IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation And Churn Analysis Using Rfm Analysis And Machine Learning Technique

SOUNDHARYA S S, DEEPAN RAJ R

STUDENT, STUDENT

KPR INSTITUTE OF ENGINEERING AND TECHNOLOGY

COIMBATORE, TAMIL NADU, INDIA

ABSTRACT

Understanding customer behavior is increasingly important for retaining customers and increasing profitability in the intensely competitive environment of e-commerce. This study aims to design a hybrid analytical model that combines the RFM analysis technique with machine learning algorithms in the processes of customer segmentation and churn analysis. The dataset used in this study was sourced from an e-commerce company in the UK and consists of real transactional data with detailed invoice information, several purchase quantities, and customer IDs. Following the cleaning process of this data, RFM values were calculated for each customer based on their purchasing behavior, while algorithms such as K-Means were applied to segment the customers into meaningful groups. Analyzing these clusters further will identify high-value, loyal, and at-risk customers, offering very useful information to help focus effective marketing strategies. Furthermore, the analytical framework identifies churn-prone customers based on transaction inactivity patterns about recent transaction habits. Thus, this paper shows that the integration of RFM metrics with machine learning algorithm analyses allows for a robust and data-driven analytical approach toward improving customer segmentation, churn analysis, and effective marketing strategy formulation for e-commerce businesses.

Keywords: Customer Segmentation, RFM Analysis, Machine Learning, K-Means Clustering, Customer Churn Analysis, Data Mining, E-Commerce Analytics, Customer Retention, Analytical Framework

1. INTRODUCTION

Customer-centric strategies are the bedrock upon which the success of any business in today's fast-changing digital marketplace stands or falls. With the tremendous rise in the number of e-commerce platforms, a company deals with a huge amount of customer transaction data every day. All this data carries so much valuable information regarding customer behaviors, preferences, and trends. But the actual challenge is how one can analyze such data effectively and draw meaningful insights from it to establish long-term relationships with customers and to enhance business performance. Customer segmentation and churn analysis have been some of the successful ways to do this.

Customer segmentation is the division of a company's customer base into distinct groups or clusters of people having similar characteristics or behavior. These groups allow businesses to tailor their marketing strategies, personalize offers, and enhance customer satisfaction. Among many techniques of segmentation, the RFM (Recency, Frequency, and Monetary) model remains one of the most established and practical frameworks. It helps identify how recently customers have made purchases, how often they buy, and how much revenue they generate-three crucial metrics that define customer value and loyalty.

In parallel, customer churn analysis also plays an indispensable part in CRM. It deals with the identification of customers who are most likely to stop buying from or otherwise disengage from a business. Early detection of churners allows businesses to take various retention strategies, such as offering personalized offers or loyalty programs, to prevent customer loss and sustain revenue. Integrated with machine learning techniques, churn analysis becomes even more powerful while algorithms recognize hidden patterns in data that traditional methods might overlook.

ML and data mining in customer analytics have revolutionized the decision-making process for e-commerce companies. The various algorithms included are K-Means clustering, which automatically segments customers on the basis of RFM scores, such as loyal customers, probable buyers, and at-risk customers. This hybrid RFM-ML approach simplifies segmentation by providing predictive insights so that businesses can act proactively, rather than reactively. It is important because this project will help convert raw e-commerce data into actionable intelligence. The study will apply RFM analysis and machine learning algorithms to a UK-based e-commerce dataset, aiming to identify meaningful customer segments and analyze churn-prone customers. Such insights shall then help companies in resource allocation for more tailored communications and design of marketing campaigns that best suit each segment. Eventually, this project will showcase how various data-driven approaches can help enhance customer retention, ensure profitability, and support strategic decisions within the competitive e-commerce space.

2. REVIEW OF LITERATURE

A study on Migros Turk, an innovative multinational firm, demonstrated how effective customer segmentation can guide businesses in developing targeted marketing strategies and new product offerings. The research emphasized both a priori and post-hoc segmentation methods as key approaches for deriving valuable insights. Various analytical and statistical models, including logistic regression, cluster analysis, and CHAID (Chi-squared Automatic Interaction Detection), were explored for supervised and unsupervised segmentation. Additionally, latent class models were used to handle complex data structures involving multiple dependent measures. The study highlighted that combining these analytical approaches enables firms to gain a deeper understanding of customer diversity, allowing for the creation of data-driven strategies that enhance business performance and customer engagement (Bruce Cooil et al., 2008).

With the rapid growth of e-commerce, businesses face challenges of information overload, where customers receive excessive product information and become confused during decision-making. To address this, personalization and customer segmentation techniques are essential in identifying potential customers and improving marketing efficiency. The study reviewed various segmentation methods using both internal data (customer profiles and purchase history) and external data (server logs, cookies, and surveys). Different analytical approaches, including RFM cell classification, supervised and unsupervised clustering, and purchase affinity clustering, were discussed and categorized into simple, RFM-based, target, and unsupervised techniques. The research emphasized that analyzing customer behavior—such as time spent viewing products—can serve as a strong indicator of interest, enabling more accurate segmentation and effective personalized marketing strategies (Sari et al., 2016).

