

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT



GROUP 5

REPORT

RFM analysis for Customer Segmentation

Subject: Fundamentals of Data Analytics

PROJECT OVERVIEW	3
I. Reasons for choosing the topic.....	3
II. Objective	4
III. Objects and Scores	4
IV. Research Methods	4
CHAPTER 1: RELATED WORK	5
1. Researches	5
2. Process Maps and Segmentations	9
3. RFM Overview	10
CHAPTER 2: THEORETICAL BACKGROUND	11
1. The RFM model	11
2. K-Means algorithm	11
3. Elbow method	12
4. Silhouette Score	13
CHAPTER 3: DATA PREPARATION	15
1. Introduce Exploratory Data Analysis (EDA)	16
2. Introduce dataset	21
CHAPTER 4: EXPERIMENT	21
1. Cleaning Data	21
2. Calculate R F M values	22
3. Use quantiles method to segmentation	25
4. K-Means Method	28
Come up with a strategy	37
Conclusion	38

Commitment

We commit that this project is done by the group itself under the guidance of lecturer Ho Trung Thanh and the reference sources as cited in the References section.

Acknowledgments

Thank you to the judges who read through this project and the authors who have published helpful research articles for us to refer to and complete this project. We will be very grateful for your comments and suggestions. Thank you very much. Best regards!

PROJECT OVERVIEW

I. Reasons for choosing the topic:

Most people who run an online business must know how important it is to gather insights from the user data generated in their platform. Can understand some usability issues, customer preferences, general buying behavior, and more.

Therefore, if you are building an electronic customer segmentation system, it is essential to save data about orders, customers and their transactions. One of the main purposes of saving that data is to analyze customer behavior and design appropriate strategies to benefit the company.

Customer segmentation allows us to:

Identify target customers: help us locate the right audience, focus on this group of objects and save maximum costs for marketing activities.

Adjust the message to reach the target customer group more quickly: the content, once the target has been identified, sticking to this target group will shorten the campaign execution time, and at the same time bring more effective results, highest results for the entire campaign.

Satisfying a specific need increases conversions: pinpointing the exact desires of a potential customer increases your chances of turning them into a real customer.

Build strong relationships with your customers and earn their loyalty.

Expand your lead file to accelerate the sales cycle.

The above ideas show us that customer segments have great importance and influence on the marketing market.

The science is growing day by day, before advanced technology customer segmentation methods were born, marketers segmented customers using traditional methods such as geolocation-based segmentation, Gender, Age, Ethnicity, Income, Education, and based on customers' past buying behavior. The traditional method will often be less accurate, the time to find the customer segment is quite long and it is also expensive to collect information about each customer. Therefore, technology customer segmentation methods were born, applied to improve accuracy, save time and effort.

II. Objective

- General goals:

Customer segmentation enables us to reduce risk in deciding where, when, how, and to whom products, services, or brands will be marketed.

To increase marketing effectiveness, benefits by focusing specific efforts on specified segments in a way that is appropriate to the characteristics of each of those segments.

- Detail goals:

Accurately identify each customer segment, how much market share each segment occupies in the overall.

Apply appropriate promotions for each segment to retain potential customers.

III. Objects and Scores

- Object:

Top Customers : are the best customers, the ones who bought most recently, the most often and the ones who spend the most on shopping.

High Value Customer: are recent, high frequency customers who have spent a lot on shopping

Medium Value Customer: are recent spenders, medium visit frequency and medium shopping spend

Low Value Customer: are customers who used to visit and purchase quite often, but haven't been visiting recently.

Lost Customer: These are valuable customers who stopped transacting a long time ago.

- *Score:*

Time: 2 Weeks

Space: From December, 14th, 2021 To December, 26th, 2021

IV. Research Methods

- *Qualitative:*

We want to apply the RFM framework to this aggregate dataset and answer the following questions:

1. Is there a correlation between Recency, Frequency, and Monetary Value?
2. How should we define our most valuable customers (MVC)? What percentage of our customers are the most valuable customers?
3. Are we able to create distinguishing segments and design CRM campaigns accordingly to improve customer engagement and/or monetization?

- *Quantitative:*

Interpret raw data from qualitative research by means of statistical analysis (Descriptive statistics and Inferential statistics) to test the proposed hypotheses.

In market research, consider using both qualitative and quantitative methods to get the most valuable results. In order to get the most perfect answer about customer behavior and attitudes and the reasons for those behaviors, the research results can contribute to more accurate management decisions.

CHAPTER 1: RELATED WORK

1. Researches

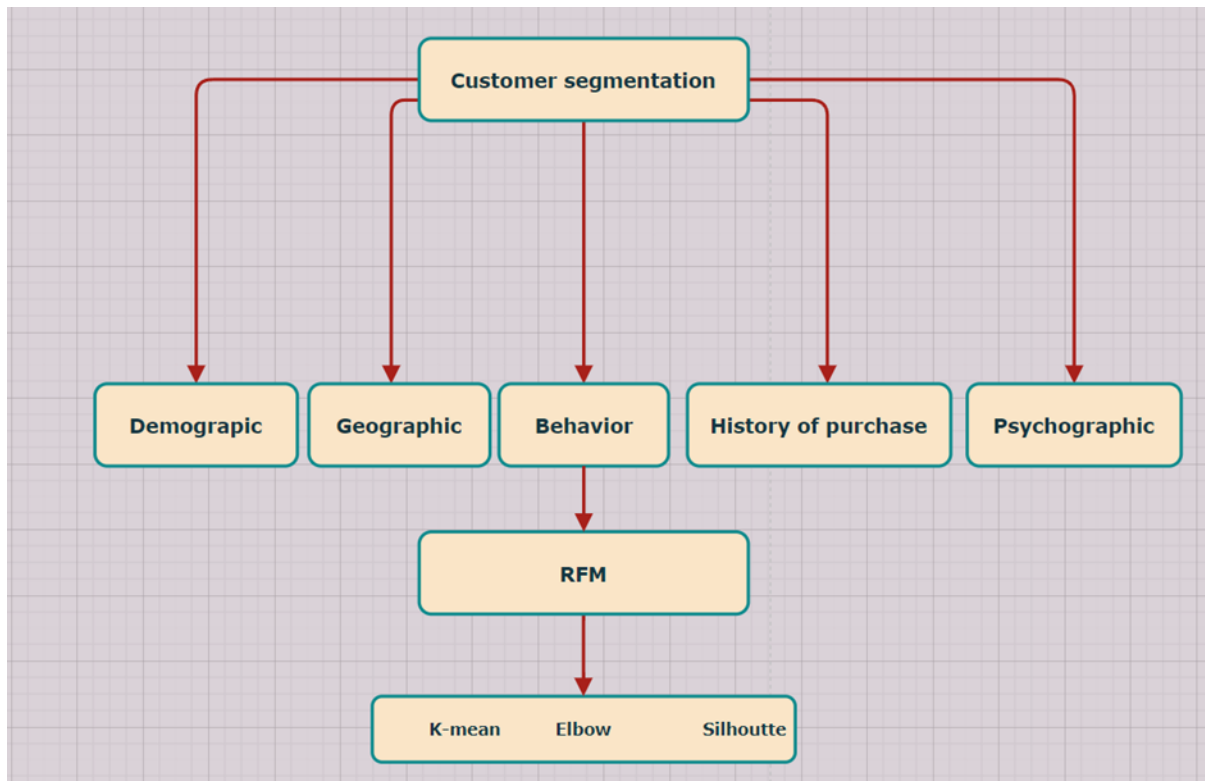
STT	Title	Author's	Year	Method	Data
1	Customer segmentation revisited: The case of the airline industry	Thorsten Teichert, Edlira Shehu, Iwan Von Wartburg	2008	latent class modeling	more than 5800 airline passengers
2	Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing	BİLİŞİM TEKNOLOJİLE Rİ DERGİSİ	2020	RFM Model, Analytical CRM	Customer data
3	Divide and Conquer: segment your customers using RFM analysis	Gabriel Signoretti	2019	- The classical RFM analysis - K-Means clustering	Customer data
4	African Journal of Business Management, 4(19), 4199–4206	J. T. Wei, S. Y. Lin, H. H. Wu	2010	A review of the application of RFM model	Customer data
5	In Proceedings of the 9th World Scientific and	V. Aggelis, D. Christodoulidis	2005	Customer Clustering Using RFM	Customer data

	Engineering Academy and Society International Conference on Computers, Athens, Greece, 2–7, July 14-16			Analysis	
6	Journal of Statistics and Management Systems, 22(6), 1049-1065	J. T. Wei, S. Y. Lin, Y. Z. Yang, H. H. Wu	2019	The application of data mining and RFM model in market segmentation	Customer data
7	CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY	Onur Dogan, Ejder Ayçin, Zeki Atıl Bulut	2018	RFM model, Clustering, K-means clustering	Customer data
8	RFM - Phân khúc khách hàng	<i>RChart</i>	2019	RFM	Is information about spending transactions via credit card

9	Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning	M Aliyev	2020	RFM model, Unsupervised Machine Learning	Customer data
10	“Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study”, Procedia Computer Science, 3, 57–63,.] M. Khajvand, K. Zolfaghar, S. Ashoori, S. Alizadeh,	2011	RFM analysis	Customer data
11	“Customer lifetime value (CLV) measurement based on RFM Model”, Iranian Accounting & Auditing Review, 14(47), 7–20	B. Sohrabi, A. Khanlari	2007	Customer lifetime value (CLV)	Customer data
12	“Approaches to customer segmentation”, Journal of Relationship Marketing, 6(3-4), 9–39	B. Coil, L. Aksoy, T. L. Keiningham	2008	Approaches customer segmentation	Customer data

13	RFM ranking – An effective approach to customer segmentation	AJ Christy	2018	RFM, K_Means	Customer data
14	Phân khúc khách hàng (Customer Segmentation) - Data Analytics for Business	Ngô Văn Khôi	2021	<i>Data Analytics in business</i>	Customer data.
15	Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining	Daqing Chen, Sai Laing Sain & Kun Guo	2012	Data mining	the customer transaction dataset

2. Process Maps and Segmentations



Demographic data refers to the socio-economic information collected about your customers. It involves data around general characteristics such as their age, gender, income, place of residence, etc. Demographic groups can be easily determined and commonly form the basis for more advanced segmentation models.

Geographic segmentation is used to divide customers based on where they live, the makeup of their local area, or the climate. Geographic segmentation is popular among businesses that have physical stores or offices.

Psychographic segmentation takes you closer to your customers' personalities, attitudes, values, and interests. Such data can be easily collected through surveys, however, interpretation can get quite challenging as a customer's interests and values keep changing with time.

Human behavior is one of those complicated things that has always puzzled even the best of marketers. However, with the right data sets at hand, you can look to solve this puzzle piece by piece. With the help of this customer segmentation approach, you can target customers by analyzing their habits and interaction behavior. For instance, if a majority of your customers prefer live chat to seek assistance, hiring more chat support professionals to keep up with the demand will only make sense. And this is the main method we use in our analysis models.

3. RFM Overview

In the customer behavior segmentation method, the group uses RFM models to conduct analysis in which analysis algorithms such as K-mean, Elbow and Silhouette will be used based on the article [14]. This to analyze our customer segmentation model.

