UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

FINAL PROJECT REPORT

DATA ANALYTICS WITH R/PYTHON COURSE

TOPIC: CUSTOMER SEGMENTATION ANALYSIS OF BRAZILIAN E-COMMERCE BUSSINESS WITH RFM MODEL USING K-MEANS CLUSTERING

Lecturer:

Nguyen Phat Dat, MA.

Tran Le Tan Thinh, TA.

Group: Num

Ho Chi Minh City, 6th June 2022

Members of Group Num

No.	Full Name	St. ID	Task	% Done
1	Vo Chi Giang (Lead)	K194111533	- General work management. - Complete Chapter 3, 4 of the project. - Support to complete sections and general check.	100%
2	Nguyen Ba Thinh An	K194111517	- Complete Chapter 2 of the project - Report formatting and slides designing. - Support to complete Chapter 3.	100%
3	Pham Minh Dat	K194111530	- Complete Chapter 1 of the project. - Support to complete Chapter 2.	100%
4	Tran Thi Thao Ly	K194111546	- Complete Chapter 4 of the project. - Support to complete Chapter 5.	100%
5	Trinh Thi Tam Oanh	K194111560	- Complete Chapter 5 of the project. - Report formatting and slides designing. - Support to complete Chapter 5.	100%

ACKNOWLEDGMENTS

First of all, the team would like to especially thank Mr. Nguyen Phat Dat (lecturer of Data Analytics with R/python course) and teaching assistant Tran Le Tan Thinh. The teachers provided knowledge, guidance and valuable suggestions to help the group complete our final project.

Stemming from the purpose of learning, deepening the knowledge of Data Analytcis with R/python, the team decided to choose the topic "CUSTOMER SEGMENTATION ANALYSIS OF BRAZILIAN E-COMMERCE BUSSINESS WITH RFM MODEL USING K-MEANS CLUSTERING". During the implementation of the project, based on the knowledge provided by the teachers in the lecture hall combined with self-study of new tools and knowledge, the team tried to perfect the report in the best way.

However, with the limited ability of the members ourselves, the project will be incomplete and there will be many errors, but it is the result of the efforts of the team members, as well as the help of all friends and teachers.

The group is looking forward to receiving suggestions from the teachers in order to draw valuable experiences and perfect our knowledge so that the group can go further in learning Data Analytics with R/python in particular, and other related studies in general, as well as serve as a basis for future work orientation.

The team sincerely thanks the two teachers!

Best regards!

Group Num

COMMITMENT

The group hereby declares that the final project on the topic " CUSTOMER SEGMENTATION ANALYSIS OF BRAZILIAN E-COMMERCE BUSSINESS WITH RFM MODEL USING K-MEANS CLUSTERING" is a stydy of the group that is conducted publicly based on the efforts of all members of the group and the great help from the lecturer Nguyen Phat Dat and teaching assistant Tran Le Tan Thinh during the course period.

The dataset group used for analysis in the project is publicly published; and the research and analysis results are self-explained and analyzed by the group in an objective, honest way, with a clear origin and not yet published in any form. All help and references for the development of the project are fully cited and clearly stated with clear origin and are allowed to be published.

The team will be fully responsible for any dishonesty in the information used in this study.

Ho Chi Minh City, 6th June 2022.

Group Num

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PROJECT OVERVIEW

Chapter summarization

In this chapter, we will introduce the reasons for doing the research about the knowledge of customer segmentation based on RFM, K-means techniques through the objectives set out with a dataset from a business. Besides, in order to make the research paper clear and useful, we will also review some previous research papers on the same topic as well as define the objects and scopes of this research and propose suitable methodology.

1.1. Reasons

In marketing analysis or jobs related to management, service, customer care, understanding customers, trying to bring the best products, services and experiences is always the goal that every business is aiming for. However, this journey will always contain many problems that are not even easy to solve. A product or a promotion program when launched on the market can hardly meet the needs of all customers. Therefore, businesses have gradually shifted to dividing customers into separate groups - called customer segments, in order to focus and take better care of customers based on the unique characteristics of each customer group.

With the strong development of current data science technology, the collection and storage of customer data is a valuable resource that is waiting to be discovered and is also a favorable basis for application. Apply mathematical models, algorithms, and machine learning methods to exploit and solve business problems. From data analysis, managers' decisions are more objective and multidimensional. Data-driven decision making is made with less emotion that is difficult to measure. The combination of data analysis based on customer segments has contributed to the success of each marketing strategy or customer care policy in particular and maintains the existence and development of the business in particular, in the context of a common market with a lot of fierce competition.

Customer segmentation helps the marketing departments easily define the pivotal solution to attract each group of customers. Based on the data segmentation, customers are classified into different groups according to distinguishing similarities such as gender,

age, income, products of interest, and purchasing behaviors. 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).

To apply all the above knowledge, our team decided to choose Olist as a business case that our team will focus on. Olist is a Brazilian department store platform which operates in the e-commerce segment (Software as Service). The service consists of management of the sales process between shopkeepers and clients, and also includes a customer satisfaction report. The advantages for the shopkeepers is a better market presence and transparent reputation metrics. The driver for the business is to attract more clients and raise the quality of the process. The motivation in this project is to support this effort.

1.2. Objectives

The analysis behind this article is part of a business case investigation that is focused on supporting Olist's business objectives. These are: Attracting more shopkeepers by enhancing the service and attracting more end-customers through a broader product spectrum and higher satisfaction.

Different methods of segmentation are applied to get different perspectives and to identify the commonly valid picture. By doing so, the following questions must be answered:

- *Marketing strategies are needed to target customer segments?*
- *What are the buying characteristics of customer segments?*
- *Which customer segments brings the most profit?*

1.3. Objects and scopes

The objects and scopes of this project consists of historical order data from 2016 to 2018 and contains 100,000 orders from over 96,000 customers of 4,119 listed cities in Brazil.

