

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
FACULTY OF INFORMATION SYSTEMS



FINAL PROJECT REPORT

FUNDAMENTALS OF DATA ANALYTICS

**TOPIC: ANALYZE CUSTOMER DATA WITH RFM, K-MEANS
AND CALCULATE BUSINESS CRITICAL METRICS**

Lecturer:

- 1. Ho Trung Thanh, Ph.D.**
- 2. Nguyen Phat Dat, MA**

- 1. Nguyen Thanh Luan**
- 2.**
- 3.**

Ho Chi Minh City, December, 2022

Members of Group 8

<i>No.</i>	Full name	Student ID	Point / 10
<i>1</i>	Nguyễn Thành Luân	K214160991	10/10
<i>2</i>			
<i>3</i>			
<i>4</i>			

Acknowledgments

First, we would like to express our sincere thanks to the University of Economics and Law for bringing basic data analysis (FDA) into the curriculum. In particular, we would like to express our deep thanks to our academic advisor and subject lecturer - Mr. Ho Trung Thanh, Mr. Nguyen Phat Dat for their dedicated guidance and imparting valuable knowledge to us during the past study period. During our time participating in the basic FDA class, we gained more useful knowledge, more analytical ability and increased the spirit of effective and serious learning. His enthusiasm and attentiveness are a great inspiration for us to complete this essay in the most thoughtful way. Without his motivation and guidance, the thesis would not have been done effectively and thoroughly.

The Department of Basic Data Analysis is an interesting, extremely useful subject and highly interdisciplinary knowledge application, highly practical, ensuring sufficient knowledge, associated with the practical needs of students. We have tried to complete this essay with all our abilities, book knowledge and experience that we have accumulated. However, due to the limited knowledge capital and the ability to absorb reality, there are still many surprises. Despite our best efforts, it is certainly difficult for the essay to avoid shortcomings and many incomplete places. Therefore, our group looks forward to receiving suggestions and sharing from teachers so that the group's plan can become more complete.

We wish you all good health and more success in your teaching career.
Thank you!

Commitment

We would like to commit that the results below are purely the application of our knowledge on the basis of the knowledge taught from the Department of Basic Data Analysis, Interdisciplinary Research Methods, Applied Statistics of Mr. Ho Trung Thanh, combined with reference resources from books, newspapers, and other media. We promise, this project does not steal or copy any sources out there. The author promises that the project will be completed on 13/12/2021, this project will be under the supervision of Dr. Ho Trung Thanh and Mr. Nguyen Phat Dat.

Table of Contents

Chapter 1 Theoretical	16
1.1 Machine learning	16
1.2 RFM (Recency – Frequency – Monetary Value)	17
1.3 KMeans	17
1.4 Customer Lifetime Value (CLV)	17
1.5 Customer Retention Rate (CRR)	18
Chapter 2 Data Preparation	19
2.1 Stages of EDA	19
2.2 Exploratory Data Analysis (EDA)	20
2.2.1 Distinguish Attribute	20
2.2.2 Univariate analysis	21
2.2.3 Pre-processing and EDA	23
2.2.4 Multivariable analysis	28
Chapter 3 Customer segmentation with machine learning method	30
3.1 Model Steps	30
3.2 Clustering Process	31
3.2.1 Elbow analyst	31
3.2.2 Review by Silhouette	31
3.2.3 Fit to Kmeans	33
3.3 Comparing clusters and Discussing business	35
3.3.1 Comparing cluster	35
3.3.2 Discussing business	35
Chapter 4 Customer Lifetime Value and Customer Retention Rate	37
4.1. Predictive analytics CLV	37
4.1.1 Customer Lifetime Value Method (CLV)	37

4.2 Customer Retention Rate (CRR)	39
4.2.1 Calculate the CRR	40
Step 1 : Create dataframe to calculate CRR and separate year from Orderdate	40
Step 2 : Find the first purchase of each customer with ‘pandas.groupby’ operation by customerkey and find the first time of Orderdate	41
Step 3 : Combine 2 data sets to create a complete dataframe to calculate CRR	42
Step 4 : Create a function to calculate CRR	42
4.2.2 CRR Visualization	43
Chapter 5 Linear Regression	45
5.1 Preparing the dataset	46
5.2 Training	48

List of Tables

Table 1.1	31
Table 1.2	37

List of Figures

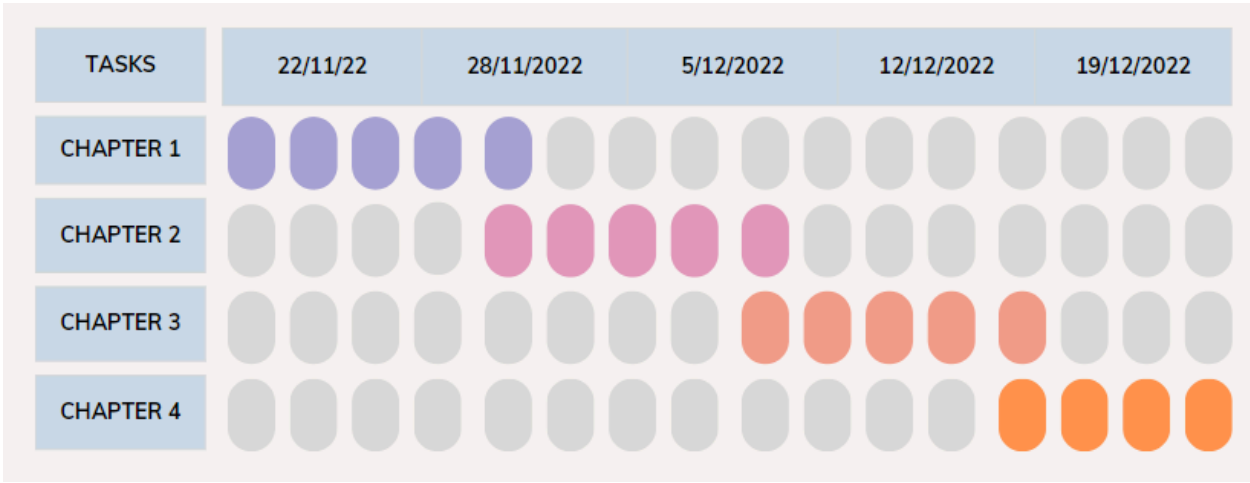
Name of Figures	Number Page
Figure 2.1 Stages of EDA	9
Figure 2.2.1.1 Overview of original Dataset	10
Figure 2.2.2.1 Data after removing value CustomerKey = -1	12
Figure 2.2.2.2 Counting non-duplicate quantities	12
Figure 2.2.2.3 Datatype after convert	13
Figure 2.2.3.1 RFM calculation	14
Figure 2.2.3.2 Descriptive statistics	14
Figure 2.2.3.3 Before Removing Outliers	15
Figure 2.2.3.4 After Removing Outliers	16
Figure 2.2.3.5 Distribution of R, F, M	16
Figure 2.2.3.6 Distribution of R, F and M (after transforming)	17
Figure 2.2.3.7 Dataset after scaling	17

Figure 2.2.4.1 Heat map showing the correlation between variables R, F and M	18
Figure 2.2.4.2 Correlation between Frequency (F) and MonetaryValue (M)	19
Figure 3.1 Model Steps	20
Figure 3.2.1 Elbow Analysis	21
Figure 3.2.2 Review by Silhouette	22
Figure 3.2.3.1 Fit Data to KMeans	23
Figure 3.2.3.2 Visualize Clusters	24
Figure 3.3.1 Comparing cluster	25
Figure 4.1.1.1 Calculate the value of AOV	27
Figure 4.1.1.2 Calculate the value of APFR	28
Figure 4.1.1.3 Calculate the value of CV	28
Figure 4.1.1.1 Calculate the value of CLV	29
Figure 5.1.1 The dataset for prediction	46

List of Acronyms

DB	Digital Business
MIS	Management Information Systems
FDA	Fundamentals of Data Analytics
CLV	Customer Lifetime Value
CRR	Customer Retention Rate

GANTT CHART



ABSTRACT

Mining customer data helps businesses improve and enhance business capabilities and generates huge benefits. The RFM model is a way for businesses to segment customers and thereby make positive changes to maximize profits. In this project, we conduct analysis that explores customer raw data. From there, detect insights in the dataset. The RFM model will then be applied to calculate the Recency, Frequency, MonetaryValue indicators combined with the Kmeans clustering method to be able to group customers with the same characteristics. Finally, calculate the Customer Lifetime Value (CLV) and Customer Retention Rate (CRR). From the results, we will propose future business directions.

Project Overview

Business problems

Vietnamese businesses increasingly want to develop in a long-term sustainable way. There are many methods researched and given by businesses to optimize profits. Many businesses are interested in attracting new target customers , who are only familiar with and use their products and services in a short time.

However, the truth is that loyal customers play a much more important role in the overall proportion of customers that businesses hold. Long-time customers are also supportive and close friends with the business. So, if you get a large number of loyal customers, you will be able to save costs while still being able to increase sales revenue effectively. Therefore, it is important to pay attention to statistics, analyzing Customer Lifetime Value (CLV) (is the value of a customer contributing to your company throughout their life. Loyal customers are the ones who bring long-term and sustainable profits to the business because of the high lifetime value) and CLV value is long-term, echoing the benefits of better ROI and unit economics. It's a completely different strategy than a short-term sales strategy. The problem is that conversion-based growth needs ongoing marketing spend, and businesses only grow at a cost.

