1. **ABSTRACT**

This study uses an *RFM (Recency, Frequency, Monetary) analysis model* on customer transaction data from an online retail store to categorize customers according to their purchasing patterns. The project uses *K-means clustering and decision tree* to identify important customer clusters, which informs strategic marketing interventions, after preprocessing and transforming the data. The model's insights are intended to enhance personalized marketing, pinpoint high-value markets, and maximize customer retention. The analysis is backed by academic frameworks and industry benchmarks and is based on the increasing importance of customer-centric strategies in the online retail sector. A set of recommendations is further provided on consumer-centric marketing. *Python and Jupyter Notebook* have been used in the study.

1. **INTRODUCTION**

Most businesses earlier used a product-centric strategy as a marketing strategy, which mainly focused on the manufacturers to create a better product and reduce the manufacturing costs rather than paying attention to the customers who used them. However, the latter half of the 20th century was the era of customer-centric marketing, which shifted from focusing on product design and delivery to one focused on individuals as customers.

In today's cutthroat online marketplace, customer retention and segmentation are essential to online retailers' success. Businesses now have the chance to use data analytics to gain a deeper understanding of customer behavior thanks to the explosive growth of e-commerce and the wealth of transactional data.

Personalized account data, comprehensive order and delivery records, and real-time tracking of customer behavior are just a few of the distinct benefits that online shopping offers over traditional retail. These characteristics have given internet merchants the ability to implement customer-centric tactics that emphasize unique tastes and habits.

Businesses are increasingly *using data mining techniques to answer important business questions, such as identifying high-value customers, comprehending loyalty patterns, and forecasting purchase behavior.* Customer lifetime value and RFM (Recency, Frequency, Monetary) are two popular models that assist businesses in making focused, well-informed decisions.

This study uses RFM analysis in conjunction with clustering techniques and decision tree to group customers with similar characteristics, enabling more effective customer relationship management and strategic marketing actions. Accordingly, a set of recommendations has been further provided on consumer-centric marketing.

The ***objectives*** of our research are:

* Which customers are the most and least valuable to the business? What are their characteristics?
* Who are the most / least loyal customers, and how are they characterized?
* What are the sales patterns in terms of various perspectives such as products / items, regions and time (weekly, monthly, quarterly, yearly and seasonally)

The rest of this paper is organized as follows:

*Section 3* - Background information about the online retailer studied along with the

associated dataset to be explored.

*Section 4* - The main steps and tasks for data pre-processing in order to create an appropriate target dataset for the required further analyses.

*Section 5* - k -means clustering analysis was conducted, and a set of meaningful clusters and segments of the target dataset were identified. Each cluster was thoroughly discussed, and decision tree induction was used to further refine the segmentation.

*Section 6* - Based on the findings of the analysis, the penultimate section provides a summary of the key consumer-centric business intelligence and offers specific suggestions to the online retailer with the goal of increasing company profits.

*Section 7* - The final section contains the closing thoughts.

1. **BUSINESS BACKGROUND AND THE ASSOCIATED DATA**

The *online retailer* examined in this study is a non-store, *UK-based company* that was established in 1981 and initially focused on selling distinctive gifts for all occasions via phone and direct mail. The business made a complete switch to e-commerce two years ago, launching its own website and complemented by sales through Amazon.co.uk. Since then, it has amassed a sizeable clientele in the UK and Europe, accumulating a wealth of transactional data that can be used to generate analytical insights.

The customer transaction data set held by the merchant has *8 variables* as shown in ***Table 1***, and it contains all the transactions occurring in the *years 2010 and 2011*. The dataset follows a sequential order (time-dependent) and includes *integer and real-valued numerical data and textual data.* There were *541,910 valid instances (record rows)* in total (including all countries), each for a particular item contained in a transaction.

Only consumers from the *United Kingdom were analyzed* and there were *495,478 valid instances and 3951 unique customers*. It is interesting to notice that the *average* number of products purchased by each unique customer during 2010 and 2011 was ***125*** (495,478/3951). This seems to *suggest that many of the consumers of the business were organizational customers* rather than individual customers.

A close-up of a list

AI-generated content may be incorrect.

