Internship Project Report

Project Title: - Customer Segmentation using K-Means Clustering

Problem Statement: -Use K-Means clustering to segment customers based on behavioral and demographic data to enable targeted marketing strategies.

Introduction: - Customer segmentation is the process of dividing a customer base into groups based on common characteristics such as age, income, and spending behavior. This helps businesses design targeted marketing strategies, improve customer satisfaction, and enhance profitability.

In this project, K-Means clustering is applied on a dataset containing customer demographic and behavioral attributes to identify distinct customer group

Week 1: - Data Collection & Preprocessing

- Collected and cleaned customer dataset.
- Performed **Exploratory Data Analysis (EDA)** to understand distributions, outliers, and trends.
- Normalized and scaled data for clustering compatibility.
- Key insights from EDA:
  - Age, Annual Income, and Spending Score are critical features.
  - Clear variation in customer spending behavior was observed.

1.Dataset Description

- For this project, we will use the **Mall Customers Dataset**. It contains customer details such as age, spending score, and income, which will help us cluster customers based on shopping patterns.
- Columns: -

1. **CustomerID**: Unique customer identifier.
2. **Gender**: Male/Female.
3. **Age**: Age of the customer.
4. **Annual Income (k$)**: Yearly income of the customer in thousands.

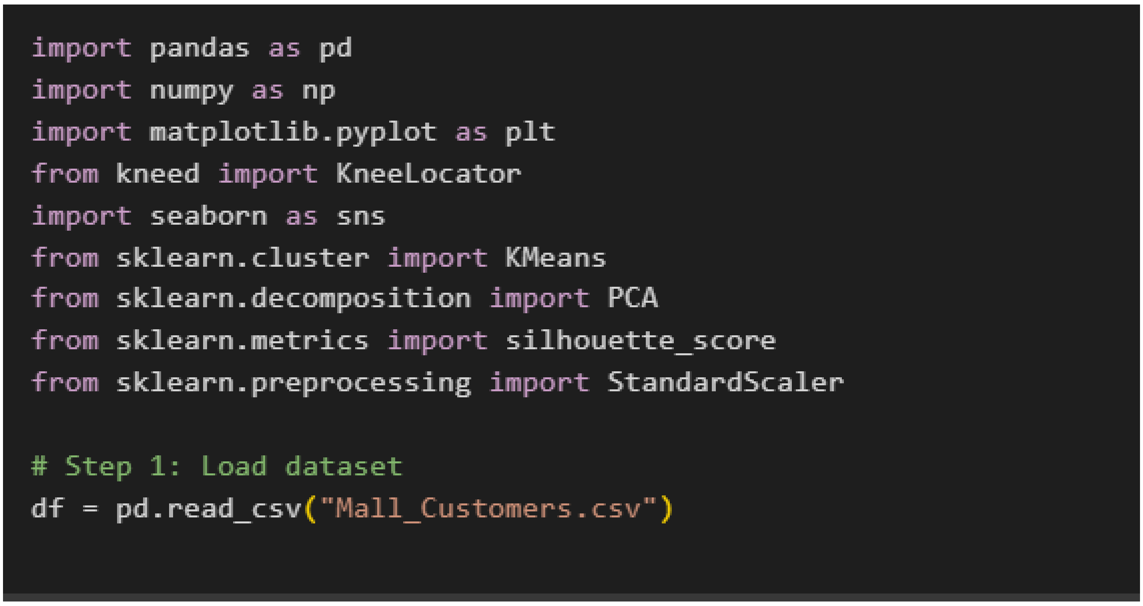
5. **Spending Score (1-100)**: Customer spending behavior score assigned by the mall.

