CUSTOMER SEGMENTATION USING MACHINE LEARNING

V.Vijilesh[1], A.Harini[2], M.Hari Dharshini[3], R.Priyadharshini[4]

[1]*Associate Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India*

[2]*Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India*

[3]*Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India*

[4]*Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India*

Abstract - *In carrying out successful E-Commerce , the most important things are innovation and understanding what customer wants. Now-a-days the ease of using ecommerce encourages the customers to buy using ecommerce. It runs on the basis of innovation having the ability to enthral the customers with the products, but with such a large raft of products leave the customers confused of what to buy and what not to. According to business , a company may create three segments like* ***High*** *(Group who buys often , spends more and visited the platform recently) ,* ***Medium*** *(Group which spends less than high group and is not that much frequent to visit the platform) and* ***Low*** *(Group which is on the verge of churning out). This is where Machine Learning provides a crucial solution , several algorithms are applied for revealing the hidden patterns in data for better decision making. In this paper we proposed a Customer segmentation concept in which the customer bases of an establishment is divided into segments based on the customers' characteristics and attributes. This idea can be used by the B2C companies to outperform the competition by developing uniquely appealing products and services and make it reach to potential customers. This approach is implemented using "k-means", an unsupervised clustering machine learning algorithm.*

Keywords: E-Commerce, Innovation, Business segments, B2C, Machine Learning, K-means clustering.

1. INTRODUCTION

Customer Segmentation means grouping the customers based on marketing groups which shares the similarity among customers. To be more exact, it means segmenting customers sharing the common characteristics which is the best way of marketing. Customer segmentation is gathering information about each customers and analysing it to identify the different patterns for creating the segments. Some of the best methods for gathering information are face-to-face interviews, Telephonic interviews, through surveys or through research using information which are published related to market categories. The basic information which includes billing information, shipping information, products purchased, promo codes, payment method etc., Beyond these some companies also collect information like reason for the purchase, advertisement channel which makes them to purchase, age, gender etc., In B2B(Business to Business) marketing customers are grouped according to numerous factors like Industries, number of employers, Products purchased from the company in earlier times and location. On other-hand, in B2C(Business to Consumer) marketing companies segment the customers based on Age, Gender of the customers, their marital status, life stage of the customers like single, married, divorced, retired etc., On of the main factor of B2C is Location of customers (Rural,suburban, urban).Customer segmentation can be practiced for all the businesses nevertheless of size or industry. Common segmentation types include Demographic, RFM (Recency, Frequency, Monetary) analysis, HVCs (High-value customer), customer status, Behavioural, psychographic etc., Some of the major benefits of customer segmentation include Marketing strategy, promotion strategy, Budget efficiency ,product development etc., In this article we applied the basic analytics functionality to provide the decision makers(in our case the business investors) with the required information to make the right decision. In this article we define a solution for reducing risk factors and also contribute to decision making for new business investments. We proposed to use K-means technique for customer segmentation. Our solution is segmenting the customers based on information analytics. Consumers can be divided into groups in relation to common behaviours they share. Such behaviours link to their knowledge of, attitude toward, use of, or spending score or response to a product. We used machine learning Clustering algorithm K-Means for this customer segmentation.

2. METHODOLOGY

There are three major sections :

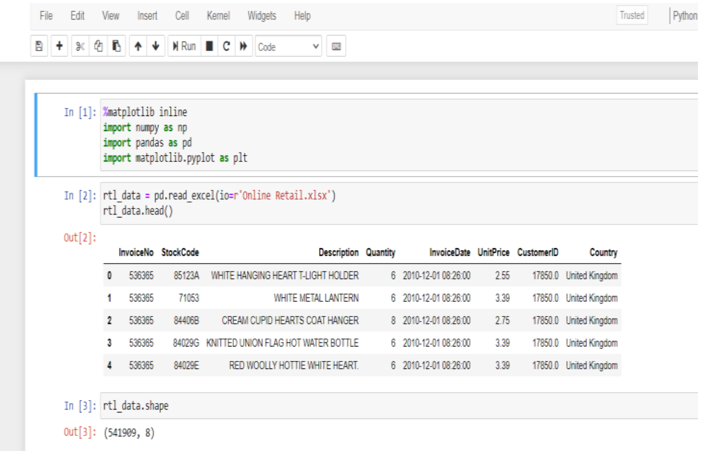
- Data Pre-processing.
- RFM score calculation.
- Cluster Creation.

