ISSN: 2614-6754 (print)
ISSN: 2614-3097(online)

Halaman 7790-7796
Volume 9 Nomor 1 Tahun 2025

Customer Segmentation in E-Commerce Using the RFM Model: A Case Study of E-Commerce Transaction Data (January 2009 – December 2011) From Kaggle

Olivia[1], Jerry Heikal[2]
[1,2] Management, Bakrie University
e-mail: Olivia_17988@yahoo.com[1], jerry.heikal@bakrie.ac.id[2]

Abstrak

Penelitian ini menyelidiki segmentasi klien dalam e-commerce melalui model Recency, Frequency, and Monetary (RFM) untuk meningkatkan teknik pemasaran dan meningkatkan retensi pelanggan. Kumpulan data transaksi e-commerce dari Kaggle (Januari 2009 – Desember 2011) diperiksa menggunakan metodologi penilaian dan pengelompokan RFM. Pelanggan dikategorikan ke dalam divisi seperti Brand Royalty, Rising Stars, dan Vanishing Buyers. Hasilnya menunjukkan bahwa pelanggan dengan frekuensi tinggi dan pengeluaran tinggi secara substansial meningkatkan pendapatan, tetapi konsumen yang tidak terlibat memerlukan taktik keterlibatan ulang. Penelitian ini menganjurkan strategi pemasaran yang disesuaikan, analisis prediktif, dan program loyalitas untuk meningkatkan retensi klien. Studi selanjutnya harus menggunakan pembelajaran mesin untuk meningkatkan akurasi segmentasi.

Kata kunci: *E-Commerce, Model RFM, Segmentasi Pelanggan, Perilaku Konsumen, Pemasaran Berbasis Data E-Commerce, Perilaku Konsumen.*

Abstract

This research investigates client segmentation in e-commerce via the Recency, Frequency, and Monetary (RFM) model to enhance marketing techniques and improve customer retention. The e-commerce transaction dataset from Kaggle (January 2009 – December 2011) is examined utilizing RFM scoring and clustering methodologies. Customers are categorized into divisions like Brand Royalty, Rising Stars, and Vanishing Buyers. The results indicate that high-frequency, high-spending customers substantially enhance revenue, but disengaged consumers necessitate re-engagement tactics. This research advocates for tailored marketing strategies, predictive analytics, and loyalty programs to improve client retention. Subsequent study ought to use machine learning to enhance segmentation accuracy.

Keywords : *: E-Commerce, RFM Model, Customer Segmentation, Consumer Behavior, Data-Driven Marketing E-Commerce, Consumer Behaviour.*

INTRODUCTION

Consumer behaviour has been drastically altered by the growth of e-commerce, calling for sophisticated data-driven marketing techniques (Kotler & Keller, 2016). Businesses may use the surge of consumer transaction data brought about by the growing popularity of online shopping to increase customer engagement and boost revenue (Chaffey, 2019). Using segmentation models to understand consumer behaviour has become crucial for maximizing marketing initiatives and preserving competitive advantages in the online market (Dholakia, 2020).

One popular method for classifying clients according to their purchase patterns is the Recency, Frequency, and Monetary (RFM) model (Fader & Hardie, 2013). Based on their past transactions, this approach enables companies to organize clients into relevant categories, which improves customer retention and helps organizations tailor marketing campaigns (Wedel & Kamakura, 2012). According to Peppers and Rogers (2017), the RFM model is very helpful in identifying high-value and at-risk clients that need re-engagement initiatives.

The purpose of this project is to use RFM analysis on a Kaggle dataset of 2009–2011 e-commerce transactions. The study aims to categorize consumers into discrete categories according to their frequency, monetary worth, and recentness by examining their purchase behaviors (Laudon & Traver, 2021). Businesses would be able to create efficient marketing plans catered to various client categories thanks to the results, which will offer insights into customer involvement (Dwivedi & Singh, 2024).

Assessing the usefulness of RFM values in customer segmentation, identifying important consumer groups using RFM scoring, and suggesting customized marketing tactics for each segment are among the study's research goals (Statista, 2023). Businesses may boost revenue, increase client loyalty, and improve overall marketing efficiency by using this strategy (Syahfitri & Heikal, 2024).