2.Tools & Libraries

We will use the following Python libraries:

- **pandas, numpy** (Data manipulation)
- **matplotlib, seaborn** (Visualization)
- **scikit-learn** (Machine learning implementation)

Importing Libraries and Loading Dataset

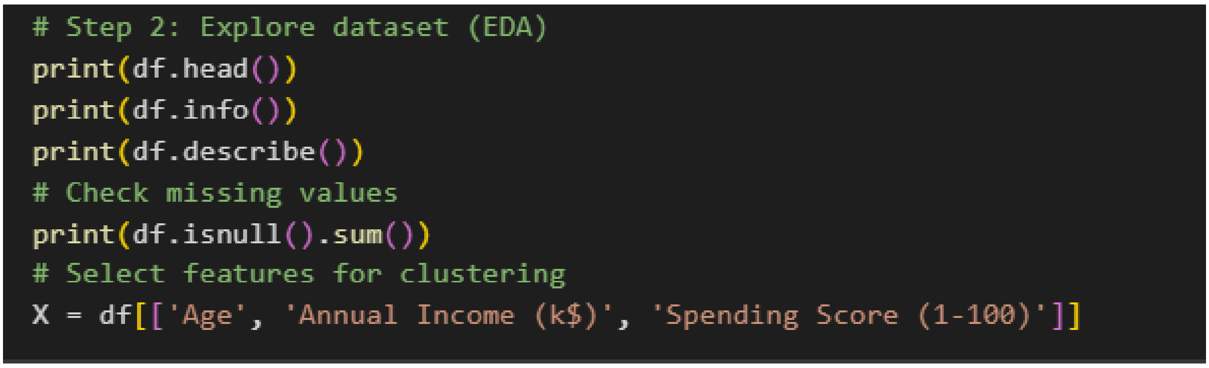


```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from kneed import KneeLocator
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import StandardScaler

# Step 1: Load dataset
df = pd.read_csv("Mall_Customers.csv")
```

3.Exploratory Data Analysis (EDA)

- Checking for Missing Values
- Descriptive Statistics
- We select the Age, Annual Income and Spending Score columns for clustering.
- Standardize the data using StandardScaler() to normalize different scales.



```
# Step 2: Explore dataset (EDA)
print(df.head())
print(df.info())
print(df.describe())
# Check missing values
print(df.isnull().sum())
# Select features for clustering
X = df[['Age', 'Annual Income (k$)', 'Spending Score (1-100)']]
```

Week 2: Model Building & Evaluation

- Applied **K-Means Clustering** algorithm.
- Determined optimal number of clusters using:
  - **Elbow Method** (WCSS curve).
  - **Silhouette Score** for cluster validation.
- Visualized clusters using:
  - **2D scatter plots** (Age vs Spending Score, Income vs Spending Score).
  - **PCA (Principal Component Analysis)** for dimensionality reduction to 2D.

