



**FOM Hochschule für Oekonomie & Management**

**Hochschulzentrum Bangalore**

**Master-Thesis**

im Studiengang Big Data & Business Analytics

zur Erlangung des Grades eines

**Master of Science (M.Sc.)**

über das Thema

**Customer Segmentation by Using RFM Model and K-means Clustering  
Technique**

von

**Agnishwar Das**

Erstgutachter

Prof. Dr. Serkan Akbay

Matrikelnummer

547965

Abgabedatum

2022-08-21

## Table of Contents

<b>Index of Figures .....</b>	<b>IV</b>
<b>Index of Tables .....</b>	<b>V</b>
<b>Index of Formula .....</b>	<b>VI</b>
<b>List of Abbreviations .....</b>	<b>VII</b>
<b>1. Introduction.....</b>	<b>1</b>
1.1. Domain.....	1
1.2. Statement of the Problems.....	6
1.3. Research Objectives.....	7
1.4. Research Questions.....	7
<b>2. Theoretical basics of customer segmentation.....</b>	<b>7</b>
2.1. Conceptual Background and related Literature.....	7
2.2. Customer Relationship Management.....	12
2.3. Customer Segmentation.....	14
2.3.1. Demographic Segmentation.....	16
2.3.2. Geographic Segmentation.....	16
2.3.3. Behavioral Segmentation.....	17
2.3.4. Psychographic Segmentation.....	18
2.4 K-means Clustering.....	19
<b>3. Recency, Frequency and Monetary (RFM) Analytic Model.....</b>	<b>24</b>
3.1. Definition of RFM analysis.....	25
3.2. Recency, Frequency and Monetary Analysis for Customer Segmentation.....	26
3.3. Steps of RFM Model.....	28
3.4. Relationship between RFM analysis and small and medium sized enterprises.....	28
<b>4. Empirical Investigation of customer segmentation.....</b>	<b>30</b>
4.1. Research Design.....	30
4.1.1 Research Purpose.....	32

4.1.2	Research Approaches.....	32
4.1.3	Research Strategy.....	34
4.1.4	Research Instrument.....	34
4.2	Data Collection.....	34
4.3	Data Preprocessing.....	35
4.4	Empirical Finding.....	38
4.4.1	Exploratory Data Analysis.....	38
4.4.2	Customer Segmentation Based on RFM model.....	40
4.4.3	Data Clustering and Customer Segmentation.....	47
4.4.4	Strategy Definition Per Segment.....	53
<b>5.</b>	<b>Discussion.....</b>	<b>54</b>
5.1.	Conclusion.....	54
5.2.	Recommendation for further Research.....	56
	<b>Appendix I: RFM Analysis Code.....</b>	<b>57</b>
	<b>Appendix II: K-means Clustering Code.....</b>	<b>58</b>
	<b>Bibliography .....</b>	<b>59</b>

**Index of Figures:**

Figure 1: Types of Customer Segmentation.....	15
Figure 2: Elbow method representing the number of clusters.....	24
Figure 3: Pyramid model.....	27
Figure 4: Categorization of marketing in SMEs.....	29
Figure 5: Research Design.....	31
Figure 6: Online Retail dataset.....	35
Figure 7: Summary statistics of retail dataset.....	36
Figure 8: Customer distribution by country.....	37
Figure 9: Dataset after preprocessing.....	38
Figure 10: Summary statistics of retail dataset.....	39
Figure 11: Information of the data frame.....	39
Figure 12: RFM value.....	40
Figure 13: Distribution plot of Recency.....	41
Figure 14: Distribution plot of Frequency.....	41
Figure 15: Distribution plot of Monetary.....	42
Figure 16: Box plot of Recency.....	43
Figure 17: Box plot of Frequency.....	44
Figure 18: Box plot of Monetary.....	44
Figure 19: Top 5 data with RFM loyalty level.....	45
Figure 20: Frequency Vs Recency.....	46
Figure 21: Monetary Vs Frequency.....	46
Figure 22: Monetary vs Recency.....	47
Figure 23: The distribution of Recency after normalization and standardization.....	48
Figure 24: The distribution of Frequency after normalization and standardization.....	49
Figure 25: The distribution of Monetary after normalization and standardization.....	49
Figure 26: Elbow method for Optimal K.....	50
Figure 27: Clusters of the customers.....	51
Figure 28: Frequency Vs Recency with three different clusters.....	51
Figure 29: Monetary Vs Recency with three different clusters.....	52
Figure 30: Monetary Vs Frequency with three different clusters.....	52

**Index of Tables**

Table 1: Dimensions of the RFM Model.....	25
Table 2: Differences between qualitative and quantitative research method.....	33
Table 3: Marketing strategy for each segment.....	53

**Index of Formula:**

Formula 1: Euclidean distance.....	22
Formula 2: Mean of the data points.....	23
Formula 3: Mean squared error.....	23
Formula 4: Formula of Standardization.....	48

**List of Abbreviations**

B2B	Business-to-business
B2C	Business-to-Consumer
CRM	Customer Relationship Management
SFA	Sales Force Automation
CSS	Customer Service System
RFM	Recency, Frequency and Monetary
DBSCAN	Density-Based Spatial Clustering of Application with Noise
WRFM	Weighted RFM
AHP	Analytic Hierarchy Process
MSE	Mean Squared Error
DB	Database
SMEs	Small and medium-sized enterprises
ROI	Return on Investment

## 1. INTRODUCTION:

The success or failure of a business depends on a variety of factors. The primary source of income for businesses is valuable, intangible assets and the revenue earned from operating from day-to-day activities for businesses that directly refers to financial outcomes. Businesses need to avoid losing existing customers and try to attract new customers. A company cannot rely on its product or service no matter how good it is because nowadays customers demand a high level of customer service.<sup>1</sup>

### 1.1. DOMAIN:

Industrial markets are facing intensive business cycles, which creates both opportunities and threats for the operating of the company in these markets. The changes have forced management to change their way of thinking. The focus of management is not only to search for and attract new customers but also to satisfy and maintain the existing customers. Keeping existing customers is less costly and more beneficial compared to attracting new customers. When existing customers are satisfied, they will buy the products again. It is more cost-effective to maintain the existing customers.<sup>2</sup>

Customer segmentation is unlimited potential as a tool that can guide any organization towards more effective ways to market their product or services. Some bases of customer segmentation that may be used include geographic, demographic, psychographic, and behavioral. Other variables may be used for segmentation such as situation and customer preferences for products or specific product attribute levels.<sup>3</sup>

Many previous studies have been executed showing how companies have successfully applied customer segmentation to enhance marketing decisions. Companies should identify valuable customers, rather than serving all customers.<sup>4</sup>

Some customers are too costly to do business with them and some customers have very little potential to become profitable customers. Grouped customers based on customer

---

<sup>1</sup> Cf. KABASAKAL, İ., Customer Segmentation, 2020, pp. 47-56.

<sup>2</sup> Cf. Maryani, I. et al., Customer Segmentation and Clustering Techniques, 2018, pp. 1-6.

<sup>3</sup> Cf. Bruce, Cooil. et al., Approaches to customer segmentation, 2007, pp. 9-39.

<sup>4</sup> Cf. ZEITHAML. et al., Services Marketing: Integrating Customer Focus Across the Firm, 2000, no page number.



segmentation, can help companies identify the customer needs and different levels of profitability. Thus, marketing strategies can be developed based on the characteristics of these customer groups.<sup>5</sup>

Customer segmentation can be explained as a game where a kid separates balls, and cubes based on their colors or shape. In general, customer segmentation isolates customers, and markets on different criteria, and groups them based on similar characteristics. Customer segmentation is based on several considerations. This includes demographic information about consumers such as ethnicity, religion, wages as well as information about their lifestyle, place, and buying habits.<sup>6</sup>

Customer segmentation is the process of dividing customers based on their characteristics so that company can market to each segmentation effectively. In B2B marketing, a company might segment customers according to some factors including Industry, number of employees, products previously purchased from the company, and the location but in B2C marketing a company might segment the customers based on some factors including Age, Gender, Location, Marital Status, etc. Customer segmentation requires a company's specific information or data about customers and analyzing it to identify patterns that can be used to create customer segmentation. Some of the typical information-gathering methods are face-to-face or telephonic interviews, Surveys, and focus groups.<sup>7</sup> In the fast-changing retail industry, there is a clear need for advanced methods to discover customer segments from sales and other data. It will empower retailers to reach consumers with specific needs and demands, by dividing the market into similar and recognizable segments. This will focus on individuals with similar characteristics.<sup>8</sup>

Companies have used several segmentation techniques to better identify and understand customer groups and provide preferable services or products to them to satisfy their needs

---

<sup>5</sup> Cf. *Tsai, Chih Fong. et al.*, Customer segmentation issues and strategies, 2015, pp.65-76.

<sup>6</sup> Cf. *Toyeb, M. et al.*, LRFMV: an efficient customer segmentation model for superstores, 2021, pp- 71-75.

<sup>7</sup> Cf. *Stormi, Kati. et al.*, Feasibility of b2c customer relationship analytics, 2018, no page number.

<sup>8</sup> Cf. *Heikkilä. et al.*, Segmenting retail customers, 2018, pp. 108-116.

and preferences. A company can create profitable segments and react to the selected segment based on its competitiveness.<sup>9</sup>

Customers have various kinds of needs and want. The company used several segmentation criteria and techniques to identify and understand the customer groups and provide them with preferable products and good services to satisfy their needs and wants. Segmentation allows businesses to better use their marketing budgets and demonstrate better knowledge of customers' needs and wants. Traditionally, marketers must identify market segmentation using a mathematical model and based on the segmentation, implement the marketing strategy to target profitable customers.<sup>10</sup> It can also help:<sup>11</sup>

- Customer and marketing services can be made effective with the help of the customer segmentation analysis that gives the company comprehensive insights into the different customer groups. All customers do not expect the same things from a brand. So, without a customer segmentation model, it is very difficult to understand the needs of customers more accurately and a few companies often struggle to meet the customers' expectations.
- Create the target audience using customer segmentation. Select the best communication channel for the segmentation which might be emailed, text messages, social media posts, radio, or TV advertising based on the segment.
- Once a company identified the key motivators such as design, price, or needs, they can brand their products appropriately.
- Improving the quality of products or services: Once the company gets feedback received from different segments, it is easier to understand how the company's offering is helping the customers. This can help the company determine how to position their solution to prospects and uncover needs that they can solve with new products or services in the future.

---

<sup>9</sup> Cf. *Onur, DOĞAN. et al.*, Customer segmentation, 2018, pp. 1-19.

<sup>10</sup> Cf. *Xiong, Weiwen. et al.*, RFM value and grey relation based customer segmentation model, 2008, pp. 1298-1301.

<sup>11</sup> Cf. *Gil-Saura. et al.*, Retail customer segmentation, 2009, pp. 253-266.

Existing customers can refer others to the business: If the existing customers are happy with products and pleased with the customer service, they will refer others to the business naturally. They can refer their family or friends to this business. The way Loyal existing customers refer others to the business:<sup>12</sup>

- Word of mouth recommendations
- Social media posts and messages about the brand.
- Positive comments and reviews about the product or services.

Customer retention statistics:<sup>13</sup>

- The probability of selling products to an existing customer is 60-70 percent. The probability of selling products to a new customer is 5-20 percent.
- 65 percent of a company's business comes from existing customers.
- 80 percent of future profits will come from just 20 percent of the existing customers.
- If the customer retention rate increases by 5 %, profit will increase from 25 % to 95%.
- Obtaining new customers costs 5 times more than keeping existing customers.

Knowledge of customer behavior is very crucial to provide them the high-quality products and services, keep them loyal customers and make big profits for the business. Customer Relationship Management (CRM) was first introduced in the United States in 1990 and has developed from the Sales Force Automation (SFA), Customer Service System (CSS) to Call Center. It combines the concepts of modern marketing and field services.<sup>14</sup> It also combines Computer Telephone Integrated Technology and Internet Technology. CRM

---

<sup>12</sup> Cf. *Lee, Chung-Shing.*, An analytical framework for evaluating e-commerce business models and strategies, 2001, no page number.

<sup>13</sup> Cf. *Georgiev, Deyan.*, Customer Retention Statistics & Predictions [Updated 2022], <https://review42.com/resources/customer-retention-statistics/., 2022>, Accessed on 12.05.2022.

<sup>14</sup> Cf. *Z, Pan. et al.*, CRM Adoption Success Factor Analysis and Six Sigma DMAIC Application, 2007, pp. 828-838.

consists of four dimensions, i.e., Customer identification, Customer attraction, Customer retention, and customer development.<sup>15</sup>

Customer segmentation was arranged in the first dimension of CRM as customer identifications that can identify groups of customers based on common characteristics and behaviours. For Customer retention and customer development dimensions, Demographic variables and Recency, Frequency, and Monetary (RFM) were used. CRM helps business to better understand their customers and allocates the resources effectively to the most important group of customers. we could not discover the meaning of information from the collected raw customer data but the data mining techniques like clustering and classification could help to find meaningful information and patterns from the raw customer data.<sup>16</sup>

RFM is an important quantitative analysis model in customer relationship management. RFM model uses three parameters to describe customers' importance and characteristics of customers. An organization uses RFM analysis to analyze the sales history and characteristics of customers and identify potential customers.<sup>17</sup>

A common problem for companies is that they have one marketing strategy for all the customers but want to know how to target specific customers with marketing strategies that are based on their purchasing behavior. In this way the response rate of customers is increasing, so more products will be sold out. It saves time and money that are invested for never returning customers. It is very interesting for B2C (Business to Consumer) companies because they have a lot of different private customers. It is hard to segment all the customers a company has. Instead of analyzing all the customers separately, it's better to look at the segmentation of customers. RFM analysis is a three-dimensional way of identifying customers to determine the best customers. It is based on the 80/20 principle that 20% of customers bring in 80 % of the revenue. RFM analysis suggested that the customer showing a high RFM score should normally conduct more transactions. This

---

<sup>15</sup> Cf. *Swift, Ronald S.*, Accelerating customer relationships, 2001, no page number.

<sup>16</sup> Cf. *Rygielski, Chris. et al.*, Data Mining techniques For Customer Relationship Management, 2002, pp. 483-502.

<sup>17</sup> Cf. *Huang, Yong et al.*, RFM customer segmentation model based on K-Means algorithm, 2020, pp.24-27.

leads to a better profit for the organization.<sup>18</sup> RFM (Recency, Frequency, and Monetary) analysis is a famous technique used for evaluating customers based on their characteristics. A scoring method is developed to evaluate scores of Recency, Frequency, and Monetary. The scores of these variables are merged as RFM scores that help to predict future patterns by analyzing the present and past records of the customer.<sup>19</sup>

Data Mining is a technique that explores some unknown and potentially useful information and knowledge from large, incomplete, noisy, and random data.<sup>20</sup> K-means Clustering is the simplest algorithm of clustering based on the partitioning principle.<sup>21</sup>

In this thesis, to segment customers, the RFM model is used. It stands for Recency, Frequency, and Monetary. Customer segmentation and clustering are carried out using the K-Means algorithm. It is based on RFM values. Python and Tableau are used for this thesis.

## **1.2. Statement of the Problems:**

Customers are different and have different needs. They are demanding more of their favourite products or service. Identification of customers' behaviours, their needs, and keeping them loyal are the main key factors for sales, marketers, and business. It is not an optimal decision to have the same strategy and marketing for every customer. It is not only just building relationships with customers but also retaining those customers and attracting new customers as well.

Retailers are the connecting link between the producer and customers. They face many problems to market their products from various dimensions. Now customers are more dynamic. They have some expectations from the products they bought from a retail store such as Quality, Price, Quantity, and product service. Many businesses have huge tons of customer data, but they cannot utilize it in a proper way that helps to grow their business. This is because of lacking skills and knowledge on how to manage, process, and transform

---

<sup>18</sup> Cf. *Pakyurek, Muhammet. et al.*, Customer clustering using RFM analysis, 2018, pp. 1-4.

<sup>19</sup> Cf. *Christy. et al.*, RFM ranking, 2021, pp. 1251-1257.

<sup>20</sup> Cf. *Zhao, Jinghua. et al.*, Improved K-means cluster algorithm, 2010, pp. 167-169.

<sup>21</sup> Cf. *Kansal, Tushar. et al.*, Customer segmentation using K-means clustering, 2018, pp. 135-139.

data into meaningful insights. For a company, it is important to segment their customers based on their purchase behaviour and make strategies for them.

### **1.3. Research Objectives:**

This research consists of RFM analysis and K-means clustering to construct a customer segmentation model, to segment customers into different groups based on their customer demographics and purchase behaviours. The insights and customer segments are very helpful to better understanding the customers and applying the right customer-centric marketing strategy.

### **1.4. Research Questions:**

1. How to apply RFM analysis to segment customers?
2. How many segments must be considered for segmentation using K-means clustering?
3. Which strategies must be defined for obtained segments?
4. What are the limitations and bugs of this thesis?

