

Customer Segmentation Project Report

1. Introduction

In today's competitive market, understanding customer behavior is essential for improving business performance. Customers differ in age, income, spending habits, loyalty, and lifestyle preferences. Treating all customers in the same way leads to ineffective marketing and wasted promotional budget.

Customer Segmentation is a data-driven method for dividing customers into smaller, meaningful groups based on shared characteristics. Once grouped, companies can design personalized marketing strategies for each group to increase sales, improve customer satisfaction, and strengthen long-term relationships.

In this project, we use K-Means Clustering to segment customers and analyze their purchasing behavior. A Streamlit dashboard is implemented for interactive visualization and real-time cluster exploration.

2. Problem Statement

Businesses often struggle to identify:

- Which customers spend more
- Which customers are loyal
- Which customers need encouragement or discounts

The aim of this project is to group customers based on demographic, financial, and shopping behavior attributes so the business can:

- Target the right customers with the right offers
- Increase customer retention
- Reduce unnecessary marketing costs
- Enhance brand loyalty

3. Dataset Description

The dataset represents retail store customers with demographic and behavioral properties.

Attributes include:

- CustomerID
- Gender
- Age
- Annual Income

- Spending Score
- Tenure
- Purchase Frequency
- Satisfaction Score
- Marital Status
- Occupation
- Credit Score
- Monthly Savings

Dataset Quality:

- No missing values
- No duplicate records
- Realistic numeric ranges
- Dataset expanded for richer clustering patterns

4. Data Preprocessing

4.1 Data Cleaning

No missing or duplicate values were present. Data was directly ready for processing.

4.2 Categorical Encoding

Categorical features like Gender, Marital Status, and Occupation were kept as-is because they were not used in final clustering. Encoding was optional based on requirement.

4.3 Feature Scaling

Numerical features were standardized using StandardScaler so that all values contribute equally to clustering.

4.4 Importance of Scaling

Without scaling, features like Annual Income would dominate clustering while smaller-scale attributes like Spending Score would be ignored.

5. Exploratory Data Analysis (EDA)

- Histograms: Most customers are aged 25–50 and income varies significantly.
- Boxplots: Spending Score shows natural behavioral separation.
- Pairplots: Some patterns visually suggest cluster structures.

- Correlation Heatmap: Spending behavior has weak correlation with age but aligns with satisfaction.

EDA confirmed that clustering was meaningful for this dataset.

6. Clustering Using K-Means

K-Means was chosen because:

- It works effectively on numeric data
- It is simple and fast
- It is commonly used in marketing segmentation

To choose the number of clusters:

- Elbow Method suggested $K = 3$
- Silhouette Score was highest at $K = 3$

Therefore, the final number of clusters selected was 3.

PCA was used for visualizing clusters in 2D and 3D space.

7. Cluster Profiling

Cluster 0: Age ~ 43, mid income, low spending → Practical buyers who purchase only when necessary.

Cluster 1: Age ~ 35, high income, medium-high spending → Premium buyers who prefer brand value.

Cluster 2: Age ~ 43, low income, high spending, highest tenure → Loyal emotional buyers who value long-term relationship.

8. Marketing and Business Strategies

Cluster 0:

- Send need-based targeted offers
- Provide loyalty points to increase retention

Cluster 1:

- Launch premium membership and exclusive offers
- Provide early access to collections and product launches

Cluster 2:

- Offer bundle discounts and festival offers

- Use relationship-based and personalized communication

9. Conclusion

The project successfully grouped customers into three meaningful clusters. This segmentation helps the business optimize marketing strategies, improve customer retention, and increase revenue efficiency. The Streamlit dashboard allows users to explore clusters interactively, making insights easier to understand and apply.