VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FINAL PROJECT REPORT

INTERDISCIPLINARY RESEARCH METHOD COURSE

TOPIC: RESEARCH PURCHASING BEHAVIOR OF CUSTOMERS BY RFM MODELS AND MACHINE LEARNING

Lecturer:

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

Group 4:

1. Lê Huỳnh Anh Thư
2. Võ Minh Thư
3. Nguyễn Cao Minh

Ho Chi Minh City, November 28th 2022

Members of Group 4

No.	Full name	Student ID	Point / 10 (Individual Contribution)	Signature
1	Lê Huỳnh Anh Thư	K214140956	10/10	
2	Võ Minh Thư	K214140957	10/10	
3	Nguyễn Cao Minh	K214140943	10/10	

Acknowledgments

During the time of working on the final project and during the study period, we received a lot of help, suggestions and enthusiastic guidance from the teacher.

I would like to express my sincere thanks to Mr. Ho Trung Thanh - an FDA lecturer and Mr. Nguyen Phat Dat - an assistant professor of FDA. During the study period, I felt the sincere sharing, enthusiasm and concern of the teacher for us. The most important thing that we learn from him is not only the knowledge but also the skills, attitudes ,...

Our final project is not perfect, but it is something we have put a lot of time, mind and heart into. Once again, we would like to thank Mr. Ho Trung Thanh and Mr. Nguyen Phat Dat for equipping us with knowledge, skills and many other things, helping us to complete our final project and prepare for the “life's projects”.

Not perfect but only one.

Group 4

Commitment

My group with three members: Le Huynh Anh Thu, Vo Minh Thu and me, Nguyen Cao Minh, commit that there is no conflict of interest regarding the submission of this project under the guidance of Dr. Ho Trung Thanh, Mr. Nguyen Phat Dat, and refer to the material sources.

Ho Chi Minh City, 19/12/2022

Group 4

Table of Content

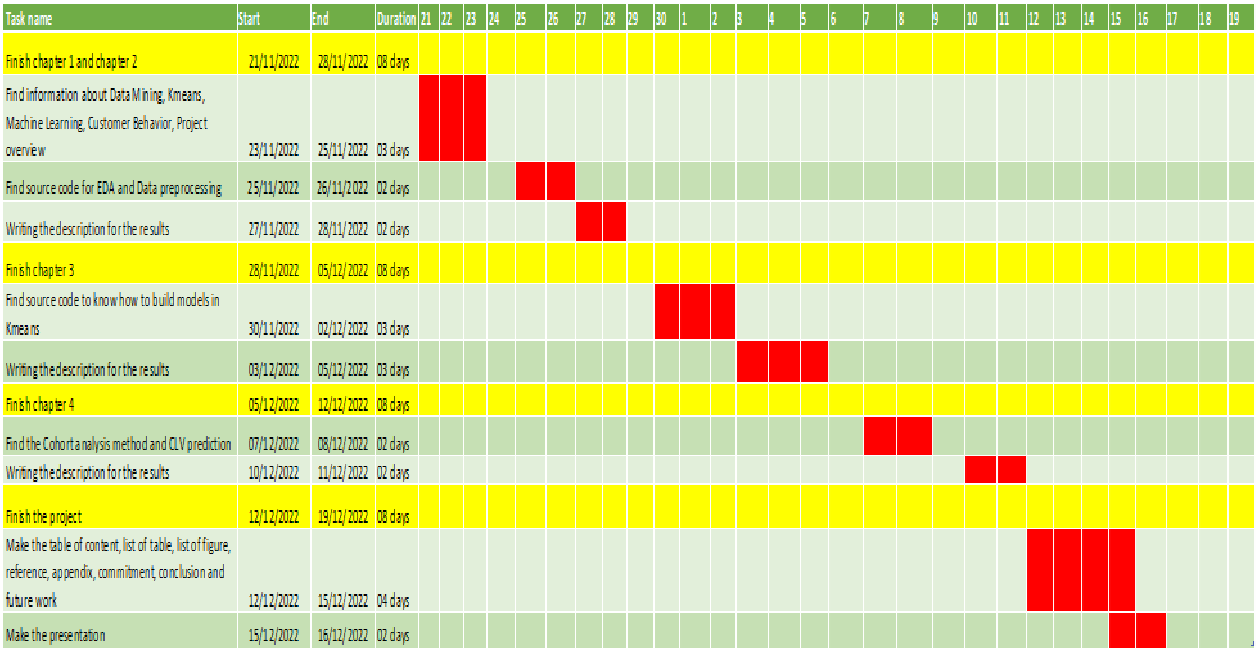
List of Tables

List of Figures

List of Acronyms

DB	Digital Business
MIS	Management Information Systems
LTV	Lifetime Value
FDA	Fundamentals of Data Analytics
RFM	Recency, Frequency and Monetary
IBM	International Business Machines
IQR	Interquartile Range
CLV	Customer Lifetime

GANTT CHART



Task name	Start	End	Duration	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
Finish chapter 1 and chapter 2	21/11/2022	28/11/2022	08 days																													
Find information about Data Mining, Kmeans, Machine Learning, Customer Behavior, Project overview	23/11/2022	25/11/2022	03 days	🟥	🟥	🟥																										
Find source code for EDA and Data preprocessing	25/11/2022	26/11/2022	02 days					🟥	🟥																							
Writing the description for the results	27/11/2022	28/11/2022	02 days							🟥	🟥																					
Finish chapter 3	28/11/2022	05/12/2022	08 days																													
Find source code to know how to build models in Kmeans	30/11/2022	02/12/2022	03 days										🟥	🟥	🟥																	
Writing the description for the results	03/12/2022	05/12/2022	03 days													🟥	🟥	🟥														
Finish chapter 4	05/12/2022	12/12/2022	08 days																													
Find the Cohort analysis method and CLV prediction	07/12/2022	08/12/2022	02 days																	🟥	🟥											
Writing the description for the results	10/12/2022	11/12/2022	02 days																				🟥	🟥								
Finish the project	12/12/2022	19/12/2022	08 days																													
Make the table of content, list of table, list of figure, reference, appendix, commitment, conclusion and future work	12/12/2022	15/12/2022	04 days																						🟥	🟥	🟥	🟥				
Make the presentation	15/12/2022	16/12/2022	02 days																									🟥	🟥			

ABSTRACT

Issues that need to be researched

Knowledge about data mining, RFM, K-Means, CLV, Machine Learning,...

Programming languages (Python)

Statistical tool (Excel)

Database query tool (SQL)

Implementation process

Learn the necessary knowledge

- Analyze the necessary theories for the project implementation
- Understand and try to apply the knowledge learned to the project

Explore Data Analysis

- Data cleaning
- Calculate and describe variable
- Draw a graph showing values and correlations

Customer segmentation with machine learning method

Visualization of experimental results from chapter 3 and predictive analysis of CLV

Project Overview

Business problems

During the operation, businesses face many problems to develop the process of production, brand positioning and extend firm size in the competitive environment. The main goal of all businesses is to know the demands of customers to forecast what happens and plan for the projects to catch up the trend of the fluctuating market. Therefore, they need to analyze data to discover the figure and research suitable methods to manage the process of products, sales and feedback from customers to import the insights for business. However, it is a challenge because of a variety of data and the large amount of money to invest in the technologies, especially small businesses.

The effective solutions for data analysis requires to overcome the challenges, such as: measuring organization performance, eliciting analytical requirements, designing the machine learning solution, ensuring the alignment between analytics initiatives and business strategies. Most businesses want to build the models to analyze data that they want to understand the efficiency of implementation process to estimate the sales quantity to product goods for the next period and forecast the present status of insights that are complicated and influenced by internal and external environments and to follow any updates in production of markets as “just-in-time” (JIT) to make scheduling and arrange the facility utilization. Some statistical methods to tackle this problem such as regression or autoregression and moving average (ARMA) or artificial neural networks (ANNs), fuzzy neural networks (FNNs) that are applied into sales forecasting. However, most of them just analyze supervised and seasonal data and get some difficulties in sudden changes or these methods are hard to apply for the programs of companies.

There are factors which are the reason why we have another solution: RFM (Recency Frequency Monetary) to understand and classify the value of customers to import beneficial insights for businesses to tackle the problems.

Objectives

We want to recommend an effective solution for businesses to build the RFM to research customer’s purchase behavior and know what happens in the operation process. With this method, businesses can assess and classify customers into levels and realize potential customers.

During the research, RFM models prove how it is possible to apply for data processing and the ways to optimize RFM research and find the suitable stages to analyze the figure and input data to result in comprehensive insights.

Objects and scopes

Objects

The object of the research is customer’s behavior and to get a better understanding of how recently they bought, how often they bought, and how much monetary value they contributed to the revenue of the business.

Scopes

Time scope: In the period between November 21, 2022 and December 19, 2022.

Space scope: RFM models and K-means clustering, CLV and Visualization.

Experimental/Research method

RFM analysis, a data mining approach, is used in the project to accomplish the aforementioned goals (Recency, Frequency and Monetary). To categorize customers for future business needs, the simplest strategy is also the most effective. In addition to the conventional approach, K-means clustering and hierarchical clustering are used to compare and select an effective model. In order to calculate the silhouette score for each cluster in K-means clustering, the ideal number of clusters must first be determined as the input. Using the Euclidean method and some connection criteria, hierarchical clustering measures the distance between groups.

Model/Process

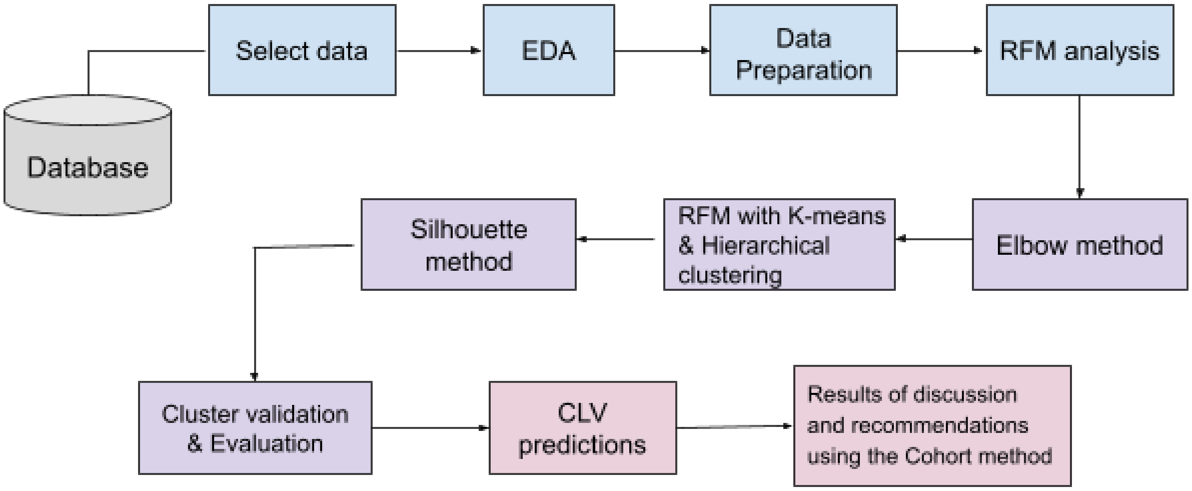

Database
Select data
EDA
Data Preparation
RFM analysis
Elbow method
RFM with K-means & Hierarchical clustering
Silhouette method
Cluster validation & Evaluation
CLV predictions
Results of discussion and recommendations using the Cohort method

Figure 0-1. Data model

Initially, select the data from the database and move it to a new data file in Excel. Using Python as a programming language and Google Colab as a tool, we started cleaning and using EDA to better understand, discover patterns and visualize data. Then the RFM value is calculated to retrieve the data goal with 3 columns (Recency score, Frequency score, Monetary score).In traditional RFM analysis, we divide each field in the target data file into 4 parts, where each part represents 25% of the observations (quartiles) and is labeled with a predefined segment. To perform K-means or hierarchical clustering, these values will be checked for any outliers, and we will see if outliers should be persisted or dropped; later, that normalizes the data to obtain useful models. After clustering similarity data into groups, we test those terms for matches, compare them, and evaluate models. With K-means phrases, the Elbow method is used to determine the number of input phrases, and the Silhouette method is used to determine how good that phrase is. Then, analyze customer groups for the best results, find out how many new customers per month, calculate retention rates according to the heat chart on the matrix distribution and use Cohort analysis to make recommendations.

Tools and Programming language

In this study, we use Python. A popular computer variable programming language for creating websites and software, automating processes, and performing data analysis.

In addition, we use Google Colab, a program that runs on Google's servers. Using Google Colab has many advantages, including the ability to run Python directly on the web without installing it on any device, and sharing for simple teamwork.

Besides, we use excel, a spreadsheet software included in the Microsoft Office suite. This software helps users to record data, present information in tabular form, calculate and process information quickly and accurately with a large amount of data.

Structure of project

This project is divided into 4 corresponding to 4 programs. Along with the main program, it also has appendices and references at the end of the project. Here are the main content sections:

Chapter 1: Theoretical Basis/Background

The theories that were used in this project to create the desired models.

Chapter 2: Data Preparation

Prepare input data for machine learning method

Chapter 3: Customer segmentation with machine learning method

The process of detailed model building, evaluation, comparison, analysis

Chapter 4: Experimental results

Visualization of experimental results and predictive analysis CLV

Chapter 1: Theoretical Basis/Background

1.1 Data mining

Data mining is the exploration and analysis of data to discover meaningful patterns or rules. It is classified as data science, a discipline in the field. Data mining techniques are used to make AI-enabled applications machine learning (ML) model. Examples of data mining in artificial intelligence include things like search engine algorithms and recommender systems.

There are many types of data mining: linear regression, logistic regression, sequentially, classification/regression trees,..

The amount of data generated each year is staggering. And, the already huge number doubles every two years. The digital world is made up of roughly 90% unstructured data, but that doesn't mean the more information, the better the knowledge. Data mining aims to change that, and with it, businesses can:

- Filter through large amounts of repetitive information in an organized manner.

- Extract relevant information and put it to good use for better results.

- Accelerate the pace of informed decision-making.

Data mining can be used in the communications industry, insurance industry, education industry, banking,…

1.1.1 The process of data mining

(1) Define the problem: The first and most important requirement before starting knowledge discovery is to understand the data and the business problem. There must be a clear and unambiguous definition of the goal, that is, to decide what exactly you want to do. For example, when you want to increase the utilization rate of e-mail, what you want to do may be "increase the user utilization rate" or "increase the value of a user's use". The models established to solve these two problems are almost completely different. , a decision must be made.

(2) Establish a data mining library: Establishing a data mining library includes the following steps: data collection, data description , selection, data quality assessment and data cleaning, merging and integration, constructing metadata , loading data mining library, and maintaining data mining library.

(3) Analyze data: The purpose of the analysis is to find the data fields that have the greatest impact on the forecast output, and to decide whether to define derived fields. If the data set contains hundreds or even thousands of fields, it will be very time-consuming and tiring to browse and analyze the data. At this time, you need to choose a tool software with a good interface and powerful functions to assist you in completing these tasks.

(4) Prepare data: This is the last step of data preparation before building a model. This step can be divided into four parts: select variables, select records, create new variables, convert variables.

(5) Build a model: Building a model is an iterative process. Different models need to be carefully examined to determine which model is most useful for the business problem at hand. A part of the data is used to build a model, and the remaining data is used to test and validate the resulting model. Sometimes there is a third data set , called a validation set , because the test set may be affected by the characteristics of the model, and an independent data set is needed to verify the accuracy of the model. Training and testing data mining models requires splitting the data into at least two parts, one for model training and the other for model testing.

(6) Evaluation model: After the model has been built, it is necessary to evaluate the results obtained and explain the value of the model. Accuracy on the test set is only meaningful on the data used to build the model. In practical applications, it is necessary to further understand the types of errors and the related costs caused by them. Experience has shown that an efficient model is not necessarily a correct model. The immediate reason for this is the various assumptions implicit in the building of the model, so it is important to test the model directly in the real world. Apply it in a small area first, obtain test data, and then promote it to a large area after you are satisfied.

(7) Implementation; Once a model has been built and validated, it can be used in two main ways. The first is to provide analysts as a reference; the other is to apply this model to different data sets.

1.1.2 Data Mining Analysis Method

Data mining is divided into supervised data mining and unsupervised data mining. Guided data mining is the use of available data to build a model, which is a description of a specific attribute. Unsupervised data mining is to find some kind of relationship among all the attributes. Specifically, classification, valuation, and prediction belong to supervised data mining; association rules and clustering belong to unsupervised data mining.

(1) Classification: It first selects the training set that has been classified from the data, uses data mining technology on the training set to establish a classification model, and then uses the model to classify the unclassified data.

(2) Valuation: Valuation is similar to classification, but the final output of valuation is a continuous value, and the amount of valuation is not predetermined. Valuation can serve as a preparation for classification.