## **2. Theoretical basics of customer segmentation:**

Decision-makers use many variables to segment customers. The easiest and most common variables are demographic variables like age, gender, education level, family, and income. Some other major variables are used for customer segmentation such as geographic, psychographic, sociocultural, and behavioral variables.<sup>22</sup>

### **2.1. Conceptual Background and Related Literature:**

There is a lot of transaction performed by the customer every day. This process generates a lot of data. This generates 82,648 transactions from January to December 2017. The author suggested customer segmentation on Nine Reload Credit by using a data mining process based on the RFM model and clustering algorithms. They used the K-Means clustering algorithm. This K-Means generates a visual cluster model with RapidMiner 5.2

---

<sup>22</sup> Cf. Onur, DOĞAN. *et al.*, Customer segmentation, 2018, pp. 1-19.

tools. This tool represents the number of customers in each cluster by using Recency, frequency, and Monetary (RFM) attributes. Based on the RFM model they calculated 102 customers in all 82,648 transactions. They analyzed clusters with the results of 63 Customers in Cluster 1 and 39 Customers in Cluster 2. The results can be used by the company to know more about customer category and to maintain the customer owned. By knowing the categories of each customer, the company will be able to take the right decision in marketing strategy.<sup>23</sup>

Customer segmentation based on the RFM model is very popular to classify the customers based on their characteristics and purchase behavior. Many clustering algorithms have been implemented to segment customers into groups to get a better clustering result. The authors proposed a comparative analysis among agglomerative, k-means, and advanced k-means that are carried out for RFM based market segmentation approach. They noticed that agglomerative clustering needs a long processing time for large datasets compared to k-means and advanced versions of k-means clustering but the crucial advantage of agglomerative clustering is, that it does not require appointing the total number of clustering initially. The results showed that the advanced version of k-means clustering can effectively reduce the total running time by 27.8% and 97.8% compared to standard k-means and agglomerative clustering respectively and increase the speed of clustering effectively. It can also create a better clustering result than standard k-means in respect of intra-cluster distance and inter-cluster distance. Their suggested method of clustering-based customer segmentation uses an advanced version of k-means clustering techniques which reduces the overall time complexity and performs well for large datasets.<sup>24</sup>

The authors discussed a proposed and new feature selection method using partitioning techniques, data mining tools, and RFM analysis. Sample data are classified based on frequency using parameters (recency, frequency, and monetary). They used these techniques to solve the customer dataset. According to them, in a future implementation, the research based on the classification rules can be designed using the demographic

---

<sup>23</sup> Cf. *Maryani, I. et al.*, Customer Segmentation and Clustering Techniques, 2018, pp. 1-6.

<sup>24</sup> Cf. *Shihab. et al.*, RFM Based Market Segmentation Approach, 2019, pp. 1-4.

variables. There is a chance to predict future customer behaviors based on the RFM values of customer segments.<sup>25</sup>

Yash Parikh and Eman Abdelfattah performed a RFM analysis on online transactions in their journal article ‘Clustering Algorithms and RFM Analysis Performed on Retail Transactions’. Along with performing RFM analysis on the retail dataset, clustering algorithms such as Mean shift, Density-Based Spatial Clustering of Application with Noise (DBSCAN), Agglomerative clustering, and K-Means were employed. The goal is to provide strategies for customer purchasing behaviors. The author examined all clustering algorithms and compared in their ability to classify the different customer groups based on RFM analysis. All clustering algorithms are examined and compared their ability to classify the different customer groups based on RFM analysis. This identification can help businesses figure out how to spend their marketing and which customers to focus on. By applying certain techniques to certain customer segments, their values can increase accordingly and have an impact such as retaining slipping customers or keeping loyal customers happy.<sup>26</sup>

Marina E. Tsoy, Vladislav Yu. Shchekoldin (2016) suggested RFM analysis, and classification methods to determine a customer’s behavior in the future based on his/her previous behavior in their article ‘RFM-analysis as a Tool for Segmentation of High-tech Products’ Consumers’. They concluded as the distribution of customer segments is nonuniform, RFM analysis implementation as a segmentation tool is an effective way to adjust the work with client groups and therefore, to increase sales and profit, and adaption of RFM analysis on the example of wholesale and retail trade of high-tech products has identified three target groups of consumers. It is recommended to improve the marketing and promotion policy.<sup>27</sup>

Peiman Alipour Sarvari, Alp Ustundag, and Hidayet Takci proposed performance evaluation of different customer segmentation based on RFM analysis and demographic

---

<sup>25</sup> Cf. *Sheshasaayee. et al.*, Implementation of clustering technique, 2019, pp. 1166-1170.

<sup>26</sup> Cf. *Parikh. et al.*, Clustering algorithms and RFM analysis performed on retail transactions, 2020, pp. 0506-0511.

<sup>27</sup> Cf. *Tsoy, Marina E. et al.*, RFM-analysis as a tool for segmentation of high-tech products' consumers, 2016, pp. 290-293.



analysis. Their goal is to determine the best approach to customer segmentation RFM considerations as well as demographic factors. Different types of scenarios were designed, performed, and evaluated under uniform test conditions. To classify accurate customer segmentation, a combination of RFM analysis and demographic attributes are recommended for clustering. The result indicates the significant importance of demographic data merged with weighted RFM (WRFM).<sup>28</sup>

Jing Wu and Zheng Lin designed a monetary matrix and fluctuate-rate matrix to study various modes. They used credit card consumption data as model-building samples and presented a modeling framework for building segment-level predictive models that utilize a pattern-based clustering approach and signature discovery techniques. As a result, they discover different customer characteristics through clustering on both matrixes. The authors can build a two-dimension Consumption-Based customer segmentation model using these characteristics.<sup>29</sup>

HAO Su-li designed the customer lifetime value evaluation indicator system, and the unascertained clustering in his article ‘The Customer Segmentation of Commercial Banks Based on Unascertained Clustering’. He wanted to establish the customer segmentation model of commercial banks. By applying the unascertained clustering, the article divides the commercial banks’ customers into four classes: quality customers, backbone customers, mass customers, and low-class customers. This unascertained clustering overcomes the deficiency of C-Mean clustering. This clustering is more scientific than C-mean value clustering and fuzzy clustering.<sup>30</sup>

İnanç KABASAKAL (2020) analyzed RFM (Recency, Frequency, and monetary) technique based on two scoring approaches. The sales data from an e-retailer has been analyzed for clustering using a prototype software and clusters uncovered from RFM analysis were compared using a cluster evaluation matrix. The target is to classify customer segmentation along with a relevant offer for marketing strategies. The resulting

---

<sup>28</sup> Cf. *Sarvari.*, Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis, 2016, no page number.

<sup>29</sup> Cf. *Wu, Jing.*, Research on customer segmentation model by clustering, 2005, pp. 316-318.

<sup>30</sup> Cf. *Su-li, Hao.*, The customer segmentation of commercial, 2010, pp. 297-300.

RFM analysis is presented with member statistics. This study provides a demonstration to identify conventional customer segments such as champions and loyalists.<sup>31</sup>

XIONG Weiwen, CHEN Liang, ZHANG Zhiyong, and QIU Zhuqiang applied the RFM and grey correlation models to evaluate proposed segment customers and identify customer behavior. The AHP (Analytic Hierarchy Process) algorithm is used to compute the weights of indicators. The purpose of this paper is present a novel approach that combines customer targeting and customer segmentation for marketing strategies. They Performed an empirical study of a logistics enterprise to segment 10 customers. The way to segment customers is effective.<sup>32</sup>

Daoud, Rachid et al., proposed a RFM model and clustering techniques to evaluate customers' values in the sector of electronic commerce. They first used the Self-organizing maps method (SOM) to decide the best number of clusters and then the K-means clustering method is applied to classify 730 customers into eight clusters when R, F, and M are the variables. Thus, they developed effective marketing strategies for each segmentation. The result shows that cluster 7 is the most crucial cluster because the average value of R, F, and M are higher than the overall average value.<sup>33</sup>

According to Namvar, Morteza et al., customer segmentation is a central issue in today's competitive commercial area. They constructed a new customer segmentation method based on RFM, demographic, and LTV data with the help of data mining tools. They used a dataset from an Iranian bank for this method. This customer segmentation includes two phases. The first phase is to classify customers into different segments regarding their RFM, with K-means clustering. Secondly, using demographic data, each cluster again is split up into new clusters. Lastly, a customer profile is created using LTV.<sup>34</sup>

According to Gustriansyah, Rendra et al., RFM is a simple but effective method that can be applied to market segmentation. In their study, RFM analysis had been used for product segmentation. This study has presented a new procedure for RFM analysis using

---

<sup>31</sup> Cf. *Kabasakal, I.*, Customer Segmentation Based on Recency Frequency Monetary Model, 2020, pp. 47-56.

<sup>32</sup> Cf. *Weiwen, Xiong. et al.*, RFM value and grey relation-based customer segmentation, 2008, pp. 1298-1301.

<sup>33</sup> Cf. *Daoud, Rachid Ait. et al.*, Combining RFM model and clustering techniques, 2015, pp.1-6.

<sup>34</sup> Cf. *Namvar, Morteza et al.*, A two phase clustering method for intelligent customer segmentation, 2010, pp. 215-219.

the k-means method and eight validity indexes to determine the optimal number of clusters. These eight validity indexes are Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski-Lai Index. The evaluation results showed that the optimal number of clusters for the k-means method applied in the RFM analysis consists of three segmentations with a variance of 0.19113.<sup>35</sup>

## **2.2. Customer Relationship Management:**

Customer Relationship Management (CRM) is one kind of practice, strategy, and technology that organizations use to understand their markets and customers to increase relationships with their customers and maximize customer value. Customer Relationship Management is described by four elements of a simple framework: Know, Target, Sell, and Service. CRM is the strategy for building, managing, and increasing the strength of loyal and long-lasting customer relationships.<sup>36</sup>

Customer segmentation is one of the basic parts of CRM. The theory of customer segmentation is a process that total customers are divided into equivalent groups based on their purchasing behavior and characteristics.<sup>37</sup>

CRM needs the firm to understand its markets and customers. This includes detailed customer data such as the most profitable customers and those customers who are no longer worth targeting. CRM also needs the development of offers such as which products to sell to which customers and through which channel. The organization use campaign management to increase the marketing department's effectiveness. CRM is a two-stage concept. The first stage is to learn the basics of building customer-centric service which follows the principle of customer-centric marketing and mass customization. The focus should be on customer orientation rather than product features. In the second stage, companies need to combine their systems with CRM by leveraging technology to achieve real-time customer management and their value proposition to customers.<sup>38</sup>

---

<sup>35</sup> Cf. *Gustriansyah, Rendra. et al.*, Clustering optimization in RFM analysis based on k-means, 2020, pp. 470-477.

<sup>36</sup> Cf. *Tsiptsis, Konstantinos K. et al.*, Data mining techniques in CRM, 2011, no page number.

<sup>37</sup> Cf. *Zhao, Jinghua. et al.*, Improved K-means cluster algorithm in telecommunications enterprises customer segmentation, 2010, pp. 167-169.

<sup>38</sup> Cf. *Rygielski, Chris.*, Data Mining techniques For Customer Relationship Management, 2002, pp. 483-502.

According to there are four dimensions of Customer Relationship Management. These are:<sup>39</sup>

1. Customer Identification: The first step of the CRM process is customer identification. This process includes targeting the population who will become customers or most profitable customers to a company and who are being lost to the competition and they can be returned as well as to identify which products to sell to which customers based on the previous purchasing behavior. For this, a company needs detailed customer intelligence, collecting the customer's data and analyzing customer's data, activities, and behavior. There are two elements for customer identification i.e. target customer analysis and customer segmentation. Target customer analysis is used to classify the specific groups of customers who are most profitable to the organization or loyal customers through the analysis of customer data. Customer segmentation is to segment the customer into a smaller group based on their similarity and purchase behavior.
2. Customer Attraction: After customer identification, this is the next phase. Since target groups were classified by target customer analysis or customer segmentation, an organization can direct effort and resources into attracting the target group. Direct marketing is an element of customer attraction. Direct marketing is described as the promotion process, which motivates customers to place an order through various channels such as coupon distribution, direct mail, and direct call
3. Customer Retention: Customer retention is the central concern for CRM. Customer satisfaction, loyalty, and commitment are the essential components of customer retention. Customer retention identifies and helps retain potential customers as well as supports the organization to encourage less profitable customers to become loyal and more profitable customers. The most important elements of customer retention are one-to-one marketing, loyalty programs, and complaints management which refers to customized marketing campaigns that are supported by analyzing customer behavior and predicting their behavior.
4. Customer Development: This involves transaction value, maximizing customer intensity, and customer profitability. The customer lifetime value analysis, up/cross-

---

<sup>39</sup> Cf. *Ngai, E, et al.*, Application of data mining techniques in customer relationship management, 2009, pp. 2592-2602.

selling, and market basket analysis are the elements of this phase. Customer lifetime value analysis is the prediction of the net profit a company expects from a customer. Up/Cross-selling refers to promotion activities that raise the value of a single transaction. Market basket analysis refers to maximizing the customer transaction intensity and additional purchased services or products.

Customer Relations Management is a philosophy in business management that includes customer-focused strategies to acquire and retain customers. CRM is an essential approach to improving customer satisfaction, customer acquisition, customer retention, and profitability. CRM strategies emphasize the importance of customers for businesses and promote customer-centric practices in marketing. Implementing CRM for business purposes involves practices and methodology to make use of customer data. Businesses try to utilize existing data sources and make data-driven decisions for competitive advantage in such a data-oriented business environment. The analysis of customer data helps businesses to establish the relationships between data elements, describe significant events, and give predictions. In this case, marketers try to divide total markets and classify customer segmentation based on geographic, psychographic, demographic, and behavioral variables. Clustering techniques in data mining are often used to analyze customer behaviors in customer segmentation.<sup>40</sup>

### **2.3. Customer Segmentation:**

In this modern era, everyone is involved in a competition to be better than others. Nowadays businesses run based on such innovations having the ability to attract customers with the products. The companies are confused about what section of customers to target to sell their products. Here machine learning plays a vital role in revealing the hidden patterns in the data for better decision-making in the future. The process of segmenting customers with similar behaviors into the same segment and with different patterns into different segments is called customer segmentation.<sup>41</sup> Customer Segmentation is the process of classifying customers into a group based on certain traits

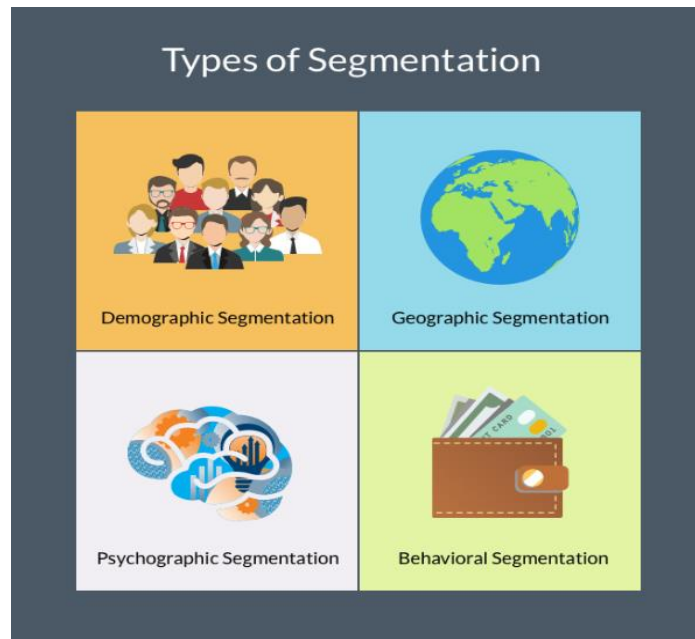
---

<sup>40</sup> Cf. *Kabasakal, I.*, Customer Segmentation Based on Recency Frequency Monetary Model, 2020, pp. 47-56.

<sup>41</sup> Cf. *Kansal, Tushar. Et.al.*, Customer segmentation using k-means clustering, 2018, pp.135-139.

(interests, habits) and factors (demographics, income) they share.<sup>42</sup> Customer segmentation requires customer data from various sources. Customer Segmentation is divided into geographic, demographic, psychographic, and behavioral segmentation.<sup>43</sup>

**Figure 1: Types of Customer Segmentation**



Source: Author generated

A combination of four types of customer segmentation criteria—demographic, geographic, behavioral, and psychographic—is often used for customer segmentation. Two-step cluster analysis tells that Education has the biggest influence on demographic criteria, low-quality clusters influence geographic criteria, frequency and monetary shows the biggest influence on behavioral criteria and the influences of psychographic criteria decreases gradually. In general, psychographic, and behavioral criteria showed the biggest influence on customer segmentation.<sup>44</sup> There are several reasons why customer segmentation is very important. The reasons are:<sup>45</sup>

<sup>42</sup> Cf. *Baker, Kristen.*, *The Ultimate Guide to Customer Segmentation: How to Organize Your Customers to Grow Better*, 2021, no page number.

<sup>43</sup> Cf. *Sari, Juni Nurma. et al.*, *Review on Customer Segmentation Technique on Ecommerce*, 2011, pp. 400-407.

<sup>44</sup> Cf. *Madzík. et al.*, *Comparison of demographic, geographic, psychographic and behavioural approach*, 2021, pp. 346-371.

<sup>45</sup> Cf. *Baker, Kristen.*, *The Ultimate Guide to Customer Segmentation*.  
<https://blog.hubspot.com/service/customer-segmentation>, Accessed on 22.05.2022.

- Learning about customers' needs and challenges.
- Improving customer service and customer support efforts.
- Understanding who are the most valuable customers and why.
- Increasing customer loyalty.
- Communicating with segments of customers via social media.
- Identifying new opportunities for products, support, and service.

