

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW



FINAL PROJECT REPORT

INTERDISCIPLINARY RESEARCH METHOD COURSE
TOPIC: CUSTOMER LIFETIME VALUE DETERMINATION BY
USING ANALYZE RFM MODEL WITH SILHOUETTE SCORE
AND ELBOW IN K-MEANS CLUSTERING

Lecturer:

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

Group 3:

- 1. Hoang Tran The Phuc**
- 2. Luc Minh Phu**
- 3. Nguyen Thi Thu Thao**
- 4. Nguyen Hoang Duy Thong**
- 5. Le Tuan Nguyen**

Ho Chi Minh City, 12/2022

Members of Group 3

<i>No.</i>	Full name	Student ID	Point / 10 (Individual Contribution)	Signature
<i>1</i>	Hoàng Trần Thế Phúc	K214162150	10	
<i>2</i>	Lục Minh Phú	K214161342	10	
<i>3</i>	Nguyễn Thị Thu Thảo	K214162154	10	
<i>4</i>	Nguyễn Hoàng Duy Thông	K214162156	10	
<i>5</i>	Lê Tuấn Nguyên	K214160993	10	

Acknowledgments

First and foremost, we would like to express our profound gratitude to all of the faculty members at the Faculty of Management Information Systems who helped us complete this project by providing the necessary knowledge, expertise, and information.

Additionally, we would like to express our sincere gratitude to PhD. Ho Trung Thanh and M.A. Nguyen Van Ho, our supervisors, for giving me the chance to do research and for their helpful guidance during this research.

Finally, despite all efforts to execute the implementation process correctly, faults will inevitably happen. As a result, we expect that teachers' and readers' suggestions will be understood and welcomed.

Ho Chi Minh City, 12/2021

Group 3

Commitment

We swore to say that the findings in the thesis were original work of ours, not a copy. The information offered throughout the entire thesis is either personal or has been assembled from a variety of sources. Every reference has been properly sourced and cited.

We will accept full responsibility for my comments and implement all necessary disciplinary measures.

Ho Chi Minh City, 12/2021

Group 3

Table of Content

Members of Group 3	1
Acknowledgments	2
Commitment	3
Table of Content	4
List of Tables	6
List of Figures	7
List of Acronyms	8
GANTT CHART	9
ABSTRACT	11
Project Overview	11
Business problems	12
Objectives	13
Objects and scopes	14
Research method	14
Process	15
Tools and Programming language	15
Structure of project	15
Chapter 1 THEORETICAL BASIS	16
1.1 Consumer behavior	17
1.1.1 Definition	17
1.1.2 Need of Studying Consumer Behavior	18
1.1.3 Evaluating the effectiveness of consumers	19
1.2 RFM model	19
1.2.1 Definition	20
1.2.2 Benefits of RFM	20
1.2.3 How to calculate RFM	21
1.3 Machine learning	22
1.3.1 What is Machine Learning?	22
1.3.1.1 Definition	22
	4

1.3.1.2 Classification	22
1.3.1.3 Machine Learning workflow	23
1.3.2 The elbow method	27
1.3.3 Silhouette Score	28
1.4. Kmean clustering	31
1.4.1. Definition	31
1.4.2. K-means process	32
1.5. Customer Lifetime Value (CLV):	34
1.5.1. Definition	34
1.5.2. Calculate customer lifetime value	34
Chapter 2 DATA PREPARATION	36
2.1 Data understanding	37
2.1.1 The table relationship	37
2.1.2 Data information tables	38
2.2 Data collection	44
2.3 Exploratory Data Analysis	46
2.3.1 Removing null and NA data	47
2.3.2 Data visualization	48
2.3.3 Calculating RFM	51
2.3.4 Removing Outliers	52
2.3.5 Transform data	54
2.3.5.1 Scaling data	54
2.3.5.2 Normalization Data	55
Chapter 3 CUSTOMER SEGMENTATION WITH MACHINE LEARNING METHOD 57	
3.1 RFM with Machine Learning methods	58
3.1.1 Elbow method	58
3.1.2 Silhouette method	59
3.2 Customer segmentation	60
3.2.1 Analysis of loyal customer groups (cluster 1)	60
3.2.2. Analysis of potential loyalist customer groups (cluster 2)	61
3.2.3. Analysis of new customer groups (cluster 0)	63

References 64

List of Tables

Table 1.1

Table 1.2

List of Figures

Figure 1.1

Figure 1.2

List of Acronyms

DB	Digital Business
MIS	Management Information Systems
FDA	Fundamentals of Data Analytics
CLV	Customer Lifetime value

GANTT CHART

ABSTRACT

The explosion of digital technologies has led to a series of changes in the way businesses and companies operate. Especially in which the retail market becomes a place of fierce competition among businesses thanks to the development of leading e-commerce sites such as Lazada, Shopee, Tiki, Sendo,... According to statistics, Vietnam has up to 58.2% of internet users shop online weekly (*e.vnexpress* - 27/11/2022). This figure shows that the level of shopping for retail goods is very high. Thus, for businesses to be successful, it is imperative that they come up with the right marketing strategies, in accordance with the segments and needs of each customer group. Therefore, the personalization of strategies for each customer is gradually getting more attention.

To achieve these above goals, the research will be based on analyzing customer behavior through general personality characteristics to segment customers into specific groups. Through qualitative and quantitative methods, the data will be analyzed on RFM (Recency, Frequency, Money) model and continue to cluster through the K-means Clustering unsupervised learning algorithm based on the determine the number of k clusters in an optimized, time-saving way with the Silhouette and Elbow method. From there, it is applied to analyze data sets of businesses to predict CLV and propose recommendations to improve customer loyalty, retain target customer groups, and help bring maximum profit depending on financial capacity core of businesses.

Project Overview

Business problems

The digital transformation of industry in recent years has included how it is providing new managerial issues. And this has brought many challenges as well as great opportunities to businesses (Hess et al, 2016)¹. It has also brought about a revolution in the way companies do business by using big data analytics in business or applying artificial intelligence algorithms and networking platforms social media (Sunil Erevelles et al, 2016)². Accordingly, some studies such as (Atis Verdenhofs, Tatjana Tambovceva, 2019)³ have also shown that the use of big data in customer data analysis has helped businesses come up with more effective and cost-effective marketing strategies. Big data analytics, like most other fields, will be a significant change for business units to devise short- and long-term strategic strategies to attract new, retain consumers. maintain existing consumers and compete with competitors (Truong Thi Hoai Linh, 2019)⁴. Thus, it can be seen that the use of data in customer segmentation analysis has brought a lot of efficiency, most notably for online sales data.

Besides that, to be able to apply big data effectively in customer marketing strategy, businesses will often use RFM model, a well-known tool for identifying customer segments. company's best by calculating and analyzing their spending habits. The RFM analysis will evaluate customer importance by scoring them on three measures, such as how recently they bought (Recency), how often they bought (Frequency), and how much they spent (Monetary) by (Jo-Ting et al, 2018)⁵. Thanks to the application of RFM model in analysis, businesses can more easily reach target customer groups. From there, it is possible to come up with specific strategies with high efficiency and low implementation costs according to (Saritha M et al, 2022)⁶. And moreover, the application of RFM model in online big data brings great efficiency to online retail businesses when the number of transactions is very large, making the analysis more accurate. and subsequent marketing campaigns also became more effective than by (Aylanur Cuce, Eda Tiryaki, 2022)⁷.

However, although K-Means is the most popular algorithm among enterprises for classifying clusters with similarities, it still has some disadvantages. Since the clusters will be randomly selected on the first run, the results may be different for each run. Moreover, determining the correct number of clusters is also a big problem for the RFM model. Accordingly, the study of (Basim Amer Jaafar.el, 2020)⁸ combined with the Elbow algorithm to be able to choose the most suitable number of clusters for the data set. However, in the study of (Thanh HT, Son N D., 2021)⁹ optimized the number of clusters K for RFM method by combining with Elbow algorithm. Then, to ensure that the number of clusters analyzed from Elbow is the most optimal, the study used the silhouette algorithm to test the group with the best score. These studies have shown the effectiveness of clustering approaches in Data Science and also perform clustering results in RFM analysis and provide different customer behaviors in specific groups.

To summary, it can be seen that customer retention for businesses is paramount. Therefore, this study was conducted under the name "CUSTOMER LIFETIME VALUE DETERMINATION BY USING ANALYZE RFM MODEL WITH SILHOUETTE SCORE AND ELBOW IN K-MEANS CLUSTERING" in order to be able to identify customer segmentation analysis goals based on behavior. purchase and thereby predict the long-term value of customers to the business based on business, marketing and information technology knowledge. The end result is to help administrators get a detailed overview of their customers. This will help managers easily decide to implement appropriate marketing strategies for each customer group as well as assess whether current customer care policies are still appropriate to retain customers.

Objectives

From project

As a result of the study, an interdisciplinary research model for assessing CLV based on the RFM approach in conjunction with machine learning algorithms like Silhouette and

Elbow to analyze customer segments will be available. When applied to real data, customer clustering will be more precise and efficient thanks to the use of algorithmic and modeling techniques.

From result

Implement and apply RFM and CLV analytical models in segmenting customer groups based on shopping behavior. From there, recommendations and solutions are given to help businesses have appropriate strategies for each customer group.

Objects and scopes

Objects

Factors that have an impact on how consumer clustering is implemented at AdventureWork businesses in the retail industry.

Scopes

Time scope: Retail market research in Adventure Work from July 2017 to July 2020.

Space scope: Use your AdventureWork business sales transaction history.

Research method

To achieve the set objectives, in this study, two methods of quantitative and qualitative research will be applied:

In the quantitative method, the group will first process the outliers and visualize the data variables of the group of interest in order to find and see the relationship between the data variables in the data set. Then research to find and analyze CLV by using RFM model combining customer clustering by K-Means, Silhouettes and elbow methods to group customers based on buying behavior characteristics.

Next, after the models are made and the resulting data are produced, the research uses qualitative methods including business and underlying theories or observable evidence to analyze the data. analyze, interpret, and describe data. From there, solving the stated

target questions by presenting methods, giving recommendations and empirical solutions to help businesses come up with appropriate strategies for each target customer group.

Process

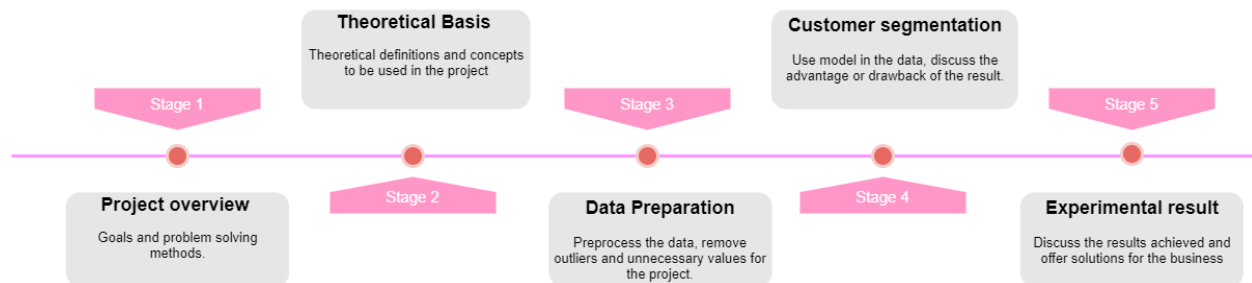


Figure 0.1: Process of study

Tools and Programming language

Tools: Trello, Google Collab, Google Sheets, Visual Studio Code.

Programming language: Python

Structure of project

CHAPTER 1: THEORETICAL BASIS

An overview of the research situation at home and abroad related to customer behavior analysis methods based on models. Then learn about the limitations of previous studies and suggest a better solution model.

CHAPTER 2: DATA PREPARATION

Pre-process the outliers, perform data visualization to find relevant variables, and then prepare the input data for application to machine learning methods including Elbow and Silhouette.

CHAPTER 3: CUSTOMER SEGMENTATION WITH MACHINE LEARNING METHOD

Presenting procedures for applying RFM model on AdventureWork dataset. From there, calculate the number of clusters using Elbow and Silhouette. And start comparing the results and assigning labels to the data variables.

CHAPTER 4: DATA VISUALIZATION

Perform result visualization, analysis, and CLV prediction. Extract data insights from RFM model implementation. After that, discuss and propose solution strategies for businesses.

Chapter 1 THEORETICAL BASIS

Chapter overview: Generalize and search for concepts and theoretical bases related to customer segmentation analysis; algorithms and models are applied in the RFM model.

1.1 Consumer behavior

1.1.1 Definition

According to Wayne D. Hoyer, Deborah J. MacInnis (2008)¹⁰, consumer behavior is understood as a series of decisions about buying what, why, how, ... that each individual or group of consumers who decide over time whether to use a product, service, or idea.

Consumer shopping behavior is the behavior that consumers display in finding, purchasing, using, and evaluating products and services that they expect will satisfy their individual needs (Peter D. .Bennett, 1995)¹¹.



1

Figure 1.1: Factors when making a purchase decision

(Source: *Consumer Behavior* by Wayne D. Hoyer, Deborah J. MacInnis, Rik Pieters, page 4)

Factors that drive customer behavior include advertising, marketing, images, sounds, habits, trends, and needs. According to Philip Kotler (2001)¹², marketers must study consumer behavior with the aim of identifying factors that lead to customer behavior to build marketing strategies to motivate consumers to purchase products and services.

1.1.2 Need of Studying Consumer Behavior

Retailers need to know who their customers are and how they make decisions. They cannot rely only on their emotions to identify customers but need to do research to better understand how consumers make decisions. Managers cannot effectively manage, and retail businesses cannot be successful without understanding how consumers make decisions and actions regarding the consumption of retail products.

¹ *Consumer Behavior* by Wayne D. Hoyer, Deborah J. MacInnis, Rik Pieters, page 4

Therefore, understanding customer behavior plays a very important role in determining the success of the business. We need to study consumer behavior to be aware of:

- How different advertising strategies work;
- The complexity of the purchasing decision-making process;
- The needs as well as the purchasing motivation of individuals;
- How demographics affect retail;
- Different market segments based on purchasing behavior;
- Be aware of risks in the retail market;
- Retailers can change the business situation of the business based on the analysis of buying behavior.

