Customer Segmentation and Purchasing Behaviour Analysis Using RFM and K-Means Clustering

Anushka Sharma
Department of Computer Applications (MCA), IIS (Deemed to be University), Jaipur, India
Email: [anushkasharma290403@gmail.com]

Dr. Preeti Tiwari,Associate Professor,International School of Informatics & Management,

Jaipur

Email:preeti.tiwari@icfia.org

Abstract — It is important for a retailer to understand their customers' purchasing behaviours to keep valuable customers and create effective marketing strategies. This paper analyses customer transactions from the Online Retail II dataset, which contains 1,048,575 records and eight variables, to understand what customers buy, how often they purchase and how much they spend. After removing missing, duplicate and cancelled records, customers were described using Recency, Frequency and Monetary (RFM) measures, and the log-transformed, standardized RFM values were grouped using K-Means clustering. The Elbow method and the Silhouette score, which reached 0.44 at K = 2, supported the choice of two clusters. Cluster 0 contains 2,301 active, high-value customers (average monetary value 6,431.31, frequency 12.62, recency 51.70 days), whereas Cluster 1 contains 3,559 lapsed, lower-value customers (606.48, 2.09 and 297.50 days respectively). Exploratory analysis also shows that the United Kingdom dominates the transactions and that purchasing activity peaks in November. Based on these segments, retention and win-back marketing strategies are proposed. The study shows that RFM-based clustering is a simple and interpretable approach to customer segmentation and business decision-making.

Keywords: Customer Segmentation, Purchasing Behaviour, RFM Analysis, K-Means Clustering, Silhouette Score, Elbow Method, Online Retail

I. INTRODUCTION

Every purchase a customer makes leaves behind a small amount of data. When online retailers record millions of purchases this can provide a whole picture of what customers are purchasing, how often and how much. Businesses that can understand this can provide their customers with a better service. Customers vary in their purchasing habits with some purchasing frequently and spending more. If all customers are treated the same this leads to wasted marketing resources and may leave valuable customers overlooked. The challenge is to take the raw data and provide meaningful segments.

Customer segmentation is a method of grouping customers into different segments where they will behave similarly. This has been a key concept in database marketing [1] as it allows customers to be described by their Recency, Frequency and Monetary (RFM). Recency describes how recently they last purchased, Frequency how often they purchased and Monetary how much they spent. Customers that are recent, frequent and have a high monetary value tend to be the most valuable. Whereas old infrequent customers are at risk of being lost. This creates an overall picture of their purchasing habits using just their transaction data. Many studies have used RFM for customer profiling [2]-[6].

However, RFM scores are rule based and depend on the thresholds chosen by the analyst. To provide a data driven approach clustering algorithms can be used. K-Means [8] is one of the most popular clustering methods [14]. This works by placing a customer into the closest of K cluster centres and re-calculating the centres repeatedly until stable cluster groups are found. The number of clusters needs to be decided in advance and the Elbow method [10] can be used to help decide this. The Silhouette score [9] can be used to validate this number. This creates an interpretable and quantifiable way to group customers using RFM.

This study looks at analysing the Online Retail II dataset [7]. The dataset is large and contains missing, duplicated and cancelled data so provides a realistic example of data cleaning. The process followed was Data Collection & Cleaning, Exploratory Analysis, RFM Calculation & Scoring, Standardisation, K determination using the Elbow method [10] and Silhouette score [9], K-Means Clustering, Cluster Profiling and Business Analysis. Questions were asked throughout the analysis.

The aim of this study is to understand customer behaviour and provide some valuable customer segments to aid marketing and customer retention. This paper is organised as follows: Section II looks at related work, Section III details the experiment setup and data, Section IV lists the questions asked during the analysis,

Section V details the results and discussion, Section VI concludes the paper and Section VII is the reference list.

II. LITERATURE REVIEW

Chen, Sain and Guo [2] applied the RFM model to the transaction data of a UK based online retailer to segment customers using data mining. Customers were scored based on RFM values and the most valuable segments were identified. This proved RFM can be practically used for customer profiling in retail. However, the scores were rule based and no validated clustering was carried out.

Christy et al [3] proposed the use of RFM ranking for customer segmentation. K-Means and Fuzzy C-Means were used to cluster on the RFM values. A novel initialisation of centroids was applied to improve the quality of clusters. This formed clear customer segments however there was no link to actual marketing strategies that could be used.

Anitha and Patil [4] created RFM features from the transaction data of retail purchases. Customers were segmented using K-Means. The number of clusters were validated using the Silhouette score. This proves that RFM and K-means work well together. However there was no explicit handling of duplicated and cancelled transactions so this could affect the RFM values.

A study in 2023 [5] applied RFM analysis and used K-Means clustering to segment customers. The number of clusters were determined using both the Elbow method and Silhouette score. This provided a more reliable way of determining the number of clusters. However the dataset was small and there was no example of the scalability of the method.

