Customer Segmentation using RFM Analysis

Karthik

RFM (Recency, Frequency, Monetary) analysis is a proven marketing model for behavior based customer segmentation. It groups customers based on their transaction history — how recently, how often and how much did they buy.

RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

The data set used contains around 400k observations. It is taken from the UCI Machine Learning Repository.

Recency (R): Days since last purchase. **Frequency** (F): Total transactions. **Monetary Value** (M): Total amount a customer spent.

Scores are assigned to each of R,F,M based on their distributions found using the ***Summary*** command. Each score ranges from 1-5.

For example, the customers with the most recent purchase dates receive a recency ranking of 5, and those with purchase dates in the distant past receive a recency ranking of 1.

A frequency ranking is assigned in a similar way. Customers with high purchase frequency are assigned a higher score (4 or 5) and those with lowest frequency are assigned a score 1.

Monetary score is assigned on the basis of the total revenue generated by the customer in the period under consideration for the analysis. Customers with highest revenue/order amount are assigned a higher score while those with lowest revenue are assigned a score of 1.

A fourth score, RFM score is generated which is simply the three individual scores concatenated into a single value.

The customers with the highest RFM scores are most likely to respond to an offer.

Loading the data

```r
library(readxl)
df = read_excel("Online Retail.xlsx")
```

The dataset is between 01/12/2010 and 09/12/2011. So, we make the analysis date as 01/01/2012.

```r
library(anytime)
df$InvoiceDate = anydate(df$InvoiceDate)
analysis_date = anydate("2012-01-01")
```

Data Cleaning

Delete all negative value in Quantity and UnitPrice. We also need to delete all NA value.

```r
library(dplyr)
df = df %>%
  mutate(Quantity = replace(Quantity, Quantity <=0, NA),
         UnitPrice = replace(UnitPrice, UnitPrice <=0, NA))

df =  df %>%
  na.omit(df)
```

Change the character variables to factors and calculate the GMV. So, we get the customer historical purchased dataset

```
df = df %>%
  mutate(InvoiceNo=as.factor(InvoiceNo),
         StockCode=as.factor(StockCode),
         Country=as.factor(Country))

df = df %>%
  mutate(sales = Quantity*UnitPrice)

df_customer = df %>%
  select(CustomerID,InvoiceDate,sales)
```

