**CUSTOMER SEGMENTATION USING RFM ANALYSIS FOR AN ONLINE RETAIL STORE**

**Khushboo Singhal :** [**khushboo16csu172@ncuindia.edu**](mailto:khushboo16csu172@ncuindia.edu)

**Khyati Bhargava :** [**khyati16csu173@ncuindia.edu**](mailto:khyati16csu173@ncuindia.edu)

**ABSTRACT:**

Online retailing (electronic retailing) for purchasing goods is increasing and so is the need to segment customers is necessary. Customer segmentation is done based on customers’ past purchase behavior and then divide them into different categories i.e. loyal customer, potential customer, new customer, customer needs attention, customers require activation. This paper uses RFM (Recency, Frequency, Monetary Value) analysis whichis one of the most widely used techniques for groupingcustomers. The implementation of an efficient customer segmentation is necessary for all retailing stores to take better management decisions.

**Keywords:** Online Retail; Customer Segmentation; RFM-model;K-means clustering; Decision Tree

**INTRODUCTION:**

Online retailing (Electronic Retailing) means the efficient usage of internet to purchase goods and services by the customer. E-Retailing is divided into 2 parts:

* [business-to-business](https://www.investopedia.com/terms/b/btob.asp)
* business-to-consumer

|  |  |
| --- | --- |
| **Business-to-Consumer E-Retailing** | **Business-to-Business E-Retailing** |
| In this type of retailing, companies sell its products and goods to the customer directly with the help of internet through their website. This helps them to maintain good relationship with customers. | In this type of retailing, companies sell its services to other companies. Wholesalers sell their goods and products directly from the manufacturing unit to the retailers. |

In this paper we are working upon business to consumer sales data to segment customers based on different categories i.e. loyal customer, potential customer, new customer, customer needs attention, customers require activation using RFM analysis.

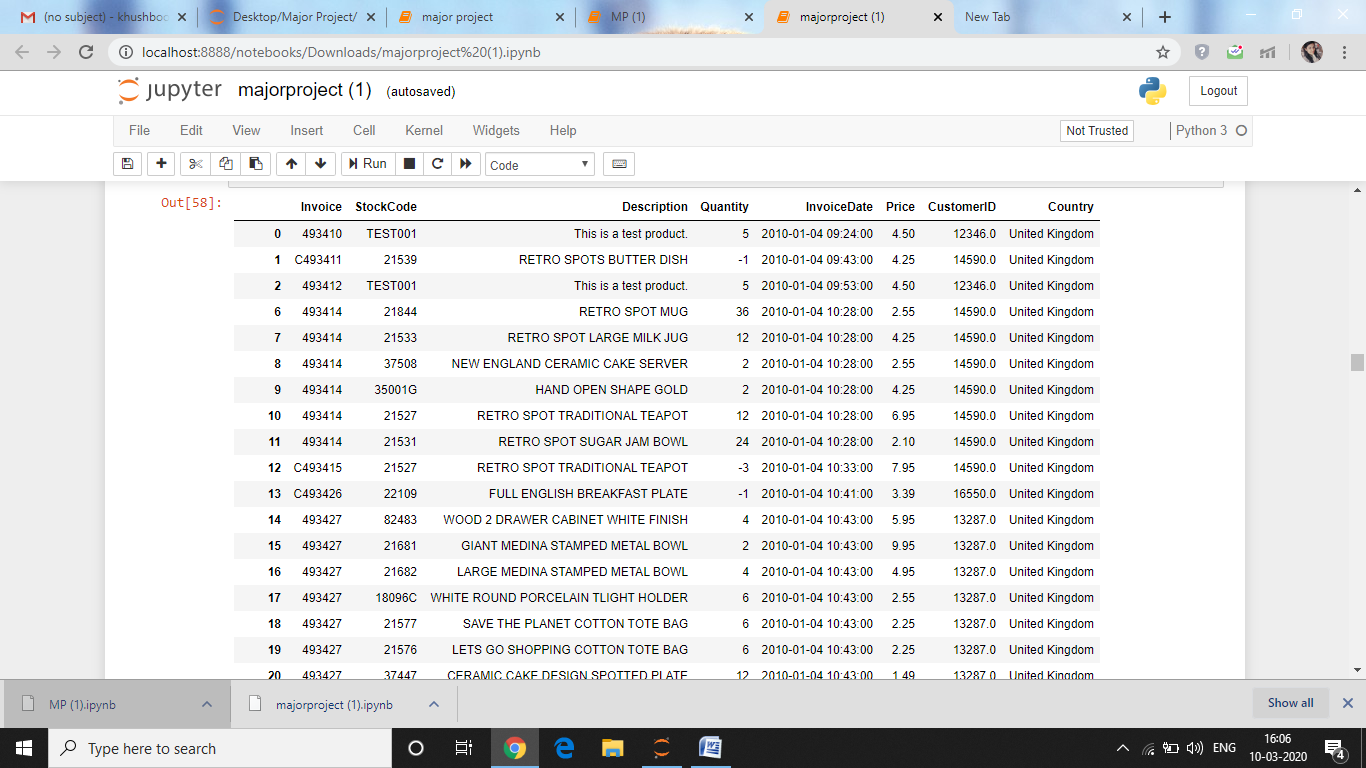
* Recency means how recent a transaction is. Recency is number of months since Dec when last transaction for the year is done by a customer.
* Frequency is frequency of buying made by a customer in the year.
* Monetary value is the total amount spent by a customer in the year covering all transactions for each customer.

|  |  |  |
| --- | --- | --- |
| **Customer Category** | **General Definition** | **Definition in terms of RFM** |
| **Loyal Customer** | This type of customer keeps coming back to purchase goods from the specific retail store. | Recency should be minimum frequency and monetary value should be maximum. |
| **Potential Customer** | This type of customer has high potential to enter into loyal customer segment. | Recency should be minimum; frequency and monetary value should be high but less than loyal customers’ frequency and monetary value. |
| **New Customer** | This type of customer is the fresh customer that just bought something from the retail store. | Recency should be minimum frequency and monetary value should be minimum as he has just bought 1 or 2 products from the store. |
| **Customer needs attention** | This type of customer has made some initial purchase but have not seen them since 5 to 6 months. | Recency should be medium and having some frequency and monetary value which indicates that he has purchased some products earlier. |
| **Customer requires activation** | This type of customer might have gone with the competitors for now. | Recency should be high and having some frequency and monetary value which is less than customers needs attention’s frequency and monetary value. |

