Data Science Customer Segmentation using RFM model

Introduction

- Different customers have different needs.
- With the increase in customer numbers and their variations, it becomes not easy to understand the requirement of each customer.
- Identifying potential customers' needs can improve the marketing campaign, which ultimately increases the sales.
- Segmentation can play a better role in grouping those customers into various segments.

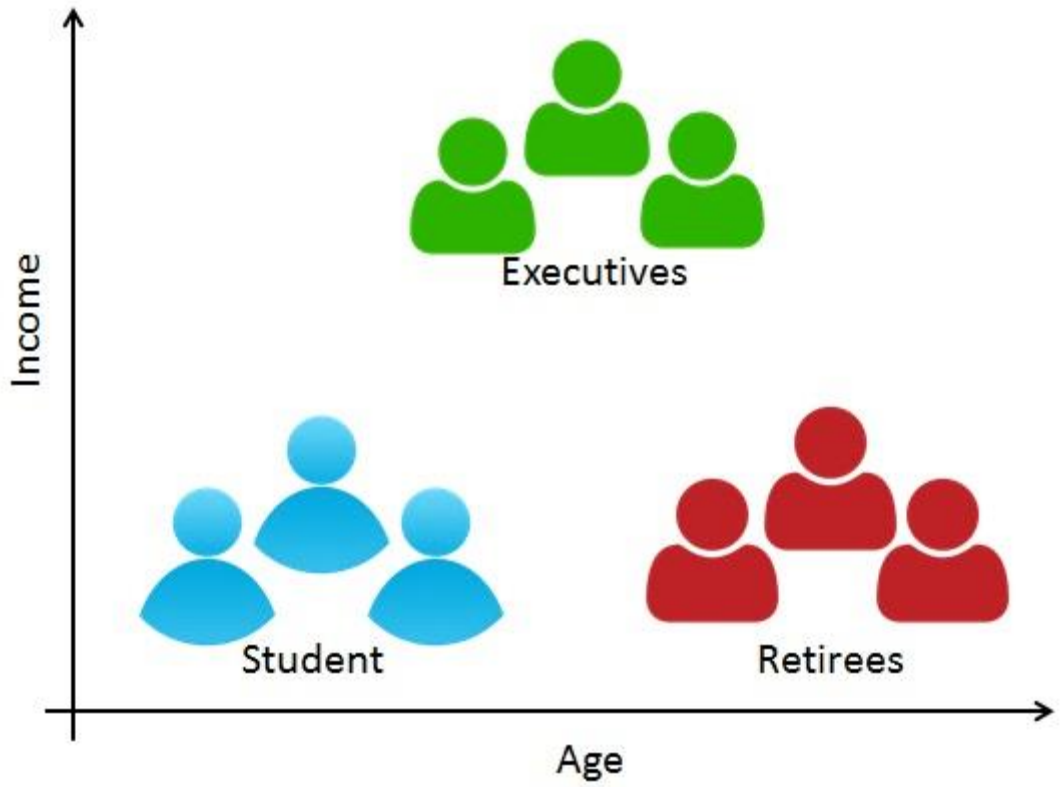
Topics of today

- What is Customer Segmentation?
- Need of Customer Segmentation
- Types of Segmentation
- Customer Segmentation using RFM Analysis
- Identify Potential Customer Segments using RFM in Python
- Conclusion

What is Customer Segmentation?

- *Customer segmentation is a method of dividing customers into groups or clusters on the basis of common characteristics.*
- We can segment customers into the B2C model using various customer's demographic characteristics such as occupation, gender, age, location, and marital status.
- Psychographic characteristics such as social class, lifestyle and personality characteristics and behavioral characteristics such as spending, consumption habits, product/service usage, and previously purchased products.

Example: Age/Income segmentation



Need of Customer Segmentation

- It helps in identifying the most potential customers.
- It helps managers to easily communicate with a targeted groups of the audience.
- It improves the quality of service, and customer loyalty, via a better understanding of segments.
- It helps managers to design special offers for targeted customers, to encourage them to buy more products.
- It also helps in identifying new products that customers could be interested in.