(3) Predict: It is performed by classification or valuation, and a model is obtained through classification or valuation training. If the model has a high accuracy rate for the test sample group, the model can be used for unknown variables of new samples. Make predictions.

(4) Relevance grouping or association rules: The goal is to discover which things always happen together.

(5) Clustering: It is a method to automatically find and establish grouping rules. It divides similar samples into a cluster by judging the similarity between samples.

1.2 RFM

1.2.1 Introduction

The RFM model is an important tool and means to measure customer value and the ability of customers to create benefits. RFM model is widely mentioned in many customer relationship management (CRM) analysis modes.

According to the research of Arthur Hughes of the American Database Marketing Institute, there are 3 magical elements in the customer database, and these 3 elements constitute the best indicators for data analysis:

Last consumption (Recency)

Consumption frequency (Frequency)

Consumption amount (Monetary)

1.2.2 Last consumption

Definition

Last purchase refers to when the last purchase was made—when was the last time the customer came to the store, what mail order catalog was the last purchase, when was the car bought, or when was the last time the customer bought breakfast at your supermarket.

Customers who have a closer consumption time closer are better customers, and they are most likely to respond to the provision of immediate goods or services. With the current saturated market, when a marketer wants to increase their customer base, they can only take away the customers of their competitors, and to do that they should take advantage of the market. the opportunity to provide good service to new customers, recent customers to make a good impression and encourage them to experience more new facilities. History shows that if we can get consumers to buy, they will keep buying. This is why customers 0-3 months receive more communications from marketers than customers 3-6 months.

Function

The function of consumption is not only the promotional information provided, the consumption report to the data analysis of business. Good marketers regularly review consumer analytics to stay on top of trends. If the monthly report shows that the number of customers who have the last purchase very recently increases, it means that the company is a steady growth company. On the other hand, if the number of customers who have last purchase for one month is decreasing, it means a symptom of the company's inefficient.

Important indicators

Consumption reports are an important indicator of customer retention. Customers who have bought your product, service, or visited a store are the ones most likely to buy. And it is much easier to attract a customer who came to the store 1-3 months ago than to attract a customer who came more than 6 months ago. Building a lasting relationship and solidifying new customer loyalty is something good marketers should pursue.

1.2.3 Consumption frequency

Definition

Consumption frequency (F) is the number of times a customer purchases in a certain period of time. The customers buying most often are also the customers who have the highest satisfaction . You should consolidate the customers' loyalty in brand and store , the consumers who buy most often will be the most loyal. Increasing the number of times customers buy means stealing market share from competitors and earning turnover from others .

Classification

According to this indicator, we divide customers into five equal parts. This five-point analysis is equivalent to a "loyalty ladder". For example, a customer who buys once is a new customer, and a customer who buys twice is a potential customer. Customers who buy three times are regular customers, customers who buy four times are mature customers, and customers who buy five times or more are loyal customers. The trick is to keep the customer moving up the ladder, and think of the sale as moving a double-buyer up into a triple-buyer, and a one-timer into a double-buyer.

Data analysis

The core factor affecting repurchase is the product, so repurchase is not suitable for cross-category comparison. For example, food category and beauty category: food is a "semi-standard product". For consumables, the consumption cycle is short, the purchase frequency is high, and repeated purchases are relatively easy to occur, so cross-category repurchases are not comparable.

1.2.4 Amount of consumption

Spending amounts are the backbone of all database reports and can also validate the "Pareto's Law" - 80% of a company's revenue comes from 20% of its customers. A natural inclination is to put more emphasis on encouraging customers who spend the most money to continue to do so. While this can produce a better return on investment in marketing and customer service, it also runs the risk of alienating customers who have been consistent but have not spent as much with each transaction.

1.2.5 Application meaning

The RFM model dynamically displays the entire outline of a customer, which provides a basis for personalized communication and services. At the same time, if the time of dealing with customers are long enough, it can accurately judge the long-term value of the customer (even is the lifetime value), by improving the status of the three indicators, thus providing support for more marketing decisions.

RFM is very suitable for enterprises that produce a variety of commodities, and the unit price of these commodities is relatively low, such as consumer goods, cosmetics, small appliances, video stores, supermarkets,...

RFM should not be used too much, and customers who cause high transactions continue to receive letters. Every enterprise should design a customer contact frequency rule, such as sending a thank you call or email within three days or within a week of purchase, and actively caring about whether consumers have problems with use, and sending out inquiries about whether they are satisfied with the use after one month, and Three months later, they will provide cross-selling suggestions, and begin to pay attention to the possibility of customer loss, and constantly create opportunities to actively contact customers. In this way, the chances of customers repurchasing will also be greatly improved.

When an enterprise implements CRM, it must understand customer differences based on the principles of the RFM model, and use this as the main axis to rebuild the enterprise process in order to innovate performance and profits. Otherwise, it will not be able to gain a foothold in the new century market.

1.3 Customer behavior

1.3.1 Definition

Customer behavior information refers to customer consumption behavior, customer preferences and lifestyles, customer satisfaction, customer loyalty, and contact records with companies and other related information

Analysis of customer behavior is not about looking at the number of customers but looking at customer preferences, how they feel about your service, find out what customers need, ... Understanding these issues can help businesses connect with customers effectively and long-term. There are 3 factors that influence customer behavior: personal, psychological, and social. The completeness of information is very important to measure customer behavior.

The main purpose of customer behavior information is to help enterprise marketing personnel and customer service personnel grasp and understand customer behavior in customer analysis. Customer behavior information reflects the customer’s consumption choices or decision-making process.

1.3.2 Analyze customer behavior

Segment the audience

The first step to a customer behavior analysis is segmenting your audience.

The core of segmentation models are demographic segmentation (sex, age , ...), psychographic segmentation (hobby, personality ...), geographic segmentation, religion or other things.

You’ll also need to identify the characteristics of customers that are most valuable to the business. We can use the RFM model (already mentioned in part 1) to segment.

Identify core values in each customer segment

Each customer will have their own reasons to choose your products/services. Therefore, you need to identify the core values in each customer segment to find out what factors influence the customer's buying decision.

For example, buy for convenience or buy for urgent needs? Are they purposeful when searching for your business brand and how much will they spend? Think about the needs that the customer may have.

Allocate quantitative data

After the above two steps, we will have a qualitative database, you need to try to get accurate sources of quantitative information. Quantitative data can help you answer what your customers do, while qualitative data helps answer why they do it.

Quantitative information can come from inside and outside the business. These sources can shape customer trends not only at the micro level but also at the macro level. For example, website subscriber metrics, social media analytics, or product usage reports, customer reviews and competitor analysis, industry statistics, etc. all come in handy when performing analysis.

Apply to strategy

Once you have identified the outstanding elements with a customer journey map, choose the right communication channels for each customer segment, personalize the brand to potential customers, "kill" obstacles. hesitate to lead them to become loyal to the product/service through every step of the purchase and experience journey.

The results from the analytics also help us know where the campaign needs to be updated.

Before launching a new campaign, businesses can use analytics to determine what customers think about the changes that will take place. Since they often act out of habit and rarely accept change, this is important if you don't want to lose customers.

Analyze the results

After applying for a project, the last part will be to observe and evaluate the results to draw lessons for the next projects. Whether the services you offer are well received by customers, and achieve the expected revenue. Should this strategy be applied for a long time,...

1.4 K-Means

1.4.1 Definition

Clustering is a process of classifying and organizing data members that are similar in some respects. Clustering is a technique for discovering this internal structure. Clustering techniques are often called unsupervised learning.

K-means clustering is the most famous partitioning clustering algorithm, and its simplicity and efficiency make it the most widely used of all clustering algorithms. Given a set of data points and the required number of clusters k, k is specified by the user, the k-means algorithm repeatedly divides the data into k clusters according to a certain distance function.

1.4.2 Algorithm steps

1. The initialized k samples as the initial cluster center a=a1,a2,…ak ;
2. For each sample in the dataset xi calculates its distance to k cluster centers and divides Select it into the class corresponding to the cluster center with the smallest distance;
3. For each category aj , recalculate its cluster center $a_j = \frac{1}{|c_i|} \sum_{x \in c_i} x$ (that is, the centroid of all samples belonging to this class);
4. Repeat the above two steps 2 and 3 until a certain termination condition (number of iterations, minimum error change, ...) is reached.

1.4.3 Advantages and disadvantages

Advantages

Easy to understand, the clustering effect is good, although it is a local optimum, but often the local optimum is enough;

When dealing with large data sets, the algorithm can guarantee better scalability;

When the cluster approximates the Gaussian distribution, the effect is very good;

Algorithm complexity is low.

Disadvantages

The K value needs to be set manually, and the results obtained with different K values are different;

Sensitive to the initial cluster center, different selection methods will get different results;

Sensitive to outliers;

Samples can only be classified into one category, which is not suitable for multi-classification tasks;

Not suitable for too discrete classification, classification of sample class imbalance, classification of non-convex shape.

1.5 CLV(Customer Lifetime Value)

CLV is understood as "customer lifetime value", which is the value that customers return to businesses during their time as customers. In other words, CLV is just a measure of the total revenue that a business can achieve thanks to a good relationship with loyal customers.

Benefits of CLV measurement work

- Measuring the CLV index helps businesses build appropriate strategies to retain customers. Customer retention is one of the important things that businesses need to do to have stable sales and profits. Increasing CLV also depends largely on whether the business can retain customers and turn them into loyal supporters of the business.
- Measuring CLV to help businesses improve ROI (Return On Investment) is an indicator that measures the percentage of revenue over total investment costs, the result of measuring profit efficiency. investment brings.
- Measure the Customer Lifetime Value Index to help businesses increase sales. Measuring the Customer Lifetime Value Index helps businesses determine whether their marketing strategy or customer retention strategy is effective, thereby saving costs and increasing sales.

4 Steps to Measure Customer Lifetime Value

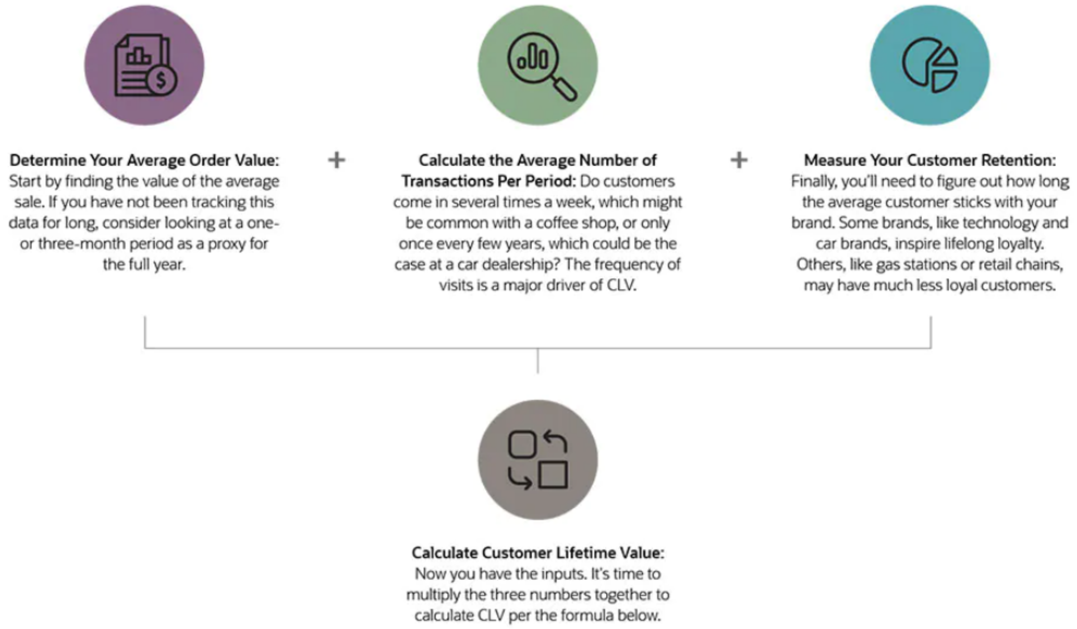

Determine Your Average Order Value: Start by finding the value of the average sale. If you have not been tracking this data for long, consider looking at a one- or three-month period as a proxy for the full year.
+
Calculate the Average Number of Transactions Per Period: Do customers come in several times a week, which might be common with a coffee shop, or only once every few years, which could be the case at a car dealership? The frequency of visits is a major driver of CLV.
+
Measure Your Customer Retention: Finally, you'll need to figure out how long the average customer sticks with your brand. Some brands, like technology and car brands, inspire lifelong loyalty. Others, like gas stations or retail chains, may have much less loyal customers.
Calculate Customer Lifetime Value: Now you have the inputs. It's time to multiply the three numbers together to calculate CLV per the formula below.

Figure 1-1. How to measure CLV

Customer Lifetime Value Calculation

LTV (Lifetime Value) = Average Value of Sale × Number of Transactions × Retention Time Period

CLV = LTV × Profit Margin

1.6 Machine Learning

1.6.1 Definition

Machine learning is part of artificial intelligence and computer science. It focuses on using data and algorithms to reimagine how people learn, absorb knowledge, and gradually improve its accuracy over time.

Machine learning is an important component in the data science field that is trending very strongly today. Through the use of statistical methods, algorithms are also used to classify or predict and find key insights in data mining projects. This information then drives decision making in applications and businesses, and the ideal impact on growth metrics for data scientists will increase. They will ultimately ask for help identifying the most relevant business-related questions and the data to answer them.

1.6.2 How machine learning works

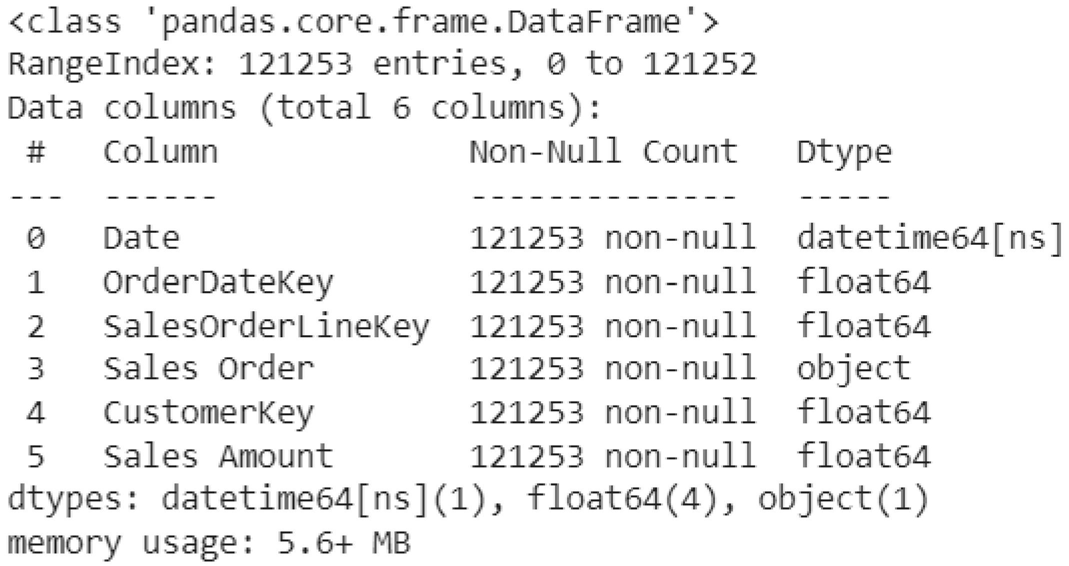
UC Berkeley (linked outside of IBM) divides the learning system of machine learning algorithms into three main parts:

- Decision process: Machine learning algorithms are used to make predictions or classifications. Thereby, through some input data, which may be labeled or unlabeled, the algorithm will make an estimate of a sample in data.
- Error function: Evaluate the model's predictions. Assuming there are known examples, an error function can be compared to evaluate the accuracy of the model.
- Model optimization process : In cases where the model can better fit the data points in the training set, the weights will be adjusted to reduce the difference between the known example and the model estimate. The algorithm then repeats this "evaluate and optimize" process and updates the weights automatically until the correct threshold is reached.

