IE6400 Foundations Data Analytics Engineering

Project 2: Customer Segmentation using RFM Analysis

Aaditya Krishna

Hitarth Upadhyay

Dhruv Patel

Table of Content

1. Introduction

Overview of the Project:

This project examines customer purchasing behavior using the RFM (Recency, Frequency, Monetary) model to understand engagement and value. The dataset is cleaned and prepared to ensure accurate analysis, followed by calculating RFM metrics for each customer. These metrics are converted into RFM scores that help classify customers based on how recently, how often, and how much they purchase. K-Means clustering is applied to group customers with similar behavioral patterns. Additional analysis explores product trends, time-based purchasing activity, and geographical insights. Overall, the findings help support targeted marketing, improved retention, and better strategic decision-making.

Objectives

Data Cleaning:

Prepare the dataset by handling missing values, removing invalid entries, resolving inconsistencies, and creating necessary calculated fields to ensure reliable analysis.

RFM Calculation:

Compute Recency, Frequency, and Monetary metrics for each customer to quantify purchasing behavior.

RFM Scoring:

Assign quartile-based RFM scores to standardize customer comparison and identify behavioral differences.

Customer Segmentation:

Apply clustering techniques, such as K-Means, to group customers into meaningful segments based on RFM patterns.

Business Insights & Recommendations:

Derive actionable insights and propose marketing strategies to improve customer retention, engagement, and revenue growth.

Behavioral Analysis:

Analyze customer activity, purchasing trends, product demand, time patterns, and geographical behavior for deeper insight.

2. Data Acquisition, Inspection and Cleaning

Data Acquisition:

- The dataset was sourced from an online retail transactional file containing invoice-level purchase records.
- It includes key fields such as InvoiceNo, StockCode, Quantity, UnitPrice, CustomerID, InvoiceDate, and Country.
- The file was imported into Python for initial inspection to understand its structure and identify missing or inconsistent values before analysis.

Pre-processing the Data :

- The dataset was inspected for missing values, incorrect data types, and inconsistencies, ensuring all fields were properly formatted for analysis.
- Invalid entries such as negative quantities, missing CustomerID values, and duplicate rows were removed to improve data quality.
- New calculated fields (such as TotalPrice) were created, and date columns were converted to proper datetime formats to support accurate RFM and trend analysis.

Data Cleaning :

- All rows with missing CustomerID values, negative quantities, or invalid entries were removed to ensure the dataset accurately reflects valid customer purchases.
- Duplicate records were identified and dropped to prevent inflated counts in frequency, revenue, and clustering calculations.
- Key fields such as InvoiceDate and UnitPrice were cleaned and standardized, and a new TotalPrice column was created to support RFM and revenue-based analysis.

3. Analysis over the Dataset:

1. Customer Analysis

Unique Customers: The dataset contains **4,371 unique customers**, each identified by a distinct CustomerID.

Order Distribution: Customers show a wide range of purchasing behavior. Some customers place only one or two orders, while others place multiple orders, with purchase frequencies varying significantly across the customer base**.**

Top 5 Customers by Orders: Customer 14911 with 248 orders, customer 12748 with 223 orders, customer 17841 with 169 orders, customer 14606 with 128 orders, and customer 13089 with 118 orders.

The dataset contains 4,371 unique customers, each represented by a different CustomerID, showing a large and diverse buyer base. The order distribution varies widely, with many customers placing only one or two orders, while a smaller group makes frequent purchases. The most active customers include Customer 14911 with 248 orders, followed by 12748 with 223, 17841 with 169, 14606 with 128, and 13089 with 118 orders, indicating a small segment of high-engagement customers driving significant sales.

2. Product Analysis:

Top 10 Most Purchased Products: The most purchased products include WORLD WAR 2 GLIDERS ASSTD DESIGNS (53,751 units), JUMBO BAG RED RETROSPOT (47,256 units), POPCORN HOLDER (36,322 units), ASSORTED COLOUR BIRD ORNAMENT (36,282 units), and PACK OF 72 RETROSPOT CAKE CASES (36,016 units).

