ABW508D Analytics Lab

Supervisor

Assoc. Prof. Dr Teh Sin Yin

Customer Segmentation and Lifetime Value Prediction Using Machine Learning Approaches

By

Ayman Mohammed Altaweel

Master of Business Analytics

P-EM0391/22

Year of Submission

2022

Table of Contents

ACKNOWLEDGEMENT

I want to express my gratitude to my supervisor, Dr. Teh Sin Yin, for giving me the chance to conduct this research. In addition, I want to thank her for her patience, helpful advice, input, and comments during the study and writing of this report. Furthermore, I want to thank my colleagues for all the information they shared throughout the discussion sessions. Finally, I want to thank my family for their unconditional support, compassion, love, and prayers. Above all, I want to thank God for His favor and grace to me.

LIST of ACRONYMS

APIs	Application Programming Interfaces
BG/NBD	Beta Geometric Negative Binomial Distribution
BI	Business Intelligence
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
CLTV	Customer Lifetime Value
CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
EDA	Exploratory Data Analysis
GMM	Gaussian Mixture Model
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
Industry 4.0	The Fourth Industrial Revolution
IS	Information System
KPI	Key Performance Indicator
ML	Machine Learning
PCA	Principal Component Analysis
RFM	Recency, Frequency and Monetary Value
SQL	Structured Query Language
XYZ	The company name in the study

LIST of TABLES

LIST of FIGURES

ABSTRACT

In the era of the industrial revolution 4.0, many businesses adopted digitalization strategies. As a result, many retail businesses shifted to operate online generating a lot of data. Marketing professionals may enhance their performance by taking advantage of this data.

This study follows the CRISP-DM framework to solve specific business problems for a real Egyptian online retail company, named: XYZ. Firstly, machine learning clustering algorithms (KMeans, Gaussian Mixture and HDBSCAN) are used to perform RFM customer segmentation. The KMeans model with 4 clusters is the best-performing model with a Silhouette score = 0.521. Secondly, BG/NBD and Gamma-Gamma models are used to predict 6-month customers' number of purchases and monetary values. Thirdly, a business intelligence dashboard is built using Tableau to track XYZ's business performance.

Finally, and based on the results, customized promotion campaigns and feasible retention campaigns are launched to increase XYZ's number of monthly orders and solve the customers' churn problem respectively.

Key Words:

CLTV Prediction - CRISP-DM - ML Clustering Algorithms - RFM Customer Segmentation - Silhouette Analysis

CHAPTER 1 INTRODUCTION

1.1 Online Retail Industry

The retail industry is a crucial business sector. It simply represents the sale of products or services to buyers. The retailers buy bulk from manufacturers, either directly or through a wholesaler, and then sell in smaller amounts for profits. It is the last link in the supply chain that connects manufacturers and consumers.

Since ancient times, there were retail marketplaces. More than 10,000 years ago, archaeological evidence for commerce was discovered, most likely including barter systems. As civilizations advanced, barter gave way to retail transactions including currency. Around the 7th millennium BCE, selling and buying are supposed to have evolved in Asia Minor (modern Turkey). In ancient Greece, markets were held in the agora, an open arena where commodities were exhibited on mats or temporary stalls on market days. Trade took place at the forum in ancient Rome. The Roman forum was likely the first permanent retail shopfront. According to a recent study, China has a long history of early retail systems. Chinese packaging and branding were used to convey family and place names as early as 200 BCE (Wikipedia, 2022).

Thanks to The Fourth Industrial Revolution (Industry 4.0) and its corresponding technical advances, the concept of digitalization evolved over time to enable companies to achieve unparalleled organizational excellence. Especially due to the COVID-19 epidemic, the retail industry was heavily affected, and some retail companies were negatively harmed and faced bankruptcy because all the operations needed to access the products and goods by physical engagement of all employees and managers, which was prohibited by some regulations (Ramírez, 2023).

Nowadays, online retailing emerged in a lot of marketplaces worldwide. Online retailing, sometimes known as electronic retailing, is the effective use of the Internet by customers to make purchases of products and services. It is split into two categories: firstly, Business-to-business. Secondly, Business-to-consumer: In this kind of retailing, businesses use the Internet and their websites to offer their goods and products to customers directly via websites and mobile applications (Chaudhary, 2022).

1.2 Research Background of RFM and CLTV Customer Segmentation

In light of Industry 4.0, machine learning (ML) and business intelligence (BI) go hand in hand with optimizing marketing activities for the online retail industry and customer-based companies. This enables marketers to run marketing activities more effectively yet efficiently to enhance overall business performance. In this paper the focus is on applying ML for Recency, Frequency and Monetary Value (RFM) based customer segmentation and customer lifetime value (CLTV) prediction for more customized and feasible online retail marketing strategies.

1.3 Problem Statements

At XYZ, which is a real Egyptian online retail start-up, the marketing team needs to run marketing more effectively yet efficiently. As a start-up, its performance is acceptable, however, it can perform rather better once the problems they face got solved:

Recently, there are some fluctuations in the number of orders made by customers. So, the marketing team needs to launch customized promotion campaigns to increase the orders again. However, they do not know their customers' personalities or purchasing behaviors.

Also, XYZ experiences low retention rates because there are many customers churned from the company. The team plans to launch retention and reactivation campaigns. However, they do not know how much to spend in order to be feasible and profitable.

Finally, marketers at XYZ don't have a tool to measure and assess the results of the campaigns. The post campaigns impact on the business performance is not measurable.

1.4 Research Objectives

In this research, an online retail marketing strategy needs to be put to run marketing more effectively and feasibly using business intelligence and machine learning approaches to make more data-driven online retail marketing decisions. Below are the research objectives:

1. To determine the best ML clustering model for XYZ's RFM-based customer segmentation.
2. To predict CLTV for planning and budgeting XYZ's retention and reactivation campaigns.
3. To keep track of XYZ's key performance indicators (KPIs) after launching the campaigns.

Results of the first 2 objectives may be cross-applied for launching effective feasible campaigns.

1.5 Research Questions

In this study, some questions are asked and need to be answered, the questions are:

1. Which is the best ML clustering algorithm for XYZ's RFM customer segmentation?
2. What are the customers' future worth/value based on their ordering behaviors?
3. How to track marketing campaigns' impact on XYZ's business performance?

1.6 Significance of Research

This research can show business analytics professionals, market researchers and marketing data analysts/scientists how to conduct end-to-end specific marketing analytics models and projects. These include some technical tasks such as: extracting customers' RFM data from transactional ordering databases, performing time-based cohort analysis to highlight churn problems, applying machine learning clustering algorithms on RFM customer segmentation data, comparing machine learning clustering algorithms using the Silhouette analysis, creating customer profiles to help in customized and accurate promotion campaigns, predicting customer lifetime value to plan and budget for feasible retention and reactivation campaigns and finally building business intelligence key performance indicators dashboard.

Also, this research contributes towards solving XYZ's problems, such as: creating customer profiles to help XYZ in launching customized and accurately targeted promotion campaigns, planning for feasible retention and reactivation campaigns based on customers' customer lifetime value that can also be used for the promotion campaigns feasibility and constructing key performance indicators dashboard to track the impact of marketing campaigns on business performance.

1.7 Organization of The Remaining Chapters

There are 5 chapters in this study. Chapter 1 presents the introduction of the study. Chapter 2 reviews the theoretical structure and related topics. In Chapter 3, the methodology used in this study is discussed. Chapter 4 presents the findings and analysis based on the data that resulted from the modeling phase. Chapter 5 has concluded all the research. All the references that were used in this study are listed. Finally, the appendices for this study will be displayed.

CHAPTER 2 LITERATURE REVIEW

2.1 Customer Segmentation Types

Customer segmentation is the process of breaking large customer groups into smaller groups (known as segments) based on some shared characteristics. Typically, this process involves existing and future customers. In order to divide or segment customers, researchers frequently look for shared traits including mutual needs, same hobbies, purchasing behaviors, comparable lifestyles, or even similar demographic profiles. Identifying high-yield segments, or those that are expected to be the most lucrative or have growth potential, is the overarching goal of segmentation since these segments may then be chosen for special attention (i.e. become target markets). There are several different approaches to market segmentation.

2.1.1 Demographic Customer Segmentation

Information on demographics and socioeconomics is technically easy to collect and measure. The collection and measurement of statistical data on socioeconomic and demographic aspects are not particularly difficult. The best demographic segmentation strategies are those that use a priori variables as their foundation. Age, gender, household size, household income, employment, education, religion, and nationality are just a few of the many factors in this list. Retailers are the group most likely to combine these factors with others. Household income, educational attainment, and socioeconomic position all have positive associations (Das, 2023).

2.1.2 Psychographic Customer Segmentation

Softer indicators like attitudes, opinions or even personality characteristics are included in psychographic segmentation. Segmenting respondents based on their attitudes, interests, and views is an approach that forecasts which customer groups a product will appeal to based on respondents' lifestyles, personalities, hobbies, and socioeconomic classifications (Fu, 2017).

2.1.3 Geographic Customer Segmentation

Geographic segmentation entails dividing up the audience according to the area in which they reside or are employed. Customers can be categorized in a variety of ways, including by their nation of residence, or by more specific geographic divisions, such as city, region, or even postal code. Also, the interactions with customers who live in the same region might influence a customer’s decision to choose a new service or brand (Kim. 2021). Even if geographic segmentation is the easiest type of market segmentation to understand, there are still many applications for it that businesses seldom consider. Depending on the company's goals, the targeted area should have a different size. Generally speaking, the locations targeted locations will be greater the larger the company. After all, it won't be economical to target each postcode separately with a larger potential audience.

2.1.4 Behavioral Customer Segmentation

Consumer behavior is essentially how individuals respond to a company's marketing activities. Where, when, and why customers make purchases are all impacted by psychological and cultural variables in addition to other considerations. For the purpose of identifying the traits of a certain section, some practitioners advise using behavioral segmentation. As future decisions might be carefully considered as a result of previously observed behavior, such as buying products from a specific brand or timing purchases for special occasions like birthdays, weddings, and births ...etc. It is crucial to think about the factors that influence client purchases because, if those factors fluctuate over the course of the year, it means that the customers' requirements are shifting (Fu, 2017).

2.2 Customer Segmentation Models

A customer segmentation model is a specialized method for grouping your customers according to shared traits. According to factors like age, gender, region, job title, and income, demographic segmentation could include dividing an audience into smaller groups. The objective is to tailor your messaging so that it resonates more strongly with each group that makes up your total customers. Going deeper there are more customer segmentation models, such as RFM and CLTV customer segmentation models.

2.2.1 RFM Customer Segmentation Model

The RFM utilizes these behavior-based strategies to cluster consumers by reviewing prior basket transactions. Additionally, it may be used to find client groups whose members share similar purchasing habits. The RFM model is based on the three values of Recency (How recently was the latest purchase made?), Frequency (How many purchases have been made overall?), and Monetary (How much money was spent overall?) correspond for every customer in the analysis. These values represent the customer's buying behavior.

One way to conduct RFM customer segmentation is by splitting the distribution of the customers' RFM into quartiles, the first stage in RFM is to score and rank the customers based on these three features. The highest 25 percent is rated as 1 (best) in each feature, while the lowest 25 percent is scored as 4 (worst). It's vital to remember that high numbers for Recency signify a bad situation where the customer spent a lot of time since the last purchase. As a result, the lowest 25% is given a ranking of 1, while the highest 25% is given a ranking of 4. These scores are put side by side after the quartile split mentioned above to get a segment bin value. The top clients are those with the lowest recency and greatest frequency-monetary amounts, indicating that their purchasing behavior is excellent for the company (Brandizzi, 2022).

Another way of conducting the RFM customer segmentation is by using unsupervised ML clustering algorithms. After extracting the RFM data of customers' transactional data, the data passes through preprocessing stage where the data is normalized and unskewed to meet the assumptions of the algorithms. Then, the algorithms start to group users based on their RFMs in a way that maximizes the similarities of each group's behaviors and minimizes the similarities to other groups' customers.

2.2.2 CLTV Customer Segmentation Model

Customer lifetime value is a crucial indicator to gauge a business's overall worth or profit from a customer over the course of the relationship with the company. While customer retention refers to a group of actions firms take to lessen the number of customers that churn, whereas client churn is defined as the termination of the business relationship between the company and the customer. Churn and retention rate are crucial metrics for every business and are regarded as the foundational elements of the long-term CLTV. The average profit that a customer is expected to make before churning is estimated by CLTV. Churn and retention are concepts that are frequently linked to the life cycle of an industry. Sales rise rapidly when an industry is in the growth phase of its life cycle. However, the retail industry's most difficult objective is reducing churn. According to this viewpoint, more knowledge is required to understand the causes of customer turnover in dynamic industries like online retailing (Yoseph, 2020).

CLTV's customer segmentation model is a potential solution for this issue. By using models like Beta Geometric Negative Binomial Distribution (BG/NBD) & Gamma-Gamma models, customer lifetime values can be calculated and predicted. Then based on the results, customers are grouped or segmented into groups according to the worth customer lifetime value. Finally, marketing retention and acquisition campaigns are launched after feasible customer lifetime value-based budgeting and planning and they are run effectively yet efficiently as the customers will still be profitable even after the retention and acquisition marketing expenses.

2.3 Application of Customer Segmentation in Online Retail Industry

Online retail companies now place a greater emphasis than ever on customer-oriented strategies due to rising competition and shifting market dynamics. Understanding the customers in this context is essential for maintaining or enhancing relationships with them. Companies from all around the world categorize customer behavior data thoroughly in order to take the appropriate steps. Customer segmentation, according to Parsell et al., is the practice of grouping customer data based on certain traits like preferences and spending patterns. Through segmentation, customer groups may be successfully targeted and marketing budgets can be distributed for maximum impact (Acar, 2021).

RFM customer segmentation is a marketing technique for more customized and accurate customer targeting that is commonly used in the online retail industry. It segments customers based on their historical purchasing behaviors: recency, frequency and monetary. Recency describes the interval of time since the last purchase made by a customer within the scope of the analysis. A lower value of recency indicates that the customer ordered recently and vice versa. Frequency describes how many purchases a customer makes within the scope of the analysis. The customer is more devoted to the business if the frequency value is higher. Monetary tells how much money is spent by the customer within the scope of the analysis. The higher the monetary, the more valuable the customer is (Das, 2023).

Also, ML clustering algorithms are applied widely in the area of customer segmentation. The aim of using machine learning in RFM customer segmentation is to find homogeneous groups of customers based on their RFM purchasing behaviors to enable marketers to create customer profiles on which they customize their customer targeting strategies. KMeans, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAB), GaussianMixture (GMM) and Hierarchical clustering are among the machine learning clustering algorithms that are used in customer segmentation. Every algorithm has its own pros and cons with which business analytics professionals make the pay-off between the clustering algorithms in addition to a clustering performance analysis that is called: Silhouette analysis. Silhouette analysis shows how well the ML algorithm is performing regarding the cohesion and separation of clusters, in other words how it finds similar homogeneous customer groups (Anitha, 2022).

Furthermore, and in the simplest terms, customer lifetime value is a measurement of a customer's profitability throughout his/her relationship with a company that is used by marketers to segment online retail buyers for specific feasibility and budgeting purposes. Because all businesses are interested in concentrating on the "cream of the crop customers" to establish long-lasting loyal relationships with chosen customers, CLTV constantly receives attention. Its value can be assessed and predicted to determine how valuable a customer is to the company. This information is crucial for making data-driven marketing and customer retention and reactivation decisions and for improving customer relationships and consequently enhancing business performance (Channa, 2019).

Finally, a widely used analytics tool in the online retail industry: is the business intelligence dashboard. It is a tool for data analysis that combines information management and data visualization. It collects and presents data in quick visual overviews, such as charts, graphs, and reports on a single screen using interactive components like filters and actions. It is used to evaluate and enhance business performance. Key performance indicators like revenue, customer turnover, and employee productivity are tracked and observed using them. Business intelligence has significantly changed business operations in recent years by offering a cloud-based solution to increase working capacity and grow the business on a large scale, such as using key performance indicators to improve Nike's company performance with conversion into analytical reports. One research demonstrated how KPIs might directly affect overall performance management when they are cleverly developed. Viewing historical performance, diagnosing judgment errors, and predicting and forecasting future demands, benefits, and the best strategy based on company facts and statistics are the main goals of business intelligence services (Alqhatani, 2022).

CHAPTER 3 METHODOLOGY

3.1 Crisp-DM Methodology

The CRIS-DM refers to the cross-industry standard process for data mining which is a common data science projects framework. This procedure has been used for 20 years. There are multiple steps to this process (see Figure 3.1), including business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Processing data with more than 1000 cells is a great fit for this strategy (Thoyyibah, 2022).

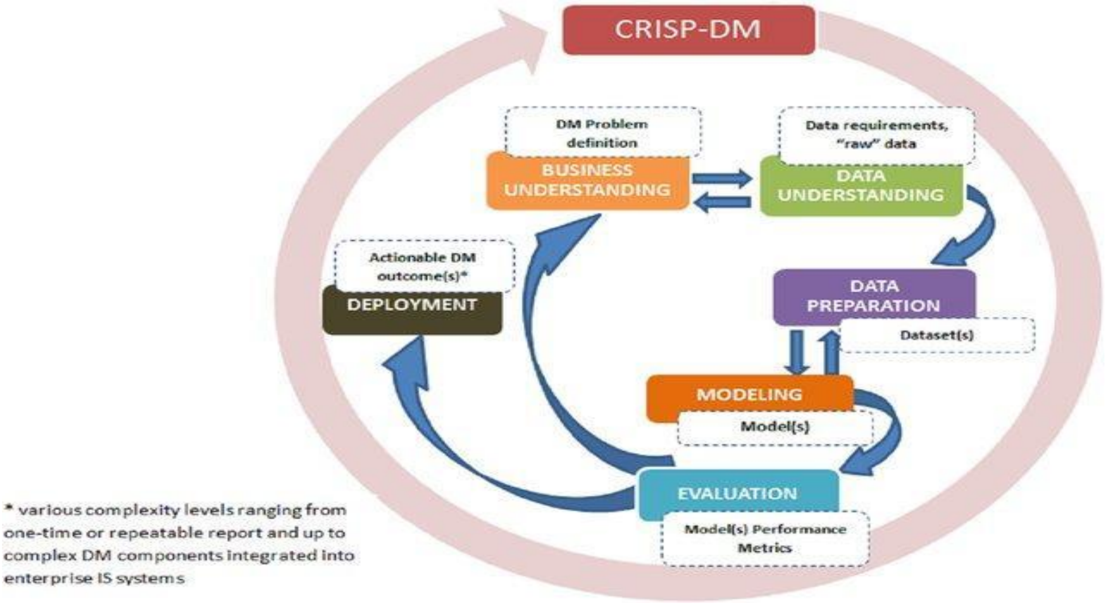

CRISP-DM
DM Problem definition
BUSINESS UNDERSTANDING
Data requirements, "raw" data
DATA UNDERSTANDING
Actionable DM outcome(s)*
DEPLOYMENT
DATA PREPARATION
Dataset(s)
MODELING
Model(s)
EVALUATION
Model(s) Performance Metrics
* various complexity levels ranging from one-time or repeatable report and up to complex DM components integrated into enterprise IS systems

Figure 3.1: CRISP-DM diagram

(Plotnikova, 2020)

3.2 Application of CRISP-DM in XYZ's Case Study

In this research, the cross-industry standard process for data mining framework is applied to deal with XYZ’s problems. ML clustering algorithms are used for RFM-based customer segmentation that enhances marketers' capabilities for launching more customized and accurately targeted marketing campaigns. Also, the customer's lifetime value is predicted for budgeting and planning feasible reactivation and retention campaigns. For tracking business performance and the impact of marketing activities on business, a business intelligence dashboard was built as well using data visualization and business intelligence software: Tableau. The end-to-end project follows a number of phases according to the CRISP-DM framework (see Figure 3.2) that is all implemented using a programming language: Python.