Chapter 2: Data Preparation

2.1 EDA (Exploratory Data Analysis)

2.1.1 Basic information about data

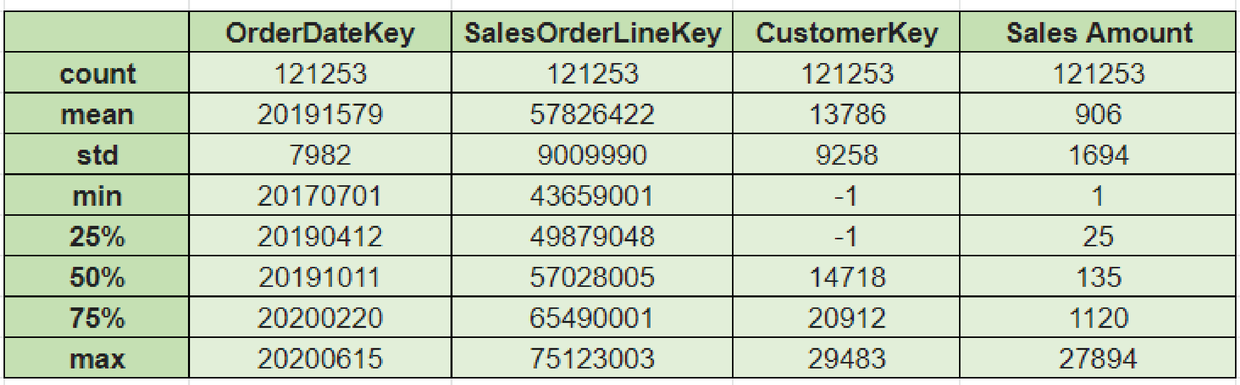


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121253 entries, 0 to 121252
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Date               121253 non-null  datetime64[ns]
 1   OrderDateKey       121253 non-null  float64
 2   SalesOrderLineKey  121253 non-null  float64
 3   Sales Order        121253 non-null  object
 4   CustomerKey        121253 non-null  float64
 5   Sales Amount       121253 non-null  float64
dtypes: datetime64[ns](1), float64(4), object(1)
memory usage: 5.6+ MB
```

Figure 2-1. Basic information results of data

Based on this information results, we can conclude the total of entries of this data. With these figures and insights, we determine that all of the columns are non-null with related values. In addition, types of data mean to realize information and link figures to compare.

2.1.2 Descriptive analysis



	OrderDateKey	SalesOrderLineKey	CustomerKey	Sales Amount
count	121253	121253	121253	121253
mean	20191579	57826422	13786	906
std	7982	9009990	9258	1694
min	20170701	43659001	-1	1
25%	20190412	49879048	-1	25
50%	20191011	57028005	14718	135
75%	20200220	65490001	20912	1120
max	20200615	75123003	29483	27894

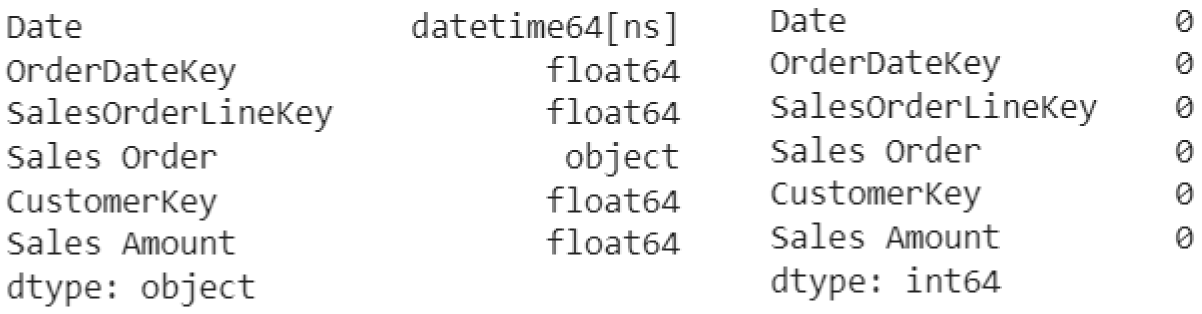
Table 2-1. Descriptive analysis result

Descriptive analysis has functions to measure and review features of data in a general chart. With min and max results , we will know the range of figures to limit the research. Other results means to visualize how data perform.

The count value of all variables is 121253 and it is no null value in this data set. All variables can be determined to analyze.

The min value of OrderDateKey is 20170701 and the max value is 20200615 that means the time range of this data set is about 3 years. The range of sales amount from 1 to 27894 helps us know the revenue of the sales is very large. The revenue is contributed by 31464002 orders.

2.1.3 Explore variables



```
Date                 datetime64[ns]    Date                0
OrderDateKey                float64    OrderDateKey        0
SalesOrderLineKey           float64    SalesOrderLineKey   0
Sales Order                  object    Sales Order         0
CustomerKey                 float64    CustomerKey         0
Sales Amount                float64    Sales Amount        0
dtype: object                          dtype: int64
```

Figure 2-2. Variables data type and missing values counting result

All of the values are determined types to review and all data in this results are analyzed, no missing values.

2.1.4 Histogram

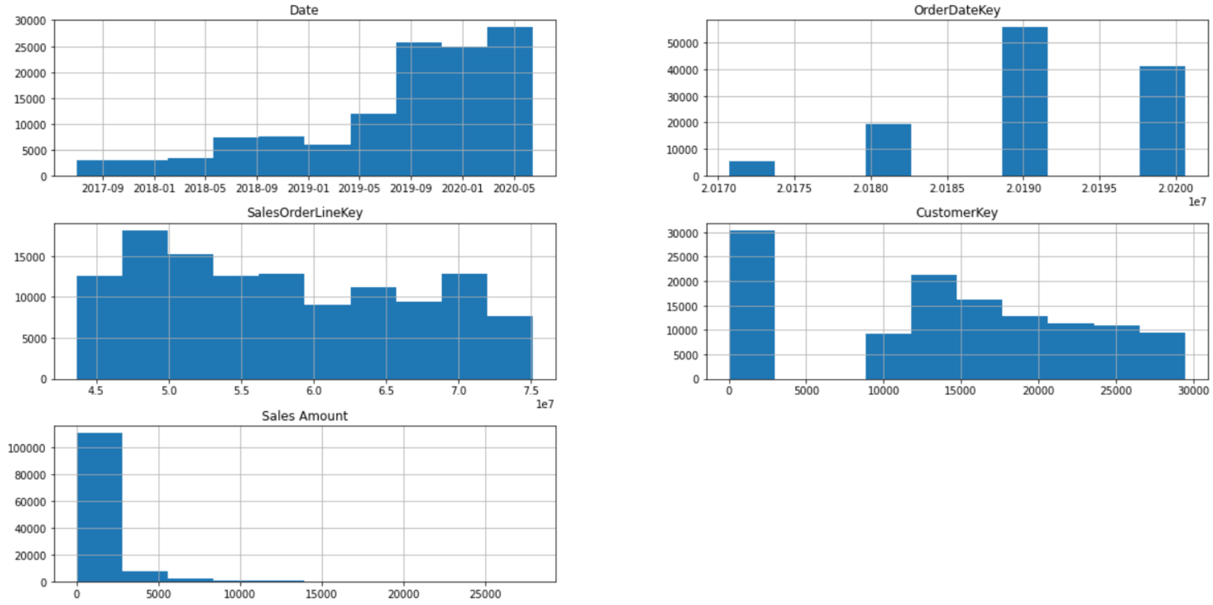


Figure 2-3. Numerical histograms

Through the shape of the distribution that compares the standard values with the distribution of the histogram, the organization can check and evaluate the capabilities of the inputs, control the process, and detect errors. Provides visual information on process variability.

The histogram tells us the following four problems:

- Most frequently occurring value (mode)

- How often each value occurs

- The shape of the part

- Relationship between data and requirements limits

2.1.5 Correlation Plot

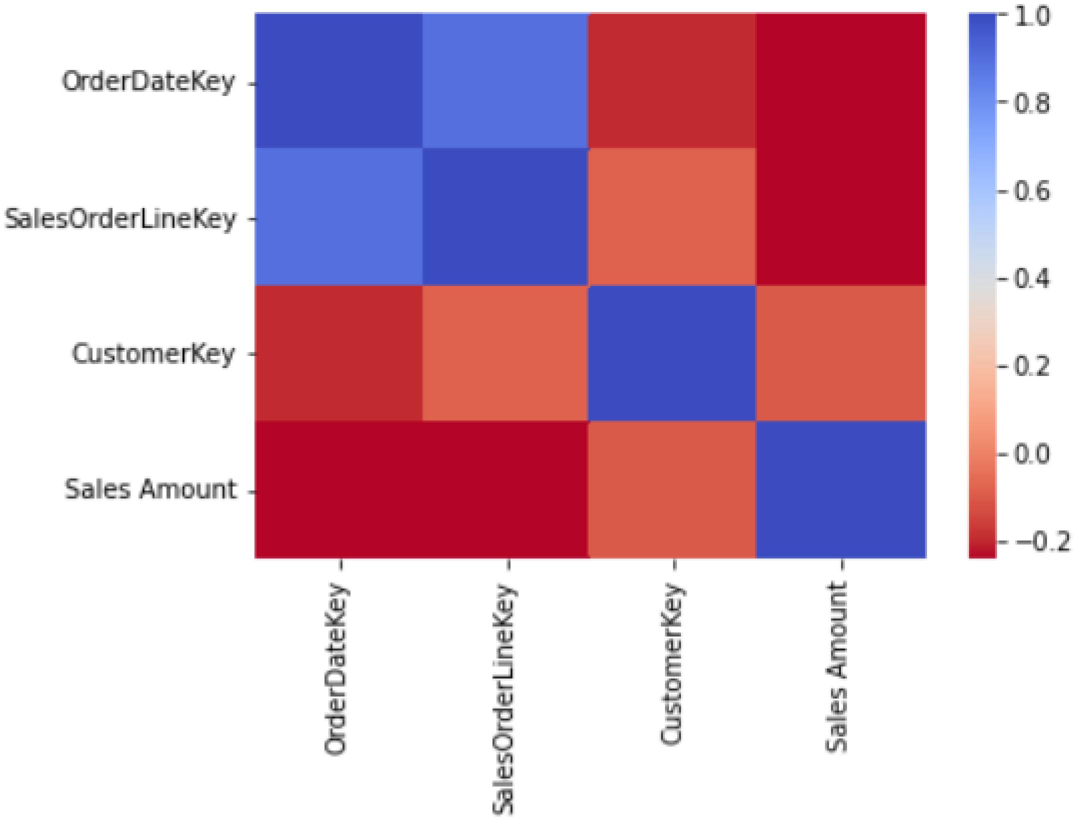


Figure 2-4. Correlation matrix

Before performing regression testing of the model, it is necessary to conduct a correlation analysis between the independent factors and the dependent factors. From there, we will choose the independent factors that are actually correlated with the dependent factor and put those factors into the regression.

2.1.6 Scatter plot

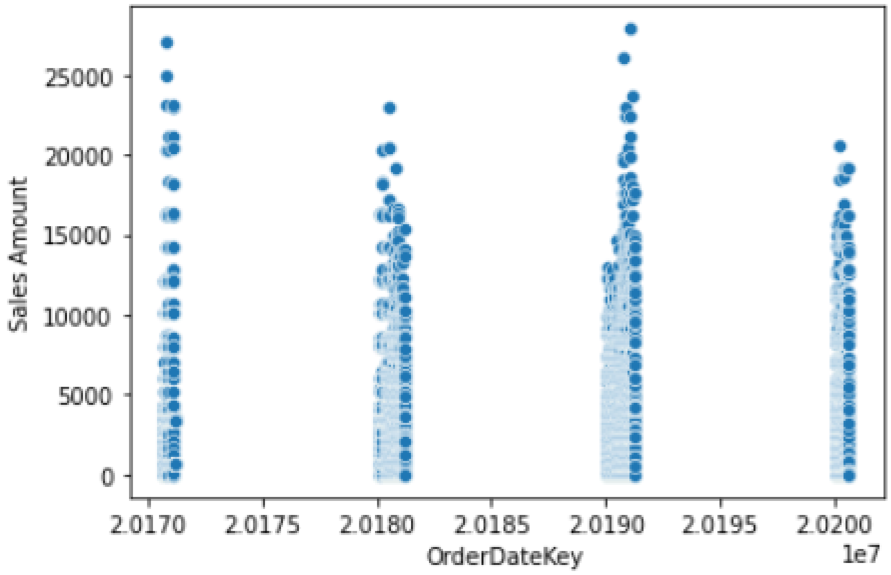


Figure 2-5. Graph illustrating the relationship between Oderdatekey and sales amount.

A quality indicator is created by the combination and impact of many factors. There is a close relationship between quality and factors.

To assess the quality situation, one can use two or more data at the same time to show the correlation between the factors on the graph. (OrderDateKey and Sales Amount)

Through which it is possible to determine the impact trend of the cause under consideration on the specific results achieved.

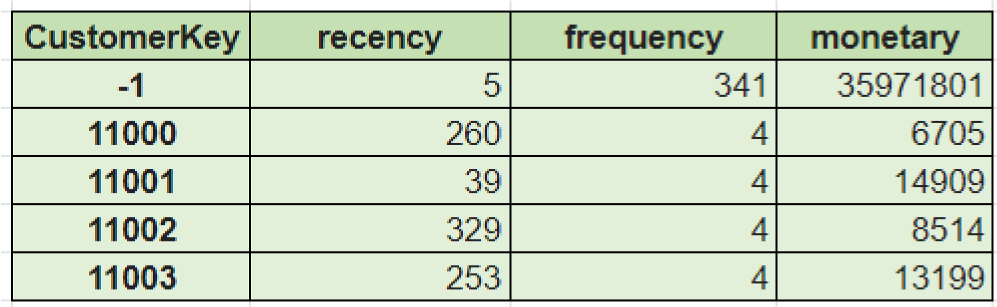
2.2 Data preprocessing

2.2.1 Calculation of R,F and M values

Calculating Recency-values, Frequency-values, and Monetary-values is essential for doing RFM analysis. Despite being extremely skewed, the distribution features are initially computed to get the target data needed for RFM analysis.

- Recency is considered as the last time the customer has made a purchase (the interval between the date of application of the method and the date of the most recent purchase by the customer). The longer a consumer went without purchasing anything from the company, the greater the R value.

- Frequency is the frequency of customer purchases or how many times the customer has purchased
- Monetary is the total amount of money that customers have spent on all shopping activities.



CustomerKey	recency	frequency	monetary
-1	5	341	35971801
11000	260	4	6705
11001	39	4	14909
11002	329	4	8514
11003	253	4	13199

Table 2-2. RFM values

The algorithm is implemented based on the AdventureWorks Sales dataset. The dataset contains 12,1253 transactions. Many of the store's customers are retailers. These transactions occurred over four years, from 2017 to 2020. After surveying and preprocessing, as well as removing unnecessary values and retaining the right ones, the RFM data model was established. With the results presented above Table 2-2

2.2.2 Outliers analysis

The difference in values is enormous. The target data may contain many outliers if the distribution is highly skewed, according to the assumption.

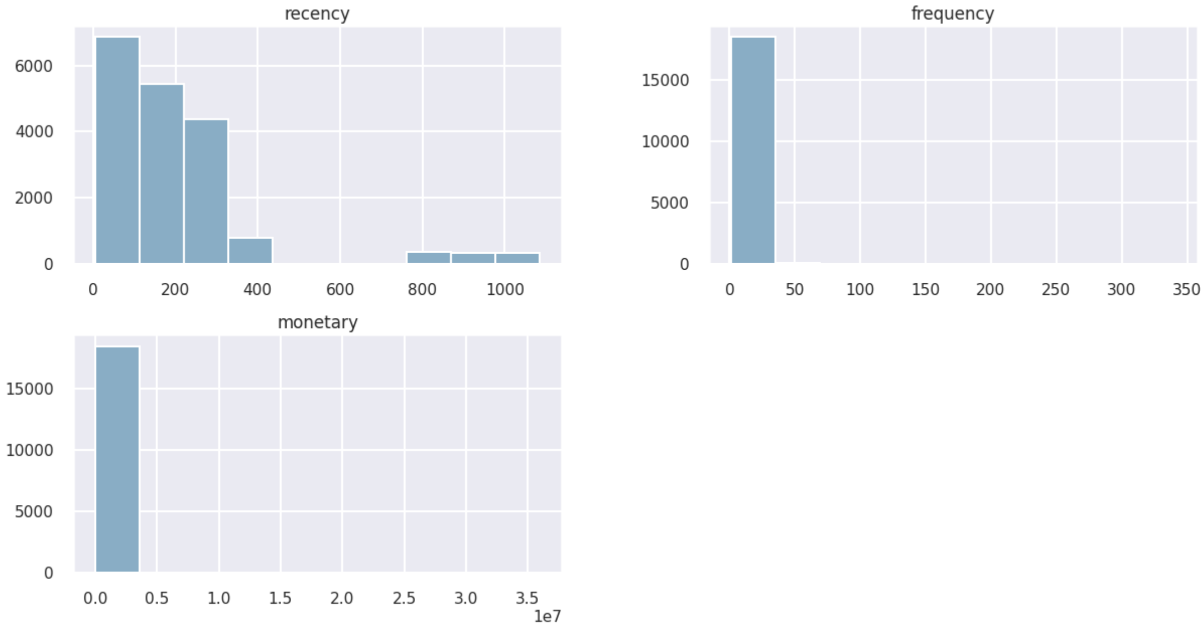


Figure 2-6. Histogram of recency, frequency, monetary

From the graph, it can be seen that Frequency and Monetary have a long tail and a right offset. The top columns contain the majority of the information. About 70% of customers buy once or twice; that is, the frequency of repeat purchases for most customers is not high, and very few customers buy again and again. Similarly, customer costs in the top columns make up the bulk of the total. Scatter values can be outliers that must be removed in order to obtain a reasonable segmentation.

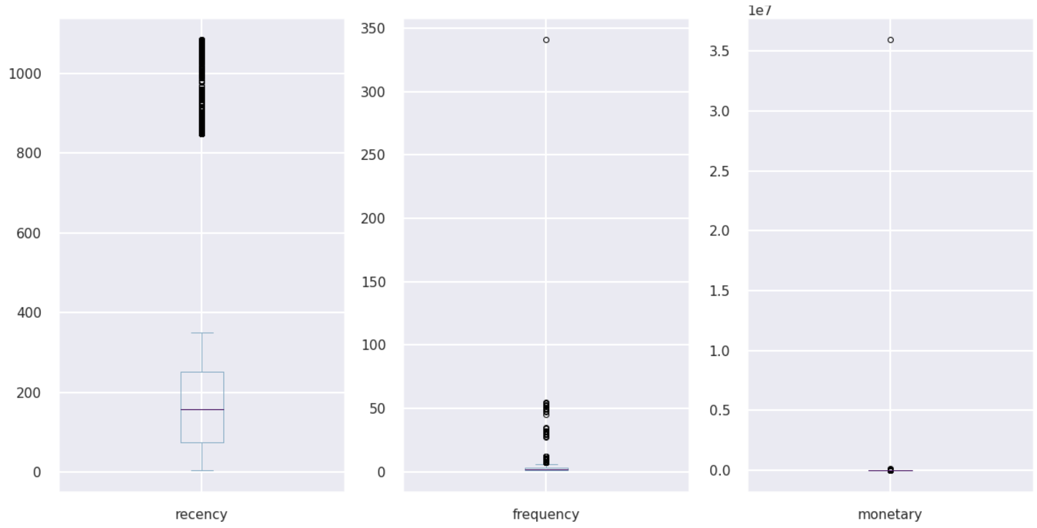
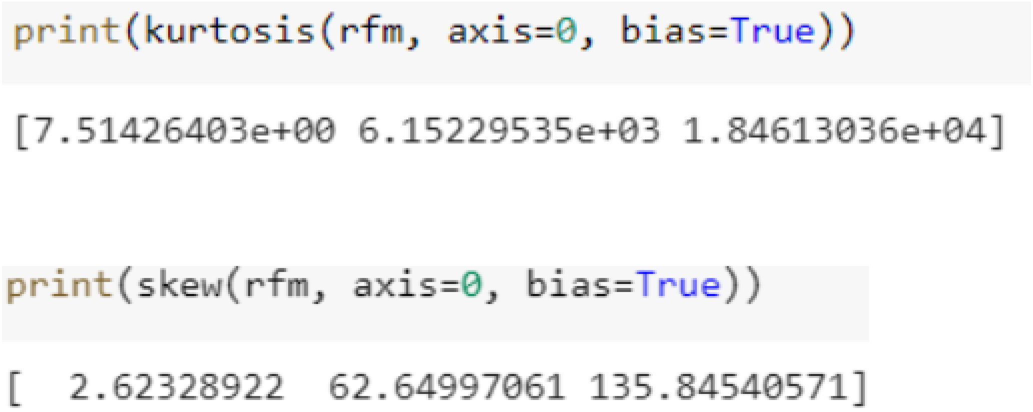


Figure 2-7. Boxplot of recency, frequency, monetary values

Many data points are outside the whisker, as indicated by the three boxplots.



