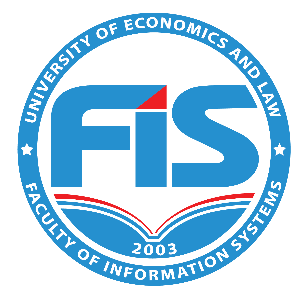
**UNIVERSITY OF ECONOMICS AND LAW**

**FACULTY OF INFORMATION SYSTEMS**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

****

**FINAL PROJECT REPORT**

**FUNDAMENTALS OF DATA ANALYTICS COURSE**

**TOPIC: TOPIC’S NAME (tên chủ đề, chẳng hạn như: Định hướng xây dựng và phát triển nghề nghiệp Business Analytics)**

**Lecturer: Ho Trung Thanh, Ph.D.**

**Group:**

**Ho Chi Minh City, December 21st, 2021­­**

# Members of Group 6

|  |  |  |
| --- | --- | --- |
| *No.* | **NAME** | **STUDENT ID** |
| *1* | Trần Triệu Vy | K204162011 |
| *2* | Bùi Thị Thu Vân | K204162007 |
| *3* | Lê Chí Hào | K204161988 |
| *4* | Hoàng Nguyễn Quỳnh Giang | K204161987 |
| *5* | Lê Thị Thùy Linh | K204161990 |

# **Acknowledgments**

My team would like to thank Mr. Ho Trung Thanh - Lecturer of Database Analysis and assistant lecturer Nguyen Phat Dat - Faculty of Information Technology who equipped us with basic knowledge and skills to complete the final project.

However, in the process of researching the topic, due to limited specialized knowledge, we still have many shortcomings when researching, evaluating and presenting the topic. We hope to receive the attention and suggestions of teachers to improve our topic.

Sincerely thanks.

Group 6

# **Commitment**

I hereby declare that the above project is the research work of our group under the guidance of lecturer Ho Trung Thanh and assistant lecturer Nguyen Phat Dat. The statements stated in the project are also the results of direct, serious, independent research of the author himself and the basis of searching, understanding and studying scientific documents or translations. other have been announced. The project will still help ensure objectivity, honesty and science.

Group 6

**Contents**

[List of Tables (Danh mục bảng biểu) 5](#_Toc91351964)

[List of Figures (Danh mục hình ảnh) 5](#_Toc91351965)

[List of Acronyms (Danh mục từ viết tắt) 5](#_Toc91351966)

[**Project overview** 6](#_Toc91351967)

[I. Reasons 6](#_Toc91351968)

[II. Objectives 6](#_Toc91351969)

[III. Objects and scopes 6](#_Toc91351970)

[1. Objects 6](#_Toc91351971)

[2. Scopes 7](#_Toc91351972)

[3. Research Methods 7](#_Toc91351973)

[3.1. Qualitative 7](#_Toc91351974)

[3.2. Quantitative 8](#_Toc91351975)

[**Chapter 1: Introduction** 9](#_Toc91351976)

[I. Background 9](#_Toc91351977)

[1. RFM Analytics Model 9](#_Toc91351978)

[2. K-Means Clustering Algorithm 10](#_Toc91351979)

[3. Elbow Method 12](#_Toc91351980)

[4. Sum Squared Error (SSE) 13](#_Toc91351981)

[5. Silhouette Index Method 13](#_Toc91351982)

[6. BG/NBD model 14](#_Toc91351983)

[7. Gamma-Gamma model 15](#_Toc91351984)

[8. Customer Retention Rate with Cohort 15](#_Toc91351985)

[9. Power BI 16](#_Toc91351986)

[10. Customer Lifetime Value 17](#_Toc91351987)

[II. Research Goals 18](#_Toc91351988)

[III. Data understanding 18](#_Toc91351989)

[IV. Data preprocessing 19](#_Toc91351990)

[**Chapter 2: RFM model with k-means clustering** 20](#_Toc91351991)

[I. EDA 20](#_Toc91351992)

[II. Cohort Analyst based on time 21](#_Toc91351993)

[III. Customer segmentation based on RFM model 23](#_Toc91351994)

[1. Calculate the RFM values. 23](#_Toc91351995)

[2. Transform the dataset. 25](#_Toc91351996)

[3. Standardize features 30](#_Toc91351997)

[4. Choose the optimal clusters number for K-means 32](#_Toc91351998)

[5. Cluster quality verification with Silhouette index 33](#_Toc91351999)

[6. Cluster customer segments and visualize analytics results 33](#_Toc91352000)

[IV. Analysis of the groups of customers based on the quartile description 37](#_Toc91352001)

[1. Analyzing the group Can’t Lose Them (cluster 0) 37](#_Toc91352002)

[2. Analyzing the group Potential Loyalists (cluster 1) 38](#_Toc91352003)

[3. Analyzing the group New customers (cluster 2) 39](#_Toc91352004)

[4. Analyzing the group At risk customers (cluster 3) 40](#_Toc91352005)

[5. Analyzing the group Champions (cluster 4) 41](#_Toc91352006)

[**Chapter 3: RFM model for customer segmentation by Power BI** 43](#_Toc91352007)

[I. Customer segment chart 43](#_Toc91352008)

[II. Explain some of the charts in the report 46](#_Toc91352009)

[III. Report about customer segmentation 48](#_Toc91352010)

[**Chapter 4: Predicting CLV** 49](#_Toc91352011)

[I. Calculating CLV 49](#_Toc91352012)

[II. Fit the BG/NBD model 52](#_Toc91352013)

[III. Fit the Gamma-Gamma model 53](#_Toc91352014)

[IV. Segment customers by their AOV 56](#_Toc91352015)

[V. Predicting Customer Lifetime Value 57](#_Toc91352016)

[**Conclusion and Future** 60](#_Toc91352017)

# **List of Tables**

[Table 1: Columns chosen from dataset 23](#_Toc91355107)

[*Table 2: Customers Segmentation based on RFM score* 48](#_Toc91355108)

# **List of Figures**

[Figure 1: Research Process 12](#_Toc91354990)

[Figure 2: First 5 values of dataset 24](#_Toc91354991)

[Figure 3: Visualization of nullity by column 24](#_Toc91354992)

[Figure 4: Quartile Description of SubTotal 25](#_Toc91354993)

[Figure 5: Retention Rate in percentage - Monthly Cohort 26](#_Toc91354994)

[Figure 6: Top 10 RFM values for each customer 28](#_Toc91354995)

[Figure 7: Plotting for the last day since the customer made a purchase 28](#_Toc91354996)

[Figure 8: Plotting the number of times the customer has made a purchase 28](#_Toc91354997)

[Figure 9: Plotting the total revenue that the particular customer brought in to the shop 29](#_Toc91354998)

[Figure 10: Plots to analyze skewness of frequency 29](#_Toc91354999)

[Figure 11: Plots to analyze skewness of recency 30](#_Toc91355000)

[Figure 12: Plots to analyze skewness of monetary values 31](#_Toc91355001)

[Figure 13: Top 10 RFM values after transformation 32](#_Toc91355002)

[Figure 14: Quartile description in RFM 33](#_Toc91355003)

[Figure 15: Scatter plot with data after transformation 33](#_Toc91355004)

[Figure 16: Illustration of the Z-Score of Frequency 35](#_Toc91355005)

[Figure 17: The Elbow Method 36](#_Toc91355006)

[Figure 18: Silhouette score for each k 37](#_Toc91355007)

[Figure 19: The number of customers in each group 38](#_Toc91355008)

[Figure 20: Scatter Plot for customers segmentation 39](#_Toc91355009)

[Figure 21: Quartile description in cluster 0 41](#_Toc91355010)

[Figure 22: Quartile description in cluster 1 42](#_Toc91355011)

[Figure 23: Quartile description in cluster 2 43](#_Toc91355012)

[Figure 24: Quartile description in cluster 3 44](#_Toc91355013)

[Figure 25: Quartile description in cluster 4 45](#_Toc91355014)

[Figure 26: Shows the customer sets and the number of customers by volume 48](#_Toc91355015)

[Figure 27: Shows a detailed table of each customer 50](#_Toc91355016)

[Figure 28: Shows the number of orders and the total price purchase value over time 51](file:///E:\UEL\Phân%20tích%20dữ%20liệu%20cơ%20bản\Final\Final%20project%20group%206.docx#_Toc91355017)

[Figure 29: Shows the total purchase value for each product and product category. 51](file:///E:\UEL\Phân%20tích%20dữ%20liệu%20cơ%20bản\Final\Final%20project%20group%206.docx#_Toc91355018)

[Figure 30: RFM model for customer segmentation 52](file:///E:\UEL\Phân%20tích%20dữ%20liệu%20cơ%20bản\Final\Final%20project%20group%206.docx#_Toc91355019)

[Figure 31: Top 10 values of the dataset1. 54](#_Toc91355020)

[Figure 32: Plot for recency in dataset1. 54](#_Toc91355021)

[Figure 33: Plot for frequency in the dataset1. 55](#_Toc91355022)

[Figure 34: Plot for monetary in the dataset1. 55](#_Toc91355023)

[Figure 35: Plot for T values in the dataset1. 56](#_Toc91355024)

[Figure 36: Customers who will return within the next 1 week 56](#_Toc91355025)

[Figure 37: Customers who will return within the next (1 month = 4 weeks) 57](#_Toc91355026)

[Figure 38: 10 customers expect to buy the most in the next 6 months 57](#_Toc91355027)

[Figure 39: Total transactions made in the next 6 months 57](#_Toc91355028)

[Figure 40: Top 10 customers that bring most values 58](#_Toc91355029)

[Figure 41: The top aov values 59](#_Toc91355030)

[Figure 42: Plot for aov scores 59](#_Toc91355031)

[Figure 43: Quartile description of 5 bin of aov 60](#_Toc91355032)

[Figure 44: 5 clusters by aov points 61](#_Toc91355033)

[Figure 45: The customers' lifetime values expected to in the next 3 months 61](#_Toc91355034)

[Figure 46: Customer segmentation by CLV 62](#_Toc91355035)

# **List of Acronyms**

|  |  |
| --- | --- |
| DB | Digital Business |
| RFM | Recency, Frequency, Monetary |
| CLV | Customer lifetime value |
| AOV | Average Order Value |
| BI | Business Intelligence |

# **Project overview**

## Reasons

Customers are separated into groups with distinct similarities as a result of data segmentation. Gender, age, lifestyle, location, purchasing, and income behavior are some of the factors relevant to client segmentation. Such characteristics are primarily classified based on previous purchasing behavior that can result in a specific outcome, such as an increase in sales and profit for the company.

In the ever-growing competition and increasing complexity of the business environment, segmentation and its systematic study improves customer loyalty and enhances enterprise-level for long lasting relationships by widening profitable customer databases (Khalili-Damghani et al., 2018). Based on the statistical results and indicators, companies in the retail industry can design various sales and marketing strategies like promotional campaigns, extending seasonal discounts or floating sales enabling coupons to increase the sales and improve customer retention (P. Anitha and M. M. Patil, 2019).

## Objectives

This study proposes a case look at the usage of statistics mining techniques, K-approach clustering, and RFM evaluation to assemble a customer segmentation model, to section clients for an IT answer and carrier issuer commercial enterprise customer demographics and buy behaviours. The findings, customer segments, are supposed to be useful for commercial enterprises to higher recognize their clients and may support their customer-centric advertising strategy.

## Objects and scopes

### Objects

Segmentation of customer groups

### Scopes

**2.1. Time scopes**

* Research time: 12.12.2021 - 25.12.2021
* Data analysis time: 06.2011 - 07.2014

**2.2. Space scopes**

* Customer segmentations for data in AdventureWorks company.
* Predict Customer Lifetime Value for data in AdventureWorks company.