1.4. Related works

- T. Jiang and A. Tuzhilin (2009), Improving personalization solutions through optimal segmentation of customer bases*:* Having identified that both customer segmentation and buyer targeting are necessary to improve the marketing performances. These two tasks are integrated into a step by step approach, but the problem faced is unified optimization. This approach focuses on distributing more resources to those customers who give more returns to the company. A sizable amount of authors had written about different methods for segmenting the customers.

- He X. and Li, C. (2016), The research and application of customer segmentation on e-commerce websites: Having suggested a three-dimensional approach to improving the customer lifetime (CLV), the satisfaction of the customer and customer behavior. The authors have concluded that the consumers are different from one another and so are their needs. Segmentation assists in finding their demand and expectations and proving a good service.

- A. Sheshasaayee and L. Logeshwari (2017), An efficiency analysis on the clustering methods for intelligent customer segmentation: Having designed a new integrated approach by segmentation with the RFM and LTV (Life Time Value) methods. They used a two-phase approach with the first phase being the statistical approach and the second phase is to perform clustering. They aim to perform K-means clustering after the two-phase model and then use a neural network to enhance their segmentation.

- Christy AJ et al. (2018), RFM ranking - An effective approach to customer segmentation: RFM method is known as a summary of customer transactions under three factors, including: Recency is considered the last time the customer made a purchase (distance between the date of application of the method). law and the date of the most recent purchase by the customer); Frequency is the frequency of customer purchases or

how many times the customer has purchased; Monetary is the total amount of money that customers have spent on all shopping activities.

- Anitha P and Patil MM (2019), RFM model for customer purchase behavior using K-Means algorithm: Having used K-means clustering method - a method in unsupervised machine learning model to divide customer groups based on three factors in RFM method. Each customer segment is now considered a cluster in K-means.

CHAPTER 2: THEORETICAL FRAMEWORK

Chapter summarization

In this chapter 2, we define some theoretical bases used in the report. With the selected topic, we need to segment the customers of a Brazilian e-commerce business. To solve this problem, we use 2 methods: RFM clustering and K-means clustering. Besides, with the K-means algorithm, it is necessary to pre-determine the number of "k" clusters that will fit the data set. To determine this number, we use Elbow method.

2.1. Customer segment

Definition: Customer segmentation is the process of dividing all customers into distinct groups with similar characteristics, such as demographics, interests, patterns, or location, in order to help a business focus marketing efforts and resources on valuable, loyal customers in order to meet business objectives. Customers' demographic, regional, behavioral, and psychological data can all be used to segment them:

- Demographic characteristics (such as age, gender, and income) are used to identify important and profitable clients for more targeted messaging.
- Geographic segmentation of the traditional grocery market leads to a more realistic business expansion site strategy.
- Behavioral factors (shopping frequency, price payment, coupon redemption, product diversification, and return rate) divide customers into five groups, allowing businesses to tailor marketing techniques to each group.
- Psychological influences on consumer purchasing intentions, such as subject norms, personal norms, and attitudes, identify the pioneer and conservative customer groupings.

In addition to these one-perspective segmentations, multisegmenting was created to disclose smaller, more defined sub-segments. Customers' behavioral and demographic segments were identified using a non-negative matrix factorization approach, and two sets of segments were then combined to produce holistic customer personas. Sub-markets are identified using a combination of psychographic and demographic segmentation to improve competitive business advantages.

Purpose: Marketers may better customize their marketing efforts to different target subgroups by segmenting their audiences. These efforts may involve both communication and product development. A corporation can benefit from segmentation in the following ways:

- Develop and distribute customized marketing messages that will resonate with select groups of customers .
- Depending on the segment, choose the appropriate communication medium, which might be email, social media postings, radio advertising, or another technique.
- Identify potential for product improvement or new product or service development.
- Develop stronger customer ties.
- Experiment with different pricing strategies.
- Concentrate on your most profitable consumers.
- Enhance client service.
- Upsell and cross-sell more goods and services.

2.2. RFM Model

Definition: RFM analysis is a marketing approach that is used to objectively evaluate and classify consumers based on the recency, frequency, and monetary amount of their most recent transactions in order to find the best customers and conduct focused marketing campaigns. To give an objective analysis, the system assigns numerical ratings to each consumer based on these parameters.

RFM analysis ranks each customer on the following critiria:

- **Recency**: How recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their mind and are more inclined to purchase or use it again. Businesses often measure recency in days. However, depending on the product, it might be measured in years, weeks, or even hours.

- **Frequency**: How frequently did this consumer make any purchases throughout a particular time period? Customers who have previously purchased are more inclined to do so again. Furthermore, first-time customers may be effective targets for follow-up advertising in order to turn them into repeat customers.
- **Monetary**: How much money did the consumer spend during a specific time period? Potential customers that spend a lot of money are more likely to spend money in the future and are valuable to a company. Monetary will have a direct impact on revenue and is indirectly affected through the remaining two factors, Recency and Frequency.

Purpose: RFM analysis helps marketers find answers to the following questions:

- *Who are your best customers?*
- *Which of your customers could contribute to your churn rate?*
- *Who has the potential to become valuable customers?*
- *Which of your customers can be retained?*
- *Which of your customers are most likely to respond to engagement campaigns?*

Each customer is ranked in each of these categories, generally on a scale of 1 to 5 (the higher the number, the better the result). The higher the customer ranking, the more likely it is that they will do business again with a firm. Essentially, the RFM model corroborates the marketing adage that "80% of business comes from 20% of the customers."

