Project Report

on

Market Segmentation

Submitted by

Student Names	Reg. Numbers
P.D.S.Suneeta	24BEC042
Anu Priya	24BEC005
Neel Gupta	24BEC032
Sharada I	24BEC056
J.S.S.R.Durgesh	24BEC016

Under the guidance of

Deepak K.T

DEPARTMENT OF ELECTRONICS AND COMMUNICATIONS
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

Contents

1 Introduction

Customer segmentation is a core task in retail analytics for understanding buying behavior and identifying revenue-driving customer groups.
This project applies **RFM** (Recency–Frequency–Monetary) modeling followed by clustering to segment customers of a large real-world online retail dataset.

1.1 Problem Definition

Retail businesses deal with customers of varied purchasing patterns. Treating all customers equally results in inefficient marketing.
The aim is to segment customers into meaningful groups using data-driven RFM metrics and machine learning clustering.

1.2 Objectives

- Clean and preprocess transactional data
- Compute RFM features per customer
- Log-transform and remove outliers
- Determine optimal number of clusters
- Compare *KMeans vs GMM* using silhouette score
- Visualize clusters using PCA
- Generate actionable customer segment insights

1.3 Dataset Description

- File used: **online_retail_II.xlsx**

- Total Records (raw): **407,664**
- Total Customers (raw RFM): **4,312**
- Customers after outlier removal: **4,239**

2 Related Work

Customer segmentation has traditionally relied on the RFM framework, which evaluates customers using Recency, Frequency, and Monetary value. RFM has been widely used in marketing because it only needs transactional data and provides meaningful behavioral indicators. Many studies combine RFM with unsupervised clustering, especially K-Means, due to its simplicity and scalability for large customer datasets. However, K-Means assumes spherical clusters, so several works also apply Gaussian Mixture Models (GMM) to capture more flexible cluster shapes.

Recent research highlights the importance of data preprocessing such as log transformation, outlier removal, and scaling—to improve clustering accuracy. The Silhouette Score and Elbow Method are commonly used to evaluate the optimal number of clusters. Visualization techniques like PCA help interpret how customer groups differ across RFM dimensions.
This project follows these established approaches and integrates them into a complete segmentation pipeline.

3 Data and Methods

3.1 Preprocessing Steps

- Removed rows with missing **Customer ID**
- Filtered out negative/zero **Quantity** and **Price**
- Excluded cancelled invoices
- Created **TotalPrice = Quantity × Price**
- Converted **InvoiceDate** to proper datetime format

Customer ID	Recency	Frequency	Monetary
12346.0	165	11	372.86
12347.0	3	2	1323.32
12348.0	74	1	222.16
12349.0	43	3	2671.14
12351.0	11	1	300.93

Table 1: First 5 rows of RFM

3.2 RFM Feature Engineering

For each customer:

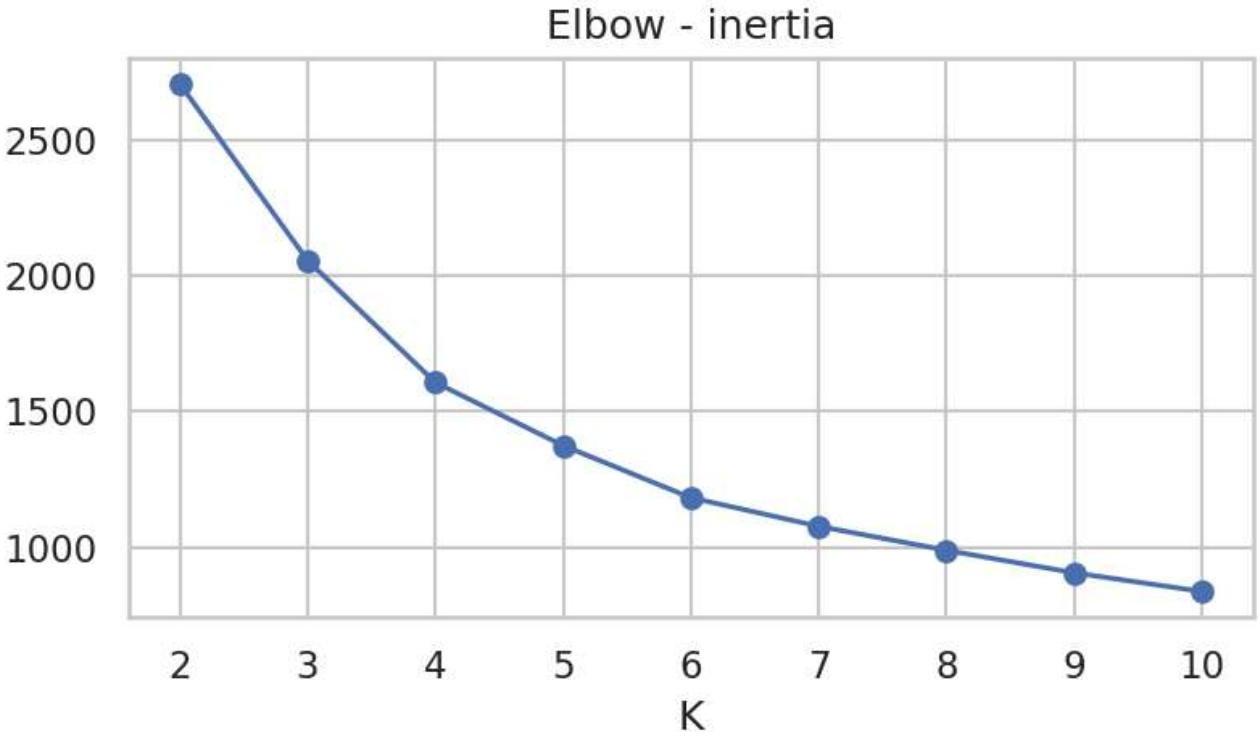
- **Recency** = Days since last purchase
- **Frequency** = Unique invoice count
- **Monetary** = Total amount spent

RFM table saved as: `rfm_raw_auto.csv`

3.3 Transformations

- Applied **log1p** on RFM to handle right-skewed distributions
- Removed outliers via **IQR (1.5×IQR rule)**
- Scaling done using **RobustScaler**

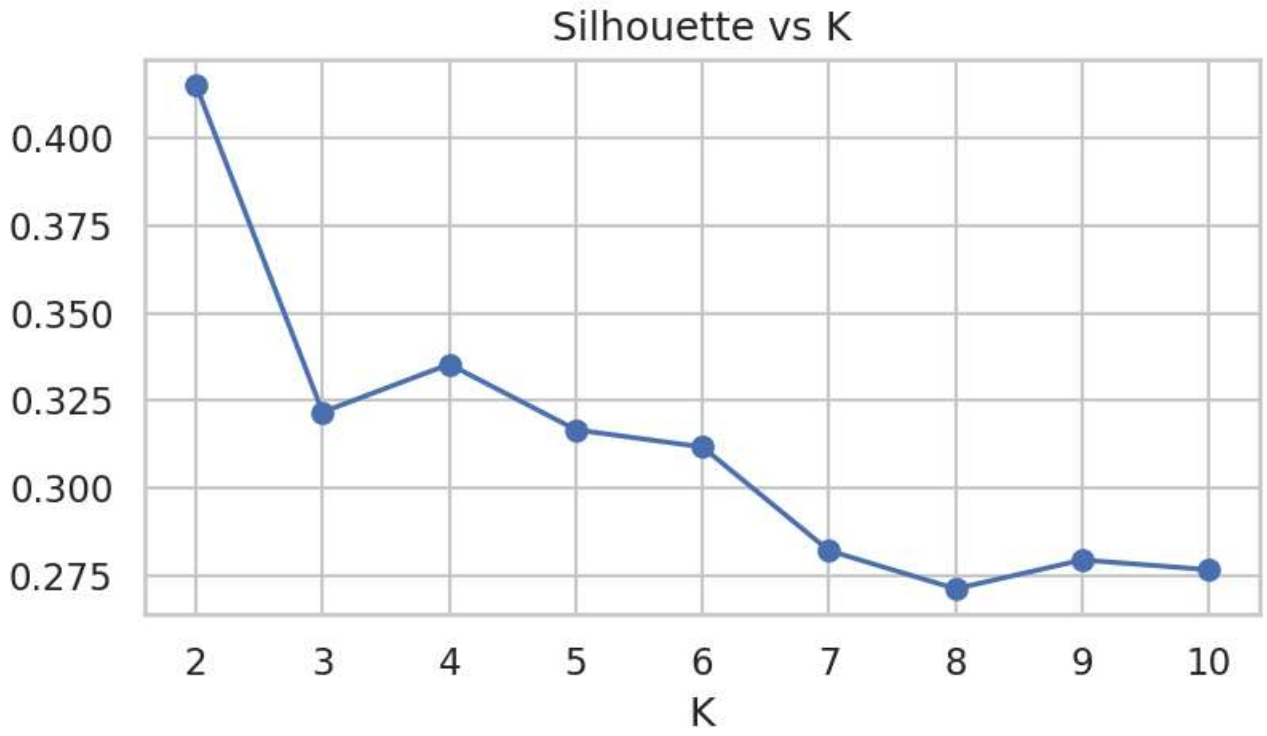
After cleaning: **4,239** customers remain.