These characteristics are analyzed and categorized based on the historical purchasing data of the business. Recency, Frequency, and Monetary (RFM) has been very famous in marketing as a tool to identify a company's best customers by calculating and analyzing their spending habits. RFM analysis weights customers' importance by scoring them in three measurements such as how recently they have made a purchase (Recency), how often they have bought (Frequency), and how much they have spent (Monetary) (Thanh & Son, 2021).

The RFM Model provides several benefits for marketers. Wei, Lin, and Wu [20] remarked that the practical use of the model is customer segmentation with the purpose of identifying valuable customers and improving response rates in direct marketing campaigns. Also, Aggelis and Christodoulakis [21] noted that customer segmentation with the RFM technique identifies customers that are more likely to respond to offers, and helps to estimate the profitability of customers according to their segments. Besides, Wei et al. [22] argued that noted that marketing strategies for businesses might involve customization of products and services according to the RFM segments, and applied the RFM technique for customer segmentation in a veterinary hospital for such purpose.

For using RFM method, the first step is EDA Data to clean dataset, Next step is normalizing data and using Elbow algorithm to select customer clusters, and K-Means algorithm to group customers and continue The next step will be to optimize customer segmentation using the Silhouette algorithm and from there will evaluate and predict customer lifetime value in the future.

Table 1. Dimensions of the RFM Model

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

After preprocessing the data and realizing the difference between the data points, the study will implement the standardization for input data, then using some methods related to K-Means to find out the optimal number of clusters for segmentation.

CHAPTER 2: THEORETICAL BACKGROUND

1. The RFM model

The RFM model is usually used to classify customers and define their behaviors. RFM records the customers' transactions under three factors:

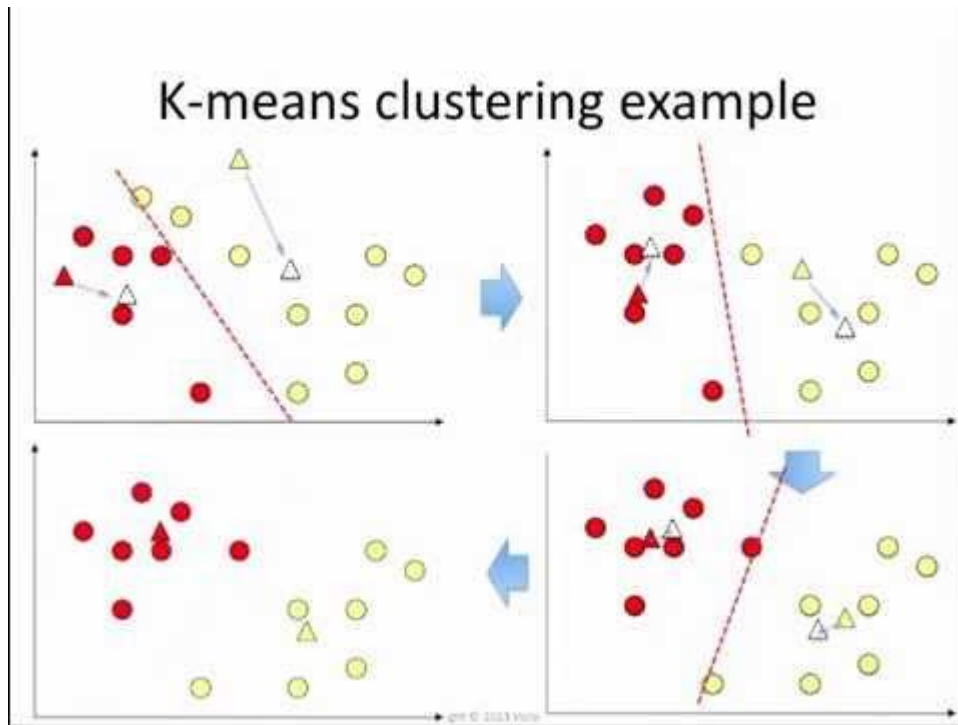
- (1) Recency is the distance between the last purchasing date of that customer and the date of implementing the model;
- (2) Frequency is the total transactions of that customer;
- (3) Monetary is the actual money that the customer had spent on businesses' products or services.

The most well-known clustering methods by RFM are customer quintiles (Miglautsch, 2000) and clustering by K-Means.

2. K-Means algorithm

Clustering using the K-Means algorithm is a method of unsupervised learning used for data analysis. It generates k points as initial centroids randomly, with k is chosen by users. Each point is assigned to the cluster with the closest centroid. Then the centroids are updated by taking the mean of the points of each cluster (Anitha & Patil, 2019; Ismail & Dauda, 2013; Madhu et al., 2010). The data points may move to different clusters after each iterative approach. The chosen centroids are defined when there are no point change clusters or the centroids remain. The algorithm uses mainly Euclidean distance to measure the distance between data points and centroids (Dwivedi et al., 2014). The formula to calculate Euclidean distance between two multi-dimensional data points $X = (x_1, x_2, x_3, \dots, x_m)$ and $Y = (y_1, y_2, y_3, \dots, y_m)$ is described as equation (1):

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_m - y_m)^2} \quad (1)$$

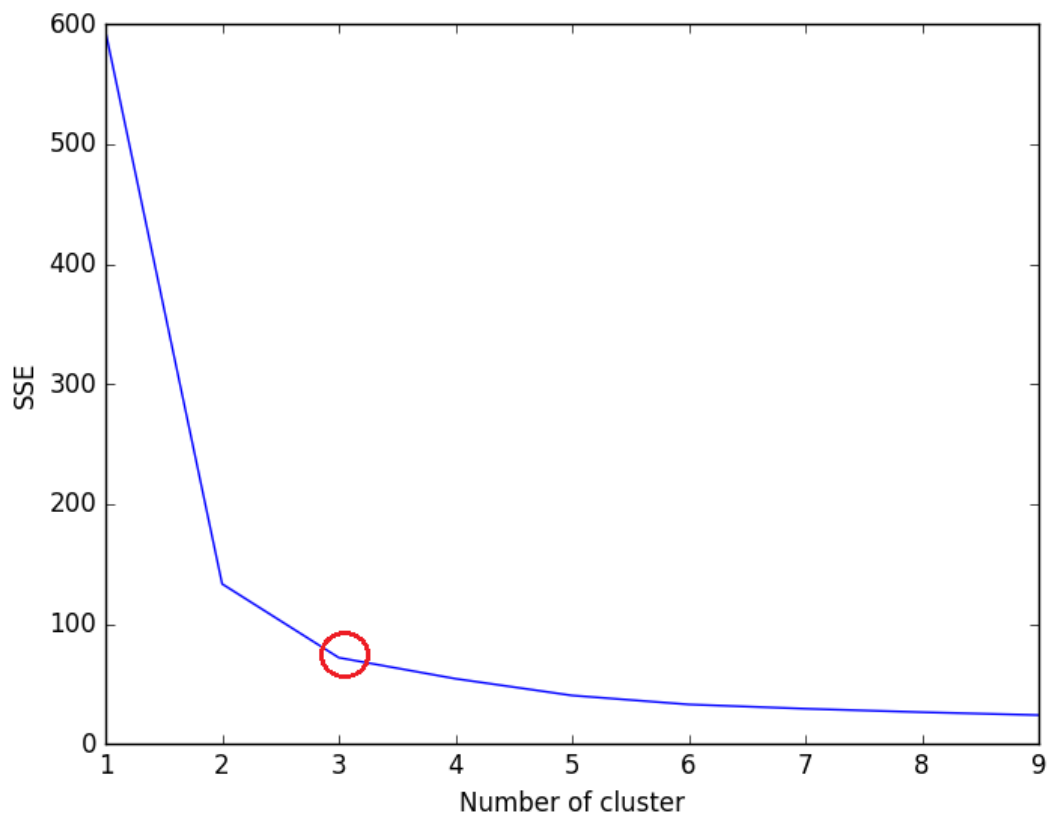


Although K-Means is the most common algorithm to classify clusters, it still has some drawbacks. Because the centroids are first chosen randomly, the results can turn out different for different runs. Besides, defining the right number of clusters is also a tremendous problem to deal with. Thanh and Son (2021) used the Elbow method to find the optimal number of clusters then used the Silhouette method to re-evaluate the results above, while Anitha and Patil (2019) only used the Silhouette score to find the optimal k. These studies pointed out the efficiency of the clustering method in Data Science and also performed the clustering results in RFM analysis and provided customers' different behaviors in specific clusters.

3. Elbow method

The Elbow method is used to determine the number of clusters of a dataset by using the visual technique. The graphic obtained the results from the Sum Squared Error (SSE) calculation, which measures the difference between points in clusters. The more the number of clusters k, the smaller the SSE value will be. If the value of the former cluster and the value of the later cluster draw an angle between them, the cluster at the elbow flexion point will be the chosen cluster or the cluster with the biggest reducing value compared with its former will be chosen (Thanh & Son, 2021; Humaira & Rasyidah, 2020; Nainggolan et al., 2015). The formula of SSE calculation is described as equation (2):

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} d(x_{ij}, m_i)^2 \quad (2)$$



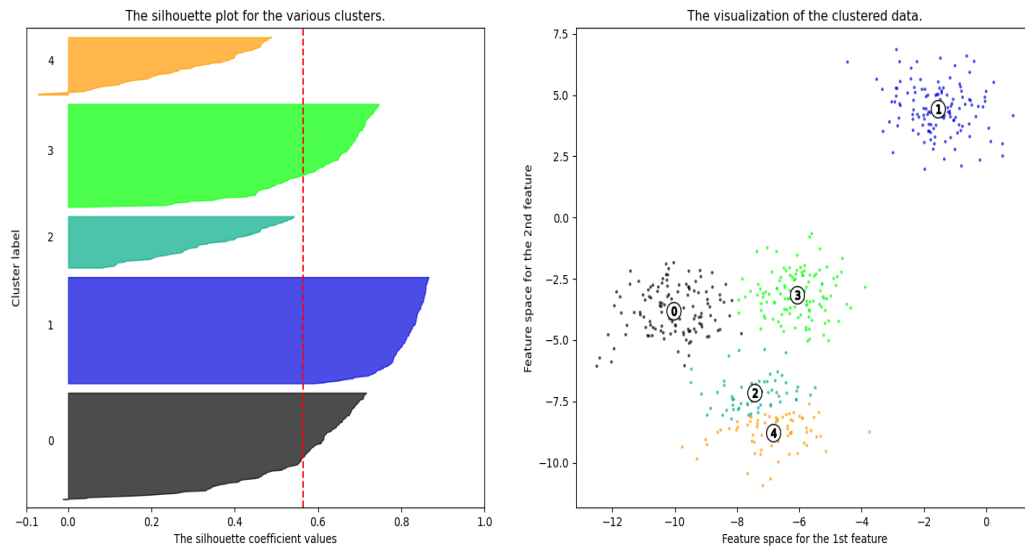
Where m is the centroid of the data point x and k is the number of clusters. The graphic which obtained the values of the SSE calculation for the different number of clusters will perform the visual looks as an elbow arm. The Elbow method is easy to implement and adequately fitted with perplexing, huge data, but its weakness is the user must choose the number of clusters based on experience (Humaira & Rasyidah, 2020).

