

# Customer Segmentation Using RFM Analysis in E-Commerce

Reet Chandra

# Problem Statement and Objective

## Problem Statement:

E-commerce businesses generate vast amounts of transactional data but often struggle to leverage it for actionable customer segmentation. Understanding purchasing patterns is crucial for predicting behavior, improving retention, reducing churn, and maximizing customer lifetime value. A structured approach is needed to identify high-value customers, detect trends, and optimize targeted marketing strategies.

## Objective:

- Utilize the RFM (Recency, Frequency, Monetary) framework to segment e-commerce customers.
- Analyze purchasing patterns to identify high-value customers and at-risk segments.
- Create targeted marketing strategies to improve customer retention and satisfaction.

# Dataset Overview

- **Source:** [[Kaggle - E-commerce Dataset](#)]
- **Size:** (541909, 8)
- **Timeframe:** 01/12/2010 to 09/12/2011

## Key Variables:

- **InvoiceNo:** Unique transaction ID
- **StockCode:** Unique product identifier
- **Description:** Product description
- **Quantity:** Number of items purchased
- **InvoiceDate:** Timestamp of purchase
- **UnitPrice:** Price per unit in GBP
- **CustomerID:** Unique customer identifier
- **Country:** Location of the customer

# Data Cleaning and Preparation

1. Handled Missing Values
2. Removed Duplicates
3. Removed Cancelled Orders
4. Removing non-product Stock-Codes

# Data Cleaning and Preparation

1. Handled Missing Values
2. Removed Duplicates
3. Removed Cancelled Orders
4. Removing non-product Stock-Codes

```
df.isnull().sum()
```

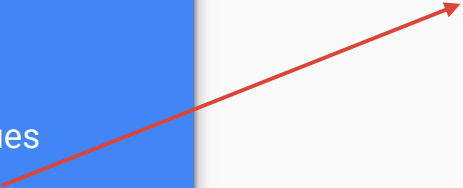
|             |        |
|-------------|--------|
| InvoiceNo   | 0      |
| StockCode   | 0      |
| Description | 1454   |
| Quantity    | 0      |
| InvoiceDate | 0      |
| UnitPrice   | 0      |
| CustomerID  | 135080 |
| Country     | 0      |

The percentage of missing values in the CustomerID column is **24.93%**.

Since the analysis will revolve around investigating customers and clustering them into categories, the missing values in the CustomerIDs were removed.

# Data Cleaning and Preparation

1. Handled Missing Values
2. Removed Duplicates
3. Removed Cancelled Orders
4. Removing non-product Stock-Codes



```
: # Check for duplicate rows
duplicates = df.duplicated()

# Count the number of duplicate rows
print(duplicates.sum())

5225
```

The number of duplicate rows in the dataset is **5525**.

These rows were removed from the dataset.

# Data Cleaning and Preparation

1. Handled Missing Values
2. Removed Duplicates
3. Removed Cancelled Orders
4. Removing non-product Stock-Codes

```
# Exploring the rows for which quantity is less than 0  
df[df["Quantity"] < 0]
```



|        | InvoiceNo | StockCode | Description                      | Quantity |
|--------|-----------|-----------|----------------------------------|----------|
| 141    | C536379   | D         | Discount                         | -1       |
| 154    | C536383   | 35004C    | SET OF 3 COLOURED FLYING DUCKS   | -1       |
| 235    | C536391   | 22556     | PLASTERS IN TIN CIRCUS PARADE    | -12      |
| 236    | C536391   | 21984     | PACK OF 12 PINK PAISLEY TISSUES  | -24      |
| 237    | C536391   | 21983     | PACK OF 12 BLUE PAISLEY TISSUES  | -24      |
| ...    | ...       | ...       | ...                              | ...      |
| 401159 | C581490   | 23144     | ZINC T-LIGHT HOLDER STARS SMALL  | -11      |
| 401243 | C581499   | M         | Manual                           | -1       |
| 401410 | C581568   | 21258     | VICTORIAN SEWING BOX LARGE       | -5       |
| 401411 | C581569   | 84978     | HANGING HEART JAR T-LIGHT HOLDER | -1       |
| 401412 | C581569   | 20979     | 36 PENCILS TUBE RED RETROSPOT    | -5       |