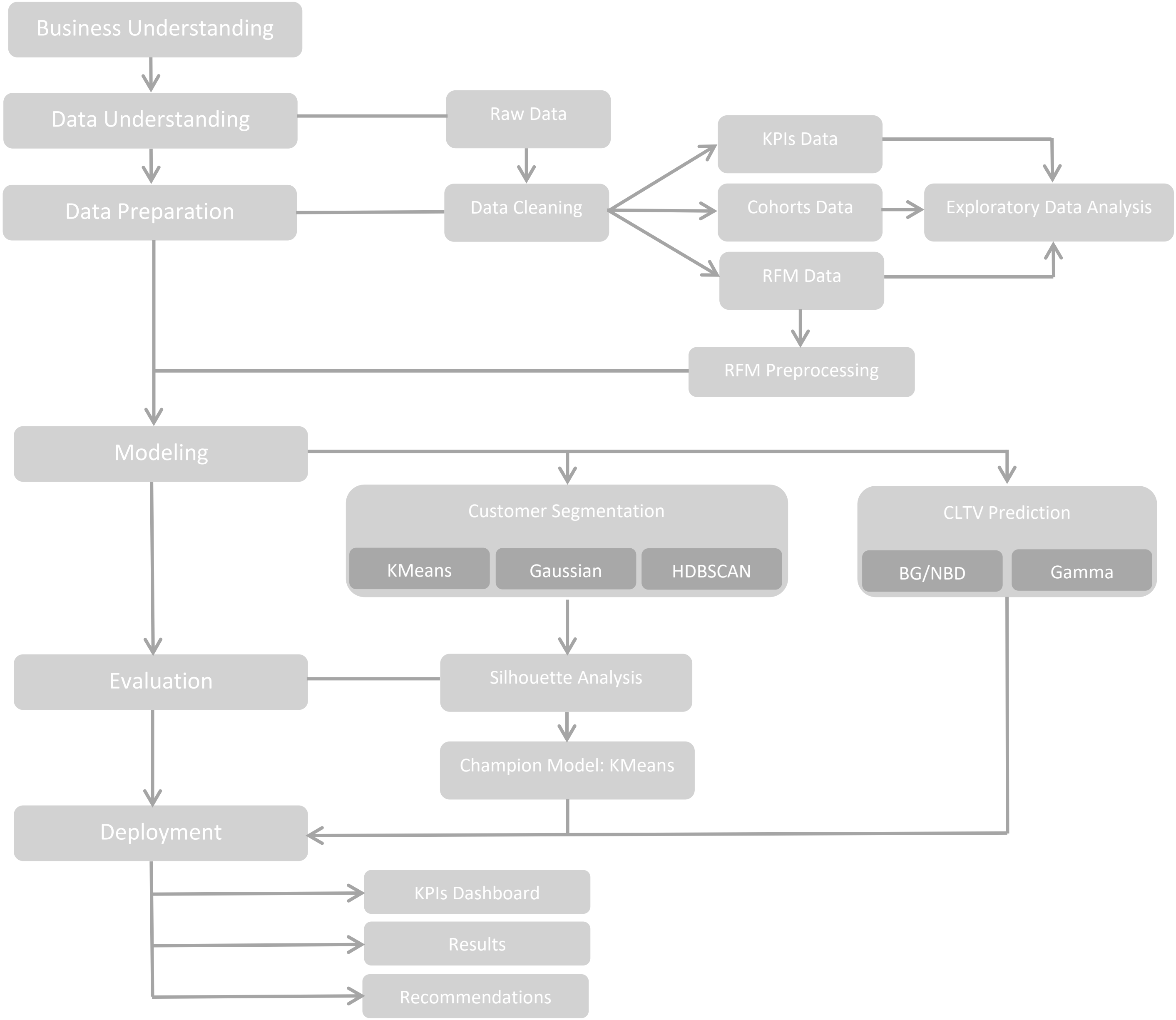


Figure 3.2: Applied CRISP-DM flowchart.

3.3 Business Understanding

Business understanding is the first phase in the CRISP-DM framework. In this phase, the business problems are approached to identify and solve them to enhance the business performance. This increases the awareness of the business problems and needs (Ziv, 2022). In XYZ's use case, there are 3 problems identified: (1) fluctuating number of purchase orders, (2) low retention rates, and (3) the lack of a business performance tracking tool. To solve these problems to meet the business objectives, specific analytics solutions and methodologies are determined and applied which will be discussed later in the following sections of the paper.

3.4 Data Understanding

Data understanding came right after the business understanding. At this phase, the real transactions data is collected from XYZ's databases using a structured query language (SQL). The retrieved data from the query above is XYZ's ordering transactional data (see Table 3.1) that contains all of the last 12 months’ completed transactions and consists of 1,276,436 rows and 5 columns. The SQL query is illustrated below:

```
SELECT  "Order Id", "Order DateTime", "Order Price", "User Id", "Quantity”

FROM "Orders"

WHERE "Order Status" = 'Completed'

AND

    "Order DataTime" >= NOW() - '12 Months'::INTERVAL
```

Column Name	Data Type	Sample Values	Description
Order Id	String	211216115328MQGI	The order’s unique Id
Order DateTime	Date/Time	2021-12-16 08:53:00	The date/time when the order made
Order Price	Float	48.00	The order’s monetary value
User Id	String	2745796d-645c	The user’s id who made the order
Quantity	Int	1	The order’s quantity

Table 3.1: Transactional data dictionary

3.5 Data Preparation

Next and in the data preparation phase, the quality of the data will then be improved. Any of the various pre-processing techniques can be used to accomplish this (Ziv, 2022). For the analysis, there are a number of specific data preparation tasks that are applied on the raw data to proceed with the analysis and provide quality results.

3.5.1 Data Cleaning

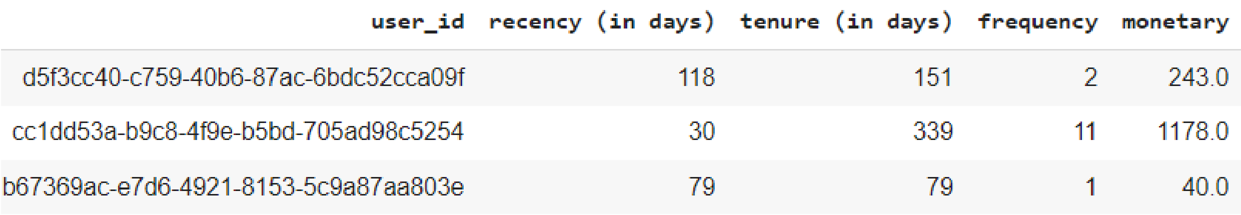
After collecting the raw data, the data is cleaned using some data cleaning techniques to meet quality data for further analysis. The data cleaning steps in the XYZ’s project includes:

1. Changing the “Order DateTime” column’s data type from string to date/time.
2. Replacing the “Order Price” column outliers by the column mean.
3. Sorting the data chronologically using the “Order DateTime”.
4. Checking the data is free from nulls and duplicates.
5. Renaming the columns to follow the same naming convention for better readability (i.e. to be in lowercase separated by '_’ instead of spaces)

3.5.2 Creating RFM

After the data cleaning, further data manipulation is performed on the data to prepare for the RFM customer segmentation analysis. The data is then aggregated with customers’ Ids as indices and for every customer 4 features are calculated: recency, tenure, frequency and monetary value. (see Table 3.2)

- Recency: indicates when a customer made his/her last order.
- Tenure: indicates the lifetime period of the customer with the analysis scope.
- Frequency: refers to the number of orders made in the scope of the analysis.
- Monetary value: refers to the amount of money spent on orders.



user_id	recency (in days)	tenure (in days)	frequency	monetary
d5f3cc40-c759-40b6-87ac-6bdc52cca09f	118	151	2	243.0
cc1dd53a-b9c8-4f9e-b5bd-705ad98c5254	30	339	11	1178.0
b67369ac-e7d6-4921-8153-5c9a87aa803e	79	79	1	40.0

Table 3.2: RFM sample data table

3.5.3 Creating Cohort Analysis Data

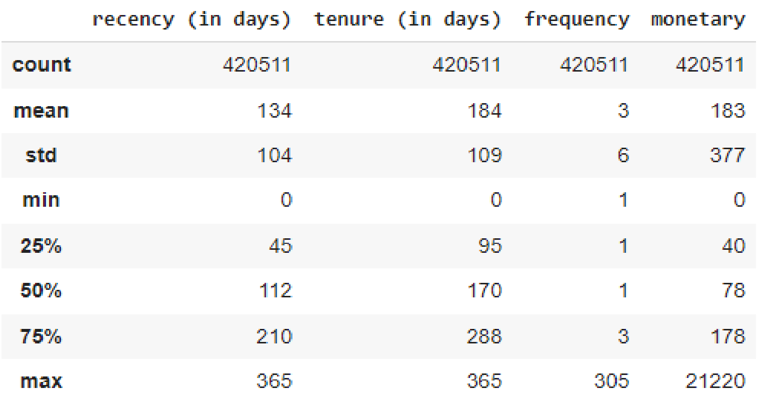
Then, further processing of the raw data is Cohort analysis data preparation. The time-based cohort analysis is a time-based behavioral customer segmentation technique that highlights common business problems such as active users, churn and retention rates. Simply the result data groups customers based on their first transaction month of date and then provides the trends of the specific metric of interest for every cohort across indexed months (i.e. 1st, 2nd months... etc.)

3.5.4 Calculating Monthly Key Performance Indicators

Also, the raw data is manipulated to create a report of XYZ's monthly KPIs which is used later for building a business intelligence dashboard. Reporting KPIs in a business user-friendly format helps executives and managers track business performance and make more data-driven accurate business decisions.

3.5.5 Exploratory Data Analysis (EDA)

Just before data preprocessing for machine learning, exploratory data analysis is performed to explore the prepared data for better understanding and engagement with the data before modeling. The RFM data is explored using various approaches: calculating descriptive summary statistics of RFM features, visualizing RFM features means and standard deviations, and visualizing RFM features distributions using a histogram grid and relations using scatter plots grid. The cohort analysis data is visualized using a colored heatmap for both active users' numbers and retention rates percentages. Finally, XYZ's key performance indicators are explored using hue-colored bar charts.



	recency (in days)	tenure (in days)	frequency	monetary
count	420511	420511	420511	420511
mean	134	184	3	183
std	104	109	6	377
min	0	0	1	0
25%	45	95	1	40
50%	112	170	1	78
75%	210	288	3	178
max	365	365	305	21220

Table 3.3: RFM summary statistics table

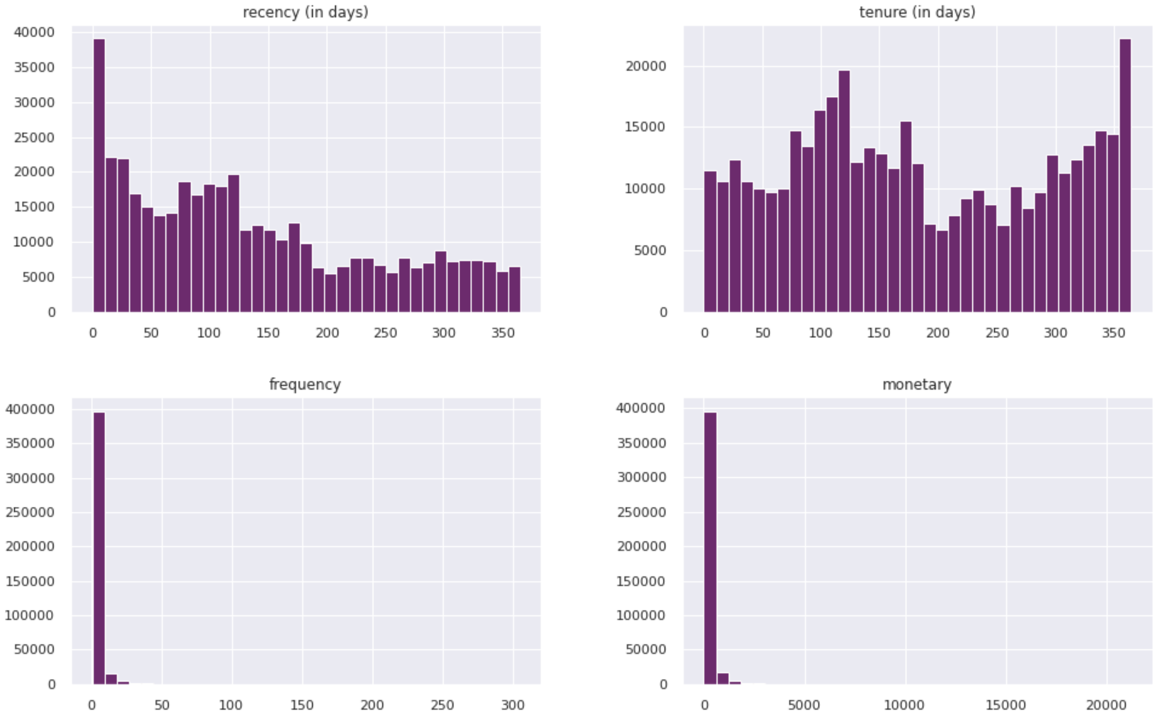


Figure 3.3: RFM distribution histograms

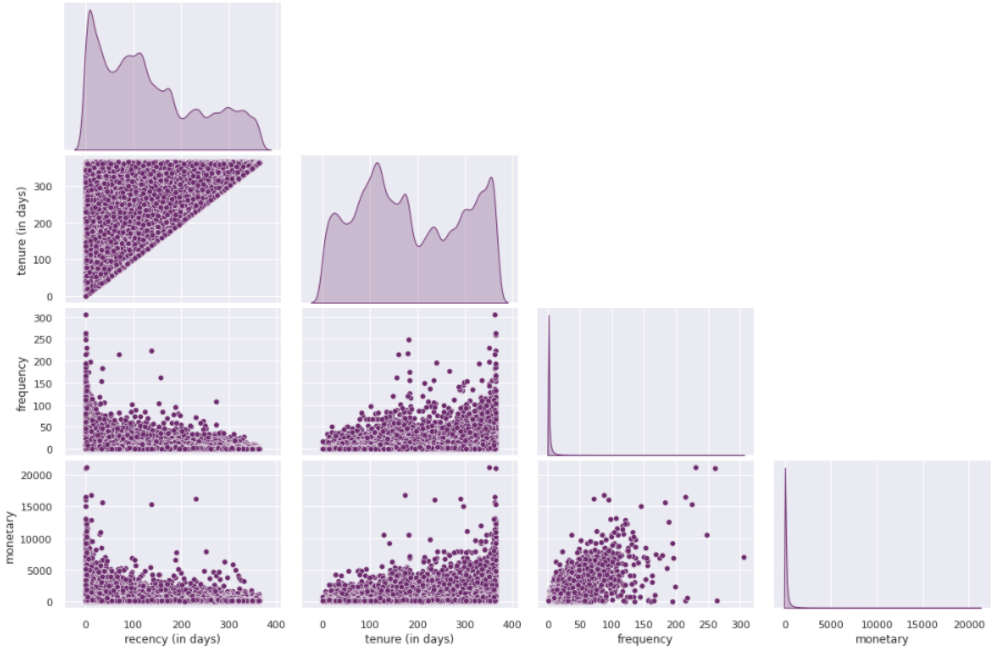


Figure 3.4: RFM pairwise relations

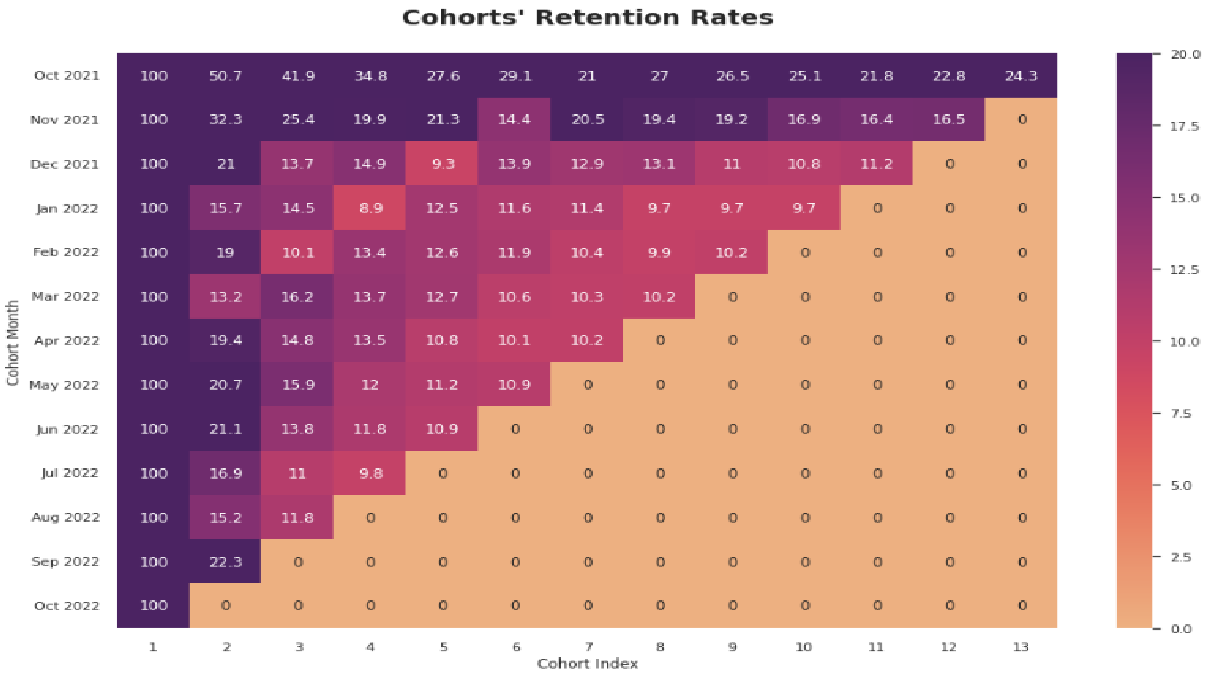


Figure 3.5: Cohorts retention rates heatmap

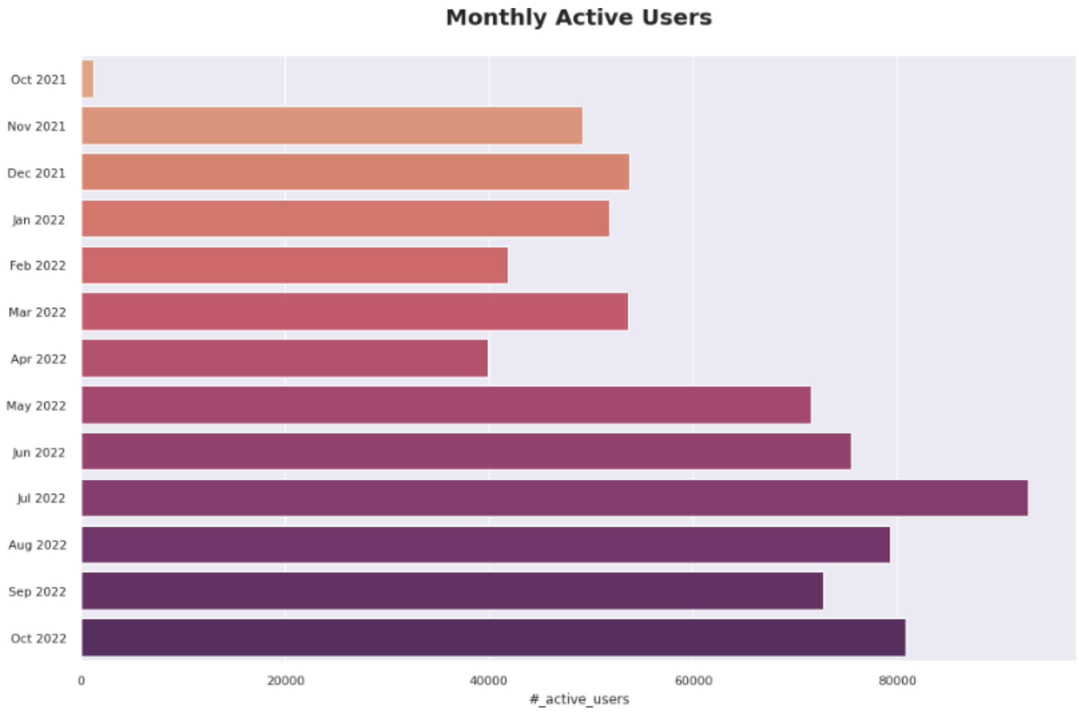


Figure 3.6: Monthly active users bar chart

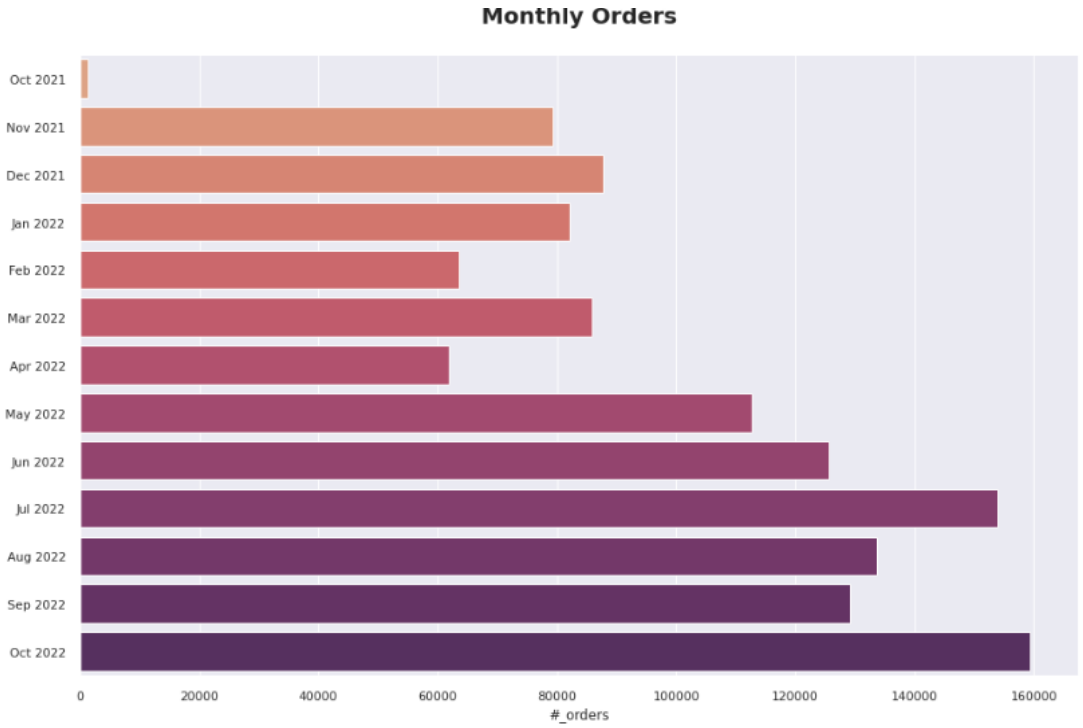


Figure 3.7: Monthly orders bar chart

3.5.6 RFM Data Preprocessing

As a final step and just after the ML modeling phase, the RFM data is preprocessed. To normalize the magnitude difference that occurred in the dataset, feature scaling is the used technique that involves scaling the numerical input variables into a standard range. This method shifts the distribution to have a mean of zero and a standard deviation of one by scaling each of the numerical input variables separately by subtracting the mean and dividing by the standard deviation (Brownlee, 2020). The magnitude difference issue in XYZ's RFM dataset for this analysis also existed. The scaling process is applied using the sklearn package in Python.

