Customer Segmentation using Clustering Algorithms

ON

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

(Information Technology)

By

Komal Sabale(45)

Krushikesh Shelar (51)

Shweta Wadhwa (58)

Under the guidance of

Dr. Ravita Mishra

Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

V.E.S.
Since 1962

Vivekanand Education Society's
Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Certificate

This is to certify that project entitled

"Customer Segmentation using Clustering Algorithms"

Group Members Names

Komal Sabale(Roll No. 45)
Krushikesh Shelar(Roll No. 51)
Shweta Wadhwa (Roll No. 58)

In fulfillment of degree of BE. (Sem. VI) in Information Technology for Project is approved.

Dr. Ravita Mishra
Project Mentor

External Examiner

Dr.(Mrs.)Shalu Chopra
H.O.D

Dr.(Mrs.) J.M.Nair
Principal

Date: / /2025
Place: VESIT, Chembur

College Seal

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Komal Sabale(Rollno. 45) **(Signature)** - - - - - - - - - - -

Krushikesh Shelar(Rollno. 51) **(Signature)** - - - - - - - - - - -

Shweta Wadhwa (Rollno. 58) **(Signature)** - - - - - - - - - - -

Contents

ACKNOWLEDGEMENT

The project report on “Customer Segmentation using Clustering Algorithms” is the outcome of the guidance, moral support and devotion bestowed on our group throughout our work. For this we acknowledge and express our profound sense of gratitude to everybody who has been the source of inspiration throughout project preparation. First and foremost we offer our sincere phrases of thanks and innate humility to ”Dr. Shalu Chopra, HOD”, ”Dr. Manoj Sabnis,Deputy HOD”, ”Dr. Ravita Mishra” for providing the valuable inputs and the consistent guidance and support provided by them. We can say in words that we must at the outset tender our intimacy for receipt of affectionate care to Vivekanand Education Society’s Institute of Technology for providing such a stimulating atmosphere and conducive work environment.

Chapter 1: Introduction

1.1 Introduction

Customer segmentation helps businesses categorize customers based on purchasing behavior. It improves marketing, loyalty, and retention strategies. This project uses unsupervised machine learning with RFM features (Recency, Frequency, Monetary) to group customers into meaningful segments.

1.2 Objectives

- To preprocess and extract RFM features from transaction data
- To apply clustering models like KMeans, DBSCAN, Agglomerative, and GMM
- To evaluate clustering performance using Silhouette Score
- To assign business-friendly labels to clusters
- To deploy an interactive segmentation tool using Streamlit

1.3 Motivation

Manual segmentation is time-consuming and inflexible. Machine learning can identify hidden customer patterns quickly and accurately. This project builds a dynamic tool for real-time behavioral segmentation.

1.4 Scope of the Work

- Focus on unsupervised clustering with RFM features
- Dataset from UCI (Online Retail Dataset)
- Visual insights using Seaborn and Plotly
- Streamlit app for live exploration and download of results

1.5 Feasibility Study

- Technical Feasibility: Uses Python, Scikit-learn, and Streamlit—open-source and reliable.
- Operational Feasibility: Lightweight and easy to integrate into business workflows.
- Economic Feasibility: Budget-friendly due to public tools and datasets.

Chapter 2: Literature Survey

2.1 Introduction

Research in customer segmentation using machine learning has advanced rapidly. This section compares key models and approaches used for grouping customers based on behavior. Recent studies have explored the use of RFM features and various clustering techniques—each offering unique advantages depending on the dataset and business goal.

2.2 Problem Definition

Customer data is often large and diverse, with no predefined labels. The challenge lies in identifying distinct customer segments from behavioral data in an unsupervised way. The model must effectively group customers into meaningful clusters while remaining interpretable and practical for business use.

2.3 Review of Literature

1. **Oluwasurefunmi Idowu et al. (2019) – "Customer Segmentation Based on RFM Model Using K-Means, Hierarchical and Fuzzy C-Means Clustering Algorithms"**

   ○ Problem Statement:
   This study focuses on segmenting e-commerce customers by analyzing their purchasing behaviors using the RFM (Recency, Frequency, Monetary) model. The goal is to identify distinct customer groups to enhance targeted marketing strategies.