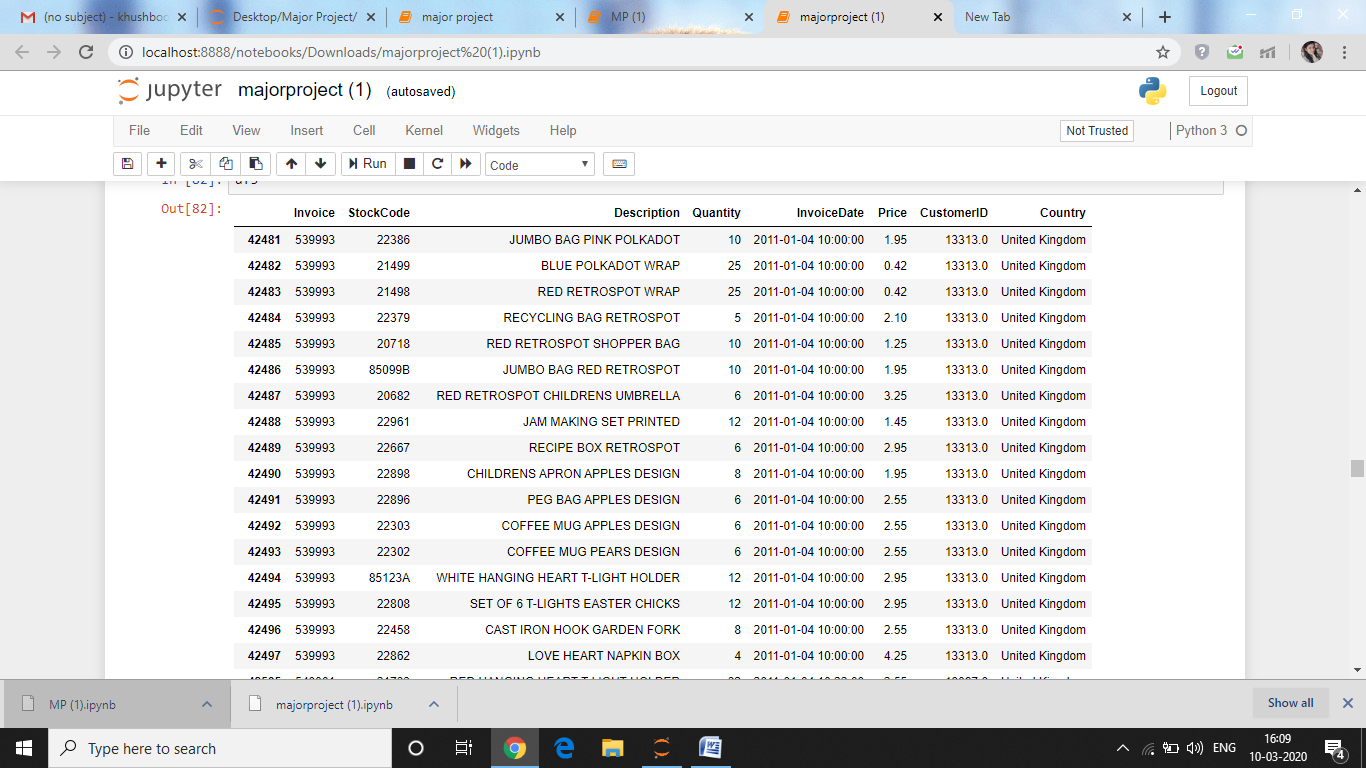
However, even the simplest of financial organization go forward with at least one unsupervised Machine Learning to segment customers. The most commonly used method for grouping customers via K-means clustering. K-means is a very efficient and easy way of clustering similar things together but the only disadvantage is that parameter differences between points of different cluster should be really big otherwise it becomes difficult for K-means to identify data points of different clusters as two different clusters. Going forward we have used K-means as the method to cluster customers according to RFM which helps in identifying which customer should be in which category on these existing cluster of data points. Also, we have used elbow curve to compute the value of ‘K’ by running K-means algorithms on different values of K and taking the minimum value before the saturation of cluster per category when plotted as a curve. The second method which we have used is Supervised Machine Learning method i.e. Decision tree to segment customers. Decision Tree splits the data with respect to a specific parameter. Decision tree is divided into two entities: decision nodes and leaves. The leaves are the final outcomes and decision node isthe entity from where the data splits.

**BUSINESS BACKGROUNDAND THE ASSOCIATED DATA**

The online based retailing dataset has 8 factors as appeared in Table 1 and Table 2, and it shows all the transactions occurred in years 2010 and 2011. The Customer variableis primary for the business to provideimportant information that makes each costumer distinct, and then it does analyses to categorize customers and business to take better decisions.