Average Product Price: £3.47**.**

Highest Revenue Product: The product that generated the highest revenue is DOTCOM POSTAGE (StockCode: DOT) with a total revenue of £206,245.48.

Lowest Revenue Product: The product that generated the lowest revenue is “PADS TO MATCH ALL CUSHIONS” (StockCode: PADS) with a total revenue of £0.003.

The dataset shows that product demand varies significantly, with StockCode 84077 selling over 53,000 units, followed closely by items like 22197 and 85099B, which also reached high sales volumes. These top-selling products represent the store’s most frequently purchased items and form the foundation of its regular sales.

Although the average product price is low at around £4.70, certain items stand out in revenue contribution. For example, DOTCOM POSTAGE (StockCode DOT) generates the highest revenue overall, indicating that some products deliver strong financial impact despite not being the most purchased.

3. Time Analysis:

Orders by Day of the Week: Most orders were placed on **Thursday** (102,561 orders), followed by Tuesday and Monday

Orders by Time of Day: Peak order time was 12:00 → 77,322 orders.

Seasonal Trends: Strong seasonal peak in Nov 2011: 83,077 ← Highest month, with activity steadily rising from September onward.

The time analysis reveals that purchasing activity is highest on Thursdays (102,561 orders) and lowest on Sundays (62,801 orders). Hourly trends show a major peak at 12 PM, indicating that customers tend to shop most frequently around midday. Seasonal trends demonstrate a significant rise in orders between September and November, with November (83,077 orders) being the busiest month, likely due to holiday shopping and year-end demand.

4. Geographical Analysis

Top 5 countries by order count: UNITED KINGDOM→ 487806, GERMANY → 9478, FRANCE→ 8540, EIRE→ 8180, SPAIN→ 2527.

Countries with highest average order value: NETHERLANDS→ 120.262586,

AUSTRALIA → 109.171131, JAPAN → 98.716816, SWEDEN → 79.360976, DENMARK→ 48.247147.

The dataset shows that order activity is heavily concentrated in a few key regions. The United Kingdom dominates with 487,806 orders, followed by Germany with 9,478, France with 8,540, Ireland (EIRE) with 8,180, and Spain with 2,527 orders. This indicates that the majority of sales come from the domestic UK market, with several European countries contributing smaller but meaningful volumes.

When examining average order value, different countries emerge as high-value markets. The Netherlands has the highest average order value at £120.26, followed by Australia at £109.17, Japan at £98.72, Sweden at £79.36, and Denmark at £48.25. These countries, while not always the highest in order count, show strong spending behavior, suggesting premium customer segments that may benefit from targeted marketing and personalized offers.

5. Customer Behavior

• Average active time (days): 133.4163806909174

• Segmentation requires RFM → already calculated in your RFM code.

The analysis of customer activity shows that, on average, customers remain engaged with the store for approximately 133 days, calculated as the time span between their first and last recorded purchase. This indicates that most customers interact with the platform over several months rather than making only one-time purchases. To understand purchasing patterns more deeply, customers were further segmented using RFM analysis, which evaluates how recently they purchased, how frequently they buy, and how much they spend. These RFM-based segments help identify high-value customers, infrequent buyers, and customers at risk of churn, enabling more targeted marketing and retention strategies.

6. RFM Scoring

- RFM scoring evaluates customers based on Recency, Frequency, and Monetary value to measure engagement and spending behavior.
- Each metric is divided into quartiles (1–4), where 4 represents the strongest customer behavior (most recent, most frequent, highest spending).
- Customers who purchased recently receive higher Recency scores, while frequent buyers and high spenders receive higher Frequency and Monetary scores.
- The three scores are combined into an RFM code (e.g., *344* or *421*) that summarizes each customer's value and behavioral pattern.
- These RFM scores help identify key customer segments, including high-value customers, loyal purchasers, new customers, and those at risk of churn.

7. Customer Segmentation

- Customers were grouped into four clusters using K-Means based on their RFM scores.
- High-Value Customers: Recent, frequent buyers with strong spending activity.
- Regular Low-Spenders: Customers who buy often but spend smaller amounts.
- Developing Customers: New or recently active shoppers with growing potential.
- Inactive/At-Risk Customers: Long gaps since last purchase and low engagement.