   ○ Models Used and Evaluation Metrics:
      1. K-Means Clustering – Formed 5 clusters; evaluated using Silhouette Width and Dunn Index.
      2. Fuzzy C-Means Clustering – Also resulted in 5 clusters; assessed with similar metrics.
      3. Hierarchical Clustering – Produced 2 clusters; achieved a Dunn Index of 1.58, indicating superior performance among the three.

   ○ Conclusion:
   The study concludes that while all three clustering methods are effective, Hierarchical Clustering outperformed the others in terms of cluster validity. This suggests its suitability for customer segmentation tasks in e-commerce settings.

   ○ Reference:
   ResearchGate

2. **Reyhan Muhammad Fauzan & Ganjar Alfian (2024) – "Customer Segmentation Using RFM Model and K-Means Clustering"**

   - Problem Statement:
     This research aims to develop a web-based application for e-commerce customer segmentation by integrating the RFM model with clustering algorithms. The objective is to assist businesses in identifying customer segments to tailor marketing efforts effectively.

   - Models Used and Evaluation Metrics:
     1. K-Means Clustering – Achieved a Silhouette Score of 0.67305, Davies-Bouldin Index of 0.51435, and Calinski-Harabasz Index of 5647.89.
     2. K-Medoids Clustering – Compared for performance but yielded lower evaluation scores.
     3. Fuzzy C-Means Clustering – Also evaluated; however, K-Means outperformed the other methods.
   - Conclusion:
     The study demonstrates that K-Means Clustering, when combined with the RFM model, effectively segments customers into categories like Loyal, Need Attention, and Promising. The developed Streamlit-based application provides a practical tool for businesses to analyze and act upon customer segmentation insights.
   - Reference:
     ResearchGate

.

Chapter 3: Design and Implementation

3.1 System Architecture

The system takes a customer transaction dataset as input, processes it into RFM features, and applies clustering algorithms to segment customers. Each model's output is evaluated and visualized. An interactive Streamlit app allows business users to explore results and download clustered data.

3.2 Data Preprocessing

The dataset was cleaned by removing null CustomerIDs, negative quantities, and cancelled transactions. New features like TotalBill were created. RFM metrics were calculated per customer:

- Recency: Days since last purchase
- Frequency: Number of unique transactions
- Monetary: Total spend

3.3 Feature Scaling

RFM values were standardized using StandardScaler to ensure equal contribution to clustering models.

3.4 Clustering Models

- KMeans: Applied with k = 3–5. Chosen for its simplicity and speed.
- DBSCAN: Density-based clustering. Tuned with eps and min_samples.
- Agglomerative Clustering: Hierarchical model with Ward linkage.
- GMM (Gaussian Mixture Model): Soft clustering model based on probability distributions.

3.5 Evaluation

Clusters were evaluated using the Silhouette Score. KMeans and Agglomerative showed the best performance (0.58–0.61). DBSCAN was useful for detecting outliers but required careful parameter tuning.

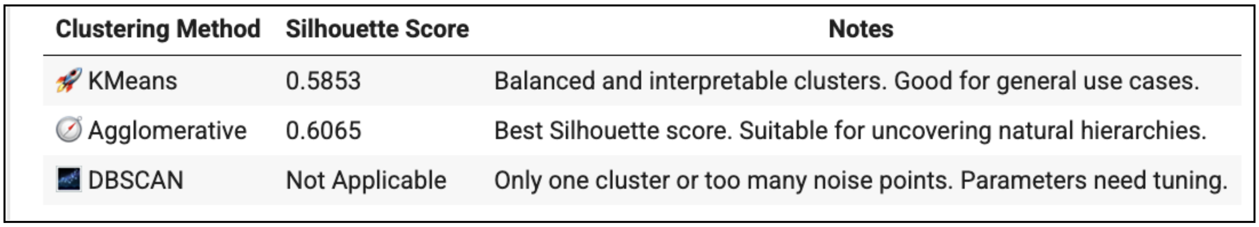
3.6 Streamlit App

A web app was built using Streamlit to upload data, choose models, view clustering results, and download the output. It helps non-technical users run and interpret the segmentation pipeline interactively.

