## Customer Segmentation using RFM Modelling and K-Means Clustering

A project report submitted for the partial fulfilment of the

**Bachelor of Technology Degree**

in

**Computer Science & Engineering**

under

**Maulana Abul Kalam Azad University of Technology**

**by**

**Sayan Banerjee**

Roll No: 10400116102, Registration Number: 161040110145

**Aritra Chakraborty**

Roll No: 1400116216, Registration Number: 161040110031

**Academic Session: 2016-2020**

Under the Supervision of,

**Prof. Shreejita Mukherjee**



**Department of Computer Science and Engineering**

**Institute of Engineering & Management**

Y-12, Salt Lake, Sector 5, Kolkata, Pin 700091, West Bengal, India

Affiliated To

**Maulana Abul Kalam Azad University of Technology**

BF 142, BF Block, Sector 1, Kolkata, West Bengal 700064

**CERTIFICATE**

**TO WHOM IT MAY CONCERN**

This is to certify that the project report titled, “**Customer Segmentation using RFM Modelling and K-Means Clustering**”, submitted by **Sayan Banerjee, Roll Number: 10400116102, Registration Number: 161040110145, and Aritra Chakraborty, Roll Number: 10400116216, Registration Number: 161040110031**, students of **Institute of Engineering & Management** in partial fulfilment of requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering**, is a bonafide work carried out under the supervision of **Prof. Shreejita Mukherjee** during the final year of the academic session of **2016-2020**. The content of this report has not been submitted to any other university or institute for the award of any other degree.

It is further certified that the work is entirely original and the performance has been found to be satisfactory.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. Shreejita Mukherjee**  **Prof. (Dr.) Sourav Saha**

Assistant Professor Head of the Department

Department of Computer Science and Engineering Department of Computer Science and Engineering

Institute of Engineering & Management Institute of Engineering & Management

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. (Dr.) Amlan Kusum Nayak**

Principal

Institute of Engineering & Management

**INSTITUTE OF ENGINEERING & MANAGEMENT**

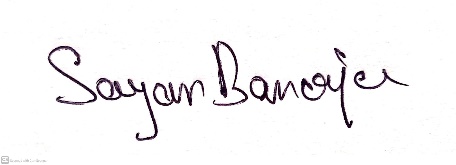


DECLARATION

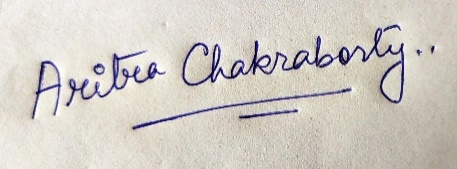
FOR NON-COMMITMENT OF PLAGIARISM

We, **Sayan Banerjee, Aritra Chakraborty** students of B.Tech. in the department of Computer Science and Engineering, Institute of Engineering & Management have submitted the project report in partial fulfilment of the requirements to obtain the above noted degree. We declare that we have not committed plagiarism in any form or violated copyright while writing the report and have acknowledged the source wherever applicable.

Name of the Student: **SAYAN BANERJEE**

Full Signature:

Name of the Student: **ARITRA CHAKRABORTY**



Full Signature:

Date : 08/07/2020

**Abstract**

The objective of this project is to apply business intelligence in identifying potential customers by providing relevant and timely data to business entities in the Retail Industry. The data furnished is based on systematic study and scientific applications in analysing sales history and purchasing behaviour of the consumers. The curated and organized data as an outcome of this scientific study not only enhances business sales and profit, but also equips with intelligent insights in predicting consumer purchasing behaviour and related patterns. In order to execute and apply the scientific approach using K-Means algorithm, the real time transactional and retail dataset are analysed. Spread over a specific duration of business transactions, the dataset values and parameters provide an organized understanding of the customer buying patterns and behaviour across various regions. This study is based on the RFM (Recency, Frequency and Monetary) model and deploys dataset segmentation principles using K-Means Algorithm**.** The results thus obtained with regard to sales transactions are compared with various parameters like Sales Recency, Sales Frequency and Sales Monetary values.

**ACKNOWLEDGEMENT**

We must not forget to acknowledge everyone who has provided constant support to us during our B.Tech course. First and foremost, we would like to express sincere gratitude to our supervisor **Prof. Shreejita Mukherjee** for her continuous support and motivation in fueling the pursuance of carrying out this project endeavour. Without her guidance and persistent encouragement, this project work would not have been possible. She has been a tremendous mentor for us throughout this academic journey. Many of her academic advises about our career growth have been priceless.