```
print(kurtosis(rfm, axis=0, bias=True))

[7.51426403e+00 6.15229535e+03 1.84613036e+04]
```

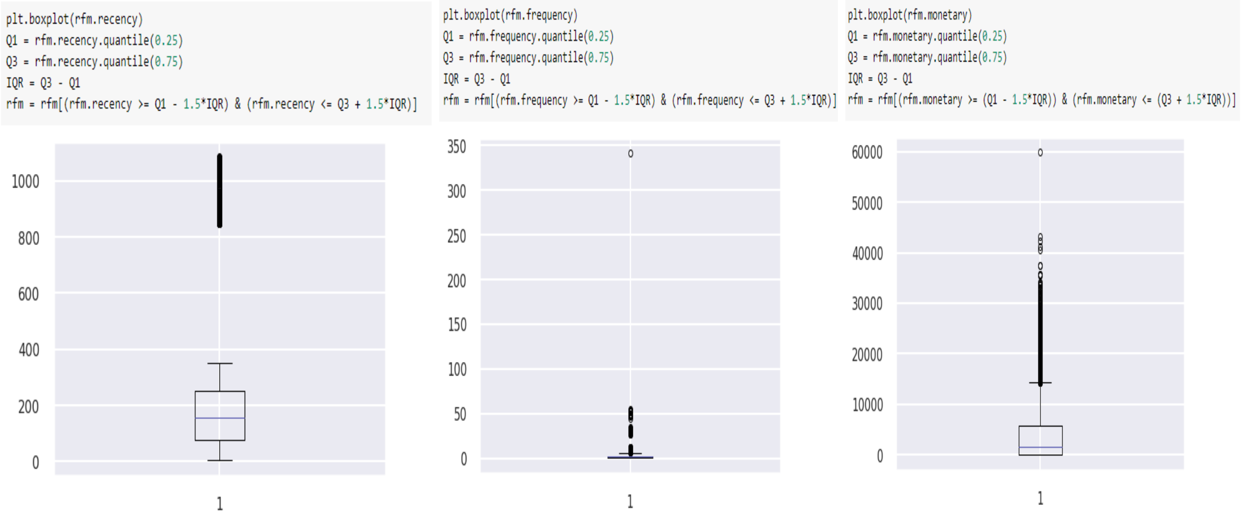
```
print(skew(rfm, axis=0, bias=True))

[  2.62328922  62.64997061 135.84540571]
```

Figure 2-8. Kurtosis and Skew value of recency, frequency and monetary

These numbers represent a large Kurtosis, often referred to as a leptokurtic or long-tailed distribution. Skew of the three large variables are all positive, and the data are highly skewed. Therefore, outliers must be permanently removed and changed to fit a normal distribution.

Interquartile Range (IQR) can be used to set the boundaries of data below the 25th percentile or above the 75th percentile to detect outliers.

After using IQR, the skewness goes towards 0 and the kurtosis is under 0.

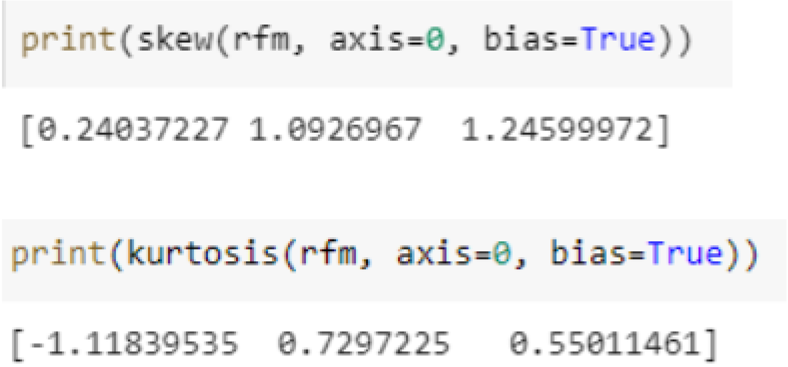


```
plt.boxplot(rfm.recency)
Q1 = rfm.recency.quantile(0.25)
Q3 = rfm.recency.quantile(0.75)
IQR = Q3 - Q1
rfm = rfm[(rfm.recency >= Q1 - 1.5*IQR) & (rfm.recency <= Q3 + 1.5*IQR)]
```

```
plt.boxplot(rfm.frequency)
Q1 = rfm.frequency.quantile(0.25)
Q3 = rfm.frequency.quantile(0.75)
IQR = Q3 - Q1
rfm = rfm[(rfm.frequency >= Q1 - 1.5*IQR) & (rfm.frequency <= Q3 + 1.5*IQR)]
```

```
plt.boxplot(rfm.monetary)
Q1 = rfm.monetary.quantile(0.25)
Q3 = rfm.monetary.quantile(0.75)
IQR = Q3 - Q1
rfm = rfm[(rfm.monetary >= (Q1 - 1.5*IQR)) & (rfm.monetary <= (Q3 + 1.5*IQR))]
```

Figure 2-9. Boxplot of recency, frequency, monetary values after using IQR



```
print(skew(rfm, axis=0, bias=True))
```

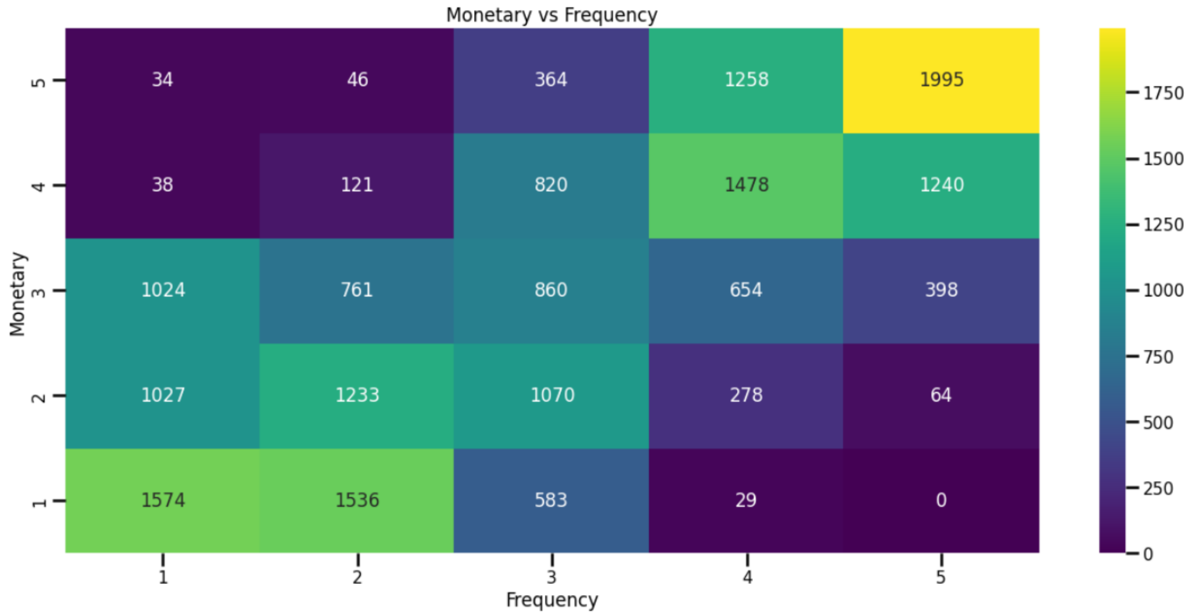
```
[0.24037227 1.0926967  1.24599972]
```

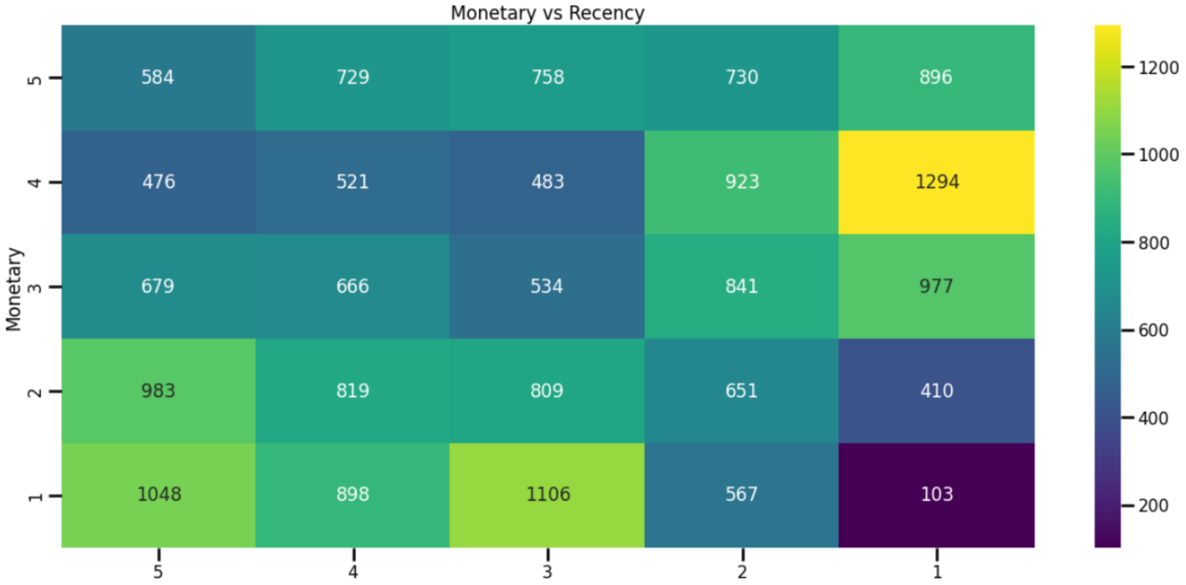
```
print(kurtosis(rfm, axis=0, bias=True))
```

```
[-1.11839535  0.7297225   0.55011461]
```

Figure 2-10. Skew and kurtosis after using IQR

2.2.3 Correlation matrix





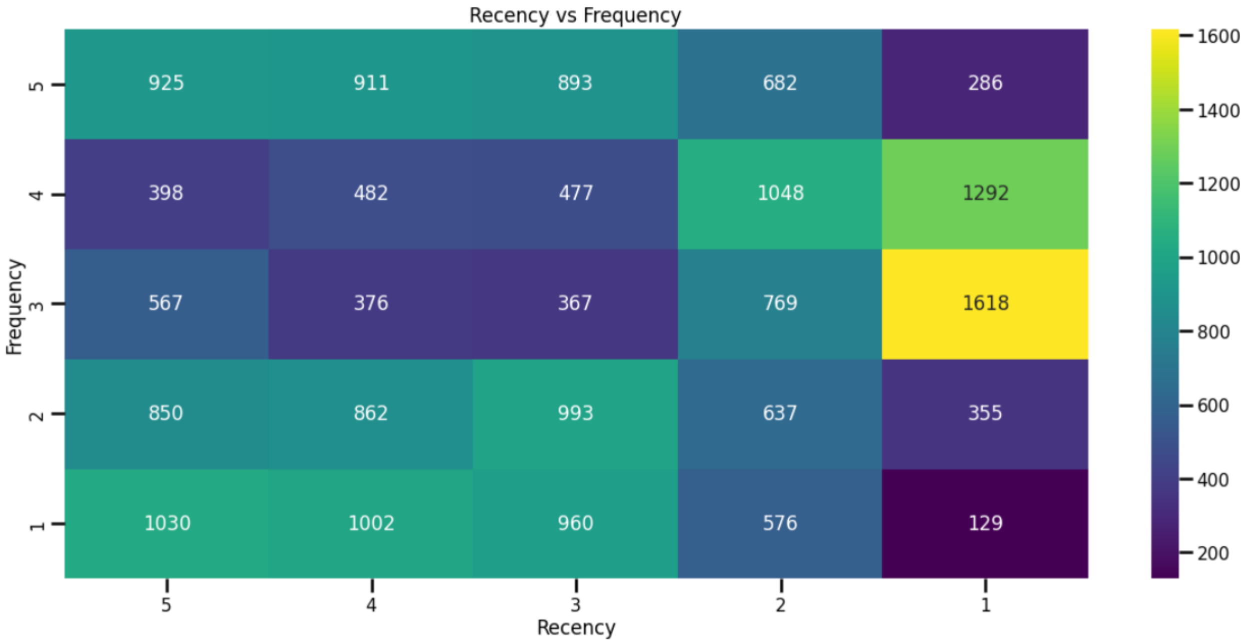


Figure 2-11. Correlation matrix

The above table shows the correlation coefficient between the pairs of variables "recency," "frequency," and "monetary."

A correlation matrix consisting of rows and columns showing the variables Each cell in the table contains the correlation coefficient between the variables. They are commonly used for normalized data sets, where the option chosen is to split the values into 5 compartments by the ordinal variable and K-Means.

Based on these Figure 2-12, we can intuitively feel that there is a change in the structure of the correlation pairs. The correlation between Monetary and Frequency is relatively low; the correlation between Monetary and Recency is relatively high; and between Recency and Frequency is relatively stable.

2.2.4 Calculate the score for R,F,M value

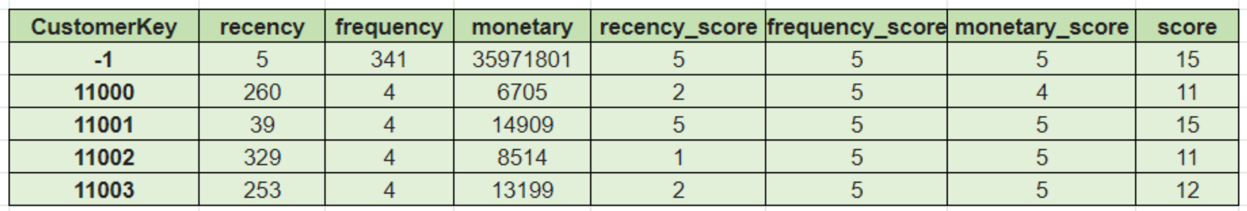
After establishing the RFM method, each customer's Recency, Frequency and Monetary factors are usually ranked on a normal scale of 1 to 5.

The columns are converted into RFM scores between 1 and 5 in the table below.

1 is the lowest and 5 is the highest.

- The score, expressed as "5", increases with increasing Monetary worth.
- Recent purchases are indicated by a smaller Recency value, hence the greater value of 5 is used.
- Frequency is the same as monetary, higher the frequency, higher the score.

Due to their high recency, frequency, and monetary scores, consumers with a score of "15" are regarded as the "best customers".



CustomerKey	recency	frequency	monetary	recency_score	frequency_score	monetary_score	score
-1	5	341	35971801	5	5	5	15
11000	260	4	6705	2	5	4	11
11001	39	4	14909	5	5	5	15
11002	329	4	8514	1	5	5	11
11003	253	4	13199	2	5	5	12

Table 2-3. Calculate the score

Based on the score, we will categorize the customer's level. We've divided them up into four levels here.

Active: High revenue generating and regular customers.

Good: Customers that make moderately frequent purchases and contribute to revenue.

Average: Less active, infrequently purchasing customers that produce little revenue.

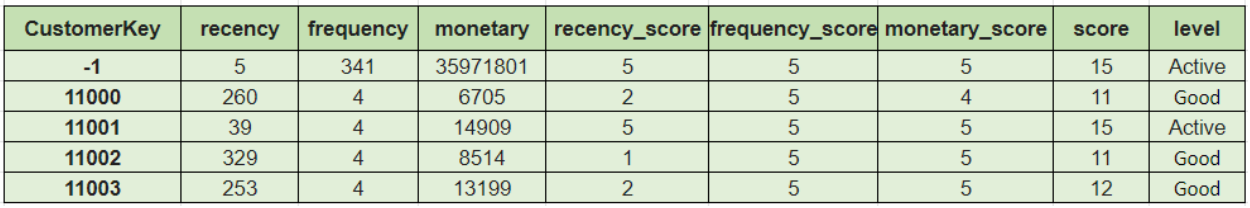
Inactive: Customers who are sporadic buyers and generate very little revenue

With: $1 < \text{Score} < 4$: Inactive

$4 \leqq \text{Score} < 8$: Average

$8 \leqq \text{Score} < 11$: Good

$11 \leqq \text{Score} \leqq 15$: Active



CustomerKey	recency	frequency	monetary	recency_score	frequency_score	monetary_score	score	level
-1	5	341	35971801	5	5	5	15	Active
11000	260	4	6705	2	5	4	11	Good
11001	39	4	14909	5	5	5	15	Active
11002	329	4	8514	1	5	5	11	Good
11003	253	4	13199	2	5	5	12	Good

Table 2-4. Segregate the levels

2.2.5 Visualizing number of customers for each level

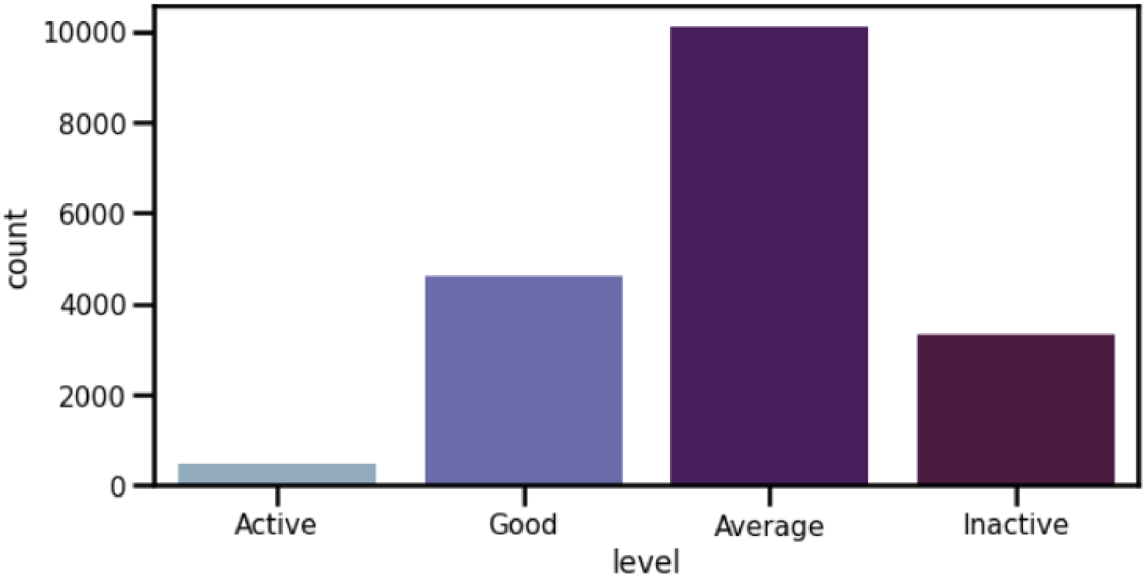


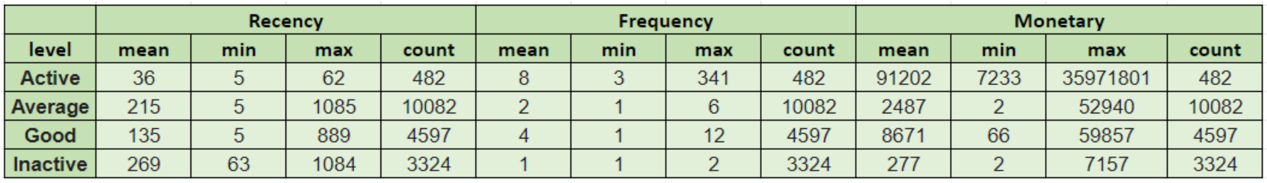
Figure 2-12. Bar chart about level

We can conclude from the bar graph above that

- Very few customers are actively shopping right now.
- The average client count is quite high, roughly doubling the Good customer.
- In comparison to the total number of active and Good customers, the number of average and inactive consumers is significantly higher

⇒ This store has very few loyal customers. The majority of customers come in just once or twice.

We will examine each consumer category in depth in order to comprehend it better.



	Recency				Frequency				Monetary			
level	mean	min	max	count	mean	min	max	count	mean	min	max	count
Active	36	5	62	482	8	3	341	482	91202	7233	35971801	482
Average	215	5	1085	10082	2	1	6	10082	2487	2	52940	10082
Good	135	5	889	4597	4	1	12	4597	8671	66	59857	4597
Inactive	269	63	1084	3324	1	1	2	3324	277	2	7157	3324

Table 2-5. Analysing each level

- Active Customers: About 482 clients made purchases totaling 91202 units 8 times, or roughly every 36 days.

  ⇒ They are the preferred customers. There should be many surprise offers for them on special and important occasions. Let them try and get access to new products first.

- Good Customers: 4597 customers have bought 8671 units by shopping 4 times every 135 days

  ⇒ They are loyal customers. Provide membership or loyalty programs, or suggest similar products to upsell to them, and turn them into your active customers.

- Average Customers: By shopping twice every 215 days, 10082 customers have purchased around 2487 units.

  ⇒ Start developing relationships with these customers by helping them through the referral process so they keep coming back to buy.

- Inactive Customers: 3324 customers have bought approximately 277 units by shopping once every 269 days.

  ⇒ Plan activities like seminars or marketing campaigns to attract and reach more customers. Enhance marketing and promotion strategies on social media or online forums to improve order frequency.

Chương 3: Customer segmentation

3.1 Machine Learning method

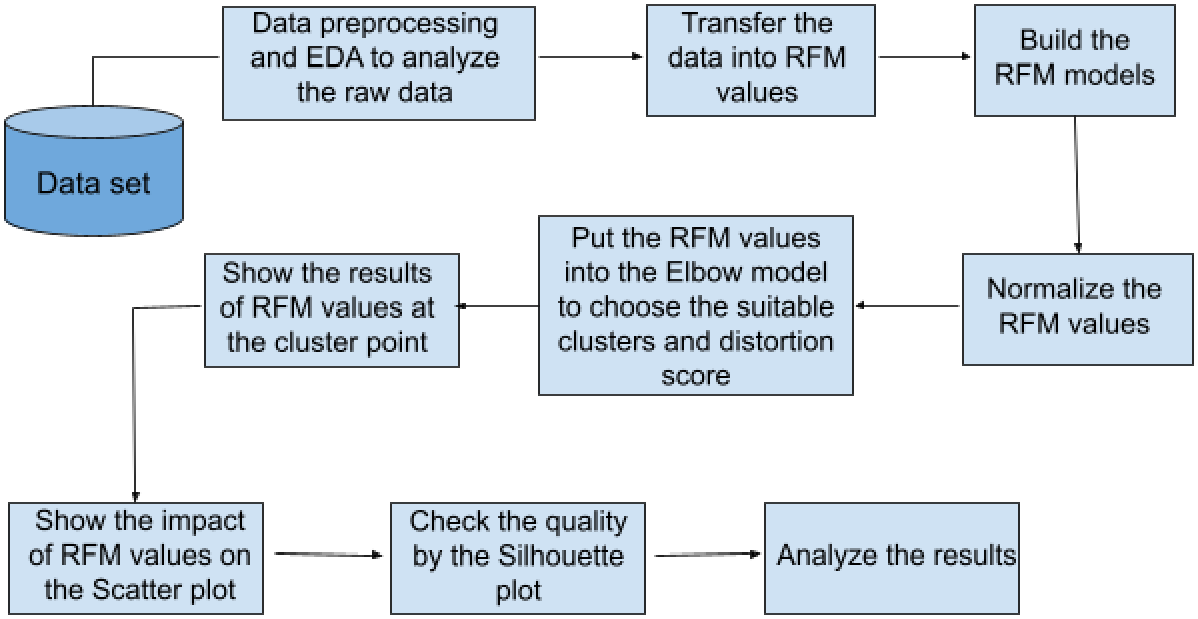
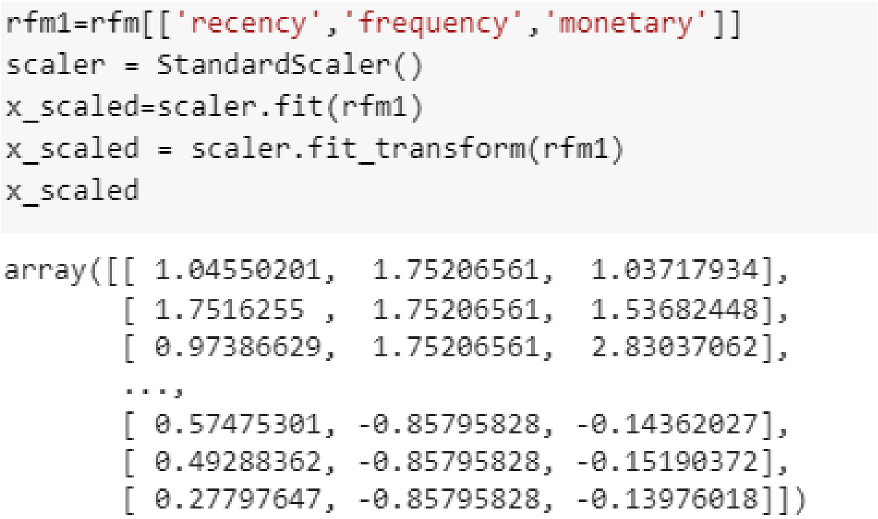

Data set
Data preprocessing and EDA to analyze the raw data
Transfer the data into RFM values
Build the RFM models
Normalize the RFM values
Put the RFM values into the Elbow model to choose the suitable clusters and distortion score
Show the results of RFM values at the cluster point
Show the impact of RFM values on the Scatter plot
Check the quality by the Silhouette plot
Analyze the results

Figure 3-0. RFM method

3.2 RFM with K-means



```
rfm1=rfm[['recency','frequency','monetary']]
scaler = StandardScaler()
x_scaled=scaler.fit(rfm1)
x_scaled = scaler.fit_transform(rfm1)
x_scaled
```

```
array([[ 1.04550201,  1.75206561,  1.03717934],
       [ 1.7516255 ,  1.75206561,  1.53682448],
       [ 0.97386629,  1.75206561,  2.83037062],
       ...,
       [ 0.57475301, -0.85795828, -0.14362027],
       [ 0.49288362, -0.85795828, -0.15190372],
       [ 0.27797647, -0.85795828, -0.13976018]])
```

Figure 3-1. The scaled figure

The x_scaled value is transferred to visualize for performance of the Elbow model. Scaler values are used as standard scales to assess and build the elbow model.

The distortion score elbow is built to review by distortion score, k as the number of clusters and the fit time of cluster performed in the chart.

K = 3 is the number of suitable clusters. It means when the line of the chart crosses the k = 3,the line is trending to reduce the slope and decrease the distortion score. At the k = 3, the score is recorded 17285.649 that the distortion score of the elbow is not high and not fluctuated at that point.

When the line crosses the k = 3, it is in tendency to raise the fit time of movement of the line.

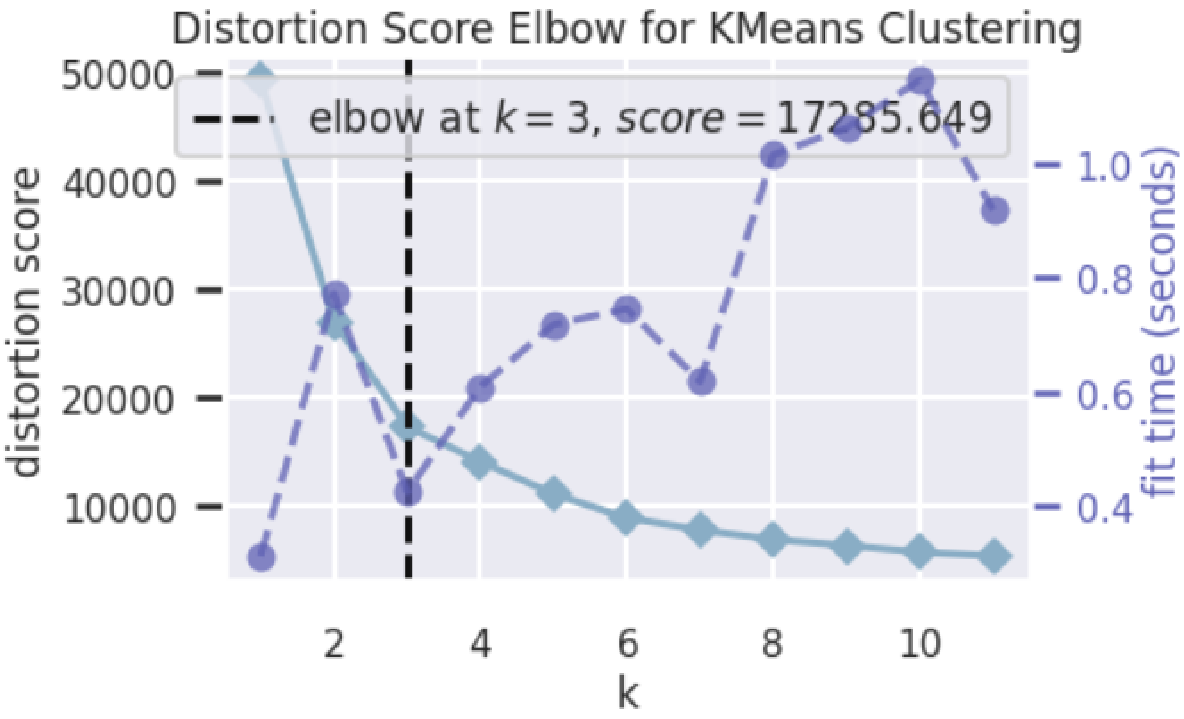


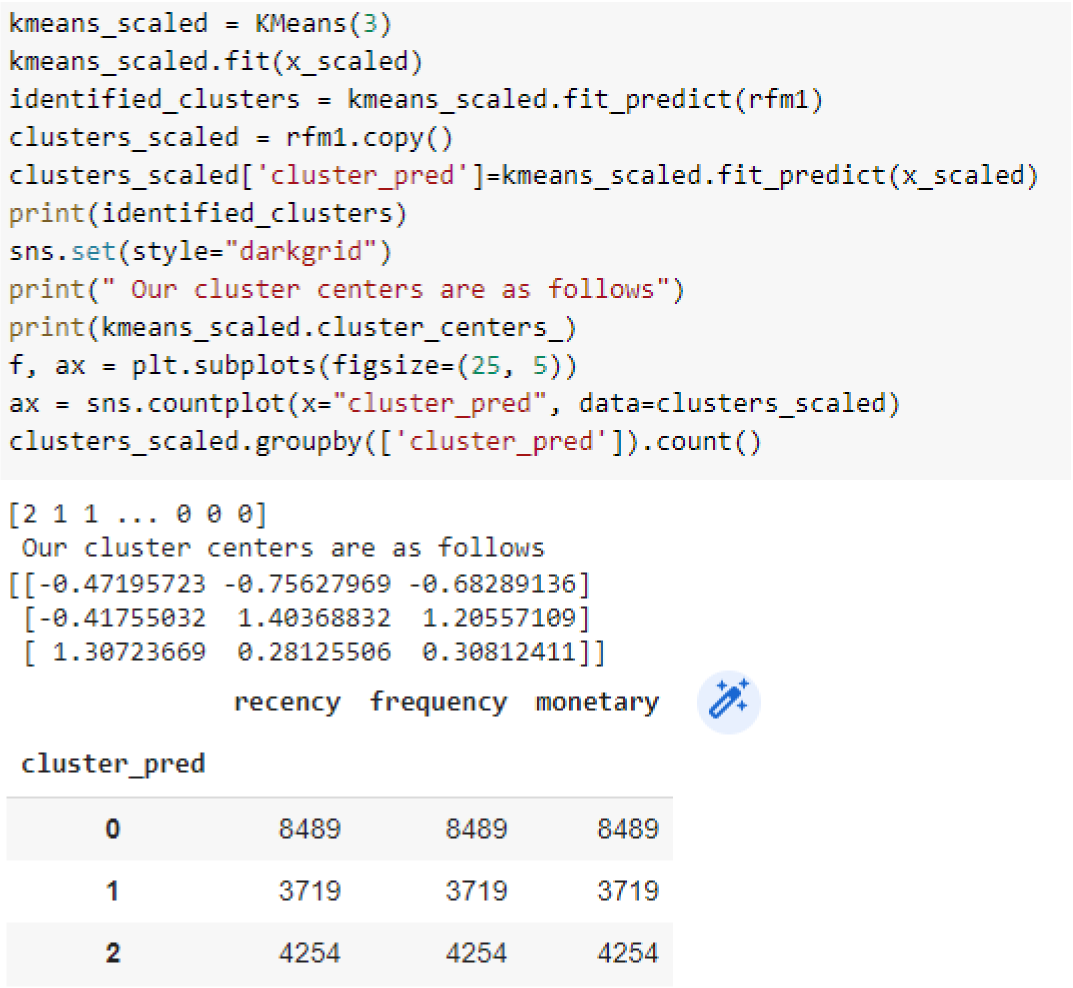
Figure 3-2. The Distortion Score Elbow for KMeans Clustering

3.3 K Means with Clustering

The RFM value shows output at the cluster predict point as 0; 1; 2 with the results. The count values of clustering perform in the chart to review the RFM value in the clustering.

At the cluster as 0; 1; 2, we can know the number of customers that we are researching. At point 0, that is the highest number of customers and at point 1, the lowest number of customers. The results of the RFM value of 3 cluster points are the same.

The RFM values are followed by the cluster centers. Most of the values of cluster centers are trending to go 0.