Chapter 4: Results and Discussion

4.1 Cluster Evaluation

Different clustering models were applied on the scaled RFM features. KMeans with k=3–5 showed strong separation. Agglomerative Clustering also performed well, with the highest Silhouette Score of 0.6065. DBSCAN detected some outliers but formed only one main cluster due to parameter sensitivity.



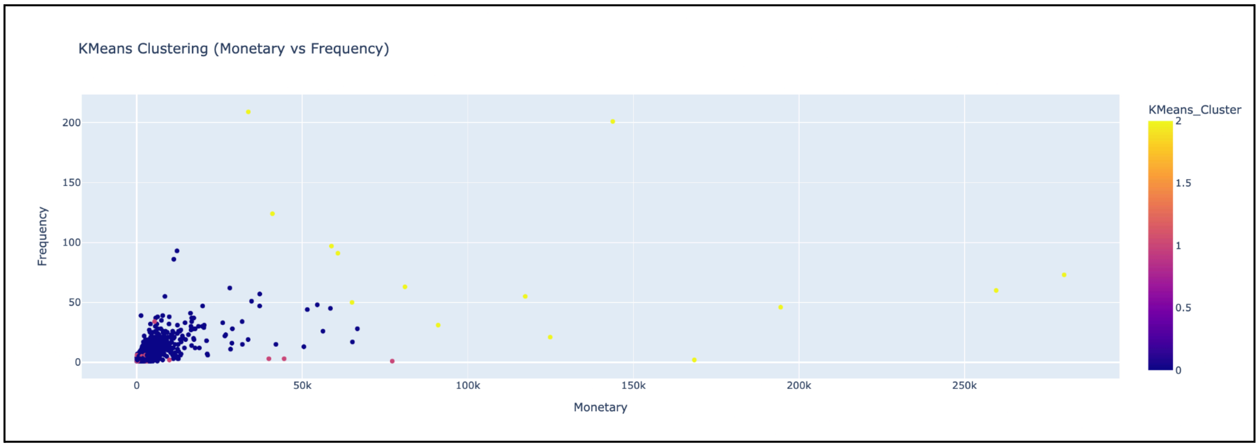
Clustering Method	Silhouette Score	Notes
🚀 KMeans	0.5853	Balanced and interpretable clusters. Good for general use cases.
🧭 Agglomerative	0.6065	Best Silhouette score. Suitable for uncovering natural hierarchies.
🌌 DBSCAN	Not Applicable	Only one cluster or too many noise points. Parameters need tuning.

4.2 KMeans Clustering

KMeans grouped customers into three primary clusters. The segments were labeled as:

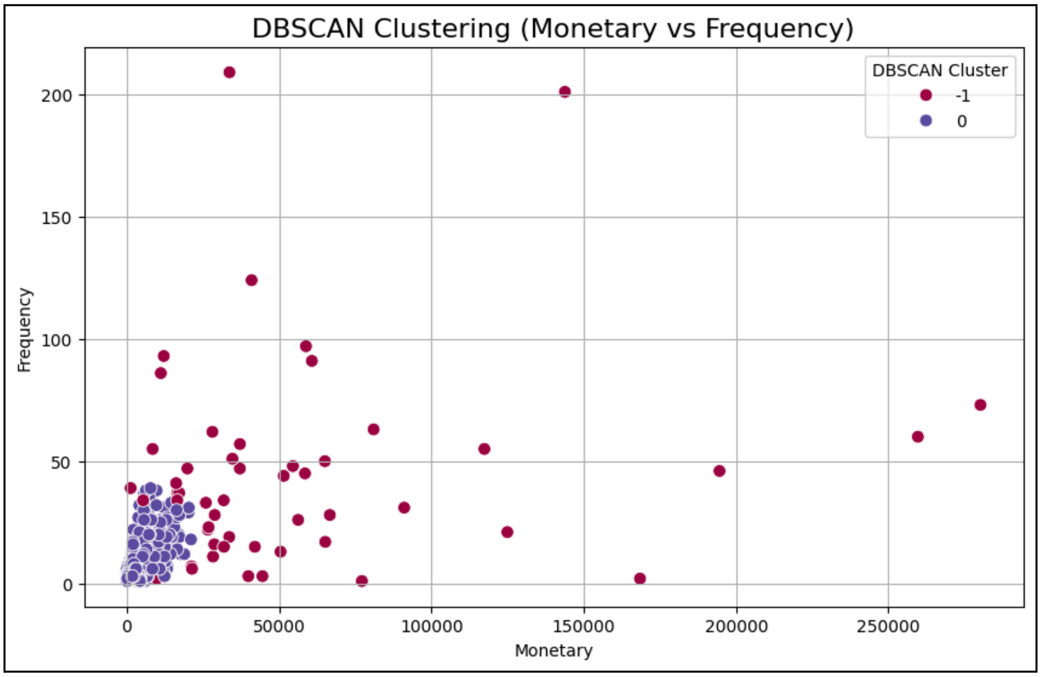
- Cluster 0: Loyal Regulars
- Cluster 1: Dormant/At-Risk
- Cluster 2: High-value VIPs

These labels were based on average Recency, Frequency, and Monetary values.



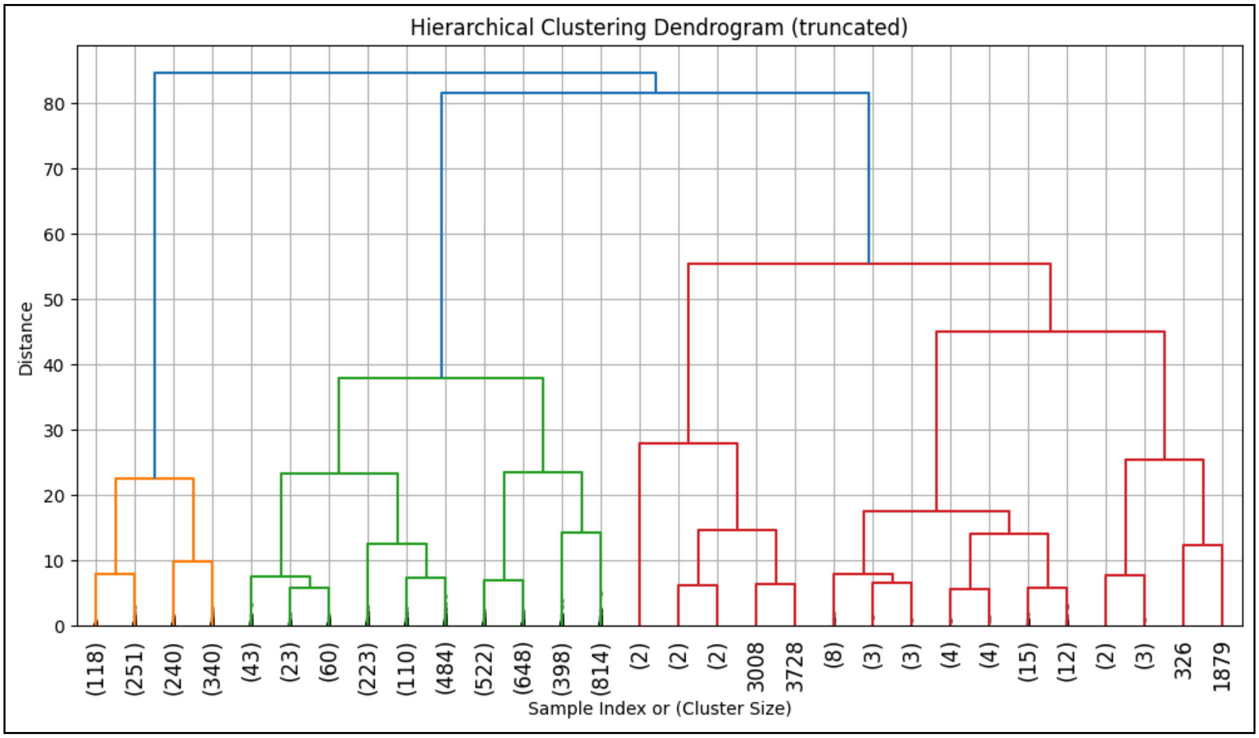
4.3 DBSCAN Clustering

DBSCAN was effective in identifying noise (-1) but struggled with sparse RFM features. It showed fewer distinct clusters under default parameters.