### **2.3.1. Demographic Segmentation:**

The marketers choose many customers, then pick up a group that has common characteristics based on age, size, gender, and choices. Demographic segmentation is often used because the variables of demographics are easily identified and measurable.<sup>46</sup> Age, gender, employment, status, ethnicity, etc., are considered as being demographic data.<sup>47</sup> The goal is to have a detailed customer purchase profile and focuses on measurable criteria of consumers and their households.<sup>48</sup> Even if the target is described according to non-demographic factors, demographic factors should know the size of target markets and reach target markets effectively. Generation Gap is a variable, that is very difficult to understand. More explanation will be given to show how the generation gap has been used to segment markets. Many researchers recently segment markets by using the generation gap. This idea comes from the fact that each generation has been affected by the environmental background such as music, sports, policies, and different kinds of events at that time. Demographers call these groups 'cohorts'.<sup>49</sup>

### **2.3.2. Geographic Segmentation:**

Geographic segmentation is one of the easiest to identify, grouping customers with the help of their physical location such as country, region, city, and postal code. For example, it is possible to group customers within a set radius of a certain location. This is the best option for marketers to launch live events looking to reach local audiences.<sup>50</sup> The

---

<sup>46</sup> Cf. Aziz, *Samer. et al.*, Demographic segmentation and its effects, 2012, pp. 361-379.

<sup>47</sup> Cf. Sarvari, *Peiman Alipour. et al.*, RFM and demographics analysis, 2016, pp. 1129-1157.

<sup>48</sup> Cf. *Chen, H. et al.*, Business intelligence and analytics, 2012, pp. 1165-1188.

<sup>49</sup> Cf. *Sun, Shili.*, An analysis on the conditions and methods of market segmentation, 2009, pp. 63-70.

<sup>50</sup> Cf. *Yieldify.*, <https://www.yieldify.com/blog/types-of-market-segmentation/>, 2020, Accessed on 05.22.2022.

advantage of geographic segmentation is that each customer can easily be assigned to a geographic unit. For this, it is easy to target communication messages, and select communication channels (such as local newspapers, local radio, and TV stations) to reach the selected geographic segments.<sup>51</sup> The organization can do business in one or more geographic regions. Different regions have different customs, The organization must execute the marketing strategy following the local situation. Geographic segmentation gives useful distinctions when regional needs exist. But it's important for marketers not only to use geographic location as a segmentation method because distinction among consumers who are in the same geographic location still exists. Therefore, using multiple segmentation is probably a better strategy for targeting a specific market.<sup>52</sup>

### **2.3.3. Behavioral Segmentation:**

Behavioral segmentation is very useful of all for all e-commerce businesses to identify, and group customers based on their spending habits, purchasing habits, loyalty to the brand, browsing habits, interaction with the brands, and previous product feedback. Behavioral segmentation includes such as areas such as purchase occasion, user status, degree of loyalty, degree of usage, and marketing factor sensitivity.<sup>53</sup> Tendencies and frequent actions, feature use or product use, and habits are the factors based on what customer segmentation happens.<sup>54</sup>

The segmentation included in behavioral segmentation is occasions. Buyers can be renowned according to the occasions when they need, purchase, or use the products. Occasional segmentation can help the companies to expand the usage scope of their products. Sometimes, the company can take the advantage of some festivals such as Father's Day and Mother's Day, to increase the sales of flowers and candy. There are so many companies that prepare products for promotion on Christmas.<sup>55</sup>

---

<sup>51</sup> Cf. Dolnicar, Sara. *et al.*, Market Segmentation Analysis, 2018, pp. 11-22.

<sup>52</sup> Cf. Sun, Shili., An analysis on the conditions and methods of market segmentation, 2009, pp. 63-70.

<sup>53</sup> Cf. Beane, T. P. *et al.*, Market Segmentation, 1987, pp. 20-42.

<sup>54</sup> Cf. Baker, Kristen., *The Ultimate Guide to Customer Segmentation*.

<https://blog.hubspot.com/service/customer-segmentation>, Accessed on 03.06.2022.

<sup>55</sup> Cf. Sun, Shili., An analysis on the conditions and methods of market segmentation, 2009, pp. 63-70.



Another one is user status. Markets can be segmented based on certain groups of non-users, ex-users, potential users, first-time users, and regular users of a product.<sup>56</sup>

The last one is loyalty status. The customers can be loyal to some brands, stores, or companies. The customers can be divided into four groups based on brand loyalty status such as hard-core loyal (customers who buy one brand all the time), split loyal (customers who are loyal to two or more brands), shifting loyal (customers who shift from one brand to another brand) and switchers (customers who show no loyalty to any brand).<sup>57</sup>

#### **2.3.4. Psychographic Segmentation:**

Psychographic segmentation is focused on customers' interests and personalities such as their hobbies, personality traits, life goals, value, lifestyles, and beliefs. The goal is to target those customers who are more budget-conscious and value a good deal.<sup>58</sup> In psychographic segmentation, customers are divided into different groups of customers based on the basics of lifestyle and personality.<sup>59</sup> Psychographic segmentation divides customers into different segments based on internal characteristics such that personality, lifestyle, values, interests, and beliefs. The factors of this segmentation are customers' value in real life, pain points they face, and how marketing can help in a way customers will find valuable. Based on these factors a company can adjust marketing messages, offers, and advertisement channels to provide maximum value to the target audience and connect with them on a more personal level.<sup>60</sup> In this segmentation, the customers are divided into different groups based on their lifestyle, personality, or values. People in the same demographic group may show differences in psychographic features. The customers are time or money constrained. Those whose time is limited intended to do two or more tasks at the same time. For example, they will call someone or eat while driving. Thus, some organizations offer suitable services for these time-constrained customers. The company can offer services or products at low cost for those customers, who are money

---

<sup>56</sup> Cf. Kotler, Philip. *et al.*, Principle of marketing, 2010.

<sup>57</sup> Cf. Rossi, P. *et al.*, The value of purchase history data in target marketing, 1996, pp. 321-340.

<sup>58</sup> Cf. Yoseph, Fahed. *et al.*, An enhanced RFM and a hybrid regression/clustering method, 2019, pp.77-82.

<sup>59</sup> Cf. Kotler, P., Marketing Management, 1997, pp. 257-60.

<sup>60</sup> Cf. Mialki, S., Psychographic Segmentation, <https://instapage.com/blog/psychographic-segmentation>, Accessed on 04.06.2022.

constrained. Another factor of psychographic segmentation is personality. Some customers are frank, some are full of enthusiasm, and some are reserved. The best example of using personality to segment customers is the automobile companies. They design different types of cars that are for conservative customers or fashionable customers.<sup>61</sup>

## 2.4 K-means Clustering:

Machine Learning techniques are mainly divided into two parts (Supervised machine learning and Unsupervised machine learning). In Supervised machine learning, the data are labeled. The algorithm learns from this labeled training data. In Unsupervised learning, the data are unlabeled. This machine learning helps us to find hidden and unknown patterns in data.<sup>62</sup>

Samuel et al., explained, ML is the process by which we teach a computer to perform a particular task by supplying it with a considerable quantity of practice data. The phrase "machine learning" is used in many different contexts. "The discipline of research that provides computers the ability to learn without being explicitly taught" is one of the most noteworthy instances defined by Arthur Samuel.<sup>63</sup>

According to Mitchell, in an alternate definition, the programs in a computer that learn with experience E if their performance in tasks T, as judged by P, increases as a result of experience E.

As an example, the chess game, E:

chess-playing experience Task:

Playing Chess

Performance: next round's chances of being won by a computer program

---

<sup>61</sup> Cf. *Sun, Shili.*, An analysis on the conditions and methods of market segmentation, 2009, pp. 63-70.

<sup>62</sup> Cf. *Majumder, Prateek*, K-Means clustering, 2021, <https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/>, Accessed on 08.06.2022.

<sup>63</sup> Cf. *Samuel, Arthur L.*, Some studies in machine learning using the game of checkers, 1959, p. 1.

### ➤ **Supervised or Discriminative learning**

F.Y et al., proposed Artificial Intelligence (AI), in which software teaches a computer on data with predefined outputs or labels, grows due to this learning. The model is aware of both the input and output data during training. Input data for a sample is converted to output data utilizing a mathematical function in this approach. After a few rounds, the model begins to comprehend the link between the input and output labels, improving over time. That is why it has been referred to as "supervised," which makes sense because the model may apply an optimization function to educate the model to perform more small errors after each such mistaken answer, culminating in fantastic outcomes on new data that has never been seen before.<sup>64</sup>

The two main problems solved by supervised learning are classification and regression.

#### *A. Classification models*

The financial data in this thesis was classified using a classification method. Ikonomakis et al., expressed there are two or more categories that need to be distinguished, and they may be done by employing a model that was developed using a supervised learning approach. When picture data is submitted to the model, it may determine whether the incoming image data depicts a cat or a dot.<sup>65</sup>

#### *B. Regression models*

Maulud et al., stated classification signifies a number output from given input data, this signifies a numerical output in contrast to classification. It has input and output mapping, but its mathematical function produces a numerical output. Different methods are utilized to compute loss since it involves numbers. House prices may be predicted by using a regression model that considers numerous input data.<sup>66</sup>

---

<sup>64</sup> Cf. *Osisanwo et al.*, Supervised machine learning algorithms: classification and comparison, 2017, pp. 1-3.

<sup>65</sup> Cp. *Ikonomakis et al.*, Text classification using machine learning techniques, 2005, pp. 1-5.

<sup>66</sup> Cp. *Maulud et al.*, A Review on Linear Regression Comprehensive in Machine Learning, 2020, p. 1.

### ➤ **Unsupervised or Generative learning**

According to Usama et al., in contrast to supervised learning, here, the distinction lies in the data. There are input values in the data, but there are no output labels; therefore, a machine learning algorithm cannot be trained on it. In the absence of any underlying output, the algorithm is left to discover patterns and new information on its own. Despite the lack of a specified goal, the model is often trained on an extensive dataset with a set of predefined objectives to guarantee that the data is comprehensible by the model.<sup>67</sup>

Fergus et al., explained that Google and other businesses have lately used this strategy in nearly every pre-trained model.<sup>68</sup> Unsupervised learning is also the subject of this thesis, which examines one such model. The thesis offers examples of how we have employed unsupervised learning and transfer learning to solve complex machine learning problems.

### ➤ **Semi-supervised learning**

According to Zhai et al., it combines the advantages of both supervised and unsupervised learning. While some data remains tagged during model training, a significant quantity of data is left unlabeled. Unlabeled data may now be processed using the model's capacity to learn from supervised data. Helpful in analyzing text data, as well. A text document classifier could make use of this technology.<sup>69</sup>

### ➤ **Reinforcement Learning**

Sutton et al., stated it as a new field of study, it was sometimes misunderstood with semi-supervised learning. Models learn to do particular activities in an environment where the agent receives a reward if it takes the proper action.<sup>70</sup> It is possible to use reinforcement in one of two ways: either through positive reinforcement or through

---

<sup>67</sup> Cp. *Usama et al.*, Unsupervised machine learning for networking: Techniques, applications and research challenges, pp. 1-3.

<sup>68</sup> Cp. *Fergus et al.*, A visual category filter for google images., 2004, pp. 1-6.

<sup>69</sup> Cp. *Zhai et al.*, S4I: Self-supervised semi-supervised learning, 2019, pp. 1-3.

<sup>70</sup> Cf. *Sutton, Richard S. et al.*, Reinforcement learning, 1999, p. 20.

punishment. Negative reinforcement is used to correct specific behaviors, whereas positive reinforcement encourages long-term improvement in overall performance.<sup>71</sup>

Example: Self-driving cars

Clustering techniques are a group of undirected data mining tools. The purpose of undirected data mining is to discover the structure of the data. There is no dependent variable to be predicted, thus there is no difference between independent and dependent variables.<sup>72</sup> Clustering techniques are used for combining observed examples into clusters or groups that satisfy two main criteria:<sup>73</sup>

- Each cluster is homogeneous. The examples that belong to the same groups are similar to each other.
- Each cluster or group should be different from other clusters. So, the examples belong to one cluster, should be different from the other examples that belong to other clusters.

K-means clustering is the simplest algorithm of clustering based on the partitioning principle.<sup>74</sup> It works on numeric attributes. There are several steps to calculate the k-mean algorithm.<sup>75</sup>

1. Define the data points as  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  and each data points are a n-dimensional real vector. The total number of data points are N.
2. Determine the number of clusters k.
3. Generate k random points as cluster centroids and define the set of centroids as  $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_k\}$ .
4. Assign each data point to the cluster where the Euclidean distance between the centroid of that cluster and the data points is the minimum.

#### **Formula 1: Minimum Euclidean distance**

$$\mathbf{x}_i^j = \min\{\mathbf{x}_i - \mathbf{c}_j\}$$

Source: Author Generated

$\mathbf{x}_i^j$  means the data point  $\mathbf{x}_i$  is assigned to the centroid  $\mathbf{c}_j$ .

---

<sup>71</sup> Cf. Ibid, pp. 20-24

<sup>72</sup> Cf. *Hand D. et al.*, Principles of Data Mining, 2001, no page number.

<sup>73</sup> Cf. *Aggelis, Vasilis, et.al.*, Customer clustering using rfm analysis, 2005, pp.2.

<sup>74</sup> Cf. *Kansal, Tushar. Et.al.*, Customer segmentation using k-means clustering, 2018, pp.135-139.

<sup>75</sup> Cf. *Maryani, I. et al.*, Customer Segmentation and Clustering Techniques, 2018, pp. 1-6.

5. Recalculate the centroids. The centroid is now the mean of the data points in that cluster.<sup>76</sup>

**Formula 2: Mean of the data points**

$$c_j = \frac{1}{M} \sqrt{\sum_{i=1}^M (x_i)}$$

Source: Author generated

M is the number of data points in a cluster.

6. Continuously do steps 4 and 5 until there is no change in the centroids. Then the clustering is optimized.

A cluster centroid is the average of all the points in the cluster.<sup>77</sup> The optimal number of clusters can be found by different methods. One of them is the Elbow method. The Elbow method can calculate the optimal number of clusters. The elbow method is based on the mean squared error between a data point and its centroid. The elbow method has also several steps.

1. Using the K-means method, calculate the clusters k.
2. For every value, k computes the mean squared error (MSE) from every data point ( $x_i$ ) to its centroid ( $c_j$ ).

**Formula 3: Mean squared error**

$$MSE = \frac{1}{N} \sum_{i=1}^N (x_i - c_j)^2$$

Source: Author generated

Where N is the number of data points in a cluster.

3. Add the MSE of the clusters together. This is called total mean squared error.
4. Plot the k-number of clusters against the total mean squared error.
5. The optimal number of clusters is where the graph bends the most in the plot.

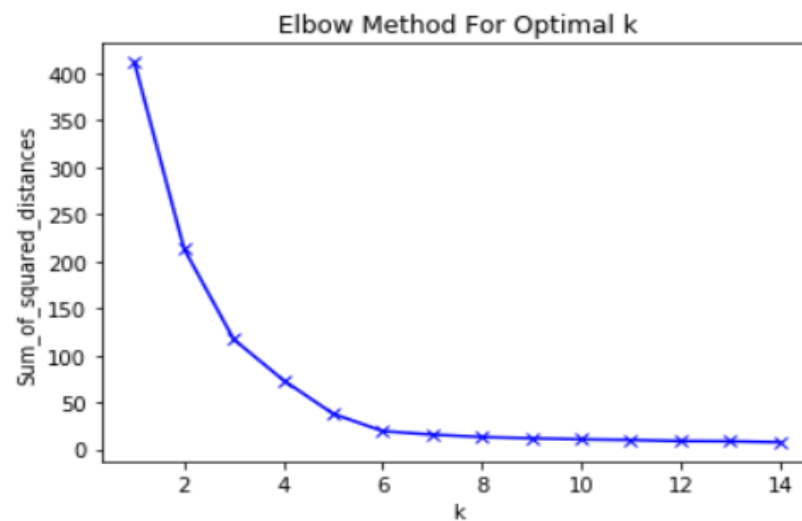
---

<sup>76</sup> Cf. *Shakur, M A. et al.*, Integration K-Means Clustering Method and Elbow Method, 2018, no page number.

<sup>77</sup> Cf. *Sheshasaayee. et al.*, Implementation of clustering technique, 2019, pp. 1166-1170.

Elbow methods generated an elbow curve plot. If the data is very large, it will take a lot of time to calculate the optimal number of clusters with all data. So, random sample data points are used.<sup>78</sup> The points where the elbow is formed are taken as the optimal value of k.

**Figure 2: Elbow method representing the number of clusters**



Source: Syakur, M.A. et.al., Integration k-means clustering method and elbow method, 2018, pp.012017.

In figure 2, a clear bent can be seen in three clusters. Hence the optimal number of clusters is three.<sup>79</sup>

### 3. Recency, Frequency, and Monetary (RFM) Analytic Model:

Stone and Bob first suggested the idea of the RFM method in 1989. The Recency, Frequency, and Monetary model is a three-dimensional analysis tool that helps an organization identify its valuable customers based on their certain characteristics. This model is based on the purchasing history of the customer. The three dimensions mentioned in the RFM analysis are introduced in the following table.

<sup>78</sup> Cf. Burg, J.M., Customer segmentation using RFM analysis, 2020, pp. 1-48.

<sup>79</sup> Cf. Patil, S. et al., Customer Segmentation using K-Means clustering and RFM modelling, 2021, pp. 556-559.

**Table 1: Dimensions of the RFM Model**

Dimension	Description
Recency	The duration since the date of the last transaction.
Frequency	The total count of purchases.
Monetary	The average amount purchased.

Source: Own representation

There are advantages and disadvantages of this analytic model.

