ECOMMERCE CUSTOMER – SEGMENTATION USING RFM ANALYSIS AND MACHINE LEARNING

21CSA697A

Final Report

Submitted by

AMEESHA C

(AA.SC.P2MCA24074012)

in partial fulfilment of the requirements for the award of the degree of

MASTER OF COMPUTER APPLICATIONS

February 2026

Acknowledgement

I would like to express my sincere thanks to my project guide Ms. Jayasree Narayanan for the guidance and support provided throughout this project.

Their suggestions and encouragement were very helpful in completing this work successfully.

I also thank the faculty members of the Department of Computer Applications, Amrita Vishwa Vidyapeetham, for their support during the project.

Abstract

Customer segmentation plays an important role in understanding customer behavior in e-commerce platforms. This project focuses on analyzing e-commerce customer data to identify different customer groups based on their buying patterns and behavior. RFM (Recency, Frequency, and Monetary) analysis is used to represent customer purchase behavior effectively.

To compare different clustering approaches and identify the most suitable method, multiple machine learning clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering are applied. These methods help analyze customer segments from different perspectives, and K-Means clustering is selected as the final method due to better interpretability of results. In addition, a Decision Tree model is used to support the interpretation and prediction of customer segments based on RFM features.

The clustered customer data is visualized using a Tableau dashboard to provide clear business insights. Additionally, a basic Streamlit application is developed to demonstrate the clustered results in an interactive format. This project helps businesses understand customer behavior and supports data-driven decision-making for personalized marketing strategies.

List of Figures

List of Tables

List of Abbreviations

RFM – Recency, Frequency, Monetary

ML – Machine Learning

K-Means – K-Means Clustering Algorithm

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

HC – Hierarchical Clustering

DT – Decision Tree

CHAPTER 1

1. Introduction

With the rapid growth of e-commerce platforms, organizations generate large volumes of customer transaction data on a daily basis. Analyzing this data is essential for understanding customer behavior, improving marketing strategies, and enhancing overall business performance. Customer segmentation plays a vital role in this process by grouping customers based on similar purchasing patterns and engagement levels.

RFM analysis is a commonly used technique for customer segmentation, which evaluates customers based on Recency, Frequency, and Monetary value. When combined with machine learning clustering algorithms, RFM analysis becomes more powerful in identifying meaningful customer groups. This project applies clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering to segment customers effectively.

The project also focuses on visualizing the segmentation results using graphs and dashboards to make insights more interpretable. By leveraging data science and machine learning methods, this work demonstrates how businesses can convert raw customer data into actionable insights for informed decision-making.

1.1 Background

The rapid expansion of e-commerce platforms has led to the generation of large volumes of customer transaction data. Each customer interaction and purchase produces valuable information that can help businesses understand customer behavior. However, raw transactional data does not provide meaningful insights unless it is properly processed and analyzed, increasing the need for data-driven analytical approaches.

Customer segmentation is a key technique used to group customers based on similar purchasing behaviors and characteristics. Effective segmentation enables organizations to design targeted marketing strategies, improve customer retention, and enhance overall business performance. Traditional segmentation methods

often struggle with large and complex datasets, making machine learning techniques more suitable for modern e-commerce environments.

RFM analysis is a widely used method for customer segmentation, evaluating customers based on Recency, Frequency, and Monetary value. In this project, clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering are applied to RFM data to identify meaningful customer groups. Using multiple algorithms allows for better comparison and validation of segmentation results. Visualization tools such as Python, Tableau, and Streamlit are used to present the results clearly, making the insights easier to interpret for business decision-making.

1.2 Problem statement and its significance

The rapid growth of e-commerce platforms has resulted in the collection of large volumes of customer transaction data. However, many organizations face challenges in converting this raw data into meaningful insights for strategic decision-making. Ineffective analysis of customer data can lead to poor understanding of customer behavior, inefficient marketing strategies, and reduced customer retention.

Traditional customer segmentation methods are often inadequate for handling large and complex datasets, as they rely on predefined rules and lack flexibility. This creates a need for intelligent, data-driven techniques that can automatically identify patterns and group customers accurately.