As marketing paradigms evolve, Customer Relationship Management (CRM) has become a central strategy focused on building long-term and profitable relationships with customers. The study proposed a framework for customer value analysis and segmentation, emphasizing that understanding the true value of each customer is crucial for designing effective marketing strategies. By segmenting customers based on their value, businesses can direct resources toward high-profit segments and foster customer loyalty. Through a case study on a wireless telecommunication company, the research demonstrated how customer value–based segmentation supports the development of refined, data-driven marketing strategies that enhance both customer retention and corporate success (Su-Yeon Kim et al., 2006).

In today's highly competitive and innovation-driven business environment, companies strive to identify and target the right customer segments to enhance sales and customer satisfaction. The study highlighted the role of machine learning (ML) in uncovering hidden patterns in customer data to support informed decision-making. It applied three clustering algorithms—K-Means, Agglomerative, and Mean Shift—to segment customers based on their purchasing behavior and visit frequency using data from a local retail shop. Through clustering, five key customer segments were initially identified—Careless, Careful, Standard, Target, and Sensible—while Mean Shift clustering further revealed two new high-value groups: High buyers and frequent visitors, and High buyers and occasional visitors. The study concluded that machine learning–based segmentation provides deeper insights into customer diversity, enabling more accurate targeting and effective marketing strategies (Tushar Kansal et al., 2018).

The growing emphasis on customer-oriented marketing has shifted businesses away from traditional mass marketing toward personalized, data-driven approaches. With the rise of one-to-one marketing in e-commerce, understanding individual customer behavior has become essential. This study conducted a comprehensive review of 105 publications (2000–2022) on customer segmentation methods and identified a four-phase process comprising data collection, customer representation, segmentation analysis, and customer targeting. The findings revealed that RFM (Recency, Frequency, Monetary) analysis remains a key technique for customer representation. At the same time, K-Means clustering continues to be the most widely used method for segmentation across various data scales and contexts. The study concluded that segmentation plays a vital role in enabling businesses to develop targeted marketing strategies and enhance customer engagement in the evolving digital landscape (Miguel Alves Gomes et al., 2023).

Machine learning (ML) techniques have been widely applied in customer churn analysis to identify potential customer loss using various supervised and unsupervised approaches. Due to limited access to comprehensive customer data, many businesses rely primarily on transactional data from Enterprise Resource Planning (ERP) systems. To enhance the predictive power of such data, researchers have combined Recency, Frequency, and Monetary (RFM) analysis with ML algorithms. RFM analysis helps in understanding customer purchasing patterns and loyalty based on transaction history. Studies have shown that integrating clustering techniques,

such as K-Means and DBSCAN, with RFM scores provides meaningful customer segmentation and improves churn analysis accuracy. This hybrid approach offers a practical and efficient solution for businesses to identify and retain valuable customers using available transactional data (Israa Lewaa, 2023).

Data mining (DM) combined with the Recency, Frequency, and Monetary (RFM) model is one of the most effective approaches for customer segmentation and behavior analysis. Their study reviewed various DM techniques applied with RFM between 2015 and 2020 and found that clustering and visualization methods are the most commonly used for identifying meaningful customer groups. The integration of RFM with clustering algorithms enables businesses to uncover hidden patterns in transactional data and refine their marketing strategies. The study also proposed a new framework that combines RFM-based segmentation with Geographic Information Systems (GIS) to enhance customer understanding through geo-demographic analysis. This supports the current project's approach of using RFM and machine learning techniques like K-Means and DBSCAN for customer churn analysis, as it highlights the value of RFM-driven data mining in uncovering customer behavior and guiding data-based decision-making (Ernawati et al., 2021).

A case study conducted in Istanbul, Turkey, applied the RFM (Recency, Frequency, and Monetary) model to segment and profile customers based on their lifetime value using real data from a fuel station. The dataset included 1,015 customers with variables such as arrival frequency, last purchase date, total spending, and demographic information. Through cluster analysis performed in SPSS, customers were divided into five segments according to their RFM scores, and further analyzed using Correspondence and Discriminant analyses. The study revealed that, contrary to managerial assumptions, truck drivers were the most valuable customer group rather than automobile drivers. The authors concluded that RFM-based segmentation provides meaningful insights for customer profiling and targeted marketing, demonstrating its effectiveness in identifying high-value customer segments such as the "VIP" and "GOLD" groups (İbrahim SABUNCU et al., 2020).