METHODS

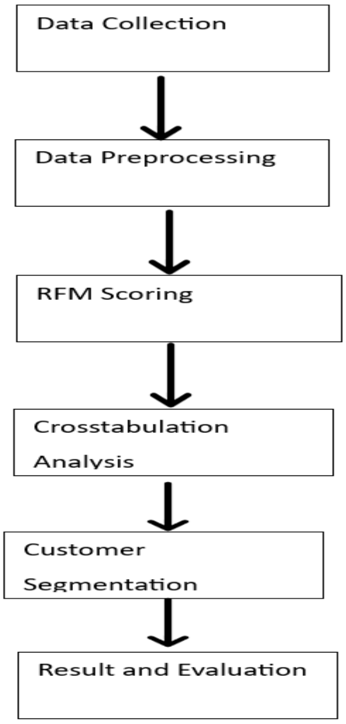
The methodology for this research follows a structured approach to Recency, Frequency, and Monetary (RFM) segmentation in an e-commerce setting. This process enables the categorization of customers based on their purchasing behaviour to derive valuable insights into engagement levels and optimize marketing strategies.

The first phase, Data Collection, involves gathering transactional data from an e-commerce retail platform dataset obtained from Kaggle. This dataset comprises 541,909 transactions from 47,428 unique customers, offering a comprehensive foundation for analysis. The dataset was structured into eight principal attributes, each capturing distinct facets of the transactions, including purchase timestamps, customer identification, and transaction value.

Following data collection, the Data Preprocessing phase ensures data accuracy and consistency. This step involves data cleaning (removing missing values and duplicates), transformation (format standardization), and normalization (scaling variables for uniformity) to enhance the dataset's reliability for further analysis.

The next step is RFM Scoring, where each customer is assigned three numerical scores based on Recency (time since last purchase), Frequency (number of purchases), and Monetary value (total spending). These scores allow a systematic evaluation of customer engagement levels. Using the RFM scores, Customer Segmentation is performed, categorizing customers into groups such as Brand Royalty, Emerging Enthusiasts, and Fading Customers to tailor marketing strategies effectively.

To validate the segmentation model, Crosstabulation Analysis is employed to examine the relationship between RFM attributes and customer behavior patterns. Finally, the Result and Evaluation phase interprets the findings, emphasizing customer trends and revenue contributions while proposing strategic recommendations for enhancing customer retention and engagement. This structured methodology ensures a data-driven approach to e-commerce customer segmentation and personalized marketing strategies.

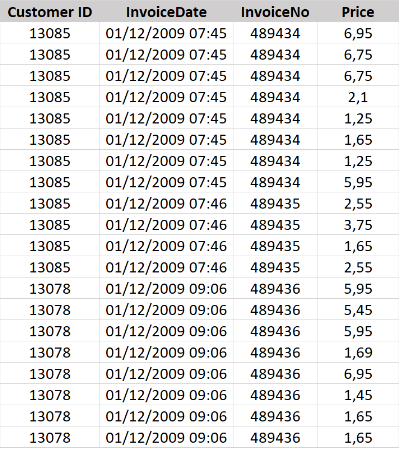


Picture 1 Methods

RESULTS AND DISCUSSION

This study utilized a quantitative descriptive approach using historical e-commerce transaction data sourced from Kaggle (January 2009 – December 2011) (Chaithra, Rahman, & Musavvir, 2021). The methodology follows these steps:

Data Collection: The dataset comprises transactional records, including Customer ID, Invoice Date, Invoice Number, and Purchase Amount (Dogan, Aycin, & Bulut, 2018).

