# CUSTOMER LIFE-TIME VALUE AND CUSTOMER SEGMENTATION


**University of North Texas**
**DSCI 5260 Section 002- Business Process Analytics**

**Project by**
**Akhila Vajeer**
**Chandrakala Moru**
**Jaikar Tridandapani**
**Mokshith Talari**
**Tarun Reddy Rapole**

**Instructor**
**Javier Rubio-Herrero, Ph.D.**

## Table of Contents

## List of Figures



## List of Tables

## ACKNOWLEDGEMENT

We would like to extend our profound gratitude to Dr. Javier Rubio Herrero for being our mentor and for guiding us throughout our research work. During “Business Process Analytics” class, we were able to acquire a key set of skills and understanding that led to the successful completion of our final research project. We would further like to thank every member of our team, “Group 3”, for their valuable input and efforts towards our project, “Customer Lifetime Value and Customer Segmentation”. We hope that this research fulfills its objectives and contributes to the data science domain, aiding data science students in maneuvering their way toward their end goal.

## ABSTRACT

This report provides a comprehensive analysis of customer segmentation using RFM (Recency, Frequency, Monetary) analysis and K-means clustering in an e-commerce setting, with a focus on Customer Lifetime Value (CLV). The study aims to identify distinct customer segments based on their purchasing behavior and preferences, integrating CLV estimation to determine the long-term profitability of customers. Through RFM analysis, customers are categorized into segments such as "Champions," "Loyal Customers," and "New Customers," each representing unique characteristics and potential value for the organization. Visualizations including scatter plots and distribution tables offer insights into customer behavior, while the Elbow Method is utilized to optimize the number of clusters for K-means clustering. By incorporating CLV estimation into the segmentation process, this study enhances the understanding of customer engagement, conversion rates, and retention strategies, thereby improving the effectiveness of marketing initiatives and fostering long-term customer satisfaction

**Keywords:** Customer segmentation, RFM analysis, K-means clustering, Customer Lifetime Value (CLV), E-commerce, Marketing strategies, Customer behavior.

## 1. INTRODUCTION

The e-commerce is a shopping space that simplifies shopping for customers by promoting convenience. The last two decades have witnessed a sharp increase in usage of e-commerce tools that today replace the traditional brick-and-mortar shops across the globe [1].The digital space is transforming the way customers shop, spending more time on their smart devices [2]. Since the digital is such a big leap in our shopping experience, there are risks involved along with the advantages. It might be the biggest issue that the customers get bored on company, that’s the probably they can choose other products. In order to keep customers, the company has to be more creative and they can make more products that are more appealing and more people tend to buy them which means more profits. What the company no need is to keep on coordinating the new customer recruitment, but what it needs to do more is to keep the existing customers loyal to the company. When on the Internet many targeted sites are available it becomes difficult to deal with clients and has a difficulty for the distribution of precious time of the company management to assess the customer lifetime value (CLV) of customer specifically because that customer can bring the most benefit to the company profit [3].

The Customer Lifetime Value (CLV) is a worth understanding term for e-commerce as it's an important indicator for businesses who're focusing on long run profits, ensures the sustainability, and profitability. CLV shows the total amount of revenue a customer is forecast to bring in during the whole course of their relationship with a business in particular relevant to their spending patterns and the way they react to the business [3]. Decoding Customer Lifetime Values (CLSV) does not only give important business insights into customers’ behaviors, but also help in the development and implementation of the strategies that keep fostering continuous growth as well as relationships between the company and clients. In addition, customer segmentation becomes very decisive in developing strong marketing strategies, and gauging customer tastes and preferences as it allows personalizing customer experiences. Hence, giving different customers a separate category based on their characteristics like spending habits, demographics and behaviors, enables companies to identify their needs and; then, design their marketing strategies to attract more customers and enhance their loyalty.

The main goal of our research is to go through the dataset from the Brazilian e-commerce sector to find what customer lifecycle value and segmentation are all about. With data-driven methods, we will research customers’ purchasing behavior for Brazilian e-commerce sites, having identified different customer groups and their drivers and factors impacting customers’ loyalty. This is our goal in our analysis as we aim to offer businesses the right tools to design smartly-shaped marketing plans, build long-lasting relationships with customers, and hence promote sustainable growth. Through an investigation of the Pivotal roles played by Customer lifetime value and customer segmentation in e-commerce landscape our research attempts to join ongoing discussion about customer relationship management and strategic operations in the age of the internet. Going through the expected analysis on the Brazilian e-commerce data set we will cover is a very important aspect.

Recency, Frequency and Monetization analysis (RFM) is a potent approach applied by enterprises for segmenting customers and grasping how they behave regarding their shopping. In the situation in which the Brazilian E-Commerce Public Data by Olist is discussed, RFM analysis can give an exceptionally significant outcomes about customer segmentation and targeting tactics. As well,

the application of segmentation and clustering methodologies is an integral part of the analysis and should add further insights into the behavior and tastes of the Brazilian E-Commerce Public Dataset. The process of segmentation is to divide customers into specific homogenous groups having the shared qualities, and clustering arranges the data by looking for natural groupings without any requirement of predefined categories.

Research outcomes are in the Brazilian E-Commerce Public Dataset with predictive modelling as the utmost priority: forecast CLV and key factors like product satisfaction and quality of which play a vital role. Segmentation strategies form the base of customer-made strategies by means of the clustering technologies which identify a customer's behaviour and feelings, encourage the targeted strategies and set personal customer experiences. Through clustering algorithms such as K-Means, firms can be able to identify quite unsalable customers groups that accumulate disparate classes of CLV levels and then, gauge their approach through this. By choosing a customer-centric approach, adequate customer engagement and retention rise, eventually promoting profitability and long-term positioning in the highly competitive e-commerce marketplace.

## 2. LITERATURE REVIEW

### 2.1 Past Research

In the dynamic online businesses, what is certain today about customer lifetime value (CLV) is that it will be the decisive factor for marketers and business strategy community. CLV, as proposed by [4], should be the basis of marketing strategy, resource facilities and the business performance. Moreover, they emphasize how CLV combines all those customer qualities that go into the choosing of the customer focused decisions for the purpose of attaining the high returns on investment. In fact, this is in line with [5] who elaborate the methodology for calculating CLV and its inferences that the customer acquisition and retention. Employing data of customers and predictive analytics, businesses can create a better way of doing business and understand what drives customer value, while emphasizing on the importance of CLV of companies to the competitive e-commerce environment [4] [5].

Predictive analysis models are also employed in e-commerce to prognosticate customers' future actions and market movements, [6] amplify the implementation of predictive models like logistic regression, neural networks and decision trees to forecast purchase patterns and product choices. With the use of these models, it looks like the data-driven marketing approaches and the intensifying level of customer engagement initiatives are set to be standardized as they are able to predict the benefits for the customer with great precision. With the utilization of predictive analytics, together with the magic of the amplification of several data sources such as transactional data, browsing history, and intercommunication in social networks which provide a vision of the patterns of customers [6].

The introduction of predictive analysis has fundamentally changed e-commerce world by creating new skills in decision-making. Deep learning techniques are used to solve problems involving the evaluation of large amounts of data and precise forecasting of customer behaviour and industry trends [7]. This highly advanced capability of analysis of businesses helps them find any action-oriented analytics from complex data patterns which, through implementation, in turn leads to process improvements and customer market response. The use of deep learning and predictive analytics is not only applying classical analytics but, on its contrary, beyond that, offering a forward-looking perspective that provides the possibility for the businesses to be proactive facing market changes and customer requests [7].

The theoretical research relating to predictive analytics in e-commerce establishes the key fundamentals such as the strategic significance of customer lifetime value, the usefulness of predictive models for comprehending customer behaviour, the change in marketing concerns by the customer segmentation and the influence of decision making by the predictive analytics among others. Furthermore, the technological advancements are shaping the predictive analytical processes. This study provides a broad exploration of predictive analytics in e-commerce focus on modelling customer lifetime value (CLV) [4], while [8] examine the real option of abandoning unprofitable customers in CLV calculations. [5] discuss Customer Relationship Management (CRM), and [9] propose a framework for customer asset management. Together, these studies offer valuable insights into the implications of predictive analytics for e-commerce businesses.

Although predictive analytics in electronic commerce appear to reap numerous benefits, the implementation of the technology itself is not absent of pitfalls. [9] reveal how the analysis can perfectly be done to help in marketing strategy optimization. Indeed, [10] report that the data quality and accessibility are the obstacles marketers face, therefore we have to focus on efficient data management techniques. Another ethical implication comes up regarding the use of predictive analytics, as stated by [11], as there is a need for responsiveness in the conduct of its implementation so as to ensure the preservation and sustainability of ethical soundness and consumer confidence.

Customer segmentation in e- commerce marketing is a very effective tool that maximizes CLV by segmenting customers and tailor specific products as per their needs. This method not only helps to attract more customers, but also to obtain more loyal ones. K-

Means clustering is used by [3] as an example of segmentation on e-commerce customers, done based on both their behavioural and preferred attributes. Thus, with segmentation, marketers can comprehensively know the different natures and needs of each customer to provide them better products and communication that are specially tailored to their tastes and there is improvement in the kind of personalization that shoppers are likely to see. Strategic customer segmentation not only enhances customers' satisfaction but also provides a significant amount of support to marketers for beneficial application of the available resources and to achieve higher conversion rate [3].

The technological revolution, the artificial intelligence (AI) and machine learning (ML), is providing increasingly higher performance for the predictive analytics in e- commerce sector, compare deep learning and AI capabilities on the basis of their ability in enhancing customer engagement and predictive models [12]. The increasing application of such technologies ensures the acquisition of the understanding of consumer behaviour, the accomplishment of intelligent campaigns for personal advertising, the development of accurate forecasting systems and it goes on. Machine learning and artificial intelligence can help organizations get access to adequate resources which will be used to their restructuring to the e-commerce realm and maintaining sustained competition in the changing markets [13].

## 2.2 Motivation:

The motivation behind this study is obtained from the increasing importance of predictive analytics in the e-commerce domain and the challenges businesses face in maintaining customer loyalty and engagement. Our primary goal is to leverage predictive modelling techniques to forecast customer lifetime value (CLV) effectively. Additionally, we aim to identify the key factors influencing CLV and translate our findings into actionable strategies for improving customer relationships and driving profitability in e-commerce businesses.

## 2.3 Project Goals:

By considering a Brazilian e-commerce dataset that includes various datasets related to customers information, their location, frequent orders, payments, reviews, concerned sellers etc. We then perform various analysis techniques to achieve the below mentioned goals, which collectively can drive businesses towards retaining customers, maintaining customer engagement and achieving profitability.

**Goal 1:** Leveraging the provided Brazilian e-commerce dataset [13] encompassing various customer attributes such as purchase history, location, and order frequency, we aim to conduct RFM analysis. By segmenting customers based on their Recency (how recently they made a purchase), Frequency (how often they make purchases), and Monetary Value (how much they spend), we can gain insights into distinct customer segments. Understanding these segments will allow us to tailor marketing strategies and promotions to effectively target and engage different groups of customers.

**Goal 2:** Another critical objective involves calculating the Customer Lifetime Value (CLV) for each identified customer segment. By assessing the long-term value that each segment contributes to the business, we can prioritize resources and efforts towards high-value customer segments. This evaluation will involve analysing various factors such as customer satisfaction, product preferences, and overall experience to determine the drivers of CLV within each segment. Ultimately, understanding CLV will guide strategic decision-making aimed at maximizing customer value and profitability.

**Goal 3:** Employing clustering techniques on the dataset, we intend to group customers based on similar characteristics and behaviours. By clustering customers into meaningful segments, we can uncover hidden patterns and preferences that may not be apparent through traditional segmentation methods. This deeper understanding of customer segments will enable us to develop targeted marketing campaigns, personalized recommendations, and tailored experiences for each cluster. Ultimately, clustering analysis will facilitate the delivery of more relevant and impactful strategies to enhance customer satisfaction and drive business growth.

## 2.4 Road Map

The plan for our work contains specific information, for each section of this plan provides certain data. The books review of related literature is mentioned in the section I, And the analysis and comparison of the e-commerce data is in section II. Discussing RFM analysis, CLV analytics [Customer Lifetime Value] in detail and some more sophisticated approaches like, clustering techniques. Succeeding section is a detailed analysis of every topic which are divided in three parts of Section, and the findings are deduced from the data which has been used. The final part, section IV, contains how the segmentation works and section V, is designed to discuss the business growth recommendations and action plans produced based on the previous sections analysis, which in turn, propagates for enhancing customer engagement.

# 3. DATA

## 3.1 Data Description:

Olist provided an interesting dataset that offered a wide array of data relating to the e-commerce business, for example, the number of customer orders, shipping & billing information, order & refund frequency per seller, shipping origin and destination, product and purchase details, promotional sales, as well as messaging, ratings and reviews. Each table in the dataset was littered with information that will be important for any researcher attempting to understand customers' behavior, order processing, product sale, and seller performance. These are the specific details about each dataset as shown in Table 1.

| Dataset | Number of Rows | Number of Columns |
|---|---|---|
| olist_customers_dataset | 99,441 | 5 |
| olist_geolocation_dataset | 1,000,163 | 5 |
| olist_order_items_dataset | 112,650 | 7 |
| olist_order_payments_dataset | 103,886 | 5 |
| olist_order_reviews_dataset | 99,224 | 7 |
| olist_orders_dataset | 99,441 | 8 |
| olist_products_dataset | 32,951 | 9 |
| olist_sellers_dataset | 3,095 | 4 |
| product_category_name_translation | 71 | 2 |

Table 1 Dataset List

**Customers Dataset:** This is the raw data set. Each row corresponds to an individual customer, with columns representing the unique ID that is generated for that customer, their zip code, city (yep, just one column for both), and state they reside in. This is basically the starting point from which you would begin to get a sense of your customer demographics and geography.

**Geolocation Dataset:** The latitude/longitude coordinates for the Brazilian zip code zones are stored in the geolocation dataset. This information is used for spatial analysis and understanding the regional differences in the customers' behavior and sellers' geographical distribution.

**Order Items Dataset:** This data sets out the list of products per order, their product price and freight in relation to the seller's ID. It can be analysed for product sales, composition of orders and performance of sellers.

**Order Payments Dataset:** Details of the payment type, installments, and payment value per order is present in the dataset, which gives an insight to the customer usually make which mode of payment while choosing their products and how they make their shopping.

