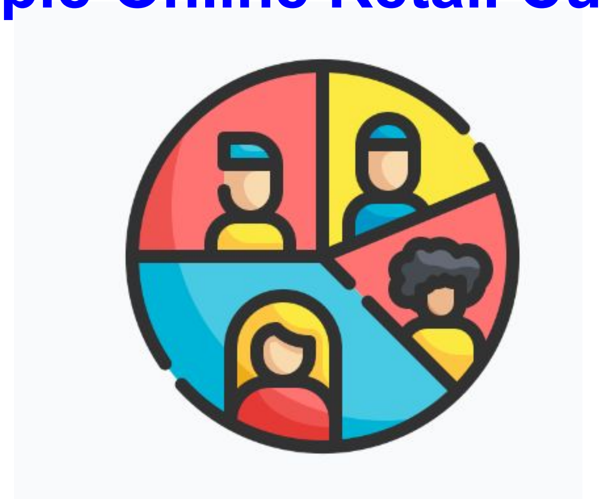
Capstone Project- 4

Topic-Online Retail Customer Segmentation



By-

Ajit Kumar Toppo

Problem Description

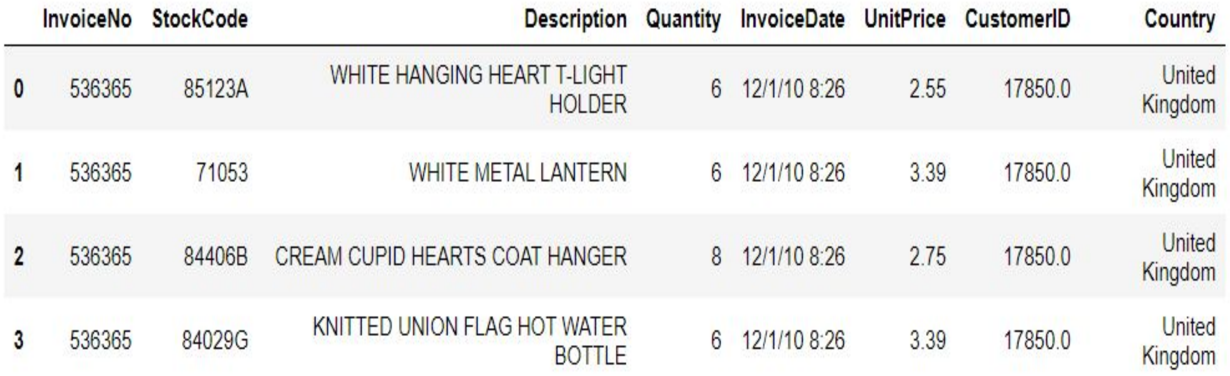
To identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Description

Attribute Information:

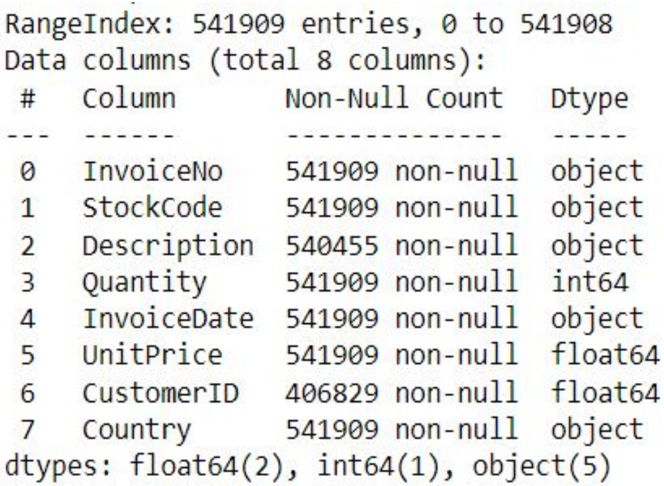
- **InvoiceNo:** Invoice number, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product (item) name. Nominal.
- **Quantity:** The quantities of each product (item) per transaction.Numeric.
- **InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice:** Unit price. Numeric, Product price per unit in sterling.
- **CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country:** Country name. Nominal, the name of the country where each customer resides.