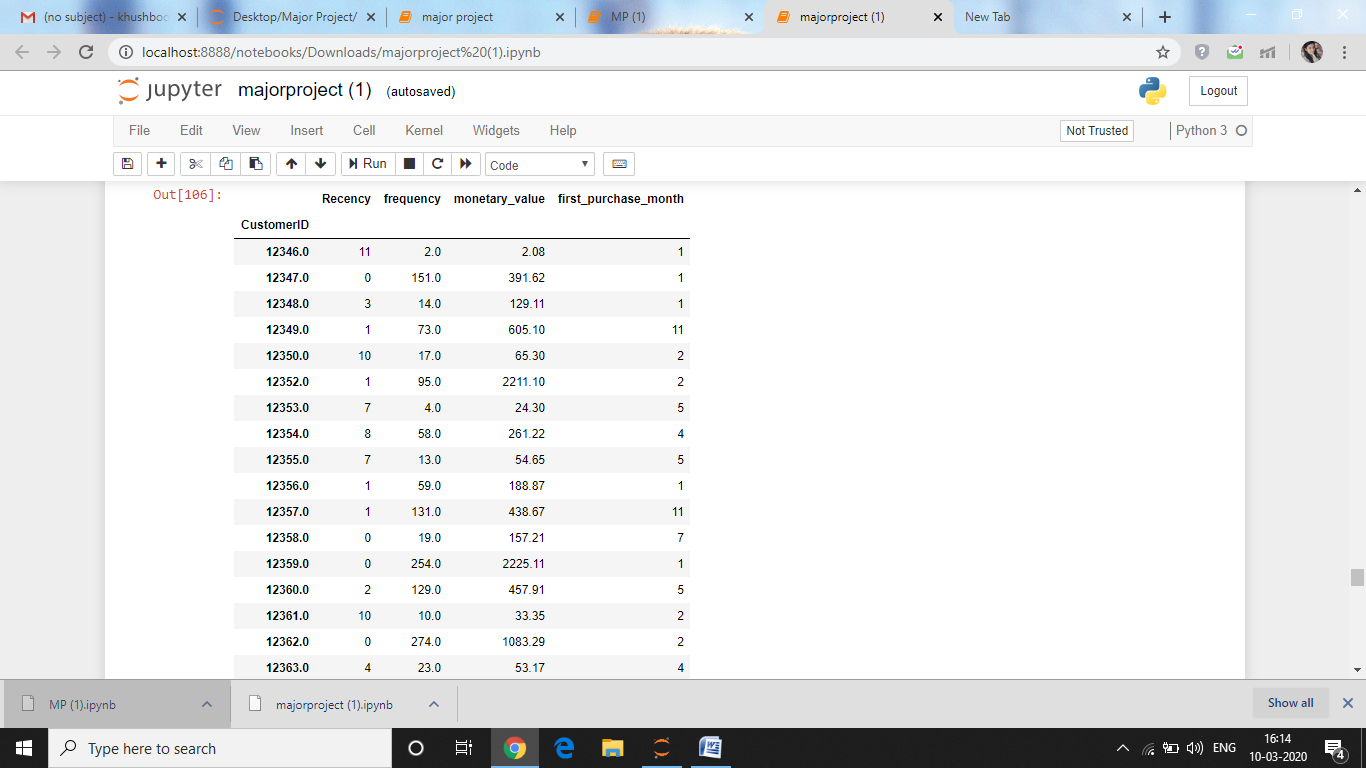
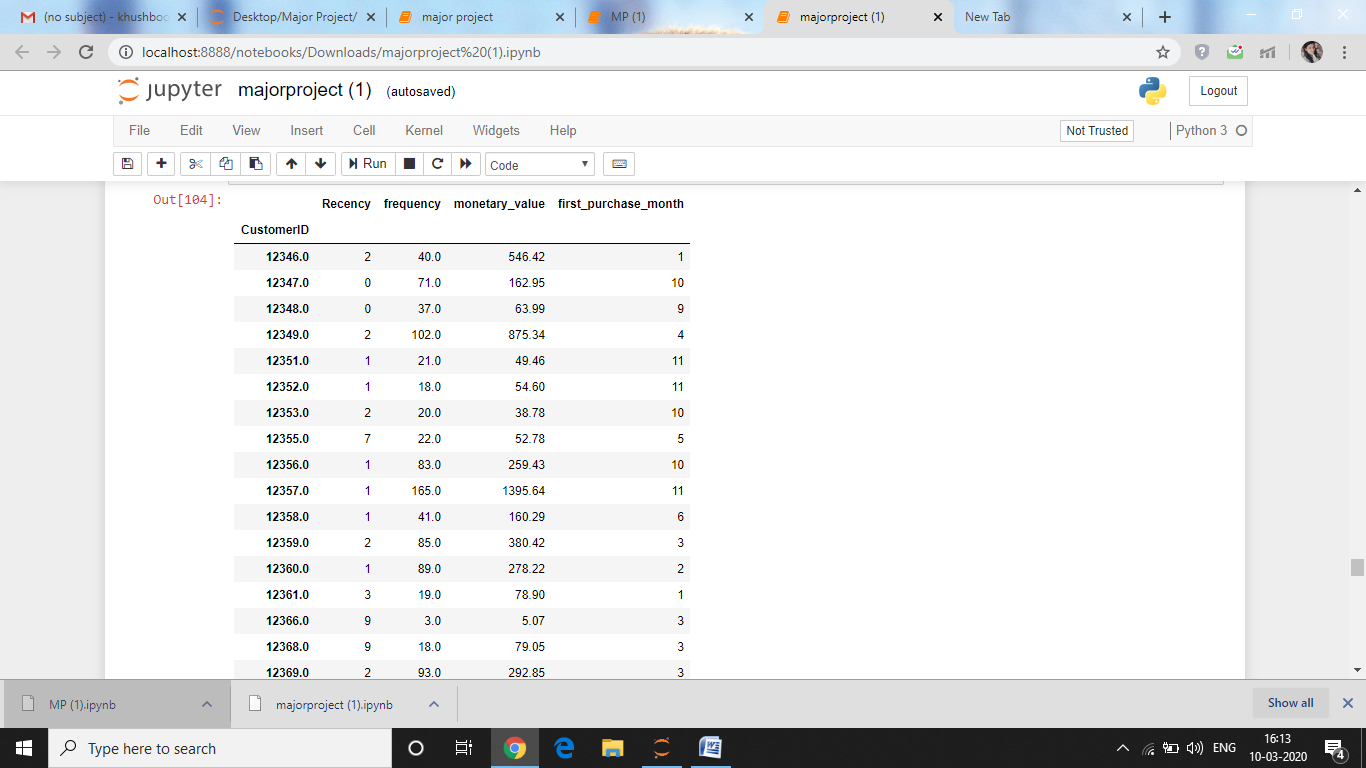
**TABLE-1**



**TABLE -2**

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Data type** | **Description** |
| **Invoice** | Numeric | Invoice variable is a 6-digit number allotted to each transaction |
| **StockCode** | Numeric | Stock code variable is a 5-digit number combined with alphabets is alloted to each distinct product. |
| **Description** | Numeric | Description variable is to describe or to name the product in words. |
| **Quantity** | Numeric | Quantity variable is assigned to each product per transaction |
| **InvoiceDate** | DateTime | The date and time at each transaction was generated;For instance: 31/05/2011 15:59 |
| **Price** | Numeric | Price variable is product price |
| **CustomerID** | Numeric | CustomerID variable is a 5-digit number assigned to each customer |
| **Country** | Numeric | Delivery address country; United Kingdom |

**TABLE-3**



**TABLE 4**

**DATA PRE-PROCESSING**

To perform RFM model-based analysis, the original dataset needs to be pre-processed. The fundamental advances associated in data preprocessing are as per following :

1. Select some specific factors for data preprocessing from the given dataset. For our situationfourfactors have been selected: Price,Quantity,Invoice Date, andCustomerID.
2. Separate the dataset into 3 dataframes according to years i.e. 2010, 2011 with the help of InvoiceDate variable and then remove duplicate and null values from the dataframes.
3. Make new column month (purchase month) with the help of variable InvoiceDate to extract the month in which customer has made purchase of the product.
4. Calculate Recency by subtracting the variable month from 12 (December = 12)
5. Calculate Frequency by counting the number of purchase made per customer by variable CustomerID.
6. Calculate Monetory Value by summing the value of variable Price per CustomerID.
7. Calculate Minimum, Maximum Amount with the help of variable Price by grouping CustomerID.
8. Calculate First Purchase Month using new column that we have created i.e. monthby grouping CustomerID and then select value having maximum recency from that.

After performing the given steps , a target dataset which contains the target features and variables has been generated .The target data set and the features has been described in Table 3and 4 and statistics calculated has been described in Table 5and 6 for each year.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2010** | **MINIMUM** | **MAXIMUM** | **MEAN** | **MEDIAN** |
| **RECENCY** | 0 | 11 | 2.75 | 2 |
| **FREQUENCY** | 1 | 5762 | 96.205 | 43 |
| **MONETORY VAUE** | 0 | 39807.85 | 373.68 | 138.91 |
| **FIRST PURCHASE MONTH** | 1 | 12 | 5.17 | 4 |

**TABLE -5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2011** | **MINIMUM** | **MAXIMUM** | **MEAN** | **MEDIAN** |
| **RECENCY** | 0 | 11 | 2.93 | 2 |
| **FREQUENCY** | 1 | 7692 | 89.53 | 41 |
| **MONETORY VAUE** | 0 | 41376.33 | 311.44 | 127.78 |
| **FIRST PURCHASE MONTH** | 1 | 12 | 5.18 | 4 |

