Customer Segmentation using RFM Analysis and K-Means Clustering to enhance Marketing Strategies

A Technical Report submitted to the Department of Engineering and Applied Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Harshil Anishkumar Patel
Spring, 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Briana Morrison, Department of Computer Science

Customer Segmentation using RFM Analysis and K-Means Clustering to enhance Marketing Strategies

CS4991 Capstone Report, 2023

Harshil Patel
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
pmb5br@virginia.edu

ABSTRACT

The competition among businesses to attract and retain customers has intensified, leading to an increased need for effective customer segmentation methods. This project focuses on developing a customer clustering approach using the k-means clustering algorithm and RFM model to create segments of customers based on their purchasing behavior. The project utilizes UK's E-commerce dataset from UCI repository for Machine Learning as the input data source. The customer clusters are based on RFM score and range from super customers to those at risk of being churned out. This approach will help businesses target customers effectively, resulting in improved performance. The project methodology comprises five stages: business understanding, data understanding, data preparation, modeling, and evaluation. The results show that the k-means clustering algorithm, coupled with RFM model, is effective in clustering customers based on their purchasing behavior. The insight from the model allows businesses to analyze their customers, enabling them to develop targeted marketing strategies that enhance customer retention and acquisition. Future work involves incorporating additional data sources, enhancing the web model's features, and conducting additional testing and evaluation to improve the clustering accuracy. A web model can also be created to enable e-commerce startups and analysts to analyze their customers using the developed model.

1. INTRODUCTION

In today's fast-paced business environment, customer loyalty is the key to success. Businesses that can retain their customers and keep them happy are more likely to succeed than those that cannot. To achieve this, companies must understand the needs, preferences and behavior of their customers and adapt their products and services accordingly. However, with the sheer volume of data available from various sources, analyzing that data and extracting meaningful insights can be daunting. This is where customer segmentation comes into play [3].

Customer segmentation is the process of dividing a customer base into smaller groups based on common characteristics such as demographics, behavior, or preferences. By grouping customers in this way, companies can better understand their customer base and develop strategies to target each group more effectively. This can increase customer satisfaction, loyalty, and revenue [7].

Data mining techniques are commonly used to solve customer segmentation problems and develop effective market strategies. It involves analyzing large datasets to uncover patterns and insights that can inform business decisions. Companies collect data from customers through various methods, including direct methods like transactions and indirect methods like website tracking. Two popular methods for customer segmentation are RFM analysis and K-means clustering. RFM analysis stands for recency,

frequency, and monetary value, and it helps companies identify customers based on their purchasing history; whereas K-means clustering is an algorithm that groups customers based on their interactions with the company [8][9].

Many data scientists prefer to use both RFM analysis and K-means clustering together for customer segmentation. RFM analysis identifies valuable customers by analyzing recency, frequency, and monetary value of transactions, while K-means clustering groups customers based on their similarities in behavior and preferences. By combining both techniques, businesses can tailor marketing strategies according to individual customer needs and preferences, as it provides a comprehensive view of the customer base [6]. This paper demonstrates the usefulness of RFM analysis and K-means clustering in achieving complete and effective customer segmentation.

2. RELATED WORKS

Demographic, socio-cultural, geographic, psychographic and behavioral factors are used to segment customers. Researchers have explored a myriad of segmentation techniques and algorithms, including RFM analysis and K-means clustering [6].

Khajvand & Tarokh (2011) developed a framework for the retail banking sector in Iran that utilized RFM analysis, K-means clustering, and a two-step algorithm to examine customers' background in different periods and estimate their future behaviors [1]. This study informed our project by highlighting the importance of RFM analysis and K-means clustering in understanding customer behavior and loyalty in the banking sector.

Khajvand, et al. (2011) worked with a health and beauty company to propose a model that clustered customers into segments based on RFM analysis and K-means clustering [2]. They found clustering customers into different groups helped decision-makers to identify market segments more clearly and develop more effective strategies for customer retention. This study is relevant to our project as it demonstrates the utility of RFM analysis and K-means clustering in identifying customer segments and improving customer engagement and loyalty.

Hu & Yeh (2014) aimed to define RFM patterns and propose a novel algorithm for discovering complete sets of RFM patterns that could approximate sets of customers in the retailing sector [4]. They evaluated the value of these patterns from the customer's perspective and determined pattern ratings based on RFM features. This study informed our project by demonstrating the efficiency of the proposed approach in discovering RFM-customer patterns and evaluating their values from a customer's point of view.