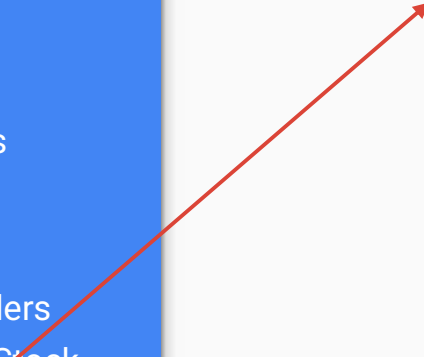
8872 rows x 8 columns

There are **8872** rows for which the quantity is negative which can be either due to data-entry errors or return orders or cancelled orders.

If we look at the InvoiceNo for all these cases, they start with the letter 'C' which indicates they are cancelled orders. Thus these rows were removed from the dataset

# Data Cleaning and Preparation

1. Handled Missing Values
2. Removed Duplicates
3. Removed Cancelled Orders
4. Removing non-product Stock-Codes



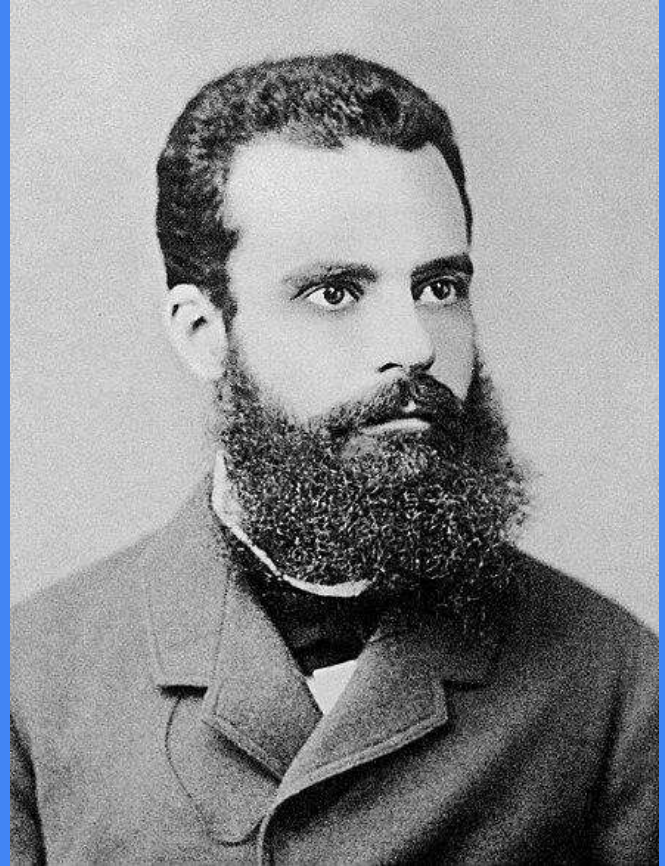
| StockCodes   | Description    |
|--------------|----------------|
| POST         | POSTAGE        |
| C2           | CARRIAGE       |
| M            | MANUAL         |
| DOT          | DOTCOM POSTAGE |
| BANK CHARGES | BANK CHARGES   |

There are certain StockCodes which do not belong to any products. All the rows containing such StockCodes were removed.



# Pareto Principle

Roughly 80% of outcomes stem  
from 20% of causes



# 26%

Customers contribute to 80% of the revenue.

# 21%

Products contribute to 80% of the revenue.