A study in 2024 [6] looked at RFM based segmentation using K-Means and Hierarchical clustering. The data was normalised using Yeo-Johnson to reduce skewness. Two different clustering methods were applied to show how preprocessing affects results. However the best normalisation method is still unclear.

An IEEE study in 2019 [8] used RFM and K-Means on a reload credit business. After filtering there was only 102 customers to analyse using K-Means. This shows the method can be applied to a telecom related business. However the sample size was very small and there was no validation index.

A 2020 arXiv pre-print [9] looked at RFM for bank customer segmentation. K-Means and K-Modes were used with demographic features added. This created a more detailed customer profile. However demographic features are rarely available in retail transaction data.

This table shows a summary of the studies analysed. The most common method was using RFM and K-Means. There were some gaps in the data cleaning, validating the number of clusters, dataset size and linking the segments to marketing. These gaps have been addressed in this study by cleaning a large dataset, validating using two criteria and suggesting strategies for each segment.

Table 1: Comprehensive summary of literature review

S. No	Paper Title	Author	Year	Tools	Methodology	Research Gap
1	RFM with clustering in e-commerce	JTDE	2024	K-Means, Hierarchical, Yeo-Johnson	RFM + normalization + two clustering methods	Best normalization method still open
2	RFM profiling via K-Means clustering	IIETA	2023	K-Means, Elbow, Silhouette	RFM + Elbow/Silhouette for K selection	Small dataset, unproven scalability
3	RFM model using K-Means algorithm	Anitha & Patil	2022	K-Means, Silhouette	RFM + K-Means, silhouette validation	No handling of duplicate/cancelled data
4	RFM ranking for customer segmentation	Christy et al.	2021	K-Means, Fuzzy C-Means	RFM + novel centroid initialization	No link to marketing action
5	RFM for bank customer segmentation	arXiv preprint	2020	K-Means, K-Modes	RFM + demographic features	Demographic data unavailable in retail sets
6	RFM + K-Means for reload-credit business	IEEE	2019	K-Means, RapidMiner	RFM reduced to 102 customers, K-Means	Very small sample, no validation index

S. No	Paper Title	Author	Year	Tools	Methodology	Research Gap
7	RFM-based customer segmentation in online retail	Chen, Sain & Guo	2012	RFM scoring	RFM model on UK retail transactions	Rule-based scoring, no validated clustering

III. EXPERIMENTAL SETUP

This section describes the hardware, software, tools and techniques used, followed by the dataset and the methodology.

Hardware and Software

The experiments do not require any specialist hardware. All that is required is a standard PC with a multi-core processor and a few gigabytes of memory to load the transaction data. The RFM table only has less than 6000 rows so this is achievable. The software and techniques used are:

- Programming Language – Python was used within a notebook.
- Pandas [12] and NumPy [15] – Used to load, clean, group and calculate.
- Matplotlib [13] – Used to create charts such as bar charts, histograms, line charts, scatter plots and the correlation heatmap.
- Scikit-learn [11] – Used to standardise, cluster using K-Means and calculate the Silhouette score.
- RFM Analysis, Log Transformation, Standardisation, K-Means Clustering [8], Elbow Method [10] and Silhouette Analysis [9].

(a) Dataset

The dataset used in this study is the Online Retail II dataset [7]. The data contains transaction records from an online retail company based in the UK. There are 1,048,575 records and 8 variables. In this study we are looking at customer purchasing behaviour. Table 2 shows the variable details and Figure 1 shows the first 5 records.

Table 2: Dataset description: fields and their meaning

Field	Description
Invoice	Unique number assigned to each transaction
StockCode	Code used to identify a product
Description	Name or description of the product
Quantity	Number of units purchased
InvoiceDate	Date and time when the transaction occurred
Price	Price of one unit of the product
Customer ID	Unique identification number of the customer
Country	Country where the customer is located

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	12/1/2009 7:45	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	12/1/2009 7:45	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	12/1/2009 7:45	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	12/1/2009 7:45	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	12/1/2009 7:45	1.25	13085.0	United Kingdom

Fig. 1: Dataset snapshot (first five records of Online Retail II)

(b) Methodology

The methodology consists of thirteen steps (shown in figure 2). The data was collected and loaded from Online Retail II. The data was cleaned removing any records with missing Customer ID, duplicated data,

cancelled transactions and invalid data. Exploratory data analysis was completed looking at the country, products, monthly trends and quantity/price distribution.
RFM analysis was performed to create one row per customer. Recency shows how many days ago they last purchased, Frequency shows how often they purchased and Monetary shows how much they spent. RFM scoring was applied to rank the customers. As the RFM values were highly skewed the values were log transformed and standardised (zero mean and unit variance). This ensures no variable dominates the distance calculation in K-Means.
The Elbow method [10] was used to calculate the Within Cluster Sum of Squares (WCSS) and the Silhouette score for K = 2 to 10 to determine the number of clusters. K-Means clustering was applied with the chosen K. A cluster profile was created showing how many customers are in each cluster and the average Recency, Frequency and Monetary value of each cluster. Business analysis was performed to identify the most valuable segment and suggest a marketing strategy for each segment.