Customers are the reason the retail market exists. Customers who go to brick-and-mortar stores, or visit a store's online sites to make purchases, are the ones driving the business. Any successful retailer understands its customers, including how they think as well as their demographics.

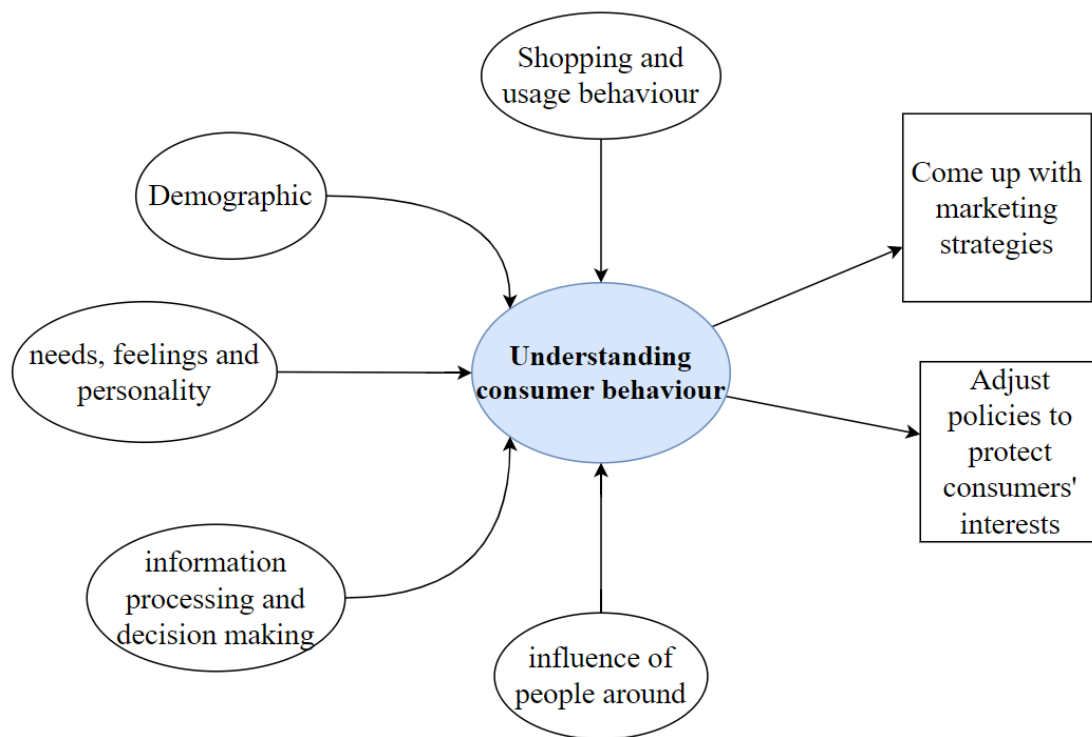


Figure 1.2: Benefits of understanding consumer behaviour

1.1.3 Evaluating the effectiveness of consumers

- Customers' reactions to promotional activities such as product discounts, service offers.
- Number of emails opened by consumers.
- Take advantage of the data gathered through website analytics.
- Pages per visit represent the average number of each user visiting your website.
- Average length of visit: The longer you stay on the website, the more likely it is to convert.
- Similarity analysis is a method that helps you find the connection between your product or service and the buyer.

1.2 RFM model

1.2.1 Definition

RFM model is a behavior-based model that is used to analyze customer behavior and then make predictions on a customer-related data set collected and stored at an enterprise. This model was first introduced widely by Arthur M. Hughes through work in 1996⁽¹³⁾ - initially called RFM analysis, later called RFM model. In which, 3 attributes of the RFM model are separated into each attribute independently, each attribute is classified into 5 groups so that each group has an equal number of observations. The results of this model are then analyzed by decision makers to determine the best customer group to implement corresponding marketing strategies.

In fact, 3 attributes in the RFM model belong to behavioral variables and can be used as categorical variables by observing customer behaviors towards brands, products, or even loyalty and based on that to conduct customer segmentation into groups of individuals with similar buying behavior.

1.2.2 Benefits of RFM

Businesses can get many benefits from adopting the RFM model, such as increased response rates, reduced ordering costs, and increased profits. In the application of RFM model, each customer needs to be distinguished by a unique identifier (account number, identifier...), and sales information needs to be stored with a unique information contained in each transaction record such as the order number. RFM analysis helps to identify important and valuable customers with customers able to be segmented into different segments.

Some of the benefits that RFM brings to businesses:

- Capture customers: You will know who your loyal customers are, who are not satisfied with the service of the business as well as which old customers have left. As a result, businesses will have plans to reduce the rate of customers leaving.
- Increase sales: Understand their needs and adapt them to the next business strategy. Moreover, it also helps businesses improve their marketing.
- Help keep old customers: With RFM, you can categorize customers into several categories. The number of categories will depend on the nature of the business, along with the analyst's instincts. Then you will know which customers are potential, which guests stay and who leave. Then, businesses will offer appropriate actions for each customer. Depending on the strategy, different priorities will be given. For example: Give vouchers to potential customers, send promotional emails to customers who are at risk of leaving, etc.

1.2.3 How to calculate RFM

RFM model proposed by Hughes in 1996⁽¹³⁾ is a model of customer segmentation with big data according to 03 variables (attributes) in the customer data set, including: customer's consumption time period, purchase frequency item and total purchase amount with detailed definition:

- R- Recency: Usually determined by calculating the time difference from last purchase to analysis time, the smaller this period, the higher the value of R. A customer with a high R means they have a high probability of making a repeat purchase. In Hughes' study, customers are classified into 5 different groups and the highest group is 5, followed by 4, ... Finally, the value of R assigned to each customer in the data set is expressed equals a number between 5 and 1.
- F-Frequency: Purchase frequency, calculated as the total number of purchases made by the customer during the analysis period. The larger the number of purchases, the greater the value of F. The value of the variable F is usually calculated as follows: The data set is sorted by descending frequency and also divided into 5 different groups. A customer with a high F means that they have a high demand for the product and have a high probability of buying the product again and again.
- M-Monetary: Sales of goods, calculated by the total amount of money that customers have spent to buy goods during the analysis period. The larger the total expenditure, the greater the M-value. There are many ways to determine M, which can be the total amount a customer has spent during this time period, or as an average total amount per transaction, or the total amount from the first purchase until analysis time. According to Marcus in a study published in 1998⁽¹⁴⁾, it was recommended to use the average amount per purchase rather than the cumulative total.

Finally, all customers are sorted by the RFM consolidation index, assigned values from 555, 554, 553... up to 111. The best customer segment is the one that corresponds to the value 555, at least 111. Based on the assigned RFM score, customers can be grouped to form different segments, each of which will be assigned a unique approach to deliver the best experience for each customer in each segment.

The larger the R and F value pair, the higher the probability that the respective client will initiate a new trade. In addition, the larger the M-value, the higher the probability that the respective customer will repeat the purchase behavior for the business.

1.3 Machine learning

1.3.1 Definition

Machine learning (ML) is a branch of artificial intelligence. By using computing, systems that can learn from data in a manner of being trained are designed. The systems might learn and improve with experience, eventually refining a model that can be used to predict the outcomes of questions based on the previous learning.

1.3.2 Classification

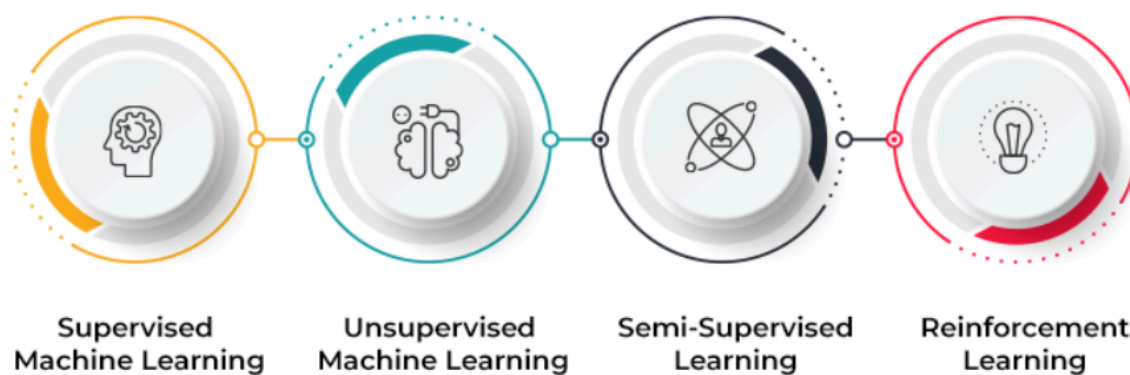


Figure 1.3: Types of Machine Learning Algorithms²

According to Vijay Kanade (2022), there are many ways to classify machine learning, usually, machine learning will be divided into two main categories:

- Supervised learning: use labeled datasets to make predictions. This learning technique is beneficial when you know the kind of result or outcome you intend to have.
- Unsupervised learning: with unlabeled data, we categorize the data or express its type, form, or structure. In this way, the result type is unknown.

In addition, machine learning can also be divided into the following categories:

² Vijay Kanade, Top 10 Machine Learning Algorithms in 2022, [Top 10 Machine Learning Algorithms \(spiceworks.com\)](https://spiceworks.com), Accessed 4/12/2022

- Semi-supervised learning (SSL): combining the above two, where labeled and unlabeled data are used with the purpose of categorizing unlabeled data based on the information derived from labeled data
- Reinforce learning: Using the result or outcome as a guideline to determine the next course of action. In other words, these algorithms learn from past results, receive feedback after each step, and then decide whether to proceed with the next step or not. During the process, the system learns whether it made the correct, incorrect, or neutral decision. Because automated systems are designed to make decisions with minimal human intervention, they can use reinforcement learning.

1.3.3 Machine Learning workflow

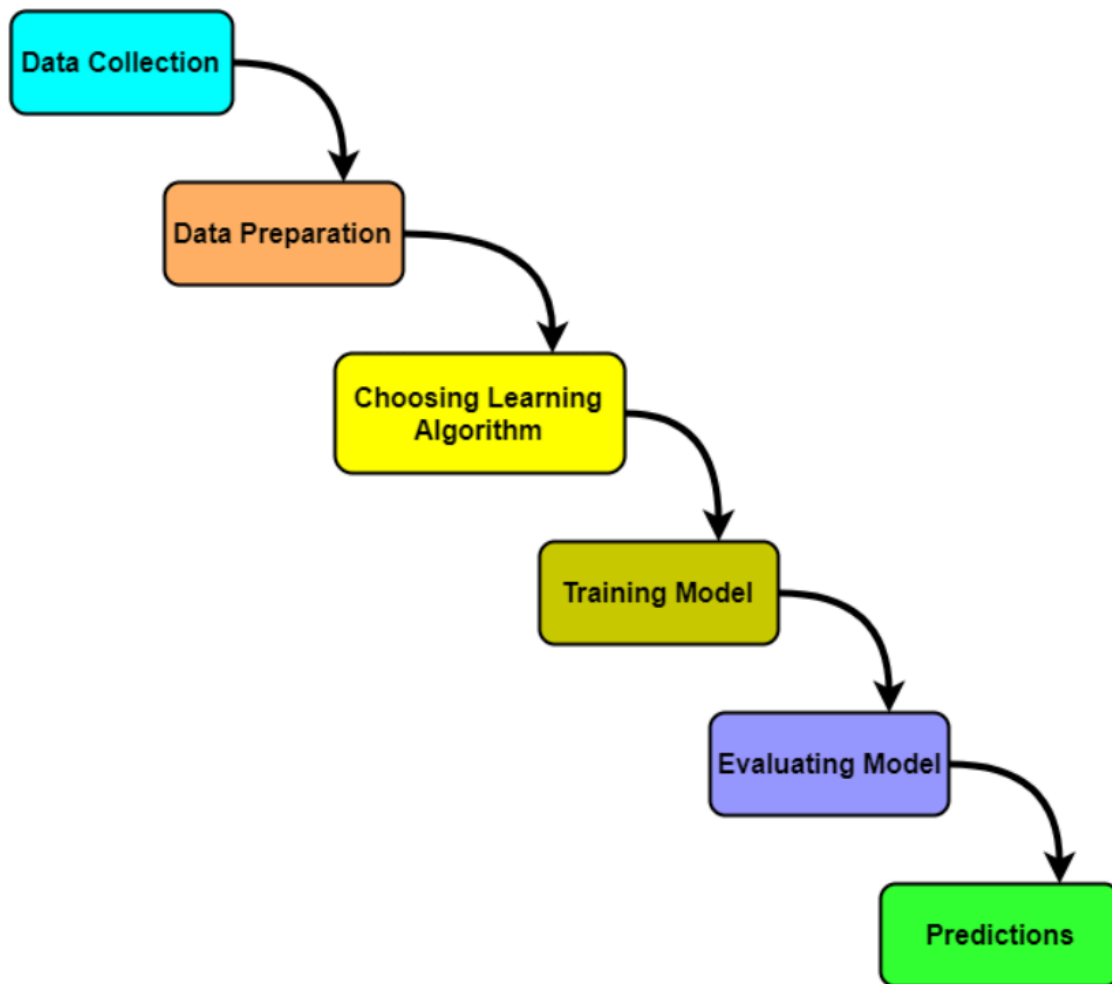


Figure 1.4: Machine learning workflow³

³ GateVidyalay, Machine Learning Workflow | Process Steps, [Machine Learning Workflow | Process Steps | GateVidyalay](#), Accessed 4/12/2022

Data collection: We need to collect data from different sources such as files, databases, sensors, etc. The type of data collected depends upon the type of desired project unstructured data and structured data. The quality and quantity of gathered data plays an important role in affecting the accuracy of the desired system. Nevertheless, this data cannot be used directly for performing the analysis process because there may be a large amount of missing data, extremely large values, unorganized text data or noisy data. To solve this problem, the next stage is Data Preparation.