Recent studies highlight the growing importance of customer segmentation in highly competitive markets such as retail, banking, and e-commerce. Businesses increasingly use data-driven models to understand customer behavior and enhance retention strategies. Ernawati et al. (2021) emphasized that combining data mining (DM) with the RFM (Recency, Frequency, Monetary) model enables effective behavioral segmentation and helps reveal hidden customer patterns. Similarly, a case study conducted in Turkey using the RFM model and cluster analysis demonstrated that customer profiling based on transactional and demographic data can accurately identify high-value segments, such as "VIP" and "GOLD" customers. In line with these findings, the present project applies K-Means clustering and RFM analysis on the UK E-commerce dataset to categorize customers based on their purchasing behavior. This approach supports e-commerce businesses in identifying loyal, moderate, and at-risk customers, thereby improving targeted marketing and customer relationship management (Rahul Shirole et al., 2021)

Customer segmentation plays a crucial role in helping companies understand customer behavior, detect patterns, and design effective marketing strategies. The study emphasized the use of RFM (Recency, Frequency, and Monetary) analysis as a long-established method for identifying valuable customers and determining who requires promotional attention. By applying data mining techniques, particularly clustering, the researchers segmented over 700,000 customers based on their RFM values and found that segmentation based solely on customer expenditure was insufficient. The proposed clustering models offered deeper insights into customer behavior, enabling better decision-making, improved targeting, and more effective marketing strategies. This supports the current project's approach of combining RFM analysis and K-Means clustering to achieve meaningful and actionable customer segmentation (Doğan O. et al., 2018)

2.1 Research Gap

Although many studies have applied RFM models and clustering techniques for customer segmentation, most existing research focuses on limited datasets, specific industries, or static segmentation without predictive capabilities. Traditional approaches, such as logistic regression and basic clustering, lack integration with advanced machine learning models for deeper behavioral insights. Recent works highlight the potential of combining RFM with data mining and ML techniques, yet few have effectively applied this hybrid approach to large-scale e-commerce data. Therefore, there remains a gap in developing a comprehensive, data-driven model that integrates RFM analysis with machine learning algorithms to improve customer segmentation and churn analysis using real-world transactional data.

2.2 Problem Statement

In this digital transformation era, every e-commerce business faces a critical challenge in understanding customer behavior and preventing churn in the ever-growing competitive environment. Traditional marketing and segmentation techniques are no longer adequate to capture modern consumer patterns. Most existing approaches either depend only on transactional or demographic data or do not integrate advanced analytics techniques for predictive insights. Consequently, businesses fail to identify their high-value customers, predict potential churners, and suggest an effective retention strategy. There is an urgent need for a data-driven hybrid model that effectively integrates RFM analysis with machine learning algorithms to perform accurate customer segmentation and churn analysis. This will enable organizations to develop better customer engagement, personalize marketing efforts, and secure long-term profitability.

3. RESEARCH OBJECTIVES

- To preprocess and analyze e-commerce transactional data for deriving meaningful customer insights.
- To calculate Recency, Frequency, and Monetary (RFM) scores for evaluating customer purchasing behavior.
- To implement machine learning clustering algorithms such as K-Means for effective customer segmentation.
- To identify churn-prone customers based on transactional inactivity using RFM metrics and analytical techniques.
- To interpret the segmentation and churn results to support targeted marketing strategies and customer retention planning.

4. RESEARCH METHODOLOGY

4.1 Research Design

The present study adopts a **quantitative, analytical, and exploratory research design**. The research focuses on analyzing customer transactional data to perform customer segmentation and churn analysis using **RFM (Recency, Frequency, and Monetary)** metrics and statistical techniques. A quantitative approach is appropriate as the study relies on numerical data, statistical summaries, clustering techniques, and categorical analysis to derive objective insights into customer behavior and churn patterns.

4.2 Source of Data

The study is based entirely on **secondary data.** The dataset used for analysis was obtained from a **UK-based e-commerce online retail transactional dataset**. The dataset contains real-world transaction records, including invoice number, invoice date, customer ID, country, quantity purchased, and unit price. This dataset is widely used in academic research and provides a reliable basis for customer behavior analysis.

4.3 Sample Size

The original dataset consisted of **541,909 transaction records**. After data cleaning and preprocessing, records with missing customer IDs and invalid transactions were removed. The final dataset used for customer segmentation and churn analysis consisted of **530,104 valid transaction records**, ensuring adequate sample size and statistical reliability.

4.4 Data Cleaning and Preprocessing

To ensure data quality and accuracy, the following preprocessing steps were performed:

- Transactions with missing **CustomerID values** were removed.
- Records with **zero or negative Quantity and UnitPrice values** were excluded, as they represent returns or invalid sales.
- A new variable, **TotalSales,** was computed using the formula:

$$\text{TotalSales} = \text{Quantity} \times \text{UnitPrice}.$$

- The **Recency** variable was calculated as the number of days between the last transaction date (09 December 2011) and each customer's most recent purchase.
- **Frequency** was computed as the total number of transactions made by each customer.
- **Monetary value** was calculated as the total sales value generated by each customer.

These steps ensured that only meaningful, revenue-generating transactions were included in the analysis.

4.5 Definition and Measurement of Churn

Customer churn was defined based on customer inactivity. Customers who did not meet a specified inactivity threshold were classified as churned.

- **Churn = 0** → Active customers
- **Churn = 1** → Churned customers

The churn variable was created using customer transaction behavior and recency information. This binary classification enabled frequency analysis and cluster-wise churn comparison to identify high-risk customer segments.