## Research Methods

### 3.1. Qualitative

***Issue 1: Construct the customer segmentation model using K-means clustering technique and RFM analysis.***

**RQ1:** How data mining is used in customer segmentation?

**RQ2:** How to use RFM evaluation to consumer segmentation?

**RQ3:** How to research RFM facts about the use of K-way clustering?

***Issue 2: RFM model for customer segmentation by Power BI***

***Issue 3: Segment customers by their CLV values***

**RQ1:** How to predict values that customers bring to a company?

**RQ2:** How to value customers based on their CLV values?

**RQ3:** How to group customers based on their CLV values?

***Issue 4: Customer retention rate***

**RQ1:** How long do your customers stay customers?

**RQ2:** How to calculate customer retention rate?

### 3.2. Quantitative



Figure 1: Research Process

# **Chapter 1: Introduction**

## Background

### RFM Analytics Model

The Recency, Frequency, and Monetary (RFM) analytic version proposed with the aid of using Strategic database advertising (Hughes, 1994), is one of the critical versions for groups to formulate advertising strategies (Hughes, 2012). The RFM version represents customers’ intake behaviours primarily based totally at the transaction database, that's simplified into 3 variables (attributes) as follows:

* Recency (R): R represents recency, which refers to the period of time starting from recent consuming behaviour happens (last purchasing) and present. The date closer to present, the higher possibility customers will make another purchase. Thus, it will has a higher value in recency variable.
* Frequency (F): F represents frequency, which refers to the number of transactions in a particular period. It is expected that the higher purchase frequency of customers, the higher loyalty customers are, and the higher customer value to the company. The more the frequency is, the higher the value in the frequency variable.
* Monetary (M): M represents monetary, which refers to the total consumption money amount in a particular period. It is expected that the higher the value of money, the higher the profit contributions made by customers to the company, and the higher customer value. (Ponlacha Rojlertjanya, 2019)

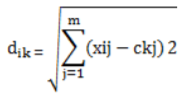
According to Wu & Lin (2005) research, the larger the fee of recency (R) and frequency (F) is, the more possibility customers made new orders with the employer. Furthermore, the larger the fee of monetary (M) is, the much more likely the corresponding customers will buy services or products with the employer again.

Many researchers confirmed that the RFM evaluation version is a great approach to phase goal clients from massive information into 3 critical variables. Nonetheless, there are distinct standards that have indicated in research approximately a way to set up with 3 variables. In the authentic RFM evaluation version developed by Hughes (1994), he described the 3 variables as similarly critical. Thus, the weight of variables are identical. However, the distinct withinside the critical idea changed into issued with the aid of using Stone (1995), he indicated that the critical of every variable is relies upon at the feature of industry. Thus, the weight of variables aren't identical withinside the Stone’s idea.

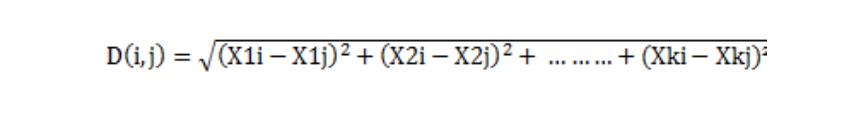
### K-Means Clustering Algorithm

K-means is one form of the simplest grouping. The procedure is simple and easy to classify data given through a number of clusters. Determination centroid is done by taking data first as the first centroid, second data as second centroid, and so on to the number of centroids required. The next step is to calculate the distance from the point to be clustered to each centroid that is available and grouped accordingly with the closest distance to the centroid. (N. H. Kristanto,2016). The K-Means algorithm is an algorithm clustering which groups data based on the cluster center point (centroid) closest to data. The purpose of K-Means is grouping data to maximize data similarity in one cluster and minimize data similarity between clusters. Similarity measures used in the cluster is the distance function. So that maximises the similarity of data obtained based on the shortest distance between data towards the centroid point (E. Muningsih and S. Kiswati, 2018).

At the beginning of the iteration, the center of each Cluster is set freely (arbitrary), ckj (k = 1, .... k, j = 1, .... m). Then the distance between each data is calculated with each Cluster center. the first in the center of the Cluster k (ck), given the name (dik) can be used Euclidean formula (1), namely:

 (1)

A data will be a member of the J-Cluster if the data distance to the center of the J-Cluster is the smallest compared to the distance to the other Cluster center. Next, group the data that is a member of each Cluster. The new Cluster center value can be calculated by finding the average value of the data that is a member of the Cluster. The one step to Cluster with the K-Means method is as follows: Select the number of Clusters k. Initialization of the center of this Cluster can be done in various ways. But the most often done is by random. Cluster centers are given an initial value with random numbers. Allocate all data / objects to the nearest Cluster. The proximity of two objects is determined based on the distance of the two objects. Likewise, the proximity of a data to a particular Cluster is determined by the distance between the data and the Cluster center. In this stage it is necessary to calculate the distance of each data to each Cluster center. The most distance between one data and one particular cluster will determine which data is included in which cluster. To count the distance of all data to each cluster center, the Cluster can use Euclidean distance theory which is formulated (2) as follows:

(2)

Where:

*D (i, j) = Data distance to i to Cluster center j*

*xki = Data to i on data attribute to k*

*Xkj = Center point to j in attribute to k*

Recalculate the Cluster center with the current Cluster membership. The Cluster Center is the average of all data / objects in a particular Cluster. If you wish you can also use the median of the Cluster. So the mean (mean) is not the only measure that can be used. Assign each object again using the new Cluster center. If the Cluster center does not change again then the Clustering process is complete. Or, return to step number 3 until the Cluster center doesn't change again. The K-Means algorithm is the best algorithm in partitional Clustering algorithm and is most often used among other Clustering algorithms, because of its simplicity and efficiency. Rena Nainggolan et al 2019 J. Phys.: Conf. Ser. 1361 012015, 2018).

### Elbow Method

Elbow method is a method used to produce information in determining the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point. Different percentage results from each cluster value can be shown using the graph as the source of the information. If the value of the first cluster with the value of the second cluster gives the angle in the graph or the value has the biggest decrease then the value of the cluster is the best. To get a comparison is to calculate SSE (Sum of Square Error) from each cluster value. Because the greater the number of cluster K, the SSE value will be smaller Following are the stages of the Elbow method algorithm in determining the k value in K-Means:

1. Initialize the initial value of k

2. Increase the value of k

3. Calculating the sum of square error results from each value of k

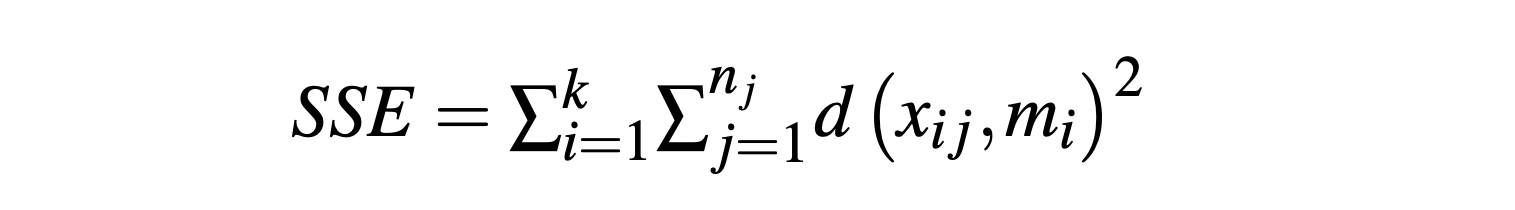
4. Analysis of the sum of square error results from the k value which has drastically decreased

5. Locate and set the elbow-shaped k value.

In the Elbow method the best cluster value will be taken from the Sum of Square Error (SSE) value which has a significant and elbow-shaped decrease. To calculate SSE using a formula (3). Rena Nainggolan et al 2019 J. Phys.: Conf. Ser. 1361 012015, 2018).

### Sum Squared Error (SSE)

SSE (Sum Square Error) is one of the statistical methods used to measure the total difference from the actual value of the value achieved (B. Yusuf Bakhtiar, A. Bima Murti Wijaya, and H. Dwi Cahyono, 2019).

(3)

Where, d is the distance between the data and the Cluster center.

Sum of Square Error (SSE) is a formula used to measure the difference between the data obtained by the prediction model that has been done previously. SSE is often used as a research reference in determining optimal clusters. (Rena Nainggolan et al 2019 J. Phys.: Conf. Ser. 1361 012015, 2018).

### Silhouette Index Method

The Silhouette Index is the most popular and widely used index of clustering results. Silhouette index analysis aims to measure the optimal level when an observation, a data point is classified into any cluster. Specifically, Silhouette's method will tell us which data points or observations fit inside the cluster (good) or close to the edge of the cluster (not good) to evaluate the clustering efficiency.

Silhouette measures the distance of a data point in the cluster to the cluster center point, and the distance of that point itself to the center point of the nearest cluster (or to the center points of the remaining clusters, and selects the shortest way). That is the measure case for K-means clustering.

If the clusters are not found based on clustering, then Silhouette will measure in the same way but instead of calculating the distance between that point and the center point, we will calculate the average distance from all the remaining points. in the cluster of that point, and the average distance from all the remaining points of the other clusters (take the shortest average distance)

If a cluster is evaluated for quality, the points in the cluster will have Silhouettes approaching 1 and vice versa.

For a quick assessment of whether a point is properly clustered we can rely on Silhouette:

The data point with high Silhouette, close to 1, is definitely in the cluster

The data point with Silhouette close to 0, is located between 2 clusters

If the data point has a low Silhouette, negative value, it is likely to be in the wrong cluster.

In addition, according to the experience of the authors in the document "Data mining and Predictive analytics" by Wiley publisher:

Silhouette average score of 0.5 or higher, proof that this cluster may be close to reality

The average Silhouette score is from 0.25 to 0.5, it is necessary to have more professional knowledge and experience to further assess the cluster's ability to exist in reality.

If the average score is less than 0.25, the cluster should not be trusted, and more evidence should be sought.

### BG/NBD model

In this research, we used to predict the expected number of transactions by using the Beta Geometric/Negative binomial distribution model which was proposed by Fader, Hardie, and Lee. It is a good model for RFM type problems by modeling discrete-time data and comparing the forecast result with the actual data. Since CLV calculation is completely based on RFM type, BG/NBD model suits this research.

* There are three steps to implement this model, they are getting ready with parameters
* Creating a sales forecast using the parameters
* Predict the future purchase of a customer base of parameters estimated and past behavior analysis. (Prathima J, Vaishnavi M, Perumalraja R and Kamalesh S, 2021)

To implement this the three needed parameters are Recency (When his last transaction occurred), Frequency (number of transactions in a specific period of time), and monetary (the amount the customer spends over the same period of time).

### Gamma-Gamma model

The Gamma-Gamma Model can predict the most likely value per transaction in the future. The Gamma-Gamma model is used to model the monetary value. The model is based on the following three general assumptions:

* The monetary value (e.g., $, £, e) of a customer’s given transaction varies randomly around their average transaction value.
* Average transaction values vary across customers but do not vary over time for any given individual.
* The distribution of average transaction values across customers is independent of the transaction process.

(Prathima J, Vaishnavi M, Perumalraja R and Kamalesh S, 2021)

### Customer Retention Rate with Cohort

Customer retention is a cornerstone of broader CRM concepts such as customer equity (Blattberg, Getz and Thomas 2001; Rust et al. 2015) and is arguably the most important component of the customer lifetime value (CLV) framework (Gupta, Lehmann and Stuart 2004). First, the central idea that customer retention is continuity – the customer continues to interact with the firm. Second, that customer retention is a form of customer behavior – a behavior that firms intend to manage. Accordingly, we propose that “Customer retention is the customer continuing to transact with the firm.” A few things are worth noting in this simple definition. It emphasizes retention as something the customer does (which possibly could be affected by the firm). Lastly, “churn” is the counterpart of retention. If the customer has decided to stop transacting with the firm, the customer has churned. In that sense churn is inferred by the cessation of the customers’ transactions with the firm. (Eva Ascarza at al., 2017).

Grouping customers according to cohort, also known as grouping customers according to the timeline from the customer's first transaction (Croll & Yoskovitz, 2013). In this study, we group customers by month. The formula to calculate the retention rate is described as equation (4):

(4)

### Power BI

#### 9.1. What is Power BI ?

Power BI is a Business Intelligence and Data Visualization tool for converting data from various data sources into interactive dashboards and analysis reports. Power BI offers cloud-based services for interactive visualizations with a simple interface for end users to create their own reports and dashboards.

