International Journal of Computing and Digital Systems
2025, VOL. 17, NO.1, 1-11

http://dx.doi.org/10.12785/ijcds/1571024871

RFM-T Model Clustering Analysis in Improving Customer Segmentation

Astrid Dewi Rana[1], Quezvanya Chloe Milano Hadisantoso[1] and Abba Suganda Girsang[1]

[1]*Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University Jakarta, Indonesia, 11480*

Received 27 April 2024, Revised 17 October 2024, Accepted 25 October 2024

Abstract: In today's fast-paced business environment, understanding and identifying customers is crucial for developing successful marketing strategies. This study explores customer segmentation, a key element of effective marketing, with a focus on the commonly used RFM (Recency, Frequency, and Monetary) model. Various adaptations of the RFM model have been investigated, with the RFM-T model being a notable extension that introduces "Time" as an additional variable. To assess their efficacy in customer segmentation, the study contrasts the performance of the enhanced RFM-T model with that of the classic RFM model. The K-Means algorithm, a well-liked technique for grouping data points, is used in the study with data from a US-based online retail platform. The ideal number of clusters can be determined using the Elbow Curve approach, and the segmentation quality is evaluated using the Silhouette Score, a measurement tool for analysing the integration and break down of clusters. This study compares the classic RFM and RFM-T models in an effort to shed light on how both models can enhance client segmentation and profiling in the online retail industry. The results provide useful advice for companies, assisting them in modifying their marketing plans to better suit the evolving demands and tastes of their customers.

Keywords: RFM, RFM-T, Time, K-Means algorithm, Customer segmentation

1. INTRODUCTION

In the business domain, a pivotal emphasis lies in identifying and understanding customers to implement effective marketing strategies and optimize their lifetime value. A robust marketing strategy involves deploying diverse and impactful tactics tailored to individual customer needs. Furthermore, database marketing is a common approach employed in customer segmentation for direct marketing endeavors. With the rapid expansion of collected data, marketers encounter the challenge of allocating their marketing communication budget judiciously, focusing on the most promising customers [1]. Leveraging advanced data insights allows us to know who our customers are and their behavior. This identification of customer profiling serves as a valuable tool in enhancing marketing strategies, facilitating more informed and precise decision-making processes.

RFM (Recency, Frequency, and Monetary) consumer segmentation has become a common approach in modern commercial operations. This method categorizes customers based on their distinct characteristics and behavior. Various clustering algorithms have been applied to effectively group customers, aiming to enhance the precision and efficacy of clustering outcomes [2]. This model evaluates customer behavior based on three key metrics: recency, which assesses how recently a customer has made a purchase; frequency, which measures the rate of customer transactions; and monetary, which gauges the total monetary value of a customer's transactions. RFM facilitates the identification of buyer characteristics that influence responses [3].

The variable "T" has been added to the RFM model, which goes by the name RFM-T, to indicate inter-purchase time. T determines the typical amount of time that passes between a customer's subsequent transactions [4]. The purpose of this T addition is to examine the connection between the ocurrency and tendency of internet purchasing [5]. As a result, there is a strong likelihood that RFM-T could surpass traditional RFM models, especially in the online retail industry.

The main aim of this study is to perform a thorough comparative analysis of the traditional RFM model and the enhanced RFM-T model. The objective is to evaluate whether the addition of the "T" variable leads to meaningful improvements in customer segmentation, ultimately increasing the accuracy and effectiveness of marketing strategies. To accomplish this, we will use a comprehensive dataset

E-mail address: astrid.rana@binus.ac.id, quezvanya.hadisantoso@binus.ac.id, abba.girsang@binus.ac.id

that includes all transactions from a US-based online retail platform. The segmentation of customers will be conducted using the K-MEANS clustering algorithm, applied to both the traditional RFM and the RFM-T models. To find the optimal number of clusters, we will utilize the elbow curve method, which identifies the point at which adding more clusters yields diminishing returns in variance explanation. Additionally, we will evaluate the quality of the segmentation using the silhouette score, a metric that assesses how closely an object relates to its own cluster compared to others, thus providing a clear measure of clustering effectiveness.

The structure of this article is as follows: Section 2 offers an in-depth review of the relevant literature, examining prior studies and methodologies in customer segmentation. Section 3 explains the methodology used in this research, including the data collection process, the implementation of the K-MEANS algorithm, and the metrics used for evaluation. Section 4 presents the findings from the comparative analysis of the RFM and RFM-T models and includes a discussion of the results. Lastly, Section 5 provides a conclusion, summarizing the key insights, implications for marketing strategies, and recommendations for future research.

2. LITERATURE REVIEW

The RFM model's first concepts were presented. The RFM model's definition was initially put out by Hughes [9]. Only transactional factors like recency, frequency and monetary are taken into account by this classic RFM model, which excludes other customer attributes [[10]. Consequently, a great deal of research has been done to enhance customer segmentation effectiveness by applying machine learning and including additional variables into the conventional RFM model.

Over the past three years (2022-2023), significant advancements have been made in the development of RFM (Recency, Frequency, Monetary) models, particularly in the application of clustering algorithms and the integration of additional factors. This is shown in Table I. For example, D. Bartine investigated how to improve RFM-based consumer segmentation using traditional supervised machine learning techniques including Naïve Bayes, Logistic Regression, SVM, and Decision Tree [1]. This approach has proven effective in refining segmentation accuracy.

On the other hand, A.J. Christy adopted a different methodology, employing unsupervised learning techniques such as K-Means, RM K-Means, and C-Fuzzy to cluster RFM variables [6]. This approach has gained popularity due to its effectiveness in identifying distinct customer segments, with clustering techniques becoming increasingly favored for RFM analysis.

As RFM models continue to evolve, researchers have explored additional variables to enhance clustering outcomes. For example, M.Y. Smaili introduced the RFM-D model, which incorporates Product Diversity as an additional attribute, significantly improving the accuracy and granularity of cluster results. Similarly, J. Zhou extended the traditional RFM model by integrating the T attribute, representing Interpurchase Time—the duration between successive transactions made by customers [4]. This extension aimed to capture more nuanced customer behavior patterns.