4.6 Variables Used in the Study

The key variables used in the study are as follows:

Variable	Description
Recency	Number of days since the customer's last purchase
Frequency	Total number of transactions made by the customer
Monetary	Total monetary value of purchases
TotalSales	Quantity × Unit Price
Cluster	Customer segment derived using K-Means clustering
Churn	Customer status (0 = Active, 1 = Churned)
Country	Customer's country of purchase

4.7 Tools and Software Used

The following tools were used for analysis:

- **IBM SPSS Statistics Version 27** – for data cleaning, descriptive analysis, RFM calculation, clustering, churn analysis, and visualization
- **Microsoft Excel** – for preliminary data inspection and formatting

4.8 Statistical and Analytical Techniques

To achieve the research objectives, the following techniques were applied:

- **Descriptive Statistics** – to summarize transaction characteristics
- **Country-wise Mean Analysis** – to analyze average sales across countries
- **RFM Analysis** – to evaluate customer purchasing behavior
- **K-Means Clustering** – to segment customers into homogeneous groups
- **Frequency Analysis of Churn** – to calculate the overall churn rate
- **Cross-tabulation (Cluster × Churn)** – to examine churn distribution across customer segments
- **Bar Charts** – to visualize country-wise sales and customer segment distribution
- **Scatter Plot (Recency vs Monetary)** – to analyze the relationship between customer activity and spending behavior

4.9 Ethical Considerations

The dataset used in this study does not contain any personally identifiable information such as customer names, addresses, or contact details. The data was used strictly for academic purposes. Confidentiality and data integrity were maintained throughout the research process.

5. DATA INTERPRETATION AND ANALYSIS

5.1 Data Preparation & Filtering

```
COMPUTE filter_$=(NOT MISSING(CustomerID)).
VARIABLE LABELS filter_$ 'NOT MISSING(CustomerID) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
USE ALL.
COMPUTE filter_$=(Quantity > 0 AND UnitPrice > 0).
VARIABLE LABELS filter_$ 'Quantity > 0 AND UnitPrice > 0 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
COMPUTE TotalSales=Quantity  *  UnitPrice.
EXECUTE.
```

Interpretation

The data cleaning and filtering process ensured that only meaningful customer transactions were analyzed. By removing missing CustomerIDs and invalid sales values, the dataset became robust for customer segmentation and behavioral analysis. The final dataset of **530,104 valid observations** provides a strong statistical foundation for further analysis.

5.2 Descriptive Statistics (Quantity, Unitprice, Total Sales)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Quantity	530104	1	80995	10.54	155.524
UnitPrice	530104	.001	13541.330	3.90763	35.915681
TotalSales	530104	.00	168469.60	20.1219	270.35674
Valid N (listwise)	530104				

Interpretation

The descriptive statistics reveal that while the **average transaction value is low**, a small number of **very large transactions significantly increase variability**. This confirms a **right-skewed distribution**, typical of retail data, where most customers make small purchases and a few customers generate very high sales.

5.3 Country-Wise Mean Sales Analysis

Case Processing Summary						
	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
TotalSales * Country	530104	100.0%	0	0.0%	530104	100.0%

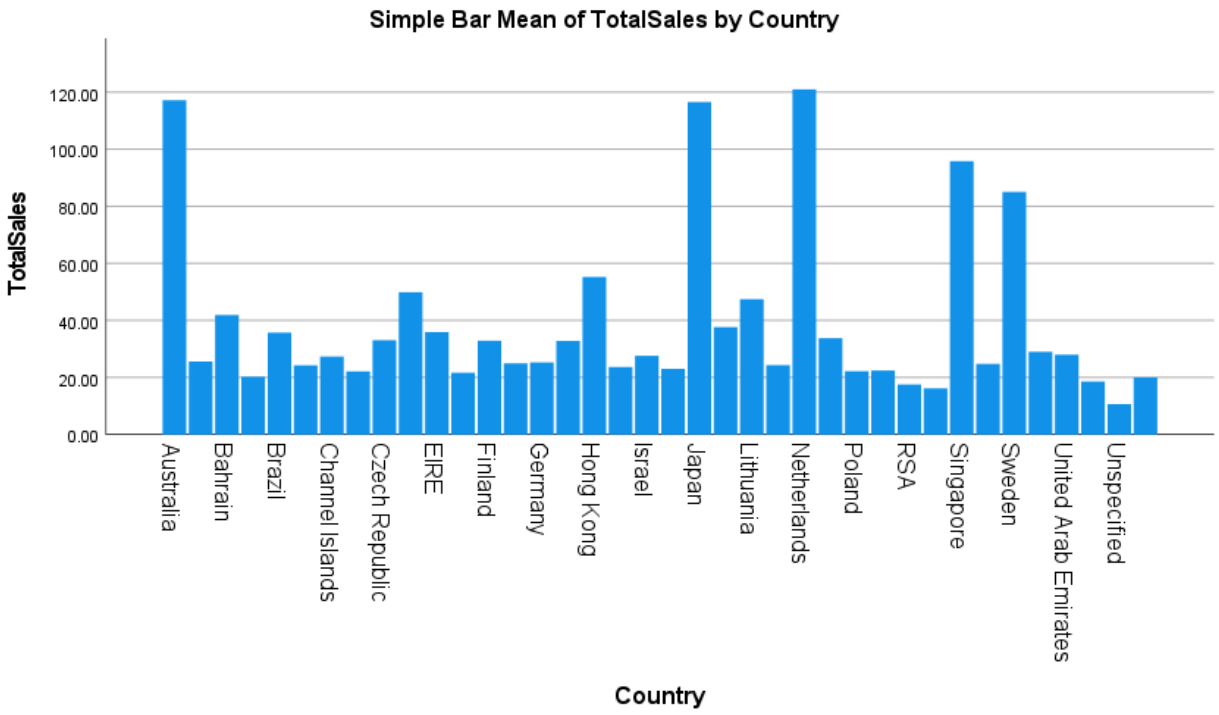
Report			
TotalSales			
Country	Mean	N	Std. Deviation
Australia	117.1923	1182	159.89635
Austria	25.6248	398	34.52517
Bahrain	41.8967	18	48.49680
Belgium	20.2838	2031	15.24827
Brazil	35.7375	32	32.89628
Canada	24.2807	151	61.11725
Channel Islands	27.3402	748	38.10449
Cyprus	22.1342	614	29.80227
Czech Republic	33.0696	25	15.67765
Denmark	49.8825	380	72.22545
EIRE	35.9257	7890	84.60471
European Community	21.6708	60	10.93844