The significance of this project lies in applying RFM analysis combined with machine learning clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering to address these challenges. Using multiple algorithms allows for better comparison and validation of customer segments, while visualization techniques help present the results clearly for improved business decision-making.

1.3 Objectives and scope of the project

Objectives of the Project

The main objectives of this project are:

To analyze e-commerce customer transaction data using RFM analysis.
To segment customers based on Recency, Frequency, and Monetary values.
To apply machine learning clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering for customer segmentation.
To compare the clustering results and identify the most suitable method.
To apply Principal Component Analysis (PCA) for dimensionality reduction and improved visualization of customer segments.
To build a Decision Tree model to predict customer segments based on RFM features.
To visualize customer segments using Tableau and Streamlit dashboards.

Scope of the Project

The scope of this project is limited to analyzing historical e-commerce transaction data for customer segmentation. The project focuses on identifying customer groups based on purchasing behavior using RFM analysis and clustering techniques. The implementation is carried out using Python and visualization tools, and the results are presented through dashboards. The project includes a basic Decision Tree model for customer segment prediction, while advanced predictive modeling and real-time deployment are beyond the scope of this project.

1.4 Organization of the Report

This report is organized into several chapters. Chapter 1 provides an introduction to the project, including background information, problem statement, objectives, and scope. Chapter 2 presents a review of related work and concepts relevant to customer segmentation and clustering techniques. Chapter 3 describes the system design and methodology used in the project. Chapter 4 explains the implementation details and experimental setup. Chapter 5 discusses the results and analysis of the customer segmentation process. Finally, Chapter 6 concludes the project and highlights possible future enhancements.

CHAPTER 2

LITERATURE REVIEW

Customer segmentation has been widely studied in the field of e-commerce analytics as an effective approach to understand customer behavior and improve business decision-making. With the increasing availability of customer transaction data, organizations have shifted from traditional demographic-based segmentation to behavior-based analytical techniques.

Customer segmentation involves grouping customers based on similarities in their purchasing behavior, preferences, and engagement levels. Several studies highlight that behavior-based segmentation helps businesses improve targeted marketing, customer retention, and personalization strategies. Transactional data provides a reliable basis for identifying high-value, loyal, and inactive customers.

RFM analysis is a commonly used technique for modeling customer behavior using three dimensions: Recency, Frequency, and Monetary value. Research indicates that RFM analysis provides a simple and effective way to summarize customer engagement and spending patterns. Due to its interpretability and ease of implementation, RFM analysis has been widely adopted in e-commerce customer segmentation studies.

Machine learning clustering algorithms have been increasingly used to enhance RFM-based segmentation. K-Means clustering is frequently applied due to its efficiency and ability to produce interpretable clusters. DBSCAN is useful for identifying dense customer groups and detecting outliers, while Hierarchical Clustering provides a structured representation of customer

relationships. Studies suggest that comparing multiple clustering techniques helps in validating segmentation results and selecting an appropriate model.

Visualization tools play a crucial role in presenting segmentation results clearly. Tools such as Python-based visualization libraries, Tableau, and interactive dashboards are commonly used to support data-driven decision-making. Combining machine learning techniques with visualization enhances the practical usability of customer segmentation models in real-world e-commerce environments.

CHAPTER 3

SYSTEM DESIGN / ARCHITECTURE

3.1 System Overview

The system is designed to perform customer segmentation on e-commerce transaction data using RFM analysis and machine learning techniques. The workflow begins with data collection and preprocessing, followed by feature engineering using RFM metrics. Machine learning clustering algorithms are then applied to group customers based on purchasing behavior.

To support interpretability and prediction, a Decision Tree model is trained using RFM features and generated cluster labels. Principal Component Analysis (PCA) is applied mainly for dimensionality reduction and visualization of customer clusters. The final results are presented using visual plots, a Tableau dashboard, and a Streamlit application for interactive analysis.

3.2 System Architecture

The system architecture consists of multiple interconnected modules arranged in a sequential flow. The process starts with historical e-commerce transaction data as input. The raw data is cleaned and preprocessed to handle missing values and inconsistencies.

