Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Customer Segmentation And Behaviour Analysis Using RFM And K-Means Clustering

K.C. Chandra Sekaran

Associate Professor, Department of Computer Science, Alagappa Govt. Arts College, Karaikudi, Sivagangai (Dt),Tamilnadu, India.
Mail: chandrasekarankc@agacollege.in

This study utilizes RFM (Recency, Frequency, Monetary) analysis and K-Means clustering techniques to explore customer behavior within a mall customer dataset. By leveraging attributes such as age, gender, annual income, and spending score, RFM scores are computed to quantify customer engagement levels. These scores are then used to categorize customers into distinct segments: 'Top Customers', 'High Value Customers', 'Mid Value Customers', and 'Low Value Customers'. Additionally, K-Means clustering is applied to identify natural groupings within the dataset based on RFM dimensions. Evaluation metrics like Silhouette Score, Davies-Bouldin Index, and Inertia assist in determining the optimal number of clusters. The findings provide actionable insights for targeted marketing strategies and customer relationship management in retail settings.

Keywords: Customer segmentation, Silhouette Score, RFM, K-means clustering.

I.INTRODUCTION

Understanding customer behaviour is crucial for businesses aiming to optimize marketing strategies and enhance customer satisfaction. One effective approach to achieve this understanding is through RFM (Recency, Frequency, Monetary) analysis, which segments customers based on their transactional history. This method categorizes customers according to three key metrics: how recently a customer has made a purchase (Recency), how often they make purchases (Frequency), and how much they spend (Monetary). In addition to RFM analysis, clustering techniques such as K-Means [1] provide further insights by grouping customers based on similarities in their RFM profiles. By identifying distinct customer segments, businesses can tailor their marketing efforts, product offerings, and customer service strategies to better meet the specific needs and preferences of each segment.

This research work focuses on applying RFM analysis and K-Means [2] clustering to a dataset of mall customers. The dataset includes demographic information such as age, gender, and annual income, as well as behavioural data such as spending scores. By computing RFM scores for each customer and clustering them based on these scores, this study aims to uncover meaningful customer segments within the mall customer base. The analysis includes

evaluating clustering performance using metrics like Silhouette Score and Davies-Bouldin Index to determine the optimal number of clusters. The insights derived from this analysis can guide mall managers and marketers in devising targeted strategies to enhance customer engagement, loyalty, and satisfaction.

Through this exploration of customer segmentation and behaviour analysis, this study contributes to the broader understanding of how data-driven approaches can inform business decisions and improve customer-centric outcomes in retail environments. This work is chaptered as chapter 1 introduction, chapter 2 related work, chapter 3 methodology, chapter 4 result and discussion and chapter 5 conclusion.

II.RELATED WORK

RFM analysis has been widely adopted across various industries due to its simplicity and effectiveness in segmentation. Research by Wan et al. [3] underscores the integration of RFM metrics into CRM systems, enhancing customer retention strategies by identifying high-value segments for targeted marketing initiatives. Raine et al. [4] further emphasize the utility of RFM analysis in optimizing resource allocation, highlighting its role in driving personalized customer interactions and improving operational efficiencies. By ranking customers based on Recency, Frequency, and Monetary values and normalizing these rankings to derive an aggregated RFM score, businesses gain actionable insights into customer behavior that inform strategic decision-making.

Clustering techniques, particularly K-Means clustering, complement RFM analysis by grouping customers into distinct clusters based on their RFM scores. Tabianan et al., [5] discuss the application of K-Means clustering in data mining for customer segmentation, illustrating its ability to identify homogeneous groups within large datasets. In retail and marketing, RFM clustering has proven instrumental in optimizing marketing campaigns, store layouts, and inventory management strategies. For instance, Liu et al. [6] apply CRM and techniques in e-commerce settings to personalize promotional offers based on customer purchase behaviours. Chen et al., used a modified RFM to understand the customer behaviour mainly they used customer index and periodicity to segment the customer.

