Customer Segmentation using Unsupervised Learning

|  |  |
| --- | --- |
| Akash Nataraj | Darshan Puttanna |
| Department of Computer Science | Department of Computer Science |
| University of North Carolina at Charlotte | University of North Carolina at Charlotte |
| anatara9@uncc.edu | dputtann@uncc.edu |
|  |  |
| Pramukh Nagendra | Supreeth Betigeri |
| Department of Computer Science | Department of Computer Science |
| University of North Carolina at Charlotte | University of North Carolina at Charlotte |
| pnagendr@uncc.edu | sbetige1@uncc.edu |

*Abstract*— Modern-era zeitgeist is creativity, where everyone is engaged in competition to be better than others. The current world is run on the potential of the organizations to captivate the customers with their products, but with vast number of products for the customers to choose from, leave the customer confounded. Also, companies are clueless about which section of customers to target to sell their products. This is where machine learning comes into play, various algorithms are applied for unravelling the hidden patterns in the data for better decision making for the future. This elude concept of which segment to target is made unequivocal by applying segmentation. The process of segmenting the customers with similar behaviors into the same segment and with different patterns into different segments is called customer segmentation. In this paper, we will use K means clustering, an unsupervised learning technique to segment the customers.

# **Introduction**

As more and more business being coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition has been cut to throat. Change or die have become the simple rule of marketing in today’s world. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer, this is where Data mining serves a very important role to unravel hidden patterns stored in the company’s database. Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customizing the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company. Customer segmentation allows companies to visualize what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them [1].

Clustering has been proven effective to implement customer segmentation. Clustering comes under unsupervised learning, having ability to find clusters over unlabeled dataset. There are several clustering algorithms over which like k-means, hierarchical clustering, DBSCAN clustering and so on [1]. In this paper, K means clustering has been implemented over a dataset with eight attributes with around 542,000 records.

# **Unsupervised Learning**

Consider a machine (or living organism) which receives some sequence of inputs x1, x2, x3…. where xt is the sensory input at time t. This input, which we will often call the data, could correspond to an image on the retina, the pixels in a camera, or a sound waveform. It could also correspond to less obviously sensory data, for example the words in a news story, or the list of items in a supermarket shopping basket [2].

One can distinguish between four different kinds of machine learning. In supervised learning the machine1 is also given a sequence of desired outputs y1, y2…. and the goal of the machine is to learn to produce the correct output given a new input. This output could be a class label (in classification) or a real number (in regression) [2].

In reinforcement learning the machine interacts with its environment by producing actions a1, a2…. These actions affect the state of the environment, which in turn results in the machine receiving some scalar rewards (or punishments) r1, r2…. The goal of the machine is to learn to act in a way that maximizes the future rewards it receives (or minimizes the punishments) over its lifetime. Reinforcement learning is closely related to the fields of decision theory (in statistics and management science), and control theory (in engineering). The fundamental problems studied in these fields are often formally equivalent, and the solutions are the same, although different aspects of problem and solution are usually emphasized. A third kind of machine learning is closely related to game theory and generalizes reinforcement learning. Here again the machine gets inputs, produces actions, and receives rewards. However, the environment the machine interacts with is not some static world, but rather it can contain other machines which can also sense, act, receive rewards, and learn. Thus, the goal of the machine is to act to maximize rewards considering the other machines’ current and future actions. Although there is a great deal of work in game theory for simple systems, the dynamic case with multiple adapting machines remains an active and challenging area of research [2].

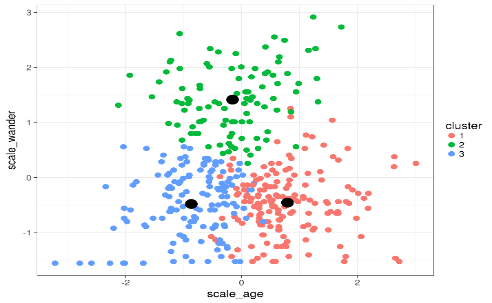
Finally, in unsupervised learning the machine simply receives inputs x1, x2…., but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it does not get any feedback from its environment. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine’s goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise [2].

# **K-Means Algorithm**

K-means cluster algorithm was proposed by J. B. MacQueen in 1967. It is based on decomposition, using K as a parameter, divide n object into K relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. The measurement of the similarity for the algorithm selection is by the reciprocal of the Euclidean distance [3]*.*

Procedure [3]:

* Distribute all objects to K number of different clusters at random.
* Calculate the mean value of each cluster and use this mean value to represent the cluster.
* Re-distribute the objects to the closest cluster according to its distance to the cluster center.
* Update the mean value of the cluster. Calculate the mean value of the objects in each cluster.
* Calculate the criterion function E, until the criterion function converges.