Data preparation: this step, which is one of the most important stages, can be called Data Pre-processing. Starting from the raw data collected in the real world, we do some methods:

- Ignoring the missing values
- Removing instances having missing values from the dataset
- Estimating the missing values of instances using mean, median or mode
- Removing duplicate instances from the dataset
- Normalizing the data in the dataset
- Detecting outliers from the dataset

The cleaner the pre - processed dataset is, the more accurate some built machine learning models are. Moreover, this stage is the most time - consuming in ML workflow, so that there is an 80/20 rule. According to the 80/20 rule, every data scientist should spend 80% of their time on data preparation and 20% of their time actually performing the analysis.

Choosing Learning Algorithm: Based on the type of problem that needs to be solved and the type of data we have, Data Scientists research the best performing learning algorithm. In the situation that the target variable is labeled and categorical (i.e the output could be classified into classes), Classification algorithms are used. If the target variable is labeled and continuous, we use Regression algorithms. On the other hand, when the problem is to create clusters and the data is unlabeled, Clustering algorithms are used. Additionally, the following chart provides the overview of learning algorithms:

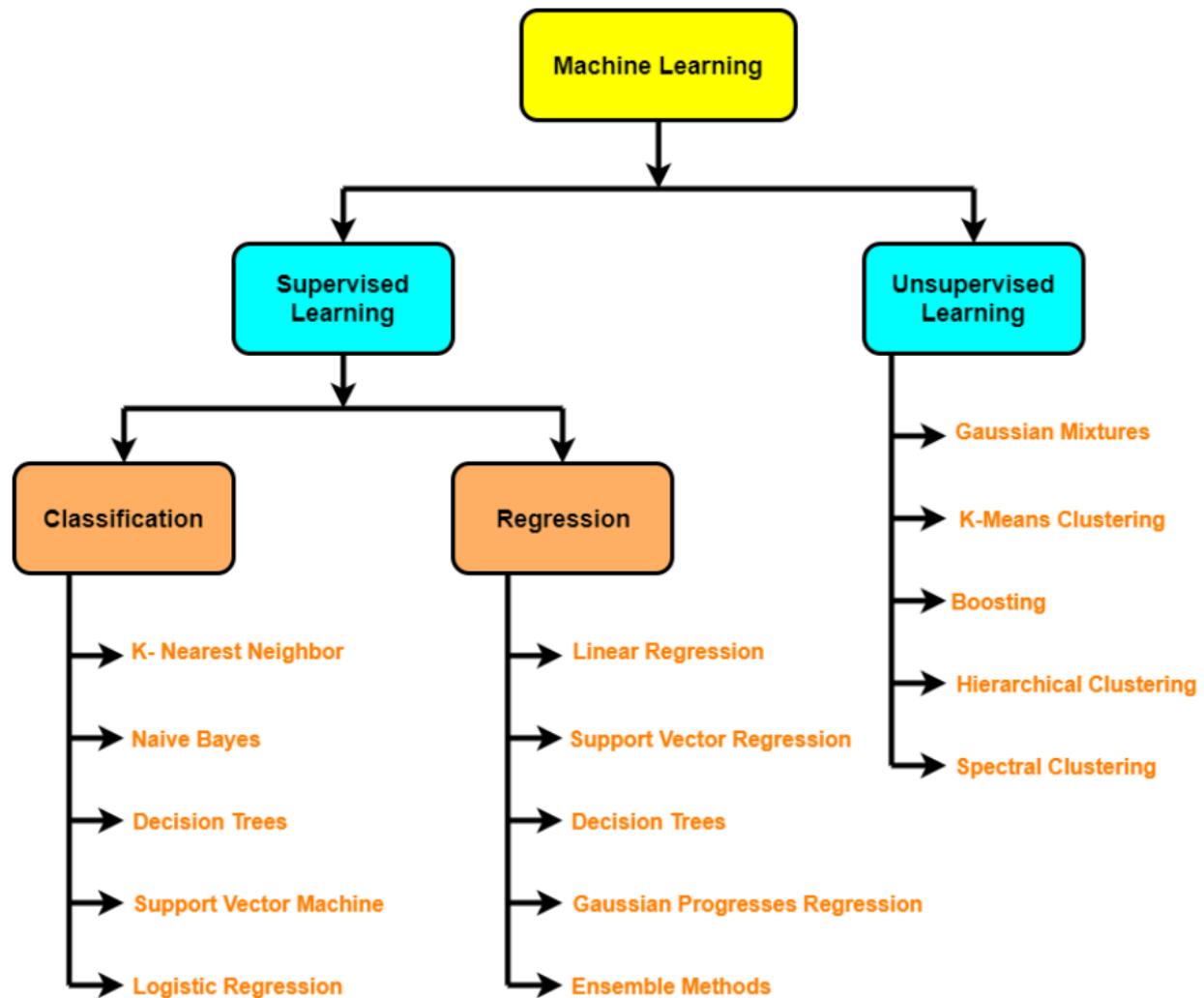


Figure 1.5: The overview of learning algorithms⁴

Training Model: For training a model, we split into 3 sections of data which are “Training data set”, “Validation data set” and “Testing data set” at the beginning.

You train the classifier using “training data set”, tune the parameters using “Validation set” and then test the performance of your classifier on unseen “Testing data set”. We only use the training and/or validation sets in training the classifier, and while testing the classifier, the test set is only available.

⁴ GateVidyalay, Machine Learning Workflow | Process Steps, [Machine Learning Workflow | Process Steps | GateVidyalay](#), Accessed 4/12/2022

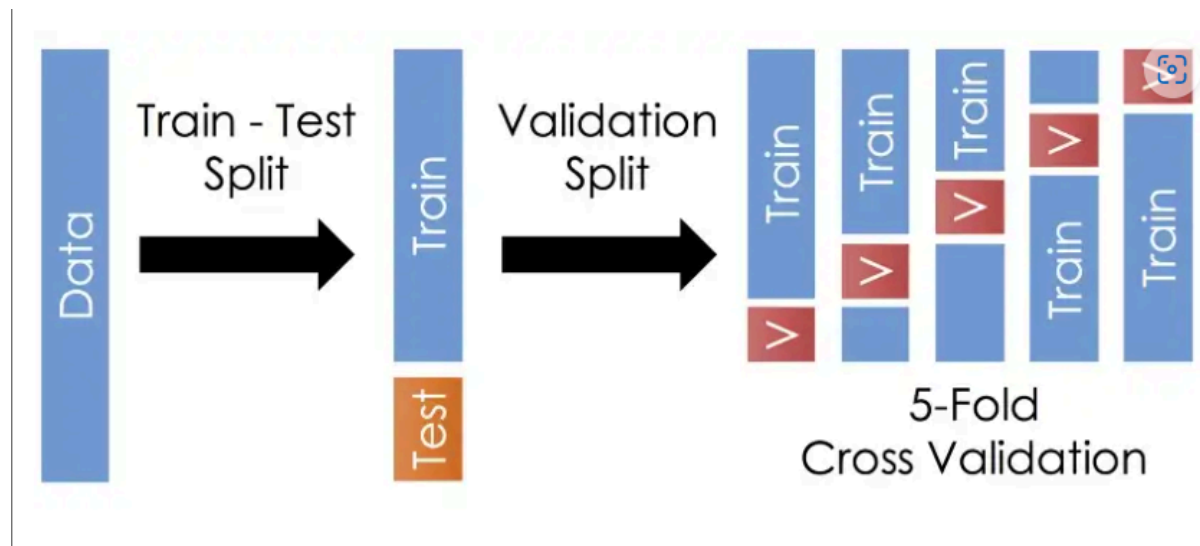


Figure 1.6: Training and testing the model on data⁵

After data set division, implementation of training data set is done to build up a model and validating the built model with a test (or validation test). Usually, in each iteration, a data set is divided into a training set, a validation set (some people use “test set” instead). or divided into 3 data sets: training set, validation set and test set in each iteration. Once this is done, this tells us how well our model is trained.

Evaluating Model: By the way, the Data Scientists are able to find the best model that represents our data and how well the chosen model will work in the future. If the model does not perform well, the model is re-built using different hyper parameters. From that, the accuracy may be improved by tuning the hyper - parameters of the model. In other words, we try to increase the number of “True positives” and “True negatives” after looking at the confusion matrix which has 4 parameters: “True positives”, “true Negatives”, “False Positives”, “False Negatives”.

Predictions: in the final stage, the built system is used to do something useful in the real world and can help to solve some troubles in different fields

⁵ Ayush Pant, Workflow of a Machine Learning project, [Workflow of a Machine Learning project | by Ayush Pant | Towards Data Science](#), Accessed 4/12/2022

1.3.4 The elbow method

In the book of Pratap Dangeti (2017), he supposed that: In K-means clustering, the elbow method is used to determine the optimal number of clusters. This method plots the value of the cost function generated by various k values. Additionally, the Elbow method measures distortion - the average squared distance between the centroid and the rest of the points (usually Euclidean distance) in each k cluster. As you may be aware, as k increases, the average distortion decreases, each cluster has fewer constituent instances, and the instances are closer to their respective centroids. Nevertheless, as k increases, the improvements in average distortion diminish. The value of k in which the improvement in distortion decreases the most is known as the elbow, and it is at this value that we should stop dividing the data into further clusters.

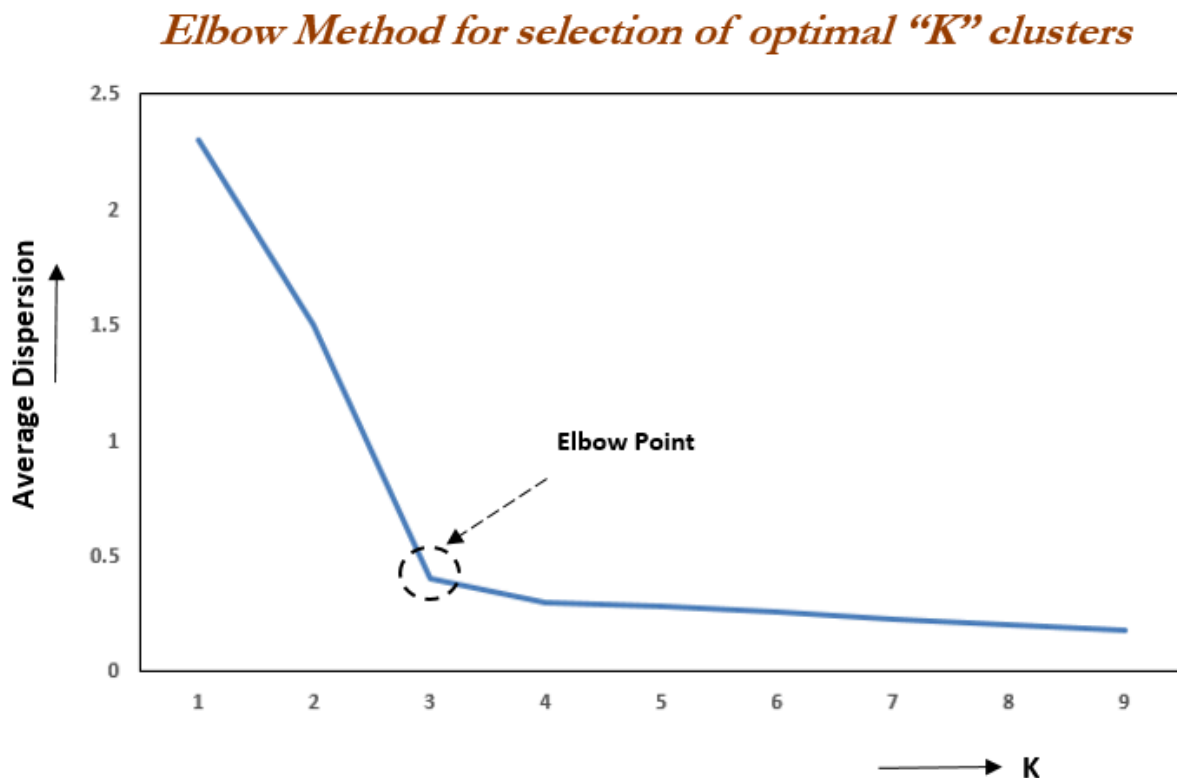


Figure 1.7: The chart of elbow method for selection⁶

⁶ Statistics for Machine Learning (Pratap Dangeti,2017)

1.3.5 Silhouette Score

Determining k data objects from the dataset in K-means Clustering can be commonly solved by Silhouette Score. Silhouette Coefficient or Silhouette score determines whether there are large gaps between each sample and all other samples within the same cluster or across different clusters, through providing a concise graphical representation of the classification level of each object. From that, this technique gives analysts some information about the validation of consistency within clusters of data [].

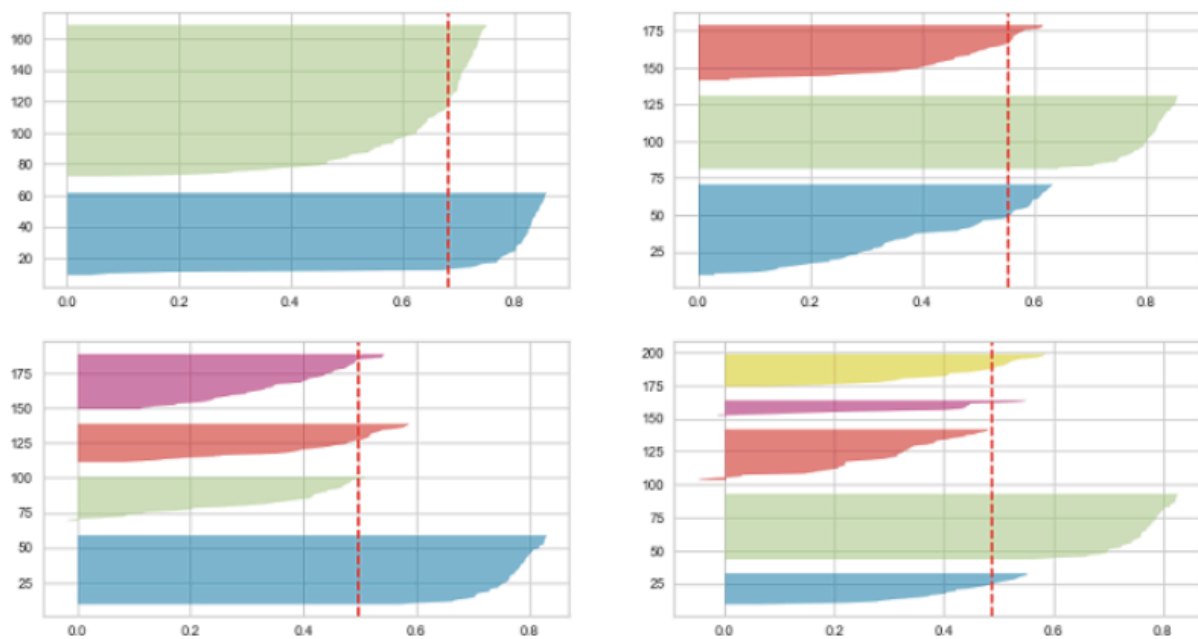


Figure 1.5: Some graphical illustrations of Silhouette Score

(Source: vitalflux.com, 9/11/2022)

In the situation that the proximities are on a ratio scale (as in the case of Euclidean distances) and someone is looking for compact and clearly separated clusters, the silhouettes constructed below are useful.