Advantages:

- RFM analysis can be done with the least set of variables. So, this analysis is cost-effective in data storage and data collection.
- This analysis is very fast and easy to implement and understand.
- This is very useful for short-term business marketing plans.

Disadvantages:

- Some of these variables are not too helpful to decision-making.
- This analysis is done based on historical data. So, this analysis is less helpful for new customers.<sup>80</sup>

### **3.1. Definition of RFM analysis:**

RFM analysis is a three-dimensional marketing technique used to rank and group customers based on their prior purchasing history such as how recently a customer has purchased, how often the customer purchases, and how much the customer spends. RFM analysis is based on the 80/20 principle that 20% of customers bring in 80% of the revenue of a company. The company should focus resources on key accounts and key products. 20% of customers who are buying 20% products equal to 4% of all customer transactions. Thus, the company can offer the highest level of service to key customers. On other hand, the company can determine to give up or pay less attention to those customers who are making less profit for the company.<sup>81</sup>

<sup>80</sup> Cf. *Hshan, T*, Exploring customer segmentation, 2020, <https://medium.com/swlh/exploring-customers-segmentation-with-rfm-analysis-and-k-means-clustering-93aa4c79f7a7>, Accessed on 25.06.2022.

<sup>81</sup> Cf. *Birant, Derya.*, Data mining using RFM analysis, 2011, pp.91-108.



The philosophy of RFM analysis is that products are to be arrayed in terms of recency (recent sales), frequency (frequent sales), and monetary (total money spent) factors. For recency, the real dataset with sales date converted to values 1 to 5 based on the date of sale. Therefore, the value of 5 is assigned to the top 20% of the data set in terms of sales date. The next 20% refers to the value of 4 and so on, while the oldest sale date refers to the value of 1. For frequency, the number of transactions in a certain period is sorted in descending order. The top 20% of the data is assigned to the value of 5, 20% of the next data is assigned to the value of 4, and so on. For monetary, the average amount of money spent per year or month for all transactions is sorted in descending order. As many as 20% of the top data from the data set are given a value of 5. The next 20% of data from the data set is assigned to the value of 4 and so on. Finally, all values of recency, frequency, and monetary are combined to rank each product.<sup>82</sup>

### **3.2. Recency, Frequency, and Monetary Analysis for Customer Segmentation:**

RFM analysis model assigns numerical scores to each customer based on some factors like Recency, Frequency, and Monetary. RFM analysis groups each customer based on the following factors:<sup>83</sup>

- Recency: How recent the customer's last purchase was? Recency is the number of days a customer takes between two purchases. The smaller value defines that the customer visits the store regularly in a short period of time and the greater value implies that the customer visits the store shortly<sup>84</sup>. Businesses generally measure recency in days but, depending on the product, they may measure recency in years, months, or hours also.
- Frequency: How often did a particular customer make a purchase in a certain period? Customers who purchased an item once, are likely to purchase it again. The higher value of frequency defines that the customer is more loyal.

---

<sup>82</sup> Cf. Wu, Hsin-Hung *et al.*, Applying RFM model and K-means method in customer value, 2009, pp. 665-672.

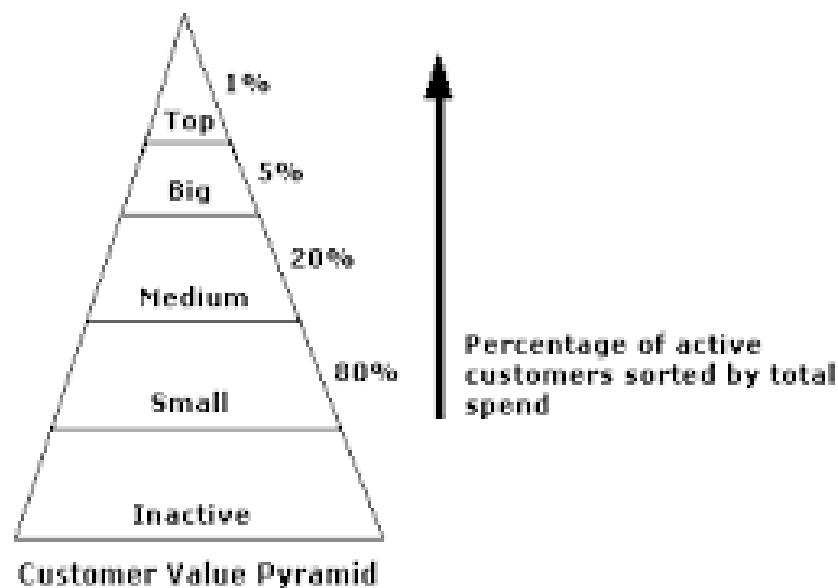
<sup>83</sup> Cf. Wright, Gavin., RFM analysis, <https://www.techtarget.com/searchdatamanagement/definition/RFM-analysis#:~:text=RFM%20analysis%20is%20a%20marketing,and%20perform%20targeted%20marketing%20campaigns.,> Accessed on 24.06.2022.

<sup>84</sup> Cf. Christy, A., *et.al.*, RFM ranking, 2021, pp.1251-1257.

- **Monetary:** How much money did the customer spend in a certain period? Customers who spend a lot of money to buy any items are more likely to spend money in the future. The higher value of monetary spent the more revenue they give to the company.

Vasilis studied the RFM value of active e-banking users. The author used clustering techniques to organize observed examples into groups based on the pyramid model which is shown following. The pyramid model is very useful to companies, banks, and financial organizations. There are some usages of a pyramid model such as decision making, customer profitability, future revenue forecasting, prediction of the alteration of customers' position in the pyramid, supervision of the most important customers, and stimulation of inactive customers. K-means algorithm and two-step clustering method were selected as clustering algorithms in this study. RFM analysis suggests that the customer, who holds a high RFM score, normally conducts more transactions and results in a higher profit for the company. Hence, they provided the results for the bank to identify the most important users-customers.<sup>85</sup>

**Figure 3: Pyramid model**



Source: Author own representation

<sup>85</sup> Cf. Aggelis, Vasilis, *et.al.*, Customer clustering using rfm analysis, 2005, pp.2.

**3.3. Steps of RFM Model:** These are the steps involved in conducting RFM analysis and these steps below provide an overview of RFM analysis and how segmentation is executed:<sup>86</sup>

- i. **Build RFM Model:** The first step is to assign recency, frequency, and monetary values to each customer. The raw data, which is a customer database from past transactions, is then compiled in a spreadsheet or database. Recency is the amount of time since the customer's most recent transaction, Frequency is the total number of transactions made by the customer during a defined period and Monetary is the total amount of money that the customer has spent.
- ii. **Divide the customer segmentation:** The second step is to divide the customer into tiered groups for each of the three dimensions (R, F, and M). Tier selection is based on the greatest to the least (for example for monetary value, tier one is assigned to the highest amount and tier five is assigned to the lowest amount of money).
- iii. **Select the targeted customers' groups:** The third step is to select a segmented customer group with high customer value. It is helpful to assign names to segments of interest such as Platinum, Gold, Silver, and Bronze levels of customers.
- iv. **Employment of a personalized marketing strategy:** The final step is to employ a personalized marketing strategy designed for each RFM segment focused on their behavioral patterns. The RFM analysis allows marketers to communicate with customers in a very way more effective manner.

**3.4. Relationship between RFM analysis and small and medium-sized enterprises:**

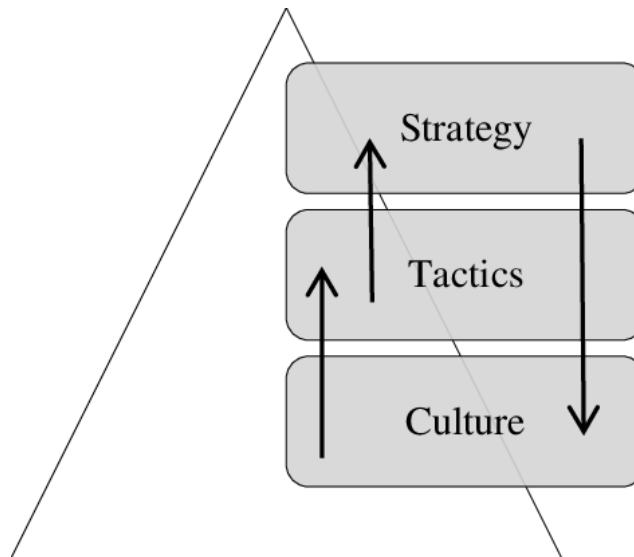
Small and medium-sized organizations can gain a competitive advantage and a sustainable business by adopting the appropriate tools. An appropriate database (DB) can bring a competitive advantage to an organization over its competitors. So, this is an essential factor of success in today's competitive global market. A key factor in SMEs (Small and medium-sized enterprises) is focused on retention and obtaining customers.

---

<sup>86</sup> Cf. *Greene, Ryan.*, What is RFM Analysis, <https://www.actioniq.com/blog/what-is-rfm-analysis/>, Accessed on 25.06.2022.

The basic principles of marketing are applicable to SMEs. Marketing in small enterprises can be classified into three-level (Culture, Strategy, and Tactics).<sup>87</sup>

**Figure 4: Categorization of marketing in SMEs**



Source: Author own presentation

Culture is the base of this pyramid, which represents the analysis of customer needs. Tactics is in the middle level in this pyramid, sustained by the analysis of 4Ps (Product, Price, Place, and Promotion) to impact the performance and growth of a business. The strategy that is at the highest level, promotes the development to increase actual and potential market position. RFM analysis is useful for small and medium-sized enterprises for many reasons:

- RFM is affordable: Businesses can apply the insights from RFM marketing analysis and increase their retention rate. Retention is 5-6 times cheaper than obtaining new customers. It is a magnificent way to increase the longevity of businesses.
- RFM analysis is very simple to use: RFM analysis is very simple to use for anyone. It does not require complex or advanced tools. There is no need to hire a data analyst to explain the data, because the principles are too easy to understand, and the results are easy to explain and act on.

<sup>87</sup> Cf. *Guarda, T. et al.*, Database marketing tools for SMEs, 2014, pp. 995-999.

- Effective in direct marketing: RFM analysis is very effective in direct marketing to platinum customers.
- Future prediction of customer behavior and minimizing marketing costs: The three variables in RFM analysis are considered powerful predictors of future behavior and are the basis of database marketing. Recency allows the prediction of future value. Frequency and Monetary value enable the evaluation of the current value. With a higher RFM score, it is more probable for a customer to respond to a marketing action. Then, the company may maximize the return on campaigns and minimize marketing costs.

#### **4. Empirical Investigation of customer segmentation:**

##### **4.1. Research Design:**

###### Step 1: Understanding business

Every step of business focuses on understanding the goal of needs based on business valuation. After understanding the business model, a data mining plan is designed to reach the goal. This is the study of an online retail E-commerce website. The transactions were made in the years 2010 and 2011.

###### Step 2: Data Understanding

After collecting the data, the whole data should be examined. In this paper, I am using online retail data which contains sales transactions from December 2010 to December 2011.

###### Step 3: Data Preparation

There are two steps in this phase. These are data preprocessing and data visualization. Missing values filling, removing errors, and dropping negative transactions are all done in this data life cycle.

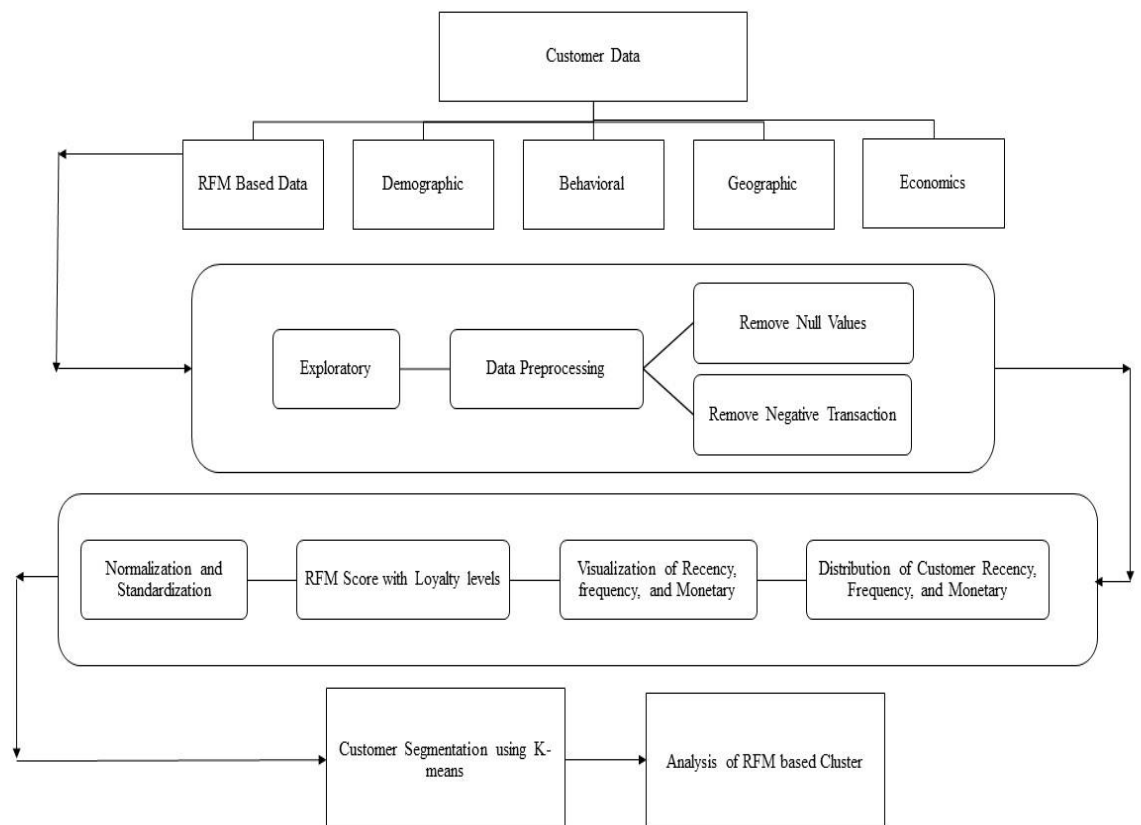
In the real world, data is not always proper. It may have incompleteness, inconsistency, and errors. If there are any kind of missing values or duplicate values, it is called uncleaned data. This uncleaned data effects the quality of the results. So, this raw data should be transformed into meaningful data and this technique is called data preprocessing.

The graphical representation of data is called data visualization. Visual elements like graphs, charts, and maps are used in data visualization. We can understand the data and analyze trends, patterns, and outliers in the data. When we have big data, data visualization can help an organization make the right data-driven decisions.

#### Step 4: Data Modelling and Evaluation

RFM is an analysis tools to identify the best customers and make the best marketing strategies for them. RFM model ranks customers in certain categories. I am using the k-means algorithm for grouping data into a certain number of clusters. Evaluation is one of the major factors in research design. The optimal number of clusters is determined by the elbow method, which is explained later in this paper.

**Figure 5: Research Design**



Source: Own Representation

#### **4.1.1. Research Purpose:**

The research purpose is to indicate what should be achieved by leading research and how the result of the research can be used. It can be arranged by its goal such as exploratory, descriptive, predictive, and explanatory. Exploratory research may be a methodological approach that investigates research questions that haven't previously been studied. The goal of exploratory research is to find out patterns or hypotheses rather than testing a hypothesis. Exploratory research is generally conducted when a researcher has begun to investigate and wants to understand the topic.<sup>88</sup> This research can be done using literature search, surveys, and case studies. Descriptive research describes phenomena because it exists. Descriptive research distinguishes and achieves information accurately about a person or the characteristics of a particular issue. The process of data collection is often quantitative and then, statistical techniques are usually used to summarize the information. Descriptive research can be used when a problem is well structured and there is no investigation regarding the cause-effect relationship. Explanatory research is a prolongation or continuation of descriptive research. The researcher describes the characteristics, analysis them, and explains why or how something is happening. Thus, the goal of explanatory research is to understand events by searching for and analyzing the cause-effect relationships.<sup>89</sup> Predictive research is to predict similar conditions. The aim of predictive research is to summarize the analysis by forecasting certain events based on a hypothesis. This research recommends solutions or new ideas beyond identifying success or performance or outcomes.<sup>90</sup>

#### **4.1.2 Research Approaches:**

There are two main traditions of approaching scientific research: quantitative and qualitative.<sup>91</sup> Quantitative research is a systematic examination of phenomena by gathering quantifiable data and performing statistical, mathematical, or computational techniques. The researcher collects information from existing and potential customers using many methods such as sampling online surveys, and questionnaires. The results can

---

<sup>88</sup> Cf. *DeCarlo, M.*, Scientific inquiry in social work, 2018, no page number.

<sup>89</sup> Cf. *Sebunje, William.*, Research Techniques, 2015, pp. 1-7.

<sup>90</sup> Cf. *Soudagar, Rana.*, Customer Segmentation and Strategy Definition in Segments: Case Study: An Internet Service Provider in Iran, 2012. pp. 1-80.

<sup>91</sup> Cf. *Madani, Samira.*, Mining changes in customer purchasing-behavior, 2009, no page number.

be portrayed in the form of numerical and the researcher can predict the future of a product or service based on the careful understanding of the results.<sup>92</sup> This research must be conducted depending on the data type and the characteristics of the gathered information. It is related to the purpose of the study and research questions. Quantitative research deals with numbers, logic, and objective. This research approach starts with a theory of general statement suggesting a general relationship between variables. This research focuses on measuring collecting and analyzing numerical data and applying statistical tests. Qualitative research requires collecting non-numerical data such as text, audio, or video, and analyzing it to understand concepts, opinions, or experiences. It can be used to generate new ideas for research or gather insights into a problem. Qualitative research is commonly used in social science and humanities. There are many types of qualitative research methods such as case study research, one-to-one interview, focus group, record keeping, ethnographic research, and qualitative observation. The below table describes the differences between qualitative research and quantitative research.<sup>93</sup>

**Table 2: Differences between qualitative and quantitative research methods.**

Attributes	Qualitative research	Quantitative research
Analytical purpose	This research method is focused on describing individual experiences.	This research method is focused on describing the characteristics of the population.
Types of questions asked	One-ended questions	Close-ended questions
Data collection	Data are collected through participant observation, focus groups, and interviews.	Data are collected through measuring things such as surveys and questionnaires.
Form of data	Descriptive data	Numerical data

<sup>92</sup> Cf. *Watson, Roger.*, Quantitative research, 2014, pp.44.

<sup>93</sup> Cf. *McLeod, Saul.*, what's the difference between qualitative and quantitative research, 2019, <https://www.simplypsychology.org/qualitative-quantitative.html>, Accessed on 01.07.2022.



Report	Data are reported in the language of an informant.	Data are reported through statistical analyses.
--------	--	---

Source: Author own representation

#### **4.1.3 Research Strategy:**

The research strategy will be the general approach of how researchers are going to answer the research questions. It describes the source from which researchers collect data and consider time, money, and location. It will contain a clear objective come from research questions. There are five research strategies in social science, i.e. experiment, survey, analysis, history, archival, and case study. The case study is a common strategy in business research, allowing the researcher to retain the holistic characteristics of real-life events while investigating empirical events. It is based on an in-depth investigation of an individual, group, or event.<sup>94</sup>

#### **4.1.4 Research Instrument:**

The research instrument is a tool used to collect, measure, and analyze data related to the researcher's research interests. Research instruments are often employed in the fields of social sciences and health sciences. A research instrument can include interviews, surveys, tests, or checklists. A good research instrument has been validated and has proven reliable. It should be able to assist in answering the research aims, objectives, and research questions and should not have any bias in the way the data is collected. The types of research instruments will depend on the format of the performed research study: qualitative, quantitative, or mixed method.

#### **4.2 Data Collection:**

Data are categorized as primary data and secondary data. Primary data are collected through interviews, observation, and questionnaires. Secondary data are collected from secondary sources such as personal records, publications, and censuses. The data needed to perform customer segmentation in this thesis were taken from Kaggle. An online retail dataset from Kaggle is used as a collected dataset in this thesis. This online retail dataset has different columns such as Invoice No, Stock Code, Description, quantity, Invoice date, Unit Price, Customer Id, and Country.

---

<sup>94</sup> Cf. *Oliva, Rogelio.*, Intervention as a research strategy, 2019, pp. 710-724.

**Figure 6: Online Retail Dataset**

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84408B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Source: Own representation

The attribute information is described below:

- InvoiceNo: This is the invoice number which is nominal. A 6-digit integer is uniquely assigned to each transaction. If it starts with the letter 'c', it means a cancellation.
- StockCode: This is a product code that is nominal. A 5-digit integer is uniquely assigned to each distinct product.
- Description: This is the product name which is nominal.
- Quantity: The quantities of each product per transaction. It has numeric values.
- InvoiceDate: This is the invoice date and time which is numeric. It indicates the day and time when a transaction was generated.
- Unit Price: This is the product price per selling. It has a numeric value.
- CustomerId: This is a customer number that is nominal. A 5-digit integer is uniquely assigned to each customer.
- Country: This is the country name that is also nominal. This indicates the name of the country where a customer resides.

### 4.3 Data Preprocessing:

As raw data are imperfect and noisy, these data are not appropriate for the data mining process. It must be cleaned data. So, it is necessary to perform preprocessing. This preprocessing includes data cleaning, integration, and transformation.

Data Cleaning and Integration: Data cleaning is one of the most important steps in the data mining process. It can be time-consuming and frustrating sometimes, but it is very essential for quantitative research. The research can show weakness if this phase of the

project does not be considered as important as other phases. In this phase, errors must be checked, detecting missing values and filling them, removing the useless attributes and abnormal items must be checked. As missing values will affect the analysis, null values should be removed. In this dataset, two columns are containing the missing value i.e., Description and CustomerID. Description columns contain 1454 missing values and the CustomerID column contains 133600 missing values. The below figure shows the number of missing values in each column.

**Figure 7: Each column containing the number of missing values**

InvoiceNo	0
StockCode	0
Description	1454
Quantity	0
InvoiceDate	0
UnitPrice	0
CustomerID	133600
Country	0
dtype:	int64

Source: Own representation

So, these missing values are removed now from the column and non-missing values are selected for the next step. In this dataset, the Country column has so many values. The Country column contains a value of the United Kingdom at 91.43%.

**Figure 8: Customer distribution by country**

United Kingdom	0.914320
Germany	0.017521
France	0.015790
EIRE	0.015124
Spain	0.004674
Netherlands	0.004375
Belgium	0.003818
Switzerland	0.003694
Portugal	0.002803
Australia	0.002323
Norway	0.002004
Italy	0.001482
Channel Islands	0.001399
Finland	0.001283
Cyprus	0.001148
Sweden	0.000853
Unspecified	0.000823
Austria	0.000740
Denmark	0.000718
Japan	0.000661
Poland	0.000629
Israel	0.000548
USA	0.000537
Hong Kong	0.000531
Singapore	0.000423
Iceland	0.000336
Canada	0.000279
Greece	0.000269
Malta	0.000234
United Arab Emirates	0.000125
European Community	0.000113
RSA	0.000107
Lebanon	0.000083
Lithuania	0.000065
Brazil	0.000059
Czech Republic	0.000055
Bahrain	0.000035
Saudi Arabia	0.000018

Name: Country, dtype: float64

Source: Own representation

So, the rest of the values must be removed from the Country column. Now United Kingdom data will be kept only. There should be no negative values in Quantity columns because Quantity can't be negative values. In this dataset, there is a negative value in the Quantity column. These negative values are removed from the Quantity column.

Data Transformation: In this step, string variables must be converted into numeric categorical variables and some codes must be replaced by text. There are two other tasks

performed in this phase. These are data aggregation and data generalization. Total purchase data of a customer in a period must be aggregated for performing a successive process. Low-level data will be replaced by high-level data in data generalization. The Invoice date column has string values. So, the string value is converted to the Date Time value.

Finally, a new column TotalAmount is created by multiplying Quantity and UnitPrice. Now, after all these phases, the dataset contains 354345 rows and 9 columns. Now the dataset looks like the following figure:

**Figure 9: Dataset after preprocessing**

	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

Source: Own representation

## 4.4 Empirical Finding

### 4.4.1 Exploratory Data Analysis :

This dataset has 541909 rows and 8 columns. Describe() method generates descriptive statistics and it is used to summarize the data. The method is used to get min, max, sum, count, mean, standard deviation, and quantile values from the data frame. The below figure is the summary statistics of the dataset.

**Figure 10: Summary statistics of the retail dataset**

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Source: Own representation

Here the Quantity column has a negative value, which does not make sense. So, it should be cleaned later.

The info() method gets an overview of the data like the number of records present in the data and the number of columns and the data type of column. It gives information about the data frame. Below the figure is the information on the data frame.

**Figure 11: Information of the data frame**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype
---  -
0   InvoiceNo       541909 non-null object
1   StockCode      541909 non-null object
2   Description    540455 non-null object
3   Quantity       541909 non-null int64
4   InvoiceDate    541909 non-null object
5   UnitPrice      541909 non-null float64
6   CustomerID     406829 non-null float64
7   Country        541909 non-null object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Source: Own representation

In this dataset, the data type of InvoiceNo, StockCode, Description, InvoiceDate, and Country are objects. The data type of UnitPrice and CustomerID is float and the data type of Quantity is an integer. The above figure describes all the information about this data frame.

#### 4.4.2 Customer Segmentation Based on RFM model

RFM metric plays an important role to understand the behavior of the customer. Recency affects retention and Frequency and monetary value affect a customer's lifetime value. Here, RFM analysis and K-means clustering techniques are used to identify the best customers and their segmentation based on their purchasing behavior. The Frequency of customers is calculated by counting the Invoice numbers of each customer. So, more frequency means the customer buys more often from the store. The subtraction between the very recent date and the transaction date of the customers is Recency. Summing up all the amount of the customer is monetary value. The below figure shows the recency, frequency, and monetary value of the customer.

**Figure 12: RFM value**

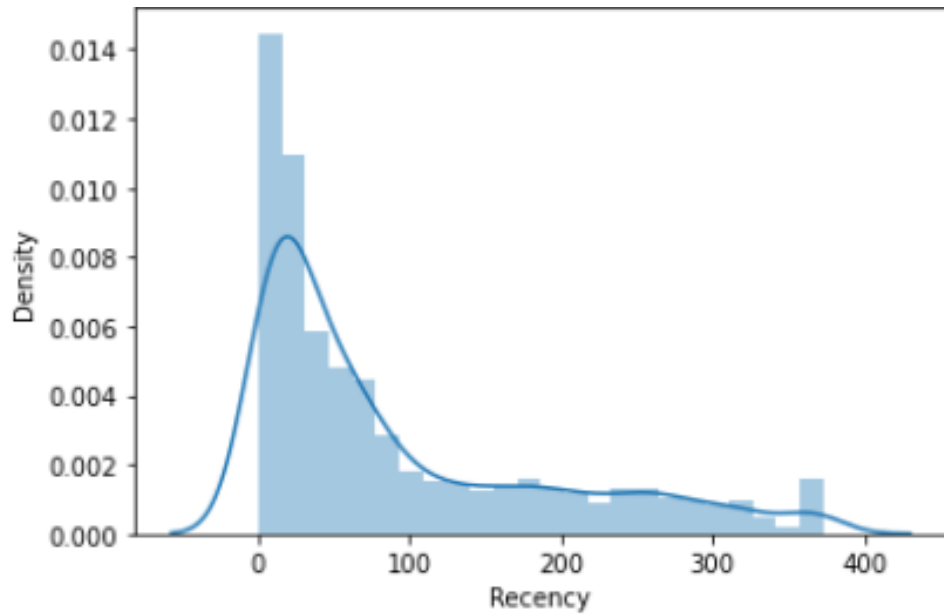
	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4596	33719.73
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34

Source: Own representation

From the above for the first row, the recency of the customer is too high, and frequency is only 1 but there is a huge monetary value. So, we can say that at that there was an offer or sales but after that, the customer never came again. It may be because of dissatisfaction

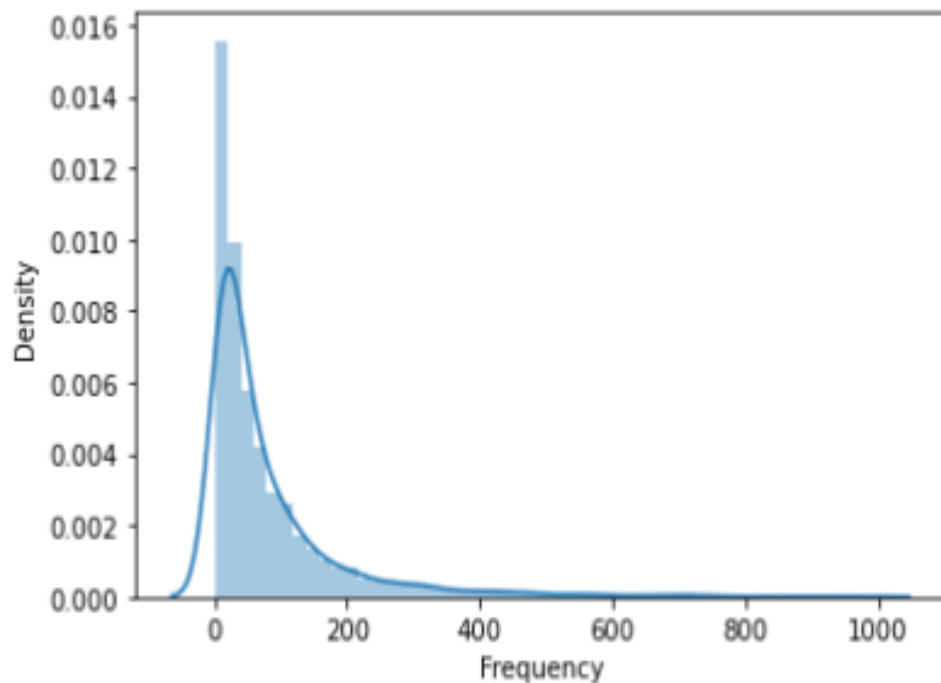
with using the product. Recency, Frequency, and Monetary values are calculated by grouping them with Customer Id. Here Distribution plots of recency, frequency, and monetary are drawn with the help of the seaborn library as shown below:

**Figure 13: Distribution plot of Recency**



Source: Own representation

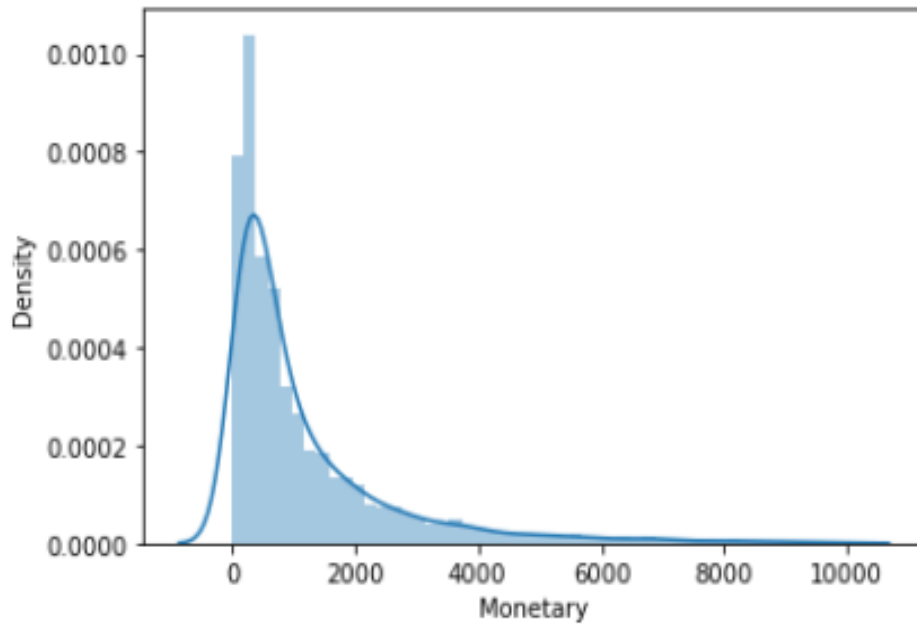
**Figure 14: Distribution plot of Frequency**





Source: Own representation

**Figure 15: Distribution plot of Monetary**



Source: Own representation

Many statistical analyses assume that the data have a normal distribution or Gaussian distribution. When the data do not have a normal distribution, they need to be transformed to make them more normal.<sup>95</sup>

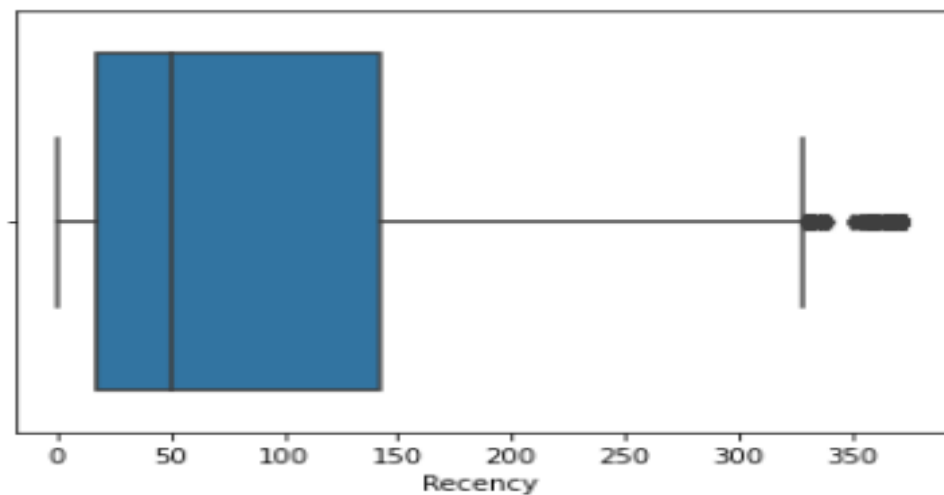
Skewness defines the asymmetry of a distribution. A skewed distribution happens when one tail is longer than the other one. Symmetry means one-half of the distribution reflects the opposite half. A normal distribution is a symmetric distribution with a bell-shaped curve. A normal distribution has a central peak where most observations occur, and the probability of events distributes equally in both the positive and negative directions on the X-axis. Both sides have an equal number of observations. Exceptional values are equally distributed in both tails. A left-skewed distribution has a long-left tail. Left-skewed distribution is also called negatively skewed distribution because there is a long tail in the negative direction on the X-axis. The mean is also to the left side of the peak. A right-skewed distribution has a long-right tail. Right-skewed distributions are also called positive-skew distributions because there is a long tail in the positive direction on

<sup>95</sup> Cf. Altman, Douglas G., et. al., Detecting skewness from summary information, 1996, pp. 1200.

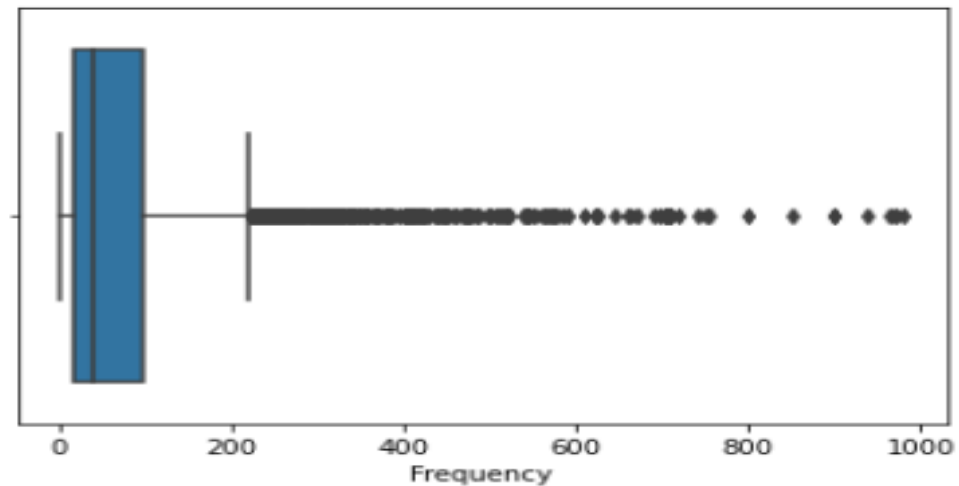
the X-axis. The mean is also to the right side of the peak. All these distribution plots do not follow a normal distribution. All these distribution plots are right-skewed. So, these distributions need to be normalized before developing the model.

In the box plot, a symmetrical distribution occurs when the box centers around the median and the right and left whiskers have approximately equal length. Right skewed distribution occurs when the median is closer to the box's left values and the right whisker is longer than the left one. Left skewed distribution occurs when the median is closer to the box's right values and the left whisker is longer than the left one. Here Box plots of recency, frequency, and monetary are drawn with the help of the seaborn library as shown below:

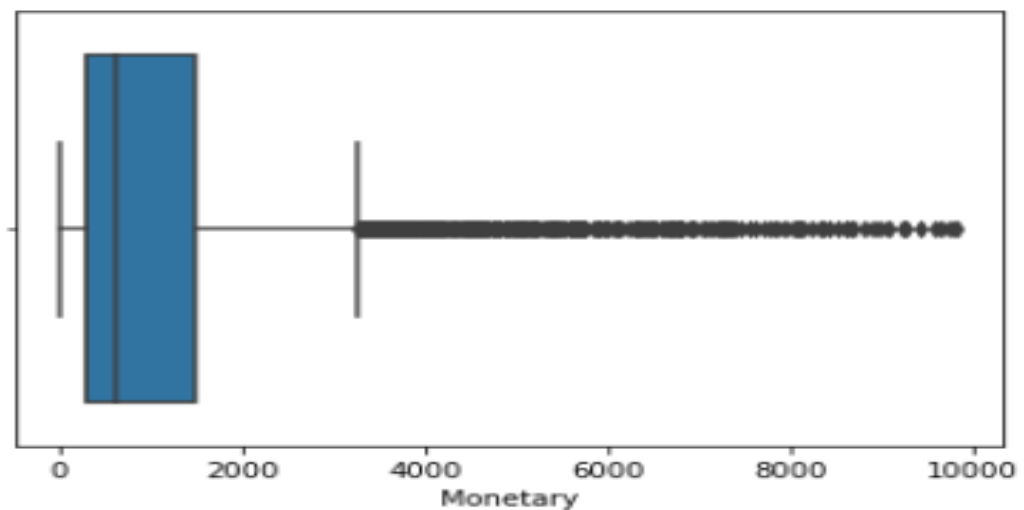
**Figure 16: Box plot of Recency**



Source: Own representation

**Figure 17: Box plot of Frequency**

Source: Own representation

**Figure 18: Box plot of Monetary**

Source: Own representation

Here all the above box plots are right skewed, and these all-black dots are outliers. It can affect accurate predictions. I am creating the quantiles (0.25,0.50,0.75) so that I can subdivide the entire dataset into four groups based on recency frequency monetary values. I also create two functions RScoring and FnMScoring to create segments that will be directed by values 1,2,3, and 4. In RScoring function, I am assigning value 1 to the lowest value of recency because the lower value of recency means that the customer is more engaged with a specific brand. On the other hand, In FnMScoring function, I am assigning value 1 to the highest value of frequency and monetary because the higher value of

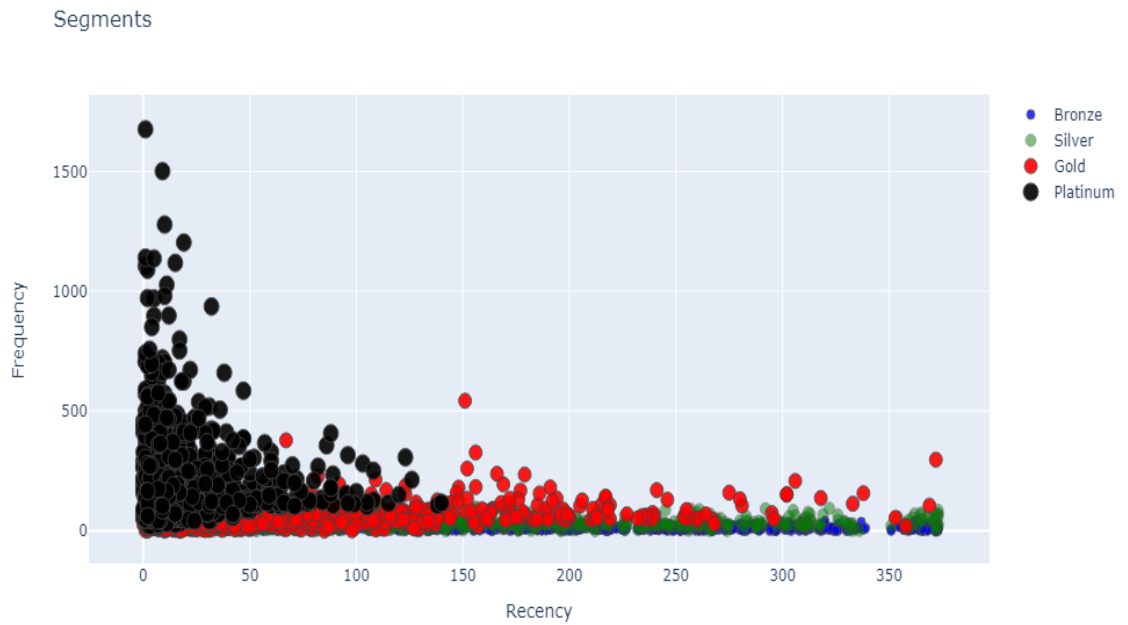
frequency and monetary means that the customer is more frequent and spends more money. we can give each buyer a RFM score based on his/her RFM data. I am scaling customers from 1-4. 64 different RFM scores ( $4*4*4$ ), ranging from 111 to 555, Each RFM cell will differ in size and vary from one another based on their key habits. Based on the RFM score, I am assigning the loyalty level (Platinum, Gold, Silver, and Bronze) to each customer. If a customer is in the Platinum group, it means that he is the most valuable customer and loyal and the company does not want to lose. The Bronze category is the one who has not purchased from the store for a quite long and he/she may be verge of churning out. I validated that customer with 111 RFM group has a loyalty level of platinum. The below figure is the top 5 data with RFM loyalty levels.

**Figure 19: Top 5 data with RFM loyalty level**

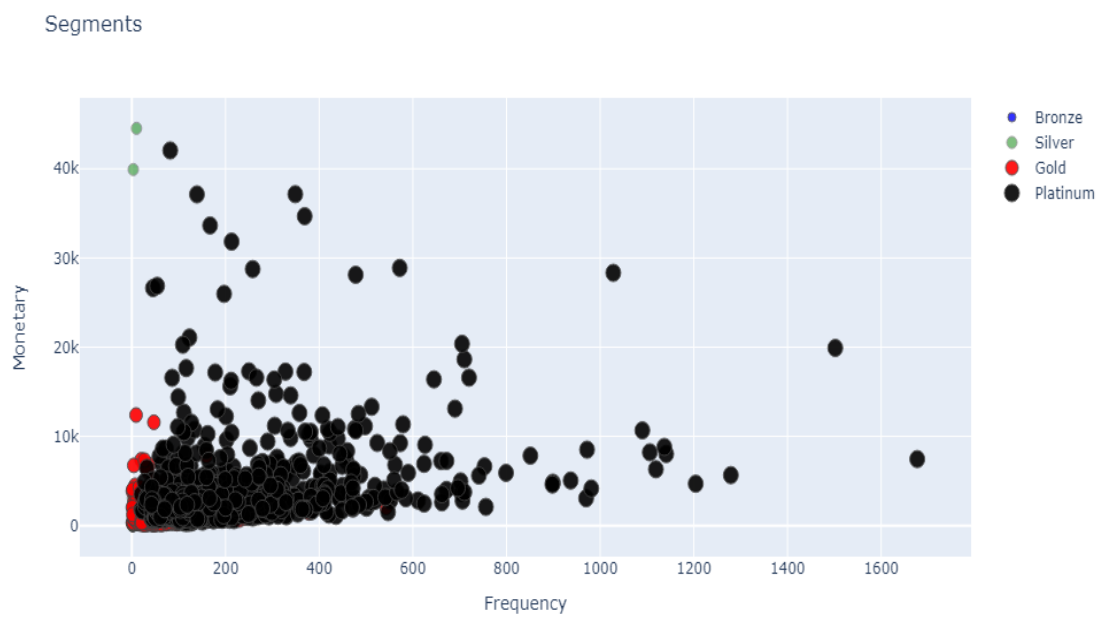
	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum
2	12748.0	0	4596	33719.73	1	1	1	111	3	Platinum
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum

Source: Own representation

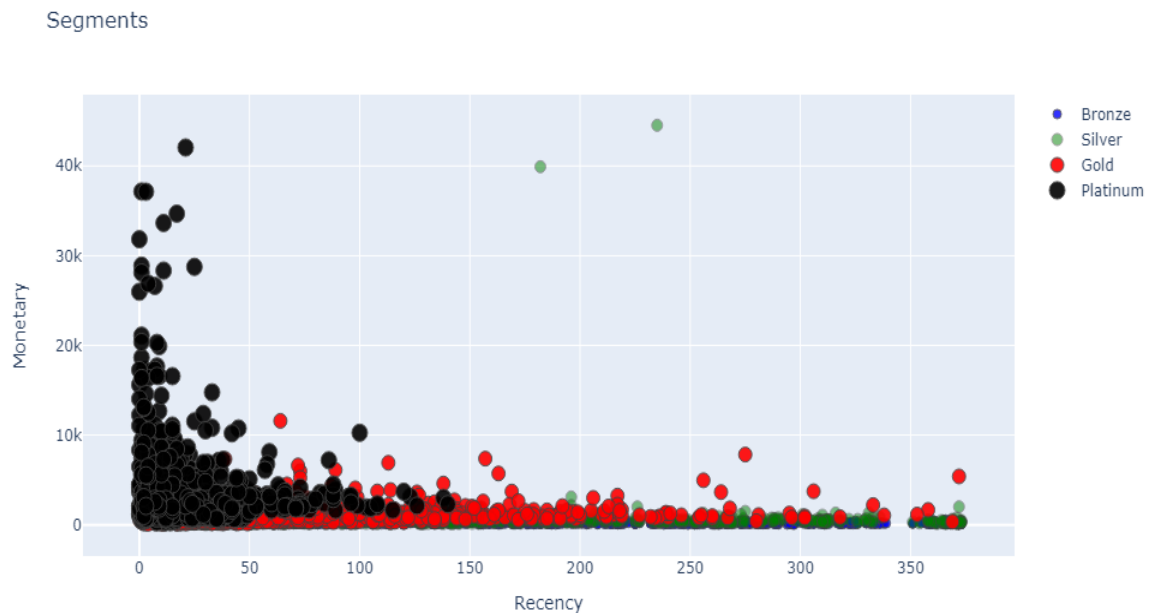
Here, I am going to plot three scatter plot graphs- Recency Vs Frequency, Frequency Vs Monetary, and Recency Vs Monetary. I choose a monetary of value less than 5000 and a frequency value of less than 2000 for these graphs. The below figures are the scatter plot of these graphs:

**Figure 20: Frequency Vs Recency**

Source: Own representation

**Figure 21: Monetary Vs Frequency**

Source: Own representation

**Figure 22: Monetary vs Recency**

Source: Own representation

#### 4.4.3 Data Clustering and Customer Segmentation:

The distribution of recency, frequency and, monetary is right-skewed, so I need to normalize it before developing the model. K-means clustering is an unsupervised clustering algorithm that makes clusters a group of data points based on the distance between the points. The data is divided based on the patterns in the data such that all the data points in a cluster should be like each other. I need to normalize and scale the data to create these clusters out of the data points. Clustering uses distances as a similarity factor hence I need to scale the data as well as normalize it if the data is right skewed or left skewed. I can make use of statistical techniques such as log transformation, Inverse transformation, and square root transformation to make the data more normal. Log transformation is employed when the information is right-skewed. Log transformation cannot be applied to zero or negative values. If log transformation is applied to zero, the results will be in a negative value. I am going to use log transformation for making this data normally distributed. The log transformation is applied in the recency and monetary columns. Recency values are ranging from 1 to 200 and Monetary values are ranging

from 100 to millions in the RFM score data frame. So, there is a need to bring the variables on the same scale to standardize them. I am scaling the values in the log-transformed data. Standardization is used to compare features that have different scales. This is done by subtracting the mean and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one. Here is the formula for standardization:

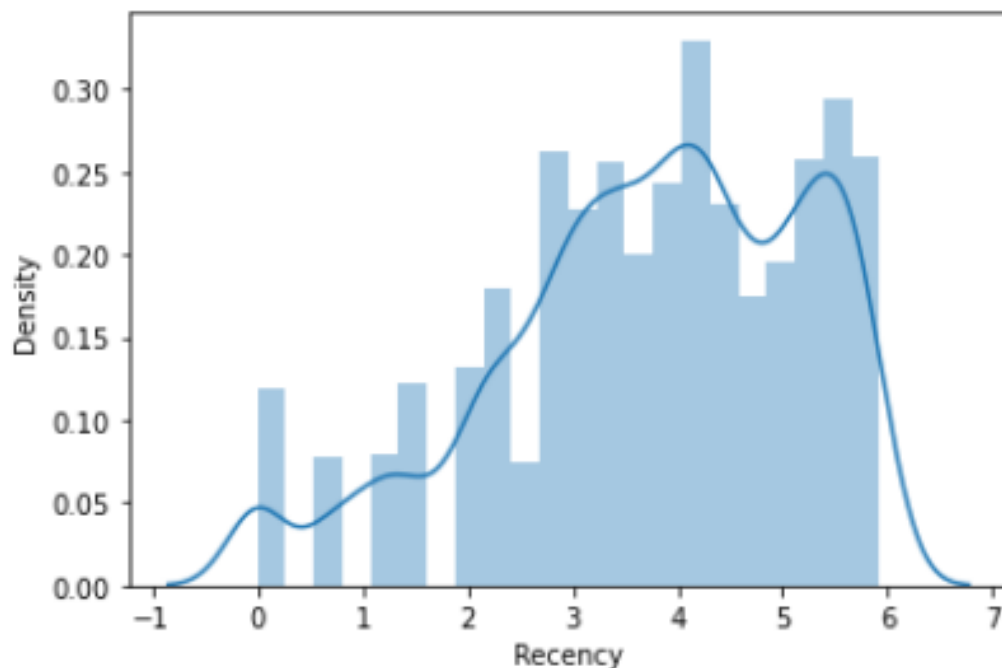
**Formula 4: Formula of Standardization**

$$Xi = \frac{X - \mu}{\sigma}$$

Source: Cf. Spiegel, Murray. Et al., Schaum's Outlines Statistics, 2008, no page number.