**Order Reviews Dataset:** Reviews dataset of order: this dataset includes an Order ID, a review ID, review scores, comments and review timestamp from customers of Toredo Bistro, which allows either sentiment analysis or evaluate of customer's satisfaction.

**Orders Dataset:** Schematic of orders dataset: This dataset contains information such as order IDs, customer IDs, order statuses, timestamps of when an order was created, and when it was delivered. This is the foundation of understanding order fulfilment.

**Products Dataset:** This data describes products for sale through Olist, with product category names, sizes and product images, allowing one to perform analysis on the campaign, date, sales, and products available.

**Sellers Dataset:** The sellers data set contains the IDs of individual sellers as well as their geolocation, and it is used to evaluate the sellers' performance, their geographic distribution and the relationships that they have with their customers.

**Product Category Name Translation Dataset:** Product category names from Portuguese were converted to English names to offer the product category information in different language settings.

These can all be linked through different keys including things like 'order_id', 'customer_id', 'product_id', 'seller_id', allowing you to do a deep analysis or joining these datasets.

The Entity Relationship Diagram (ERD), a graphical presentation of the database system related to the Brazilian e-commerce system highlights how different datasets of the database are interconnected with one another and the relationships between them. The ERD maps out the relationship structure of the datasets, helping to represent searches that can discover different aspects of the ecommerce site at once, for example, linking the customers' orders to the corresponding payments details, items purchased, sellers of the purchased items, and the geographical location of both the customers and sellers. By using this ERD structure, we will create joins and merge the whole dataset together to get a final dataset that contains all information from all datasets into one single dataset for each customer ID and for the order ID.

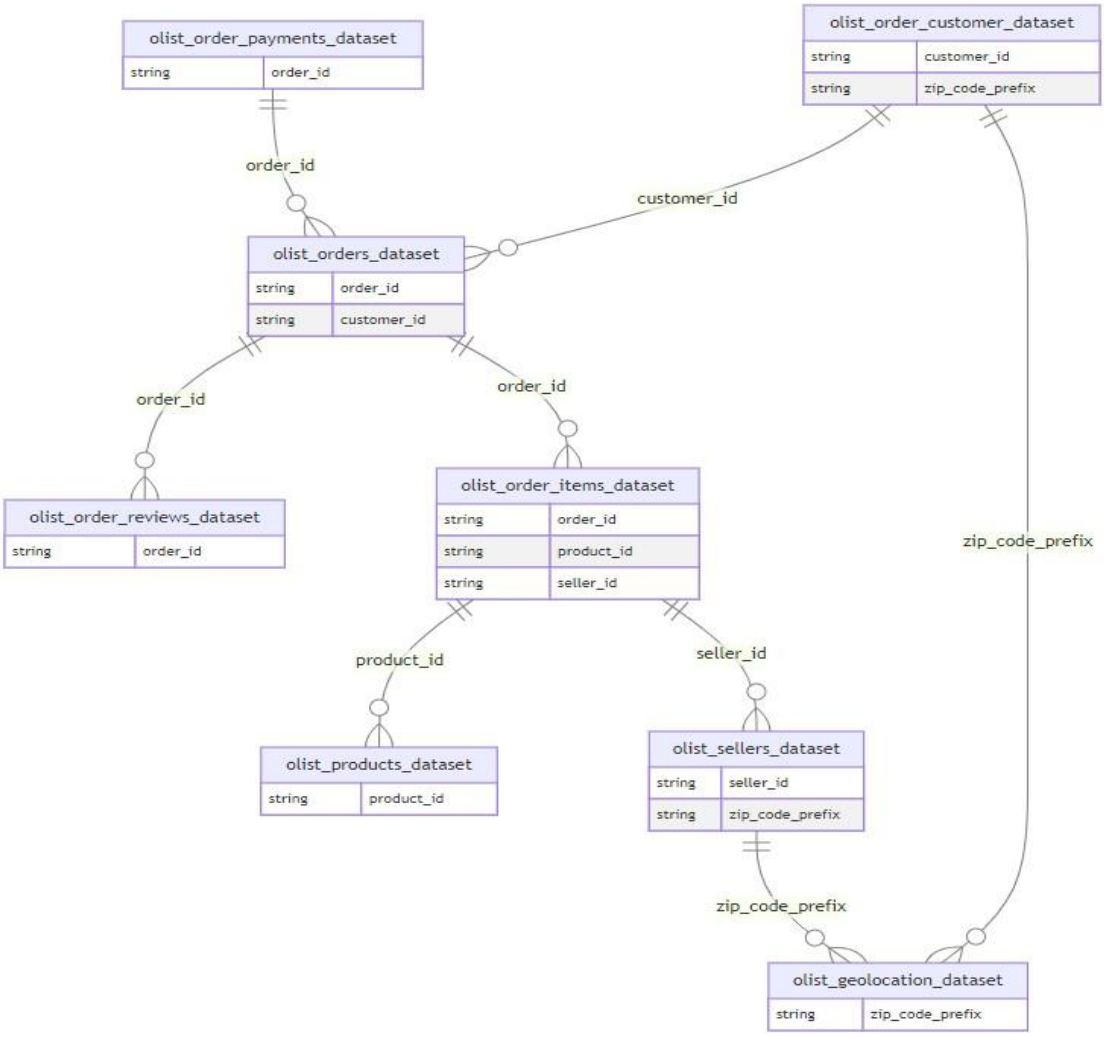


**Figure 1 Entity Relationship Diagram**

After performing all the joins as mentioned above a Final Data frame has been created which consists of all information as we can see that it has 39 features.

## 3.2 Descriptive statistics of interest:

Following all the detailed datasets integration, the Final merged dataset has a shape of (119143, 39), meaning it contained the total of 119,143 rows and 39 columns. This table has many columns like: 'order_id', 'customer_id', 'order_status' and 'order_purchase_timestamp' and so on. The data columns have some missing values. There are also columns for the "order_approved_at" and "order_delivered_carrier_date" that are missing, as well as for the "order_delivered_customer_date" and others. Rows have on duplicate key. We will learn through descriptive statistics whether or not each of the variables have outliers or missing values, in order to understand each one of them, and some insights. Therefore, we shall look at the statistics and decipher which kind of plot will fit each variable for examining the data visually and find any hidden significance in the dataset.

### 3.2.1 Descriptive Statistics (Numerical)

Furthermore, based on stellar metrics, we derived certain numbers, and we saw something intriguing. Total results were useful to the company from different perspectives of its e-commerce business. That's something, indeed, because, these figures show that the average payment value is $172.74, and it is divergently distributed with the standard deviation being $267.78, which mean that varies between customer expenditures. It is overwhelming that majority of the clients opt for single-payments and some got up to 24 instalments. However, it is very interesting to note the same sample has a paid amount as low as $0.0, and this could sensitize one of the possible uses of the non-traditional payment ways or vouchers. Continuing the review part, average score of 4.02 showed mostly favourable responses, however, a minimum score of 1.0 depicts where they had not met expectations. Actually, the greater portion of our ratings are 5.0s (75th percentile), thus the customers were evidently very satisfied with our works. Product attributes, including product name length, description length, and photograph quantity, are variable in the extent of differences that these attributes display to factors such as the type of the products offered. Furthermore, product size and dimension are far from being the same, reminding us products differ in size and types. Despite the fact that majority of the orders contain only one item (translate as mean order item count 1.20), the maximum order count of 21 items demonstrates the biggest shopping imagination scale. This overview of research helps to identify opportunities for doing a better job with pricing, product description, etc. and hence improve products marketability and responsibility.

### 3.2.2 Descriptive Statistics (Categorical)

This category data has helped us to do some explorations. To deliver you the statistics it was observed same way that unique values are in particular most of the variables, so visualization is not possible with each variable and you can only visualize top 10 categories for each variable. The study of e-commerce data describes the statistics of object-type parameters and uncovers many interesting aspects of the e-commerce dataset. Firstly, we can say for sure that "delivered" status became a major order status, and then we can detect other stats. Nevertheless, analysing the orders with the status of cancelled or on hold might narrow the scope of the problem of the fulfilment process. Such aggregated data points to São Paulo (SP) as the customer city and state that prevail, leading to concluding that a large proportion of the customers is in this region which impacts the marketing strategies and logistics planning. Other than that, the frequency of "credit card" almost as a single type of payment emphasizes the fact that arrangements have to be made to improve how customers are served. To additionally demonstrate, inspecting responding commentary consumers grant useful feedback and the high frequency of the "Recomendo" (Recommend) and "Muito bom" (Very good) title reveals high customer satisfaction. In this case, an enthusiastic tone reveals that the clients are happy with services received, but, from another point of view, it signifies the necessity to possess the instruments to deal with negative responses effectively. While this data is not available on specific product category frequencies in totality, it is possible to dig into product category names appearing frequently as the top-class segments, which office supply these days, even provide both inventory management service and marketing effort. In addition, finally at the end SP manages to become again the most dominant seller city and state, and this consideration of where the sellers are physically located can be relevant both for the relationship between the supplier and the company as well as for the logistical management. As a result, these observations from descriptive statistics provide to an e-commerce company fuel it into a data informed mode where it makes calculated decisions in all operational aspects

## 3.3 Exploratory data analysis:

After checking through the dataset and after performing necessary EDA now data is visualized for finding hidden insights.

### 3.3.1 Visualizations:

#### 3.3.1.1 Pie Chart for Payment Type Distribution:

The credit cards type of payment marked a clear majority preference in the Payments Distribution Pie Chart, over 70% of transactions, showing that customers are comfortable with these features because of the way they have been dealing with it as shown in Fig 2. As a result, 'Boleto' follows 'Cartão Suíte' coming as the second chosen method and highlighting the necessity to back up alternative methods. Vouchers and debit cards are active about which can result in extra pairs of advertising to widen their implementation. We could strengthen our offerings through credit card companies' alliances and thus secure both new clients and transactions. Besides that, introducing alternative payment methods may be the way of reducing the operational risk associated with processing network malfunctions or changes in credit supply.

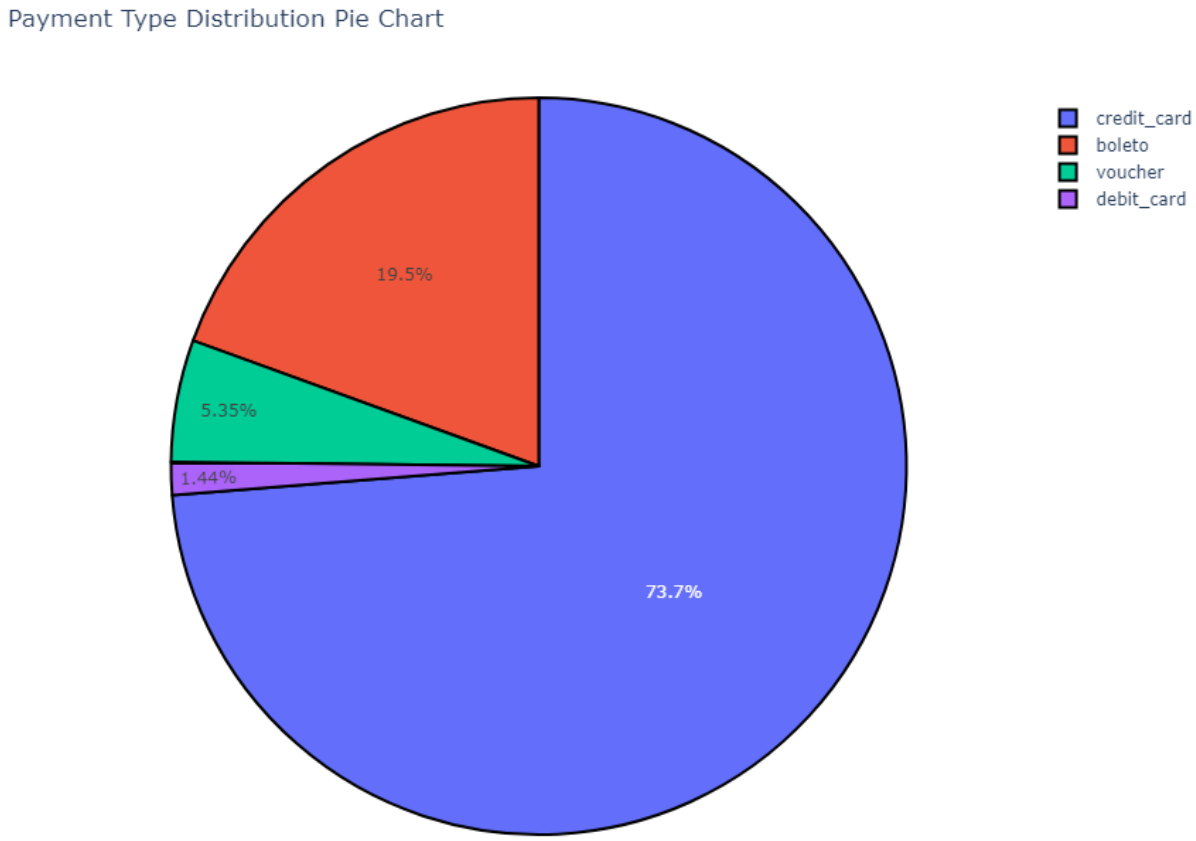


**Figure 2 Payment Type Distribution Pie Chart**

### 3.3.1.2 Line Chart for Order Purchase Timeline

The Order Purchase Timeline Line Chart as shown in Fig. 3 portrays a major trend of order constant, in due course, mark by waves that might be linked with promotional events or seasonal sales that can inspire future marketing plans. On the other hand, an unexpected spike in purchase ratio makes me vulnerable to questions related to the event like a sale, promotion, or data entry error for the recurrence of the success and rectification of the issue. The number of stabilities in orders is indicative of customers' stable demand, allowing us to consider efficient strategies which stabilize the milieu as well as possible improvements. A slow down can show off the existence of competitors, the market reaching saturation, or a behavioral shift on the part of customer, all of which deserve researching so that strategic planning is now stressed. Through the in-depth research of the impacts on demand, an efficient fashion brand can achieve decreasing shifts and aligning production and resource allocation successfully. There are higher orders on Nov 24th and it is understandable from the visualization that it's a Thanksgiving Day and thus orders are more in that day when comparing to the rest of days.