4. Silhouette Score

Along with the Elbow method, the Silhouette score is also an effective way to see how well each cluster is separated from the others. In the two studies (Anitha & Patil, 2019; Humaira & Rasyidah, 2020), the authors give two different theories about the range values of the Silhouette score. After researching more deeply, the Silhouette score is informed to be in the range $[-1, +1]$, if it is scored near $+1$, the clustering quality performed well, if it is valued at 0 , we can say there is no distinction between the clusters, and if it is near -1 , the clusters were not distributed well (Ogbuabor & Ugwoke, 2018). The formula to calculate Silhouette score is written as equation (3):

$$\text{Silhouette score}_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (3)$$

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



With a is the average intra-cluster distance (the mean distance between i and the data points in the same cluster), and b is the average inter-cluster distance (the mean distance between the i to all the data points outside its cluster). To ensure the selection of the optimal number of clusters from Elbow method because the selection of k in Elbow method is that we rely on eyesight and estimation to choose, So we can use Silhouette method to verify that. . For the Silhouette method, in most cases taking the highest Silhouette Coefficient yields the optimal number of clusters.

Calculation:

The average distance between the observation and all other data points in the same cluster. This distance can also be called the mean distance in the cluster. The average distance is denoted by a .

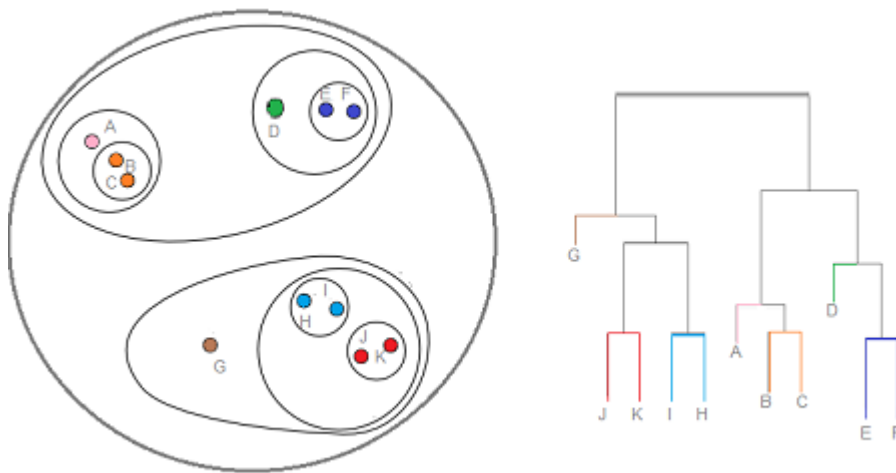
The average distance between the observation and all other data points of the next closest cluster. This distance can also be called the mean nearest cluster distance. The mean distance is denoted by b .

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

A Silhouette score close to 1 means that the clusters are very dense and distinct, and a score of 0 means that the clusters overlap. A score less than 0 means that the data belonging to the clusters may be wrong/inaccurate. ata point i to all the data points outside its cluster).

In our project, we use another clustering method called **Hierarchical Clustering** to compare with K-means clustering.

The result of hierarchical clustering is a collection of nested clusters arranged in a hierarchical tree. It's possible to see it as a dendrogram, a diagram that looks like a tree and records the sequences of mergers and splits.

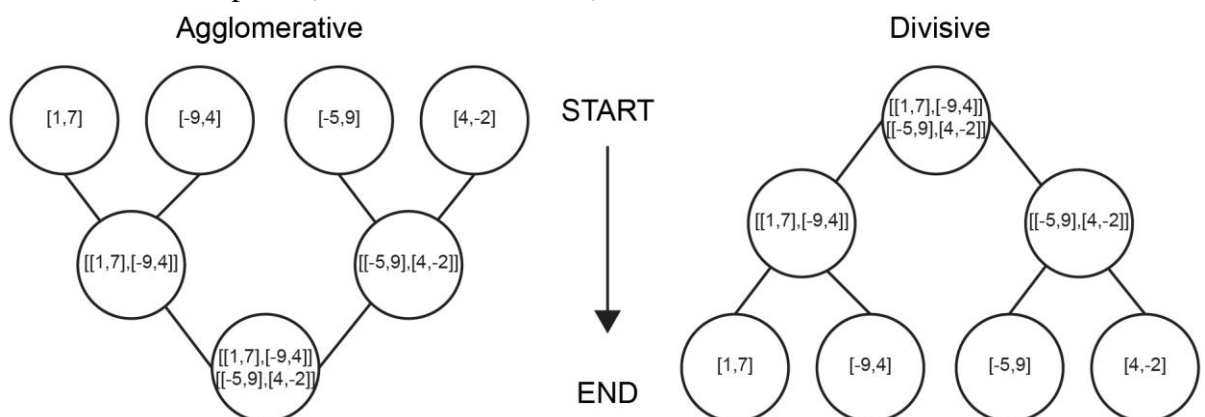


Hierarchical Clustering's Advantages:

- No need to assume a specific number of clusters - by 'cutting' the dendrogram at the appropriate level, any desired number of clusters can be obtained.
- It's possible that they match to relevant taxonomies. Biological sciences (e.g., animal kingdom, phylogeny reconstruction, etc.) are an example.

Two main types of hierarchical clustering:

- Agglomerative: Start with the points as individual clusters. At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive: Start with one, all-inclusive cluster. At each step, split a cluster until each cluster contains a point (or there are k clusters)



At one point or another, a variety of agglomerative hierarchical clustering techniques have been developed. Such hierarchical algorithms can be divided into two kinds of methods for ease of use. The single, full, weighted, and unweighted average linkage methods are the first group of linkage methods.

The second set of hierarchical clustering methods allows the cluster centers to be specified (as an average or a weighted average of the cluster's member vectors). The centroid, median, and minimum variance approaches are among them.

If points (objects) i and j are grouped into cluster ij , the new dissimilarity between the cluster and all other points (objects or clusters) must be specified. The formula is as follows: (17)

$$d(i \cup j, k) = \alpha_i d(i, k) + \alpha_j d(j, k) + \beta d(i, j) + \gamma |d(i, k) - d(j, k)| \quad [17]$$

CHAPTER 3: DATA PREPARATION

1. Introduce Exploratory Data Analysis (EDA)

In this chapter we use an example dataset to EDA and demonstrate parts of EDA technique that used in chapter 4 to evaluate customer segmentation.

Example dataset:

```
[ ] from google.colab import files
    uploaded = files.upload()

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving Kiên.xlsx to Kiên.xlsx

[ ] df=pd.read_excel('/content/Kiên.xlsx')

[ ] df.head()
```



	CustomerID	TerritoryID	Name	CountryRegionCode	ProductDescriptionID	Name.1	Description
0	1	1.0	Northwest	US	3.0	LL Bottom Bracket	Chromoly steel.
1	2	1.0	Northwest	US	4.0	ML Bottom Bracket	Aluminum alloy cups; large diameter spindle.
2	3	4.0	Southwest	US	5.0	HL Bottom Bracket	Aluminum alloy cups and a hollow axle.
3	4	4.0	Southwest	US	8.0	Mountain-500	Suitable for any type of riding, on or off-roa...
4	5	4.0	Southwest	US	64.0	Mountain-400-W	This bike delivers a high-level of performance...

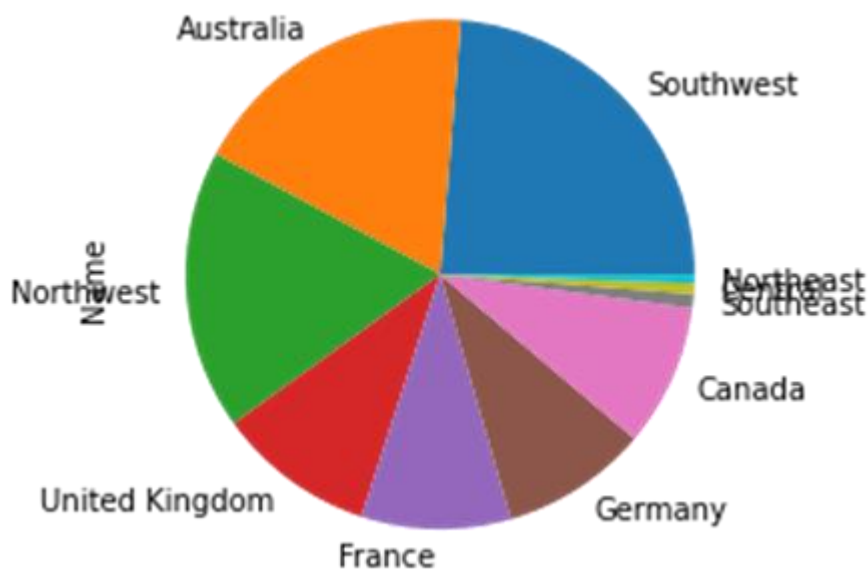
Import Necessary libraries

df.describe()

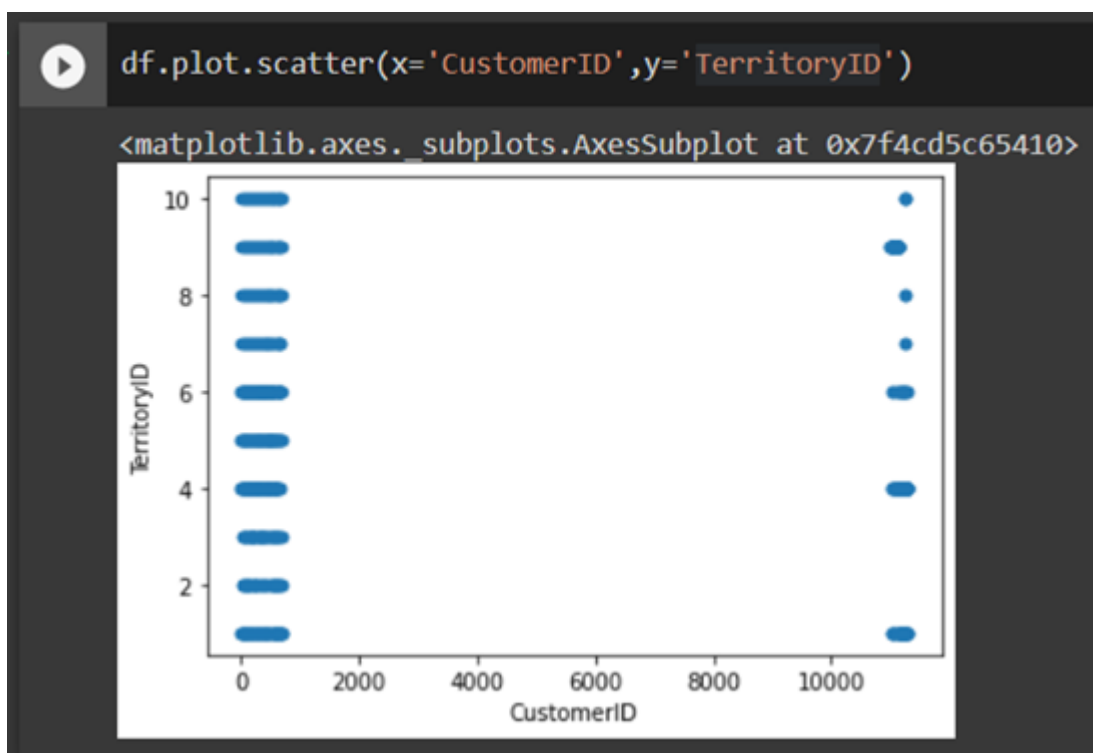
	CustomerID	TerritoryID	ProductDescriptionID
count	19820.000000	1000.000000	762.000000
mean	19844.277094	4.981000	1542.582677
std	6581.785914	2.763743	388.528836
min	1.000000	1.000000	3.000000
25%	15253.750000	3.000000	1429.250000
50%	20208.500000	4.000000	1620.500000
75%	25163.250000	7.000000	1814.750000
max	30118.000000	10.000000	2010.000000