RFM is a renowned and often used strategy for customer segmentation based on the value delivered to the firm. We may build company strategies and organizational structures based on consumer segmentation to better serve each category. Simultaneously, analyzing the change in customer structure in each segment over time aids in assessing the company's customer base's development level. Consequently, the firm has the essential data to make modifications and plans aimed at raising the proportion of potential consumers, retaining customers returning to shop more frequently, and boosting the value of their orders each transaction.

- The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer
- The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F and M) and calculate rank for a customer. Rank of a customer will be calculated by the average of 3 rank values of R-score, F-score and M-score. The higher this value, the more valuable our customers are to the company.
- The third step is to select and label groups of customers to whom specific types of communications will be sent, based on the RFM segments.

Using RFM modeling can provide valuable insights about customers. However, it does not take into consideration many other aspects of the buyer. In-depth targeted marketing may also include criteria such as the sort of goods purchased or consumer campaign replies. Customer demographics like as age, gender, and ethnicity are also not considered in RFM analysis.

Additionally, RFM only examines customers' historical data and may not anticipate future consumer action. Predictive approaches, unlike RFM analysis, may be able to predict future client behavior.

2.3. Clustering

Definition:

Clustering, as the name implies, involves splitting data points into several clusters having comparable values. In other words, the goal of clustering is to separate groups with similar characteristics and put them into various clusters.

It is ideally the application of human cognitive capability in machines, allowing them to recognize and discriminate between various objects based on their natural qualities. Unlike humans, machines have a tough time distinguishing between various things unless properly educated on a large relevant dataset. Unsupervised learning algorithms, notably clustering, are used to accomplish this training.

The common clustering algorithms include K-means clustering , Expectation-Maximization (EM) clustering and Hierarchical clustering (Zhou J. et al.,2021).

Purpose:

When working with enormous datasets, dividing the data into logical groupings, or clusters, is an effective approach to study. You could extract value from a big quantity of unstructured data in this manner. It allows you to quickly scan the data for patterns or structures before delving further into the data for particular results.

Data clustering aids in discovering the underlying structure of the data and applications across industries. Clustering, for example, may be used to categorize sickness in the realm of medical study, as well as in consumer categorization in marketing research.

In certain applications, data partitioning is the end aim; however, clustering is also a prerequisite to prepare for other artificial intelligence or machine learning problems. It is an effective technique for knowledge discovering in data in the form of recurring patterns, underlying rules, and more.

2.4. K-means

Definition: K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a straightforward procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand. The clusters are then positioned as points, and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process is repeated until the desired result is obtained.

K-means clustering tries to make the intra-cluster data points as similar as possible while also keeping the clusters as far as possible. As a consequence, the less variation we have within clusters, the more homogeneous data points inside the same cluster are.

The algorithm:

- K points are placed into the object data space representing the initial group of centroids.
- Each object or data point is assigned into the closest k.
- After all objects are assigned, the positions of the k centroids are recalculated.

- Steps 2 and 3 are repeated until the positions of the centroids no longer move.

Purpose: K-means clustering is used mainly in statistics and can be applied to almost any branch of study. For example, in marketing, it can be used to group different demographics of people into simple groups that make it easier for marketers to target. Astronomers use it to sift through huge amounts of astronomical data; since they cannot analyze each object one by one, they need a way to statistically find points of interest for observation and investigation.

2.5. Elbow method

A measure to express the accuracy of the clustering method is the Sum of Squared Errors. It describes how well the clustering model fits the data. The SSE in K-Means Clustering is depending on the number of selected clusters. The goal is to select a number of clusters that is as small as possible, but one that still archives a significant improvement if fitting the data.

If there is no idea about the optimal value of k, then there are various methods to find the optimal value of k. In this article, we will cover **Elbow Method:** A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of K. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. To find the optimal number of clusters (k), observe the plot and find the value of k for which there is a sharp and steep fall of the distance. This is will be an optimal point of k where an elbow occurs.

CHAPTER 3: DATASET AND PROPROSED RESEARCH MODEL

Chapter summarization

In this study, we proposed a research model with 2 phases: Data preprocessing and RFM model setup, Clustering with K-means algorithm. We introduce to the dataset "**Brazilian E-Commerce Public Dataset by Olist**" with 116,581 records of customer transactions from 2016 to 2018.

3.1. Research model

Figure below describes methodology and proposed research model with two main stages:

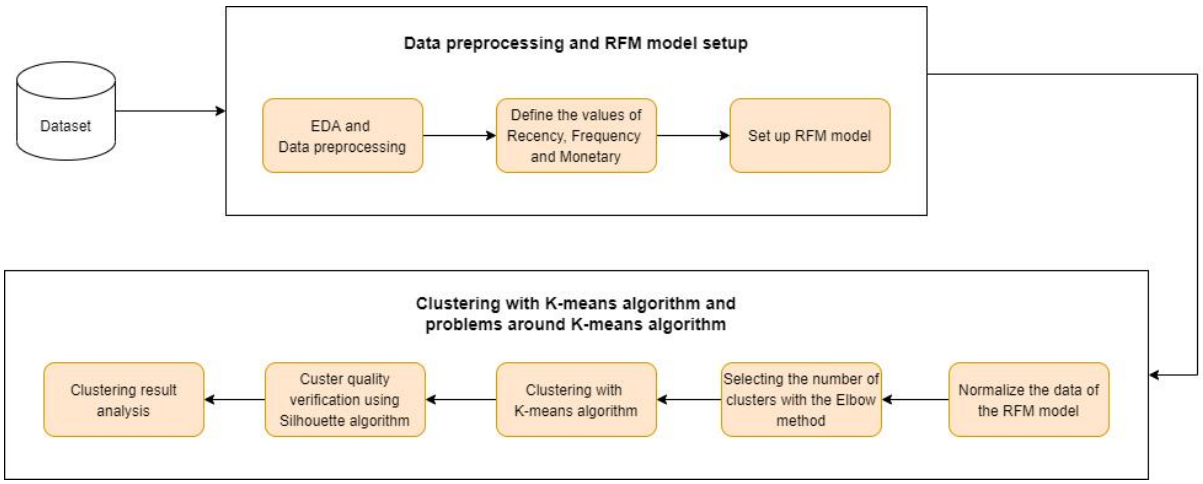

Figure 1. Overview of research model.