RFM Analysis

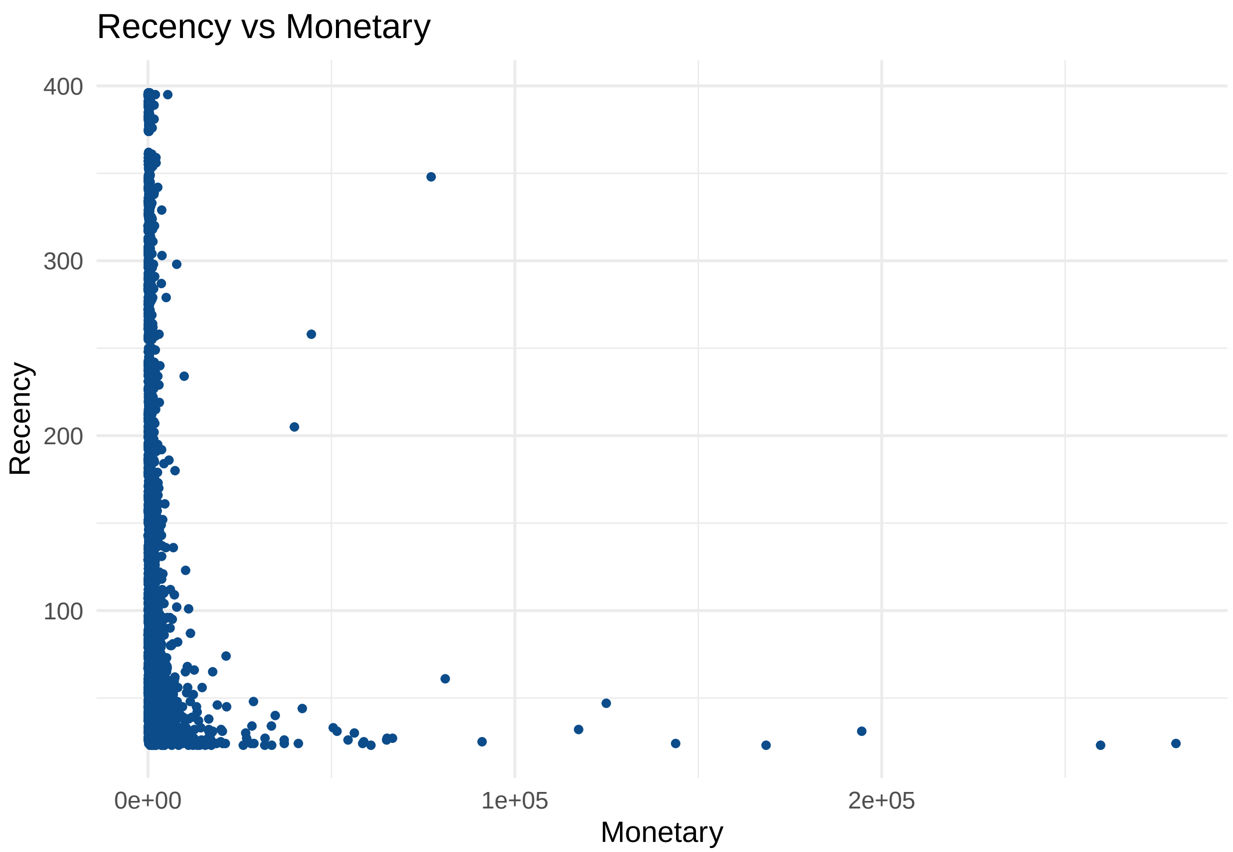
We have to process the data to get Recency,Frequency, and Monetary from each customers.

```
df_RFM = df %>%
  group_by(CustomerID) %>%
  summarise(recency = as.numeric(analysis_date - max(InvoiceDate)),
            frequency = n_distinct(InvoiceNo),
            monetary = sum(sales))
```

The best customers are those who: i) bought most recently ii) most often iii) spend the most Now let us examine the relationship between the above. ***Recency vs Monetary Value***

Customers who visited more recently generated more revenue compared to those who visited in the distant past. The customers who visited in the recent past are more likely to return compared to those who visited long time ago as most of those would be lost customers. As such, higher revenue would be associated with most recent visits.