Different Power BI versions like Desktop, Service-based (SaaS), and mobile Power BI apps are used for different platforms. It provides multiple software connectors and services for business intelligence.

#### 9.2 Why do we choose Power BI ?

* ***Access to Volumes of Data from Multiple Sources***

Power BI can access vast volumes of data from multiple sources. It allows you to view, analyze, and visualize vast quantities of data that cannot be opened in Excel. Some of the important data sources available for Power BI are Excel, CSV, XML, JSON, pdf, etc. Power BI uses powerful compression algorithms to import and cache the data within the.PBIX file.

* ***Interactive UI/UX Features***

Power BI makes things visually appealing. It has an easy drag and drops functionality, with features that allow you to copy all formatting across similar visualizations.

* ***Accelerate Big Data Preparation with Azure***

Using Power BI with Azure allows you to analyze and share massive volumes of data. An azure data lake can reduce the time it takes to get insights and increase collaboration between business analysts, data engineers, and data scientists.

* ***Turn Insights into Action***

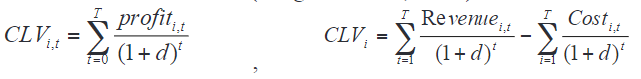
Power BI allows you to gain insights from data and turn those insights into actions to make data-driven business decisions.

* ***Real-time Stream Analytics***

Power BI will enable you to perform real-time stream analytics. It helps you fetch data from multiple sensors and social media sources to get access to real-time analytics, so you are always ready to make business decisions.

### Customer Lifetime Value

Customer Lifetime Value has been studied under the name of LTV, Customer Value, Customer Equity and Customer Profitability. The concept is defined as the sum of the revenues gained from a company's customers over the lifetime of transactions after deduction of the total cost of attracting, selling and servicing customers, taking into account the time value of money (Hwang et al, 2004). The basic formula (5) for calculating CLV for customer i at time t for a finite time horizon T (Berger & Nasr, 2004) is:

(5)

Theoretically, CLV models should estimate the value of a customer over the entire customer’s lifetime. However, in practice most researchers use a finite time horizon of three or four years (Donkers et al, 2007; Rust et al, 2000, Beniot and Poel, 2009). Three to four years is a good estimate for the horizon over which the current business environment would not substantially change and even then, there is significant uncertainty in predicting customer behavior (Venkatesan et al, 2007). Moreover, some research considers an even shorter time horizon (Hwang et al, 2004). (Mohammad Safari Kahreh et al., 2014)

## Research Goals

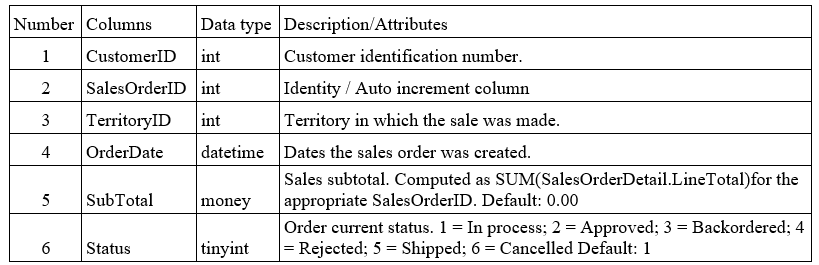
1. Calculate customer retention rate monthly by cohort analyst
2. Construct the customer segmentation model using K-means clustering technique and RFM analysis.
3. Conduct the customer segmentation based on their customer purchase behaviours.
4. Predicting CLV using NB/NBD and Gamma-Gamma model
5. Analyze results and provide recommendations to the business on marketing strategy for each customer segment.

## Data understanding

The study makes use of a customer transaction dataset taken from the Adventure Works Cycles dataset. This is a worldwide corporation that produces and sells bicycles in North America, Europe, and Asia. From 06/2011 to 07/2014, the retrieved dataset contains 121,317 transactions from the company. This applies to both individuals and businesses. To be more specific, the percentage of customers in each country is different and the US is the biggest market. According to that, the transactions in the US (United States) market were excluded from the analysis in terms of analyzing the optimal customer segment for each different market.

After defining the concept of the model, we choose 6 variables / columns from Sales.SalesOrderHeader table to use in modelling, which is shown in table 1. We choose customers from TerritoryID from range 1 to 5, all belonging to the US. Customers that have status equal 6 will be dismissed, since we choose only orders that have been successful to calculate.

Table 1: Columns chosen from dataset



## Data preprocessing

First, we select appropriate features from the main data to build the model, bring it into a dataset. The main task in this step is to clean the data set and remove missing values and null attributions. Then transform it from raw data to the useful format that has normal distribution. We will use this data to build the model.

# **Chapter 2: RFM model with k-means clustering**

## EDA

**Descriptive Analysis**

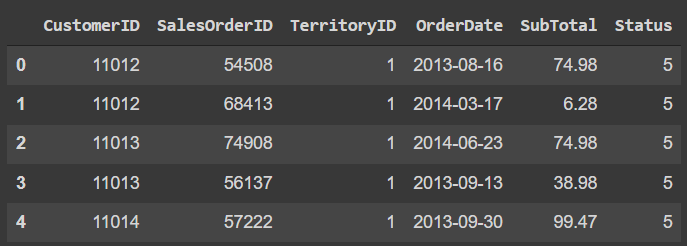


Figure 2: First 5 values of dataset

Customer ID ‘11012’ has 2 rows of orders, meaning that one customer can order more than one product and each order for each type of product represents a row.

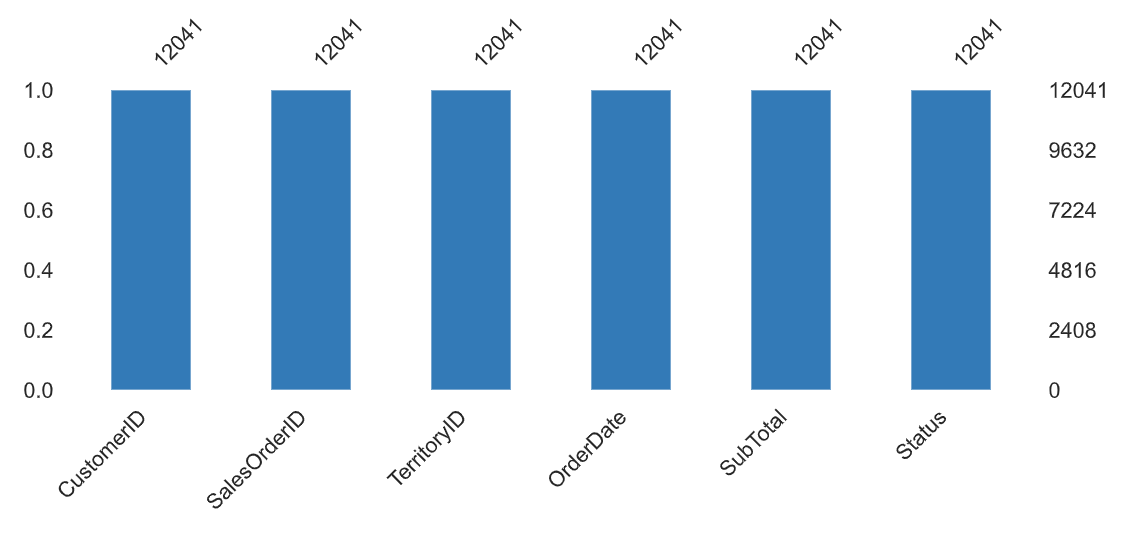


Figure 3: Visualization of nullity by column

According to the chart, the dataset has 12041 rows, with no missing valuesor null attributes.

We can see the quartile description of the SubTotal column in figure 3

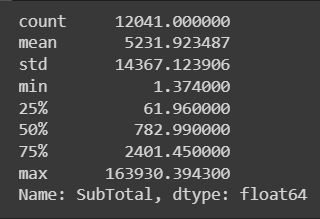


Figure 4: Quartile Description of SubTotal

The average revenue of the US market is 5231.923 (dollars), while the revenue range is from 1.374 (dollars) to 163930.394 (dollars).

## Cohort Analyst based on time

Before going to the RFM model, we considered how the business was going in the company in this period. Therefore, we calculated the percentage of customer retention, whether the customer continues to interact with the firm. Customer retention is a form of customer behavior – a behavior that firms intend to manage. Accordingly, we propose that “Customer retention is the customer continuing to transact with the firm.” (Eva Ascarza, Scott A. Neslin, Rom Schrift, 2017).

In this part, customers will be divided into acquisition cohorts depending on the month of their first purchase. The cohort index would then be assigned to each of the customer’s purchases, which will represent the number of months since the first transaction.

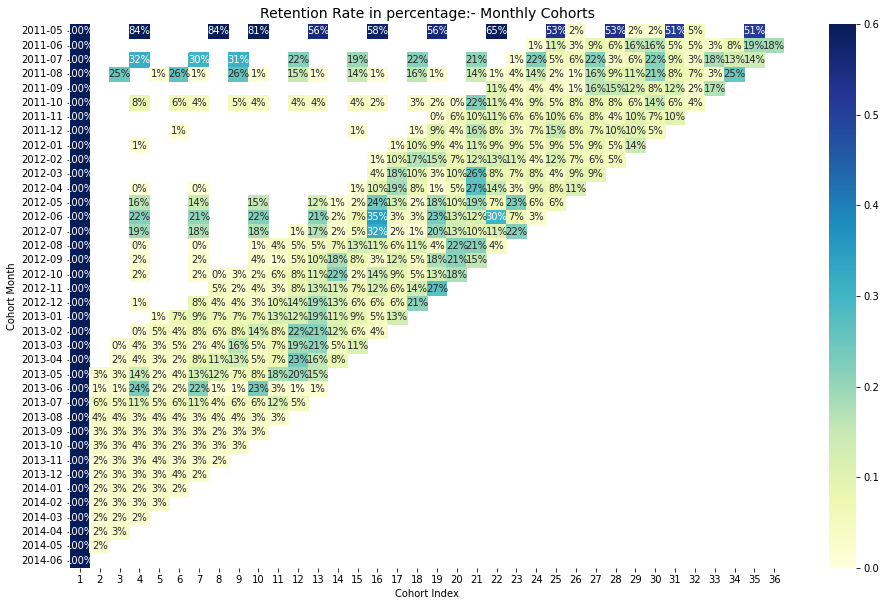


Figure 5: Retention Rate in percentage - Monthly Cohort

Retention Rate 1st index i.e 1st month is 100% as all the customers for that particular customer signed up in 1st Month. The retention rate may increase or decrease in subsequent Indexes.

Here, we have the first order made in May 2011 and the last order was recorded in June 2014 and there are 36 cohort indexes. The darker the blue shades higher the values. The highest percentage can be seen in the row of 2011-05 with 84% in 4th cohort indexes and 8th cohort indexes, which means 84% of customers that signed in May 2011 were active 4 or 8 months later.

With customers' first order in June 2011, only 1% of them returned after 25 months, more than 2 year after. However, this group continued to come back and 18% of them maintained transactions after 36 months, the largest index of all.

Customers did not transact regularly after the first 3 months, but after the 4th month, most customers started reorder again. Customers that ordered the first one from 05-2013 to 05-2014 all returned to transact from the second month, while the customers who transacted from 06-2014 did not come back the second time.

Generally, the company's customer retention policy was appropriate for the period before 2012 and was able to retain this group of loyal customers until the end of the period. However, it seemed to be no longer suitable for new customer groups, especially when the business in later periods promoted marketing and attracted more customers but could not keep them.