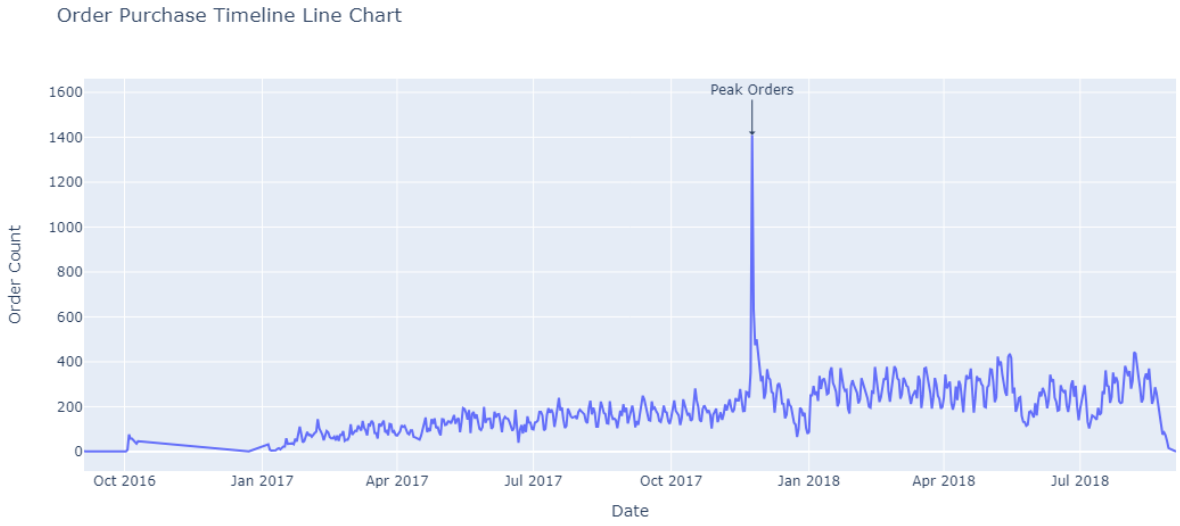


**Figure 3 Order Purchase Timeline Line Chart**

### 3.3.1.3 Histogram for Review Score

We wanted to check how top 10 products with different product categories with respect to review scores as shown in Fig. 4 and We can clearly see that Review Score of the Top10 products are almost 9 to 10. To put it simply, most of the customers seem very happy with the product or service provided by the business and thus, tend to give high rated review scores. Although, customers with low review scores appear to be in an isolated circumstance, their complaints should be inspected closely by the company to ensure that these are only outlier incidents and will not be repeated on a larger scale. The disproportionately higher positive review scores should be utilized by the business extensively while marketing any products and services in the future. More specifically, while predominantly high scores by the customers will help to establish a reputable and authentic relationship between the business and the customers, the company should try to maintain the exceptional levels of customer satisfaction visible through scores of 4 and 5 by ensuring that only top-quality products and services are provided to them on a continuous basis. On the other hand, responding to the concerns and complaints of the customers who have left low scores has the potential to improve these clients' perception of the company and thereby reduce the chances of repeat churn. But these are the top products which are more bought are scored as well.

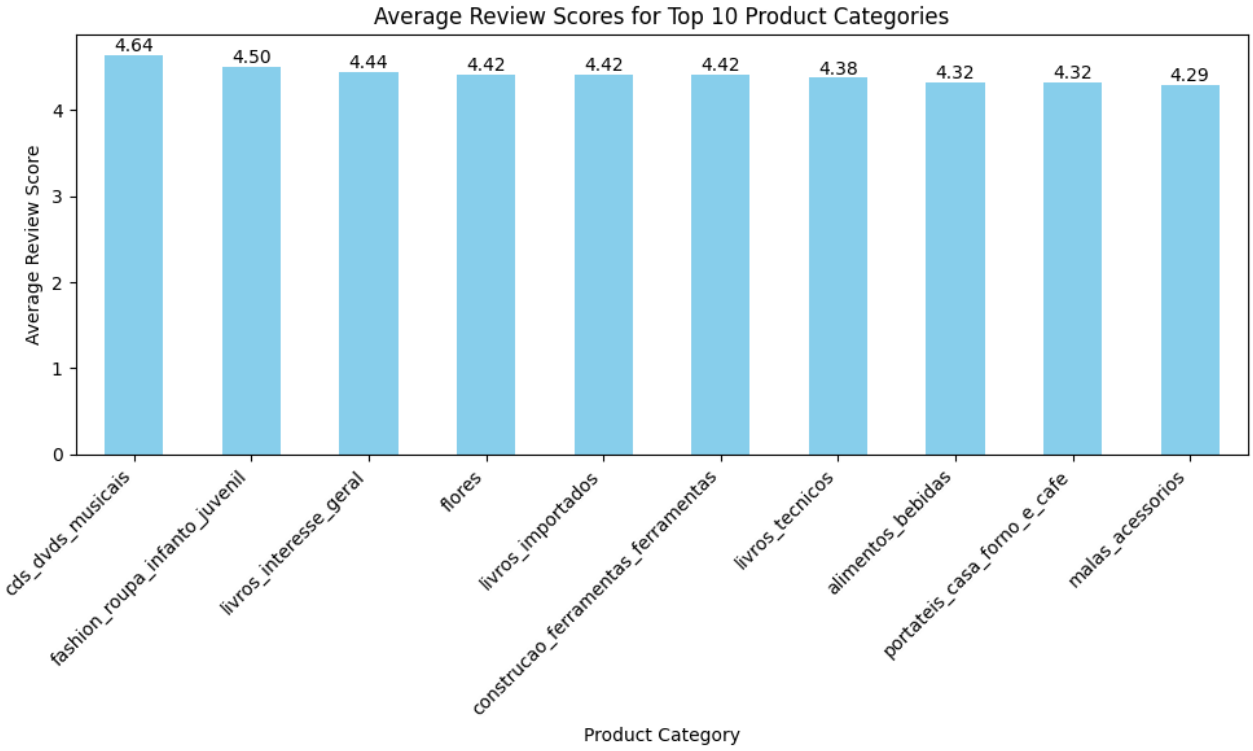


**Figure 4 Create a Review Score Histogram**

### 3.3.1.4 Donut Chart for Average Order Distribution for each Payment type

The Payment Type of Average Order Value Donut Chart as shown in Fig.5 indicates a diverse distribution of average order values across different payment methods, highlighting a heterogeneous payment system catering to various customer preferences. Generally, customers exhibit similar spending patterns for credit card and boleto payments, suggesting a level of comfort with both methods, particularly for transactions of similar magnitudes. The relatively lower average value associated with voucher payments may indicate a strategy to attract new customers or reserve this method for smaller transactions. Conversely, debit card payments show a slightly lower share, prompting further investigation into whether this reflects customer behavior or potential limitations on transaction amounts for this payment method.

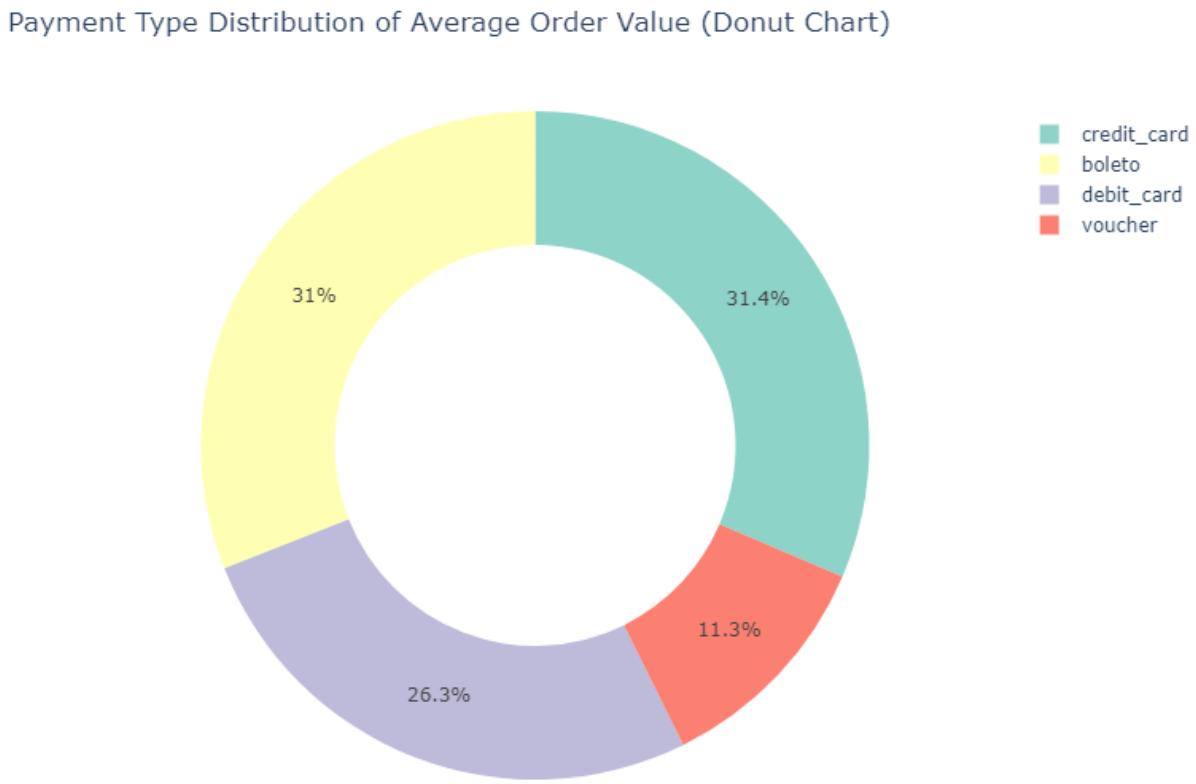


Figure 5 **Donut Chart for each payment type of Average Order Value**

**3.3.1.5 Horizontal Bar Graph for finding Order value by payment type:**

The order value by Payment Type, portrayed through a Horizontal Bar Graph as shown in Fig. 6, highlights the significant contribution of credit cards to total order value, signalling their importance in revenue generation and prompting potential negotiations for better transaction rates. Boletos also hold a substantial share, underscoring their popularity and the necessity of maintaining a seamless process for this payment method. Debit cards and vouchers contribute less to the total order value, suggesting opportunities for growth through targeted marketing efforts. Developing loyalty programs or payment-specific promotions could incentivize customers to increase their order value, particularly for underutilized methods like debit cards and vouchers. These insights inform decisions regarding payment infrastructure investments and negotiations with payment service providers to optimize transaction costs. For higher transactions i.e., order value Credit card is mostly used and then boleto is used more and boleto is a Brazilian cash-based payment method.

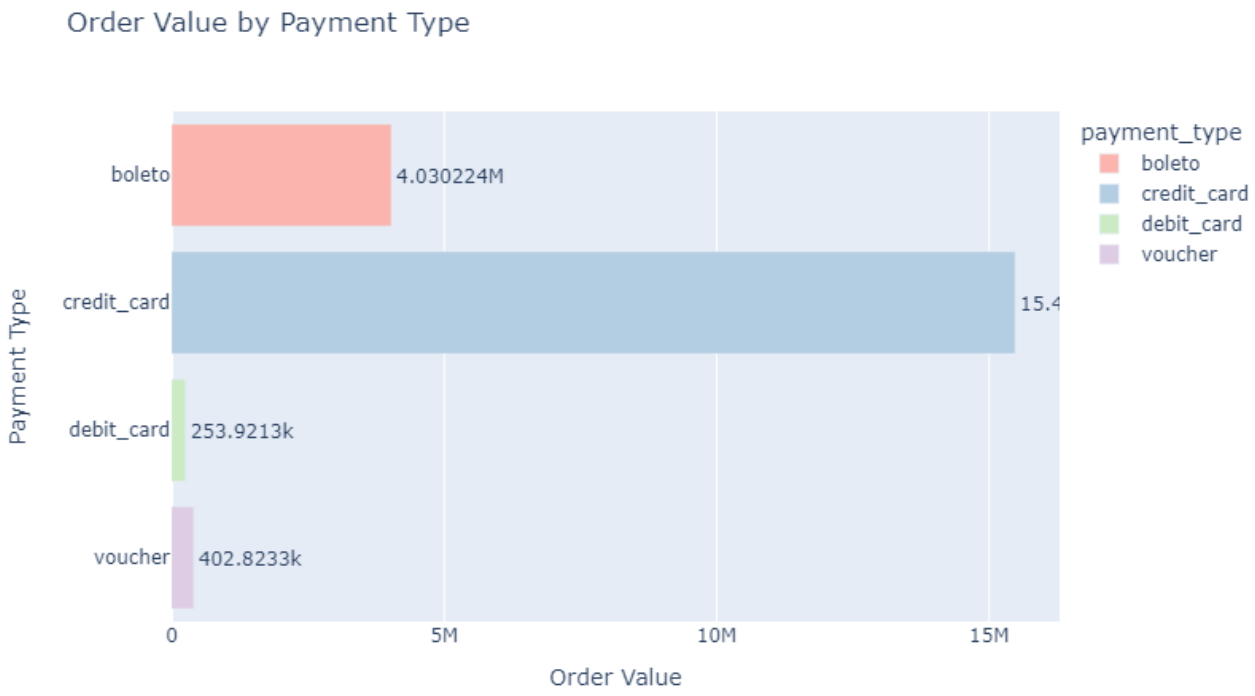


Figure 6 **Horizontal bar graph to visualize the order value by payment type**

### 3.3.1.6 Order Count by Month, and Year using Table:

The given Order Count Table, arranged by Month, Year as shown in Fig.7 helps identify trends, seasonal variations and marketing peak months for possible inventory change and marketing campaigns. For instance, after comparison, the Table shows high counts for May and September in 2017 and 2018, implying a probable event or promo that could be related to the tourism or culture sector, which draws more customers in warmer seasons. Therefore, further analysis might be needed to identify the underlying reason, which could be internal or external marketing efforts. Moreover, this data could possibly be used for inventory change as well, as these days reflect increased demands as in marketing occasions or celebrations, so authorities can prepare for such peaks. Overall, the data suggests potential cyclical or seasonal variations in orders at some months compared to others such as January, noted by their lower count in 2017. This becomes a good reference for future preparations for sales and operations to reduce disappointment, boost sales and improve inventory control. Looking at the data, the overall wide variation among years might reflect the company's growth in terms of market expansion, hence the need to forecast future growth and employ this data as a benchmark to make strategic decisions. It is evident from the Graph that with each year the orders are increasing gradually.