We would like to convey sincere gratitude to **Prof. Sourav Saha** for providing us constant inspiration to stand firm against several setbacks throughout the course. Additionally, we would like to thank all the technical, non-technical and office staffs of our department for extending facilitating cooperation wherever required. We also express gratitude to all of our friends in the department for providing the friendly environment to work on the project work.

We would also like to thank our Director **Prof. Satyajit Chakraborti** for providing us an outstanding platform in order to develop our academic career. In addition, we also preserve a very special thankful feeling about our Principal **Prof. Amlan Kusum Nayak** for being a constant source of inspiration.

A special thank is due to our family. Words cannot express how grateful we are to our parents for all the sacrifices that they have made while giving us necessary strength to stand on our own feet.

Finally, we would like to thank everybody who has provided assistance, in whatever little form, towards successful realization of this project but with an apology that we could not mention everybody’s name individually.

**Content**

**List of Figures** 6

**List of Tables** 7

**1. Introduction** 8

1.1 Motivation .…………………………………………………………………………………….8

1.2 Literature Survey ……………………………………………………………………………9

1.3 Objective of the Proposed Work ……………………………………………………11

**2. Proposed Methods and Models** 12

2.1 Overview ……………………………………………………………………………………….12

2.2 Selection of Clustering Algorithm ………………………………………………….12

2.2.1 Proposed Methodology ………………………………………………………………12

2.2.2 Mathematical Model ………………………………………………………………….15

2.2.3 Attributes in Dataset ………………………………………………………………….16

**3. Results and Observations** 17

3.1. Dataset Description……………………………………………………………………….17

3.2. RFM Modelling………………………………………………………………………………18

3.3. K-Means Clustering……………………………………………………………………….20

3.3.1. Data Normalization………………………………………………………….20

3.3.2. Elbow Method………………………………………………………………….21

**4. Conclusion and Scope of Future Work**………………………………………………………23

**Bibliography** 24

## List of Figures

## 2.1 Flow chart for different steps in methodology…………………………14

## 3.1 Recency vs Frequency scatter plot……………………………………………18 3.2 Frequency vs Monetary scatter plot…………………………………………19 3.3 Recency vs Monetary scatter plot…………………………………………….19 3.4 Recency distribution plot before and after normalization…………20 3.5 Frequency distribution plot before and after normalization……..20 3.6 Monetary distribution plot before and after normalization………21 3.7 Elbow Curve……………………………………………………………………………..21

## List of Tables

## 2.1 Attributes in the dataset ………………………………………………………….16

## 3.1 First 10 rows in dataset ……………………………………………………………17 3.2 RFM values of first 10 customers in dataset ……………………………18 3.3 Customer’s RFM dataset after K-Means clustering ………………….22

## 1

## Introduction

* 1. **Motivation**

In the ever-growing competition and increasing complexity of business environment, segmentation and its systematic study improves customer loyalty and enhances enterprise-level for long lasting relationship by widening profitable customer database. The two most prominent types of segmentations used in K-Means Algorithm are the Qualitative and Quantitative insights. In the scope of the current study, Quantitative insight is used for the purpose of segmentation clustering Well-defined customer segmentation helps in effective allocation of marketing resources, enables the companies to target the specific group of customers and also helps in building healthy long-term relationship with the customers. The major industries wherein customer segmentation and for data mining can be applied the Retail Industry because it requires a vast amount of data on sales, transportation, consumption ratio, redelivery service and many others. Also, Retail data mining helps in identifying and effectively mapping customer behaviour and related patterns during the entire life-cycle of business transactions. This ultimately, leads to improved customer service, effective sales and distribution strategies and many more. This work mainly focuses on tracking the historical purchasing behaviour of customers with the aim to find maximum amount of sale possible in the specific area. Based on the statistical results and indicators, companies in the retail industry can design various sales and marketing strategies like promotional campaigns, extending seasonal discounts or floating sales enabling coupons to increase the sales and improve customer retention.

To achieve the above objectives, customer clustering and segmentation is carried out using the K-Means algorithm. It is based on RFM values for different regions. RFM can be defined as segmentation of customer analysis which not only gives information on frequent purchasing pattern of the customer, but also recent purchase and the profit obtained.

## Literature Survey

## Group-specific marketing is common and much needed from the traditional mass marketing perspective. Customer segmentation is a part of various activities under Customer Relationship Management(CRM) value chain ([Kolarovszki et al., 2016](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0080), [Khalili-Damghani et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0075), [Khajvand et al., 2011](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0070)). The author has proposed a novel modelling in the field of postal area using the multidimensional segmentation. This design of CRM is useful in companies of postal services ([Kolarovszki et al., 2016](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0080)).