## Customer segmentation based on RFM model

### Calculate the RFM values.

The traditional RFM model used to segment customers based on R (Recency), F (Frequency) and M (Monetary). In this study, to make RFM, we choose OrderDate (Dates the sales order was created) column to get the number of days for recency column, SalesOrderID (Identity / Auto increment column) column to count how much transactions by each customer, SubTotal (Sales subtotal) column to sum all transactions for each customer and other column that appropriate. Missing values, negative transactions, mismatch instock code and description are handled using data preprocessing. For the modified dataset, apply RFM analysis and K-Means clustering.

To calculate the Recency, we calculated the distance between the order date for each customer in the dataset and the next day of the last transaction day that was recorded, then we group by CustomerID to have the dataset in figure 6.

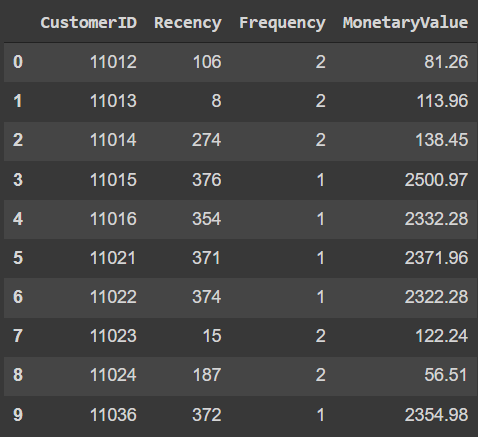


Figure 6: Top 10 RFM values for each customer

Visualization of this results in figure 7, figure 8, figure 9

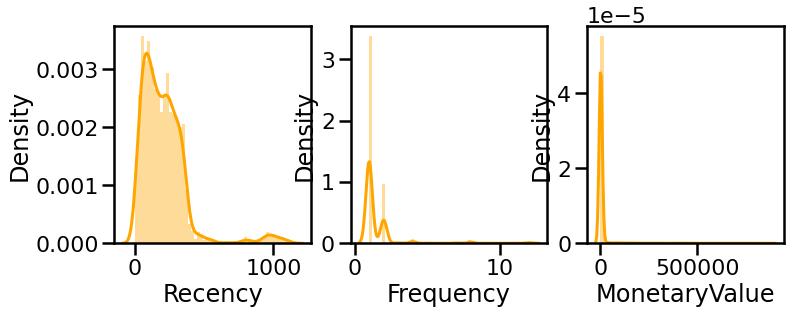


Figure 7: Plotting for the last day since the customer made a purchase

Figure 8: Plotting the number of times the customer has made a purchase

Figure 9: Plotting the total revenue that the particular customer brought in to the shop

Unfortunately, this model is not an optimal model, as we can see three plots are right-skewed or positively-skewed so it does not follow a normal distribution, leading to this data not being appropriate for building a useful model. There are 3 methods of [transformation](https://calculushowto.com/transformations/) of non-normal [dependent variables](https://www.statisticshowto.com/dependent-variable-definition/) into a normal shape: Log, Square and Box Cox. We will compare the results and choose one that most fits.

### Transform the dataset.

First, analyze the skew of frequency values:

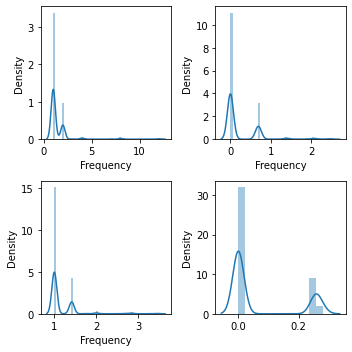


Figure 10: Plots to analyze skewness of frequency

The first one in the upper left-handed is the plot with original data, the next one used log transformation, the third one on the left-handed corner used squared root transformation and the last one used Box Cox transformation.

The skewness are: 5.22, 2.47, 3.72, 1.12 relatively.

Analyze the skew of recency values:

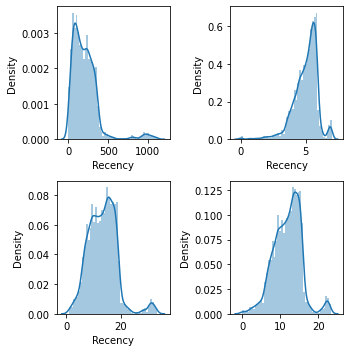


Figure 11: Plots to analyze skewness of recency

The skewness are: 2.56, -1.18, 0.78, 0.06 relatively.

Analyze the skew of monetary values:

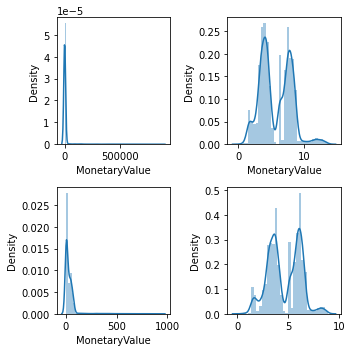


Figure 12: Plots to analyze skewness of monetary values

The skewness are: 10.23, 0.37, 6.27, 0.04 relatively.

We choose a method that has the result more approximately 0, which is the Box Cox method to transform the data for achieving normality. After that, we had the new dataset as in figure…

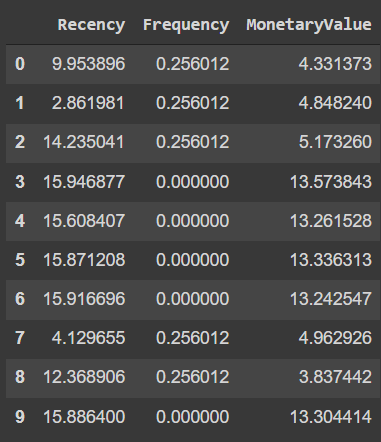


Figure 13: Top 10 RFM values after transformation

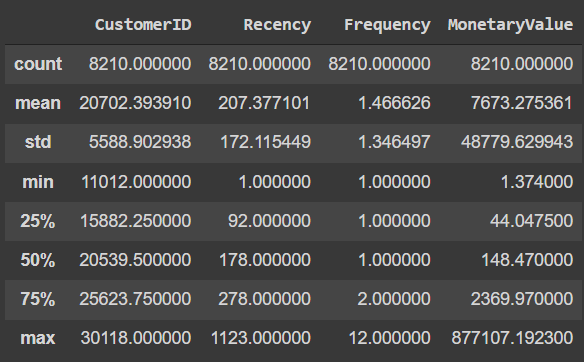


Figure 14: Quartile description in RFM

We illustrates the data with scatter plot in figure 15:

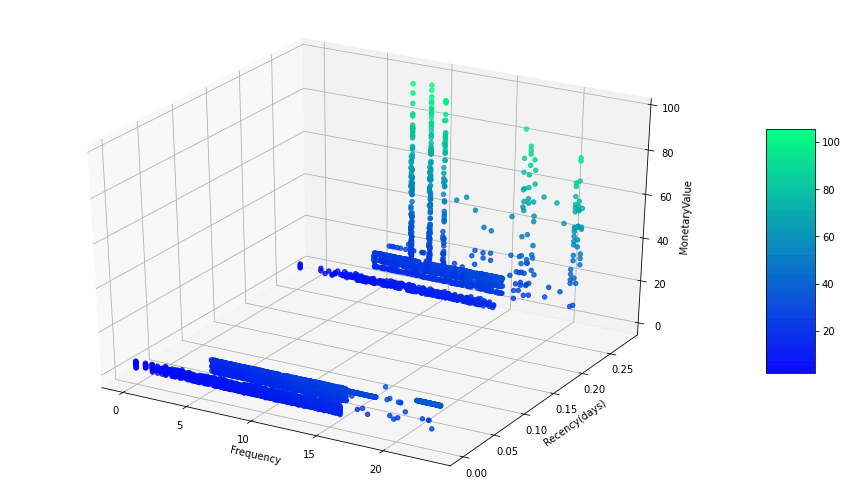


Figure 15: Scatter plot with data after transformation

According to the chart, we can see that the customers have been divided into different groups clearly. The more money spent by each customer, their point is more green and the one with less monetary value has blue points. There are no unexpected gaps in the data and there are any outlier points as we had preprocessed the data before.

### Standardize features

Input variables may have different units (in this data are days, dollars and number) that, in turn, may mean the variables have different scales. Because of the distribution of the values ​​of the elements in the data set and the effects of outliers on the clustering results, the solution is to convert the above values ​​to

the same units as the standard score distribution also known as Z-score.

Normally, Monetary will be very large compared to Recency and Frequency, so in Euclidean space, the distance factor between points representing a customer will be less affected by Recency and Frequency than Monetary. If the data is not normalized, the distance will be mostly affected by Monetary and less affected by the other two variables. We standardize features by removing the mean and scaling to unit variance and the results like in figure…:

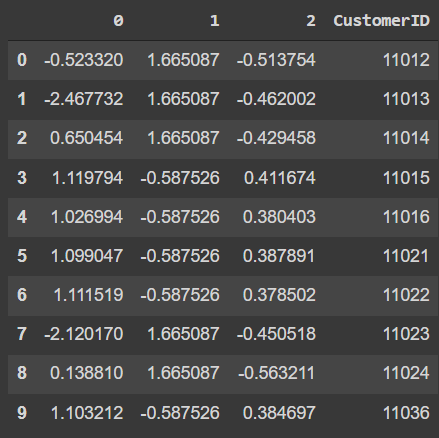


Figure 16: Illustration of the Z-Score of Frequency

Column 0 is the results for Recency Z-score, column 1 is the results for Frequency Z-score, column 2 is the results for Monetary Values. The explanation for this result: the average purchase frequency per customer is 1.46 times. When comparing with the Frequency Z-score of customer 11015 and 11012: Customer 11015 has 0.58 times less purchases than the general market (on average). This is the reason for the presence of a negative sign in this value; Customer 11012 has 1.6 times higher and more frequent purchases on average (1.66).

The average revenue per customer is 7673.27 dollars. When comparing with the Monetary Z-score customer 11012 and 11036: Customer 11012 brings 0.51 times less revenue than the general market (on average). This is the reason for the presence of a negative sign in this value; Customer 11011 has 0.38 times higher and more money spent on average (0.384).

### Choose the optimal clusters number for K-means

The Elbow method is illustrated as a curve graph with the horizontal axis being the number of K clusters (that is, the number of customer segments based on values ​​from the RFM data model), the vertical axis being the SSE (Sum of Errors) index. –i.e. an index that measures the difference between points in a cluster.

Carrying out the Elbow method with the number of clusters from 1 to 10 on the RFM model, the results are as follows:

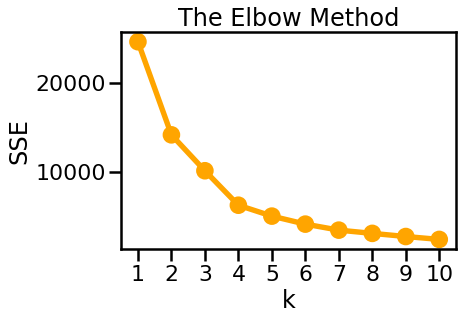


Figure 17: The Elbow Method

With the above SSE curve, the team found that the elbow bend with K = 5 and K = 4 would be an appropriate number of clusters. To explain this, as the basis mentioned above, when increasing the number of clusters, the value of the SSE curve also increases almost steadily, meaning that the difference between points in the cluster is almost unchanged. In other words, the SSE curve tends to gradually decrease in slope after the "elbow" point, and this position on the SEE curve is considered as the optimal point for the input parameter in the K-means clustering method. That satisfies both points K = 4 and K = 5 if observed normally.

### Cluster quality verification with Silhouette index

For accreditation which is the most optimal K cluster when we choose k from the elbow method, we value the Silhouette index from k = 2 to k =10 and get the results in the figure… . Silhouette score for cluster 5 is 0.5082 and is the highest in the range for k from 3 to 10.

Which means that with k = 5, the distance from the point of the cluster to the core of the cluster was optimal and it is close to reality. As a result, we choose K = 5 to cluster, which will be more optimal than K= 4.