Dataset



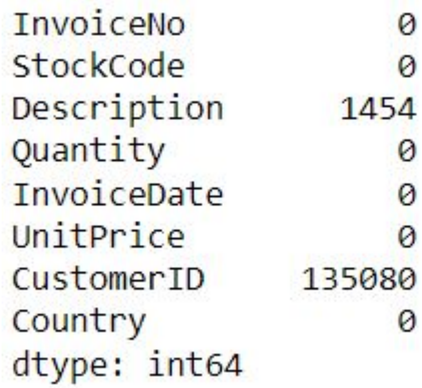
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom

Information about dataset



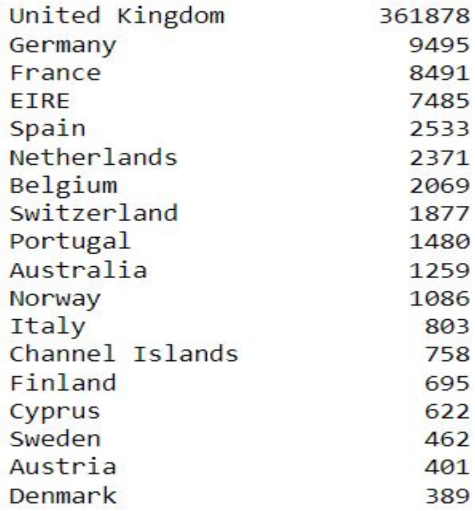
```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
```