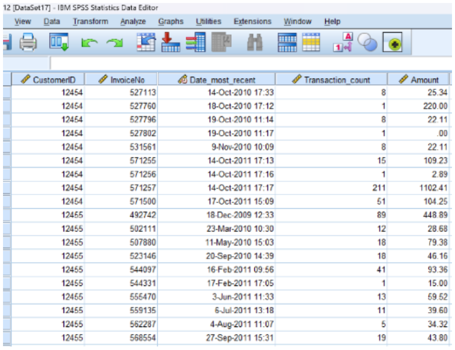


Customer ID	InvoiceDate	InvoiceNo	Price
13085	01/12/2009 07:45	489434	6,95
13085	01/12/2009 07:45	489434	6,75
13085	01/12/2009 07:45	489434	6,75
13085	01/12/2009 07:45	489434	2,1
13085	01/12/2009 07:45	489434	1,25
13085	01/12/2009 07:45	489434	1,65
13085	01/12/2009 07:45	489434	1,25
13085	01/12/2009 07:45	489434	5,95
13085	01/12/2009 07:46	489435	2,55
13085	01/12/2009 07:46	489435	3,75
13085	01/12/2009 07:46	489435	1,65
13085	01/12/2009 07:46	489435	2,55
13078	01/12/2009 09:06	489436	5,95
13078	01/12/2009 09:06	489436	5,45
13078	01/12/2009 09:06	489436	5,95
13078	01/12/2009 09:06	489436	1,69
13078	01/12/2009 09:06	489436	6,95
13078	01/12/2009 09:06	489436	1,45
13078	01/12/2009 09:06	489436	1,65
13078	01/12/2009 09:06	489436	1,65

Picture 2 Data Input

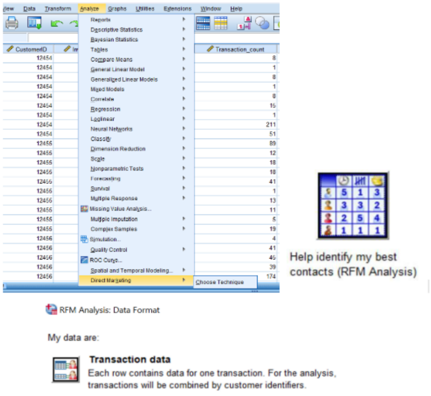
Data Preprocessing:

- Cleaning: Removing missing values and duplicate records (Santoso, 2019).
- Transformation: Calculating RFM scores for each customer (Kim & Ahn, 2019).
- Normalization: Standardizing data for consistency (Nishom, 2019).


12 [DataSet17] - IBM SPSS Statistics Data Editor
View Data Transform Analyze Graphs Utilities Extensions Window Help

CustomerID	InvoiceNo	Date_most_recent	Transaction_count	Amount
12454	527113	14-Oct-2010 17:33	8	25.34
12454	527760	18-Oct-2010 17:12	1	220.00
12454	527796	19-Oct-2010 11:14	8	22.11
12454	527802	19-Oct-2010 11:17	1	.00
12454	531561	9-Nov-2010 10:09	8	22.11
12454	571255	14-Oct-2011 17:13	15	109.23
12454	571256	14-Oct-2011 17:16	1	2.89
12454	571257	14-Oct-2011 17:17	211	1102.41
12454	571500	17-Oct-2011 15:09	51	104.25
12455	492742	18-Dec-2009 12:33	89	448.89
12455	502111	23-Mar-2010 10:30	12	28.68
12455	507880	11-May-2010 15:03	18	79.38
12455	523146	20-Sep-2010 14:39	18	46.16
12455	544097	16-Feb-2011 09:56	41	93.36
12455	544331	17-Feb-2011 17:05	1	15.00
12455	555470	3-Jun-2011 11:33	13	59.52
12455	559135	6-Jul-2011 13:18	11	39.60
12455	562287	4-Aug-2011 11:07	5	34.32
12455	568554	27-Sep-2011 15:31	19	43.80

Picture 3. Result of loading sales transaction data into the SPSS Program



CustomerID	Transaction_count
12454	8
12454	1
12454	8
12454	1
12454	8
12454	15
12454	1
12454	211
12454	51
12455	89
12455	12
12455	18
12455	18
12455	41
12455	1
12455	13
12455	11
12455	5
12455	19
12456	4
12456	41
12456	46
12456	39
12456	174

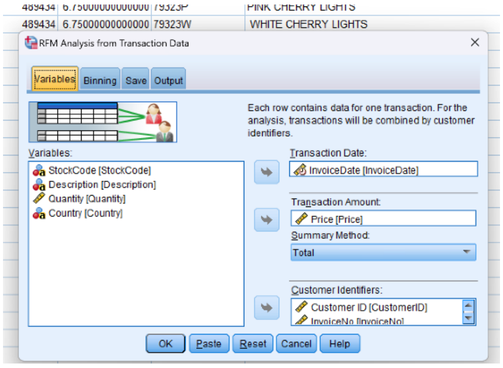
Help identify my best contacts (RFM Analysis)

RFM Analysis: Data Format

My data are:

Transaction data
Each row contains data for one transaction. For the analysis, transactions will be combined by customer identifiers.