.

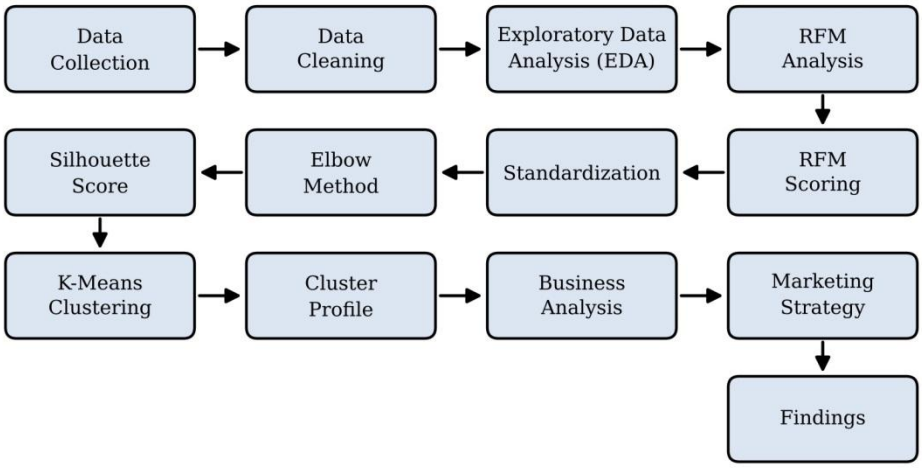


Fig. 2: Methodology flowchart of the proposed customer segmentation study

IV. EXPERIMENTS

The analysis is organised as a sequence of questions that move from understanding the data to preparing it, describing it, computing RFM values, clustering customers and finally drawing business conclusions. Each question is answered with a chart or table in Section V. Table 3 lists the questions, the section they belong to and why each one matters.

Table 3: Table of analysis questions

S. No	Section Header	Question	Significance
Q1	Data Understanding	Which countries have the most transactions?	Shows the geographic concentration of the business
Q2		Which products have the highest total quantity sold?	Identifies high-demand products and customer preferences
Q3		How do transaction volumes change over time?	Reveals seasonal peaks and low-activity periods
Q4	Data Quality & Preprocessing	How much missing data is present in the dataset?	Decides how records are handled before customer-level analysis
Q5		How many duplicate and cancelled transactions are there?	Prevents distortion of sales and RFM values
Q6	Descriptive Analysis	What can we learn from the quantity and price distribution?	Detects skewness and extreme outliers
Q7		Which customers contribute the most to total purchase value?	Highlights dependence on a few high-value customers

S. No	Section Header	Question	Significance
Q8	RFM Analysis	How recent have customers made a purchase? (Recency)	Measures how active customers are at present
Q9		How frequently do customers make purchases? (Frequency)	Separates occasional from repeat customers
Q10		How much do customers spend? (Monetary)	Separates lower-spending from high-value customers
Q11		How are the RFM variables related to each other?	Checks whether the variables give complementary information
Q12	Customer Segmentation	What is the optimal number of clusters according to the Elbow Method?	Provides the first estimate of K for K-Means
Q13		Which cluster number gives the best separation?	Validates the choice of K using the Silhouette score
Q14		How can customers be grouped based on purchasing behaviour?	Converts RFM measures into customer groups
Q15	Cluster Analysis	How do the clusters differ in profile and purchasing behaviour?	Interprets what each segment represents
Q16	Business Analysis	Which customer segment has the highest business value?	Directs marketing effort to the right segment
Q17		What marketing strategy suits each customer segment?	Turns the segments into practical actions
Q18	Advanced Analysis	Can we forecast future sales based on historical transaction trends?	Supports inventory and sales planning

V. RESULTS AND DISCUSSIONS

Q1. Which countries have the most transactions?

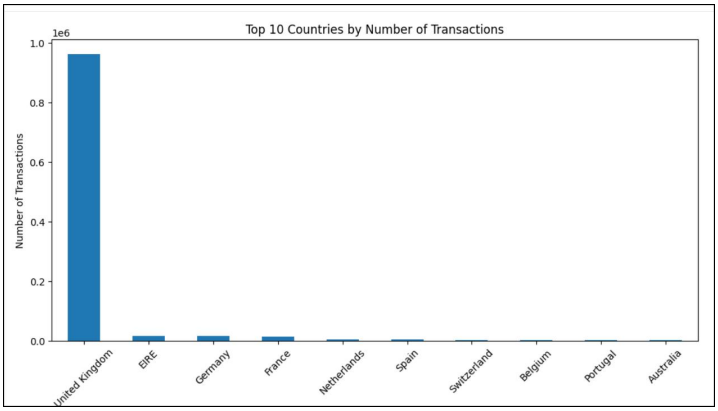


Fig. 3: Top 10 countries by number of transactions

The United Kingdom had the most transactions (963,819) which was around 92% of the data. After the UK, EIRE had the next most transactions (17,581) followed by Germany (17,327), France (14,100), Netherlands (5,054), Spain (3,740), Switzerland (3,189), Belgium (3,056), Portugal (2,566) and Australia (1,913). We can see that most of the business comes from customers within the home market. The segments will mainly be based on UK customers.