Let's analyze the Territory Variable from the dataset. Since we've already seen a bar plot, let's see how a Pie Chart looks like.

	df.Name.value_counts(normalize=True) df.Name.value_counts(normalize=True).plot.pie()	
Southwest	0.236932	
Australia	0.184914	
Northwest	0.177598	
United Kingdom	0.100454	
France	0.095055	
Germany	0.093441	
Canada	0.090363	
Southeast	0.008880	
Central	0.006660	
Northeast	0.005701	



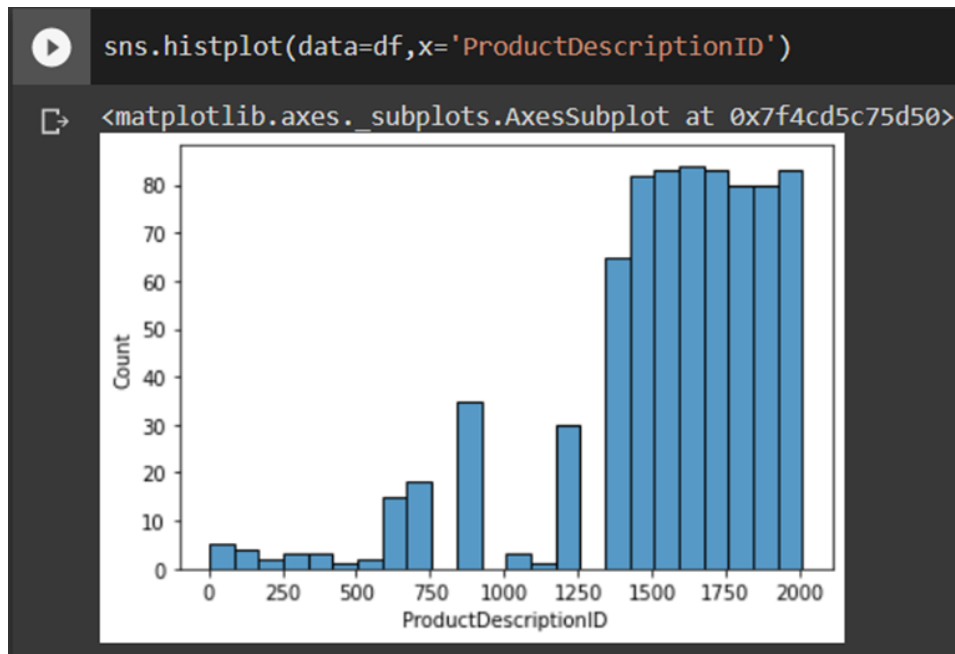
- As we can see that The Pie Chart shows the number of customers from the Southwest are the highest, at about 23.7%. The proportion of customers from Australia and the Northwest are about 18.5% and 17.8%, respectively. Meanwhile, Customers from France, Germany and Canada are approximately the same, accounting for about 9.3% to 10%. At the end, The Southeast, the Central and the Northeast have the fewest customers, occupying about 0.88%, 0.67% and 0.57%, respectively.



- As we can see in the scatter plot, most of the customers who have CustomerID < 1000 belong to the same number of customers of each TerritoryID. However, there are no

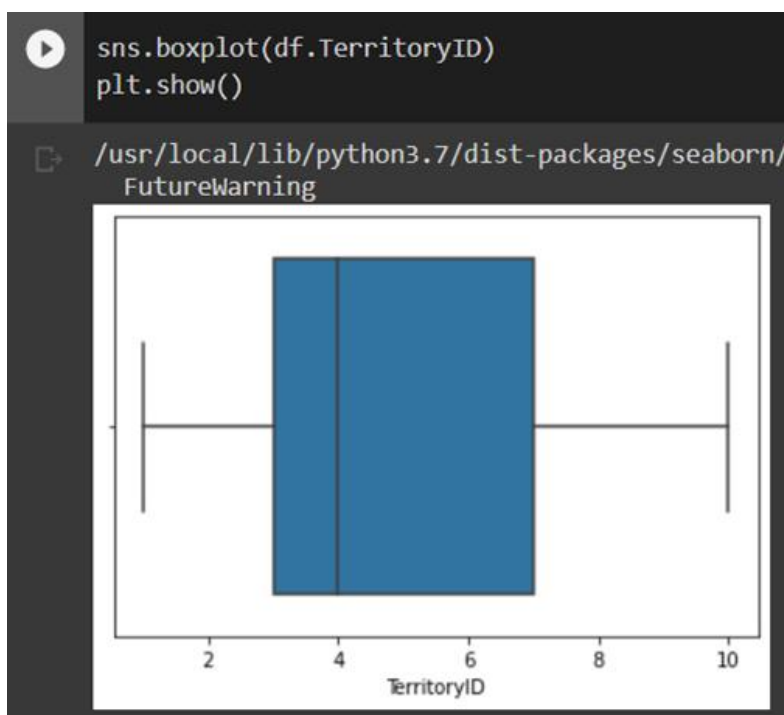
customers who have CustomerID from 1000 to 11000, while the number of CustomerID > 11000 are not the same as each TerritoryID.

Let's use historical plot to understand by example datas:



- As we can see from the plot that most of ProductDescriptionID are > 1400

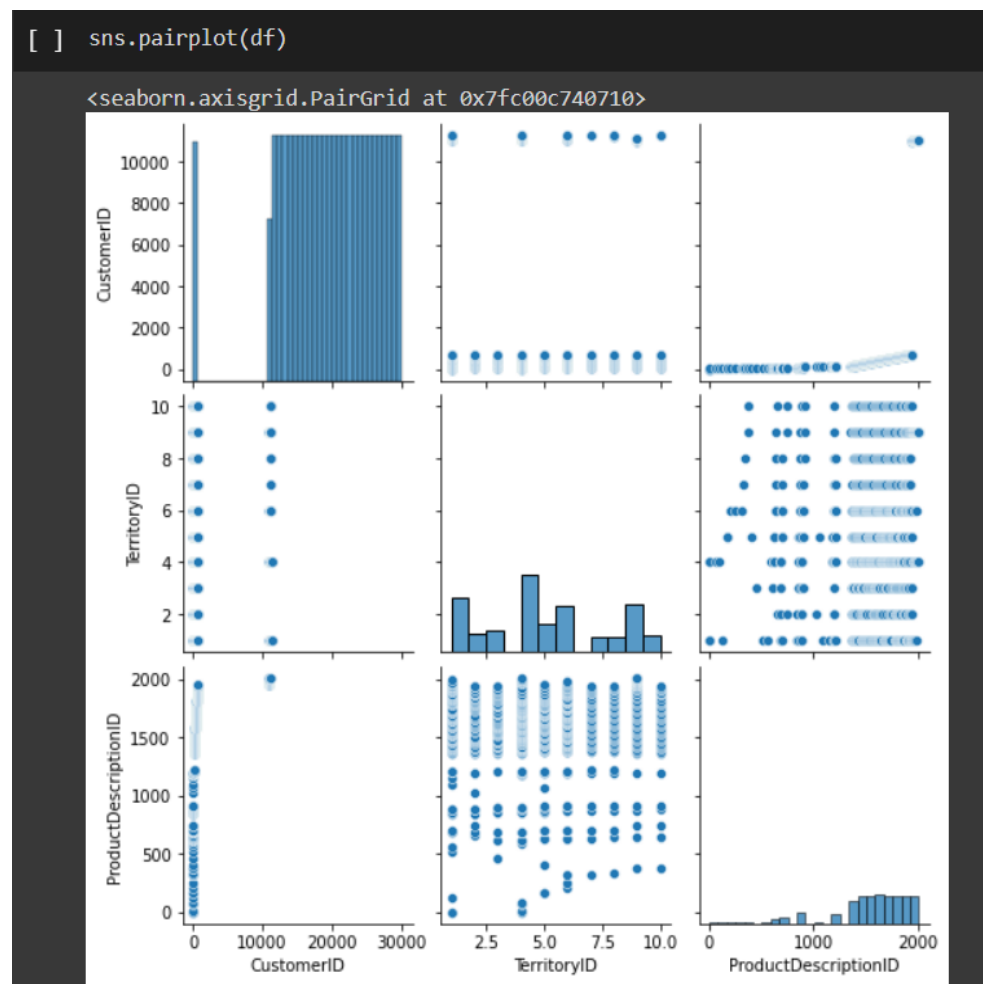
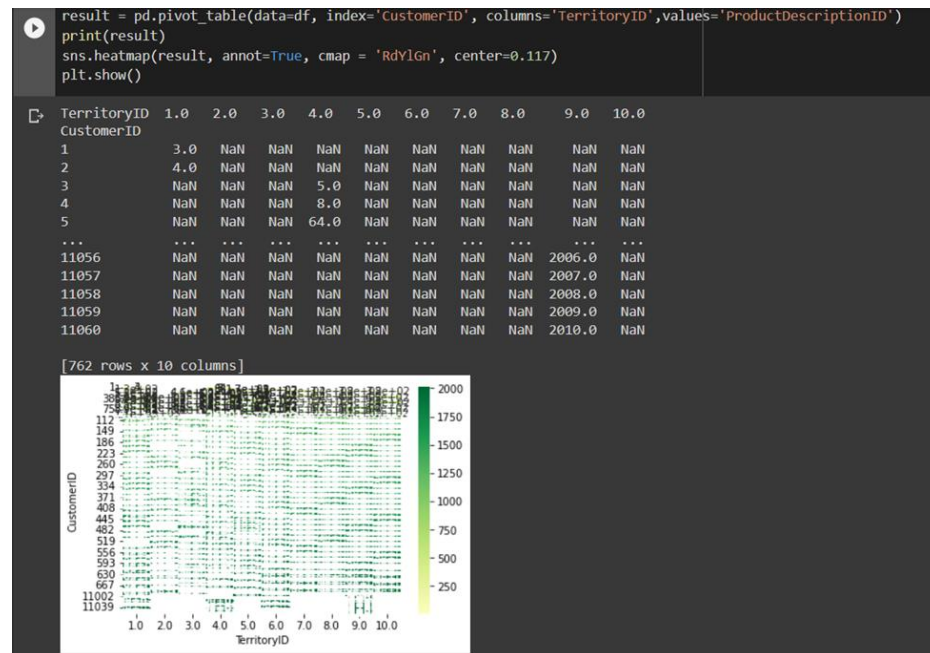
We can also use boxplot to evaluate the dataset



- . As we can see from the boxplot, the minimum value is 1 and the maximum is 10. The median is 4. And the Interquartile range is between 3 and 7.