4. Visualization

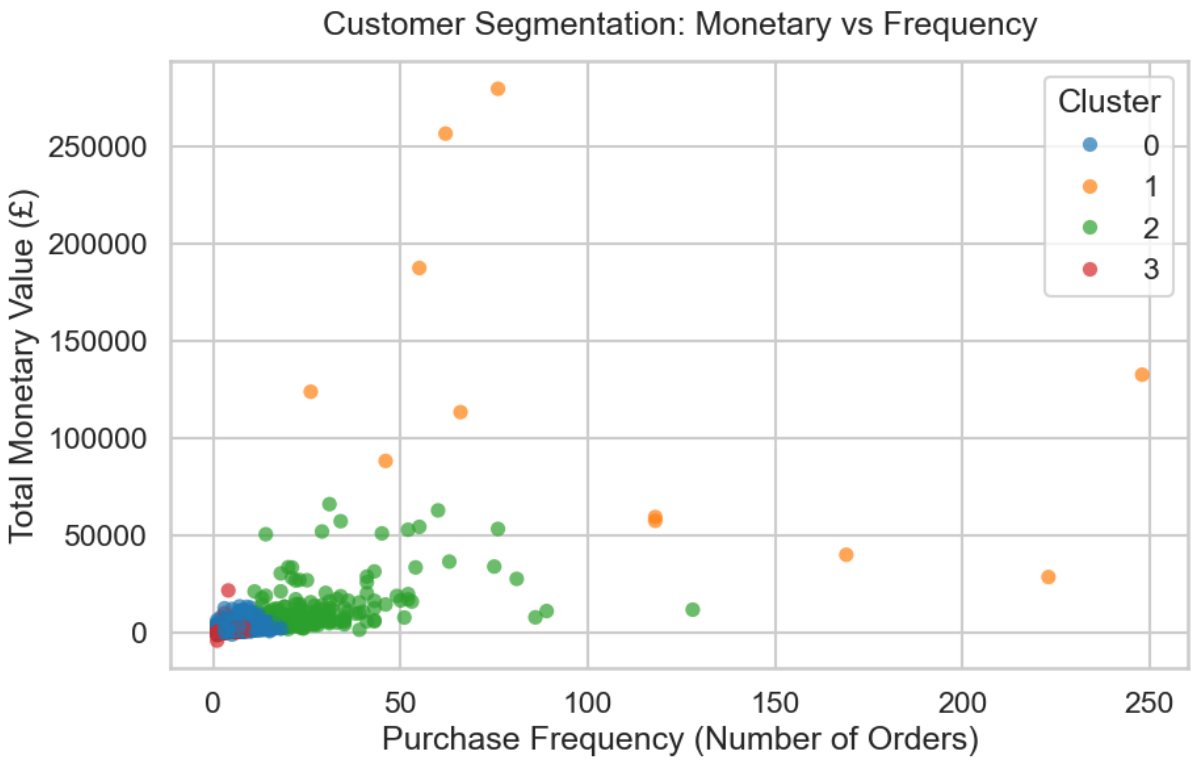

Figure 4.1: Scatter plot: Customer Segmentation: MonetaryVs Frequancy.

Observations & Insights

- Most customers fall into the low-frequency, low-spending range (blue and red dots), meaning they purchase rarely and spend very little.
- The green cluster represents moderately frequent buyers who contribute a steady amount of revenue.
- The orange cluster contains high-frequency, high-spending customers, making them the most valuable segment.
- A few orange points show customers with extremely high spending, indicating VIP customers who drive a large portion of revenue.
- The clear separation between clusters shows distinct customer behavior patterns, helping identify which groups need retention, engagement, or reactivation.