Order Count Table by Month and Year

| Month | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|
| Jan | 0 | 1012 | 8513 | 9525 |
| Feb | 0 | 2057 | 7923 | 9980 |
| Mar | 0 | 3168 | 8500 | 11668 |
| Apr | 0 | 2838 | 8182 | 11020 |
| May | 0 | 4400 | 8166 | 12566 |
| Jun | 0 | 3776 | 7340 | 11116 |
| Jul | 0 | 4845 | 7289 | 12134 |
| Aug | 0 | 5147 | 7417 | 12564 |
| Sep | 3 | 5089 | 1 | 5093 |
| Oct | 379 | 5565 | 0 | 5944 |
| Nov | 0 | 8992 | 0 | 8992 |
| Dec | 1 | 6506 | 0 | 6507 |
| Total | 383 | 53395 | 63331 | 117109 |

Figure 7 **Order Count Table by Month, and Year**

### 3.3.1.7 Order count for each Day of Week and Hour using Multi-Line Chart:

The graph in Fig 8 that shows the order numbers by each hour of the day for every day of the week, will provide useful information on customer demand patterns, that will guide decision-making, especially in matters such as manning of the business to fit the demand, work planning and peak hour management. Indeed, the visualization highlights a all-the-time/constant increase in orders from late mornings through to early afternoons demonstrating that peak hours likely occur during such periods. This could therefore mean that flah sales and promotional offers can be put in place in such periods. On Tuesdays, there is a highest count of orders, and the other weekdays follow subsequently and almost in the same range, which could imply that the shopping motivation rises during the week reaching the highest peak on the weekend when sales should be targeted. Another aspect of the chart is the periods of the day with smaller order counts, which is particularly visible in the early morning hours. Website services for example the maintenance or the inventory updates, may be conducted at such hours. Although fluctuations in time could exist, a number crunching shows the general trajectory is an unflinching demand that presents a timely marketing strategic competitors with the chance to introduce high-ticket and customer service improvement initiatives.

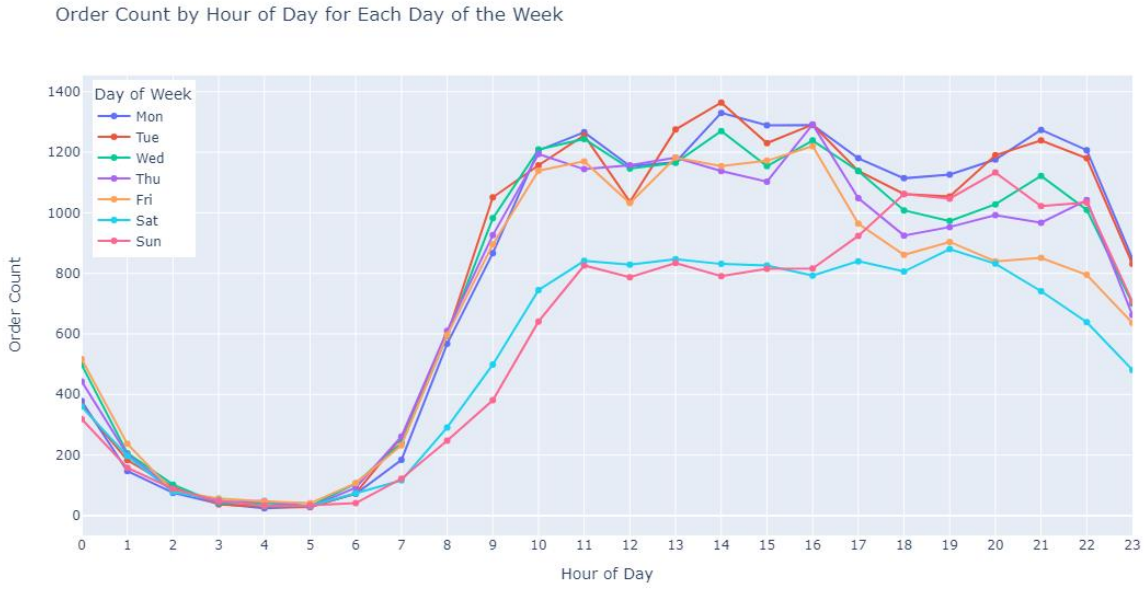


Figure 8 **Order Count Table by Day of Week and Hour**

### 3.3.1.8 Product Category Distribution using Bar Chart:

A closer look at the Product Category Distribution bar graph as shown in Fig 9 reveals that some categories are far more popular than others, which could indicate opportunities for strategic positioning and marketing. This might involve featuring the most frequently ordered category in the company's marketing materials, as well as in a more visible position on the website, which could help to raise its profile and generate additional sales for the company. Meanwhile the clearance of many niche categories in the long tail of the distribution makes it clear that there could also be opportunities to exploit in terms of increased exposure and sales through a more targeted marketing strategy. The breadth of popularity across all of the available categories also serves to emphasize the very real need to market to different buyers with different interests, while even those in which there's a high order count ought to be prioritized for stock, availability and potentially faster shipping options, again to capitalize on their popularity. When product category distribution information can be effectively used for targeted marketing strategies, optimized CLV, efficient customer segmentation, strategic decisions of positioning and stock management, there is an enhancement of the business growth and multiplied success.


Figure 9 **Product Category Distribution**

### 3.3.1.9 Product Price Distribution using Bar Chart:

In the histogram of the product price distribution as shown in Fig 10, we see the largest number of products are the ones that cost lower prices. This can mean, there is more attention paid to affordability. Distribution long tail forecast suggests the products portfolio is a mix of the higher priced items, contributing significantly to revenue despite the purchases that are not made frequently. Providing customers with the option of financing and payment instalments will undoubtedly increase product affordability for people who cannot afford expensive products in one-time payment. The bulk of the product mix in the lowest price range may imply aggressive pricing approach with discounts to attract customers. This principle plays an important role in inventory management, which basically implies to offset the cost of lower priced goods by virtue of the high turnover, while using a conservative approach for costly pieces.

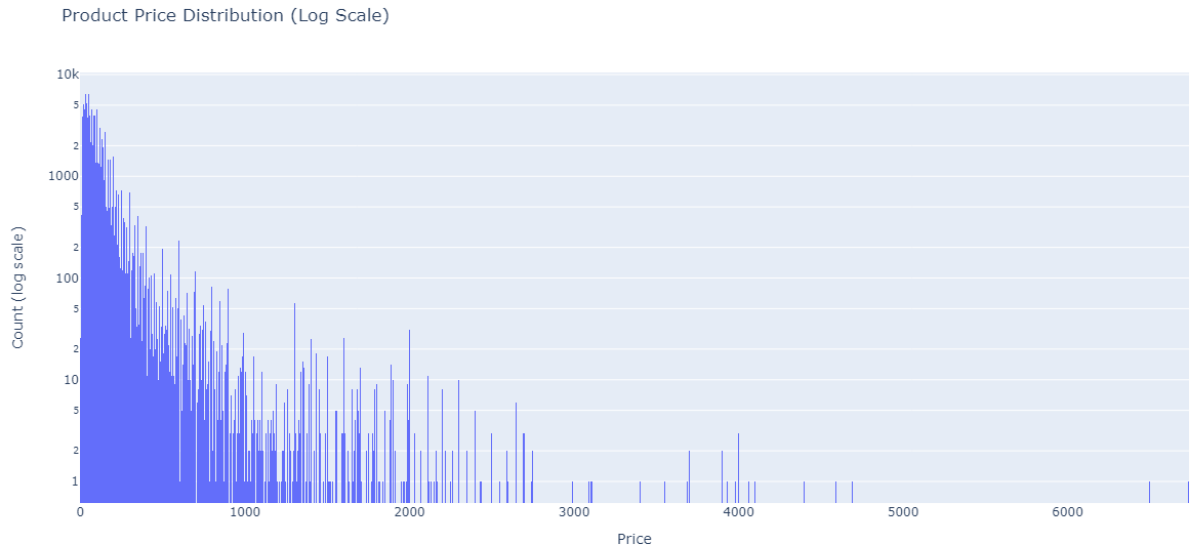


Figure 10 **Product Price Distribution**

### 3.3.1.10 Correlation between Volume and Weight using Scatter Plot

The Scatter Plot that is correlated between Volume and Weight show a moderate to strong positive relation as shown in Fig 11. This demonstrates that, in general, the bigger the item, the more weight it has, which is useful for the understanding of logistics and pricing strategies concerning shipping costs. The concentration of data points at the bottom part of the scale indicates that majority of the products sold are of smaller size and lowest weight helpful for an efficient utilization of space as well as lower freight costs. Outliers exhibit both high volume and high weight attributes. Hence, these products may necessitate special handling and shipping provisions that would only make shipping fees reflect the true cost. The knowledge of volume-weight ratio is amongst the key factors in achieving efficient packaging solution and save the space by reducing the shipping cost to the consumer as well as thus improve the profits. Information from the correlation metric can be used in bargaining with shipping carriers, you might be able to have bulk rate shipping discounts for some manufactured products falling into standard or average volume and weight boundaries.

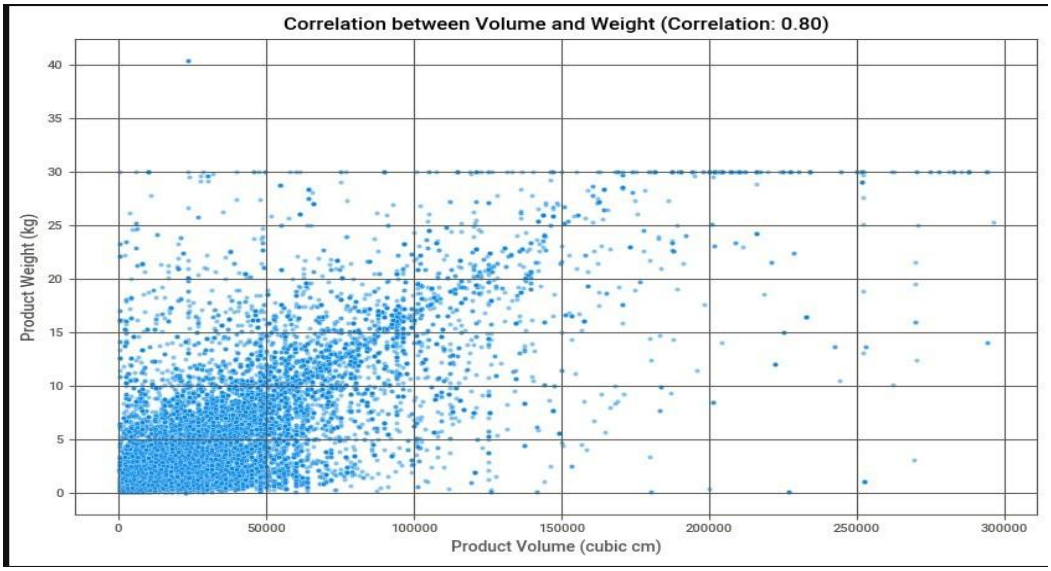


Figure 11 **Correlation between Volume and Weight (Scatter Plot)**

### 3.3.1.11 Correlation between Volume and Freight Value using Scatter Plot:

The Scatter Chart of the Freight Value vs. Product Volume ($m^3$) shows a moderate positive relationship as shown in Fig 12, compared to the stronger association of weight and volume. The reason for this might be in the case of the non-linearity of the pricing structure for freight or the other factors like the distance and speed of the delivery impacting it. The graph has a set of clustered points at the bottom part of both axis, which means that majority of products have lower volumes and hence lower freight charges, a good business strategy for the budget conscious consumers. The results of the plot demonstrate that more volume does not always result in direct proportional increase in freight values as changes in volumetric weight pricing and benefits from flat shipping rates. The absence of the consistent linear pattern may reflect the influence of number of variables such as the length of shipping, package size or shipping mode, pointing towards other ways of cost-cutting.

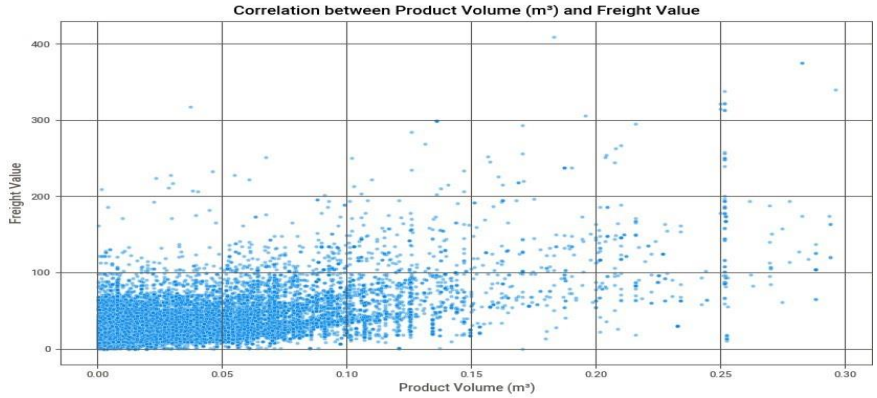


Figure 12 **Correlation between Volume and Freight Value (Scatter Plot)**

### 3.3.1.12 Distribution of Seller Shipping Time using Histogram:

The Histogram that shows the Distribution of Seller Shipping Time depicts a cluster that is evident around the range of shipping days as shown in Fig 13, meaning that most of the sellers have been efficient in fulfilment process. There are some potential outliers corresponding to high shipping times, warn of potential logistic errors or inefficiencies that is dealt and create a visualization that shows distribution of seller shipping time by implementing a tight delivery time requirement of offer a support to weed out outliers will ensure consistent customer experience. The high shipping times within a lower range implies that logistics handling is efficient; thus, useful during customer engagement & feedback. The application of this mechanism in determining reasonable shipping expectations and adjusted delivery time-lines based on actual performance possibly enhance customer satisfaction.