## The proposed hybrid method is to predict the new entities for the customer centric companies. This study is based on the methods like decision tree approach, clustering, rule extraction and many more. The author has used K-Means for predicting future transactions based on the historical behaviour of the customers under various segmentation. To facilitate this, the hybrid feature selection for filtering making and decision making method is used. The current study can also be applied for Telecommunication and Insurance industries to predict the sales volume and projecting business profit of a company. This method is very effective not only in predicting profitable customers, but also in identifying behaviour of new customer. ([Khalili-Damghani et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802#b0075)).

The concept of customer segmentation can be applied for beauty and healthcare companies, which indirectly leads to CRM. The author used two approaches for segmentation, where one is based on RFM and second one is extended RFM by addition of count item parameter. Sales and marketing strategies are explained by calculating Customer Lifetime value using weighted RFM ([Khajvand et al., 2011](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0070)).

Multiple analysis that is based on integration of CRM and RFM model is essential for exploring CRM in large scale data ([Song et al., 2017](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0115)). RFM model is employed to predict the supply quantity per month by clustering the customers using K-Means algorithm. Each group is distinguished using CHAID decision trees based on attribute values ([You et al., 2015](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0125)).

The relationship between consumer behaviour and order fulfilment in the field of marketing and operations is identified using various marketing tools, which enhances the consumer service levels ([Nguyen et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0090)). Clustering technique is used to group the retailers using RFM model based on Electronic Funds Transfer at Point of Sale (EFTPOS) in businesses ([Singh et al., 2014](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0110)).

Data analytics approach is proposed for customer segmentation based on the customer visit to the store, collected from the overall sales data. Also, feature selection approach is proposed, which takes product taxonomy as input and categories of customers as output ([Griva et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0045), [Hu and Yeh, 2014](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0060)).

In the current study, the RFM analysis is executed using transactional dataset for evaluating customers on their purchase behaviour and analysing the same using unsupervised algorithms like K-Means and Fuzzy C – Means. Also, the author has introduced a novel idea of selection of centroids in K-Means Algorithm and comparing the results ([Christy et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0030)).

Customer segmentation requires descriptive variables for identifying behavioural patterns. However, in some domains, descriptive variables are not adequate. Considering this, the author has proposed the segmentation method to solve the problem leading to better performance using data mining methods ([Murray et al., 2017](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0085)).

K-Means and self-organizing SOM algorithms are used to cluster the customer characteristics using RFM model for insurance dataset. It helps in identifying the customer needs, their understanding and characteristics ([Qadadeh and Abdallah, 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0100)). The study suggests that fuzzy and SOM based clustering methods are more efficient compared to the traditional methods ([Arunachalam and Kumar, 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0005)). Sequential exploratory design is adopted which combines both qualitative and quantitative methods ([Arunachalam and Kumar, 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802#b0005)).

Customer segmentation can also be applied for e-pharmacy clients to increase the retention of e-customers and it is one of the prerequisite for the essential CRM in the area of customer loyalty ([Patak et al., 2014](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0095)). Due to high volatility of the market, customer segments change over a period of time. The author has introduced the concept of ‘Stream Clustering’ as a tool and has proposed a new algorithm to overcome this problem. An important aspect in Stream Clustering is to identify the new clusters or emerging clusters and replacing the older ones ([Carnein and Trautmann, 2019](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0020)).

Swarm based algorithms like Flower Pollination Algorithm, Black Hole Algorithm, Bat Algorithms and others are proposed to overcome the problem of slow convergence for larger datasets. Also Comparative analysis of algorithms is carried out using four performance parameters ([Kaur et al., 2019](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0065)). A framework is designed, where e-commerce research is categorized or segmented into three phases. In each phase, the author has gather the issues reported by the practitioners and has suggested solutions to overcome them. Conceptual framework consists of service relationships, business models and technologies ([Yoo and Jang, 2019](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0120)).

Data mining technique called Clustering Approach can also be used to address various road-blocks in the manufacturing and marketing problems in fashion industry. Needless to say, segmentation is very important for finding the patterns of customer preferences ([Brito et al., 2015](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0015)).

Business Intelligence refers to the intelligent technologies that help in improving the business performance. Sometimes, this concept difficult to apply for small size companies, due to high cost, limited availability of resources and many other factors. In this view, the author has introduced not only he Business Intelligence role for major companies, but also how implement the concept for small size enterprises to increase their profitability and productivity ([D’Arconte, 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0035)).

To retain the online-gaming customers, the author has proposed an innovative model of segmentation. Various features such as performance, engagement and social interactions are considered to segment the players ([Fu et al., 2017](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0040)). Data driven approach is used to cluster the retail products based for the market basket data. Results are compared using k-means, SOM and hierarchical clustering approaches ([Holý et al., 2017](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0055)).

Impact of Big data on CRM is reviewed based on the critical success factors. Results shows three contributions like past reviews, five propositions and previous contributions ([Zerbino et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0130)). Naïve Bayes and neural networks classification approach is used to design a data mining CRM framework to enhance the decision making for retaining the customers ([Bahari and Sudheep Elayidom, 2015](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0010)). Streams of research in multiple fields like marketing, operations, information management and other areas are linked to develop an integrated framework ([Sheng et al., 2017](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0105)).

* 1. **Objective of the Proposed Work**

The primary objective of the proposed work is to divide customers of a company into groups on the basis of different factors so that the company can run individual marketing schemes on targeted customers. In the light of data segmentation, customers are divided into set of individuals with distinct similarities. Some of the attributes relevant to customer segmentation are gender, age, lifestyle, location, purchase and income behaviour. Such attributes are mainly categorized based on the historical purchasing behaviour that can lead to a specific outcome, for example, an increase in sales and the profit for the company.

**2**

**Proposed Methods and Models**

**2.1 Overview**

This section discusses our proposed methods and models in details. The proposed model divides the customers in the dataset into groups or clusters on the basis of their RFM scores. After that the dataset is fed into a clustering model that will be suitable for our project.

## 2.2. Selection of Clustering Algorithm

A cluster is understood as a conceptually meaningful group of objects that have common characteristics. Clustering can be used for customer segmentation for additional analysis. K-Means clustering algorithm is a prototype-based partition clustering technique that finds the user specified number of clusters, which are represented by their centroids. K-Means is computationally faster and performs well on large datasets compared to other clustering methods. Another advantage of using K-Means is that the algorithm requires only one input parameter ‘K‘, unlike other algorithms. Also, it decreases the rate of misclassification of data. One of the major applications of K-Means is customer segmentation. The present work uses K-Means algorithm.

## 2.2. Proposed Methodology

The proposed methodology can be broadly divided into the following steps as shown in the fig 1.1. The corresponding details are explained as below:

**Step 1:** **Exploratory Analysis and Data Pre-processing**Exploratory data analysis (EDA) refers to initial exploration of data in order to extract or discover the patterns with the help of statistics or graphical representations. In this activity, EDA helps in identifying unique customers, percentage of orders by top 10 or more, information about the data, mismatch in description, stock code and to check null values. Further, data pre-processing is applied to identify and remove missing customer identification number, negative transactions and so on.  
We have also added a new feature depicting total amount by multiplying unit price with quantity in order to find the monetary value for each customer.

**Step 2: Execution of RFM Analysis**  
After data is pre-processed, check for recent transactions, frequency and the amount spent by the customers. In order to create recency variable, decide the reference date - that is one day prior to the last transaction. RFM analysis is a very popular customer segmentation and identifiable technique in database marketing ([Christy et al., 2018](https://www.sciencedirect.com/science/article/pii/S1319157819309802#b0030)). It is significant especially in Retail Industry. Each customer under RFM is scored based on three factors.

* **Recency**: It refers to the number of days before the reference date when a customer made the last purchase. Lesser the value of recency, higher is the customer visit to a store.
* **Frequency**: It is the period between two subsequent purchases of a customer. Higher the value of Frequency, more is the customer visit to the company.
* **Monetary**: This refers to the amount of money spent by a customer during a specific period of time. Higher the value, more is the profit generated to the company.

**Step 3: Normalization of the Recency, Frequency and Monetary curve**

**Before feeding data to our model it is necessary to check whether the data is normally distributed in order to achieve maximum accuracy. If the distribution plot is skewed (not normal), we can try to remove the skewness by applying either square-root, cube or logarithmic functions to our features. In our case we have used logarithmic function since it gave us the best result.**

**Step 4: Evaluation of number of clusters using Elbow Method**

In cluster analysis, the elbow method is a heuristic method used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

**Step 5: K-Means clustering model building**Once we get the number of clusters or ‘k’ from the elbow curve, we can put it in our K-means clustering model and find the clusters in our dataset. After the analysis of clusters, we compare the sales recency with sales amount and sales frequency with sales amount from one cluster to another cluster respectively. This helps in identifying the group of customers having highest sales recency, sales frequency and the sales amount.

Fig 2.1 : Flow chart for different steps in methodology

Data Set

Remove null values

EDA and Data Preprocessing

RFM Modelling

Divide the customers into loyalty levels on the basis of their RFM

Calculate Recency, Frequency and Monetary values of each customer

Remove negative transactions

Normalization of Recency, Frequency and Monetary curves

Evaluate number of clusters using elbow method

K-Means clustering model building

## 2.3. Mathematical Model

The K-Means clustering algorithm separates data into the best suited group based on the information the algorithm already has. Data is separated in k*k* different clusters, which are usually chosen to be far enough apart from each other spatially, in Euclidian distance, to be able to produce effective data mining results. Each cluster has a centre, called the **centroid**, and a data point is clustered into a certain cluster based on how close the features are to the centroid.

K-means algorithm iteratively minimizes the distances between every data point and its centroid in order to find the most optimal solution for all the data points.

1. Krandom points of the data set are chosen to be centroids.
2. Distances between every data point and the k centroids are calculated and stored.
3. Based on distance calculates, each point is assigned to the nearest cluster
4. New cluster centroid positions are updated: similar to finding a mean in the point locations
5. If the centroid locations changed, the process repeats from step 2, until the calculated new centre stays the same, which signals that the clusters' members and centroids are now set.

Finding the minimal distances between all the points implies that data points have been separated in order to form the most compact clusters possible, with the least variance within them. In other words, no other iteration could have a lower average distance between the centroids and the data points found within them.

The K-means algorithm defined above aims at minimizing an [objective function](https://brilliant.org/wiki/objective-function/), which in this case is the [squared error](https://brilliant.org/wiki/squared-error/?wiki_title=squared%20error) function.

The objective function for the K-means clustering algorithm is the squared error function:  
 **J = \sum\_{i=1}^{k} \sum\_{j=1}^{n} (||x\_ i - v\_j||)^2 = 1*J*=∑*i*=1*k*​∑*j*=1*n*​(∣∣*xi*​−*vj*​∣∣)2=1**

where,  
||x\_i - v\_j||∣∣*xi*​−*vj*​∣∣ is the Eucledian distance between a point, x\_i*xi*​, and centroid, v\_j*vj*​, iterated over all *k* points in the i^{th}*ith* cluster, for all *n* clusters.

n simpler terms, the objective function attempts to pick centroids that minimize the distance to all points belonging to its respective cluster so that the centroids are more symbolic of the surrounding cluster of data points.

[[5]](https://brilliant.org/wiki/k-means-clustering/#citation-5)

## 2.4. Attributes in the Dataset

Description of synthetic dataset is shown in [Table 2.1](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "t0005). This dataset contains all purchases made for an online retail company based in the United Kingdom during an eight month period.

Table 2.1 : Attributes in the dataset

| **Sl. No** | **Name of the Attribute** | **Type of the Attribute** | **Description of the Attribute** |
| --- | --- | --- | --- |
| 1 | Invoice Number | Nominal | Six-digit number uniquely assigned for each transaction. |
| 2 | StockCode | Nominal | Five-digit unique number assigned to each distinct product. |
| 3 | Description | Nominal | Name of the product |
| 4 | Quantity | Numeric | Quantities of each product per transaction |
| 5 | InvoiceDate | Numeric | Date and time of each transaction generated of x attribute |
| 6 | UnitPrice | Numeric | Product price per unit of |
| 7 | Customer Id | Numeric | Five digit unique number assigned to a customer. |
| 8 | Country | Character | Name of the country |

## 3

## Results and Observations

## 3.1. Dataset Description

## The proposed methodology is implemented on the synthetic dataset of one-year customer transactions obtained UCI repository ([Chen et al., 2012](https://www.sciencedirect.com/science/article/pii/S1319157819309802" \l "b0025)). The dataset consists of 8492 instances of information on customer purchase from 1-12-2010 to 09-12-2011 with eight attributes. Missing values, negative transactions, mismatch in stock code and description are handled using data pre-processing. For the modified dataset, we applied RFM analysis and K-Means clustering.

## Table 3.1 : First 10 rows in the dataset

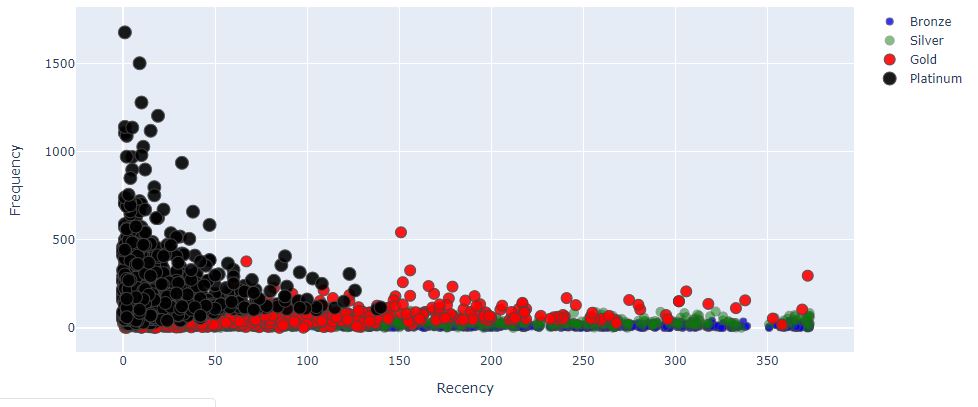
## 

## 3.2 RFM Modelling

## The result of Recency, Frequency and Monetary values are calculated and each customer is assigned to his/her loyalty level. The result is shown in table 3.2 below.

## Table 3.2 : RFM Values of first 10 customers

The visualization of scatter plots for Recency vs Frequency, Frequency vs Monetary and Recency vs Monetary are shown in figures 3.1, 3.2 and 3.3 given below.

Fig 3.1 : Recency vs Frequency scatter plot

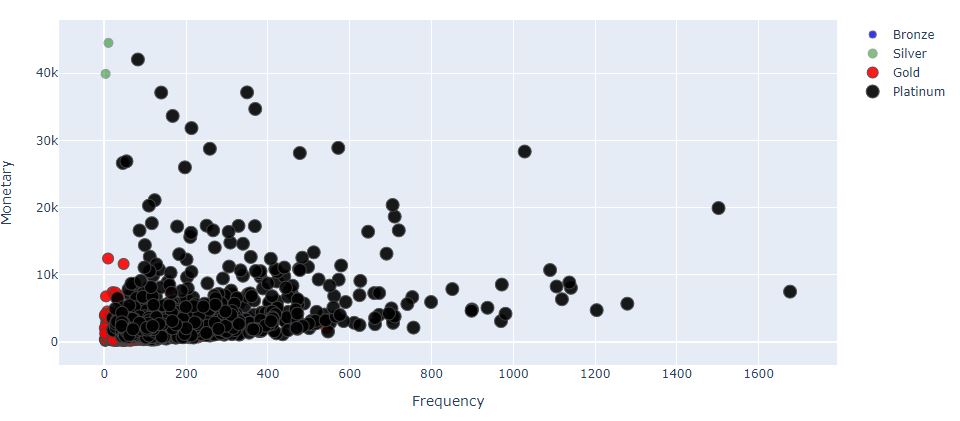
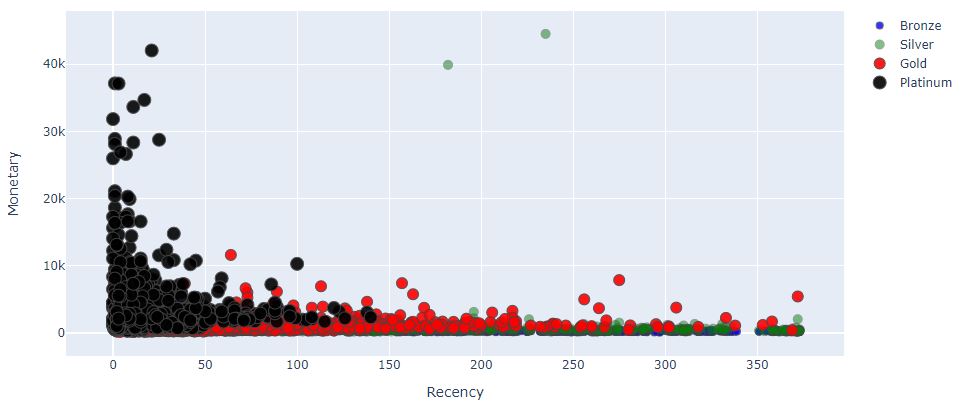
Fig 3.2 : Frequency vs Monetary scatter plot

Fig 3.3: Recency vs Monetary scatter plot



In the above figures we can see that the platinum valued customers (black dots) have higher frequency and monetary values and low recency values. This verifies that our RFM modelling has worked accurately.

## 3.3. K-Means Clustering

## 3.3.1. Data Normalization :

## In the figures 3.4, 3.5 and 3.6 shown below the distribution plot on the left shows data before normalization and the distribution plot on the right shows the same data after normalization using logarithmic function for Recency, Frequency and Monetary values respectively.

## Fig 3.4 : Recency distribution plot before and after normalization

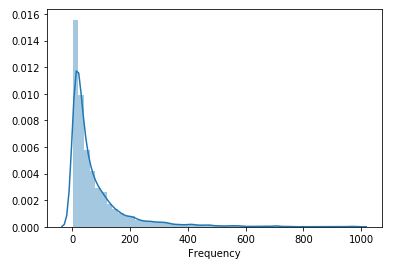
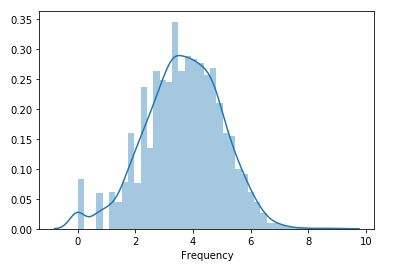


Fig 3.5 : Frequency distribution plot before and after normalization

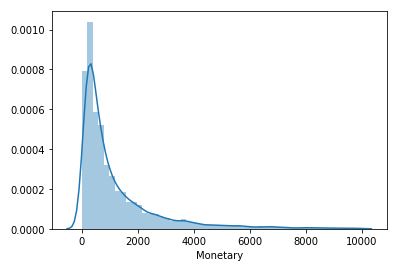
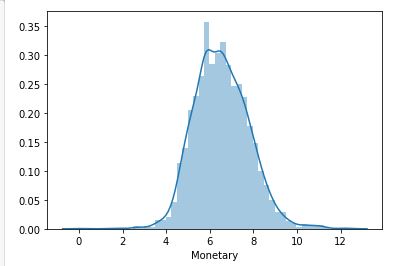


Fig 3.6 : Monetary distribution plot before and after normalization

All the three features were left skewed before normalization. So, we applied logarithmic function to each of the features to obtain normalization.

## 3.3.2. Elbow Method :

To find the number of clusters, we used elbow method. The result is shown in figure below.

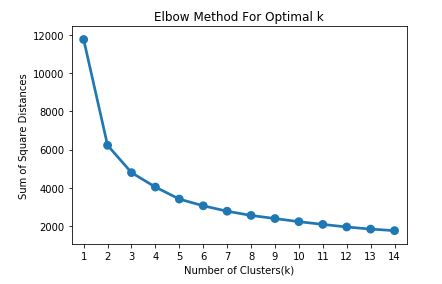


Fig 3.7 : Elbow Curve

In the fig we can see that there is a steep decrease in sum of square distance when number of clusters is 3. So, we can use the value of k as 3 while building our K-means clustering model.

Table below shows the first 10 rows of dataset after applying K-Means clustering. Every customer has been assigned to a cluster value (0, 1, 2).

0 is for the bronze level customers, 1 for silver and gold level and 2 for platinum.

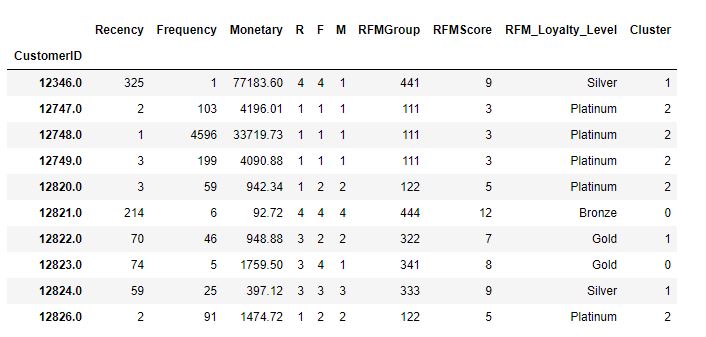


Table 3.3 : Customers RFM dataset after K-Means clustering

**4**

**Conclusion and Scope of Future Work**

Customer segmentation based on the buying pattern of customers though strategical algorithms, is a challenging task. Customer retention is another major concern for both online and physical enterprises. In the present work, the RFM model is implemented for synthetic and real datasets, to analyse customer segmentation. Also, clusters are evaluated using elbow method for K-Means clustering algorithm with different number of clusters. Based on the Sales Recency, Sales Frequency and Sales Monetary can be analysed and an optimal solution is found.

The scope of future work in this area lies in the study and analysis of specific categories of products, for example, Mobile Phones and Accessories. Various other business parameters such as the most preferred product or the most effective sales technique during as specific event, or some threshold parameters in different regions can be studied for designing effective business enhancement. Such advancements and deliberations in this area will help the enterprises to improve businesses by offering promotions and designing innovative strategies that can prove cutting edge against the competitors.

**Bibliography**

Claudio Marcus, (1998), A practical yet meaningful approach to customer segmentation, Journal of Consumer Marketing, Volume 15, issue 5, pp.494–504.