3.6 Customer Segmentation Using ML

Insights and trends can be extracted by studying consumer data using machine learning approaches. Models using machine learning are effective tools for decision-makers. They are able to accurately identify customer groups, which is far more difficult to achieve manually or using traditional analytical techniques. There are several machine learning algorithms, each of which is appropriate for a certain class of issues. The k-means clustering approach is one of the most popular machine-learning algorithms that are appropriate for customer segmentation issues. HDBSCAN, GaussianMixture, Agglomerative Clustering, balanced iterative reducing and clustering using hierarchies (BIRCH), and more clustering methods are available as well. Following are the applied clustering algorithms applied in this study.

3.6.1 KMeans Model

KMeans is an unsupervised ML clustering algorithm where the non-overlapping data points are assigned to each of the closest 'K' clusters by this approach from the dataset 'D'. In the K-means algorithm, the intra-cluster distance (distance from one cluster to another) is maximized while the inter-cluster (distance from one point to another within the same cluster) distance is minimized as much as possible. This approach is iterative, data points are relocated to various clusters based on the centroids computation till finding the optimal clustering of the data points which are the customers in this research based on various customer features (Anitha, 2022). In the case study, KMeans is applied and the number of 'K' clusters is determined using the "Elbow" method (see Figure 3.8) and validated using the Silhouette score which is 4 clusters.

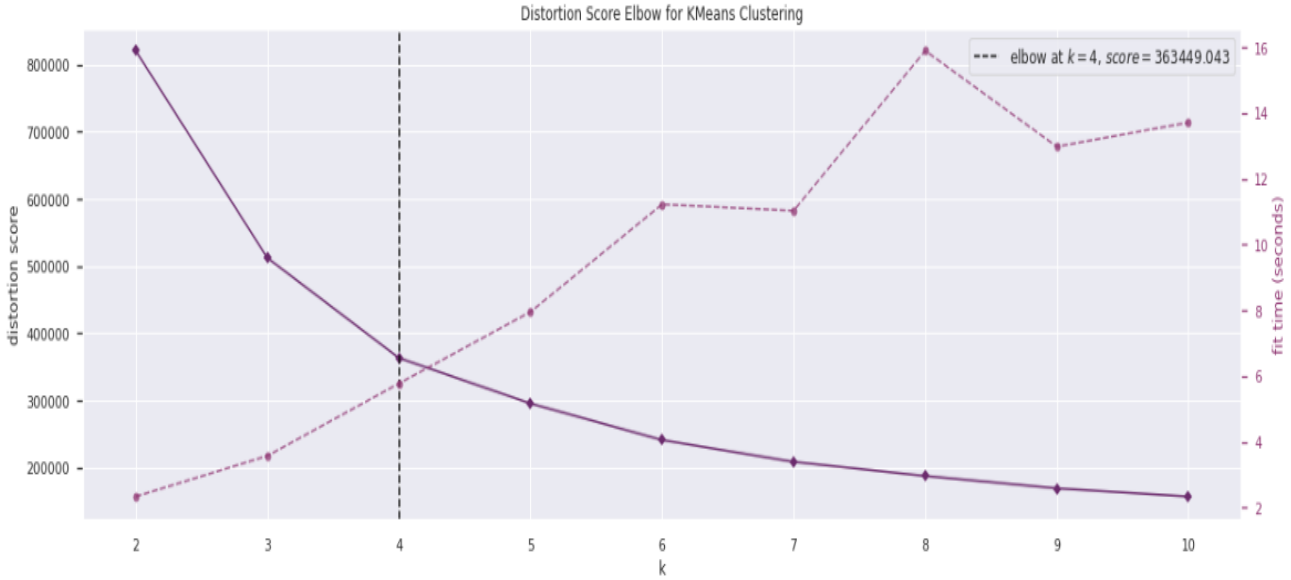


Figure 3.8: KMeans Elbow method

3.6.2 GaussianMixture Model

Gaussian Mixture Model is an unsupervised ML probabilistic model that uses the soft clustering approach for distributing the points in different clusters. The assumption behind the Gaussian Mixture Model is that there exists a certain number of Gaussian distributions and that each of these distributions represents a cluster. Because of this, a Gaussian Mixture Model tends to combine data points that correspond to the same distribution. Consider the scenario where we have three Gaussian distributions, denoted as GD1, GD2, and GD3. The mean (μ1, μ2, μ3) and variance (σ1, σ2, σ3) of these are each a specific value. The Gaussian Mixture Model would determine the likelihood of each data point falling into each of these distributions for a certain group of data points which are the customer who will be segmented according to their corresponding features. For multivariate density estimation and probabilistic clustering, Gaussian mixture models offer a robust approach (McCaw, 2022).

3.6.3 HDBSCAN Model

Finally, Hierarchical Density-Based Spatial Clustering of Applications with Noise is abbreviated as HDBSCAN. It is run with different epsilon values, and the results are integrated to discover the clustering that provides the highest stability. As a result, HDBSCAN is more resilient to parameter selection and may locate clusters with a range of densities (unlike DBSCAN). In practice, this implies that HDBSCAN produces a suitable clustering immediately and requires little to no parameter adjusting. The key parameter, minimum cluster size, is very simple to choose. HDBSCAN is a quick and reliable method that you can rely on to generate useful clusters which are features-based similar customer segments (McInnes, 2017).

3.6.4 Models Evaluation

Simply, the purpose of the evaluation phase in the CRISP-DM is to evaluate modeling results by evaluating the extent to which models satisfy the business objectives and, if time and budget allowed, testing the models on test applications. Some metrics are used during the evaluation phase to assess the performance of trained models. To further highlight the outcomes, metrics are also visualized. Additionally, a discussion and statistical analyses might support the findings (Schröer, 2021).

Silhouette is one of the most popular and effective internal measures for the evaluation of clustering validity. The distances between each data point and its closest adjacent cluster are examined using Silhouette (as the average distance of a data point to all other data points in its own cluster and that to all other data points in its closest neighbor cluster). In contrast to most other internal measures, Silhouette may be used to determine if a single cluster or even a single data point is well clustered as well as to assess the validity of a comprehensive clustering. Each data point is used to calculate the silhouette, and the silhouette value for a cluster or whole clustering is just the average of the point silhouettes. Simply, the Silhouette score can be defined as the degree of cohesion and separation (Wang, 2017). In XYZ's customer segmentation case study, it shows how well each customer is segmented, the closer to 1 the better the model performance is.

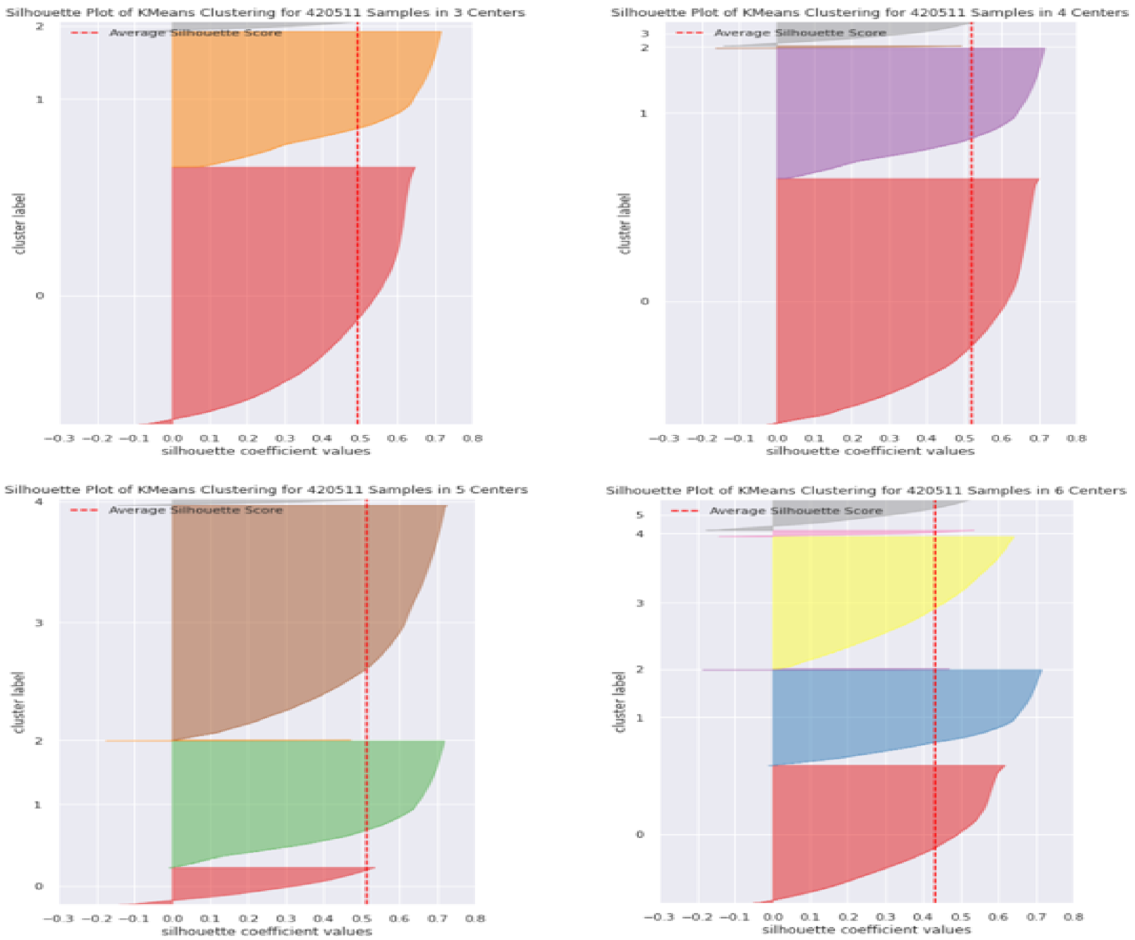


Figure 3.9: KMeans Silhouette score visualizations

3.6.5 Champion Model

After comparing the model performance, mainly based on the Silhouette score, and other computational efficiency criteria like model fit time and memory usage (see Table 3.4), it is clear that KMeans with 4 clusters is the best clustering model to XYZ’s customer segmentation case study (see Figure 3.4), in other words, it has the highest cohesion and separation segmentation performance with Silhouette score = 0.521.

Model	Silhouette Score	Memory Usage	Time Efficiency
KMeans	**0.521**	Low	High
Gaussian Mixture	0.099	High	Low
HDBSCAN	0.251	Medium	Medium

Table 3.4: Clustering models performance comparison

3.7 Customer Lifetime Value Prediction

Regarding the customer lifetime value prediction modeling, there are 2 models have been applied: The Beta Geometric Negative Binomial Distribution, abbreviated as BG-NBD, to predict the number of future orders that will be made by a customer and the Gamma-Gamma model to predict customers' future monetary values. The BG/NBD model is developed from the first principles and presents the expressions required for making individual-level statements about future buying behavior (Fader, 2005).

3.8 Specific Tools

Business intelligence's goal is to transform raw data into business insights so that managers and executives make data-driven accurate decisions. Business intelligence tools and software are used by companies to produce decision-support solutions for enhancing business management (Cruz, 2023). Tableau is used in the case study to build a key performance indicators dashboard that is illustrated in Figure 3.10.

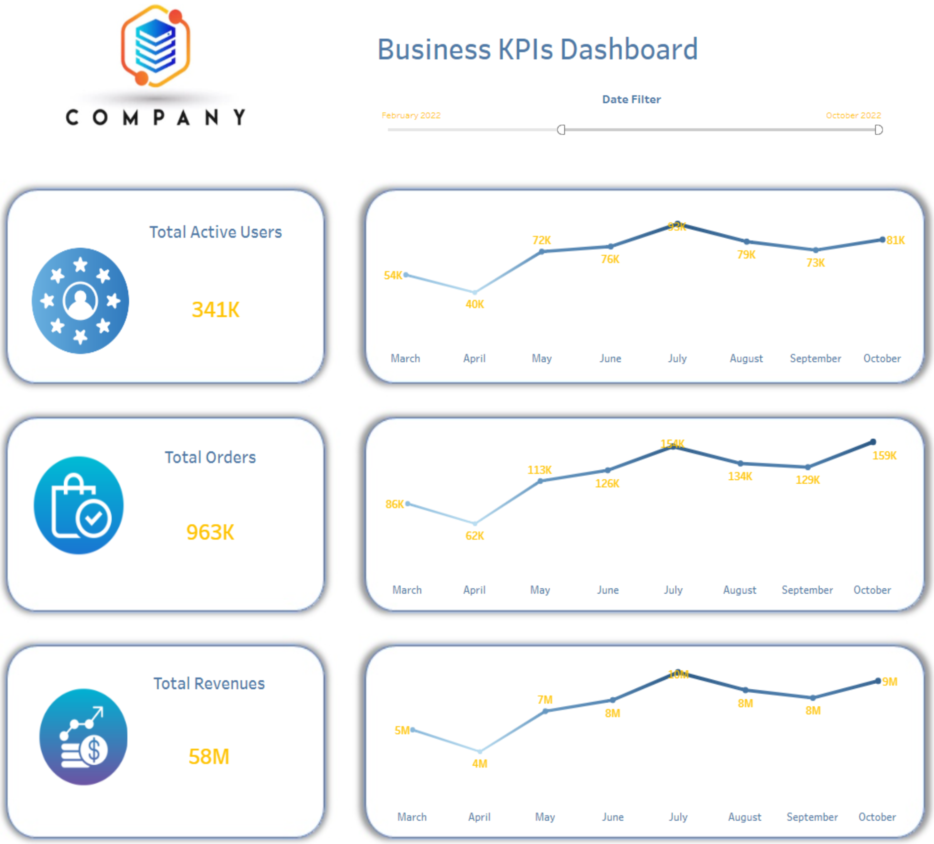


Figure 3.10: KPIs dashboard

CHAPTER 4 RESULTS & DISCUSSION

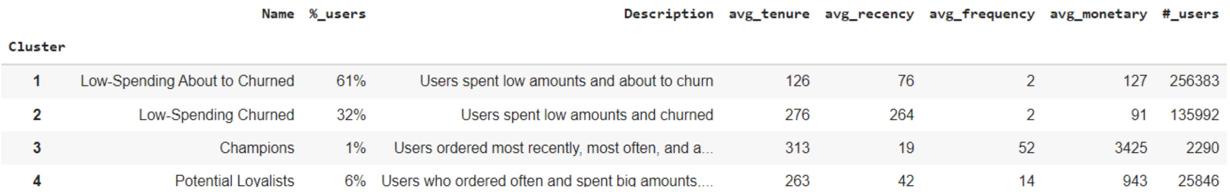
4.1 RFM Customer Segmentation Profiling

The last and most important step in any customer segmentation is customer profiling. Customer profiling involves using the same variables that are used in the clustering models corresponding to every customer group, in XYZ's case: recency, frequency and monetary value. This may be accomplished by locating the centroids of each cluster that the model produced, researchers tried to define the centroid's metrics in terminologies that businesses would understand (Harish, 2023).

4.1.1 Insights of RFM Customer Segmentation Profiling

According to RFM customer segmentation results or what is called customer profiles. It is found that at XYZ, there are 4 main customer groups: Low-Spending about to Churn, Low-Spending Churned, Champions and Potential Loyalists as illustrated in table 5. Low-Spending about to Churn group represents 61% of the customers and simply they can be described as the customers who spent low amounts of money and are about to churn from XYZ. Low-Spending Churned customers represent 32% of the customers and simply they can be described as the customers who spent low amounts of money and churned from XYZ. Potential Loyalists are 6% of customers who ordered often with due amounts of money but have not ordered recently. Finally, Champions who are 1% of the customers, are the best customers at XYZ who ordered most recently, most often and are heavy spenders (see Table 4.1, Figure 4.1 and Figure 4.2).

Also, there are features that are more distinctive in some customer profiles than other profiles. Frequency and monetary value highly contribute to defining the Champions group. Recency also distinguishes Low-Spending Churned customers. Eventually, the number of users is distinctive for Low-Spending about to Churn customers (see Figure 4.3).



Cluster	Name	%_users	Description	avg_tenure	avg_recency	avg_frequency	avg_monetary	#_users
1	Low-Spending About to Churned	61%	Users spent low amounts and about to churn	126	76	2	127	256383
2	Low-Spending Churned	32%	Users spent low amounts and churned	276	264	2	91	135992
3	Champions	1%	Users ordered most recently, most often, and a...	313	19	52	3425	2290
4	Potential Loyalists	6%	Users who ordered often and spent big amounts....	263	42	14	943	25846

Table 4.1: Customer Segmentation Profiles Summary

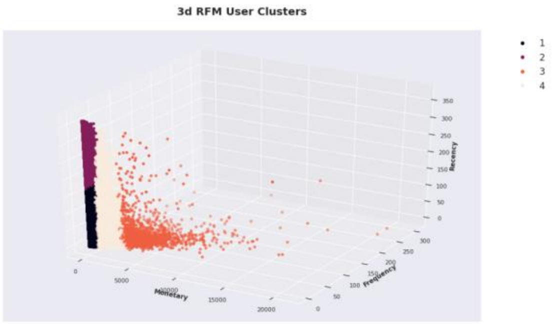


Figure 4.1: 3D RFM customer profiles

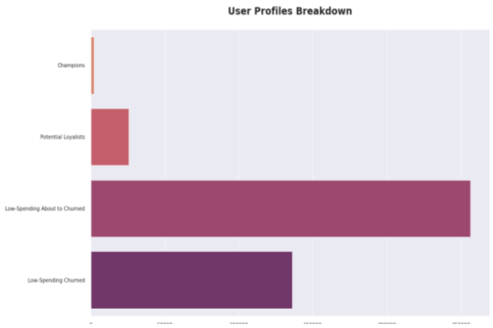


Figure 4.2: RFM customer profiles breakdown

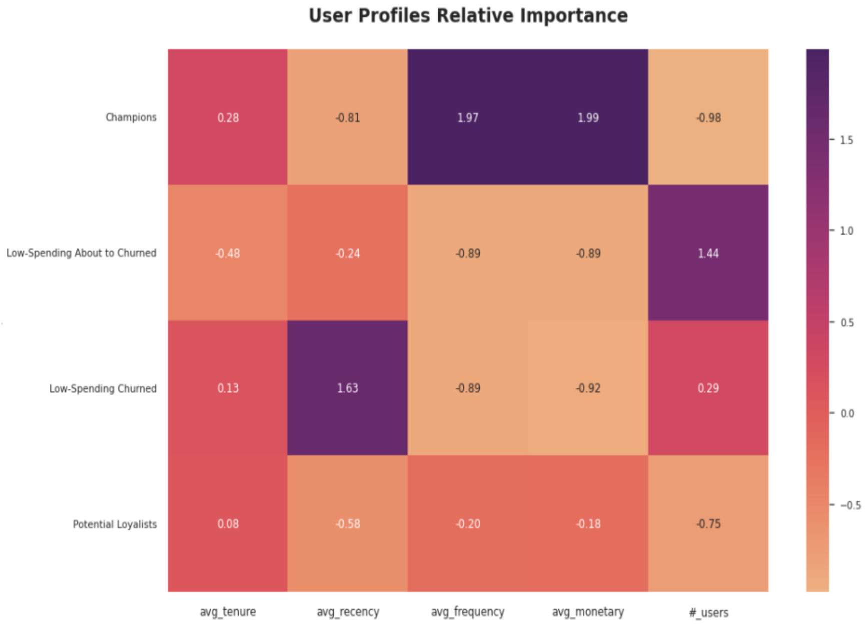


Figure 4.3: Customer profiles RFM relative importance

4.1.2 Business Recommendations Based on RFM Customer Segmentation Profiling

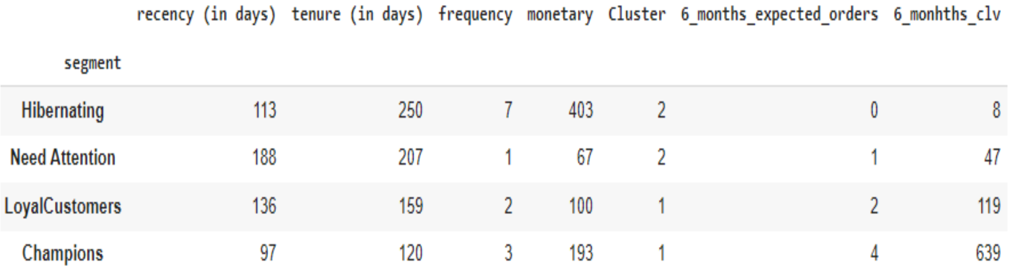
After RFM-based customer profiling and getting insights out of the segmentation results, business recommendations should be articulated to enhance business performance. Recommendations for XYZ’s marketing management can be made to strategize and customize marketing activities to fit the different customer profiles. For champions, they may be rewarded to stay loyal to our company. Communication with this group should make them feel valued and appreciated Also, they can become early adopters of new products and will help promote the brand. And the Potential Loyalists may be offered membership, loyalty programs or recommend related products to upsell them and help to make them Champions. Finally, for Low-Spending Churned & About to Churned Customers, retention & reactivation campaigns may be launched to make them order again, but it's also important to take their CLTVs into consideration to be still profitable customers.