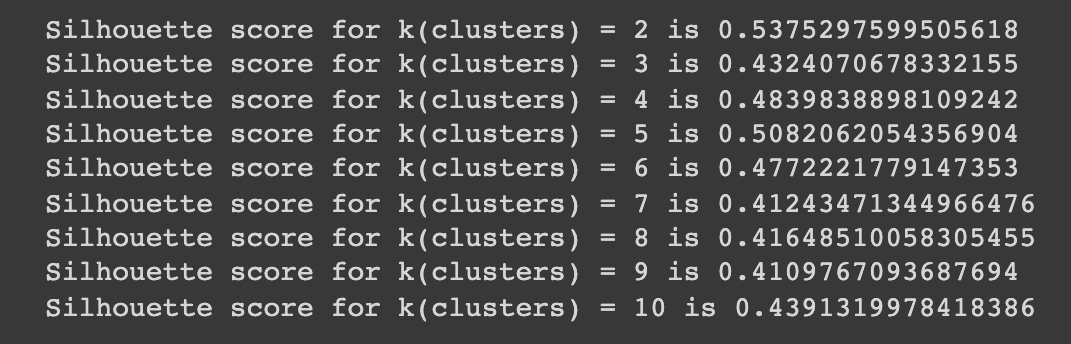


Figure 18: Silhouette score for each k

### Cluster customer segments and visualize analytics results

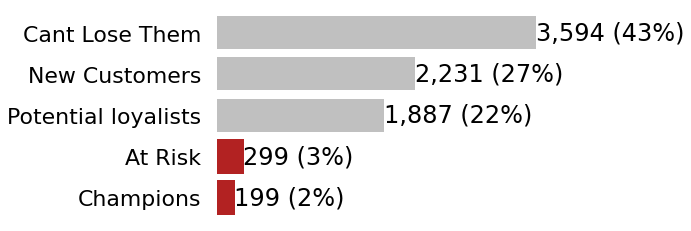
****

Figure 19: The number of customers in each group

We give each cluster a segment based on their characteristic, which will be discussed in detail in the next part. Overall, the most significant percentage of customers was accounted for by the group Can't Lose Them, at 43%, almost twice as high as the number of Potential loyalists customers (22%). Then followed by the percentage of New customers with 27%. On the contrary, the two groups At Risk and Champions amounted to 3% and 2% respectively. It can be seen that the business of this company in the US market has not been steady in the long term.

Analyze and visualize clustering results with scatter plots on three-dimensional space. The results show 5 clusters of customer segments with features in each cluster.

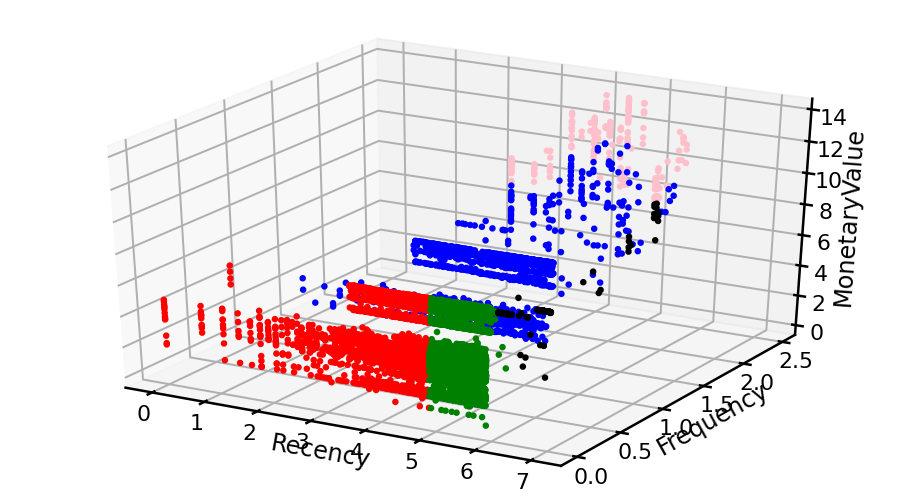


Figure 20: Scatter Plot for customers segmentation

Cluster 0 points is green with 3594 points - can’t lose them

Cluster 1 points is blue with 1887 points - Potential loyalists

Cluster 2 points is red with 2231 points - new customers

Cluster 3 points is black with 299 points - At risk customers

Cluster 4 points is pink with 199 points - Champions

The clustering results are visualized on the figure 20, with the density of points of clusters 0 and 2 being considered the most stable among the 5 divided clusters with the highest number of points, thick and many similarities in the other cluster. However, in both these clusters, the frequency of purchase (Frequency) is low as well as the amount of value (Monetary) brought to the company is not high. In the Recency feature, cluster 0 shows that the customer group has rarely purchased recently, while cluster 2 is the opposite, with a large number of recent purchases. Thus, seeing that despite the similarities, cluster 0 seems to indicate that the customer group seems to be gradually leaving the company. As for cluster 2, it shows a new group of customers, they are attracted but still do not have much trust in the company's campaigns.

Next, in cluster 4, although the number of points is the smallest, only 199, but when it comes to Monetary attribute, it has a rather high positive value, the highest among the 5 clusters. Besides, the frequency (Frequency) of buying is the highest among the 5 clusters, but the index showing the level of recent purchase (Recency) is also at an average level that is not too high. As such, this can be a loyal and potential customer group that can consider if the company has new campaigns or **may** help strengthen the company's brand.

In cluster 1, represented as the blue points with 1887 points, many points are far from the core of the cluster. Frequency of transactions is just relative, but most of them bring much value to the company. This group of customers is also considered to be returning to buy in the near future. Thus, this customer group is considered potential and can become loyal customers if there are reasonable campaigns.

Finally, cluster 3 with 299 black points, many points is far from the core, with frequency and monetary value is considered high. However, it takes a long time for this group to come back for transactions, as can be seen in Recency. This can be a customer group that presents a challenge to the company with the problem of how to attract them to buy regularly in the near future.

In the following results, we will focus on analyzing the characteristics of the respective clusters of customer groups. The names are labeled for the following customer groups (segments) based on the quartile descriptor characteristics and are the most general description of the characteristics of each customer segment. Detailed characteristics of each group of customers are labeled and analyzed in the next part.

## Analysis of the groups of customers based on the quartile description

### Analyzing the group Can’t Lose Them (cluster 0)

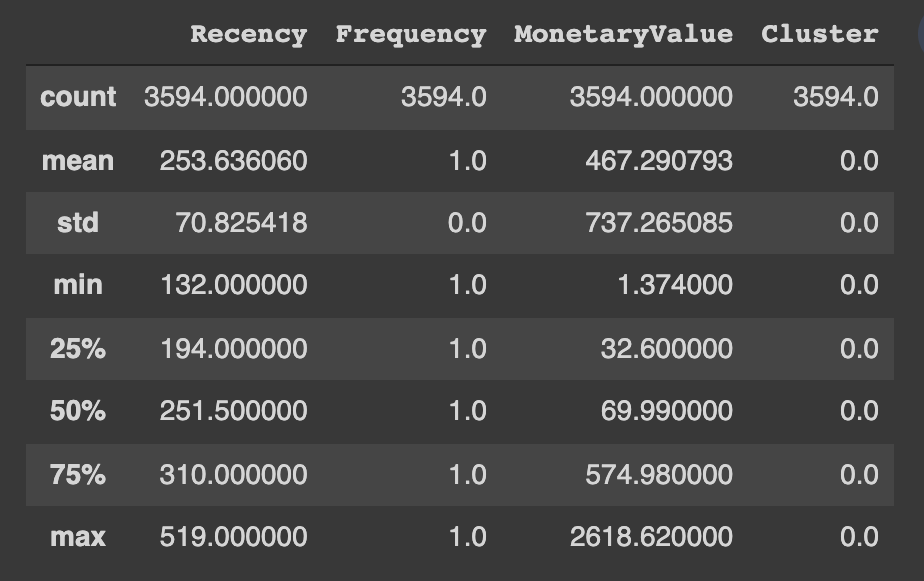
****

Figure 21: Quartile description in cluster 0

Some characteristics in this customer group are worth paying attention to compared to the other two customer groups shown in the figure above.

In which, this customer segment has the number of 3594 customers, accounting for 29.85% of the total number of customers, which is the largest customer group. From the above results, some characteristics of this customer group can be drawn. In which, the latest purchase date is high, this group of customers has not purchased for a long time. The frequency of purchases is very low and tends to stay low, specifically, the average Rencency is very high, on average it has been more than 253 days, equivalent to more than 8 months of no shopping activity. Spending is the lowest of all segments, accounting for 2.67% of revenue. The average frequency is very low, about once in the years of data collection, even in which 75% of customers in this group only shop at most once a year. This can be seen as a customer group that brings many risks as well as challenges to businesses. The value contribution of this group of customers is not high and not outstanding, but accounts for 28.98% of the total number of customers of the enterprise.

Because of the largest customers, the company should not ignore this Cluster. Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor. This can be applied together with up-selling and cross-selling strategies.

### Analyzing the group Potential Loyalists (cluster 1)

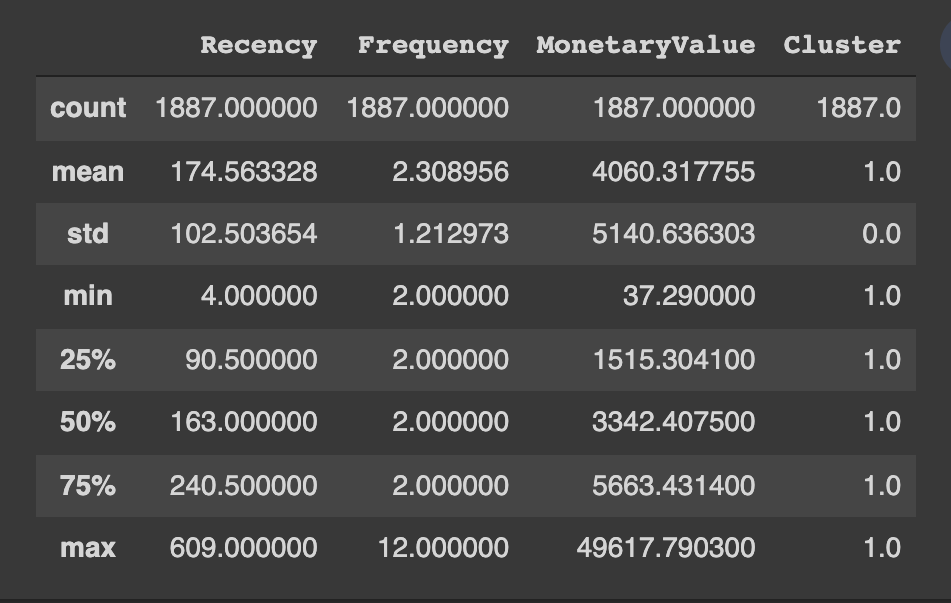
****

Figure 22: Quartile description in cluster 1

According to quartile description, this group has 1887 customers, accounting for 15,67% of all customers. On average, this group of customers usually has the nearest purchase date of approximately 175 days, the average purchase frequency is 2.3 times higher than the other groups, and the customer group may be willing to spend a lot of money on shopping.

With the characteristics of Recency, Frequency and Monetary, we can see that this is not only a loyal customer group and even a customer group that brings great potential to businesses. Although this group of customers only accounted for 15.67% of the customers, the revenue they brought in accounted for 12.16%. Plus another advantage is that the Recency of this cluster is still low compared to the common ground, which means they still tend to come back for the next purchase. The company should offer membership or loyalty programs or recommend related products to upsell them and help them become your Loyalists or Champions.