Example figure K-Means Clustering Source: https://rpubs.com/cyobero/k-means

# **Related Works**

Ying Li and Feng Lin, in the paper "*Customer segmentation analysis based on SOM clustering* [4]" utilizes a two-stage integration of SOM network and K-means algorithm to carry out segmentation analysis for security clients. After demonstration, this kind of two-stage method can produce better performance than the integration of factor analysis and K-means method. Availing of above results, the customers are divided into groups according to customer value and behavior, and the detailed features of each customer group are described. Thus, the company can be familiar with their clients and is able to forecast the purchasing behavior of their clients in the near future [4].

T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, in the paper "*Customer Segmentation using K-means Clustering* [1]", 3 different clustering algorithms (k-Means, Agglomerative, and Mean shift) have been implemented to segment the customers and finally compare the results of clusters obtained from the algorithms. A python program has been developed and the program is been trained by applying standard scaler onto a dataset having two features of 200 training sample taken from local retail shop. Both the features are the mean of the amount of shopping by customers and average of the customer's visit into the shop annually. By applying clustering, 5 segments of cluster have been formed labelled as Careless, Careful, Standard, Target, and Sensible customers. However, two new clusters emerged on applying mean shift clustering labelled as High buyers and frequent visitors and High buyers and occasional visitors [1].

G. Lefait and T. Kechadi, in the paper "*Customer Segmentation Architecture Based on Clustering Techniques* [5]" present a data mining architecture based on clustering techniques to help experts to segment customer based on their purchase behaviors. In this architecture, diverse segmentation models are automatically generated and evaluated with multiple quality measures. Some of these models were selected for given quality scores. Finally, the segments are compared. This paper presents experimental results on a real-world data set of 10000 customers over 60 weeks for 6 products. These experiments show that the models identified are useful and that the exploration of these models to discover interesting trends is facilitated using the architecture implemented [5].

Z. Lu, W. Peiyi, C. Ping, L. Xianglong, Z. Baoqun and M. Longfei, in the paper "*Customer Segmentation Algorithm Based on Data Mining for Electric Vehicles* [6]" discusses the significance of Customer segmentation for charge service operators to obtain the features of charging behaviors and individual differences between various Electric Vehicle (EV) users. Based on large numbers of charging service data fast accumulated by Operations Management System, exploratory data analysis was applied to all the historical data in the database. Firstly, the paper screened out key variables to the segmentation model, and then presented the EV customer segmentation method by data mining technique and k-means algorithm. Secondly the customer value evaluation method was proposed and charging behavioral features and customer values were analyzed. At last, conclusions and suggestions are given, which would provide data supports to the improvement of operation and maintenance management and decision-making of the precision marketing [6].

X. Qin, S. Zheng, Y. Huang and G. Deng, in the paper "*Improved K-Means Algorithm and Application in Customer Segmentation* [7]" uses an improved K-means algorithm for segmentation of customers. K-means algorithm is by far the most commonly used method for clustering. Although, the time consumption is fairly high when faced with larger-scale data. The improvement of the K-means algorithm is based on the triangle inequality theorem. The improved algorithm is used to carry out a case study in the customer classification. The experimental results show that the improved method indeed lead to lower time consumption, and therefore more effective for large-scale dataset [7].

X. He and C. Li, in the article "*The Research and Application of Customer Segmentation on E-Commerce Website* [8]" constructs a three dimensional customer segmentation model based on customer lifetime value, customer satisfaction and customer activity, which more accurately divides customers into different groups. The corresponding variables are obtained by RFM model, Kano model and BG/NBD model. The customer segmentation model provides ten groups of customers with corresponding marketing strategies, so that it can help enterprises maximize profits [8].

# **Implementation**

## RFM Analysis for Customer Segmentation

Customer segmentation is a process of dividing the customers into several segments or categories in a way such that they are very similar to one another relevant to marketing, interests, and spending habits. Customer loyalty and contributions are accurately measured by RFM analysis. Customer segmentation or clustering based techniques are very important to calculate RFM analysis. Various groups of customers divided according to customer segmentation technique, to identify which customers are most likely to respond to promotion.[9]

### Recency: It is about when was the last order of a customer. It means the number of days since a customer made the last purchase. If it’s a case for a website or an app, this could be interpreted as the last visit day or the last login time.