Objectives

In this study, the goal is to build an RFM model combined with KMeans methodology and CLV and CRR indicators to group customers based on their own characteristics. From there, understanding customer lifecycle value will help businesses optimize marketing strategies to improve revenue, market share and profit.

With the above objectives, the following questions will be identified as directions and research objectives:

- How to build an RFM model that incorporates K-Means to get customer groups with corresponding characteristics?
- How should businesses focus policies to maximize revenue based on separate customer groups?

Object

This research focuses on analyzing and mining data on the subjects then segmenting the corresponding customers combined with clustering. From clustering, companies can plan and optimize marketing costs for marketing campaigns.

Besides, determining the CLV of each customer will also help businesses come up with appropriate marketing strategies for each of them. For customers with high CLV, businesses need to maintain customer care activities, as well as provide more value for each purchase to retain them longer. For customers with low CLV, businesses need to have appropriate promotions and stimulus to encourage them to come back in the future.

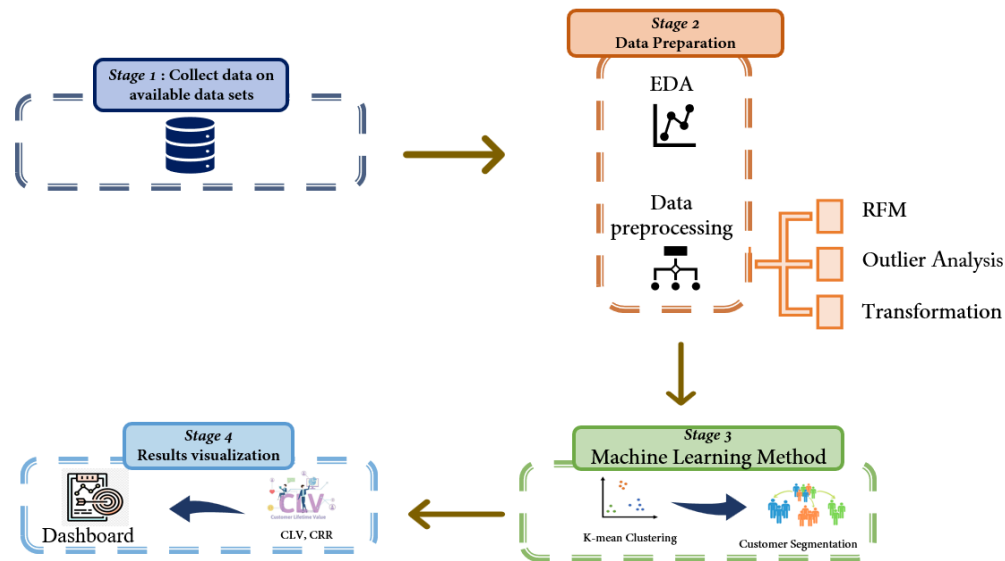
Project duration: The project is implemented from 21/11/2022 - 21/12/2022.

The data duration of the dataset is from 2015 to 2016, from the AdventureWorks dataset. Besides, spatial range is about RFM analysis with traditional methods, K-means clustering and some other support methods (BoxCox, Elbow,...)

Research Methodology

To accomplish specific objectives, we implemented the project using data mining methodology - RFM analysis (Average purchase value, Average purchase frequency, average customer value index) combined with KMeans, CRR and CLV clustering methods to determine customer lifecycle value.

Proposed Model



The project that we are going to present consists of 4 phases and is executed from top to bottom as shown in the figure. Firstly, in this research we use the AdventureWorks database supported by Microsoft, we will define the objective of this study and collect data that is relevant to the goal. In stage 2, we will clean the data first by removing outliers, and faulty values then transform and use exploratory data analysis (EDA) for the data set. Through the above 2 phases, the data of our input is relatively good, it is easy to deploy the 3rd stage, which is to put the data into the model for analysis. And in the final stage, we will use this result to evaluate and develop a suitable strategy for the business through the proposed model.

Structure of Project

Chapter 1 - Theoretical

Chapter 2 - Data Preparation

Chapter 3 - Customer segmentation with machine learning method

Chapter 4 - Customer Retention Rate (CRR) and Customer Lifetime Value (CLV)

Chapter 5 - Predict CLV by Linear Regression

Chapter 1 Theoretical

1.1 Machine learning

Machine learning (ML) or machine learning is a branch of artificial intelligence (AI), it is a field of study that allows computers to have the ability to improve themselves based on training data or training data. based on experience (what has been learned). Machine learning can predict or make decisions on its own without being specifically programmed.

Machine learning workflow:

- + Data collection: in order for the computer to learn you need a data set, you can collect them yourself or get previously published data sets. Note that you must collect from an official source, so the new data is accurate and the machine can learn properly and achieve higher efficiency.
- + Preprocessing: this step is used to normalize data, remove unnecessary attributes, assign data labels, encode some features, extract features, reduce data but still ensure results...
- + Training model: this step is the step where you train the model or let it learn on the data you have collected and processed.
- + Evaluating model: after training the model, we need to use the metrics to evaluate the model, depending on the different metrics, the model is also evaluated well or not. Model accuracy above 80% is considered good.
- + Improve: after the model has been evaluated, the models with poor accuracy need to be retrained, we will repeat from step 3, until the expected accuracy is reached.

1.2 RFM (Recency – Frequency – Monetary Value)

RFM is a model used to analyze customer value, thereby helping businesses to analyze each customer group they have, from which to have marketing campaigns or make business decisions.

R (Recency): The time period in which the customer made the most recent purchase.

F (Frequency): Frequency of customer purchases. This index is often interested in businesses with services and products with low profit value.

M (Monetary Value): This index is used to calculate the material value that the business has every time a customer uses the service.

1.3 KMeans

K-means clustering is a vector quantization method used to classify given data points into different clusters. K-means clustering has many applications, but is most used in Artificial Intelligence and Machine Learning (specifically Unsupervised Learning).

The input condition of the algorithm must specify the value of k , but in practice it is not always possible to know in advance how many groups there are. This problem can be improved by the following methods: Elbow, Silhouette

1.4 Customer Lifetime Value (CLV)

Customer Lifetime Value - CLV is the time period that a customer starts buying/using a product/service from a business until it stops using it.

Thus, (Customer Lifetime Value - CLV for short) is the total profit earned from customers during the time they buy or use the products/services of the business, used as an indicator to reflect the profitability. information of a customer to a business, thereby helping businesses determine which customers are worth investing in and exploiting.

Determining the CLV of each customer will also help businesses come up with appropriate marketing strategies for each of those objects. Usually, for customers with

high CLV, businesses need to maintain customer care activities, as well as provide additional value for each purchase to retain them longer. For customers with low CLV, businesses need to have appropriate promotions and stimulus programs to encourage them to return in the near future.

1.5 Customer Retention Rate (CRR)

Customer Retention Rate (CRR) is understood as the number of customers that a business can successfully retain for a certain period of time as soon as they have or use their products or services. Specifically, compared to the entire customer that the business has.

Customer retention rate is considered an important factor because it shows the level of interest and interaction as well as customer loyalty to the business. From there, businesses can quickly capture the groups of loyal customers to come up with measures to motivate customers to return to buy and save costs.