### Analyzing the group New customers (cluster 2)

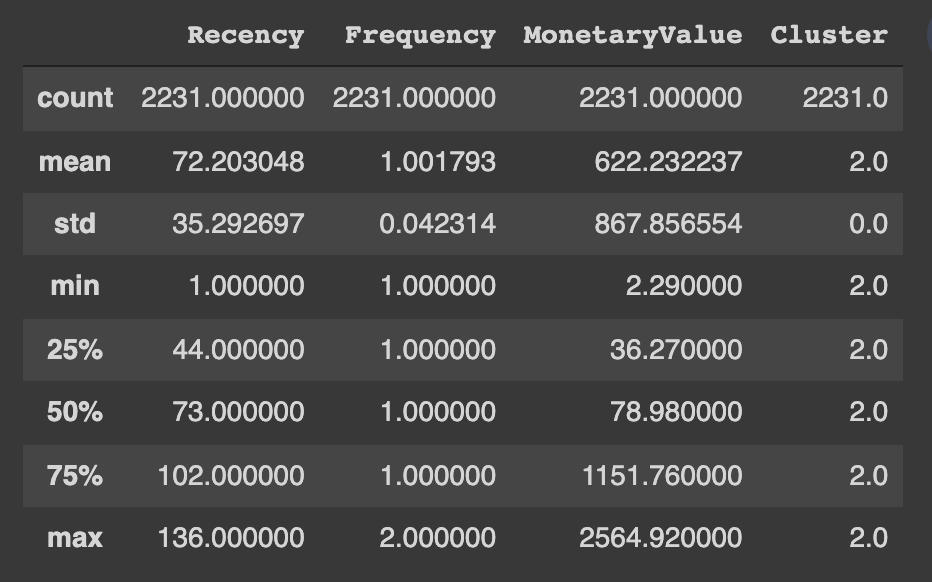
****

Figure 23: Quartile description in cluster 2

This is a customer group with 2231 people accounting for 18% of the total number of customers. In which, according to the analysis results in the figure, the spending level is not too high, although the total number of customers in the group is still quite large, the revenue only accounts for 2.2% of the total revenue. The frequency of purchase is still once, low, but the recency index of this group is at the highest level, which means this group of customers very often come to the store. With this group of customers, businesses can continue to improve their current sales policies to retain the group. Besides finding out potential customers in this group and promoting them to become loyal customers. Start building relationships with these customers by providing onboarding support and special offers to increase their visits.

### Analyzing the group At risk customers (cluster 3)

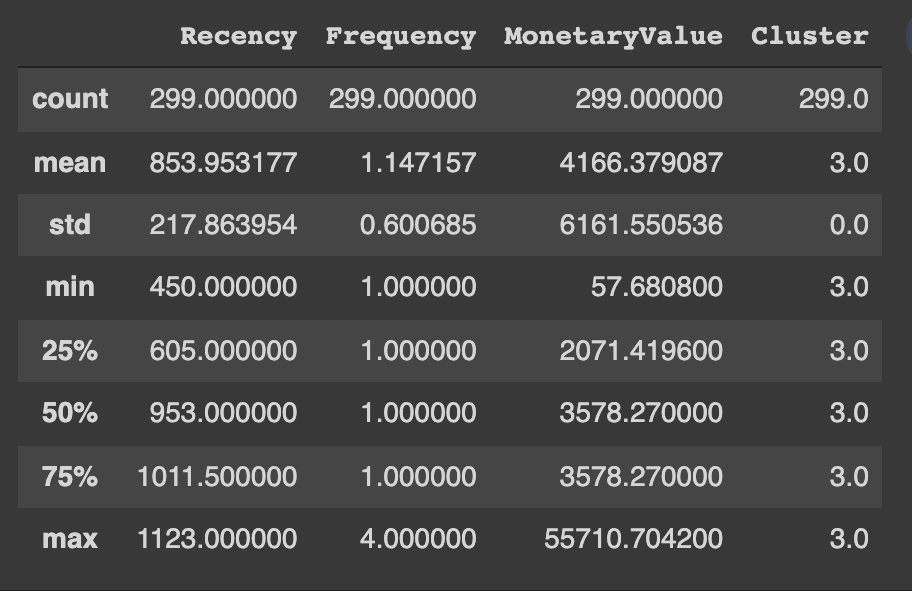
****

Figure 24: Quartile description in cluster 3

In this customer group, according to the results, the number of customers is only 299 with the revenue from this customer group reaching 1.97% of the total revenue of all customers. Besides, the recency index is quite high, which means that the probability of customers coming to the store is quite low, but there are still customers who spend very high up to 55710.7 units. This is also an opportunity and a challenge for businesses to be able to attract potential customers from the group and keep them coming back again. Send them personalized reactivation campaigns to reconnect, and offer renewals and helpful products to encourage another purchase.

### Analyzing the group Champions (cluster 4)

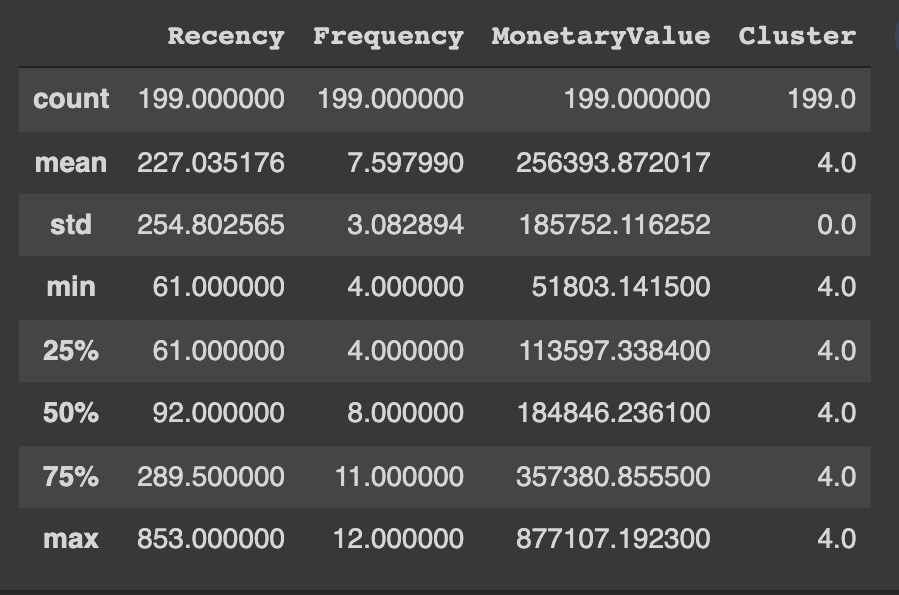


Figure 25: Quartile description in cluster 4

This cluster is your best customer, who bought most recently, most often, and are heavy spenders. With consumption up to 80.99% of total revenue, although the number of customers in this group only accounts for 1.65% of the total number of visitors. The average buying frequency is over 7 times as well as the average business spending is also very high, about 256393.87 money units. According to the Pareto principle that 20% of customers will bring 80% of sales, this group of customers is called VIP, Priority or high-class customers, depending on the business. This company should reward these customers to retain them. Moreover, they can become early adopters for new products and will help promote your brand. However, the number of customers in this cluster only accounts for a small part of the total number of customers, the company may need to have policies to change products to catch up with market tastes or strengthen marketing to attract customers. consume more. A retention strategy for the most profitable customers, such as delivering exceptional proactive customer service, loyalty and rewards programs, and more frequent follow-up, as well as personalized marketing.

# **Chapter 3: RFM model for customer segmentation by Power BI**

## Customer segment chart

Model RFM applies metrics to each customer based on their purchase history, so we will use DAX in Power BI to calculate Recency, Frequency and Monetary values ​​for each customer . We will calculate the metrics in turn according to the following formula.

* ***Recency***: calculate the number of days from the last order of that customer to the current time (6/30/2014).
* ***Frequency***: calculate the number of orders of customers in a certain period of time (e.g. half a year back, 1 year back, or on sales history). In this model, it will simply calculate based on the sales history. row.
* ***Monetary***: calculate the total purchase value of customers in a certain period of time or on the whole sales history. In this model, it will simply calculate the sales history.

After we've created these key metrics, we'll categorize each metric. There are many ways to classify each metric and from each classification each metric can be created into a different set of customers. For example, there are many industries where repeat customers may have to make weekly purchases, and there are others that define this as monthly. We can apply customer insights to make logical rules when classifying each value, or we can also rely on statistics to classify. Here, this model is based on Business Rules to classify customer sets.

After each customer has an RFM classification score from 1 to 4, the higher the score, the more valuable the customer is according to each indicator. From here we can combine this set of points to create customer sets.

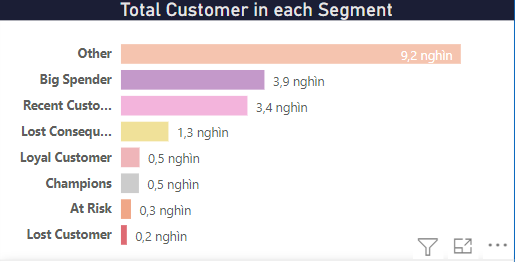


Figure 26: Shows the customer sets and the number of customers by volume

Including:

*Table 2: Customers Segmentation based on RFM score*

|  |  |  |
| --- | --- | --- |
| Segment | Description | R-F-M |
| Big Spender | Customers who spend the most | 1-1-4, 1-2-4, 1-3-4, 2-1-4  2-2-4, 2-3-4, 3-1-4, 3-2-4  3-3-4, 3-4-4, 4-1-4, 4-2-4  4-3-4 |
| Recent Customer | Customers who have purchased recently | 4-1-1, 4-1-2, 4-1-3,  4-2-1, 4-2-2, 4-2-3,  4-3-1, 4-3-2, 4-3-3 |
| Lost Consequential Customer | Customers who bought a long time ago, may have bought a few times and spent little | 1-1-1 |
| Loyal Customer | Customers who buy most often | 1-4-1, 1-4-2, 1-4-3, 2-4-1  2-4-2, 2-4-3  3-4-1, 3-4-2, 3-4-3  4-4-1, 4-4-2, 4-4-3 |
| Champions | Customers who have purchased most recently, spent the most and most often | 4-4-4 |
| At risk | Customers didn't buy often, but they bought a lot and spent a lot in the past | 2-4-4 |
| Lost Customer | Customers have not bought often for a long time, but before they bought a lot and spent a lot | 1-4-4 |
| Others | Remaining customers |  |

We can see that the customers with the highest score 4-4-4 are rated the highest because they have just ordered a large value, bought often and recently. This is a group of customers who bring a lot of profit to the company and need good customer service or send many appropriate purchase recommendations to maintain.

Some new customers can be easily identified with a set of points with R = 1. And can be classified as potential new customers when the order volume is high or the purchase value is large. They may also just be new customers trying out and buying some essential small products.

Customers who have not placed orders for a long time and have R = 4 can also be classified if they are important customers who have made many large purchases. Or if R =3, it means that the customer is in a situation that needs attention, maybe with discount packages to encourage purchase.

## Explain some of the charts in the report

The RFM model visualization chart focuses on understanding individual customer groups or from customer groups to find more detailed purchasing behaviors to create tailored marketing campaigns or find out which customer groups are suitable for which form of marketing. In this model, we will create 3 more charts to visualize customer behaviors and information in a more understandable and objective way:

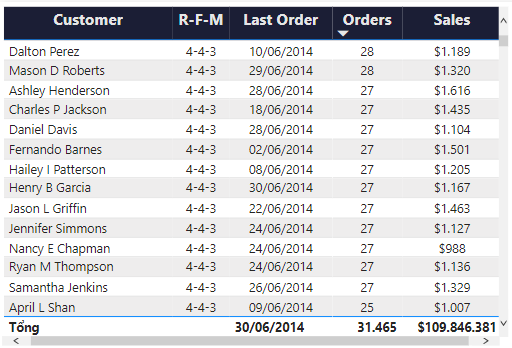


Figure 27: Shows a detailed table of each customer

Figure 27 shows us know with indicators of purchasing behavior such as:

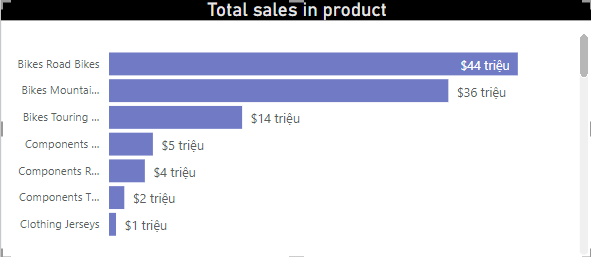
* Have the customers purchased recently?
* Do customers buy often?
* Is the product purchased by the customer of great value?