Let's see how "TerritoryID", "CustomerID", "ProductDescriptionID" vary with each other.

First, we'll create a pivot table with the three columns and after that, we'll create a heatmap.



- Pair plot also helps visualize data with many charts of CustomerID, TerritoryID and ProductDescription.

2. Introduce dataset:

CustomerID	OrderDate	Quantity	UnitPrice	SaleOrderID	SalesOrderNumber
29825	2011-05-31 00:00:00.000	1	2024.994	43659	SO43659
29825	2011-05-31 00:00:00.000	3	2024.994	43659	SO43659
29825	2011-05-31 00:00:00.000	1	2024.994	43659	SO43659
29825	2011-05-31 00:00:00.000	1	2039.994	43659	SO43659
29825	2011-05-31 00:00:00.000	1	2039.994	43659	SO43659
29825	2011-05-31 00:00:00.000	2	2039.994	43659	SO43659
29825	2011-05-31 00:00:00.000	1	2039.994	43659	SO43659
29825	2011-05-31 00:00:00.000	3	28.8404	43659	SO43659
29825	2011-05-31 00:00:00.000	1	28.8404	43659	SO43659
29825	2011-05-31 00:00:00.000	6	5.7	43659	SO43659

In our project we use this dataset to analyze in chapter 4. By this dataset we will combine RFM (Recency, Frequency, Monetary) and Machine Learning (K-means) to analyze customer segmentation.

To export dataset, we use these codes in Sql server take data from Adventureworks:

“Select SOH.CustomerID, SOH.SalesOrderID, SOH.OrderDate, SOD.OrderQty AS 'Quantity', SOD.UnitPrice

From Sales.SalesOrderHeader SOH

Full outer join Sales.SalesOrderDetail SOD on SOH.SalesOrderID = SOD.SalesOrderID”.

CHAPTER 4: EXPERIMENT

1. Cleaning Data

```
# Import thư viện
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Import Necessary libraries

```
[ ] df.describe()
```

	CustomerID	Quantity	UnitPrice	SaleOrderID
count	121317.000000	121317.000000	121317.000000	121317.000000
mean	24345.630505	2.266080	465.093496	57827.363782
std	6689.110387	2.491323	751.885081	9009.147902
min	11000.000000	1.000000	1.328200	43659.000000
25%	18177.000000	1.000000	21.490000	49884.000000
50%	29485.000000	1.000000	49.990000	57029.000000
75%	29795.000000	3.000000	602.346000	65490.000000
max	30118.000000	44.000000	3578.270000	75123.000000

Brief evaluation of the data set


```
df_new=df[df["CustomerID"].notna()]
#df_new = df_new.sample(10000, random_state=42)

df_new.head(10)
```

	CustomerID	OrderDate	Quantity	UnitPrice	SaleOrderID	SalesOrderNumber
0	29825	2011-05-31 00:00:00.000	1	2024.9940	43659	SO43659
1	29825	2011-05-31 00:00:00.000	3	2024.9940	43659	SO43659
2	29825	2011-05-31 00:00:00.000	1	2024.9940	43659	SO43659
3	29825	2011-05-31 00:00:00.000	1	2039.9940	43659	SO43659
4	29825	2011-05-31 00:00:00.000	1	2039.9940	43659	SO43659
5	29825	2011-05-31 00:00:00.000	2	2039.9940	43659	SO43659
6	29825	2011-05-31 00:00:00.000	1	2039.9940	43659	SO43659
7	29825	2011-05-31 00:00:00.000	3	28.8404	43659	SO43659
8	29825	2011-05-31 00:00:00.000	1	28.8404	43659	SO43659
9	29825	2011-05-31 00:00:00.000	6	5.7000	43659	SO43659

```
df_new.shape
```

```
(121317, 6)
```

2. Calculate R F M values

Here we are calculating:

- The Recency for customers who had made a purchase with a company.
- The frequency of frequent transactions of the customer in ordering/buying some product from the company.
- The monetary value of customers spent on purchasing products from the company.

And Then, we are merging all the data frame columns in a single entity using the merge function to display the recency, frequency, monetary value.

```
df_new["OrderDate"]=pd.to_datetime(df_new["OrderDate"], format='%Y-%m-%d %H:%M:%S')

# Lay ngay lon nhat trong InvoiceDate + 1
import datetime
current_date = max(df_new['OrderDate']) + datetime.timedelta(days=1)


# ----- Tinh M - MoneytaryValue
df_new['TotalPay'] = df_new['Quantity'] * df_new['UnitPrice']

# Group by CustomerID de tinh R, F, M

df_customers = df_new.groupby(['CustomerID']).agg(
    {
        'OrderDate': lambda x: (current_date- x.max()).days,
        'SalesOrderNumber': 'nunique',
        'TotalPay': 'sum'
    }
)


df_customers.head()
```

Now, we have a dataframe like this picture below.




	OrderDate	SalesOrderNumber	TotalPay
CustomerID			
11000	271	3	8248.99
11001	50	3	6383.88
11002	340	3	8114.04
11003	264	3	8139.29
11004	273	3	8196.01

And, change old name columns into Recency, Frequency and Monetary Value



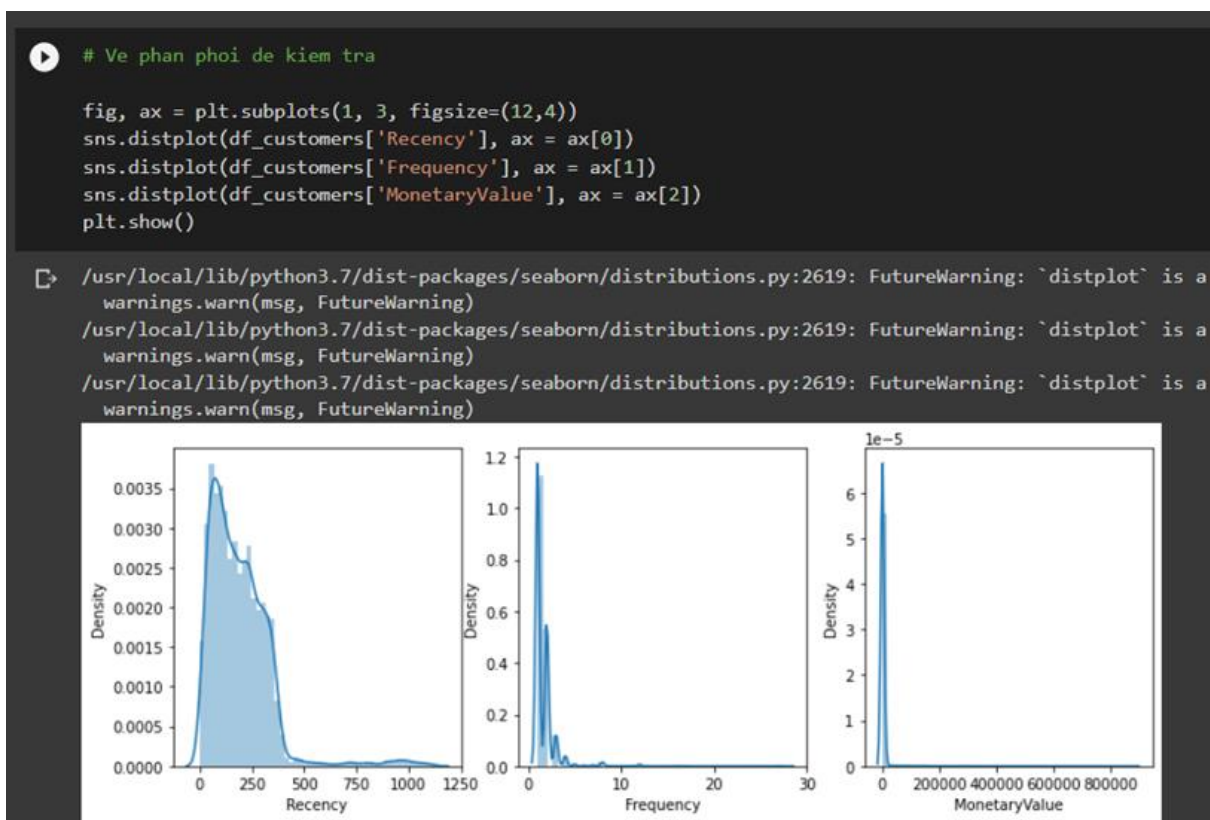
```
df_customers.describe()
```



	OrderDate	SalesOrderNumber	TotalPay
count	19119.000000	19119.000000	19119.000000
mean	191.267483	1.645745	5772.994891
std	150.423605	1.457054	39030.612897
min	1.000000	1.000000	1.374000
25%	86.000000	1.000000	54.980000
50%	166.000000	1.000000	548.980000
75%	264.000000	2.000000	2822.386050
max	1127.000000	28.000000	882276.496600

	Recency	Frequency	MonetaryValue
CustomerID			
11000	271	3	8248.99
11001	50	3	6383.88
11002	340	3	8114.04
11003	264	3	8139.29
11004	273	3	8196.01
11005	272	3	8121.33

- Data of Recency is skewed to the right.
- Data of Frequency and Monetary not only is it skewed to the right, but it's also pointed at the top.
- So, we convert the above values into the same unit with the standard score distribution method, also known as Z-score.



df_customers_t=df_customers.apply(stats.zscore)
df_customers_t

	CustomerID	Recency	Frequency	Monetary
0	-1.731960	0.530067	0.929472	0.064526
1	-1.731779	-0.939156	0.929472	0.016456
2	-1.731598	0.988784	0.929472	0.061048
3	-1.731417	0.483531	0.929472	0.061699
4	-1.731235	0.543363	0.929472	0.063161
...
19114	1.731235	-0.460495	4.361143	0.152260
19115	1.731417	-0.659937	4.361143	0.081758
19116	1.731598	-0.460495	1.615806	4.674530
19117	1.731779	-0.659937	7.106481	20.902664
19118	1.731960	-0.866027	4.361143	7.031639

19119 rows × 4 columns

3. Use quantiles method to segmentation

- Divide the Recency class with 1 being recent and 4 being furthest. For Frequency and Monetary, 4 is the lowest and 1 is the highest.