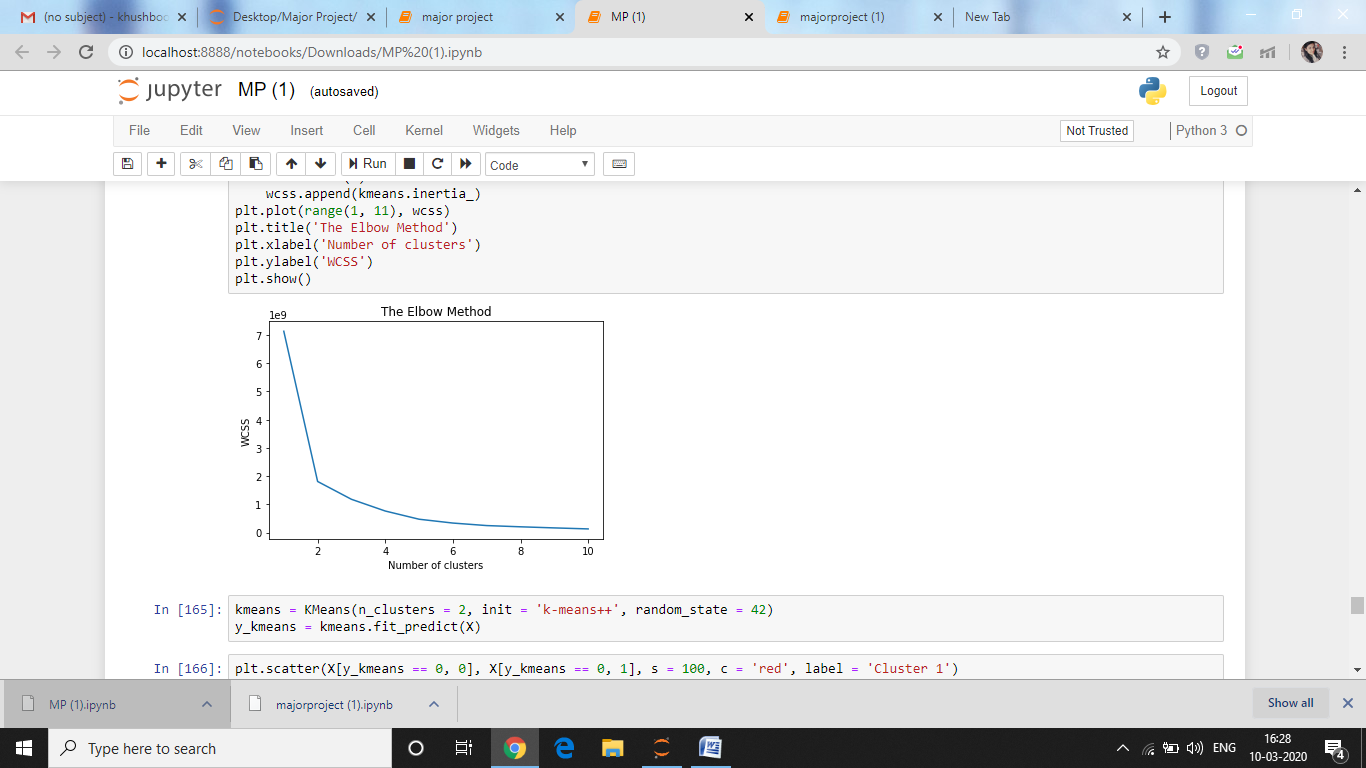
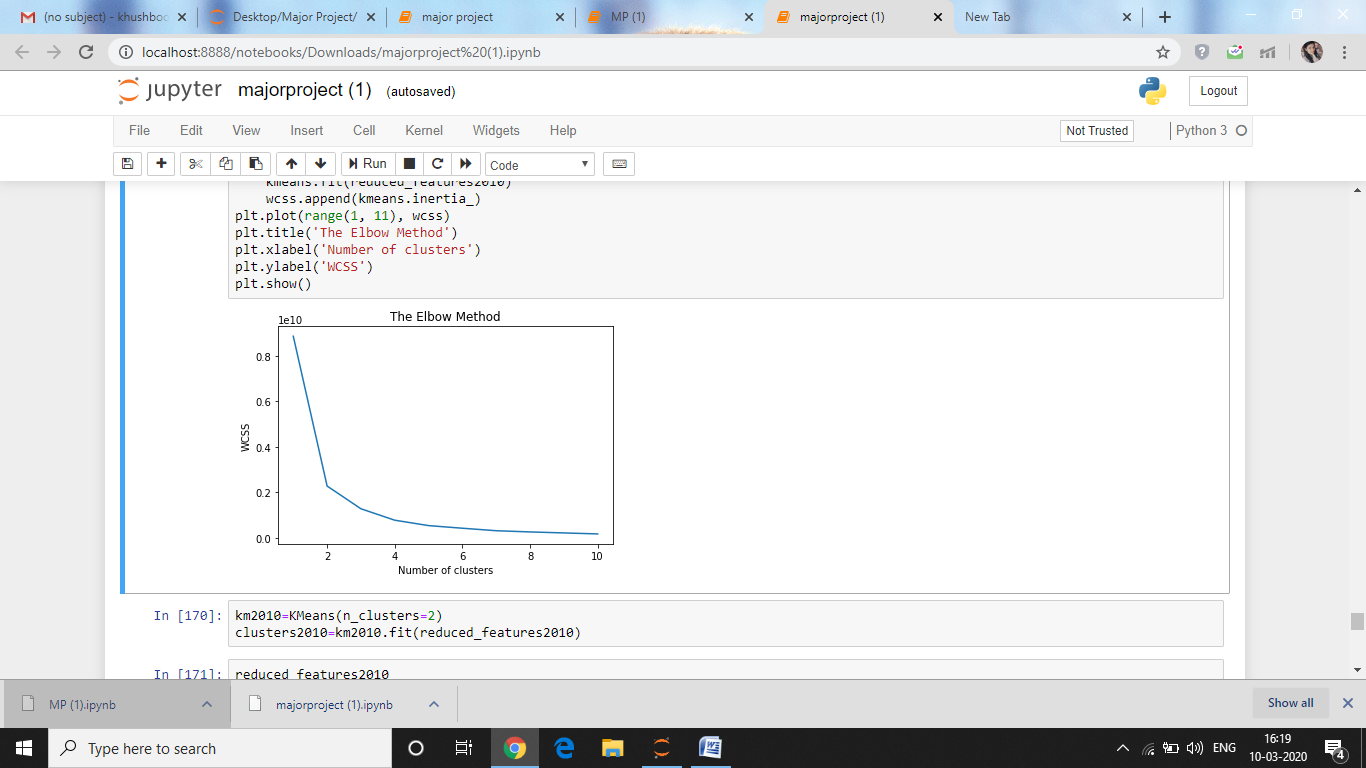
**TABLE-6**

**RFMMODELBASED CLUSTERING ANALYSIS**

With the prepared datasetin the wake of preprocessing we need to distinguish whether consumers can be segmented meaningfully in terms of recency, frequency and monetary values. The k-means clustering algorithm is applied to find the optimal value of K for clustering the Elbow Curve method was implemented. Whenwe apply k-meansclustering algorithm on a dataset, it shows outliers or variables having incomparable magnitudes.