Figure 28: Shows the number of orders and the total price purchase value over time

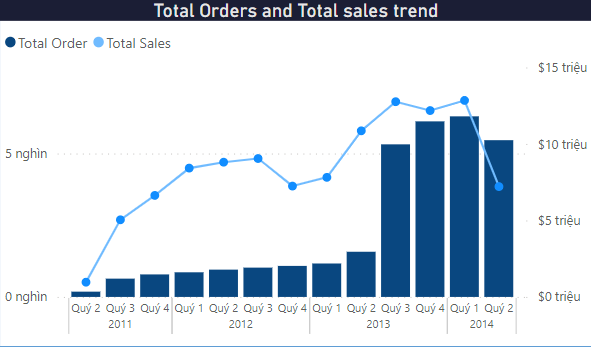
From Figure 28, we can see each customer group about the development or learn the purchase history of a certain customer.

Figure 29: Shows the total purchase value for each product and product category.

Figure 29 helps us to know what each of those customers bought and often bought on the system.

## Report about customer segmentation

After completing the above charts, we have created a Report. Looking at the report table, we can determine general information such as the number of customers, total sales, total orders, quantity sold

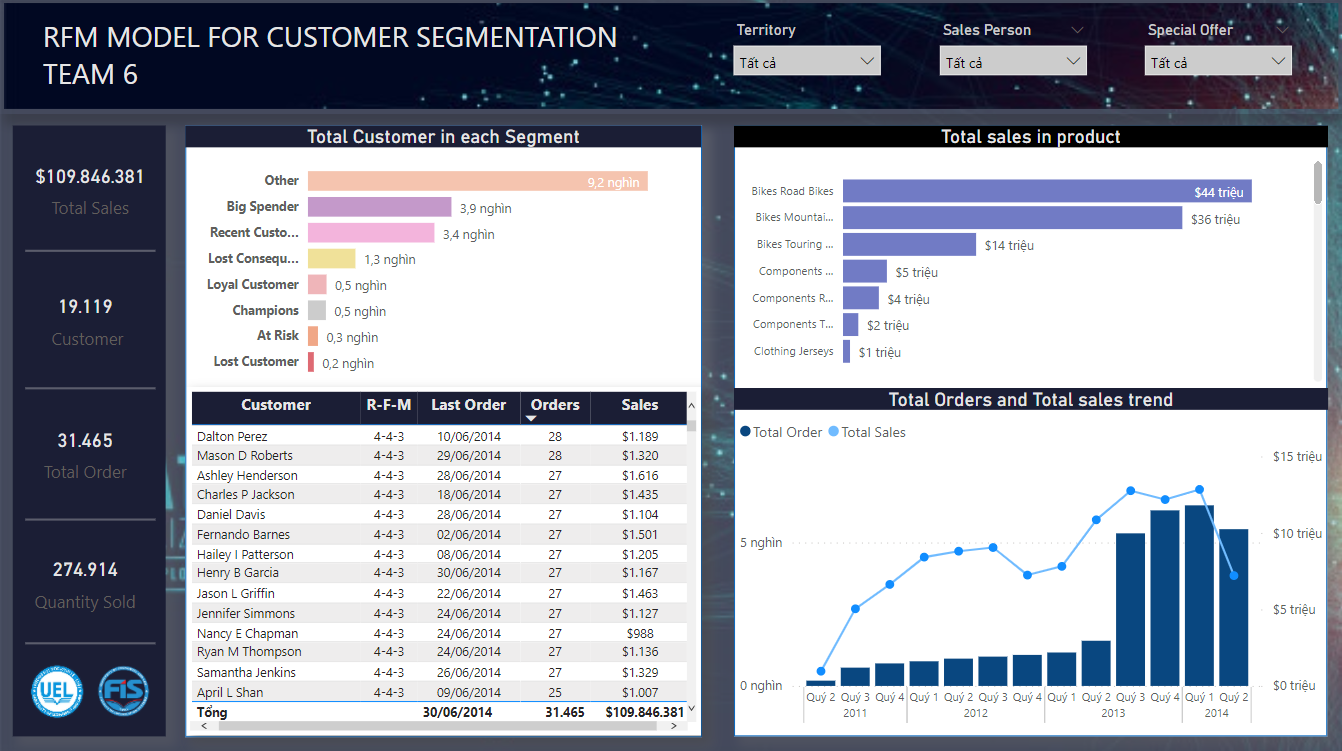


Figure 30: RFM model for customer segmentation

In addition, we can extract information from the report such as:

* Which segment does that customer belong to?
* What products did the customer buy?
* How many times has that customer ordered and in what time period (Quarterly)
* In that quarter, what products did you sell and how much money did you sell?
* What products do customer segmentation usually buy and like?

From all the information above, we can analyze and give appropriate insights to have more effective marketing campaigns.

# **Chapter 4: Predicting CLV**

Calculating Customer Lifetime Value or CLV is considered a really important thing in marketing and ecommerce, yet most companies can’t do it properly. This clever metric tells you the predicted value each customer will bring to your business over their lifetime, and as such requires the ability to detect which customers will churn as well as what they’re likely to spend if they’re retained.

## Calculating CLV

Here we’re going to use models BG/NBD model and the Gamma-Gamma model - to predict the following:

1. *Which customers are still customers*
2. *Who will order again in the next period*
3. *The number of orders each customer will place*
4. *The average order value of each customer’s order*
5. *The total amount each customer will generate for the business*

Calculate the raw recency, frequency, and monetary metrics

The next step is to turn our raw transactional data into the recency, frequency, monetary and tenure (T) metrics we need to provide to the CLV models. While it’s relatively straightforward to do this manually using Pandas. After using, we have a table like that

*With T is the tenure of the customers or when they were acquired*

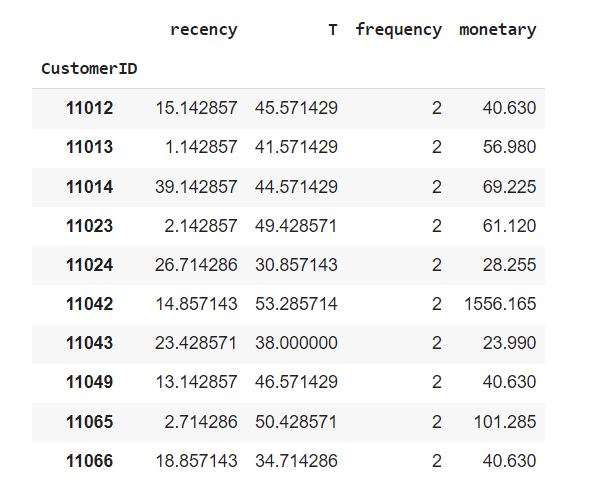


Figure 31: Top 10 values of the dataset1.

***Examine the statistical distribution of the data***

Next, it’s a good idea to take a look at the [***statistical distribution***](https://practicaldatascience.co.uk/data-science/how-to-visualise-statistical-distributions-with-seaborn) of your data. This dataset is relatively typical for an ecommerce retailer. The data are strongly skewed, with most customers having placed a single low value order, and a long tail of higher value repeat customers contributing the rest. In this data, recency is valued by weeks, not days, and we will call it dataset1.

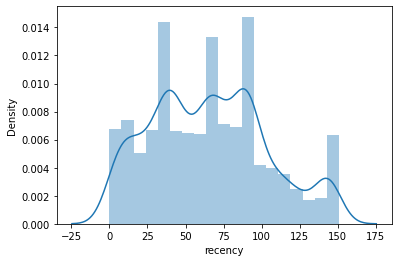


Figure 32: Plot for recency in dataset1.

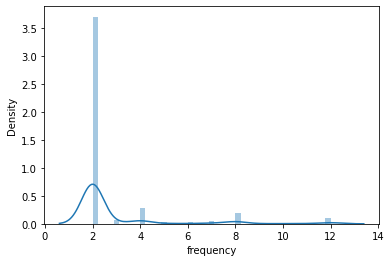


Figure 33: Plot for frequency in the dataset1.

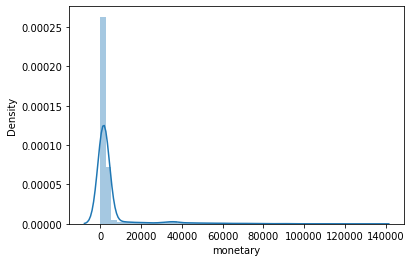


Figure 34: Plot for monetary in the dataset1.

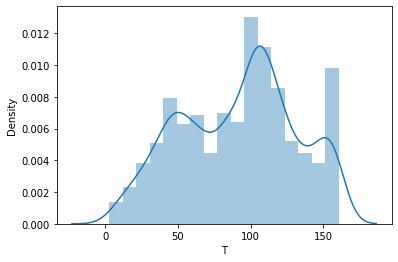


Figure 35: Plot for T values in the dataset1.

## Fit the BG/NBD model

We can fit the model using the frequency, recency, and T data and then view a summary. The model will help us predict which customers will return in the near future

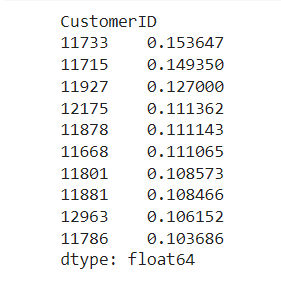
**

Figure 36: Customers who will return within the next 1 week

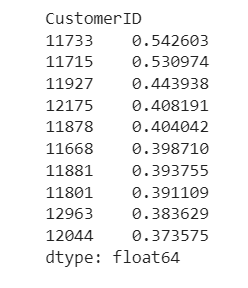
**

Figure 37: Customers who will return within the next (1 month = 4 weeks)

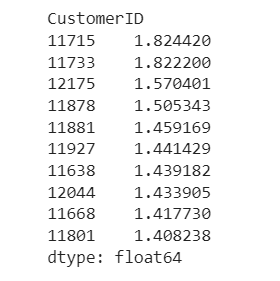
**

Figure 38: 10 customers expect to buy the most in the next 6 months

**

Figure 39: Total transactions made in the next 6 months

## Fit the Gamma-Gamma model

First we will collect a list of 10 customers that bring the most value

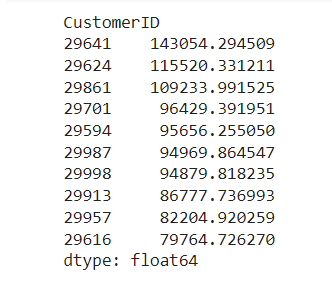
**

Figure 40: Top 10 customers that bring most values

***Predict average order value***

If you run the conditional\_expected\_average\_profit() function on the fitted ggf model you get back the predicted average order value for each customer. If you have customer-level profit data, you can also fit the model using this data instead of their revenue, providing it’s the mean and is consistently calculated.

* **Definition*:****Average Order Value (AOV) is an ecommerce metric that measures the average total of every order placed with a merchant over a defined period of time.*[*AOV is one of the most important metrics*](https://www.bigcommerce.com/blog/4-tips-increase-average-order-value-boost-revenue-video/)*for online stores to be aware of, driving key business decisions such as advertising spend, store layout, and*[*product pricing*](https://www.bigcommerce.com/blog/value-based-ecommerce-pricing-strategy/)*.*

Now we’ll take the Series returned by the above function and assign it to a dataframe, placing the predicted AOV value in its own column and leaving the customer\_id value in the index. If you sort the aov column in descending order and reassign the sorted data back to the dataframe, you can then look at the top and bottom of your customer list.