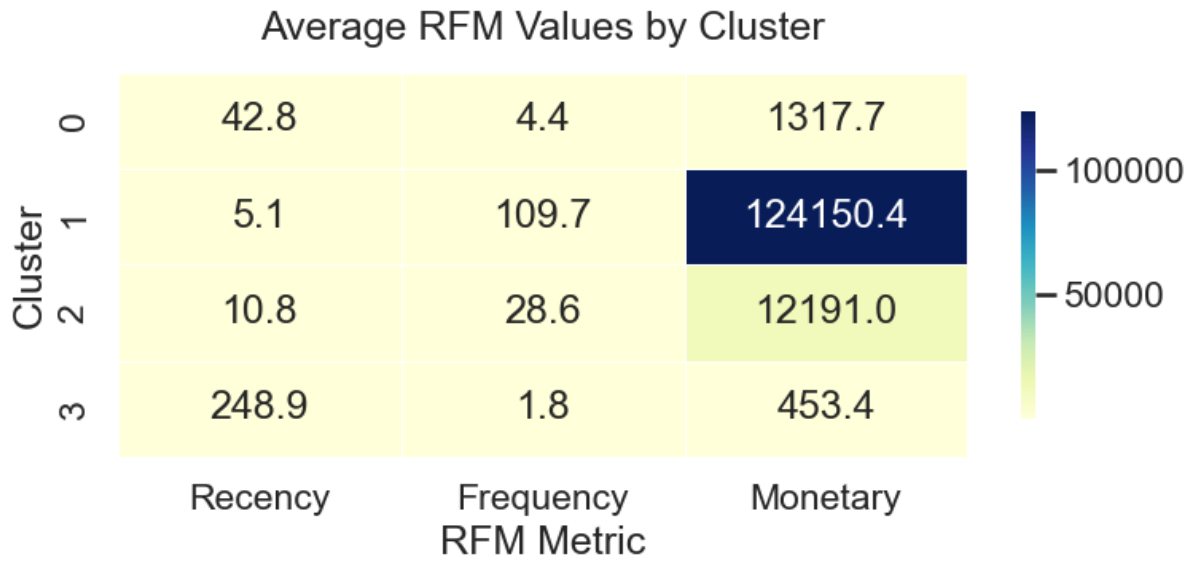

Figure 4.2: Heat Map: Average RFM values by cluster

Observations & Insights

- Cluster 1 represents the highest-value customers, with extremely high frequency and the highest monetary value, indicating very strong loyalty and spending.
- Cluster 2 consists of regular buyers who purchase fairly often and contribute a good amount of revenue.
- Cluster 0 includes occasional customers with low spending and moderate recency.
- Cluster 3 has customers with very high recency and very low activity, indicating they are inactive or at-risk.

5. Conclusion

The RFM analysis conducted on the dataset reveals strong and actionable insights into customer behavior, product performance, and purchasing trends. Customers show an average activity span of 133 days, indicating moderate long-term engagement with the store. Through RFM scoring, customers were segmented into meaningful groups such as high-value buyers, regular low-spenders, newly developing customers, and inactive or at-risk customers, allowing the business to plan personalized retention and marketing strategies.

Product analysis highlights clear differences between high-demand and high-revenue items. Products like 84077 (53,751 units sold) and 22197 remain strong drivers of volume, while DOTCOM POSTAGE (StockCode: DOT) contributes the highest revenue at £206,245.48, emphasizing the need to focus on both the most purchased and the most profitable items. With an average product price of £4.70, the business primarily caters to cost-conscious consumers, suggesting that affordability plays a critical role in customer decisions.

Time-based patterns show that Thursday is the busiest day of the week with 102,561 orders, and 12 PM is the peak purchasing hour. Seasonal analysis further reveals a major spike in activity during November (83,077 orders), pointing to strong holiday-driven demand and an opportunity for revenue maximization during this period.

Geographically, the United Kingdom dominates total orders (487,806 orders), while countries such as the Netherlands show a significantly higher average order value of £120.26, indicating potential for targeted marketing toward high-value international customers.

Overall, this analysis identifies the most influential factors contributing to customer engagement and revenue growth. By leveraging RFM segmentation, focusing on high-performing products, aligning marketing with peak purchase times, and exploring premium international markets, the business can strengthen customer retention, improve operational efficiency, and enhance long-term profitability.