1. Determining the Optimal Number of Clusters

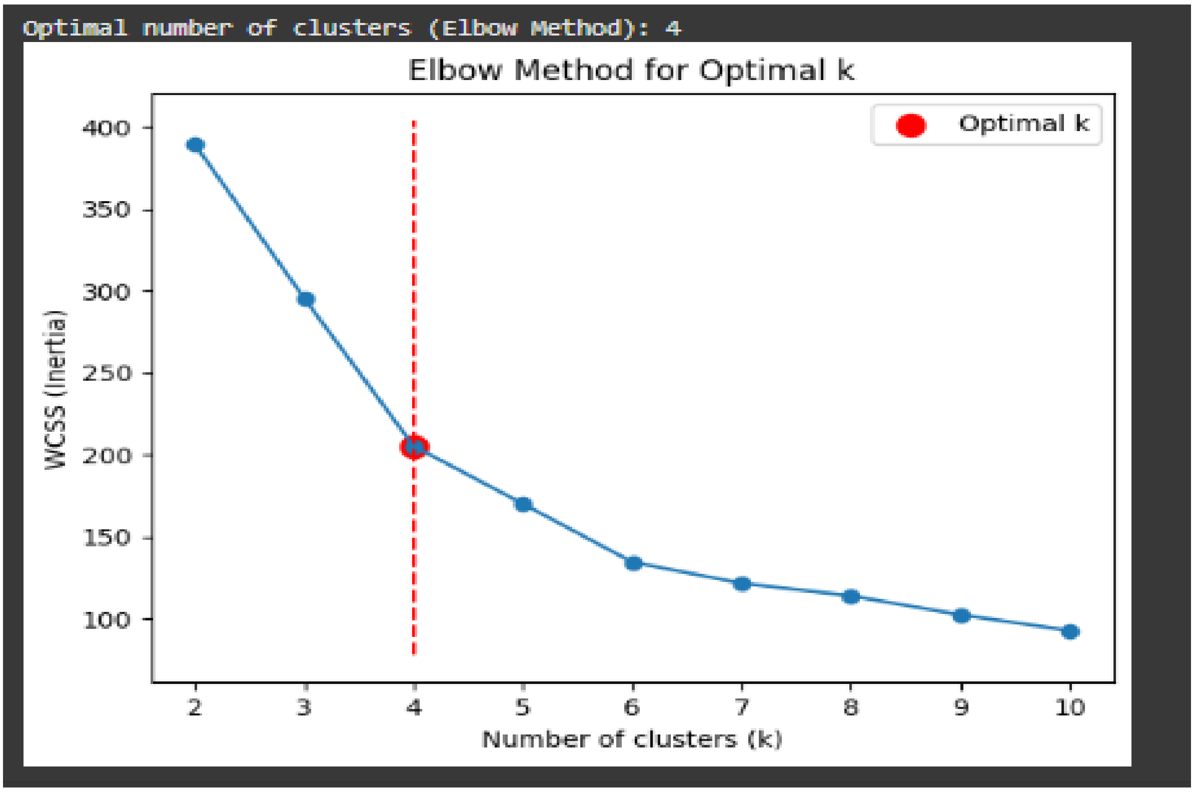
Determining the optimal number of clusters means finding the best number of groups that naturally exist in your dataset, so that your clustering results are both accurate and useful.

A) Elbow Method:-

The elbow method works by plotting the WCSS against the number of clusters and finding the point where the decrease in WCSS slows down, creating an elbow-like shape in the plot. This elbow-like shape indicates the optimal number of clusters, with the region before the elbow being under-fitting and the region after being over-fitting.

- **Idea**: Increase number of clusters (k) and compute **WCSS (Within-Cluster Sum of Squares)**.
- **Plot**: WCSS vs. k.

- **Elbow Point**: The spot where WCSS reduction slows down (curve bends like an elbow).
- **Usefulness**: Provides a quick visual way to guess the best k.