Additionally, A. Ullah experimented with a number of clustering analysis variables, including the Dunn index, Davies-Bouldin, Calinski-Harabasz, and Silhouette, in addition to techniques like K-Means, Hierarchical Clustering, Gaussian Mixture Models, and DBSCAN [8]. Although this study did not conclusively prove the effectiveness of the T variable compared to the classic RFM model, it underscored the potential importance of incorporating the T variable.

Given the importance of the T variable in the RFM model, we chose to compare the performance of the extended RFM-T model with that of the regular RFM model in order to better examine its impact. This comparison aims to determine the extent to which the T variable impacts clustering outcomes and overall model effectiveness.

For the model we consider to use K-Means as it is on of the most popular clustering technique [11]. RFM models and other customer segmentation models could potentially be used with the K-Means algorithm. A non-hierarchical clustering technique called K-Means divides data into clusters with a focus on low inter-cluster similarity and high intra-cluster similarity [12]. The optimal number of clusters is still difficult to determine, despite K-Means' popularity due to its computational simplicity and quickness in choosing the cluster (k) centre (centroid) [13]. According to research by Subbalakshmi et al., choosing the right initial value and cluster selection can increase the K-Means method's accuracy [14].

Plotting the number of clusters against a metric, usually the within-cluster sum of squares (WCSS) or inertia, allows the Elbow Method (EM) to determine the ideal number of clusters by measuring the curvature of the curve that results [15]. This technique helps identify the spot on a curve, which frequently has an elbow-like appearance, where the slope changes significantly. This crucial point, represented by 'k', shows the ideal number of clusters. Typically, the elbow point signifies the balance between minimizing within-cluster variance and avoiding overfitting, making it a vital step in cluster analysis.

The Silhouette Score stands out among various methods used to evaluate clustering results. Unlike most other performance assessment techniques, the Silhouette Score does not necessitate a training set for evaluating clustering outcomes. This characteristic makes it particularly well-suited for assessing RFM clustering [16]. Since RFM clustering aims to categorize customers based on their transaction behavior, the Silhouette Score's ability to evaluate clustering quality without requiring labeled data makes it a valuable tool in

TABLE I. COMPARATIVE ANALYSIS OF RFM-BASED CUSTOMER SEGMENTATION MODELS AND METHODS

Author (Year)	Model	Method
D. Bratina (2023) [1]	RFM	Naïve Bayes, Logistic Regression, SVM, and Decision Tree
A. J. Christy (2021) [6]	RFM	K-Means, RM K-Means, and C-Fuzzy
M. Y. Smaili (2023) [7]	RFM-D (Product Diversity)	K-Means
J. Zhou (2021) [4]	RFM-T	Hierarchical Clustering
A. Ullah (2023) [8]	RFM-T	K-Means, Hierarchical Clustering, Gaussian, and DBSCAN

this context. This approach enables businesses to effectively measure the coherence and separation of clusters generated by the RFM clustering algorithm, providing insights into the effectiveness of the segmentation process.

3. RESEARCH METODOLOGY

This section describes the suggested approach for identifying customer segments utilizing the RFM and RFM-T model then using clustering algorithms (Elbow Curve and K-means) in order to maximize benefits and compare both of the models.

The proposed systematic framework in Figure **??**, the dataset undergoes pre-processing and normalization to handle missing values and standardize scales. Feature extraction follows, generating Recency (R), Frequency (F), Monetary (M) and Time (T) attributes that encapsulate essential customer behavior patterns. The K-Means clustering algorithm is then applied to both RFM (Recency, Frequency, Monetary) and RFM-T (including Time) models for segmentation. Silhouette analysis evaluates the quality of clusters, facilitating a comparative study between RFM and RFM-T models. The final stage involves detailed customer analysis within each segment, providing insights into customer characteristics and preferences. This comprehensive approach, spanning pre-processing, feature extraction, clustering and customer analysis, forms a structured framework for effective customer segmentation using the RFM and RFM-T models.

A. *Dataset*

This research utilized the US based online retail dataset encompassing one-year transactions recorded from 01/01/2019 to 31/12/2019 contained 52,955 entries across 11 attributes as seen on Table II. The company specializes in selling electronic, office and apparel products. They also offer coupons for the customers to use on each product they bought. On the dataset, customer age and location are provided for better customer profiling [17].

B. *Data Pre-processing*

From the dataset, it is evident that the data contained in 52,955 entries requires data processing. The main goal of this step is data pre-processing, which removes erroneous data that may affect the analysis and final segmentation.

TABLE II. DATASET ATTRIBUTES DESCRIPTION

No	Attribute	Description
1	Customer ID	A distinct identity for every customer
2	Gender	Customer's gender
3	Location	Customer's location
4	Invoice ID	A distinct identity for every invoice
5	Date	The date of the transaction
6	Product ID	A distinct identity for every product
7	Product Description	An overview of the product
8	Product Category	Categorization of the product
9	Quantity	The amount of products purchased throughout a transaction
10	Price	The cost of the item per unit
11	Coupon Status	Indicates whether a coupon was applied

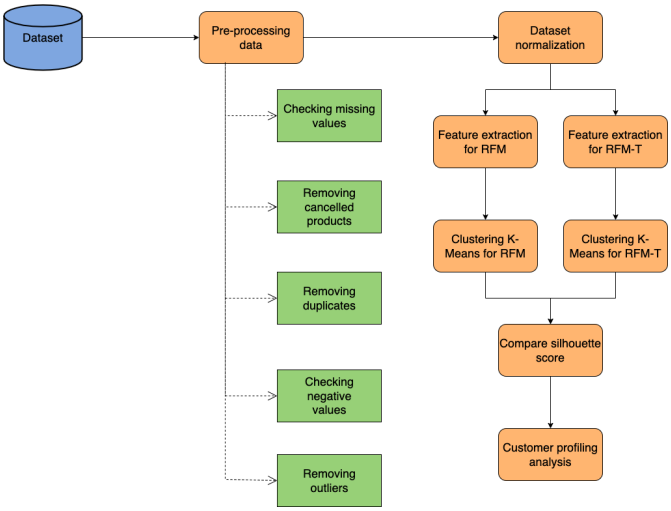

Figure 1. After the outliers were removed.

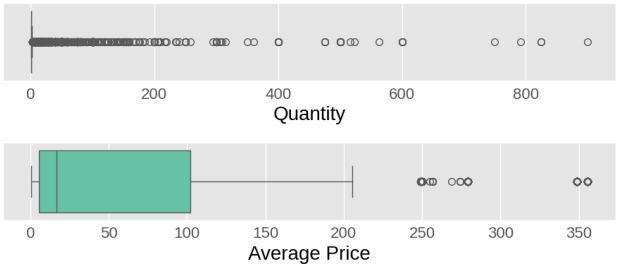

Figure 2. Before the outliers were removed.

Figure 3. After the outliers were removed.

Data with null stock codes, transactions with negative values, and lines without customer numbers are impacted by this step.

Upon analysis, the dataset exhibits missing values for certain Customer IDs which need to be removed. This process has removed a total of 31 records from 52, 955 to 52, 924.

After removal, outliers need to be handled by replacing the outlier values with a specified threshold value of the up and low limit. The outliers can be detected on attributes with numeric values. In this case it is quantity and price. The before and after transformation can be seen as visualized in Figure 2 and Figure 3.

Additionally, Table III outlines the distribution of customers by gender and location, with 62.3% being female and 37.6% male. The majority of customers are located in Chicago (65.1%), followed by New York, New Jersey, and Washington.

TABLE III. DISTRIBUTION OF CUSTOMERS BY GENDER AND LOCATION

Attribute		Percentage
Gender	Female	62.3%
	Male	37.6%
Location	New York	21.1%
	Chicago	65.1%
	New Jersey	8.5%
	Washington	5.1%

C. RFM and RFM-T Model Score

The When it comes to segmenting and analyzing potential customers, the RFM model is highly used. It is a model that primarily analyzes customer behavior with regard to transactions and purchases before making a database forecast. This model consists of three measures: monetary, frequency and recency, which are combined into the concept of RFM [18].

In this study, Time variable is added to analyze its effect on improving the performance of RFM in customer segmentation. According to the suggested RFM-T model, Time is calculated by adding up and averaging the number of days between consecutive transactions. This means it takes into account not just how often and how much a customer spends, but also how quickly they repeat purchases. The ”Time” component focuses on measuring the time interval between successive transactions for each customer. It captures their purchasing rhythm and allows for identifying customers who tend to buy frequently within a short timeframe or those with longer intervals between purchases [18].

To build the model, each variable's score needs to be calculated first. Here, let's examine the meanings of R, F, M and T:

- Recency Score: The number of days that have passed between the customer's most recent transaction and their last purchase made within the analysis period. The defined last purchase date in this study is 01/06/2011. The recurring visits from satisfied customers are indicative from a modest value of recency.
- Frequency Score: The frequency throughout the study time indicates how many visits the consumer made. As long as frequency has a high value, he is regarded as loyal.

- Monetary Score: This metric indicates the total expenditure by a customer over the test dataset period. A higher monetary score suggests that a customer is likely more satisfied with the store, as they are willing to spend more. Equation (1) illustrates the calculation of the Monetary Score, where Q represents quantities multiplied by the price per unit (P).

$$M = \sum_{i=1}^{n} (Q_i x P_i) \tag{1}$$

- Monetary Time Score: The sum of days between all consecutive transactions made by a customer is calculated. This sum is then divided by the total number of transactions to obtain the average time interval between purchases. Equation (2) details how the Time Score is calculated, where L represents the shopping cycle, obtained by summing transaction date gaps (T_i). Then, L is divided by the number of frequencies (F) minus one, considering only transactions with a frequency greater than 1.

$$T = \frac{L}{F-1} = \frac{\sum_{i=2}^{n} (T_i + T_{i-1})}{F-1} \tag{2}$$

A statistical summary generated on Table IV after determining the values of R, F, M, and T. Each variable contains 1,468 data points. The mean values indicate that, on average, the most recent purchase was 145.29 days ago, customers made 2.19 purchases, and the average monetary value per customer was $2546.61 over a typical timeframe of 19.50 units (days, months, etc., depending on context).

The data shows considerable variation, particularly in Monetary values, with a standard deviation of 3641.88, suggesting significant differences in customer spending. The range in Monetary values is also notably wide, from a minimum of $1 to a maximum of $42,433.25, highlighting the presence of outliers.

For Recency, the values range from 1 to 365 days, with quartile data showing that 25% of customers have made a purchase within the last 56 days, and 75% within the last 221 days. Frequency and Time distributions are also detailed, with max values indicating the highest observed purchase frequencies and time periods.

Given the variability and skew in these statistics, normalizing the data is before using it in clustering or machine learning models to ensure a common range, preventing the creation of biased models.

D. Normalization

Table IV shows that the variables have different scales that can lead to issues in the algorithm later. Especially, K-Means involves distance calculations, hence features with larger scales could disproportionately influence the results. Furthermore, when the skewness of the data is checked, Figure 4 proves that the variables are right-skewed. This distribution is identified by the the x-axis that represents the range of values for the variable, and each variable has the long tail that extends to the right or positive side of the x-axis.

TABLE IV. R, F, M AND T VARIABLE STATISTICS

	Recency	Frequency	Monetary	Time
count	1468.00	1468.00	1468.00	1468.00
mean	145.29	2.19	2546.61	19.50
std	101.94	2.24	3641.88	30.14
min	1.00	1.00	1.00	0.00
25%	56.00	1.00	561.40	0.00
50%	132.00	1.50	1468.16	0.00
75%	221.00	3.00	3311.86	36.00
max	365.00	34.00	42433.25	175.00

For normalization, this study used the Quantile Transformer method. It is a technique used to transform the probability distribution of a dataset into a specific known distribution or a uniform distribution. It is particularly useful when the data has a skewed distribution. After normalization, the variables are distributed normally as shown in Figure 5.