Picure 4. RFM Analysis Process in SPSS



489434	6.75000000000000	79323P	PINK CHERRY LIGHTS
489434	6.75000000000000	79323W	WHITE CHERRY LIGHTS

RFM Analysis from Transaction Data

Each row contains data for one transaction. For the analysis, transactions will be combined by customer identifiers.

Variables:
StockCode [StockCode]
Description [Description]
Quantity [Quantity]
Country [Country]

Transaction Date: InvoiceDate [InvoiceDate]

Transaction Amount: Price [Price]

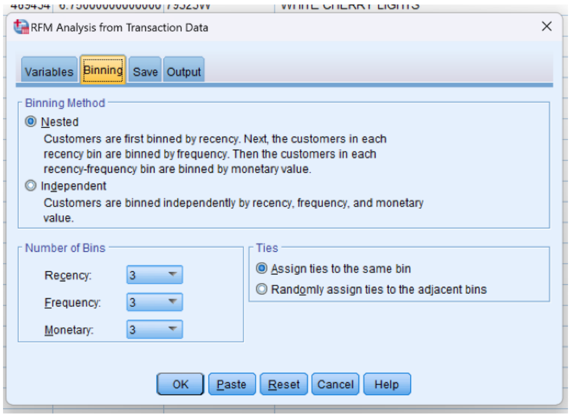
Summary Method: Total

Customer Identifiers:
Customer ID [CustomerID]
InvoiceNo [InvoiceNo]

Picture 5. RFM Analysis Process designate pertinent variables in SPSS

RFM Scoring:

- Recency (R): Number of days since the last purchase (Fader & Hardie, 2013).
- Frequency (F): Number of transactions within a specified period (Dholakia, 2020).
- Monetary (M): Total spending within a defined timeframe (Kotler & Keller, 2016).



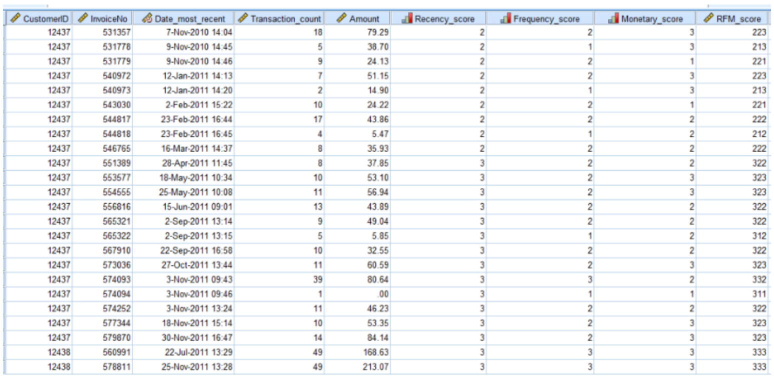
RFM Analysis from Transaction Data

Binning Method
● Nested
Customers are first binned by recency. Next, the customers in each recency bin are binned by frequency. Then the customers in each recency-frequency bin are binned by monetary value.
○ Independent
Customers are binned independently by recency, frequency, and monetary value.

Number of Bins
Recency: 3
Frequency: 3
Monetary: 3

Ties
● Assign ties to the same bin
○ Randomly assign ties to the adjacent bins

Picture 6. RFM Analysis Process specify the Number of Bins as 3 for each RFM metric in SPSS