4.2 Customer Lifetime Value Prediction

After using both the BG/NBD and the Gamma-Gamma model, the number of orders that are more likely to be made by the customers in addition to their corresponding customer lifetime value is predicted for a period of 6 months. Then, the customers are classified into segments based on customer lifetime value. This enables the marketers at XYZ to plan the retention and reactivation campaigns more feasibly, in other words, to spend for every customer what still makes him/her profitable for XYZ.

4.2.1 Insights of Customer Lifetime Value Segments

After the customers CLTVs are predicted and segmented for a period of 6 months into 4 main segments based on their CLTVs: Hibernating, Need Attention, Loyal Customers and Champions. It is found that Hibernating, Need Attention, Loyal Customers and Champions are customers whose 6-month future values are approximately 8 EGPs, 47 EGPs, 119 EGPs and 639 EGPs respectively (see Table 4.2 and Figure 4.4).



	recency (in days)	tenure (in days)	frequency	monetary	Cluster	6_months_expected_orders	6_monhths_clv
segment							
Hibernating	113	250	7	403	2	0	8
Need Attention	188	207	1	67	2	1	47
LoyalCustomers	136	159	2	100	1	2	119
Champions	97	120	3	193	1	4	639

Table 4.2: CLTV customer segments summary

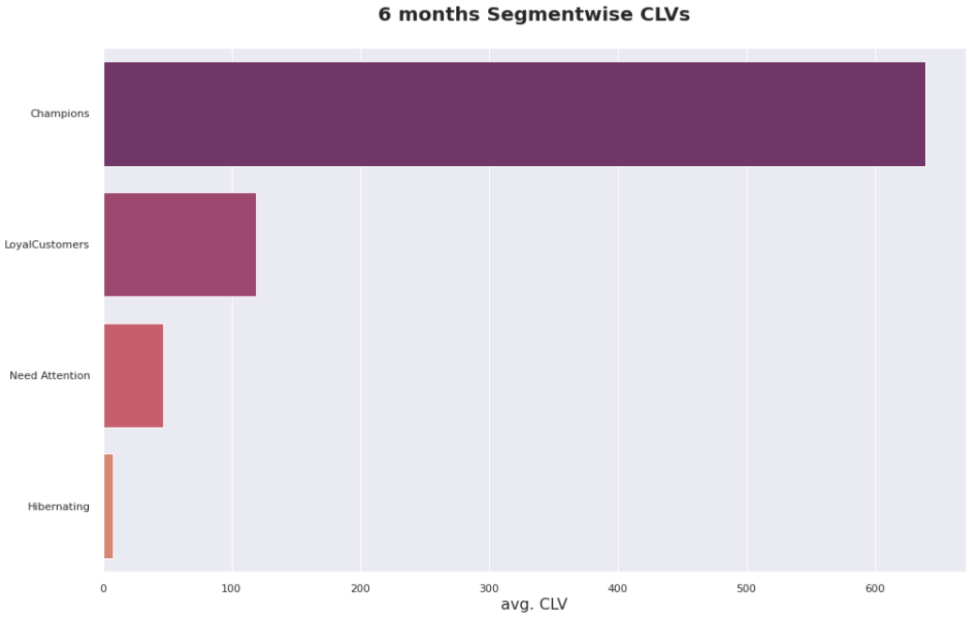


Figure 4.4: Customer segments average CLTV

4.2.2 Business Recommendations Based on Customer Lifetime Value Segments

Referencing the retention rates heatmap in figure 10, it's obvious that after the Nov 2021's cohort, the company's retention rate is decreasing drastically, in other words, a lot of users are churned. So, the marketing team needs to launch activation & retention campaigns to improve the company’s retention rates. But to still be operating efficiently, CLTVs should be used in the campaigns' budgeting and planning. The latter is to determine to which extent the company may spend to either retain or reactivate a customer and still be profitable for the company. Also, customer lifetime value can be further improved for the sake of XYZ using different marketing activities. This includes: automating customized marketing actions using softwares based on customer behaviors, increasing customer satisfaction to prolong their customer lifespan and encouraging additional purchases without increasing the marketing budget, which can be done by selling complementary services.

Both customer lifetime value prediction and RFM customer segmentation profiles can be cross-applied for every XYZ's marketing campaign: the first, customer lifetime values, for the feasibility, while the latter, RFM customer profiles for the targeting and effectiveness of the campaigns, such as retention, reactivation and promotion campaigns.

CHAPTER 5 CONCLUSION

5.1 Summary

To wrap up, RFM customer segmentation using ML clustering algorithms is used in this study to group XYZ's customers according to their ordering behaviors and build customer profiles. Also, BG/NBD & Gamma-Gamma models are used to predict XYZ's customers' 6-month future worth and their expected number of orders.

This study uses XYZ's real ordering transactions data for 12 months. The dataset includes all transactions from 30th September to 31st October 2022. It consists of 1,276,436 rows and 5 columns: (1) Order Id, (2) Order DateTime, (3) Order Price, (4) User Id and (5) Order Quantity.

The technical work in the study is held mainly using Python. It started by collecting the raw data. Then, the data went through specific manipulations for further analysis purposes: (1) data cleaning, (2) extracting the RFM data, (3) creating cohort analysis data, (4) calculating monthly KPIs data and (5) preprocessing the RFM data for clustering ML models. Afterward, the manipulated data are explored both numerically and visually, this exploratory data analysis included computing descriptive summary statistics, numerical variables distributions histograms, relations' scatterplots, heatmaps and bar charts. Finally, three ML clustering models were built for RFM customer segmentation: (1) KMeans, (2) GaussianMixture and (3) HDBSCAN. Then, the models were evaluated based on the Silhouette score to end up with KMeans with 4 clusters as the champion model for RFM segmentation. While BG/NBD & Gamma-Gamma were used to predict future 6-month customers' worth and the number of orders. Based on the models, specific business recommendations were articulated to solve XYZ's problems.

Regarding research question (1), " Which is the best ML clustering algorithm for XYZ's RFM customer segmentation?", it is concluded that KMeans with 4 clusters is the best clustering model to fit the data in this study. This is because it shows the highest Silhouette score of 0.521 among the clustering models in this study. Therefore, KMeans is used to group the customers into 4 groups/profiles. Therefore, the marketing team can customize their campaigns and targeting strategies relevant to each customer group to increase the number of monthly orders.

Regarding research question (2), " What are the customers' future worth/value based on their ordering behaviors?", the BG/NBD and Gamma-Gamma models are used to predict 6-month customers' number of orders and their monetary values. As a result, the marketing team can use this information to plan for more feasible reactivation and retention marketing campaigns by determining how much the amount to spend for every customer and still be profitable for the company. Therefore, the churned customers from XYZ will be compensated in an efficient way.

Regarding research question (3), " How to track marketing campaigns' impact on XYZ's business performance?", it is suggested to build a business intelligence dashboard. This is to keep track of XYZ's key performance indicators to assess XYZ's performance and the impact of marketing campaigns on it. Therefore, a business KPIs dashboard is built to track the main KPIs: the number of active users, the number of orders made and the revenues generated.

5.2 Limitations

There is a limitation in this study: when implementing the BG/NBD & Gamma-Gamma models, they raise errors when running the code for customers with either recency, frequency or a monetary value less than 1. Therefore, the customers' RFM data were filtered to include only customers with more than 1 for recency, frequency and monetary values. There are 16,591 customers who were filtered, approximately 4% of the customers.

5.3 Potential Future Work

In this study, a transaction data extract from XYZ databases is used to build and apply the analytical solutions discussed. For future work, it is highly recommended, instead of using data extracts, the models may be deployed to a live environment using application programming interfaces (APIs) for real-time continuous tracking of customers ordering behaviors, customer segmentation and future monetary values.

For more enhanced customer segmentation, other customer features such as tenure (the length of a customer's relationship with the company), gender and other demographics may be added to the clustering models and to reduce the variability in the data due to the added features the principal component analysis (PCA) may be used in the data preprocessing phase.

REFERENCES

Acar, S., Köroğlu, F., Duyuler, B., Kaya, T., & Özcan, T. (2021, August). Customer segmentation using RFM model and clustering methods in online retail industry. In *International Conference on Intelligent and Fuzzy Systems* (pp. 69-77). Springer, Cham.

Alqhatani, A., Ashraf, M. S., Ferzund, J., Shaf, A., Abosaq, H. A., Rahman, S., ... & Alqhtani, S. M. (2022). 360° retail business analytics by adopting hybrid machine learning and a business intelligence approach. *Sustainability*, *14*(19), 11942.

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences, 34*(5), 1785-1792. https://doi.org/10.1016/j.jksuci.2019.12.011 Cc

Brandizzi, N., Russo, S., Galati, G., & Napoli, C. (2022). Addressing vehicle sharing through behavioral analysis: A solution to user clustering using recency-frequency-monetary and vehicle relocation based on neighborhood splits. *Information*, *13*(11), 511.

Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. *Machine Learning Mastery*.

Channa, H. S. (2019). Customer lifetime value: An ensemble model approach. In *Data Management, Analytics and Innovation* (pp. 353-363). Springer, Singapore.

Chaudhary, P., Kalra, V., & Sharma, S. (2022). A hybrid machine learning approach for customer segmentation using RFM analysis. In *International Conference on Artificial Intelligence and Sustainable Engineering* (pp. 87-100). Springer, Singapore.

Cruz, J. Q., & Tapia, F. (2023). Data mining prospective associated with the purchase of life insurance through predictive models. In *International Conference on Software Process Improvement* (pp. 165-179). Springer, Cham.

Das, P., & Singh, V. (2023). Knowing your customers using customer segmentation. In *Computational Methods and Data Engineering* (pp. 437-451). Springer, Singapore.

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

Fu, X., Chen, X., Shi, Y. T., Bose, I., & Cai, S. (2017). User segmentation for retention management in online social games. *Decision Support Systems*, 101, 51-68.

Harish, A. S., & Malathy, C. (2023). Customer segment prediction on retail transactional data using K-Means and Markov model. *Intelligent Automation and Soft Computing, 36*(1), 589-600

Kim, S., DeSarbo, W. S., & Chang, W. (2021). Note: A new approach to the modeling of spatially dependent and heterogeneous geographical regions. *International Journal of Research in Marketing*, *38*(3), 792-803.

McCaw, Z. R., Aschard, H., & Julienne, H. (2022). Fitting Gaussian mixture models on incomplete data. *BMC bioinformatics*, *23*(1), 1-20.

McInnes, L., & Healy, J. (2017, November). Accelerated hierarchical density-based clustering. In *2017 IEEE International conference on Data Mining Workshops (ICDMW)* (pp. 33-42). IEEE.

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science,* 6, e267.

Retail. (2022, October 28). In Wikipedia. https://en.wikipedia.org/wiki/Retail

Ramírez-Asís, H., Vílchez-Vásquez, R., Huamán-Osorio, A., Gonzales-Yanac, T., & Castillo-Picón, J. (2023). Digitalization and success of Peruvian micro-enterprises in the retail 4.0 sector. In *The Implementation of Smart Technologies for Business Success and Sustainability* (pp. 225-236). Springer, Cham.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526-534.

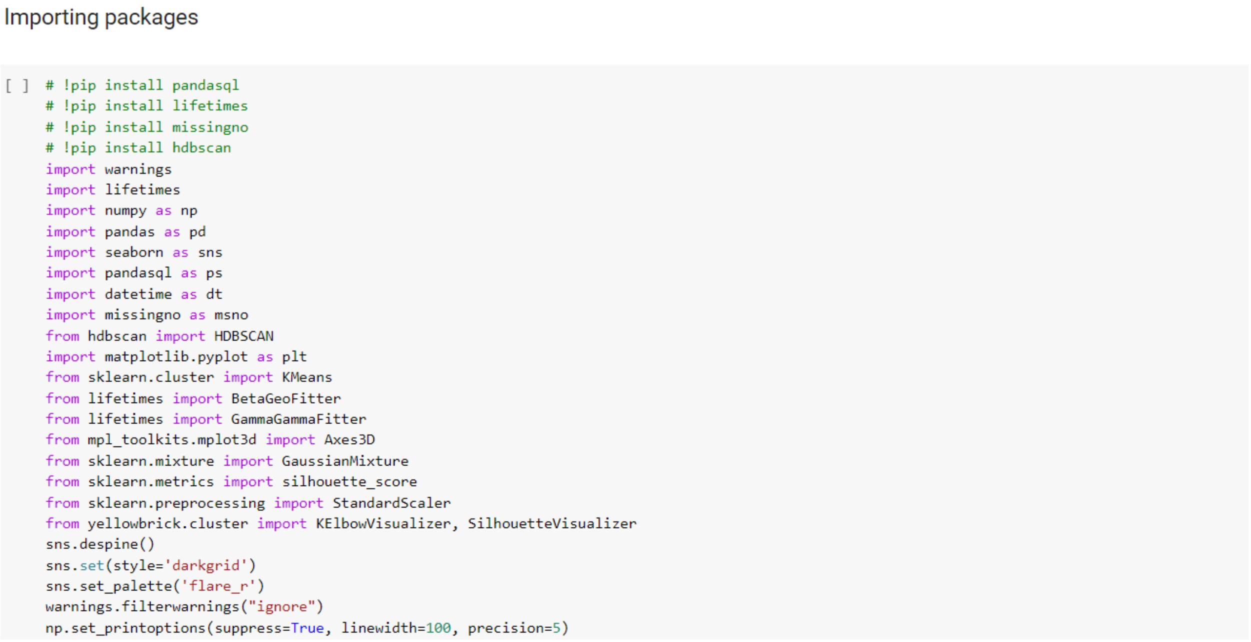
Thoyyibah. T. (2022). CRISP-DM method for mood classification in Indonesian music 70 and 80 era. *International Journal of Applied Engineering & Technology, 4*(1), 51–55. https://doi.org/10.5281/zenodo.7266878

Wang, F., Franco-Penya, H. H., Kelleher, J. D., Pugh, J., & Ross, R. (2017, July). An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 291-305). Springer, Cham.

Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent & Fuzzy Systems, 38*(5), 6159-6173.

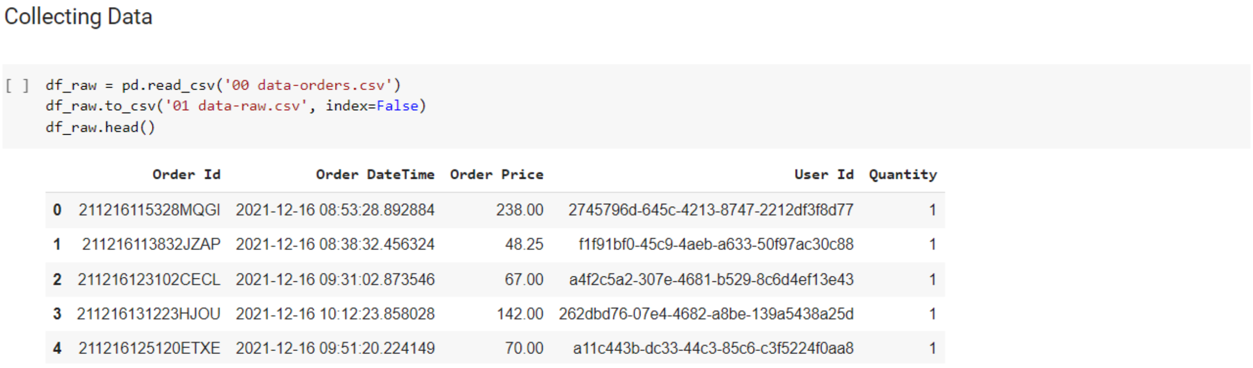
Ziv, B., & Parmet, Y. (2022). Improving nonconformity responsibility decisions: A semi-automated model based on CRISP-DM. *International Journal of System Assurance Engineering and Management*, *13*(2), 657-667.