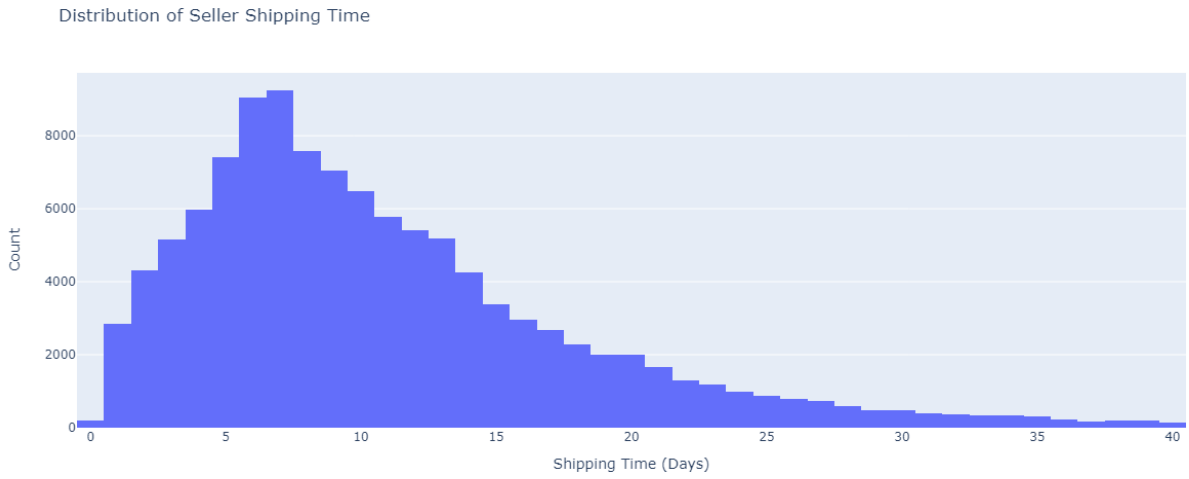


Figure 13 **Distribution of Seller Shipping Time (Histogram)**

### 3.3.1.13 Average Shipping Time by Seller State using Bar Chart:

The Bar Chart of Average Shipping Duration by Seller State portrays the significant level of differences in the performance of shipping/delivery among different regions as shown in Fig 14, which shows the geographical impact on shipping performance. The states Acre and Amazonas show remarkably high average shipping times, which depict presumable logistical problems. Focused developments of these areas, like optimizing the routes or associating with local delivery agents, might turn helpful in tackling these difficulties. The states with the lowest shipping period could be used as a benchmark for other states and provide information about the effective procedures they can adopt.

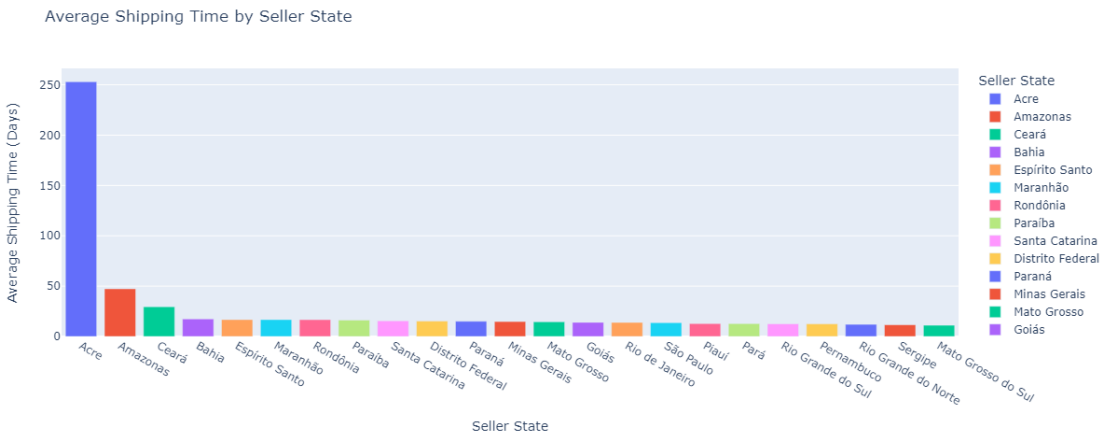


Figure 14 **Average Shipping Time by Seller State (Bar Chart)**

### 3.4 Strategies for Handling Missing Values and Outliers:

Regarding handling of missing values, we took the standpoint to distinguish the variables having more than 50% missing values. Herein, the portion of missing values was determined for every variable and those variables where the portion of missing values surpassed 50% were omitted. These variables were found to have nearly no values resolved to fill the gaps and further dropped from the dataset. Following removation of the variables which have been identified as missing, we then used strategies to impute these missing values, for numeric and categorical variables. For the case of arrays of numbers, we replaced, the absent ones with the average of each column. This model yields the imputed values which are aligned with the overall pattern of the data and becomes a succinct feature as well. Additionally, our imputation strategy for categorical variables involved mode imputation, which implies that the missing values are filled up with the most frequency occurring value (mode) of every given column. Therefore, we make sure that the data we are adding agrees with the most frequently occurring values of the classification parameters within the dataset in order to preserve the properties of categorical features. Finally, we performed the last stage that was imputation processes. Then we completed a final stage check to confirm that there was no missing value for the dataset. This all-encompassing approach to handling the missing data makes it possible for the dataset to be suiting for the subsequent analysis and modelling jobs fully therefore lessening the effect that missing data could have on the results' accuracy and integrity. A function called remove_outliers is setup to remove outlier from a Data Frame based on their z-scores. Thus, performing a function with the stats.zscore yields z-scores for each column. The threshold value, which is often at 3, deems particular data points to be considered as outliers. Next, it will capture the Data Frame into a variable and then bypass the remaining rows where z-score is above the threshold to remove outliers. This is used to extract only 3 particular columns from original RFM Data Frame (i.e. 'Recency', 'Frequency', 'Monetary Value'). Hence, Data Frame that is before and after outlier removal is shown so that people can compare the dataset the level of impact of outlier removal. In essence, this methodology provides a means to identify and minimise the adverse effect of outliers, thus the subsequent statistical analysis and machine learning equipment's will be more trustworthy.

## 4. METHODOLOGY

### 4.1 Customer Lifetime Value (CLV):

A key CLV forecasting tool which is used to determine the long-term profitability of customers is CLV estimation. CLV calculation algorithm requires to be performed at first determining each average purchase value and purchase frequency rate. Those arithmetical operations provide customers with pricing. It is being supposed that a typical customer will live for so few years, and CLV will be counted as multiplying a customer's value with the length of his / her life. Thus, this yielding mechanism allows companies to determine the revenue amount from each customer that is forecasted to have a multi-year lifetime relationship with the business and shape business strategy and resource allocation.

### 4.2 RFM Analysis:

RFM analysis is a fundamental tool that assists marketers in understanding consumer behavior and is the basis of a way of segmentation. Being a case of a retrospective look at Recency, Frequency, and Monetary Value metrics, Treasure Data dices customers into groups depending on their transaction history. Quintile calculations assign RFM scores to customers based on the quintiles they fall in. These score outputs customer behaviour that is then translated into numbers, with the data used to form a segmentation strategy. Thus, this approach works best for any organization to identify the needs/lifestyle and then come up with marketing strategies that suit each of the customer segments.

### 4.3 Segmentation:

Identifying buyer categories is a part of the process that targets customers by character and behaviour. At first, the analysis of RFM, the customer segmentation is performed on the basis of Recency, Frequency, and Monetary Values of customers. Consequently, these evaluations provide the room to make a researcher run the segments, which will present the unique behaviour groups and attributes. The Information one avails from RFM segmentation helps vendor empower their marketing efforts in a strategic manner that has a high probability of achieving its business goals. Businesses are able to optimize strategies tailoring to the unique needs and inclinations of each segment, as a result, enhance customers' engagement and allegiance.

### 4.4 Cluster Analysis:

Clustering analysis has been used in this case to get through groups of customers which have similar behavior patterns. Targeting features that incorporate RFM as input for clustering analysis thus becomes a key approach. The design of these features is rooted in the aim that their significance will be of equal weight during clustering. The K-Means clustering algorithm is thereafter implemented in which customers are grouped into clusters on the basis of their respective standardized RFM scores. K-medoids clustering technique with optimal cluster determination processes, such as elbow method, is employed to filter the number of clusters. Thanks to this approach businesses can start receiving in-depth understanding on customer behaviour and needs, which in turn translates into more effective marketing initiatives and increased customers engagement.

#### 4.1.1 Distribution of Customer Lifetime Value (CLV) by Product Category using Bar chart:

The Customer Lifetime Value (CLV) by Product Category bar chart as shown in Fig 15 provides us with information regarding the categories that are profitable to us of a customer over his/ her lifetime. Categories placed in top of the CLV are those that offer fancy items or prompt for repeat purchase. Thus, this data helps in defining trajectory of marketing and development. The process of directing marketing resources to highly lucrative categories is made possible by emphasizing on CLV. The diagram ultimately defines the cross-selling and upselling strategies by determining the additional products to be included. Categories with low CLV tend to require modifications from pricing to costs or engagement tactics to improve long CLV.

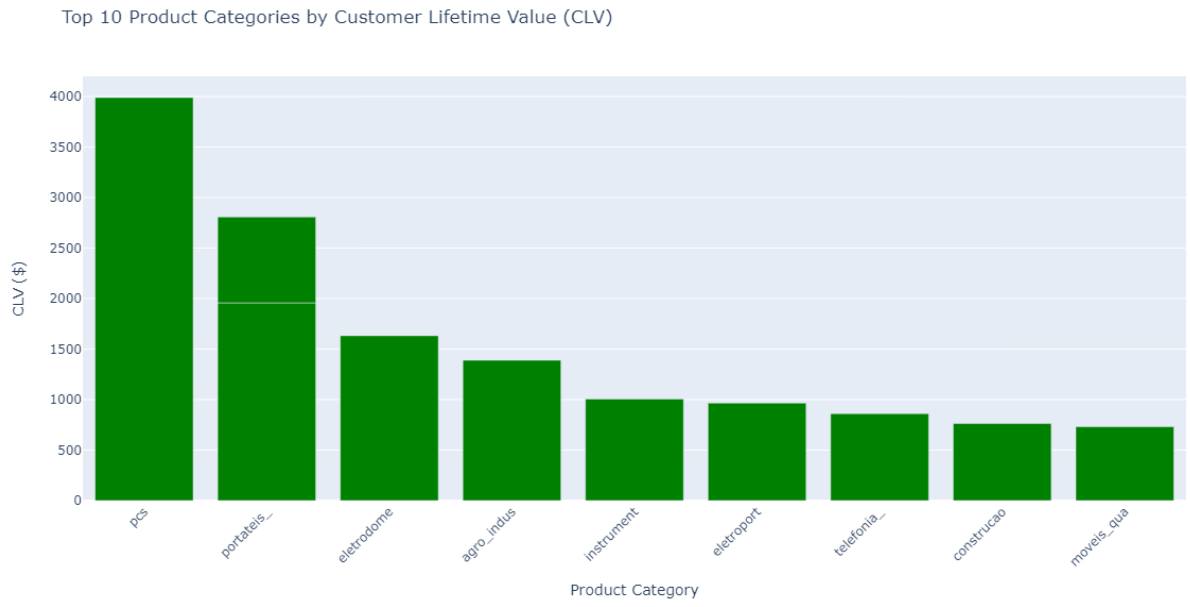


Figure 15 **Customer Lifetime Value (CLV) by Product Category (Bar Chart)**

### 4.1.2 Distribution of Customer Lifetime Value (CLV) by State using Bar chart:

The Customer Lifetime Value (CLV) by State bar chart is based on where customers spend more over their entire relationship with the company as shown in Fig 16. The ranking of these states provides us with a geographical perspective on where the customer spending is at its peak. Mature CLV states may need extra marketing strategies; below-par CLV states could look for ways to effectively enhance customer spending and loyalty. Various CLV differences may be government from economic divisions, cultural shopping habits, or just how strong the local marketing initiatives are. These data characteristics will help determine whether space optimization, like opening new stores or warehouses in states with greater customer value is beneficial. The target has customized offerings composed of local product assortment, promotion, and loyalty program which altogether, can add more value to the customers located in each state.

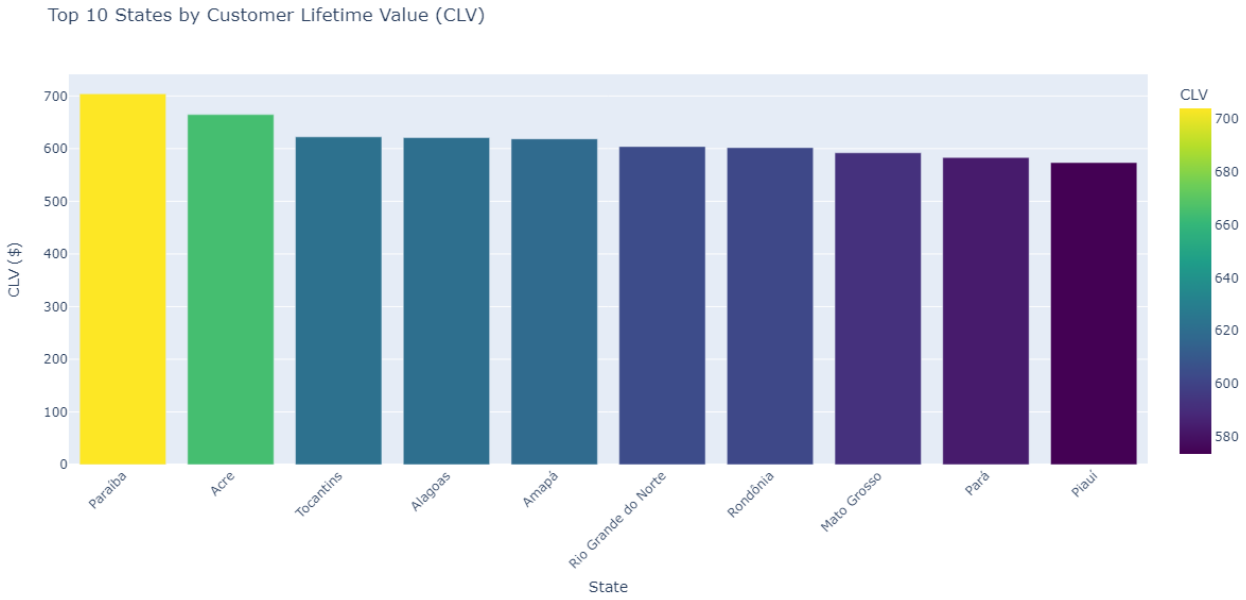


Figure 16 **Customer Lifetime Value (CLV) by State (Bar Chart)**

### 4.2.1 New, Existing, and Cumulative Customers Over Timetable

The data in the table demonstrate growing number of customers recorded per year over time that illustrate positive outcomes of customer acquisition tactics as shown in Fig 17. A consistent monthly number of customers existing gives an impression of steady customer base that can provide stable pipeline for further growth. The fact that more and more individuals are becoming customers indicates having both broadened your market as well as successful marketing campaigns. The number of new customers may be hugely affected according to whether there is a seasonal trend or not as well as oiling of different promotional tools. The use of these data in establishing customer relationship management policies can be seen in twofold by many organizations. While ensuring customer retention, another important motive is to increase retention rates.