Chapter 2 Data Preparation

2.1 Stages of EDA

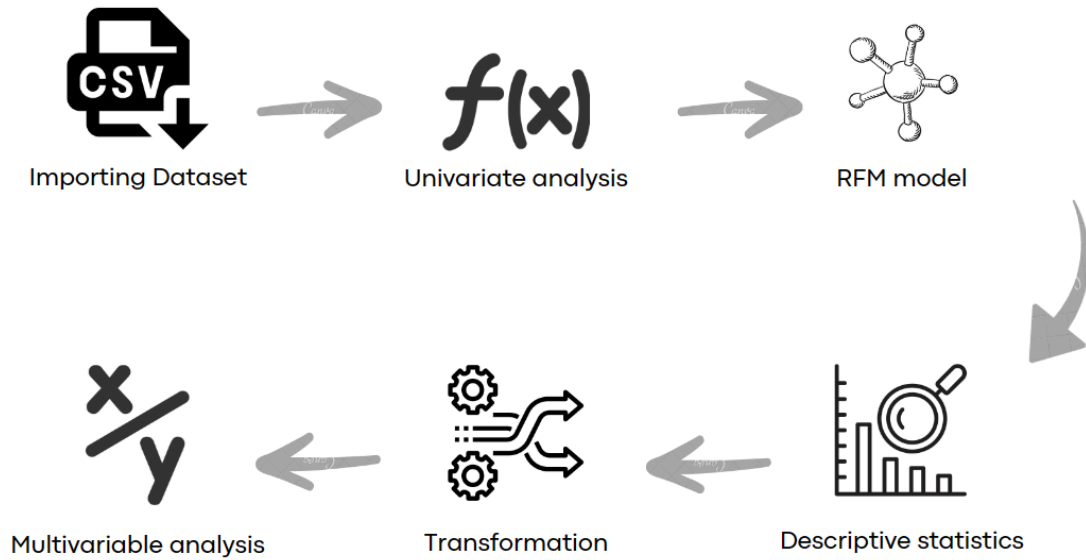


Figure 2.1 Stages of EDA

2.2 Exploratory Data Analysis (EDA)

2.2.1 Distinguish Attribute

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 121253 entries, 0 to 121252
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  -
0   CustomerKey            121253 non-null float64
1   Orderdate              121253 non-null datetime64[ns]
2   ProductKey             121253 non-null float64
3   Total Product Cost     121253 non-null float64
4   Sales Amount           121253 non-null float64
5   Sales Order            121253 non-null object
dtypes: datetime64[ns](1), float64(4), object(1)
memory usage: 6.5+ MB
```

	CustomerKey	Orderdate	ProductKey	Total Product Cost	Sales Amount	Sales Order
0	-1.0	2017-07-02	349.0	1898.0944	2024.994	SO43659
1	-1.0	2017-07-02	350.0	5694.2832	6074.982	SO43659
2	-1.0	2017-07-02	351.0	1898.0944	2024.994	SO43659
3	-1.0	2017-07-02	344.0	1912.1544	2039.994	SO43659
4	-1.0	2017-07-02	345.0	1912.1544	2039.994	SO43659

Figure 2.2.1.1 Overview of original Dataset

The added dataset consists of 121253 lines and 6 columns (no null values). These data will be used to calculate RFM model metrics, CLV, CRR indicators and put into the model to perform predictive analysis.

Column	Type	Table	Description
Sales Order	int	Sales_data	Auto increment
CustomerKey	int	Sales_data	Customer identification number
Orderdate	date	Sales_data	Date that an order was created
Total Products cost	float	Sales_data	The total cost each product
ProductKey	int	Sales_data	Product identification number
Sale Amount	float	Sales_data	The total amount paid by each customer with each product

Figure 2.2.1.2 Distinguish Attribute

2.2.2 Univariate analysis

Remove -1

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 60398 entries, 60855 to 121252
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  -
0   CustomerKey            60398 non-null  float64
1   Orderdate              60398 non-null  datetime64[ns]
2   ProductKey             60398 non-null  float64
3   Total Product Cost     60398 non-null  float64
4   Sales Amount           60398 non-null  float64
5   Sales Order            60398 non-null  object
dtypes: datetime64[ns](1), float64(4), object(1)
memory usage: 3.2+ MB

```

	CustomerKey	Orderdate	ProductKey	Total Product Cost	Sales Amount	Sales Order
60855	21768.0	2017-07-01	310.0	2171.2942	3578.2700	SO43697
60856	28389.0	2017-07-01	346.0	1912.1544	3399.9900	SO43698
60857	25863.0	2017-07-01	346.0	1912.1544	3399.9900	SO43699
60858	14501.0	2017-07-01	336.0	413.1463	699.0982	SO43700
60859	11003.0	2017-07-01	346.0	1912.1544	3399.9900	SO43701

Figure 2.2.2.1 Data after removing value CustomerKey = -1

After removing the line containing the -1 value in the CustomerKey column, we see that they exist quite a bit in this dataset (from 121253 lines to 60398 lines). This has a great influence on the later analysis of the project, so the immediate task is to eliminate them.

Counting non-duplicate quantities

```

1 dt.nunique()

```

CustomerKey	18484
Orderdate	1081
ProductKey	158
Total Product Cost	45
Sales Amount	42
Sales Order	27659
dtype: int64	

Figure 2.2.2.2 Counting non-duplicate quantities

Looking at the image above, we see that CustomerKey only has 18484 values, but Sales Order has 276659 different values. This represents duplication of the Customerkey variable.