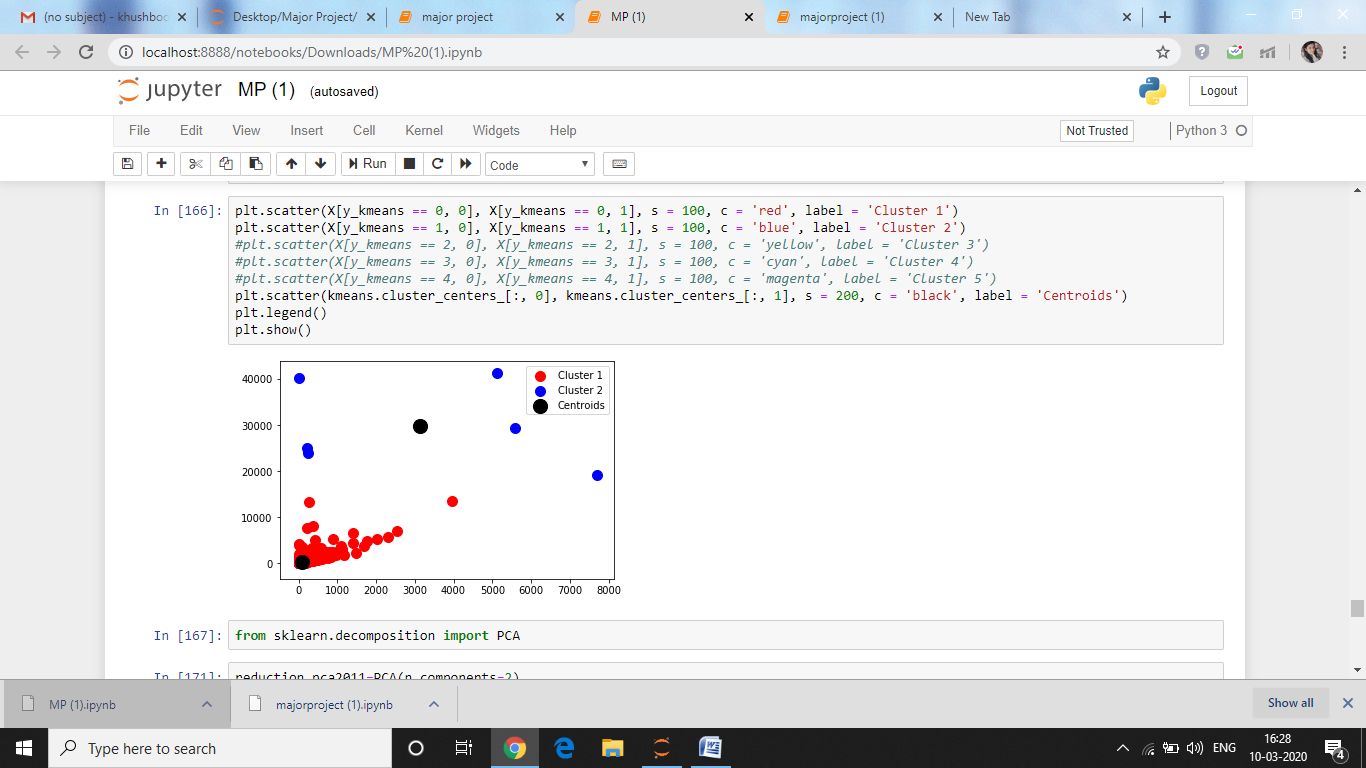
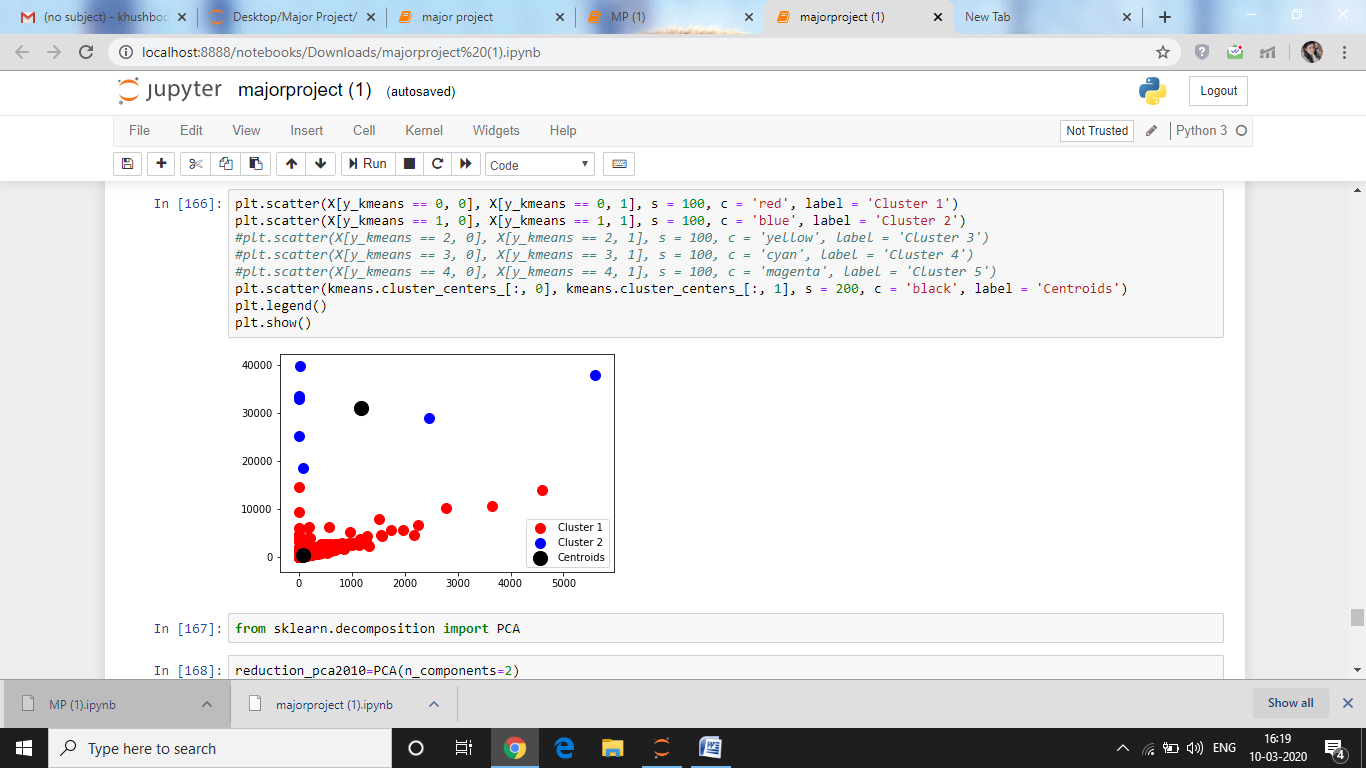
The K means clustering method required hit and miss trial for finding out the values of K that must be selected to find the best solution. Thus,so as to find the optimal solution of K ,the Elbow curve technique is used.