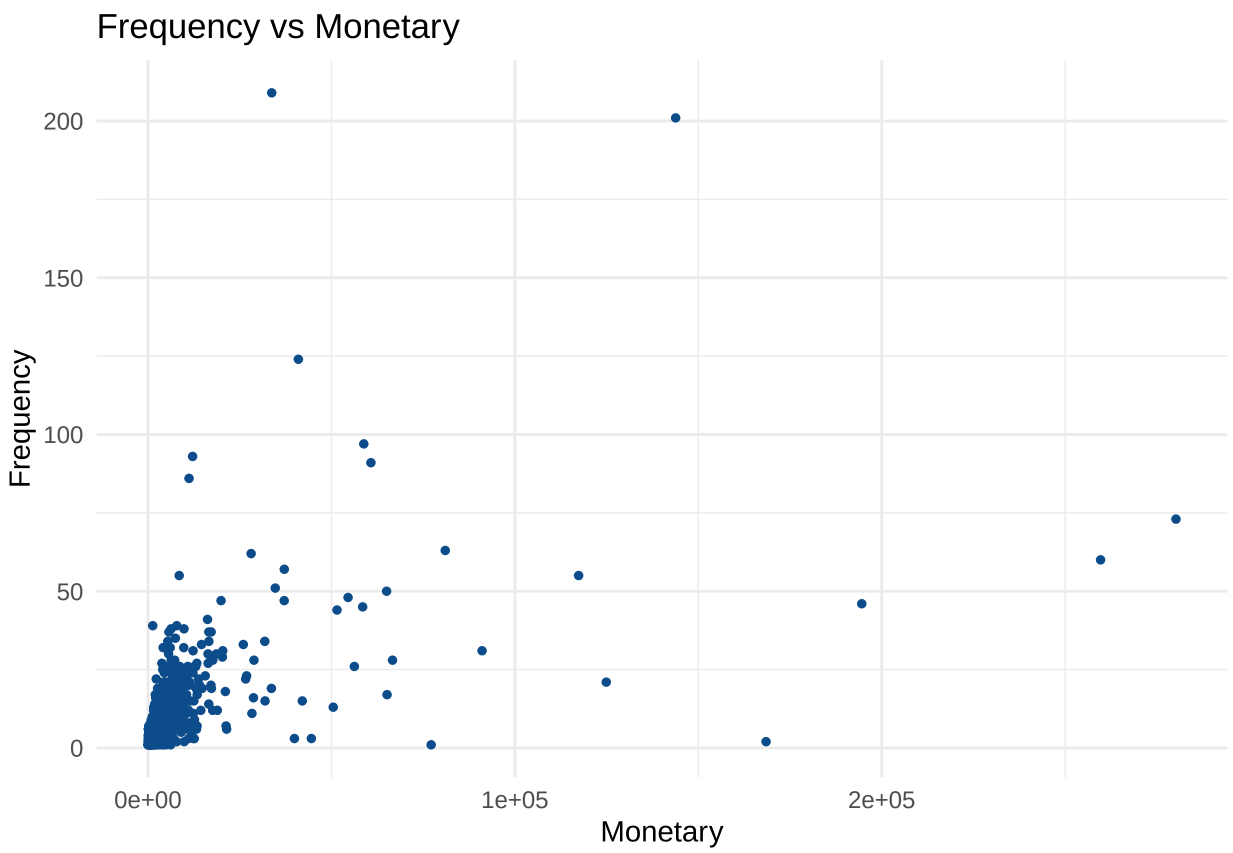
```
library(ggplot2)
ggplot(df_RFM) +
  aes(x = monetary, y = recency) +
  geom_point(size = 1L, colour = "#0c4c8a") +
  labs(x = "Monetary", y = "Recency", title = "Recency vs Monetary") +
  theme_minimal()
```



Frequency vs Monetary Value

As the frequency of visits increases, the revenue generated also increases. Customers who visit more frquently are your champion customers, loyal customers or potential loyalists and they drive higher revenue.