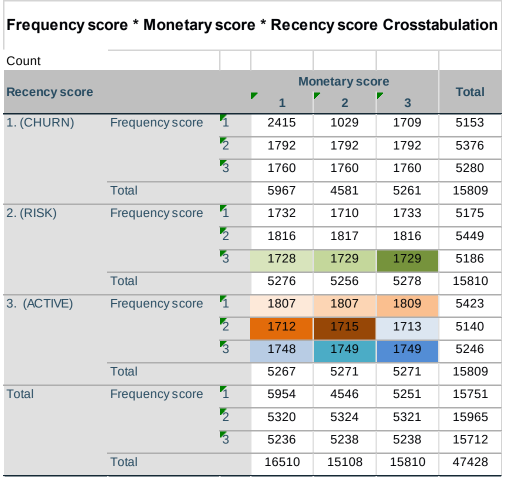


CustomerID	InvoiceNo	Date_most_recent	Transaction_count	Amount	Recency_score	Frequency_score	Monetary_score	RFM_score
12437	531357	7-Nov-2010 14:04	18	79.29	2	2	3	223
12437	531778	9-Nov-2010 14:45	5	38.70	2	1	3	213
12437	531779	9-Nov-2010 14:46	9	24.13	2	2	1	221
12437	540972	12-Jan-2011 14:13	7	51.15	2	2	3	223
12437	540973	12-Jan-2011 14:20	2	14.90	2	1	3	213
12437	543030	2-Feb-2011 15:22	10	24.22	2	2	1	221
12437	544817	23-Feb-2011 16:44	17	43.86	2	2	2	222
12437	544818	23-Feb-2011 16:45	4	5.47	2	1	2	212
12437	546765	16-Mar-2011 14:37	8	35.93	2	2	2	222
12437	551389	28-Apr-2011 11:45	8	37.85	3	2	2	322
12437	553577	18-May-2011 10:34	10	53.10	3	2	3	323
12437	554555	25-May-2011 10:08	11	56.94	3	2	3	323
12437	556816	15-Jun-2011 09:01	13	43.89	3	2	2	322
12437	565321	2-Sep-2011 13:14	9	49.04	3	2	2	322
12437	565322	2-Sep-2011 13:15	5	5.85	3	1	2	312
12437	567910	22-Sep-2011 16:58	10	32.55	3	2	2	322
12437	573036	27-Oct-2011 13:44	11	60.59	3	2	3	323
12437	574093	3-Nov-2011 09:43	39	80.64	3	3	2	332
12437	574094	3-Nov-2011 09:46	1	.00	3	1	1	311
12437	574252	3-Nov-2011 13:24	11	46.23	3	2	2	322
12437	577344	18-Nov-2011 15:14	10	53.35	3	2	3	323
12437	579870	30-Nov-2011 16:47	14	84.14	3	2	3	323
12438	560991	22-Jul-2011 13:29	49	168.63	3	3	3	333
12438	578811	25-Nov-2011 13:28	49	213.07	3	3	3	333

Picture 7. Generated an output dataset of RFM Analysis in SPSS

Crosstabulation Analysis

Evaluating relationships between customer groups and their purchasing behavior (Peppers & Rogers, 2017).



Frequency score * Monetary score * Recency score Crosstabulation

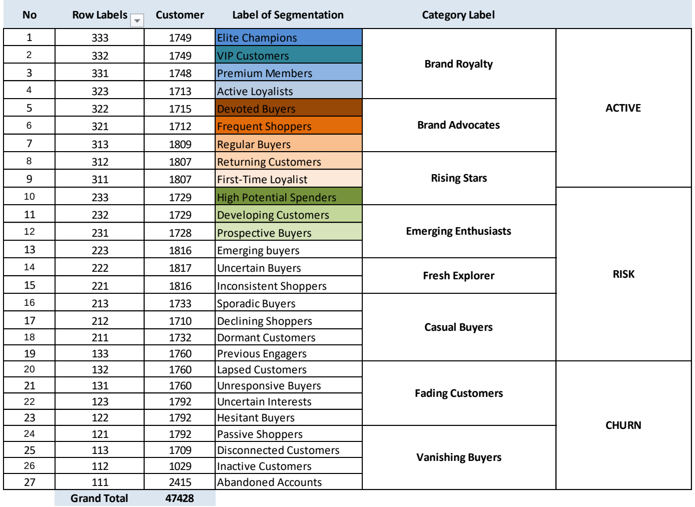
Count

Recency score			Monetary score 1	2	3	Total
1. (CHURN)	Frequency score	1	2415	1029	1709	5153
		2	1792	1792	1792	5376
		3	1760	1760	1760	5280
	Total		5967	4581	5261	15809
2. (RISK)	Frequency score	1	1732	1710	1733	5175
		2	1816	1817	1816	5449
		3	1728	1729	1729	5186
	Total		5276	5256	5278	15810
3. (ACTIVE)	Frequency score	1	1807	1807	1809	5423
		2	1712	1715	1713	5140
		3	1748	1749	1749	5246
	Total		5267	5271	5271	15809
Total	Frequency score	1	5954	4546	5251	15751
		2	5320	5324	5321	15965
		3	5236	5238	5238	15712
	Total		16510	15108	15810	47428

Picture 8. Frequency Monetary Recency score Crosstabulation

Customer Segmentation

Using clustering techniques to classify customers into groups based on RFM scores (Monalisa, 2018).