For the year 2010 , The Elbow curve as shown in Figure 1 , determines the value of K as 2 or 3 whichever be more suitable.Similarly for the year 2011,the elbow curve as shown in Figure 2 ,determines the value of K as 2.



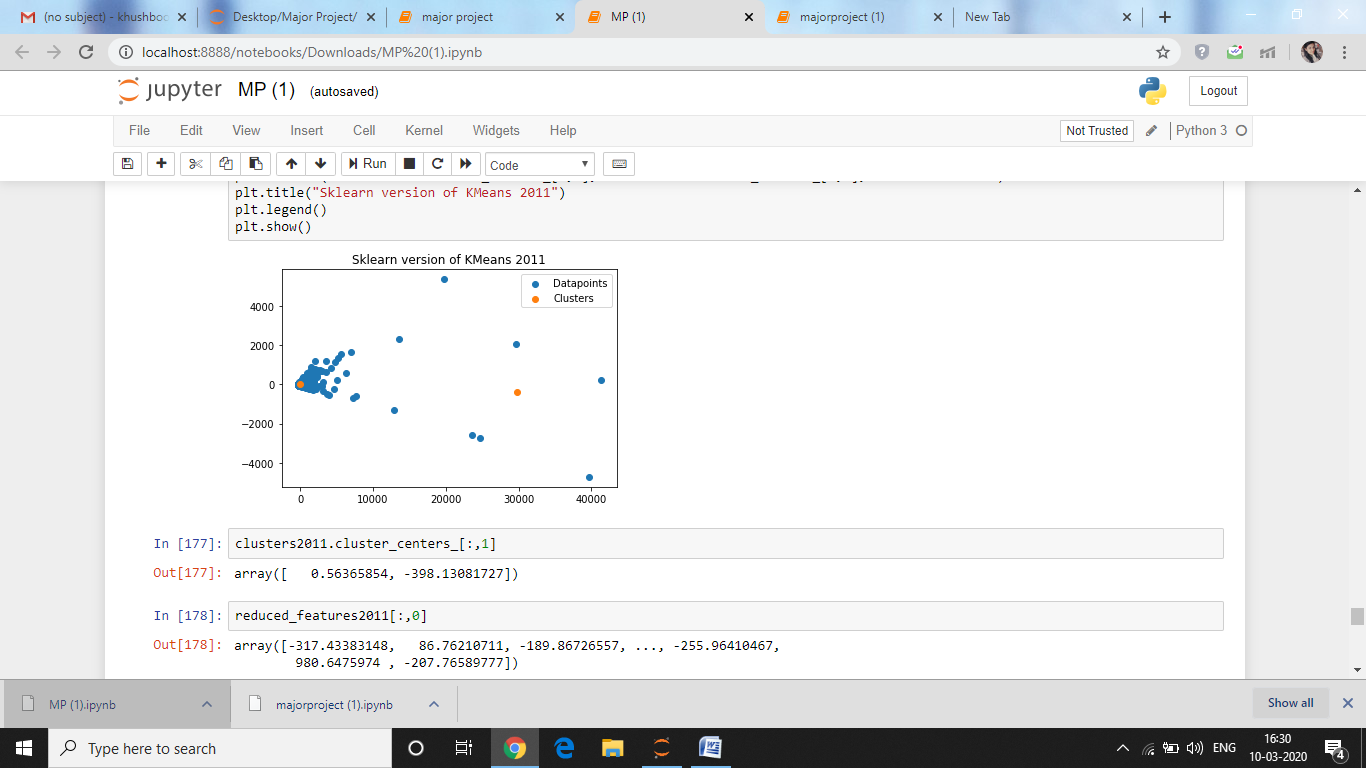
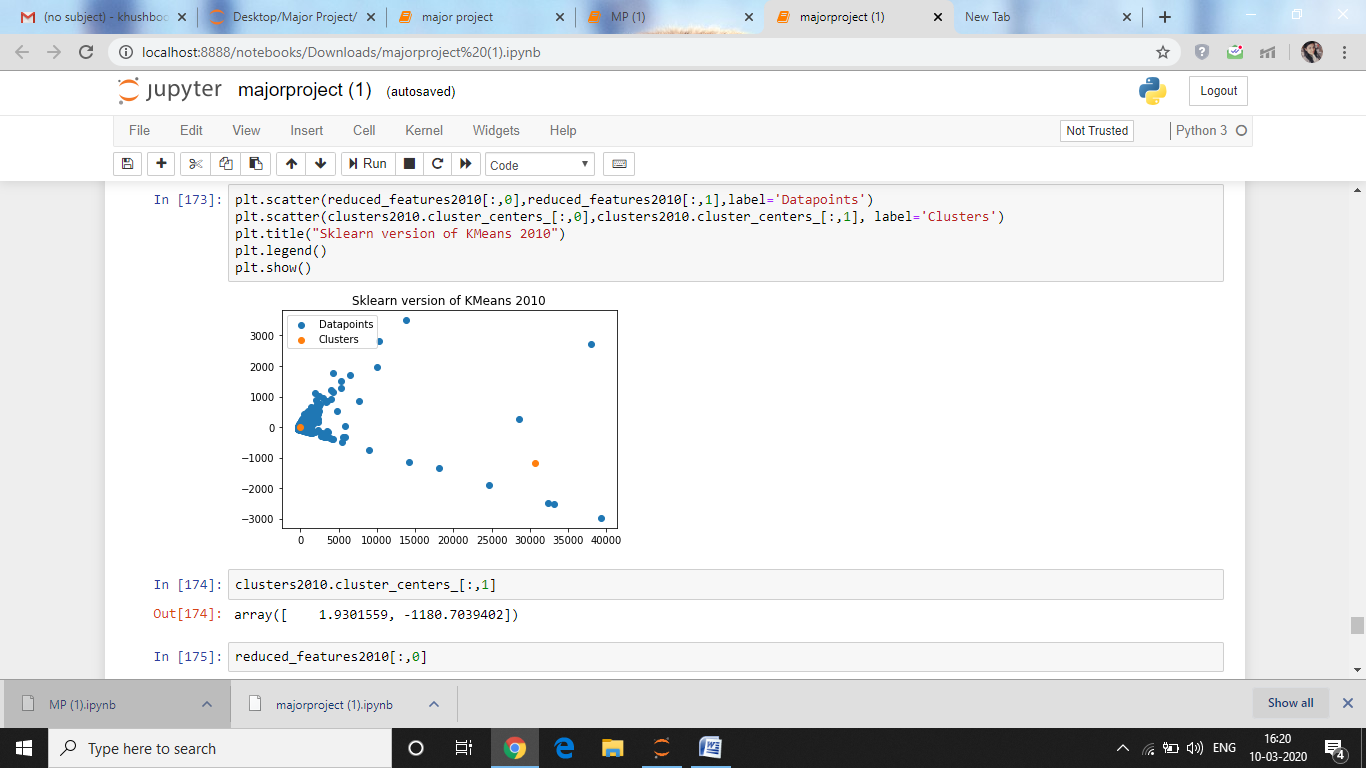
**Figure-1 Figure-2**

For applying K means clustering we need to apply PCA (Principle Component Analysis) so as to reduce the features that can be easily used in clustering process. The three important features Recency, Frequency and Monetary Value were taken and reduced features were generated according to their correlation with each other.