```

# Arguments (x = value, p = recency, monetary_value, frequency, k = quantiles dict)
def R_Class(x,p,d):
    if x <= d[p][0.25]:
        return 1
    elif x <= d[p][0.50]:
        return 2
    elif x <= d[p][0.75]:
        return 3
    else:
        return 4

# Arguments (x = value, p = recency, monetary_value, frequency, k = quantiles dict)
def FM_Class(x,p,d):
    if x <= d[p][0.25]:
        return 4
    elif x <= d[p][0.50]:
        return 3
    elif x <= d[p][0.75]:
        return 2
    else:
        return 1

quantiles = pd.DataFrame()
quantiles = df_customers.quantile(q=[0.25, 0.5, 0.75])
# convert quantiles to a dict, easier to use.
quantiles = quantiles.to_dict()
RFM_Seg = df_customers.copy()
RFM_Seg['R_Quartile'] = RFM_Seg['Recency'].apply(R_Class, args=('Recency',quantiles,))
RFM_Seg['F_Quartile'] = RFM_Seg['Frequency'].apply(FM_Class, args=('Frequency',quantiles,))
RFM_Seg['M_Quartile'] = RFM_Seg['Monetary'].apply(FM_Class, args=('Monetary',quantiles,))
RFM_Seg['RFMClass'] = RFM_Seg.R_Quartile.map(str) \
    + RFM_Seg.F_Quartile.map(str) \
    + RFM_Seg.M_Quartile.map(str)

```

- Show the Data Frame with RFM class which helps us to know which group each customer will be

RFM_Seg

	CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMClass
0	11000	271	3	8248.990000	4	1	1	411
1	11001	50	3	6383.880000	1	1	1	111
2	11002	340	3	8114.040000	4	1	1	411
3	11003	264	3	8139.290000	3	1	1	311
4	11004	273	3	8196.010000	4	1	1	411
...
19114	30114	122	8	11652.991100	2	1	1	211
19115	30115	92	8	8917.559400	2	1	1	211
19116	30116	122	4	187114.201050	2	1	1	211
19117	30117	92	12	816755.576276	2	1	1	211
19118	30118	61	8	278568.569942	1	1	1	111

19119 rows × 8 columns

- Next, assign names to each Customer's Class that we evaluate through the R F M.


```

segm_map = {
    r'111': 'Best Customers',
    r'444': 'Lost Cheap Customers',
    r'411': 'Lost Customers',
    r'311': 'Almost Lost Customers',
    r'[1-4]1[1-4]': 'Loyal Customers',
    r'[1-4][1-4]1': 'Big Spenders',
    r'144': 'New Customer',
    r'[2-3]44': 'Unsteady Customers',
    r'[1-4]4[3-4]': 'Lost Customers'
}

RFM_Seg['Segment'] = RFM_Seg['R_Quartile'].map(str) + RFM_Seg['F_Quartile'].map(str) + RFM_Seg['M_Quartile'].map(str)
RFM_Seg['Segment'] = RFM_Seg['Segment'].replace(segm_map, regex=True)
RFM_Seg

```

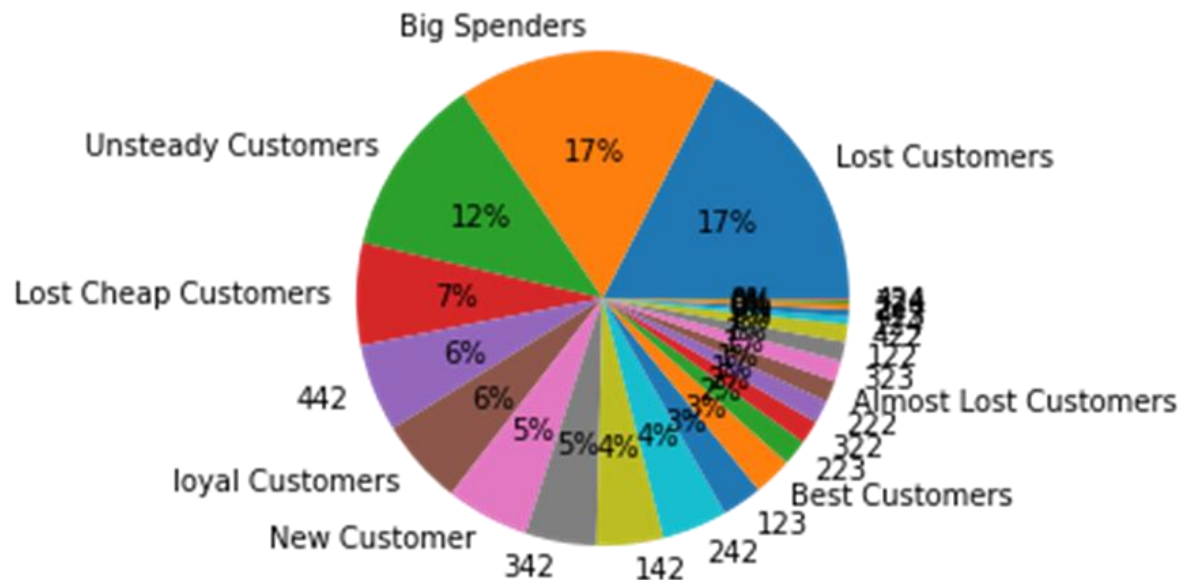
- Now, we have Table of Customer with customer groups have been clustered and now subclass with names

RFM_Seg

	CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMClass
0	11000	271	3	8248.990000	4	1	1	411
1	11001	50	3	6383.880000	1	1	1	111
2	11002	340	3	8114.040000	4	1	1	411
3	11003	264	3	8139.290000	3	1	1	311
4	11004	273	3	8196.010000	4	1	1	411
...
19114	30114	122	8	11652.991100	2	1	1	211
19115	30115	92	8	8917.559400	2	1	1	211
19116	30116	122	4	187114.201050	2	1	1	211
19117	30117	92	12	816755.576276	2	1	1	211
19118	30118	61	8	278568.569942	1	1	1	111

19119 rows × 8 columns

- But in this way, we face the problem because this method is quite manual when we have to manually label each set of numbers such as missing 342, 142,... and to do it all is very time-consuming, so this method seems to only be used to evaluate directly. can't evaluate overall.



So now, we have another method capable of self-selecting the number of clusters, evaluating an overview of the customer group, that is K Means, combined with Elbow and other parameters.

4. K-Means Method

```
# Chosen số cụm bằng Elbow
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

sse = {}
range_n_clusters = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]

for num_clusters in range_n_clusters:

    # initialise kmeans
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(df_customers_t)
    sse[num_clusters] = kmeans.inertia_
    cluster_labels = kmeans.labels_

    # silhouette score
    silhouette_avg = silhouette_score(df_customers_t, cluster_labels)
    print("For n_clusters={0}, the silhouette score is {1}".format(num_clusters, silhouette_avg))

plt.title('The Elbow Method')
plt.xlabel('k')
plt.ylabel('SSE')
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.show()
```

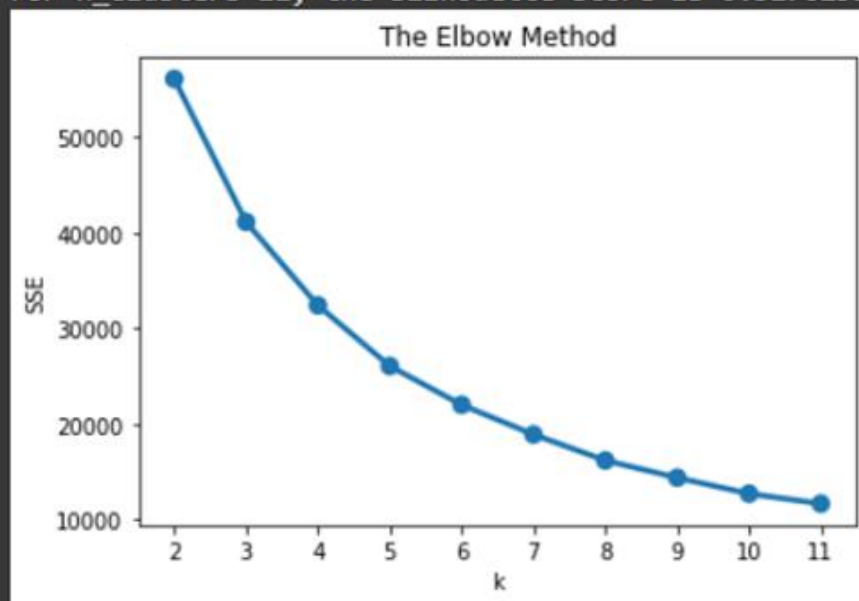
Draw Elbow by Sklearn library

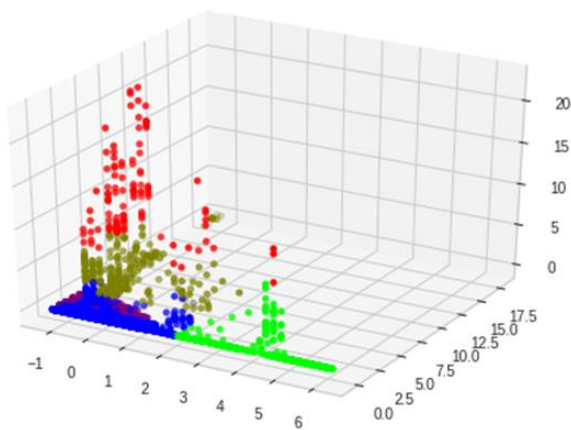
The elbow bend with K = 5 (points between 4 and 6 on the horizontal axis) is the proper number of clusters for the SSE curve resembling the elbow. As the number of clusters grows,

the value of the SSE curve declines fairly uniformly, implying that the difference between cluster points remains almost static. In other words, after the "elbow" point, the slope of the SSE curve steadily decreases, and this point on the SEE curve is regarded as the ideal location for the input parameter in the K-means clustering algorithm. Additionally, the Silhouette score at 5 is higher than the rest.

So we choose $k=5$;

```
↳ For n_clusters=2, the silhouette score is 0.8105734602847686  
For n_clusters=3, the silhouette score is 0.34560687578549604  
For n_clusters=4, the silhouette score is 0.3744006713296172  
For n_clusters=5, the silhouette score is 0.38370581591431135  
For n_clusters=6, the silhouette score is 0.34345605478821695  
For n_clusters=7, the silhouette score is 0.34577523057869336  
For n_clusters=8, the silhouette score is 0.3259682794427628  
For n_clusters=9, the silhouette score is 0.3234892315672541  
For n_clusters=10, the silhouette score is 0.324513901624545  
For n_clusters=11, the silhouette score is 0.3176255141078553
```





- We can see 5 clusters divided quite clearly, only a few values are skewed, called outline.