A python program has been developed and the program is executed in Jupyter Notebook by importing the following necessary packages :

- **Pandas** – Loading & Pre-processing of data.
- **Numpy** – numeric data calculation.
- **seaborn, matplotlib** – Visualization.
- **scikit**-learn Machine Learning library.

2.1 Data Pre-processing :

The dataset which is an excel file is loaded using Pandas.



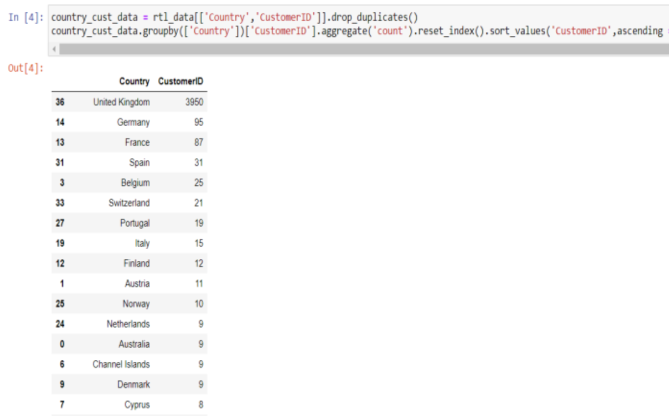
```
In [1]: %matplotlib inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt

In [2]: rtl_data = pd.read_excel(io=r'Online Retail.xlsx')
        rtl_data.head()

Out[2]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

```
In [3]: rtl_data.shape

Out[3]: (541909, 8)
```

Fig-1: Dataset loaded using Pandas

We dropped all the duplicate entries in **CustomerID** and **Country** column. Count of customers for each country is calculated and sorted the count in descending order to see from which country the maximum number of customers purchases.

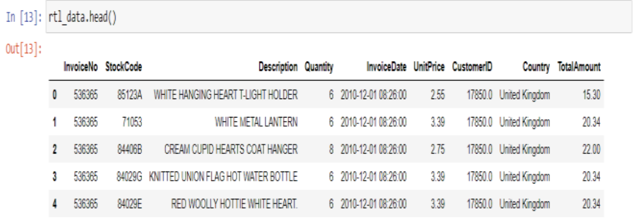


```
In [4]: country_cust_data = rtl_data[['Country','CustomerID']].drop_duplicates()
        country_cust_data.groupby(['Country'])['CustomerID'].aggregate('count').reset_index().sort_values('CustomerID',ascending

Out[4]:
```

	Country	CustomerID
36	United Kingdom	3950
14	Germany	95
13	France	87
31	Spain	31
3	Belgium	25
33	Switzerland	21
27	Portugal	19
19	Italy	15
12	Finland	12
1	Austria	11
25	Norway	10
24	Netherlands	9
0	Australia	9
6	Channel Islands	9
9	Denmark	9
7	Cyprus	8

Fig-2: Count of customers in descending order

From the first few rows we observed that maximum number of customers are from UK in Country. So we grouped the customers based on country. So we filtered out other countries using "query" method. In the consolidated dataset, only the **Description** and **CustomerID** columns had null values. Those entries are dropped off. Also the negative entries in **Quantity** as it can not be negative. **Invoice_Data** which is in string format and that is converted to date and time format as it is necessary while calculating Recency. A new column **Total_Amount** is added as a product of Quantity and Unit_Price for each customer. This is useful in Monetary calculation.



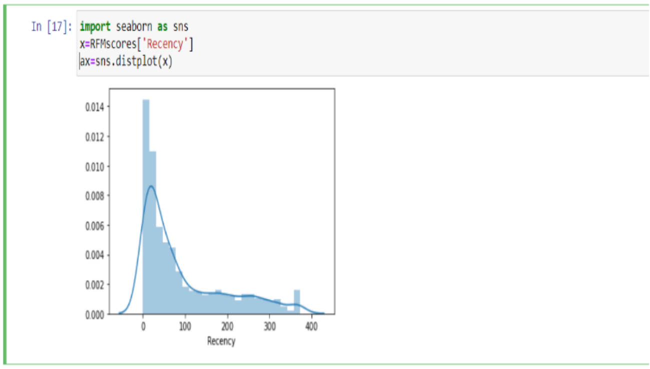
```
In [13]: rtl_data.head()