Datatype conversion

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 60398 entries, 60855 to 121252
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  -
0   CustomerKey            60398 non-null  object
1   Orderdate              60398 non-null  datetime64[ns]
2   ProductKey             60398 non-null  object
3   Total Product Cost    60398 non-null  float64
4   Sales Amount           60398 non-null  float64
5   Sales Order            60398 non-null  object
dtypes: datetime64[ns](1), float64(2), object(3)

```

Figure 2.2.2.3 Datatype after convert

Looking at Figure 2.2.1.1, we can see that the number of variables with datetime64, float, and object data types is 1, 4, 1 respectively. But now they have been changed to 1, 2, 3 because we have changed the data type of the CustomerKey and ProductKey variables from float64 to object.

2.2.3 Pre-processing and EDA

RFM calculation

RFM is a model used to identify a company or organization's best customers by measuring and analyzing spending habits to improve low-scoring customers and retain high-scoring customers.

- Last time (R) represents the number of days since each customer's last purchase.
- Frequency (F) is the number of times a customer buys over a specified period of time.
- Currency (M) illustrates the customer's total cost for all previous purchases.

	Recency	Frequency	MonetaryValue
CustomerKey			
11000.0	256.0	3	8248.9900
11001.0	35.0	3	6383.8800
11002.0	325.0	3	8114.0400
11003.0	249.0	3	8139.2900
11004.0	258.0	3	8196.0100
...
29479.0	497.0	1	2049.0982
29480.0	181.0	1	2442.0300
29481.0	885.0	1	3374.9900
29482.0	483.0	1	2049.0982
29483.0	492.0	1	2049.0982

18484 rows × 3 columns

Figure 2.2.3.1 RFM calculation

After the calculation process, the number of lines is only 18484 instead of 60398 as before. Demonstrate that duplicate data have been grouped to calculate the RFM model in the most accurate way.

Descriptive statistics

	Recency	Frequency	MonetaryValue
count	18484.00	18484.0	18484.00
mean	175.67	1.5	1588.33
std	145.64	1.1	2124.23
min	1.00	1.0	2.29
25%	72.00	1.0	49.97
50%	154.00	1.0	270.26
75%	249.00	2.0	2511.28
max	1081.00	28.0	13295.38

Figure 2.2.3.2 Descriptive statistics

Above is a descriptive table of statistics for the values R, F and M. Notice that the range (max – min) of the 3 variables is quite large but the mean value tends to be closer to the mean value. We can initially assume that this dataset has outlier values.

Outliers

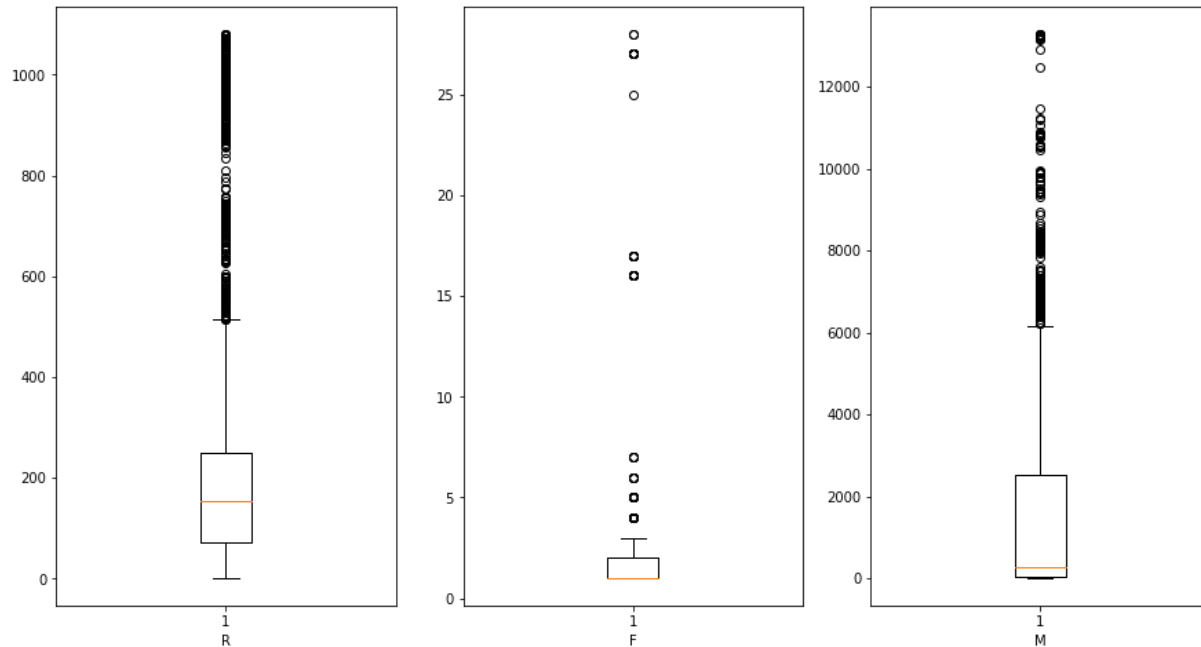


Figure 2.2.3.3 Before Removing Outliers

After drawing the box-plot, we immediately confirm that there are many outlier values in the data set. These values will cause discrepancies when we calculate CLV indicators for the general level of the data. Therefore it is our duty to delete them.

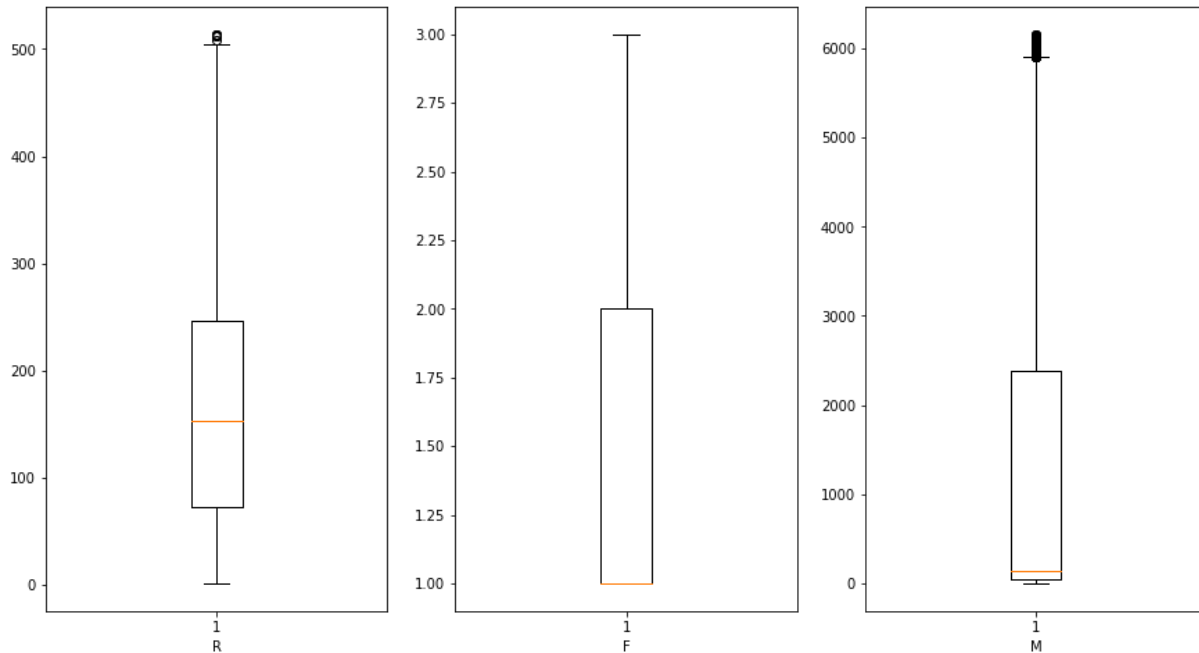


Figure 2.2.3.4 After Removing Outliers

Transformation

We also found another problem that the distribution of data is currently very skewed. We need to convert them so that the distribution is more balanced so that it is easy to compare and calculate.

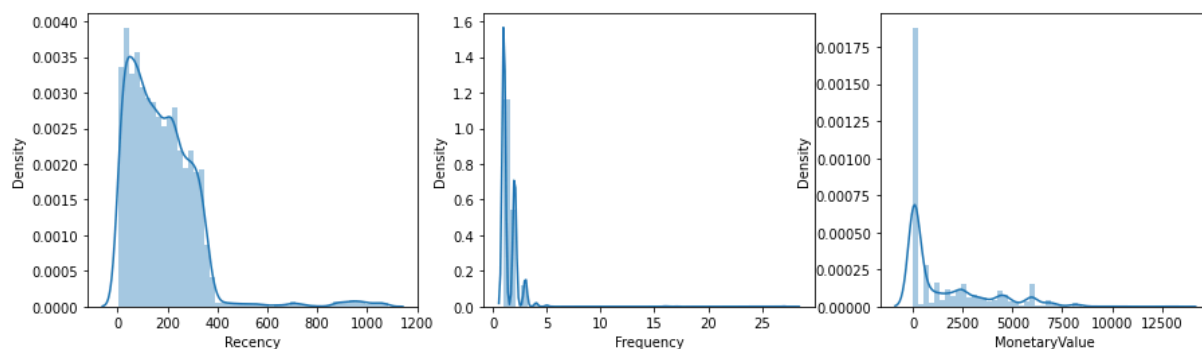
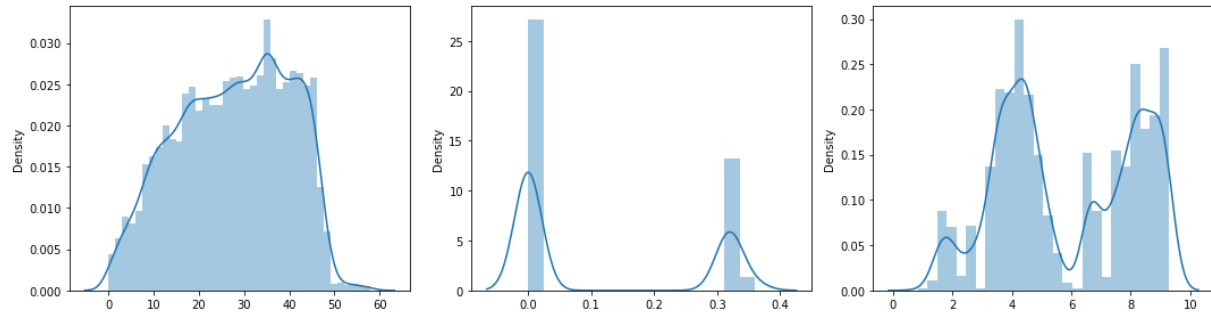


Figure 2.2.3.5 Distribution of R, F, M

Figure 2.2.3.1 shows the unevenness of data distribution. Although exceptions have been eliminated, there are still distribution disparities such as Recency and MonetaryValue