```

model = KMeans(n_clusters=5, random_state=42)
model.fit(df_customers_t)

KMeans(n_clusters=5, random_state=42)

df_customers['Cluster'] = model.labels_
df_customers.head()

```

	CustomerID	Recency	Frequency	Monetary	Cluster
0	11000	271	3	8248.99	1
1	11001	50	3	6383.88	1
2	11002	340	3	8114.04	1
3	11003	264	3	8139.29	1
4	11004	273	3	8196.01	1

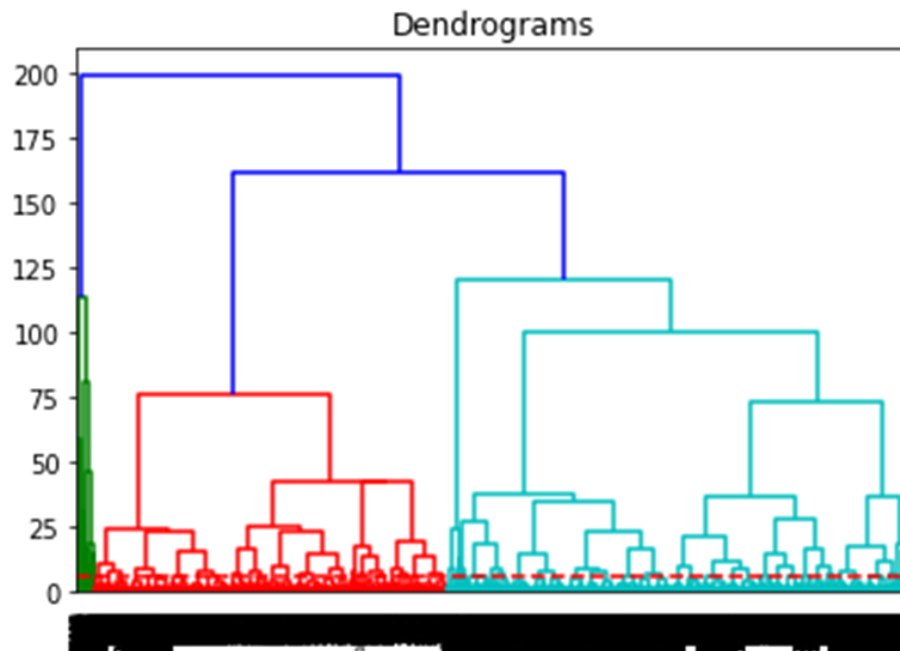
Then, we use the K-means model of sklearn to divide Customer into 5 segments.

- Next, Group by 'Cluster'

		Recency	Frequency	Monetary
Cluster				
0		844.65	1.32	10200.48
1		160.43	1.79	2154.58
2		189.60	1.21	1278.81
3		116.03	8.85	76174.82
4		147.78	8.78	435822.25

Cluster	CustomerID count	Recency mean	Frequ mean	mean	count
0	477	844.651992	1.322851	10200.480503	477
1	8702	160.431395	1.793151	2154.575459	8702
2	9525	189.603255	1.214173	1278.812765	9525
3	297	116.030303	8.851852	76174.815825	297
4	118	147.779661	8.779661	435822.248475	118

- It can be seen that the 4 clusters have the fewest clients but have the most transactions, latest purchases, and generate the most money for the company.
- 2 groups of Cluster 1 and 2 have the most significant number of customers but have a low average value.
- Cluster 0 looks like they haven't come back in a long time, even though their average monetary value is much higher than that of Cluster 1 and 2. But still much lower than Cluster 3, the average economic value average to \$76,174, despite double the number of customers, the Cluster group is still marginally better with an average monetary value of \$435,822.



- After adding the Hierarchical method to once again be sure that k=5 is the number of clusters we need then

```

df_customers['R_rank'] = df_customers['Recency'].rank(ascending=False)
df_customers['F_rank'] = df_customers['Frequency'].rank(ascending=True)
df_customers['M_rank'] = df_customers['Monetary'].rank(ascending=True)

# normalizing the rank of the customers
df_customers['R_rank_norm'] = (df_customers['R_rank']/df_customers['R_rank'].max())*100
df_customers['F_rank_norm'] = (df_customers['F_rank']/df_customers['F_rank'].max())*100
df_customers['M_rank_norm'] = (df_customers['M_rank']/df_customers['M_rank'].max())*100

df_customers.drop(columns=['R_rank', 'F_rank', 'M_rank'], inplace=True)
df_customers.head()

```

	CustomerID	Recency	Frequency	Monetary	Cluster	R_rank_norm	F_rank_norm	M_rank_norm
0	11000	271	3	8248.99	1	23.691720	92.708633	92.706208
1	11001	50	3	6383.88	1	88.337391	92.708633	92.706208
2	11002	340	3	8114.04	1	10.021205	92.708633	92.706208
3	11003	264	3	8139.29	1	25.094898	92.708633	92.706208
4	11004	273	3	8196.01	1	23.225739	92.708633	92.706208

This step, we normalized the rank of the customers

```
df_customers['RFM_Score'] = 0.15*df_customers['R_rank_norm']+0.28 * \
df_customers['F_rank_norm']+0.57*df_customers['M_rank_norm']
df_customers['RFM_Score'] *= 0.05
df_customers = df_customers.round(2)
df_customers[['RFM_Score']].head(7)
```

	RFM_Score
0	4.12
1	4.60
2	4.02
3	4.13
4	4.11
5	4.12
6	4.13

RFM score is calculated based upon recency, frequency, monetary value and normalized ranks. Based upon this score we divide our customers. Here we rate them on a scale of 5. Formula used for calculating rfm score is : $0.15 \times \text{Recency score} + 0.28 \times \text{Frequency score} + 0.57 \times \text{Monetary score}$. [11]

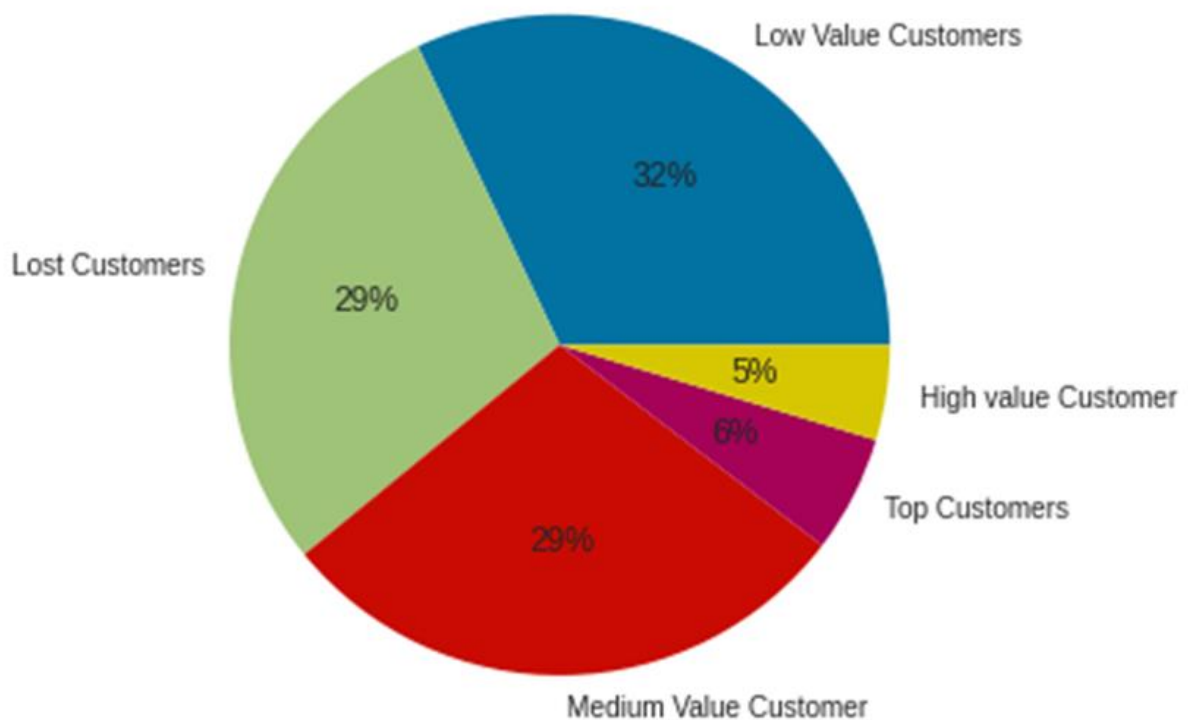
Way to assign names to each customer segment rfm score :

- >4.5 : Top Customer
- $4.5 > \text{rfm score} > 4$: High Value Customer
- $4 > \text{rfm score} > 3$: Medium value customer
- $3 > \text{rfm score} > 1.6$: Low-value customer
- $\text{rfm score} < 1.6$:Lost Customer

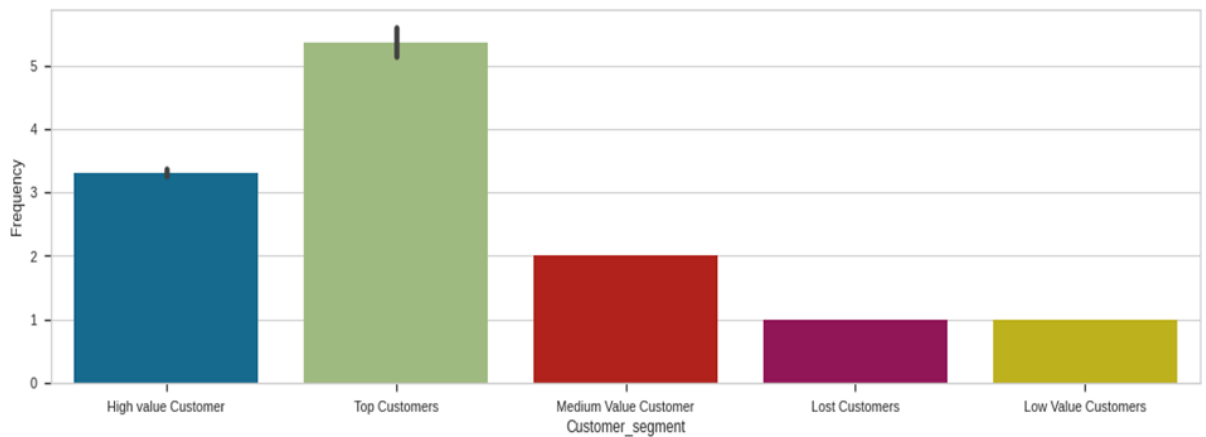
```

df_customers["Customer_segment"] = np.where(df_customers['RFM_Score'] > 4.5, "Top Customers", (np.where(df_customers['RFM_Score'] > 4,
"High value Customer",
(np.where(
df_customers['RFM_Score'] > 3,
"Medium Value Customer",
np.where(df_customers['RFM_Score'] > 1.6,
'Low Value Customers', 'Lost Customers'))))))))
df_customers[['RFM_Score', 'Customer_segment']].head(20)