1. **DATA PRE-PROCESSING**

The original dataset must be pre-processed in order to perform the necessary RFM model-based clustering analysis. The main steps and relevant tasks involved in data preparation were as follows:

1. Relevant variables were chosen from the data set. The seven variables that were selected for our case are *Invoice, StockCode, Quantity, Price, Country, InvoiceDate, and Customer ID.*
2. Removed rows with *missing CustomerID values*, as these are crucial for customer-level analysis.
3. To maintain data completeness and ensure consistency during analysis, missing values in the *'Description' column were replaced with the placeholder text 'no description'*, allowing each transaction record to retain a value in that field even when product details were unavailable.
4. Created an aggregated variable named *Amount* , by multiplying Quantity with Price , which gives the total amount of money spent per product / item in each transaction.
5. Converted InvoiceDate *from string to datetime* format for accurate calculation of recency.
6. Separated the variable *InvoiceDate into two variables Date and Time* . This allows different transactions created by the same consumer on the same day but at different times to be treated separately.
7. Retained only *records with positive quantity and price* to reflect genuine purchases.
8. 'Customer ID' column was converted *from a float to an object (string) data type* to accurately represent it as a categorical identifier rather than a numeric value, ensuring proper handling during grouping, segmentation, and customer-level analysis.
9. Filtered out the transactions for *UK customers*.
10. Used the most recent invoice date in the dataset to *compute recency* (days since last purchase), *frequency* (number of transactions), and *monetary* (total spend).
11. Applied *Min-Max Scaling to standardize* the RFM variables before clustering.

Following the above steps a target dataset for the analysis was generated. Part of the target dataset is shown in ***Figure 1***, and the variables in the target dataset and their statistics are described in ***Tables 2 and 3***. Python codes were used to transform the dataset and to calculate the values for the variables Recency, Frequency and Monetary, for each given CustomerID,

respectively.

**Figure 1: Samples of the target dataset**

A screenshot of a spreadsheet

AI-generated content may be incorrect.

A table of data type

AI-generated content may be incorrect.

A table with numbers and a number of data

AI-generated content may be incorrect.

**5. RFM MODEL BASED CLUSTERING**

The Recency, Frequency, and Monetary (RFM) model is a well-established framework in customer analytics, used to understand and segment customers based on their purchasing behavior. In this study, we applied the RFM model to the *Online Retail II* dataset using CustomerID as the unique identifier.

* *Recency (R)* refers to how recently a customer made a purchase. It is calculated by measuring the number of days between the customer’s most recent transaction and the reference date (typically the latest transaction date in the dataset).
* *Frequency (F)* measures how often a customer has purchased over the period. It reflects customer loyalty and engagement.
* *Monetary (M)* quantifies the total amount spent by a customer, representing their financial value to the business.

After preparing the final dataset, the objective was to determine whether customers could be meaningfully segmented based on their recency, frequency, and monetary values. To achieve this, we applied the K-Means clustering algorithm using Python. Constructing the RFM table involved transforming raw transaction data into three key variables: **Recency**, **Frequency**, and **Monetary**. Frequency was calculated by counting how many times a customer made purchases, based on invoice records. Recency was computed as the number of days since a customer’s most recent purchase relative to a reference date. Monetary value was derived by summing up the total purchase amounts made by each customer over the selected time period.

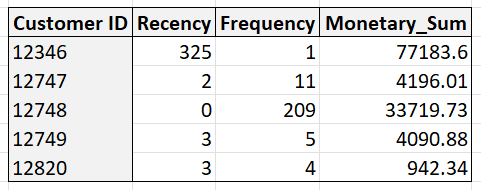


Table 4: Output for RFM values generated for each Customer.