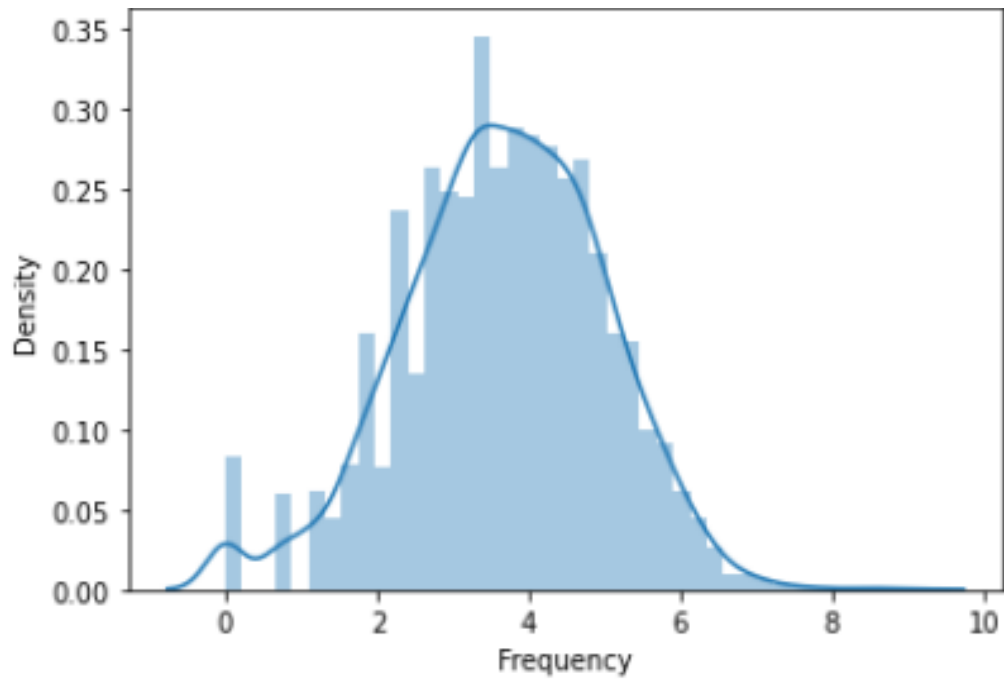
$\mu$  is the mean of the feature values and  $\sigma$  is the standard deviation of the feature value. Here the distribution of Recency, Frequency, and Monetary values, after normalization and standardization are shown below:

**Figure 23: The distribution of Recency after normalization and standardization**



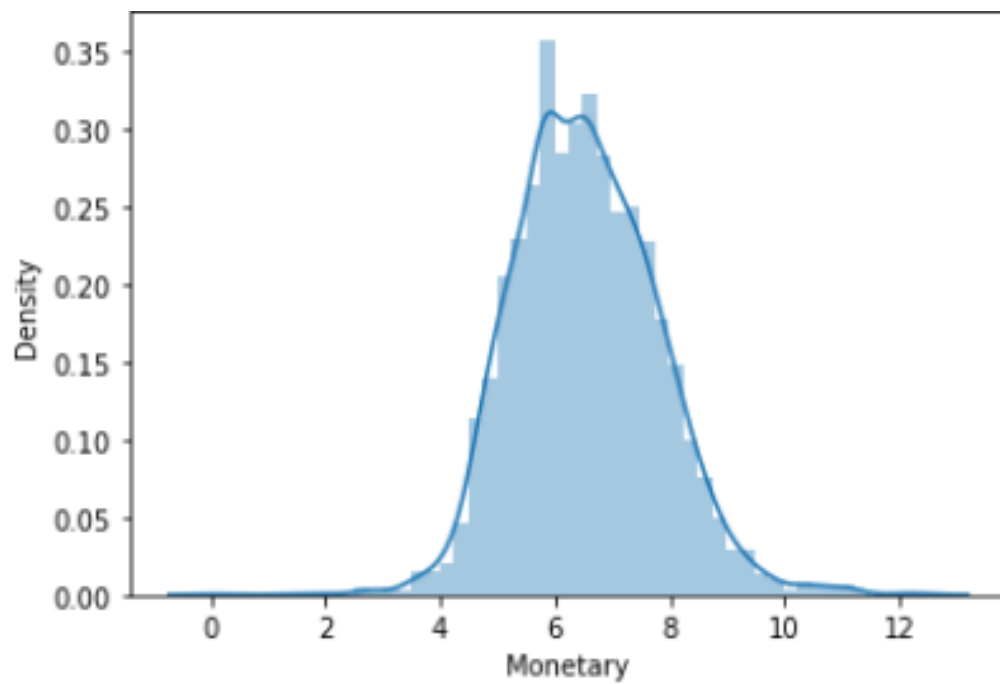
Source: Own representation

**Figure 24: The distribution of Frequency after normalization and standardization**



Source: Own representation

**Figure 25: The distribution of Monetary after normalization and standardization**



Source: Own representation

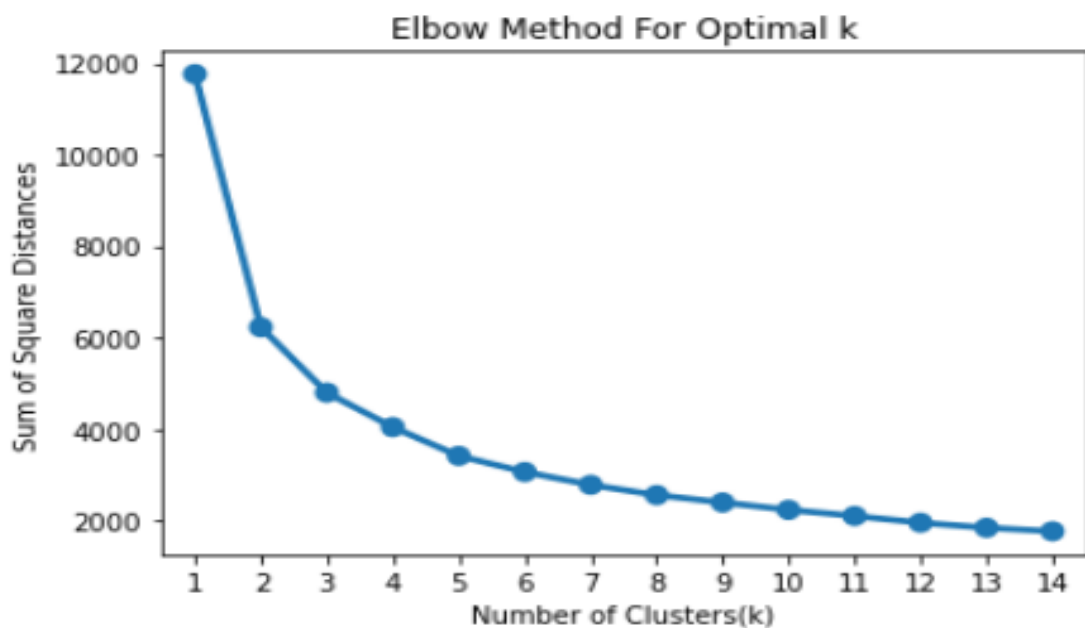


After normalization and standardization, the distribution of recency, frequency, and monetary values are close to normal distribution.

K-means clustering algorithm is an iterative algorithm that tries to partition the dataset into  $k$  distinct clusters. It tries to make intra-cluster data points as similar as possible while also keeping the clusters as far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the centroid of the cluster is at a minimum. If clusters have less variation, the more homogeneous the data points are within the same cluster.

I used the elbow method to decide the  $k$ -value for the  $k$ -means clustering. I take the value of  $k$  where there is a sudden decrease at a specific value of the sum of square distances.

**Figure 26: Elbow method for Optimal K**



Source: Own representation

$K$  dramatically decreases at  $k$  equals 3 of the elbow of this line. So, 3 is the optimal value of  $k$  in this case. Now I can build my model since I have found the number of clusters. Now I am developing a  $k$ -means cluster model by defining the number of clusters as 3 and the maximum iteration has been assigned as 1000. Now, I am applying the operation on scaled and normalized data.

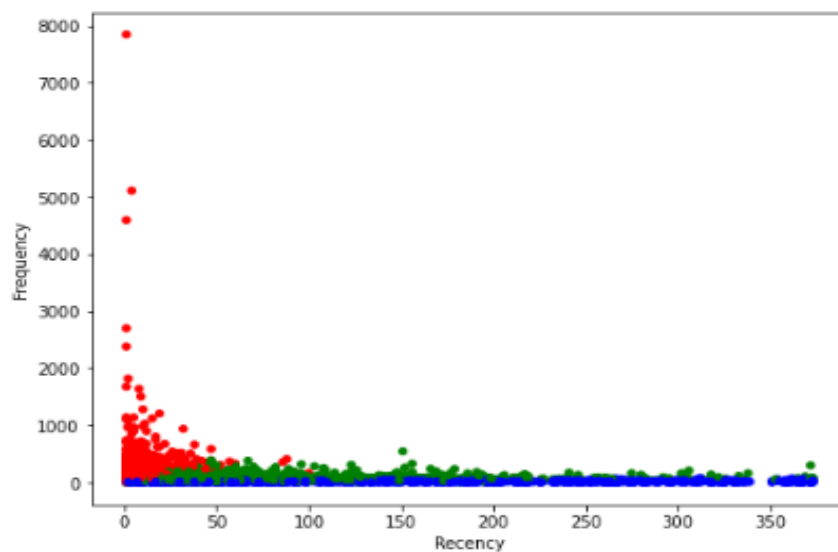
**Figure 27: Clusters of the customers**

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

Source: Own representation

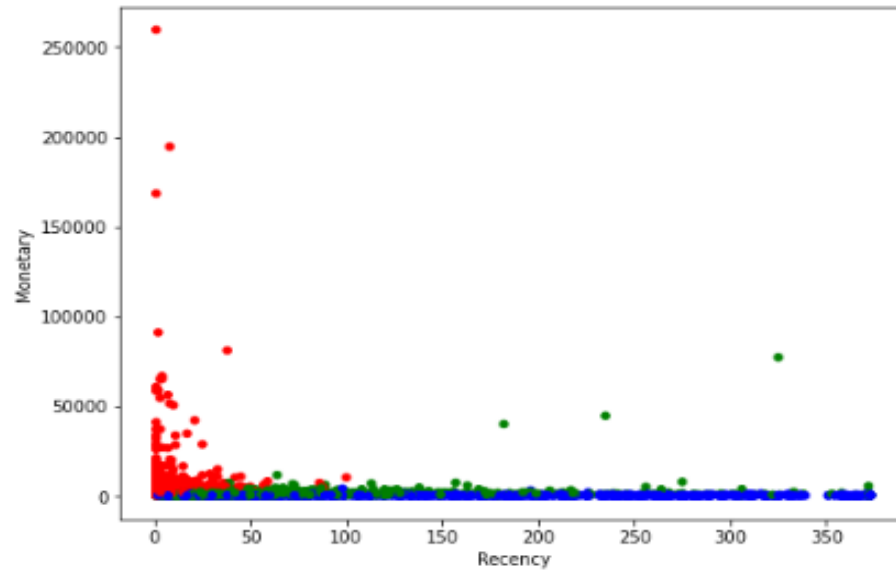
A cluster number is assigned for each customer. In this case, I have only 3 clusters and In RFM modeling, I have four clusters of platinum, gold, silver, and bronze. I can infer that in the case of k-means, out of four clusters, two clusters are merged and the remaining two are left as it is. Bronze and silver groups are merged into Group number 2.

Now, I am plotting the data points in the form of clusters using the matplotlib library. I am creating the scatter plot below as well such that access on the x-axis we want recency data and I want frequency data on the y-axis. I assigned three specific colors (Red, green, and blue) to three specific clusters. In this case, blue is assigned to silver and bronze, green is assigned to gold, and red is assigned to platinum.

**Figure 28: Frequency Vs Recency with three different clusters**

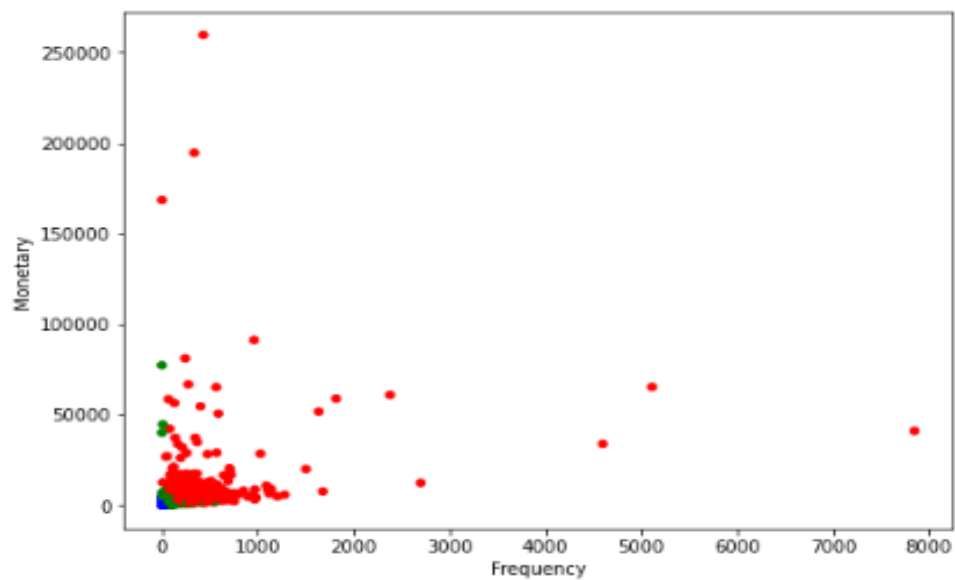
Source: Own representation

**Figure 29: Monetary Vs Recency with three different clusters**



Source: Own representation

**Figure 30: Monetary Vs Frequency with three different clusters**



Source: Own representation

#### 4.4.4 Strategy Definition Per Segment

There are 3 clusters (0, 1, and 2) in the K-means algorithm. Cluster 0 refers to Platinum, Cluster 1 refers to gold, and Cluster 2 refers to silver and bronze. The table shows the marketing strategy for each segmentation

**Table 3: Marketing strategy for each segment**

Cluster	Customer Type	RFM Characteristics	Marketing strategy
0	Platinum	They are recent and frequent shoppers. They spend huge money to buy products or services.	They are target customers. The company can try to cross-sell other products as well as they can be encouraged to sign up for a loyalty program to enjoy some elite experiences like free shipping, same-day shipping, priority access to newly launched products, etc.
1	Gold	They are average frequent and recent customers.	The company should emphasize Customer Relationship Management (CRM) to increase shopping

			experience and customized marketing plans encouraging purchases again
2	Silver and Bronzed	They have not purchased from the store for a quite long and they are low frequent shoppers. He/ She may be on the verge of churning out.	The company needs to figure out why they are on verge of churning out and the company may try to give rewards or coupons to trigger the spending from the almost churned out customer.

Source: Own representation

## **5. Discussion:**

### **5.1. Conclusion**

Customer segmentation is a method for grouping customers based on their similarities, whether it is customer needs, interest in certain products, channel preferences, etc. Customer segmentation and strategy definition per segment can provide enormous returns for companies.

This thesis aims to classify customer segmentation based on their purchasing history and analyze the segmentation of the customers.

In chapter 1 customer segmentation is introduced with the research problem, research questions, and research objectives.

In chapter 2, literature research is done. Customer segmentation is explained along with 4 types of segmentation (Demographic, Geographic, behavioral, and psychographic

segmentation). After this, K-means clustering is described with several steps to calculate the k-means algorithm.

In chapter 3, RFM analytical model is described in steps. The relationship between RFM analysis and small and medium-sized enterprises is described.

Research design with research purpose, approaches, strategy, and instrument are explained in chapter 4. Data collection and data preprocessing steps are also discussed in this chapter. Exploratory data analysis (summary statistics of the dataset and Information of the data frame) is done in the 4.4.1 subchapter. In 4.4.2, the RFM model is created, and the variable loyalty is added. The four groups of customers that came out of this model, are 'platinum', 'Gold', 'silver', and 'Bronze'. The skewness of recency, frequency, and monetary variable is handled in the next subchapter and the K-means clustering method is created using the scaled data. The three clusters (Clusters 0,1, and 2) of customers came out from this method. The conclusion is that silver and bronze customers are merged into cluster 2. The customer who is in cluster 0, are the best customers.

#### 1. How to apply RFM analysis to segment customers?

First, recency, frequency, and monetary values are assigned to each customer. Based on that a RFM score is created for each customer. The customers are divided into 4 tiered groups for each of the three dimensions (R, F, and M). At last Loyalty levels (platinum, gold, silver, and Bronze) are assigned to each customer.

#### 2. How many segments must be considered for segmentation using K-means clustering?

The Elbow method is used to calculate the optimal number of clusters. In this thesis, the optimal value of k is 3. So, three segments must be considered for customer segmentation using k-means clustering. The result of this research is 3921 customers. Out of 3921 customers, 1417 customers are in Cluster 0, 840 customers are in Cluster 1, and 1664 customers are in Cluster2.

#### 3. Which strategies must be defined for obtained segments?

Cluster 0 from the k-means model are the platinum customers. This is an interesting and biggest group for the company. This group of customers is the target customer, and the company can encourage them to enjoy elite experiences. Some special offers are to be provided to achieve the marketing target for the company.

Cluster 1 is the gold level customers who are average recent and average frequent customers. The company can plan some schemes to increase the shopping experience

Cluster 2 is the least active customer. Companies can avoid the customers or find out the reason why the customers are on the verge of churning out.

By knowing the categories of each customer, the company will be able to take the right decision in marketing strategy.

#### 4. What are the limitations and bugs of this thesis?

This thesis suffers from some limitations. In this thesis, the dataset covers 91.4 % of customers who are from the United Kingdom, So, I could this segmentation only for United Kingdom customers. I have limited and incomplete information about customers. Because of incomplete and limited data, I could do the basic Exploratory data analysis and basic segmentation. This same project can be implemented into real-time data, which will be very helpful for companies to determine their marketing strategies.

By being aware of customers' needs and placing priority on the most important customers, each department of the company will be able to quickly optimize the business. By tracking the behavior of each segment, the company will be able to calculate the ROI on changes that impact a particular segment.

### **5.2. Recommendation for Further Research:**

While this paper provides a step-by-step process for identifying and targeting the best customer segments, simply following it does not guarantee success. To be effective, the company must prepare and plan for various challenges in each step and always make sure to adopt new information or feedback that might change its output.

Future work includes studying the performance of the shoppers in each segment like the products which are bought frequently by the members of every segment. This can help better for the company in providing better promotional offers for specific products. Besides that, future work is on the payment method in each segment, customer profiles, and other general characteristics of every customer segmentation.