Overall, these studies emphasize the importance of RFM analysis and K-means clustering in understanding customer behavior and loyalty, identifying customer segments, and developing effective marketing and sales strategies. While our project builds on the methodologies of these studies, it takes a different direction by focusing on the application of RFM analysis and K-means clustering in the e-commerce sector, specifically in the context of improving customer engagement, loyalty, and revenue growth.

3. METHODOLOGY

The project methodology is described below.

3.1 Business Problem and Objectives

The cost of acquiring new customers is higher than retaining existing ones, sometimes vice versa. To extend business, customer retention is crucial. Marketers focus on maximizing the impact of customized plans for targeted customers. After understanding the business, there is an initial data mining plan designed to collect data to achieve a goal. This paper focuses on customer segmentation on a UK-based online retail company's dataset using RFM analysis and K-means clustering. The dataset as shown in Table 1 contains information of over 540,000 transactions made by over 20,000 customers between December 2010 and December 2011.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Table 1. Sample Transaction Dataset

3.2 Data Understanding

After collecting the data, it is essential to get familiar with what the data actually represents. The dataset included 8 attributes which are listed below in Table 2.

Attribute	Description
InvoiceNo	6-digit auto generated number, assigned to each transaction. If code starts with 'c', it's a cancellation. **Nominal**
StockCode	5-digit number assigned to each product. **Nominal**
Description	Name of product. **Nominal**
Quantity	Number of products per transaction. **Numeric**
InvoiceDate	Date and time of transaction. **Numeric**
UnitPrice	Product price per unit in sterling. **Numeric**
CustomerID	5-digit number assigned to each customer. **Nominal**
Country	Name of country where customer resides. **Nominal**

Table 2. Data Attributes

3.3 Data Cleaning and Preprocessing

Data preparation is a crucial step in data analysis that helps to ensure the accuracy, quality, and usefulness of the data. These steps must be performed carefully and systematically to obtain clean data that is suitable for analysis.

Handle missing values - Null or missing values can impact data analysis accuracy and quality. Null values can be removed using dropna or filled with suitable values.

Remove duplicate values - Duplicate values can impact data analysis quality. Identifying and removing duplicates results in clean data.

Transform data into understandable format - Real-world data can be inconsistent or contain errors [5]. Preprocessing involves transforming data into an understandable format by removing null, missing attributes, duplicates, incorrect data, and outlier values as shown in Fig 1.

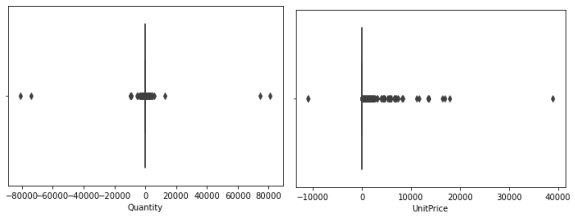

Figure 1. Negative Quantity and Unit Price could be observed in the box plot.

Add new features or columns - In cases where data is not in the desired format for analysis, new columns or features can be added. For instance, a new column named 'TotalSpent' was added to the dataset that represents the total price of a product.

Normalize data - Normalizing data involves transforming numeric column values to a common scale to avoid biases and ensure that all variables have equal importance during analysis [5].

Visualize data - As shown in Fig 2, Visualization of data using graphs and charts provides insights into patterns, trends, and outliers. This can aid in data analysis and decision-making.

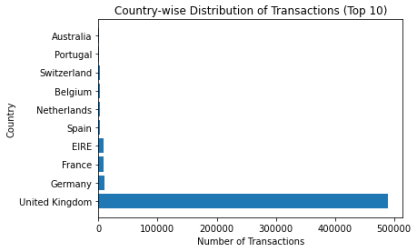

Figure 2. UK customers accounted for more than 90% of the transactions

3.4 RFM Analysis

RFM analysis was used to identify the best, or ideal, customers. RFM analysis is a method for analyzing customer value based on three factors: recency (how recently a customer made a purchase), frequency (how often a customer makes purchases), and monetary value (how much a customer spends) [11]. The analysis produced RFM scores for each customer as shown in Table 3, which were then labeled based on their score as shown in Table 4.