Q2. Which products have the highest total quantity sold?

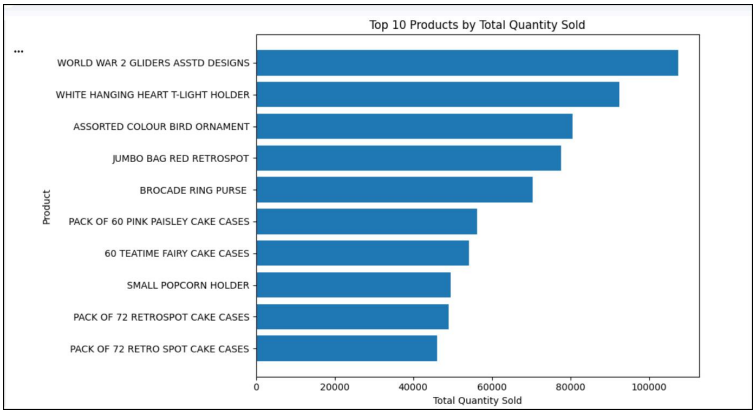


Fig. 4: Top 10 products by total quantity sold

The most popular item sold was WORLD WAR 2 GLIDERS ASSTD DESIGNS with around 107,000 units sold. This was followed by WHITE HANGING HEART T-LIGHT HOLDER and ASSORTED COLOUR BIRD ORNAMENT. The most popular products are all toys, bags, cake cases and home decoration items. There is a large variation in total quantity sold per product meaning it would be useful to look at individual products to understand better what customers want.

Q3. How do transaction volumes change over time?

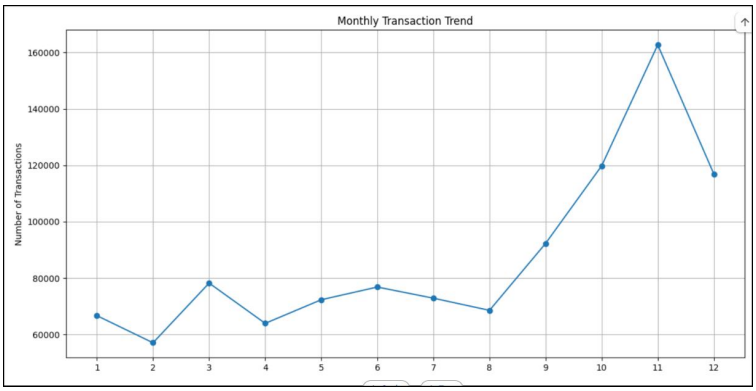


Fig. 5: Monthly transaction trend (months 1–12)

Overall we can see that the number of transactions is increasing steadily from January to August. The number of transactions then increases significantly from September. There were around 92,000 transactions in September and this increased to 120,000 in October. November had the highest number of transactions of around 163,000. There is then a decrease in the number of transactions in December to around 117,000. There is a large spike in transactions just before the festive period. This can show us when customers are most active and when they are not.

Q4. How much missing data is present in the dataset?

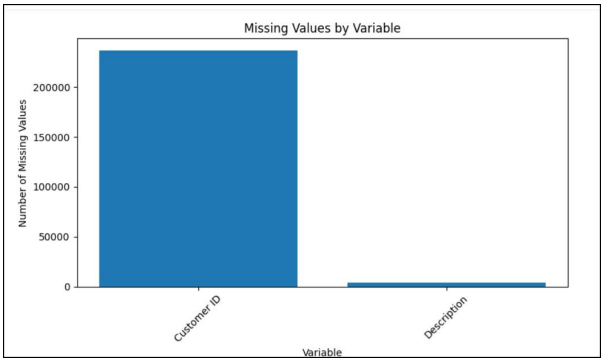


Fig. 6: Missing values by variable

Customer ID had the most missing data with over 230,000 missing entries. Description only had a few thousand missing entries. As we are analysing the data at a customer level we need to know who the

customers are. Any records with no Customer ID will not be able to be assigned to any customers. This has been dealt with prior to performing RFM and segmenting the data.

Q5. How many duplicate and cancelled transactions are there?

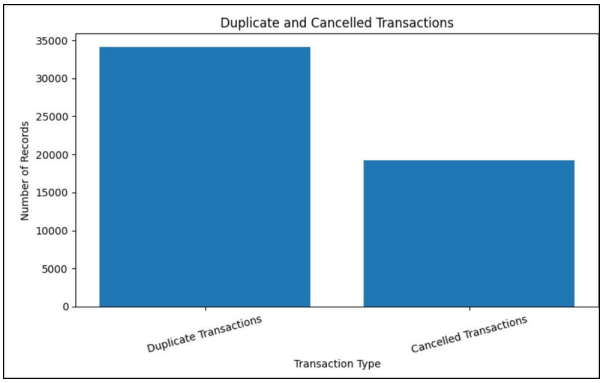


Fig. 7: Duplicate and cancelled transactions

There are 34,150 duplicate transactions and 19,261 cancelled transactions. This would mean we are double counting purchases and some purchases have been cancelled. This can cause problems when analysing sales and RFM values. These have been identified and removed.