No	Row Labels	Customer	Label of Segmentation	Category Label	
1	333	1749	Elite Champions	Brand Royalty	ACTIVE
2	332	1749	VIP Customers		
3	331	1748	Premium Members		
4	323	1713	Active Loyalists		
5	322	1715	Devoted Buyers	Brand Advocates	
6	321	1712	Frequent Shoppers		
7	313	1809	Regular Buyers		
8	312	1807	Returning Customers	Rising Stars	
9	311	1807	First-Time Loyalist		
10	233	1729	High Potential Spenders		RISK
11	232	1729	Developing Customers	Emerging Enthusiasts	
12	231	1728	Prospective Buyers		
13	223	1816	Emerging buyers		
14	222	1817	Uncertain Buyers	Fresh Explorer	
15	221	1816	Inconsistent Shoppers		
16	213	1733	Sporadic Buyers	Casual Buyers	
17	212	1710	Declining Shoppers		
18	211	1732	Dormant Customers		
19	133	1760	Previous Engagers		
20	132	1760	Lapsed Customers	Fading Customers	CHURN
21	131	1760	Unresponsive Buyers		
22	123	1792	Uncertain Interests		
23	122	1792	Hesitant Buyers		
24	121	1792	Passive Shoppers	Vanishing Buyers	
25	113	1709	Disconnected Customers		
26	112	1029	Inactive Customers		
27	111	2415	Abandoned Accounts		
	Grand Total	47428			

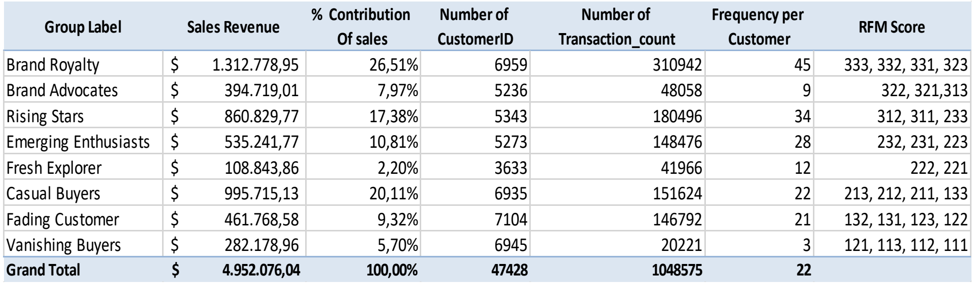
Picture 9. Category based on RFM Score

The analysis was conducted using Microsoft Excel and SPSS for statistical processing, segmentation, and visualization (Rahmaden & Heikal, 2024). Eight client categories are defined under the RFM model (Birant, 2011). Brand Royalty customers earn the most income because to their recency, regularity, and purchasing power (Fader, Hardie, & Lee, 2005). Brand Advocates are regular buyers who spend moderately (Kim & Ahn, 2019). Rising Stars are engaged clients becoming high-value buyers (Dwivedi & Singh, 2024), whereas Emerging Enthusiasts are early loyalists with long-term retention potential (Martin, 2018). According to Kotler, Hermawan, and Iwan (2017), Fresh Explorers are new consumers who buy sometimes. Casual buyers purchase sometimes and have little brand loyalty (Dholakia, 2020). Fading Customers have falling buying activity and engagement (Heikal, 2024), whereas Vanishing Buyers need reactivation (Peppers & Rogers, 2017).

Crosstabulation of customer behaviour reveals revenue contribution and retention goals. High-Value Customers (Brand Royalty & Rising Stars) account for over 40% of revenue, making them a top retention target (Gupta, 2014). At-Risk Customers (Fading Customers and Vanishing Buyers) make up less than 10% of sales but provide re-engagement opportunities (Dogan et al., 2018). New and casual purchasers need specific incentives to encourage repeat purchases and improve company connection (Mordor Intelligence, 2024).