New, Existing, and Cumulative Customers Over Time

| order_month | existing | new | Total | Cumulative_Customers |
|---|---|---|---|---|
| 2016-09 | 0 | 2 | 2 | 2 |
| 2016-10 | 0 | 300 | 300 | 302 |
| 2016-12 | 0 | 1 | 1 | 303 |
| 2017-01 | 1 | 744 | 745 | 1048 |
| 2017-02 | 3 | 1693 | 1696 | 2744 |
| 2017-03 | 6 | 2568 | 2574 | 5318 |
| 2017-04 | 20 | 2316 | 2336 | 7654 |
| 2017-05 | 28 | 3519 | 3547 | 11201 |
| 2017-06 | 40 | 3076 | 3116 | 14317 |
| 2017-07 | 51 | 3799 | 3850 | 18167 |
| 2017-08 | 62 | 4108 | 4170 | 22337 |
| 2017-09 | 77 | 4048 | 4125 | 26462 |
| 2017-10 | 89 | 4361 | 4450 | 30912 |
| 2017-11 | 123 | 7138 | 7261 | 38173 |
| 2017-12 | 112 | 5376 | 5488 | 43661 |
| 2018-01 | 131 | 6911 | 7042 | 50703 |
| 2018-02 | 110 | 6351 | 6461 | 57164 |
| 2018-03 | 146 | 6866 | 7012 | 64176 |
| 2018-04 | 163 | 6639 | 6802 | 70978 |
| 2018-05 | 187 | 6556 | 6743 | 77721 |

Table 2 **Table for New, Existing, and Cumulative Customers**

### 4.2.2. Trend Analysis of New vs Cumulative Customers Graph using Line Chart

The line graph as shown in Fig 18, reflects the consistent upwards rise in cumulative customers, an indication that the business is still experiencing business growth. Nevertheless, the sluggish emergence of new customer fills shows that the process of market penetration is slowly starting to reach the plateau implying the need for a strategic evaluation. Such a gap (between one off and returning customers) evidences the necessity of developing customer retention strategies in order to proceed with company's sales and success. This visualization provides easy viewing of the noteworthy trends, which in turn allows in later in detection of the anomalies or the fluctuations. The patterns extracted from big data informs the entire process of furnishing the future forecasting and busting marketing and customer making endeavors of the small or the big enterprises in the market.


Figure 17 **Trend Analysis of New vs Cumulative Customers Graph**

### 4.2.3. Monthly User Retention Analysis

The Customer Engagement Analysis Table on a Monthly Basis presents the overall data for user engagement month over month as shown in Fig 19 whereas the brand illustrates progress in customer attracting, the relatively low retention rates reveal that there could be problems with keeping the engagement of customer's longer term. The total retention rate curve indicates overall customer loyalty trends, displaying the fact that the rate of customer retention is increasing slowly over time. Such mounting could be the result of the improvements that have been made in customer delight services, product quality, or customer loyalty initiatives. Base on range of activities we can have targeted retention campaign and programs that will ensure that the appearance will be enhanced. Moreover, the same programs will aid in increasing the retention rates dramatically.

Monthly User Retention Analysis

| Month | Total Users | Retained Users | Retention Rate (%) | Cumulative Retention Rate (%) |
|---|---|---|---|---|
| 2016-09 | 2 | 0 | 0 | 0 |
| 2016-10 | 300 | 0 | 0 | 0 |
| 2016-12 | 1 | 0 | 0 | 0 |
| 2017-01 | 745 | 1 | 0.13 | 0.13 |
| 2017-02 | 1696 | 3 | 0.18 | 0.31 |
| 2017-03 | 2574 | 4 | 0.16 | 0.47 |
| 2017-04 | 2336 | 13 | 0.56 | 1.03 |
| 2017-05 | 3547 | 14 | 0.39 | 1.42 |
| 2017-06 | 3116 | 17 | 0.55 | 1.97 |
| 2017-07 | 3850 | 17 | 0.44 | 2.41 |
| 2017-08 | 4170 | 23 | 0.55 | 2.96 |
| 2017-09 | 4125 | 32 | 0.78 | 3.74 |
| 2017-10 | 4450 | 32 | 0.72 | 4.46 |
| 2017-11 | 7261 | 37 | 0.51 | 4.97 |
| 2017-12 | 5488 | 40 | 0.73 | 5.699999999999999 |
| 2018-01 | 7042 | 15 | 0.21 | 5.909999999999999 |
| 2018-02 | 6461 | 26 | 0.4 | 6.31 |
| 2018-03 | 7012 | 25 | 0.36 | 6.67 |
| 2018-04 | 6802 | 32 | 0.47 | 7.14 |
| 2018-05 | 6743 | 44 | 0.65 | 7.79 |
| 2018-06 | 6079 | 38 | 0.63 | 8.42 |

Table 3 **Monthly User Retention Analysis Table**

### 4.2.4. Retention Rate Trend Over Time Graph

The Cumulative Retention Rates Trend over Time Graph as shown in Fig 20. demonstrates a trend of continuous increase, evidence of better customer retention and loyalty practices. Although the site maintains a higher monthly retention rate, there are still room for advancement in the strategies of continuity management. Perhaps in a month, but when we diligently assess the relationship between the monthly and cumulative retention rates, a clear observation is that the sustained customer relationships facilitate higher retention success rates. Probably, the slew of people engaging in it is an effect of long-term efforts like loyalty programs and high level of customer service. The graph put forward that the effective retention of the costumers needs continuous expanses of the efforts throughout time. Hence, such efforts result in hierarchical positive impact over time.

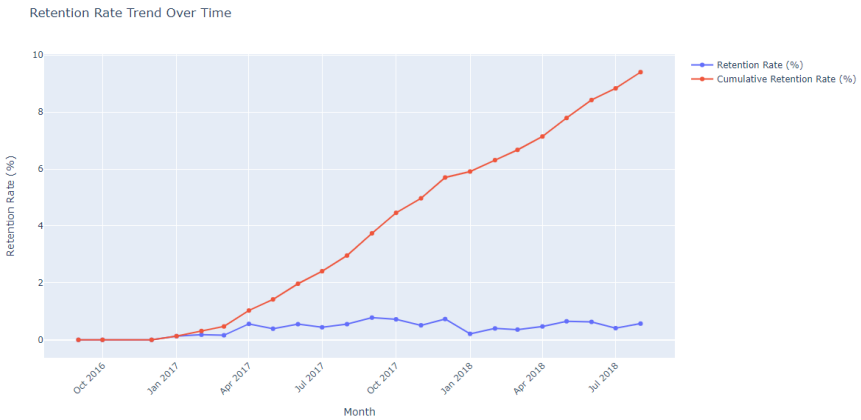


Figure 18 **Retention Rate Trend Over Time Graph**

### 4.2.5 RFM Scatter Plot

With the RFM Scatter Plot at hand, we can see the comprehensive picture of customer buying behavior as shown in Fig 21, thus we can identify distinguish patterns that can inform the execution of specific marketing approaches. Firstly, a quite impressive number of users whom purchased for only one time signify a market potential of converting users to first-time buyers with help of proper retention tactics. In addition to that, the customer profile of the spotty high-frequency customers tends to be differentiated, showing the loyal customer group that buys on a regular basis. Such customers would make up a major target group which it is necessary to build individualized links with to ensure their staying loyal. Lastly, this chart that shows customers' recency values distributed throughout different time scales reveals that different customers have diverse purchasing pattern which calls for personalized marketing approaches to effectively engage customer bases according to their specific needs and demands.

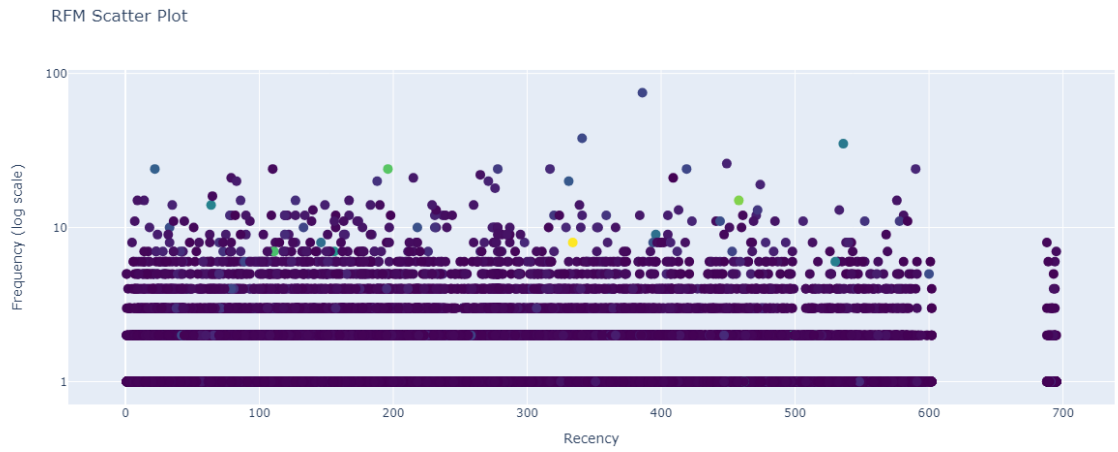


Figure 19 **RFM Scatter Plot**

### 4.2.6. Relationship between Recency and Monetary Value with Frequency

The correlation of recency and currency value with all frequency demonstrate interesting peculiarities in customers' purchasing behavior as shown in Fig 22. To start with, there is no straight reference between recent purchases of individual customers and higher monetary values. This shows the incongruousness between the two. The next point concerns the fact that the visualization reveals the low frequency of high-end customer, hinting to the possibility of going for potential engagement strategies able to capitalize on the remarkable spending power of highly committed purchasers. Ultimately, the frequency gradient, portrayed by color, demonstrates one major trend: the frequency is lowering as it is not constantly increased as seen amongst all customers regardless of the extent of their recency or their pay level. This insight reiterates the purpose of the aiming campaigns to retain the buyers and thus engage them, as different market segments undergo different purchase cycles.

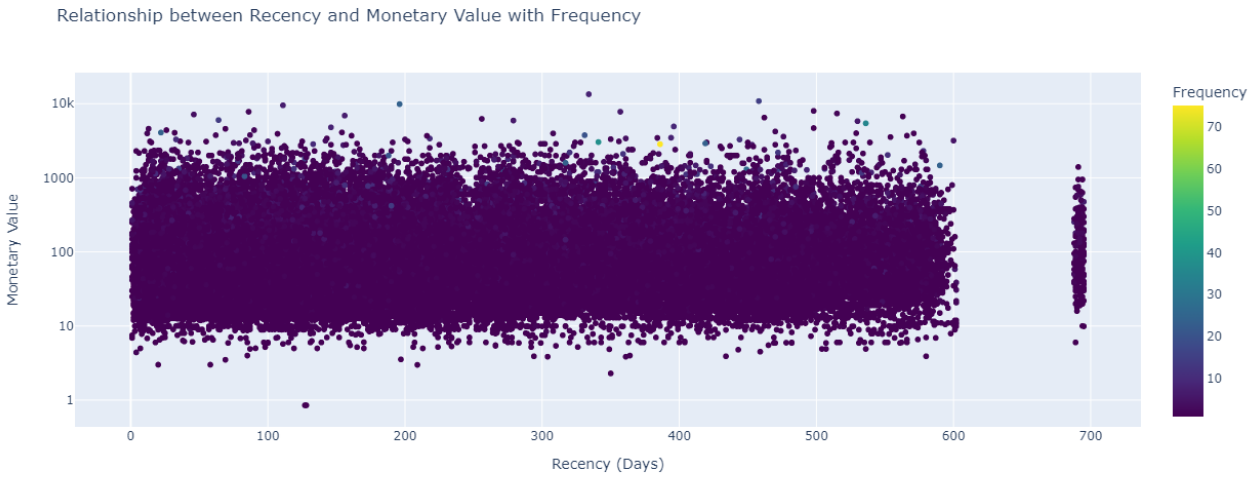


Figure 20 **Relationship between Recency and Monetary Value with Frequency**

### 4.3.1 Customer Segmentation using RFM Analysis

At this time of the project, the company employed the segmentation tool of RFM (Recency, Frequency, Monetary) to identify the right customer base for its products. As far as the segmentations were concerned these were done by choosing and studying people who have recently purchased the product(s) or the value they brought to our organization. Through the help of that segmentation, the business could clearly distinguish between each demographic group that was a potential customer. Each group had individual characteristics and potential value for the organization. The team's analysis helped the company build a customized marketing plan designed to improve customer engagement, drive conversion, and spark a high rate of retention. This part of the report describes the RFM analysis methodology and its results in more detail as shown in Table 2. After that, the group members carry on with the customer segmentation prospects. Aim is to divide the customer base into actionable segments such as those demanding additional price incentives, those requiring cross-selling scheme and those very frequent shoppers.

| customer_unique_id | Recency | Frequency | Monetary Value ($) | R_Score | F_Score | M_Score | RFM_Score |
|---|---|---|---|---|---|---|---|
| 0000366f3b9a7992bf8c76cfdf3221e2 | 112 | 1 | 129.90 | 2 | 1 | 4 | 214 |
| 0000b849f77a49e4a4ce2b2a4ca5be3f | 115 | 1 | 18.90 | 2 | 1 | 1 | 211 |
| 0000f46a3911fa3c0805444483337064 | 537 | 1 | 69.00 | 5 | 1 | 2 | 512 |
| 0000f6ccb0745a6a4b88665a16c9f078 | 321 | 1 | 25.99 | 4 | 1 | 1 | 411 |
| 0004aac84e0df4da2b147fca70cf8255 | 288 | 1 | 180.00 | 4 | 1 | 4 | 414 |

Table 4 **Customer Segmentation using RFM Analysis**

**4.3.2. Distribution of Customers by Segment**

The segmentation process categorizes customers into distinctive groups based on their RFM score combinations, each bearing unique characteristics: The segmentation process categorizes customers into distinctive groups based on their RFM score combinations as shown in Table 3, each bearing unique characteristics:

1. **Champions** cardholders (555) are often viewed as the benchmark of recent buyers, the so-called active, and bring rather impressive amounts of money to the company, to name a few. We are discussing about the consumers who are interested in novel products and who keep the brand name fresh when absence of tangible reasons.
2. **Loyal customers** are found even in [444, 454, 455, 445] clusters and they tend to repeat ordering and stern response to promotions. It is these clients who are ideal participants of upselling endeavors, and those among them who are likely to rave about you and the services through reviews are the most essential.
3. If the loyalty score card displays [355, 345, 354, 344, 255, 245, 254, 244, 455, 445], having a good history of recent spending, such individuals would show **loyalty** positions Strategies like rewards programs and personalized offerings would help make them stay interested for a longer duration.
4. **New Customers** are determined by a ranking system which uses scores [155, 145, 154, 144,255, 245, 254, 244, 455, 445] represent the customers that joined our customer base on the most recent point in time. In particular, establishing and introducing warm collegial relationships as well as the onset process affects these groups.
5. Re-engagement strategies are necessary because **promising** customers having scores of [254, 244, 253, 243, 352, 342, 351, 341, 452, 442, 451, 441, 553, 543, 552, 542, 554, 544, 555] indicate intermediate to significant buy frequency as well as value who haven't made any purchases recently.
6. Scores hold **attention** such as [154, 144, 153, 143, 252, 242, 251, 241, 352, 342, 351, 452, 442, 451, 441, 553, 543, 552, 542, 554, 544] are attributed to those who need focused effort and call for methods to increase their participation and value contributions.
7. Customers who will shortly **sleep**, scoring [541, 542, 543, 544], show a great deal of recency, with a low level of involvement and purchasing, indicating that wake-up ads are necessary.
8. Clients we **Cannot Lose but Losing**, rated [555, 545, 554, 544], have demonstrated strong consumption and involvement but are beginning to show symptoms of dwindling contact.
9. The **At-Risk** group, denoted by scores [541, 542, 543, 544], consists of individuals that haven't spent money or engaged in activities lately and may need extra care to prevent attrition.
10. Customers who are in **hibernation** and have scores of [451, 441, 351, 341, 352, 342, 251, 241, 252, 242, 154, 144, 153, 143] may require revival techniques as they are close to going into inactive status.
11. Others which doesn't come under any of the above combinations with a scores of [131,123, 121...etc] who has least recency and moderate frequency and least monetary value.

| Segment | RFM Score Combination | Description |
|---|---|---|
| **Champions** | [555] | Bought recently, over and over, and place the biggest orders. Probably first to buy a new product of you. Likely to advertise your brand, as well. |
| **Loyal Customers** | [455, 445, 454, 444] | Buys frequently from the brand and has a habit of responding to promotions. Can recommend them to related products with higher value. Their feedback can be also collected. |
| **Potential Loyalists** | [355, 345, 354, 344, 255, 245, 254, 244, 455, 445] | Some recent customers who paid a fair lot of money. Have membership or loyalty programs which will be designed to keep your visitors interested and offer them personalized recommendations. |
| **New Customers** | [155, 145, 154, 144, 255, 245, 254, 244, 355, 345, 354, 344] | Bought most recently. Offer onboarding assistance, let them try out your product, and start developing that connection with them. |
| **Promising** | [254, 244, 253, 243, 352, 342, 351, 341, 452, 442, 451, 441, 553, 543, 552, | Moderate and high frequency plus money amount value, low recency. |

| | | |
|---|---|---|
| | 542, 554, 544, 555] | |
| **Need Attention** | [154, 144, 153, 143, 252, 242, 251, 241, 352, 342, 351, 341, 452, 442, 451, 441, 553, 543, 552, 542, 554, 544] | Low contact points (visit rate/recency), less than medium frequency purchase and medium or low price. |
| **About to Sleep** | [541, 542, 543, 544, 541, 542, 543, 544] | High recency, low bundling diameter, low price and Qty. |
| **Cannot Lose Them But Losing** | [555, 545, 554, 544, 554, 544, 554, 544] | Small time horizon, big money volume. |
| **At Risk** | [541, 542, 543, 544, 541, 542, 543, 544] | Reflection of the recency, frequency or monetary value which is not too high or low. |
| **Hibernating Customers** | [451, 441, 451, 441, 351, 341, 352, 342, 251, 241, 252, 242, 154, 144, 153, 143, 154, 144, 153, 143] | Proximity effect, low or moderate frequency at minimum cost |
| **Losing But engaged** | [554, 544, 554, 544, 553, 543, 553, 543] | Moderate refrains, high words and money rates. |
| **Lost Customers** | [145, 135, 144, 134, 245, 235, 244, 234, 145, 135, 144, 134, 245, 235, 244, 234] | Regular purchases, at low recency, with a high frequency and a high monetary value.er modifications! |
| **Others** | [141,123,131] etc | Any combination that doesn't fit it into the given RFM scores combinations |

Table 5 **Distribution of Customers by Segment**

### 4.3.3. Distribution of Customers by Segment

The distribution analysis of customers by the segment, we find several leading points about the customer categorization as shown in Fig 23. At first, there are two biggest clusters - 'Other' and 'New Customers' - these show that a number of new customers joined the group and a wide range of different other categories inside the customers' base. Other indicates the other possible of RFM scores which makes it least valuable as it has low frequeny and recency with least to moderate monetary value, On the other hand, the segments like 'Potential Loyalists' and 'Champions' indicates the areas where development is possible and the strategies applied could enhance customer loyalty and increase their spending. In line with this, the fact that the 'At Risk' and 'Hibernating Customers' segments are filled up with a considerable number of customers points towards a potential for re-engagement initiatives not only to save these customers but also to turn them into valuable ones, which, in turn, suggests the significance of targeted retention strategies for building up customer lifetime value and the general health of the enterprise.

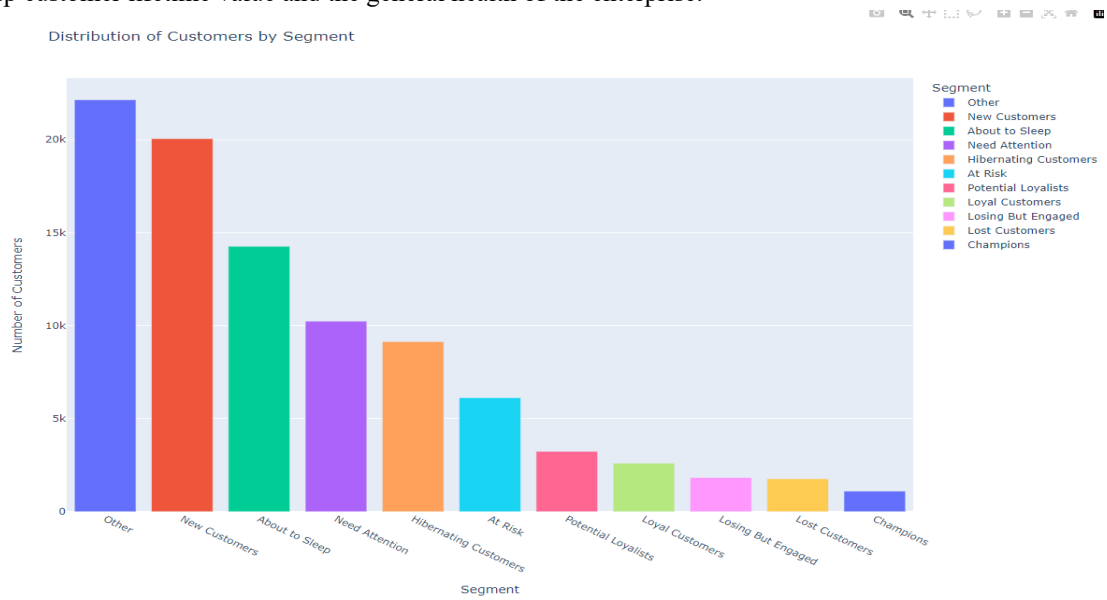


Figure 21 **Distribution of Customers by Segment**

#### 4.4.1. Customer Segmentation using K means Clustering:

We divide customers into clusters by their recency, frequency, and monetary value (RFM), the clusters are made using the K-means algorithm. Initially, the data is being prepped by picking informative options and converting them into same units of measurement. The Elbow Method and Distortion Score, which are implemented, make it possible to arrive at the best quantity of classes required for segmentation and reveal the "trade-off" between detail and simplicity.

## 5. RESULTS

### 5.1. Elbow Method to find best K-Value

After choosing features of RFM, we performed clustering, and to find the optimum K, the Elbow Method is, indeed, the approach suitable for determining the optimum number of groups in a particular dataset. The curve depicted in this analysis shows the classical WCSS drop as the number of clusters goes up, a typical situation when we speak about the clustering schemes as shown in Fig 24. But it is seen when the WCSS is reduced abruptly near 4 clusters that the rate of decrease from this point on begins to be a little slower. This insinuates that the adoption of new clusters may not be able to bring so much of new dependent variables. Hence, the elbow implies the number of clusters to be optimized at 4 when data detailing is a concern, but clustering simplicity should not be neglected. It results in the best clustering analysis for the purpose of the presented data.

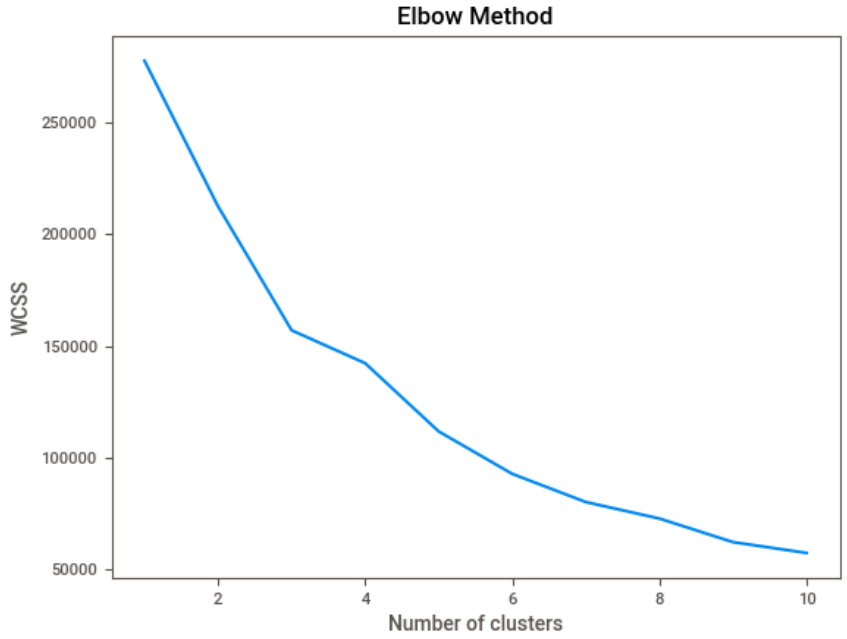


Figure 22 **Elbow Method**

### 5.2. Distortion Score Elbow for K-Means Clustering

The Distortion Score Elbow method also is a great tool that allows us to determine the optimal number of clusters in K-Means clustering. Fig. 25 shows the method depicted in detail. Plot demonstrates 8 as an inflection point, where the curve starts to flatten down. This suggests that the chance of reducing within-cluster variance will be rather low when the number of clusters exceeds this stage. And the elbow-corresponding score, 71122 approximately, becomes a clue for the intra-cluster variance at that position. Additionally, adding a second y-axis showing the computation time, it becomes evident that the computation costs increase as the number of clusters grow, thus, the proper levels of model complexity and efficiency have to be balanced.

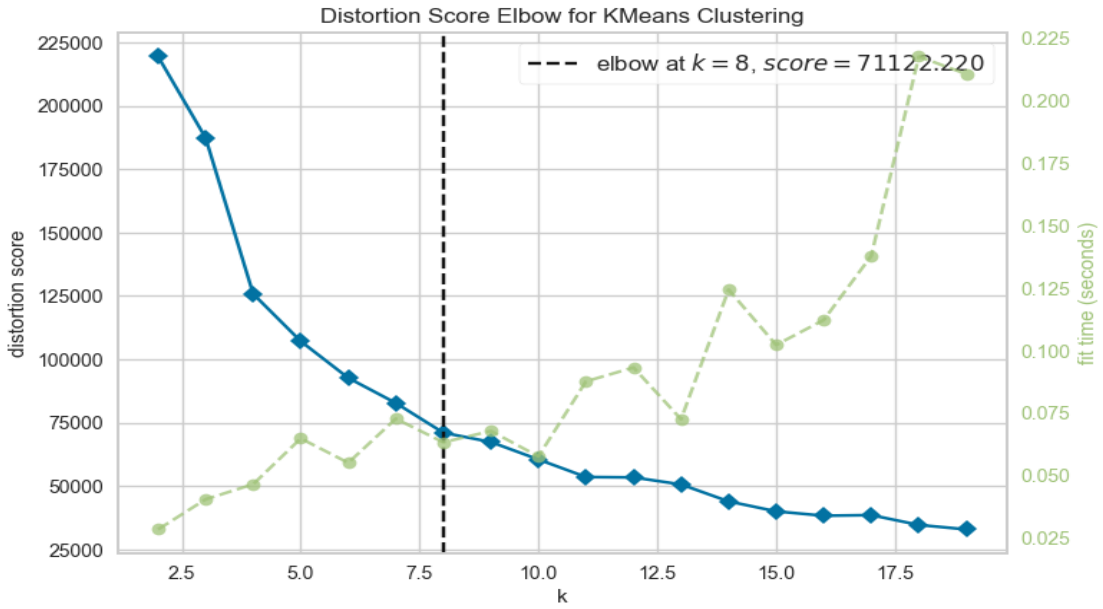


Figure 23 **Distortion Score Elbow for K-Means Clustering**

### 5.3. Clustering using Silhouette Score and Davies-Boulding Score

This proves that with 2 clusters, there is a substantial overlap between them, and they are not well separated in the result too, as shown by a relatively low silhouette score (0.40) and a high Davies-Bouldin score (1.16). If the k-means algorithm goes from 3 to 3 clusters, both scores are raised; the silhouette score reaches up to 0.42 and the Davies-Bouldin score drops down to 1.06 this way, thus, maintain the separation well and minimize the overlapping among clusters. At the beginning, 4 clusters serve better the purpose. With an improved silhouette score at 0.43 and a Davies-Bouldin score 0.87, there is a fine clustering structure. We work better with five clusters and maximum silhouette (0.45) and minimum Davies-Bouldin (0.81) measures which have highest value of silhouette and minimum value of Davies-Bouldin scores that are required to enhance the clusters' class identifications. One thing that fails to tally is that the silhouette score is similar to what it is in 6 clusters, but the Davies-Bouldin score skyrockets to 0.94 when the number of clusters are incremented, indicating that the extra cluster is useless. In the case of crossing 5 steps above, it will result in blurring of the clusters and hence, a drop in the silhouette score will be from 0.37 to 0.31 and an increase in the Davies-Bouldin score will be from 0.31 to 0.91. 8 clusters better themselves, a bit, in the silhouette score and go down to 0.41 and D-B score of 0.80, but the silhouette score is still lower than the population of 5 clusters, so the optimal number of clusters for our data set might be 5, that distinctly strikes the balance between silhouette and D-B scores.

```
Number of clusters: 2, Silhouette score: 0.40071756114177215, Davies-Bouldin score: 1.1622255123476308
Number of clusters: 3, Silhouette score: 0.42248355078371447, Davies-Bouldin score: 1.0564002548685387
Number of clusters: 4, Silhouette score: 0.434034041787554, Davies-Bouldin score: 0.8733769556738606
Number of clusters: 5, Silhouette score: 0.4479354598517632, Davies-Bouldin score: 0.8158748222336616
Number of clusters: 6, Silhouette score: 0.44774895747210885, Davies-Bouldin score: 0.9444632306588717
Number of clusters: 7, Silhouette score: 0.3739230569656783, Davies-Bouldin score: 0.9127374387861243
Number of clusters: 8, Silhouette score: 0.41160559517376044, Davies-Bouldin score: 0.8053612366140266
```