### Frequency: It is about the number of purchase in a given period. It could be 3 months, 6 months or 1 year. So we can understand this value as for how often or how many times a customer used the product of a company. The bigger the value is, the more engaged the customers are. Could we say them as our VIP? Not necessary. Cause we also have to think about how much they actually paid for each purchase, which means monetary value.

### Monetary: It is the total amount of money a customer spent in that given period. Therefore big spenders will be differentiated with other customers such as MVP or VIP.

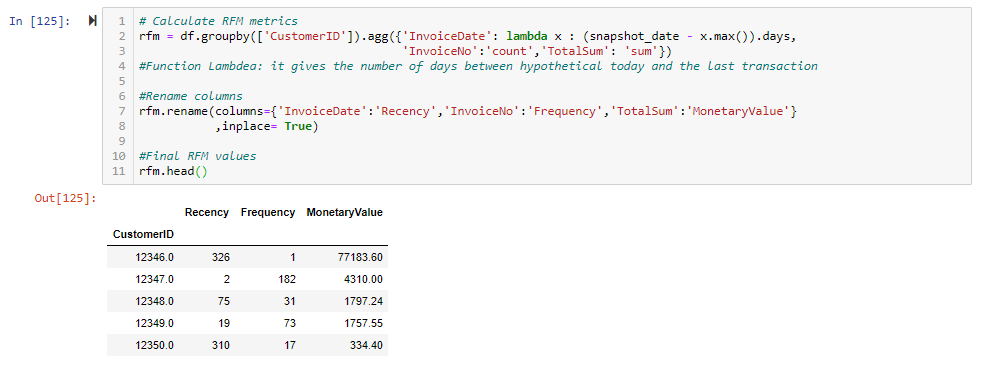
Once we have calculated these numbers, next step is to categorize them into some sort of categorization such as high, medium, or low.

We can break the customers into groups of equal size based on percentile value of each metric i.e. (RFM). Process of calculating percentiles is fairly simple:

# Sort customers based on that metric

# Break customers into a pre-defined number of groups of equal size.

# Assign a label to each group

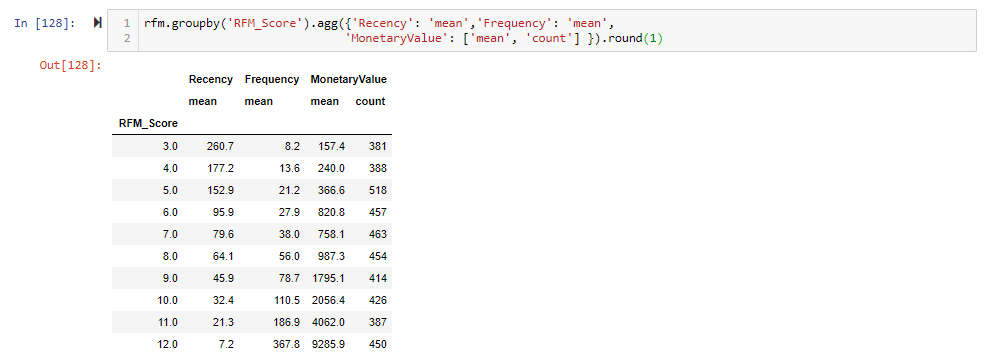


Example figure: Customers Transaction(sample dataset)

An analysis RFM measured some sample dataset based on customer transactions. The customer’s data are segmented based on recency, then frequency and the last monetary values.