Samira Madani, (2009), Mining Changes in Customer Purchasing Behavior - a Data Mining Approach. Master Thesis, Lulea University, Department of Business Administration and Social Sciences Division of Industrial marketing and e-commerce

Brent A. Gloy Jay T. Akridge Paul V. Preckel, (1997), Customer Lifetime Value: An Application in the Rural Petroleum Market, Wiley & Sons, Inc. Agribusiness, Volume 13, No. 3, pp.335– 347.

Jinghua Zhao, Wenbo Zhang and Yanwei Liu, (2010), Improved K-Means Cluster Algorithm in Telecommunications Enterprises Customer Segmentation. Information Theory and Information Security (ICITIS), IEEE International Conference on, pp.167 – 169

Xiaoping Qin, ShijueZheng, Ying Huang and Guangsheng Deng, (2010), Improved K-Means algorithm and application in customer segmentation. IEEE International Conference on Web Information Systems and Mining, pp. 13 – 16

Yuerong Chen and Xueping Li, (2009), The Effect of Customer Segmentation on an Inventory System in the Presence of Supply Distributions. IEEE International Conference on Winter Simulation Conference (WSC), pp.2343 – 2352.

Zhang xiao-bin and Gaofeng, Huang hui, (2009), Customer-churn Research Based on Customer Segmentation. IEEE International Conference on Electronic Commerce and Business Intelligence, pp.443 – 446.

Burcu Sag ̆lam, F. Sibel Salman, SerpilSayın and MetinTu ̈rkay, (2006), A mixed-integer programming approach to the clustering problem with an application in customer segmentation. Elsevier Conference on European Journal of Operational Research 173, pp. 866–879.

Zhou Zhongding, Miao Xuemei and Liu Guangcan, (2009), Customer Segmentation Algorithm of Wireless Content Service Based on Ant K-Means. IEEE Conference on International Forum on Computer Science Technology and Applications, Volume 1, pp.267 – 269.

Izak Benbasat, David K. Goldstein and Melissa Mead (1987) The case research strategy in studies of information systems. MIS Quarterly, 11(3), pp. 369-386.

Mike McGuirk, (2007), Customer Segmentation and Predictive Modeling. Behavioral Sciences, marketing and analytic consulting firm located in Burlington, Massachusetts. Available at : www.emarketingpapers.com/content10131

Osmar R. Zaïane, (1999), CMPUT690 Principles of Knowledge Discovery in Databases, Available at: [www.di.ubi.pt/~ddg/aulas/licenciatura/dwdm/dw/ch1s.pdf](http://www.di.ubi.pt/~ddg/aulas/licenciatura/dwdm/dw/ch1s.pdf).

Xi Zhang and Yu Tang, (2006), Customer Perceived E-service Quality in Online Shopping. Master Thesis, Lulea University, Department of Business Administration and Social Sciences Division of Industrial marketing and e-commerce.

LalehNosrati, (2008), The Impact of Website Quality on Customer Satisfaction. A Research on Iranian Online Bookstores, Master Thesis, Lulea University, Department of Business Administration and Social Sciences Division of Industrial marketing and e-commerce.

Chun Wang Zheng Wang, (2006), The Impact of Internet on Service Quality in the Banking Sector. Master Thesis, Lulea University, Department of Business Administration and Social Sciences Division of Industrial marketing and e-commerce. Yin, R. K. (2003), Case study Research Design and Methods (3rd ed.) California: Sage Publications.

Vasilis Aggelis and Dimitris Christodoulakis, (2005), Customer Clustering using RFM analysis, 9th WSEAS International Conference on Computers, Special Session Data Mining, Techniques and Application.

Chu Chai Henry Chan, (2008), Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. Elsevier Conference on Expert Systems with Applications Volume 34, pp. 2754–2762.

Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh and Hyun-Seok Hwang (2006), Customer segmentation and strategy development based on customer lifetime value: A case study, Elsevier Conference on Expert Systems with Applications Volume 31, pp. 101–107.

Cheng Li (2008) Research on Segmentation implementation process of air cargo customer based on Data Mining, IEEE 4th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-4.

Huaping Gong, Qiong Xia (2009), Study on Application of Customer Segmentation Based on Data Mining Technology, IEEE Conference on ETP International Conference on Future Computer and Communication, pp. 167 – 170.

Xin-an Lai (2009), Segmentation Study on Enterprise Customers Based on Data Mining Technology, IEEE First International Workshop on Database Technology and Applications, pp. 247 – 250.

Tian Yuanli and shaoliangshan (2010), Customer segmentation based on Ant clustering Algorithm, IEEE Second Conference On Computational Intelligence and Neural computing (CINC) Volume 1: pp. 133 – 136.

Babak Sohrabi, Amir Khanlari (2007), Iranian Accounting & Auditing Review, Volume 14 No. 47, pp. 7- 20.