(1) Stage 1 from the input data that is explored and preprocessed (Data Preprocessing) to find out the inappropriate features. Then, the necessary features from the latent customer's consumption behavior in the data are selected in accordance with the calculation of Recency, Frequency, Monetary values and finally complete the RFM data model.

(2) Phase 2 is the stage with the largest proportion as well as the most complicated in the whole study. From the EDA in Stage 1, problems and characteristics related to the values in the RFM model are also found and this affects the input data for the K-means as well as ensuring accuracy in clustering results when the method is executed. Therefore, in phase 2, the research will select methods and models suitable for data objects to solve the normalization of input data and verification methods related to K-

means method to achieve the best results and analyze customer groups, make decisions to select customer groups based on analysis results.

3.2. Introduction to dataset

The study uses a dataset of customer transactions extracted from the dataset of the company Olist e-commerce website, the largest department store in Brazilian marketplaces. The dataset has information of 116,581 orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: order status, order id, order status, order purchase timestamp, ... But for our study, we only use some features such as: order id, order status , order purchase timestamp, payment value, customer unique id, customer zip code prefix, customer city, customer state.

Table 1. Description of variables

Variable name	Types of variable	Description
order_id	text	unique identifier of the order
order_status	text	status of the order
order_purchase_timestamp	datetime	the date that customer purchased
payment value	numerical	total value of an order
customer unique id	text	key to the customer dataset
customer zip code prefix	numerical	zip code of customer
customer city	text	customer city name
customer state	text	customer state name

```
[7]  df_orders.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116581 entries, 0 to 116580
Data columns (total 8 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   order_id                116581 non-null  object
 1   order_status            116581 non-null  object
 2   order_purchase_timestamp  116581 non-null  object
 3   payment_value           116581 non-null  float64
 4   customer_unique_id      116581 non-null  object
 5   customer_zip_code_prefix  116581 non-null  int64
 6   customer_city           116581 non-null  object
 7   customer_state          116581 non-null  object
dtypes: float64(1), int64(1), object(6)
memory usage: 7.1+ MB
```

Figure 2. Information about Dataframe df_orders (1)

We have 116,581 records in the dataset with no missing value. But after checking duplicate records we realize that there are 15,588 duplicate records. These records must be dropped then.

```
df_orders[df_orders.duplicated()].count()

order_id                  15588
order_status              15588
order_purchase_timestamp  15588
payment_value             15588
customer_unique_id        15588
customer_zip_code_prefix  15588
customer_city             15588
customer_state            15588
dtype: int64
```

Figure 3. Information about Dataframe df_orders (2)

Some columns in dataframe have wrong datatype such as order_purchase_timestamp. We need to convert it to datetime format.

```
df_orders['order_purchase_date'] = df_orders.order_purchase_timestamp.apply(lambda x: pd.to_datetime(x, format="%Y-%m-%d"))
```

Figure 4. Coding for converting data into datetime format

In this study, we selected record with delivery status

```
[ ]  df_orders = df_orders[df_orders['order_status']=='delivered']
```

Figure 5. Selecting record with delivery status

CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION

Chapter summarization

Clustering with traditional RFM method by calculating Recency, Frequency, Monetary value and ranking based on the quartile and labels it from 1 - 4. After mapping the RFM ratings, we divided customers into 7 groups with RFM segment and RFM score of each customer. Based on the categorical group, we suggest suitable marketing action. Using unsupervised machine learning algorithm K-means to segment customers with 3 features: Recency, Frequency, Monetary and compare number of clusters to traditional RFM method. The result shows that the machine learning algorithm K-means has not been fully effective because with K=4, we miss some groups of customers.

4.1. Customer segmentation with traditional RFM Score

First, in the RFM model, we need to calculate the Recency value. The last invoice date is 29/08/2018, we will use the day after. Recency means the last date to the snapshot day, here we subtracts the snapshot date with the last transaction date of the customers. The Frequency means number of purchases by each customer, calculating by counting order id of each customer. The last one, Monetary is amount of total purchase value, taking the sum of all orders of each customer.

```
[ ]  snapshot_date = df_orders['InvoiceDate'].max() + timedelta(days=1)

     data_process = df_orders.groupby(['customer_unique_id']).agg({
          'InvoiceDate': lambda x: (snapshot_date - x.max()).days,
          'order_id': 'count',
          'payment_value': 'sum'})
```

Figure 6. Calculating recency, frequency and monetary

After calculating recency, frequency, and monetary for RFM analysis, the characteristics of the statistical distribution of these factors such as average, minimum value, maximum value, as well as quartiles are described in table below. The average last purchased date is 161 days ago with nearly 2.38 purchases and 367 revenue in total. The Recency value ranges from 1 to 696 (last purchase date), Frequency ranges from 1 to 23 (purchase). In particular, Monetary is the value with the largest range from 9.59 to 14130.57 (currency unit). When looking at the distribution of the quartiles in Monetary, we can see that Monetary has a much larger value than the other two factors.

```
] data_process.describe()
```

	Recency	Frequency	Monetary
count	41431.000000	41431.000000	41431.000000
mean	161.390939	2.383964	367.073389
std	129.779263	1.371866	391.168004
min	1.000000	1.000000	9.590000
25%	57.000000	1.000000	135.590000
50%	130.000000	2.000000	258.620000
75%	232.000000	3.000000	459.580000
max	696.000000	23.000000	14130.570000