Q6. What can we learn from the quantity and price distribution?

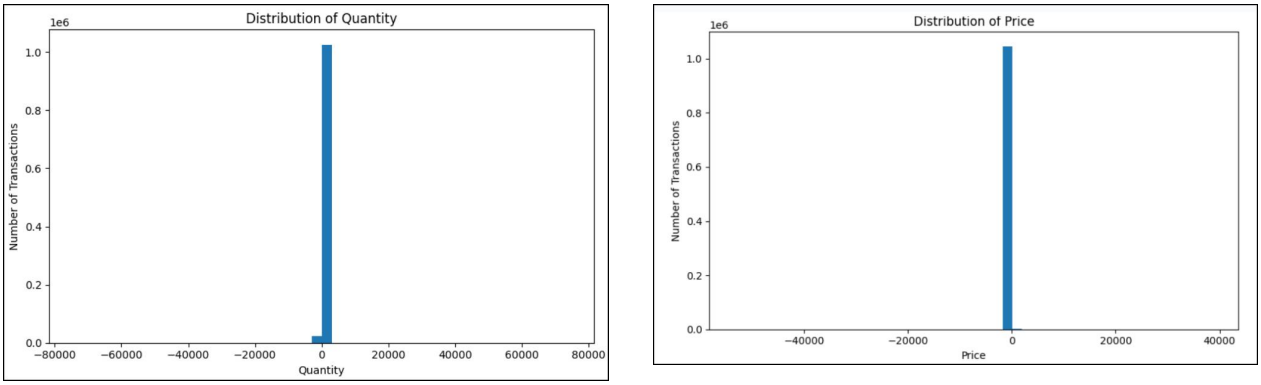


Fig. 8: Distribution of (a) quantity and (b) price

From this we can see that both have a high mean value compared to the median. Quantity has a mean of 9.96 and a median of 3. Price has a mean of 4.63 and a median of 2.10. From the histograms we can see that there is a large spread of values with extreme outliers. There are negative values which are due to returns and adjustments. Overall most transactions have a low quantity and price however there are a few with very high values.

Q7. Which customers contribute the most to total purchase value?

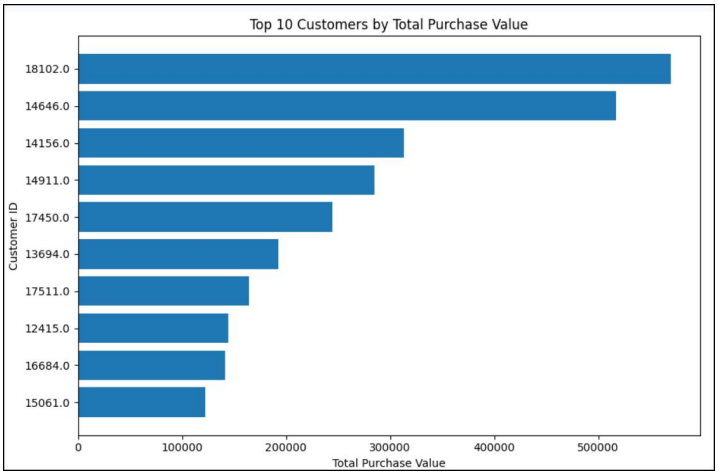


Fig. 9: Top 10 customers by total purchase value

The two largest customers (IDs 18102 and 14646) each spent over 500,000, and the top ten customers range from roughly 120,000 to 570,000. Compared with the typical customer (Q10), a small group of customers contributes a substantial part of the total purchase value. Identifying such high-value customers is important for targeted retention and marketing.

Q8. How recent have customers made a purchase?

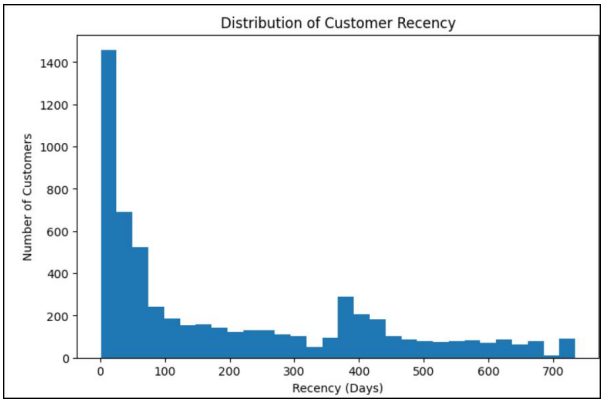


Fig. 10: Distribution of customer recency

There are a large number of customers with low recency values. Around 1,450 customers are in the first bin meaning they have made a recent purchase. As recency increases the number of customers decreases creating a right skewed distribution with values up to around 730. There are another group of customers around the 350-450 mark. This means there is a group of customers who have not made a purchase in quite some time. The lower the recency the more recent a purchase has been made meaning the customer is more active.