APPENDIX: PYTHON PROGRAMMING CODE



Importing packages

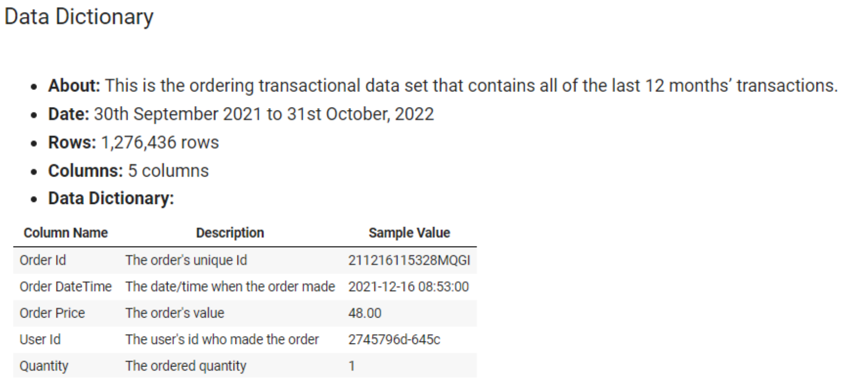
```
[ ]  # !pip install pandasql
     # !pip install lifetimes
     # !pip install missingno
     # !pip install hdbscan
     import warnings
     import lifetimes
     import numpy as np
     import pandas as pd
     import seaborn as sns
     import pandasql as ps
     import datetime as dt
     import missingno as msno
     from hdbscan import HDBSCAN
     import matplotlib.pyplot as plt
     from sklearn.cluster import KMeans
     from lifetimes import BetaGeoFitter
     from lifetimes import GammaGammaFitter
     from mpl_toolkits.mplot3d import Axes3D
     from sklearn.mixture import GaussianMixture
     from sklearn.metrics import silhouette_score
     from sklearn.preprocessing import StandardScaler
     from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer
     sns.despine()
     sns.set(style='darkgrid')
     sns.set_palette('flare_r')
     warnings.filterwarnings("ignore")
     np.set_printoptions(suppress=True, linewidth=100, precision=5)
```



Collecting Data

```
[ ]  df_raw = pd.read_csv('00 data-orders.csv')
     df_raw.to_csv('01 data-raw.csv', index=False)
     df_raw.head()
```

```
           Order Id             Order DateTime  Order Price                               User Id  Quantity
0  211216115328MQGI  2021-12-16 08:53:28.892884       238.00  2745796d-645c-4213-8747-2212df3f8d77         1
1  211216113832JZAP  2021-12-16 08:38:32.456324        48.25  f1f91bf0-45c9-4aeb-a633-50f97ac30c88         1
2  211216123102CECL  2021-12-16 09:31:02.873546        67.00  a4f2c5a2-307e-4681-b529-8c6d4ef13e43         1
3  211216131223HJOU  2021-12-16 10:12:23.858028       142.00  262dbd76-07e4-4682-a8be-139a5438a25d         1
4  211216125120ETXE  2021-12-16 09:51:20.224149        70.00  a11c443b-dc33-44c3-85c6-c3f5224f0aa8         1
```

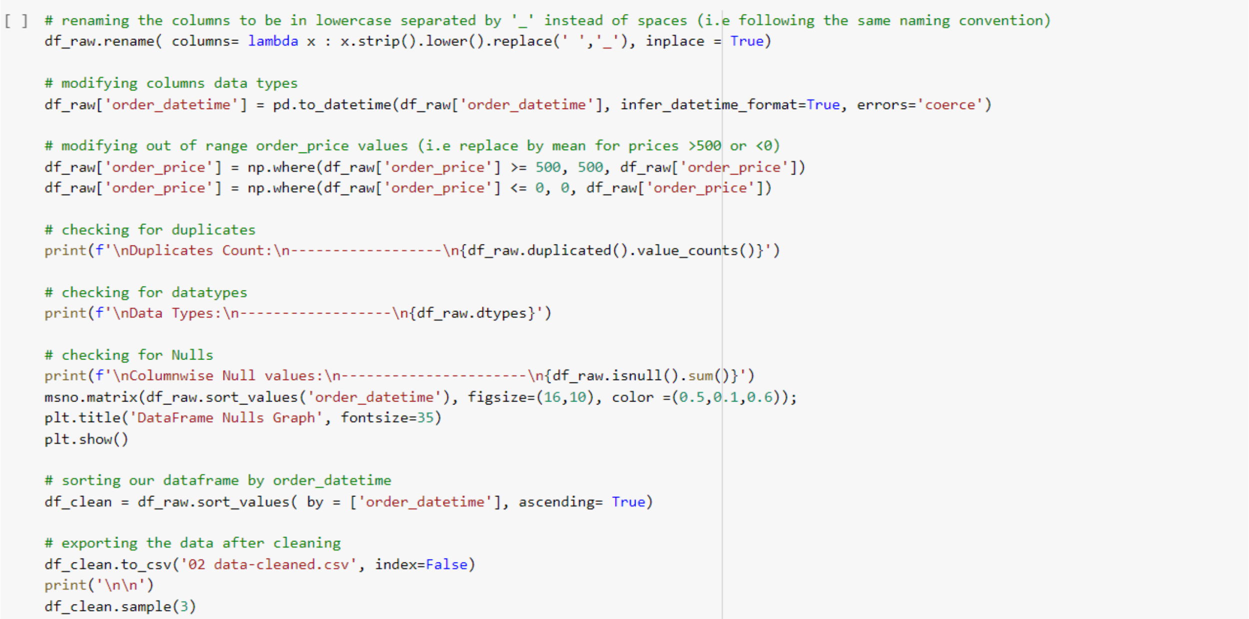


Data Dictionary

- **About:** This is the ordering transactional data set that contains all of the last 12 months' transactions.
- **Date:** 30th September 2021 to 31st October, 2022
- **Rows:** 1,276,436 rows
- **Columns:** 5 columns
- **Data Dictionary:**

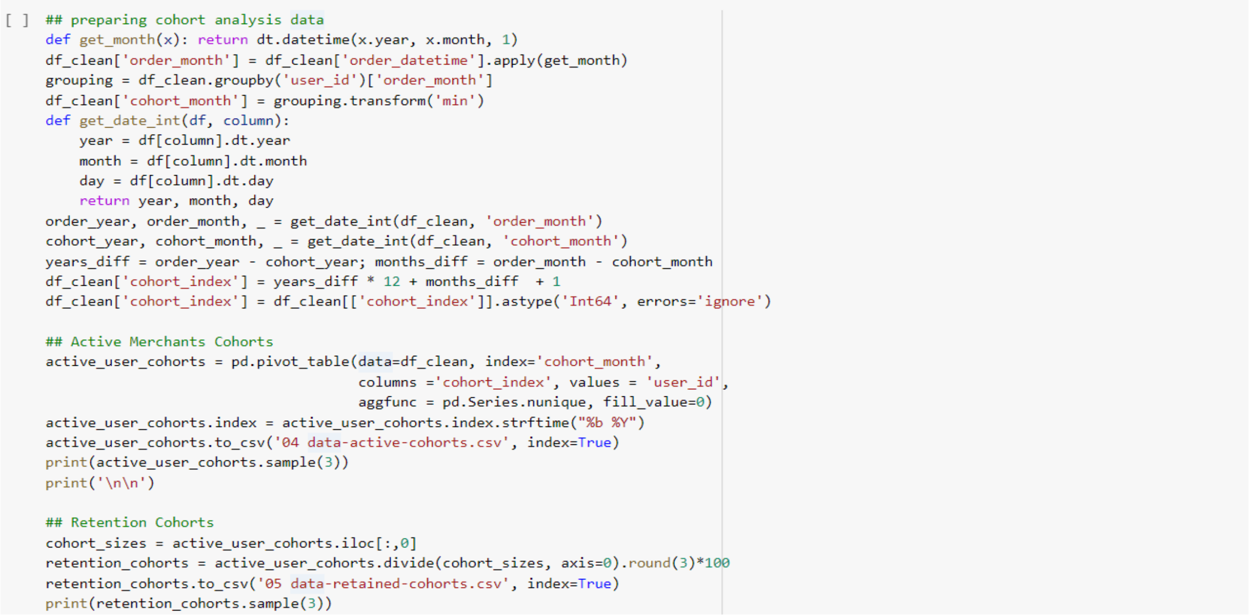
Column Name	Description	Sample Value
Order Id	The order's unique Id	211216115328MQGI
Order DateTime	The date/time when the order made	2021-12-16 08:53:00
Order Price	The order's value	48.00
User Id	The user's id who made the order	2745796d-645c
Quantity	The ordered quantity	1

Data Cleaning



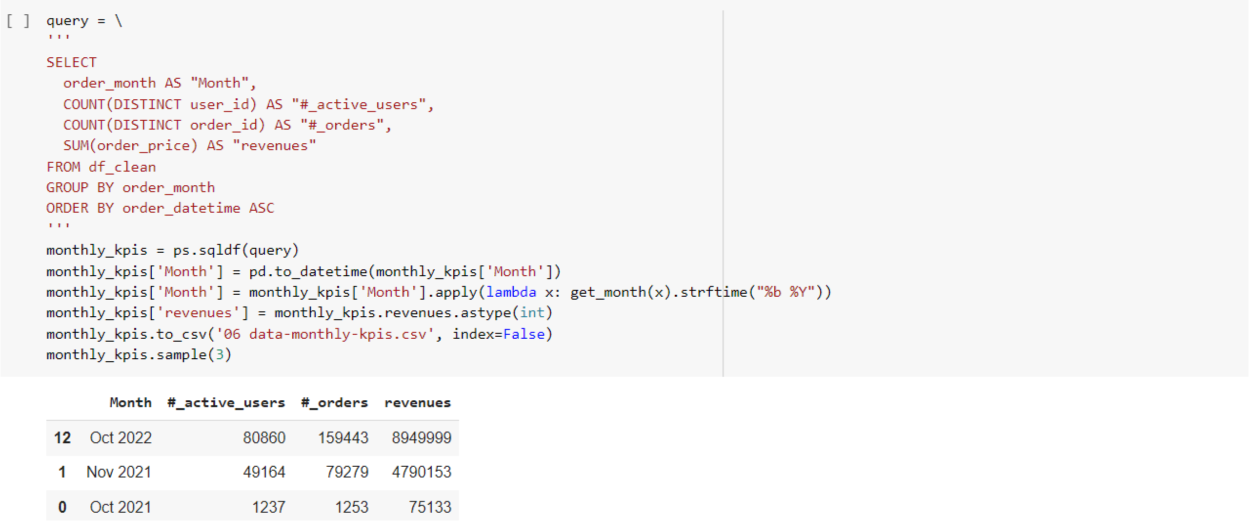
```
[ ] # renaming the columns to be in lowercase separated by '_' instead of spaces (i.e following the same naming convention)
    df_raw.rename( columns= lambda x : x.strip().lower().replace(' ','_'), inplace = True)

    # modifying columns data types
    df_raw['order_datetime'] = pd.to_datetime(df_raw['order_datetime'], infer_datetime_format=True, errors='coerce')

    # modifying out of range order_price values (i.e replace by mean for prices >500 or <0)
    df_raw['order_price'] = np.where(df_raw['order_price'] >= 500, 500, df_raw['order_price'])
    df_raw['order_price'] = np.where(df_raw['order_price'] <= 0, 0, df_raw['order_price'])

    # checking for duplicates
    print(f'\nDuplicates Count:\n-----------------\n{df_raw.duplicated().value_counts()}')

    # checking for datatypes
    print(f'\nData Types:\n-----------------\n{df_raw.dtypes}')

    # checking for Nulls
    print(f'\nColumnwise Null values:\n---------------------\n{df_raw.isnull().sum()}')
    msno.matrix(df_raw.sort_values('order_datetime'), figsize=(16,10), color =(0.5,0.1,0.6));
    plt.title('DataFrame Nulls Graph', fontsize=35)
    plt.show()

    # sorting our dataframe by order_datetime
    df_clean = df_raw.sort_values( by = ['order_datetime'], ascending= True)

    # exporting the data after cleaning
    df_clean.to_csv('02 data-cleaned.csv', index=False)
    print('\n\n')
    df_clean.sample(3)
```

Cohort Data



```
[ ] ## preparing cohort analysis data
    def get_month(x): return dt.datetime(x.year, x.month, 1)
    df_clean['order_month'] = df_clean['order_datetime'].apply(get_month)
    grouping = df_clean.groupby('user_id')['order_month']
    df_clean['cohort_month'] = grouping.transform('min')
    def get_date_int(df, column):
        year = df[column].dt.year
        month = df[column].dt.month
        day = df[column].dt.day
        return year, month, day
    order_year, order_month, _ = get_date_int(df_clean, 'order_month')
    cohort_year, cohort_month, _ = get_date_int(df_clean, 'cohort_month')
    years_diff = order_year - cohort_year; months_diff = order_month - cohort_month
    df_clean['cohort_index'] = years_diff * 12 + months_diff  + 1
    df_clean['cohort_index'] = df_clean[['cohort_index']].astype('Int64', errors='ignore')

    ## Active Merchants Cohorts
    active_user_cohorts = pd.pivot_table(data=df_clean, index='cohort_month',
                                         columns ='cohort_index', values = 'user_id',
                                         aggfunc = pd.Series.nunique, fill_value=0)
    active_user_cohorts.index = active_user_cohorts.index.strftime("%b %Y")
    active_user_cohorts.to_csv('04 data-active-cohorts.csv', index=True)
    print(active_user_cohorts.sample(3))
    print('\n\n')

    ## Retention Cohorts
    cohort_sizes = active_user_cohorts.iloc[:,0]
    retention_cohorts = active_user_cohorts.divide(cohort_sizes, axis=0).round(3)*100
    retention_cohorts.to_csv('05 data-retained-cohorts.csv', index=True)
    print(retention_cohorts.sample(3))
```

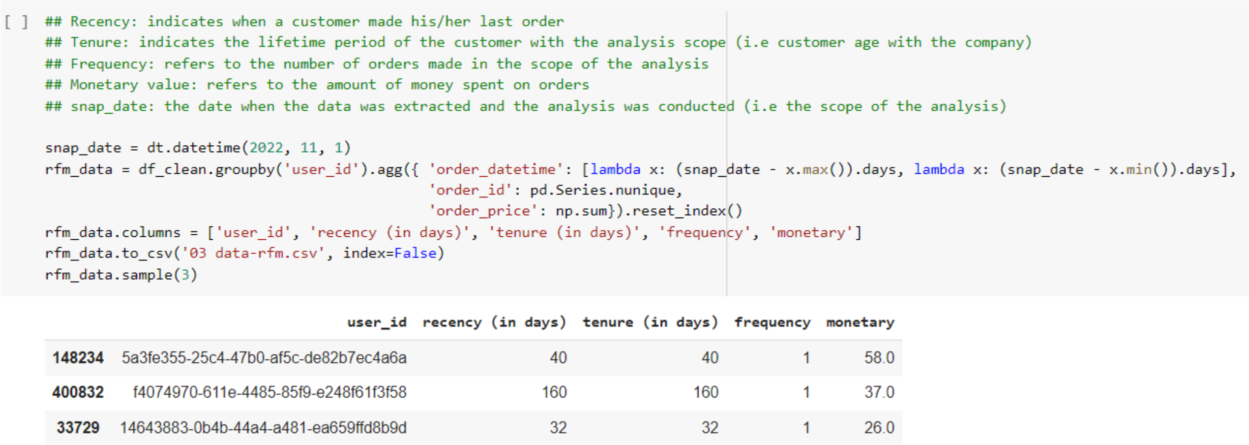
Monthly KPIs



```
query = \
'''
SELECT
  order_month AS "Month",
  COUNT(DISTINCT user_id) AS "#_active_users",
  COUNT(DISTINCT order_id) AS "#_orders",
  SUM(order_price) AS "revenues"
FROM df_clean
GROUP BY order_month
ORDER BY order_datetime ASC
'''
monthly_kpis = ps.sqldf(query)
monthly_kpis['Month'] = pd.to_datetime(monthly_kpis['Month'])
monthly_kpis['Month'] = monthly_kpis['Month'].apply(lambda x: get_month(x).strftime("%b %Y"))
monthly_kpis['revenues'] = monthly_kpis.revenues.astype(int)
monthly_kpis.to_csv('06 data-monthly-kpis.csv', index=False)
monthly_kpis.sample(3)
```

	Month	#_active_users	#_orders	revenues
12	Oct 2022	80860	159443	8949999
1	Nov 2021	49164	79279	4790153
0	Oct 2021	1237	1253	75133

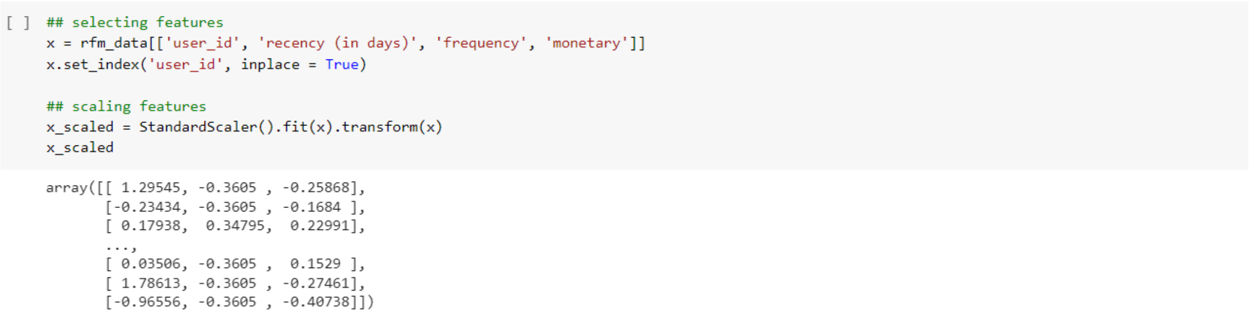
RFM Data



```
## Recency: indicates when a customer made his/her last order
## Tenure: indicates the lifetime period of the customer with the analysis scope (i.e customer age with the company)
## Frequency: refers to the number of orders made in the scope of the analysis
## Monetary value: refers to the amount of money spent on orders
## snap_date: the date when the data was extracted and the analysis was conducted (i.e the scope of the analysis)

snap_date = dt.datetime(2022, 11, 1)
rfm_data = df_clean.groupby('user_id').agg({ 'order_datetime': [lambda x: (snap_date - x.max()).days, lambda x: (snap_date - x.min()).days],
                                             'order_id': pd.Series.nunique,
                                             'order_price': np.sum}).reset_index()
rfm_data.columns = ['user_id', 'recency (in days)', 'tenure (in days)', 'frequency', 'monetary']
rfm_data.to_csv('03 data-rfm.csv', index=False)
rfm_data.sample(3)
```

	user_id	recency (in days)	tenure (in days)	frequency	monetary
148234	5a3fe355-25c4-47b0-af5c-de82b7ec4a6a	40	40	1	58.0
400832	f4074970-611e-4485-85f9-e248f61f3f58	160	160	1	37.0
33729	14643883-0b4b-44a4-a481-ea659ffd8b9d	32	32	1	26.0

Preprocessing Data

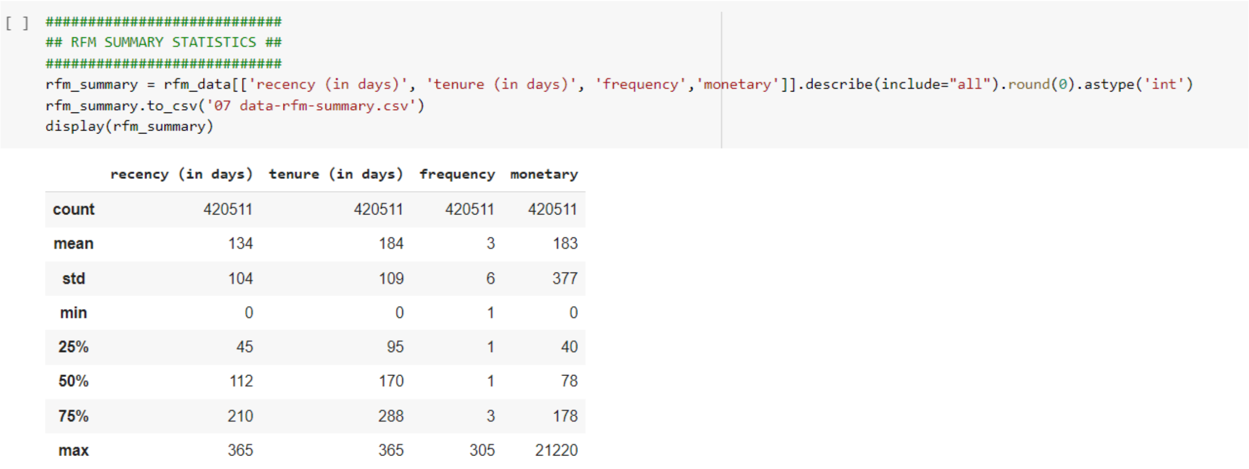


```
## selecting features
x = rfm_data[['user_id', 'recency (in days)', 'frequency', 'monetary']]
x.set_index('user_id', inplace = True)

## scaling features
x_scaled = StandardScaler().fit(x).transform(x)
x_scaled
```

```
array([[ 1.29545, -0.3605 , -0.25868],
       [-0.23434, -0.3605 , -0.1684 ],
       [ 0.17938,  0.34795,  0.22991],
       ...,
       [ 0.03506, -0.3605 ,  0.1529 ],
       [ 1.78613, -0.3605 , -0.27461],
       [-0.96556, -0.3605 , -0.40738]])
```

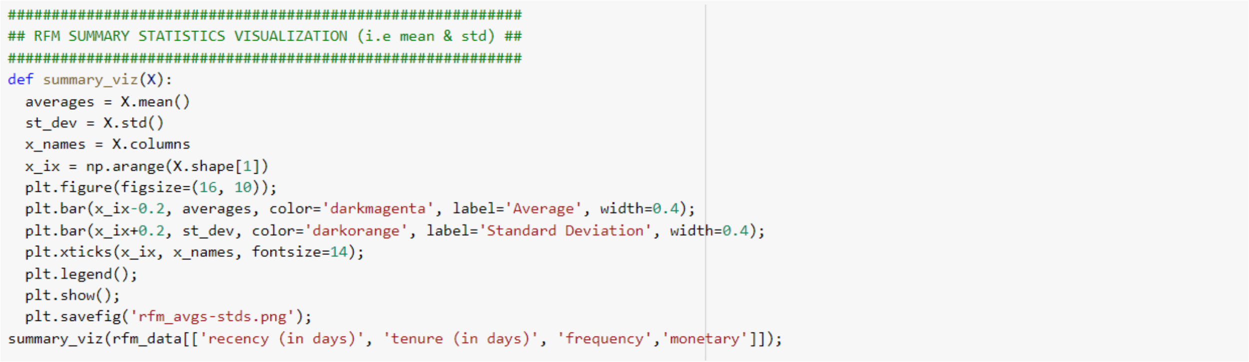
Exploratory Data Analysis (EDA)