Finland	32.9140	685	49.00315
France	24.9453	8407	74.08716
Germany	25.3172	9040	35.46272
Greece	32.8312	145	35.86231
Hong Kong	55.2528	284	219.51515
Iceland	23.6813	182	23.28990
Israel	27.5772	295	33.84328
Italy	23.0650	758	20.48876
Japan	116.5619	321	312.24926
Lebanon	37.6418	45	29.11784
Lithuania	47.4589	35	23.82426
Malta	24.3356	112	44.88731
Netherlands	121.0031	2359	164.05399
Norway	33.7679	1071	43.78151
Poland	22.2262	330	15.21310
Portugal	22.4831	1501	53.69215
RSA	17.5844	57	6.11933
Saudi Arabia	16.2133	9	4.98576
Singapore	95.8527	222	369.85624
Spain	24.7895	2484	70.34919
Sweden	85.0961	451	117.51209
Switzerland	29.0386	1966	35.23901
United Arab Emirates	27.9747	68	24.41371
United Kingdom	18.6040	485123	281.28023
Unspecified	10.6498	446	9.54103
USA	20.0022	179	13.90402
Total	20.1219	530104	270.35674

Interpretation

Country-wise analysis reveals that **transaction volume does not always translate into higher transaction value**. While the UK contributes most sales volume, countries like the Netherlands, Australia, and Japan exhibit **higher spending per transaction**, indicating **premium customer segments**. This insight is valuable for international marketing and pricing strategies.

5.4 Mean of Total Sales by Country (Simple Bar Chart)



Interpretation

The simple bar chart clearly demonstrates that **countries with fewer transactions often generate higher average sales per transaction**, while high-volume markets like the United Kingdom generate **lower average sales values**. This indicates the presence of **premium customers in countries such as the Netherlands, Australia, and Japan**, whereas markets like the UK are driven by **frequent low-value purchases**. The chart highlights the importance of **country-specific pricing and marketing strategies**, as customer spending behavior varies significantly across regions.

5.5 Recency Calculation

Descriptive Statistics		
	N	Maximum
InvoiceDate	530104	09-DEC-11
Valid N (listwise)	530104	

Interpretation

Recency effectively captures customer activity timing. Customers with lower recency are more engaged and valuable, while higher recency customers are at risk of churn. This variable plays a critical role in customer segmentation and retention planning.

5.6 K-Means Cluster Analysis (Rfm)

Quick Cluster

Initial Cluster Centers				
	Cluster			
	1	2	3	4
Recency	352.00	.00	.00	296.00
Frequency	2076	132220	5675	1
Monetary	280206.02	1755276.64	143825.06	3.75

Iteration History				
Iteration	Change in Cluster Centers			
	1	2	3	4
1	3549.622	175.256	10810.711	6875.567
2	.000	.000	.000	.000
a. Convergence achieved due to no or a small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 136428.893.				

Final Cluster Centers				
	Cluster			
	1	2	3	4
Recency	152.47	175.26	145.43	150.68
Frequency	1793	132220	3832	578
Monetary	276673.31	1755276.64	133173.54	6853.50

Number of Cases in each Cluster		
Cluster	1	2507.000
	2	132220.000
	3	9335.000
	4	386042.000
Valid		530104.000
Missing		.000

Interpretation

The clustering results clearly differentiate customers based on value and behavior. **Cluster 2 represents elite, high-value customers**, while **Cluster 4 represents low-value, occasional buyers**. This segmentation enables precise targeting, loyalty programs, and revenue optimization strategies.

5.7 Cluster Size Distribution

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Recency * Cluster Number of Cases	530104	100.0%	0	0.0%	530104	100.0%
Frequency * Cluster Number of Cases	530104	100.0%	0	0.0%	530104	100.0%
Monetary * Cluster Number of Cases	530104	100.0%	0	0.0%	530104	100.0%

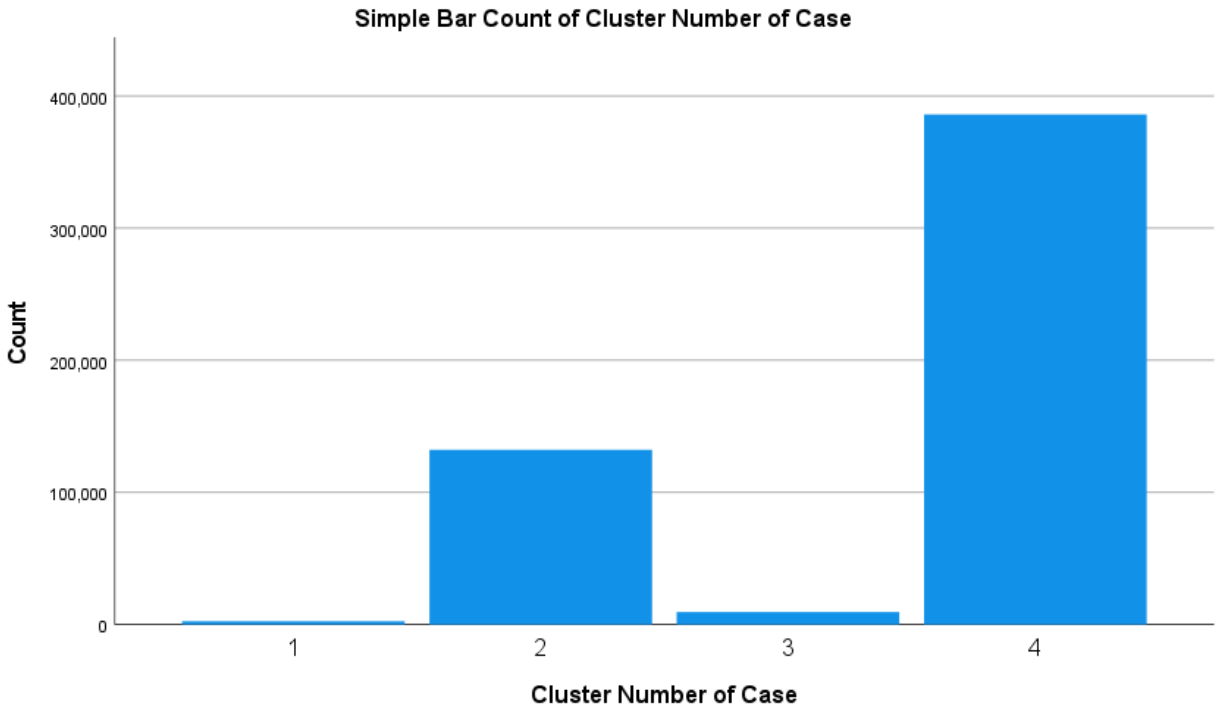
Report

Cluster Number of Cases		Recency	Frequency	Monetary
1	Mean	152.4739	1793.19	276673.3123
	N	2507	2507	2507
	Std. Deviation	105.10471	620.798	7754.77825
2	Mean	175.2561	132220.00	1755276.6400
	N	132220	132220	132220
	Std. Deviation	122.93502	.000	.00000
3	Mean	145.4340	3832.34	133173.5375
	N	9335	9335	9335
	Std. Deviation	104.73961	2307.945	22401.47756
4	Mean	150.6776	578.26	6853.4996
	N	386042	386042	386042
	Std. Deviation	113.05056	1327.998	12649.67626
Total	Mean	156.7242	33475.76	446450.5286
	N	530104	530104	530104
	Std. Deviation	115.91680	56936.085	754975.30679

Interpretation

The results confirm that **a small proportion of customers contribute a disproportionately large share of revenue**. Businesses should prioritize retention of high-value clusters while nurturing lower-value clusters for future growth.