Let us first define the numbers $s(i)$ in the case of dissimilarities. Take any object i in the data set and denote by A the cluster to which it has been assigned. (For a concrete

illustration, see Figure 2.3.4.2). When cluster A contains other objects apart from i , then we can compute $a(i)$ = average dissimilarity of i to all other objects of A.

In Figure 2.3.4.2, this is the average length of all lines within A. Let us now consider any cluster C which is different from A, and compute $d(i, C)$ = average dissimilarity of i to all objects of C.

In Figure 2.3.4.2, this is the average length of all lines going from i to C. After computing $d(i, C)$ for all clusters $C \neq A$, we select the smallest of those numbers and denote it by

$b(i) = \text{minimum } d(i, C) (C \neq A)$

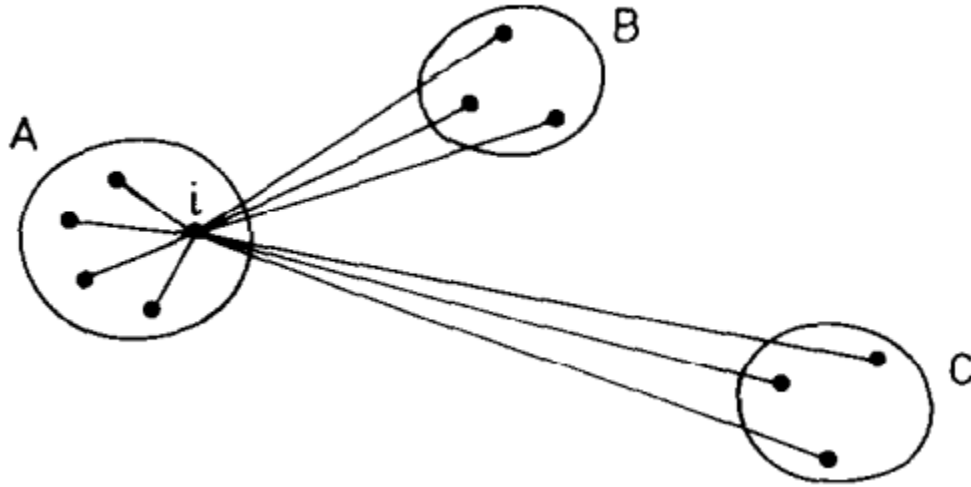


Figure 1.6: An illustration of the elements involved in the computation of $s(i)$, where the object i belongs to cluster A.

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i). \end{cases}$$

It is even possible to write this in one formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

When cluster A contains only a single object it is unclear how $u(i)$ should be defined, and then we simply set $s(i)$ equal to zero. This choice is of course arbitrary, but a value of zero appears to be most neutral. Indeed, from the above definition we easily see that $-1 \leq s(i) \leq 1$ for each object i

1.4. Kmean clustering

1.4.1. Definition

In data mining, clustering is a common unsupervised file learning approach (Data Mining-DM). It is used to locate classes or groups of data files that have the greatest number of similarities in the same cluster, whilst distinct items are located in various clusters (Đỗ, 2022, 15). Clustering is a DM approach used to split data into similar groups without knowing the definition of the groups. There are two types of cluster clustering techniques: hard clusters and open clusters. Each point in an extended or complex cluster can belong to two or more clusters

K-means is the most often used tough clustering approach for dividing data into groups. Kmeans is an unsupervised ML approach. So, what exactly is unsupervised learning? The simple explanation is that they employ ML algorithms to clustertuent and explore datasets without the need for labeling data. All the data points with similar characteristics will be grouped together.

There are two types of clustering: exclusive clustering and overlapping clustering. Exclusive clustering is a type of grouping in which a data point can only reside in one cluster. The opposite to overlap clustering, they allow data points to be in multiple clusters. Kmeans is an example of exclusive clustering, and the items in each cluster are very homogenous and not comparable to those in other clusters. Each data point exclusively belongs to one cluster.

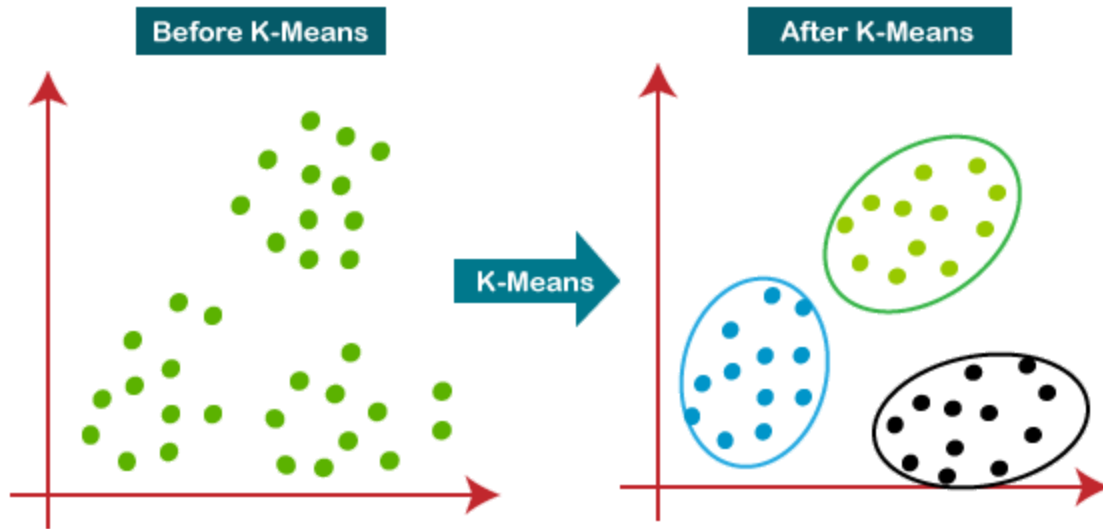


Figure 1.7: How K-Means Algorithm Works⁷

The diagram above shows how the K-Means algorithm works. It takes an unlabeled dataset as input and separates it into k-groups (K is the number of clusters) to identify the best clusters. Methods such as Elbow's curve or Silhouette analysis can be used to determine the K value; these methods assist us in determining the optimal number of clusters for the model.

⁷ JavaTpoint, K-Means Clustering Algorithm, <https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning>, Accessed 5/12/2022.

1.4.2. K-means process

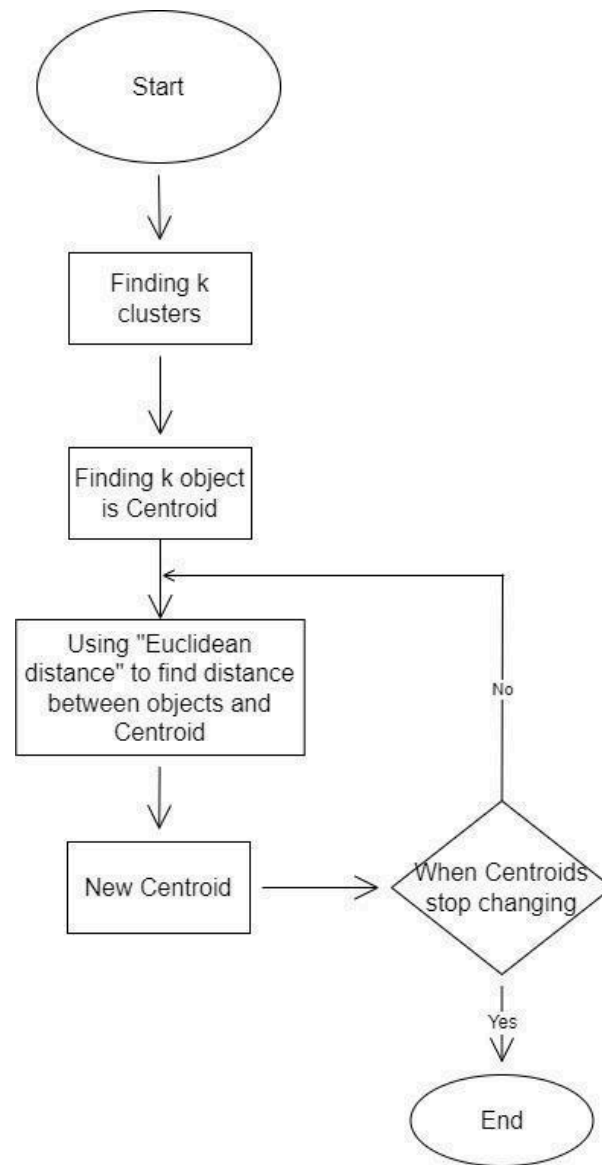


Figure 1.8: K-Means process step by step⁸

First, we determine the number of clusters to be used for clustering. Then, we choose K centroid points at random and use "Euclidean distance" to compute the distance between the object and it. We replaced the original centroid with the nearest distance data point.

⁸ BigDataUni, K-means Clustering và mô hình RFM, <https://bigdatauni.com/tin-tuc/k-means-clustering-va-mo-hinh-rfm-p-1.html>, Accessed 5/12/2022

The method will be repeated until there is no significant difference in centroid location. Final results will be K centroid with optimized distance to all object of its cluster.

1.5. Customer Lifetime Value (CLV):

1.5.1. Definition

Customer lifetime value is the entire value of a customer to a business over the course of their relationship. It's a significant measure since keeping existing customers is less expensive than acquiring new ones, thus boosting the value of your existing customers is a fantastic method to drive growth. Understanding the CLV enables organizations to establish strategies for acquiring new consumers and retaining existing ones while preserving profit margins. While intuitive, such strategies presume that a firm can accurately predict the future profitability of customers (Edward C. Malthouse & Robert C. Blattberg, 2005). In addition, long-term relationships are emphasized, thus leading to an increase in CLV.

Companies may use CLV to separate successful consumers from unprofitable clients, which leads to more effective decision making. Since the mid-1980s, most scholars have recognized the notion of "Customer Lifetime Value". They feel that clients who stay with the firm for a longer period of time will profit more. have worth, and the organization must be able to quantify and assess that worth. In reality, assessing the financial return of the client and the business relationship.

1.5.2. Calculate customer lifetime value

Before we go into the CLV formula, we need a few pieces of data to hand⁹:

⁹ Qualtrics. (n.d.). How to Calculate Customer Lifetime Value (CLV). Qualtrics. Retrieved December 5, 2022, from <https://www.qualtrics.com/uk/experience-management/customer/calculate-clv/?rid=ip&prevsite=en&newsite=uk&geo=GB&geomatch=uk>

- **Average purchase value** — the value of all customer purchases over a particular timeframe (a year is usually easiest), divided by the number of purchases in that period
- **Average purchase frequency** — divide the number of purchases in that same time period by the number of individual customers who made a transaction over the same period
- **Customer value** — the average purchase frequency multiplied by the average purchase value
- **Average customer lifespan** — the average length of time a customer continues buying from you.

CLV = customer value X average customer lifespan

Traditional CLV formula: $GML * R / (1 + D - R) = CLV$

GML – gross margin per customer lifespan

R – retention rate

D – discount rate

The resultant CLV is a monetary number (depending on the currency you work in) that illustrates how much the average client may reasonably anticipate spending with you during their lifetime.

In actuality, however, CLV will vary depending on the consumer niche. One or two segments are likely to have a substantially greater CLV than others, whether because they spend more each transaction or because they remain with you longer.

Understanding your CLV by segment is beneficial since it allows you to identify what is causing a higher CLV (i.e. making those higher-value customers more valuable to you) and find ways to make fewer valuable clients more valuable (i.e., identify actions that will increase CLV with segments that are currently spending less over their lifetime)

The formula for calculating CLV at the individual level is the same, but a little easier to compute - simply multiply the amount the client spends each year by the number of years (hence, no average value for frequency of purchases). Purchase or claim multiplied by the number of years you can expect them to remain with you. This method is appropriate for circumstances in which the data are projected to stay reasonably consistent year after year.

Customer revenue per year X Duration of the relationship in years – Total costs of acquiring and serving the customer = CLV

Once you know your CLV – whether it's an average or broken down by segment – there are numerous ways to use it to not only track the return on your investment in customer experience, but also to identify new opportunities to design experiences that work to the bottom, such as: optimizing marketing spend, reducing churn and driving loyalty, designing experiences for new business development, identify costly experience gaps, and so on.