Q9. How frequently do customers make purchases?

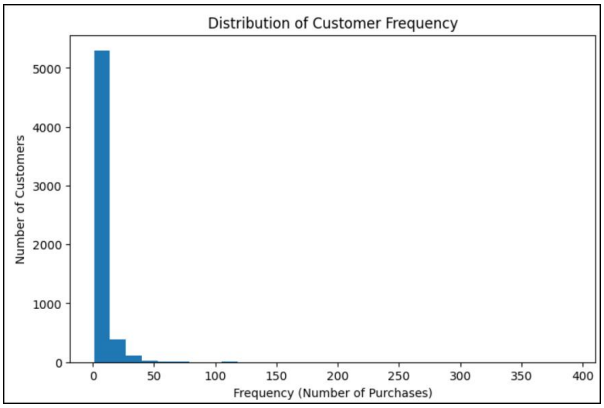


Fig. 11: Distribution of customer frequency

The vast majority of customers have a low frequency of purchases (around 5,300). As frequency increases the number of customers decreases rapidly. Only a small number of customers have a very high frequency with the highest being around 400. This means most customers do not purchase very often however there are a few who purchase frequently..

Q10. How much do customers spend on their purchases?

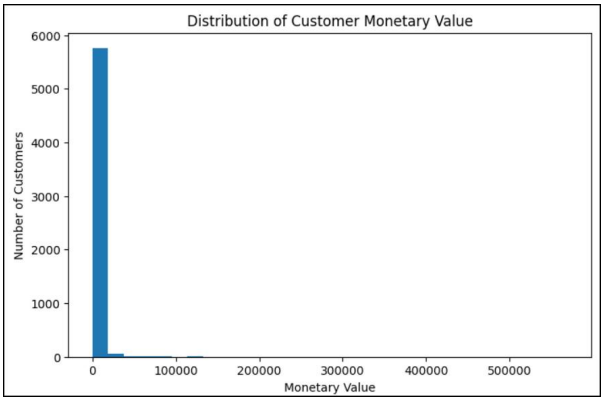


Fig. 12: Distribution of customer monetary value

Most customers fall into the lowest spending bracket (around 5,750). As monetary value increases the number of customers decreases. There are a small number of customers who spend a lot more (up to around 570,000). This means there are a number of low value customers however there are also high value customers.

Q11. How are the RFM variables related to each other?

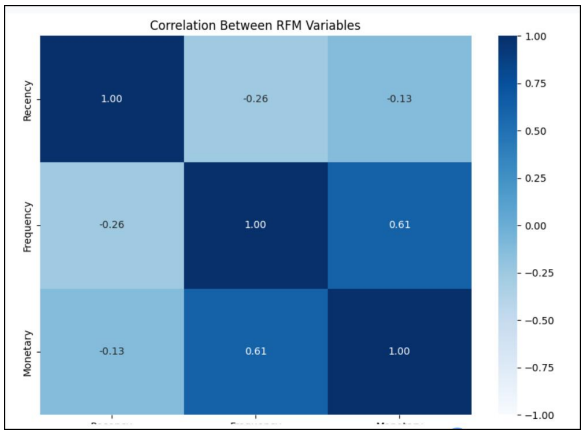


Fig. 13: Correlation between RFM variables

Frequency and Monetary have a strong positive correlation (0.61). Customers who purchase frequently will generally contribute more to the purchase value. Recency has a negative correlation with Frequency (-0.26) and Monetary (-0.13). This is expected as customers with a low recency have purchased recently. Overall the variables have a moderate correlation meaning we can gain complementary information.

Q12. What is the optimal number of clusters according to the Elbow Method?

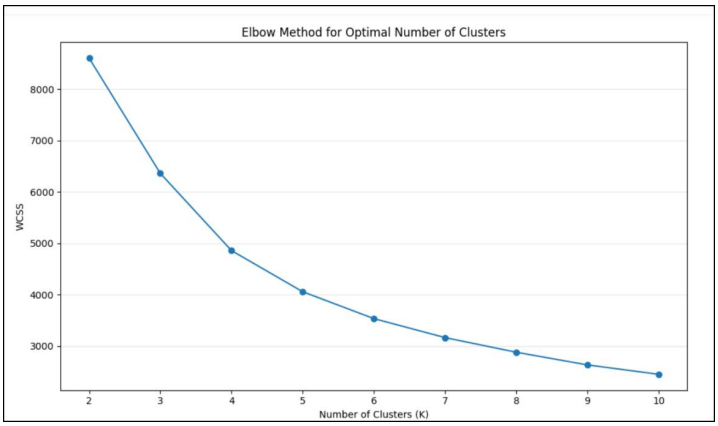


Fig. 14: Elbow method for the optimal number of clusters

The WCSS starts at around 8,600 with K = 2 clusters and ends at around 2,450 with K = 10 clusters. We can see there is a noticeable elbow at K = 2. After this point adding more clusters does not improve the model as much. We decided to go with K = 2 clusters. However after this point we can see the curve starts to decrease slowly meaning we used the Silhouette score.