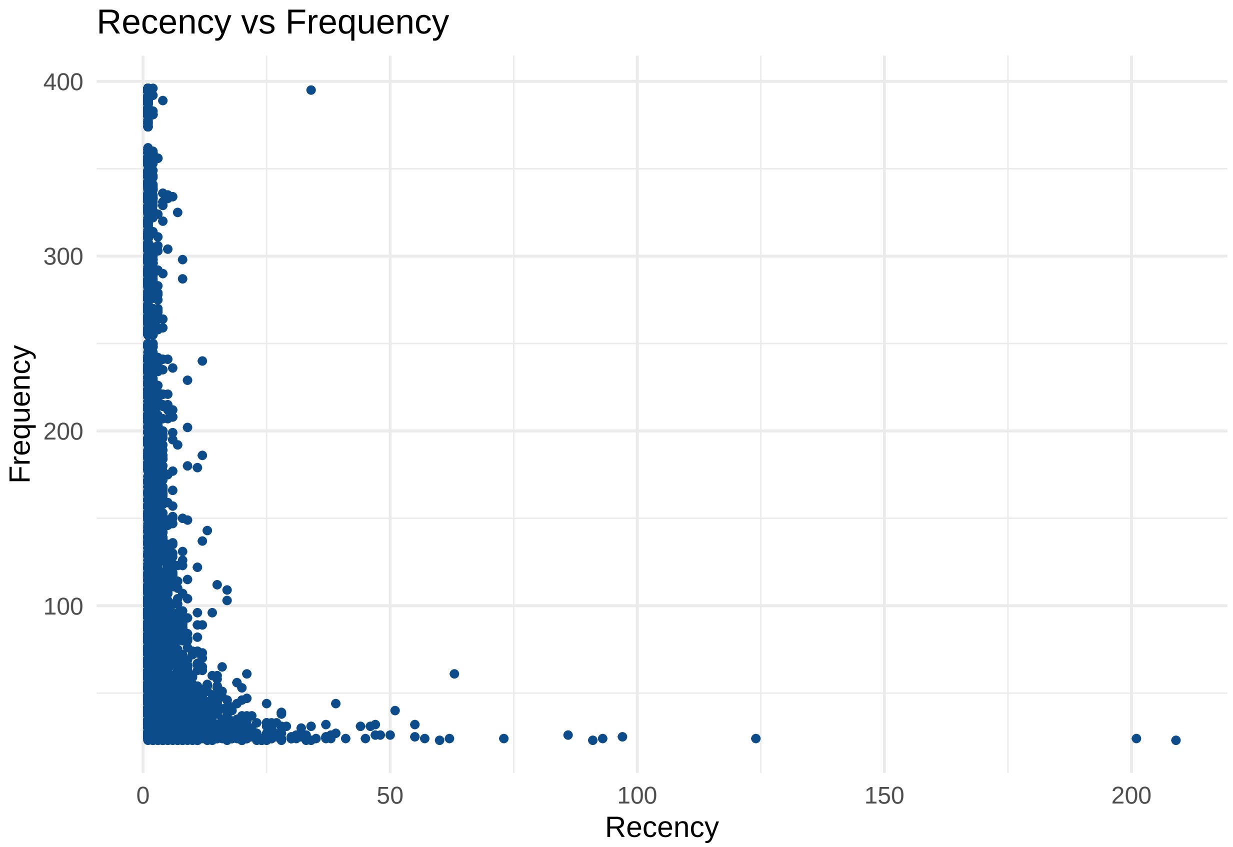
```r
library(ggplot2)
ggplot(df_RFM) +
  aes(x = monetary, y = frequency) +
  geom_point(size = 1L, colour = "#0c4c8a") +
  labs(x = "Monetary", y = "Frequency", title = "Frequency vs Monetary") +
  theme_minimal()
```



Recency vs Frequency

Customers with low frequency visited in the distant past while those with high frequency have visited in the recent past. Again, the customers who visited in the recent past are more likely to return compared to those who visited long time ago. As such, higher frequency would be associated with the most recent visits.

```r
library(ggplot2)
ggplot(df_RFM) +
  aes(x = frequency, y = recency) +
  geom_point(size = 1L, colour = "#0c4c8a") +
  labs(x = "Recency", y = "Frequency", title = "Recency vs Frequency") +
  theme_minimal()
```



Check the distribution of R,F,M

```
summary(df_RFM)
```

```
##    CustomerID       recency        frequency          monetary
##  Min.   :12346   Min.   : 23.0   Min.   :  1.000   Min.   :     3.75
##  1st Qu.:13813   1st Qu.: 40.0   1st Qu.:  1.000   1st Qu.:   307.42
##  Median :15300   Median : 73.0   Median :  2.000   Median :   674.49
##  Mean   :15300   Mean   :115.1   Mean   :  4.272   Mean   :  2054.27
##  3rd Qu.:16779   3rd Qu.:164.8   3rd Qu.:  5.000   3rd Qu.:  1661.74
##  Max.   :18287   Max.   :396.0   Max.   :209.000   Max.   :280206.02
```