Types of Segmentation

Customer Segmentation using RFM analysis

- RFM (Recency, Frequency, Monetary) analysis is a behavior-based approach grouping customers into segments.
- It groups the customers on the basis of their previous purchase transactions.
- How recently, how often, and how much did a customer buy.
- RFM filters customers into various groups for the purpose of better service.
- There may be a segment of customers who are the big spenders but what if they purchased only once! How recently they purchased? Do they often purchase our product?

RFM Model

- Recency (R): Who have purchased recently? Number of days since last purchase (least recency)
- Frequency (F): Who has purchased frequently? It means the total number of purchases. (high frequency)
- Monetary Value(M): Who has high purchase amount? It means the total money customer spent (high monetary value)

RFM Model

- Here, Each of the three variables(Recency, Frequency, and Monetary) can be divided into four equal quartiles (Q1-Q4), which creates 64 (4x4x4) different customer segments.

Steps of RFM(Recency, Frequency, Monetary):

1. Calculate the Recency, Frequency, Monetary values for each customer.
2. Add segment bin values to RFM table using quartile.
3. Sort the customer RFM score in ascending order.

- RFM analysis helps marketers find answers to the following questions:
- Who are your best customers?
- Which of your customers could contribute to your churn rate?
- Who has the potential to become valuable customers?
- Which of your customers can be retained?
- Which of your customers are most likely to respond to engagement campaigns?

Example

- To conduct RFM analysis for this example, let's see how we can score these customers by ranking them based on each RFM attribute separately.
- Assume that we rank these customers from 1-5 using RFM values.
- Let's begin with ranking customers on recency first, as shown in the below table:

CUSTOMER ID	RECENCY	RANK	R SCORE
12	1	1	5
11	3	2	5
1	4	3	5
15	5	5	4
2	6	5	4
7	7	6	4
10	10	7	3
5	15	8	3
14	18	9	3
4	23	10	2
13	27	11	2
6	32	12	2
9	34	13	1
3	46	14	1
8	50	15	1

- As seen in the above table, we have sorted customers by recency, with the most recent purchasers at the top. Since customers are assigned scores from 1-5, the top 20% of customers (customer 12, 11, 1) receive a recency score of 5, the next 20% (the next 3 customers 15, 2, 7) a score of 4, and so on.

CUSTOMER ID	FREQUENCY	F SCORE
9	15	5
2	11	5
12	10	5
11	8	4
1	6	4
10	5	4
5	4	3
13	3	3
7	3	3
4	3	2
14	2	2
6	2	2
15	1	1
8	1	1
3	1	1

CUSTOMER ID	MONETARY	M SCORE
9	2630	5
12	1510	5
8	950	5
2	940	4
11	845	4
1	540	4
10	191	3
5	179	3
7	140	3
4	65	2
6	56	2
13	54	2
14	40	1
3	35	1
15	25	1

RFM Score

- Finally, we can rank these customers by combining their individual R, F, and M rankings to arrive at an aggregated RFM score. This RFM score, displayed in the table below, is simply the average of the individual R, F, and M scores, obtained by giving equal weights to each RFM attribute.

CUSTOMER ID	RFM CELL	RFM SCORE
1	5,4,4	4.3
2	4,5,4	4.3
3	1,1,1	1.0
4	2,2,2	2.0
5	3,3,3	3.0
6	2,2,2	2.0
7	4,3,3	3.3
8	1,1,5	2.3
9	1,5,5	3.7
10	3,4,3	3.3
11	5,4,4	4.3
12	5,5,5	5.0
13	2,3,2	2.3
14	3,2,1	2.0
15	4,1,1	2.0

Recency, Frequency, and Monetary Analysis

- The next question that arises is: Is it fair to average out the individual R, F, and M scores for each customer and assign them to RFM segment, as per their purchase or engagement behavior?
- Depending on the nature of your businesses, you might increase or decrease the relative importance of each RFM variable to arrive at the final score. For example:

- In a consumer durables business, the monetary value per transaction is normally high but frequency and recency is low. For example, you can't expect a customer to purchase a refrigerator or air conditioner on a monthly basis. In this case, a marketer could give more weight to monetary and recency aspects rather than the frequency aspect.
- In a retail business selling fashion/cosmetics, a customer who searches and purchases products every month will have a higher recency and frequency score than monetary score. Accordingly, the RFM score could be calculated by giving more weight to R and F scores than M.
- For content apps like Shahid or Netflix, a binge watcher will have a longer session length than a mainstream consumer watching at regular intervals. For bingers, engagement and frequency could be given more importance than recency, and for mainstreamers, recency and frequency can be given higher weights than engagement to arrive at the RFE score.