RFM values are computed to represent customer behavior numerically. The normalized RFM data is used as input for clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering to generate customer segments. PCA is then applied to reduce dimensionality for visualization purposes. The clustered dataset is further used to train a Decision Tree model to predict customer segments. Visualization tools are used to present the output in a clear and interpretable format.

3.3 Data Flow Description

The data flow begins with loading raw transaction data into the system. After preprocessing, RFM features are calculated for each customer. Clustering algorithms are applied on normalized RFM data to generate customer segments. PCA is then applied to the clustered data for visualization and interpretation. Cluster labels generated from the clustering process are used to train and test a Decision Tree model using a simple train–test split for basic evaluation and understanding of prediction behavior. The final outputs include customer segments, prediction results, and visual dashboards.

3.4 Modules Description :

3.4.1 Data Preprocessing Module

This module handles data cleaning tasks such as removal of missing values, filtering invalid transactions, and formatting data to ensure consistency and reliability for further analysis.

3.4.2 RFM Analysis Module

This module computes Recency, Frequency, and Monetary values for each customer based on transaction history. These features serve as the primary input for clustering and prediction models.

3.4.3 Clustering Module

This module applies machine learning clustering algorithms including K-Means, DBSCAN, and Hierarchical Clustering to segment customers based on RFM features. Using multiple clustering techniques allows comparison and validation of segmentation results.

3.4.4 PCA Visualization Module

This module applies Principal Component Analysis (PCA) to reduce the dimensionality of RFM data for visualization purposes. PCA is used to project customer data into two dimensions, enabling clear visual representation of customer clusters. It is primarily employed to improve interpretability of clustering results and is not used as a core preprocessing step for clustering.

3.4.5 Decision Tree Prediction Module

This module trains a Decision Tree model using RFM features and cluster labels to predict customer segments. The model provides a simple and interpretable approach to understanding customer behavior.

3.4.6 Visualization Module

This module presents the results of clustering and prediction using visual plots, Tableau dashboards, and a Streamlit application, enabling interactive exploration of customer segments.

3.5 Algorithms Used

The following algorithms and techniques are used in this project:

RFM Analysis

K-Means Clustering

DBSCAN Clustering

Hierarchical Clustering

Principal Component Analysis (PCA)

Decision Tree Classification

3.6 Pseudocode

Pseudocode for Customer Segmentation System

1. Load e-commerce transaction data

2. Clean and preprocess the dataset

3. Compute RFM values for each customer

4. Normalize RFM features

5. Apply clustering algorithms on RFM data

6. Assign cluster labels to customers

7. Apply PCA for dimensionality reduction and visualization

8. Train Decision Tree model using RFM features and cluster labels

9. Predict customer segments

10. Visualize results using plots and dashboards

CHAPTER 4

IMPLEMENTATION DETAILS

4.1 Introduction

This chapter explains how the proposed customer segmentation system was implemented in practice. It describes the steps followed to preprocess data, compute RFM values, apply clustering algorithms, perform visualization using PCA, and build a basic prediction model using a Decision Tree. The implementation was carried out using Python in a Jupyter Notebook environment, and the results were visualized using plots and dashboards.

4.2 Tools and Technologies Used

The following tools and technologies were used for implementing the project:

Programming Language: Python

Development Environment: Jupyter Notebook

Libraries: Pandas, NumPy, Scikit-learn

Visualization Libraries: Matplotlib, Seaborn

Dashboard Tools: Tableau and Streamlit

These tools were selected due to their simplicity, efficiency, and suitability for data analysis and visualization tasks.

4.3 Data Preprocessing Implementation

The implementation begins with preprocessing the raw e-commerce transaction dataset. The dataset is examined for missing values, invalid records, and inconsistencies. Transactions with missing customer identifiers and invalid purchase quantities are removed. Date-related fields are converted into appropriate formats to enable time-based calculations.

Only the necessary attributes required for customer analysis are retained. This preprocessing step ensures that the dataset is clean and reliable for further analysis.