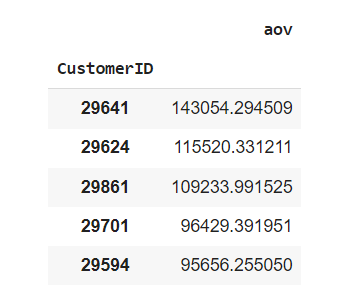
**

Figure 41: The top aov values

*Here’s a plot of their AOVs, which follows the pattern you’d expect to see on a long-tailed dataset like this one. Most of our customers are going to spend only a modest amount, but there’s a long-tail of higher value customers and a little group of really big spenders on the far right.*

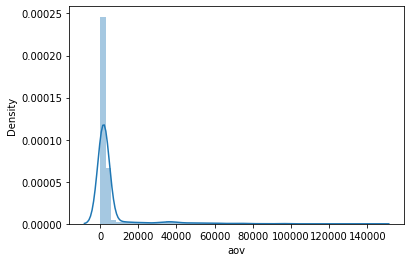


Figure 42: Plot for aov scores

## Segment customers by their AOV

To get an idea of the spread of AOV across customers we can use [quantile-based discretization](https://practicaldatascience.co.uk/data-science/how-to-assign-rfm-scores-with-quantile-based-discretization) or binning via qcut(). We can take the data frame above and create five bins labeled 1 to 5, reset the index, and then create an aggregation to calculate the summary statistics based on the aov\_bin.

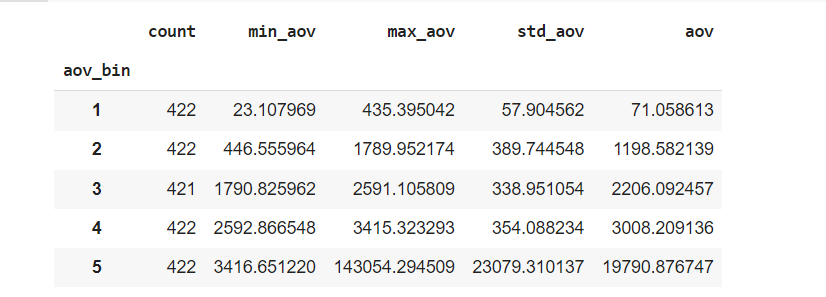


Figure 43: Quartile description of 5 bin of aov

Bin 1 contains the low CLV customers who have an average value of 71.06, the averages go up in each bin, rising to an average of 19790.88 in bin 5. However, these include customers from 3415.32 to 143054.29, so there’s quite a spread at the top end.

***Applying K-means clustering***

Using [K-means clustering](https://practicaldatascience.co.uk/machine-learning/how-to-use-k-means-clustering-for-customer-segmentation) is another way to get a picture of the data. While the cluster numbers aren’t sorted in order, you can clearly see the differences in AOV across the five clusters. Cluster zero contains low spenders, averaging ~1921 per order, while cluster 3 includes some very big spenders.

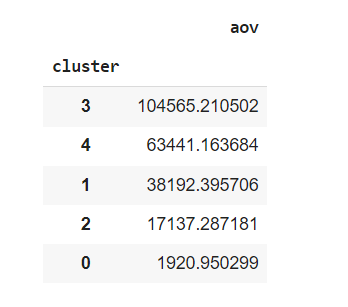


Figure 44: 5 clusters by aov points

## Predicting Customer Lifetime Value

Finally, we’ll get to the good bit - predicting Customer Lifetime Value. To do this we’ll use both models - the BG/NBD model to predict the number of orders and the Gamma-Gamma model to predict their values. First we’ll re-fit the BetaGeoFitter BG/NBD model to our dataset for the returning customers, which includes the monetary data.

Next, we’ll use the Gamma-Gamma model’s customer\_lifetime\_value() function to predict their value. We’ll pass in the bgf model, along with the recency, frequency, monetary\_value, and T data for each customer. The time parameter is in months and defines how many months into the future we wish to predict.

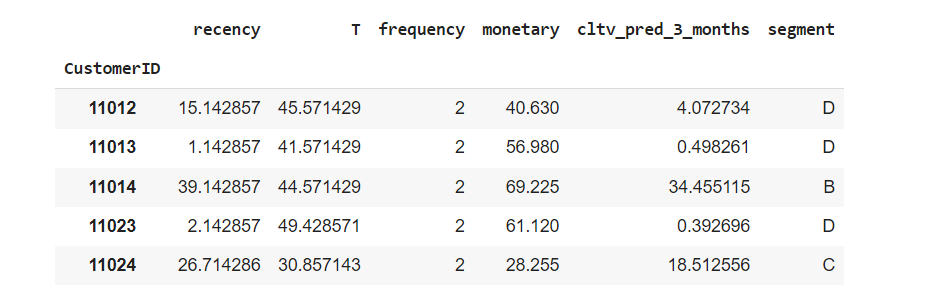


Figure 45: The customers' lifetime values expected to in the next 3 months

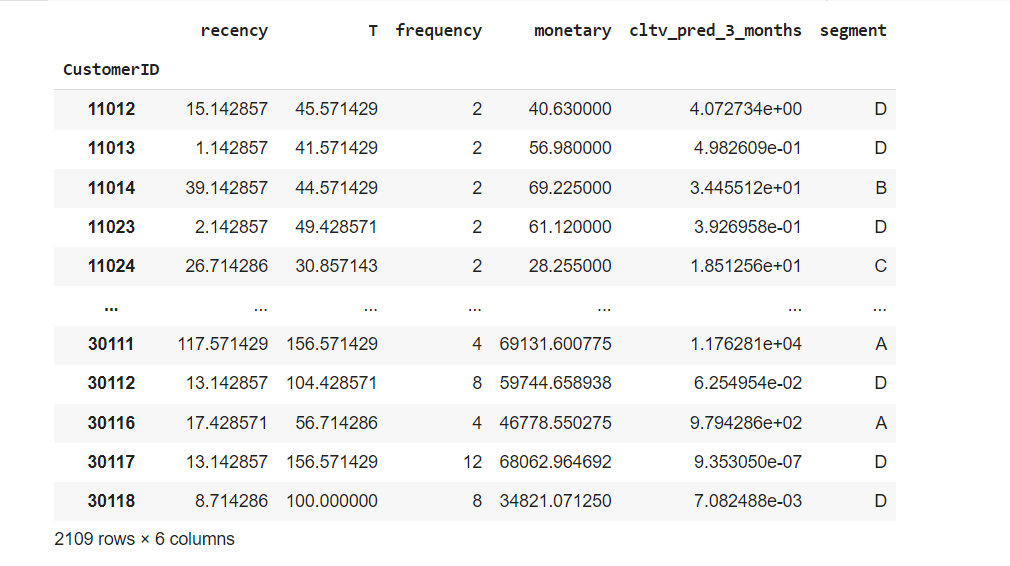


Figure 46: Customer segmentation by CLV

***Insight:***

* Here we see, depending on the business rules of each business, there are different customer segments
* For this time, I divided the customer segment into 4 groups ( A, B, C,D )
* With:

A: Customers that have highest values

B: Loyal customers

C: Customers that have average values

D: Customers that have lowest values

***Advantages***

* Easy to predict
* Specification
* Clear segment

***Disadvantages:***

* Customers often vary dramatically in their behaviour - some order weekly and others order annually. There’s no such thing as “typical customer.” If a customer normally orders from you every week and hasn’t shopped for a year, do you really think they’re still alive?
* Because customer behavior is unpredictable

# 

# **Conclusion and Future works**

From the results of the customer segmentation built on these studies, the principle traits of every customer organization are provided in conjunction with the advertising and marketing techniques for every specific organization. We should see that the five clusters and five customer organizations assist the agency lots to apprehend the clients withinside the experience of clients demographic and the behaviours withinside the methods that RFM evaluation became presented, which simply best the bought date frequency, and quantity of spending. Many businesses additionally stated that the advised advertising and marketing techniques are affordable primarily based totally at the evaluation; nevertheless, if the customer’s desires and motivations may be defined, it's going to assist lots for growing the advertising and marketing method.

According to the segmentation version, this studies became composed with the idea of the demographic and behavioural segmentation. Hence, it no longer offers the insights of customers in phrases of desires and motivations at the back of the acquisition behaviour, which it's far vital for the advertising and marketing method development.

According to this example have a look at research, the primary issue to recollect while acting the client segmentation is the shape of records, due to the fact if the gathered records changed into saved withinside the non-based format, it'll devour numerous time withinside the records processing system. The 2d issue is the exceptional and the variety of times withinside the pattern records may be very critical for the clustering method, due to the fact clustering method is an unmanaged getting to know set of rules which means the set of rules classify each tangible and intangible styles and relationships among the records with none previous records. Therefore, the accuracy of clustering relies upon the exceptional and the quantity of records. Lastly, the researchers should sincerely recognize the commercial enterprise of your case, due to the fact withinside the interpretation system the in-intensity client records should be used to investigate and interpret in conjunction with the segmentation results, mainly while the advertising and marketing approach wishes to be developed, the whole thing approximately products, customers, and income records want to be understood.

# **References**

**[1]** <https://www.sciencedirect.com/science/article/pii/S1319157819309802> (RFM model for customer purchase behavior using K-Means algorithm)

**[2]** <https://www.sciencedirect.com/science/article/pii/S1877050910003868> (Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study) -> Cái này lấy làm form viết báo cáo

**[3]** <https://www.sciencedirect.com/science/article/pii/S1319157818304178> (RFM ranking – An effective approach to customer segmentation)

**[4]** <https://www.researchgate.net/publication/340356463_Clustering_optimization_in_RFM_analysis_Based_on_k-Means> (Clustering optimization in RFM analysis based on k-means )

**[5]** <http://www.tailieumienphi.vn/doc/mot-nghien-cuu-lien-nganh-giua-phan-tich-phan-khuc-khach-hang-trong-marketing-va-8hciuq.html> (Một nghiên cứu liên ngành giữa phân tích phân khúc khách hàng trong marketing và phương pháp học máy)

**[6]** <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0E7F08C76E43D509EC7F621124397DB2?doi=10.1.1.736.3182&rep=rep1&type=pdf&fbclid=IwAR0hH1Vd1g0RVok9F1XfBB2Xsf54oioCV0DIqm6ZuKW-XpxdjtbPYWd-cqg> (Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services)

**[7]** <http://brucehardie.com/notes/025/gamma_gamma.pdf> (Fader, P.S. and Hardie, B.G., 2013. The Gamma-Gamma model of monetary value.)

**[8]** Fader, P.S., Hardie, B.G. and Lee, K.L., 2005. “Counting your customers” the easy way: An alternative to the Pareto/NBD model. Marketing science, 24(2), pp.275-284.

**[9]** <https://reader.elsevier.com/reader/sd/pii/S1877042812035707?token=4AA4C0162CB170EB74B7553DC177B55D7C34E5FEEBDD30AF3ADBDE2A5413F547C45A01150F0C58F982542E09213D344D&originRegion=eu-west-1&originCreation=20211222141424> (Investigating two customer lifetime value models from segmentation perspective)

**[10]** <https://practicaldatascience.co.uk/data-science/how-to-calculate-clv-using-bgnbd-and-gamma-gamma>(Calculating CLV using BG/NBD and Gamma - Gamma)

**[11]** <http://www.ijmo.org/papers/362-M011.pdf> (How to calculate CLV)

**[12]** <https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1238&context=pare> Improving your data tr our data transformations: Applying the Bo ansformations: Applying the Box-Cox transformation

**[13]** <https://www.frontiersin.org/articles/10.3389/fams.2015.00012/full> (A new approach to the Box–Cox transformation

**[14]** <https://iopscience.iop.org/article/10.1088/1742-6596/1361/1/012015/pdf?fbclid=IwAR1TGyFjDEsWRrX4EHuumlWBalsyv55bOj1pUzvqkMKVIk7YU_a2TeCfZxE> (Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method To cite this article: Rena Nainggolan et al 2019 J. Phys.: Conf. Ser. 1361 012015 )

**[15]** <http://dspace.bu.ac.th/bitstream/123456789/4003/1/ponlacha_rojl.pdf?fbclid=IwAR0hek-SBYhQd42Va49rLUsfn7F-nN5qeu7o4On-6O6R71QjI1sUknDiUCM> (CUSTOMER SEGMENTATION BASED ON THE RFM ANALYSIS MODEL USING K-MEANS CLUSTERING TECHNIQUE: A CASE OF IT SOLUTION AND SERVICE PROVIDER IN THAILAND)