

# Customer Segmentation System

**Project Title:** Advanced Customer Analytics using K-Means Clustering

**Objective:** To identify distinct consumer groups within a retail environment based on income and spending patterns using Unsupervised Machine Learning.

## 1. Project Planning & Design

The system follows a **modular functional architecture**, separating data ingestion, mathematical optimization, and graphical reporting. This ensures the code is maintainable, scalable, and easy to debug.

### Program Structure

The application is divided into five logical layers:

1.  **Data Ingestion Layer:** Loads raw .csv data into memory using the Pandas library.
2.  **Preprocessing Layer:** Filters features and applies **Z-score Standardization** to ensure unit variance.
3.  **Optimization Layer:** Automatically determines the mathematical "Best K" using the **Silhouette Coefficient**.
4.  **Training Layer:** Executes the \$K\$-Means clustering algorithm on the scaled dataset.
5.  **Visualization Layer:** Generates a suite of 5+ analytical charts for stakeholder review.

## 2. Algorithms & Logic

The program relies on three core computational processes to ensure segmentation accuracy.

### A. Feature Scaling (Standardization)

To prevent features with higher numerical ranges (like Annual Income) from dominating the distance calculation, we use:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- $x$  = original value (example: customer income)
- $\mu$  (**mu**) = mean (average) of all values
- $\sigma$  (**sigma**) = standard deviation (spread of data)
- $z$  = standardized value (scaled value)

Where  $x$  is the value,  $\mu$  is the mean, and  $\sigma$  is the standard deviation.  $\square$

## B. Automatic Cluster Detection (Silhouette Score)

The system iterates through k values (2 to 10) and calculates the **Silhouette Coefficient**.

- **Goal:** Maximize the score to find the point where clusters are most distinct and dense.
- **Interpretation:** A score closer to 1 indicates that the sample is far away from the neighboring clusters.

## C. K-Means Clustering

The algorithm minimizes the **Within-Cluster Sum of Squares (WCSS)** through an iterative process:

1. **Initialize:** Randomly pick K cluster centers (centroids).
2. **Assignment:** Assign each customer to the nearest centroid using **Euclidean Distance**.
3. **Update:** Calculate the new mean of each cluster and move the centroid.
4. **Convergence:** Repeat until centroids no longer move significantly.  $\square$

## 3. $\square$ System Flowchart (Pseudocode)

The following logic represents the execution flow of the program:

Step	$\square$ Phase	$\square$ Description	$\square$ Output / Result
------	-----------------	-----------------------	---------------------------

<b>1</b>	<b>Data Ingestion</b>	Load Mall_Customers.csv using Pandas.	Raw DataFrame
<b>2</b>	<b>Feature Selection</b>	Isolate Annual Income and Spending Score for analysis.	Feature Subset
<b>3</b>	<b>Data Scaling</b>	Apply StandardScaler to normalize features (Mean=0, Std=1).	Scaled Matrix
<b>4</b>	<b>Hyperparameter Tuning</b>	Iterate \$k\$ from 2 to 10 and calculate <b>Silhouette Scores</b> .	Optimal \$K\$ Value
<b>5</b>	<b>Model Training</b>	Initialize and fit the <b>K-Means</b> algorithm using the best \$k\$.	Trained Model
<b>6</b>	<b>Label Assignment</b>	Map cluster IDs back to the original customer dataset.	Segmented Data
<b>7</b>	<b>Visual Analysis</b>	Generate Scatter, Box, and Pairplots for business insights.	Graphical Reports

## 4. □ Class & Data Flow Design

The following table describes the internal data flow and responsibilities of each module:

Module Name	Responsibility	Input	Output
<b>DataLoader</b> □	CSV I/O Management	File Path	Pandas DataFrame
<b>FeatureEngineer</b> □	Scaling & Transformation	Raw Features	Scaled Matrix
<b>ModelOptimizer</b>	Automatic K-Selection	Scaled Matrix	Optimal K-Value
<b>ClusterModel</b> □	Machine Learning Training	Scaled Matrix + K	Cluster Labels
<b>ReportingEngine</b> □	Data Visualization	Data + Labels	Matplotlib/Seaborn Plots

## 5. □ Visual Analysis Plan

The output provides four key analytical perspectives to ensure the business can act on the data:

- **Segmentation Map:** A scatter plot identifying the 5 types of customers (e.g., "Sensible," "Careless," "Target").
- **Income Distribution:** Boxplots comparing the financial health and "spread" of each segment. □
- **Volume Analysis:** Count plots identifying which segment holds the most customers for resource allocation.
- **Correlation Heatmap:** Identifying if income and spending have a hidden linear relationship. □