Analyzing RFM Segmentation

- Let’s delve into few interesting segments:
- **Champions** are your best customers, who bought most recently, most often, and are heavy spenders. Reward these customers. They can become early adopters for new products and will help promote your brand.
- **Potential Loyalists** are your recent customers with average frequency and who spent a good amount. Offer membership or loyalty programs or recommend related products to upsell them and help them become your Loyalists or Champions.
- **New Customers** are your customers who have a high overall RFM score but are not frequent shoppers. Start building relationships with these customers by providing onboarding support and special offers to increase their visits.
- **At Risk Customers** are your customers who purchased often and spent big amounts, but haven’t purchased recently. Send them personalized reactivation campaigns to reconnect, and offer renewals and helpful products to encourage another purchase.
- **Can’t Lose Them** are customers who used to visit and purchase quite often, but haven’t been visiting recently. Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor.

Example

1.Calculate the Recency, Frequency, Monetary values for each customer.

CustomerID	monetary	frequency	recency
12346.0	325	77183.60	1
12747.0	2	4196.01	103
12748.0	0	33719.73	4596
12749.0	3	4090.88	199
12820.0	3	942.34	59

Example

2. Add segment bin values to RFM table using quartile.

CustomerID	monetary	frequency	recency	r_quartile	f_quartile	m_quartile
12346.0	325	77183.60	1	1	1	1
12747.0	2	4196.01	103	4	1	4
12748.0	0	33719.73	4596	4	1	4
12749.0	3	4090.88	199	4	1	4
12820.0	3	942.34	59	3	2	4

Example

3. Concatenate all scores in single column (RFM_Score).

CustomerID	monetary	frequency	recency	r_quartile	f_quartile	m_quartile	RFM_Score
12346.0	325	77183.60	1	1	1	1	111
12747.0	2	4196.01	103	4	1	4	414
12748.0	0	33719.73	4596	4	1	4	414
12749.0	3	4090.88	199	4	1	4	414
12820.0	3	942.34	59	3	2	4	324

Identify Potential Customer Segments using RFM in Python

- #import modules

```
import pandas as pd # for dataframes
import matplotlib.pyplot as plt # for plotting graphs
import seaborn as sns # for plotting graphs
import datetime as dt
```

- **Loading Dataset**
- Let's first load the required HR dataset using the pandas read CSV function.

```
data = pd.read_excel("Online_Retail.xlsx")

df=pd.DataFrame(data)
print(df.head()) #first 5 rows
print(df.tail()) #last 5 rows
```

Identify Potential Customer Segments using RFM in Python

```
data = pd.read_excel("Online_Retail.xlsx")
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   lower        1816 non-null    object
 3   Description  540455 non-null  object
 4   Quantity     541909 non-null  int64
 5   InvoiceDate  541909 non-null  datetime64[ns]
 6   UnitPrice    541909 non-null  float64
 7   CustomerID   406829 non-null  float64
 8   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(5)
memory usage: 37.2+ MB
```

Identify Potential Customer Segments using RFM in Python

```
data = pd.read_excel("Online_Retail.xlsx")
data.info()
```