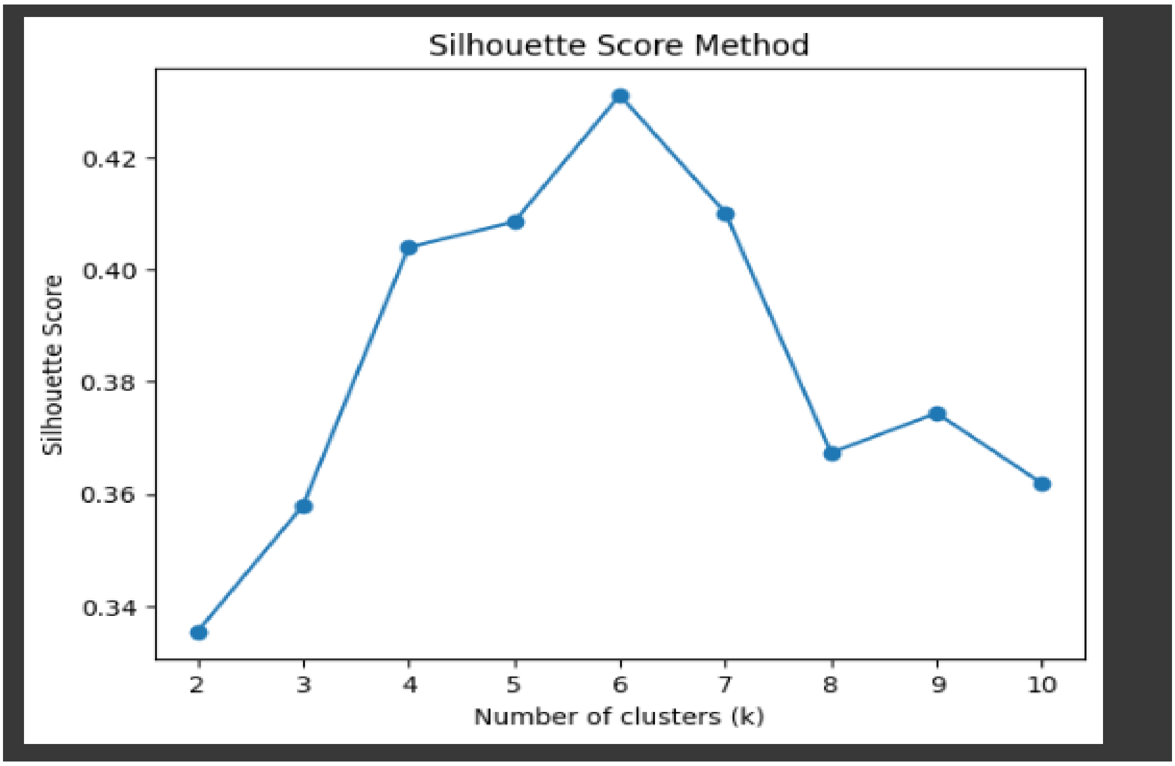
```
Optimal number of clusters (Elbow Method): 4
```

A) **Silhouette Score for cluster validation: -**

Silhouette analysis refers to a method of interpretation and validation of consistency within clusters of data. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

- **Formula**: Measures how similar a point is to its own cluster vs. other clusters.
- Score range: -1 to +1.
- **+1** → well-separated clusters.
- **0** → overlapping clusters.

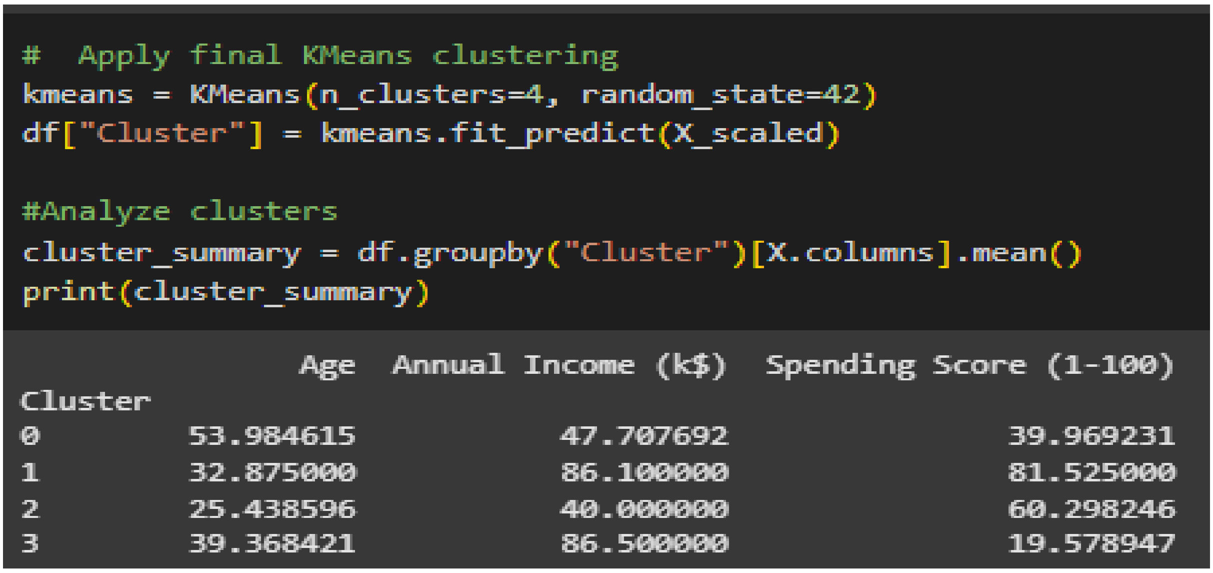
- **-1** → wrong clustering (misclassified).