skewed to the right. This affects quite a bit of data analysis in the following steps. Therefore, they need to be converted so that they follow the standard distribution or their graph is bell-shaped as shown in Figure 2.2.3.2



2.2.3.6 Distribution of R, F and M (after transforming)

Scale

	0	1	2
0	1.442304	3.338898	1.427349
1	-0.074828	3.338898	1.583286
2	2.913521	3.338898	1.673381
3	3.593296	1.268247	3.042437
4	3.454623	1.268247	3.008737

Figure 2.2.3.7 Dataset after scaling

Since the units of the 3 columns R, F and M are different and the spread between these 3 variables is also quite large, we need to scale them to the same interval to facilitate comparison and analysis.

2.2.4 Multivariable analysis

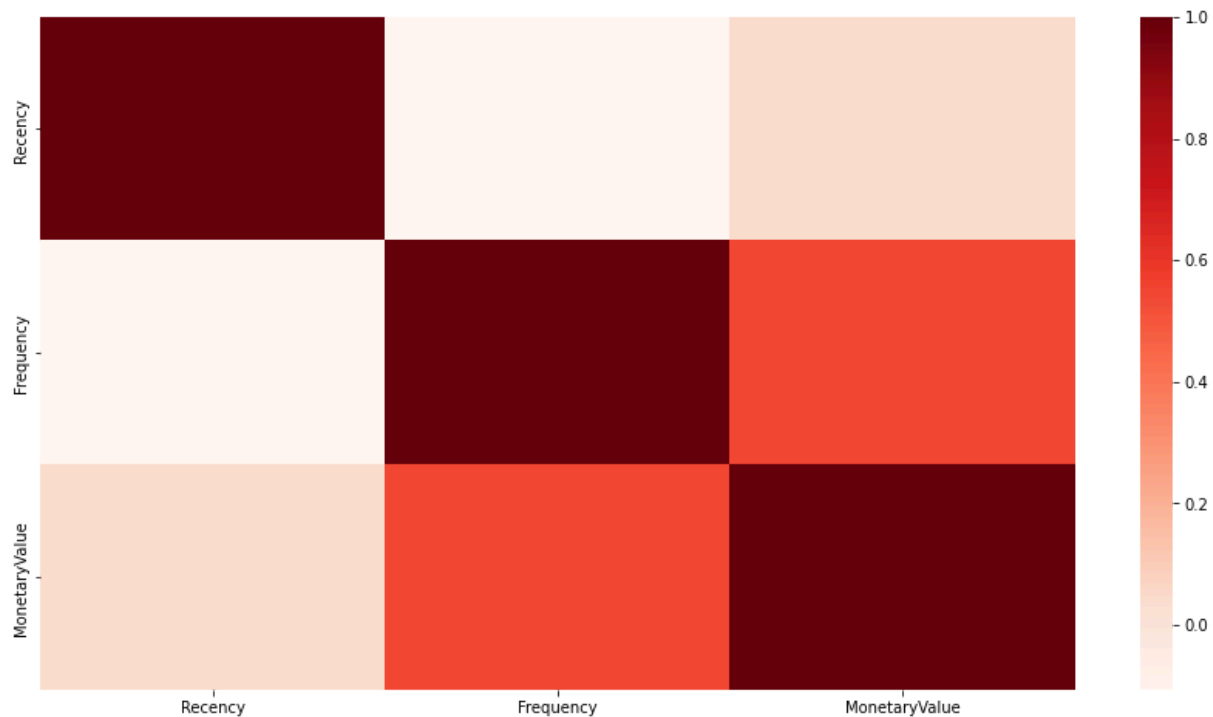


Figure 2.2.4.1 Heat map showing the correlation between variables R, F and M

Looking at the heat-map, the 3 variables R, F and M seem to have very little effect on each other except for the pair F and M.

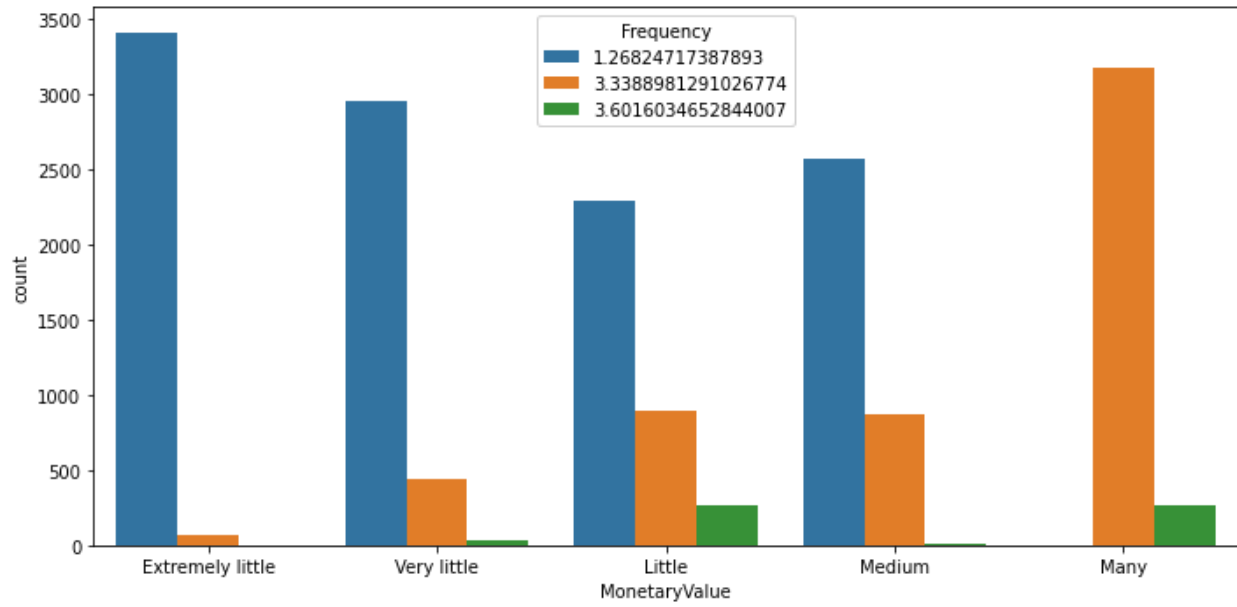


Figure 2.2.4.2 Correlation between Frequency (F) and MonetaryValue (M)

It's easy to see that the more often we buy, the greater the amount of money spent.

Chapter 3 Customer segmentation with machine learning method

3.1 Model Steps

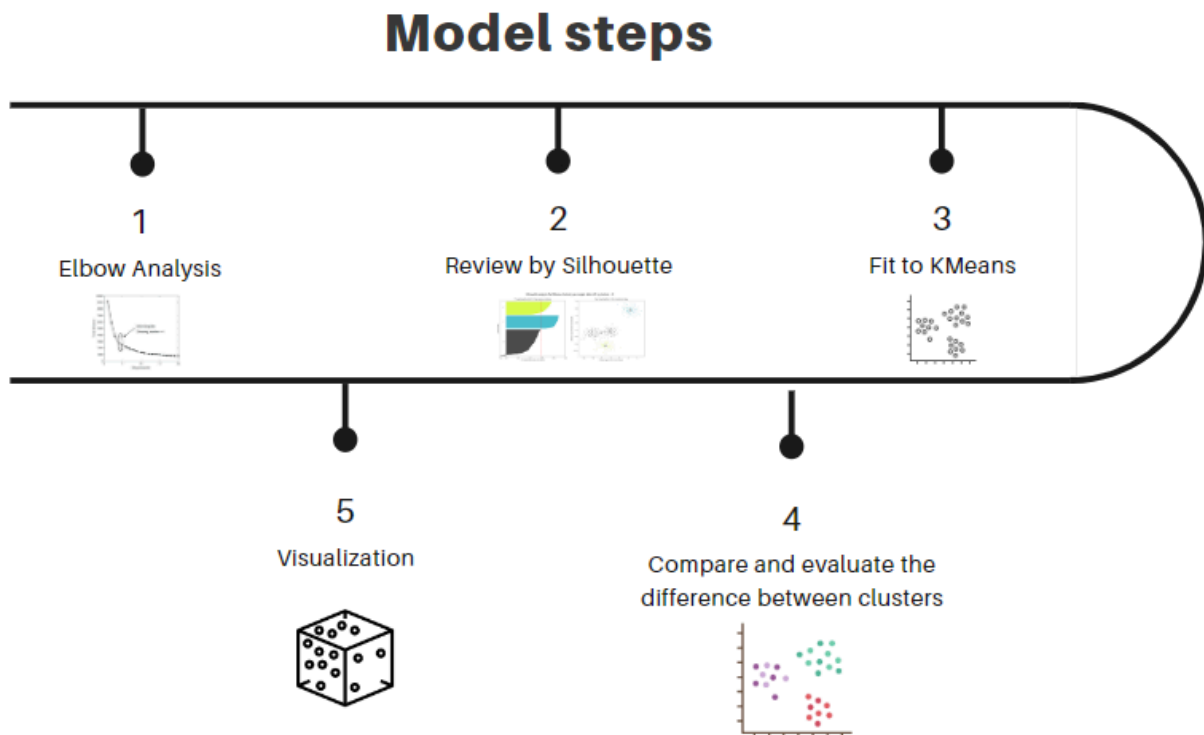


Figure 3.1 Model Steps

This chapter consists of 5 steps. First, once the data has been scaled to a certain interval, we begin the analysis of cluster selection with the Elbow method. Step 2 we will go to the evaluation by scoring the selected clusters to give the most optimal number of clusters using Silhouette. Now that we have selected an elite number, the data will be fit to KMeans to conduct clustering. By the fourth step, we need to analyze and compare the different characteristics between the phrases and thereby perform the labeling. Finally, data from clusters will be visualized on a 3D scatter plot.

3.2 Clustering Process

3.2.1 Elbow analyst

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k . As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters. Here, we let K run from 2 to 15 as figure below.

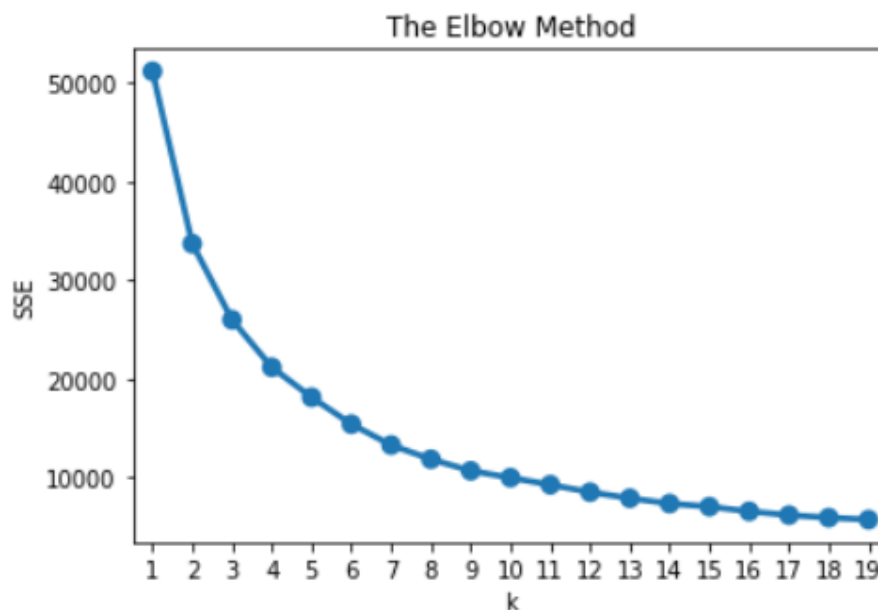


Figure 3.2.1 Elbow Analysis

3.2.2 Review by Silhouette

When performing partial clustering (partition) in general and K-meaning clustering in particular, we need to pre-determine the number of clusters (clusters). To do this, we will

use the Silhouette icon. A Silhouette chart is a method of representing and determining the stability of data points (samples) in the same phrase. Each data point (sample) will be calculated Silhouette value, it has a value from -1 to 1, with the higher value (closer to 1) indicating its similarity with individuals in the same the higher the cluster, and at the same time it is different from the individuals in other clusters.

With the above data set, we will try to divide it into 4, 5, 6, 7 clusters and analyze which number of clusters is appropriate thanks to the Silhouette chart.

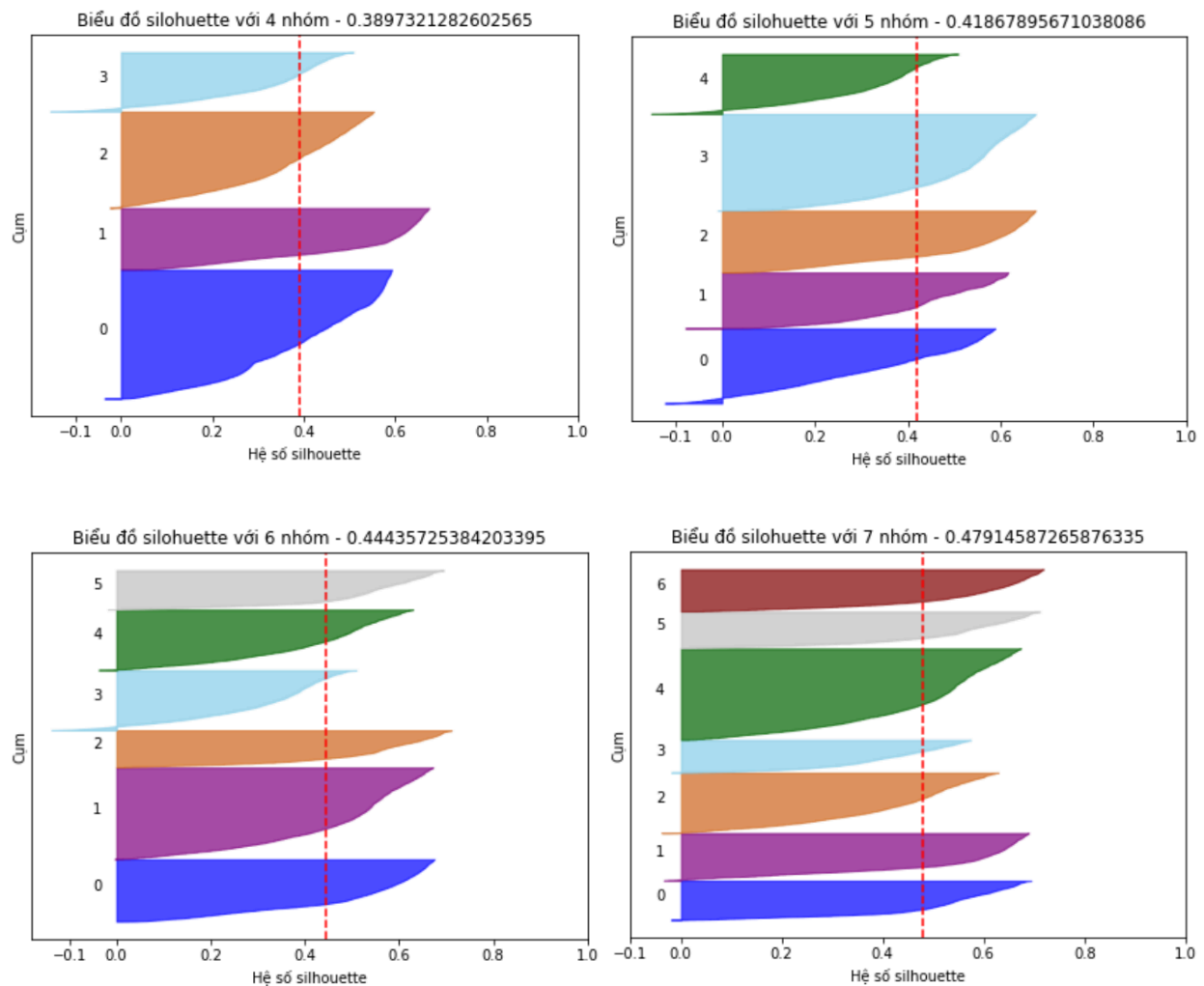


Figure 3.2.2 Review by Silhouette

3.2.3 Fit to Kmeans

According to the K-means algorithm we just presented above, the library of sklearn is available because scientists have built it to serve the analysis and data processing more conveniently. All of the above algorithms are easily handled with the model Kmean. After referring to two methods of cluster segments: Elbow and Silhoutte, we divided customers into 5 segments.