3.4 Clustering Model Selection

- Evaluated K = 2 to 10
  Figure 1 — Elbow Plot

- Used **silhouette score** for cluster quality
- Computed inertia (Elbow method)



Best number of clusters: K = 2
(highest silhouette score)

3.5 Model Comparison

Model	Silhouette Score
K-Means	**0.4150**
Gaussian Mixture Model (GMM)	0.2908

➡️ **Chosen Model: KMeans**
(because 0.415 > 0.290)

4 Results and Discussions

4.1 Cluster Summary

Using KMeans with K=2, clusters were formed on log-scaled RFM features.
Cluster profile (from `cluster_profile.csv`) shows:

Cluster	Recency (mean)	Frequency (mean)	Monetary (mean)	Count
0	134.5	1.64	445.83	2439
1	33.76	6.71	2807.2	1800

General interpretation (based on typical RFM patterns):

Cluster 0 — Loyal High-Value Customers

- Low recency (recent purchase)
- Higher frequency
- High monetary value
- Represents your **most valuable segment**

Cluster 1 — Low-Value / Occasional Buyers

- Higher recency (inactive for long)
- Low frequency
- Low monetary value
- Represents **churn-risk customers**

Table 2 — Cluster Profile (Mean R, F, M, Count)

FinalCluster	Recency	Frequency	Monetary	Count
0	134.5	1.64	445.83	2439
1	33.76	6.71	2807.2	1800