Q13. Which cluster number gives the best separation?

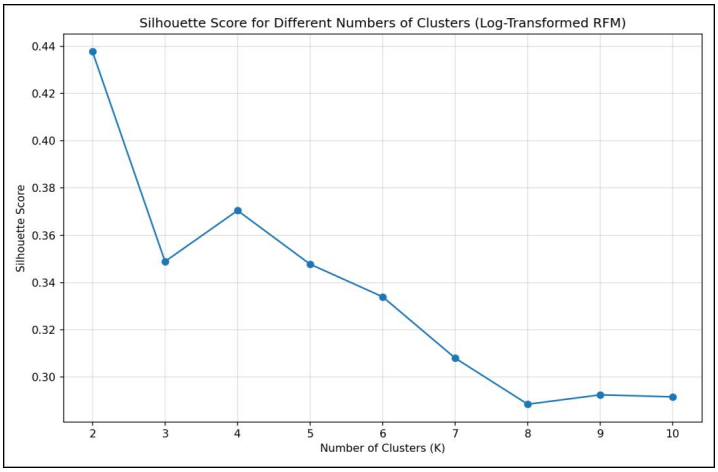


Fig. 15: Silhouette score for different numbers of clusters (log-transformed RFM)

K = 2 had the highest Silhouette score (0.44) for log transformed RFM data. K = 3 and K = 4 have lower scores (around 0.35 and 0.37). K = 8 have a much lower score (around 0.29). This means there is a good balanced split between two customer groups. We decided to go with K = 2 clusters for our final K-Means model. This agrees with the Elbow method.

Q14. How can customers be grouped based on their purchasing behaviour?

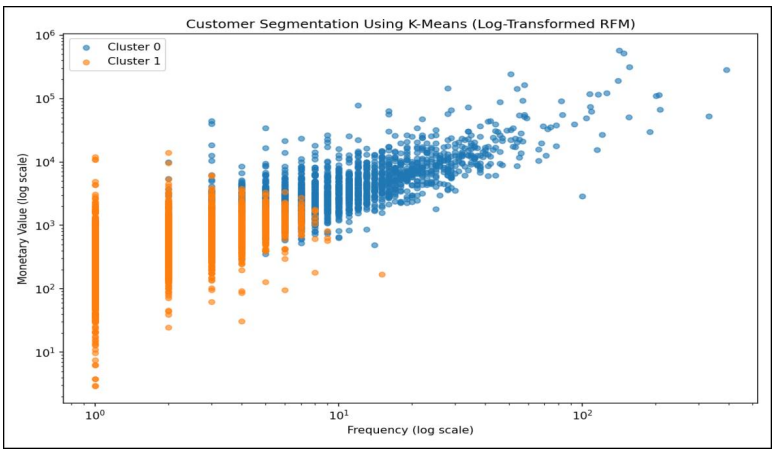


Fig. 16: Customer segmentation using K-Means (log-transformed RFM)

Cluster 0 contains customers with a high frequency and monetary value. Cluster 1 contains customers with a lower frequency and monetary value. Looking at the scatter plot we can see that Cluster 1 are concentrated in the lower frequency (mostly between 1- a few purchases) and monetary value. Cluster 0 has a spread across the higher frequency and monetary values. We have created two meaningful groups of customers based on their RFM.

Q15. How do customer clusters differ in their profile and purchasing behaviour?

Table 4: Cluster profile and customer behaviour

Cluster	Customer Count	Avg. Recency (Days)	Avg. Frequency	Avg. Monetary
0	2,301	51.70	12.62	6,431.31
1	3,559	297.50	2.09	606.48

Cluster 0 is a smaller group of active customers (2,301 – around 39%). They purchase frequently and spend a lot of money. This can be seen from the low recency (51.70 days). Cluster 1 is a larger group of customers (3,559 – around 61%). They purchase infrequently and spend little. They have a high average recency (297.50 days) meaning they have mostly become inactive. On average Cluster 0 purchase around 6 times more and spend 10 times more than Cluster 1.

Q16. Which customer segment has the highest business value?