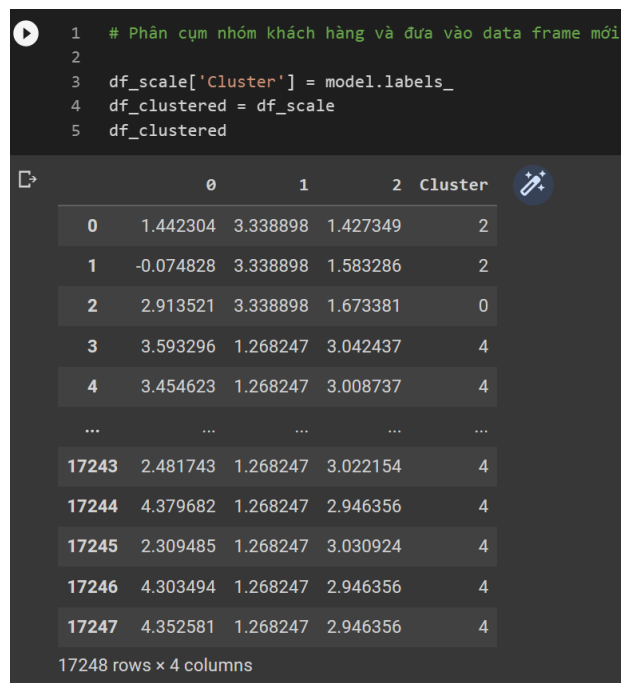


Figure 3.2.3.1 Fit Data to KMeans

Predictions about customer groups after processing will be put in an array and printed like that in the order of the data set. But it doesn't look that good. We'll visualize it on an interactive 3d chart. We'll use different colors for clustering for clarity.

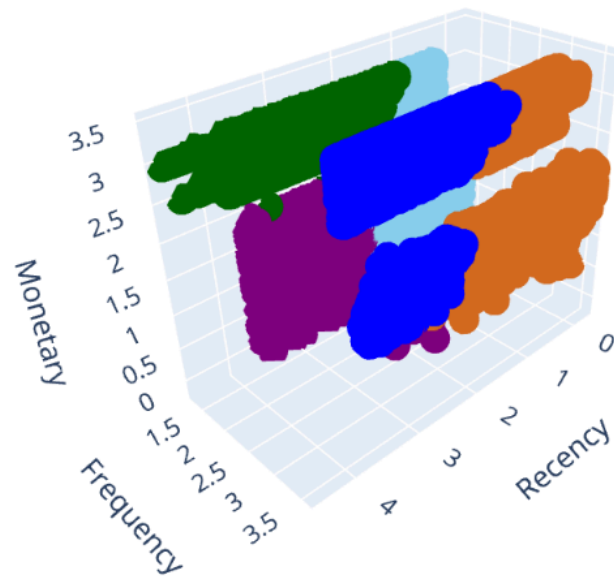


Figure 3.2.3.2 Visualize Clusters

3.3 Comparing clusters and Discussing business

3.3.1 Comparing cluster

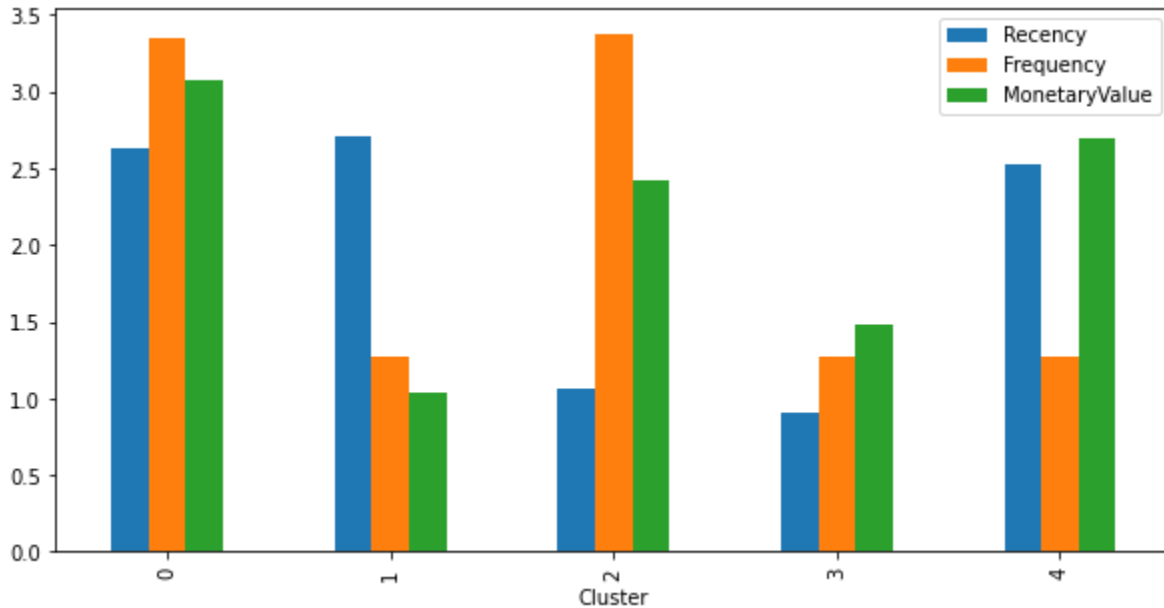


Figure 3.3.1 Comparing cluster

3.3.2 Discussing business

Based on the RFM score, we labeled for each cluster as below:

- Cluster 0: Cannot lose them (3076)
- Cluster 1: Hibernating (4750)
- Cluster 2: Champions (2942)
- Cluster 3: Loyal customer (3672)
- Cluster 4: Customers needing attention (2808)

Cannot lose them

With the number of groups that can't lose them at around 17.83%, businesses need to care more about them so they can make a purchase again. At this time, businesses will gain a lot of benefits because they have spent a huge amount of money.

Hibernates

The group of customers hibernating accounts for a very large number (27.54%). These must be customers who have made a purchase but they are not very interested in our business. Therefore, businesses need to improve and add more interesting incentives and purchasing policies.

Champions

17.06% is a bit small compared to the general level, but this is an extremely important customer group for businesses. We need to keep them as long as possible. This requires businesses to maintain their interest and also improve services so that customers have a better experience.

Loyal customers

This is a group of loyal customers for businesses. Maybe they have made good use of promotions or maybe they are only interested in low-cost products so the amount of money they spend is not much, but they have a very positive frequency and recency index. It's important for businesses to get this group of customers to spend more money on each of their purchases.

Customers needing attention

If this group of customers increases the RF index and maintains M, it can be said that this is an extremely potential group. Attention should be paid to them and implement stronger marketing.

Chapter 4 Customer Lifetime Value and Customer Retention Rate

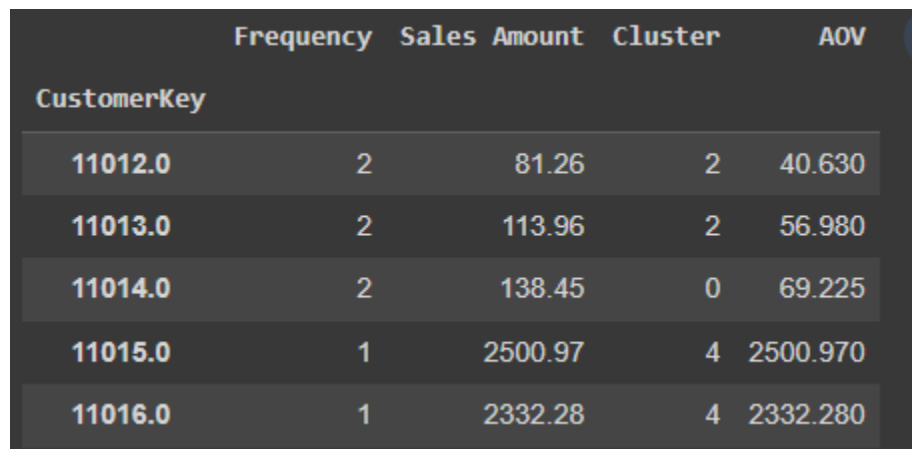
4.1. Predictive analytics CLV

4.1.1 Customer Lifetime Value Method (CLV)

Step 1: Calculate the average purchase value of each customer (Average Order Value - AOV)

Average purchase value (APV) is the average temporary profit earned per order sold. APV can be determined by using the following formula:

$$AOV = \frac{(Total\ revenue)}{Number\ of\ orders}$$



CustomerKey	Frequency	Sales Amount	Cluster	AOV
11012.0	2	81.26	2	40.630
11013.0	2	113.96	2	56.980
11014.0	2	138.45	0	69.225
11015.0	1	2500.97	4	2500.970
11016.0	1	2332.28	4	2332.280

Figure 4.1.1.1 Calculate the value of AOV

Step 2: Calculate the average frequency of purchases (Average Purchase Frequency Rate - APFR)

Average Purchase Frequency Rate (APFR) is the average number of purchases a customer makes with your company over a defined period of time. APFR can be determined using the following formula:

$$APFR = \frac{\text{Total Orders}}{\text{Total number of customers}}$$

