Customer Segmentation using RFM Analysis

TEAM 7:
AARYA SANTOSH GADEKAR
ANDREW ROTH
POOJA PATIL

Table of Contents

Introduction

Data Overview

Customer Analysis

Product Analysis
4.1 Top-Selling Products
4.2 Revenue-Contributing Products

Time-Based Trends
5.1 Daily Trends
5.2 Hourly Trends
5.3 Monthly Seasonality

Geographical Insights
6.1 Orders by Country
6.2 Average Order Value by Region

Customer Behavior
7.1 Customer Lifespan
7.2 Purchase Patterns

Returns & Refunds Overview
8.1 Return Rates
8.2 Most Returned Items

Revenue Summary
9.1 Total Revenue
9.2 Revenue Distribution

RFM Analysis
10.1 Recency, Frequency, Monetary Scores
10.2 RFM Interpretation

Customer Segmentation
11.1 Clustering Approach
11.2 Cluster Profiles

Recommendations & Conclusion
12.1 Marketing Strategies
12.2 Operational Improvements
12.3 Final Summary

1. Introduction

This report summarizes an entire study of RFM (Recency, Frequency, Monetary) Analysis and customer segmentation based on exploratory data analysis performed on an e-commerce transaction dataset. This analysis includes assessment of customer's purchasing trends and patterns observed over a period of time. Customer behavioural insights are evaluated based on analysing the seasonality trends, product's performance across geographical aspects, returns/refunds assessment, RFM scoring and K-means segmentation.

RFM is a technique used to detect how recently a customer has bought some product or service, how often is the customer buying, and how much has the customer spent in total. Using these scores we can segment the buyers based on their purchasing habits and interests in order to predict future purchases and customize the marketing strategies.

2. Data Overview

The dataset is taken from Kaggle, it had nearly 53K rows and 8 columns after cleaning. The time frame ranged from December 2010 to December 2011. We identified 4373 unique customers in total. The dataset consisted of attributes like InvoiceNo, StockCode, Description, CustomerID, Country, etc. The data is preprocessed by handling the missing values in fields like CustomerID and Description, removing duplicates, and converting relevant datatypes. A new column 'Revenue' is introduced by computing 'Quantity' x 'Unit Price' in order to understand the amount spent during each transaction. Additional features such as day of week, hour, and month were created from the 'InvoiceDate' column to support time-based analysis.

3. Customer Analysis

There is a huge chunk of data(nearly 25%) where customer ID is unknown. This means there are many such transactions which are made from un-registered accounts. This indicates that either this could be guest checkouts or internal system testing trial transactions. There is a good amount of diversity in the customer base due to which we are able to differentiate between high-frequency power buyers, occasional customers and one-time or inactive customers. The high concentration of orders from a small subset of customers indicates that a VIP segment significantly contributes to overall revenue.

4. Product Analysis

4.1 Top-Selling Products

If we observe the quantities of products sold, we understand that there are few goods that make up most of the sales. World War 2 Gliders Assorted Designs, with more than 53,000 units sold, is the most popular item in the entire dataset. Other products such as the Jumbo Bag Red Retrospot, Popcorn Holder, Assorted Colour Bird

Ornament, and Pack of 72 Retrospot Cake Cases also have a significant amount of purchases. These items could be categorised as generally low-priced and lightweight, and therefore supposedly purchased frequently either for gifting, decoration, or resale. This pattern suggests that everyday novelty and decorative goods drive the majority of unit sales.

4.2 Revenue-Contributing Products

The analysis of revenue highlights a notable difference between the products that sell the most and the ones that contribute the highest monetary value. The item generating the greatest total revenue is ‘Dotcom Postage’ accounting for over 206,000 of the amount. As per the e-commerce standards this entry generally does not represent a physical product; instead, it reflects postage or shipping charges applied to online customer orders. Because it appears frequently across many invoices, it accumulates a huge revenue. Following this, premium items such as the ‘Regency 3-Tier Cake Stand’ and the ‘White Hanging Heart T-Light Holder’ also contribute significantly to overall revenue as these are some expensive stuff. With our in-depth analysis we can conclude that when people buy cheap goods in large quantities, we may assume that these are the products that attract the most revenue, but this is not always true. Because there is a segment of expensive products which only a handful of elite customers could afford, but these products are the ones that drive the actual revenue.

5. Time-Based Trends

5.1 Daily Trends

Timely patterns uncover that Tuesdays and Thursdays experience large traffic of customers, while weekends, especially Sundays see lower activity. This indicates that customers are more active during weekdays, suggesting stronger engagement with the store during typical business days.

5.2 Hourly Trends

Hourly distribution reveals a spike in purchase activity from 10 AM to 3 PM, with the highest number of orders placed around midday. During the rest of the day, early mornings and late evenings show minimal activity.

5.3 Monthly Seasonality

As per the seasonality observed, there is a surge in the purchasing rate of customers during the month of November, mostly driven by holiday shopping(gifts and decoration). This particular trend indicates that there is a high rate of consumerism during pre-Christmas period.

6. **Geographical Insights**

6.1 Orders by Country

The geographical distribution of orders suggests that the country that dominates overwhelmingly is the United Kingdom, with approximately 23.5K unique orders, far higher than any other region. We can conclude that the majority of the data is UK-centric and the primary customer base is from the UK. European countries such as Germany (603 orders), France (461 orders), Ireland (360 orders), and Belgium (119 orders) are the next most active markets, though their contribution remains significantly smaller.

6.2 Average Order Value by Region

We can observe a particular pattern while assessing the average order value(AOV), countries with fewer total orders often have higher per-order expenditure. Regions such as the Netherlands, Australia, Japan, Sweden, and Denmark have the largest total amount spent transactions. This difference between how often customers order and how much they spend suggests that international shoppers may be buying fewer but higher-value items, possibly because of shipping costs or simply because they are more selective. These higher average order values show that there's real potential to focus more marketing efforts on these regions, where each purchase already carries more value.

7. **Customer Behavior**

7.1 Customer Lifespan

Customer lifespan tells us how long a person keeps buying from the store. In this dataset, some customers bought something only once, so their active time is just 0 days. Others kept buying for almost the whole year, with the longest lasting 373 days. On average, customers stay active for about 133 days. This shows that the store has a mix of one-time shoppers and loyal repeat buyers, with only a smaller group staying connected over a long period of time.

7.2 Purchase Patterns

Customers in this dataset show very different buying habits. Many people only placed one or two orders, while some people were frequent buyers. Some customers made purchases regularly and recently, while others went long stretches without buying anything. These patterns match the RFM results, which separate loyal and high-value customers from those who shop rarely or not at all. Overall, the data shows a mix of casual shoppers and highly active, repeat buyers.

8. Returns & Refunds Overview

8.1 Return Rates

About 20% of all orders in the dataset include returned items roughly one out of every five orders. That’s a fairly high return rate, suggesting that some customers may not be fully satisfied with what they received, or there may be issues with product quality, packaging, or order mistakes.

8.2 Most Returned Items

A few products are returned much more often than others. Items like Paper Craft Little Birdie, the Medium Ceramic Top Storage Jar, and several products marked as damaged, unsaleable, or with printing issues show very high negative quantities. These returns mostly revolve around fragile items, misprinted products, or items with quality problems. This indicates that certain product types may need better quality checks or improved packaging to reduce return rates.

9. Revenue Summary

9.1 Total Revenue

The total revenue for the period from December 2010 to December 2011 is $9,726,006.95. This value is calculated by multiplying the quantity of each product sold with its respective unit price and summing the revenue across all transactions. This data underlines a robust revenue foundation supported by a regular order flow throughout the year and especially during the peak shopping months like November. All in all, this revenue figure reflects extensive sales volume and a wide range of products purchased across various regions.

9.2 Revenue Distribution

The total revenue for December 2010 to December 2011 stands at $9,726,006.95. This is obtained by multiplying the quantity of the product sold by the unit price and summing up revenues in all transactions. This information underlines a strong revenue base supported by a regular order flow throughout the year and particularly in peak shopping months such as November. All in all, this revenue figure reflects extensive sales volume and a wide range of products purchased across various regions.

10. RFM Analysis

10.1 Recency, Frequency, Monetary Scores

To calculate RFM, we used the Columns InvoiceDate, InvoiceNo, and UnitPrice. Recency is defined as days since last purchase, frequency is the number of purchases, and monetary is the unit price total. We then ranked customers, grouped by CustomerID, and assigned them a quartile for each of the RFM categories. A total

RFM score was also calculated in RFM Segmentation; however, the individual scores were more useful for clustering.

10.2 RFM Interpretation

Recency is sorted by days since last purchase, with a 4 representing the 25% of customers who had made purchases most recently and a 1 for those who had rarely made purchases, with scores of 3 and 2 for the middle 50% of customers. For frequency, a 4 represents the top 25% of purchases, while a 1 is the bottom 1% of purchases. Lastly, a 4 for monetary represents the top 25% spending customers, while a 1 for monetary shows those who had spent the least.

11. Customer Segmentation

11.1 Clustering Approach

Segmentation of customers was done by using the RFM scoring framework, where each customer was rated on three dimensions: Recency, Frequency, and Monetary. After calculating individual R, F, and M scores, standardizing the values provided inputs for running K-Means to find distinct groups of customers. The optimal number of clusters was determined using the Elbow Method, with curve analysis indicating that three clusters (k = 3) best described the meaningful segmentation. Experiments were repeatedly run for 2, 3, and 4 clusters; segmentation with three clusters yielded sharp separation between every group of customers and also most closely mirrored the natural distribution of the RFM score. This method enabled customers to be segmented into meaningful behavioral categories based on their purchasing habits.

11.2 Cluster Profiles

The final segmentation yielded three clear customer clusters, each representing different purchasing behaviours.

- Cluster 0 consists of the most valuable customers, showing the highest average Recency, Frequency, and Monetary scores. Customers in this cluster purchase often, spend more, and have bought recently, making them strong revenue contributors for the store.
- Cluster 1: These are the mid-tier customers who buy occasionally, spend a moderate amount, and neither fall into highly frequent nor into inactive categories. They have the potential to move upward into the high-value segment with targeted engagement.
- Cluster 2 represents the least active customers, characterized by low frequency, low spending, and longer gaps since their last purchase. They make up the largest part of the customer base but contribute minimally in terms of revenue.

This segmentation offers a clear behavioral structure that can further enhance personalized marketing strategy, targeted offering, and long-term customer relationship planning.

12. Recommendations & Conclusion

12.1 Marketing Strategies

Targeted marketing strategies, based on customer segmentation and RFM insights, can yield a much improved customer engagement and increased revenue. Applying loyalty programs, VIP-exclusive offers, and personalized appreciation messages can help reinforce the purchasing behavior of high-value customers in Cluster 0 and create long-term retention. Mid-tier customers, or Cluster 1, can be further triggered toward becoming high-value buyers by using limited-time promotions, incentives to increase spending, and notifications that highlight progress toward VIP status. Lastly, low-engagement customers in Cluster 2 can be targeted with broader awareness campaigns, free shipping periods, and introductory discounts that might help rekindle interest in making repeat purchases. Such tiered strategies will enable the business to tailor marketing efforts to the behavior and potential value of each customer group.

12.2 Operational Improvements

Operational improvements can further fortify customer experience and support further sales growth. As a significant portion of revenues come from a limited number of high-value products and shipping charges, inventory planning to stock more premium items and ensuring that the logistics process is efficiently managed will have direct implications for profitability. Moreover, the high returns of certain product categories indicate that there is a need to focus more on quality control or product descriptions/packaging in order to minimize dissatisfaction. The registration process should also be improved, which will cut the high level of "Unknown" customer IDs, thus enabling better tracking and effective personalization. These will jointly help reduce operational inefficiencies and increase overall customer satisfaction.

12.3 Final Summary

The e-commerce dataset offers a good view of customer behavior, the distribution of revenues, the performance of the products, and purchasing trends. Using an RFM-based segmentation and clustering approach resulted in three clear groups of customers that differed significantly from one another in recency, frequency, and spending. Revenue analysis showed how much power lies with premium products and service charges when it comes to bringing in returns, and time-based trends identify some key seasonal and hourly activity patterns. Geographical insights further demonstrate market potential beyond the core base of customers in the UK. Combined, these findings provide for strategic decisions on marketing, operations, and customer engagement. A focus on high-value segments, leveraging improvements in product and process quality, and targeted promotions would be better at increasing customer retention, driving sales, and fueling business growth in a more sustainable manner.