# RFM Analysis and Customer Segmentation

## What is RFM?

**Recency (R):** Days since last purchase

**Frequency (F):** Number of purchases

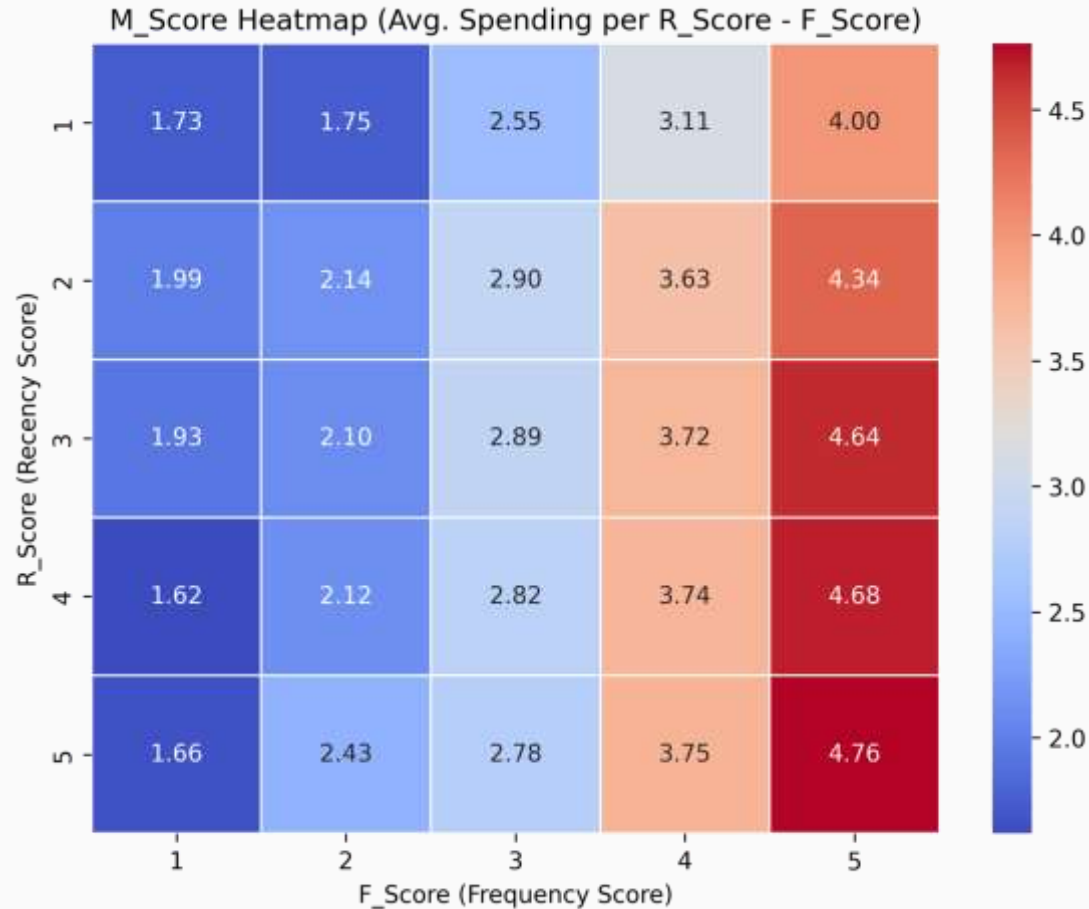
**Monetary (M):** Total spending

Customers are segmented into **five equal buckets** based on Recency, Frequency, and Monetary values. Each customer is ranked for each metric, assigned a **score from 1 to 5**, and their scores are summed to derive an overall **RFM score** for analysis.