RFM EDA

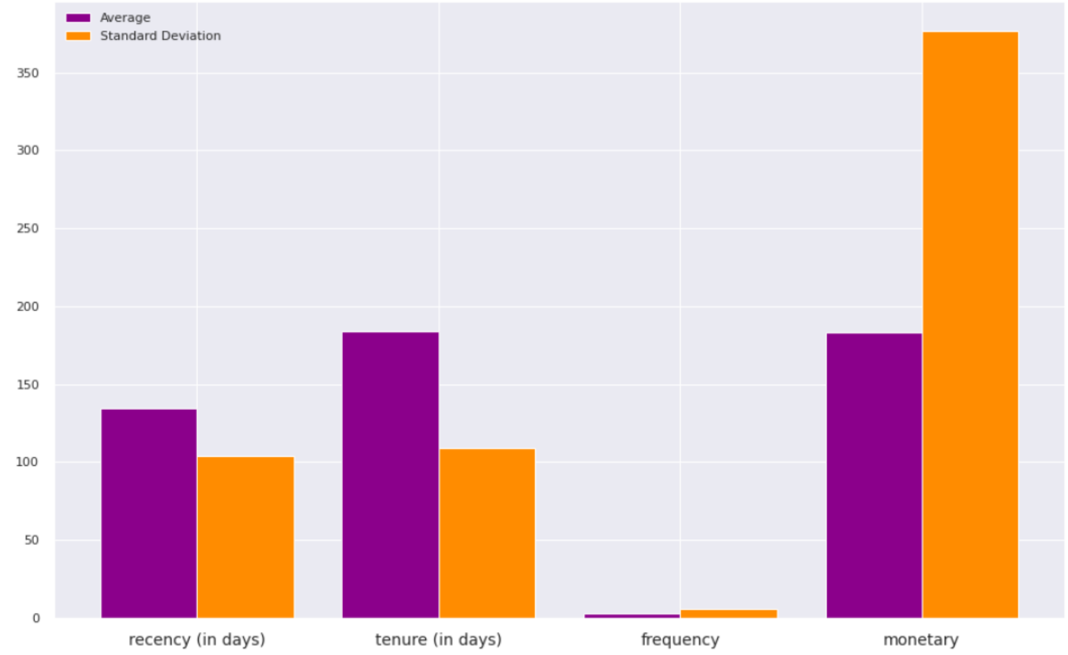


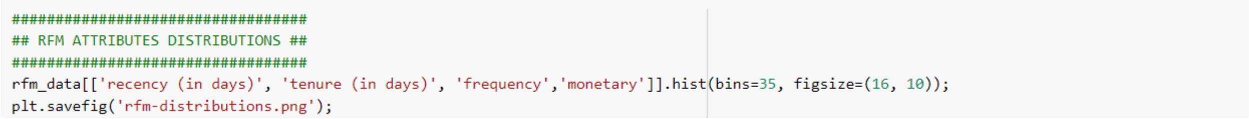
```
[ ] ###########################
    ## RFM SUMMARY STATISTICS ##
    ###########################
    rfm_summary = rfm_data[['recency (in days)', 'tenure (in days)', 'frequency','monetary']].describe(include="all").round(0).astype('int')
    rfm_summary.to_csv('07 data-rfm-summary.csv')
    display(rfm_summary)
```

	recency (in days)	tenure (in days)	frequency	monetary
count	420511	420511	420511	420511
mean	134	184	3	183
std	104	109	6	377
min	0	0	1	0
25%	45	95	1	40
50%	112	170	1	78
75%	210	288	3	178
max	365	365	305	21220

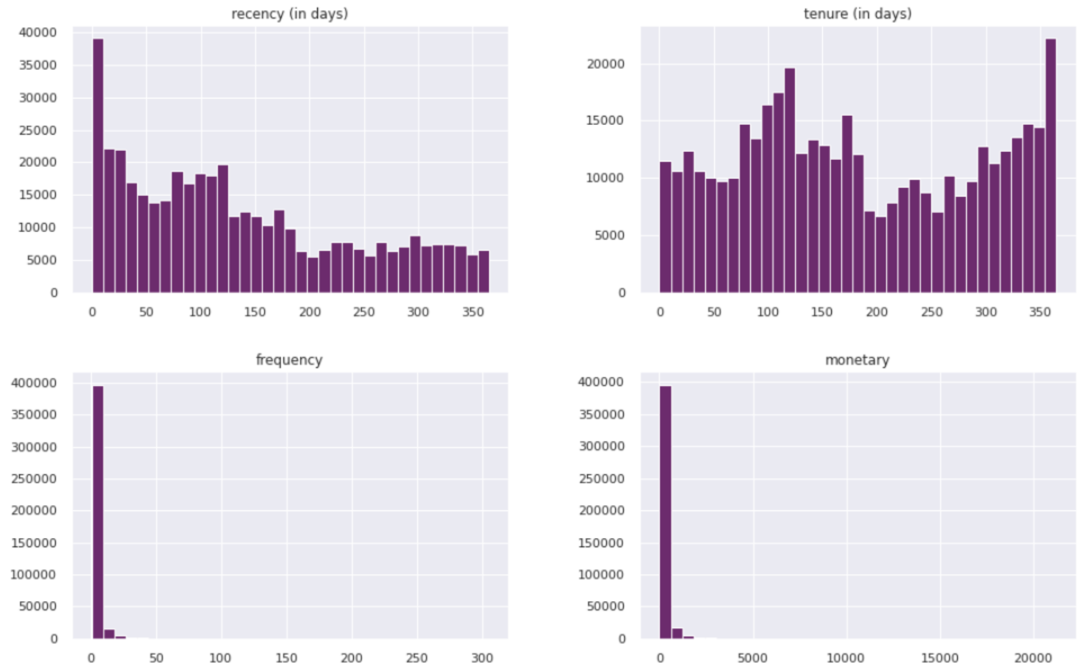


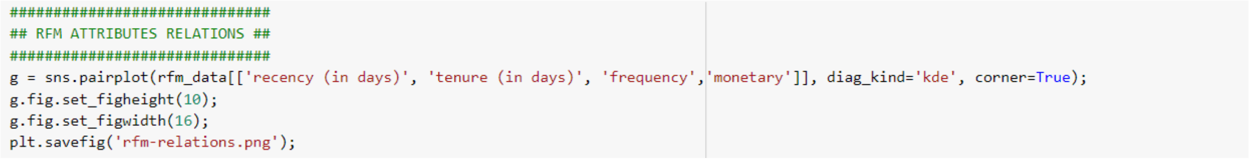
```
###########################################################
## RFM SUMMARY STATISTICS VISUALIZATION (i.e mean & std) ##
###########################################################
def summary_viz(X):
  averages = X.mean()
  st_dev = X.std()
  x_names = X.columns
  x_ix = np.arange(X.shape[1])
  plt.figure(figsize=(16, 10));
  plt.bar(x_ix-0.2, averages, color='darkmagenta', label='Average', width=0.4);
  plt.bar(x_ix+0.2, st_dev, color='darkorange', label='Standard Deviation', width=0.4);
  plt.xticks(x_ix, x_names, fontsize=14);
  plt.legend();
  plt.show();
  plt.savefig('rfm_avgs-stds.png');
summary_viz(rfm_data[['recency (in days)', 'tenure (in days)', 'frequency','monetary']]);
```



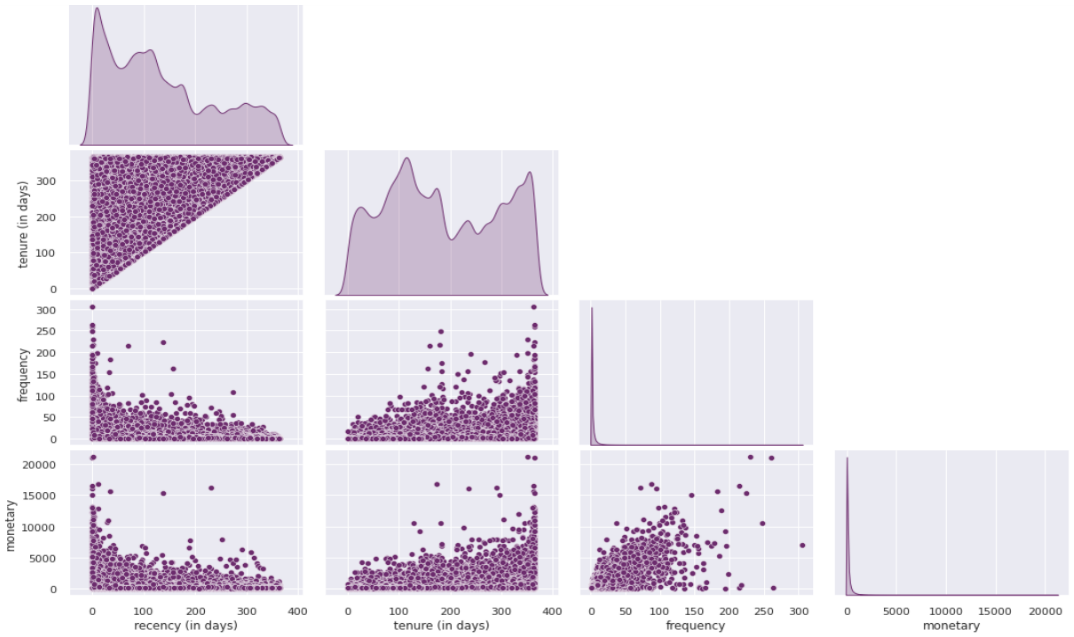


```
###################################
## RFM ATTRIBUTES DISTRIBUTIONS ##
###################################
rfm_data[['recency (in days)', 'tenure (in days)', 'frequency','monetary']].hist(bins=35, figsize=(16, 10));
plt.savefig('rfm-distributions.png');
```



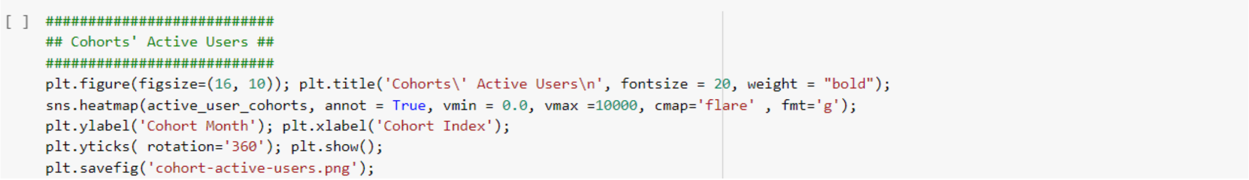


```
###############################
## RFM ATTRIBUTES RELATIONS ##
###############################
g = sns.pairplot(rfm_data[['recency (in days)', 'tenure (in days)', 'frequency','monetary']], diag_kind='kde', corner=True);
g.fig.set_figheight(10);
g.fig.set_figwidth(16);
plt.savefig('rfm-relations.png');
```

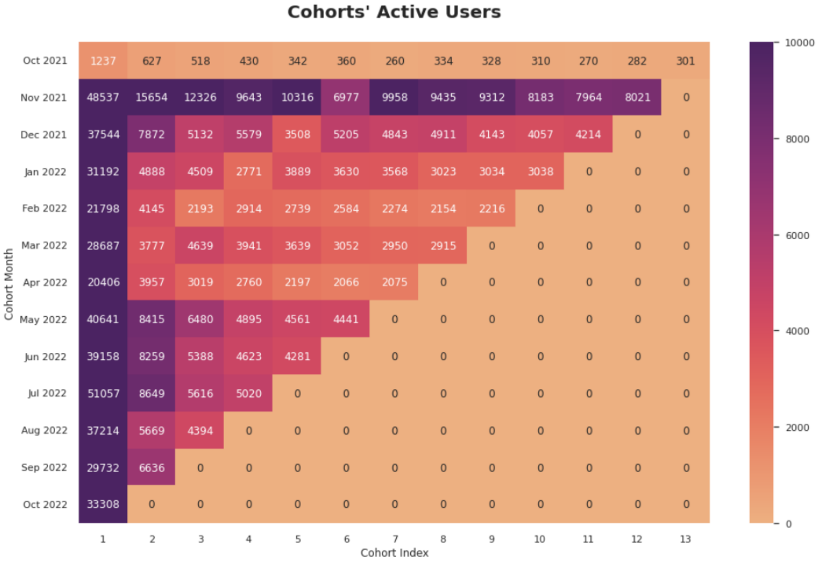


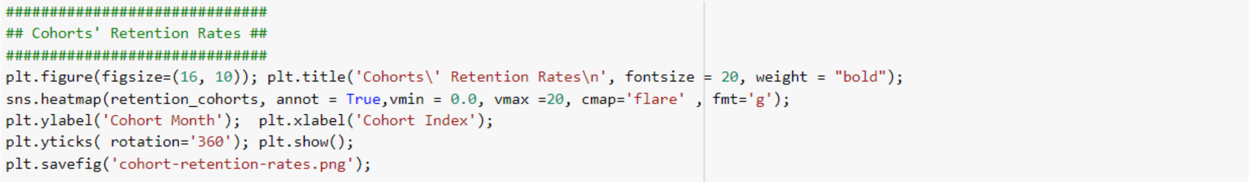
Cohort Analysis EDA

> Highlighting **Churn** Problems using Time-Based **Cohort Analysis**

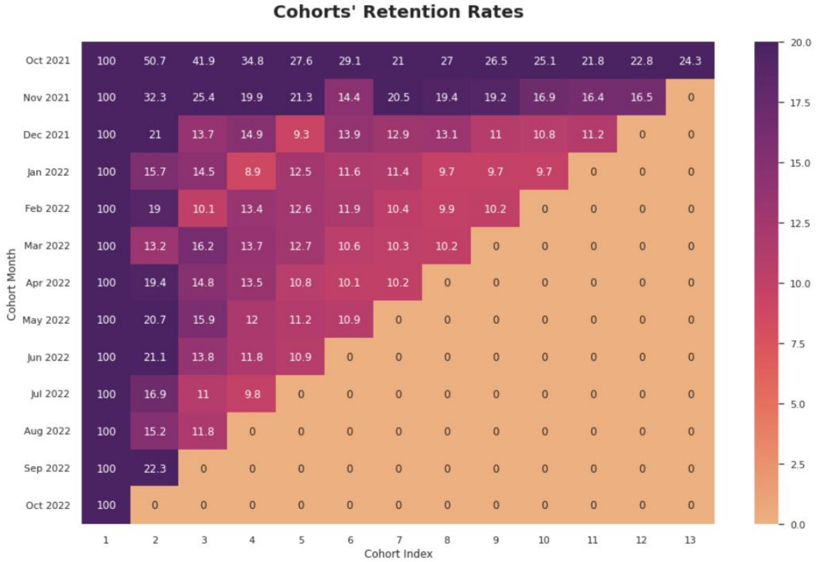


```
[ ] #############################
    ## Cohorts' Active Users ##
    #############################
    plt.figure(figsize=(16, 10)); plt.title('Cohorts\' Active Users\n', fontsize = 20, weight = "bold");
    sns.heatmap(active_user_cohorts, annot = True, vmin = 0.0, vmax =10000, cmap='flare' , fmt='g');
    plt.ylabel('Cohort Month'); plt.xlabel('Cohort Index');
    plt.yticks( rotation='360'); plt.show();
    plt.savefig('cohort-active-users.png');
```

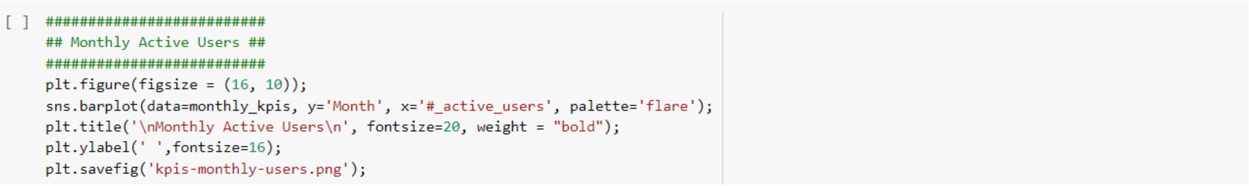




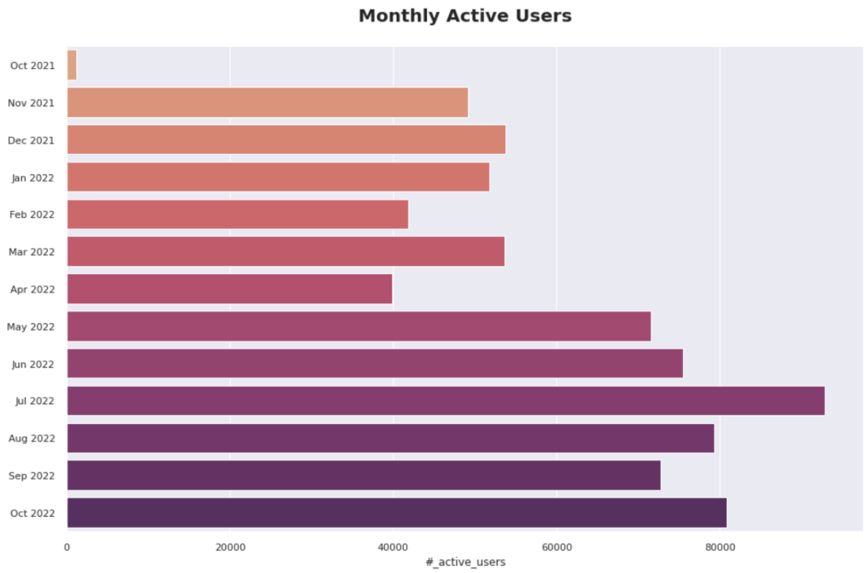
```
#################################
## Cohorts' Retention Rates ##
#################################
plt.figure(figsize=(16, 10)); plt.title('Cohorts\' Retention Rates\n', fontsize = 20, weight = "bold");
sns.heatmap(retention_cohorts, annot = True,vmin = 0.0, vmax =20, cmap='flare' , fmt='g');
plt.ylabel('Cohort Month');  plt.xlabel('Cohort Index');
plt.yticks( rotation='360'); plt.show();
plt.savefig('cohort-retention-rates.png');
```

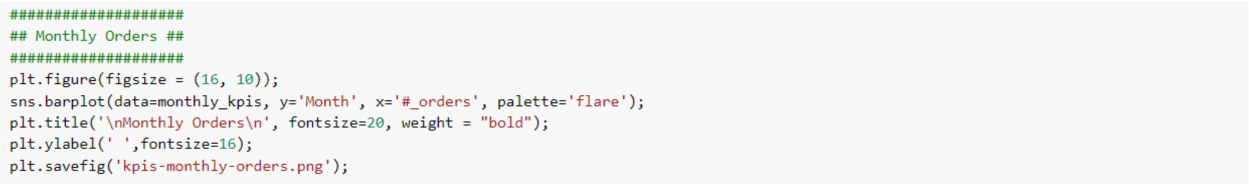


Monthly KPIs EDA

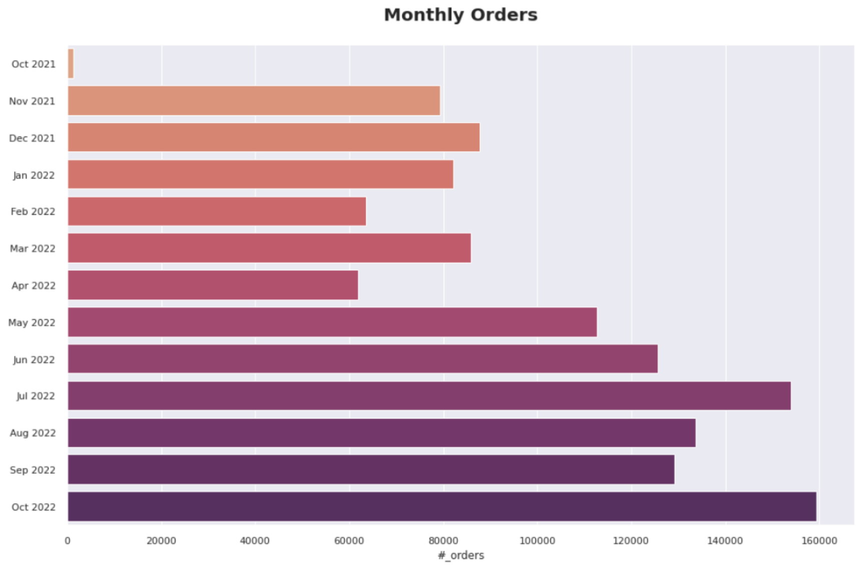


```
[ ] ##########################
    ## Monthly Active Users ##
    ##########################
    plt.figure(figsize = (16, 10));
    sns.barplot(data=monthly_kpis, y='Month', x='#_active_users', palette='flare');
    plt.title('\nMonthly Active Users\n', fontsize=20, weight = "bold");
    plt.ylabel(' ',fontsize=16);
    plt.savefig('kpis-monthly-users.png');
```