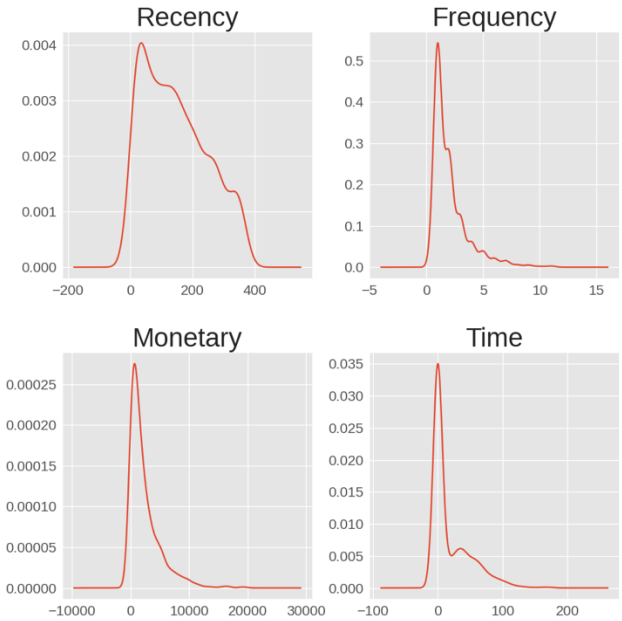

Figure 4. R, F, M and T skewness

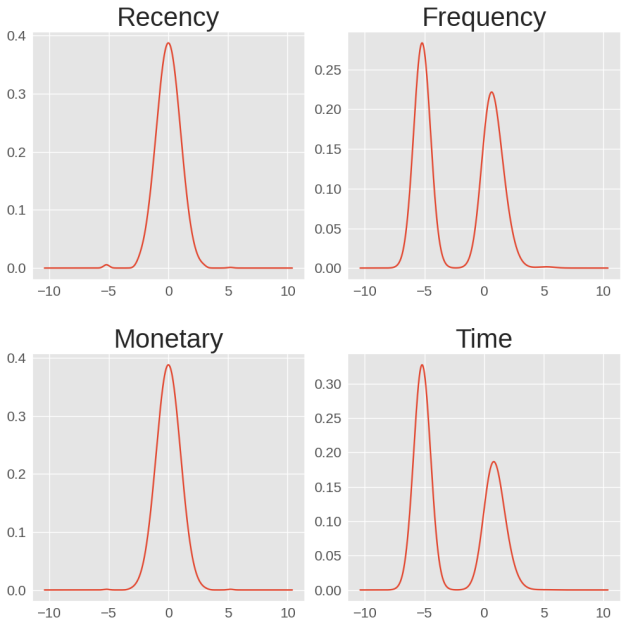

Figure 5. R, F, M and T skewness after normalization.

E. Clustering K-Means

K-Means used as the algorithm for clustering in this study. A few advantages K-Means assures convergence, scales huge data sets, is quite easy to construct, and is easily adaptable to new instances [18].

K-Means determines how many clusters to form by calculating the Euclidean Distance. Euclidean Distance as a method to calculate the distance between data points and centroids. The formula for Euclidean Distance shown as on Equation (3). In this case, xi denotes the ith point in the dataset, k stands for K cluster center, and μk for the kth center. This calculation will be assisted by implementing the Scikit-Learn library.

$$d = \sum_{k=1}^{K} \sum_{i=1}^{n} (x_i - \mu_k)^2 \tag{3}$$

The computer calculates the centroid value before each iteration step in the K-Means algorithm. The Euclidean Distance metric is employed to determine the cluster with the closest centroid for each data point. The iterative process continues until the results of the cluster obtained are comparable to the results of the previous iteration, the procedure as mentioned earlier will be repeated.

The potential of the algorithm to recognize inherent patterns in the dataset is demonstrated by the examination of the K-means clustering findings in RFM and RFM-T. The resulting clusters show distinct separation, suggesting that the data contains significant patterns of groups of customers. It is crucial to recognize that the selection of K classifications has carefully assessed how many clusters are acceptable given the specifics of their study.

F. Elbow Curve

To determine the number of clusters, the ”Elbow Method” is employed. The Elbow Curve methodology is considered the most reliable and effective method for figuring out the optimal number of clusters for RFM and RFM-T segmentation. The graph’s slope is used to determine how many clusters to create [14]. The Elbow Curve rule produces a range of possible values for K by taking the square of the distance between the centroid of each cluster and the data points. Known as the distortion or inertia score, the sum of squared errors (SSE) is employed as a performance metric (Equation 4). As long as the SSE values are low, clusters converge. The SSE shows a sharp decrease when the number of clusters gets closer to the ideal number. The SSE declines, although very slowly, if the ideal number of clusters is surpassed [7].

$$\text{Inertia} = \sum_{i=1}^{n} \min_{k} \|x_i - \mu_k\|^2 \tag{4}$$

The distance between the data points in a cluster and its centroid, or centre, is known as the cluster’s inertia. In the context of K-means clustering, the formula for inertia is the sum of squared distances, over all clusters, between each data point in a cluster and its centroid. The term , which denotes the mean of a cluster’s sample—that is, the average value of the data points within a particular cluster—is used mathematically to calculate the inertia in Equation (4). Each distinct data point inside a cluster is represented by the symbol , while the total number of clusters in the dataset is indicated by the number n. Therefore, to provide a representative measure for that specific set of data points, is computed as the mean of all xi values within the corresponding cluster.

G. Silhouette Score

The optimal cluster number discovered using the Elbow technique was verified using the silhouette methodology. This method was first introduced by Peter J. Rousseeuw in 1987 to aid in the interpretation and validation of cluster analysis [19]. The Silhouette score quantifies an object’s cohesion—how similar it is to its own cluster—as opposed to separation—how similar it is to other clusters. The item is poorly matched to nearby clusters and well matched to its own cluster if it has a high Silhouette score. On a scale of -1 to 1, a score getting close 1 denotes significant clustering, a score of 0 means the object is close to the boundary between two clusters, and a score below 1 implies misclassification. Equation (5) provides the formula for determining the Silhouette score for a data point (i).

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{5}$$

The average distance between point (i) and every other point in the same cluster is represented by (a(i)), while the

smallest average distance between point (i) and points in the closest cluster is represented by (b(i)). For a given number of clusters, the Silhouette score serves as an effective tool to evaluate the quality of clustering, helping to measure the consistency and separation between clusters.

4. RESULTS AND DISCUSSION

The retail business may implement more targeted marketing tactics to particular customer groups for improved retention by analyzing the features of each grouped cluster and comparing the quality and performance of the RFM and RFM-T models to determine which is the best. In the sections that follow, those results are addressed.

A. Time as Extended Variable

The variables R, F, M and T of the model underwent normalization using Quantile Transformers to address any asymmetry in their values. Then the data undergo a correlation matrix analysis using Pearson Correlation Coefficient formula shown as on Equation (4). Where Y and X represent the two variables being analyzed for their correlation together.

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{6}$$

The correlation analysis based on RFM-T variables revealed a significant correlation. In Figure 6, the heatmap of R, F, M and T variables exhibited a strong positive correlation between the Time and Frequency variables (0.86). Additionally, a moderate positive correlation (0.40) was observed between the Time and Frequency variables. Thus, these findings underscore the importance of considering variable Time when analyzing customer behavior in the context of RFM-based segmentation.

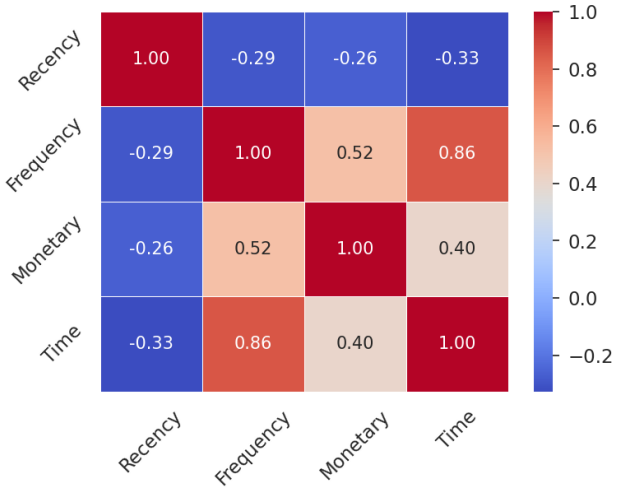

Figure 6. R, F, M and T correlations

B. RFM vs RFM-T

The purpose is to compare the segmentation based on the two models and provide evidence that enhancing the quality of the clustering requires using the variable "Time" in the customer segmentation.

Both the RFM and RFM-T models are used in the application of the K-Means method to carry out the segmentation. The Elbow Curve method is used to segment data according to the two models in order to determine the ideal number of clusters. As Figure 7 and Figure 8 illustrate, the segmentation using the RFM model yields five clusters, while the segmentation using the RFM-T model creates three distinct clusters.

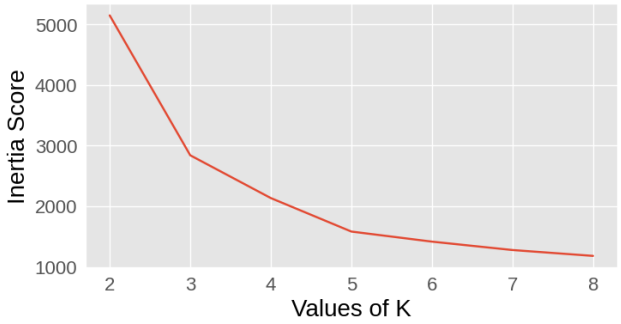

Figure 7. RFM Elbow Curve

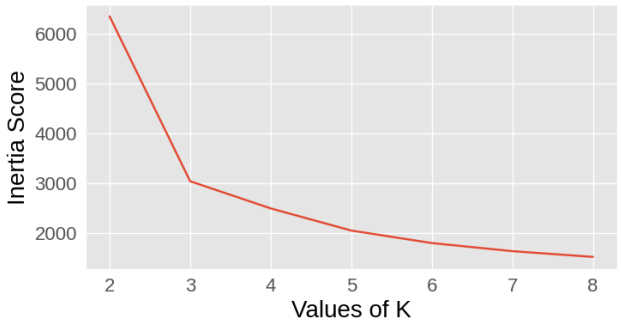

Figure 8. RFM-T Elbow Curve

When analyzing the score on Table V, RFM is leading by its lower inertia value of 1.584 and faster fit time by only 0.486. While inertia value and fit time of RFM-T are 3.043 and 1.556 respectively. However, upon analyzing the Silhouette Score, RFM achieves a score of 0.3347, while RFM-T attains a score of 0.7096. This suggests that RFM-T has better quality and high performance in clustering than the RFM model alone without T.

Additionally, Table VI and Table VII reveal distinct mean values for each variable. In the RFM Analysis, Cluster 4 has the highest mean values for Recency, Frequency, and Monetary metrics, while other clusters show lower values. In the RFM-T Analysis, Cluster 0, the largest group, has

TABLE V. K-MEANS AND SILHOUETTE SCORE ANALYSIS

Model	Clusters	Inertia	Fit Time (s)	Silhouette Score
RFM	5	1.584	0.486	0.3347
RFM-T	3	3.042	1.556	0.7096

low mean values across all variables, whereas Cluster 1 has moderate values, and Cluster 2 shows the lowest Time metric. This highlights differences in the average values of Recency, Frequency, Monetary, and Time across clusters.

TABLE VI. ANALYSIS FOR EACH CLUSTER ON RFM

Cluster	Mean values of				Count	Percentage (%)
	R	F	M	T		
Cluster 0	0.26	-5.19	-0.49	-5.19	734	50.3%
Cluster 1	-0.38	0.90	0.48	0.94	608	41.67%
Cluster 2	0.24	0.33	0.57	-519	117	8.02%
			Total		1.459	100%

TABLE VII. ANALYSIS FOR EACH CLUSTER ON RFM-T

Cluster	Mean Values of			Count	Percentage(%)
	R	F	M		
Cluster 0	-0.93	1.14	1.05	333	22.82%
Cluster 1	0.42	-5.19	-1.35	245	16.79%
Cluster 2	-0.82	-5.19	-0.01	247	16.92%
Cluster 3	1.22	-5.199	-0.11	242	16.58%
Cluster 4	0.27	0.52	0.02	392	26.85%
	Total		1.459		100%

Furthermore, Figure 9 of RFM Monetary-Frequency-Recency and Figure 10 of RFM-T Monetary-Time-Frequency demonstrate that RFM-T clusters exhibit a much clearer separation between clusters compared to RFM.