Calculate the score based on the quartiles

Scoring

```
#Scoring
#Recency_score
df_RFM$R_Score[df_RFM$recency > 164.8] = 1
df_RFM$R_Score[df_RFM$recency > 115.1 & df_RFM$recency <= 164.8 ] = 2
df_RFM$R_Score[df_RFM$recency > 73 & df_RFM$recency <= 115.1 ] = 3
df_RFM$R_Score[df_RFM$recency > 40 & df_RFM$recency <= 73 ] = 4
df_RFM$R_Score[df_RFM$recency <= 40] = 5

#Frequency_score
df_RFM$F_Score[df_RFM$frequency < 1] = 1
```

```
df_RFM$F_Score[df_RFM$frequency >= 1 & df_RFM$frequency < 2] = 2
df_RFM$F_Score[df_RFM$frequency >= 2 & df_RFM$frequency < 4 ] = 3
df_RFM$F_Score[df_RFM$frequency >= 4 & df_RFM$frequency < 5 ] = 4
df_RFM$F_Score[df_RFM$frequency >= 5] = 5

#Monetary_score
df_RFM$M_Score[df_RFM$monetary <= 307.42] = 1
df_RFM$M_Score[df_RFM$monetary > 307.42 & df_RFM$monetary <= 674.49] = 2
df_RFM$M_Score[df_RFM$monetary > 674.49 & df_RFM$monetary <= 1661.74 ] = 3
df_RFM$M_Score[df_RFM$monetary > 1661.74 & df_RFM$monetary < 2054.27 ] = 4
df_RFM$M_Score[df_RFM$monetary >= 2054.27] = 5

#RFM_score
df_RFM = df_RFM %>%
  mutate(RFM_Score = 100*R_Score + 10*F_Score + M_Score)
```

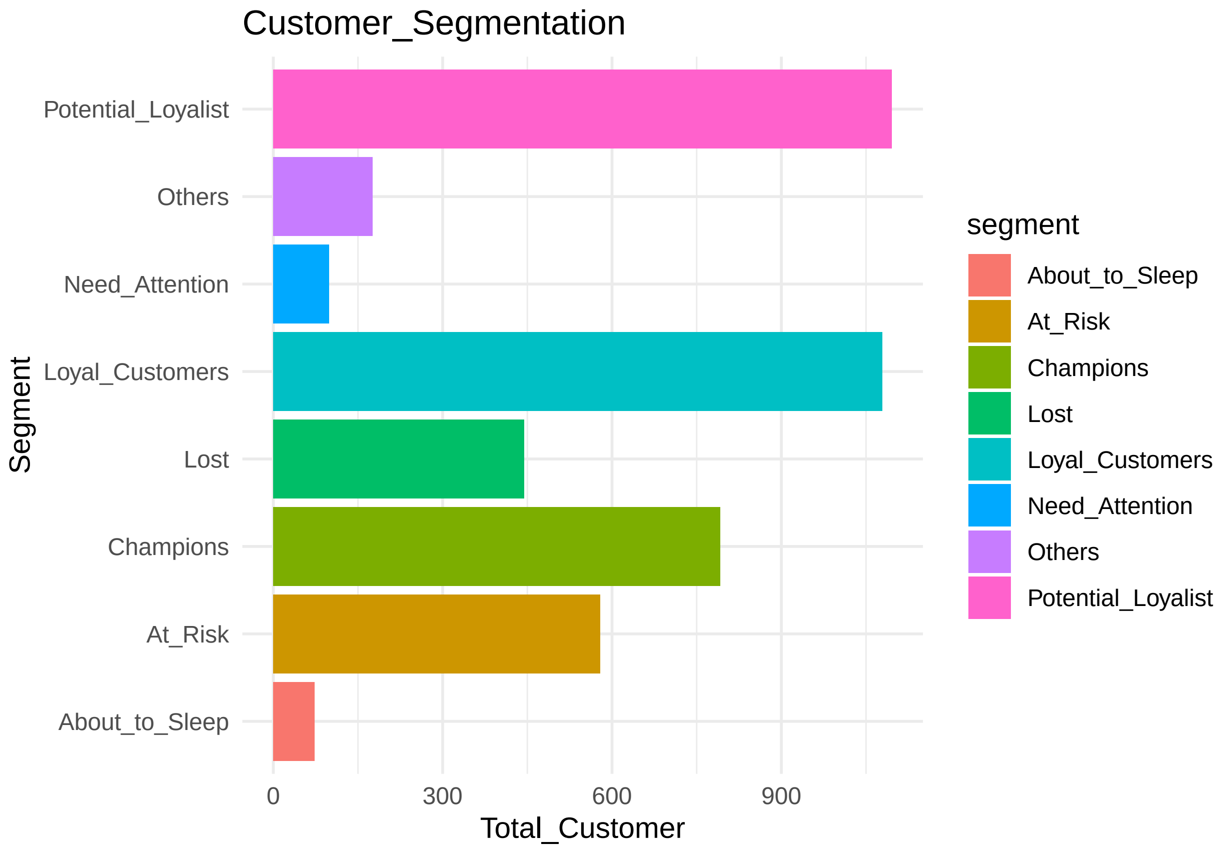
Let's classify the customers based on the recency, frequency and monetary scores. ***Segments***

```
df_RFM$segment = ifelse(between(df_RFM$R_Score,4,5) & between(df_RFM$F_Score, 4,5) &
                        between(df_RFM$M_Score, 4,5),"Champions",

                 ifelse(between(df_RFM$R_Score,2,5) & between(df_RFM$F_Score, 3,5) &
                               between(df_RFM$M_Score, 3,5),"Loyal_Customers",

                ifelse(between(df_RFM$R_Score,3,5) & between(df_RFM$F_Score, 1,3) &
                               between(df_RFM$M_Score,1,3),"Potential_Loyalist",

                ifelse(between(df_RFM$R_Score,4,5) & between(df_RFM$F_Score, 1,1) &
                              between(df_RFM$M_Score,1,1),"New_Customers",

                ifelse(between(df_RFM$R_Score,3,4) & between(df_RFM$F_Score, 1,1) &
                              between(df_RFM$M_Score,1,1),"Promising",

                ifelse(between(df_RFM$R_Score,2,3) & between(df_RFM$F_Score, 2,3) &
                              between(df_RFM$M_Score,2,3),"Need_Attention",

                ifelse(between(df_RFM$R_Score,2,3) & between(df_RFM$F_Score, 1,2) &
                              between(df_RFM$M_Score,1,2),"About_to_Sleep",

               ifelse(between(df_RFM$R_Score,1,2) & between(df_RFM$F_Score, 2,5) &
                              between(df_RFM$M_Score,2,5),"At_Risk",

              ifelse(between(df_RFM$R_Score,1,1) & between(df_RFM$F_Score, 4,5) &
                             between(df_RFM$M_Score,4,5),"Can't_Lose_them",

              ifelse(between(df_RFM$R_Score,1,2) & between(df_RFM$F_Score, 1,2) &
                            between(df_RFM$M_Score,1,2),"Lost", "Others")))))))))
```