Out[13]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalAmount
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

Fig-3: Customer Grouping based on Country

2.2 RFM Score Calculation

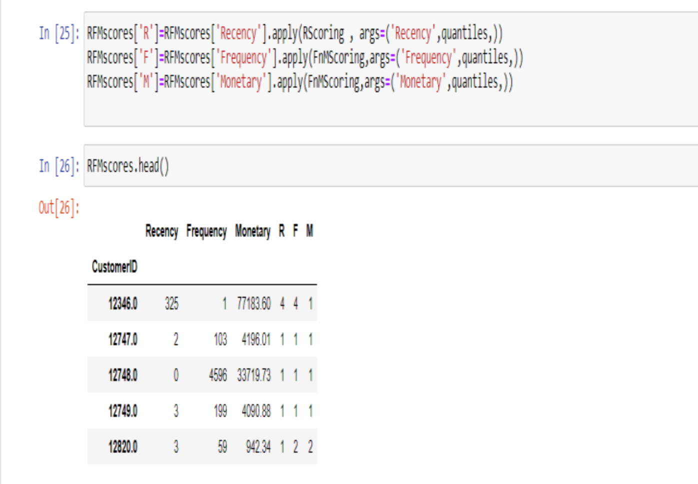
Our dataset is limited to sales record, we can use a **RFM** based model for finding segments where R is Recency (how recently a purchase happened), F is Frequency (how frequent transactions are made), M is Monetary value(Value of all transactions). Recency, Frequency and Monetary score for each customer is calculated. The latest date is assigned as a placeholder to calculate recent purchases. All the transactions are grouped using **CustomerID** and then aggregate lambda operations are performed. As a result of this operation numbers will be obtained which depicts the recency., frequency and how much a specific customer spent till date. All these are stored in a new dataframe **RFMscores.** To note, the distribution for recency is right skewed.



```
In [17]: import seaborn as sns
         x=RFMscores['Recency']
         ax=sns.distplot(x)
```

Chart-1: Recency plot

Using **quantiles** data is divided into four groups (we can choose quantile values of our own). After RFM calculation the newly added columns are as follows:

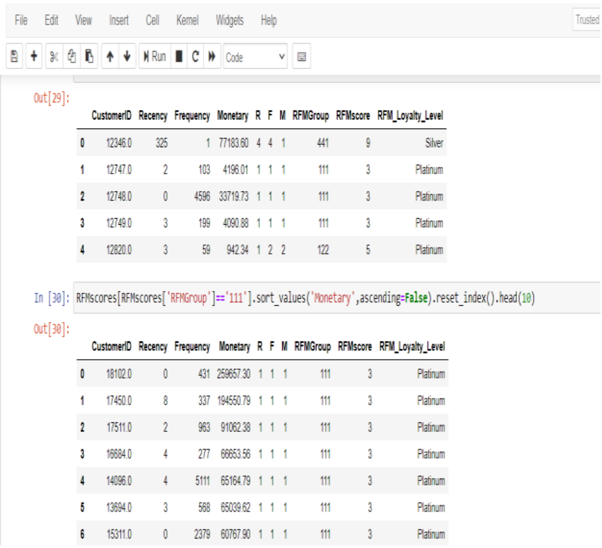


```
In [25]: RFMscores['R']=RFMscores['Recency'].apply(RScoring , args=('Recency',quantiles,))
         RFMscores['F']=RFMscores['Frequency'].apply(FnMScoring,args=('Frequency',quantiles,))
         RFMscores['M']=RFMscores['Monetary'].apply(FnMScoring,args=('Monetary',quantiles,))

In [26]: RFMscores.head()

Out[26]:
```

CustomerID	Recency	Frequency	Monetary	R	F	M
12346.0	325	1	77183.60	4	4	1
12747.0	2	103	4196.01	1	1	1
12748.0	0	4596	33719.73	1	1	1
12749.0	3	199	4090.88	1	1	1
12820.0	3	59	942.34	1	2	2