Figure 7. RFM variables descriptive statistics

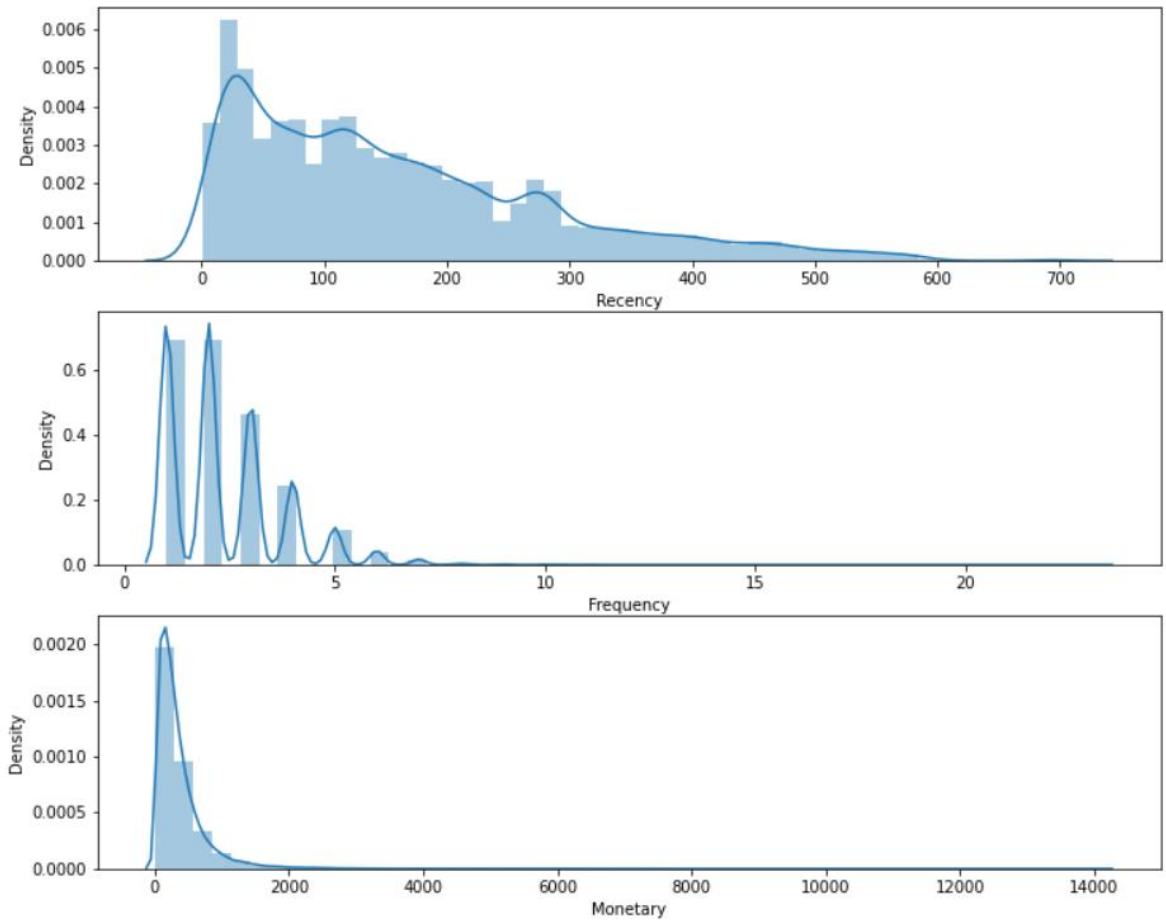

Figure 8. Distribution of R, F, M variables

For each value of Recency, Frequency, Monetary, we rank based on the quartile and labels it from 1 - 4 (from lowest to highest). The customers with the lowest recency labeled with R = 4 and the most recency value received 1 R. This phase was repeated for

frequency and monetary, but in the opposite order, with the greatest frequency and monetary receiving 4 points and the lowest receiving one.

After mapping the RFM ratings, we found the customer with RFM score of 111 is the worst customer value but RFM score of 444 is the most valuable customer for company.

customer_unique_id	Recency	Frequency	Monetary	R	F	M	RFM_Segment
0000b849f77a49e4a4ce2b2a4ca5be3f	170	2	136.26	2	1	2	212
0000f46a3911fa3c0805444483337064	126	3	583.87	3	2	4	324
0004bd2a26a76fe21f786e4fbd80607f	8	3	336.11	4	2	3	423
00050ab1314c0e55a6ca13cf7181fecf	220	1	80.18	2	1	1	211
0005ef4cd20d2893f0d9fbd94d3c0d97	482	1	187.91	1	1	2	112

Figure 9. Sample of RFM Segment

After some calculations on the RFM data, base on three customer data points: the recency of purchase (R), the frequency of purchases (F) and the mean monetary value of each purchase (M) and the RFM_Segment, RFM_Score, we can create customer segments that are actionable and easy to understand - like the ones below:

- Customers with an RFM score greater than 9 or a large RFM segment 434 mean the most recent purchases, frequent purchases, and largest amount of spend. This group is classified as VVIP - Bought recently, buy often and spend large amounts.
- Champions Big Spenders is the group defined as having an RFM score greater than 8 and a Monetary rating of 4. These customers buy on a regular basis and spend the most.
- Loyal Customers are customers whose RFM segment is greater than or equal to 6 and Frequency rating is greater than or equal to 2. People buy on a regular basis and responsive to promotions.
- The Potential Loyalists is a group of clients with an RFM Segment greater than or equal to 221 and an RFM score greater than or equal to 6. This group has characteristics such as: recent shoppers, but haven’t spent much.