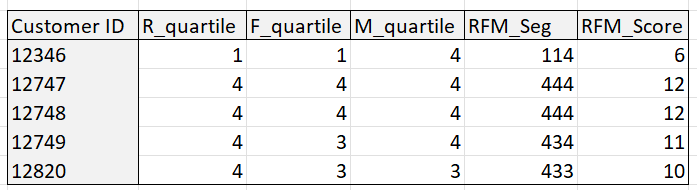


Table 5: Output for RFM quartile for the RFM scores

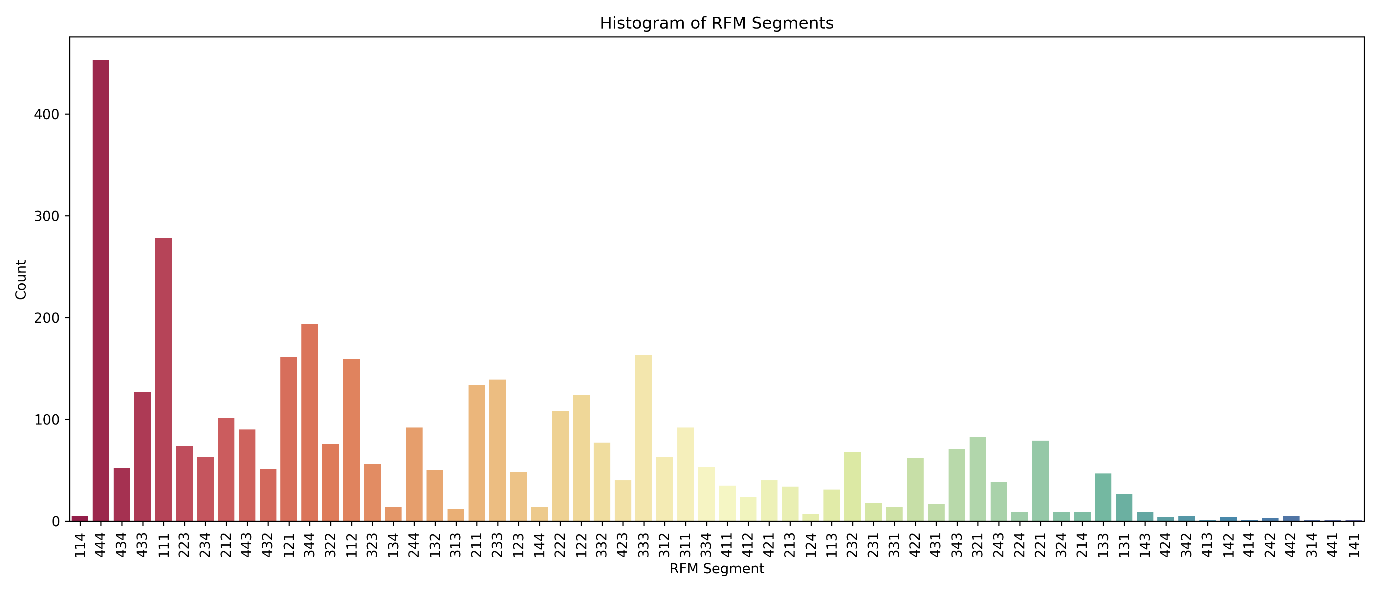


Figure 2. Histograms for Segmentation for RFM Groups

To gain deeper insights, we categorized customers into different tiers based on their total RFM Scores, ranked from highest to lowest. Following this, we applied K-Means Clustering, an unsupervised learning technique, to further segment the customer base. Clustering involves grouping similar data points together, and in this context, it helps identify distinct customer profiles.

Before applying the K-Means algorithm, we ensured that the RFM features—Recency, Frequency, and Monetary—were appropriately prepared. Since K-Means assumes the input variables are normally distributed, we examined each feature for skewness. When skewness was detected, particularly in Recency or Monetary values, we applied logarithmic transformation to normalize the data. The transformed distributions of the RFM features are shown in the following graphs.

Figure 3. Positive (Right) Skewness graph of the Frequency value.

Figure 4. Positive (Right) Skewness graph of the Recency value.

Figure 5. Positive (Right) Skewness graph of the Monetary value.

From the density plots shown earlier, it is evident that the Recency and Monetary variables exhibit positive (right) skewness. To address this, we applied a logarithmic transformation, which helps in normalizing the data and improving the performance of clustering algorithms like K-Means.

The next step in our analysis involved determining the optimal number of clusters for K-Means segmentation. Choosing the right value of *k* is a critical aspect of partition-based clustering. Several methods have been proposed in the literature for this purpose, and among them, the Elbow Method is one of the most widely used.

In this technique, we plot the within-cluster sum of squares (WCSS) for different values of *k*. The point at which the curve starts to flatten (i.e., the "elbow") indicates the ideal number of clusters. It reflects a balance between model complexity and explained variance. The elbow point for our dataset is illustrated in the following chart.