Fig-4: RFM Calculation

The individual recency , frequency and monetary values are concatenated and converted to string using map function. This is done to easily check which group the customer belongs to. This RFMscore column shows the loyalty of engagement of the customer. In our case, the lower the value of RFMscore, more loyal the customer will be as well as more engaged he/she would be. Based on this in the next step **Loyalty_Level** like Platinum, Gold, Silver and Bronze levels are assigned to each customer. From this we could derive a conclusion that if the customer is in platinum group we can say that they are the best customers whereas in bronze group, the customer haven't purchased for a longer time. With this a company can decide like providing special attention, offers and priority access to newly launched products to their platinum customers. On the other hand, if the customer falls into the bronze group, the company can give some rewards or coupons to encourage the spending score of them.



```
Out[29]:
```

	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMscore	RFM_Loyalty_Level
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum
2	12748.0	0	4596	33719.73	1	1	1	111	3	Platinum
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum

```
In [30]: RFMscores[RFMscores['RFMGroup']=='111'].sort_values('Monetary',ascending=False).reset_index().head(10)

Out[30]:
```

	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMscore	RFM_Loyalty_Level
0	18102.0	0	431	259657.30	1	1	1	111	3	Platinum
1	17450.0	8	337	194550.79	1	1	1	111	3	Platinum
2	17511.0	2	963	91062.38	1	1	1	111	3	Platinum
3	16684.0	4	277	66653.56	1	1	1	111	3	Platinum
4	14096.0	4	5111	65164.79	1	1	1	111	3	Platinum
5	13694.0	3	568	65039.62	1	1	1	111	3	Platinum
6	15311.0	0	2379	60767.90	1	1	1	111	3	Platinum

Fig-5: RFM Loyalty level

To visualize, following is the scatter plot of Loyalty level and RFM score of recency against frequency. It is observed that levels are grouped together.

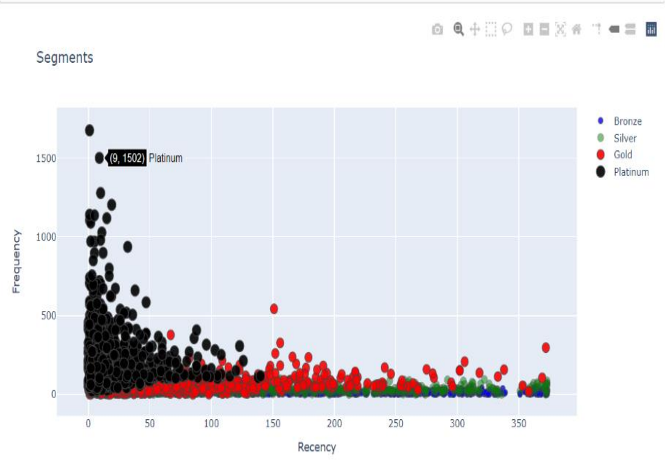


Fig-6: Scatter plot of Loyalty level & RFM score of recency against Frequency

2.3 Clustering

K-Means is an unsupervised learning algorithm and used for clustering tasks which works really well with complex dataset. It is an iterative algorithm that partitions the dataset into "k" pre-defined non-overlapping subgroups (clusters) where each data point belongs to **only one group**.

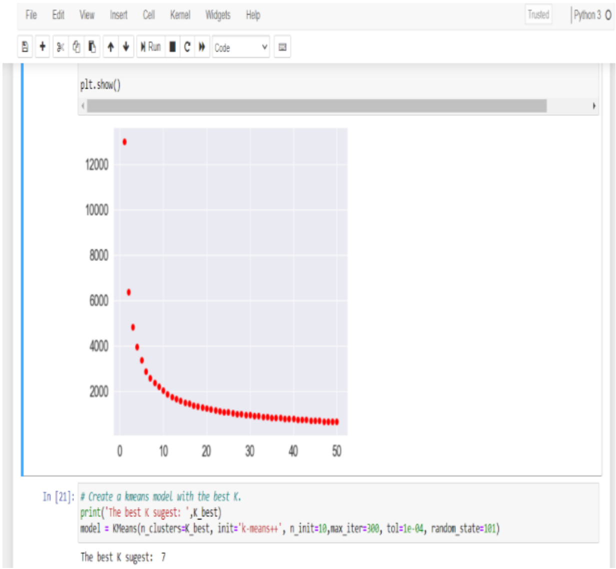
The algorithm works as follows:

Step-1 : Specifying the number of clusters – k value.