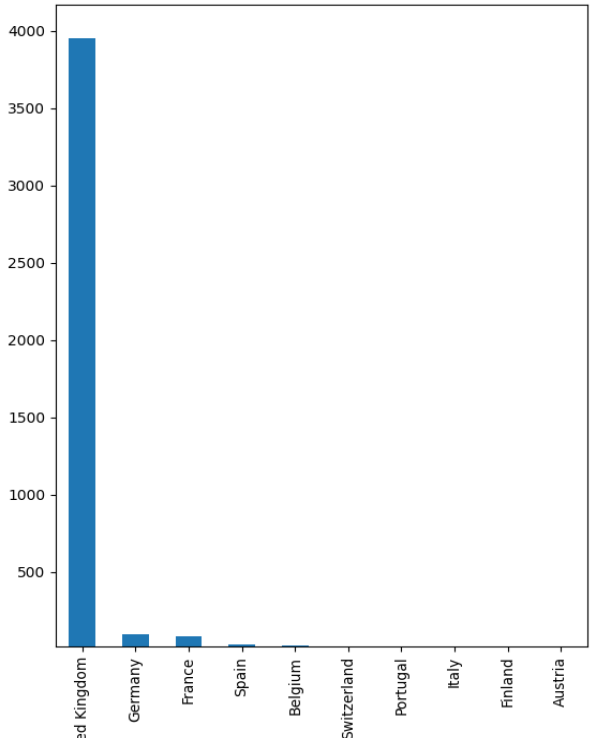
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   lower        1816 non-null    object
 3   Description  540455 non-null  object
 4   Quantity     541909 non-null  int64
 5   InvoiceDate  541909 non-null  datetime64[ns]
 6   UnitPrice    541909 non-null  float64
 7   CustomerID   406829 non-null  float64
 8   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(5)
memory usage: 37.2+ MB
```

Identify Potential Customer Segments using RFM in Python

- **Removing Duplicates**
- Sometimes you get a messy dataset.
- You may want to remove duplicates, which will skew your analysis.
- In python, pandas offer function drop_duplicates(), which drops the repeated or duplicate records.
- `filtered_data=data[['Country','CustomerID']].drop_duplicates()`

Identify Potential Customer Segments using RFM in Python

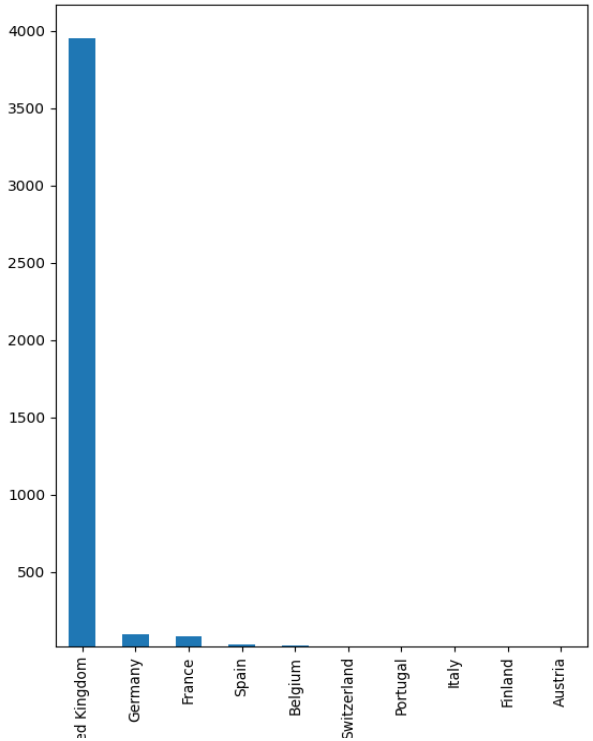
- **Data Insights**
- `#Top ten country's customer filtered_data.Country.value_counts()[:10].plot(kind='bar')`



Identify Potential Customer Segments using RFM in Python

- **Data Insights**
- In the shown graph, you can observe most of the customers are from the "United Kingdom". So, we can filter data for United Kingdom customers.

```
uk_data=data[data.Country=='United Kingdom']
```



Identify Potential Customer Segments using RFM in Python

- **Data Insights**

```
uk_data=data[data.Country=='United Kingdom’]
uk_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 361878 entries, 0 to 541893
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    361878 non-null  object
 1   StockCode    361878 non-null  object
 2   lower        1297 non-null    object
 3   Description  361878 non-null  object
 4   Quantity     361878 non-null  int64
 5   InvoiceDate  361878 non-null  datetime64[ns]
 6   UnitPrice    361878 non-null  float64
 7   CustomerID   361878 non-null  float64
 8   Country      361878 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(5)
memory usage: 27.6+ MB
```

Identify Potential Customer Segments using RFM in Python

- **Data Insights**

The nunique() function in pandas is convenient in getting summary statistics about the uniqueness of the columns. This function returns the count, of unique values in each column of the data.

```
print(df.nunique ())
```

```
InvoiceNo      25900
StockCode       4070
lower            953
Description     4223
Quantity         722
InvoiceDate    23260
UnitPrice       1630
CustomerID      4372
Country           38
dtype: int64
```

Identify Potential Customer Segments using RFM in Python

- **Data Insights**

The describe() function in pandas is convenient in getting various summary statistics. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.