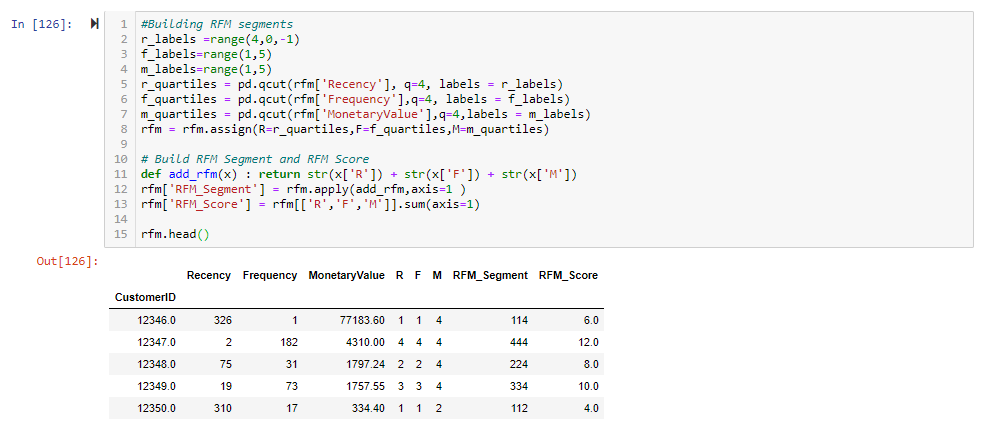
First step starts with customer’s data and sorted based on recency (last transaction customers), the next the data are ordered low to high, then divided into five equal groups. The first 20% a recency top score 5, next 20% score 4 and so on. The second step the frequency (number of transactions) customers are sorted most to least, and frequency 20% top score started 5 and then next scores 4,3,2,1. The final step process undertaken for monetary values. [10]

Finally, the RFM values measured for some customer transactions (sample dataset).



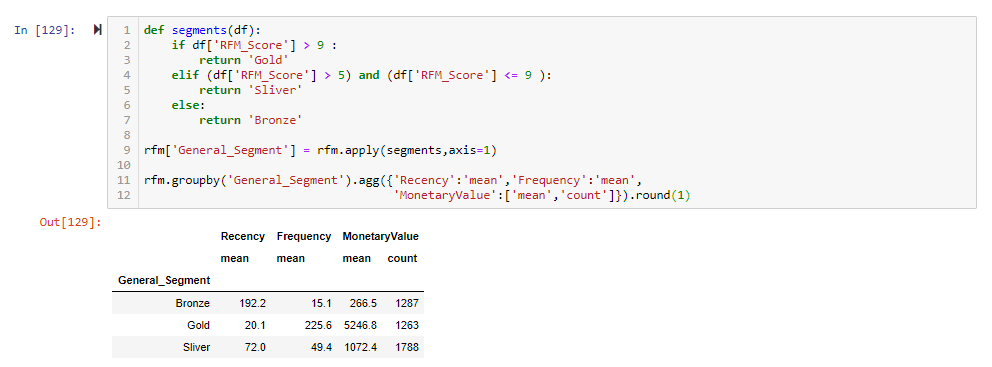
Example figure: Finding mean RFM values (sample dataset) and assigning scores

The values of recency, frequency and monetary are combined to form RFM scores. For example, in a five-category ranking system, there are about 125 possible RFM scores and the highest RFM score is 555. RFM scores clearly shows the categories of different consumers. The best customers are chosen with the highest RFM scores. In this paper, the ranking 1-5 is used to evaluate the customer retention. [10]



Example figure: Assigning RFM scores and calculating total RFM values

Based on the total RFM Scores customers are categorized into Gold, Silver and Bronze.



Example figure: Customer Categorization

If the total RFM score of a customer is greater than 9, then that customer is categorized as Gold. If RFM value is in-between 5 to 9 categorized as Silver. If value is less than 5 then categorized as bronze.

## Implementation of k-means

We implement k-means algorithm in 4 key steps:

* Data pre-processing
* Choosing the number of clusters
* Running k-means clustering on pre-processed data
* Analyzing RFM values of each other

**Data Pre-processing**

The data we have used as our input will need labels to be created before we proceed with the implementation part. There are 3 labels created they are

1. Invoice Period – It is a string representation of the year and month of a single transaction/invoice.
2. Cohort Group - A string representation of the year and month of a customer’s first purchase. This label is common across all invoices for a particular customer.
3. Cohort Period/Cohort Index - An integer representation of a customer’s stage in the “lifetime”. The number represents the number of months since the first purchase.

After processing the RFM values we proceed to check with the mean and standard distribution of each value in the data. Upon calculation it is found that the values of mean and variance are unequally distributed. Since, the data values must not be skewed, skewness is removed by applying logarithmic transformations. The data values are standardized to some average values and scaled to the same standard deviation so that it is fitted into a common scale. This is done using the scikit-learning library. We import scikit-learning library and the data values are scaled using Standard Scalar function applied on the output of the logarithmically transformed data. This data preprocessing is required so that the input values exits as a symmetric data and it gets easy for comparison.

**Choosing the number of clusters**

There are several general methods to define the number of clusters, they are:

1. *Visual method - Elbow Criterion*
2. Mathematical methods - Silhouette coefficient
3. Experimentation and interpretation.