Step-2 : Centroids are initialized by shuffling the dataset and then randomly selecting k data points for the centroids without replacement.

Step-3: Repeat the iteration until there is no change to the centroids. i.e, assignment of data points to the clusters does not change.

Recency, Frequency and Monetary are brought to the same scale and the data is normalized before clustering process. It is important to determine the optimum number of clusters i.e, "k value". For this we used **Elbow method.** It involves running the algorithm multiple times over a loop with an increasing number of cluster choice and then plotting a score as a function of the number of clusters. When "k" increases, the centroids are closer to cluster centroids. The improvement will decline at some point rapidly creating an elbow-like shape in graph and that is the whole reason this method is called as elbow. We take the count of cluster, k-value at the point where this elbow is bending. When we executed this metric, the result was not obvious and the bend is not clear as there was rapid decline at three values **– 3, 5 and 7.**



```
plt.show()
```

```
In [21]: # Create a kmeans model with the best K.
         print('The best K sugest: ',K_best)
         model = KMeans(n_clusters=K_best, init='k-means++', n_init=10,max_iter=300, tol=1e-04, random_state=101)

The best K sugest:  7
```

Chart-2: K value using Elbow method

Silhouette method which is considered as a better metric than elbow is used to determine the optimum number of clusters. Silhouette score for each sample is calculated using the formula:

Calculate the silhouette s(i) as follows , ratio of the difference between cluster cohesion and separation to the greater of the two :

s(i) = b(i) – a(i) / [max{a(i),b(i)}]

Silhouette co-efficient ranges from [-1,1]. The calculated average score is 0.3. Any value higher than this means it is well matched to its own cluster.

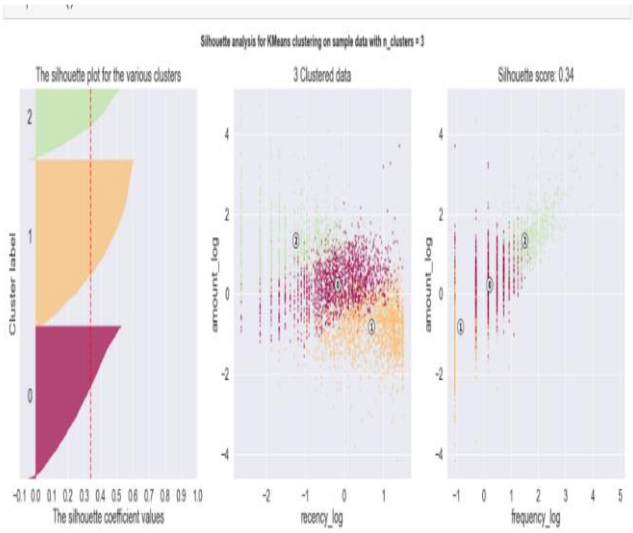
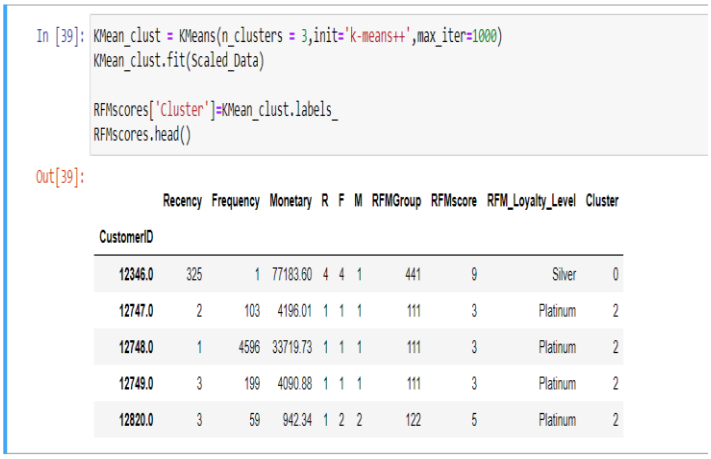


Fig-7: Silhouette analysis for KMeans clustering on sample data with n_clusters=3