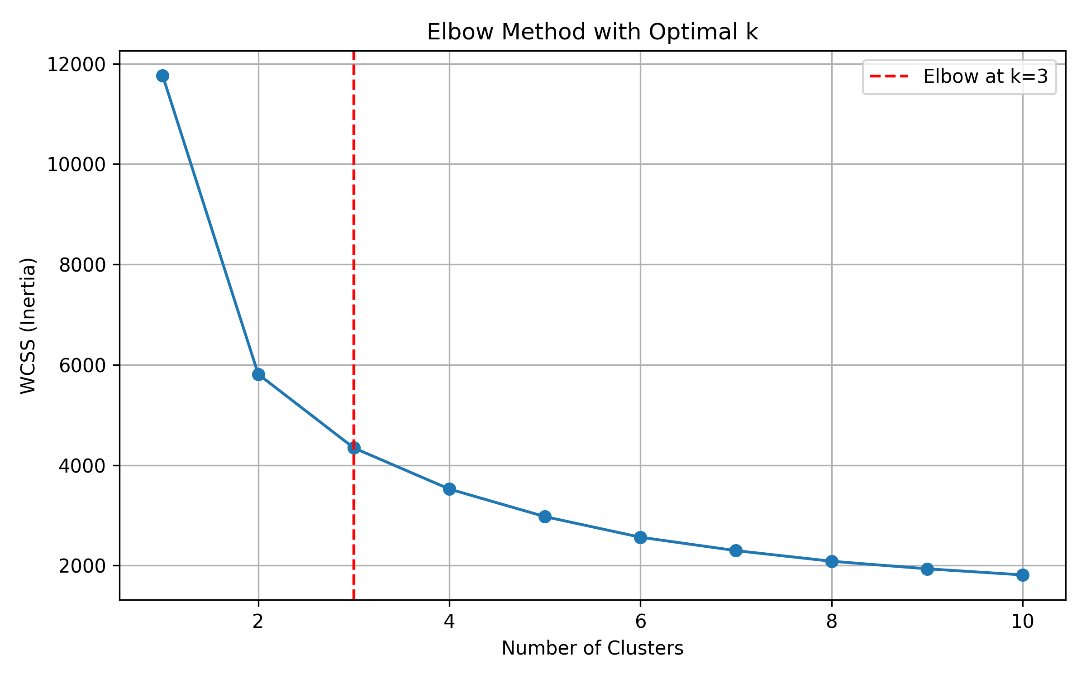


Figure 6. Elbow Method for finding number of clusters

We identified 3 as the optimal number of clusters for our dataset. The table below presents the clustering results and the distribution of customer profiles across each segment, offering a clear interpretation of the data based on their purchasing behavior.

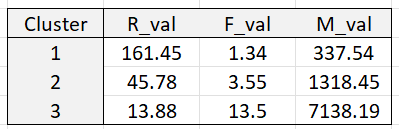


Table 6: Cluster Summary

Cluster 1

* Recency: 161.45 — Customers made their last purchase a long time ago.
* Frequency: 1.34 — These customers purchase very infrequently.
* Monetary: 337.54 — Their total spending is relatively low.

Cluster 2

* Recency: 45.78 — Customers made a purchase moderately recently.
* Frequency: 3.55 — These are occasional buyers.
* Monetary: 1,318.45 — Their spending is moderate.

Cluster 3

* Recency: 13.88 — Customers purchased very recently.
* Frequency: 13.50 — These are frequent buyers.
* Monetary: 7,138.19 — These customers are high spenders.

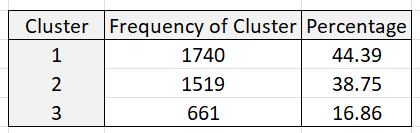


Table 7: Instances in each cluster

**Understanding the Clusters**

Interpreting the composition and behavioural traits of each identified cluster is essential for generating meaningful, customer-centric business intelligence. As shown in Table 7, the dataset was segmented into three distinct clusters. Each cluster reflects a specific type of consumer behavior based on transaction volume and share of the total customer base.

*Cluster 1* – Broad Base of Moderate Shoppers

This cluster contains 1,740 customers, which accounts for 44.39% of the total population making it the largest group. While they may not exhibit extreme values in recency or monetary contribution, their large numbers suggest a steady revenue base for the business. These customers may represent consistent, though not high-value, purchasing behavior. They are ideal candidates for upselling or loyalty programs to improve their overall contribution.