CustomerID	Recency	Frequency	MonetaryValue	R	F	M	RFM_Score	RFM_Label
12346.0	325	1	77183.60	1	1	4	6	Promising
12747.0	1	103	4196.01	4	4	4	12	Champions
12748.0	0	4413	33053.19	4	4	4	12	Champions
12749.0	3	199	4090.88	4	4	4	12	Champions
12820.0	2	59	942.34	4	3	3	10	Champions

Table 3. RFM Score Table for each Customer with their respective RFM Labels

RFM_Label	Recency mean	Frequency mean	MonetaryValue mean	MonetaryValue count
Champions	25.7	185.5	3919.2	1525
Loyal Customers	59.8	54.3	903.9	409
Potential Loyalists	77.9	38.3	731.9	390
Promising	96.4	27.9	820.6	428
New Customers	151.9	20.6	355.2	467
Require Attention	175.1	13.7	233.3	360
Need Activation	257.9	8.0	151.8	342

Table 4. RFM Labels & Counts based on Average Recency, Frequency, and Monetary Values.

3.5 K-means Clustering

K-means is efficient and easy to implement for large datasets, making it a popular choice for

clustering. K-means clustering is a popular unsupervised machine learning technique used for clustering data points into groups to get the right number of clusters based on their similarity [8]. Data needs to be preprocessed for K-means Clustering to segment customers into groups based on their RFM scores.

Preprocessing - The data should have symmetrical distribution of variables, similar mean values of variables, and similar standard deviation values of variables. As seen in Fig. 3, all the variables do not have a symmetrical distribution.

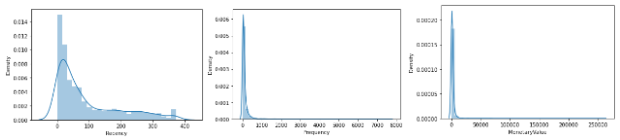

Figure 3. Distribution of Recency, Frequency and Monetary Value Metrics

Log transformation is a data transformation technique that involves taking the logarithm of each value in a dataset. It is used to reduce the variability and compress the range of data values, especially when the data has a skewed distribution or contains extreme values (outliers) [12]. To remove the skewness, log transformation is used because the dataset does not have any negative values since we are dealing with customer transactions Fig.4.

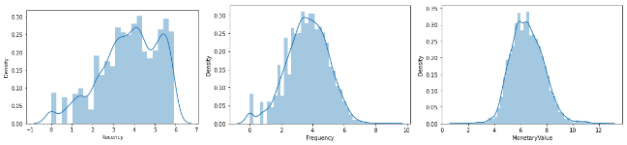

Figure 4. Distribution of Recency, Frequency and Monetary Value Metrics on Preprocessed data

Clustering - The algorithm iteratively assigns each data point to its closest cluster based on the Euclidean distance between the data point and the centroid of the cluster. The centroid of a cluster is the arithmetic mean of all the data points in the cluster. The algorithm continues until the data points are assigned to their respective clusters and the centroids no longer change.

The following steps are involved in during Clustering:

1. Determine the number of clusters (k) to be formed.
2. Initialize k cluster centroids randomly.
3. Assign each data point to the closest centroid.
4. Calculate the new centroid of each cluster.
5. Repeat steps 3 and 4 until convergence is achieved, that is centroids no longer move significantly.

The optimal number of clusters was determined using the elbow method and silhouette score.

Elbow Curve - In an elbow graph, we can determine the optimal number of clusters, k, by identifying the point of inflection, the elbow, where increasing the number of clusters no longer significantly reduces the sum of squared distances between data points and their assigned cluster centroids [13]. From Fig 5. k = 4 can be observed.

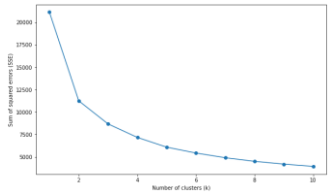

Figure 5. Elbow Plot Showing Optimal Number of Clusters for K-Means Clustering

Silhouette Analysis - The silhouette score measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters [10].