Figure 24 **Results of clustering using Silhouette Score and Davies-Boulding Score**

### 5.4 Customer Segmentation with Clustering

The inclusion of cluster assignment has helped to enhance customer segmentation by adding a separate dimension on customers categorization based on their similarities to own each other on within the dataset. Like, the customers belonging to 'e2', who between them have an RFM Score of 214, and are apparently 'Hibernating Customers', which is a cluster group that makes significant purchases, despite not having done so very often or recently, shows a lot in common with others in the same group. In this regard, customers like '3f' in Cluster 1 with a score of 211, which is defined as 'New Customers', give us a clue that, among the potential customers, there may be those who recently joined us and those who are less active with considerable future income potential. Particularly, customers like '064' from Cluster 0 of 'About to Sleep' set with the score of 512, being characterised by higher recency with lower frequency and monetary dimensions, relate to those with decreasing engagement. Customers with '078' are classified to the Cluster 7; although bear the same label but with a different composition, which depict a 'About to Sleep' segmentated in several groups with varied features. This is followed by visitors to '255', a part of Cluster 3 and classified as 'Hibernating Customers', but with a remarkably higher worth when compared to all the customers in that cluster, thus, determination of their segmentation. Finally, the consolidation confirms the intricate tendencies that exist in different segments, and thus marketers' efforts of creating more particular marketing strategies and customer engagement programs that are targeted at specific value and behavioral clusters will lead to improved customer retention and value optimization.

| customer_unique_id | Recency | Frequency | Monetary Value ($) | R_Score | F_Score | M_Score | RFM_Score | Segment | Cluster |
|---|---|---|---|---|---|---|---|---|---|
| 0000366f3b9a7992bf8c76cfdf3221e2 | 112 | 1 | 129.90 | 2 | 1 | 4 | 214 | Hibernating Customers | 1 |
| 0000b849f77a49e4a4ce2b2a4ca5be3f | 115 | 1 | 18.90 | 2 | 1 | 1 | 211 | New Customers | 1 |
| 0000f46a3911fa3c0805444483337064 | 537 | 1 | 69.00 | 5 | 1 | 2 | 512 | About to Sleep | 0 |
| 0000f6ccb0745a6a4b88665a16c9f078 | 321 | 1 | 25.99 | 4 | 1 | 1 | 411 | About to Sleep | 7 |
| 0004aac84e0df4da2b147fca70cf8255 | 288 | 1 | 180.00 | 4 | 1 | 4 | 414 | Hibernating Customers | 3 |

Table 6 **CUSTOMER SEGMENTATION WITH CLUSTERING**

### 5.5. Cluster Distribution by Segment

A targeted division of the SHARE data reveals that clusters form around certain topics that otherwise might not be clustered together. Clusters 0, 2, and 4 differ in the composition of their customer segments: Fig26 promotes. Listen to the given audio and preview the given sentence to get its contextual meaning. One of the top type that are used for the motion picture is "Other" together with "New Customers". Notwithstanding, the substantially represented sections, 'Potential Loyalists' and 'Loyal Customers,' of the 'Core Consumers' category are the signs of methods for retention-oriented actions. Cluster 6 stands out from all the other clusters since it consists of a higher proportion of the customers from the two more serious categories, namely, 'At Risk' and 'Need Attention'. The actual message that reactive churn prevention activities and re-engagement are required is given level-headedly by the aforementioned clustering.

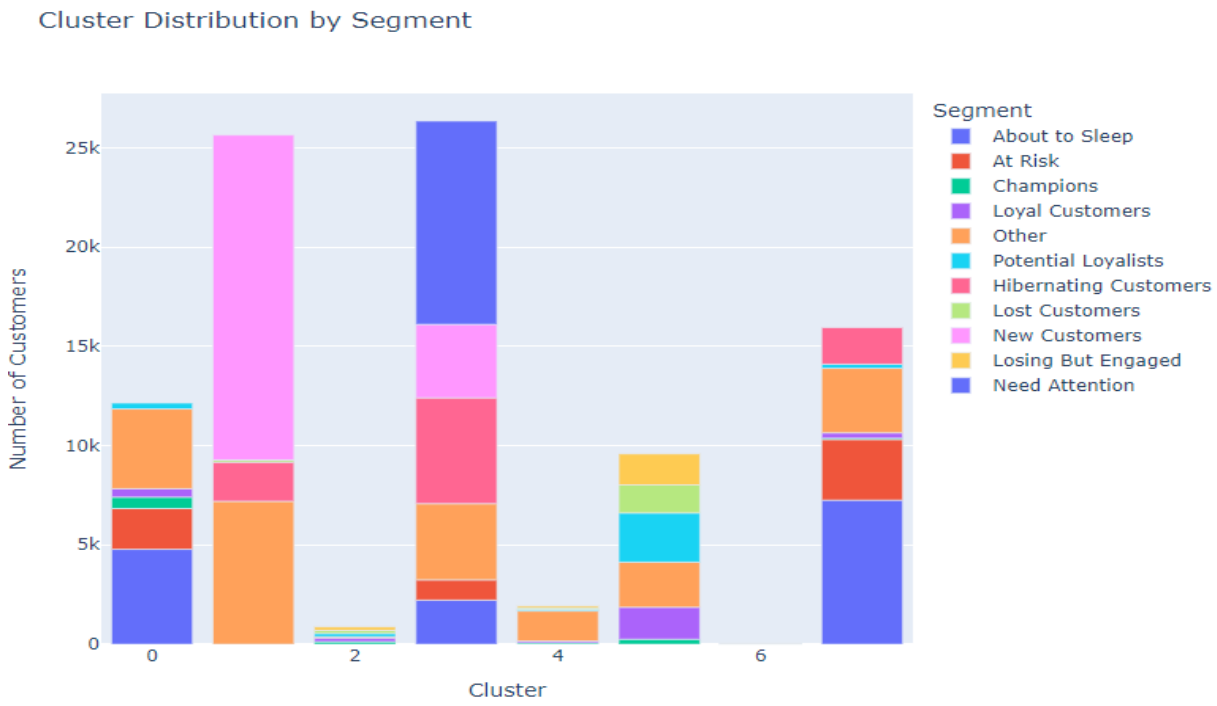


Figure 25 **Results of clustering using Silhouette Score and Davies-Boulding Score**

## 6. CONCLUSION AND FUTURE RESEARCH QUESTIONS

### 6.1. Summarizing Results

RFM client's segmentation and clustering analytics yielded an explanation of varied behavior types which is in line with segmentation outcomes. By clustering people in RFM groups, we have made our message much more targeted and meaningful to particular customer segments, hence the latter become more possible customers to be engaged with specific marketing campaigns. Synergetic segmentation revealed some segments of consumers such as 'Champions' who are loyal, frequent customers who make recent purchases, and 'Loyal Customers' who respond to good offers though they may make only some purchases. Besides, there is the 'Potential Loyalists'who have the possibility for substantial spending and have just bought recently. Retailers can also concentrate on like eClusters such as 'About To Sleep' which contains most recency data but less engagement versus 'Dormant' Cluster customers which are less engaged than 'Hibernating' Customers but they have a high financial value. Our research directions will include segmenting the customers, crafting in a way marketing campaigns and retaining strategies for required customer segments, in order to strengthen customer engagement and customer lifetime value, which leads to improved customer relationships. RFM analysis in particular is a very efficient technique being combined in clustering methods to diagnose customer behaviours and to customize customer relationship management strategies.

### 6.2 Conclusion

In our study, we focused on Customer Lifetime Value Analysis and the Customer Segmentation with K-means clustering of the data based on the recency (when the client is close to a market), frequency (when the client purchases more than one product) and monetary value (RFM). Through Elbow method and distortion score elbow strategy, we identified the ultimate figure of clusters that is a piece of the mixture of particular segmenting and analysis simplicity. The results stemming the analysis indicated intricate differences within them, for example, high spenders that made fewer buys and new customers that demonstrated immense growth potential. Thus, this strategy allows us to craft the best marketing strategies possible, distributing resources most rationally and ensuring a long-term relationship with customers. Our project, in the end, prepare us with practical principles to help us understand and govern our customers well. Excellent performance highly relates to sustainability and growth of the businesses.

### 6.3. Future work

The exploratory analytical part of this project has revealed a lot of fascinating questions that I'd like to find more information about. Likewise, we would like to focus on the features of a specific group, for example, what triggers them to shop in a given store or what conditions make them loyal to certain brands. Also, we always want to know which marketing channels work better for a specific group of people. Additionally, we are interested in how specific external factors which for instance are change in economy or competitor's advertisement affect consumer interaction. In case we had more time, we would additionally check how the patterns of clients change over time and measure whether our strategies deliver to successful performance measures indeed. That'd be intriguing to, to observe how the segmentation of customers could be successfully integrated with other elements of the business i.e. product development or improvement in the delivery process. To summarize, we can carry out more so that we have more insights on our customers and help us to improve our business as well.

## REFERENCES:


1. Smith, M. D., & Brynjolfsson, E. (2001). Consumer decision-making at an Internet shopbot: Brand still matters. The Journal of Industrial Economics, 49(4), 541-558. https://doi.org/10.1111/1467-6451.00162

2. Brynjolfsson, Erik and Hu, Yu Jeffrey and Smith, Michael D., From Niches to Riches: Anatomy of the Long Tail. Sloan Management Review, Vol. 47, No. 4, pp. 67-71, Summer 2006, Available at SSRN: https://ssrn.com/abstract=918142

3. Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of marketing research, 42(4), 415-430. https://doi.org/10.1509/jmkr.2005.42.4.415

4. Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., ... & Sriram, S. (2006). Modeling customer lifetime value. Journal of service research, 9(2), 139-155. https://doi.org/10.1177/1094670506293810

5. Kumar, V., & Reinartz, W. (2018). Customer relationship management. Springer-Verlag GmbH Germany, part of Springer Nature 2006, 2012, 2018. https://link.springer.com/book/10.1007/978-3-662-55381-7

6. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining,inference, and prediction (Vol. 2, pp. 1-758). New York: springer. https://link.springer.com/book/10.1007/978-0-387-21606-5

7. Oh, S., Ji, H., Kim, J. et al. Deep learning model based on expectation-confirmation theory to predict customer satisfaction in hospitality service. Inf Technol Tourism 24, 109–126 (2022). https://doi.org/10.1007/s40558-022-00222-z

8. Haenlein, M., Kaplan, A. M., & Schoder, D. (2006). Valuing the real option of abandoning unprofitable customers when calculating customer lifetime value. Journal of marketing, 70(3), 5-20. https://doi.org/10.1509/jmkg.70.3.005

9. Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The theoretical underpinnings of customer asset management: A framework and propositions for future research. Journal of the academy of marketing science, 32(3), 271-292. https://doi.org/10.1177/0092070304263341

10. Martin, K.D., Murphy, P.E. The role of data privacy in marketing. J. of the Acad. Mark. Sci. 45, 135–155 (2017). https://doi.org/10.1007/s11747-016-0495-4

11. P. W. Laksono and D. A. Wulansari, "K-means clustering for e-commerce customer segmentation," Journal of Big Data, vol. 9, no. 1, p. 2, 2022.

12. T. Z. Win and N. N. Bo, "Deep learning for customer churn prediction," Journal of Big Data, vol. 7, no. 1, p. 25, 2020.

13. Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021, July). Deep learning for customer churn prediction in e-commerce decision support. In Business Information Systems (pp. 3-12). https://doi.org/10.52825/bis.v1i.42

14. Brazilian E-Commerce Public Dataset by Olist. (2021, October 1). Kaggle. https://www.kaggle.com/datasets/olistbr/brazilianecommerce/data

# CONTIRBUTION OF MEMBERS ON PROJECT

**Akhila Vajeer:**

- Led the data collection process by sourcing and organizing the relevant customer data.
- Contributed to the initial brainstorming sessions to define the problem statement and objectives.
- Actively participated in the RFM analysis, interpreting the results, and deriving insights.
- Assisted in drafting the final project report, summarizing key findings and recommendations.

**Chandrakala Moru:**

- Conducted extensive literature review on customer lifetime value (CLV) and segmentation techniques.
- Played a key role in preprocessing the dataset, including data cleaning and feature engineering.
- Implemented the K-means clustering algorithm for customer segmentation and analyzed the results.
- Collaborated with other team members to create visually appealing presentations and slides.

**Jaikar Tridandapani:**

- Conducted statistical analysis on the RFM data, calculating metrics such as recency, frequency, and monetary value.
- Utilized Python libraries for data visualization, creating insightful graphs and charts for presentation.
- Actively participated in discussions on the interpretation of clustering results and their implications for business strategy.
- Worked on final project documentation and references with APA style.

**Mokshith Talari:**

- Worked on usage of the Elbow Method and Distortion Score Elbow techniques for determining the optimal number of clusters.
- Contributed to the visualization of clustering results and the comparison of different cluster configurations.
- Collaborated with team members to synthesize findings and draw actionable conclusions from the analysis.
- Played a role in delivering the final presentation, explaining technical aspects in a clear and accessible manner.

**Tarun Reddy Rapole:**

- Led the formulation of the problem statement and project objectives, ensuring alignment with business goals.
- Conducted in-depth discussions on the interpretation of RFM analysis and clustering results.
- Reviewed and edited the project report, focusing on clarity, coherence, and adherence to formatting guidelines.
- Presented key findings and recommendations during team meetings and contributed to the synthesis of insights.