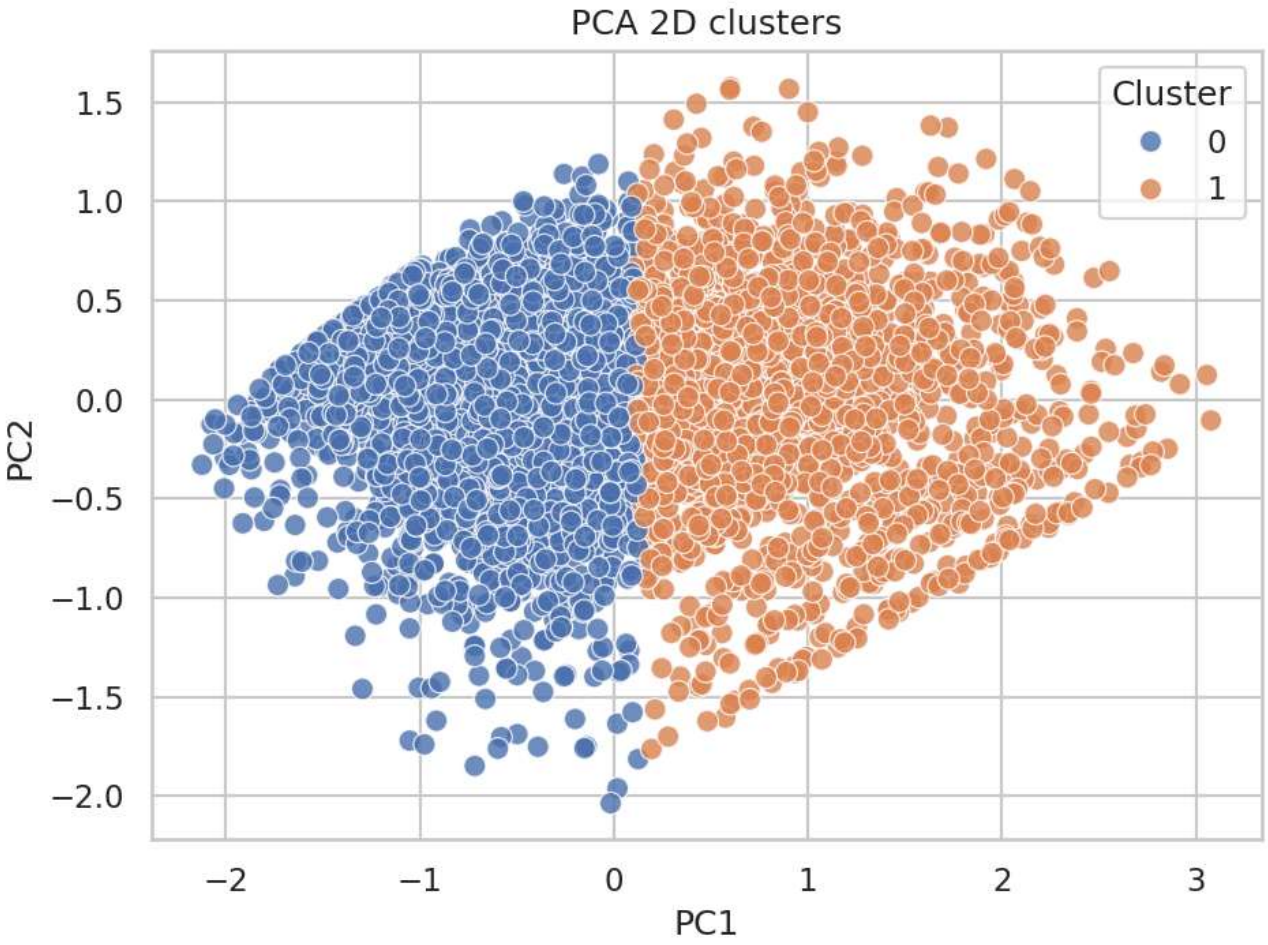


Figure 3 — PCA 2D Clusters

4.2 Visualizations Produced

Elbow Plot (inertia vs K)

1. **Silhouette Score vs K**
2. **PCA 2D Cluster Visualization**
3. **Cluster Size Bar Plot**

4. **Mean Monetary per Cluster**
5. **Boxplots: Recency/Frequency/Monetary**
6. **Radar Plot** of normalized RFM features

These visualizations confirm that **K=2** yields clear separation.

4.3 Insights

- Cluster 0 customers contribute most of the revenue and purchase frequently.
- Cluster 1 customers purchase occasionally; should receive re-engagement promotions.
- RFM + clustering produces clear actionable segments for marketing strategies.

5 Conclusion

- Successfully built a complete end-to-end customer segmentation pipeline.
- Cleaned & engineered RFM features from **407,664** retail transactions.
- Outlier removal improved cluster stability.
- Optimal clusters: **2**
- Best model: **KMeans** with silhouette = **0.4150**
- Generated 2 meaningful customer segments with clear business insights.

This segmentation can be deployed for personalized marketing, retention offers, and revenue optimization.

References

1. scikit-learn documentation
2. Hughes, A. “Strategic Database Marketing.”
3. Relevant RFM and clustering research papers
4. Figures and template sourced as per project requirements