- Customers with condition of an RFM segment greater than 121 and Recency ranking of 1 or RFM score is 5 is Need Attention group - Above average recency, frequency and monetary values. May not have bought very recently though.
- Similarly we have Hibernating group with RFM score greater than or equal to 4 and Recency rating of 1. They are below average recency and frequency. We will lose them if not reactivated.
- Customers who do not belong to any of the groups above will be classified as Lost Customers - Last purchase of these customers was long back and low number of orders.

Corresponding to each customer group and their characteristics, we set up separate upselling and cross selling strategies to bring the best results.

customer_unique_id	Recency	Frequency	Monetary	R	F	M	RFM_Segment	RFM_Score	Customer Segment	Marketing Action
0000b849f77a49e4a4ce2b2a4ca5be3f	170	2	136.26	2	1	2	212	5	Needs Attention	Price incentives and Limited time offer
0000f46a3911fa3c0805444483337064	126	3	583.87	3	2	4	324	9	VVIP - Can't Loose Them	No Price Incentives; Offer Limited edition and...
0004bd2a26a76fe21f786e4fbd80607f	8	3	336.11	4	2	3	423	9	VVIP - Can't Loose Them	No Price Incentives; Offer Limited edition and...
00050ab1314c0e55a6ca13cf7181fecf	220	1	80.18	2	1	1	211	4	Lost Customers	Don't spend too much trying to re-acquire
0005ef4cd20d2893f0d9fbd94d3c0d97	482	1	187.91	1	1	2	112	4	Hibernating - Almost Lost	Aggressive price incentives
000949456b182f53c18b68d6babc79c1	106	1	116.90	3	1	1	311	5	Potential Loyalists	Cross Sell Recommendations and Discount coupons
000a5ad9c4601d2bbdd9ed765d5213b3	168	1	462.72	2	1	4	214	7	Potential Loyalists	Cross Sell Recommendations and Discount coupons
000c8bdb58a29e7115cfc257230fb21b	227	2	852.54	2	1	4	214	7	Potential Loyalists	Cross Sell Recommendations and Discount coupons
000de6019bb59f34c099a907c151d855	24	1	96.66	4	1	1	411	6	Potential Loyalists	Cross Sell Recommendations and Discount coupons
000ec5bff359e1c0ad76a81a45cb598f	53	3	262.05	4	2	3	423	9	VVIP - Can't Loose Them	No Price Incentives; Offer Limited edition and...

Figure 10. Sample of customer segment and marketing actions

From here, we can see that a sufficient percentage (~45%) of our customers are in the top tier RFM levels, which are VVIP - Can't Lose Them, Loyal Customers and Champions Big Spenders. Olist must be doing something right to be maintaining their loyalty!

	Customer Segment	Recency	Frequency	Monetary		Marketing Action
		mean	mean	mean	count	unique
0	Champions Big Spenders	160.5	2.8	816.2	1563	[Upsell most expensive items]
1	Hibernating - Almost Lost	349.1	1.4	188.0	2541	[Aggressive price incentives]
2	Lost Customers	298.5	1.2	78.2	6856	[Don't spend too much trying to re-acquire]
3	Loyal Customers	148.6	3.2	319.8	5335	[Loyality programs;Cross Sell]
4	Needs Attention	245.6	1.8	241.5	3735	[Price incentives and Limited time offer]
5	Potential Loyalists	102.8	1.6	279.4	11492	[Cross Sell Recommendations and Discount coupons]
6	VVIP - Can't Loose Them	61.6	4.0	716.5	9909	[No Price Incentives; Offer Limited edition an...

Figure 11. RFM statistics of each customer group

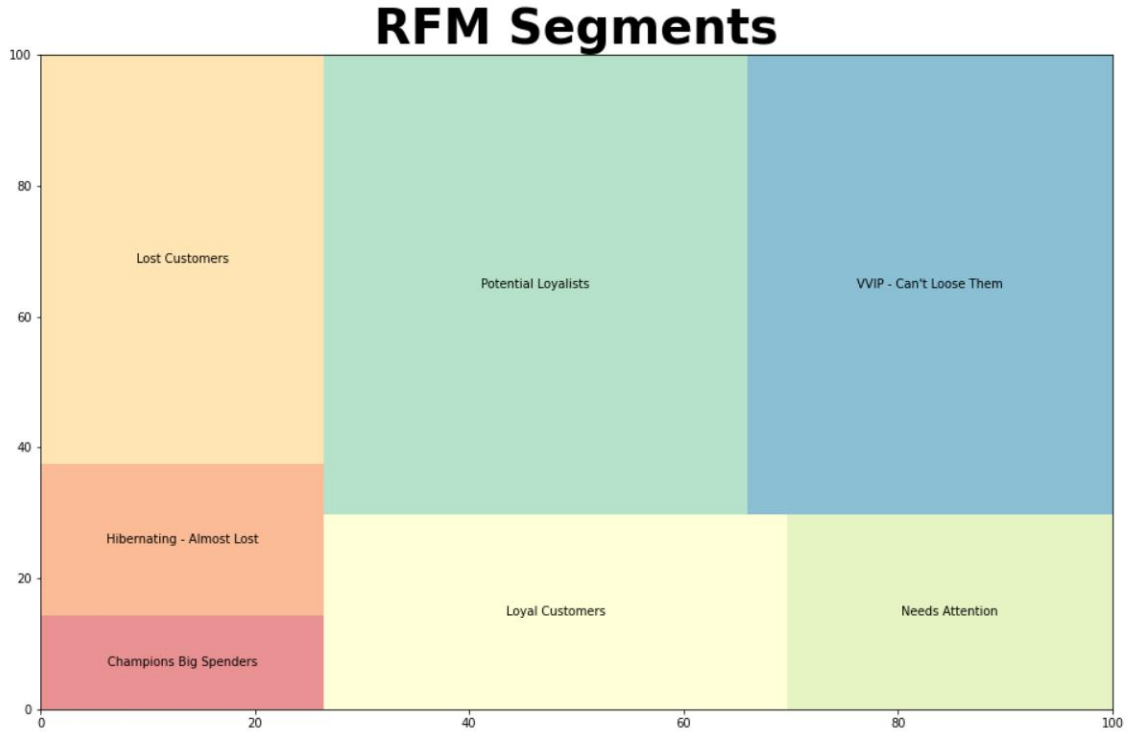

Figure 12. Treemap of RFM Segment

So, we suggest some targeted recommendations to win the other 55%.