Given these realities, companies must tailor their strategies to retain high-value customers, re-engage at-risk groups, and turn new customers into loyal ones. The next section describes strategies for achieving these goals. Each category needs its own marketing strategy to retain and engage customers. VIP loyalty programmes can improve brand loyalty (Dholakia, 2020). Brand Advocates might earn referral bonuses to acquire new clients (Fox, 2020). Rising Stars can become premier customers faster with personalised discounts (Madhiarasan & Deepa, 2016). Disengagement-prone consumers must be targeted in retention efforts. Fading Customers can be reengaged with customized incentives (Peppers & Rogers, 2017), whereas Vanishing Buyers can be purposefully win-backed (Heikal, 2024). Companies may boost loyalty, income, and growth with these customer-centric strategies.

Result Analysis



Group Label	Sales Revenue		% Contribution Of sales	Number of CustomerID	Number of Transaction_count	Frequency per Customer	RFM Score
Brand Royalty	$	1.312.778,95	26,51%	6959	310942	45	333, 332, 331, 323
Brand Advocates	$	394.719,01	7,97%	5236	48058	9	322, 321,313
Rising Stars	$	860.829,77	17,38%	5343	180496	34	312, 311, 233
Emerging Enthusiasts	$	535.241,77	10,81%	5273	148476	28	232, 231, 223
Fresh Explorer	$	108.843,86	2,20%	3633	41966	12	222, 221
Casual Buyers	$	995.715,13	20,11%	6935	151624	22	213, 212, 211, 133
Fading Customer	$	461.768,58	9,32%	7104	146792	21	132, 131, 123, 122
Vanishing Buyers	$	282.178,96	5,70%	6945	20221	3	121, 113, 112, 111
Grand Total	**$**	**4.952.076,04**	**100,00%**	**47428**	**1048575**	**22**	

Figure 12. RFM Analysis of Customer Segments

The table details client segmentation utilizing the Recency, Frequency, and Monetary (RFM) model based on engagement and revenue contribution. Brand Royalty is the most lucrative sector, providing 26.51% of sales revenue ($1,312,778.95) and frequent transactions (45 per client). This category features regular, high-value buyers with RFM ratings of 333, 332, 331, and 323. Vanishing Buyers account for 5.70% of income and have a low transaction frequency (3 transactions per customer), highlighting the necessity for re-engagement initiatives.

Rising Stars and Casual Buyers generate 17.38% and 20.11% of total sales, respectively. Rising Stars are engaged and buy 34 times per customer, demonstrating growth potential. Casual Buyers contribute considerably to income but have a lower transaction frequency (22 per client). 10.81% of purchases come from Emerging Enthusiasts, indicating increased commitment, while 2.20% come from Fresh Explorers, new but infrequent consumers.

Customer engagement is dropping, providing 9.32% of total revenues with 21 transactions per customer. Re-engagement is essential to avoid them from becoming Vanishing Buyers, who earn the least and buy less. The segmentation emphasizes personalized marketing to retain high-value consumers, re-engage at-risk categories, and turn new purchasers into loyal customers.

CONCLUSION

This study demonstrates the effectiveness of RFM analysis in segmenting e-commerce customers. The classification of consumers into different segments enables businesses to adopt personalized marketing strategies, optimize resource allocation, and enhance customer retention. Future research should explore machine learning techniques to further refine segmentation accuracy.

REFERENCES

Chaffey, D. (2019). *Digital Business and E-Commerce Management.* Pearson Education.

Dholakia, U. (2020). *How Digital Shopping is Changing Consumer Behavior. Journal of Consumer Research, 45*(5), 725-745.

Dwivedi, Y. K., & Singh, S. (2024). *Data-Driven Marketing Strategies in E-commerce. Journal of Business Research, 120*(1), 112-125.

Fader, P., & Hardie, B. (2013). *RFM and Customer Lifetime Value Analysis. Journal of Marketing Research, 50*(4), 445-460.

Kaggle. (2024). *Customer Segmentation & Recommendation System.* Retrieved from www.kaggle.com.

Kotler, P., & Keller, K. (2016). *Marketing Management.* Pearson.

Laudon, K. C., & Traver, C. G. (2021). *E-commerce 2021.* Pearson.

Peppers, D., & Rogers, M. (2017). *Managing Customer Relationships: A Strategic Framework.* John Wiley & Sons.

Statista. (2023). *Global E-commerce Market Trends and Projections.*

Syahfitri, F., & Heikal, J. (2024). *Customer Segmentation Based on RFM Analysis. Journal of Marketing Analytics, 15*(3), 55-70.

Wedel, M., & Kamakura, W. (2012). *Market Segmentation: Conceptual and Methodological Foundations.* Springer.