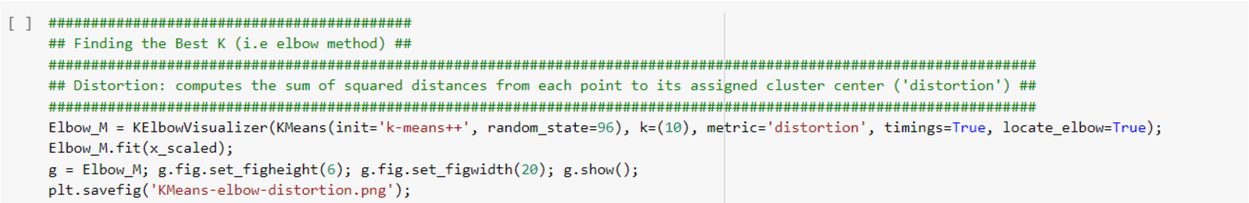
```
####################
## Monthly Orders ##
####################
plt.figure(figsize = (16, 10));
sns.barplot(data=monthly_kpis, y='Month', x='#_orders', palette='flare');
plt.title('\nMonthly Orders\n', fontsize=20, weight = "bold");
plt.ylabel(' ',fontsize=16);
plt.savefig('kpis-monthly-orders.png');
```



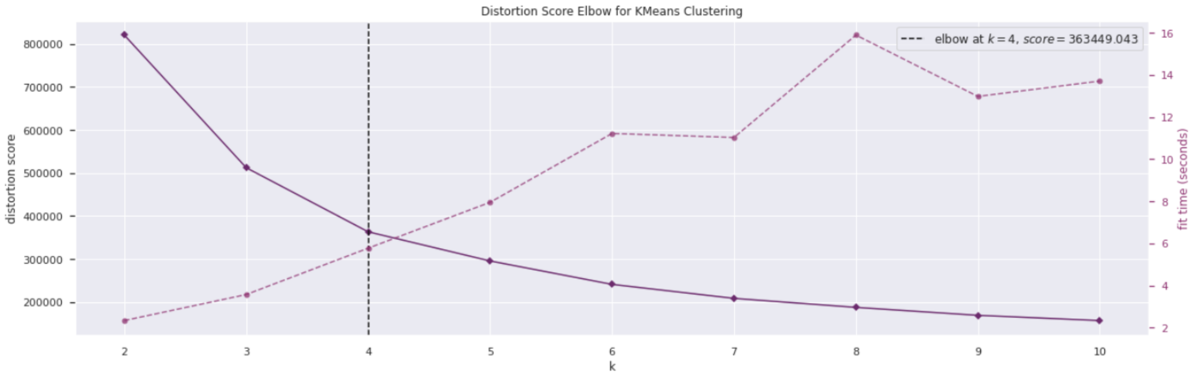
Modeling & Evaluation

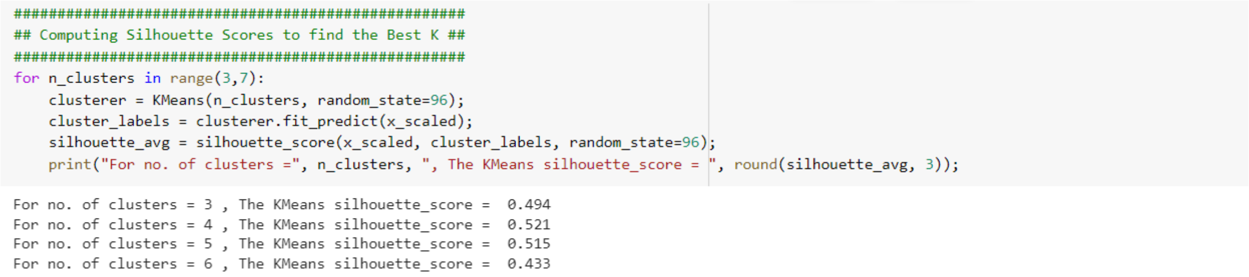
Customer Segmentation Using ML Algorithm

KMeans Clustering Algorithm



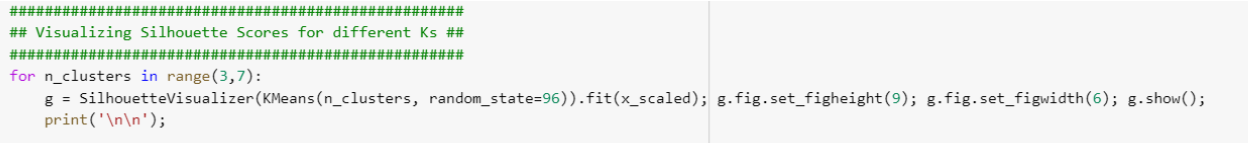
```
[ ] ###########################################
    ## Finding the Best K (i.e elbow method) ##
    #####################################################################################################################
    ## Distortion: computes the sum of squared distances from each point to its assigned cluster center ('distortion') ##
    #####################################################################################################################
    Elbow_M = KElbowVisualizer(KMeans(init='k-means++', random_state=96), k=(10), metric='distortion', timings=True, locate_elbow=True);
    Elbow_M.fit(x_scaled);
    g = Elbow_M; g.fig.set_figheight(6); g.fig.set_figwidth(20); g.show();
    plt.savefig('KMeans-elbow-distortion.png');
```



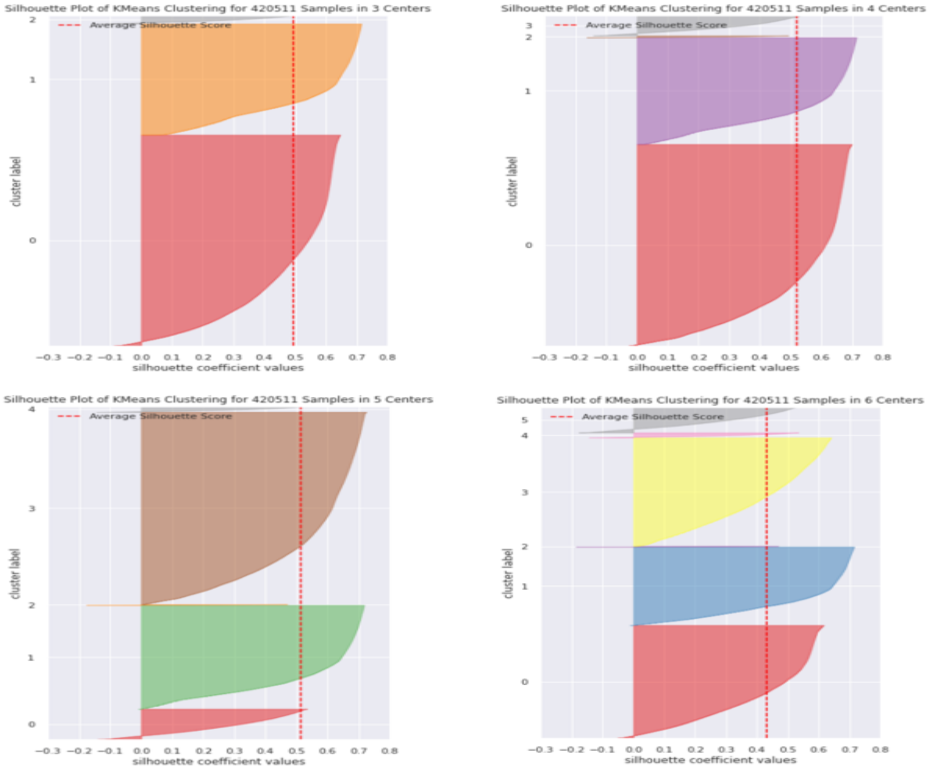


```
####################################################
## Computing Silhouette Scores to find the Best K ##
####################################################
for n_clusters in range(3,7):
    clusterer = KMeans(n_clusters, random_state=96);
    cluster_labels = clusterer.fit_predict(x_scaled);
    silhouette_avg = silhouette_score(x_scaled, cluster_labels, random_state=96);
    print("For no. of clusters =", n_clusters, ", The KMeans silhouette_score = ", round(silhouette_avg, 3));
```

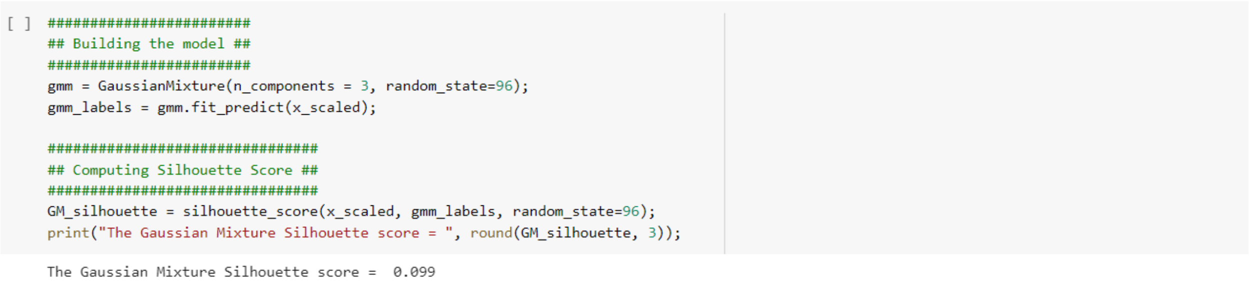
```
For no. of clusters = 3 , The KMeans silhouette_score =  0.494
For no. of clusters = 4 , The KMeans silhouette_score =  0.521
For no. of clusters = 5 , The KMeans silhouette_score =  0.515
For no. of clusters = 6 , The KMeans silhouette_score =  0.433
```



```
########################################################
## Visualizing Silhouette Scores for different Ks ##
########################################################
for n_clusters in range(3,7):
    g = SilhouetteVisualizer(KMeans(n_clusters, random_state=96)).fit(x_scaled); g.fig.set_figheight(9); g.fig.set_figwidth(6); g.show();
    print('\n\n');
```

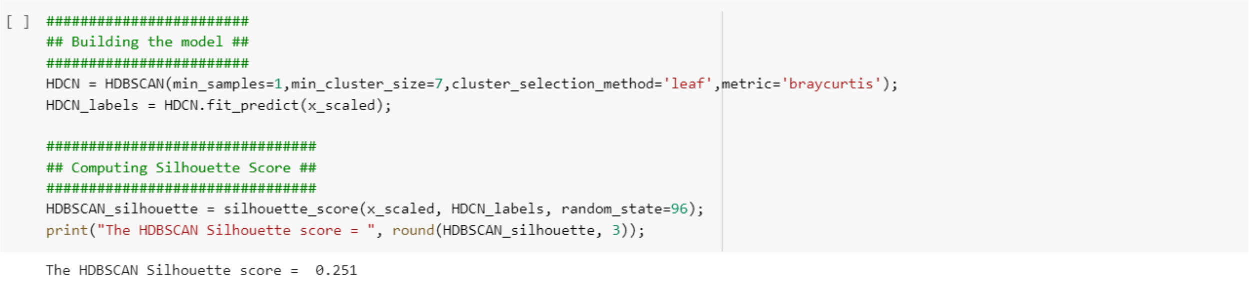


GaussianMixture Clustering Algorithm



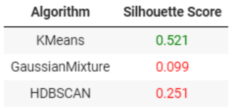
```
[ ] ########################
    ## Building the model ##
    ########################
    gmm = GaussianMixture(n_components = 3, random_state=96);
    gmm_labels = gmm.fit_predict(x_scaled);

    ################################
    ## Computing Silhouette Score ##
    ################################
    GM_silhouette = silhouette_score(x_scaled, gmm_labels, random_state=96);
    print("The Gaussian Mixture Silhouette score = ", round(GM_silhouette, 3));

    The Gaussian Mixture Silhouette score =  0.099
```

HDBSCAN Clustering Algorithm



```
[ ]  ########################
     ## Building the model ##
     ########################
     HDCN = HDBSCAN(min_samples=1,min_cluster_size=7,cluster_selection_method='leaf',metric='braycurtis');
     HDCN_labels = HDCN.fit_predict(x_scaled);

     ################################
     ## Computing Silhouette Score ##
     ################################
     HDBSCAN_silhouette = silhouette_score(x_scaled, HDCN_labels, random_state=96);
     print("The HDBSCAN Silhouette score = ", round(HDBSCAN_silhouette, 3));

     The HDBSCAN Silhouette score =  0.251
```

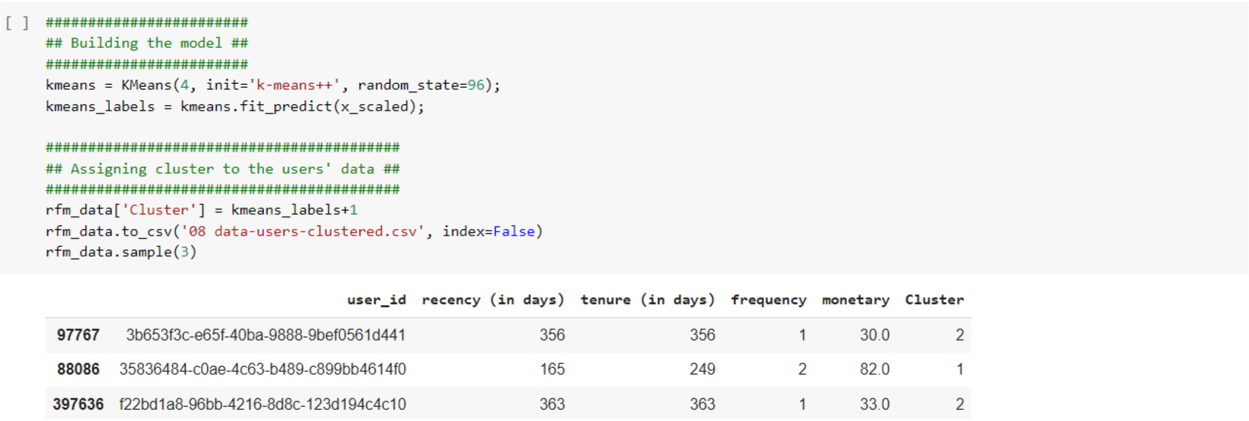
Evaluation



Algorithm	Silhouette Score
KMeans	0.521
GaussianMixture	0.099
HDBSCAN	0.251

The **KMeans** model with **4 clusters** shows **better** performance than the other algorithms according to the **Silhouette** score

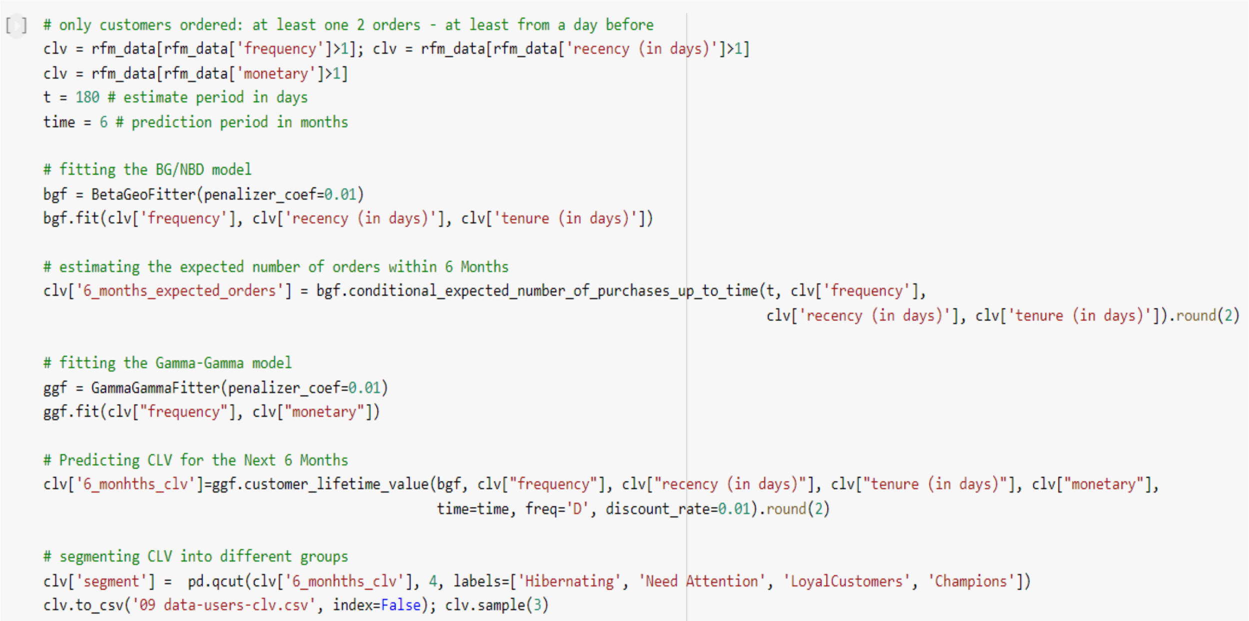
Deployment



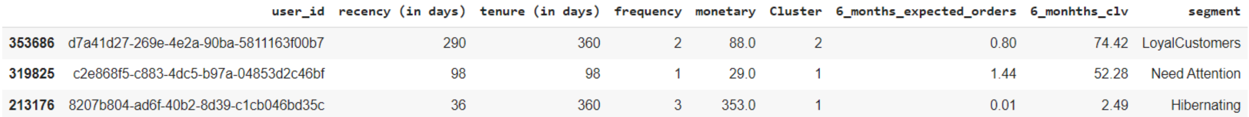
```
[ ]  ########################
     ## Building the model ##
     ########################
     kmeans = KMeans(4, init='k-means++', random_state=96);
     kmeans_labels = kmeans.fit_predict(x_scaled);

     ##########################################
     ## Assigning cluster to the users' data ##
     ##########################################
     rfm_data['Cluster'] = kmeans_labels+1
     rfm_data.to_csv('08 data-users-clustered.csv', index=False)
     rfm_data.sample(3)
```

	user_id	recency (in days)	tenure (in days)	frequency	monetary	Cluster
97767	3b653f3c-e65f-40ba-9888-9bef0561d441	356	356	1	30.0	2
88086	35836484-c0ae-4c63-b489-c899bb4614f0	165	249	2	82.0	1
397636	f22bd1a8-96bb-4216-8d8c-123d194c4c10	363	363	1	33.0	2

Customer Lifetime Value (**CLTV**) Prediction



```
[ ] # only customers ordered: at least one 2 orders - at least from a day before
    clv = rfm_data[rfm_data['frequency']>1]; clv = rfm_data[rfm_data['recency (in days)']>1]
    clv = rfm_data[rfm_data['monetary']>1]
    t = 180 # estimate period in days
    time = 6 # prediction period in months

    # fitting the BG/NBD model
    bgf = BetaGeoFitter(penalizer_coef=0.01)
    bgf.fit(clv['frequency'], clv['recency (in days)'], clv['tenure (in days)'])

    # estimating the expected number of orders within 6 Months
    clv['6_months_expected_orders'] = bgf.conditional_expected_number_of_purchases_up_to_time(t, clv['frequency'],
                                                                                              clv['recency (in days)'], clv['tenure (in days)']).round(2)

    # fitting the Gamma-Gamma model
    ggf = GammaGammaFitter(penalizer_coef=0.01)
    ggf.fit(clv["frequency"], clv["monetary"])

    # Predicting CLV for the Next 6 Months
    clv['6_monhths_clv']=ggf.customer_lifetime_value(bgf, clv["frequency"], clv["recency (in days)"], clv["tenure (in days)"], clv["monetary"],
                                                 time=time, freq='D', discount_rate=0.01).round(2)

    # segmenting CLV into different groups
    clv['segment'] =  pd.qcut(clv['6_monhths_clv'], 4, labels=['Hibernating', 'Need Attention', 'LoyalCustomers', 'Champions'])
    clv.to_csv('09 data-users-clv.csv', index=False); clv.sample(3)
```

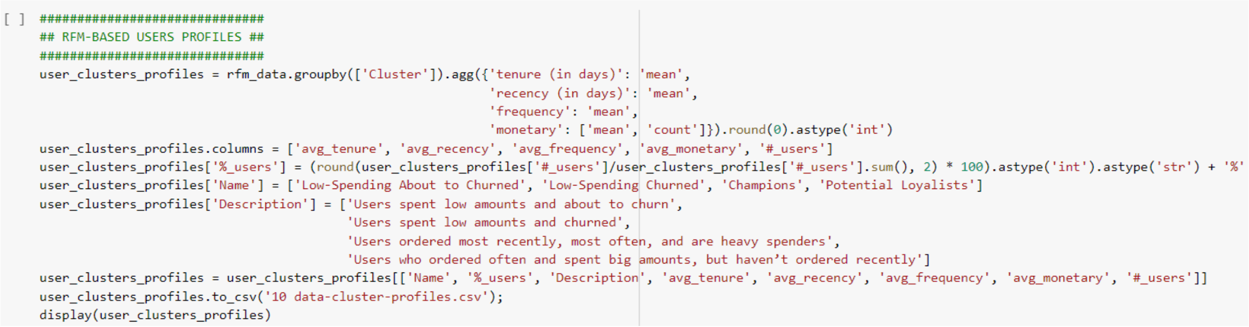


	user_id	recency (in days)	tenure (in days)	frequency	monetary	Cluster	6_months_expected_orders	6_monhths_clv	segment
353686	d7a41d27-269e-4e2a-90ba-5811163f00b7	290	360	2	88.0	2	0.80	74.42	LoyalCustomers
319825	c2e868f5-c883-4dc5-b97a-04853d2c46bf	98	98	1	29.0	1	1.44	52.28	Need Attention
213176	8207b804-ad6f-40b2-8d39-c1cb046bd35c	36	360	3	353.0	1	0.01	2.49	Hibernating