1. Potential Loyalists

This segmentation has high potential to enter our loyal customer segments. We can throw in some freebies on their next purchase to show that you value them and engage in loyalty programs.

2. Needs Attention

This group shows promising signs with the quantity and value of their purchase but it has been a while since they last bought sometime from you. Olist can target them with their wishlist items and a limited time offer discount.

Besides, we can focus on targeted customer service and personalize their experiences.

3. Hibernating Almost Lost

The hibernating customer made some initial purchases but have not seen them since. Maybe they had customer service experience or our products don't meet their requirements, or they forgot our brand. We could spend some resources building our brand awareness with mass advertising campaign, customer service with them.

4. Lost Customers

Poorest performers of our RFM model. They might have gone with our competitors for now and will require a different activation strategy long term to win them back. Maybe this segmentation won't be our priority for business development now.

4.2. Customer segmentation using K-means clustering

Let's apply a machine learning approach to identify if there are any hidden segments we can find from clusters. To find which 'k' value is more suitable for our data we will use the elbow method.

The elbow method runs k-means clustering on the dataset for a range of values of k:

- Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
- Plot these points and find the point where the average distance from the centroid falls suddenly ("Elbow").

Step 1: Normalize the data to have all features on the same scale. This is to avoid that one feature dominates others.

```
data_log = np.log(data_rfm)
scaler = StandardScaler()
scaler.fit(data_log)
data_normalized = scaler.transform(data_log)
data_normalized = pd.DataFrame(data=data_normalized, index=data_rfm.index, columns=data_rfm.columns)
```