The RFM model has evolved over the last 20 years by adding more factors, including time since first purchase [7], churn likelihood [8], product category group information [8], and customer relationship duration. RFM models are extensively used in customer analysis and segmentation in numerous industries due to their quick implementation time and ability to capture consumer attributes efficiently. The customer quintile differentiation approach [9] divides consumers into five equal quintiles based on RFM factors. This method can be used to evaluate customers or actual values. Depending on the features of the industry, the weighting of the model variables may be equal or different [10].

This survey highlights the foundational principles, methodologies, and applications of RFM analysis and clustering techniques in customer segmentation and behaviour analysis. It underscores the importance of leveraging data-driven approaches to enhance customer

engagement, loyalty, and business profitability in diverse industry contexts. Future research directions should focus on addressing scalability challenges, integrating advanced machine learning techniques, and adapting segmentation models to evolving consumer behaviours and market dynamics.

III. Methodology

1. Data Collection and Preparation

Data Collection: The dataset used in this study is derived from a mall customer dataset from kaggele[11], which includes attributes such as customer ID, gender, age, annual income, and spending score. The dataset was loaded from a CSV file and basic statistics and structure were analysed to understand the data.

Data Preparation: Missing values can skew analysis [12] results and must be addressed. In this dataset, missing values were handled by using an appropriate imputation method (e.g., mean, median, or mode imputation) or by removing rows/columns with a significant amount of missing data. Ensuring that data types are consistent across variables is crucial for accurate analysis. Data types were corrected where necessary to match the expected types for each attribute. Data inconsistencies such as outliers, duplicate entries, and incorrect values were identified and resolved through appropriate cleansing techniques. This step ensures the integrity and quality of the data. Before clustering, it is important to standardize the data to ensure that all features contribute equally to the analysis. Standardization was performed using the StandardScaler from the sklearn library.

2. RFM Calculation

Recency, Frequency, Monetary (RFM) [13] analysis is a marketing technique that classifies and measures the worth of customers according to their purchase patterns. It assists companies in determining which clients are the most valuable and informing marketing strategy accordingly. Thorough breakdown of each element and the methodology is used to calculate it in this work.

Recency (R):

Regency measures how recently a customer has made a purchase. Customers who have purchased more recently are considered more likely to buy again compared to those who have not purchased in a while. In this study, the "Spending Score (1-100)" was used as a proxy for Regency, assuming that a higher score indicates more recent engagement with the mall.

Frequency (F):

Frequency measures how often a customer makes a purchase within a given time period. Frequent buyers are usually more engaged and loyal. Due to the lack of transactional frequency data in this dataset, a default frequency value of 1 was assigned to each customer, indicating that each customer has made at least one purchase.

Monetary Value (M):

Monetary value measures how much money a customer spends on purchases. Higher spending customers are generally more valuable to a business. In this study, "Annual Income (in $k)" was used as a proxy for the Monetary component, assuming that customers with higher incomes are likely to spend more.

RFM Score Assignment: The RFM score is a composite score derived from the Recency, Frequency, and Monetary values. The steps to calculate the RFM scores for each customer in this study are as follows:

Score Assignment:

- Each RFM dimension (R, F, M) was assigned scores based on quartiles:
    - Recency Score: Higher scores indicate more recent transactions.
    - Frequency Score: Higher scores indicate more frequent transactions.
    - Monetary Value Score: Higher scores indicate higher spending.

RFM Score Calculation:

- Combined RFM scores were calculated using the formula [14] RFM_Factor = Rnew+F+M.

Clustering Algorithm: K-means clustering algorithm was employed to group customers based on RFM scores and additional demographic or behavioral features.

Evaluation Metrics:

- **Silhouette Score:** Measures cluster cohesion and separation.
- **Davies-Bouldin Index:** Evaluates cluster compactness.
- **Inertia:** Sum of squared distances from each point to its assigned cluster centroid.

3. Cluster Analysis and Interpretation

Cluster Profiles:

- Average RFM scores and relevant attributes were analyzed for each cluster.

- Customer characteristics, preferences, and behaviors were identified within each cluster.

4. Marketing Strategies

Segment-Specific Strategies:

- **Top Customers:** Offer exclusive rewards and personalized experiences.
- **High Value Customers:** Focus on upselling and cross-selling premium products.
- **Mid Value Customers:** Implement targeted promotions and loyalty programs.
- **Low Value Customers:** Launch reactivation campaigns and offer discounts.

5. Implementation and Evaluation

Implementation:

- Marketing strategies were implemented based on identified customer segments and RFM profiles.

Evaluation:

- Continuous monitoring of strategy effectiveness using customer feedback and behavior analytics.

IV Result and discussion

This section presents the findings from the analysis of customer data using RFM (Regency, Frequency, Monetary) segmentation and clustering techniques. The dataset, consisting of 200 customers, was thoroughly examined to derive valuable insights into customer behaviours and characteristics. Initial descriptive statistics provided a comprehensive overview of the dataset's composition. Then the RFM score is discussed, followed by cluster analysis and finally cluster profile.

Table 1 presents the descriptive statistics for the dataset comprising 200 customers. CustomerID ranges from 1 to 200, with a mean of 100.5, indicating a balanced dataset distribution. Age has an average of 38.85 years, with a standard deviation of 13.97, suggesting a diverse age range (min: 18, max: 70). The median age is 36, with 25th and 75th percentiles at 28.75 and 49, respectively. Annual Income averages $60,560, with a standard deviation of 26.26, indicating varying income levels among customers (min: $15,000, max: $137,000). The median income is 61.5, with 25th and 75th percentiles at 41.5 and 78, respectively. Spending Score averages 50.2, with a standard deviation of 25.82, reflecting moderate customer engagement. The scores range from 1 to 99, with a median score of 50 and 25th and 75th percentiles at 34.75 and 73, respectively.

Table 1: Descriptive statistics from dataset

Statistic	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
Count	200	200	200	200
Mean	100.5	38.85	60.56	50.2
Standard Deviation (std)	57.88	13.97	26.26	25.82
Minimum (min)	1	18	15	1
25th Percentile (25%)	50.75	28.75	41.5	34.75
Median (50%)	100.5	36	61.5	50
75th Percentile (75%)	150.25	49	78	73
Maximum (max)	200	70	137	99

Table 2: RFM Scores: Example of the first five customers' RFM scores

CustomerID	RFM_score	RFM_segment
1	95.71	Top Customers
2	87.22	Top Customers
3	98.75	Top Customers

4	87.06	Top Customers
5	94.35	Top Customers

Table 2 presents the RFM scores for the first five customers in the dataset, highlighting their segmentation based on purchasing behavior.

- All five customers have high RFM scores, with values ranging from 87.06 to 98.75.
- Each of these customers falls into the Top Customers segment, indicating strong engagement and value to the business.

These results emphasize the presence of a loyal customer base, crucial for targeted marketing strategies and customer retention efforts.

RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique used to identify and understand the most valuable customers by examining three key factors: how recently a customer made a purchase (Recency), how often they make purchases (Frequency), and how much they spend (Monetary).

- **CustomerID 1**: This customer has an RFM score of 95.71 and falls into the "Top Customers" segment. This high RFM score indicates that Customer 1 has made recent purchases, does so frequently, and spends a significant amount of money. As a top customer, they are likely highly engaged and valuable to the business.
- **CustomerID 2**: This customer has an RFM score of 87.22, also placing them in the "Top Customers" segment. While their score is slightly lower than Customer 1, it still reflects strong engagement, frequent purchasing behavior, and high spending. Customer 2 is also considered highly valuable and likely contributes significantly to revenue.
- **CustomerID 3**: This customer boasts the highest RFM score of 98.75 and is categorized as one of the "Top Customers". This score suggests extremely recent activity, very frequent purchases, and substantial spending. Customer 3 represents an exceptionally valuable customer, demonstrating the highest level of engagement and spending among the three listed.

Table 3: Cluster Evaluation Metrics

K	Silhouette Score	Davies-Bouldin Index	Inertia
2	0.351	1.196	249.87

3	0.477	0.665	139.7
4	0.499	0.685	89.1
5	0.564	0.564	47.18
6	0.509	0.697	38.82
7	0.461	0.783	34.12
8	0.438	0.783	30.02
9	0.413	0.852	25.57
10	0.411	0.806	24.08

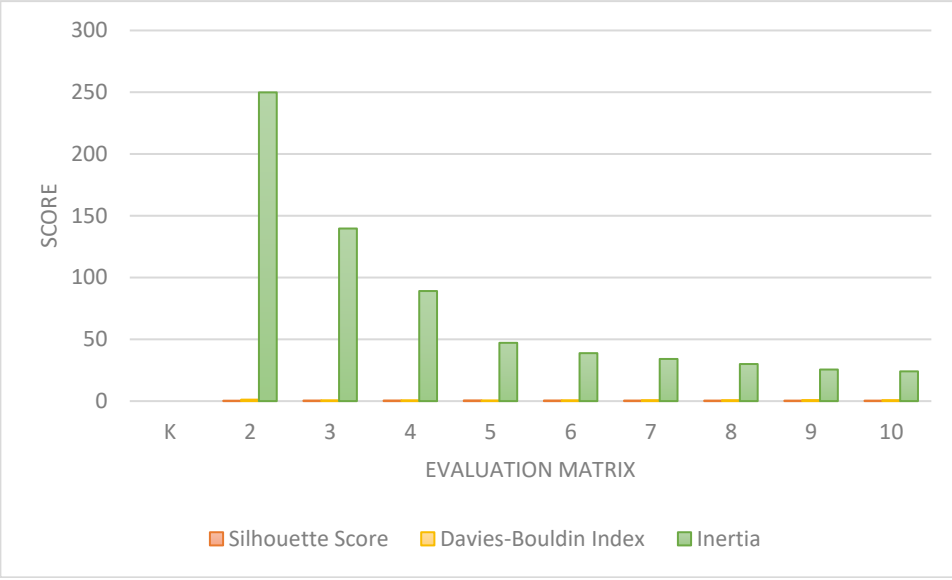

Fig : **Cluster Evaluation Metrics**

RFM Segmentation: Customers in this segment have high RFM scores is top cutomer, indicating that they are the most engaged and valuable. They purchase frequently, have recently made purchases, and spend more than other customers. Businesses should focus on retaining these customers by offering them exclusive deals, loyalty rewards, and personalized experiences to ensure their continued patronage and to maximize their lifetime value. And the middle level RFM score is the moderate purchaser and the lowest RFM score indicate the customer with less purchase capacity. Overall, these results from the RFM analysis highlight the top customers who are crucial for the business. Recognizing and nurturing such customers can lead to increased customer loyalty, higher sales, and sustained business growth.

Table 3 summarizes the evaluation metrics for different values of kkk in the KMeans clustering analysis, providing insights into the effectiveness of the clustering results The clustering results for various values of kkk (the number of clusters) are evaluated using the

Silhouette Score, Davies-Bouldin Index, and Inertia. These metrics help determine the optimal number of clusters for the data.

- **Silhouette Score**: This metric assesses the quality of clustering by measuring the cohesion and separation of clusters. A higher Silhouette Score indicates better-defined clusters. The scores range from 0.351 for k=2k = 2k=2 to 0.564 for k=5k = 5k=5, after which the score declines, indicating that five clusters provide the most distinct and well-separated grouping.
- **Davies-Bouldin Index**: This index evaluates clustering performance based on the average similarity between clusters. Lower values of the Davies-Bouldin Index signify better clustering performance. The index shows a general improvement as kkk increases from 2 to 5, with the lowest value of 0.564 at k=5k = 5k=5, suggesting optimal clustering at this point. Beyond k=5k = 5k=5, the index increases, indicating less distinct clusters.
- **Inertia**: Inertia measures the sum of squared distances of samples to their nearest cluster center, indicating cluster compactness. Lower inertia values reflect more tightly grouped clusters. Inertia decreases significantly from 249.87 at k=2k = 2k=2 to 24.08 at k=10k = 10k=10, showing that increasing kkk results in more compact clusters.

Table 4: Cluster Profiles

Cluster	Average Age	Average Annual Income (k$)	Average Spending Score (1-100)
0	46.55	49.27	45
1	30.63	94.38	70.75
2	44.16	98.79	20.68
3	25.33	25.1	80.05
4	45.22	26.3	20.91
5	42.86	49.29	56.04
6	40.14	63.28	48.55
7	33.42	74.63	82.74
8	36.63	75.21	16.58
9	32.92	100.17	88.75

Fig 2: Cluster Profile

From these metrics, the optimal number of clusters appears to be k=5k = 5k=5. At this point, the Silhouette Score reaches its peak, and the Davies-Bouldin Index is at its lowest, indicating well-separated and distinct clusters with good compactness. While inertia continues to decrease with more clusters, the diminishing returns in Silhouette Score and the increasing Davies-Bouldin Index beyond k=5k = 5k=5 suggest that additional clusters may not provide significantly better segmentation.

The cluster profiles present in table 4 is a detailed view of the characteristics of each customer segment based on their average age, average annual income, and average spending score. Here are the insights for each cluster:

- **Cluster 0**: This group has an average age of 46.55 years, an average annual income of $49.27k, and an average spending score of 45. These individuals are middle-aged with moderate income and balanced spending habits.
- **Cluster 1**: With an average age of 30.63 years and an average annual income of $94.38k, this cluster has a relatively high spending score of 70.75. These younger individuals are high earners and tend to spend more.
- **Cluster 2**: The average age here is 44.16 years, with an annual income of $98.79k and a low spending score of 20.68. Despite their high income, these middle-aged individuals are conservative in their spending.
- **Cluster 3**: This cluster comprises young adults with an average age of 25.33 years, a low income of $25.1k, but a high spending score of 80.05. They are value seekers, prioritizing spending despite limited financial resources.

- **Cluster 4**: Members of this cluster are older, with an average age of 45.22 years and an annual income of $26.3k. Their average spending score is 20.91, indicating budget-conscious behavior.
- **Cluster 5**: With an average age of 42.86 years and an annual income of $49.29k, this cluster has a moderate spending score of 56.04. These individuals are middle-aged with balanced income and spending habits.
- **Cluster 6**: This group has an average age of 40.14 years, an income of $63.28k, and a spending score of 48.55. They represent middle-aged individuals with moderate income and average spending behavior.
- **Cluster 7**: With an average age of 33.42 years and a high income of $74.63k, this cluster exhibits the highest spending score of 82.74. These younger, high-income individuals are significant spenders.
- **Cluster 8**: This cluster has an average age of 36.63 years, an income of $75.21k, but a low spending score of 16.58. They have a relatively high income but prefer to spend conservatively.
- **Cluster 9**: The members of this cluster are around 32.92 years old, with the highest average annual income of $100.17k and the highest spending score of 88.75. These young, affluent individuals are top spenders.

These profiles can be used to develop targeted marketing strategies tailored to the specific characteristics and behaviors of each cluster, optimizing engagement and customer satisfaction.

The RFM analysis provided insights into customer behavior and value based on Recency, Frequency, and Monetary metrics. Customers were segmented into distinct groups:

- **Top Customers**: These customers exhibited high RFM scores, indicating recent purchases, frequent transactions, and high spending. They represent a valuable segment for personalized marketing strategies aimed at enhancing loyalty and increasing average transaction values.
- **Other Segments**: Mid-tier and lower-tier customers were also identified, each characterized by varying levels of RFM scores. These segments require tailored approaches to improve engagement and retention.

Clustering Analysis Insights

Clustering analysis using K-means revealed optimal cluster solutions based on evaluation metrics:

- **Silhouette Score** and **Davies-Bouldin Index** indicated that k=5 produced the most cohesive and distinct clusters, with moderate inertia suggesting compact and well-separated clusters.
- **Cluster Profiles**: Each cluster exhibited unique demographic and behavioral traits. For instance, Cluster 7 comprises younger individuals with high incomes and

spending scores, suggesting a segment ripe for premium product offerings and targeted promotional campaigns.

Marketing Strategies

Based on the RFM and clustering results, the following strategies are recommended:

- **Top Customers**: Implement exclusive rewards programs and personalized marketing initiatives to foster loyalty and increase customer lifetime value.
- **High Value Customers**: Focus on upselling and cross-selling premium products to maximize average transaction values and enhance profitability.
- **Mid and Low Value Customers**: Tailor promotions and discounts to encourage repeat purchases and improve retention rates. Implement reactivation campaigns for low-engagement segments to stimulate renewed interest.

V.CONCLUSION

In this study, we employed RFM analysis and clustering techniques to segment customers of a retail mall based on their transactional behavior and demographics. The findings provided valuable insights into customer segmentation, which can guide targeted marketing strategies aimed at enhancing customer engagement and maximizing business profitability. By analyzing Recency, Frequency, and Monetary metrics, we identified distinct customer segments such as Top Customers, High Value Customers, and segments with varying engagement levels. These insights enable personalized marketing strategies tailored to different customer behaviors and preferences. Utilizing K-means clustering, we further categorized customers into cohesive groups based on RFM scores and additional attributes. The optimal number of clusters was determined using metrics like Silhouette Score and Davies-Bouldin Index, highlighting clear and meaningful segmentation patterns. Each cluster exhibited unique characteristics in terms of age, income, and spending behavior. This granular understanding allows for targeted marketing initiatives that resonate with specific customer segments, thereby improving campaign effectiveness and customer satisfaction.

VI.REFERENCES

[1] Velmurugan, T. "Evaluation of k-Medoids and Fuzzy C-Means clustering algorithms for clustering telecommunication data." In 2012 international conference on emerging trends in science, engineering and technology (INCOSET), pp. 115-120. IEEE, 2012.

[2] Dawane, Vinit and Waghodekar, Prajakta and Pagare, Jayshri, RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention Proceedings of the International Conference on Smart Data 2021 http://dx.doi.org/10.2139/ssrn.3852887

[3] Wan, Shicheng, Jiahui Chen, Zhenlian Qi, Wensheng Gan, and Linlin Tang. "Fast RFM model for customer segmentation." In Companion Proceedings of the Web Conference 2022, pp. 965-972. 2022.

[4] Rane, Nitin, Saurabh Choudhary, and Jayesh Rane. "Hyper-personalization for enhancing customer loyalty and satisfaction in Customer Relationship Management (CRM) systems." Available at SSRN 4641044 (2023).

[5] Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." Sustainability 14, no. 12 (2022): 7243.
[6] Fang, Chu, and Haiming Liu. "Research and application of improved clustering algorithm in retail customer classification." Symmetry 13, no. 10 (2021): 1789.
[7] Yeh, I.-C.; Yang, K.-J.; Ting, T.-M. Knowledge Discovery on RFM Model Using Bernoulli Sequence. Expert. Syst. Appl. **2009**, 36, 5866–5871. [**Google Scholar**] [**CrossRef**]
[8] Chang, H.-C.; Tsai, H.-P. Group RFM Analysis as a Novel Framework to Discover Better Customer Consumption Behavior. Expert. Syst. Appl. **2011**, 38, 14499–14513. [**Google Scholar**] [**CrossRef**]
[9] Miglautsch, J.R. Thoughts on RFM Scoring. J. Database Mark. Cust. Strategy Manag. **2000**, 8, 67–72. [**Google Scholar**] [**CrossRef**]
[10] Peker, S.; Kocyigit, A.; Eren, P.E. LRFMP Model for Customer Segmentation in the Grocery Retail Industry: A Case Study. Mark. Intell. Plan. **2017**, 35, 544–559. [**Google Scholar**] [**CrossRef**]
[11] https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
[12] Sridevi P C, T. Velmurugan, "Impact of Preprocessing on Twitter Based Covid-19 Vaccination Text Data by Classification Techniques", IEEE International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2022).
[13] Shirole, Rahul, Laxmiputra Salokhe, and Saraswati Jadhav. "Customer segmentation using rfm model and k-means clustering." Int. J. Sci. Res. Sci. Technol 8 (2021): 591-597.
[14] Aggelis, Vasilis, and Dimitris Christodoulakis. "Customer clustering using rfm analysis." In Proceedings of the 9th WSEAS International Conference on Computers, p. 2. 2005.