**Number of clusters in 2010 Number of clusters in 2011**

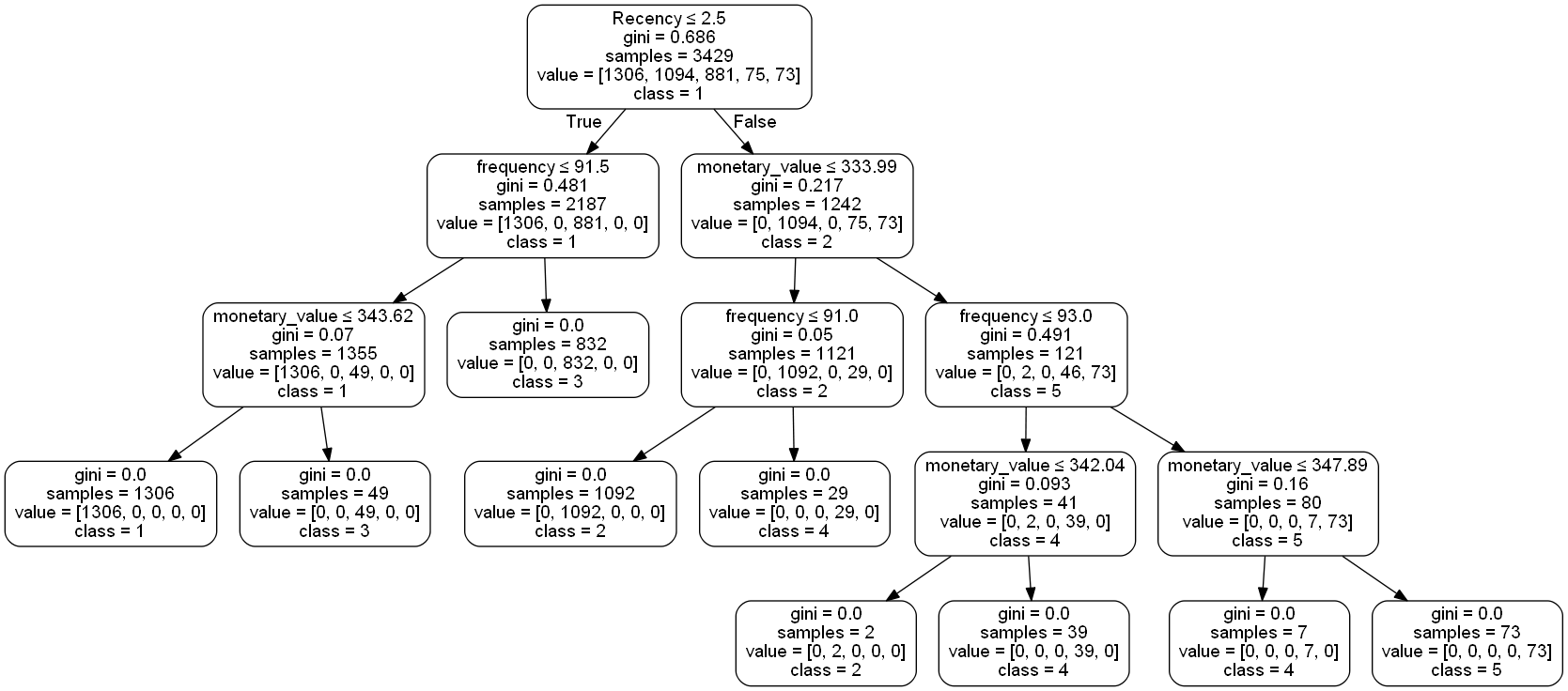
Accordingly, the clusters (Figure 1 and 2) were generated and their distribution has been shown for each year in Figure 3 and 4.



**Figure 3 Figure 4**

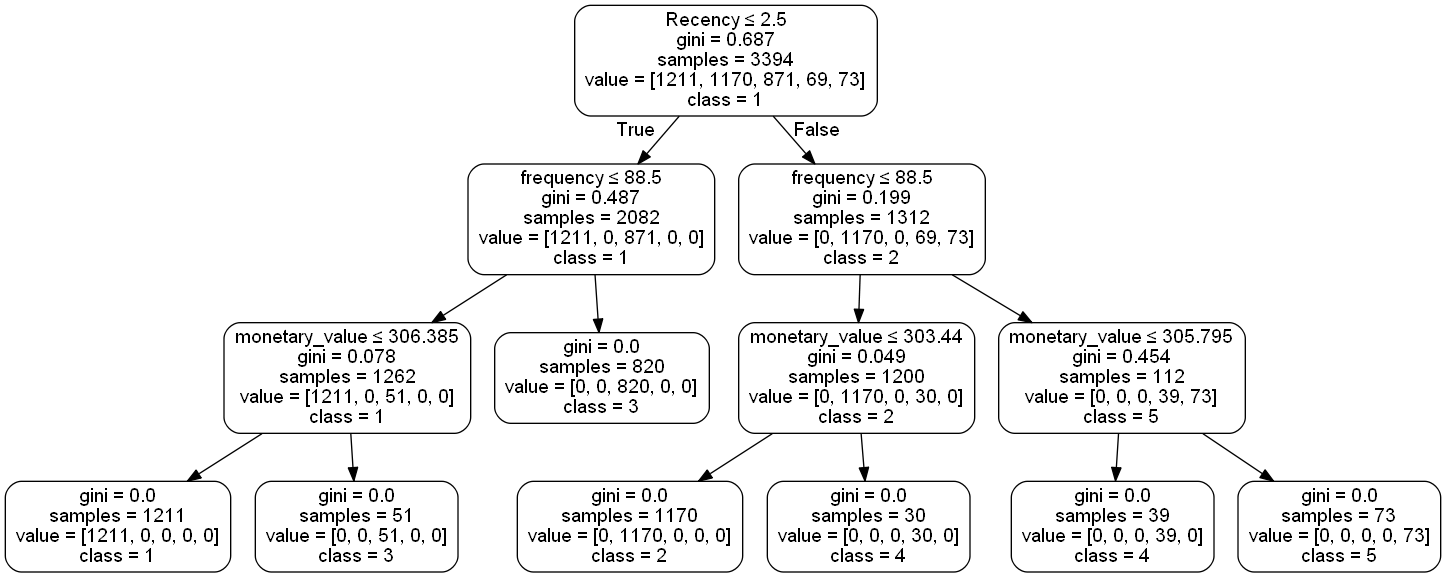
**ENHANCING K-MEANS CLUSTERING ANALYSIS USING DECISION TREE**

From k-means clustering, we analyzed that cluster 2 is the most appropriate cluster and it has been selected as the optimal because it contains both customers: new as well as old and that too for both the years.To refine segmentation of the occurrence in the cluster, a decision tree is used to create nested splitting inside the cluster. Thesedivisions form some sub-clusters inside cluster 2, andhelps to differentiate the consumers into some categories. For instance, for the year 2010, as shown in Figure 5 , the customers can be dividedinto categories such as frequency less than 91.5 and monetary value less than 333.9 ; frequency less than 93 and monetary value less than 347 with initial splitting node as recency less than 2.5.