2.KMeans Clustering

K-Means is a widely used unsupervised machine learning algorithm designed to group data points into clusters based on their similarity. It is particularly effective for segmenting data into distinct groups when labels are not available. The algorithm minimizes the distance between data points and their respective cluster centroids.

After deciding on the optimal number of clusters, we instantiate a KMeans object with 4 clusters and fit it to the scaled data.



```
#  Apply final KMeans clustering
kmeans = KMeans(n_clusters=4, random_state=42)
df["Cluster"] = kmeans.fit_predict(X_scaled)

#Analyze clusters
cluster_summary = df.groupby("Cluster")[X.columns].mean()
print(cluster_summary)
```

```
               Age  Annual Income (k$)  Spending Score (1-100)
Cluster
0        53.984615           47.707692               39.969231
1        32.875000           86.100000               81.525000
2        25.438596           40.000000               60.298246
3        39.368421           86.500000               19.578947
```

3. Visualisation of Clusters

After applying K-Means clustering and determining the optimal number of clusters using the Elbow Method and Silhouette Score, the clusters were visualized to understand customer distribution and validate the model.

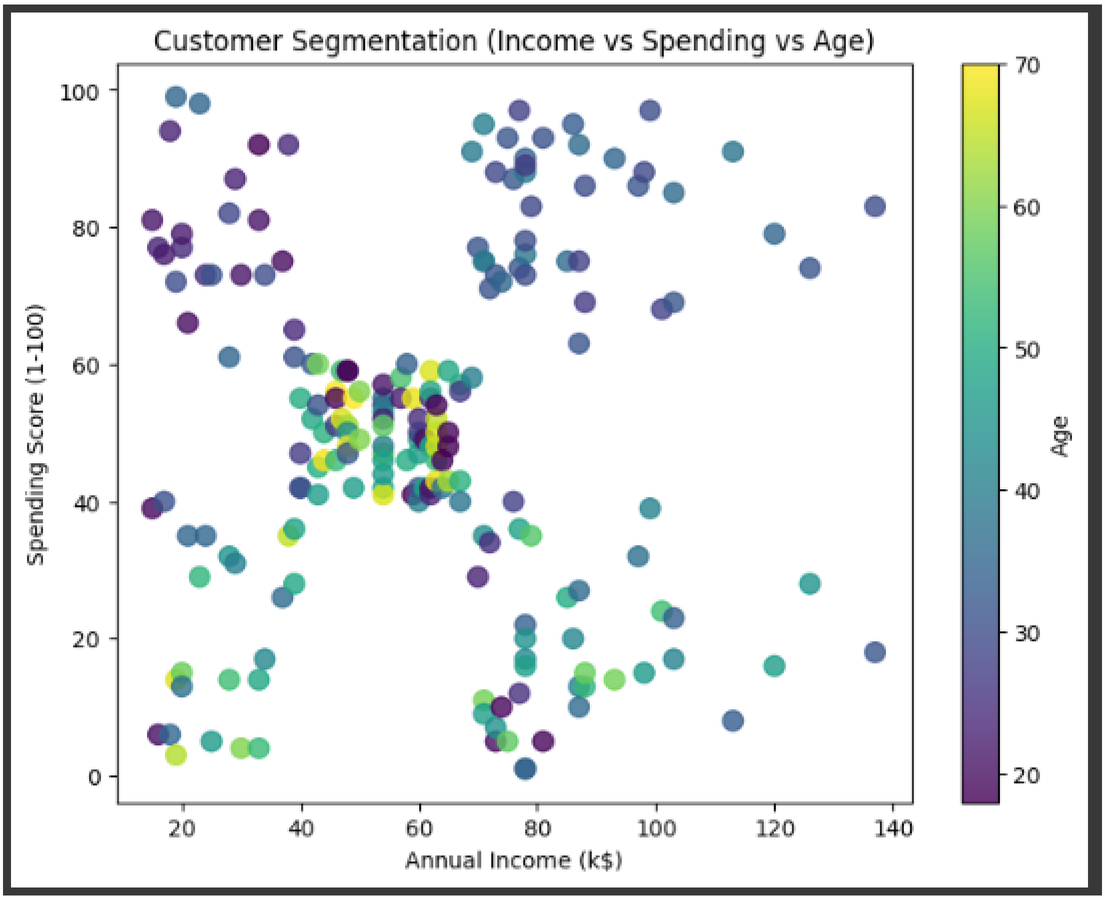
A) **2D scatter plots** (Age vs Spending Score, Income vs Spending Score).