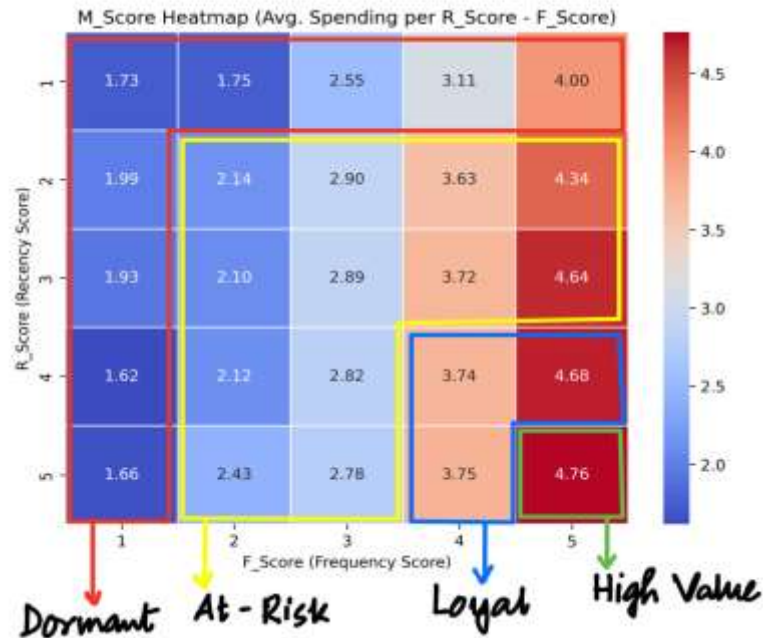
Customer Segments Based on RFM Score



# Heat Map based on RFM Score



# RFM Analysis and Customer Segmentation



Based on this RFM Score Analysis, we have four customer segments:

1. High Value Customers
2. Loyal Customers
3. At-Risk Customers
4. Dormant Customers

# Dashboard

Country  
All

Month  
All

8.74M

Total Revenue

4335

Active Customers

5M

Sales Volume

Qtr 1

Qtr 2

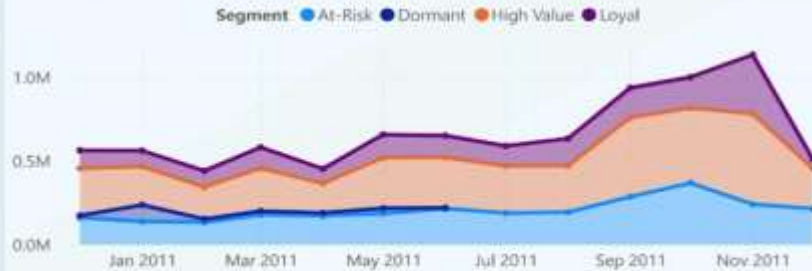
Qtr 3

Qtr 4

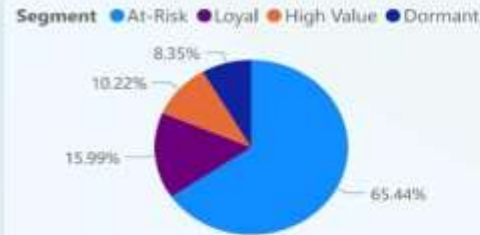
2010

2011

Revenue Contributions by Segment



Customers by Segment



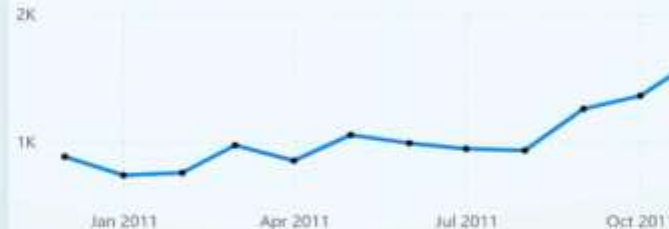
RFM Heatmap

|   | 1    | 2    | 3    | 4    | 5    |
|---|------|------|------|------|------|
| 1 | 1.73 | 1.75 | 2.55 | 3.11 | 4.00 |
| 2 | 1.99 | 2.14 | 2.90 | 3.63 | 4.34 |
| 3 | 1.93 | 2.10 | 2.89 | 3.72 | 4.64 |
| 4 | 1.62 | 2.12 | 2.82 | 3.74 | 4.68 |
| 5 | 1.66 | 2.43 | 2.78 | 3.76 | 4.76 |

