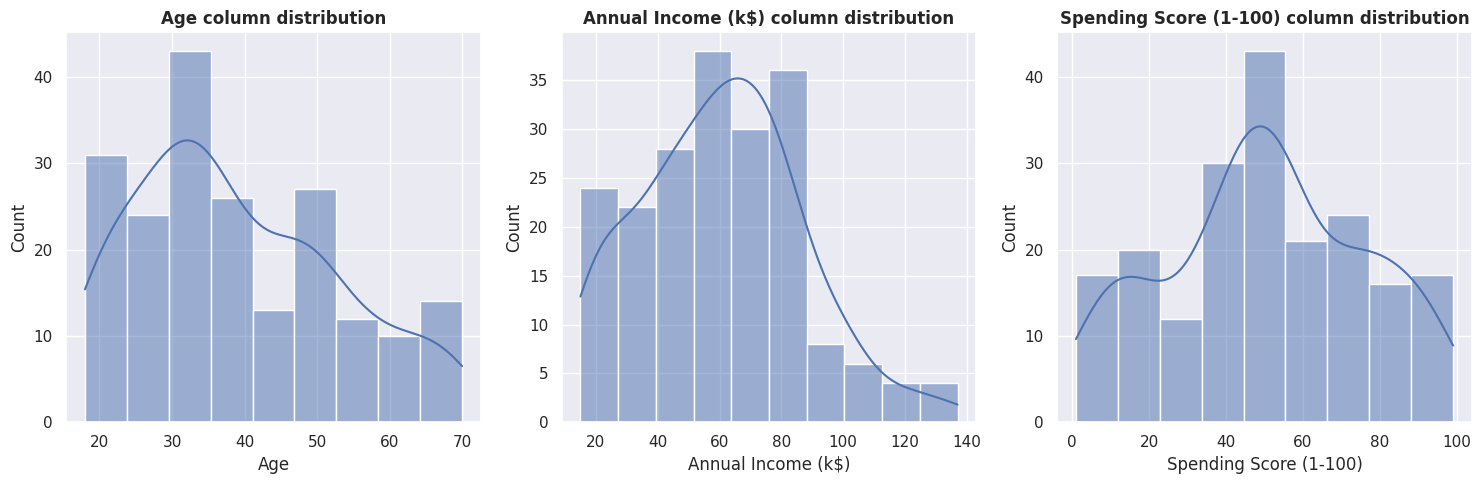
| **Column Names** | **Description** |
| --- | --- |
| *CustomerID* | Unique ID assigned to each customer |
| *Gender* | Gender of each customer (Male/ Female) |
| *Age* | Age of each customer |
| *Annual Income (k$)* | Approximate annual income in thousands of dollars |
| *Spending Score (1-100)* | Score assigned based on customer spending behavior and shopping patterns (higher score = higher spending tendency) |

**Data Exploration**

The data has no missing and duplicated information.

* ***Annual Income (k$)*** and ***Spending Score (1-100)*** are used as major columns to identify types of customers.
* ***Gender*** and ***Age*** columns are used to find some hidden insights.
* ***CustomerID*** column is not useful here.

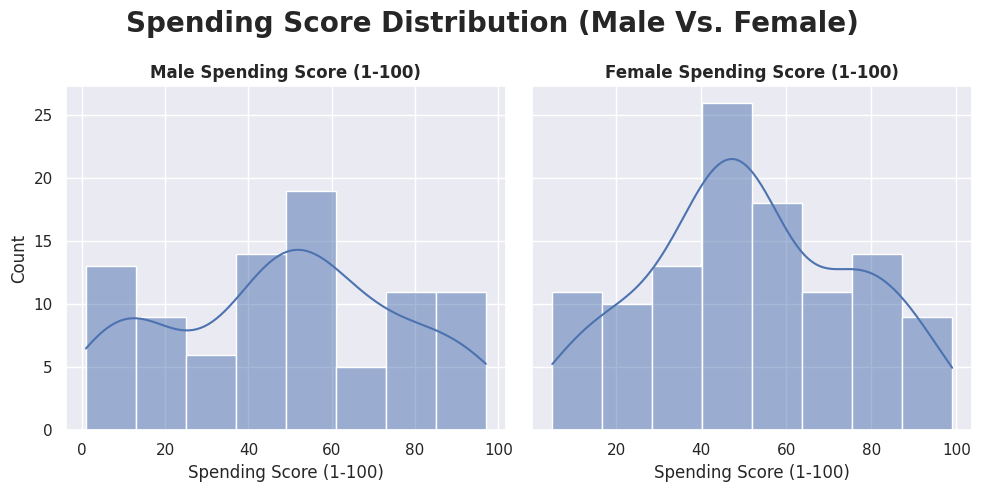
**Analysis**



**Age**: Ranging from 18 to 70 years, with a mean of approximately 38.85. The distribution shows a higher concentration of younger customers (around 20-35 years old).

**Annual Income (k$):** Ranges from $15k to $137k, with a mean of about $60.56k. The distribution is right-skewed, meaning few customers’ annual incomes are with higher incomes.

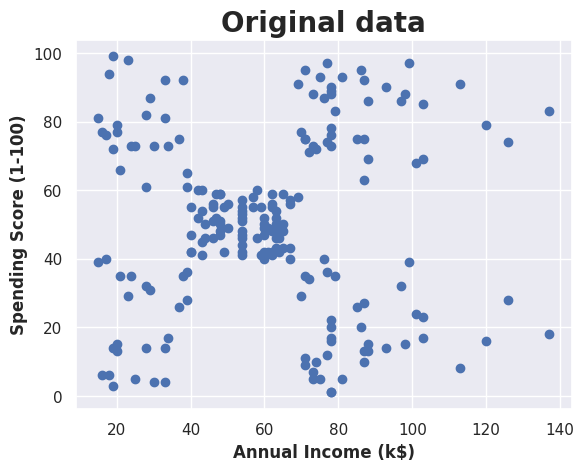
**Spending Score (1-100)**: Ranges from 1 to 99, with a mean of approximately 50.20. The distribution appears relatively uniform across the range.



**Gender Vs. Spending Score:** Female spending score (ranging from 5 to 99, mean value: 51.53) is higher than male spending score(ranging from 1 to 97, mean value: 48.51).

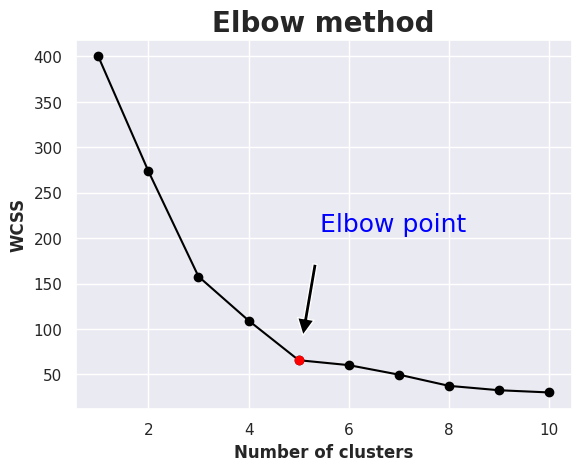
**Original Data Visualization**

The original data before scaling based on majors columns (with 200 records).

