UNIVERSITY OF ECONOMICS AND LAW

FINAL PROJECT REPORT

FUNDAMENTAL DATA ANALYSIS COURSE

TOPIC:

CUSTOMER SEGMENTATION BY USING ANALYZE RFM MODEL WITH K-MEANS CLUSTERING

Lecturers:

1. Ho Trung Thanh, PhD
2. Nguyen Phat Dat, MA

Group 2:

1. Le Ngoc Quynh Giang
2. Tran Thu Hien
3. Nguyen Dang Tue Thi
4. Bui Vo Kim Ngan
5. Nguyen Ngoc Nhu Yen

Ho Chi Minh City, November 28, 2022

MEMBERS OF GROUP 2

No	Full name	Student ID	Individual contribution	Assignment of duties
1	Lê Ngọc Quỳnh Giang	K214140936	100%	Word, **Slide**, Readme
2	Trần Thu Hiền (leader)	K214140938	100%	Word, **Slide**, Code
3	Bùi Võ Kim Ngân	K214140945	100%	Word, Slide, Code
4	Nguyễn Đăng Tuệ Thi	K214140954	100%	**Word**, **Slide**
5	Nguyễn Ngọc Như Yến	K214141339	100%	Word, Slide, Readme

Note: Text in bold means that person is the leader of this section.

Leader says:

I am feel appreciate that I have opportunity to working with these members in group 2. Each of them always dedicated, thoughtful, caring and committed to the final project. Thanks to all the group members for working smoothly together and finished this course.

ACKNOWLEDGEMENTS

We would like to start by sending our most sincere gratitude to PhD. Ho Trung Thanh and M.A. Nguyen Phat Dat, who helped and personally taught us in the Fundamental Data Analysis course this semester. We have welcomed the teacher's attention, assistance, direction with great excitement as we have studied and understood this subject. We gained an overview and a deeper understanding of data analysis from the teacher's knowledge.

While it's possible that knowledge is endless, there are certain restrictions on what each person can learn. Because of this, we will inevitably make mistakes when writing the research paper. In order to make our research more thorough, we eagerly await instructor recommendations.

Finally, we send our best wishes for continued health, joy, and professional success in teaching!

Group 2

COMMITMENT

We certify that the findings in the thesis are our own creations and not the work of others. The information offered throughout the entire thesis is either firsthand or has been assembled from a variety of sources. Every reference has been properly referenced and credited. We shall accept full responsibility for our actions and apply any necessary disciplinary measures.

Ho Chi Minh City, December 19, 2022

Group 2

TABLE OF CONTENTS

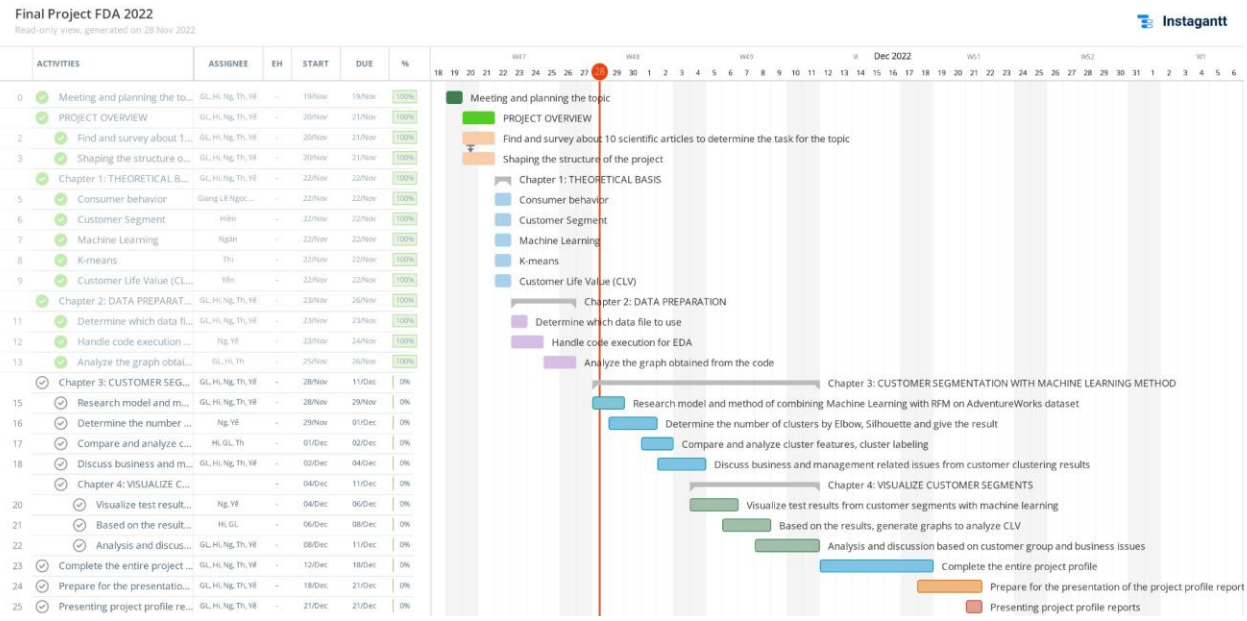
LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

RFM	Recency - Frequency - Monetary
CLV	Customer Life Value

GANTT CHART



Final Project FDA 2022

Read-only view, generated on 28 Nov 2022

Instagantt

	ACTIVITIES	ASSIGNEE	EH	START	DUE	%
0	Meeting and planning the to...	GL, Hi, Ng, Th, Yế	-	19/Nov	19/Nov	100%
	PROJECT OVERVIEW	GL, Hi, Ng, Th, Yế	-	20/Nov	21/Nov	100%
2	Find and survey about 1...	GL, Hi, Ng, Th, Yế	-	20/Nov	21/Nov	100%
3	Shaping the structure o...	GL, Hi, Ng, Th, Yế	-	20/Nov	21/Nov	100%
	Chapter 1: THEORETICAL B...	GL, Hi, Ng, Th, Yế	-	22/Nov	22/Nov	100%
5	Consumer behavior	Giang Lê Ngọc...	-	22/Nov	22/Nov	100%
6	Customer Segment	Hiền	-	22/Nov	22/Nov	100%
7	Machine Learning	Ngân	-	22/Nov	22/Nov	100%
8	K-means	Thi	-	22/Nov	22/Nov	100%
9	Customer Life Value (CL...	Yến	-	22/Nov	22/Nov	100%
	Chapter 2: DATA PREPARAT...	GL, Hi, Ng, Th, Yế	-	23/Nov	26/Nov	100%
11	Determine which data fi...	GL, Hi, Ng, Th, Yế	-	23/Nov	23/Nov	100%
12	Handle code execution ...	Ng, Yế	-	23/Nov	24/Nov	100%
13	Analyze the graph obtai...	GL, Hi, Th	-	25/Nov	26/Nov	100%
	Chapter 3: CUSTOMER SEG...	GL, Hi, Ng, Th, Yế	-	28/Nov	11/Dec	0%
15	Research model and m...	GL, Hi, Ng, Th, Yế	-	28/Nov	29/Nov	0%
16	Determine the number ...	Ng, Yế	-	29/Nov	01/Dec	0%
17	Compare and analyze c...	Hi, GL, Th	-	01/Dec	02/Dec	0%
18	Discuss business and m...	GL, Hi, Ng, Th, Yế	-	02/Dec	04/Dec	0%
	Chapter 4: VISUALIZE C...		-	04/Dec	11/Dec	0%
20	Visualize test result...	Ng, Yế	-	04/Dec	06/Dec	0%
21	Based on the result...	Hi, GL	-	06/Dec	08/Dec	0%
22	Analysis and discus...	GL, Hi, Ng, Th, Yế	-	08/Dec	11/Dec	0%
23	Complete the entire project ...	GL, Hi, Ng, Th, Yế	-	12/Dec	18/Dec	0%
24	Prepare for the presentatio...	GL, Hi, Ng, Th, Yế	-	18/Dec	21/Dec	0%
25	Presenting project profile re...	GL, Hi, Ng, Th, Yế	-	21/Dec	21/Dec	0%

W47
W48
W49
W
Dec 2022
W51
W52
W1
18 19 20 21 22 23 24 25 26 27 28 29 30 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 1 2 3 4 5 6
Meeting and planning the topic
PROJECT OVERVIEW
Find and survey about 10 scientific articles to determine the task for the topic
Shaping the structure of the project
Chapter 1: THEORETICAL BASIS
Consumer behavior
Customer Segment
Machine Learning
K-means
Customer Life Value (CLV)
Chapter 2: DATA PREPARATION
Determine which data file to use
Handle code execution for EDA
Analyze the graph obtained from the code
Chapter 3: CUSTOMER SEGMENTATION WITH MACHINE LEARNING METHOD
Research model and method of combining Machine Learning with RFM on AdventureWorks dataset
Determine the number of clusters by Elbow, Silhouette and give the result
Compare and analyze cluster features, cluster labeling
Discuss business and management related issues from customer clustering results
Chapter 4: VISUALIZE CUSTOMER SEGMENTS
Visualize test results from customer segments with machine learning
Based on the results, generate graphs to analyze CLV
Analysis and discussion based on customer group and business issues
Complete the entire project profile
Prepare for the presentation of the project profile report
Presenting project profile reports

PROJECT OVERVIEW

1. Business Problem

The term "customer segmentation" is crucial in the realm of marketing. More than 30 papers in the Journal of Marketing, more than 50 articles in the Journal of Marketing Research, and other publications are returned when this keyword is searched in article titles. Businesses must maximize the worth of their customers while conducting a marketing strategy to boost sales and successfully compete with rivals. To achieve this, businesses must comprehend their target market, learn the behaviors that consumers use, and identify the appropriate customer groups and potential customer segments (Jedid-Jah Jonker et al, 2004). This is the foundation for providing guidance and recommendations to those who promote marketing activities and developing successful business strategies to ensure the long-term and sustainable growth of businesses. Companies need to consider client retention strategies if they want to compete and last over the long run. (DWK.Khong 2021)

Businesses used to often adopt mass-market marketing methods in the past. Due to this, corporate expenses rise while efficiency does not. As a result, new marketing techniques emerged that offered better client services and divided the target market into manageable segments. Additionally, focusing promotions simply on timing rather than the type and scope of customer incentives will lessen the danger of pricing discrimination in terms of public relations. (Sally Dibb, 2010)

For the reasons listed above, a company that wishes to run successfully must apply the appropriate marketing strategy for each consumer category. In this situation, Adventureworks needs to know how they have client segmentation and what the purchasing characteristics of each section are. From there, provide relevant marketing solutions to maximize clients' capabilities and potential, enhance revenue, and decrease expenses associated with acquiring new customers.

How to divide clients into the appropriate categories? We may use the RFM (Recency, Frequency, Monetary Value) model based on the amount of orders they have made. Companies place a high value on RFM data in order to be able to give them the most complete information and consumer data. Since the RFM variable is the instrument that makes it simplest to identify client segments for shopping, it is frequently required to find segments that depend on it. (Jun Wu et al, 2006) Apply clustering (K-means method) to group consumers into clusters using information from the derived RFM model, then use CLV to divide these clusters into customer segments that businesses need to know. (Dr. Yanka Aleksandrova, 2018, Saharon Rosse et al, 2002)

2. Objectives

2.1. Overall objective

The goal of this project is to study consumer behavior and identify useful customer subgroups using an unsupervised machine learning model. Create a development strategy for the company's future based on that.

2.2. Research questions

The project's main goal was attained by utilizing the following research questions.

1. What information may be utilized to pinpoint the qualities of a customer?

2. What percentage of customers tend to leave the company? How should businesses retain customers?

3. What techniques are available for segmenting and predicting future client purchasing patterns?

2.3. Research objectives

The above research questions then advised on the following research objectives.

1. Locate information about clients and sales for businesses.

2. List the elements that influence how consumers respond to the company's offerings.

3. Select the appropriate algorithms for client segmentation and customer grouping.

4. Create mental images of outcomes, conclusions, and solutions.

3. Objects and scopes

3.1. Objects

Sales statistics from AdventureWorks' database, providing detailed information about the company's online transactions.

3.2. Scopes

Time scope: The data used for the dataset had a time range from July 1, 2017 to June 30, 2021.

Space scope: The data comes from the United Kingdom, Australia, France, Germany, United State of America and Canada regions.

4. Research method

The study used two methods: Qualitative and Quantitative.

5. Process

We will start with the quantitative approach. The study, which was carried out to comprehend consumer behavior, used RFM and K-means Clustering models to cluster each client group.

After normalizing the data, we employ qualitative techniques to describe and analyze the models that were created.

The utilization of the aforementioned forms will serve as a foundation for some of the solutions that are suggested for firms who are struggling to utilize and benefit from client data sources based on the study's objectives.

6. Tools and Programing language

6.1. Tools

Google Colab, Google Sheets, SQL Server.

6.2. Programing language

Python.

7. Structure of project

There are 4 chapters in the project.

I. CHAPTER 1: THEORETICAL BASIS

Chapter overview: The background information to assist the research is presented in this chapter includes Consumer’s behavior ideas, the RFM customer segmentation model, machine learning, and other algorithms like K-means, CLV, etc.

1.1. Consumer behavior

1.1.1. The meaning

Consumer behavior is defined as the behaviors that consumers exhibit while learning about, assessing, buying goods and services. These behaviors also include psychological responses to accept or reject, think about, or compare these things to other goods,... This serves as the foundation for decisions that consumers make on the use of their resources (money, time, effort, etc.) for the acquisition and utilization of products and services to meet personal requirements.

“Selling to individuals who actually want to hear your viewpoint is more effective than interrupting strangers who don't want to hear it”, according to American author and former dot com sales executive Seth Godin. As a result, organizations must not only build internal values but also pay close attention to client behavior if they hope to increase the quality of their customer service purchase from customers to obtain the most precise evaluations for each distinct client; this aids organizations in gaining a better understanding of the industry and simultaneously inspires them to develop new solutions. Identify prospective clients, uphold fidelity, and keep existing ones. Additionally, organizations may readily recognize beneficial and successful business strategies, produce high-quality goods to meet the wants of clients, and become more competitive with rival industry participants. With the aforementioned, the company will always grow responsibly and over time.

1.1.2. Factors influencing consumer behavior

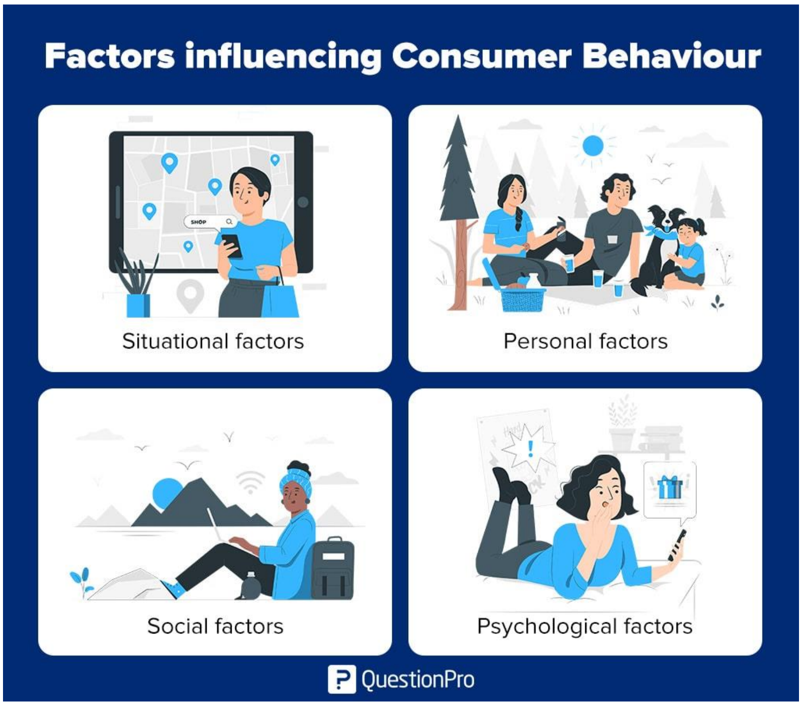


Figure 1: Factors influencing consumer behavior

Cultural, social, personal, and psychological aspects have a big impact on what consumers buy. Even though it is nearly impossible to control these elements, it is nonetheless important to thoroughly examine and take them into account in order to affect consumer behavior.

Firstly, about Cultural factors. A society, government, or religion will establish a system of values, beliefs, traditions, and conventions known as culture, a specific race or religion that has been passed down through the generations. Culture is "a complex consisting of knowledge, belief, art, morals, law, custom, and all other talents and habits of which man as a member of society has gained" (E. B. Tylor, 1871). Therefore, marketers must research ethnic identity, belief, and religion when developing a communication plan because these characteristics all significantly affect the assessment and selection of consumer products.

The following are Social factors. Social factors including Reference Groups, Family, Roles and social status; they influence customer behavior. Firstly, reference groups seem to be associations which either directly or indirectly shape an individual's behavior, beliefs and viewpoints. Consumer behavior is influenced by reference groups including information searching, personal attitudes, and other factors. In addition, family is the closest factor and also the most influential factor on the consumption of goods. In the family, usually the person who generates the main income will play a decisive role in shopping and spending. In addition, the educational environment and habits from a young age also influence the decision to use products and services. The last one is roles and social status, it can be considered as factors representing income levels, profoundly affecting consumer behavior. Most people who are willing to pay for all goods and services are people of social status. People with different roles and statuses will have different needs later on products to use as well as forms of entertainment,...

Furthermore, Personal factors play an important role in influencing the behavior of customers to use the product. The fact that individual buying habits therefore play a significant role. Age, employment, education level, economic standing, personality, and lifestyle are all significant personal aspects that influence consumer behavior. It is clear that psychological traits, requirements, and interests can alter with time as well as with age and stage of life. Due to income and job obligations, profession affects consuming behavior. The consumer trend is more cutting-edge and contemporary the more education one has. Just as someone who is financially secure is more inclined to purchase more and more expensive goods,...

Last but not least are Psychological factors, including Motivation, Perception, Experience, Beliefs and attitudes. About motivation, everybody might have a range of needs at any one period in their life. Hunger, thirst, and weariness are a few physiologically demanding physical conditions that give rise to instinctive urges. Others stem from psychological pressures such as the need for respect, adoration, or recognition. The need can only be made into a motivator if it is intensified enough. Contrasted with, perception is influenced by a variety of elements, including a person's various characteristics, the influence of influencing factors, and the interactions between those factors and external context and internal traits. There are three cognitive processes: Selective Attention, Selective Distortion, and Selective Memory. Additionally, the changes and considerations of customers in shopping behavior stem from their experiences. Experiences emerge through the interplay of urges, stimuli, cues, response, and reinforcement. Moreover, beliefs and attitudes also have a significant effect on purchasing behavior. Trust is based on the experience of buying and using a product or service, which is a comparison between the results received when buying versus the desired expectations. Attitude is expressed in the evaluation and action of the company's products or services.

1.1.3. The importance of analyzing the factors that influence consumer behavior

Firstly, it is simpler for businesses to accurately position their brands in the direction of "personalization" when they are able to pinpoint the elements that have an impact on customers. In the age of the information explosion, positioning in a "personalized" manner is the ideal approach for businesses to connect with the correct consumer insight, directly address the requirements of the user, and encourage the user to purchase the product or brand.

Secondly, companies may create objectives, business plans, and workable marketing programs. Marketers will provide reliable information regarding sales, traffic, and the time period during which customers buy by analyzing consumer behavior. From there, define precise and doable business and marketing goals by making accurate estimates about sales, the amount of potential clients that can be reached, etc.

Additionally, by better understanding customers, firms may find new client segments and foster brand loyalty. Businesses will develop the finest strategies to meet customers' requirements and wishes with products when they are aware of the features of their behavior. Make them into devoted brand ambassadors and "propagandists" for the company.

Moreover, doing this can help firms become more competitive versus rival companies. Businesses will be able to develop more effective product and business strategies as they have a better understanding of client behavior. The product will at this point satisfy the customer's demands and wants. Beyond only offering answers, the goal is to surpass clients' expectations. Businesses may become more competitive by monitoring client behavior beyond just their products.

1.2. Customer segment

1.2.1. The meaning

Client segmentation is the process of distinguishing customer groups based on distinct traits. Each distinct consumer category will have unique qualities and purchasing habits, which will impact the company's business strategy. A marketing and sales plan that is appropriate for the market segment can significantly increase the effectiveness of the company's commercial operations.

1.2.2. Traditional customer segmentation

1.2.2.1. Customer segmentation by demographic

With variables like age, income level, employment, and geography that can all be utilized to separate groups of clients, demographic segmenting is the most well-liked and often employed method by organizations.

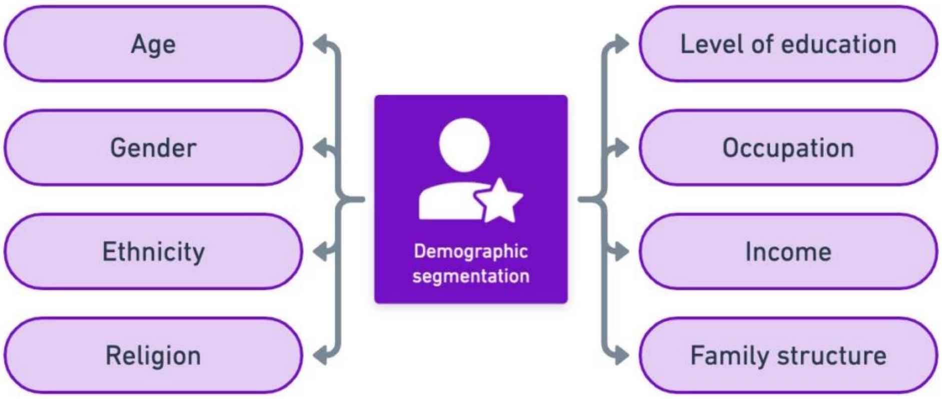


Figure 2: Customer segmentation by demographic

The following benefits and drawbacks of this approach:

Regarding advantages:

Data collection: is made simple for businesses by employing systems that allow data statistics, surveys, and other methods.

Easy to measure: Because demographic data is based on basic facts about persons and has a straightforward character, it is often updated, evaluated, and measured more easily.

Follow social trends easily: By identifying trends, organizations may track, monitor, and evaluate the customer journey, which allows them to forecast future market conditions.

Maximizing time and resources: A customized marketing message will be more effective than a generic one when it is sent to each segment of the target market.

About the disadvantages:

Uncertain data: While demographic information can reveal a person's age or annual income, it cannot give specifics about that person's personality or goals and their purchasing habits.

Data misinterpretation: The demographic information is largely out of date. The way of life of people has drastically altered. What was once commonplace has little worth today.

Rapidly changing data: Over time, demographic data is likely to change quickly. For instance, a customer's age, income, marital status, level of education, and employment all fluctuate with time. Therefore, firms will need to conduct frequent surveys if it becomes vital to pinpoint the precise client groupings based on demographics.

1.2.2.2. Segmenting customers by buying behavior

To get the answer to the question: What did consumers buy? How frequently do they buy? What is the cost of each purchase for the consumer? Why do consumers select that good or service? The detailed process of customer behavior segmentation aids firms in more precise client identification.

Customers will be split into the following groups using this methodology: first-time purchasers, potential buyers, regular customers, and those who have moved to another brand,...

1.2.3. Customer segmentation by RFM model

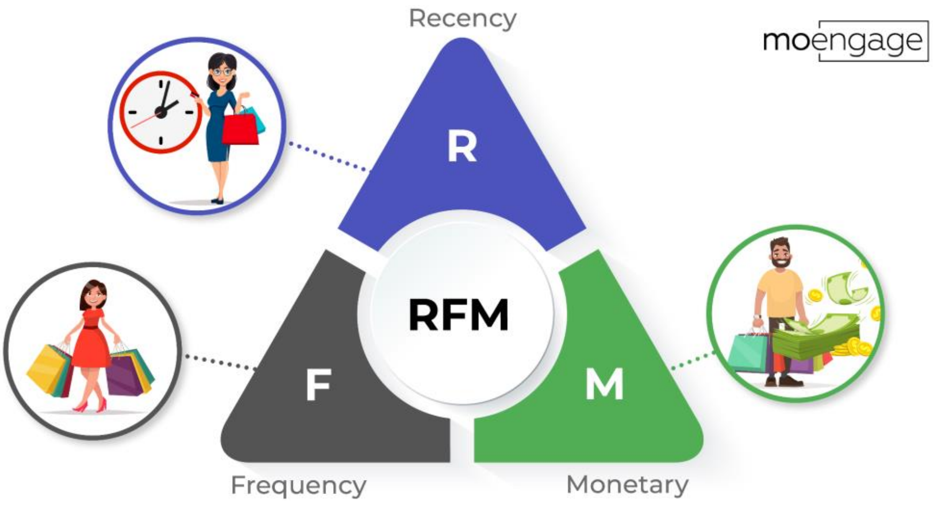


Figure 3: RFM model

With the development of advanced technology, there are more and more modern methods to efficiently segment huge data from customers. One of them is the RFM model - a useful and popular tool for customer segmentation.

RFM is composed of 3 elements: Recency, Frequency and Monetary. In there:

Recency answers the question “When was the most recent purchase? The time since last purchase indicates whether a consumer was actually active during the review period. The higher this index, the greater the likelihood that consumers will leave. This is the basis for businesses to timely adjust their business strategies to ensure conformity and customer retention.

Frequency represents how often customers make purchases. Purchase frequency helps businesses assess whether customers have potential or not; thereby promoting the quality of products and services to increase the number of potential customers for businesses.

Monetary shows the amount of money that consumers have spent to buy products of businesses. This is a factor that directly affects sales. Monetary will have a direct impact on revenue and indirectly through the remaining two factors, Recency and Frequency.

Thanks to the effectiveness of the RFM model, businesses have more opportunities to better understand their customers, create conditions for businesses to improve the quality of products and services, and promote sustainable development. . However, the growing number of large and small businesses and the rapid increase in the number of customers make the RFM model problematic for complex and huge data. Therefore, the exploitation of methods in the field of Machine Learning becomes essential in assisting customer segmentation. Accordingly, data analysts used K-means Clustering unsupervised machine learning method combined with RFM model to increase the efficiency of customer clustering.

1.3. K-means clustering with Elbow method

1.3.1. K-means clustering

K-means is an unsupervised clustering algorithm first proposed by McQueen as early as 1967. The practice of clustering in large data sets is very common nowadays.

The input and result of the K-means clustering algorithm:

- Input: number k, expected data set.
- Output: k classified clusters.

K-means clustering includes the following steps:

Step 1: Select the k-point data from the dataset.

Step 2: Group the data points into the nearest centroid using Euclidean distance.

Step 3: Update the new centroid by averaging the distances of the data points belonging to that centroid.

Step 4: Repeat Step 2 until it satisfies the loss function and the centroids do not change.

Formula to calculate Lost Function:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2 \qquad (1)$$

In there:

E: SSE of all data points

k: number of clusters

x: data point

xi: average of Ci

Euclidean distance formula:

$$d(\mathbf{x}_i, \mathbf{y}_i) = \left[\sum_{i=1}^{n} (x_i - \mathbf{y}_i)^2 \right]^{\frac{1}{2}} \quad (2)$$

1.3.2. Elbow method

We must decide how many clusters there will be before running the K-means method. The elbow method is a heuristic used in cluster analysis to estimate the number of clusters present in a data set. Plotting the explained variation as a function of the number of clusters, the procedure entails choosing the elbow of the curve as the appropriate number of clusters.

The Elbow method, which is frequently used to choose a manageable number of split clusters based on the histogram, occasionally makes it difficult to determine where the Elbow is located, particularly in the case of data sets where it is difficult to identify the clustering algorithm. However, in general, the Elbow method is still the best approach to use for determining how many clusters should be separated.

Implementing the Elbow approach involves the following steps:

Step 1: Use a clustering technique with a configurable number of clusters, such as K-means (eg from 1 to 10).

Step 2: Determine the value of WSS for each k value.

Step 3: Using the k values, draw the elbow curve.

Step 4: Select the proper k, which is the location of the bend, based on the Elbow curve.

1.4. Customer Life Value (CLV)

CLV is understood as the full amount that a customer has to pay for services or products during their lifetime. Understanding CLV is essential for businesses to come up with business strategies to retain customers longer. The customer CLV index helps businesses understand and evaluate the loyalty of existing customers.

Enterprises can easily calculate CLV through the following steps:

Step 1: Calculate Average Purchase Value (APV). APV is the average temporary profit earned on each sale order. It can be determined using the following formula:

APV = (Total Revenue - Total Cost)/Number of Orders

Step 2: Calculate the Average Purchase Frequency (APFR). It is the average number of purchases a customer makes with your company during a specified time period. APFR can be determined using the following formula:

APFR = Total Orders/Total Customers

Step 3: Calculate the average Customer Value (CV). Average customer value is an indicator that reflects the average net profit of a customer for a company. CV can be determined using the following formula:

CV = APV*APFR

Step 4: Calculate the Average Customer Lifespan (ACL). It is the average length of time a customer maintains a purchase/service from a business.

ACL = Total Customer CL/Total Customer

Customer CL = Last purchase date - First purchase date

Step 5: Calculate Customer Lifetime Value (CLV)

CLV = CV*ACL

Enterprises must use CLV to shape their overall business strategy. If a customer's CLV is increasing, it means the business should continue to invest in product development and meeting customer needs. If CLV decreases, businesses should promptly change their business strategies to better suit their customers.

1.5. Cohort Analytics

Cohort analysis is the process of examining a group of users' behavior. In it, separates the data in a data collection into similar groups before examination. These cohorts, or groups, typically have similar traits or experiences across a specific time period. It can be said that cohort analysis is a tool to measure user engagement over time. From there, user groups may be compared to identify discrepancies and glaring trends.

Because it enables marketers to go beyond the bounds of averages, cohort analysis is significant. It helps businesses to have more precise information and hence make better judgments. We are able to see the growth rate of each region or country more clearly thanks to cohort analysis. Furthermore, data analysis using cohorts is preferable. Its use is not constrained to a particular field or job function.

Cohort analytics come in two varieties. There are Acquisition cohorts and Behavioral cohorts. Depending on when a user was found or registered for a product, users are divided into cohorts called acquisitions. User acquisition may be monitored daily, weekly, or monthly depending on your product. Users are categorized into behavioral cohorts according to the actions they do while using the product over a specific time frame.

Cohort analysis is frequently used in all of these sectors to determine why consumers depart and what can be done to keep them from doing so. This brings up the question of how to calculate the Customer Retention Rate (CRR). This formula is used to compute the customer retention rate:

CRR = ((E - N)/S)*100

Three elements make up the formula:

E: The total number of clients at the conclusion of the time frame.

N: The total number of new clients attracted at that time.

S: The total number of clients at the commencement (or beginning) of the time period.

Customer loyalty is higher when the CRR is higher. The business may determine customer retention level by comparing their CRR to the industry standard. Cohort analysis can be helpful in this situation if CRR paints a gloomy image since it allows for the data analysis that will be needed to take remedial action.

CHAPTER 2: DATA PREPARATION

Chapter overview: After learning the fundamentals in Chapter 1, Chapter 2 will start to process data. The Adventureworks dataset will then be used to test RFM model analysis along with K-means clustering.

2.1. Data Introduction

The data set of interest with the given objectives is: Sales_data, Customer_data, Sales Order_data, Sales Territory_data, Product_data. The study and assessment will be carried out using these as the primary data sets. We shall do data processing and normalize the data of the relevant datasets for the ease of a subsequent study.

The data was obtained between July 1, 2017 and June 30, 2021. Sales data is used for RFM analysis and clustering.

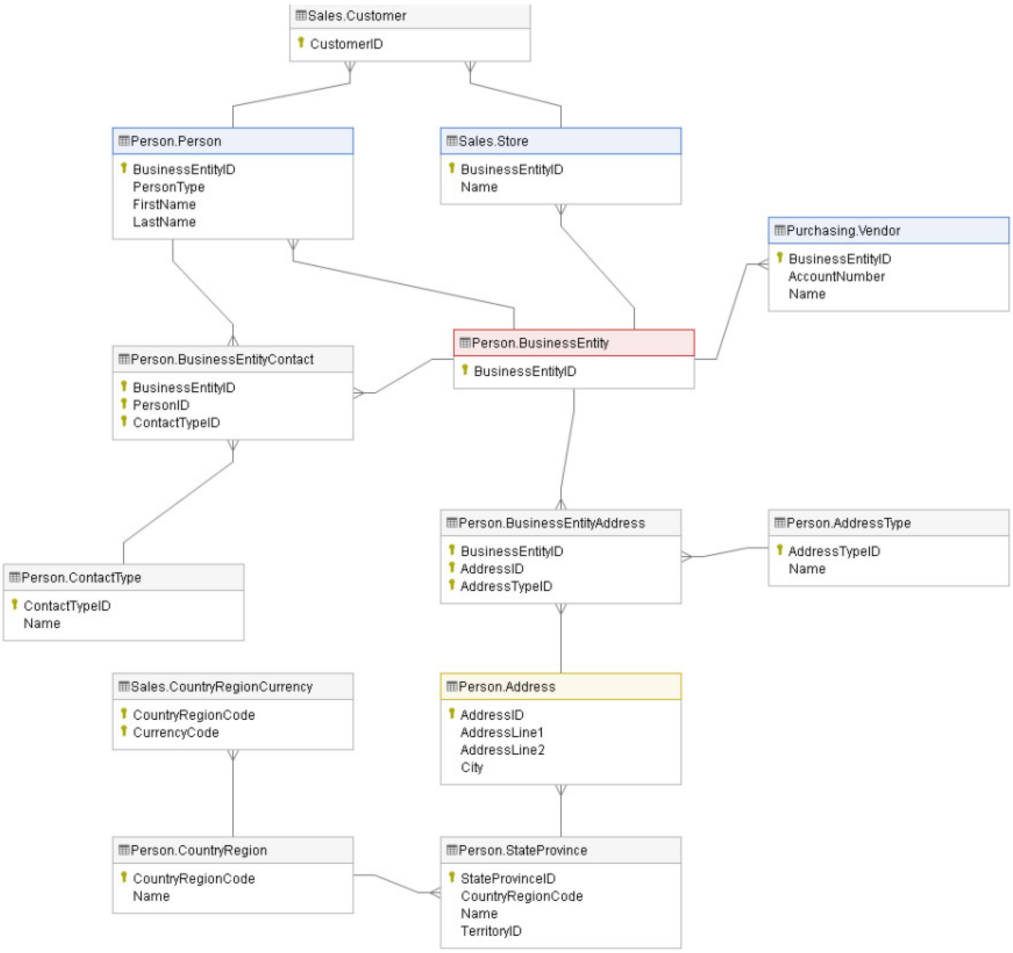
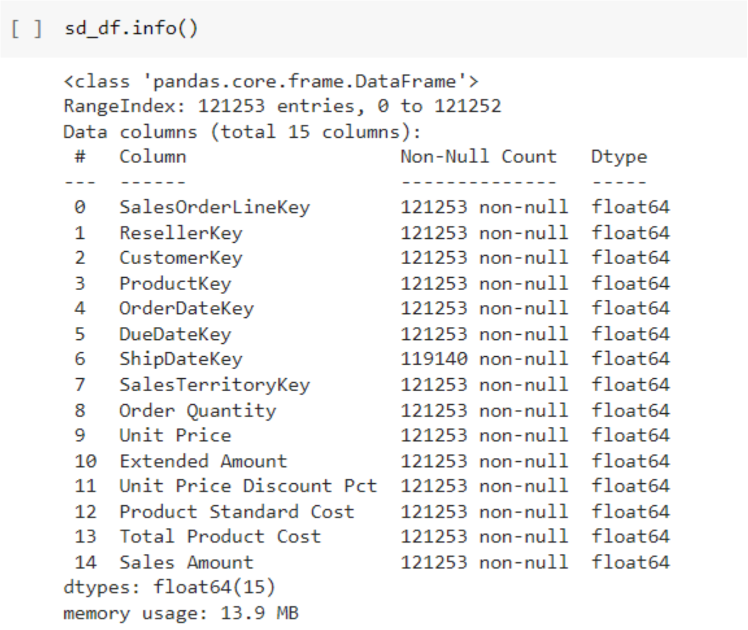


Figure 4: SQL diagram

2.2. Data understanding

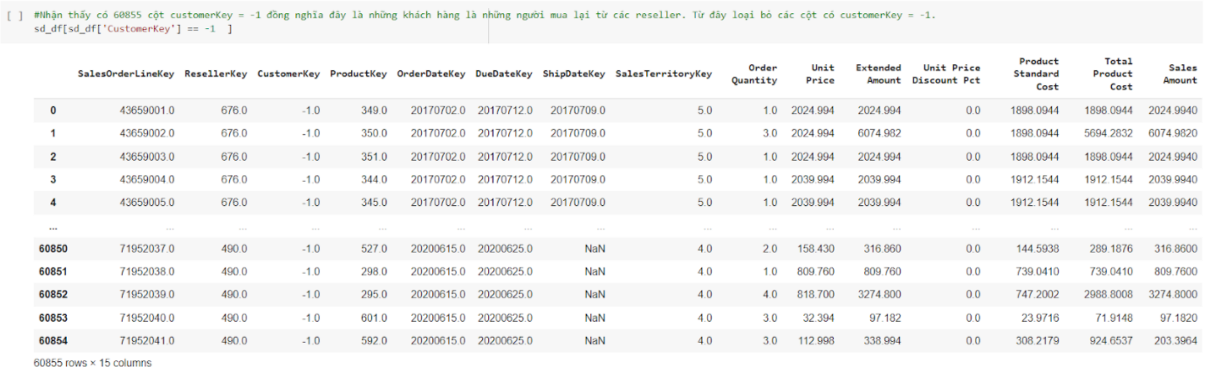
CustomerKey has two values in this data set: equal to -1 and other than -1. This implies that there will be two categories of company clients: direct purchasers (resellers), and indirect clients (those who buy from resellers). The following processing stages let us see more clearly:



```
[ ]  sd_df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 121253 entries, 0 to 121252
     Data columns (total 15 columns):
      #   Column                    Non-Null Count   Dtype
     ---  ------                    --------------   -----
      0   SalesOrderLineKey         121253 non-null  float64
      1   ResellerKey               121253 non-null  float64
      2   CustomerKey               121253 non-null  float64
      3   ProductKey                121253 non-null  float64
      4   OrderDateKey              121253 non-null  float64
      5   DueDateKey                121253 non-null  float64
      6   ShipDateKey               119140 non-null  float64
      7   SalesTerritoryKey         121253 non-null  float64
      8   Order Quantity            121253 non-null  float64
      9   Unit Price                121253 non-null  float64
      10  Extended Amount           121253 non-null  float64
      11  Unit Price Discount Pct   121253 non-null  float64
      12  Product Standard Cost     121253 non-null  float64
      13  Total Product Cost        121253 non-null  float64
      14  Sales Amount              121253 non-null  float64
     dtypes: float64(15)
     memory usage: 13.9 MB
```

Figure 5: Data set details prior to processing

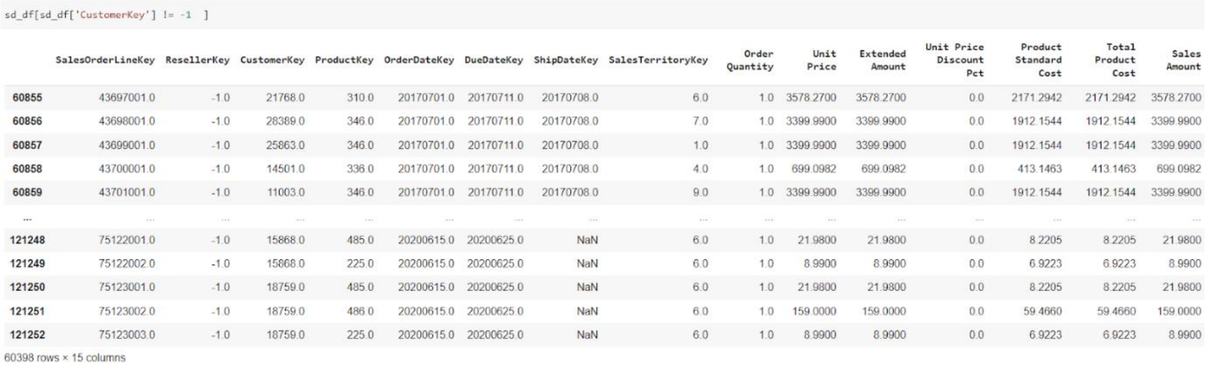


```
[ ]  #Nhận thấy có 60855 cột customerKey = -1 đồng nghĩa đây là những khách hàng là những người mua lại từ các reseller. Từ đây loại bỏ các cột có customerKey = -1.
     sd_df[sd_df['CustomerKey'] == -1  ]
```

	SalesOrderLineKey	ResellerKey	CustomerKey	ProductKey	OrderDateKey	DueDateKey	ShipDateKey	SalesTerritoryKey	Order Quantity	Unit Price	Extended Amount	Unit Price Discount Pct	Product Standard Cost	Total Product Cost	Sales Amount
0	43659001.0	676.0	-1.0	349.0	20170702.0	20170712.0	20170709.0	5.0	1.0	2024.994	2024.994	0.0	1898.0944	1898.0944	2024.9940
1	43659002.0	676.0	-1.0	350.0	20170702.0	20170712.0	20170709.0	5.0	3.0	2024.994	6074.982	0.0	1898.0944	5694.2832	6074.9820
2	43659003.0	676.0	-1.0	351.0	20170702.0	20170712.0	20170709.0	5.0	1.0	2024.994	2024.994	0.0	1898.0944	1898.0944	2024.9940
3	43659004.0	676.0	-1.0	344.0	20170702.0	20170712.0	20170709.0	5.0	1.0	2039.994	2039.994	0.0	1912.1544	1912.1544	2039.9940
4	43659005.0	676.0	-1.0	345.0	20170702.0	20170712.0	20170709.0	5.0	1.0	2039.994	2039.994	0.0	1912.1544	1912.1544	2039.9940
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
60850	71952037.0	490.0	-1.0	527.0	20200615.0	20200625.0	NaN	4.0	2.0	158.430	316.860	0.0	144.5938	289.1876	316.8600
60851	71952038.0	490.0	-1.0	298.0	20200615.0	20200625.0	NaN	4.0	1.0	809.760	809.760	0.0	739.0410	739.0410	809.7600
60852	71952039.0	490.0	-1.0	295.0	20200615.0	20200625.0	NaN	4.0	4.0	818.700	3274.800	0.0	747.2002	2988.8008	3274.8000
60853	71952040.0	490.0	-1.0	601.0	20200615.0	20200625.0	NaN	4.0	3.0	32.394	97.182	0.0	23.9716	71.9148	97.1820
60854	71952041.0	490.0	-1.0	592.0	20200615.0	20200625.0	NaN	4.0	3.0	112.998	338.994	0.0	308.2179	924.6537	203.3964

60855 rows × 15 columns

Figure 6: Indirect customer data



```
sd_df[sd_df['CustomerKey'] != -1  ]
```

	SalesOrderLineKey	ResellerKey	CustomerKey	ProductKey	OrderDateKey	DueDateKey	ShipDateKey	SalesTerritoryKey	Order Quantity	Unit Price	Extended Amount	Unit Price Discount Pct	Product Standard Cost	Total Product Cost	Sales Amount
60855	43697001.0	-1.0	21768.0	310.0	20170701.0	20170711.0	20170708.0	6.0	1.0	3578.2700	3578.2700	0.0	2171.2942	2171.2942	3578.2700
60856	43698001.0	-1.0	28389.0	346.0	20170701.0	20170711.0	20170708.0	7.0	1.0	3399.9900	3399.9900	0.0	1912.1544	1912.1544	3399.9900
60857	43699001.0	-1.0	25863.0	346.0	20170701.0	20170711.0	20170708.0	1.0	1.0	3399.9900	3399.9900	0.0	1912.1544	1912.1544	3399.9900
60858	43700001.0	-1.0	14501.0	336.0	20170701.0	20170711.0	20170708.0	4.0	1.0	699.0982	699.0982	0.0	413.1463	413.1463	699.0982
60859	43701001.0	-1.0	11003.0	346.0	20170701.0	20170711.0	20170708.0	9.0	1.0	3399.9900	3399.9900	0.0	1912.1544	1912.1544	3399.9900
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
121248	75122001.0	-1.0	15868.0	485.0	20200615.0	20200625.0	NaN	6.0	1.0	21.9800	21.9800	0.0	8.2205	8.2205	21.9800
121249	75122002.0	-1.0	15868.0	225.0	20200615.0	20200625.0	NaN	6.0	1.0	8.9900	8.9900	0.0	6.9223	6.9223	8.9900
121250	75123001.0	-1.0	18759.0	485.0	20200615.0	20200625.0	NaN	6.0	1.0	21.9800	21.9800	0.0	8.2205	8.2205	21.9800
121251	75123002.0	-1.0	18759.0	486.0	20200615.0	20200625.0	NaN	6.0	1.0	159.0000	159.0000	0.0	59.4660	59.4660	159.0000
121252	75123003.0	-1.0	18759.0	225.0	20200615.0	20200625.0	NaN	6.0	1.0	8.9900	8.9900	0.0	6.9223	6.9223	8.9900

60398 rows × 15 columns

Figure 7: Direct customer data

We can observe that 60855 customers have customerkey values of -1, and 60398 customers have customerkey values of -1. Following is a description of the two types of customer component ratio:

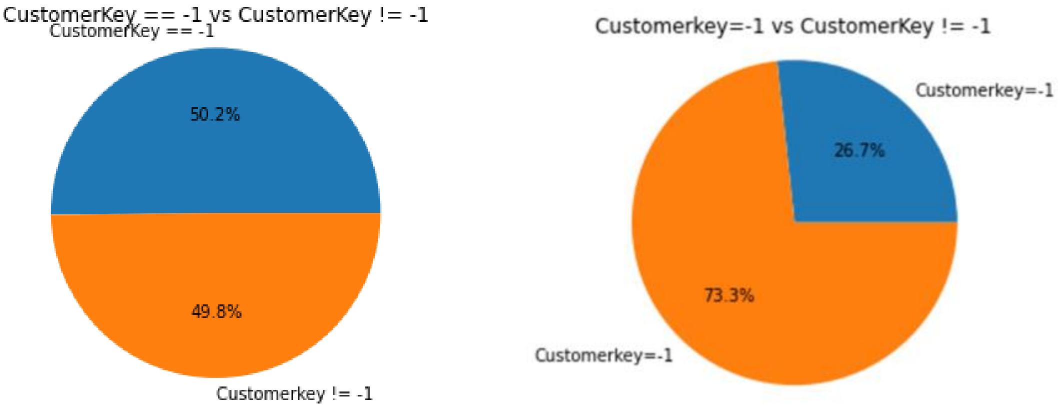


Figure 8: Two types of customer component ratio and sale amount of them

Since clients with customerkey equal to -1 are indirect customers and are the ones who purchase via resellers, we categorize and delete them as outliers of the business.

2.3. Data collection

The dataset from a renowned retail store is reviewed. The data consists of many pieces of information, such as details on customers, sales, orders, and products. Due to the variety of characteristics about the company and its clients that were available, the company served as a source of survey data for study.

The study uses 5 types of data, including: Sale Order data, Sale Territory data, Sales data, Product data, Customer data.

About the Sale Order data:

- 'Customer': this attribute includes 2 types of customer which are Individual (amount: 60398) and Reseller (amount: 60855).
- 'Status': Order Date, Due Date, Ship Date (Status: Ship Date).

About Products in this business:

- Number of products: There are 4 types of items with 397 orders.
- Number of categories of product: There are 4 different categories of products. The general categories are mentioned: Components, Accessories, Bikes, Clothing.

About the data in Sale Territory data:

- Country: Data is available from 6 countries: UK, Australia, France, Germany, USA and Canada.

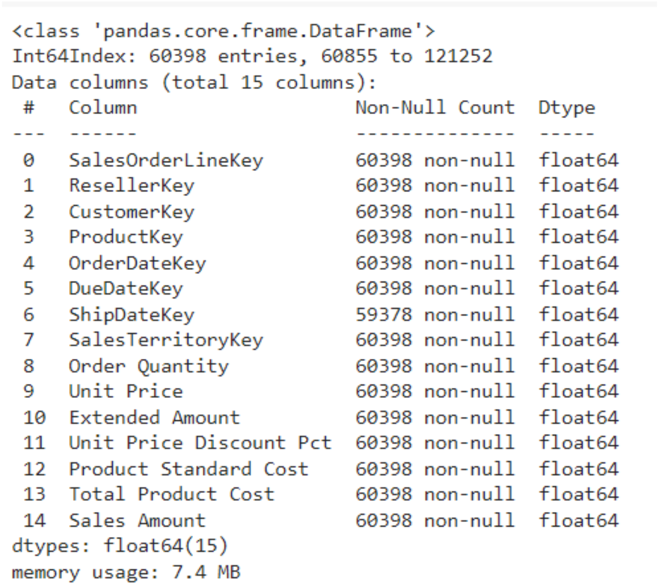
- Region: Divide the nation into 3 major categories: North America, Europe, Pacific and Corporate HQ.

About the data in Customer data:

- Number of customers: 18485.

2.4. Data preprocessing

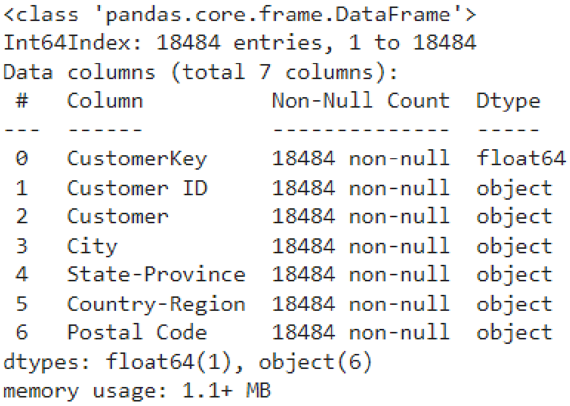
We proceed to remove the customerkey = -1 in Sales data set. Then, we can see clearly that the entire dataset is fit with data of the customerkey ≠ -1.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 60398 entries, 60855 to 121252
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   SalesOrderLineKey       60398 non-null  float64
 1   ResellerKey             60398 non-null  float64
 2   CustomerKey             60398 non-null  float64
 3   ProductKey              60398 non-null  float64
 4   OrderDateKey            60398 non-null  float64
 5   DueDateKey              60398 non-null  float64
 6   ShipDateKey             59378 non-null  float64
 7   SalesTerritoryKey       60398 non-null  float64
 8   Order Quantity          60398 non-null  float64
 9   Unit Price              60398 non-null  float64
 10  Extended Amount         60398 non-null  float64
 11  Unit Price Discount Pct 60398 non-null  float64
 12  Product Standard Cost   60398 non-null  float64
 13  Total Product Cost      60398 non-null  float64
 14  Sales Amount            60398 non-null  float64
dtypes: float64(15)
memory usage: 7.4 MB
```

Figure 9: After remove the customerkey = -1 at the Sales data set

And then proceed to remove the customerkey = -1 at the Customer data set.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 18484 entries, 1 to 18484
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   CustomerKey     18484 non-null  float64
 1   Customer ID     18484 non-null  object
 2   Customer        18484 non-null  object
 3   City            18484 non-null  object
 4   State-Province  18484 non-null  object
 5   Country-Region  18484 non-null  object
 6   Postal Code     18484 non-null  object
dtypes: float64(1), object(6)
memory usage: 1.1+ MB
```

Figure 10: Remove the customerkey = -1 at the Customer data set

We will proceed to explain the revenue in accordance with the real customers of the firm using the boxplot chart after we have excluded clients with a customerkey of -1 as follows:

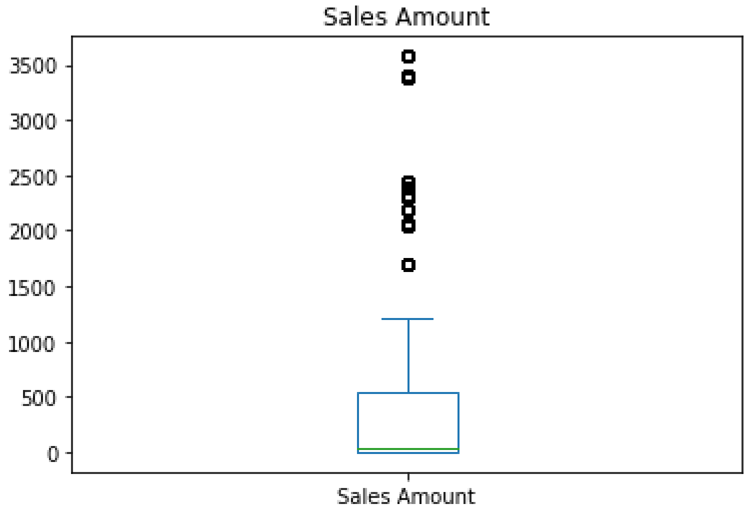


Figure 11: Revenue by actual customers of the business

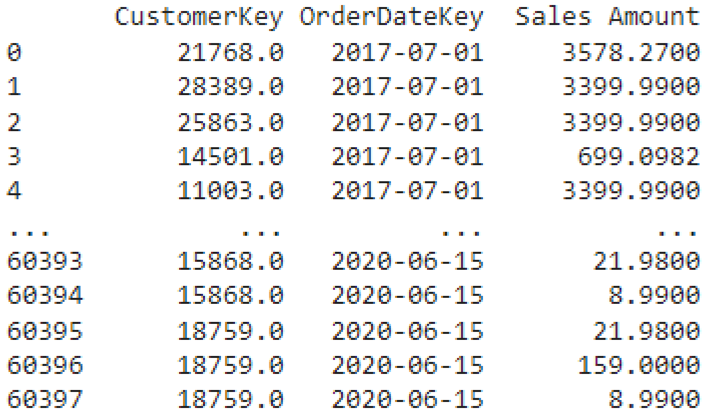
The data collection has a large number of outliers. The accuracy of the statistics in the next sections will suffer by the abundance of outliers. Outliers can cause inaccurate inferences to be drawn regarding the relationship between independent and dependent variables if they are not dealt with.

Descriptive statistics procedures should then be carried out to eliminate pointless outliers for the benefit of the ensuing research process.

	Before	**After**
Count	60398.000000	60398.000000
Mean	486.086911	477.455877
Std	928.489892	900.794937
Min	2.290000	2.290000
25%	7.950000	7.950000
50%	29.990000	29.990000
75%	539.990000	539.990000
Max	3578.270000	3271.556586

Table 1: Sales amount data before and after outliers have been removed

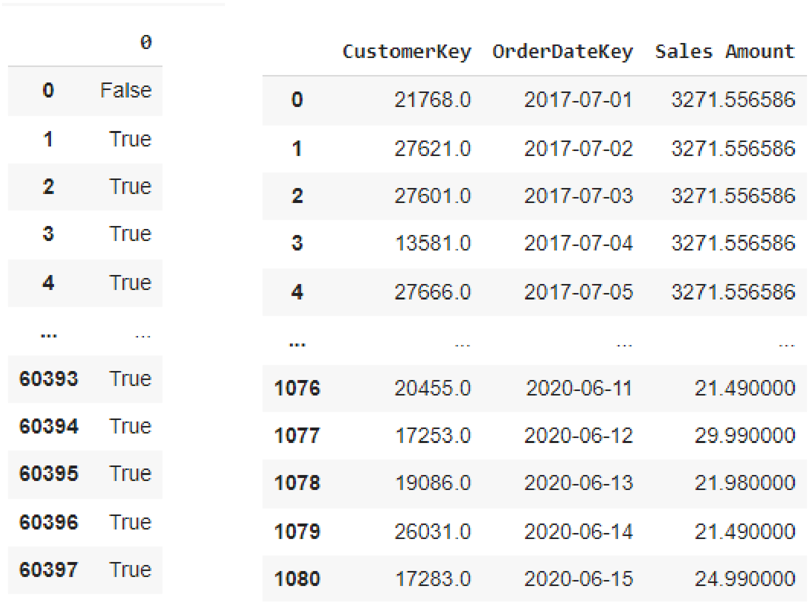
From 3 key columns like Customerkey, OrderDateKey and Sales Amount, we create a Data Frame used to analyze RFM models.



```
       CustomerKey OrderDateKey  Sales Amount
0          21768.0   2017-07-01     3578.2700
1          28389.0   2017-07-01     3399.9900
2          25863.0   2017-07-01     3399.9900
3          14501.0   2017-07-01      699.0982
4          11003.0   2017-07-01     3399.9900
...            ...          ...           ...
60393      15868.0   2020-06-15       21.9800
60394      15868.0   2020-06-15        8.9900
60395      18759.0   2020-06-15       21.9800
60396      18759.0   2020-06-15      159.0000
60397      18759.0   2020-06-15        8.9900
```

Figure 12: The Data Frame of RFM model

To prevent data duplication in the Sales Order field. We will examine the data and eliminate duplicates.



	0
0	False
1	True
2	True
3	True
4	True
...	...
60393	True
60394	True
60395	True
60396	True
60397	True

	CustomerKey	OrderDateKey	Sales Amount
0	21768.0	2017-07-01	3271.556586
1	27621.0	2017-07-02	3271.556586
2	27601.0	2017-07-03	3271.556586
3	13581.0	2017-07-04	3271.556586
4	27666.0	2017-07-05	3271.556586
...	...	...	...
1076	20455.0	2020-06-11	21.490000
1077	17253.0	2020-06-12	29.990000
1078	19086.0	2020-06-13	21.980000
1079	26031.0	2020-06-14	21.490000
1080	17283.0	2020-06-15	24.990000

Figure 13: Eliminate duplicates

2.5. Exploratory Data Analysis

One must analyze clients and sales using a variety of elements in order to comprehend both the present status of the business and the consumers.

Firstly, we will identify purchases from different regions:

Name of country	Math_Score	Cum_percent
United States	7819	42.301450
Australia	3591	61.729063
United Kingdom	1913	72.078554
France	1810	81.870807
Germany	1780	91.500757
Canada	1571	100.000000

Table 2: Describe purchase from different regions

The US region has the largest sales volume with nearly 8000 products, followed by Australia with about 3500 products. The remaining regions including the UK, France, Germany and Canada have roughly the same number of products, ranging from 1000 to 2000 products. The red line represents the 80-20 rule, whereby businesses derive most of their revenue from regions such as the US and Australia, the rest also bring in revenue but a smaller percentage.

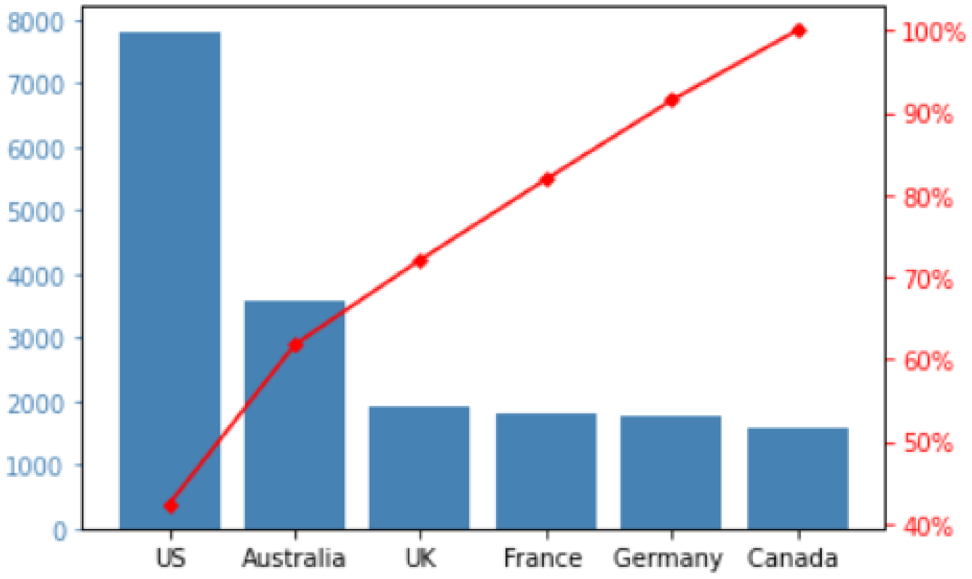


Figure 14: Combination chart of purchase from different regions

It can be seen that the United State is the country with the most concentrated number of customers. Therefore, the proportion of customers by region in the US also accounts for the highest proportion.

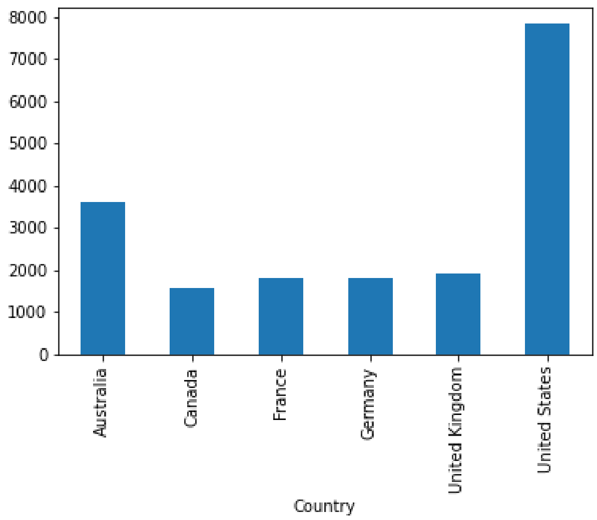


Figure 15: Bar chart describe of sales by country

Customers are primarily from the US. Especially, North America has the largest market share in terms of the total number of clients from all continents.

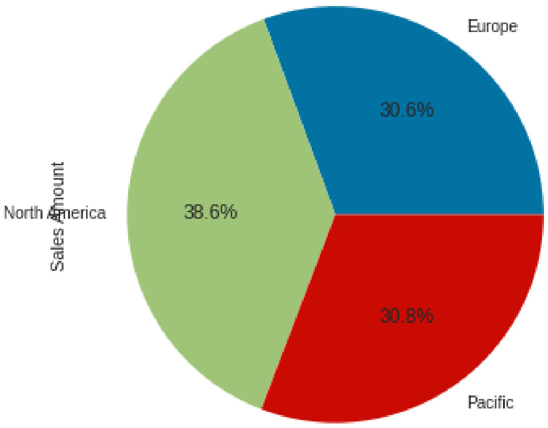


Figure 16: Pie chart describe of sales by region

Once you know the distribution of customers by country and region, understanding the product is the next thing to do.

The most purchased product types after analysis are as follows:

Category	**Sales Amount**
Accessories	7.007600e+05
Bikes	2.779685e+07
Clothing	3.397726e+05

Table 3: The most purchased product types

Revenue is not evenly distributed. Bikes are the main product category of the business with the highest proportion of revenue. Therefore, if you want to increase sales even higher, businesses should continue to promote the quality of Bikes' products, in addition, businesses also need to develop other items evenly to ensure stable revenue.

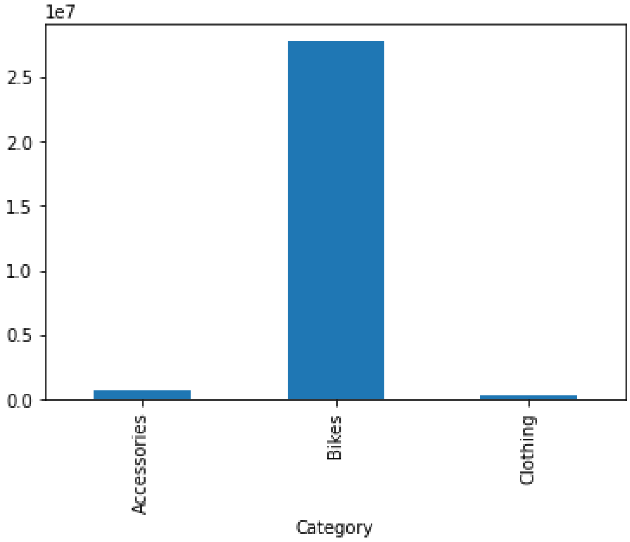


Figure 17: Revenue by product category

Revenue is most concentrated in 5 products of Bikes, businesses should cut unprofitable products to reduce costs, instead businesses should focus on developing products that bring more profit. Or businesses have to increase advertising costs for potential products, such as Mountain Helmets, Road Frames, Touring Bikes,... to bring more positive revenue.

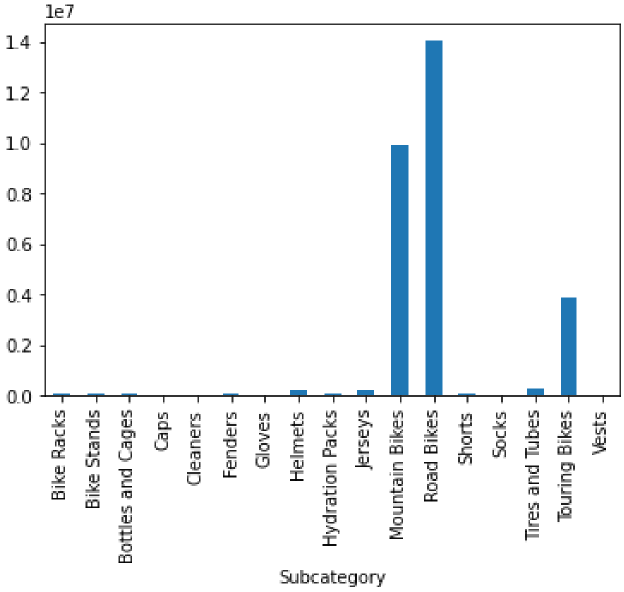


Figure 18: Revenue by product

Bikes continue to be the most popular goods marketed in other nations, in which the United States is the best-selling market share of Bikes, followed by Australia. Businesses must decide whether to concentrate their efforts on investing in the primary specialty or to continue evenly distributing other items in order to further develop the market. This will be examined following the segmentation of the client base and the development of effective clustering tactics.

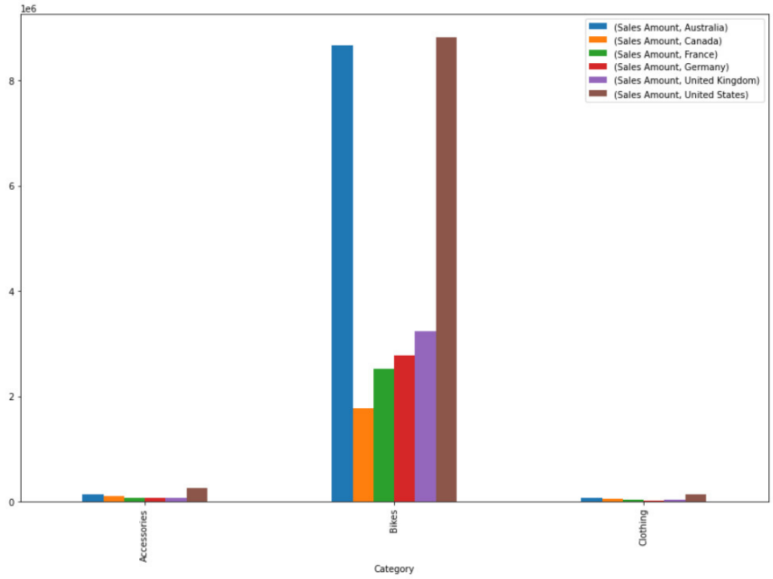


Figure 19: Total sales by Category and Country

High revenue tends to be centered mostly in 2019 and 2020. The month with the most income in 2019 is September. The revenue was minimal in the early years when the company was just getting started, in July 2017.

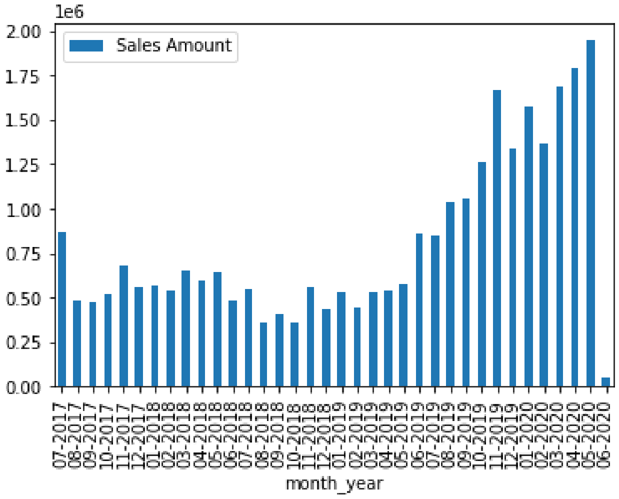


Figure 20: Revenue by month

The cost of the product is one of the factors that determines a client segment. However, firms must take earnings into account. Low selling prices and high manufacturing expenses won't help enterprises turn a profit.

Category	**Total Product Cost**	**Sale Amount**	**Profit**
Accessories	2.620854e+05	7.007600e+05	4.386746e+05
Bikes	1.681235e+07	2.779685e+07	1.09845e+07
Clothing	2.033600e+05	3.397726e+05	1.364126e+05

Table 4: Profit description of the items

Recognize that although clothing and accessories both have the same production costs, accessories generate larger profits. Businesses think about advertising accessories.

CHAPTER 3: CUSTOMER SEGMENTATION WITH MACHINE LEARNING METHOD

Chapter overview: When we use EDA, we find that the Customerkey variable has an outlier portion of -1, meaning that the customers for whom Customerkey = -1 are the Reseller's indirect customers. Whether a consumer purchased a product directly or indirectly, the business must assess and maximize those sales, but it is also hard to exclude outliers from the overall analysis when using common factors. Group 2 will thus examine both categories of customers: those who make direct purchases from the business (Customerkey ≠ 1) and those who make indirect purchases through resellers (Customerkey = 1). The K-means approach will be used in this chapter to segment consumers within each of these two client groups after the RFM model has been used to examine each group. Take informative data from there to identify the traits of each client group that has bought your goods, then label it accordingly. We will also offer a viewpoint based on the model's findings, giving organizations the right business and management solutions.

3.1. Implement RFM model

3.1.1. Variable preprocessor

We will construct a Data Frame containing the required elements, such as CustomerKey, OrderDateKey, and Sales Amount, in order to run RFM. We discovered in the previous EDA phase that the variables are not normally distributed and have a rather large distribution. In order for the subsequent stages to be stable and reliable, we scaled the dataset using the Min-Max approach and then returned the data to its normalized form using Box-cox.

After re-scaling, the data distribution level has fallen to within the range [0, 1.0], whereas the data used for RFM before scale had a rather substantial fluctuation amplitude in the range [0, 9.5].

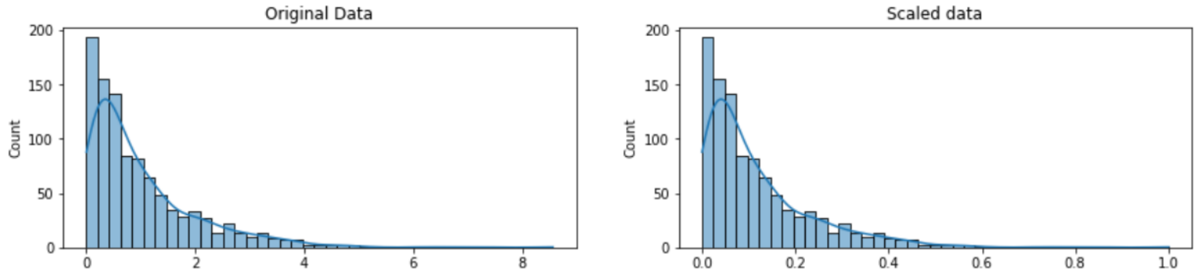


Figure 21: Before and after data partitioning during data scale processes

Although the dataset has been shrunk, it is still clearly strongly separated to the right. Therefore, to transform the data set into a distributed form, we use normalization using the Box-cox technique.

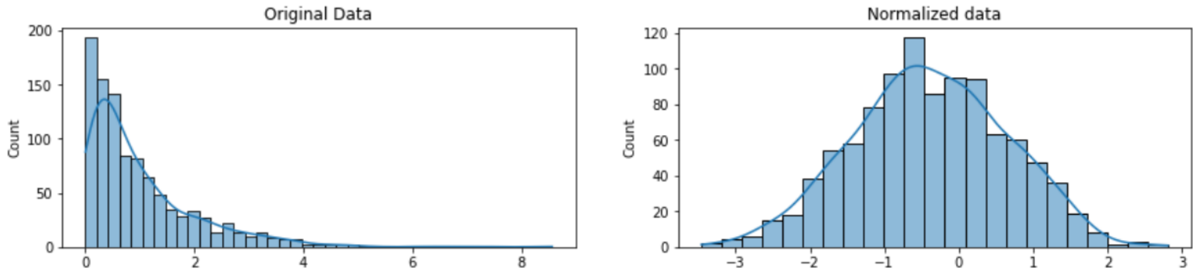
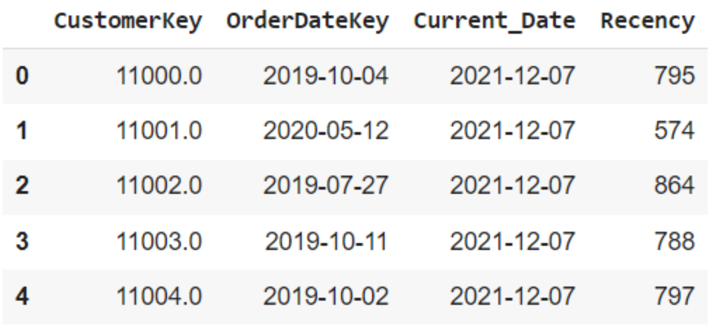


Figure 22: Distribution of data before and after data normalization

3.1.2. Determine the RFM model value

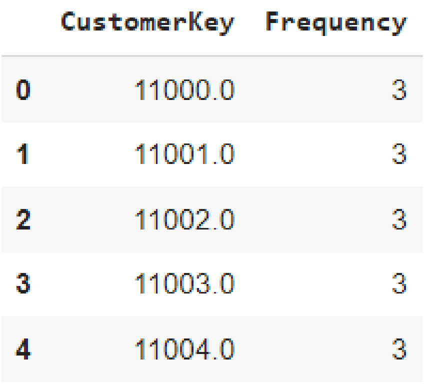
The first is the R - Recency section, which asks for the customer's most recent purchase date. According to the data set, the earliest order was placed on July 1, 2017; and the most recent order was placed on June 15, 2020. The milestone we use to calculate the customer's final purchase date is **December 7, 2021**. As a result, the greater the number in the Recency column is, the longer the client has gone without making a purchase; conversely, the smaller the number will be, the more recently the consumer has made a purchase.



	CustomerKey	OrderDateKey	Current_Date	Recency
0	11000.0	2019-10-04	2021-12-07	795
1	11001.0	2020-05-12	2021-12-07	574
2	11002.0	2019-07-27	2021-12-07	864
3	11003.0	2019-10-11	2021-12-07	788
4	11004.0	2019-10-02	2021-12-07	797

Figure 23: Recency factor

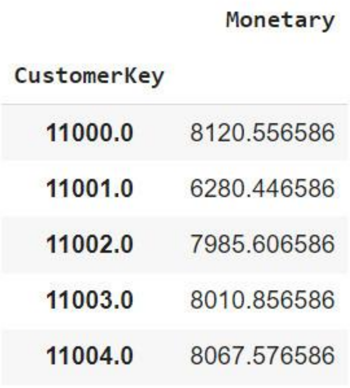
The next section is F, which stands for frequency. Here, we will tally the times consumers have made purchases between July 1, 2017, and December 7, 2021; the higher the number, the more frequently customers have made purchases from the business, and vice versa.



	CustomerKey	Frequency
0	11000.0	3
1	11001.0	3
2	11002.0	3
3	11003.0	3
4	11004.0	3

Figure 24: Frequency factor

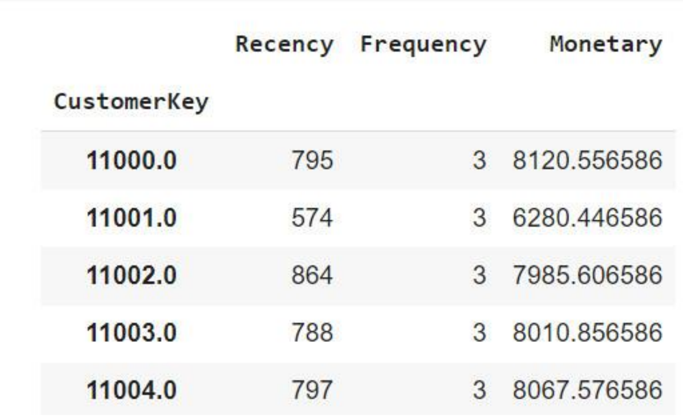
The last variable is M, or Monetary value, which represents the total sum of money that consumers spent within the time frame under consideration. Additionally, this represents the total of each customer's sales.



	Monetary
CustomerKey	
11000.0	8120.556586
11001.0	6280.446586
11002.0	7985.606586
11003.0	8010.856586
11004.0	8067.576586

Figure 25: Monetary factor

Following the computation of the three variables Recency, Frequency, and Monetary, we will join the three tables mentioned above to obtain the following RFM table:



	Recency	Frequency	Monetary
CustomerKey			
11000.0	795	3	8120.556586
11001.0	574	3	6280.446586
11002.0	864	3	7985.606586
11003.0	788	3	8010.856586
11004.0	797	3	8067.576586

Figure 26: Whole RFM model

3.1.3. Create the RFM score

We will score the aforementioned data to determine where the value of R - F - M is located. The grading system we use is based on the data's quartiles. Create a 4 - point scale using the following measuring frame to get the R – F - M values:

Score	Recency	Frequency	Monetary
1	> 788	< 2	0 - 49.97
2	788 - 693	1	49.97 - 270.265
3	693 - 611	1 - 2	270.265 - 2511.275
4	< 611	> 2	> 2511.275

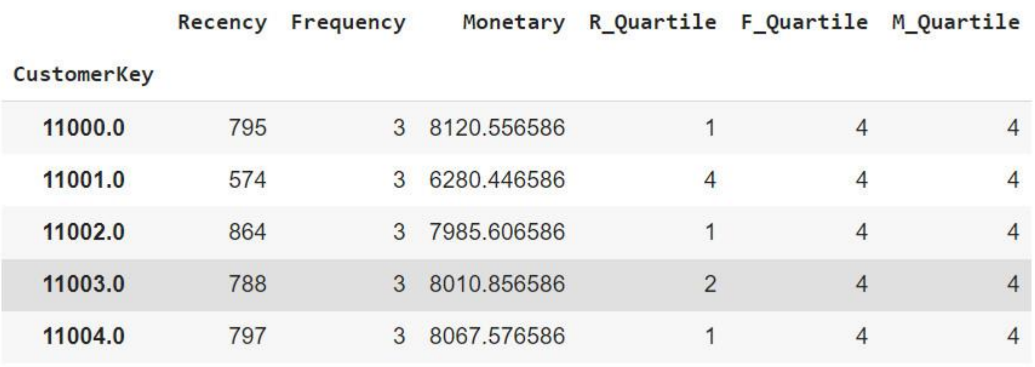
Table 5: RFM value scale

R_Quartile: Recency value as high, the number of points as small.

F_Quartile: The higher the frequency value, the higher the score.

M_Quartile: The higher the Currency Value, the higher the score.

The best RFM score is a score of 444 - the score every store wants its customers to achieve - and the worst RFM score is a score of 111 - the score a business should promptly rework its strategy. Based on the measurement frame above, we will have an RFM score consisting of 3 digits representing the score of R - F - M.



	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile
CustomerKey						
11000.0	795	3	8120.556586	1	4	4
11001.0	574	3	6280.446586	4	4	4
11002.0	864	3	7985.606586	1	4	4
11003.0	788	3	8010.856586	2	4	4
11004.0	797	3	8067.576586	1	4	4

Figure 27: The score of the RFM model

3.2. Implement machine learning to classify customers

3.2.1. Preparing for the K-means method

It is important to do the scale and normalize the data since the RFM findings have an excessively huge variance. Here, the input variable is too right-skewed to be processed in a way that is more like the normal distribution, thus the Standard Scaler approach is used.

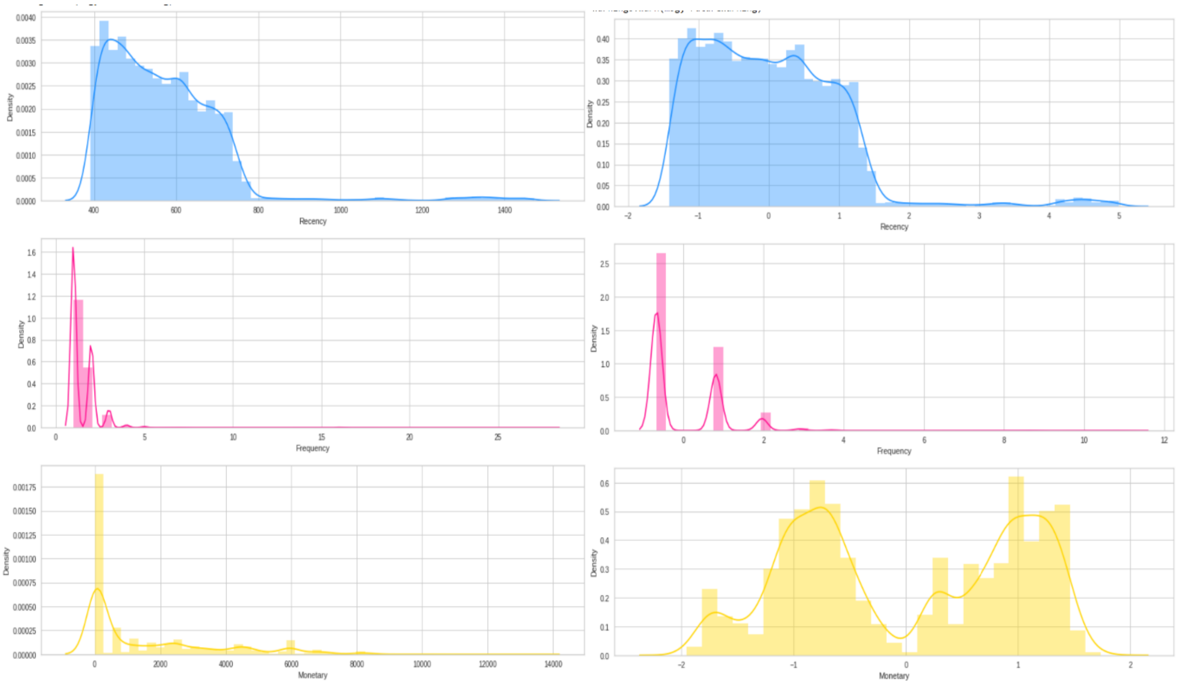
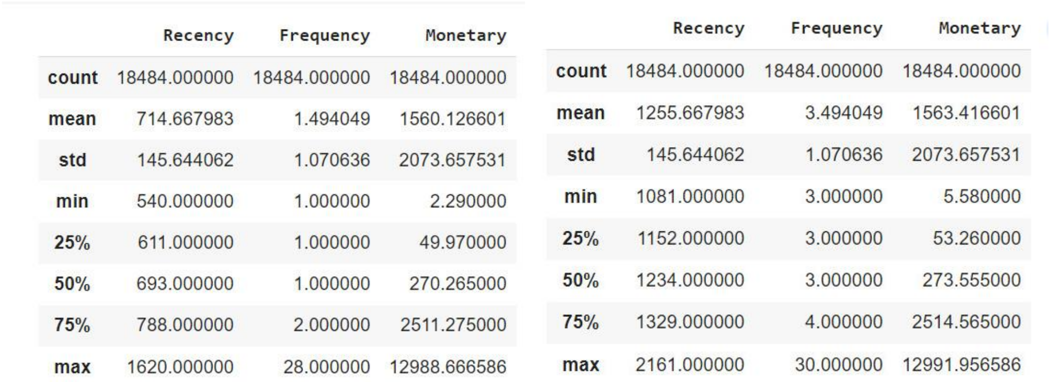


Figure 28: Histogram showing RFM before and after data scaling



	Recency	Frequency	Monetary
count	18484.000000	18484.000000	18484.000000
mean	714.667983	1.494049	1560.126601
std	145.644062	1.070636	2073.657531
min	540.000000	1.000000	2.290000
25%	611.000000	1.000000	49.970000
50%	693.000000	1.000000	270.265000
75%	788.000000	2.000000	2511.275000
max	1620.000000	28.000000	12988.666586

	Recency	Frequency	Monetary
count	18484.000000	18484.000000	18484.000000
mean	1255.667983	3.494049	1563.416601
std	145.644062	1.070636	2073.657531
min	1081.000000	3.000000	5.580000
25%	1152.000000	3.000000	53.260000
50%	1234.000000	3.000000	273.555000
75%	1329.000000	4.000000	2514.565000
max	2161.000000	30.000000	12991.956586

Figure 29: Table describing RFM data before and after data scaling

The number of clusters to be separated must equal the size of the data set in order to use the K-means method. Here, we use two techniques to determine how many clusters should be separated by K-means. The elbow is used as the initial technique to determine how many clusters there are. This elbow is operated using the number k, which ranges from 1 to 10. When we look at the graph, we can see that the "elbow" - the section of the line that sticks out the most - is at point 3, which indicates that the K-means will be separated into clusters of k = 3.

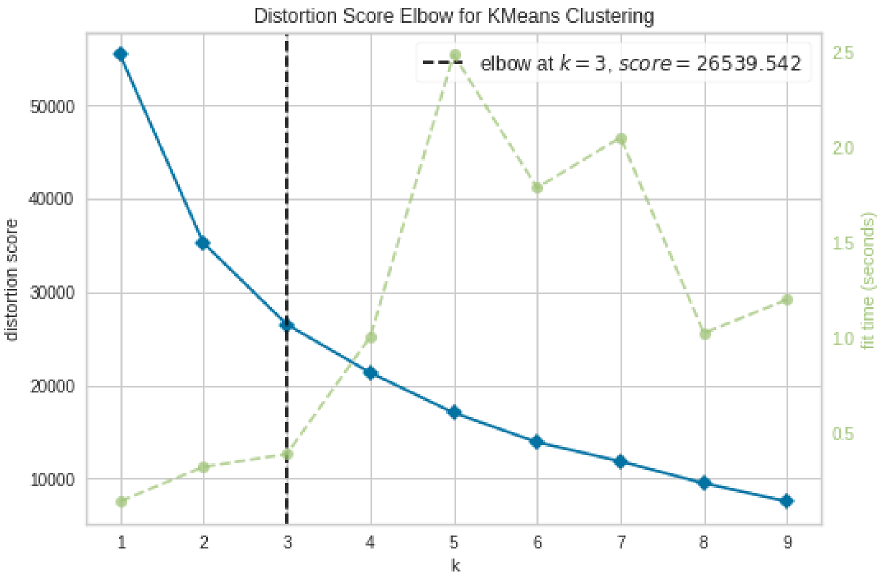
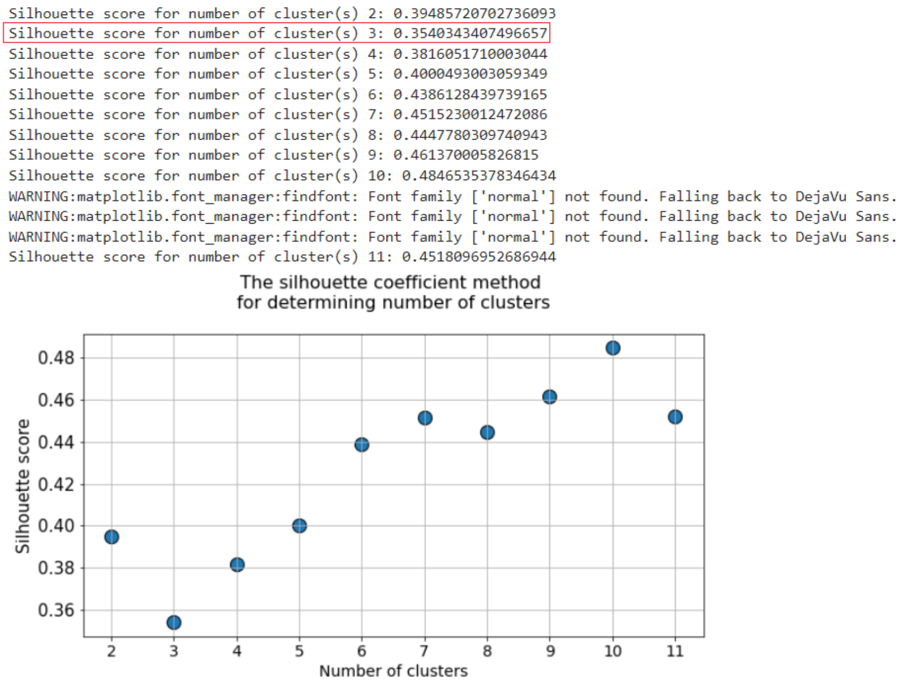


Figure 30: SSE curve results in Elbow method

Apply the Silhouette calculation technique to validate that the findings produced from the Elbow method are accurate in order to establish the right quantity of k to locate. The least Silhouette Score at point k = 3 is 0.3545, as can be seen in the chart below. As a result, pick k = 3 as the most advantageous point and 3 as the ideal number of clusters to separate.

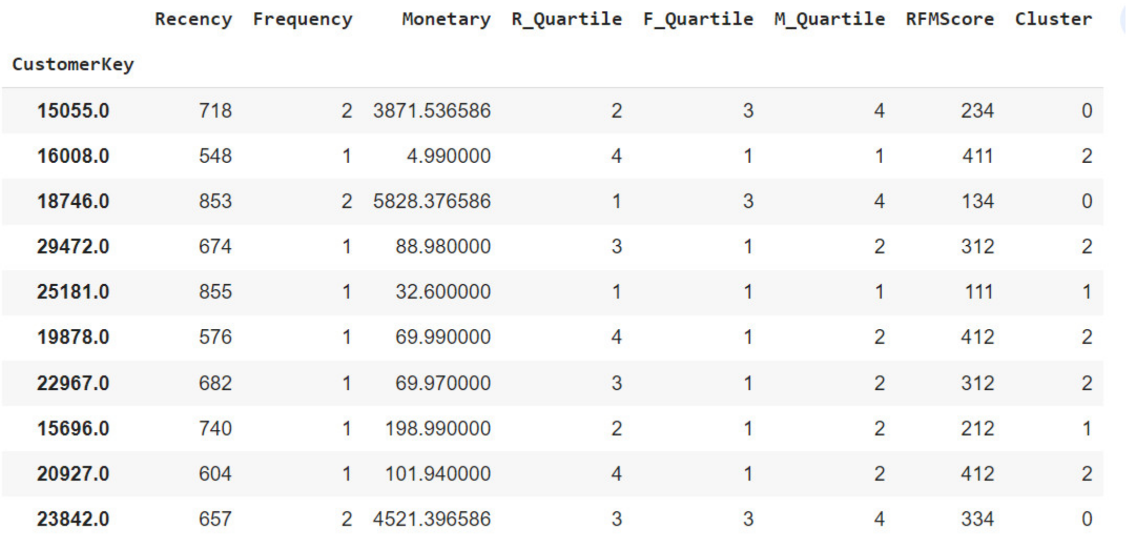


```
Silhouette score for number of cluster(s) 2: 0.39485720702736093
Silhouette score for number of cluster(s) 3: 0.3540343407496657
Silhouette score for number of cluster(s) 4: 0.3816051710003044
Silhouette score for number of cluster(s) 5: 0.4000493003059349
Silhouette score for number of cluster(s) 6: 0.4386128439739165
Silhouette score for number of cluster(s) 7: 0.4515230012472086
Silhouette score for number of cluster(s) 8: 0.4447780309740943
Silhouette score for number of cluster(s) 9: 0.461370005826815
Silhouette score for number of cluster(s) 10: 0.4846535378346434
WARNING:matplotlib.font_manager:findfont: Font family ['normal'] not found. Falling back to DejaVu Sans.
WARNING:matplotlib.font_manager:findfont: Font family ['normal'] not found. Falling back to DejaVu Sans.
WARNING:matplotlib.font_manager:findfont: Font family ['normal'] not found. Falling back to DejaVu Sans.
Silhouette score for number of cluster(s) 11: 0.4518096952686944
```

Figure 31: Distortion Score Elbow for K-means clustering

3.2.2. Perform customer classification by K-means method

K-means is divided into three groups corresponding to Cluster = 0, 1, 2 based on the findings of Elbow.

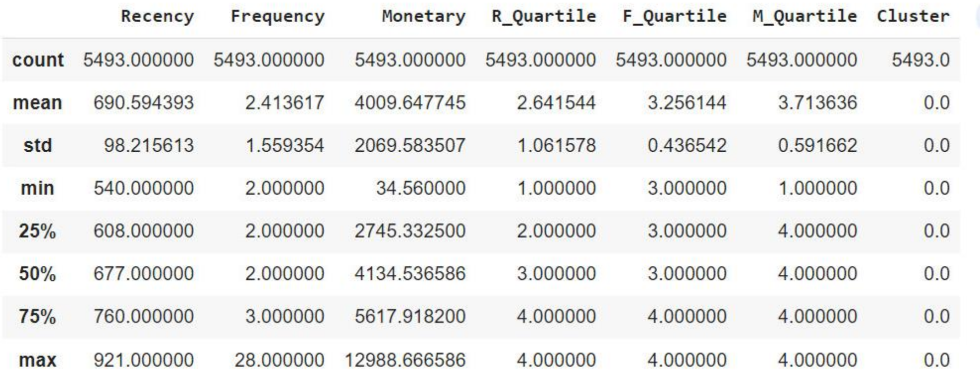


	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore	Cluster
CustomerKey								
15055.0	718	2	3871.536586	2	3	4	234	0
16008.0	548	1	4.990000	4	1	1	411	2
18746.0	853	2	5828.376586	1	3	4	134	0
29472.0	674	1	88.980000	3	1	2	312	2
25181.0	855	1	32.600000	1	1	1	111	1
19878.0	576	1	69.990000	4	1	2	412	2
22967.0	682	1	69.970000	3	1	2	312	2
15696.0	740	1	198.990000	2	1	2	212	1
20927.0	604	1	101.940000	4	1	2	412	2
23842.0	657	2	4521.396586	3	3	4	334	0

Figure 32: Customer division table by cluster

What are the traits of the three clusters that K-means' findings revealed, and how can we identify them appropriately? In order to resolve the aforementioned issues, we give labels to each cluster based on the average of the words in each cluster's description table.

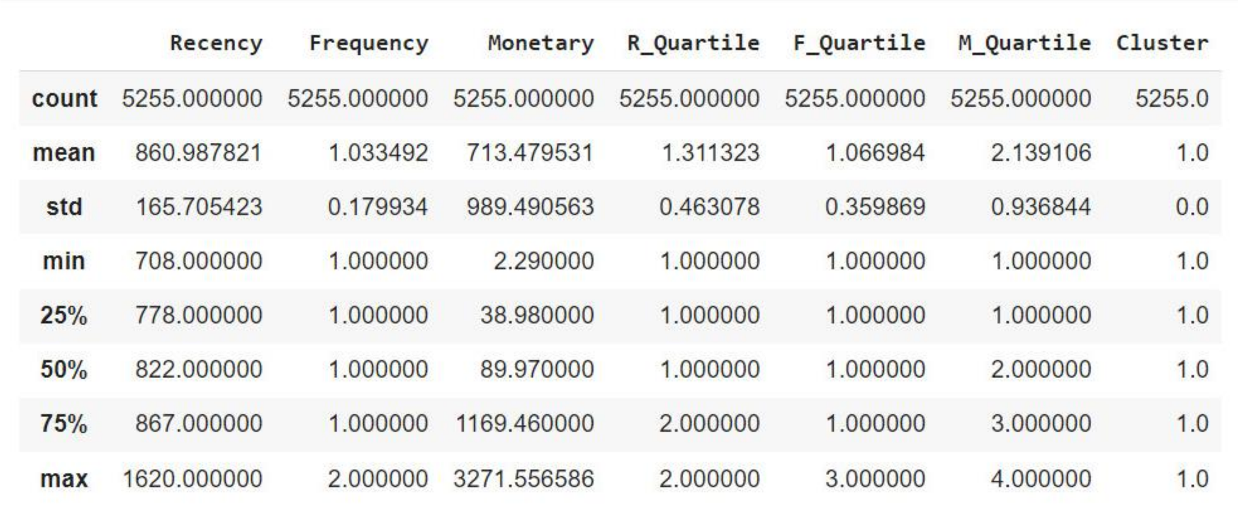
As can be seen, cluster 1 has the fewest clients, making up 28.43% of the total set of direct consumers. This customer cluster also has the lowest score among the three clusters and the lowest Recency and Frequency score on the scale (R_Quartile = F_Quartile ≈ 1). Point M, however, has a mean score (M_Quartile ≈ 2). One may say that this is the Irregular Customer. For this particular consumer group, they only sometimes visit the store. When they do, they suddenly decide that the company's product is appropriate, and they buy it without any particular purpose in mind.



	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	Cluster
count	5493.000000	5493.000000	5493.000000	5493.000000	5493.000000	5493.000000	5493.0
mean	690.594393	2.413617	4009.647745	2.641544	3.256144	3.713636	0.0
std	98.215613	1.559354	2069.583507	1.061578	0.436542	0.591662	0.0
min	540.000000	2.000000	34.560000	1.000000	3.000000	1.000000	0.0
25%	608.000000	2.000000	2745.332500	2.000000	3.000000	4.000000	0.0
50%	677.000000	2.000000	4134.536586	3.000000	3.000000	4.000000	0.0
75%	760.000000	3.000000	5617.918200	4.000000	4.000000	4.000000	0.0
max	921.000000	28.000000	12988.666586	4.000000	4.000000	4.000000	0.0

Figure 33: Data description table of cluster = 1

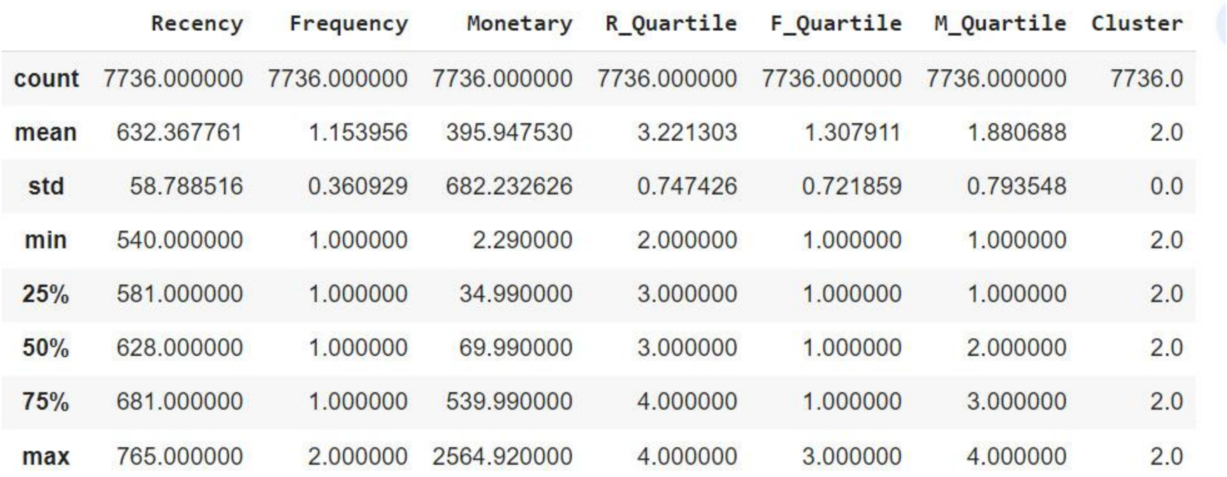
Despite only having 29.7% of the market, Cluster 0 generates 78.43% of its income from direct consumers. The average RFM score for cluster 1 is fairly high (RFM Score = 334), demonstrating that customers have a great deal of confidence in the business and select it as their go to retailer. This cluster might be viewed as a collection of Loyal Customer. These are clients who depend on the firm to uphold the standards of excellence and the principles it has engendered in clients.



	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	Cluster
count	5255.000000	5255.000000	5255.000000	5255.000000	5255.000000	5255.000000	5255.0
mean	860.987821	1.033492	713.479531	1.311323	1.066984	2.139106	1.0
std	165.705423	0.179934	989.490563	0.463078	0.359869	0.936844	0.0
min	708.000000	1.000000	2.290000	1.000000	1.000000	1.000000	1.0
25%	778.000000	1.000000	38.980000	1.000000	1.000000	1.000000	1.0
50%	822.000000	1.000000	89.970000	1.000000	1.000000	2.000000	1.0
75%	867.000000	1.000000	1169.460000	2.000000	1.000000	3.000000	1.0
max	1620.000000	2.000000	3271.556586	2.000000	3.000000	4.000000	1.0

Figure 34: Data description table of cluster = 0

Although it has a sizable client base (41.86%), the final cluster only contributes 10% of the overall income from direct consumers. Customers in this cluster have only made an average of one single transaction (Frequency ≈ 1), although their most recent purchase was quite near (R Quartile ≈ 3), this is the New Customer. Businesses must pay attention to completely utilize this consumer segment to efficiently grow income.



	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	Cluster
count	7736.000000	7736.000000	7736.000000	7736.000000	7736.000000	7736.000000	7736.0
mean	632.367761	1.153956	395.947530	3.221303	1.307911	1.880688	2.0
std	58.788516	0.360929	682.232626	0.747426	0.721859	0.793548	0.0
min	540.000000	1.000000	2.290000	2.000000	1.000000	1.000000	2.0
25%	581.000000	1.000000	34.990000	3.000000	1.000000	1.000000	2.0
50%	628.000000	1.000000	69.990000	3.000000	1.000000	2.000000	2.0
75%	681.000000	1.000000	539.990000	4.000000	1.000000	3.000000	2.0
max	765.000000	2.000000	2564.920000	4.000000	3.000000	4.000000	2.0

Figure 35: Data description table of cluster = 2

CHAPTER 4: VISUALIZATION OF EXPERIMENTAL RESULTS AND PREDICTIVE ANALYSIS OF CLV

Chapter overview: In order to optimize consumer insights, we continue to display the experimental results and forecast CLV analysis after receiving the findings of the RFM model analysis. Understanding consumers is the foundation for drawing correct conclusions about the performance of the business, which then suggests a means to address unresolved issues and advance business. These conclusions may be drawn using RFM and CLV models develop a sustainable world.

4.1. Customer lifetime value

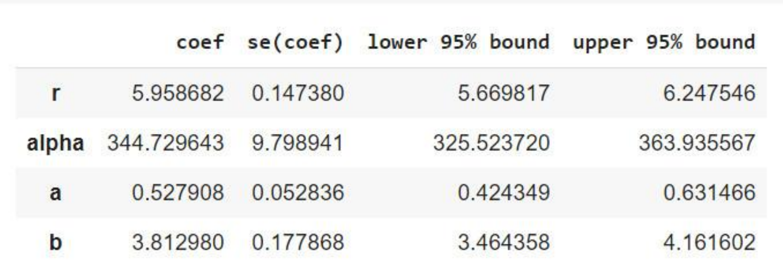
CLV will develop and visualize two models:

BG/ NBD model: A predictive model based on how frequently and how many times clients have recently made purchases (Frequency and Recency).

Gamma-Gamma Submodel: Predictive model based on client financial worth and buy frequency.

4.1.1. Frequency/ Recent usage analysis

BG/ NBD model: Predictive model based on frequency and number of recent purchases by customers (Frequency and Recency). The BG/ NBD model may be built with 6860 acceptable values.



	coef	se(coef)	lower 95% bound	upper 95% bound
r	5.958682	0.147380	5.669817	6.247546
alpha	344.729643	9.798941	325.523720	363.935567
a	0.527908	0.052836	0.424349	0.631466
b	3.812980	0.177868	3.464358	4.161602

Figure 36: BG/ NBD model

After visualizing the customer's frequency/recent purchase matrix, we have the following observations:

Customers recently stopped buying from your business, can we tell that they will no longer use our products and services? On the other hand, can we assume that clients who now make frequent purchases from the firm will continue to do so in the future?

To get a correct assessment of customer loyalty as well as the tendency to leave, we can visualize this relationship using the frequency/ recent matrix, which calculates the number of the expected number of transactions a customer will make over the next period, based on most recent (age of last purchase) and frequency (number of repeat transactions that person makes).

As we can see, our top clients are in the bottom right corner; because they have lately made several purchases, we have high hopes that they will continue to do so.

Our least promising clients, on the other hand, may be seen in the upper right corner; they frequently make purchases before ceasing to visit, and we haven't seen them in months. They most likely found another store (or were shut down) at that point.

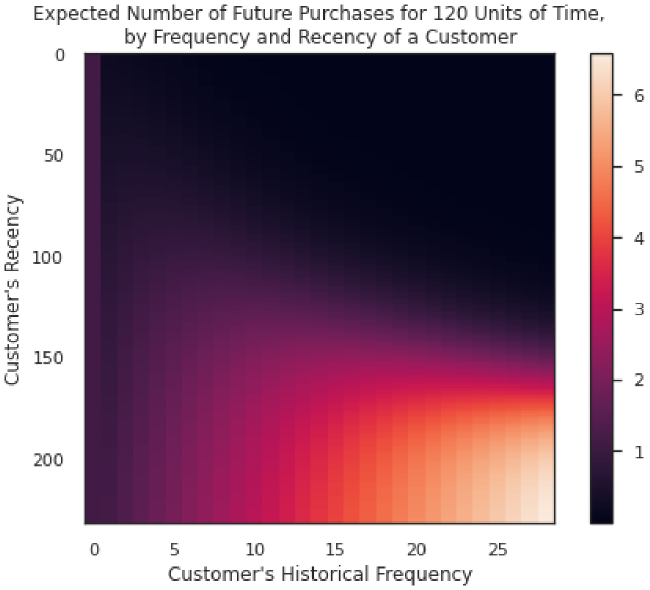


Figure 37: Frequency/ recent matrix based on lifetime of customers

To make sure that customers are still using your services and products, you can consider them through:

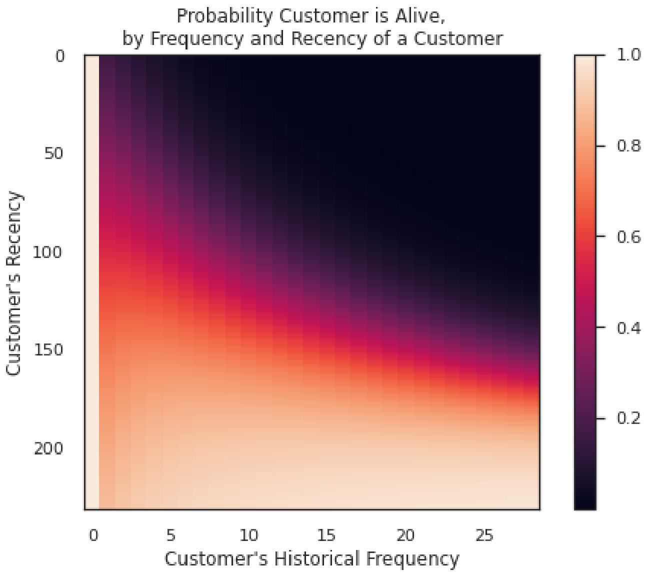
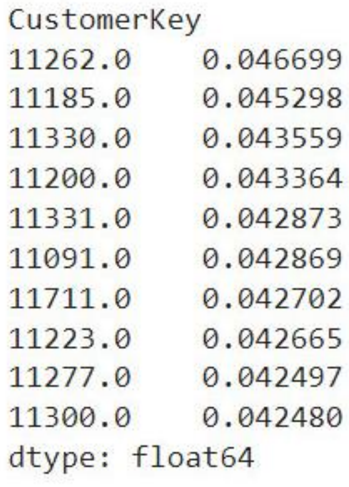


Figure 38: Probability alive matrix based on lifetime of customers

It makes sense that the greatest and worst clients are located in the top right and bottom right corners of the matrix, respectively.

After visualizing the purchase frequency and building the next BG/ NBD model, we will proceed to predict some expected values as follows:

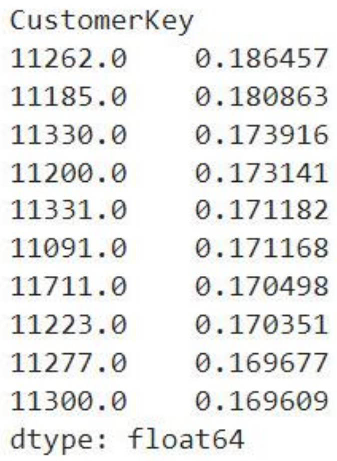
Firstly, we can see top 10 customers expected to make the most purchases in a week:



```
CustomerKey
11262.0    0.046699
11185.0    0.045298
11330.0    0.043559
11200.0    0.043364
11331.0    0.042873
11091.0    0.042869
11711.0    0.042702
11223.0    0.042665
11277.0    0.042497
11300.0    0.042480
dtype: float64
```

Figure 39: Top 10 customers make the most purchases in 1 week

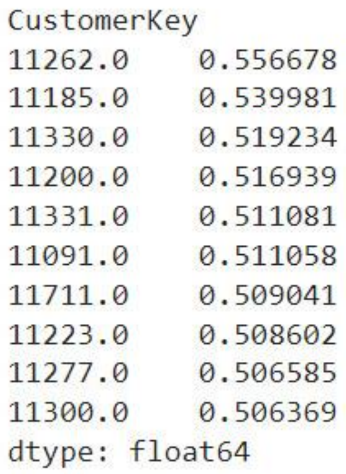
Here are the top 10 customers who are expected to make the most purchases in the month:



```
CustomerKey
11262.0    0.186457
11185.0    0.180863
11330.0    0.173916
11200.0    0.173141
11331.0    0.171182
11091.0    0.171168
11711.0    0.170498
11223.0    0.170351
11277.0    0.169677
11300.0    0.169609
dtype: float64
```

Figure 40: Top 10 customers make the most purchases in 1 month

We can also find the top 10 customers who are expected to buy the most in 3 months:



```
CustomerKey
11262.0    0.556678
11185.0    0.539981
11330.0    0.519234
11200.0    0.516939
11331.0    0.511081
11091.0    0.511058
11711.0    0.509041
11223.0    0.508602
11277.0    0.506585
11300.0    0.506369
dtype: float64
```

Figure 41: Top 10 customers make the most purchases in 3 months

Next, we proceed to evaluate the fit of the model. We find that the actual values are quite different from what the BG/ NBD model predicts. Therefore, the model's predictions are for reference only. If businesses want to make the right marketing plans, they need to consider many other factors.

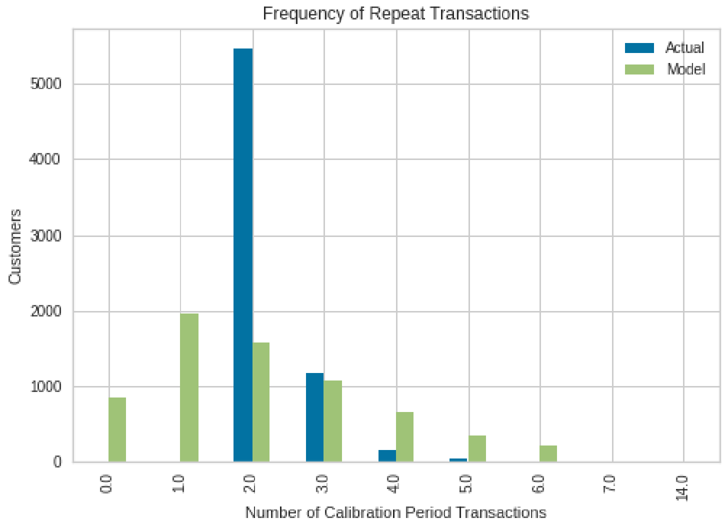
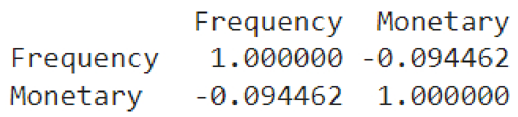


Figure 42: Correlation between BG/ NBD model and reality

4.1.2. Estimating customer lifetime value using a Gamma-Gamma model of monetary value

We evaluate the correlation between Monetary factor and Frequency factor:



```
           Frequency  Monetary
Frequency   1.000000 -0.094462
Monetary   -0.094462  1.000000
```

Figure 43: Correlation between Monetary factor and Frequency factor

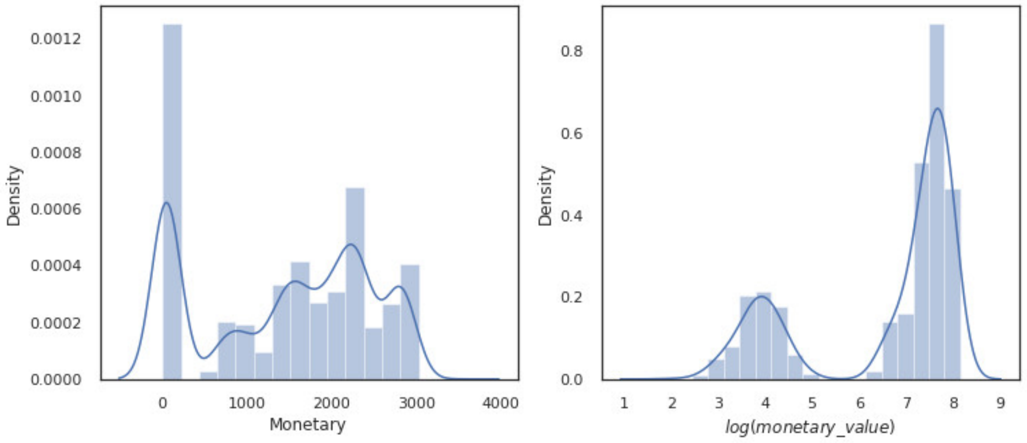
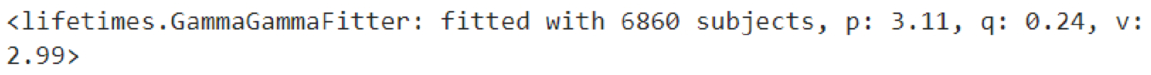


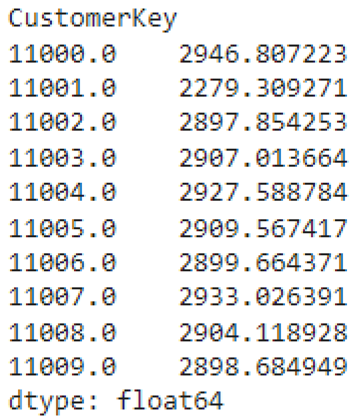
Figure 44: The graph shows the correlation between Monetary factor and Frequency factor

Similar to the BG/ NBD model, the Gamma-Gamma submodel is also available to 6860 customer objects.



```
<lifetimes.GammaGammaFitter: fitted with 6860 subjects, p: 3.11, q: 0.24, v:
2.99>
```

After applying the Gamma-Gamma model, we can now estimate the average transaction value per customer.

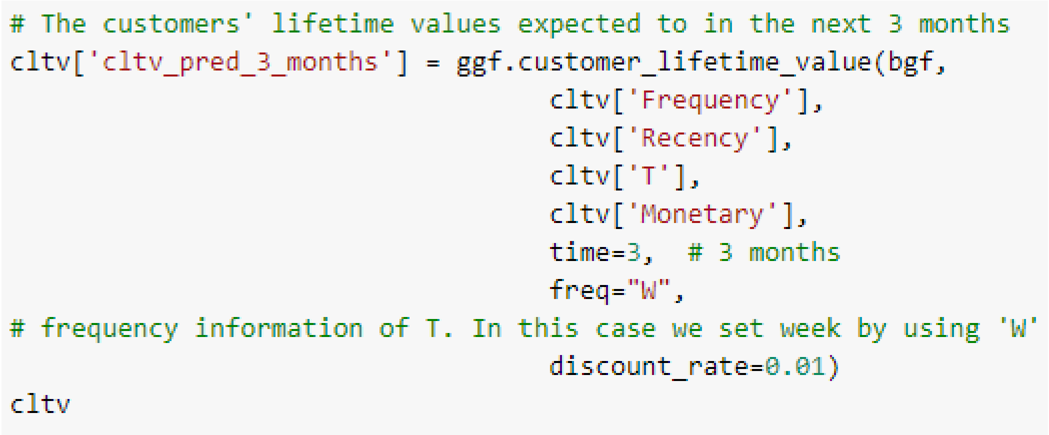


```
CustomerKey
11000.0    2946.807223
11001.0    2279.309271
11002.0    2897.854253
11003.0    2907.013664
11004.0    2927.588784
11005.0    2909.567417
11006.0    2899.664371
11007.0    2933.026391
11008.0    2904.118928
11009.0    2898.684949
dtype: float64
```

Figure 45: Average transaction value per customer

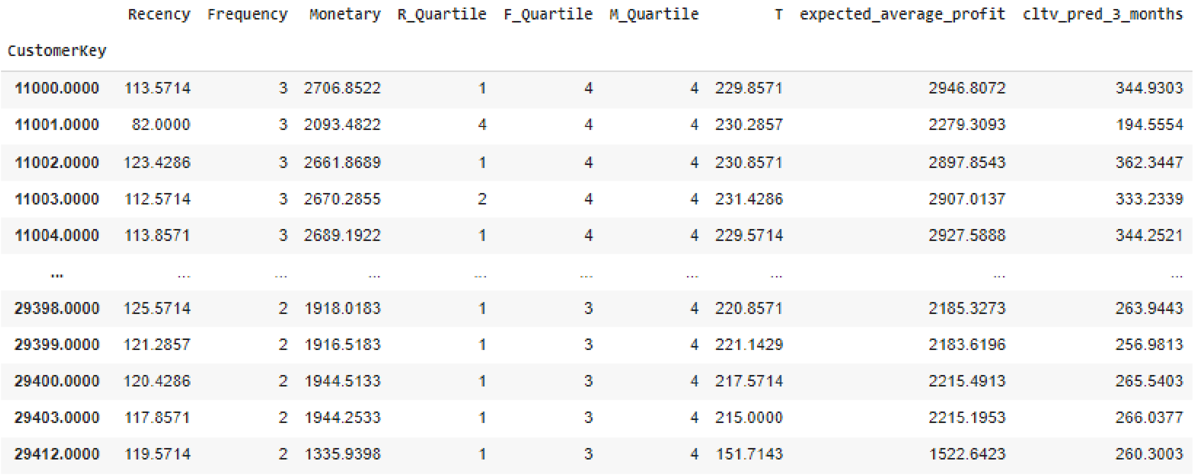
4.1.3. Predicting CLV by using BG/ NBD and Gamma-Gamma models

To build a general and accurate CLV model requires a combination of two models: BG/ NBD and Gamma-Gamma models.



```
# The customers' lifetime values expected to in the next 3 months
cltv['cltv_pred_3_months'] = ggf.customer_lifetime_value(bgf,
                                   cltv['Frequency'],
                                   cltv['Recency'],
                                   cltv['T'],
                                   cltv['Monetary'],
                                   time=3,  # 3 months
                                   freq="W",
# frequency information of T. In this case we set week by using 'W'
                                   discount_rate=0.01)
cltv
```

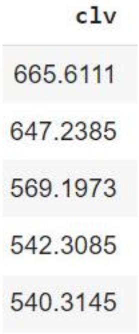
Figure 46: Building CLV models based on BG-NBD and Gamma-Gamma models



	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	T	expected_average_profit	cltv_pred_3_months
CustomerKey									
11000.0000	113.5714	3	2706.8522	1	4	4	229.8571	2946.8072	344.9303
11001.0000	82.0000	3	2093.4822	4	4	4	230.2857	2279.3093	194.5554
11002.0000	123.4286	3	2661.8689	1	4	4	230.8571	2897.8543	362.3447
11003.0000	112.5714	3	2670.2855	2	4	4	231.4286	2907.0137	333.2339
11004.0000	113.8571	3	2689.1922	1	4	4	229.5714	2927.5888	344.2521
...	...	...	...	...	...	...	...	...	...
29398.0000	125.5714	2	1918.0183	1	3	4	220.8571	2185.3273	263.9443
29399.0000	121.2857	2	1916.5183	1	3	4	221.1429	2183.6196	256.9813
29400.0000	120.4286	2	1944.5133	1	3	4	217.5714	2215.4913	265.5403
29403.0000	117.8571	2	1944.2533	1	3	4	215.0000	2215.1953	266.0377
29412.0000	119.5714	2	1335.9398	1	3	4	151.7143	1522.6423	260.3003

Figure 47: CLV model

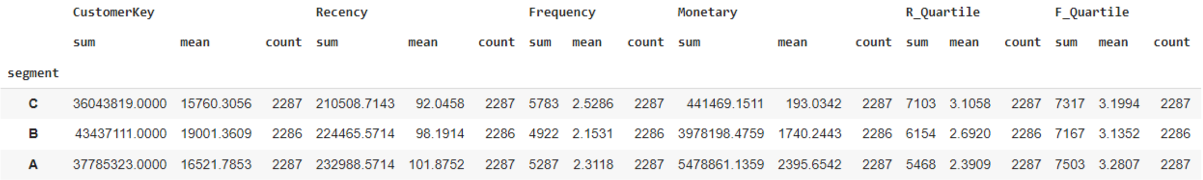
In addition, we can also use CLV values to conduct customer segmentation. With the CLV data just calculated:



clv
665.6111
647.2385
569.1973
542.3085
540.3145

Figure 48: CLV value

Next, we label and classify customer groups by CLV values:



	CustomerKey			Recency			Frequency			Monetary			R_Quartile			F_Quartile		
	sum	mean	count	sum	mean	count	sum	mean	count	sum	mean	count	sum	mean	count	sum	mean	count
segment																		
C	36043819.0000	15760.3056	2287	210508.7143	92.0458	2287	5783	2.5286	2287	441469.1511	193.0342	2287	7103	3.1058	2287	7317	3.1994	2287
B	43437111.0000	19001.3609	2286	224465.5714	98.1914	2286	4922	2.1531	2286	3978198.4759	1740.2443	2286	6154	2.6920	2286	7167	3.1352	2286
A	37785323.0000	16521.7853	2287	232988.5714	101.8752	2287	5287	2.3118	2287	5478861.1359	2395.6542	2287	5468	2.3909	2287	7503	3.2807	2287

Figure 49: Label and classify customers based on CLV value

4.2. 3D Scatter plot with Plotly

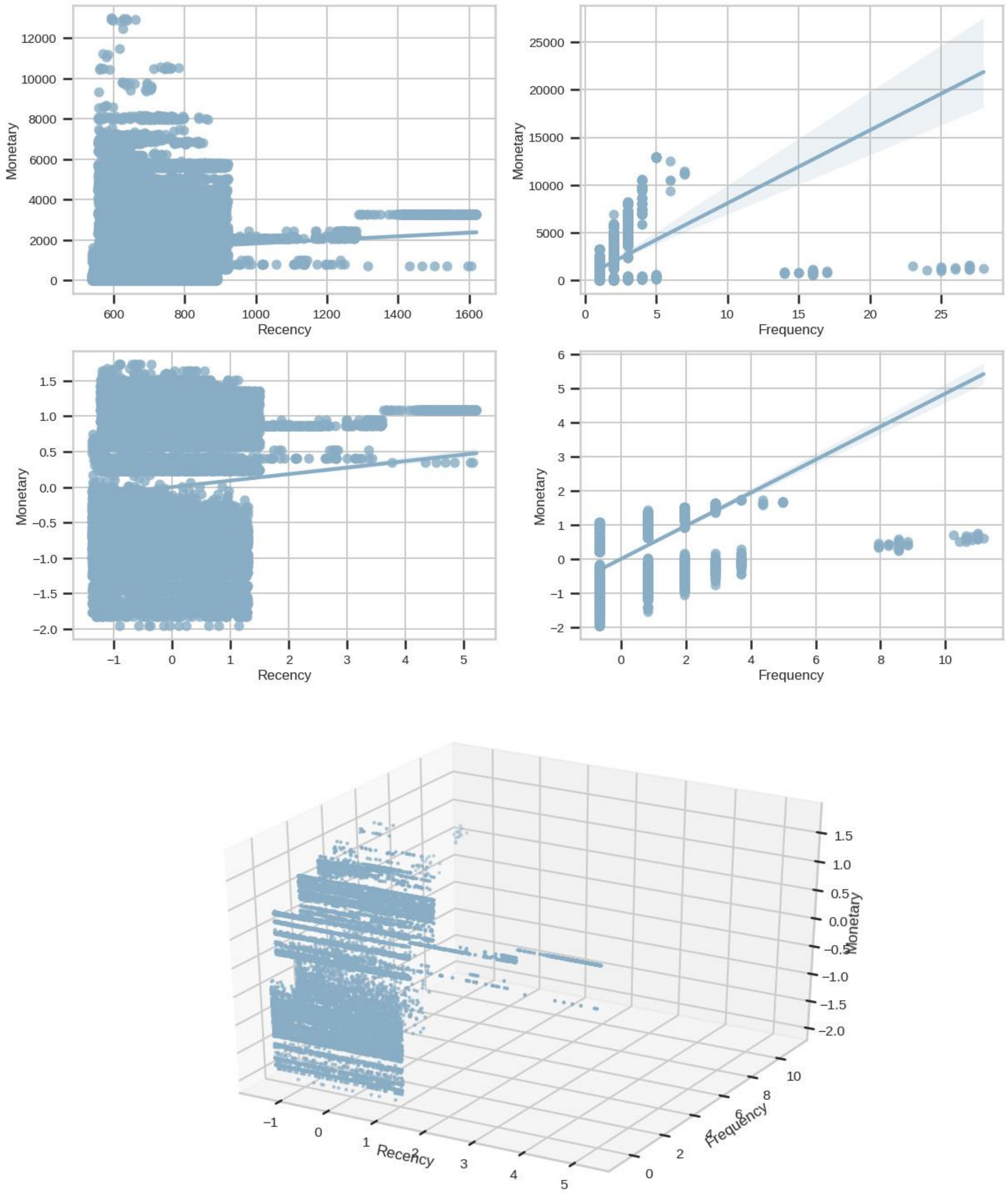


Figure 50: 3D Scatter plot with Plotly

The clear patterns we can notice from the plots above are that customers who buy with a greater frequency and more recency prefer to spend more based on the increasing trend in Monetary (amount value) with a matching increasing and declining trend for Frequency and Recency, respectively.

4.3. Cluster visualization

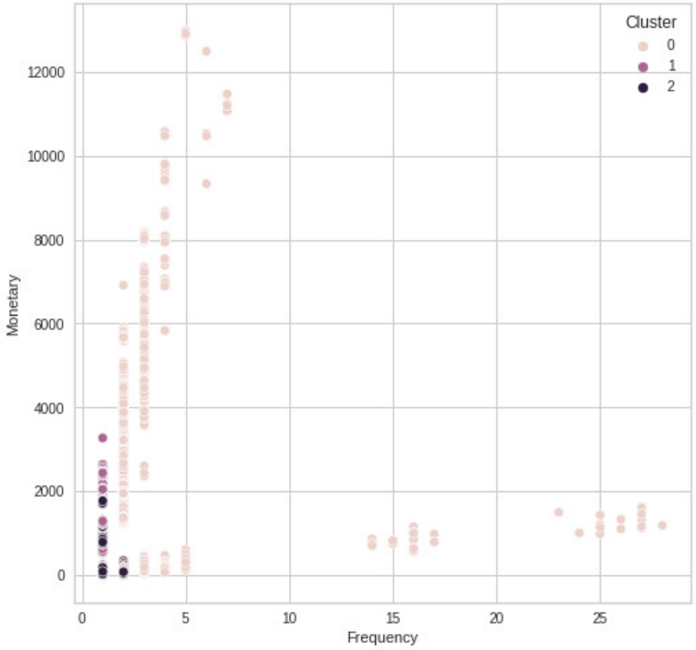


Figure 51: Scatterplot of Clusters

The highest frequency and monetary worth is Cluster 0. Cluster 2 has the lowest frequency and monetary value among the clusters because it only represents a small portion of new and emerging clients. Cluster 1 lies halfway between the two poles, it denotes a tiny percentage of consumers who do not frequent the store and do not spend a lot of money there. The scatter plot demonstrates the high average monetary value and high frequency of occurrence of the group of loyal customers. As a result, it is reasonable to assume that the business's client retention index is extremely high. From here, we will come up with the right marketing strategy analyzed in the next section to not only keep the percentage of loyal customer groups high, but also drive the remaining two groups of customers to increase.

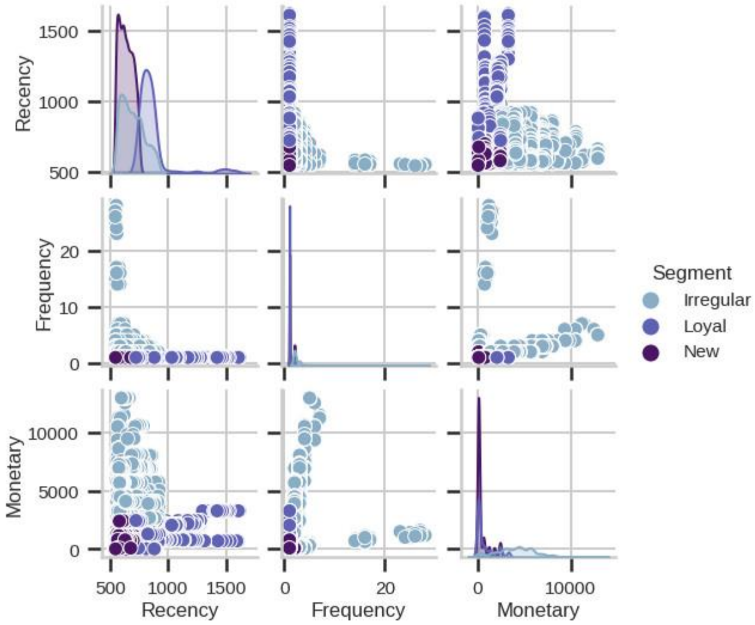


Figure 52: Pairplot of Segment

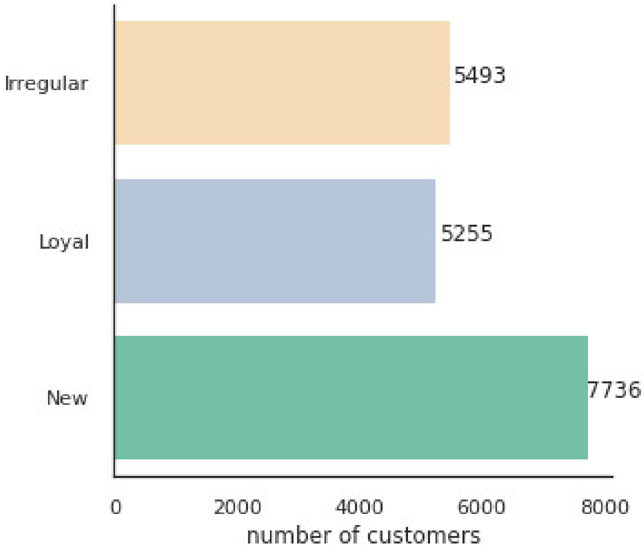


Figure 53: Barplot of number of customers

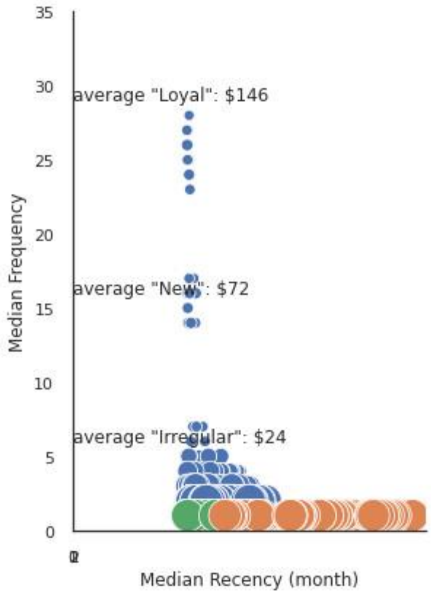


Figure 54: Scatterplot of average of segment based on Median Frequency and Median Recency

Our customers were divided into three categories:

(1) Irregular Consumers: These customers buy frequently yet spend the least amount of money.

(2) Loyal Consumers: This group shopped frequently (though not as frequently as irregular customers) and spent a lot of money.

(3) New Consumers: We were shocked by the enormous number of new customers. They did not begin shopping recently, therefore they do not buy frequently and do not spend a lot of money.

4.4. Snake Plot

We use Snake Plot to:

- Technique for comparing different market segments.
- Each segment's properties are shown graphically.
- Plot the average normalized values of each attribute for each cluster.

We start by separating each client cluster:

- Cluster 1: Irregular Customers.
- Cluster 0: Loyal Customers.
- Cluster 2: New Customers.

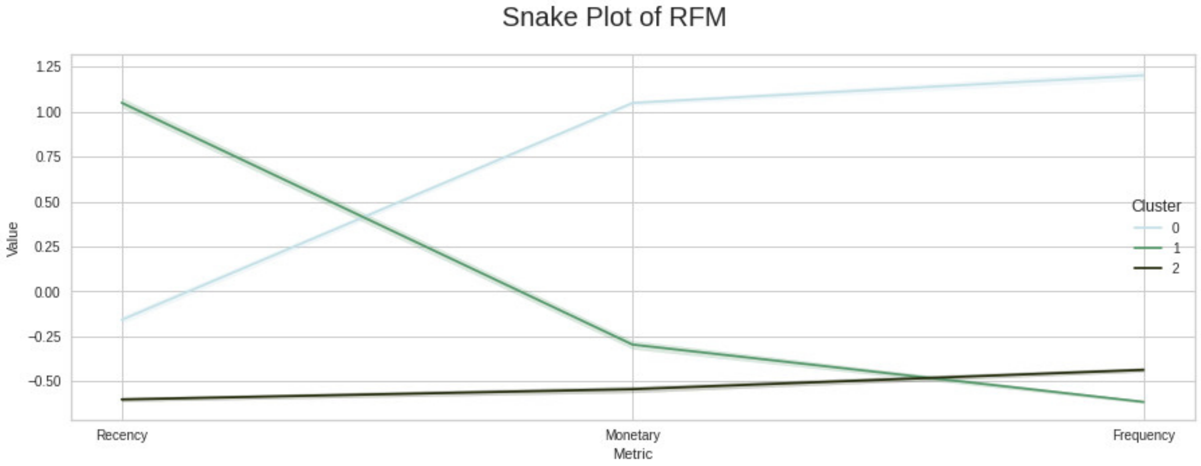


Figure 55: Snake Plot of RFM

The above graph clearly shows the behavior of customers in different clusters:

Cluster 0 customers have the highest frequency and monetary value, as well as a medium recency value. These are the company's most valuable customers.

Cluster 2 customers have lower recency and monetary rates than clusters 0 and 1, but not the lowest frequency rate. These are new clients that have shopped with the company. Although the aggregate worth is currently low, it is clear that they have future potential. To raise the residual value, it is vital to have a strategy to promote this group of clients so that they can make a purchase on their first visit to the firm.

Cluster 1 customers differ in several ways, including the highest recency rate, lowest frequency ratio, and low average monetary value. This client group shops at the business the fewest frequently of the remaining clusters, but when they do, their orders are very large. Despite the large number of purchases, the monetary value of those orders is fairly low; this is a group of customers that do not generate much income for the company.

4.5. Heat Map

We will use a heat map to show the relative relevance of each attribute in each of the three client groupings, or clusters. It computes significance scores by dividing them by one and subtracting one (ensures 0 is returned when cluster average equals population average).

The greater the deviation from zero, the more essential that feature is for a segment in comparison to the whole population.

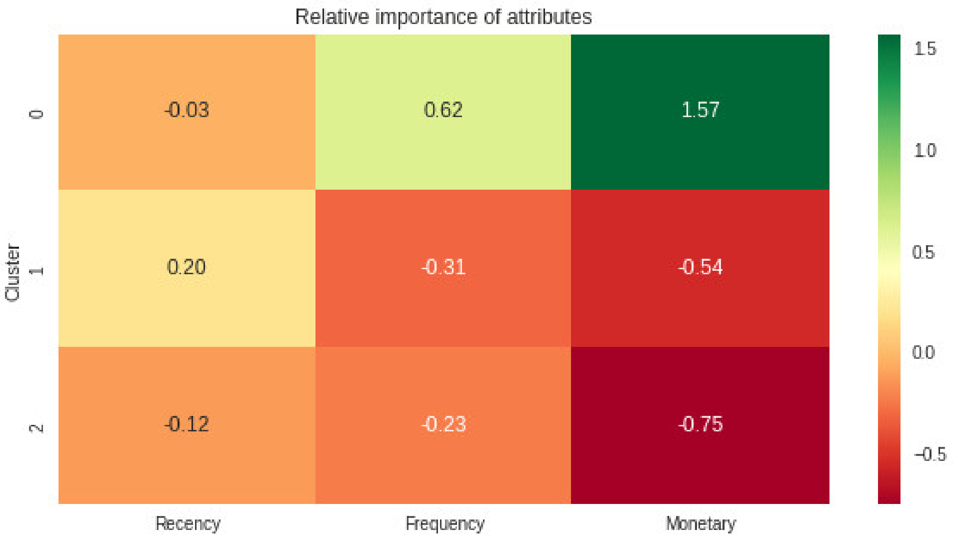


Figure 56: Relative importance of attributes through Heatmap

Through the graph above, we can see the following:

The frequency and monetary properties in cluster 0 are both of high relative importance.

Again evident that the cluster 2 which is a segment of potential customers shows an average similar importance to all three attributes as shown in the snake plot.

In cluster 1 the all attributes are relatively opposite in valuation as compared to cluster 0.

Conclusion and Future works

After researching RFM model and other models and algorithms such as: K-means, Silhouette, CLV promotes the process of analyzing, evaluating and classifying customers optimally. Additionally, it enables us to maximize the value of unknowable data and transform it into known knowledge. These data become more accurate and assist us in creating the best company outline after being normalized, examined, and checked for outliers.

We may create marketing strategies and customer service programs tailored to different demographics using calculated and studied data:

Irregular Customer (Cluster 1) are more likely to have left the company in the past or be on the verge of doing so. Businesses in this situation need to make enquiries, exercise caution, and develop a plan to entice customers back to their operations. Firstly, a survey to determine the cause is required. Once the reason is identified, the business assesses, analyzes, and provides solutions for the majority of customers' concerns. Businesses need to use measures to encourage customers to purchase more expensive goods and more frequently, such as offering limited-time deals, once they have identified the causes of customers leaving term, advice given on the basis of previous purchases. Additionally, to leave a positive impression on their customers, the business need to have a solid marketing and customer service plan. The first impression that the consumer group has of this store is crucial since it affects the company's income from this customer group.

For resolving the issues faced by this client group, we have a suggestion. Make a marketing plan first, then adorn the shop and use various sources to draw customers in. Businesses also need to set up a division that gives consumers the best care and assistance possible. In particular, allow customers to experience the company's high-quality items. There is a basis for a consumer who does not initially want to purchase to become confused and ultimately decide to do so.

Adventureworks has to understand how to respect the values that this consumer segment appreciates when they make purchases from the business for the group of Loyal Customers (Cluster 0). We may create campaigns to keep clients by providing them with promotional coupons or thank-you presents on key occasions, as well as frequent visits and help when required. Customers in this demographic will feel valued as a result and be more inclined to patronize the company for an extended period of time. The business has to understand the purchasing preferences of this consumer segment in order to do that.

Because this particular client group spends an exceptionally high amount of money on average, they are also the simplest to influence to buy if they employ the proper strategy to influence what they desire (what firms are doing well). They'll always be eager to pay a fair price for the goods. Give thanks to this particular set of clients by sending them presents like texts and gift cards on key occasions, which will help them remember the company. Encourage sales by, for example, giving away presents to consumers who spend more than $4200 on merchandise (with Mean Monetary = 4093.39, then when paying for a little greater value (12%)) so that buyers may think again. simpler to pay). Additionally, never forget that taking care of these particular consumers is crucial since they account for the majority of business’s earnings. As a result, business must do all in the power to provide them a great shopping experience new things exist. The business must not only uphold its core beliefs but also cultivate strengths in a variety of areas so that each time a consumer shops, they are unique and distinct from others purchasing previously and eventually get bored.

The New Customers (Cluster 2) have a sizable client concentration but are still hesitant to pay for bulk purchases and are prone to switching brands quickly. We will need to take steps to boost demand and win over this demography trust in our company. The program encourages consumers to make purchases by offering discounts for bulk purchases and actively caring for and offering goods and services. Additionally, firms need to watch their clients closely and provide them specific attention. Change plans and procedures if aberrant points are discovered to prevent these clients from leaving and win their loyalty to the company. This is a prospective consumer base that might help the business grow. This group wants to have a memorable first-time retail experience just like regular consumers do. Make it possible for this set of clients to purchase the store's goods.

After classifying the data, we discovered that one issue is that customer purchases are highly varied and rely too much on revenue from a group with insufficient numbers of consumers. More sensible marketing tactics must be used, and they must be expanded to include a wider range of client segments. Due to this imbalance and reliance, if the business makes a mistake in the future, such as losing its desired values, applying the incorrect marketing plan for the devoted client base would cause this group of consumers to depart at the same time. As a result, the firm is put at danger of experiencing a recession, insolvency, etc.

It is necessary for the constructed model to be reliable and have accurate evaluation capabilities in order to be able to appropriately discover insights. Making the correct strategy, which is a challenging step and requires a lot of effort from managers, also has a significant role in the effectiveness of the firm. Therefore, using the RFM model and having a prediction tool will save time and result in the best efficiency when evaluating, monitoring, and developing company plans.

References

[1] Jedid-Jah Jonker, Nanda Piersma, & Dirk Van den Poel. (2004). Expert Systems with Applications. J*oint Optimization of Customer Segmentation and Marketing Policy to Maximize Long-Term Profitability*, 159–168. https://doi.org/10.1016/j.eswa.2004.01.010

[2] Dennis W K Khong. (2021). Asian Journal of Law and Policy. Rents: *How Marketing Causes Inequality by Gerrit De Geest*, Vol.01(No.01). https://doi.org/https://journals.mmupress.com/index.php/ajlp/article/view/177

[3] Sally Dibb. (2010, December 9). Journal of Strategic Marketing. *New Millennium, New Segments: Moving Towards the Segment of One?*, 193–213. https://doi.org/10.1080/713775742

[4] Edward C Malthouse, & Ralf Elsner. (2006, December 20). Journal of Database Marketing & Customer Strategy Management. *Customisation With Crossed-Basis Sub-Segmentation*, 01 October 2006, 40–50. https://doi.org/10.1057/palgrave.dbm.3250035

[5] Dr. Yanka Aleksandrova. (2018). 18 th International Multidisciplinary Scientific GeoConference SGEM 2018. *APPLICATION OF MACHINE LEARNING FOR CHURN PREDICTION BASED ON TRANSACTIONAL DATA (RFM ANALYSIS)*. https://bom.so/boEBGX

[6] Saharon Rosset, Einat Neumann, Uri Eick, Nurit Vatnik, & Yizhak Idan. (2002, July 23). Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. *Customer Lifetime Value Modeling and Its Use for Customer Retention Planning*, 332–340. https://doi.org/10.1145/775047.775097

[7] Valentin Radu. (2022, October 3). *Consumer behavior in marketing – patterns, types, segmentation*. Omniconvert. Retrieved November 27, 2022, from https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/

[8] C.A. Cole. (2007). Encyclopedia of Gerontology (Second Edition). *Consumer Behavior*, 307–315. https://doi.org/10.1016/B0-12-370870-2/00040-8

[9] Adam Uzialko. (2022, January 29). BUSINESS NEWS DAILY. *What Is Customer Segmentation*? Retrieved November 28, 2022, from https://www.businessnewsdaily.com/15973-what-is-customer-segmentation.html

[10] Kristen Baker. (2022, November 25). HubSpot. *Customer Segmentation: How to Effectively Segment Users & Clients*. Retrieved November 28, 2022, from https://blog.hubspot.com/service/customer-segmentation

[11] Pham Dinh Khanh. (2019, November 8). Khoa học dữ liệu - Khanh's blog. *Model RFM Phân Khúc Khách Hàng*. Retrieved November 28, 2022, from https://phamdinhkhanh.github.io/2019/11/08/RFMModel.html

[12] Casey Murphy. (2022, November 19). Investopedia. Recency, Frequency, Monetary Value (RFM) Definition. Retrieved November 28, 2022, from https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp

[13] Vũ Hữu Tiệp. (n.d.). Machine Learning cơ bản. *K-means Clustering*. Retrieved November 28, 2022, from https://machinelearningcoban.com/2017/01/01/kmeans/

[14] Pulkit Sharma. (2019, August 19). Analytics Vidhya. *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. Retrieved November 30, 2022, from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

[15] AlindGupta. (2022, August 22). GeeksforGeeks. Elbow Method for Optimal Value of K in KMeans. Retrieved November 30, 2022, from https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

[16] Austin Caldwell. (2022, July 21). Oracle Netsuite. What Is Customer Lifetime Value (CLV) & How to Calculate? Retrieved November 29, 2022, from https://www.netsuite.com/portal/resource/articles/ecommerce/customer-lifetime-value-clv.shtml

[17] Pushpa Makhija. (2020, July 30). CleverTap. Cohort Analysis: Beginners Guide to Improving Retention. Retrieved November 30, 2022, from https://a1digihub.com/cohort-analytics-la-gi/