```
print(uk_data.describe ())
```

```
            Quantity      UnitPrice     CustomerID
count  361878.000000  361878.000000  361878.000000
mean       11.077029       3.256007   15547.871368
std       263.129266      70.654731    1594.402590
min    -80995.000000       0.000000   12346.000000
25%         2.000000       1.250000   14194.000000
50%         4.000000       1.950000   15514.000000
75%        12.000000       3.750000   16931.000000
max     80995.000000   38970.000000   18287.000000
```

Identify Potential Customer Segments using RFM in Python

- ## Data Insights

We observed some of the customers have ordered in a negative quantity, which is not possible. So, we need to filter Quantity greater than zero.

```
uk_data = uk_data[(uk_data['Quantity']>0)]
uk_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 354345 entries, 0 to 541893
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype         
---  ------       --------------   -----         
 0   InvoiceNo    354345 non-null  object        
 1   StockCode    354345 non-null  object        
 2   lower        1285 non-null    object        
 3   Description  354345 non-null  object        
 4   Quantity     354345 non-null  int64         
 5   InvoiceDate  354345 non-null  datetime64[ns]
 6   UnitPrice    354345 non-null  float64       
 7   CustomerID   354345 non-null  float64       
 8   Country      354345 non-null  object        
dtypes: datetime64[ns](1), float64(2), int64(1), object(5)
memory usage: 27.0+ MB
```

Identify Potential Customer Segments using RFM in Python

- **Filter required Columns**

We can filter the necessary columns for RFM analysis. We only need her five columns CustomerID, InvoiceDate, InvoiceNo, Quantity, and UnitPrice.

CustomerId will uniquely define your customers, InvoiceDate help us calculate recency of purchase, InvoiceNo helps us to count the number of time transaction performed(frequency). Quantity purchased in each transaction and UnitPrice of each unit purchased by the customer will help us calculate the total purchased amount.

```
uk_data=uk_data[['CustomerID','InvoiceDate','InvoiceNo','Quantity','UnitPrice']]
uk_data['TotalPrice'] = uk_data['Quantity'] * uk_data['UnitPrice']
uk_data['InvoiceDate'].min(),uk_data['InvoiceDate'].max()
(Timestamp('2010-12-01 08:26:00'), Timestamp('2011-12-09 12:49:00'))
PRESENT = dt.datetime(2011,12,10)
uk_data['InvoiceDate'] = pd.to_datetime(uk_data['InvoiceDate'])
print(uk_data.head())
```

Identify Potential Customer Segments using RFM in Python

- **Filter required Columns**

```
uk_data=uk_data[['CustomerID','InvoiceDate','InvoiceNo','Quantity','UnitPrice']]
uk_data['TotalPrice'] = uk_data['Quantity'] * uk_data['UnitPrice']
uk_data['InvoiceDate'].min(),uk_data['InvoiceDate'].max()
(Timestamp('2010-12-01 08:26:00'), Timestamp('2011-12-09 12:49:00'))
PRESENT = dt.datetime(2011,12,10)
uk_data['InvoiceDate'] = pd.to_datetime(uk_data['InvoiceDate'])
print(uk_data.head())
```

```
dtypes: datetime64[ns](1), float64(2), int64(1), object(5)
memory usage: 27.0+ MB
   CustomerID         InvoiceDate InvoiceNo  Quantity  UnitPrice  TotalPrice
0     17850.0 2010-12-01 08:26:00    536365         6       2.55       15.30
1     17850.0 2010-12-01 08:26:00    536365         6       3.39       20.34
2     17850.0 2010-12-01 08:26:00    536365         8       2.75       22.00
3     17850.0 2010-12-01 08:26:00    536365         6       3.39       20.34
4     17850.0 2010-12-01 08:26:00    536365         6       3.39       20.34
```

Identify Potential Customer Segments using RFM in Python

- **RFM Analysis**
- Here, we are going to perform following operations:
  - *For Recency*, Calculate the number of days between present date and date of last purchase each customer.
  - *For Frequency*, Calculate the number of orders for each customer.
  - *For Monetary*, Calculate sum of purchase price for each customer.