```
APF = sum(df_pre_CLV['Frequency'])/df_pre_CLV.shape[0]
APF
1.3822472170686457
```

Figure 4.1.1.2 Calculate the value of APFR

Step 3: Calculating customer value (Customer Value - CV)

Average customer value is an indicator that reflects the average net profit of a customer for a company. The CV can be determined using the following formula:

$$CV = AOV \times APFR$$

CustomerKey	Frequency	Sales Amount	Cluster	AOV	CV
11012.0	2	81.26	2	40.630	56.160704
11013.0	2	113.96	2	56.980	78.760446
11014.0	2	138.45	0	69.225	95.686064
11015.0	1	2500.97	4	2500.970	3456.958822
11016.0	1	2332.28	4	2332.280	3223.787539

Figure 4.1.1.3 Calculate the value of CV

Step 4: Calculate average customer lifecycle (Customer Lifecycle - CL) by using Churn Rate (CR)

Average customer lifecycle (ACL) is the average amount of time a customer maintains a purchase of a business's goods/services. The ACL can be determined using the following formula:

$$ACL = \text{Total } CL_{\text{Customer}} / \text{Total Customer}$$

and Total $CL_{Customer}$ = Last Purchase Date - First Purchase Date

```
# Repeat Rate
repeat_rate = df_pre_CLV[df_pre_CLV.Frequency > 1].shape[0]/df_pre_CLV.shape[0]

# Churn Rate
churn_rate = 1 - repeat_rate

# Customer Lifespan
CL = 1 / churn_rate
CL

1.5365701559020046
```

Figure 4.1 Calculate the value of ACL

Step 5: Calculate Customer Lifetime Value (Customer Lifetime Value - CLV)

Finally, the customer lifecycle value (CLV) can be determined by the customer value multiplied by the average customer lifecycle, specifically with the formula:

$$CLV = CV * ACL$$

	Customer Lifespan	Frequency	Sales Amount	Total Product Cost	Cluster	AOV	CV	CLV
CustomerKey								
11012.0	213	2	81.26	30.3914	2	40.630	56.160704	86.294862
11013.0	272	2	113.96	42.6212	2	56.980	78.760446	121.020951
11014.0	38	2	138.45	51.7805	0	69.225	95.686064	147.028350
11015.0	0	1	2500.97	1333.3060	4	2500.970	3456.958822	5311.859757
11016.0	0	1	2332.28	1265.9278	4	2332.280	3223.787539	4953.575722

Figure 4.1.1 Calculate the value of CLV

4.2 Customer Retention Rate (CRR)

4.2.1 Calculate the CRR

Step 1 : Create dataframe to calculate CRR and separate year from Orderdate

```
[ ] 1 df_CRR = dt.drop(['ProductKey','Total Product Cost','Sales Amount'], axis = 1)
    2

▶ 1 df_CRR['year'] = df_CRR['Orderdate'].apply(lambda x: x.strftime('%Y'))
  2 df_CRR['year'] = pd.to_numeric(df_CRR['year'])
  3 df_CRR
```

	CustomerKey	Orderdate	Sales Order	year
60855	21768.0	2017-07-01	SO43697	2017
60856	28389.0	2017-07-01	SO43698	2017
60857	25863.0	2017-07-01	SO43699	2017
60858	14501.0	2017-07-01	SO43700	2017
60859	11003.0	2017-07-01	SO43701	2017
...
121248	15868.0	2020-06-15	SO75122	2020
121249	15868.0	2020-06-15	SO75122	2020
121250	18759.0	2020-06-15	SO75123	2020
121251	18759.0	2020-06-15	SO75123	2020
121252	18759.0	2020-06-15	SO75123	2020

60398 rows x 4 columns

Figure 4.2.1.1 Create dataframe to calculate CRR

Step 2 : Find the first purchase of each customer with ‘pandas.groupby’ operation by customerkey and find the first time of Orderdate

```
1 # Tìm lần mua hàng đầu tiên của từng khách hàng
2 df_1st_buy = df_CRR.groupby('CustomerKey', as_index=False).agg({'Orderdate': 'first'})
3 df_1st_buy['Customer type'] = 'First-time'
4 df_1st_buy
```

	CustomerKey	Orderdate	Customer type
0	11000.0	2017-07-12	First-time
1	11001.0	2017-07-09	First-time
2	11002.0	2017-07-05	First-time
3	11003.0	2017-07-01	First-time
4	11004.0	2017-07-14	First-time
...
18479	29479.0	2019-02-05	First-time
18480	29480.0	2019-12-18	First-time
18481	29481.0	2018-01-13	First-time
18482	29482.0	2019-02-19	First-time
18483	29483.0	2019-02-10	First-time

18484 rows x 3 columns

Figure 4.2.1.2 Find the first purchase

Step 3 : Combine 2 data sets to create a complete dataframe to calculate CRR

```
] 1 df_CRR = pd.merge(df_CRR,df_1st_buy, how='outer')
2 df_CRR['Customer type'].fillna("2nd onwards", inplace = True)
3 df_CRR
```

	CustomerKey	Orderdate	Sales Order	year	Customer type
0	21768.0	2017-07-01	SO43697	2017	First-time
1	28389.0	2017-07-01	SO43698	2017	First-time
2	25863.0	2017-07-01	SO43699	2017	First-time
3	14501.0	2017-07-01	SO43700	2017	First-time
4	11003.0	2017-07-01	SO43701	2017	First-time
...
60393	15868.0	2020-06-15	SO75122	2020	2nd onwards
60394	15868.0	2020-06-15	SO75122	2020	2nd onwards
60395	18759.0	2020-06-15	SO75123	2020	2nd onwards
60396	18759.0	2020-06-15	SO75123	2020	2nd onwards
60397	18759.0	2020-06-15	SO75123	2020	2nd onwards

60398 rows x 5 columns

Figure 4.2.1.3 Combine 2 data sets

Step 4 : Create a function to calculate CRR

You can calculate the Customer Retention Rate by using a simple formula that has three variables. The idea is to take the difference between the number of customers at the end of the period (E) and the new customer acquired (N). To that difference, we will divide by the number of customers at the start of the period (S).

The CRR's mathematical representation is **$CRR = ((E-N)/S) \times 100$**

For example, let's assume you had 200 customers at the end of 2019 (S), and you want to know your Customer Retention Rate for 2020. At the end of 2020, you had 375 customers (E), where 175 were new ones (N).

Consequently, we will have $((375 - 235))/200 = 140/200 = 0.7 \times 100 = 70\%$. The Customer Retention Rate for that store is therefor 70%.