```
kmeans_scaled = KMeans(3)
kmeans_scaled.fit(x_scaled)
identified_clusters = kmeans_scaled.fit_predict(rfm1)
clusters_scaled = rfm1.copy()
clusters_scaled['cluster_pred']=kmeans_scaled.fit_predict(x_scaled)
print(identified_clusters)
sns.set(style="darkgrid")
print(" Our cluster centers are as follows")
print(kmeans_scaled.cluster_centers_)
f, ax = plt.subplots(figsize=(25, 5))
ax = sns.countplot(x="cluster_pred", data=clusters_scaled)
clusters_scaled.groupby(['cluster_pred']).count()
```

```
[2 1 1 ... 0 0 0]
 Our cluster centers are as follows
[[-0.47195723 -0.75627969 -0.68289136]
 [-0.41755032  1.40368832  1.20557109]
 [ 1.30723669  0.28123506  0.30812411]]
```

cluster_pred	recency	frequency	monetary
0	8489	8489	8489
1	3719	3719	3719
2	4254	4254	4254

Table 3-1. The figure of cluster prediction

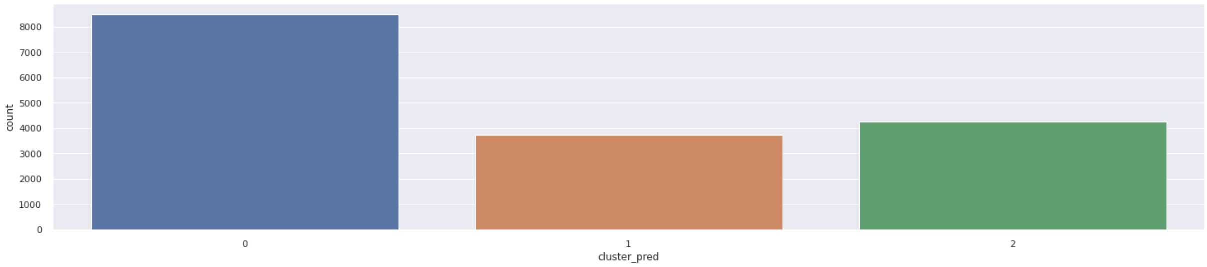
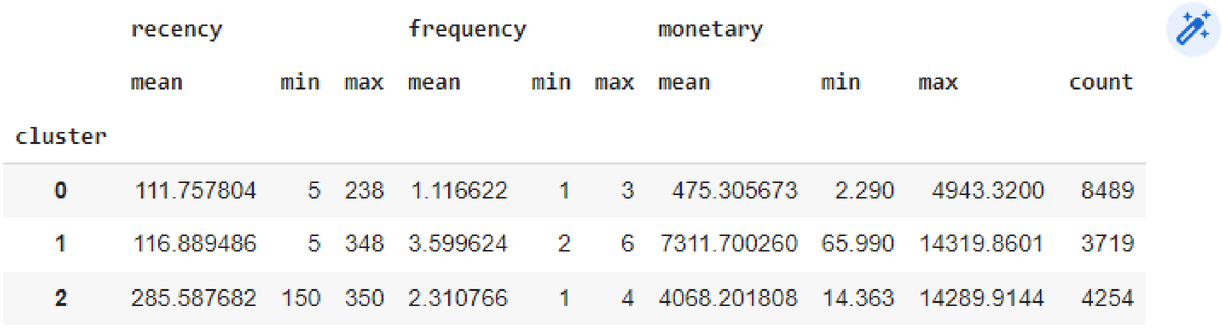


Figure 3-3. The cluster charts perform the predicted figure

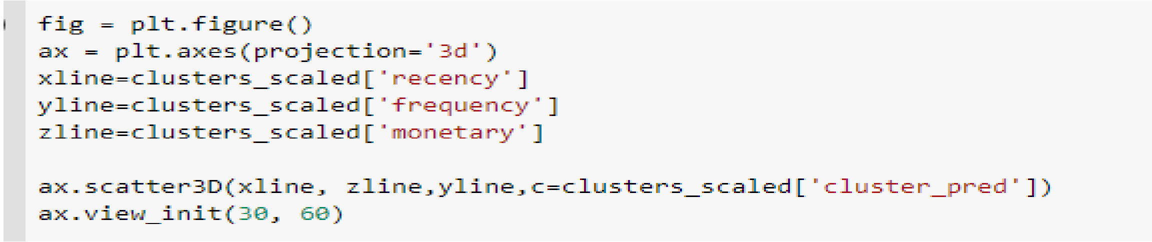
The table shows the mean, min, max and count value at 3 clusters of the RFM value to describe the figure of RFM value in the cluster chart. In addition, it can get to know the range of the RFM charts.



	recency			frequency			monetary			
	mean	min	max	mean	min	max	mean	min	max	count
cluster										
0	111.757804	5	238	1.116622	1	3	475.305673	2.290	4943.3200	8489
1	116.889486	5	348	3.599624	2	6	7311.700260	65.990	14319.8601	3719
2	285.587682	150	350	2.310766	1	4	4068.201808	14.363	14289.9144	4254

Table 3-2. The range of cluster results

3.4 Scatter plot



```
fig = plt.figure()
ax = plt.axes(projection='3d')
xline=clusters_scaled['recency']
yline=clusters_scaled['frequency']
zline=clusters_scaled['monetary']

ax.scatter3D(xline, zline,yline,c=clusters_scaled['cluster_pred'])
ax.view_init(30, 60)
```

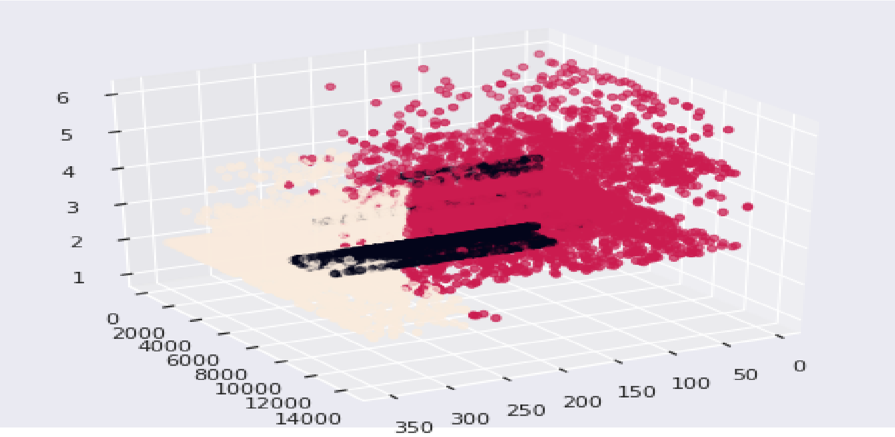
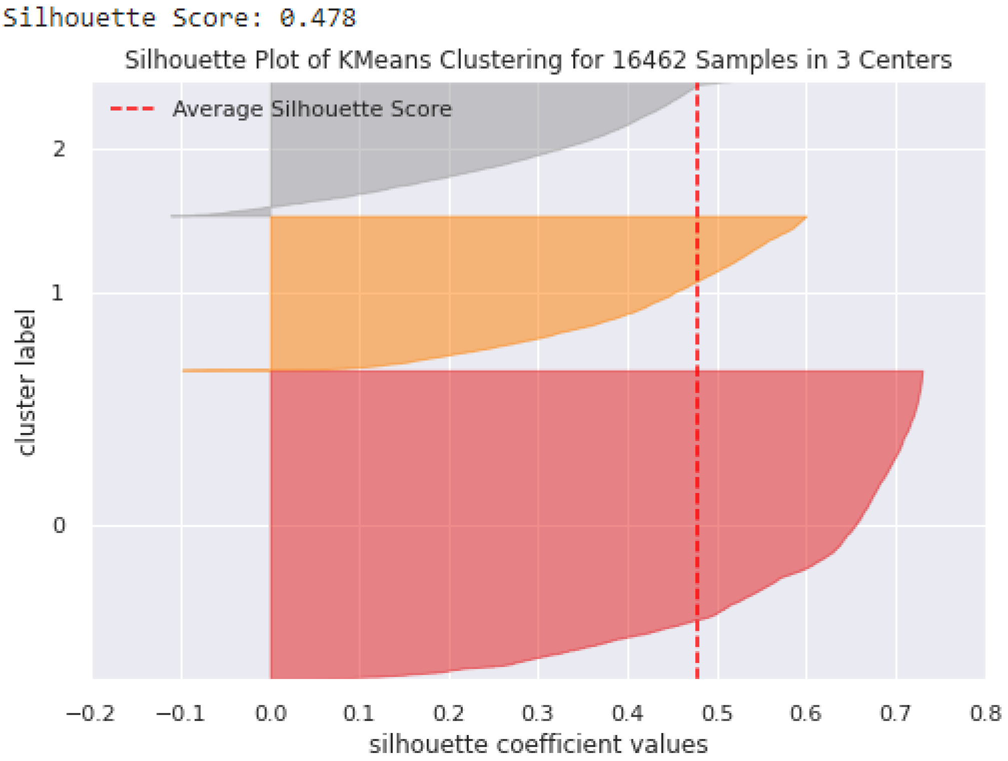


Figure 3-4. The scatter plot

The graph shows the combination and impact of the three R, F and M values.

By analyzing and evaluating the chart above, we can see the influence of the values on the research results.

3.5 Silhouettes



```
Silhouette Score: 0.478
```

Figure 3-5. The silhouette plot

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

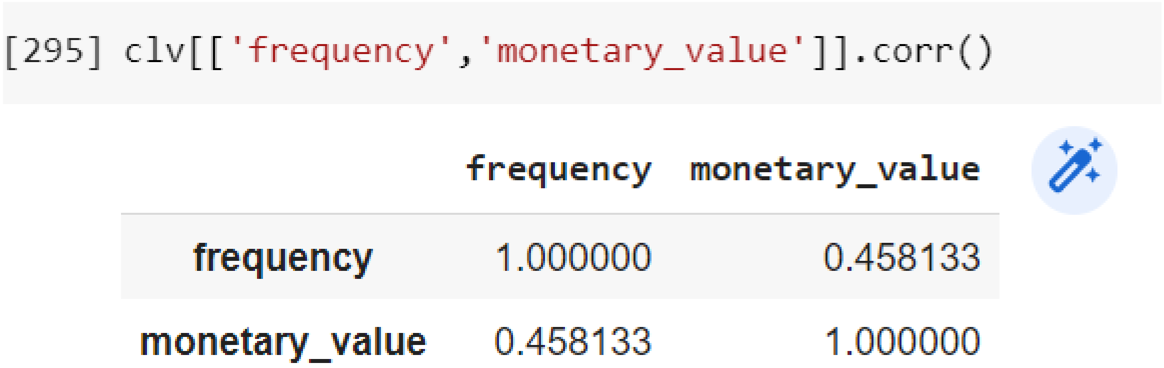
The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). With K = 3, the silhouette score = 0,478 shows that the object is well matched to its own cluster and poorly matched to neighboring clusters. Most objects have a high value, then the clustering configuration is appropriate.

The silhouette score = 0,478 is the distance from the assessing point to the central point that is taken really closely to the other central point of other clusters. It means with this silhouette score, we need more information to analyze and understand. K = 3 is the turning point that the highest point to generally decrease value K helps to gain the silhouette score to reduce the analysis information to increase accuracy of data.

Chapter 4: Normalization and Customer Lifetime Value

4.1 Customer Life Value (CLV)

4.1.1 Correlation of frequency and monetary



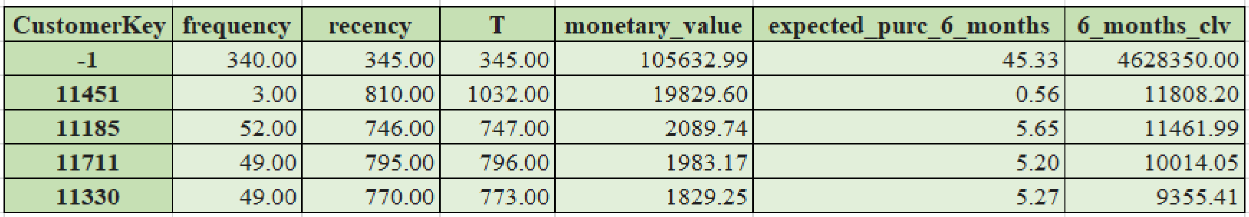
```
[295] clv[['frequency','monetary_value']].corr()
```

	frequency	monetary_value
frequency	1.000000	0.458133
monetary_value	0.458133	1.000000

Figure 4-1. Correlation of frequency and monetary

Since the correlation between the frequency value and the monetary value is weak, we build a gamma-gamma model to predict the values.

4.1.2 Expected and actual purchases in 6 months



CustomerKey	frequency	recency	T	monetary_value	expected_purc_6_months	6_months_clv
-1	340.00	345.00	345.00	105632.99	45.33	4628350.00
11451	3.00	810.00	1032.00	19829.60	0.56	11808.20
11185	52.00	746.00	747.00	2089.74	5.65	11461.99
11711	49.00	795.00	796.00	1983.17	5.20	10014.05
11330	49.00	770.00	773.00	1829.25	5.27	9355.41

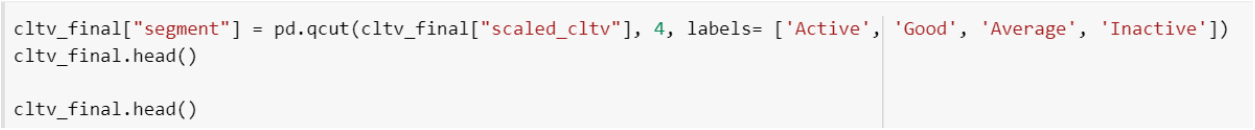
Table 4-1. Expected and actual purchases in 6 months

We add T value represents the age of the customer at whatever time units are chosen (weekly, in the above dataset). This is equal to the duration between a customer's first purchase and the end of the period under study.

Customer Lifetime Value prediction is a great way to get valuable insights about customer acquisition, marketing efforts, and company's financial future.

We will be able to know which marketing method is appropriate for each customer group, how effective it is and can we find a better method. See next section for more details

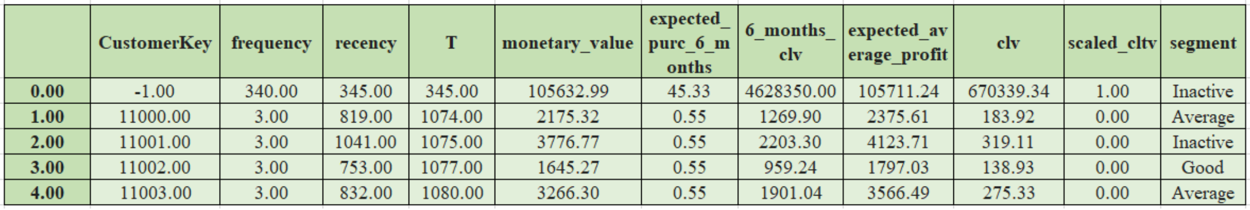
4.1.3 Segment and Final result