The Elbow criterion is a simple but yet effective method to determine the number clusters. Whereas the Silhouette coefficient needs additional information such as intra-cluster distance and nearest cluster distance, the Elbow Criterion is best suited since the data we have is most suited. Elbow criterion is decided by plotting a number of clusters against within sum of squared errors (SSE). SSE is basically the sum of squared distances from every data point to their cluster center. In the line graph plotted, the point where the line appears to be an elbow represents the optimal number of clusters. The optimal value of k thus obtained from this analysis is equal to three.

**Running k-means clustering on pre-processed data**

The k-means algorithm is run over the normalized data in this stage. Cluster Labels are assigned to each of the RFM values. After clustering since there are three clusters, average RFM values for each cluster is computed and their size is calculated, accordingly.

**Analyzing RFM values of each other**

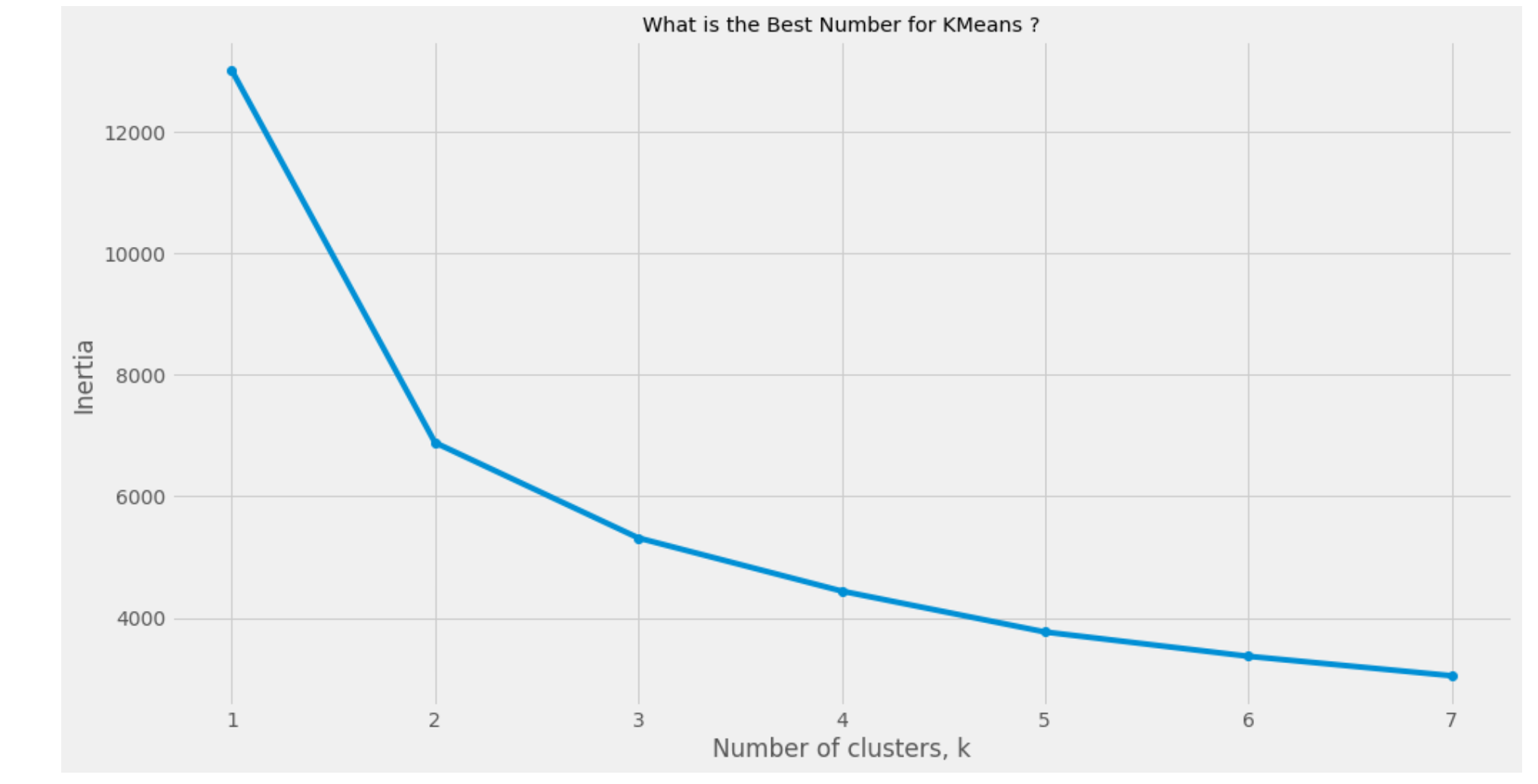
To better visualize the data, snake plots of the normalized data and cluster’s average normalized values of each attributes is calculated. To analyze each customer, general segmentation is done which basically groups each customer using the customer ID into three categories: gold, silver, and bronze. These scores are assigned based on the RFM score calculated. Those customers with RFM scores greater than 9 are given gold, those between 5 to 9 are given silver and those below 5 are assigned bronze. Using the snake plot the data is melted into one single metric to better analyze each customer based on their RFM scores.

After calculating the average values of each cluster and average of population of dataset we can calculate the RFM score by dividing them and subtracting 1. The further this calculated value is from 0, the more important that attribute is for a segment relative to the total population. After the values are calculated based on the number of clusters and segmentation of each customer, this data is used to plot a heat map. This heat map gives the information about which customer with respective RFM values are clustered in which segment and how each customers are categorized based their activities and interaction with the system/product.

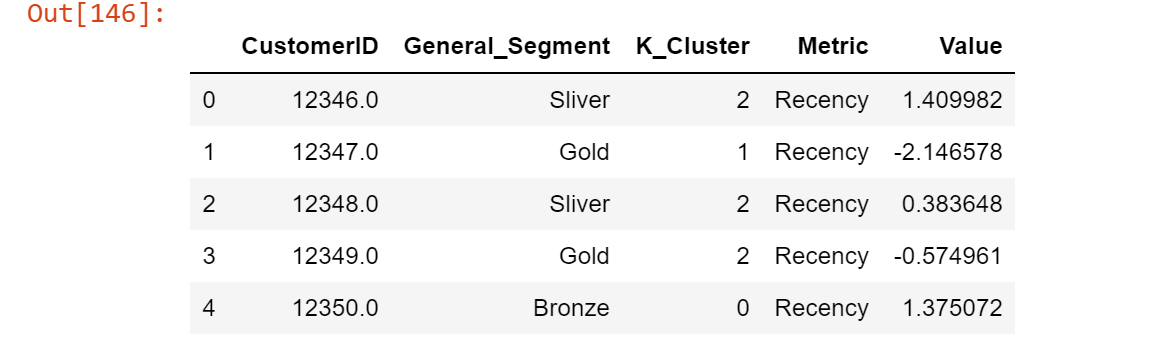
# **Results**

Post implementing k-Means clustering method on a dataset, a clear understanding of the cohort behavior with respect to time, loyalty, and lifetime of a product can be observed.

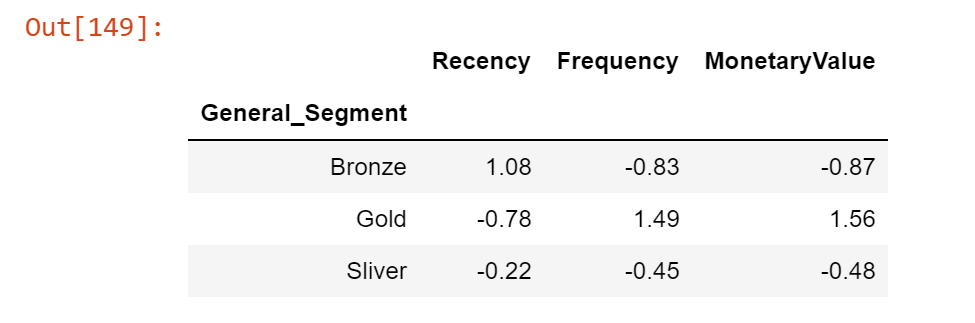
* Post data-preprocessing, the number of clusters is chosen. Here, the number of clusters chosen is 3 using the elbow criterion, as seen in the figure below.