*Cluster 2* – Strong Contributors with Growth Potential

With 1,519 customers (38.75%), this segment reflects a substantial portion of the customer base. Their shopping behavior likely includes moderate-to-high frequency and spending, suggesting they are engaged and responsive to marketing efforts. While they may not be as active as the top segment, their potential to move into the highest-value group is significant. Focused campaigns and personalized offers may further enhance their lifetime value. It is ideal for decision tree refinement due to its substantial size (~39%) and diverse behavior across recency, frequency, and spending. This allows for identifying high-potential customers, low-engagement outliers, and emerging loyalists.

*Cluster 3* – High-Value, Loyal Customers

Comprising only 661 customers (16.86%), this is the smallest but most impactful segment. Customers in this cluster likely show very recent activity, frequent purchases, and high monetary value — traits that define them as top-tier or VIP customers. Despite being a minority in size, their contribution to total revenue is likely the highest. Retention of this group is critical, and they should be prioritized for exclusive promotions, early access to products, and premium engagement strategies.

**Enhancing Clustering Analysis Using Decision Tree**

As previously discussed, Cluster 2 emerged as one of the most behaviorally diverse groups in our segmentation analysis. This cluster includes customers with varied purchasing patterns and monetary contributions, making it a strong candidate for deeper sub-segmentation. To refine the understanding of customer behavior within this group, we applied a decision tree regression model using Frequency and Recency as predictors and Monetary value as the target variable.

The resulting tree, illustrated in Figure X, reveals a clear structure of nested customer segments. For instance, the first major split occurred at Frequency ≤ 2.5, separating low-activity customers from more frequent buyers. Among the high-frequency group, additional splits based on Recency and higher Frequency thresholds further distinguished customers by their spending potential.

The model uncovered sub-groups such as:

* Customers with Frequency > 2.5 and Recency ≤ 10, showing average monetary values above $19,000.
* Conversely, customers with low frequency and older recency tended to spend below $4,000 on average.

This tree structure not only provides interpretable insights but also supports a monotonic positive relationship between frequency and monetary value, aligning with expected consumer behavior. By segmenting within Cluster 2, the business can now design more precise strategies, such as identifying potential VIPs or targeting mid-tier customers for upsell opportunities.

**6. RECOMMENDATIONS**

The clustering and decision tree analyses enabled a deeper understanding of customer behavior, revealing clear distinctions in purchasing patterns across segments. Notably, Cluster 2 showed significant internal diversity, which was further explored through decision tree regression. This sub-segmentation highlighted variations in frequency and monetary value, allowing the identification of high-potential customers, occasional buyers, and valuable loyalists. These insights form a solid foundation for customer-centric strategies that prioritize retention, growth, and revenue optimization.

To operationalize these insights, the business should implement targeted marketing campaigns tailored to each sub-segment’s behavior. High-frequency, high-spend customers can be nurtured through loyalty programs and exclusive offers, while moderate buyers could be incentivized with personalized promotions to increase engagement. Further exploration of product purchase patterns—such as commonly bought items or seasonal trends—can enhance cross-selling strategies. Additionally, distinguishing between individual and organizational customers, predicting future high-value customers, and examining potential geographic influences will allow the business to craft more personalized and impactful customer experiences.

**7. CONCLUSION**

This study demonstrates how customer segmentation through RFM analysis and k-means clustering, followed by refinement using decision trees, can offer actionable insights into consumer behavior. These insights support the development of targeted marketing strategies and customer relationship initiatives.

The project underscores the importance of data preparation and model interpretation, which are often the most time-intensive stages but critical to achieving business value. Looking forward, applying association rule mining and lifecycle value prediction can further deepen understanding of consumer behavior.

By adopting these data-driven practices, the business can shift toward more personalized, efficient, and profitable customer engagement.

**8. REFERNCES**

**1. Chen, D., Sain, S. L., & Guo, K.** (2012). *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, 19(3), 197–208. <https://doi.org/10.1057/dbm.2012.17>

**2. Bhupathiraju, G. V., & Raghavendra, V. R. T. S.** (2022). *Data mining for the online retail industry: Customer segmentation and assessment of customers using RFM and k-means* [Student research project, Oklahoma State University].