❖ **Age vs Spending Score**

- Clear grouping of customers based on spending behavior across age brackets.
- **Insights:**
  - Younger customers (20–35 years) formed clusters with high spending scores → *potential premium segment*.
  - Middle-aged customers showed moderate spending behavior.
  - Older customers were clustered with lower spending scores → *budget-conscious group*.

❖ **Income vs Spending Score**

- Showed how spending varies across income groups.
- **Insights:**
  - A segment of **high-income but low-spending customers** → *untapped potential*.
  - A segment of **high-income and high-spending customers** → *ideal target group*.
  - Customers with lower income but moderate/high spending → *value-driven shoppers*.



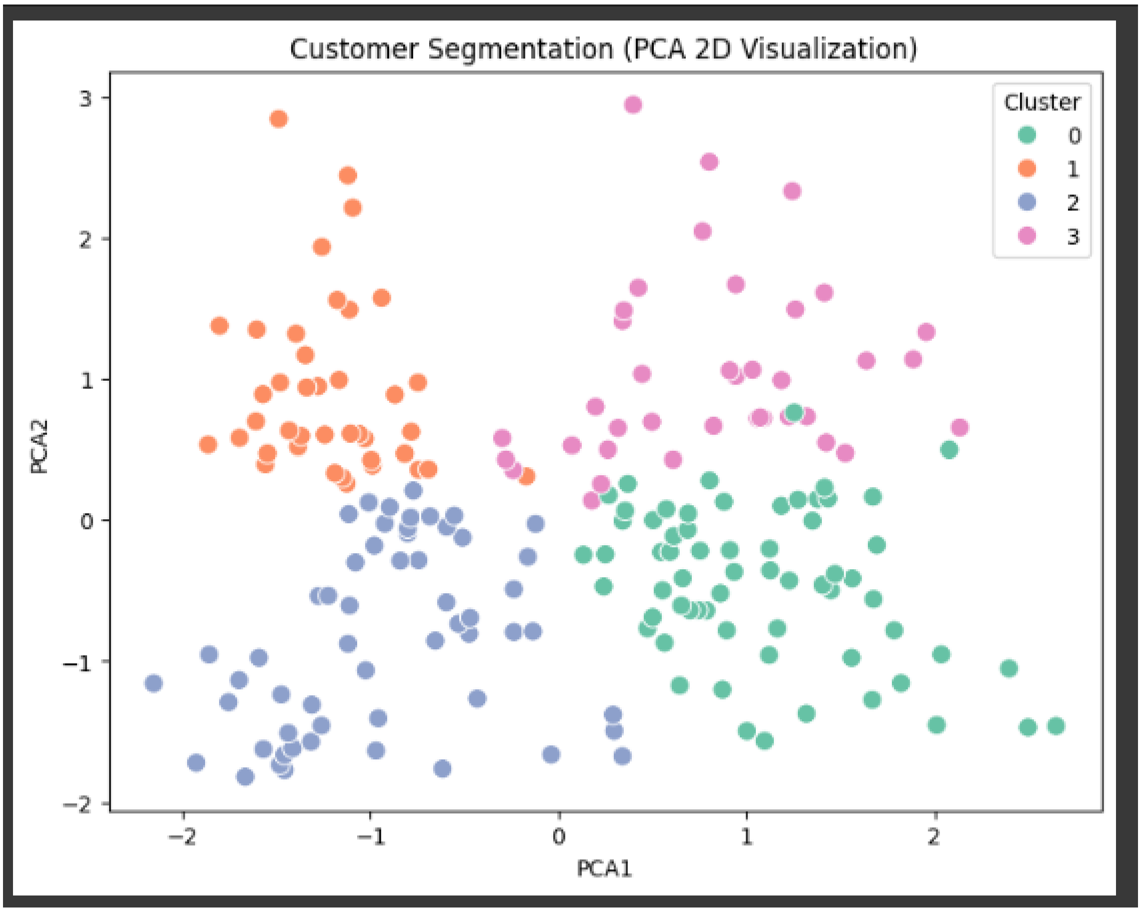
B) **PCA (Principal Component Analysis)** for dimensionality reduction to 2D.