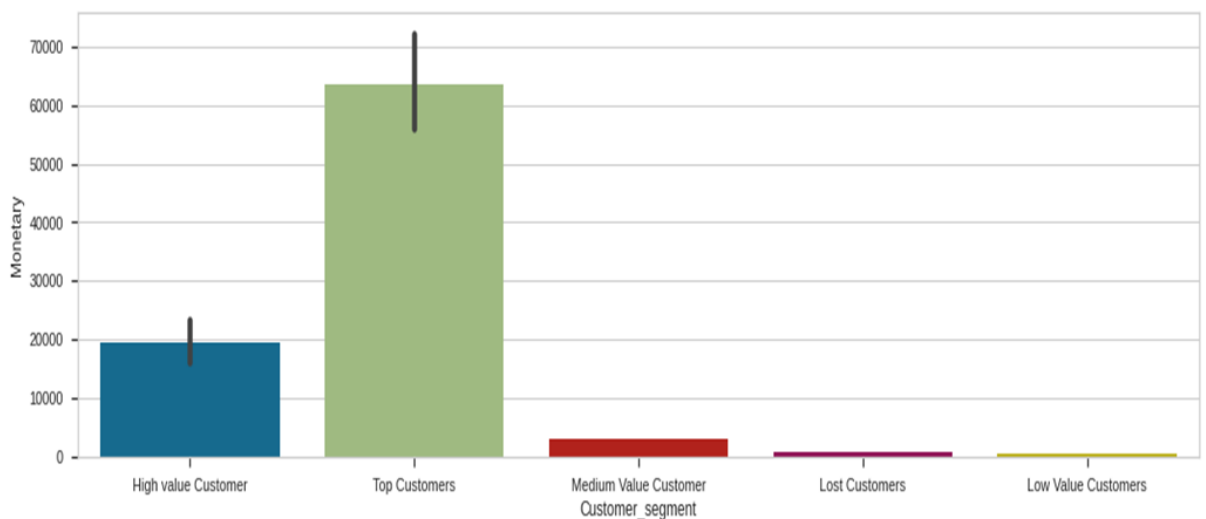
```



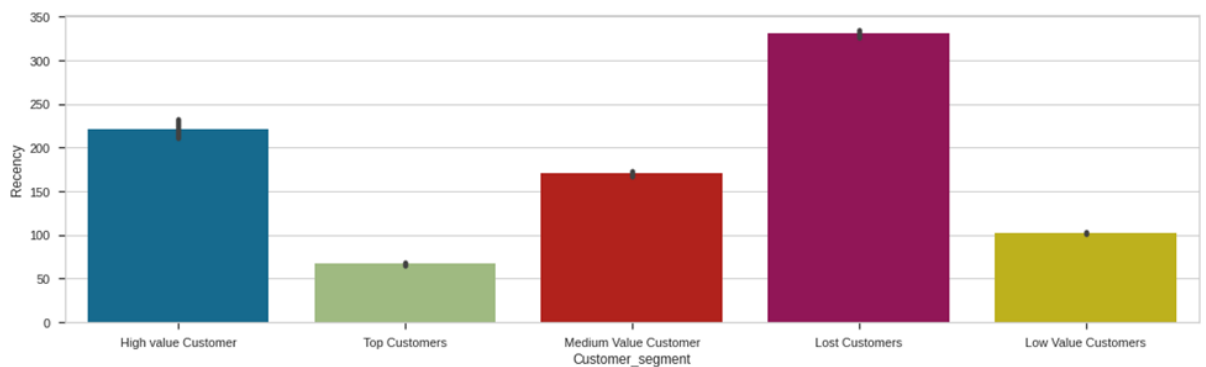
- As we can see that The Pie Chart shows Low Value Customers occupies the largest area with 32%, Top, High value Customers accounted for 6% and 5% respectively. At the same time, Medium Value Customers and Lost Customers together accounted for 29%.



- Top Customers has the highest purchasing frequency, followed by High Customers, Medium Customers and followed by Lost, Low Value Customers



- Top Customers spent the most total amount, followed by High Customers, Medium Customers and followed by Lost, Low Value Customers



- Top Customers have the most time to shop, for Lost Customers it seems that they have not come for a long time, followed by High Value Customers, even though they have high profit value for the business. They didn't come back for a long time. Low

Value Customers they also often come to us but only buy a few things of little value or come to have entertainment space

- >4.5 : Top Customer

```
df_filter=df_customers.loc[(df_customers['RFM_Score'] > 4.5) ]  
df_filter[['Cluster','Recency','Recency','Monetary']].describe()
```

	Cluster	Recency	Recency	Monetary
count	1096.000000	1096.000000	1096.000000	1096.000000
mean	1.864051	66.655109	66.655109	63549.574480
std	1.074717	32.795863	32.795863	140781.819681
min	1.000000	1.000000	1.000000	34.560000
25%	1.000000	45.000000	45.000000	455.957500
50%	1.000000	61.000000	61.000000	6772.140000
75%	3.000000	92.000000	92.000000	29900.355000
max	4.000000	182.000000	182.000000	877107.190000

- 4.5 > rfm score > 4 : High Value Customer

```
df_filter=df_customers.loc[(df_customers['RFM_Score'] > 4) & (df_customers['RFM_Score'] <= 4.5 )]  
df_filter[['Cluster','Recency','Recency','Monetary']].describe()
```

	Cluster	Recency	Recency	Monetary
count	888.000000	888.000000	888.000000	888.000000
mean	1.128378	221.530405	221.530405	19512.056926
std	0.627593	161.722336	161.722336	56799.256361
min	0.000000	86.000000	86.000000	42.930000
25%	1.000000	126.000000	126.000000	5151.602500
50%	1.000000	176.000000	176.000000	6830.605000
75%	1.000000	238.000000	238.000000	8153.020000
max	4.000000	853.000000	853.000000	585516.430000

- 4>rfm score >3 : Medium value customer

Businesses can continue to strengthen their present sales policies to maintain this core customer group with this set of customers. In addition to identifying potential clients in this demographic and encouraging them to become loyal customers

- 3>rfm score>1.6 : Low-value customer


```
[65] df_filter=df_customers.loc[(df_customers['RFM_Score'] > 1.6) & (df_customers['RFM_Score'] <= 3 )]
df_filter[['Cluster','Recency','Recency','Monetary']].describe()
```

	Cluster	Recency	Recency	Monetary
count	6132.000000	6132.000000	6132.000000	6132.000000
mean	1.660144	102.471298	102.471298	493.664847
std	0.473699	53.743187	53.743187	754.945086
min	1.000000	1.000000	1.000000	2.290000
25%	1.000000	57.000000	57.000000	32.600000
50%	2.000000	100.000000	100.000000	69.970000
75%	2.000000	148.000000	148.000000	603.490000
max	2.000000	199.000000	199.000000	2860.880000

- rfm score<1.6 :Lost Customer

```
[66] df_filter=df_customers.loc[(df_customers['RFM_Score'] <1.6) ]
df_filter[['Cluster','Recency','Recency','Monetary']].describe()
```

	Cluster	Recency	Recency	Monetary
count	5352.000000	5352.000000	5352.000000	5352.000000
mean	1.599215	334.830717	334.830717	688.152997
std	0.629630	167.649380	167.649380	1019.328065
min	0.000000	206.000000	206.000000	1.370000
25%	1.000000	247.000000	247.000000	35.000000
50%	2.000000	296.000000	296.000000	75.480000
75%	2.000000	342.250000	342.250000	1145.480000
max	2.000000	1127.000000	1127.000000	6214.260000

Come up with a strategy:

- For groups of customers who bring high value but have not returned for a long time, we need to find out why (maybe due to poor customer care, unprofessional service,...), come up with strategies to entice them to come back because the cost of retaining customers is much cheaper than finding new customers
- For vip customers, there should be incentives, showing the business's gratitude to them, making them feel worthy for choosing to use our services.
- For customers who bring low value but have a high frequency of using business services, they should also give them a certain amount of respect because each customer is a piece of the puzzle that creates the success of the business.

Conclusion:

With the results given by models, techniques from dry, rigid algorithms, data analysts or managers should also evaluate the results from a practical, business, depending perspective. time. From there, make smart, creative, highly effective decisions when combining both technical analysis and sensory analysis of the manager. To understand and capture customer behavior is a very difficult thing, like people say "So many men, so many minds", so businesses need to combine many business models, as well as business models. Smart strategies, methods of changing to adapt to the market, have separate approaches for each customer. Only then can new businesses succeed and develop sustainably.

REFERENCES

- [1] Đặng Thế Lâm, Scikit Learn - K-Means - Elbow - tiêu chí; 2017
- [2] Big Datauni, Các phương pháp đánh giá trong thuật toán Clustering;
- [3] Đỗ Khánh Ngọc, Giải thích về điểm số hình bóng của K-Means bằng ví dụ Python; 2020
- [4] Ichi.pro, Phương pháp Elbow so với Silhouette Đồng hiệu quả trong việc xác định số lượng cụm;
- [5] Ichi.pro, Điểm Z hoặc Điểm chuẩn hóa;
- [6] Thanh Trung Ho and Son Dang Nguyen, An interdisciplinary research between analyzing customer segmentation in marketing and machine learning methods. SCIENCE & TECHNOLOGY DEVELOPMENT JOURNAL - ECONOMICS - LAW AND MANAGEMENT; 2021
- [7] Cleverlap, RFM analysis for Customer Segmentation; 2021
- [8] Gabriel Signoretti, Divide and Conquer: segment your customers using RFM analysis; 2019
- [9] SimERP, Phân khúc khách hàng là gì? Các phân khúc khách hàng phổ biến; 2021
- [10] Geeksforgeeks.org, RFM Analysis Analysis Using Python
- [11] Surefunmi Idowu, Customer Segmentation Based on RFM Model Using K-Means, Hierarchical and Fuzzy C- Means Clustering Algorithms; 2019
- [12] Optimove, RFM Segmentation; 2021
- [13] Surendra Tanniru, Customer Segmentation Using RFM Analysis. was published as a part of the Data Science Blogathon; 2021
- [14] BİLİŞİM TEKNOLOJİLERİ DERGİSİ, Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing. Araştırma Makalesi/Research Article; 2020
- [15] Omniconvert, Effective Customer Segmentation through RFM Analysis
- [16] Kim-Giao Tran, Van-Ho Nguyen, Thanh Ho, Customer segmentation analysis and customer lifetime value prediction using Pareto/NBD and Gamma-Gamma model, University of Economics and Law, Ho Chi Minh City, Vietnam;

SỐ CẤU TRÚC PHÂN CHIA CÔNG VIỆC (WBS)	TIÊU ĐỀ CÔNG VIỆC	NGƯỜI PHỤ TRÁCH CÔNG VIỆC	NGÀY BẮT ĐẦU	NGÀY ĐẾN HẠN	THỜI GIAN	PHẦN TRĂM CÔNG VIỆC HOÀN THÀNH
1	Phrase 1: Chuẩn bị					
	Team Data: Huy + Linh					
1.1	Tham Khảo các bài báo	Cả nhóm	19/12/21	20/12/21	1	100%
1.1.1	Tìm cách chạy View, Import Feature	Huy	17/12/21	18/12/21	1	100%
1.2	Hiểu kết cấu đồ án, lập sơ đồ mô hình	Linh, Huy	18/12/21	20/12/21	2	100%
1.3	Hoàn thành chương 1, chương 2	Linh, Huy	20/12/21	25/12/21	5	100%
	Team Model : Kiên + Tuyến + Vinh					
2.1	Hiểu được thuật toán RFM, K_Means, Silhouette	Kiên	18/12/21	20/12/21	2	100%
2.2	Lý thuyết Z_Score, Elbow	Vinh	18/12/21	20/12/21	2	100%
2.3	Tìm hiểu mô hình RFM (Code + Lý Thuyết)	Tuyến	18/12/21	20/12/21	2	100%
2	Phrase 2: Thực nghiệm					
	Team Data: Huy + Linh					
3.1	Format lại Word	Vinh	23/12/21	25/12/21	2	100%
3.2	Lấy Feature, tạo Dataset	Huy	20/12/21	21/12/21	1	100%
3.2.1	Phân chia làm PPT	Cả nhóm	24/12/21	26/12/21	2	100%
3.2.2	Đánh giá, viết báo cáo	Cả nhóm	24/12/21	25/12/21	1	100%
	Team Model : Tuyến + Kiên					
4.1	Chạy Model, xuất kết quả	Tuyến	21/12/21	24/12/21	3	100%
4.2	Kết luận, Giải Pháp	Tuyến	24/12/21	25/12/21	1	100%
4.3	EDA Data of AdventureWork	Kiên	23/12/21	25/12/21	2	100%
4.4	EDA RFM dataframe	Tuyến	24/12/21	25/12/21	1	100%