```
cltv_final["segment"] = pd.qcut(cltv_final["scaled_cltv"], 4, labels= ['Active', 'Good', 'Average', 'Inactive'])
cltv_final.head()

cltv_final.head()
```

Figure 4-2. Segment

Additionally, we can divide up our clientele into many categories such as Active, Good, Average, Inactive.



	CustomerKey	frequency	recency	T	monetary_value	expected_purc_6_months	6_months_clv	expected_average_profit	clv	scaled_cltv	segment
0.00	-1.00	340.00	345.00	345.00	105632.99	45.33	4628350.00	105711.24	670339.34	1.00	Inactive
1.00	11000.00	3.00	819.00	1074.00	2175.32	0.55	1269.90	2375.61	183.92	0.00	Average
2.00	11001.00	3.00	1041.00	1075.00	3776.77	0.55	2203.30	4123.71	319.11	0.00	Inactive
3.00	11002.00	3.00	753.00	1077.00	1645.27	0.55	959.24	1797.03	138.93	0.00	Good
4.00	11003.00	3.00	832.00	1080.00	3266.30	0.55	1901.04	3566.49	275.33	0.00	Average

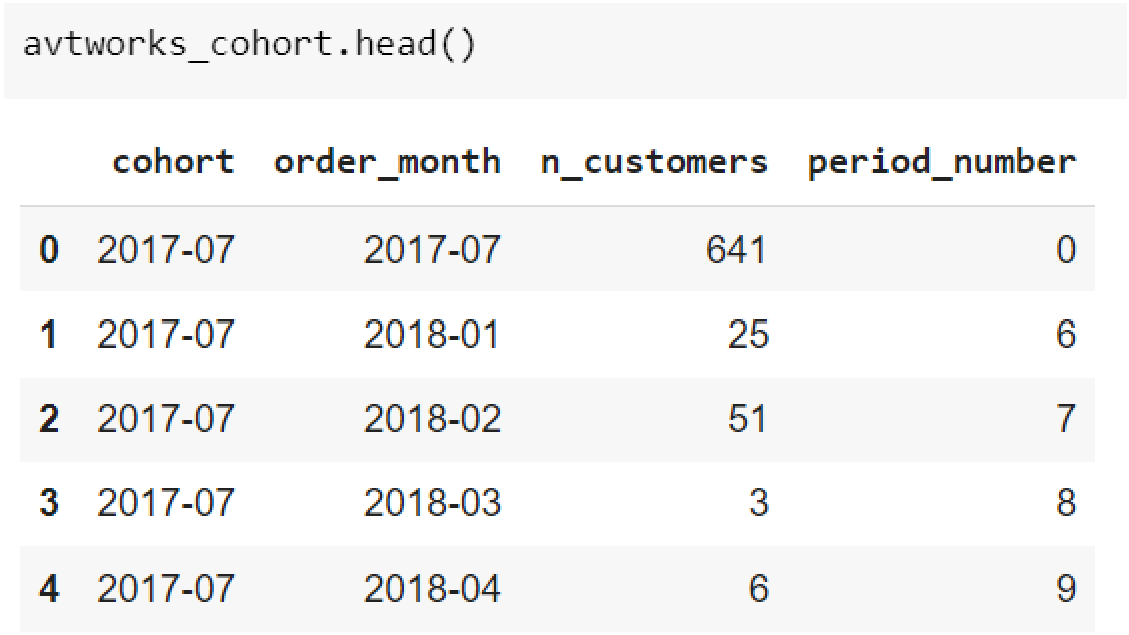
Table 4-2. The segment results

After segmenting our customers by CLV, we can:

- Offer specific products to each segment
- Create a marketing plan to increase CLV for lower segment
- Try to focus on the higher segments in order to decrease customer acquisition costs.

4.2 Cohort Analysis

4.2.1 Cohort table



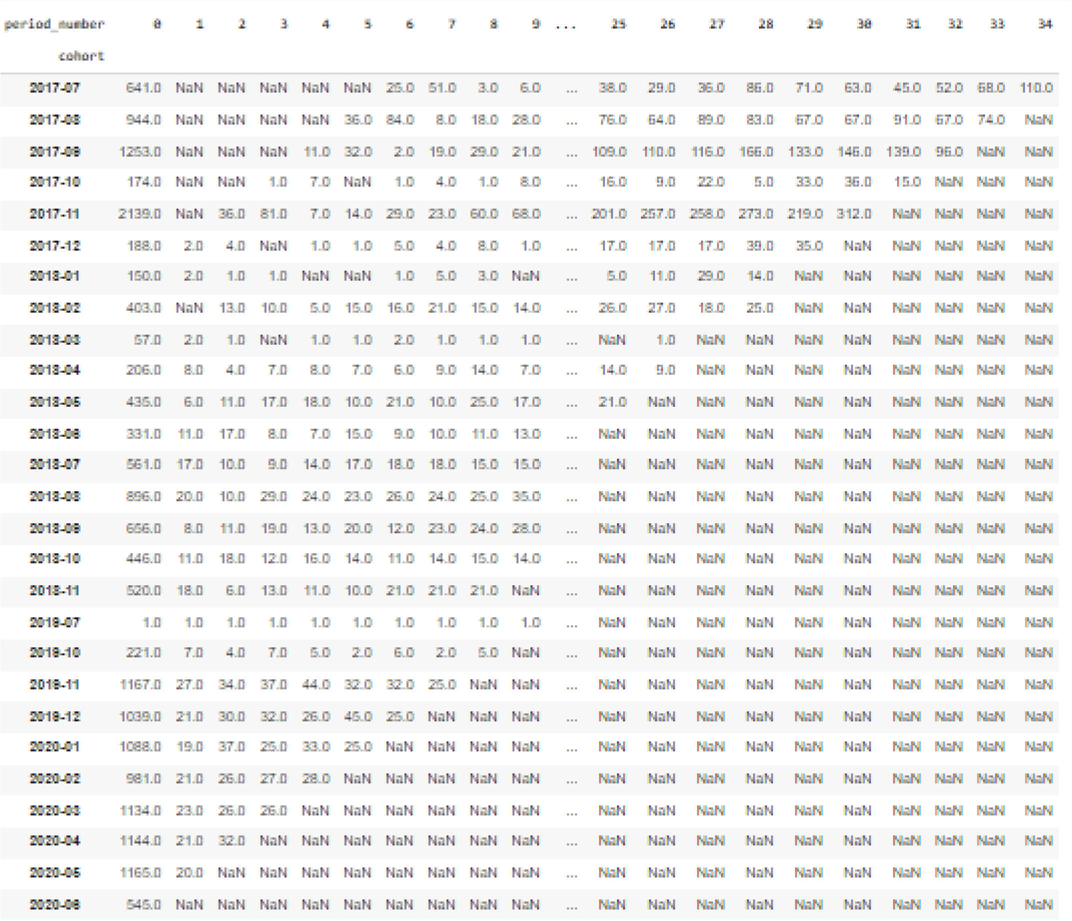
```
avtworks_cohort.head()
```

	cohort	order_month	n_customers	period_number
0	2017-07	2017-07	641	0
1	2017-07	2018-01	25	6
2	2017-07	2018-02	51	7
3	2017-07	2018-03	3	8
4	2017-07	2018-04	6	9

Table 4-3. The cohort table with 5 first rows

Next we will come to cohort analytics, focusing on analyzing the behavior of a group of customers who share a common characteristic over a certain period of time helps us gain insights into the customer experience. products to improve those experiences. It helps to go beyond the limitations of the average indicators analyzed in the EDA section, helping us to have clearer insights and thereby make more accurate decisions.

4.2.2 Pivot



period_number	0	1	2	3	4	5	6	7	8	9	...	25	26	27	28	29	30	31	32	33	34
cohort																					
2017-07	641.0	NaN	NaN	NaN	NaN	NaN	25.0	51.0	3.0	6.0	...	38.0	29.0	36.0	86.0	71.0	63.0	45.0	52.0	68.0	110.0
2017-08	944.0	NaN	NaN	NaN	NaN	36.0	84.0	8.0	18.0	28.0	...	76.0	64.0	89.0	83.0	67.0	67.0	91.0	67.0	74.0	NaN
2017-09	1253.0	NaN	NaN	NaN	11.0	32.0	2.0	19.0	29.0	21.0	...	109.0	110.0	116.0	166.0	133.0	146.0	139.0	96.0	NaN	NaN
2017-10	174.0	NaN	NaN	1.0	7.0	NaN	1.0	4.0	1.0	8.0	...	16.0	9.0	22.0	5.0	33.0	36.0	15.0	NaN	NaN	NaN
2017-11	2139.0	NaN	36.0	81.0	7.0	14.0	29.0	23.0	60.0	68.0	...	201.0	257.0	258.0	273.0	219.0	312.0	NaN	NaN	NaN	NaN
2017-12	188.0	2.0	4.0	NaN	1.0	1.0	5.0	4.0	8.0	1.0	...	17.0	17.0	17.0	39.0	35.0	NaN	NaN	NaN	NaN	NaN
2018-01	150.0	2.0	1.0	1.0	NaN	NaN	1.0	5.0	3.0	NaN	...	5.0	11.0	29.0	14.0	NaN	NaN	NaN	NaN	NaN	NaN
2018-02	403.0	NaN	13.0	10.0	5.0	15.0	16.0	21.0	15.0	14.0	...	26.0	27.0	18.0	25.0	NaN	NaN	NaN	NaN	NaN	NaN
2018-03	57.0	2.0	1.0	NaN	1.0	1.0	2.0	1.0	1.0	1.0	...	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-04	206.0	8.0	4.0	7.0	8.0	7.0	6.0	9.0	14.0	7.0	...	14.0	9.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-05	435.0	6.0	11.0	17.0	18.0	10.0	21.0	10.0	25.0	17.0	...	21.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-06	331.0	11.0	17.0	8.0	7.0	15.0	9.0	10.0	11.0	13.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-07	561.0	17.0	10.0	9.0	14.0	17.0	18.0	18.0	15.0	15.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-08	896.0	20.0	10.0	29.0	24.0	23.0	26.0	24.0	25.0	35.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-09	656.0	8.0	11.0	19.0	13.0	20.0	12.0	23.0	24.0	28.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-10	446.0	11.0	18.0	12.0	16.0	14.0	11.0	14.0	15.0	14.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-11	520.0	18.0	6.0	13.0	11.0	10.0	21.0	21.0	21.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2019-07	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2019-10	221.0	7.0	4.0	7.0	5.0	2.0	6.0	2.0	5.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2019-11	1167.0	27.0	34.0	37.0	44.0	32.0	32.0	25.0	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2019-12	1039.0	21.0	30.0	32.0	26.0	45.0	25.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-01	1088.0	19.0	37.0	25.0	33.0	25.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-02	981.0	21.0	26.0	27.0	28.0	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-03	1134.0	23.0	26.0	26.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-04	1144.0	21.0	32.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-05	1165.0	20.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-06	545.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 4-4. the retention and acquisition of customers.

The aforementioned table displays customer acquisition and retention.

The first column, or "0," in the vertical format, indicates how many new clients the company added during a specific month. For illustration, 641 represents the number of customers the company had in July 2017 and 944 represents the number of customers the company will have in the next month (August 2017), see the same with the next months

Looking horizontally, the first row shows the number of customers who have continued to be part of the business since their first purchase, i.e. July 2017. For example, 25 is the number of customers. out of 641 people who shopped six months after the first time. 51 is the number of customers who continue to buy after 7 months from the first purchase,...

4.2.3 Heatmap

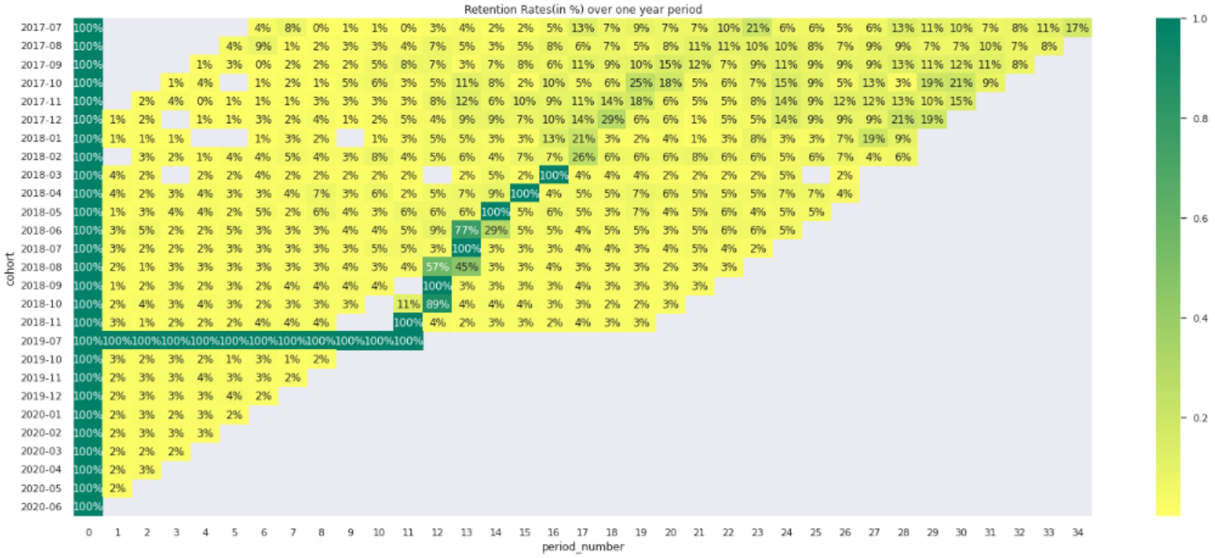


Figure 4-3. Visualize customer retention in the form of matrices and heat charts.

The above table merely displays value as percentages.

We can see how customers interact with companies from 2017 to 2020. Looking at the chart horizontally, the customer retention rate based on the first point of July 2017, the number of customers reached 4% in the sixth month and doubled (8%) in the seventh month. However, it decreased sharply and there was no significant change in the following months. The highlight is that in the 23rd month, there was a strong increase of up to 21%. Similarly for other timelines, we can completely re-check the objectivity at different times of the year.

In December 2017, the company received several new customers, however, only 1% of them were received. supposed to stay or come back. When this rate hits 2% the next month, it means certain customers come back and make new purchases.

Especially in July 2019, the number of customers completely remained at 100% in the following months. This could be because an invitation or an offer has been made to that particular customer base. However, the customer retention rate in the remaining months of the year is still not good and quite low.

Conclusion and Future Works

The research result shows that K-means clustering is an appropriate model for the current business situation. In this research, we also recommend the company the suitable marketing for each customer group campaigns idea to better perform targeted marketing.

However, there are still many limitations in our research. We haven't been able to fully exploit customer insight, we also have some difficulties in EDA and K-means.

In the future, we will also consolidate our knowledge of customers. Segmentation and marketing strategies to better experience in labeling and analyze the insight of customer groups.

References

[1] Kim-Giao Tran, Van-Ho Nguyen, Thanh Ho, University of Economics and Law, Ho Chi Minh City, Vietnam. Customer segmentation analysis and customer lifetime value prediction using Pareto/NBD and Gamma-Gamma model

[2] Thanh Trung Ho, Sơn Đăng-Nguyễn. An interdisciplinary research between analyzing customer segmentation in marketing and machine learning method

[3] Arpan Mishra, EDA and Customer Segmentation using RFM Analysis, last access date 15/12/2022, from: https://www.kaggle.com/code/arpanmishra/eda-and-customer-segmentation-using-rfm-analysis/notebook?fbclid=IwAR06fCSwpk17RBiVYZzEE9FwmL9xQXFoOHZo3oNPe-RVz-HjUe4xTQT6wv4

[4] Craig Stedman, what is data mining, last access date 15/12/2022, from: https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

[5] Pulkit Sharma - Published On August 19, 2019 and Last Modified On June 15th, 2022, The Most Comprehensive Guide to K-Means Clustering You'll Ever Need, last access date 15/12/2022, from: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

[6] Valentin Radu, Consumer behavior in marketing – patterns, types, segmentation, last access date 15/12/2022, from: https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/

[7] Keerthana Nithyakumar, What is Customer Lifetime Value (CLV) – Definition, Formula, Calculation with Examples, last access date 15/12/2022, from: https://www.zoho.com/subscriptions/guides/what-is-customer-lifetime-value-clv.html

[8] Pushpa Makhija, Cohort Analysis: Beginners Guide to Improving Retention, last access date 15/12/2022, from: https://clevertap.com/blog/cohort-analysis/

[9] Austin Caldwell, What Is Customer Lifetime Value (CLV) & How to Calculate ?, last access date 15/12/2022, from: https://www.netsuite.com/portal/resource/articles/ecommerce/customer-lifetime-value-clv.shtml

[10] Jay Selig, What Is Machine Learning? A Definition, last access date 15/12/2022, from: https://www.expert.ai/blog/machine-learning-definition/

[11] Uğur Savcı, Customer Lifetime Value Prediction in Python, last access date 15/12/2022, from: https://medium.com/@ugursavci/customer-lifetime-value-prediction-in-python-89e4a50df12e

[12] Silhouette (clustering), last access date 15/12/2022, from: https://en.wikipedia.org/wiki/Silhouette_(clustering)

[13] Arash Howaida, Sklearn kmeans equivalent of elbow method, last access date 15/12/2022, from: https://stackoverflow.com/questions/41540751/sklearn-kmeans-equivalent-of-elbow-method

[14] Anitha, P., & Patil, M. M. (2019). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University-Computer and Information Sciences

[15] Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, December). Customer segmentation using K-means clustering. In 2018 international conference on

computational techniques, electronics and mechanical systems (CTEMS) (pp. 135-139). IEEE

[16] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208

Appendix

Source code

```
from google.colab import drive

drive.mount('/contentdrive/')

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import datetime as dt

from scipy.stats import kurtosis

from scipy.stats import skew

from scipy.stats import boxcox

# Commented out IPython magic to ensure Python compatibility.

import warnings

warnings.filterwarnings('ignore')

# %matplotlib inline

# Import libraries

import os

for dirname, _, filenames in
os.walk('/contentdrive/Shareddrives/FDA-Ca1 T4 /Final
project/Final/data final.xlsx'):

    for filename in filenames:

        print(os.path.join(dirname, filename))

import io

avtworks = pd.read_excel('/contentdrive/Shareddrives/FDA-Ca1 T4
/Final project/Final/data final.xlsx')
```

Chapter 2: Data preparation

#EDA

```
avtworks.sample
avtworks.head()
avtworks.shape
avtworks.describe()
avtworks.info()
avtworks.dtypes
avtworks['OrderDateKey'].unique()
avtworks.isnull().sum().sort_values(ascending=False)
avtworks.hist(figsize =(20,10))
plt.show();
corravtworks = avtworks.corr()
sns.heatmap(corravtworks,
        xticklabels=corravtworks,
        yticklabels=corravtworks.columns, cmap='coolwarm_r')
sns.scatterplot(data=avtworks, x='OrderDateKey', y='Sales Amount')
plt.xlabel('OrderDateKey')
plt.ylabel('Sales Amount');
avtworks.plot(kind = 'box', subplots = True, figsize = (20,10))
fig = plt.figure(figsize =(10, 7))
```

#Data preprocessing

```
avtworks=avtworks.drop_duplicates()
avtworks.shape
avtworks.Date.max()
avtworks['Sales Amount']
pin_date = dt.datetime(2020, 6,20)
```

```
rfm = avtworks.groupby('CustomerKey').agg({'Date': lambda Date:
(pin_date - Date.max()).days,
                                     'OrderDateKey': lambda
OrderDateKey: OrderDateKey.nunique(),
                                     'Sales Amount': lambda
Sales_Amount: Sales_Amount.sum()})

rfm.head()

rfm.rename(columns={'Date': 'recency', 'OrderDateKey': 'frequency',
'Sales Amount': 'monetary'}, inplace=True)

rfm.head()

rfm.hist(figsize =(20,10))

plt.show();

rfm.plot(kind = 'box', subplots = True, figsize = (20,10))

from scipy.stats import kurtosis

print(kurtosis(rfm, axis=0, bias=True))

from scipy.stats import skew

print(skew(rfm, axis=0, bias=True))

plt.boxplot(rfm.recency)

Q1 = rfm.recency.quantile(0.25)

Q3 = rfm.recency.quantile(0.75)

IQR = Q3 - Q1

rfm = rfm[(rfm.recency >= Q1 - 1.5*IQR) & (rfm.recency <= Q3 +
1.5*IQR)]

plt.boxplot(rfm.frequency)

Q1 = rfm.frequency.quantile(0.25)

Q3 = rfm.frequency.quantile(0.75)

IQR = Q3 - Q1

rfm = rfm[(rfm.frequency >= Q1 - 1.5*IQR) & (rfm.frequency <= Q3 +
1.5*IQR)]

plt.boxplot(rfm.monetary)
```

```
Q1 = rfm.monetary.quantile(0.25)

Q3 = rfm.monetary.quantile(0.75)

IQR = Q3 - Q1

rfm = rfm[(rfm.monetary >= (Q1 - 1.5*IQR)) & (rfm.monetary <= (Q3 +
1.5*IQR))]

from scipy.stats import kurtosis

from scipy.stats import skew

print(kurtosis(rfm, axis=0, bias=True))

print(skew(rfm, axis=0, bias=True))

corrrfm = rfm.corr()

sns.heatmap(corrrfm,

        xticklabels=corrrfm,

        yticklabels=corrrfm.columns, cmap='coolwarm_r')

rfm["recency_score"] = pd.qcut(rfm['recency'], 5, labels=[5, 4, 3, 2,
1])

rfm["frequency_score"] =
pd.qcut(rfm['frequency'].rank(method="first"), 5, labels=[1, 2, 3, 4,
5])

rfm["monetary_score"] = pd.qcut(rfm['monetary'], 5, labels=[1, 2, 3,
4, 5])

rfm['score']=rfm['recency_score'].astype(int)+rfm['frequency_score'].
astype(int)+rfm['monetary_score'].astype(int)

rfm.head()

rfm[rfm['score']== 15].sort_values('monetary',
ascending=False).head()

rfm[rfm['score']==15].count()

def rfm_level(score):

    if  ((score >= 3) and (score < 7)):

        return 'Inactive'

    elif ((score >= 7) and (score < 11)):

        return 'Average'
    elif ((score >= 11) and (score <15)):
        return 'Good'
    else:
        return 'Active'
rfm['level'] = rfm['score'].apply(lambda score : rfm_level(score))
rfm.head()
plt.figure(figsize=(10,5))
sns.set_context("poster", font_scale=0.7)
sns.set_palette('twilight')
sns.countplot(rfm['level'])
rfm.groupby('level').agg({
    'recency' : ['mean', 'min','max','count'],
    'frequency' : ['mean', 'min','max','count'],
    'monetary' : ['mean','min','max','count']})
cross_table1 = pd.crosstab(index=rfm['monetary_score'],
columns=rfm['frequency_score'])
cross_table2 = pd.crosstab(index=rfm['monetary_score'],
columns=rfm['recency_score'])
cross_table3 = pd.crosstab(index=rfm['frequency_score'],
columns=rfm['recency_score'])
plt.figure(figsize=(20,30))
plt.subplot(311)
ax1 = sns.heatmap(cross_table1, cmap='viridis', annot=True,
fmt=".0f")
ax1.invert_yaxis()
ax1.set_ylabel('Monetary')
ax1.set_xlabel('Frequency')
ax1.set_title('Monetary vs Frequency')
plt.subplot(312)
```

```
ax2 = sns.heatmap(cross_table2, cmap='viridis', annot=True,
fmt=".0f")

ax2.invert_yaxis()

ax2.set_ylabel('Monetary')

ax2.set_xlabel('Recency')

ax2.set_title('Monetary vs Recency')

plt.subplot(313)

ax3 = sns.heatmap(cross_table3, cmap='viridis', annot=True,
fmt=".0f")

ax3.invert_yaxis()

ax3.set_ylabel('Frequency')

ax3.set_xlabel('Recency')

ax3.set_title('Recency vs Frequency')

plt.show()

active = rfm[rfm['level'] == 'Active']

average = rfm[rfm['level'] == 'Average']

good = rfm[rfm['level'] == 'Good']

inactive = rfm[rfm['level'] == 'Inactive']

active_df = pd.DataFrame()

active_df["customer_id"] = rfm[rfm["level"] == "Active"].index

active_df.to_excel("active_customers.xlsx", sheet_name='Active
Customers Index')

average_df = pd.DataFrame()

average_df["customer_id"] = rfm[rfm["level"] == "Average"].index

average_df.to_excel("average_customers.xlsx", sheet_name='Average
Customers Index')

good_df = pd.DataFrame()

good_df["customer_id"] = rfm[rfm["level"] == "Good"].index

good_df.to_excel("good_customers.xlsx", sheet_name='Good Customers
Index')
```

```
inactive_df = pd.DataFrame()
inactive_df["customer_id"] = rfm[rfm["level"] == "Inactive"].index
inactive_df.to_excel("inactive_customers.xlsx", sheet_name='Inactive
Customers Index')
rfm.describe()
```

Chapter 3: Customer segmentation

```
from yellowbrick.cluster import KElbowVisualizer
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
rfm1=rfm[['recency','frequency','monetary']]
scaler = StandardScaler()
x_scaled=scaler.fit(rfm1)
x_scaled = scaler.fit_transform(rfm1)
x_scaled
model = KMeans()
visualizer = KElbowVisualizer(model, k=(1,12))
visualizer.fit(x_scaled)
visualizer.show()
kmeans_scaled = KMeans(3)
kmeans_scaled.fit(x_scaled)
identified_clusters = kmeans_scaled.fit_predict(rfm1)
clusters_scaled = rfm1.copy()
clusters_scaled['cluster_pred']=kmeans_scaled.fit_predict(x_scaled)
print(identified_clusters)
sns.set(style="darkgrid")
print(" Our cluster centers are as follows")
print(kmeans_scaled.cluster_centers_)
```

```
f, ax = plt.subplots(figsize=(25, 5))
ax = sns.countplot(x="cluster_pred", data=clusters_scaled)
clusters_scaled.groupby(['cluster_pred']).count()
fig = plt.figure()
ax = plt.axes(projection='3d')
xline=clusters_scaled['recency']
yline=clusters_scaled['frequency']
zline=clusters_scaled['monetary']
ax.scatter3D(xline, zline,yline,c=clusters_scaled['cluster_pred'])
ax.view_init(30, 60)
from sklearn.metrics import silhouette_samples, silhouette_score
sil_score = silhouette_score(x_scaled, kmeans_scaled.labels_,
metric='euclidean')
print('Silhouette Score: %.3f' % sil_score)
from yellowbrick.cluster import SilhouetteVisualizer
model = KMeans(3)
visualizer = SilhouetteVisualizer(model)
visualizer.fit(x_scaled)
visualizer.poof()
rfm1['cluster']= clusters_scaled['cluster_pred']
rfm1['level']=rfm['level']
rfm1.groupby('cluster').agg({
    'recency' : ['mean','min','max'],
    'frequency' : ['mean','min','max'],
    'monetary' : ['mean','min','max','count']})
rfm1.head()
rfm1.groupby(['cluster']).size()
rfm_scaled=pd.DataFrame()
```

```
rfm_scaled=rfm1.copy()

scaler=StandardScaler()

rfm_scaled[['recency', 'frequency','monetary']] =
scaler.fit_transform(rfm_scaled[['recency', 'frequency','monetary']])

rfm_scaled['cust_id']=rfm1.index

rfm_scaled.head()

rfm_melted = pd.melt(frame= rfm_scaled, id_vars= ['cust_id',
'cluster','level'], var_name = 'metrics', value_name = 'value')

rfm_melted.head()

sns.lineplot(x = 'metrics', y = 'value', hue = 'level', data =
rfm_melted)

plt.title('Snake Plot of RFM')

plt.legend(loc = 'upper right')

sns.lineplot(x = 'metrics', y = 'value', hue = 'cluster', data =
rfm_melted)

plt.title('Snake Plot of Clusters')

plt.legend(loc = 'upper right')
```

Chapter 4: Normalization and Customer Lifetime Value

#PREDICTION CLV

```
!pip install lifetimes

import lifetimes

# BG/NBD

from lifetimes import BetaGeoFitter

# Gamma-Gamma Model

from lifetimes import GammaGammaFitter

from lifetimes.plotting import plot_frequency_recency_matrix

pd.set_option('display.max_rows', 500)

pd.set_option('display.max_columns', 500)

pd.set_option('display.width', 1000)
```

```
clv =
lifetimes.utils.summary_data_from_transaction_data(avtworks,'Customer
Key','Date','Sales Amount',observation_period_end='2020-06-15')

clv = clv[clv['frequency']>1] # we want only customers shopped more
than 2 times

bgf = BetaGeoFitter(penalizer_coef=0.001)

bgf.fit(clv['frequency'], clv['recency'], clv['T'])

# 30 day period

t = 180

clv['expected_purc_6_months'] =
bgf.conditional_expected_number_of_purchases_up_to_time(t,
clv['frequency'], clv['recency'], clv['T'])

clv.sort_values(by='expected_purc_6_months',ascending=False).head(5)

clv[['frequency','monetary_value']].corr()

# Creating Gamma-Gamma Model

ggf = GammaGammaFitter(penalizer_coef=0.01) # model object

ggf.fit(clv['frequency'], clv['monetary_value']) # model fitting

clv['6_months_clv']=ggf.customer_lifetime_value(bgf,

                                  clv["frequency"],

                                  clv["recency"],

                                  clv["T"],

                                  clv["monetary_value"],

                                  time=6,

                                  freq='D',

                                  discount_rate=0.01)

clv.sort_values('6_months_clv',ascending=False).head()

ggf.summary

ggf.conditional_expected_average_profit(clv['frequency'],
clv['monetary_value']).sort_values(ascending=False).head(10)
```

```
clv['expected_average_profit'] =
ggf.conditional_expected_average_profit(clv['frequency'],
clv['monetary_value'])

cltv = ggf.customer_lifetime_value(bgf, clv['frequency'],
clv['recency'], clv['T'], clv['monetary_value'], time=6, freq='W')

cltv = cltv.reset_index()

cltv_final = clv.merge(cltv, on='CustomerKey', how='left')

cltv_final.sort_values(by='clv', ascending=False).head()

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature_range=(0, 1))

cltv_final.head()

scaler.fit(cltv_final[['clv']])

cltv_final['scaled_cltv'] = scaler.transform(cltv_final[['clv']])

cltv_final.sort_values(by='scaled_cltv', ascending=False).head()

cltv_final.sort_values(by='scaled_cltv', ascending=False).tail()

cltv_final["segment"] = pd.qcut(cltv_final["scaled_cltv"], 4, labels=
['Active', 'Good', 'Average', 'Inactive'])

cltv_final.head()

cltv_final.head()
```

```
#COHORT

avtworks['Date'] = pd.to_datetime(avtworks['Date'], format='%m/%d/%Y
%H:%M')

avtworks['order_month'] = avtworks['Date'].dt.to_period('M')

avtworks['cohort'] =
avtworks.groupby('CustomerKey')['Date'].transform('min').dt.to_period
('M')

avtworks_cohort = avtworks.groupby(['cohort',
'order_month']).agg(n_customers=('CustomerKey',
'nunique')).reset_index(drop=False)

from operator import attrgetter
```

```
avtworks_cohort['period_number'] = (avtworks_cohort.order_month -
avtworks_cohort.cohort).apply(attrgetter('n'))

avtworks_cohort.head()

cohort_pivot = avtworks_cohort.pivot_table(index='cohort',
columns='period_number', values='n_customers')

cohort_pivot

cohort_size = cohort_pivot.iloc[:, 0]

retention = cohort_pivot.divide(cohort_size,axis=0) #axis=0 to ensure
the divide along the row axis

retention_matrix = cohort_pivot.divide(cohort_size, axis=0)

import matplotlib.colors as mcolors

#Build the heatmap or pictorial representation of above table

plt.figure(figsize=(25, 10))

plt.title('Retention Rates(in %) over one year period', size=12)

sns.heatmap(data=retention, annot = True, fmt = '.0%',
cmap="summer_r")

plt.show()
```