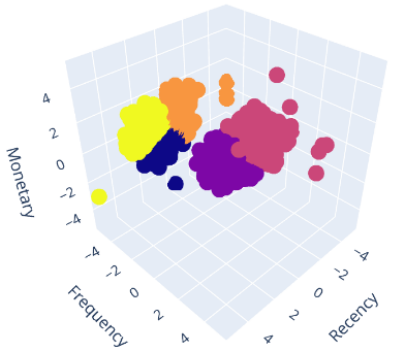

Figure 9. RFM cluster visualization

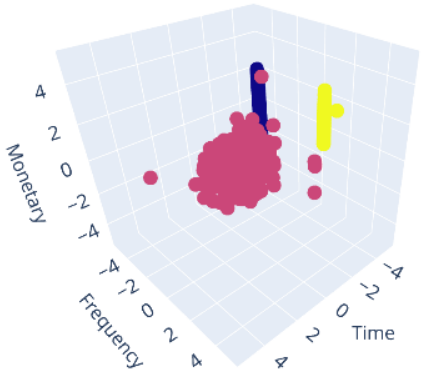

Figure 10. RFM-T cluster visualization

C. *RFM-T Customer Profiling Analysis*

From the previous section, it's evident that the RFM-T clustering model outperforms the RFM model. So, this study will continue the customer profiling on RFM-T model, as a result there will be 3 customer segments across the data. When counted on Table VI, Cluster 0 has the highest number of customers, then Cluster 2 and Cluster 3 came second and third respectively.

The distribution of R, F, M and T variables for each cluster is clearly visualized in Figure 11. Customers belonging to Cluster 0 exhibit the highest frequency scores ranging from 0 to 4, visit the store frequently based on their inter-purchase time scores and spend monetarily at a mid to high level. Similarly, customers in Cluster 2 have moderate frequency and monetary scores, but their visits to the store are less frequent due to a low inter-purchase time score. In contrast, Cluster 1 has low scores for all variables, indicating that customers in Cluster 1 visit the store only once.

Customer identities within each cluster exhibit similar distributions, as illustrated in Table VIII and Table IX, most of each cluster is composed of female customers, primarily originating from Chicago or New York. Examining Figure 12 reveals distinct purchase behaviors for each cluster. In Cluster 0, customers tend to buy a large quantity and variety of products but with low value. They frequently use coupons and predominantly visit the store on weekends. Cluster 1, on the other hand, consists of customers who typically purchase a small quantity of products but with high value. These customers are high spenders, as indicated by their average monthly spending and rarely use coupons, often visiting the store on weekdays. Lastly, Cluster 2 customers exhibit purchase behaviors similar to those in Cluster 0. They prefer using coupons, resulting in lower average transaction values and monthly spending compared to Cluster 1. However, their product diversity and quantity are considered low, and they tend to shop on weekdays.

