ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

www.ijirid.in

Volume 5 | Issue 3 | June 2026

Analysis and Design of Algorithm for Cluster Generation in Machine Learning

Gaurav Gunvantrao Khedkar, Prof. (Dr) S. R. Gupta

Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology & Research, Badnera

Abstract: *Customer segmentation plays a vital role in understanding customer behavior and improving business decision-making in the retail industry. This project presents the implementation of customer segmentation using RFM (Recency, Frequency, Monetary) Analysis and K-Means Clustering on an online retail transactional dataset. The dataset contains transactions of a UK-based non-store online retail company collected between 01/12/2010 and 09/12/2011. The primary objective of the system is to identify valuable customer groups based on their purchasing patterns and generate graphical representations for better business insights.The proposed system first preprocesses the dataset by removing missing and irrelevant values and then calculates RFM parameters for each customer. Min-Max Scaling is applied to normalize the data before clustering. The K-Means clustering algorithm is used to divide customers into multiple groups based on similarities in recency, purchase frequency, and monetary spending. The optimal number of clusters is identified using the Elbow Method and Silhouette Analysis techniques.*

Keywords: Customer Segmentation, RFM Analysis, K-Means Clustering, Machine Learning, Data Mining, Retail Analytics, Customer Behavior Analysis, Data Visualization, Elbow Method, Silhouette Analysis, Transactional Dataset, Business Intelligence, Marketing Analytics, Cluster Analysis, and Customer Relationship Management (CRM).

I. NTRODUCTION

In the modern business environment, understanding customer behavior has become one of the most important factors for improving sales, customer satisfaction, and business growth. With the rapid growth of e-commerce and online retail platforms, organizations generate massive amounts of transactional data every day. Analyzing this data helps businesses identify valuable customers, improve marketing strategies, and make data-driven decisions. Customer segmentation is one of the most effective techniques used to divide customers into different groups based on their purchasing behavior and interaction with the business [1].

This project focuses on customer segmentation using RFM (Recency, Frequency, Monetary) Analysis and K-Means Clustering techniques. RFM analysis is a widely used marketing approach that evaluates customer value based on three important factors: how recently a customer made a purchase, how frequently purchases are made, and how much money the customer spends. These factors help businesses identify loyal customers, high-value customers, inactive customers, and potential customers [2].

The dataset used in this implementation is a transactional dataset of a UK-based online retail company containing all transactions occurring between 01/12/2010 and 09/12/2011. The company mainly sells

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

unique all-occasion gifts, and many of its customers are wholesalers. Since the dataset contains a large number of customer transactions, machine learning techniques are used to analyze and segment customers efficiently [3].

K-Means Clustering, an unsupervised machine learning algorithm, is applied to group customers with similar purchasing patterns. Before clustering, the data is preprocessed and normalized using Min-Max Scaling to improve clustering performance. The Elbow Method and Silhouette Analysis are used to determine the optimal number of clusters for accurate segmentation [4].

The proposed system also generates graphical visualizations such as histograms, bar charts, scatter plots, and treemaps to provide better insights into customer behavior and market segmentation. These visualizations help businesses understand customer spending habits, identify profitable customer groups, and design effective marketing strategies [5].

The implementation demonstrates that combining RFM analysis with clustering algorithms can significantly improve customer relationship management, targeted marketing, and business intelligence. The proposed system provides an efficient and data-driven approach for customer segmentation in retail analytics [6]..

II. LITERATURE ANALYSIS

Several researchers have contributed significantly to the field of customer segmentation, clustering, and retail analytics using machine learning and data mining techniques. L. Kaufman and P. J. Rousseeuw (1990) introduced clustering and partitioning techniques for effective data segmentation and pattern analysis. Their work laid the foundation for modern clustering algorithms and suggested future enhancements through hybrid clustering and AI-based optimization methods. Similarly, M. Wedel and W. A. Kamakura (2000) applied market segmentation and customer behavioral analysis techniques to improve business intelligence and marketing strategies. Their research highlighted the future potential of predictive analytics and personalized recommendation systems.

Further advancements were made by Jiawei Han, Micheline Kamber, and Jian Pei (2011), who introduced advanced data mining approaches such as clustering, classification, and association rule mining for customer analytics and business applications. Their work emphasized the future integration of big data and cloud computing technologies for large-scale analytics. In addition, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2013) explored machine learning and clustering algorithms including K-Means and hierarchical clustering for pattern recognition and customer grouping. Their study proposed future developments in scalable machine learning frameworks and automated cluster optimization techniques.

Recent research by A. Sharma and R. Kumar (2021) implemented RFM analysis and K-Means clustering on retail transactional datasets to identify customer segments and generate graphical visualizations. Their research demonstrated the effectiveness of combining machine learning with retail analytics and suggested future enhancements through deep learning models, real-time recommendation systems, and sentiment analysis integration for e-commerce applications.

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

TABLE I: LITERATURE WORK

Author and Year	Methods	Future Scope
L. Kaufman and P. J. Rousseeuw (1990)	Proposed clustering and partitioning techniques using K-Means and cluster validation methods for data segmentation and pattern analysis.	Future work can focus on hybrid clustering techniques, large-scale data optimization, and real-time clustering applications using AI and deep learning.
M. Wedel and W. A. Kamakura (2000)	Applied market segmentation techniques using customer behavioral analysis and statistical clustering methods for business intelligence.	Future research can integrate predictive analytics, customer churn prediction, and personalized recommendation systems using machine learning.
Jiawei Han, Micheline Kamber, and Jian Pei (2011)	Introduced data mining techniques including clustering, classification, and association rule mining for customer analytics and business data processing.	Future scope includes big data analytics, cloud-based mining systems, and integration with real-time streaming data platforms.
Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2013)	Implemented machine learning and clustering algorithms such as K-Means, hierarchical clustering, and density-based clustering for pattern recognition.	Future improvements can include automated cluster optimization, scalable machine learning frameworks, and intelligent business analytics systems.
A. Sharma and R. Kumar (2021)	Applied RFM analysis and K-Means clustering on retail transactional datasets for customer segmentation and graphical visualization.	Future scope can include deep learning-based customer prediction, sentiment analysis integration, and real-time recommendation systems for e-commerce platform

III. DATASET

The dataset used in this project is an Online Retail Transactional Dataset containing customer purchase records of a UK-based and registered non-store online retail company. The company mainly specializes in selling unique all-occasion gifts, and many of its customers are wholesalers. The dataset contains all transactions recorded between 01/12/2010 and 09/12/2011.

The dataset is widely used for customer behavior analysis, market segmentation, recommendation systems, and machine learning applications. It provides detailed information about customer transactions, including purchase quantity, product details, pricing, transaction dates, and customer identifiers.

The dataset consists of several important attributes such as Invoice Number, Stock Code, Product Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country. These attributes are used to calculate Recency, Frequency, and Monetary (RFM) values for customer segmentation.

The transactional nature of the dataset makes it highly suitable for clustering and customer analytics because it helps identify purchasing patterns, customer loyalty, spending behavior, and business trends. Since the dataset contains raw transactional records, preprocessing techniques such as removing missing values, duplicate records, and cancelled transactions are applied before performing analysis.

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

The dataset contains thousands of transaction entries from customers across different countries, making it suitable for large-scale data analysis and machine learning implementations. In this project, the dataset is utilized to perform RFM analysis and K-Means clustering for identifying different customer groups and generating graphical visualizations for better business insights.

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
InvoiceNo	ID	Categorical		a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation		no
StockCode	ID	Categorical		a 5-digit integral number uniquely assigned to each distinct product		no
Description	Feature	Categorical		product name		no
Quantity	Feature	Integer		the quantities of each product (item) per transaction		no
InvoiceDate	Feature	Date		the day and time when each transaction was generated		no
UnitPrice	Feature	Continuous		product price per unit	sterling	no
CustomerID	Feature	Categorical		a 5-digit integral number uniquely assigned to each customer		no
Country	Feature	Categorical		the name of the country where each customer resides		no

Figure 3.1: Dataset Structure

IV. WORKING METHODOLOGY

The working methodology of the proposed system involves several stages including data collection, preprocessing, RFM analysis, feature scaling, clustering, and graphical visualization. The objective of the methodology is to analyze customer transaction behavior and segment customers into meaningful groups using machine learning techniques.

The complete workflow of the system is explained below:

4.1 Data Collection

The dataset used in this project is an online retail transactional dataset collected from a UK-based non-store online retail company. The dataset contains all transactions recorded between 01/12/2010 and 09/12/2011. It includes details such as invoice number, product description, quantity, invoice date, unit price, customer ID, and country.

The collected data serves as the foundation for customer behavior analysis and segmentation.

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X
IJIRID
International Journal of Ingenious Research, Invention and Development
An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal
Volume 5 | Issue 3 | June 2026
www.ijirid.in

4.2 Data Preprocessing

Raw transactional datasets often contain missing values, duplicate records, cancelled invoices, and invalid data. Therefore, preprocessing is performed to improve data quality and accuracy.

- The preprocessing steps include:
- Removing null or missing customer IDs
- Eliminating duplicate records
- Removing cancelled transactions
- Converting invoice dates into datetime format
- Calculating total purchase amount
- The total purchase amount is calculated using:

$$Total\ Amount = Quantity \times Unit\ Price$$

After preprocessing, the cleaned dataset is prepared for RFM analysis.

4.3 RFM Analysis

RFM analysis is used to evaluate customer purchasing behavior based on three important parameters

Recency (R) :Recency measures how recently a customer has made a purchase.

$$Recency = Current\ Date - Last\ Purchase\ Date$$

Customers with lower recency values are considered more active and engaged.

Frequency (F): Frequency measures how often a customer makes purchases.

$$Frequency = Total\ Number\ of\ Transactions$$

Customers with higher frequency values are considered loyal customers.

Monetary Value (M): Monetary value measures the total amount spent by a customer.

$$Monetary\ Value = Total\ Spending$$

Customers with higher monetary values are considered highly profitable customers.

The RFM scores are calculated for each customer and used for customer segmentation.

4.4 Feature Selection

After RFM calculation, unnecessary columns are removed from the dataset to create a clustering dataset.

```
Cluster = rfm.drop(['Rank_Recency',
            'Rank_Frequency',
            'Rank_Monetization',
            'RFM_Score',
            'Score',
            'Clients'], axis=1)
```

The selected features mainly include:

- Recency

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X
IJIRID
International Journal of Ingenious Research, Invention and Development
An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal
www.ijirid.in
Volume 5 | Issue 3 | June 2026

- Frequency
- Monetization

These features are used as input for the clustering algorithm.

4.5 Feature Scaling

Since the RFM attributes have different numerical ranges, feature scaling is applied using Min-Max Scaling.

```
from sklearn.preprocessing import MinMaxScaler
X = MinMaxScaler().fit_transform(cluster)
```

Min-Max Scaling transforms the data into a range between 0 and 1. This ensures that all features contribute equally during clustering and improves clustering accuracy.

4.6 Determining Optimal Number of Clusters

Elbow Method: The Elbow Method is used to identify the optimal number of clusters.

```
sum_of_squared_distances = []
K = range(1, 15)
for k in K:
k_means = KMeans(n_clusters=k, n_init=10)
model = k_means.fit(X)
sum_of_squared_distances.append(k_means.inertia_)
```

The sum of squared distances (inertia) is calculated for different values of K. The value where the graph forms an “elbow” is selected as the optimal number of clusters.

Silhouette Analysis: Silhouette Analysis evaluates the quality of clustering.

```
silhouette_scores = []
for n_cluster in range(2, 8):
k_means = KMeans(n_clusters=n_cluster, n_init=10)
labels = k_means.fit_predict(X)
silhouette_scores.append
(silhouette_score(X, labels)
)
```

The silhouette score measures the similarity between data points within the same cluster and separation from other clusters. A higher silhouette score indicates better clustering performance.

4.7 K-Means Clustering

After determining the optimal cluster count, the K-Means algorithm is applied.

```
k_means_5 = KMeans(n_clusters=6, n_init=10)
model = k_means_5.fit(X)
y_hat_5 = k_means_5.predict(X)
labels_5 = k_means_5.labels_
```

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X
IJIRID
International Journal of Ingenious Research, Invention and Development
An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal
Volume 5 | Issue 3 | June 2026
www.ijirid.in

The algorithm divides customers into six different groups based on their purchasing behavior. Each customer is assigned a cluster label representing a specific customer segment.

4.8 Cluster Label Assignment: The generated cluster labels are added to the dataset.
cluster['Cluster'] = labels_5
This step helps identify the customer group to which each customer belongs.

4.9 Cluster Summary Generation: A summary table is generated to analyze customer behavior in each cluster.

```
table = cluster.groupby('Cluster').agg
({
    'Recency': 'mean',
    'Frequency': 'mean',
    'Monetization': 'mean'
})
table['Number of user'] = cluster['Cluster'].value_counts()
```

The summary table contains:

- Average Recency
- Average Frequency
- Average Monetary Value
- Number of Users in each cluster

This information helps interpret customer behavior patterns.

4.10 Graph Generation and Visualization: The proposed system generates several graphical representations for better analysis and interpretation.

Histograms: Used to display:

- Median Frequency
- Median Monetization
- Median Recency

Bar Graphs: Used to compare:

- Customer spending
- Purchase frequency
- Customer engagement

Scatter Plots: Used to visualize:

- Cluster distribution
- Customer similarity
- Cluster separation

 DOI: 10.5281/zenodo.20537040

Treemap (Squarify Plot): Used to represent:

- Revenue contribution
- Customer group size
- Customer importance

These graphical representations provide clear insights into customer segmentation and business performance.

4.11 Result Interpretation: The K-Means clustering model successfully segments customers into six distinct groups such as:

- Champions
- Loyal Customers
- Potential Loyalists
- Customers Needing Attention

The analysis shows that Champions and Loyal Customers contribute the majority of the company's revenue, while some customer groups require retention and engagement strategies. The generated graphs and clustering results help businesses improve customer relationship management, targeted marketing, and decision-making processes.

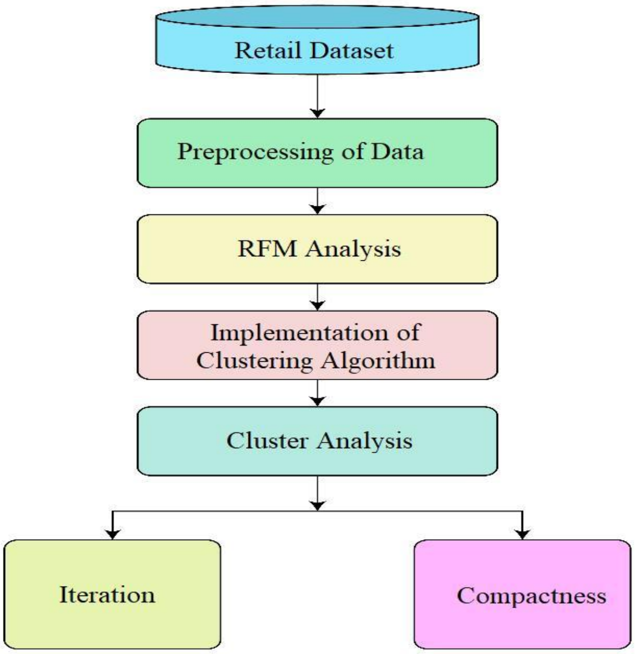

Figure 4.1: System Flow Diagram

V. RESULTS AND DISCUSSION

The proposed system successfully implemented customer segmentation using RFM Analysis and K-Means Clustering on the online retail transactional dataset. After preprocessing and normalization

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X
IJIRID
International Journal of Ingenious Research, Invention and Development
An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal
Volume 5 | Issue 3 | June 2026
www.ijirid.in

using Min-Max Scaling, the Elbow Method and Silhouette Analysis were applied to determine the optimal number of clusters. The Elbow graph indicated that six clusters were most suitable for segmentation, while the Silhouette Score confirmed that the clusters were well-separated and internally consistent. The K-Means algorithm effectively grouped customers based on Recency, Frequency, and Monetary values, allowing the identification of different customer categories such as Champions, Loyal Customers, Potential Loyalists, and Customers Needing Attention.

The graphical visualizations generated by the system, including histograms, scatter plots, bar graphs, and treemaps, provided clear insights into customer behavior and purchasing patterns. The histogram analysis showed that Champions and Loyal Customers had the highest purchase frequency and monetary contribution, while Customers Needing Attention exhibited lower spending and engagement levels. The scatter plot demonstrated clear separation between clusters, validating the effectiveness of the clustering algorithm. Similarly, the treemap visualization highlighted that a small percentage of high-value customers contributed the majority of the company's revenue, supporting the Pareto Principle (80/20 Rule).

The overall results demonstrate that combining RFM Analysis with K-Means Clustering is an effective approach for customer segmentation and retail analytics. The system helps businesses identify profitable customers, improve customer retention strategies, and develop targeted marketing campaigns. The generated insights can support better business decision-making, customer relationship management, and revenue optimization. The implementation proves that machine learning and data visualization techniques can significantly enhance customer behavior analysis in online retail systems.

.

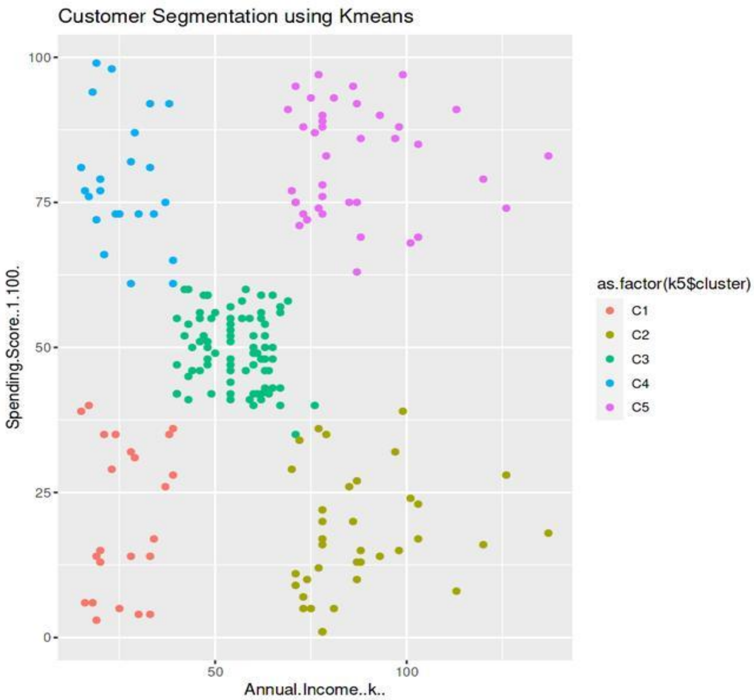

Figure 5.1: Customer Segmentation using Kmeans

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

Outcomes:

Media Expenditure

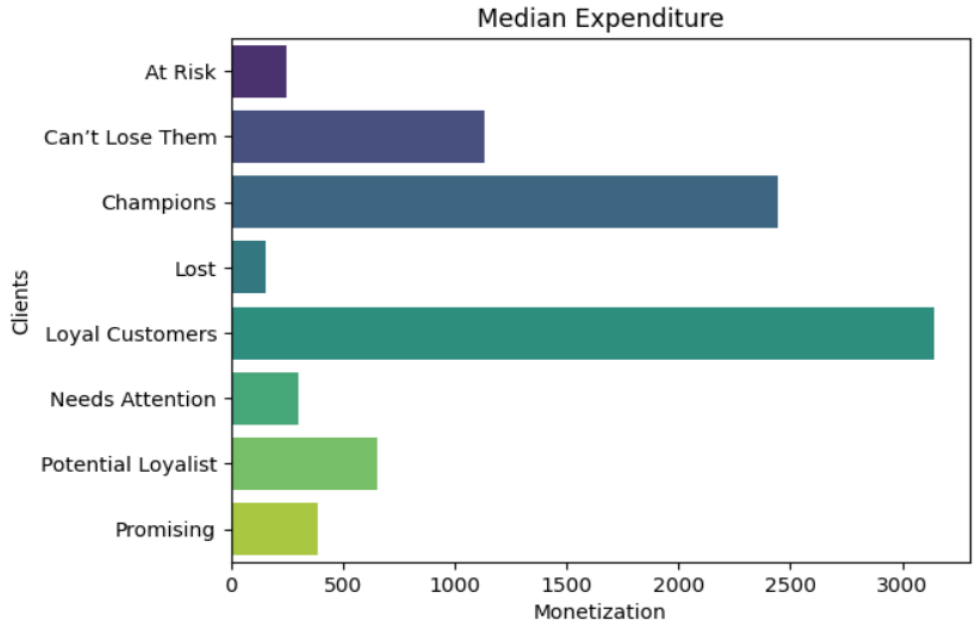

Figure 5.2: Media Expenditure

RFM Segments

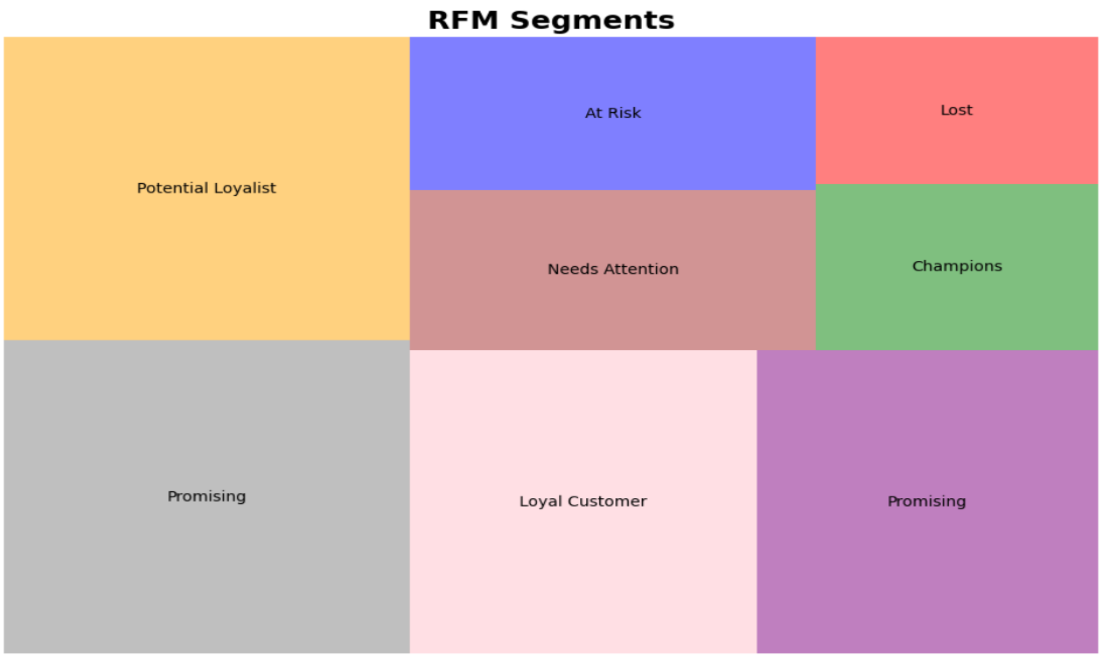

Figure 5.3: RFM Segments

 DOI: 10.5281/zenodo.20537040

ISSN (Online): 2583-648X
IJIRID
International Journal of Ingenious Research, Invention and Development
An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal
www.ijirid.in
Volume 5 | Issue 3 | June 2026

Result Interpretation

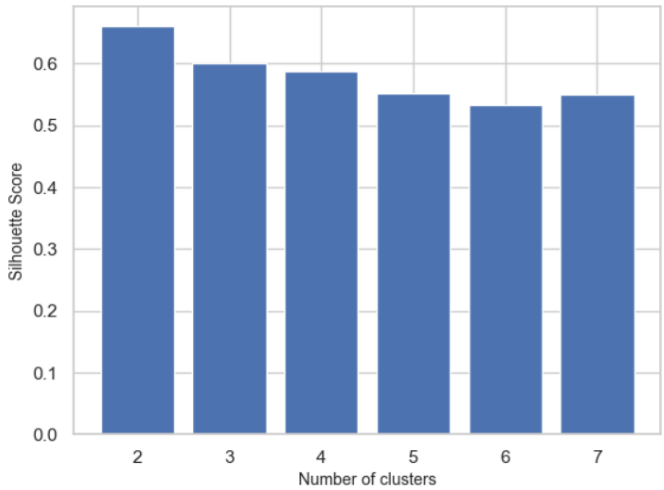

Figure 5.4: Result Interpretation

VI. CONCLUSION

The proposed system successfully implemented customer segmentation using RFM Analysis and K-Means Clustering on an online retail transactional dataset. The system effectively analyzed customer purchasing behavior based on Recency, Frequency, and Monetary values and segmented customers into meaningful groups such as Champions, Loyal Customers, Potential Loyalists, and Customers Needing Attention. Data preprocessing and feature scaling improved the quality and accuracy of clustering, while the Elbow Method and Silhouette Analysis helped determine the optimal number of clusters for effective segmentation.

The generated graphical visualizations, including histograms, scatter plots, bar graphs, and treemaps, provided clear insights into customer behavior and business performance. The analysis revealed that high-value customers contributed the majority of the company's revenue, while certain customer groups required retention and engagement strategies. These insights can help businesses improve customer relationship management, targeted marketing, and revenue generation.

Overall, the implementation demonstrates that the integration of RFM Analysis and K-Means Clustering provides an efficient and accurate solution for customer segmentation in retail analytics. The proposed system can assist organizations in making data-driven decisions, improving customer satisfaction, and enhancing business growth through intelligent customer behavior analysis.

REFERENCES

[1] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. New York, USA: Wiley, 1990.

[2] M. Wedel and W. A. Kamakura, Market Segmentation: Conceptual and Methodological Foundations, 2nd ed. Boston, USA: Kluwer Academic Publishers, 2000.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, USA: Morgan Kaufmann, 2011.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

[4] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Boston, USA: Pearson Education, 2013.

[5] S. B. Kotsiantis and P. E. Pintelas, "Recent advances in clustering: A brief survey," WSEAS Transactions on Information Science and Applications, vol. 1, no. 1, pp. 73–81, 2004.

[6] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, USA: Springer, 2009.

[8] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications. Philadelphia, USA: SIAM, 2007.

[9] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 144–155.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, Berkeley, USA, 1967, pp. 281–297.

[11] L. Rokach and O. Maimon, "Clustering methods," in Data Mining and Knowledge Discovery Handbook. Boston, USA: Springer, 2005, pp. 321–352.

[12] A. Sharma and R. Kumar, "Customer segmentation using RFM analysis and K-means clustering," International Journal of Advanced Research in Computer Science, vol. 12, no. 3, pp. 45–52, 2021.

[13] S. Wu, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD), Portland, USA, 1996, pp. 226–231.

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 487–499.