4.4 RFM Analysis Implementation

RFM analysis is implemented to represent customer purchasing behavior using three key metrics:

Recency: The number of days since the customer's most recent purchase

Frequency: The total number of purchases made by the customer

Monetary Value: The total amount spent by the customer

These RFM values are calculated for each customer using transaction history. To ensure equal contribution of all features, the RFM values are normalized using standard scaling techniques. The normalized RFM dataset serves as the primary input for clustering and prediction models.

4.5 Clustering Implementation

Clustering algorithms are applied to the normalized RFM data to segment customers into distinct groups based on purchasing behavior. The following clustering techniques are implemented:

K-Means Clustering

DBSCAN

Hierarchical Clustering

K-Means clustering is used as the primary method due to its simplicity and interpretability. DBSCAN helps in identifying dense customer groups and detecting outliers, while Hierarchical Clustering provides insights into relationships between customer segments. Using multiple clustering techniques allows comparison and validation of segmentation results. The final cluster labels are assigned to each customer.

4.6 PCA Visualization Implementation

Principal Component Analysis (PCA) is applied after clustering to reduce the dimensionality of RFM data for visualization purposes. PCA projects the high-dimensional RFM data into two principal components, enabling clear two-dimensional visualization of customer clusters.

PCA is used primarily to improve the interpretability of clustering results and is not used as a core preprocessing step for clustering.

4.7 Decision Tree Prediction Implementation

A Decision Tree model is implemented to support basic prediction and interpretability. The model is trained using RFM features as input and the cluster labels generated from K-Means clustering as the target variable.

The Decision Tree learns simple decision rules that explain how customers are assigned to different segments based on their purchasing behavior. This model provides an easy-to-understand approach for predicting customer segments.

4.8 Visualization and Dashboard Implementation

Visualization is used to present the results in a clear and meaningful way. Cluster distributions, RFM comparisons, and PCA scatter plots are generated using Python visualization libraries.

In addition, dashboards are created using Tableau to present cluster-wise customer insights, and a Streamlit application is developed to provide an interactive interface for exploring clustered customer data. These dashboards help convert analytical results into business-friendly insights.

4.9 Pseudocode of the System

The overall implementation process of the system is summarized below:

Load e-commerce transaction data

Clean and preprocess the dataset

Compute RFM values for each customer

Normalize RFM features

Apply clustering algorithms on normalized RFM data

Assign cluster labels to customers

Apply PCA for visualization of customer clusters

Train a Decision Tree model using RFM features and cluster labels

Predict customer segments

Visualize results using plots and dashboards

CHAPTER 5

TESTING, VALIDATION AND RESULTS

5.1 Introduction

This chapter presents the testing process, validation approach, and results obtained from the customer segmentation system. Since the project mainly focuses on unsupervised learning techniques, validation is performed using visual analysis and logical interpretation of clusters. In addition, a basic Decision Tree model is evaluated using a train–test split to observe prediction behavior.

5.2 Testing Methodology

The system was tested using historical e-commerce transaction data. After data preprocessing and RFM value computation, clustering algorithms were applied to segment customers based on purchasing behavior. The generated cluster labels were then used to train a basic Decision Tree model.

For the Decision Tree model, the dataset was split into training and testing sets to perform basic evaluation.

Testing focused on:

Correct execution of each module

Proper flow of data between modules

Logical formation of customer segments

Clear visualization of results

All modules were tested successfully, and the system produced consistent outputs.

5.3 Validation of Clustering Results

K-Means, DBSCAN, and Hierarchical Clustering are unsupervised learning techniques, so traditional accuracy metrics are not directly applicable. Therefore, clustering results were validated using the following methods:

Visual inspection of clusters using PCA-based scatter plots

Analysis of RFM value differences across clusters

Comparison of customer distribution among clusters

The clusters formed showed clear differences in customer purchasing behavior. Using multiple clustering algorithms helped in validating the segmentation results and improved confidence in the identified customer groups.

Bar charts were used to analyze the number of customers in each cluster formed by K-Means, DBSCAN, and Hierarchical Clustering. The results showed that K-Means produced clearly separated clusters with meaningful customer distribution, while DBSCAN identified a small number of outliers. Hierarchical clustering also provided a similar grouping pattern, supporting the consistency of clustering results.

5.4 Decision Tree Model Evaluation

A Decision Tree model was trained using RFM features as input and K-Means cluster labels as the target output. The dataset was divided into training and testing sets using a train–test split approach.

The model evaluation included accuracy score and classification report to understand prediction performance. The objective of this evaluation was not to achieve high predictive accuracy, but to verify whether the model could learn cluster patterns and provide interpretable decision rules based on RFM values.

5.5 Results and Observations

The main results obtained from the system are summarized below:

• Customers were successfully segmented into meaningful groups based on purchasing behavior.

• K-Means clustering produced clearly interpretable customer segments.

• DBSCAN identified dense customer groups and detected outliers.

• Hierarchical Clustering provided additional insights into customer relationships.

• PCA visualization clearly showed separation between customer clusters.

•Scatter plots such as Monetary Value vs Purchase Frequency helped in understanding spending behavior across different customer groups.

•Cluster-wise bar charts clearly showed differences in average Recency, Frequency, and Monetary values, confirming meaningful customer segmentation.

• Bar charts and scatter plots helped analyze customer distribution and RFM patterns.

• Tableau dashboards presented cluster-wise summaries of Recency, Frequency, and Monetary values.

• The Streamlit application displayed clustered customer data in an interactive tabular and visual format.

• The Decision Tree model supported customer segment prediction in an interpretable manner.

Overall, the system effectively transformed raw transaction data into useful customer insights.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This project implemented customer segmentation using RFM (Recency, Frequency, and Monetary) analysis on e-commerce transaction data. After cleaning and preprocessing the dataset, RFM values were calculated to represent customer purchasing behavior.

Clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering were applied to group customers into meaningful segments. K-Means produced clearly interpretable clusters, DBSCAN helped identify dense groups and outliers, and Hierarchical Clustering supported understanding relationships between customer groups.

The clustering results were validated using visual analysis. Bar charts, scatter plots, and PCA-based visualizations showed clear differences between customer segments. Tableau dashboards provided cluster-wise summaries of RFM values, and a Streamlit application displayed the clustered customer data in an interactive format.

A Decision Tree model was also used to support customer segment prediction based on RFM features, helping in better interpretation of customer behavior.

Overall, the project successfully converted raw transaction data into useful customer insights using clustering and visualization techniques.

6.2 Future Work

The project can be extended in the future by including additional customer attributes to improve segmentation. More advanced models and evaluation methods can be explored. The Streamlit application can be enhanced with more interactive features, and the approach can be applied to larger or real-time datasets.

CHAPTER 7

REFERENCES

1.Kaggle, Online Retail Dataset

[Online]. Available: https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset

2.Python Software Foundation, Python Documentation.

[Online]. Available: https://www.python.org/

3.Scikit-learn Developers, Scikit-learn Documentation.

[Online]. Available: https://scikit-learn.org/

4.Jupyter Project, Jupyter Notebook Documentation.

[Online]. Available: https://jupyter.org/

5.Tableau Software, Tableau Desktop User Guide.

[Online]. Available: https://www.tableau.com/support/help

6.Streamlit Inc., Streamlit Documentation.

[Online]. Available: https://streamlit.io/

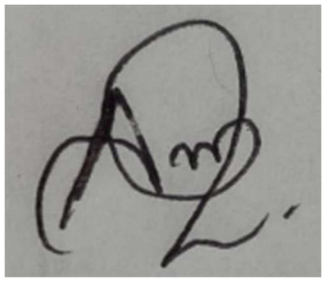
CHAPTER 8

APPENDIX

GitHub Repository Link:

https: https://github.com/Ameeshac2003/ecommerce-customer-segmentation-rfm-ml

The repository contains the complete implementation of the project, including Jupyter Notebook files for data preprocessing, RFM analysis, clustering techniques, visualizations, and model development. The dataset and related supporting files are also included for reference.

Date	**02/02/2026**
Student Name and Signature	AMEESHA C 

Name and Signature of the Evaluator.

Date