Thus, the silhouette score can be used to evaluate the quality of clustering for different values of k. Generally, a higher silhouette score indicates better clustering and a value closer to 1 indicates that the clusters are well-separated.

```
For k = 2, the average silhouette score is 0.394
For k = 3, the average silhouette score is 0.306
For k = 4, the average silhouette score is 0.307
For k = 5, the average silhouette score is 0.283
For k = 6, the average silhouette score is 0.281
For k = 7, the average silhouette score is 0.267
For k = 8, the average silhouette score is 0.266
For k = 9, the average silhouette score is 0.264
For k = 10, the average silhouette score is 0.258
```

Figure 6. Silhouette Scores for K-Means Clustering with k Ranging from 2 to 10

4. RESULTS

After applying the K-means clustering algorithm on the RFM model-based transactional dataset of a UK based online retail store, we identified four distinct clusters, as shown in Table 5.

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
Cluster				
0	200.0	16.0	308.0	1193
1	79.0	90.0	1596.0	1212
2	25.0	29.0	459.0	782
3	9.0	271.0	6310.0	733

Table 5. Cluster Analysis Results: Mean Recency, Frequency, and Monetary Value with Total Count per Cluster

The results of the analysis are presented in the form of visualizations, such as scatter plots and snake plots, which enabled the identification of customer segments with different purchasing patterns and preferences. From the scatterplots below, Fig 8, we observe:

- The customers in orange group, cluster 3, are the ones who like to spend more, and they are recent customers. They are more frequent customers.
- The customers in green group, cluster 2, are the ones who shopped recently. They are less frequent and spend less money.
- The customers in yellow group, cluster 1, are the customers who haven't shopped recently. They were frequent and spent quite a lot.
- The customers in the blue group, cluster 0, are the customers who haven't shopped recently. They are less frequent and spend the least amount of money.

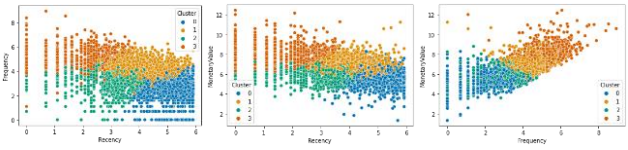

Figure 8. Scatterplots to Visualize each Cluster.

The snake plot in Fig. 9 displays the distribution of recency, frequency, and monetary values across the four clusters. It appears that the four clusters are distinct from each other, suggesting a favorable heterogeneous mix of clusters.

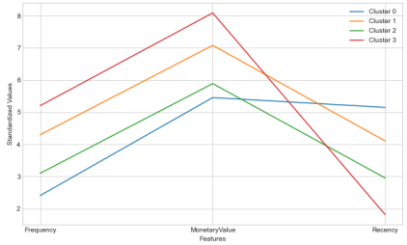

Figure 9. Snake Plot for 4 Clusters

Plotting Recency, Frequency, and Monetary Value on a 3D plot with clusters number as the color of the plot provides a visual representation of the customer segmentation results. It allows us to see the distribution of customers across the three dimensions, as well as the separation of the clusters. This visualization can help in understanding the characteristics of each cluster and identifying patterns that might not be evident from a tabular representation of the data.

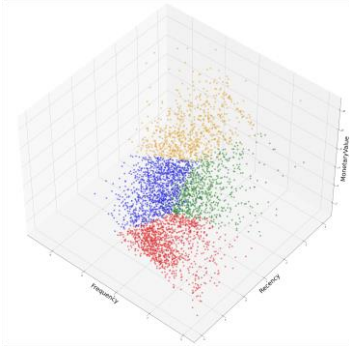

Figure 10. Visualization of Segmented Customers in 3D Space

Insights from the analysis were used to create strategies, shown in Table 6, that could be used to improve business performance, such as targeted marketing campaigns, personalized offers, and product recommendations.

Cluster	Customer Type	RFM Interpretation	Recommended Action
0	Lost Customer	No order placed recently, low frequency, and least monetary spending.	Already lost them. Try to understand why they left. Enhance product's quality.
1	At risk of leaving	Last transaction was a while ago. Were frequent and heavy spenders.	Need attention urgently. Figure out the reasons they're leaving. Discounts. Customized marketing.
2	New Customer	Transacted recently. Low frequency and monetary spending.	Handle with care. Enhance their purchasing experience. Good product and customer care service.
3	Best Customer	Heavy spenders. Frequent and recent shoppers.	Potential targets for new products. Discounts not required.