## Appendix I: RFM Analysis Code

### Importing Libraries:

```
In [1]: #Import necessary Libraries
%matplotlib inline
import numpy as np
import pandas as pd

#Visualization library
import matplotlib.pyplot as plt
import seaborn as sns

import datetime as dt

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

### Importing dataset in jupyter notebook

```
In [2]: #Import Online Retail Data containing transactions
data = pd.read_csv('data/retail.csv', encoding = 'unicode_escape')
data.head()
```

```
Out[2]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84408B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

### RFM Modelling:

```
In [22]: #Recency = Latest Date - Last Invoice Date, Frequency = count of invoice no. of transaction(s), Monetary = Sum of Total
#Amount for each customer
import datetime as dt

#Set Latest date 2011-12-10 as Last invoice date was 2011-12-09. This is to calculate the number of days from recent purchase
Latest_Date = dt.datetime(2011,12,10)

#Create RFM Modelling scores for each customer
RFMScores = data.groupby('CustomerID').agg({'InvoiceDate': lambda x: (Latest_Date - x.max()).days, 'InvoiceNo': lambda x: len(x)})

#Convert Invoice Date into type int
RFMScores['InvoiceDate'] = RFMScores['InvoiceDate'].astype(int)

#Rename column names to Recency, Frequency and Monetary
RFMScores.rename(columns={'InvoiceDate': 'Recency',
                          'InvoiceNo': 'Frequency',
                          'TotalAmount': 'Monetary'}, inplace=True)

RFMScores.reset_index().head()
```

### R, F and M Segmentation:

```
In [35]: #Functions to create R, F and M segments
def RScoring(x,p,d):
    if x <= d[p][0.25]:
        return 1
    elif x <= d[p][0.50]:
        return 2
    elif x <= d[p][0.75]:
        return 3
    else:
        return 4

def FmScoring(x,p,d):
    if x <= d[p][0.25]:
        return 4
    elif x <= d[p][0.50]:
        return 3
    elif x <= d[p][0.75]:
        return 2
    else:
        return 1

In [36]: #Calculate Add R, F and M segment value columns in the existing dataset to show R, F and M segment values
RFMScores['R'] = RFMScores['Recency'].apply(RScoring, args=('Recency',quantiles,))
RFMScores['F'] = RFMScores['Frequency'].apply(FmScoring, args=('Frequency',quantiles,))
RFMScores['M'] = RFMScores['Monetary'].apply(FmScoring, args=('Monetary',quantiles,))
RFMScores.head()
```



## Appendix II: K-means clustering Code

### Log Transformation and Standardization:

```
In [46]: #Handle negative and zero values so as to handle infinite numbers during Log transformation
def handle_neg_n_zero(num):
    if num <= 0:
        return 1
    else:
        return num
#Apply handle_neg_n_zero function to Recency and Monetary columns
RFMScores['Recency'] = [handle_neg_n_zero(x) for x in RFMScores.Recency]
RFMScores['Monetary'] = [handle_neg_n_zero(x) for x in RFMScores.Monetary]

#Perform Log transformation to bring data into normal or near normal distribution
Log_Tfd_Data = RFMScores[['Recency', 'Frequency', 'Monetary']].apply(np.log, axis = 1).round(3)
```

```
In [47]: from sklearn.preprocessing import StandardScaler

#Bring the data on same scale
scaleobj = StandardScaler()
Scaled_Data = scaleobj.fit_transform(Log_Tfd_Data)

#Transform it back to dataframe
Scaled_Data = pd.DataFrame(Scaled_Data, index = RFMScores.index, columns = Log_Tfd_Data.columns)
```

### Elbow Method:

```
In [51]: from sklearn.cluster import KMeans

sum_of_sq_dist = {}
for k in range(1,15):
    km = KMeans(n_clusters= k, init= 'k-means++', max_iter= 1000)
    km = km.fit(Scaled_Data)
    sum_of_sq_dist[k] = km.inertia_

#Plot the graph for the sum of square distance values and Number of Clusters
sns.pointplot(x = list(sum_of_sq_dist.keys()), y = list(sum_of_sq_dist.values()))
plt.xlabel('Number of Clusters(k)')
plt.ylabel('Sum of Square Distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

### K-Mean Clustering:

```
In [52]: #Perform K-Mean Clustering or build the K-Means clustering model
KMean_clust = KMeans(n_clusters= 3, init= 'k-means++', max_iter= 1000)
KMean_clust.fit(Scaled_Data)

#Find the clusters for the observation given in the dataset
RFMScores['Cluster'] = KMean_clust.labels_
RFMScores.head()
```

## **Bibliography:**

- Aggelis, Vasilis, and christodoulakis Dimitris. 2005. "Aggelis, Vasilis, and Dimitris Christodoulakis. "Customer clustering using rfm analysis." In Proceedings of the 9th WSEAS International Conference on Computers 2
- Altman , Douglas G, and J. Martin Bland. 1996. "Statistics Notes: Detecting skewness from summary information." *Bmj* 313 1200
- Aziz, Samer, and Dr. Zekeriya Nas. 2012. "Demographic segmentation and its effects on customer satisfaction." *International Journal of Contemporary Business Studies* 361-379
- Baker, Kristen. 2021. *The Ultimate Guide to Customer Segmentation: How to Organize Your Customers to Grow Better*. HubSpot
- Beane, T. P., D. M. Ennis, and Philip Morris. 1987. "Market Segmentation: A Review." *European Journal of Marketing* 20-42
- Birant, Derya. 2011. "Data mining using RFM analysis." In *Knowledge-oriented applications in data mining*
- Bruce, Cooil, Aksoy Lerzan, and Keiningham Timothy. 2007. "Approaches to customer segmentation." 6 (3-4): 9-39
- Chen, H., R.H. Chiang, and V.C. Storey. 2012. "Business intelligence and analytics: from big data to a big impact." *MIS Quarterly* 1165-1188
- Christy, A. Joy, A Umamakeswari, L. Priyatharsini, and A. Neyaa. 2021. "RFM ranking – An effective approach to customer segmentation." *Journal of King Saud University - Computer and Information Sciences* 1251-1257
- Daoud, Rachid Ait, Amine Abdellah, Bouikhalene Belaid, and Lbibb Rachid. 2015. "Combining RFM model and clustering techniques for customer value analysis of a company selling online." In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)* 1-6
- DeCarlo, Matthew. 2018. *Scientific inquiry in social work*. Open Social Work Education

- Dolnicar, Sara, Bettina Grün, and Friedrich Leisch. 2018. *Market Segmentation Analysis*. Singapore: Austrian Science Fund (FWF)
- Georgiev, Deyan. 2022. "Customer Retention Statistics & Predictions [Updated 2022]." Review42. <https://review42.com/resources/customer-retention-statistics/>
- Gil-Saura, Irene, and Maria Eugenia Ruiz-Molina. 2009. "Retail customer segmentation based on relational benefits." *Journal of Relationship Marketing* 253-266
- Gordini, Niccolo , and Valerio Veglio. 2014. *Customer Relationship Management and Data Mining: A Classification Decision Tree to Predict Customer Purchasing Behavior in Global Market*. IGI-Global, Hershey: Pennsylvania, USA
- Guarda, Teresa, Maria Augusto, Carlos Silva, João Lourenço, Andrea Sousa, and Aquilino Costa. 2014. "Database marketing tools for SMEs: The case of RFM model." *International Conference on Logistics, Engineering, Management and Computer Science, LEMCS 2014* 995-999
- Gustriansyah, Rendra, Suhandi Nazori , and Antony Fery. 2020. "Clustering optimization in RFM analysis based on k-means." *Indones. J. Electr. Eng. Comput. Sci* 18 470-477
- Heikkilä, Fahed Yoseph and Markku. 2018. "Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method." *International Conference on Machine Learning and Data Engineering (iCMLDE)* 108-116
- Huang, Yong, Zhang Mingzhen, and He Yue. 2020. "Research on improved RFM customer segmentation model based on K-Means algorithm." In *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)* 24-27
- Kabasakal, I. 2020. "Customer Segmentation Based On Recency Frequency Monetary Model : A Case Study in E-Retailing." *Bilişim Teknolojileri Dergisi* 13 (1): 47-56
- Kansal, Tushar, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. 2018. "Customer segmentation using K-means clustering." *international conference on computational techniques, electronics and mechanical systems (CTEMS)* 135-139

- Kotler, P. 1997. "Marketing Management Analysis , Planning, Implementation, and Control." Prentice-Hall International 257-60
- Kotler, Philip, and Armstrong Gary. 2010. Principles of marketing. Pearson education
- Lee, Chung-Shing. 2001. "An analytical framework for evaluating e-commerce business models and strategies." Internet Research
- Madani, Samira. 2009. "Mining changes in customer purchasing behavior: a data mining approach."
- Madzík, Peter, Karol Čarnogurský, Miroslav Hrnčiar, and Dominik Zimon. 2021. "Comparison of demographic, geographic, psychographic and behavioural approach to customer segmentation." International Journal of Services and Operations Management 346-371
- Majumder, Prateek. 2021. K-Means clustering with Mall Customer Segmentation Data | Full Detailed Code and Explanation. 05 25. Accessed 06 08, 2022. <https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/>
- Maryani, Ina , Dwiza Riana, Rachmawati Darma Astuti, Ahmad Ishaq, Sutrisno, and Eva Argarini Pratama. 2018. "Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm." 2018 Third International Conference on Informatics and Computing (ICIC) 1-6
- McLeod, Saul. 2019. What's the difference between qualitative and quantitative research? Accessed 07 01, 2022. <https://www.simplypsychology.org/qualitative-quantitative.html>
- Mialki, Stephanie. 2022. Instapage. 05 11. Accessed 06 94, 2022. <https://instapage.com/blog/psychographic-segmentation>
- Namvar, Morteza, Mohammad R. Gholamian, and Sahand KhakAbi. 2010. "A two phase clustering method for intelligent customer segmentation." In 2010 International Conference on Intelligent Systems, Modelling and Simulation 215-219

- Ngai, E. W.T., LI Xiu, and D.C.K Chau. 2009. "Application of data mining techniques in customer relationship management:A literature review and classification." *Expert Systems with Applications* 36 (2 PART 2): 2592-2602
- Oliva, Rogelio. 2019. "Intervention as a research strategy." *Journal of Operations Management* 710-724
- Onur, DOĞAN, AYÇİN Ejder , and BULUT Zeki Atıl. 2018. "CUSTOMER SEGMENTATION BY USING RFMMODEL AND CLUSTERING METHODS: A CASESTUDY IN RETAIL INDUSTRY." *International Journal of Contemporary Economics and Administrative Sciences* 1-19
- Pakyurek, Muhammet, Mehmet Selman Sezgin, Sedat Kestepe, Busra Bora, Remzi Duzagac, and Olcay Taner Yildiz. 2018. "Customer clustering using RFM analysis." *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018* 1-4
- Pan, Zhedan, Hoyeon Ryu, and Jongmoon Baik. 2007. "A Case Study: CRM Adoption Success Factor Analysis and Six Sigma DMAIC Application." *5th ACIS International Conference on Software Engineering Research, Management & Applications (SERA 2007)* 828-838. doi:10.1109/SE
- Parikh, Yash, and Eman Abdelfattah. 2020. "Clustering algorithms and RFM analysis performed on retail transactions." In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* 0506-0511
- Patil, Saurabh, Hasnath Khan, Sachin Mehta, and Prof Umakant Mandawkar. 2021. "STUDY OF CUSTOMER SEGMENTATION USING k-MEANS CLUSTERING AND RFM MODELLING." *Journal of engineering sciences* 556-559
- Rocca, Baptiste, and Joseph Rocca. 2019. *Introduction to recommender systems. Towards Data Science*

- Rojlertjanya, Ponlacha. 2019. "Customer Segmentation Based on the RFM Analysis Model Using K-Means Clustering Technique: A Case of IT Solution and Service Provider in Thailand."
- Rossi, P., E. McCulloch, and G. Allenby. 1996. "The Value of Purchase History Data in Target Marketing." *Marketing Science* 321-340
- Rygielski, Chris , Jyun-Cheng Wang, and David C. Yen. 2002. "Data mining techniques for customer relationship management." *Technology in Society (Technology in Society)* 483-502
- Sari, Juni Nurma , Lukito Nugroho, Ridi Ferdiana, and Paulus Insap Santosa. 2011. "Review on Customer Segmentation Technique on Ecommerce." (Amercian Scientific publishers) 400-407
- Sarvari, Peiman Alipour, Alp Ustundag, and Hidayet Takci. 2016. "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis." *Kybernetes* 1129-1157
- Sciences, Administrative. 2018. "Customer Segmentation By Using Rfm Model and Clustering Methods : a Case." *International Journal of Contemporary Economics and Administrative Sciences* 1-19
- Sebunje, William. 2015. " Research Techniques."
- Sheshasaayee, Ananthi, and L. Logeshwari. 2018. "IMPLEMENTATION OF CLUSTERING TECHNIQUE BASED RFM ANALYSIS FOR CUSTOMER BEHAVIOUR IN ONLINE TRANSACTIONS." 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) ( IEEE.) 1166-1170
- Shihab, Sabbir Hossain , Shyla Afroge, and Sadia Zaman Mishu. 2019. "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) 1-4
- Stormi, Kati, Teemu Laine, and Tapio Elomaa. 2018. "Feasibility of b2c customer relationship analytics in the b2b industrial context."

- Su-li, Hao. 2010. "The customer segmentation of commercial banks based on unascertained clustering." In 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM) 297-300
- Sun, Shili. 2009. "An analysis on the conditions and methods of market segmentation." International Journal of Business and Management 63-70
- Swift, Ronald S. 2001. Accelerating customer relationships: Using CRM and relationship technologies. Upper Saddle River: Prentice Hall Professional
- Syakur, M A , B K Khotimah, E M S Rochman, and B D Satoto. 2018. "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster." IOP Conference Series
- Toyeb, Md., Rezwana Mahfuza, Nafisa Islam, and Md Asaduzzaman Faisal Emon. 2021. "LRFMV: an efficient customer segmentation model for superstores." Brac University 71-75
- Tsai, Chih Fong, Ya Han Hu, and Yu Hsin Lu. 2015. "Customer segmentation issues and strategies for an automobile dealership with two clustering techniques." Expert Systems 32 (1): 65-76
- Tsiptsis, Konstantinos K., and Antonios Chorianopoulos. 2011. "Data mining techniques in CRM: inside customer segmentation." John Wiley & Sons
- Tsoy, Marina E., and Vladislav Yu Shchekoldin. 2016. "RFM-analysis as a tool for segmentation of high-tech products' consumers." In 2016 13th International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE) 290-293
- Watson, Roger. 2014. "Quantitative research." Nursing Standard 44
- Weiwen, Xiong, Chen Liang, Zhang Zhiyong, and Qiu Zhuqiang. 2008. "RFM value and grey relation based customer segmentation model in the logistics market segmentation." In 2008 International Conference on Computer Science and Software Engineering 1298-1301

- Wu, Hsin-Hung, En-Chi Chang, and Chiao-Fang Lo. 2009. "Applying RFM model and K-means method in customer value analysis of an outfitter." In *Global Perspective for Competitive Enterprise* 665-672
- Wu, Jing, and Zheng Lin. 2005. "Research on customer segmentation model by clustering." In *Proceedings of the 7th international conference on Electronic commerce* 316-318
- Xiong, Weiwen, Liang Chen, Zhiyong ZhanG, and Zhuqiang Qiu. 2008. "RFM value and grey relation based customer segmentation model in the logistics market segmentation." *Proceedings - International Conference on Computer Science and Software Engineering, CSSE 2008* 1298-1301
2020. Yieldify. 06 23. Accessed 05 22, 2022. <https://www.yieldify.com/blog/types-of-market-segmentation/>
- Yoseph , Fahed, and Markku Heikkilä. 2019. "Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method." *International Conference on Machine Learning and Data Engineering (iCMLDE)* 77-82
- Z. Pan, H. Ryu and J. Baik. 2007. "A Case Study: CRM Adoption Success Factor Analysis and Six Sigma DMAIC Application." *5th ACIS International Conference on Software Engineering Research, Management & Applications* 828-838
- ZEITHAML, V.A. and M.J. BITNER. 2000. *Services Marketing: Integrating Customer Focus Across the Firm*. London: McGraw-Hill
- Zhao, Jinghua, Wenbo Zhang, and Yanwei Liu. 2010. "Improved K-means cluster algorithm in telecommunications enterprises customer segmentation." *IEEE International Conference on Information Theory and Information Security* 167-169
- Zhou, Xiaojing, Zhuo Zhang, and Yin Lu. 2011. "Review of Customer Segmentation method in CRM." In *2011 International Conference on Computer Science and Service System (CSSS)* 4033-4035

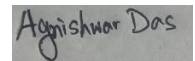


**Declaration in lieu of oath**

I hereby declare that I produced the submitted paper with no assistance from any other party and without the use of any unauthorized aids and, in particular, that I have marked as quotations all passages, which are reproduced verbatim or nearly-verbatim from publications. Also, I declare that the submitted print version of this thesis is identical with its digital version. Further, I declare that this thesis has never been submitted before to any examination board in either its present form or in any other similar version. I herewith agree/disagree that this thesis may be published. I herewith consent that this thesis may be uploaded to the server of external contractors for the purpose of submitting it to the contractors' plagiarism detection systems. Uploading this thesis for the purpose of submitting it to plagiarism detection systems is not a form of publication.

Essen, 22/08/2022

(Location, date)

A handwritten signature in black ink on a light gray rectangular background. The signature reads "Agnishwar Das" in a cursive script.

(genuine signature)