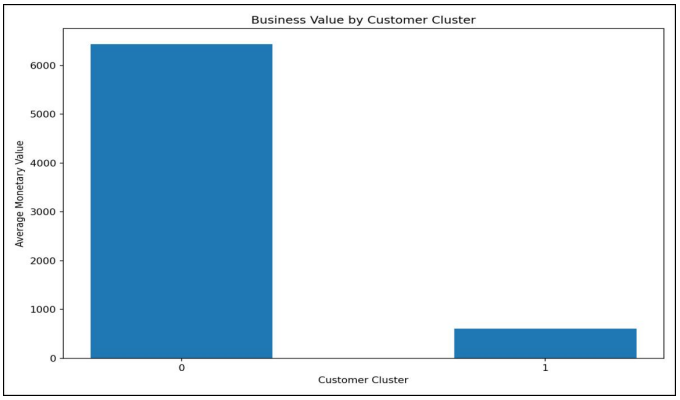


Fig. 17: Business value by customer cluster (average monetary value)

Cluster 0 has the highest value with an average monetary value of 6,431.31 compared to Cluster 1 (606.48). Cluster 0 also has a much higher average frequency (12.62 vs 2.09) and lower recency (51.70). Looking at the size and averages we can see Cluster 0 (2,301) make up roughly 87% of the total spend between the two clusters. Cluster 0 are the high value active customers whilst Cluster 1 are the larger group meaning they should be the focus of win back efforts.

Q17. What marketing strategy suits each customer segment?

Table 5: Marketing strategy for each customer segment

Segment	Recommended marketing strategy
Cluster 0 – Active / High-Value Customers	Give rewards and exclusive offers to keep these customers. Offer VIP benefits and personal services. Recommend relevant products as they purchase frequently. Focus on keeping them.
Cluster 1 – Lapsed / Lower-Value Customers	Give personalised discounts and offers to encourage purchases. Send reminders and recommendations. Give limited time offers to encourage less recent customers. Gradually increase purchase frequency and spend.

The strategies follow directly from the cluster profiles. For Cluster 0 the goal is to protect an existing source of revenue, so rewards and personal service are appropriate. For Cluster 1 the goal is reactivation, so time-limited incentives and reminders are appropriate, with the aim of moving customers gradually towards higher frequency and spending.

Q18 . Can we forecast future sales based on historical transaction trends?

We can use historical transaction data to identify trends in sales. Looking at the sales per month we can see if sales are increasing, decreasing or following a pattern. Looking at Figure 5 we can see there is an increase in sales seasonally in November. We could use a forecasting technique such as moving average, exponential smoothing or ARIMA to forecast future sales based on historical trends. This can help businesses plan their stock, sales targets and future operations. Forecasting has been discussed in this study but can be expanded on.

VI. CONCLUSION

In this study we have analysed the Online Retail II transactions to understand customer purchasing behaviour. Some key findings from the study are: The United Kingdom had the most transactions. Data was cleaned removing missing, duplicated, cancelled and invalid records. RFM analysis showed there was a clear difference between customer recency, frequency and monetary. The Elbow method and Silhouette score (0.44) suggested there should be 2 clusters. Cluster 1 had the majority of customers (3,559). Cluster 0 had fewer customers (2,301) who were very active with the highest average frequency and monetary value. This means they are the most valuable segment.

Overall we have seen there is a large difference in purchasing behaviour between different customer groups. Using descriptive analysis, RFM analysis and K-Means we can separate a small portion of highly active high value customers from a larger relatively lower value segment. This can help with targeted marketing, customer retention and making better business decisions. Understanding customer behaviour can help businesses target the right customers with the correct strategy.

Some limitations to this study are that we only looked at one retailers data and mostly from one country. The Elbow curve decreased gradually meaning the number of clusters was mostly confirmed by the Silhouette score. Future work could look at more clusters, compare K-Means with hierarchical clustering, look at other features such as product category and forecasting sales.

VII. REFERENCES

[1] A. M. Hughes, Strategic Database Marketing. Probus Publishing, Chicago, IL, USA, 1994.
[2] D. Chen, S. L. Sain and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining". Journal of Database Marketing & Customer Strategy Management, vol. 19, no. 3, pp. 197-208, 2012.
[3] 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.
[4] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm". Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 1785-1792, 2022.
[5] "The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering". Mathematical Modelling of Engineering Problems, vol. 10, no. 1, 2023.
[6] "Exploring Customer Segmentation in E-Commerce using RFM Analysis with Clustering Techniques". Journal of Telecommunications and the Digital Economy, vol. 12, no. 3, 2024.
[7] D. Chen, "Online Retail II", UCI Machine Learning Repository, 2019. [Online]. Available: https://archive.ics.uci.edu/dataset/502/online+retail+ii
[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, Vol. 1, 1967, pp. 281-297.
[9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987.
[10] R. L. Thorndike, "Who belongs in the family?", Psychometrika, Vol. 18, No. 4, pp. 267-276, 1953.
[11] F. Pedregosa et al., "Scikit-learn: Machine learning in Python". Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[12] W. McKinney, "Data structures for statistical computing in Python", in Proc. 9th Python in Science Conf., pp. 56-61, 2010.
[13] J. D. Hunter, "Matplotlib: A 2D graphics environment". Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
[14] A. K. Jain, "Data clustering: 50 years beyond K-means". Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.
[15] C. R. Harris et al., "Array programming with NumPy". Nature, vol. 585, pp. 357-362, 2020.