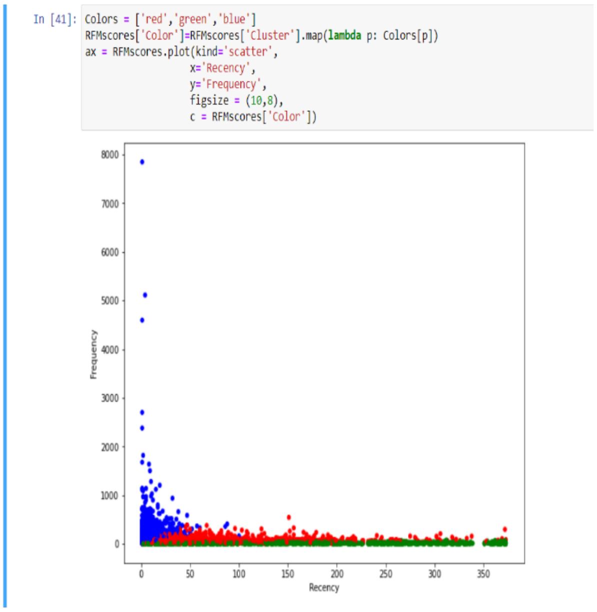
From the outputs , we can infer that cluster 3 has a score 0.34 and the fluctuation size is similar. To note that the thickness of the silhouette plot representing each cluster also contributes to the decision. Thus the optimum value of k is 3. k-means is imported from sklearn libray and the arguments are set and the operation is applied on the scaled data. In RFM we got 4 groups whereas in k-means the silver and bronze groups are merged as follows :



```
In [39]: KMean_clust = KMeans(n_clusters = 3,init='k-means++',max_iter=1000)
         KMean_clust.fit(Scaled_Data)

         RFMscores['Cluster']=KMean_clust.labels_
         RFMscores.head()

Out[39]:
```

	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMscore	RFM_Loyalty_Level	Cluster
CustomerID										
12346.0	325	1	77183.60	4	4	1	441	9	Silver	0
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	2
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	2
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	2
12820.0	3	59	942.34	1	2	2	122	5	Platinum	2

Fig-8: Silver and Bronze merged to Silver group.



```
In [41]: Colors = ['red','green','blue']
         RFMscores['Color']=RFMscores['Cluster'].map(lambda p: Colors[p])
         ax = RFMscores.plot(kind='scatter',
                             x='Recency',
                             y='Frequency',
                             figsize = (10,8),
                             c = RFMscores['Color'])
```

Fig-9: Graph representation of the 3 Clusters.

3. FUTURE WORK

The proposed basic cluster model given the lack of details about the changes in customer behaviours. Therefore different rules and strategies are necessary to find the hidden patterns and shopping trends of the customers. RFM and K-means helped to find clusters of potential customers. In addition to this **Cross Selling** and **Market Basket Analysis** techniques can be used to analyse and offer additional products to customers as a suggestion in the hope that they would buy benefiting the customer and the retail establishment.

4. CONCLUSION

This paper presented an implementation of the k-Means clustering algorithm for customer segmentation using data collected from an online retail outfit. Our model has partitioned customers into mutually exclusive groups, three clusters in our case. This will be useful for applying further data mining strategies and the derived insights are helpful in decision making for the business wings.

REFERENCES

[1] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K - means Clustering," International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.

[2] Nurma Sari, Juni & Nugroho, Lukito & Ferdiana, Ridi & Santosa, Paulus. Review on Customer Segmentation Technique on Ecommerce. Advanced Science Letters. 2016.

[3] ChanC, Swatman PMC. Management and business issues for B2B ecommerce implementation. Proc 35th Annual Hawaii IntConfSystSci.

[4] Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.Silverman, and A. Y. Wu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.

[5] Aloise, A. Deshpande, P. Hansen, and P. Popat, “The Basis Of Market Segmentation”Euclidean sum-of-squares clustering,” Machine Learning, vol. 75, pp.245-249, 2009.

[6] Effective Cross Selling, Springer Science and Business Media LLC, 2018.

[7] Chinedu Pascal, Simeon Ozuomba. "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services", International Journal of Advanced Research in Artificial Intelligence, 2015.

[8] F.Daniel, Customer Segmentation: classification, clustering , marketing . www.kaggle.com

[9] Rachel Blasucci. Event triggered Customer Segmentation. DZone, July 23, 2018.