Segment Size

```
Segmented_df = df_RFM %>%
  select(names(df_RFM)) %>%
  group_by(segment) %>%
```

```r
  summarize(Total_Customers = n(),
            Avg_Recency = mean(recency),
            Avg_frequency = mean(frequency),
            Avg_monetary_value = mean(monetary))

library(ggplot2)
ggplot(Segmented_df) +
  aes(x = segment, fill = segment, weight = Total_Customers) +
  geom_bar(position = "dodge") +
  scale_fill_hue() +
  coord_flip() +
  labs(title = "Customer_Segmentation",
       x = "Segment",
       y = "Total_Customer") +
  theme_minimal()
```



So, we have segmented our dataset. We need to also plan on how to target these individual segments.

Marketing Strategies

i) ***Champions***: Reward them. Can be early adopters for new products. Will promote our brand.
ii) ***Loyal Customers***: Up-sell higher value products. Ask for reviews. Engage them.
iii) ***Potential Loyalist***: Offer membership / loyalty program, recommend other products.
iv) ***Promising***: Create brand awareness, offer free trials
v) ***Customers Needing Attention***: Make limited time offers, Recommend based on past purchases. Reactivate them.
vi) ***At Risk***: Send personalized emails to reconnect, offer renewals, provide helpful resources.

vii) ***Can't Lose Them***: Win them back via renewals or newer products, don't lose them to competition, talk to them.

viii) ***Lost***: Offer other relevant products and special discounts. Recreate brand value.