Revenue by Days of Week



Active Customers Over Time



Top/Bottom Customers

| CustomerID | TotalSales   | SalesVolume |
|------------|--------------|-------------|
| 14646      | 2,79,138.02  | 197420      |
| 18102      | 2,59,657.30  | 64124       |
| 17450      | 1,94,390.79  | 69973       |
| 16446      | 1,68,472.50  | 80997       |
| 14911      | 1,36,161.83  | 80404       |
| 12415      | 1,24,564.53  | 77669       |
| 14156      | 1,16,560.08  | 57755       |
| Total      | 87,37,227.64 | 5155661     |

# Recommendations Based on Segments

## High Value Customers

**Characteristics:** Brand advocates with exceptional engagement and spending.

**Behaviour:** Low Recency (recent purchases), High Frequency, High Monetary.

**Strategy:** Reward with loyalty programs and exclusives.

Customer Segments Based on RFM Score





# Recommendations Based on Segments

## Loyal Customers

**Characteristics:** Consistent buyers with moderate activity.

**Behaviour:** Average Recency, Moderately High Frequency, High Monetary.

**Strategy:** Upsell/cross-sell opportunities to boost value.

Customer Segments Based on RFM Score



# Recommendations Based on Segments

## At-Risk Customers

**Characteristics:** Not-so Consistent buyers

**Behaviour:** Moderately High Recency, Low Frequency, Variable Monetary.

**Strategy:** Personalized emails and Limited-time promotions to regain interest.

Customer Segments Based on RFM Score



# Recommendations Based on Segments

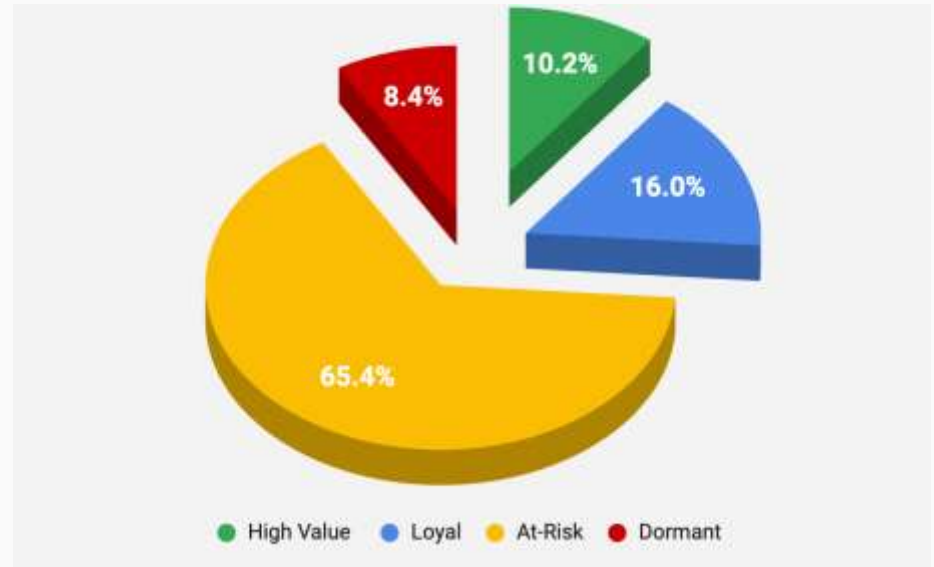
## Dormant Customers

**Characteristics:** Declining engagement with potential churn risk.

**Behaviour:** High Recency, Below Average Frequency, Low Monetary.

**Strategy:** Win-back campaigns or surveys to address dissatisfaction.

Customer Segments Based on RFM Score



# Thanks!

Contact:

Reet

+91-7980623308  
reetphy@gmail.com