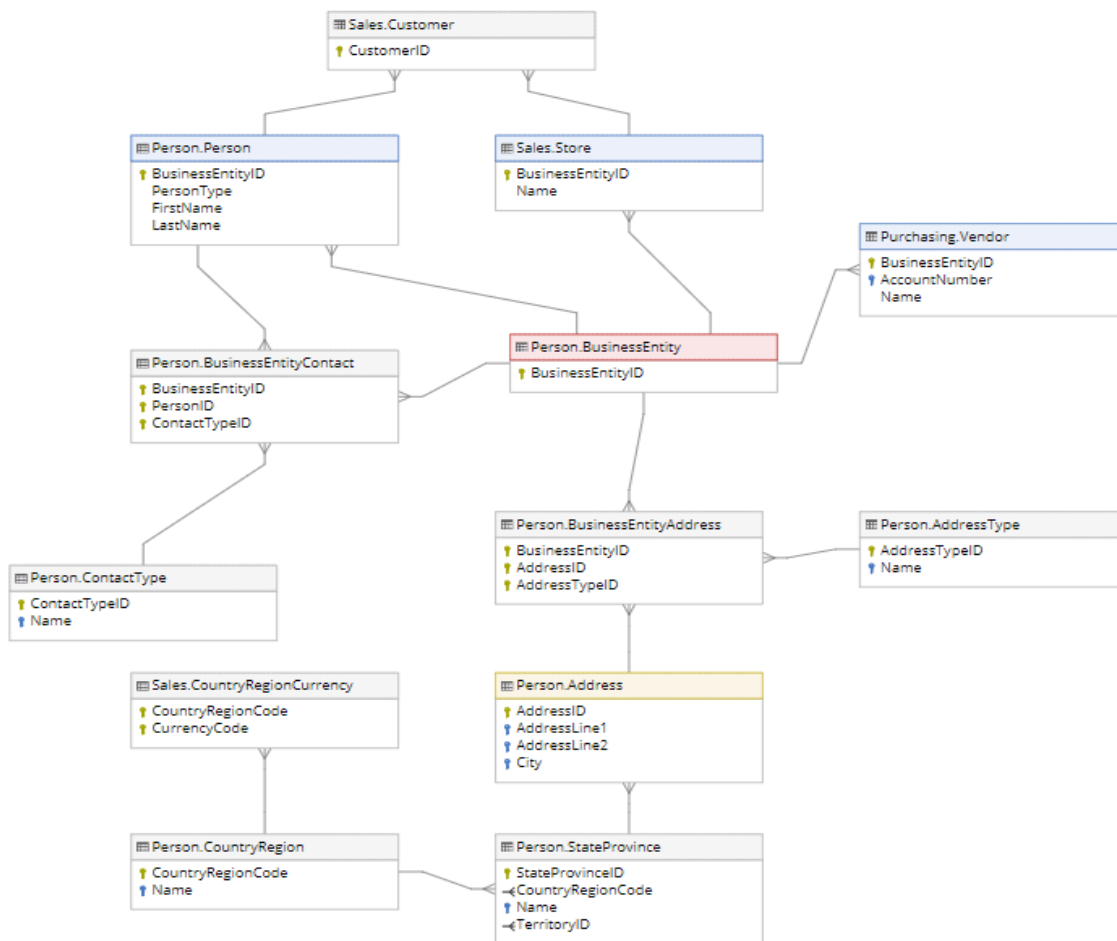
Chapter 2 DATA PREPARATION

Chapter overview: Pre-process the outliers, perform data visualization to find relevant variables, and then prepare the input data for application to machine learning methods including Elbow and Silhouette.

2.1 Data understanding

2.1.1 The table relationship

Business Entities



10

Figure 2.1: Business entities of AdventureWork company

¹⁰ <https://dataedo.com/samples/html/AdventureWorks/>, 2/12/2022

To better understand the relationship between the data tables, the team used a link table like Figure 2.1.1. Here we have a more overview of how to access data in the tables together. For example, to be able to find out the customer's location, we need to access the data in the table Preson.Address. Those tables are linked to one table - BusinessEntity that holds ID for all vendors, customers, and employees tables.

2.1.2 Data information tables

- **Sales_data**

```
1 df_sales.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121253 entries, 0 to 121252
Data columns (total 15 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   SalesOrderLineKey                     121253 non-null int64
1   ResellerKey                           121253 non-null int64
2   CustomerKey                           121253 non-null int64
3   ProductKey                            121253 non-null int64
4   OrderDateKey                           121253 non-null int64
5   DueDateKey                            121253 non-null int64
6   ShipDateKey                           119140 non-null float64
7   SalesTerritoryKey                     121253 non-null int64
8   Order Quantity                         121253 non-null int64
9   Unit Price                            121253 non-null float64
10  Extended Amount                       121253 non-null float64
11  Unit Price Discount Pct               121253 non-null int64
12  Product Standard Cost                  121253 non-null float64
13  Total Product Cost                     121253 non-null float64
14  Sales Amount                           121253 non-null float64
dtypes: float64(6), int64(9)
memory usage: 13.9 MB
```

In the Sales_data dataset there are 7 columns representing 7 elements, and each of these columns has a total of 121253 rows. And the data information of each column will be presented in the table below:

Column	Datatype
SalesOrderLineKey	int64
ResellerKey	int64
CustomerKey	int64
ProductKey	int64

OrderDateKey	int64
DueDateKey	int64
ShipDateKey	float64
SalesTerritoryKey	int64
Order Quantity	int64
Unit Price	float64
Extended Amount	float64
Unit Price Discount	int64
Product Standard Cost	float64
Total Product Cost	float64
Sales Amount	float64

Table 2.1: Information of Sales_data

- **Sales Order_data**

```
[29] 1 df_sales_order.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121253 entries, 0 to 121252
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Channel                121253 non-null object
1   SalesOrderLineKey      121253 non-null int64
2   Sales Order            121253 non-null object
3   Sales Order Line       121253 non-null object
dtypes: int64(1), object(3)
memory usage: 3.7+ MB
```

In the Sales Order_data dataset there are 4 columns representing 4 elements, and each of these columns has a total of 121253 rows. And the data information of each column will be presented in the table below:

Column	Datatype
Channel	object
SalesOrderLineKey	int64
Sales Order	object
Sales Order Line	object

Table 2.2: Information of Sales Order_data

- **Date_data**

```
[30] 1 df_date.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1461 entries, 0 to 1460
Data columns (total 7 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   OrderDateKey    1461 non-null  int64
1   Date            1461 non-null  datetime64[ns]
2   Fiscal Year     1461 non-null  object
3   Fiscal Quarter  1461 non-null  object
4   Month           1461 non-null  object
5   Full Date       1461 non-null  object
6   MonthKey        1461 non-null  int64
dtypes: datetime64[ns](1), int64(2), object(4)
memory usage: 80.0+ KB
```

In the Date_data dataset there are 7 columns representing 7 elements, and each of these columns has a total of 1461 rows. And the data information of each column will be presented in the table below:

Column	Datatype
OrderDateKey	int64
Date	datetime64
Fiscal Year	object
Fiscal Quarter	object
Month	object

Full Date	object
MonthKey	int64

Table 2.3: Information of Date_data

- **Product_data**



```
1 df_product.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 397 entries, 0 to 396
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  -
0   ProductKey      397 non-null    int64
1   SKU             397 non-null    object
2   Product         397 non-null    object
3   Standard Cost   397 non-null    float64
4   Color           341 non-null    object
5   List Price      397 non-null    float64
6   Model           397 non-null    object
7   Subcategory     397 non-null    object
8   Category        397 non-null    object
dtypes: float64(2), int64(1), object(6)
memory usage: 28.0+ KB
```

In the Date_data dataset there are 9 columns representing 9 elements, and each of these columns has a total of 397 rows. And the data information of each column will be presented in the table below:

Column	Datatype
ProductKey	int64
SKU	object
Product	object
Standard Cost	float64
Color	object
Model	object
Subcategory	object

Category	object
----------	--------

Table 2.4 : Information of Product_data

- **Customer_data**

```
[32] 1 df_customer.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18485 entries, 0 to 18484
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   CustomerKey            18485 non-null  int64   
1   Customer ID            18485 non-null  object  
2   Customer                18485 non-null  object  
3   City                   18485 non-null  object  
4   State-Province         18485 non-null  object  
5   Country-Region         18485 non-null  object  
6   Postal Code            18485 non-null  object  
dtypes: int64(1), object(6)
memory usage: 1011.0+ KB
```

In the Date_data dataset there are 7 columns representing 7 elements, and each of these columns has a total of 18485 rows. And the data information of each column will be presented in the table below:

Column	Datatype
CustomerKey	int64
Customer ID	object
Customer	object
City	object
State-Province	object
Country-Region	object
Postal Code	object

Table 2.5 : Information of Customer_data

- **Sales Territory_data**

```
[33] 1 df_territory.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11 entries, 0 to 10
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype  
---  -
0   SalesTerritoryKey 11 non-null    int64  
1   Region            11 non-null    object  
2   Country           11 non-null    object  
3   Group             11 non-null    object  
dtypes: int64(1), object(3)
memory usage: 480.0+ bytes
```

In the Date_data dataset there are 4 columns representing 4 elements, and each of these columns has a total of 11 rows. And the data information of each column will be presented in the table below:

Column	Datatype
SalesTerritoryKey	int64
Region	object
Country	object
Group	object

Table 2.6: Information of Sales Territory_data

- **Reseller_data**

```
[34] 1 df_reseller.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 702 entries, 0 to 701
Data columns (total 8 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   ResellerKey            702 non-null    int64  
1   Reseller ID            702 non-null    object  
2   Business Type          702 non-null    object  
3   Reseller               702 non-null    object  
4   City                   702 non-null    object  
5   State-Province         702 non-null    object  
6   Country-Region         702 non-null    object  
7   Postal Code            702 non-null    object  
dtypes: int64(1), object(7)
memory usage: 44.0+ KB
```

In the Date_data dataset there are 8 columns representing 8 elements, and each of these columns has a total of 702 rows. And the data information of each column will be presented in the table below:

Column	Datatype
ResellerKey	int64
Reseller ID	object
Business Type	object
Reseller	object
City	object
State-Province	object
Country-Region	object
Postal Code	object

Table 2.7: Information of Reseller_data

2.2 Data collection

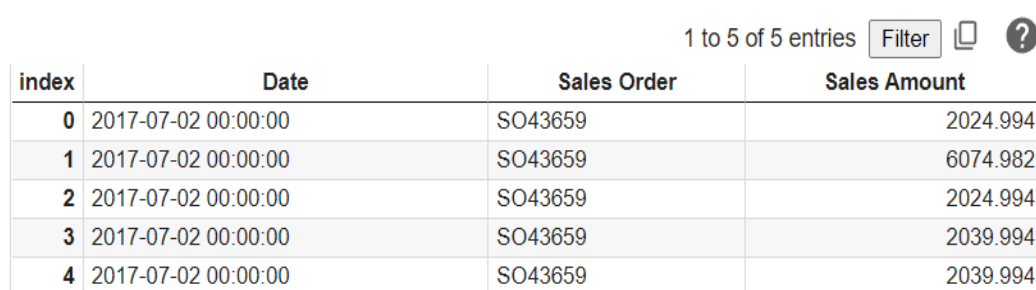
After doing research on the information of the data set, the team will filter and select variables that are related to each other and have a great impact on customer behavior in choosing products.

First, to be able to calculate Recency, we need to know when the most recent transaction date of the customer is. So here we need to use the Sales_data table with the OrderDateKey column referencing the Data_data table to know when the customer's last transaction date is.

Next, to retrieve the purchase frequency of each customer during the transaction history from June 2017 to June 2020 is how many groups make a reference from the Sales_data table with the SalesOrderLineKey column to the Sales Order_data table with the Sales Order_data table with the SalesOrderLineKey column. Sales Order column. From there, it is possible to calculate how many times each customer buys at the company.

And finally, to calculate the total amount of money each customer has spent on the business, we need to retrieve the Sales Amount column in the Sales_data data table. Then calculate the total amount spent by the customer in each transaction.

So from the three factors above we get a new table like the picture below:



The image shows a screenshot of a data table interface. At the top right, it says "1 to 5 of 5 entries" followed by a "Filter" button, a copy icon, and a help icon. The table has four columns: "index", "Date", "Sales Order", and "Sales Amount". There are five rows of data, all with the same date (2017-07-02 00:00:00) and sales order (SO43659), but different sales amounts.

index	Date	Sales Order	Sales Amount
0	2017-07-02 00:00:00	SO43659	2024.994
1	2017-07-02 00:00:00	SO43659	6074.982
2	2017-07-02 00:00:00	SO43659	2024.994
3	2017-07-02 00:00:00	SO43659	2039.994
4	2017-07-02 00:00:00	SO43659	2039.994

and the objects in the table will include:

```
[42] 1 df_rfm.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 121253 entries, 0 to 121252
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype  
---  -
0   Date            121253 non-null  datetime64[ns]
1   Sales Order     121253 non-null  object  
2   Sales Amount    121253 non-null  float64  
dtypes: datetime64[ns](1), float64(1), object(1)
memory usage: 3.7+ MB
```

In the RFM_dataset there are 3 columns representing 3 elements R-F-M, and each of these columns has a total of 12153 rows. And the data information of each column will be presented in the table below:

Column	Datatype
Date	datetime64[ns]
Sales Order	object
Sales Amount	float64

Table 2.8: Information of RFM_dataset

2.3 Exploratory Data Analysis

EDA (Exploratory Data Analysis) is an important step before doing any ML problem with tabular data. Before building the model, you need to build the feature. Before building the feature, you must do the data discovery step. This EDA step gives us a first look at the data. You need to have a certain feel for what you have in hand before you have modeling strategies. EDA helps you visualize the complexity of the problem and outlines the first steps to take.

Data exploration should not only stop at the first time before feature building but should also be done throughout the system development process. After building the features, you also need to do the EDA again to see if the processed data is clean. In addition, after building and analyzing the model, we often need to return to EDA to continue to discover what is still hidden in the problem data. The deeper you understand

the data, the sooner you will be able to interpret the behavior of the model and make the appropriate changes.

2.3.1 Removing null and NA data

The first step before performing a search for outliers is to ensure that the data is complete by removing nulls and NA values:

```
[189]  1 df_rfm = df_rfm.dropna()
      2 df_rfm.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 121253 entries, 0 to 121252
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Date            121253 non-null  datetime64[ns]
1   Sales Order     121253 non-null  object
2   Sales Amount    121253 non-null  float64
dtypes: datetime64[ns](1), float64(1), object(1)
memory usage: 3.7+ MB
```

```
[190]  1 df_rfm = df_rfm.isnull()
      2 df_rfm.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 121253 entries, 0 to 121252
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Date            121253 non-null  bool
1   Sales Order     121253 non-null  bool
2   Sales Amount    121253 non-null  bool
dtypes: bool(3)
memory usage: 1.3 MB
```

Then, when performing the data collection, the receiving group at CustomerKey has a large number of -1 values. Explained according to the business's business, these values are customers who buy intermediaries through a 3rd party business, or these customers are temporary customers, without identifiable information. So, in order to ensure the customer clustering is based on common characteristics and to keep the data from being biased, giving the best results, in this study, the CustomerKey -1 values will be removed.