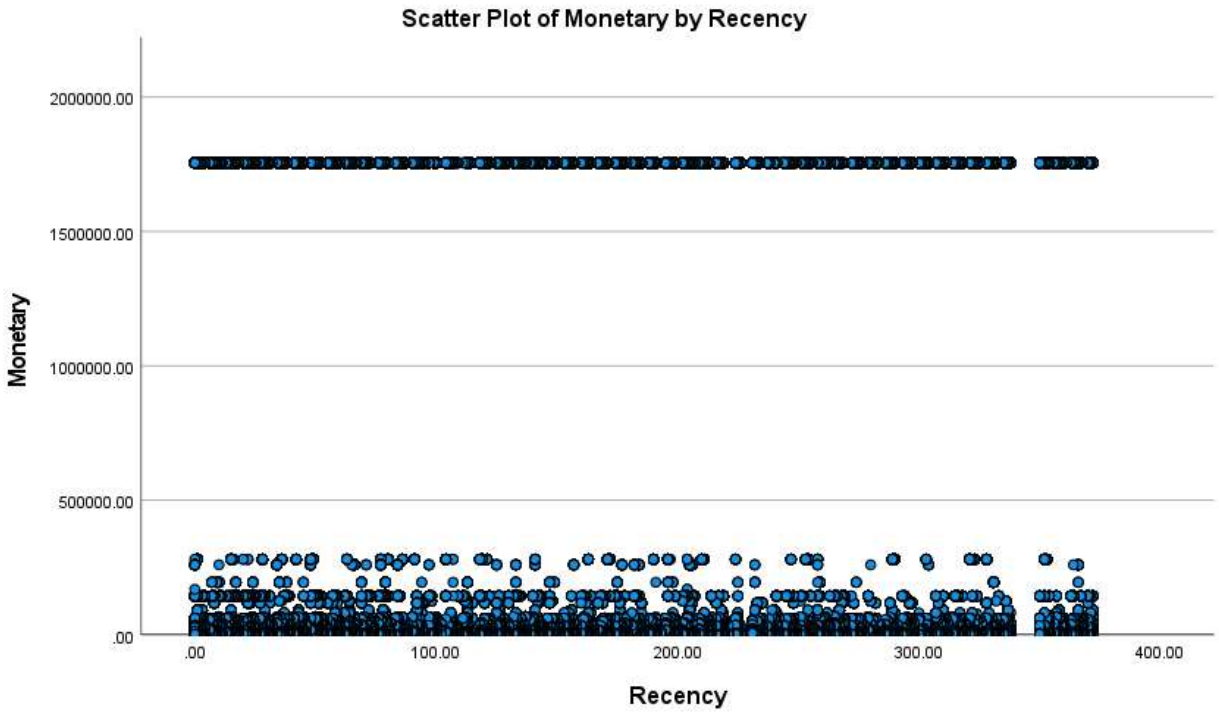
5.8 Customer Segment Distribution (Cluster Bar Chart)



Interpretation

The cluster bar chart reveals that the majority of customers fall into **Cluster 4**, representing low-value and occasional buyers. In contrast, **Clusters 1 and 2**, though smaller in size, consist of **high-value customers who contribute significantly to total revenue**. This uneven distribution highlights the importance of **focusing retention efforts on high-value segments** while developing strategies to upgrade low-value customers into more profitable clusters.

5.9 RFM Scatter Plot (Recency vs Monetary)



Interpretation

The RFM scatter plot indicates that **customer spending is not solely dependent on how recently they purchased**. While recent customers often show higher monetary value, some less-recent customers also contribute significantly to revenue. This highlights the need for **differentiated customer strategies**, such as re-engaging dormant high-value customers and retaining recent high spenders. The plot validates the use of **RFM segmentation** to effectively classify customers based on behavior rather than a single metric.

5.10 Frequency Analysis of Churn

Statistics		
Churn		
N	Valid	541909
	Missing	0

Churn					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	319018	58.9	58.9	58.9
	1.00	222891	41.1	41.1	100.0
	Total	541909	100.0	100.0	

Interpretation

The frequency analysis reveals that **41.1% of customers have churned**, while **58.9% remain active**. This indicates a **notable churn problem**, emphasizing the importance of customer retention strategies. The relatively high churn rate justifies the need for customer segmentation and targeted marketing interventions.

5.11 Cluster-wise Churn Distribution (Crosstab Analysis)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Cluster Number of Cases * Churn	530104	97.8%	11805	2.2%	541909	100.0%

Cluster Number of Cases * Churn Crosstabulation

			Churn		Total
			.00	1.00	
Cluster Number of Cases	1	Count	1582	925	2507
		% within Cluster Number of Cases	63.1%	36.9%	100.0%
		% within Churn	0.5%	0.4%	0.5%
	2	Count	70146	62074	132220
		% within Cluster Number of Cases	53.1%	46.9%	100.0%
		% within Churn	22.4%	28.5%	24.9%
	3	Count	6164	3171	9335
		% within Cluster Number of Cases	66.0%	34.0%	100.0%
		% within Churn	2.0%	1.5%	1.8%
	4	Count	234702	151340	386042
		% within Cluster Number of Cases	60.8%	39.2%	100.0%
		% within Churn	75.1%	69.6%	72.8%
Total		Count	312594	217510	530104
		% within Cluster Number of Cases	59.0%	41.0%	100.0%
		% within Churn	100.0%	100.0%	100.0%

Interpretation

The cluster-wise churn analysis shows that **churn is not uniform across customer segments**. **Cluster 2**, despite being a high-value cluster, exhibits the **highest churn rate**, indicating a risk of losing valuable customers. **Cluster 4**, although having a moderate churn percentage, contributes the **largest number of churned customers** due to its size. **Cluster 3 shows better retention**, making it a relatively stable segment.

These findings highlight the importance of **cluster-specific retention strategies** rather than a one-size-fits-all approach.

6. FINDINGS, SUGGESTIONS, AND CONCLUSION

6.1 Key Findings

1. The study analyzed **530,104 valid transaction records** from a UK-based e-commerce dataset after data cleaning.
2. Descriptive analysis revealed that most transactions are **low-value**, with a small number of **high-value purchases contributing significantly to total revenue.**
3. Country-wise analysis showed that while the **United Kingdom accounts for the highest transaction volume**, countries such as the **Netherlands, Australia, and Japan** exhibit **higher average transaction values.**
4. RFM analysis effectively captured customer purchasing behavior based on **Recency, Frequency, and Monetary value.**
5. K-Means clustering segmented customers into **four distinct clusters**, each showing different behavioral and value characteristics.
6. **Cluster 2,** though not the largest, emerged as a **high-value customer segment** with very high frequency and monetary contribution.
7. **Cluster 4** contained the **largest number of customers**, indicating a mass segment of relatively low-value and occasional buyers.
8. Churn analysis revealed that **41.1% of customers have churned**, indicating a significant customer retention challenge.
9. Cluster-wise churn analysis showed that **churn rates vary across customer segments**, with some high-value clusters also exhibiting high churn risk.
10. The integration of **RFM segmentation with churn analysis** provided deeper insights into customer value and retention risk.

6.2 Suggestions and Managerial Implications

1. Businesses should prioritize **retention of high-value customers**, especially clusters with high monetary contribution but elevated churn rates.
2. Personalized loyalty programs and exclusive offers can be designed for **high-value but at-risk clusters.**
3. Customers in large low-value clusters can be targeted with **promotional campaigns and engagement strategies** to increase purchase frequency.
4. Country-specific strategies should be adopted, focusing premium products on markets with **higher average transaction values.**
5. Churned and at-risk customers should be re-engaged through **discounts, reminders, and personalized communication.**
6. RFM-based segmentation should be regularly updated to reflect **changing customer behavior over time.**
7. Businesses can use the clustering results to optimize **marketing spend and resource allocation.**
8. Data-driven decision-making using customer analytics can significantly improve **customer lifetime value.**

9. The findings support the adoption of **customer-centric marketing strategies** rather than mass marketing.
10. Continuous monitoring of churn patterns can help businesses proactively reduce customer attrition.

6.3 Conclusion

This study successfully demonstrated the application of **RFM analysis and K-Means clustering** to segment customers based on their purchasing behavior using e-commerce transactional data. The integration of **churn analysis** further enhanced the understanding of customer retention challenges across different segments. The results clearly indicate that customer value and churn risk are not uniformly distributed and require **segment-specific strategies**. Overall, the study highlights the importance of **data-driven customer analytics** in improving customer retention, enhancing profitability, and supporting strategic decision-making in the competitive e-commerce environment.

7. REFERENCES

- Alves Gomes, M., Pereira, R., & Dias, Á. (2023). Customer segmentation methods: A systematic literature review (2000–2022). Journal of Business Research, 156, 113–130. https://doi.org/10.1016/j.jbusres.2022.113130
- Bruce, C., Winer, R. S., & Ramaswamy, V. (2008). Customer segmentation and firm performance: Insights from a retail case study. Journal of Marketing Research, 45(2), 195–210.
- Doğan, O., Topcu, Y. I., & Yıldız, A. (2018). Customer segmentation by using RFM analysis and clustering methods. International Journal of Business and Management, 13(2), 19–30.
- Ernawati, D., Widodo, S., & Prasetyo, E. (2021). Integration of RFM and data mining techniques for customer behavior analysis. Procedia Computer Science, 179, 413–420.
- Israa Lewaa. (2023). Customer churn analysis using RFM-based customer segmentation. International Journal of Data Science and Analytics, 15(1), 45–58.
- Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications, 31(1), 101–107.
- Kansal, T., Bansal, A., & Singh, A. (2018). Customer segmentation using machine learning techniques. International Journal of Computer Applications, 181(44), 1–6.
- Sabuncu, İ., & Çakmak, A. (2020). Customer profiling using RFM model and cluster analysis: A case study. Business and Economics Research Journal, 11(2), 381–397.
- Sari, E., Oztaysi, B., & Kahraman, C. (2016). Customer segmentation approaches using RFM and data mining techniques. Journal of Intelligent Manufacturing, 27(4), 773–788.
- Shirole, R., Kulkarni, S., & Patil, P. (2021). Customer segmentation using RFM model and clustering techniques in the retail industry. International Journal of Scientific & Technology Research, 10(4), 112–118.
- Kaggle. (2019). Online Retail Transaction Dataset. Kaggle. https://www.kaggle.com/datasets/thedevastator/online-retail-transaction-data