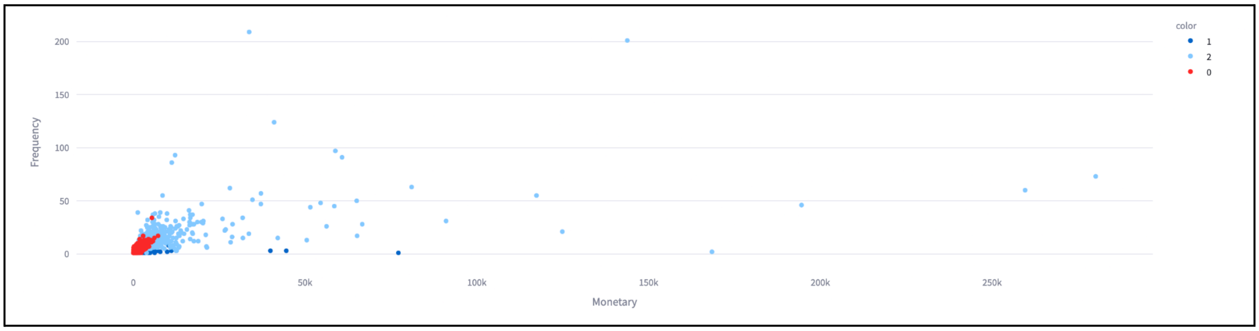
4.4 Agglomerative Clustering

Agglomerative Clustering revealed clear hierarchy among customers. The dendrogram helped identify the natural number of clusters (suggesting 4). Final cluster summaries showed tight behavioral groupings.



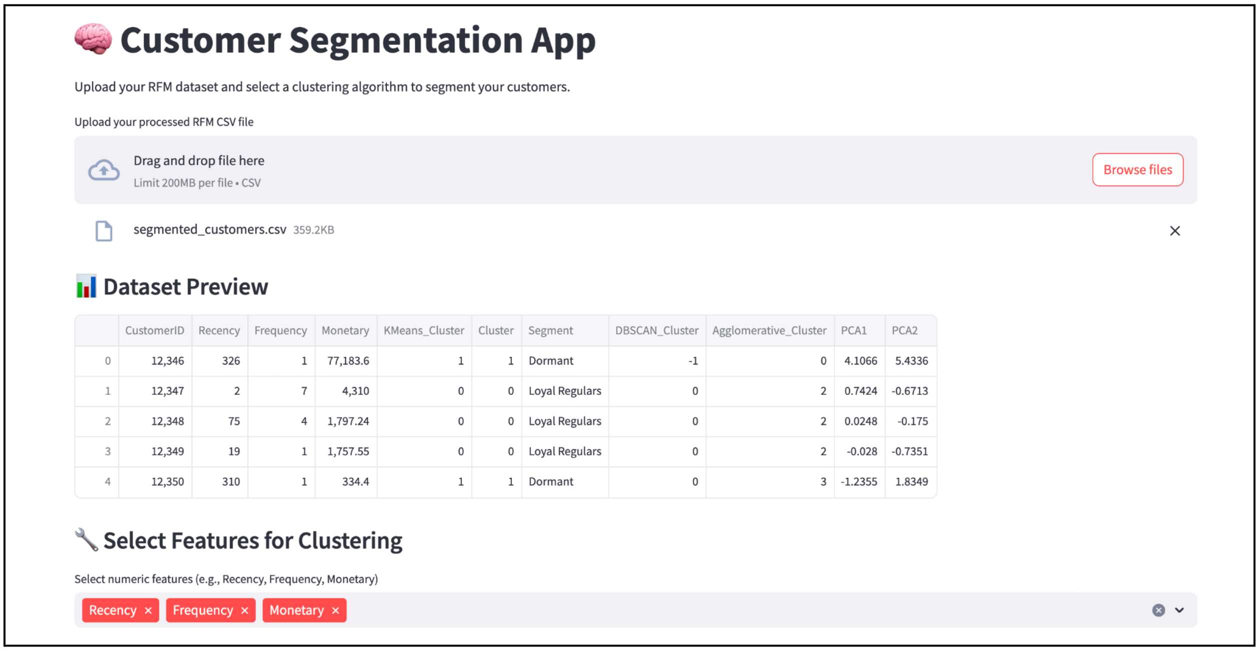
4.5 Gaussian Mixture Model (GMM)

GMM provided a probabilistic approach to clustering, where customers were grouped based on likelihood of belonging to a distribution. It produced soft cluster boundaries and worked well on scaled RFM features. GMM detected high-value clusters similar to KMeans, but with more overlap between boundaries.



4.6 Streamlit App Output

The Streamlit app allowed interactive selection of algorithms, visualization of clusters, and downloading results. It simplified deployment for marketing teams.



🧠 Customer Segmentation App

Upload your RFM dataset and select a clustering algorithm to segment your customers.

Upload your processed RFM CSV file

segmented_customers.csv 359.2KB

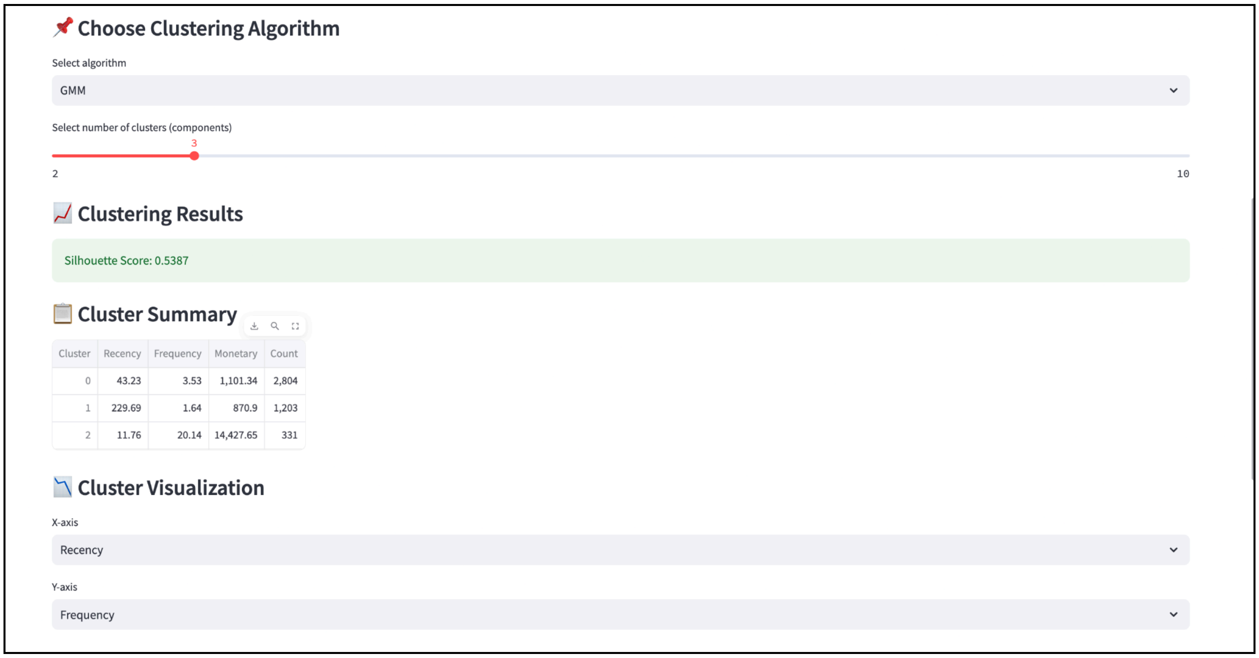
📊 Dataset Preview

	CustomerID	Recency	Frequency	Monetary	KMeans_Cluster	Cluster	Segment	DBSCAN_Cluster	Agglomerative_Cluster	PCA1	PCA2
0	12,346	326	1	77,183.6	1	1	Dormant	-1	0	4.1066	5.4336
1	12,347	2	7	4,310	0	0	Loyal Regulars	0	2	0.7424	-0.6713
2	12,348	75	4	1,797.24	0	0	Loyal Regulars	0	2	0.0248	-0.175
3	12,349	19	1	1,757.55	0	0	Loyal Regulars	0	2	-0.028	-0.7351
4	12,350	310	1	334.4	1	1	Dormant	0	3	-1.2355	1.8349

🔧 Select Features for Clustering

Select numeric features (e.g., Recency, Frequency, Monetary)

Recency × Frequency × Monetary ×



📌 Choose Clustering Algorithm

Select algorithm

GMM

Select number of clusters (components)

3

📈 Clustering Results

Silhouette Score: 0.5387

📋 Cluster Summary

Cluster	Recency	Frequency	Monetary	Count
0	43.23	3.53	1,101.34	2,804
1	229.69	1.64	870.9	1,203
2	11.76	20.14	14,427.65	331

📉 Cluster Visualization

X-axis

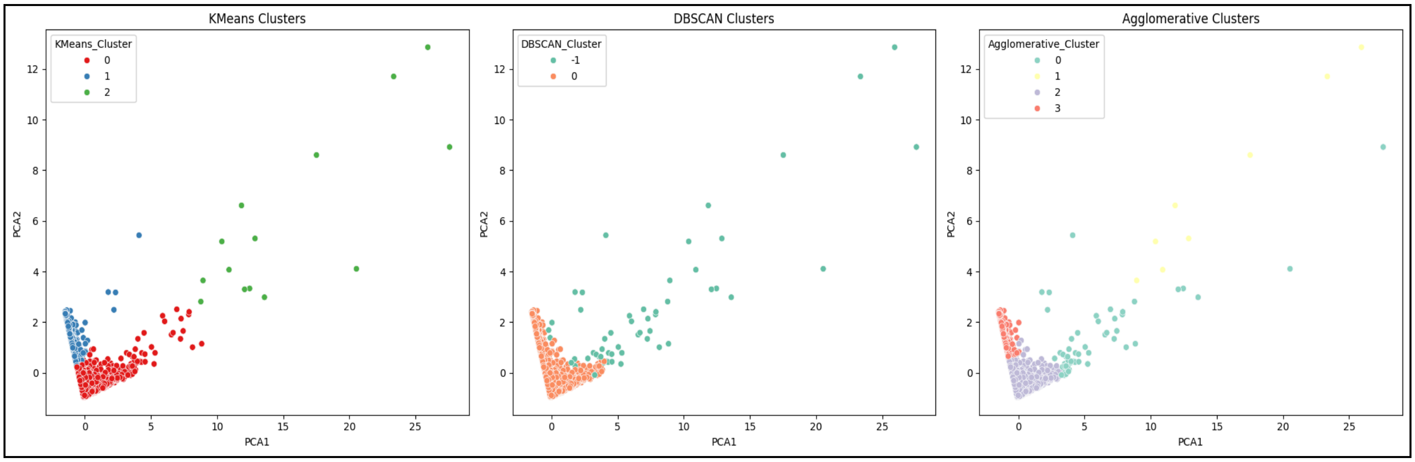
Recency

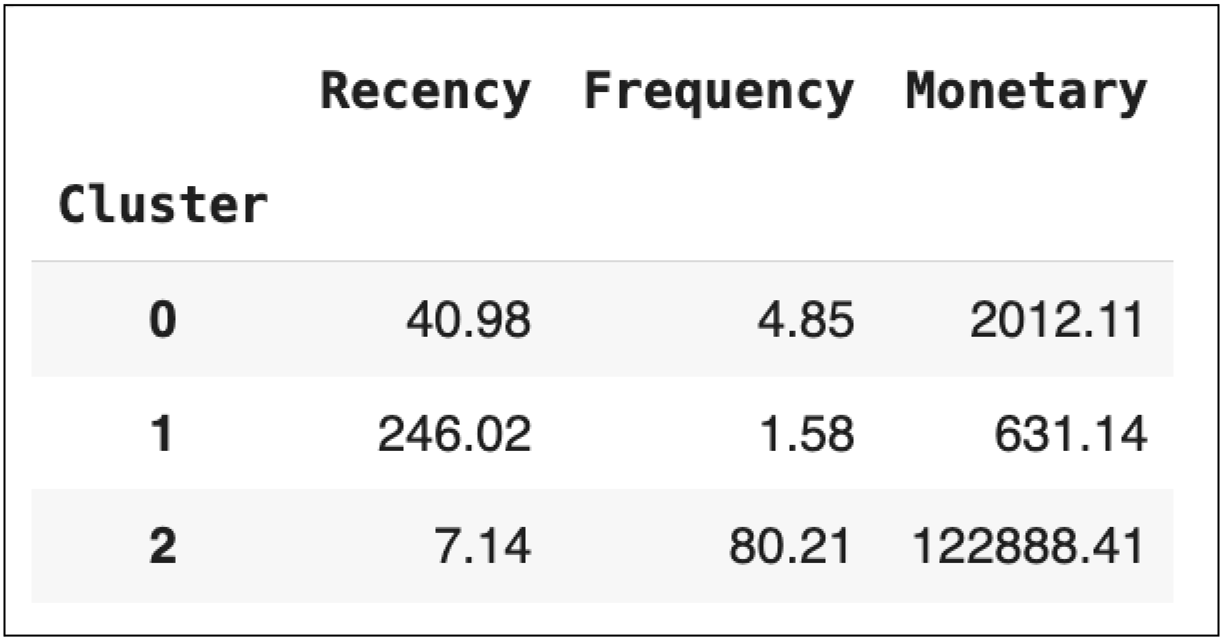
Y-axis

Frequency

4.7 Final Summary Table

The final clusters were summarized with average Recency, Frequency, and Monetary values, along with customer count





Cluster	Recency	Frequency	Monetary
0	40.98	4.85	2012.11
1	246.02	1.58	631.14
2	7.14	80.21	122888.41

Chapter 5: Conclusion and Future Scope

5.1 Conclusion

This project demonstrates the use of unsupervised learning for customer segmentation using RFM features. By applying models like KMeans, DBSCAN, Agglomerative Clustering, and GMM, we identified distinct customer groups for targeted strategies. The Streamlit app provides a simple interface to explore and apply these insights, making the system practical and business-ready.

5.2 Future Scope

- Supports personalized marketing and retention strategies: Enables businesses to target customer groups with customized offers, improving satisfaction and conversion.
- Can be extended with demographic or location-based data: Enhances segmentation depth by integrating data such as age, region, or product category.
- Enables real-time segmentation on live data: Can be integrated with dashboards or CRM tools for dynamic targeting.
- Can be scaled across industries: Applicable to retail, banking, hospitality, or e-commerce with minimal adjustments.

5.3 Societal Impact

- Reduces spam and irrelevant marketing: Personalized communication respects consumer attention and privacy.
- Supports small businesses with accessible insights: Open-source tools make customer intelligence achievable for startups and SMEs.
- Promotes ethical data use: Encourages transparency and fairness in data-driven decision-making.

5.4 Bibliography

[1] O. Idowu, E. Olaniyi, and O. Oyelade, "Customer Segmentation Based on RFM Model Using K-Means, Hierarchical and Fuzzy C-Means Clustering Algorithms," ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/342571209_Customer_Segmentation_Based_on_RFM_Model_Using_K-Means_Hierarchical_and_Fuzzy_C-_Means_Clustering_Algorithms

[2] R. M. Fauzan and G. Alfian, "Customer Segmentation Using RFM Model and K-Means Clustering," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/352393770_Customer_Segmentation_using_RFM_Model_and_K-Means_Clustering