Here are some suggestions based on valuable segments

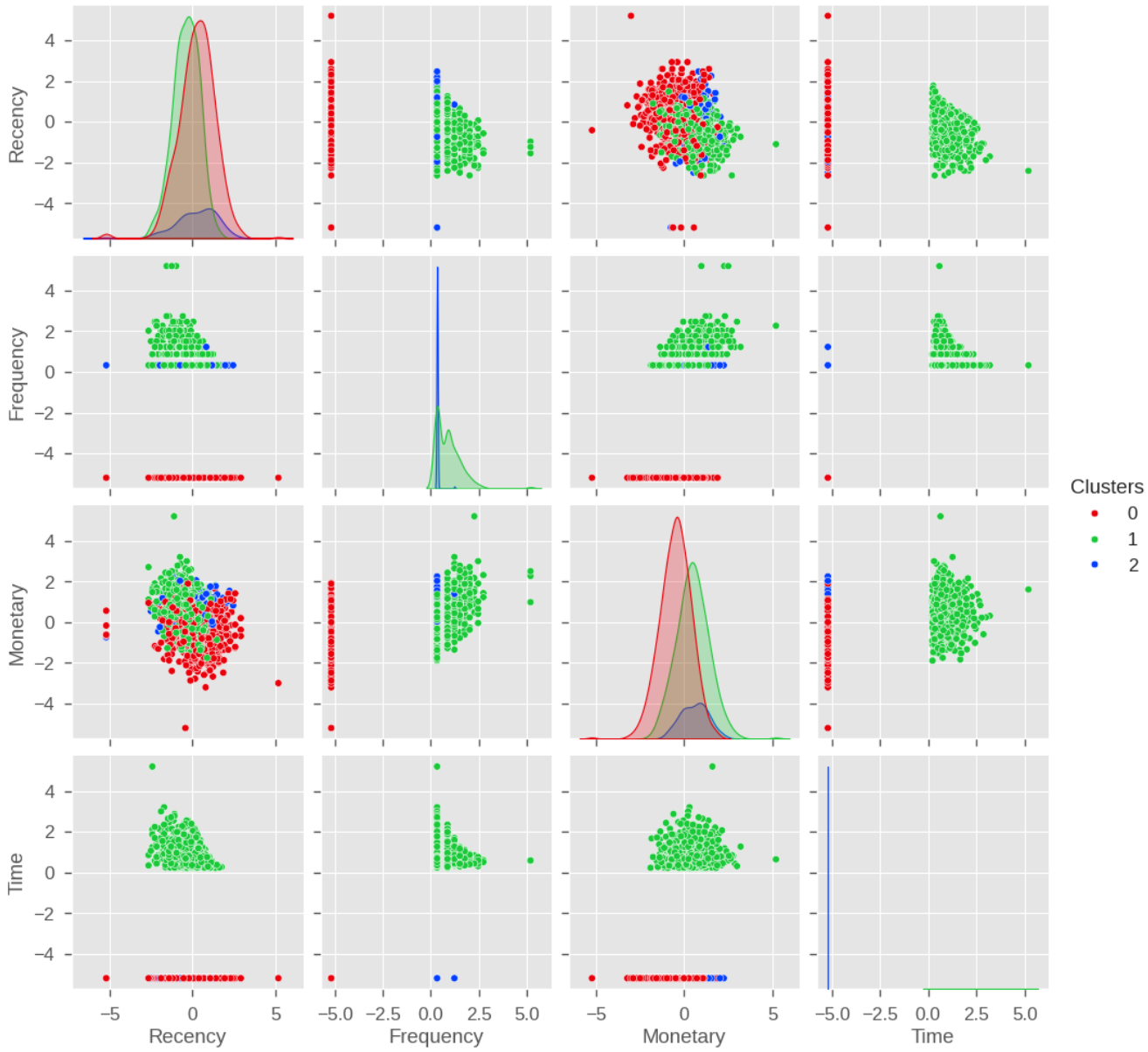

Figure 11. Cluster R, F, M and T distribution

profiles:

- **Cluster 0 is defined as Loyal Customers** by Author because they are regular patrons of the store and have lower spending habits. This group requires maintenance through loyalty promotions and is an ideal target for bulk coupons, given their preference for purchasing high quantities of products at low prices.
- **Cluster 1 is defined as Thrifters** by Author because they are infrequent visitors who only come to the store occasionally for specific high-value items. This group is interested in niche luxury items, making them suitable targets for new trending products in the market.
- **Cluster 2 is defined as Attention-Needed Customers** by Author because this group has high potential to become regular customers like those in Cluster 0, but additional effort is required to increase their frequency, quantities and diversity of purchased products. Similar to Cluster 0, they appreciate coupons, so offering them flash sale promotions on a variety of products on weekdays might be more effective.

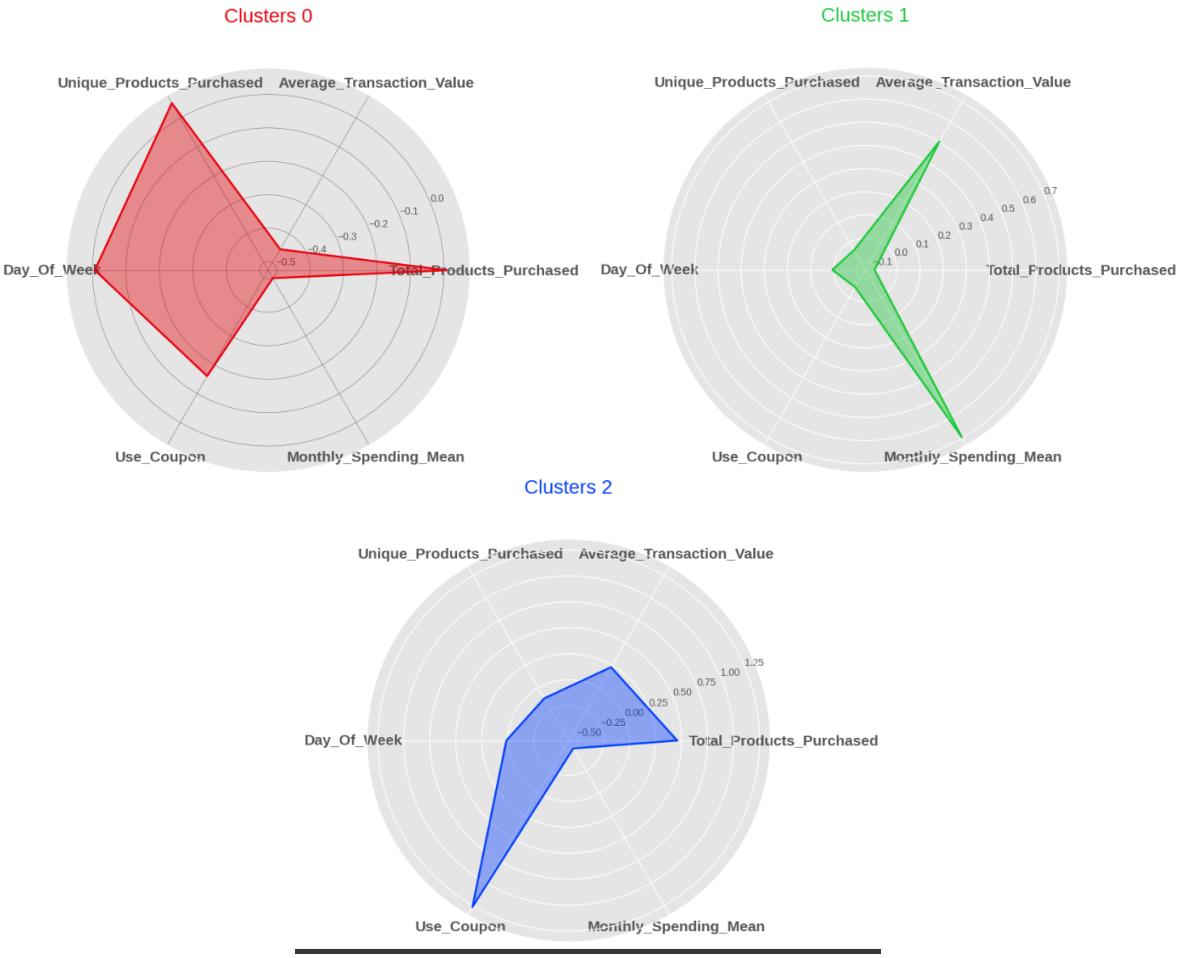

Figure 12. RFM-T cluster behavior.