Null values

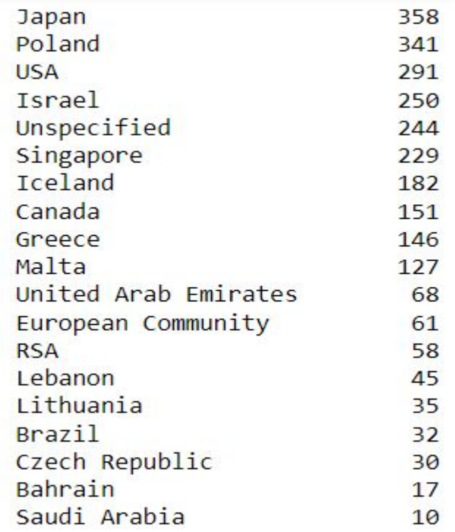


```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

number of customers from different countries



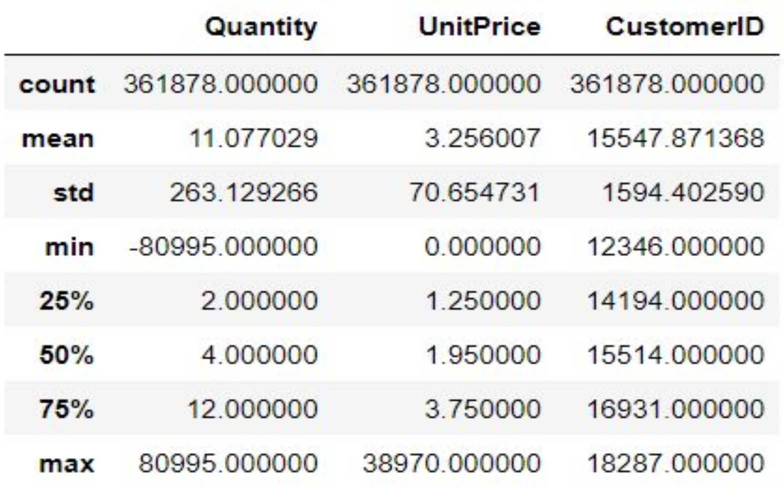
```
United Kingdom          361878
Germany                   9495
France                    8491
EIRE                      7485
Spain                     2533
Netherlands               2371
Belgium                   2069
Switzerland               1877
Portugal                  1480
Australia                 1259
Norway                    1086
Italy                      803
Channel Islands            758
Finland                    695
Cyprus                     622
Sweden                     462
Austria                    401
Denmark                    389
```



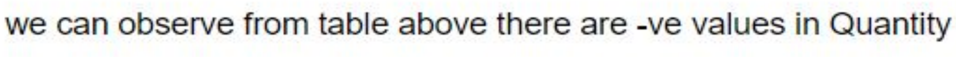
```
Japan                      358
Poland                     341
USA                        291
Israel                     250
Unspecified                244
Singapore                  229
Iceland                    182
Canada                     151
Greece                     146
Malta                      127
United Arab Emirates        68
European Community          61
RSA                         58
Lebanon                     45
Lithuania                   35
Brazil                      32
Czech Republic              30
Bahrain                     17
Saudi Arabia                10
```

Most of the customers in the data are from the United Kingdom. Customer clusters vary by geography, so here we’ll restrict the data to the United Kingdom only.

Data summary of United Kingdom customers



	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

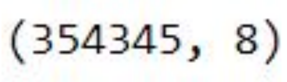


we can observe from table above there are -ve values in Quantity

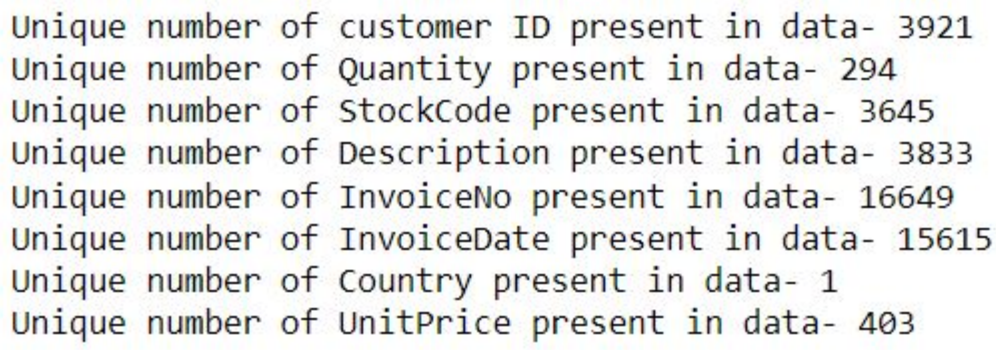
Because quantity cannot be -ve for purchase done by customers we will delete all the rows that contains negative values.

For the segmentation we will consider only the transactions that are done between 9/12/2010 to 9/12/2011 because it's better to use a metric per Months or Years in RFM.

Dataset after processing

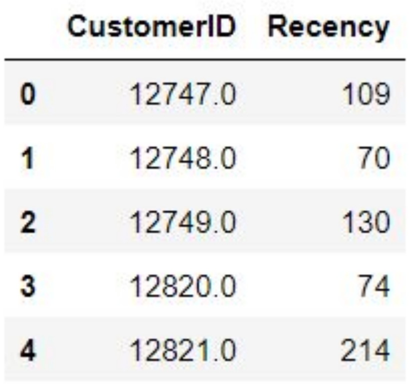