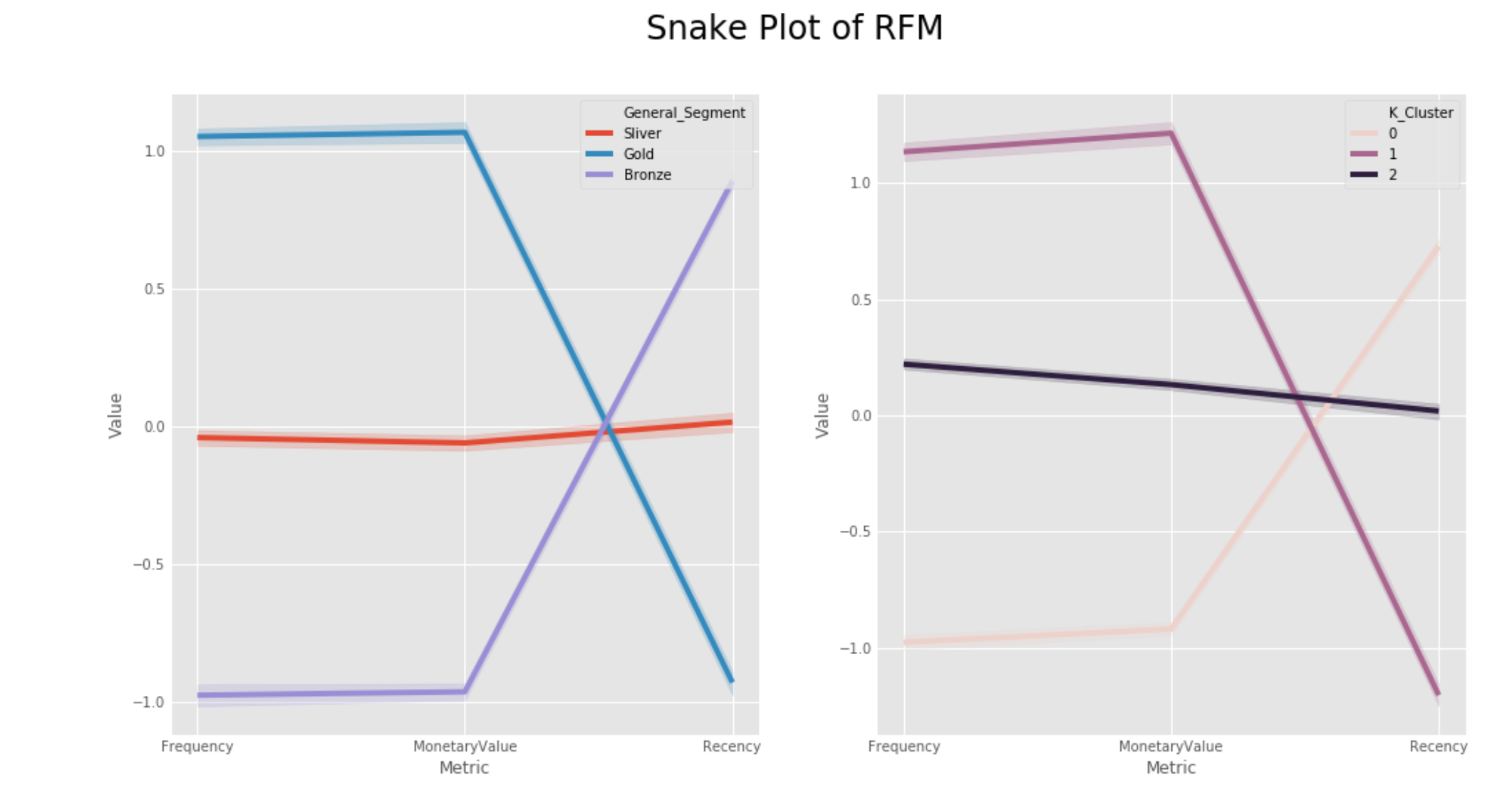
* Implementation of k-Means clustering on the data helps us segment the customers in clusters and assign them an RFM score and value based on which they will be segmented.