Table 6. Clusters Analysis Results and Corresponding Business Recommendations

5. CONCLUSION

This paper highlights the importance of customer segmentation in understanding customer behavior and developing effective marketing strategies. By utilizing the RFM model and K-means clustering, we were able to group customers into four distinct clusters based on their recency, frequency, and monetary value. This segmentation enabled us to identify high-value and low-value customer segments and those that required targeted marketing efforts. Our findings emphasize the need for businesses to have a detailed understanding of their customers and use customer segmentation as a tool to gain insights into their behaviors and preferences. By doing so, businesses can better cater to the needs and wants of their target market, thereby increasing customer satisfaction and loyalty. Overall, customer segmentation using RFM model and K-means clustering is a simple yet effective technique for businesses to optimize their marketing strategies and enhance their bottom line.

6. FUTURE WORK

Further research can be done to compare the performance of different clustering algorithms on the same data set and explore the potential of other customer segmentation models. Investigating the impact of external factors such as seasonality, promotions, and discounts on customer behavior and segmentation results would be interesting. The insights gained from customer segmentation can be used by businesses to personalize their marketing campaigns and offer better customer experiences. Incorporating machine learning algorithms and big data analysis into customer segmentation can provide more accurate and detailed insights into customer behavior, leading to better business decisions and increased profitability. To make the customer segmentation process more accessible to e-commerce startups and analysts, a user-friendly web model can be created that incorporates the RFM model and K-means clustering algorithm. The web model could provide valuable insights to businesses looking to improve their marketing strategies and customer retention efforts and could be continuously updated with new data sources and features for accuracy and relevance.

REFERENCES

[1] Khajvand, M. and Tarokh, M.J. (2011) "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," *Procedia Computer Science*, 3, pp. 1327–1332. Available at: https://doi.org/10.1016/j.procs.2011.01.011.

[2] Khajvand, M. *et al.* (2011) "Estimating customer lifetime value based on RFM analysis of Customer Purchase Behavior: Case Study," *Procedia Computer Science*, 3, pp. 57–63. Available at: https://doi.org/10.1016/j.procs.2010.12.011.

[3] LeBlanc, D. (2019) *Creating actionable customer segmentation models | google cloud blog, Google*. Google. Available at: https://cloud.google.com/blog/products/data-analytics/creating-actionable-customer-segmentation-models (Accessed: January 22, 2023).

[4] Hu, Y.-H. and Yeh, T.-W. (2014) "Discovering valuable frequent patterns based on RFM analysis without customer identification information," *Knowledge-Based Systems*, 61, pp. 76–88. Available at: https://doi.org/10.1016/j.knosys.2014.02.009.

[5] Optimove Inc. (2023) *RFM segmentation, Analysis & Model Marketing, Optimove*. Available at: https://www.optimove.com/resources/learning-center/rfm-segmentation (Accessed: March 12, 2023).

[6] Reddy, R. (2020) *Who's who: Understanding your business with customer segmentation, The Intercom Blog*. Available at: https://www.intercom.com/blog/customer-segmentation/ (Accessed: January 22, 2023).

[7] Leonard, J. (2022) *4 types of customer segmentation all marketers should know, Business 2 Community*. Available at: https://www.business2community.com/customer-experience/4-types-of-customer-segmentation-all-marketers-should-know-02120397 (Accessed: February 8, 2023).

[8] *K-means clustering* (2023) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/K-means_clustering (Accessed: March 18, 2023).

[9] *RFM (market research)* (2023) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/RFM_%28market_research%29 (Accessed: March 12, 2023).

[10] *Silhouette (clustering)* (2023) *Wikipedia*. Wikimedia Foundation. Available at:

https://en.wikipedia.org/wiki/Silhouette_%28clustering%29 (Accessed: March 18, 2023).
[11] Makhija, P. (2021) *RFM analysis for Customer Segmentation*, *CleverTap*. Available at: https://clevertap.com/blog/rfm-analysis/#:~:text=RFM%20is%20a%20data%2Ddriven,improves%20user%20engagement%20and%20retention (Accessed: January 19, 2023).
[12] Feng, C. *et al.* (2014) *Log-transformation and its implications for data analysis*, *Shanghai archives of psychiatry*. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/#:~:text=The%20log%20transformation%20is%2C%20arguably,normal%20or%20near%20normal%20distribution (Accessed: March 24, 2023).
[13] Saji, B. (2023) *Elbow method for finding the optimal number of clusters in K-means*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/ (Accessed: March 12, 2023).