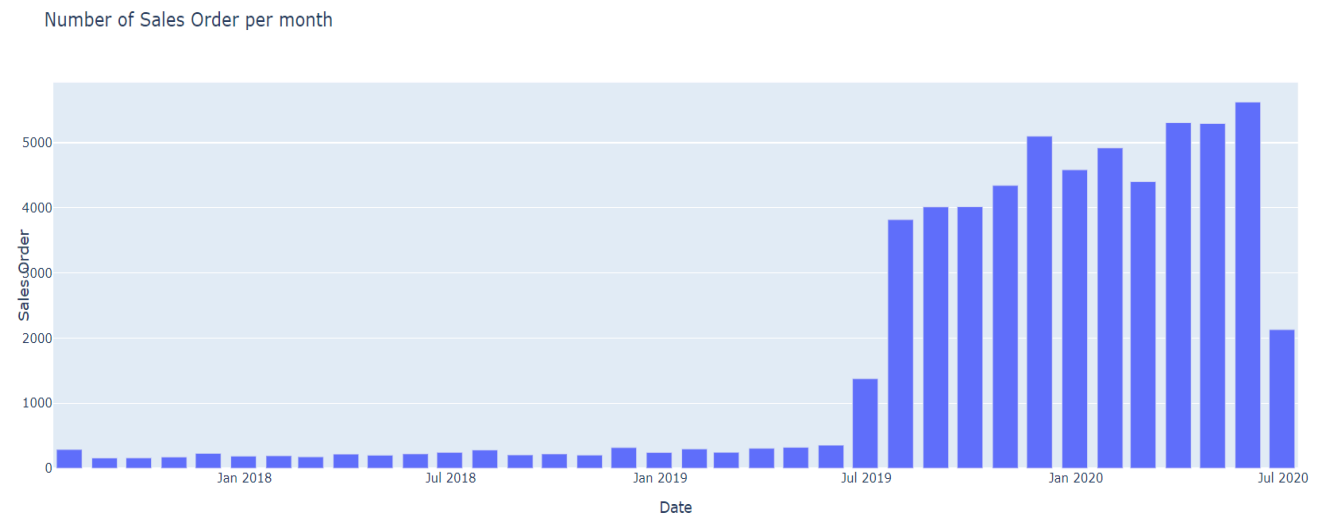
```
[191] 1 df_merge = df_merge[(df_merge['CustomerKey'] != -1)]
      2 df_rfm = df_merge.iloc[:, [18, 16, 14]]
      3 df_rfm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 60398 entries, 60855 to 121252
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Date            60398 non-null  datetime64[ns]
1   Sales Order     60398 non-null  object
2   Sales Amount    60398 non-null  float64
dtypes: datetime64[ns](1), float64(1), object(1)
memory usage: 1.8+ MB
```

After removing redundant values, the dataset will be left with 60 389 transactions. And these data are all non-null values.

2.3.2 Data visualization

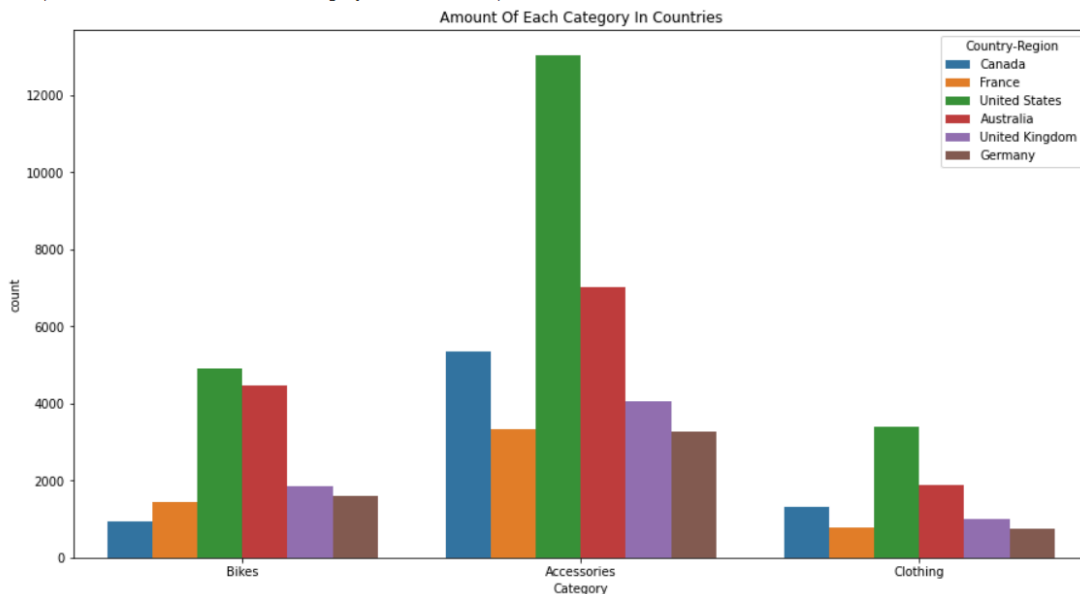
First, we need to look at the number of orders the company achieves each month through the chart below



So, we can see that the number of orders started to increase rapidly from July 2019 and reached the milestone of 5624 orders by the end of May 2020. It can be concluded that during this period, customers are quite fond of our products. business, so the number of orders is very high. Or it could be that the company's marketing campaigns have

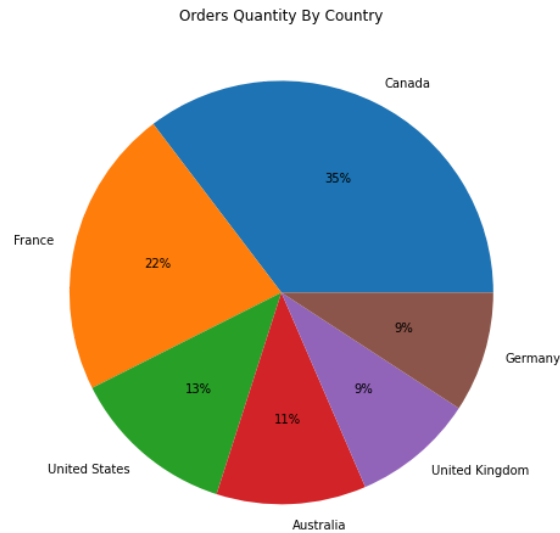
worked, and there could be other reasons. There is a need to review and review the products offered at this time.

To learn about the above problem, the team visualized the number of products consumed the most through the chart below:



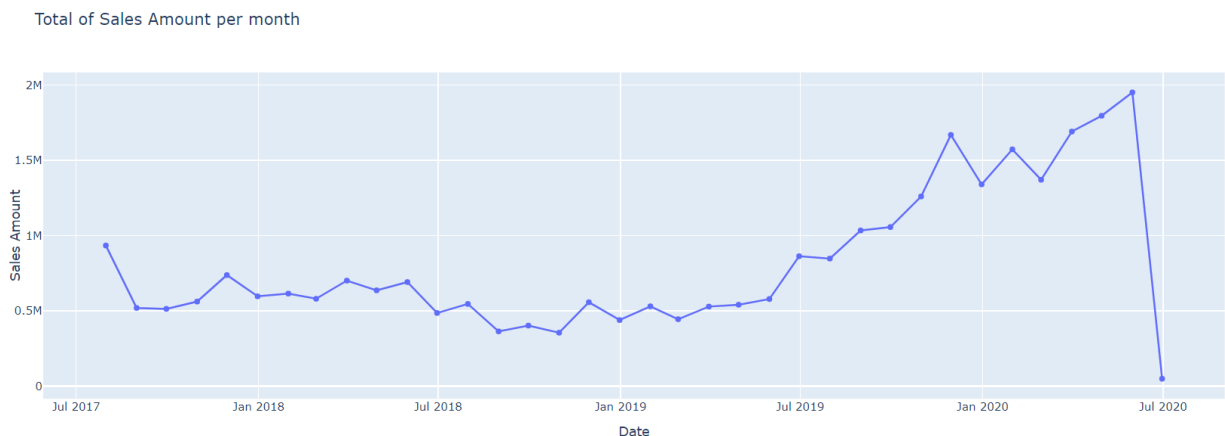
It can be seen that the United State market is where the largest number of products are consumed. In which, the product that all 6 markets consume the most is Accessories, followed by Bike and finally Clothing. Thus, Accessories products are very popular in many markets, especially United State market.

So, to be able to see where the potential market is, the number of customers shopping the most, the team visualized through the chart below.



Through this pie chart, we can see that the market with the most customers for businesses is Canada with the total number of visitors accounting for 35%, followed by the market in France with the number of customers up to 22%. So, businesses need to pay attention to the production of products that are suitable for the tastes of customers in these two markets.

Next, to be able to see the total sales of the business through each transaction month, we can identify it through the line chart below:



From the above line chart, we can see that the revenue of the business fluctuated insignificantly over each month from 6/2017 to 5/2019. But from the above time onwards until the beginning of June 2020, the revenue has grown significantly from 578 000 USD,

skyrocketing to nearly 2 000 000 USD. This increase can be compared with the increase in the number of orders during the same period. Thus, it can be said that in this period, enterprises have had very successful business strategies and brought great results.

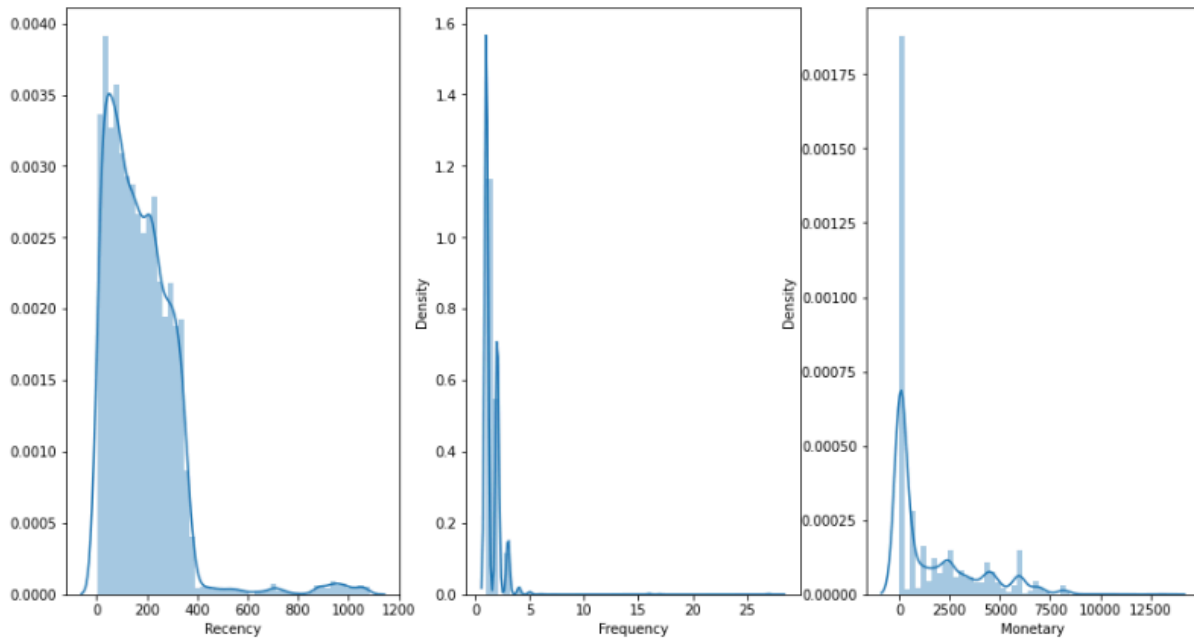
2.3.3 Calculating RFM

Then the next step is to calculate the variables Recency, Frequency and Monetary. With Recency we can calculate by taking the most recent purchase date of the data set (Max_date) and then subtracting the latest purchase date of each order in each customer. Next with Frequency, we can use the "Count distinct" command to count the number of orders that customers place in that period. And finally, Monetary, we will use the "Sum" function to calculate the total amount of money that customers have spent in the period. So, after performing the calculation, we will have the following data set:

```
[115] 1 df_rfm.head(5)
```

	Recency	Frequency	Monetary
CustomerKey			
11000.0	256	3	8248.99
11001.0	35	3	6383.88
11002.0	325	3	8114.04
11003.0	249	3	8139.29
11004.0	258	3	8196.01

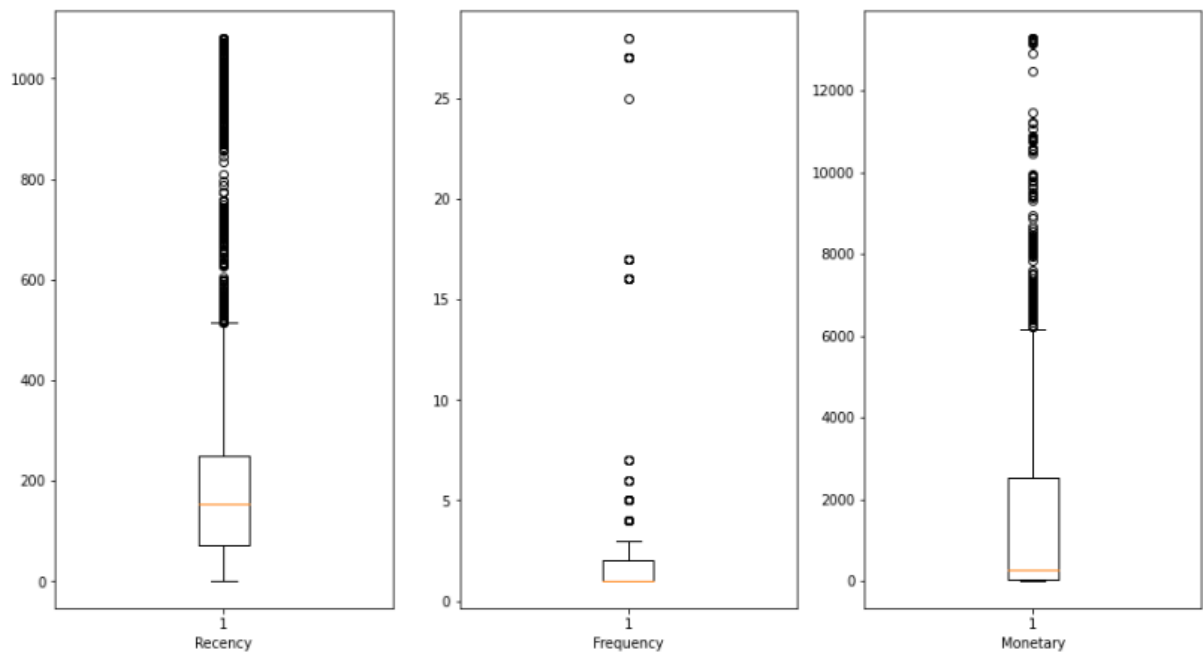
In these data variables, we visualize the frequency distribution chart of 3 factors Recency, Frequency and Monetary.