```
(354345, 8)
```

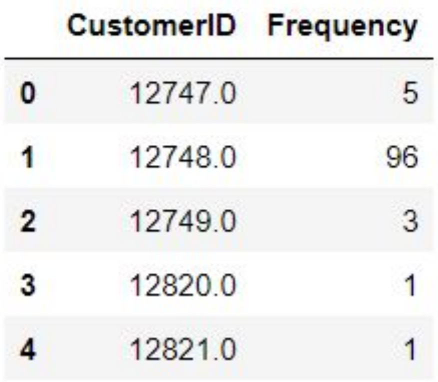


```
Unique number of customer ID present in data- 3921
Unique number of Quantity present in data- 294
Unique number of StockCode present in data- 3645
Unique number of Description present in data- 3833
Unique number of InvoiceNo present in data- 16649
Unique number of InvoiceDate present in data- 15615
Unique number of Country present in data- 1
Unique number of UnitPrice present in data- 403
```

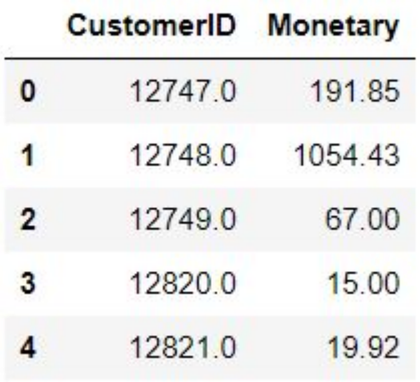