Figure 13. Normalization of data

Step 2: Calculate and visualize

```
sse = {}
for k in range(1, 8):
    kmeans = KMeans(n_clusters=k, random_state=1)
    kmeans.fit(data_normalized)
    sse[k] = kmeans.inertia_
plt.figure(figsize=(18,9))

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

Figure 14. Using Elbow method to find K value

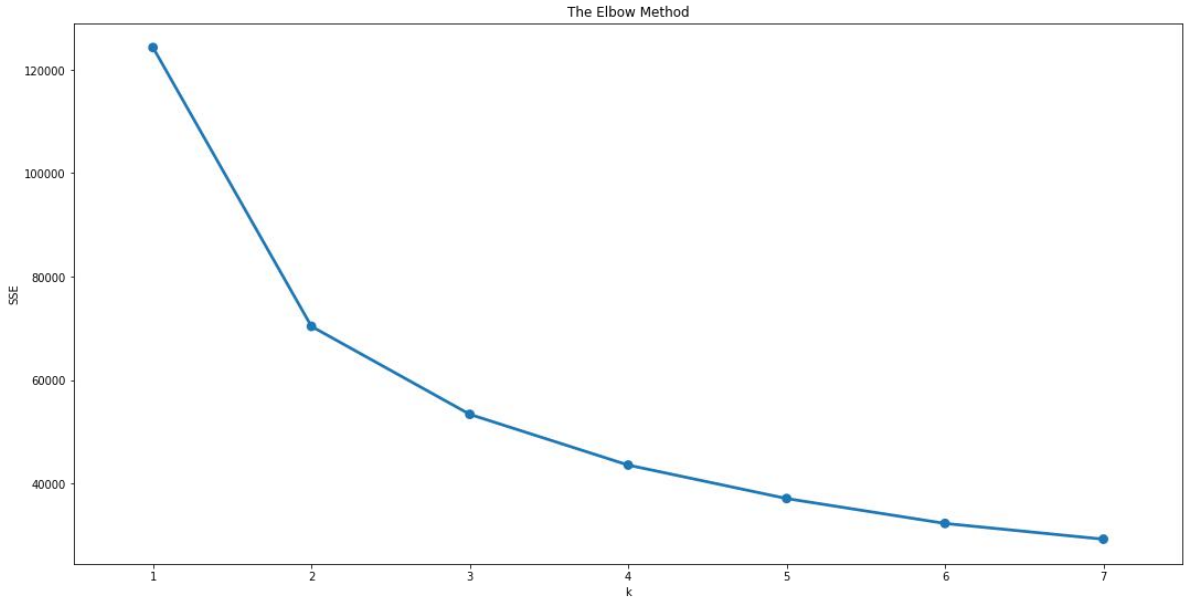

Figure 15. Visualizing Elbow method

=> The Elbow Method revealed that a cluster number of 4 is a suitable value.

K-means clustering table with k = 4

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	77.4	4.2	788.9	9284
1	262.3	1.1	113.9	11325
2	27.6	2.2	255.5	6336
3	194.8	2.4	343.5	14486

Figure 16. Result of K-means clustering

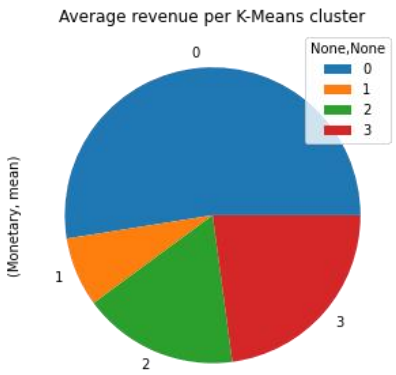

Figure 17. Average revenue per K-means cluster

- Cluster 0 dominates this chart. That means it has the highest monetary value mean.
- Cluster 1 has the lowest monetary value mean.

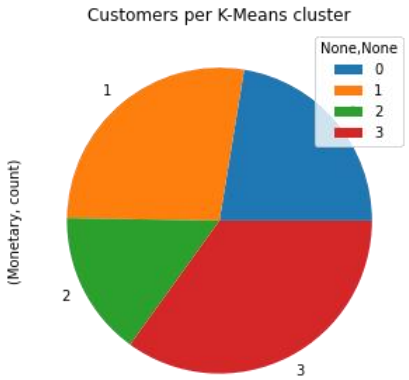

Figure 18. Customers per K-means cluster

- There's no significant difference between the numbers of customers of four clusters. Cluster 3 has the highest number of customers.

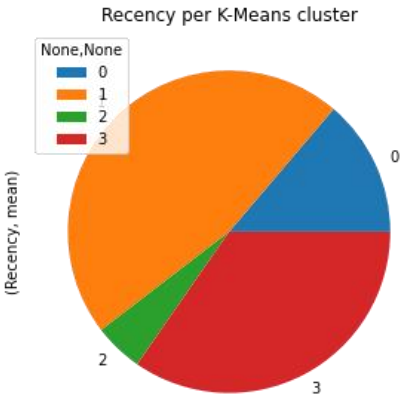

Figure 19. Recency per K-means cluster

- Cluster 2 and cluster 0 have the low recency while cluster 1 and 3 are the opposite.

We can see that our grouped summary of the mean of R, F, M that each cluster of customers places a different emphasis on our 4 features:

1. **Cluster 0 (Champions)**

It has the highest MontaryValue mean and low Recency mean and the highest frequency mean — This is our ideal customer segment. We can consider customers in this group as Champions. Champions are our best customers, who bought most recently,

most often, and are heavy spenders. The best strategy for this segment is rewarding. They can become early adopters for new products and will help promote our brand.

2. **Cluster 1 (Hibernating - Almost lost)**

It performs poorly across R, F, and M. This **Hibernating - Almost lost segment,** we will need to design campaigns to activate them again. Olist can recreate brand value. Offer relevant products and good offers.

3. **Cluster 2 (Potential Loyalist)**

Potential Loyalists with us recently but have not spent as much or as frequently as we would like them to — perhaps some personalization of products targeted at them can help to maximize their lifetime-value and come back to purchase? Offer membership or loyalty programs or recommend related products to upsell them and help them become your Loyalists or Champions. Besides, Olist can recommend other products and engage in loyalty programs.

4. **Cluster 3 - At Risk Customers**

At risk customers have spent quite a fair amount with us but have not shopped with us in the 3–4 months — We will need to do something before we lose them! Send them personalized reactivation campaigns to reconnect, and offer renewals and helpful products to encourage another purchase.

CHAPTER 5: CONCLUSION

5.1. Conclusion and implications

Customer segmentation based on the buying pattern of customers through strategically important, is an equally challenging task. Customer retention is another major concern for both online and physical enterprises. In the present work, the RFM model is implemented for synthetic and real datasets, to analyze customer segmentation. Also, clusters are evaluated using Elbow method for K-Means clustering algorithm with different numbers of clusters.

This study has produced a procedure for RFM analysis (in customer segmentation) using the K-Means method, where in the basic concept of RFM analysis to segment customers from brazil's biggest online marketplace, to segment customers. Marketing Analytics at Olist helps in measuring, managing and analyzing marketing performance to maximize its effectiveness and optimize ROI. Understanding Marketing analytics allows Olist to minimize wasted web marketing dollars by attributing budget to the targeted campaign, missed opportunities by not being able to show recommendations personalized as per the user's preferences.

Customer segmentation to identify key customers based on RFM model by using data mining techniques, and it gives them a chance to choose goal customers and invest in them, but it doesn't mean that Olist doesn't pay attention to the other customers, and they limited their effort to satisfy their profitable customers but it means that they allocated a suitable budget for performing the relationship programs with customers to increase their loyalty and satisfaction.

5.2. Limitations

- However, with the clustering results obtained based on technical factors, businesses and managers need to re-validate the above results with business and practical perspectives to be able to make optimal decisions.

- The K-means machine learning algorithm has not yet clearly divided the customer groups from which is a disadvantage for the company to evaluate the result.

- The large data difference leading to calculating the Recency, Frequency and Monetary values affects the prediction process of the machine learning algorithm.

- The dataset has not mentioned a specific category which is difficult for suggesting a new strategy for each product.

REFERENCES

1. T. Jiang and A. Tuzhilin (2009), *Improving personalization solutions through optimal segmentation of customer*, in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 3, pp. 305-320.
2. X. He and C. Li (2016), *The Research and Application of Customer Segmentation on E-Commerce Websites*, 2016 6th International Conference on Digital Home (ICDH), pp. 203-208.
3. A. Sheshasaayee and L. Logeshwari (2017), *An efficiency analysis on the TPA clustering methods for intelligent customer segmentation*, 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 784-788.
4. Christy AJ et al.(2021), *RFM ranking – An effective approach to customer segmentation*, Journal of King Saud University - Computer and Information Sciences, Volume 33, Issue 10, pp. 1251-1257.
5. Anitha P and Patil MM (2019), *RFM model for customer purchase behavior using K-Means algorithm*, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 5, May 2022, pp. 1785-1792.
6. Zhou J. et al. (2021), *Customer segmentation by web content mining,* Journal of Retailing and Consumer Services, Volume 61, 102588.