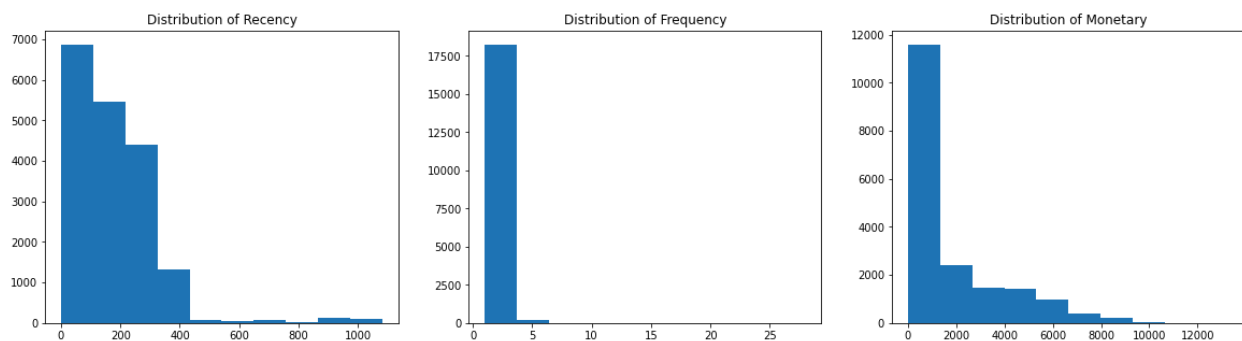
Through the visualization of the above data set, we can see that most of the data is skewed to the right. This shows that the dataset is unevenly distributed and highly concentrated at low values. Part of these factors occur due to the fact that the dataset has many large deviation values, and many Outliers have not been found.

2.3.4 Removing Outliers

Text(0.5, 0, 'Monetary')



Through the box-plot chart above, we can see that the values are outside the marginal area quite a lot, the Outliers data accounts for large data and is outside Q3 of the dataset. This will be shown more clearly through the histogram.



Through the chart above, we can see that in Recency, the customer's most recent purchase date accounted for the highest percentage from 0-250 days.

For the chart of Frequency, the frequency data of customers buying is mainly concentrated from 1-2 times to buy the most.

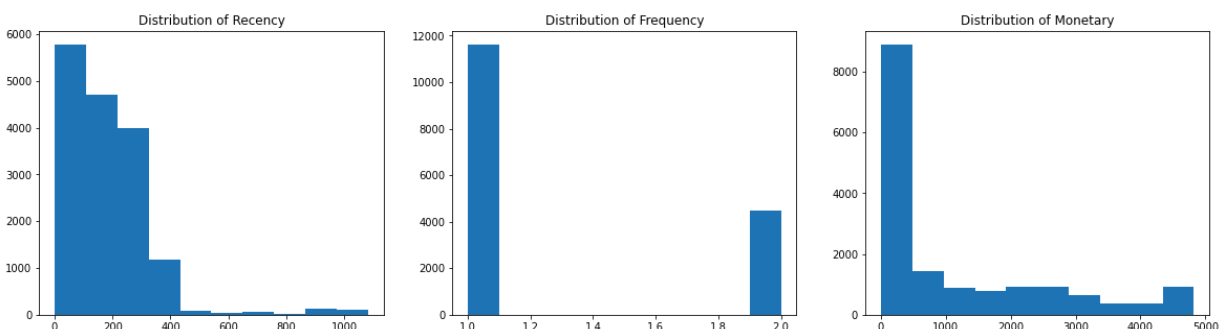
For Monetary, the total amount of money each customer spends for businesses will mainly focus on 100-2000 USD.

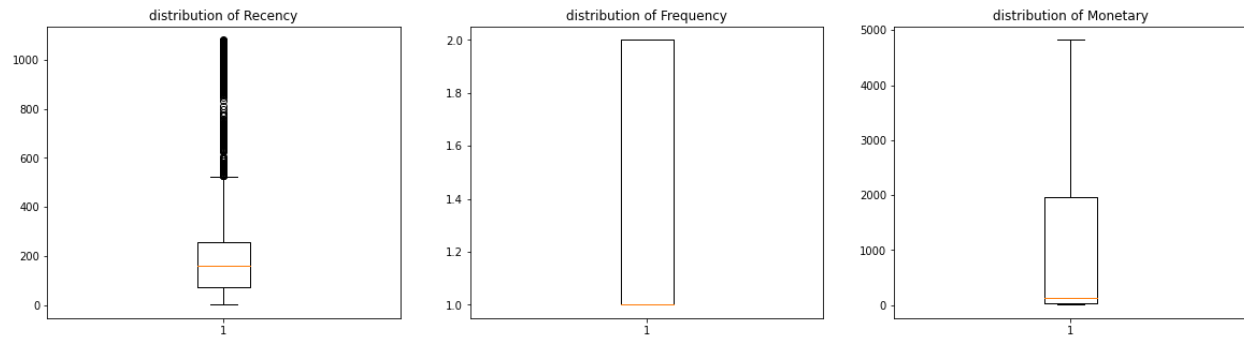
However, it can be said that the above data, if we do not remove the outliers, will greatly affect the application to the customer segmentation model because the data is unevenly distributed. So, in order to optimize the data and help the model build high efficiency, we need to remove outliers.

count	18484.000000	count	18484.000000
mean	1588.329216	mean	1.496375
std	2124.231717	std	1.101139
min	2.290000	min	1.000000
25%	49.970000	25%	1.000000
50%	270.265000	50%	1.000000
75%	2511.275000	75%	2.000000
85%	4337.560000	85%	2.000000
90%	4826.839000	90%	2.000000
100%	13295.380000	100%	28.000000
max	13295.380000	max	28.000000
Name: Monetary, dtype: float64		Name: Frequency, dtype: float64	

Through the overview information about Monetary and Frequency we can see that more than 90% of customer spending on business is approximately \$4,800; and more than 90% of the times customers come back to shop at the business is about 2 times. So the remaining 10% of the two datasets above will be outliers. 10% of these customers will be analyzed later.

So, after removing these outliers, we will have a more regular frequency distribution table:





Thus, after visualizing R F and M after processing the outliers, we see that the dataset has a more uniform distribution among the variables. And clearly see the milestones for each data variable.

2.3.5 Transform data

2.3.5.1 Scaling data

▶

1 df_rfm.describe()

📄

	Recency	Frequency	Monetary
count	18484.000000	18484.000000	18484.000000
mean	175.667983	1.496375	1588.329216
std	145.644062	1.101139	2124.231717
min	1.000000	1.000000	2.290000
25%	72.000000	1.000000	49.970000
50%	154.000000	1.000000	270.265000
75%	249.000000	2.000000	2511.275000
max	1081.000000	28.000000	13295.380000

✎

Looking at the picture above, the Recency value ranges from 1 to 1081 (the latest purchase date), the Frequency ranges from 1 to 28 (the time of purchase). In particular, Monetary is the value with the largest range from 2.29 to 13295.3 (currency unit). When

looking at the distribution of the quartiles in Monetary, it can be seen that Monetary has a much larger value than the other two factors.

Because of the distribution of the above factors, there are many inadequacies and unevenness among the elements in the data set that can affect the results of the model building. To be able to help the data variables not be different because the scale of each variable is different, we need to convert the units of each data variable into a common scale so that it can be included in machine learning to calculate the number of clusters k . Also known as Min-Max scaling.

	Recency	Frequency	Monetary	minmax_R	minmax_F	minmax_M
CustomerKey						
11012.0	91	2	81.2600	0.083333	1.0	0.016381
11013.0	4	2	113.9600	0.002778	1.0	0.023165
11014.0	259	2	138.4500	0.238889	1.0	0.028245
11015.0	361	1	2500.9700	0.333333	0.0	0.518321
11016.0	339	1	2332.2800	0.312963	0.0	0.483328
...
29479.0	497	1	2049.0982	0.459259	0.0	0.424586
29480.0	181	1	2442.0300	0.166667	0.0	0.506095
29481.0	885	1	3374.9900	0.818519	0.0	0.699626
29482.0	483	1	2049.0982	0.446296	0.0	0.424586
29483.0	492	1	2049.0982	0.454630	0.0	0.424586

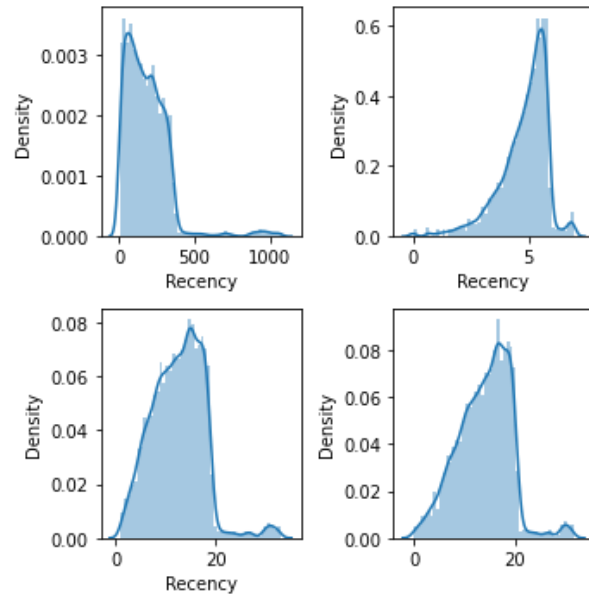
[16120 rows x 6 columns]

The goal of the method is to bring the values closer to the meaning of the features. This method returns the values to a special interval, usually $[0, 1]$, $[0, 1]$ or $[-1, 1]$. One of the limitations of this method is that when applied to a small range of values, we get a smaller standard deviation, which reduces the weight of the outliers in the data.

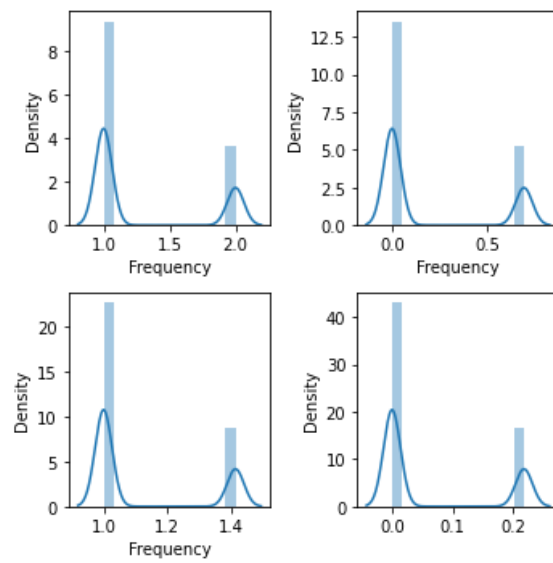
2.3.5.2 Normalization Data

We have 3 values to normalize; all three values need to be transformed. My team will offer 3 usage models: log transformation, square root transformation and box cox transformation.

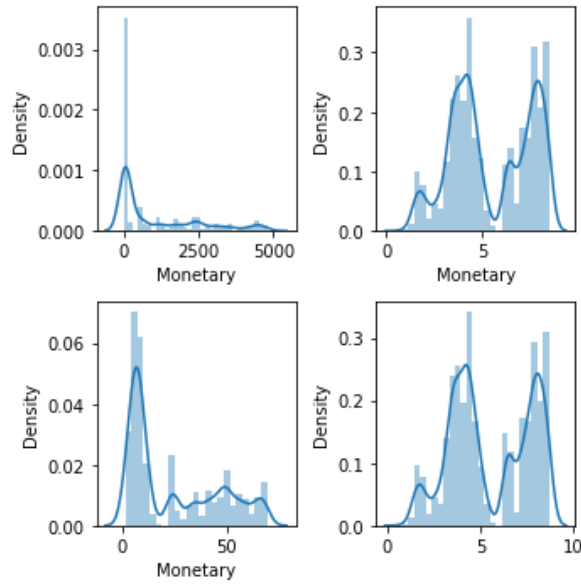
- Recency



- Frequency



- Monetary



Therefore, after data transformation, we find that using the box cox method is the most optimal because the results are asymptotically close to 0. Shows the lowest error in the 3 methods used and has the following result:

	Recency	Frequency	MonetaryValue
0	11.298616	0.217658	4.447710
1	1.797100	0.217658	4.793932
2	17.726654	0.217658	4.993488
3	20.327109	0.000000	7.983831
4	19.810562	0.000000	7.911148

In summary, after performing data collection, visualization of variables and preprocessing of variables, the data set is basically complete so that it can be applied to machine learning calculation methods in the next chapter.

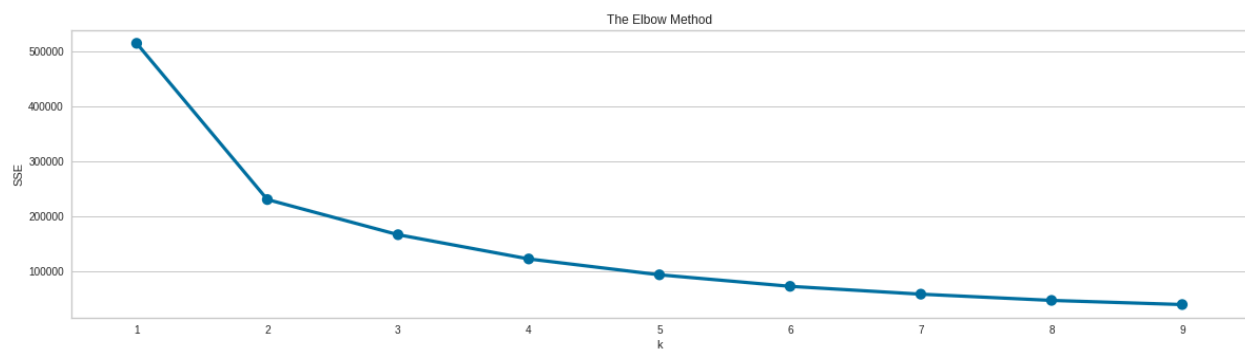
Chapter 3 CUSTOMER SEGMENTATION WITH MACHINE LEARNING METHOD

Chapter overview: Presenting procedures for applying RFM model on AdventureWord dataset. From there, calculate the number of clusters using Elbow and Silhouette. And start comparing the results and assigning labels to the data variables.