TABLE VIII. ANALYSIS FOR EACH CLUSTER ON RFM

Cluster	Gender		Location			
	Female	Male	New York	Chicago	New Jersey	Washington DC
Cluster 0	64.6%	35.4%	30.8%	44.7%	19.5%	5.0%
Cluster 1	61.3%	38.7%	30.3%	47.5%	13.4%	8.8%
Cluster 2	63.9%	36.1%	35.9%	43.6%	14.1%	6.4%
Cluster 3	62.2%	37.8%	35.4%	42.7%	13.5%	8.4%
Cluster 4	67.2%	32.8%	29.2%	48.5%	15.2%	7.0%

TABLE IX. ANALYSIS FOR EACH CLUSTER ON RFM-T

Cluster	Gender		Location			
	Female	Male	New York	Chicago	New Jersey	Washington DC
Cluster 0	61.5%	38.5%	32.3%	44.2%	14.4%	9.1%
Cluster 1	65.3%	34.7%	31.9%	45.7%	16.3%	6.2%
Cluster 2	63.2%	36.8%	36.6%	48.8%	8.5%	6.1%

5. CONCLUSION

A Time (T) variable was added to the conventional RFM (Recency, Frequency, Monetary) model in this study in order to assess its effect on outcomes. R, F, M, and T values were retrieved using preprocessing techniques on a dataset from a US-based online shop to guarantee data balance. The ideal number of clusters (K) was found using the Elbow Curve approach in conjunction with K-

Means clustering. Subsequently, the Silhouette Scores for both RFM and RFM-T models were evaluated. The results revealed a significant improvement in the new model with the inclusion of T, with scores of 0.3347 and 0.7096, respectively (as shown in Table III). For customer profiling, demographic information was applied to each RFM-T cluster, providing insights into tailored marketing strategies to enhance customer relationships. Specifically, Cluster 0 consists of lower spenders who require maintenance and bulk coupons for high-quantity, low-price purchases. Cluster 1 includes infrequent visitors interested in niche luxury items, ideal targets for introductions to new products. Cluster 2 comprises potential regulars who could be encouraged to increase purchases through coupons and flash sales.

The primary advantage of this research is the empirical demonstration of the Time variable's impact on the RFM model, showing how T can enhance customer segmentation and predictive accuracy. This knowledge enables companies to more precisely predict consumer behaviour and better customise marketing campaigns, which may improve client engagement and boost sales.

The study's primary issue is that it only uses K-Means clustering. Despite being a reliable and widely used technique, K-Means' efficacy varies based on the distribution of the data and the initial cluster centers selected. Furthermore, clusters of various forms and densities may not perform as well using K-Means since it assumes spherical clusters. Future research could explore other clustering techniques, such as hierarchical clustering or DBSCAN, to potentially uncover more insights and validate the robustness of incorporating the Time variable.

ACKNOWLEDGMENT

The authors would like to extend their appreciation to the individuals who contributed to the online-shopping-dataset, as well as to Jackson Divakar R for his valuable contributions. They also acknowledge the reviewers and peers for their insightful feedback, and express gratitude to the research participants who indirectly supported this study through Kaggle comments.

REFERENCES

[1] D. Bratina and A. Faganel, "Using supervised machine learning methods for rfm segmentation: A casino direct marketing communication case," *Market-Tržište*, vol. 35, no. 1, pp. 7–22, 2023.

[2] S. H. Shihab, S. Afroge, and S. Z. Mishu, "Rfm based market segmentation approach using advanced k-means and agglomerative clustering: a comparative study," in *2019 International conference on electrical, computer and communication engineering (ECCE)*. IEEE, 2019, pp. 1–4.

[3] J. R. Bult and T. Wansbeek, "Optimal selection for direct mail," *Marketing Science*, vol. 14, no. 4, pp. 378–394, 1995.

[4] J. Zhou, J. Wei, and B. Xu, "Customer segmentation by web content mining," *Journal of Retailing and Consumer Services*, vol. 61, p. 102588, 2021.

[5] D. Vakratsas and F. M. Bass, "The relationship between purchase regularity and propensity to accelerate," *Journal of Retailing*, vol. 78, no. 2, pp. 119–129, 2002.

[6] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "Rfm ranking–an effective approach to customer segmentation," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, 2021.

[7] M. Y. Smaili and H. Hachimi, "New rfm-d classification model for improving customer analysis and response prediction," *Ain Shams Engineering Journal*, vol. 14, no. 12, p. 102254, 2023.

[8] A. Ullah, M. I. Mohmand, H. Hussain, S. Johar, I. Khan, S. Ahmad, H. A. Mahmoud, and S. Huda, "Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time," *sensors*, vol. 23, no. 6, p. 3180, 2023.

[9] X. He and C. Li, "The research and application of customer segmentation on e-commerce websites," in *2016 6th International Conference on Digital Home (ICDH)*. IEEE, 2016, pp. 203–208.

[10] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in rfm analysis based on k-means," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470–477, 2020.

[11] S. Idowu, A. Annam, E. Rangarajan, and S. Kattukottai, "Customer segmentation based on rfm model using k-means, hierarchical and fuzzy c-means clustering algorithms," *Jerarchical y Fuzzy C-Means, August 2019*, 2019.

[12] C. Subbalakshmi, G. R. Krishna, S. K. M. Rao, and P. V. Rao, "A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set," *Procedia Computer Science*, vol. 46, pp. 346–353, 2015.

[13] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, "Applying data mining and rfm model to analyze customers' values of a veterinary hospital," in *2016 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, 2016, pp. 481–484.

[14] M. A. Syakur, B. K. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336. IOP Publishing, 2018, p. 012017.

[15] M. Aliyev, E. Ahmadov, H. Gadirli, A. Mammadova, and E. Alasgarov, "Segmenting bank customers via rfm model and unsupervised machine learning," *arXiv preprint arXiv:2008.08662*, 2020.

[16] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021.

[17] A. Dewi Rana, Q. Chloe Milano Hadisantoso, and A. Suganda Girsang, "Rfm-t model clustering analysis in improving customer segmentation," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 1–11, 2024.

[18] A. Amine, B. Bouikhalene, R. Lbibb *et al.*, "Customer segmentation model in e-commerce using clustering techniques and lrfm model: The case of online stores in morocco," *International Journal of Computer and Information Engineering*, vol. 9, no. 8, pp. 1993–2003, 2015.