**Figure-5**

Similarly, decision tree was constructed for year 2011 as shown in Figure-6.



**Figure-6**

According the classes were assigned which divided the customers into the 5 sets of classes.

**CUSTOMER-CENTRIC BUSINESS INTELLIGENCE AND RECOMMENDATIONS**

For each of thecostumer groups, it is necessary to find out which of the items have been purchased by the customer of each group and how frequently they have purchased the product along with in which order they have purchased it. With the assistance of customer groups and the products they have purchased, the business can take better management decisions. Customer classification can also help in maintaining and engaging in business. Most of the customers were the organization who have been purchased products in higher quantity per transaction. Examining the frequency or how recent the products have been purchased would help the business for better understanding. Customer classification is done on the basis of different types of customers and their pattern to purchase the product which lead them to divide into some specific categories.

Customer categories helps to monitor the most to least potential customer and how much profit they earn from each category which helps to run the business more efficiently. Identifying appropriate categories for such predictions is extremely useful.It might enable the business to investigate different components that may influence client's purchasing expectation and inclinations.

**CONCLUDING REMARKS**

A contextual investigation that has been introducedin this paper is used to determine the methods in which customer-centric business for online retailers can be created by means of machine learning algorithms.The distinctive client bunches that has been described help the business for better comprehension as far as their benefit, and as needs be, take proper promoting techniques and choices for every client classification. It has been shown in this analysis that there are two most crucial and time-consuming steps in the whole process includes: information preprocessing and model understanding and assessment.Further research for the business includes: categorize customers according to their buying patterns in terms of recency, frequency and monetory value into different categories that we have mentioned above; and to enhance retailer’s website so that they can track and capture customers’ shopping activities more accurate.

**REFERENCES**

1. He, Xixi, and Chen Li. "The Research and Application of Customer Segmentation on E-commerce Websites." 2016 6th International Conference on Digital Home (ICDH*)*. IEEE, 2016.
2. Hossain, ASM Shahadat. "Customer segmentation using centroid based and density based clustering algorithms." 2017 3rd International Conference on Electrical Information and Communication Technology (EICT). IEEE, 2017.
3. Bhade, Kalyani, et al. "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT*)*. IEEE, 2018.
4. Santana, Clodomir J., Pedro Aguiar, and Carmelo JA Bastos-Filho. "Customer Segmentation in a Travel Agency Dataset using Clustering Algorithms." 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE, 2018.
5. Maulina, NadhiraRiska, IstiSurjandari, and Annisa Marlin Masbar Rus. "Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering." 2019 16th International Conference on Service Systems and Service Management (ICSSSM). IEEE, 2019.
6. M. Pakyürek, M. S. Sezgin, S. Kestepe, B. Bora, R. Düzağaç and O. T. Yıldız, "Customer clustering using RFM analysis," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.
7. Sheshasaayee and L. Logeshwari, "Implementation of Clustering Technique Based RFM Analysis for Customer Behaviour in Online Transactions," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, 2018, pp. 1166-1170.
8. Sheshasaayee and L. Logeshwari, "IMPLEMENTATION OF RFM ANALYSIS USING SUPPORT VECTOR MACHINE MODEL," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, Palladam, India, 2018, pp. 760-763.
9. Tavakoli, M. Molavi, V. Masoumi, M. Mobini, S. Etemad and R. Rahmani, "Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study," 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), Xi'an, 2018, pp. 119-126.
10. S. H. Shihab, S. Afroge and S. Z. Mishu, "RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-4.
11. M. R. Maulani, S. F. Pane, R. M. Awangga, D. A. Wijayanti and W. Caesarendra, "An Analysis of Customer Agrotourism Resort Behaviour based on RFM and Mean Shift Clustering," 2018 International Conference on Applied Engineering (ICAE), Batam, 2018, pp. 1-5.
12. Mehar, Arshad Muhammad, Kenan Matawie, and Anthony Maeder. "Determining an optimal value of K in K-means clustering." 2013 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2013.
13. Bholowalia, Purnima, and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN." International Journal of Computer Applications 105.9 (2014).
14. Matsui, Tomohiro, et al. "Cluster validation in k-Means clustering based on PCA-guided k-Means and procrustean transformation of PC scores." 2009 IEEE International Conference on Fuzzy Systems. IEEE, 2009.
15. Atabay, Habibollah Agh, Mohammad Javad Sheikhzadeh, and Mehdi Torshizi. "A clustering algorithm based on integration of K-Means and PSO." 2016 1st Conference on Swarm Intelligence and Evolutionary Computation *(CSIEC)*. IEEE, 2016.
16. Qiao, Jia, and Yong Zhang. "Study on K-means method based on Data-Mining." 2015 Chinese Automation Congress *(CAC)*. IEEE, 2015.
17. Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." *J—*Multidisciplinary Scientific Journal 2.2 (2019): 226-235.
18. Dzulhaq, Muhammad Iqbal, et al. "Customer Segmentation Based on RFM Value Using K-Means Algorithm." *2019* Fourth International Conference on Informatics and Computing *(ICIC)*. IEEE, 2019.
19. Navada, Arundhati, et al. "Overview of use of decision tree algorithms in machine learning." 2011 IEEE control and system graduate research colloquium. IEEE, 2011.
20. Zhong, Yurong. "The analysis of cases based on decision tree." 2016 7th IEEE International Conference on Software Engineering and Service Science *(ICSESS)*. IEEE, 2016.
21. Zhang, He, and Runjing Zhou. "The analysis and optimization of decision tree based on ID3 algorithm." 2017 9th International Conference on Modelling, Identification and Control *(ICMIC)*. IEEE, 2017.
22. Leiva, Rafael García, et al. "A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design." IEEE Access 7 (2019): 99978-99987
23. Tirenni, Giuliano, Christian Kaiser, and Andreas Herrmann. "Applying decision trees for value-based customer relations management: Predicting airline customers' future values." Journal of Database
24. Sarma, V. Mani, and Ato Yoseph Abate. "A New Decision Tree Approach to Image Data Mining and Segmentation." International Journal of Information Technology & Computer Sciences Perspectives 3.2 (2014): 928.Marketing & Customer Strategy Management 14.2 (2007): 130-142.