```
rfm= uk_data.groupby('CustomerID').agg({'InvoiceDate':
lambda date: (PRESENT - date.max()).days, 'InvoiceNo':
lambda num: len(num), 'TotalPrice': lambda price:
price.sum()})
```

Identify Potential Customer Segments using RFM in Python

- **RFM Analysis**

```
print(rfm.head())
```

CustomerID	InvoiceDate	InvoiceNo	TotalPrice
12346.0	325	1	77183.60
12747.0	2	103	4196.01
12748.0	0	4596	33719.73
12749.0	3	199	4090.88
12820.0	3	59	942.34

Identify Potential Customer Segments using RFM in Python

- **RFM Analysis**

```
# Change the name of columns
rfm.columns=['monetary','frequency','recency']
rfm['recency'] = rfm['recency'].astype(int)

print(rfm.head())
```

```
             monetary  frequency  recency
CustomerID
12346.0           325          1    77183
12747.0             2        103     4196
12748.0             0       4596    33719
12749.0             3        199     4090
12820.0             3         59      942
```

Identify Potential Customer Segments using RFM in Python

- **Computing Quartile of RFM values**

- Customers with the lowest recency, highest frequency and monetary amounts considered as top customers.
- qcut() is Quartile-based discretization function. qcut bins the data based on sample quartiles.
- For example, 1000 values for 4 quartiles would produce a categorical object indicating quartile membership for each customer.

Identify Potential Customer Segments using RFM in Python

- **Computing Quartiles of RFM values**

```
rfm['r_quartile'] = pd.qcut(rfm['recency'], 4, ['1','2','3','4'])
rfm['f_quartile'] = pd.qcut(rfm['frequency'], 4, ['4','3','2','1'])
rfm['m_quartile'] = pd.qcut(rfm['monetary'], 4, ['4','3','2','1’])
print(rfm.head())
```

CustomerID	monetary	frequency	recency	r_quartile	f_quartile	m_quartile
12346.0	325	1	77183	4	4	1
12747.0	2	103	4196	4	1	4
12748.0	0	4596	33719	4	1	4
12749.0	3	199	4090	4	1	4
12820.0	3	59	942	3	2	4

Identify Potential Customer Segments using RFM in Python

- **RFM result interpretations**
- concatenate all three quartiles(r_quartile,f_quartile,m_quartile) in a single column, this rank will help you to segment the customers well group.
-
```
rfm['RFM_Score'] = rfm.r_quartile.astype(str)+
rfm.f_quartile.astype(str) + rfm.m_quartile.astype(str)
print(rfm.head())
```

```
            monetary  frequency  recency  r_quartile f_quartile  m_quartile  RFM_Score
CustomerID
12346.0          325          1    77183           4          4           1        441
12747.0            2        103     4196           4          1           4        414
12748.0            0       4596    33719           4          1           4        414
12749.0            3        199     4090           4          1           4        414
12820.0            3         59      942           3          2           4        324
```

Identify Potential Customer Segments using RFM in Python

- **RFM result interpretations**
- We need to see the highest priority customers (RFM_Score=111). To do that,
- ```
  # Filter out Top/Best customers
  print(rfm[rfm['RFM_Score']=='111'].sort_values('monetary', ascending=False).head())
  ```
- ```
  print(rfm.columns)
  ```

```
            monetary  frequency  recency  r_quartile f_quartile  m_quartile  RFM_Score
CustomerID
16754.0          372     2002.4        2           1          1           1        111
12346.0          325    77183.6        1           1          1           1        111
15749.0          235    44534.3       10           1          1           1        111
16698.0          226     1998.0        5           1          1           1        111
13135.0          196     3096.0        1           1          1           1        111
[5 rows x 7 columns]
Empty DataFrame
Columns: [monetary, frequency, recency, r_quartile, f_quartile, m_quartile, RFM_Score]
Index:[]
```

RFM Model Conclusion

- We have learned customer segmentation, Need of Customer Segmentation, and Types of Segmentation.
- We have focused on Customer segmenation via RFM model.
- We have learned RFM implementation via Python. Also, we have seen some basic concepts of pandas module such as handling duplicates, groupby, and qcut() for bins based on sample quartiles.

Assignment 2

- Find another data set, with at least 50,000 records
- What else does the Lambda function for aggregation in the Pandas module provide?
- Give examples of its use on the downloaded data set you found
- Perform RFM model on it