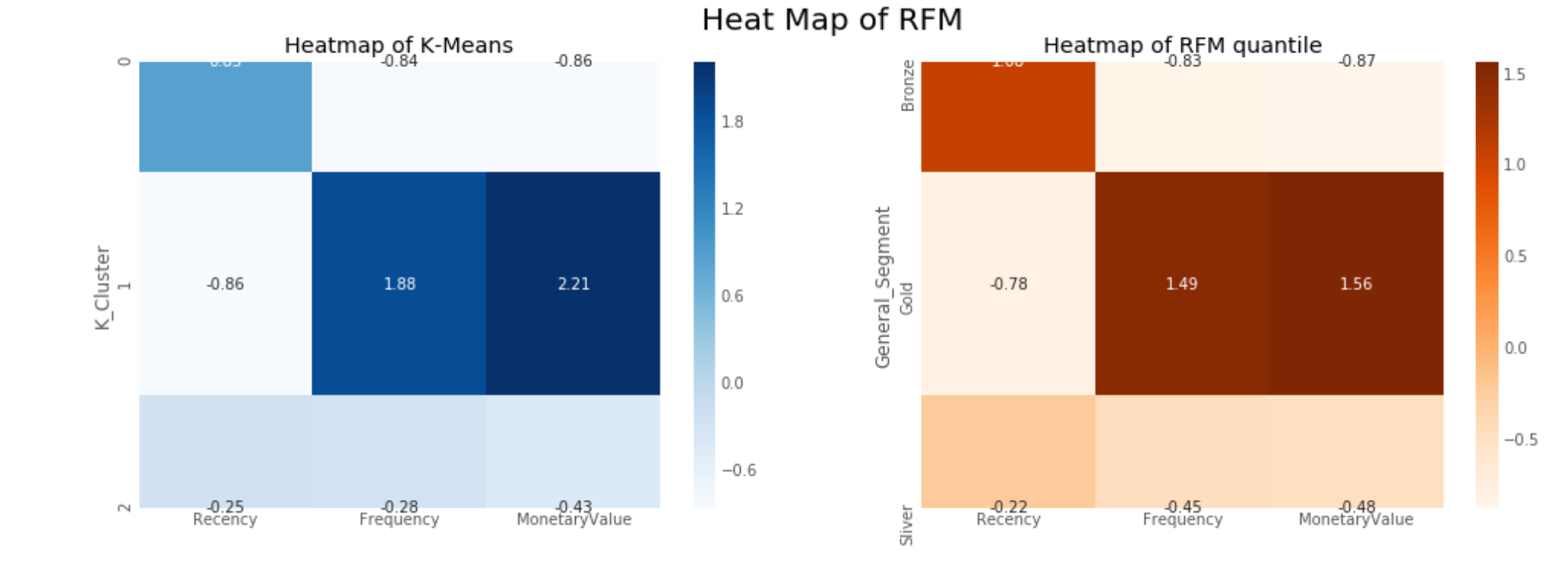
* The segments are created based on the Recency, Frequency, and Monetary Value according to the respective Gold, Silver or Bronze ranking/scoring.



* Once the customers are clustered and segmented according to their respective metrics, we can visualize the trend of cohorts and their behavioral aspects towards to the lifetime of a product and loyalty.



* To visualize the trend, we segment them using ratios relative segment value and total population within the cluster.
* Once the ratios are created, we plot a heat map as an easier visual representation of the trends in cohorts.
* The darker regions represent the higher affinity towards a product or organization during its lifetime, while the lighter regions represent otherwise.



# **Conclusion**

The main purpose of this research was to categorize the customers from the transaction data of 542,000 based on RFM model, and furthermore clustering analysis was performed using K-Means algorithm.

With the clustering and segmentation of customers using RFM, and RFM metrics, we are able to find behavioral aspects of the customers towards a product, and the loyalty towards the product, along the lines of the lifetime of a product.

##### **References**

1. T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 135-139.
2. Ghahramani Z. (2004) Unsupervised Learning. In: Bousquet O., von Luxburg U., Rätsch G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg
3. Chunfei Zhang and Zhiyi Fang, “An Improved K-means Clustering Algorithm”, Journal of Information & Computational Science 10: 1 (2013) 193–199.
4. Ying Li and Feng Lin, "Customer segmentation analysis based on SOM clustering," 2008 IEEE International Conference on Service Operations and Logistics, and Informatics, Beijing, 2008, pp. 15-19.
5. G. Lefait and T. Kechadi, "Customer Segmentation Architecture Based on Clustering Techniques," 2010 Fourth International Conference on Digital Society, St. Maarten, 2010, pp. 243-248.
6. Z. Lu, W. Peiyi, C. Ping, L. Xianglong, Z. Baoqun and M. Longfei, "Customer Segmentation Algorithm Based on Data Mining for Electric Vehicles," 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2019, pp. 77-83.
7. Sadegh Bafandeh Imandoust And Mohammad Bolandraftar, X. Qin, S. Zheng, Y. Huang and G. Deng, "Improved K-Means Algorithm and Application in Customer Segmentation," 2010 Asia-Pacific Conference on Wearable Computing Systems, Shenzhen, 2010, pp. 224-227.
8. X. He and C. Li, "The Research and Application of Customer Segmentation on E-Commerce Websites," 2016 6th International Conference on Digital Home (ICDH), Guangzhou, 2016, pp. 203-208.
9. A. 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.
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.