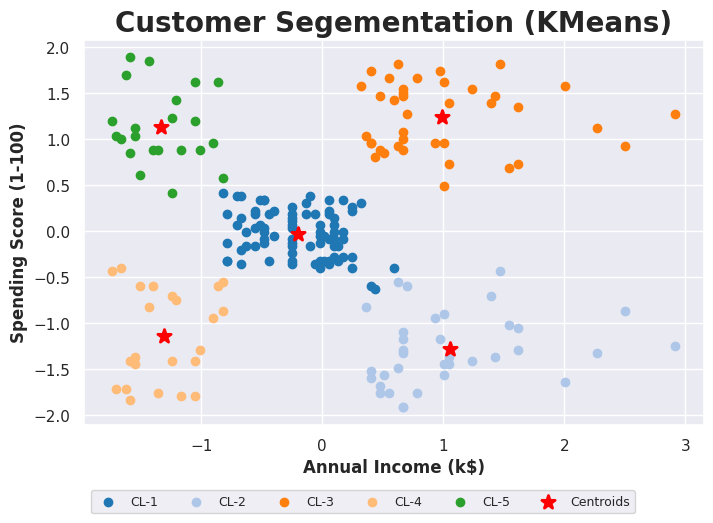


**Building Models**

**K-Means (K = 5)**

****

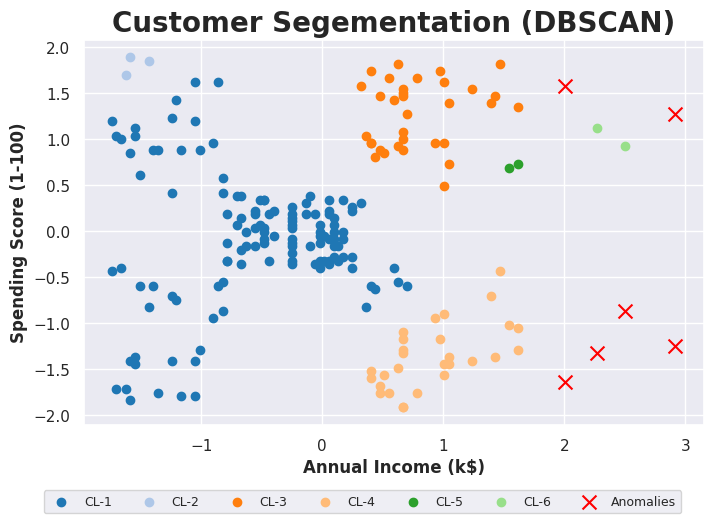
* **Elbow method** is used to find the best number of clusters K, ranging from 1 to 10.
* **K = 5** is chosen based on the result.
* KMeans(n\_clusters=5, init='k-means++')



Silhouette Score = 0.554657

**DBSCAN**

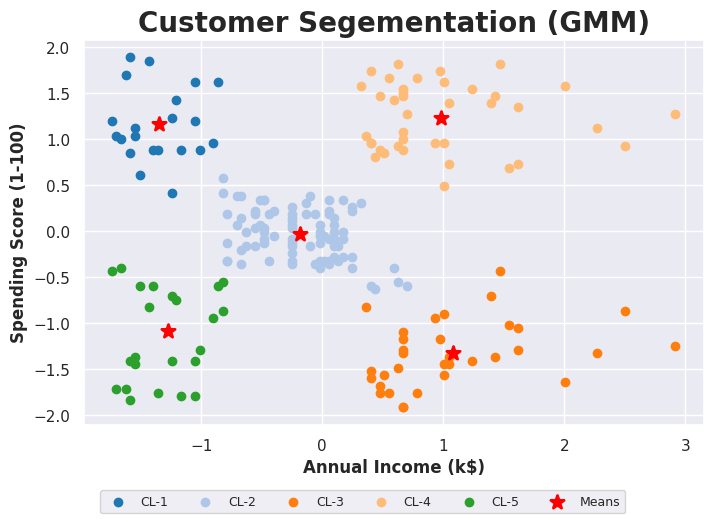
**DBSCAN(min\_samples=2, eps=10)**

****

Silhouette Score = 0.322194

**Gaussian Mixture Model**

**GaussianMixture(n\_components=5)**

****

Silhouette Score = 0.553689

**Model Evaluations and Recommendation**

KMeans is the best (~ near GMM) to choose for customer segmentation in clustering in this case according to model evaluations using **Silhouette Score**.

| **Model Names** | **Silhouette Score** |
| --- | --- |
| **KMeans** | 0.554657 |
| **GMM** | 0.553689 |
| **DBSCAN** | 0.322194 |

The final representation for seven type of customers are as per following:

1. High Income, High Spending
2. Moderate Income and Spending
3. Low Income and Spending
4. Low Income, High Spending
5. High Income, Low Spending

| **Customer\_Type** | **Annual Income (k$**) | **Spending Score (1-100)** |
| --- | --- | --- |
| High Income, High Spending | 55.296296 | 49.518519 |
| High Income, Low Spending | 25.727273 | 79.363636 |
| Low Income and Spending | 86.538462 | 82.128205 |
| Low Income, High Spending | 26.304348 | 20.913043 |
| Moderate Income and Spending | 88.200000 | 17.114286 |

**Key findings**

The KMeans clustering algorithm was found to be the best model for the given features of annual income and spending scores of the mall customer dataset. And it was effective in segmenting five different groups of customers as discussed earlier.

**Completed code file**: [[LINK](https://drive.google.com/file/d/1PSqZSfYPdlzPkYMW1nFMKCMJSB1EC0a_/view?usp=sharing)]