❖ **Purpose:** Since clustering was performed on multiple features (Age, Gender, Income, Spending Score), PCA was applied to reduce dimensions into two principal components for holistic visualization.

❖ **Outcome:**

- The PCA scatter plot showed **well-separated clusters**, validating that K-Means captured meaningful customer segments.
- Some overlap was observed, which is natural given similarities between certain customer groups.

❖ **Benefit:** Unlike scatter plots restricted to two variables, PCA visualized how clusters are distributed considering all features together.



Week 3: Cluster Profiling & Marketing Strategy

1) Profiled clusters based on Age, Income, and Spending behavior.

2) Example:

- Cluster 1: *Young high-spenders* → premium products.
- Cluster 2: *High-income but low-spending* → potential targets for upselling.
- Cluster 3: *Low-income moderate spenders* → budget product segment.

3) Recommended targeted marketing strategies for each cluster.

4) Drafted initial cluster summary report.

1. Cluster Profiling

Cluster ID	Age Range	Income Range	Spending Behavior	Cluster Traits
Cluster 1	20–35 years	Medium to High	High Spending Score	Young, enthusiastic shoppers; trend-seekers
Cluster 2	30–50 years	High	Low Spending Score	Affluent but conservative; untapped potential
Cluster 3	25–40 years	Low to Medium	Moderate Spending	Price-sensitive but engaged; value seekers
Cluster 4	40–65 years	Medium	Low Spending Score	Budget-conscious; prefer essentials
Cluster 5	Mixed ages	High	High Spending Score	Loyal premium customers; high-value group

2. Marketing Strategy Recommendations

- **Cluster 1 (Young high-spenders):**
  - Focus on lifestyle and premium product promotions.
  - Leverage digital marketing and seasonal offers.
- **Cluster 2 (High-income low-spenders):**
  - Use personalized offers and loyalty programs to increase engagement.
  - Highlight exclusivity and premium services.
- **Cluster 3 (Value seekers):**
  - Offer discounts, bundles, and referral programs.
  - Position products as affordable but high-quality.
- **Cluster 4 (Budget-conscious group):**
  - Promote essential items and economy packs.
  - Engage through family-oriented or needs-based marketing.
- **Cluster 5 (Premium loyal customers):**
  - Provide VIP membership, early access to new products, and luxury perks.
  - Maintain relationships with strong after-sales support.

Thank You