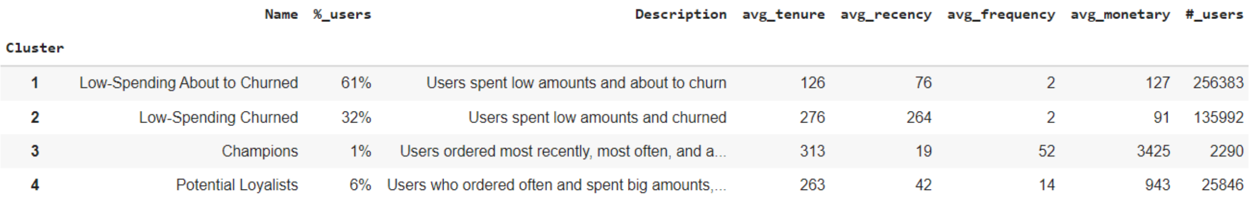
Results

RFM-Based Customer Segmentations

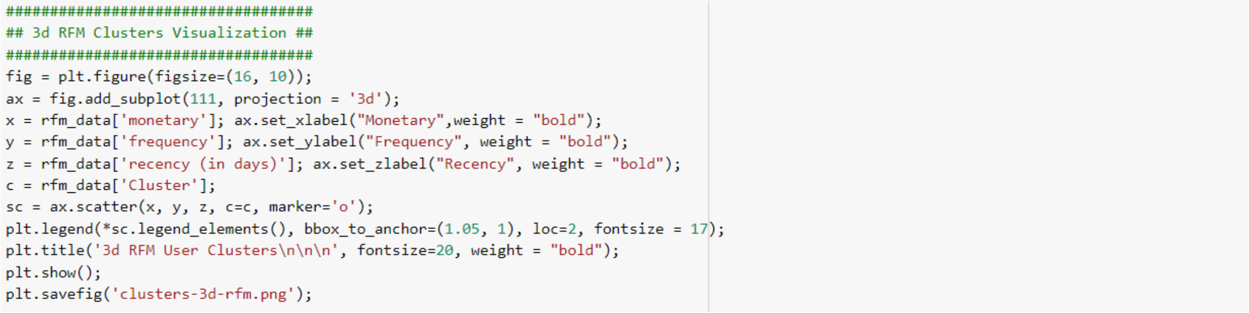
Insights



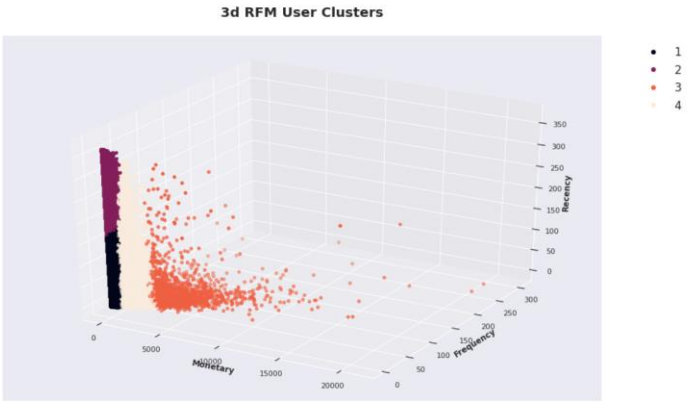
```
[ ] ##############################
    ## RFM-BASED USERS PROFILES ##
    ##############################
    user_clusters_profiles = rfm_data.groupby(['Cluster']).agg({'tenure (in days)': 'mean',
                                                                'recency (in days)': 'mean',
                                                                'frequency': 'mean',
                                                                'monetary': ['mean', 'count']}).round(0).astype('int')
    user_clusters_profiles.columns = ['avg_tenure', 'avg_recency', 'avg_frequency', 'avg_monetary', '#_users']
    user_clusters_profiles['%_users'] = (round(user_clusters_profiles['#_users']/user_clusters_profiles['#_users'].sum(), 2) * 100).astype('int').astype('str') + '%'
    user_clusters_profiles['Name'] = ['Low-Spending About to Churned', 'Low-Spending Churned', 'Champions', 'Potential Loyalists']
    user_clusters_profiles['Description'] = ['Users spent low amounts and about to churn',
                                             'Users spent low amounts and churned',
                                             'Users ordered most recently, most often, and are heavy spenders',
                                             'Users who ordered often and spent big amounts, but haven’t ordered recently']
    user_clusters_profiles = user_clusters_profiles[['Name', '%_users', 'Description', 'avg_tenure', 'avg_recency', 'avg_frequency', 'avg_monetary', '#_users']]
    user_clusters_profiles.to_csv('10 data-cluster-profiles.csv');
    display(user_clusters_profiles)
```

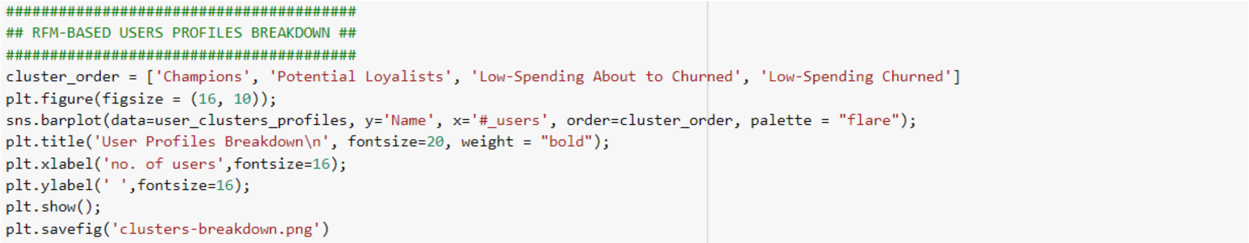


	Name	%_users	Description	avg_tenure	avg_recency	avg_frequency	avg_monetary	#_users
Cluster								
1	Low-Spending About to Churned	61%	Users spent low amounts and about to churn	126	76	2	127	256383
2	Low-Spending Churned	32%	Users spent low amounts and churned	276	264	2	91	135992
3	Champions	1%	Users ordered most recently, most often, and a...	313	19	52	3425	2290
4	Potential Loyalists	6%	Users who ordered often and spent big amounts,...	263	42	14	943	25846

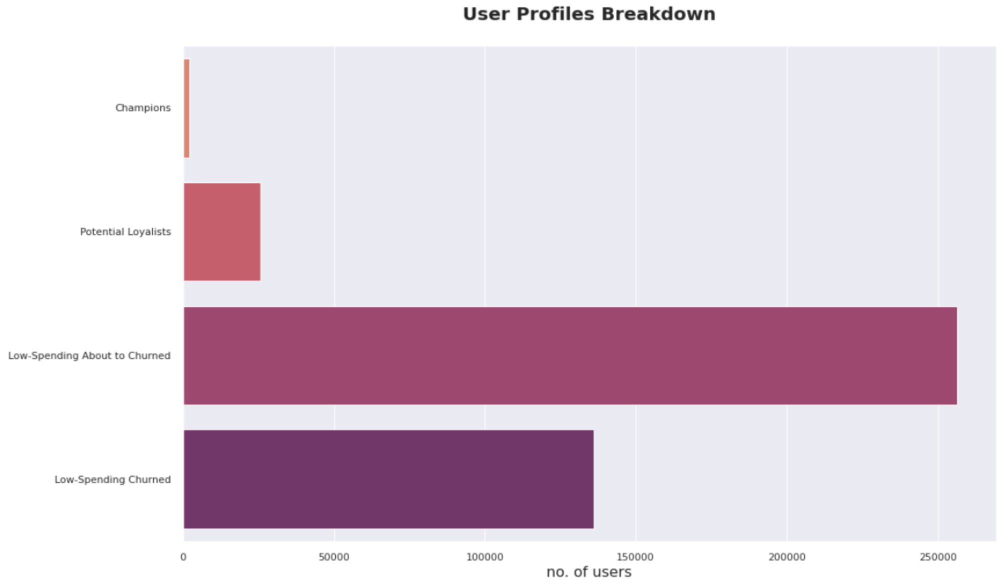


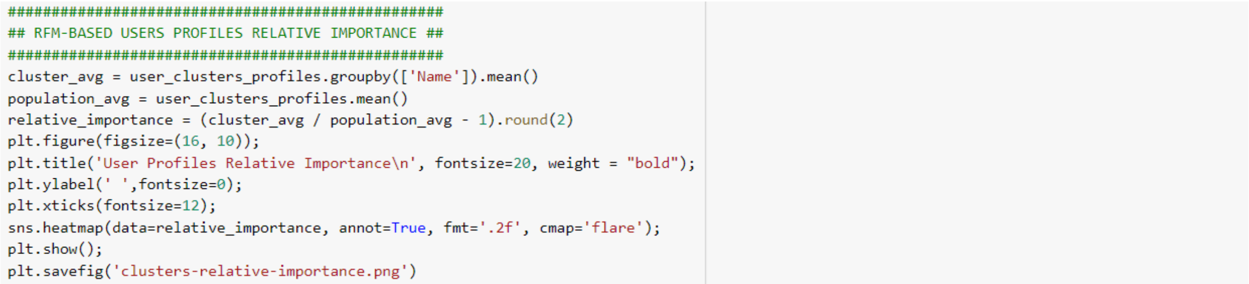
```
####################################
## 3d RFM Clusters Visualization ##
####################################
fig = plt.figure(figsize=(16, 10));
ax = fig.add_subplot(111, projection = '3d');
x = rfm_data['monetary']; ax.set_xlabel("Monetary",weight = "bold");
y = rfm_data['frequency']; ax.set_ylabel("Frequency", weight = "bold");
z = rfm_data['recency (in days)']; ax.set_zlabel("Recency", weight = "bold");
c = rfm_data['Cluster'];
sc = ax.scatter(x, y, z, c=c, marker='o');
plt.legend(*sc.legend_elements(), bbox_to_anchor=(1.05, 1), loc=2, fontsize = 17);
plt.title('3d RFM User Clusters\n\n\n', fontsize=20, weight = "bold");
plt.show();
plt.savefig('clusters-3d-rfm.png');
```



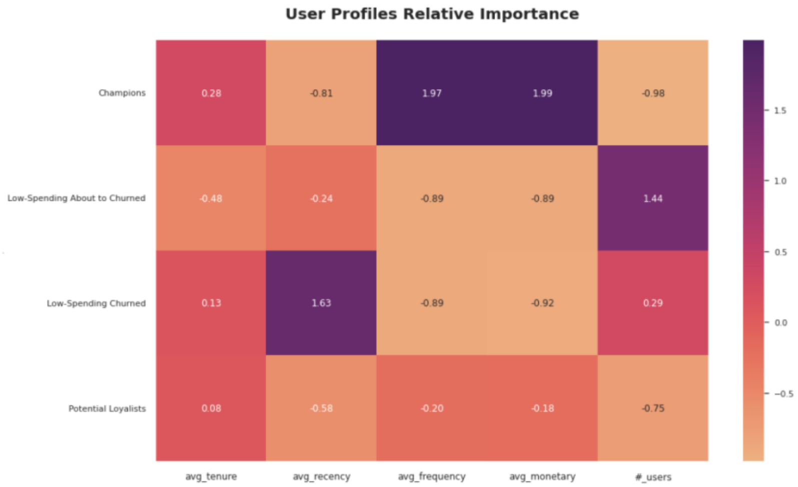


```
########################################
## RFM-BASED USERS PROFILES BREAKDOWN ##
########################################
cluster_order = ['Champions', 'Potential Loyalists', 'Low-Spending About to Churned', 'Low-Spending Churned']
plt.figure(figsize = (16, 10));
sns.barplot(data=user_clusters_profiles, y='Name', x='#_users', order=cluster_order, palette = "flare");
plt.title('User Profiles Breakdown\n', fontsize=20, weight = "bold");
plt.xlabel('no. of users',fontsize=16);
plt.ylabel(' ',fontsize=16);
plt.show();
plt.savefig('clusters-breakdown.png')
```





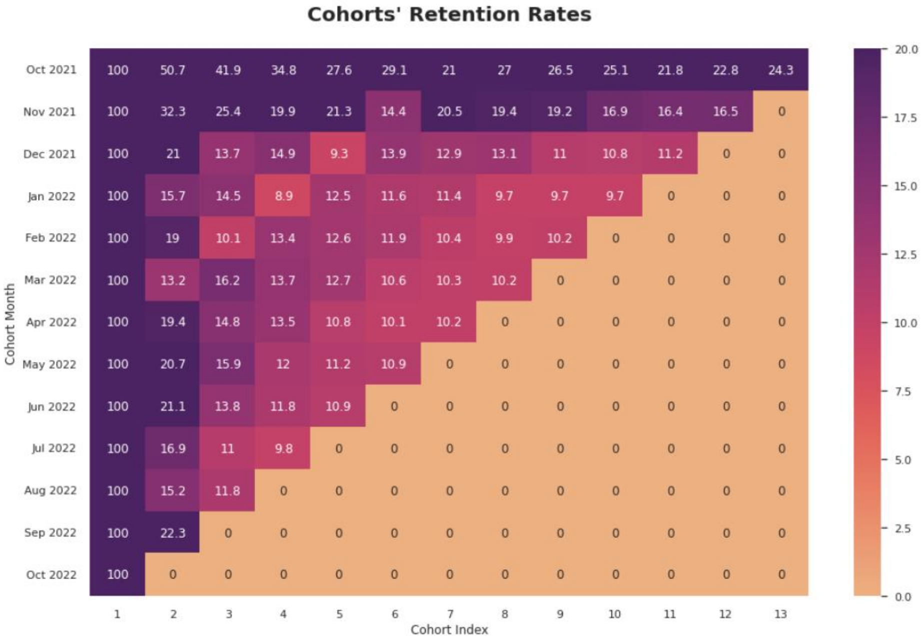
```
##################################################
## RFM-BASED USERS PROFILES RELATIVE IMPORTANCE ##
##################################################
cluster_avg = user_clusters_profiles.groupby(['Name']).mean()
population_avg = user_clusters_profiles.mean()
relative_importance = (cluster_avg / population_avg - 1).round(2)
plt.figure(figsize=(16, 10));
plt.title('User Profiles Relative Importance\n', fontsize=20, weight = "bold");
plt.ylabel(' ',fontsize=0);
plt.xticks(fontsize=12);
sns.heatmap(data=relative_importance, annot=True, fmt='.2f', cmap='flare');
plt.show();
plt.savefig('clusters-relative-importance.png')
```



CLTV-based Predictions

Insights

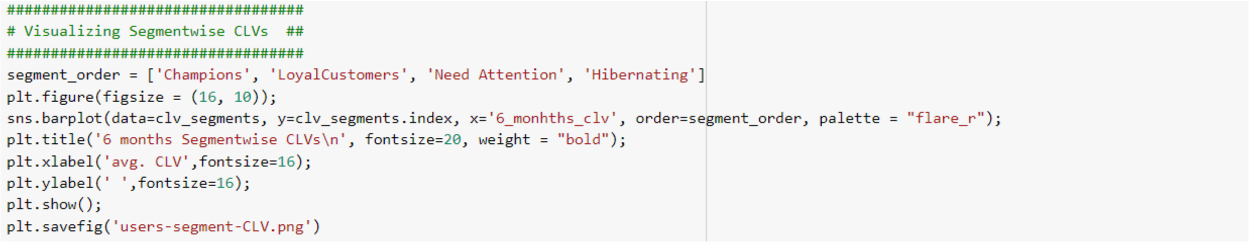
```
[ ] ##############################
    ## Cohorts' Retention Rates ##
    ##############################
    plt.figure(figsize=(16, 10)); plt.title('Cohorts\' Retention Rates\n', fontsize = 20, weight = "bold");
    sns.heatmap(retention_cohorts, annot = True,vmin = 0.0, vmax =20, cmap='flare' , fmt='g');
    plt.ylabel('Cohort Month');  plt.xlabel('Cohort Index');
    plt.yticks( rotation='360'); plt.show();
```



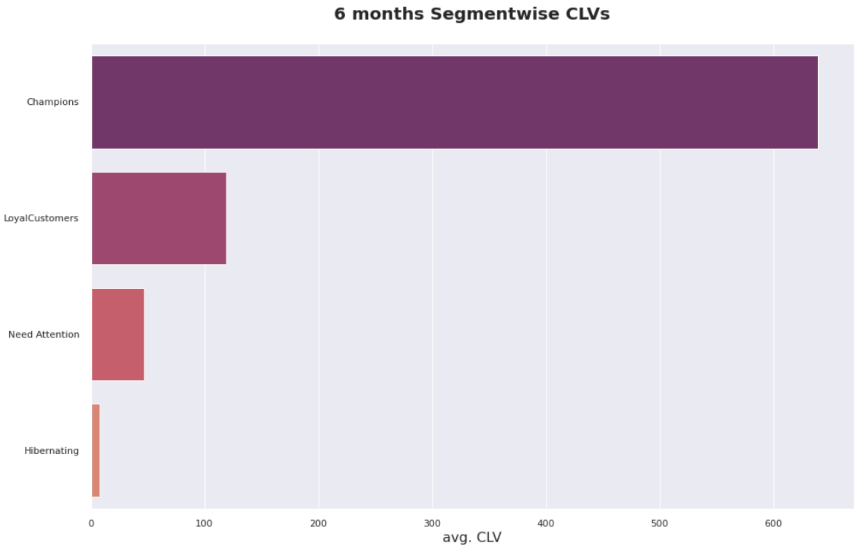
	1	2	3	4	5	6	7	8	9	10	11	12	13
Oct 2021	100	50.7	41.9	34.8	27.6	29.1	21	27	26.5	25.1	21.8	22.8	24.3
Nov 2021	100	32.3	25.4	19.9	21.3	14.4	20.5	19.4	19.2	16.9	16.4	16.5	0
Dec 2021	100	21	13.7	14.9	9.3	13.9	12.9	13.1	11	10.8	11.2	0	0
Jan 2022	100	15.7	14.5	8.9	12.5	11.6	11.4	9.7	9.7	9.7	0	0	0
Feb 2022	100	19	10.1	13.4	12.6	11.9	10.4	9.9	10.2	0	0	0	0
Mar 2022	100	13.2	16.2	13.7	12.7	10.6	10.3	10.2	0	0	0	0	0
Apr 2022	100	19.4	14.8	13.5	10.8	10.1	10.2	0	0	0	0	0	0
May 2022	100	20.7	15.9	12	11.2	10.9	0	0	0	0	0	0	0
Jun 2022	100	21.1	13.8	11.8	10.9	0	0	0	0	0	0	0	0
Jul 2022	100	16.9	11	9.8	0	0	0	0	0	0	0	0	0
Aug 2022	100	15.2	11.8	0	0	0	0	0	0	0	0	0	0
Sep 2022	100	22.3	0	0	0	0	0	0	0	0	0	0	0
Oct 2022	100	0	0	0	0	0	0	0	0	0	0	0	0

```
####################################
## Segmenting Users Based on CLVs ##
####################################
clv_segments = clv.groupby('segment').mean().round(0).astype('int')
clv_segments.to_csv('11 data-clv-segments-profiles.csv')
display(clv_segments)
```

	recency (in days)	tenure (in days)	frequency	monetary	Cluster	6_months_expected_orders	6_monhths_clv
segment							
Hibernating	113	250	7	403	2	0	8
Need Attention	188	207	1	67	2	1	47
LoyalCustomers	136	159	2	100	1	2	119
Champions	97	120	3	193	1	4	639



```
##################################
# Visualizing Segmentwise CLVs  ##
##################################
segment_order = ['Champions', 'LoyalCustomers', 'Need Attention', 'Hibernating']
plt.figure(figsize = (16, 10));
sns.barplot(data=clv_segments, y=clv_segments.index, x='6_monhths_clv', order=segment_order, palette = "flare_r");
plt.title('6 months Segmentwise CLVs\n', fontsize=20, weight = "bold");
plt.xlabel('avg. CLV',fontsize=16);
plt.ylabel(' ',fontsize=16);
plt.show();
plt.savefig('users-segment-CLV.png')
```


6 months Segmentwise CLVs
Champions
LoyalCustomers
Need Attention
Hibernating
avg. CLV