RFM



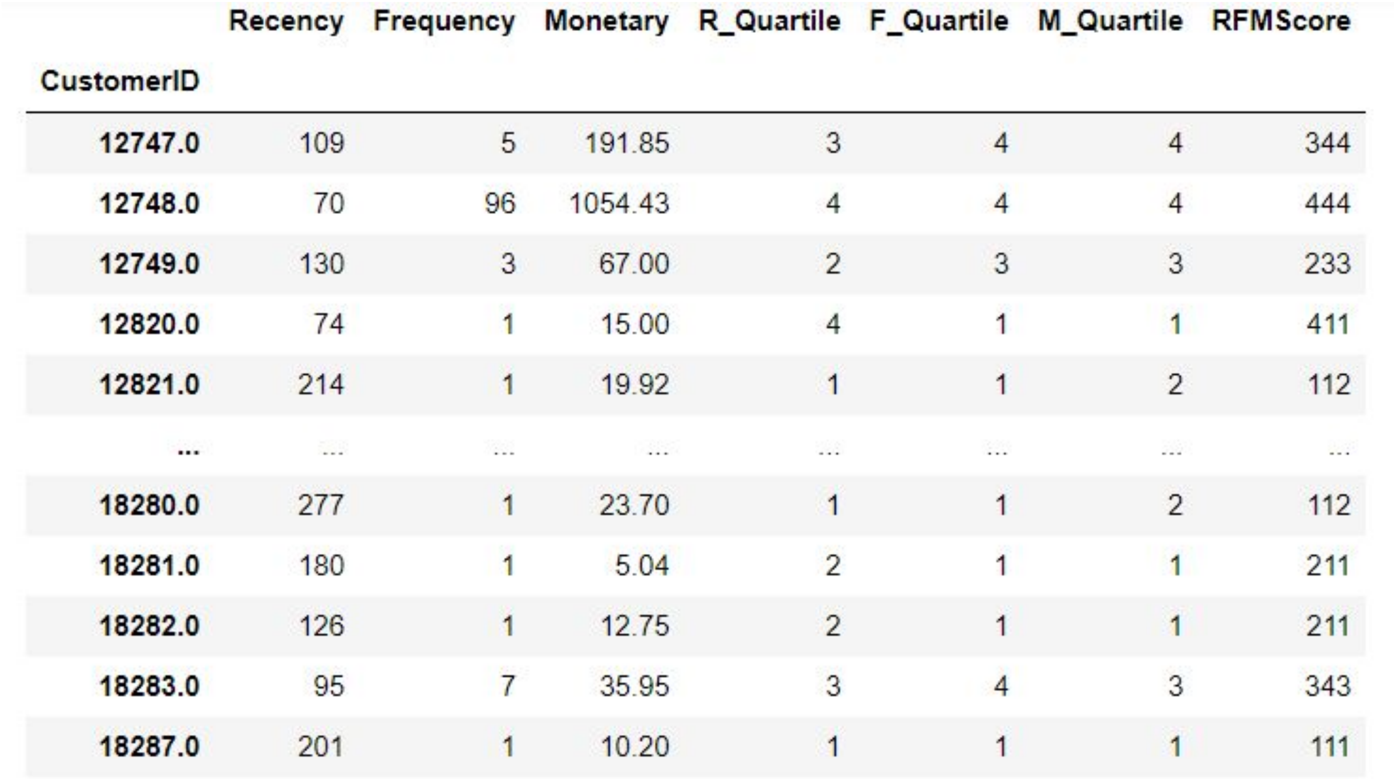
	CustomerID	Recency
0	12747.0	109
1	12748.0	70
2	12749.0	130
3	12820.0	74
4	12821.0	214



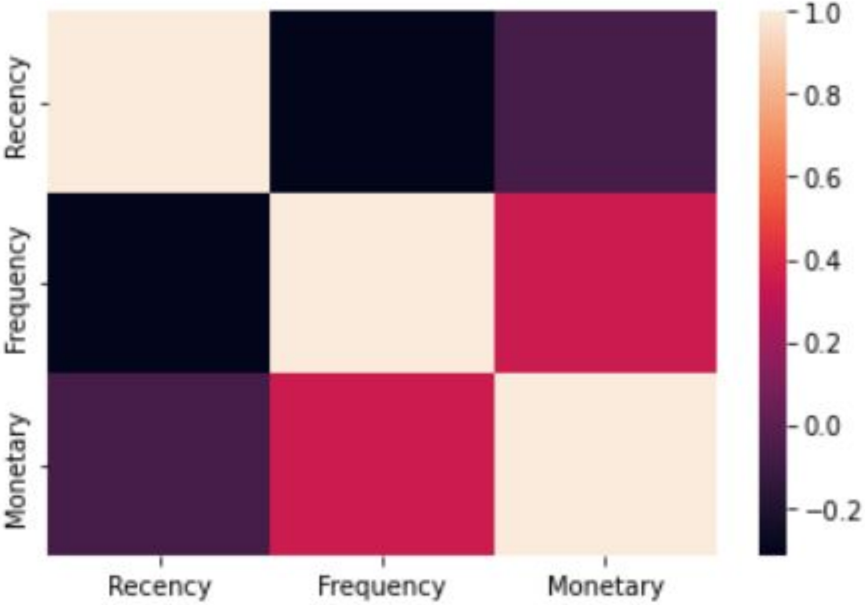
	CustomerID	Frequency
0	12747.0	5
1	12748.0	96
2	12749.0	3
3	12820.0	1
4	12821.0	1



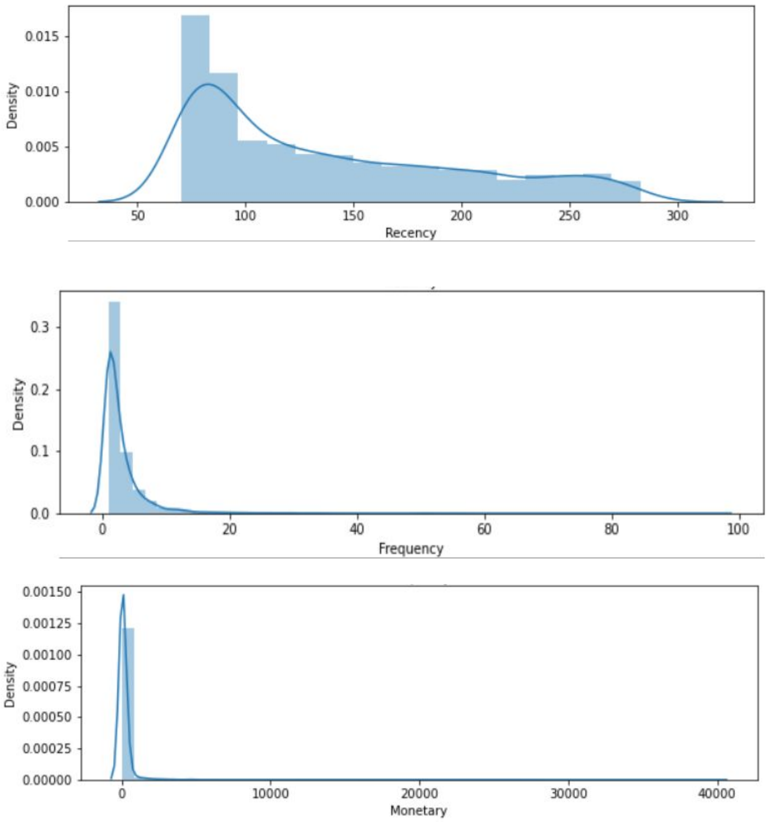
	CustomerID	Monetary
0	12747.0	191.85
1	12748.0	1054.43
2	12749.0	67.00
3	12820.0	15.00
4	12821.0	19.92



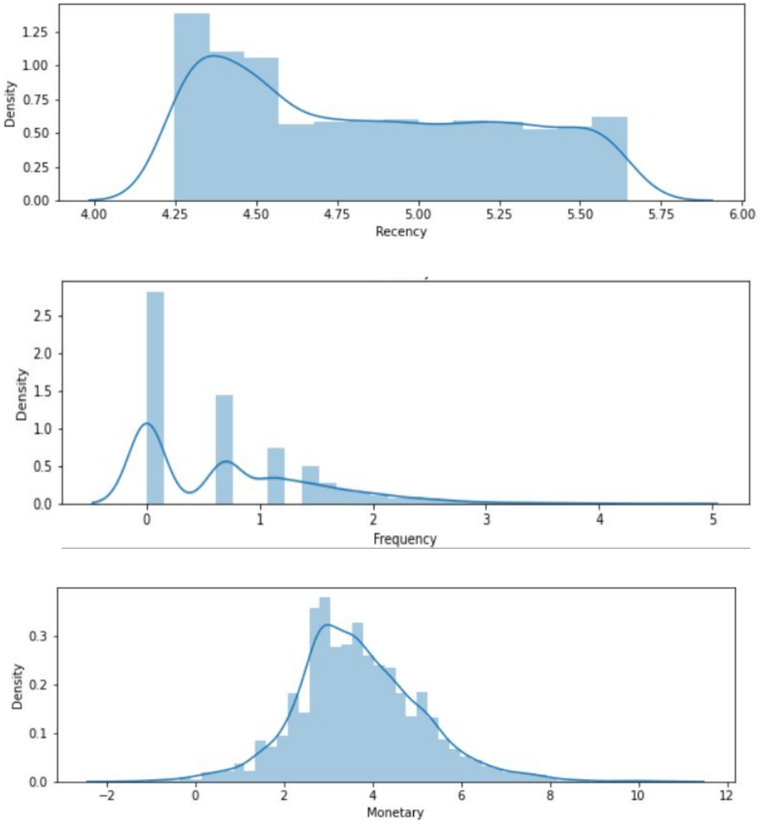
	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore
CustomerID							
12747.0	109	5	191.85	3	4	4	344
12748.0	70	96	1054.43	4	4	4	444
12749.0	130	3	67.00	2	3	3	233
12820.0	74	1	15.00	4	1	1	411
12821.0	214	1	19.92	1	1	2	112
...	...	...	...	...	...	...	...
18280.0	277	1	23.70	1	1	2	112
18281.0	180	1	5.04	2	1	1	211
18282.0	126	1	12.75	2	1	1	211
18283.0	95	7	35.95	3	4	3	343
18287.0	201	1	10.20	1	1	1	111

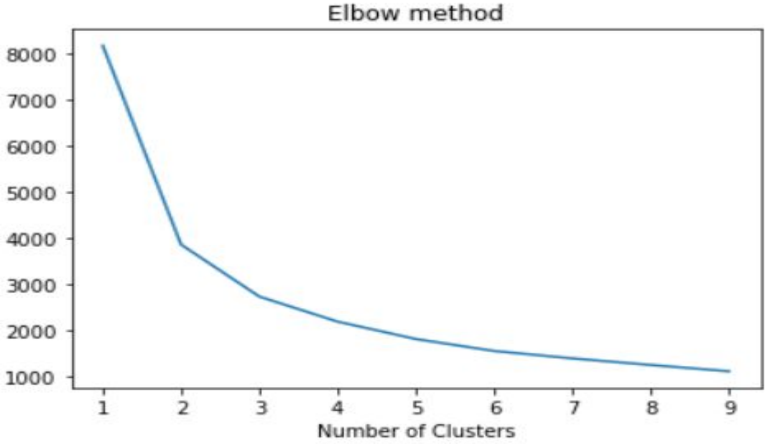

Recency
Frequency
Monetary
Recency
Frequency
Monetary
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Before log transformation

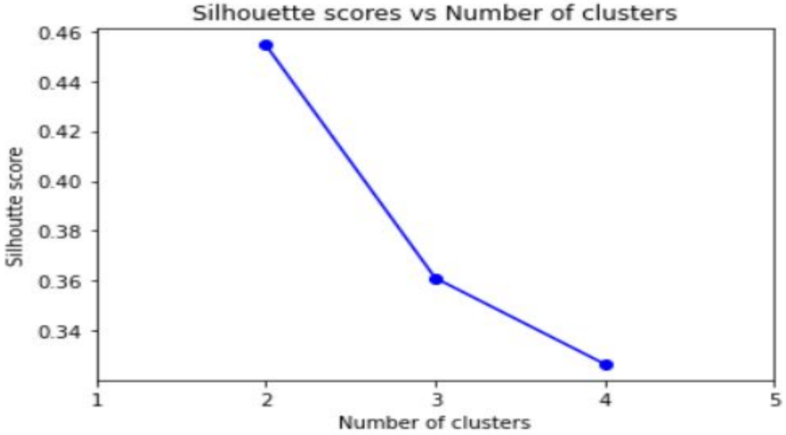


After log transformation



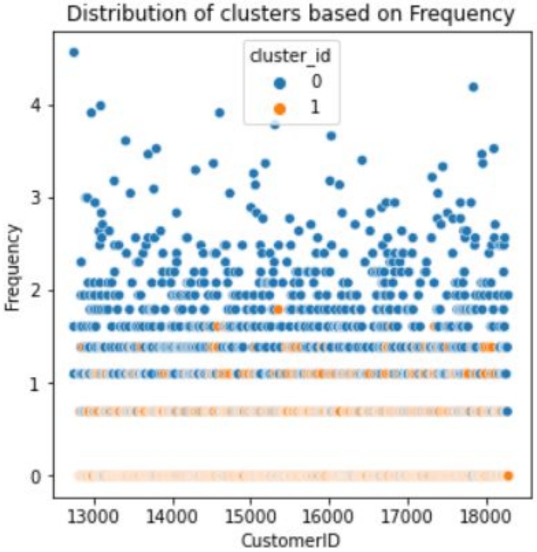


Elbow method

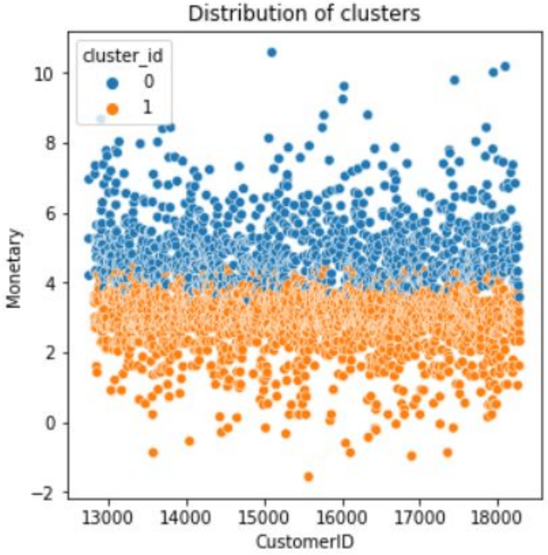


Silhouette scores vs Number of clusters

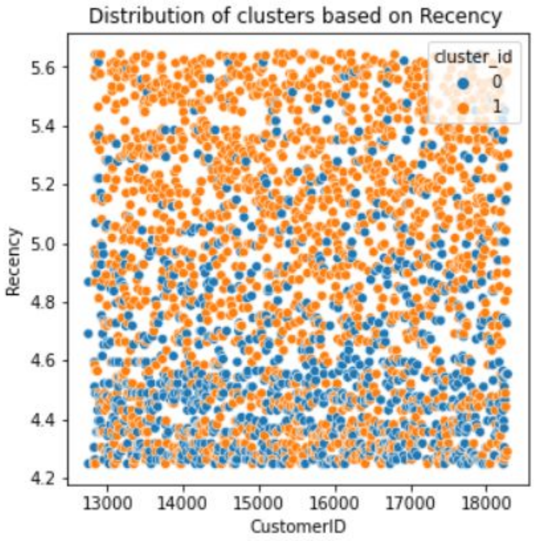
Applying K-means with k=2



Distribution of clusters based on Frequency



Distribution of clusters



Distribution of clusters based on Recency

```
1     1764
0     1100
```

THANK
YOU