3.1 RFM with Machine Learning methods

After performing visualization of the data variables R F and M, the next step is to cluster and calculate the number of groups K of the dataset with the Elbow method.

3.1.1 Elbow method

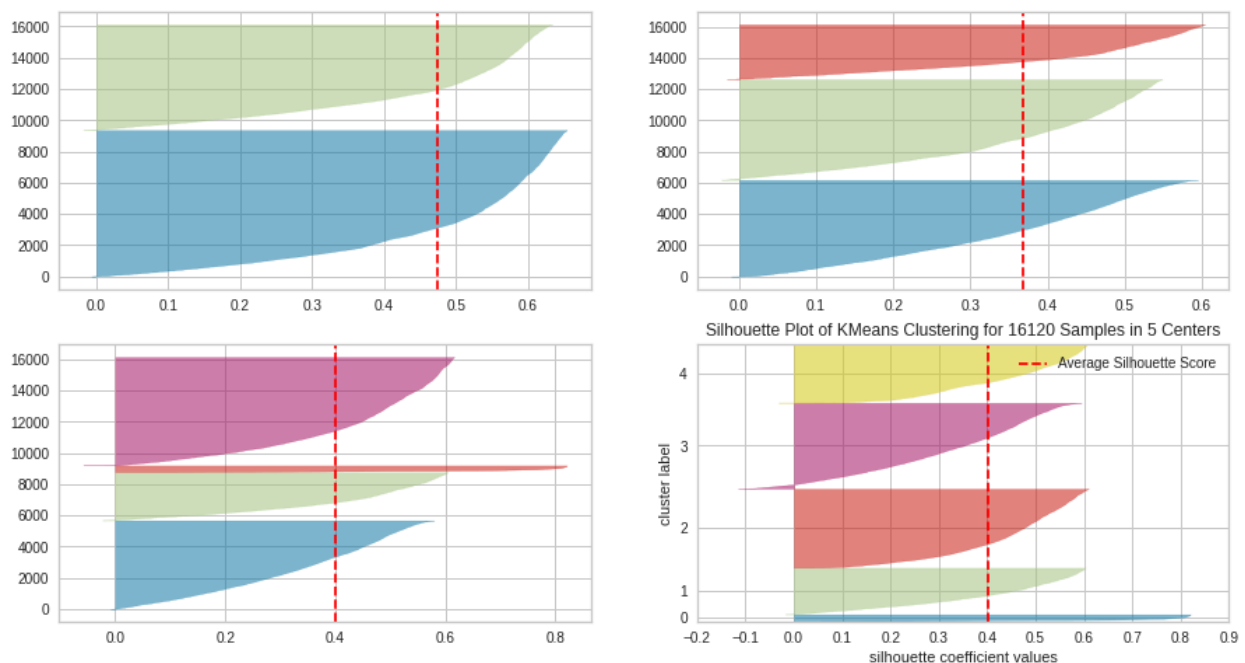


The elbow is the point at which the deceleration rate of the distortion function will change significantly. In other words, from this position, increasing the number of clusters does not help reduce the distortion function so much. If the algorithm divides according to the number of clusters at this position, it will achieve the most general clustering without overfitting phenomena. In the above figure, we see the position of the elbow is $k=2$ because when the number of clusters increases, the deceleration rate of the distortion function seems to be negligible compared to the previous numbers of clusters. Nevertheless, there are some situations in which we easily find the position of the Elbow, especially for datasets having the rule of clustering is not really easy to detect. From that,

we need to consider $k=3$ or $k=4$ - positions are close to the elbow. To find the number of clusters more exactly, we implemented the second method that is Silhouette Score.

3.1.2 Silhouette method

The term "silhouette analysis" refers to yet another metric for assessing the quality of clustering. Other clustering algorithms can also use silhouette analysis. The silhouette coefficient ranges from -1 to 1, with a higher silhouette coefficient indicating a model with more coherent clusters. In other words, a silhouette coefficient near +1 indicates that the sample is far from the nearby clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters. Negative results also suggest that the samples may have been misassigned to a cluster.



The uniformity between clusters in groups of two clusters and groups of three clusters is evident in the image above. Although both are greater than the ASS (Average Silhouette Value), the performance for group 3 does not decline despite increasing by one cluster, which can help us delve deeper into the characteristics of that customer group. We underestimate their efficacy for groups of 4 and 5 clusters because the differences between the groups are stark and unbalanced. In conclusion, we think that the k means

algorithm, with $k=3$, will produce the best results for aiding in the analysis and labeling of each customer cluster.

3.2 Customer segmentation

	Recency	Frequency	Monetary
Cluster			
0	163.98	1.14	137.72
1	249.97	1.89	3932.03
2	186.91	1.31	1957.86

```
0    64.3%
2    21.7%
1    14.1%
Name: Cluster, dtype: object
```

3.2.1 Analysis of loyal customer groups (cluster 1)

	Recency	Frequency	Monetary	minmax_R	minmax_F	minmax_M	cluster
count	6446.000000	6446.000000	6446.000000	6446.000000	6446.000000	6446.000000	6446.0
mean	312.191126	1.207260	1116.671787	0.288140	0.207260	0.231165	1.0
std	152.553508	0.405375	1419.504360	0.141253	0.405375	0.294459	0.0
min	199.000000	1.000000	2.290000	0.183333	0.000000	0.000000	1.0
25%	233.000000	1.000000	39.980000	0.214815	0.000000	0.007818	1.0
50%	278.000000	1.000000	131.950000	0.256481	0.000000	0.026896	1.0
75%	325.000000	1.000000	2181.562500	0.300000	0.000000	0.452064	1.0
max	1081.000000	2.000000	4820.310000	1.000000	1.000000	0.999440	1.0

According to the quartile description of the quartiles of the loyal customer group, the customer segment of this group has the number of 6446, accounting for 14.1% total customers. In which, the average date of the latest purchase in the group is about 312

days, this is quite high compared to the other groups but, the average purchase frequency is about 2 times and the average amount spent per transaction is 1117. This group of customers are willing to spend the most tar for each shopping activity.

With the characteristics of Recency, Frequency and Monetary, we can see that this group of customers is not just a loyal group of customers, but even a group of customers with great potential for business.

To increase the revenue of the business, retaining this group of customers is absolutely necessary. Some ways to increase their loyalty to the business:

Always bring convenience to them: Any customer wants to experience a product or service in the most convenient way. Convenience here can be customer experience, quick payment, convenience or easy shopping. Therefore, businesses should optimize the convenience for customers both before and after shopping. Examples: Improve web browser (online business), Expand multiple payments, integrate many shipping units for customers to choose easily.

Optimizing personalized customer experience: Capturing the right customer insight is one of the most important goals from which to come up with appropriate sales strategies, attracting customers to use your products more often.

Create dynamic membership levels: each seller rank establishes the benefits and incentives that customers will receive when reaching that level in order to attract and motivate customers to buy products. This form applies to shops with many different customer levels, and each level has its own incentives.

3.2.2. Analysis of potential loyalist customer groups (cluster 2)

	Recency	Frequency	Monetary	minmax_R	minmax_F	minmax_M	cluster
count	3494.000000	3494.000000	3494.000000	3494.000000	3494.000000	3494.000000	3494.0
mean	33.463938	1.321122	900.293397	0.030059	0.321122	0.186280	2.0
std	18.552183	0.466974	1310.346230	0.017178	0.466974	0.271815	0.0
min	1.000000	1.000000	2.290000	0.000000	0.000000	0.000000	2.0
25%	18.000000	1.000000	39.980000	0.015741	0.000000	0.007818	2.0
50%	33.000000	1.000000	102.570000	0.029630	0.000000	0.020802	2.0
75%	49.000000	2.000000	1622.470000	0.044444	1.000000	0.336087	2.0
max	67.000000	2.000000	4822.110000	0.061111	1.000000	0.999813	2.0

Figure 3....: Describe the quartiles of the potential loyalist customer group

The number of potential loyalist customers accounted for 21.7% of the total number of customers analyzed. This client group is the most recent; their time period is around 185 days, and they are nearing the new customer group (clustering 0 is 165 days). Customers spend a significant amount of money, around \$2,000, on firm items, and their frequency of purchases is also extremely high when they return to buy more than once.

Businesses must create trust and encourage this set of consumers back by creating a loyalty program. Furthermore, it is vital to care for and solicit clients in order to generate interest in them. They are the ones who are eager to spend money, thus firms.

When we look at this set of clients, we notice that they are people who are willing to spend money on the company's products, and their spending is not insignificant. They are also likely to be new consumers, therefore your approach to them requires special consideration. They also routinely purchase from the firm; therefore, the company has previously pleased them.

Offer membership/loyalty program: Businesses must build trust and encourage this set of clients to return to the firm by implementing a loyalty program.

Recommend other products: Furthermore, it is critical to care for and engage consumers in order to build interest in them. They are enthusiastic spenders, therefore

businesses only need to provide them with the chance to spend money, such as by informing them about new items or great products from the firm.

3.2.3. Analysis of new customer groups (cluster 0)

	Recency	Frequency	Monetary	minmax_R	minmax_F	minmax_M	Cluster	cluster
count	10358.00000	10358.000000	10358.000000	10358.000000	10358.000000	10358.000000	10358.0	10358.0
mean	163.97876	1.135740	137.721821	0.150906	0.135740	0.028094	0.0	0.0
std	107.59335	0.342529	203.149912	0.099623	0.342529	0.042141	0.0	0.0
min	1.00000	1.000000	2.290000	0.000000	0.000000	0.000000	0.0	0.0
25%	73.00000	1.000000	30.970000	0.066667	0.000000	0.005949	0.0	0.0
50%	156.00000	1.000000	60.470000	0.143519	0.000000	0.012069	0.0	0.0
75%	249.00000	1.000000	108.850000	0.229630	0.000000	0.022105	0.0	0.0
max	1066.00000	2.000000	1000.437500	0.986111	1.000000	0.207054	0.0	0.0

This group of customers' accounts for 64.3% of the total number of customers of the business with an average of 164 days of return visits, a purchase frequency of only 1, and a spend per purchase of 138, these indicators are at a relatively low level, but their percentages account for the most.

With this group of customers, businesses can continue to improve their current sales policies to retain this key customer group. Besides finding potential customers in this group and promoting them to become potential loyalist customers.

Some ways to motivate them:

Send customer engagement emails: Emails provide businesses with a great opportunity to build and deepen relationships with customers both before and after they shop. It's important that every email you send adds value to your customer's shopping experience, or you risk losing this customer. Sending them notifications of new product launches, recommendations, or ongoing promotions are ways you can keep your customers engaged.

Offer new customers discounts and promotions: Today's consumers are still searching for good deals and value. Offer introductory discounts, specials like buy two,

get one free or free gift wrapping for the first three purchases to entice customers to your store. Deals like these can draw in new clients who were considering doing business with you but required a push to alter their purchasing patterns. Then keep tabs on their purchases and the promotions they took advantage of so you can more effectively target them in the future with messages that will win their loyalty.

Chapter 4

4.1

References

- [1] Sunil Erevelles, Nobuyuki Fukawa, Linda Swayne (2016), “Big Data consumer analytics and the transformation of marketing”, *Journal of Business Research*, Volume 69, Issue 2, February 2016, p897-904.
- [2] Sunil Erevelles, Nobuyuki Fukawa, Linda Swayne (2016), “Big Data consumer analytics and the transformation of marketing”, *Journal of Business Research*, Volume 69, Issue 2, February 2016, p897-904
- [3] Atis Verdenhofs, Tatjana Tambovceva, (2019), "Evolution of Customer Segmentation in the Era of Big Data", *Marketing and Management of Innovations*, pages 1-20.
- [4] Truong Thi Hoai Linh, Le Thi Nhu Quynh (2019), “Big Data and applications in banking”, *Banking Magazine*, No. 17/2019
- [5] Jo-Ting, W., Shih-Yen, L., & Hsin-Hung, W. (2010). “A review of the application of RFM model”. *African Journal of Business Management*, 4(19), 4199-4206.
- [6] Saritha M, Manoj B R, Neola Sendril Dias, Nisha Joshal Pinto and Padma Prasad H M (2022), ‘Segmentation of Mall customers using RFM Analysis and K-means Algorithm’, *Journal of Data Mining and Management* (7).
- [7] Aylanur Cuce and Eda Tiryaki (2022), "*Data Analytics in Customer Segmentation and RFM Method*", Master's thesis, Istanbul Technical University.
- [8] Basim Amer Jaafar, Methaq Talib Gaata and Mahdi Nsaif Jasim (2020), "Home appliances recommendation system based on weather information using combined modified k-means and elbow algorithms", *Indonesian Journal of Electrical Engineering and Computer Science*, pages 19(3):1635.
- [9] Ho Trung Thanh, Nguyen Dang Son (2021), "An interdisciplinary study between customer segmentation analysis in marketing and machine learning methods", *Journal of Science and Technology Development - Economics - Law and manage*, 6 (1): 2005-2015.
- [10] Wayne D. Hoyer & Deborah J. MacInnis (2008), *Customer Behavior*, Cengage Learning, UK.

- [11] Peter D. Bennett (1995), *Dictionary of Marketing Terms*, NTC Business Books, US.
- [12] Philip Kotler (2001), *Marketing Management*, Pearson Education Canada, Canada.
- [13] Hughes, A.M., 1996. Boosting response with RFM. *American Demographics*, 4-10.
- [14] Claudio Marcus (1998), "A practical yet meaningful approach to customer segmentation", *Journal of Consumer Marketing*, vol. 15, no. 5, page 494-504.