```
1 # Hàm tính CRR theo năm
2 def customer_retention_rate_year(df, end_year):
3     try:
4         # Get all the customers in 2019 ("E")
5         S = df_CRR.loc[df_CRR.year == end_year-1].CustomerKey.nunique()
6         # Get all the first time customers in 2020 ("N")
7         N = df_CRR.loc[df_CRR.year == end_year].loc[df_CRR['Customer type'] == "First-time"].CustomerKey.nunique()
8         # Get the customer_id number of customers in 2020
9         E = df_CRR.loc[df_CRR.year == end_year].CustomerKey.nunique()
10        return ((E-N)/S)*100
11    except ZeroDivisionError:
12        return 0

1 print(customer_retention_rate_year(df_CRR,2020))
```

Figure 4.2.1.4 Create a function to calculate CRR

4.2.2 CRR Visualization

The number of customers in the first 2 years is very low and there is a spike in 2019 and 2020

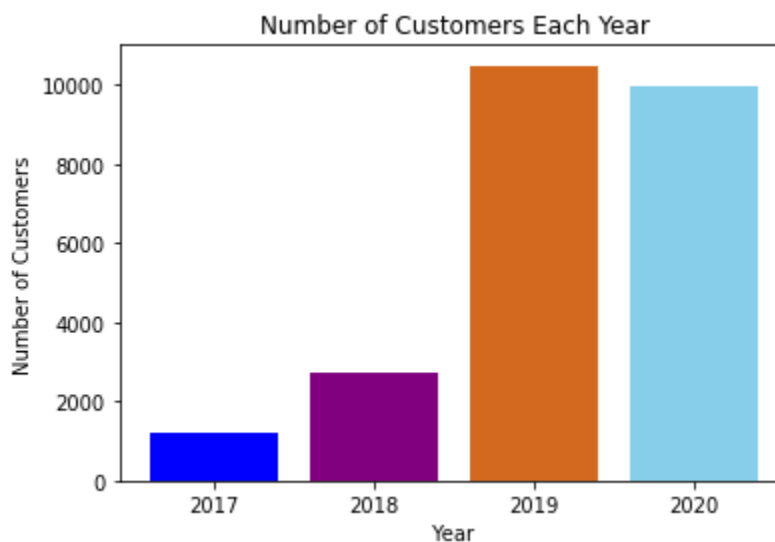


Figure 4.2.2.1 Number of Customer Each Year

In 2018 the number of new customers is 100% and no customer has returned since 2017

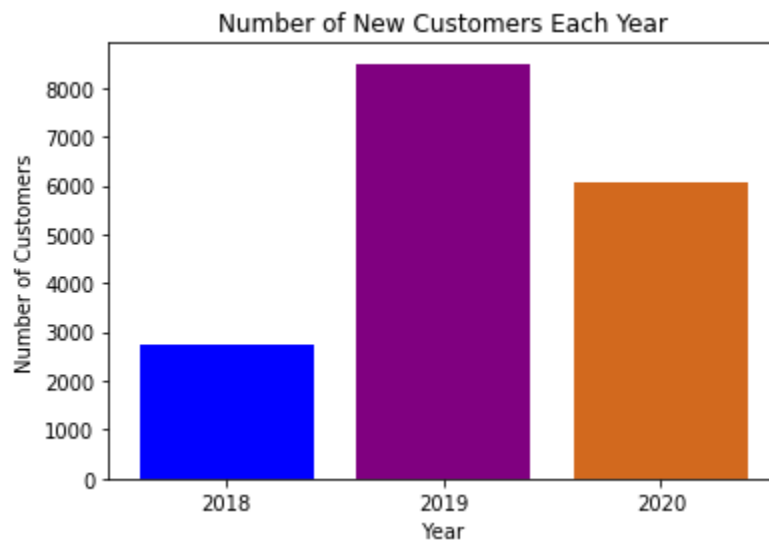


Figure 4.2.2.2 Number of New Customers Each Year

Customer Retention Rate of 2018 was zero and accounted for quite high in 2019 and 2020

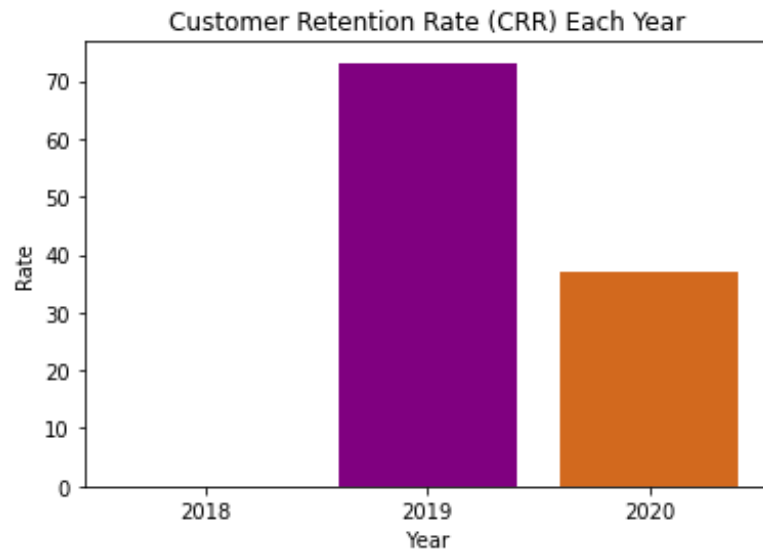


Figure 4.2.2.3 Customer Retention Rate Each Year

Chapter 5 Linear Regression

5.1 Preparing the dataset

```
1 df_linear = df_pre_CLV.drop(['AOV', 'Total Product Cost', 'CV'], axis = 1)
2 df_linear
```

CustomerKey	Frequency	Sales	Amount	CLV
11012.0	2	81.2600	86.294862	
11013.0	2	113.9600	121.020951	
11014.0	2	138.4500	147.028350	
11015.0	1	2500.9700	5311.859757	
11016.0	1	2332.2800	4953.575722	
...	
29478.0	1	2398.0500	5093.265929	
29479.0	1	2049.0982	4352.120284	
29480.0	1	2442.0300	5186.675922	
29482.0	1	2049.0982	4352.120284	
29483.0	1	2049.0982	4352.120284	

Figure 5.1.1 The dataset for prediction

The dataset is inherited from the dataset after the CLV index is calculated. We also need to remove unnecessary columns 'AOV', 'Total Product' and 'CV'

	Frequency	Sales Amount
CustomerKey		
25684.0	1	23.7800
16021.0	1	539.9900
26191.0	2	1938.4700
12281.0	3	5169.3582
25304.0	1	67.4900

Figure 5.1.2 Independent variables

```


CustomerKey
25684.0      50.506813
16021.0     1146.895465
26191.0     2058.577429
12281.0     3659.767449
25304.0      143.343349
Name: CLV, dtype: float64

```

Figure 5.1.3 Dependent variable

From the original data set, we will split them into 2 parts with a ratio of 8:2. With 80% we will use it to train the Linear Regression model with independent variables is Frequency and Sales Amount and dependent on them is CLV. The remaining 20% will be used to evaluate the accuracy.

5.2 Training



	Predictions	y(actual)	x1	x2
CustomerKey				
16516.0	5350.059258	5027.052728	2	4733.75
24676.0	3468.308366	5228.134837	1	2461.55
17903.0	479.515072	68.538893	1	32.27
19386.0	1910.968615	2058.577429	2	1938.47
22841.0	482.873848	74.337194	1	35.00
...
26501.0	500.688890	105.091553	1	49.48
19731.0	536.983349	167.747188	1	78.98
29256.0	3384.043706	5082.667569	1	2393.06
20233.0	525.898160	148.610670	1	69.97
23243.0	482.849241	74.294715	1	34.98

Figure 5.2.1 Prediction Table of CLV

Looking at figure 5.2, we see that the predicted CLV has a slight difference compared to the correctly calculated CLV. Therefore, we need to check their accuracy with R2 (coefficient of determination) regression score function


```
[315] 1    r2_score(predictions,y_test)
0.8978976930021869
```

Figure 5.2.2 R2_Score

89.8% is a really greatable number. This shows that the training process is quite smooth for making such accurate predictions.

Conclusion and Future Works

The study uses machine learning methods like KMeans to identify lead segments and give us an overview of a company's current customer retention and customer lifecycle. The article also offers a suitable marketing plan based on the above two facts to find more loyal customers. However, the dissertation still has some shortcomings. First, the analysis still doesn't identify customer behavior because it only provides retention rates, not specific behaviors. Moreover, due to time constraints based on technical analysis and customer factors, the goal of supply chain optimization has not fully solved the research objectives. Anyway, the research paper can be done with full technical background. In the future, the team will further analyze customer behavior through behavioral queues and optimize the supply chain based on this. Combine CRM methods to find solutions that improve the customer experience.

References

- [1] Hồ Trung Thành, Nguyễn Đăng Sơn (2021) An interdisciplinary research between analyzing customer segmentation in marketing and machine learning method
- [2] Rendra Gustriansyah, Nazori Suhandi, Fery Antony (2020) Clustering optimization in RFM analysis based on k-means
- [3] Daqing Chen, Sai Laing Sain, Kun Guo (2012) Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining