Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

XYZ Sports Company: Customer Segmentation through RFM and Cluster Analysis

Group 82

Cláudia Beiral, number: 20230387

Diogo Pimenta, number: 20230498

João Maia, number: 20230746

January, 2024

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Index

Introduction

XYZ Sports Company is a longstanding fitness facility committed to gain insight into its user base. With that, the company will be able to optimize its marketing efforts, create personalized services, and enhance overall customer satisfaction by segmenting their customer base.

Customer segmentation involves dividing a customer base into distinct groups based on certain characteristics, behaviors, or other relevant criteria. The goal is to better understand the various needs and preferences of different segments, which enables companies to customize their offerings, services and marketing approaches to address the demands of each group more effectively.

This project aims to develop the customer segmentation from different perspectives and using two approaches: RFM analysis and the clustering algorithm K-means. To achieve that, we will leverage the data from the company's ERP system between the 1st of June of 2014 and the 31st of October of 2019.

Finally, we also want to give some advice if something is going out of the rails and to formulate a few marketing strategies that we think that can be helpful for the company.

Exploratory Data Analysis (EDA)

XYZ Sports Company's dataset provides a wealth of customer-related data and it has 14942 records and 30 features associated with the customer´s profile. These entail demographic information (age, gender, income), the different time periods by which the clients have been enrolled, the many activities that the customers participate, date of the last visit, days without frequency, lifetime value, enrolment status (whether the customer left the company or not), among many others.

First, in order to have better access to our data, we set the column that identifies the customer records, ID, as index. Then, we checked the data types of each variable in the dataset and noticed that gender had all its values in string type. As such, for a better evaluation of our data in the future we decided to change its types later on.

We checked for possible missing values with isna().sum(), and we found several missing values in the categorical variables, especially in activity columns.

Data Incoherence

- Individuals Under 16 Years Old Earning Income

We observed that individuals under 16 typically lack regular income due to legal restrictions on employment. To reflect this reality, we uniformly set their reported income to zero.

- Lifetime Value at Zero

Three customers exhibited a 'Lifetime Value' of 0. Considering this as an inconsistency in the dataset—there being no feasible scenario for a customer with zero payments—we decided to remove these cases.

- Inconsistent number of references

There were some instances where ‘Number of References’ is zero but the variable ‘Has References’ was marked as 1. This is contradictory and was rectified by setting the variable ‘Has references’ to zero in such cases. We also checked if there were any cases where the ‘Number of References’ was not zero, yet ‘HasReferences’ was marked as zero, but there were no such cases.

- Enrollment Start and Enrollment Finish

We noticed numerous instances where the enrollment start date coincided with the enrollment finish date. In all these cases, the 'Dropout' variable was marked as 0, indicating that these clients did not discontinue their enrollment. To align these records with others who hadn't dropped out and had their enrollment finish set to the dataset's closing date (2019-10-31), we decided to update the enrollment finish date for these specific clients to match that closing date. This adjustment aimed to maintain consistency within the dataset for clients who completed their enrollment without dropping out.

Duplicates

Upon inspecting the dataset for duplicate entries, we identified only a singular duplicated record, which we subsequently removed.

Handling missing values

This is an important task in our project, as treating the missing values properly could lead to better clustering, which would ultimately make sure our findings are the most accurate and reliable possible.

- Income

In this variable, we considered the influence of age on income. We assumed that individuals under 18 years old most likely do not have an income yet, so we set their income value to 0. For those over 18, we replaced missing income values with the mean income of individuals in that age group. There was deliberation about setting the age threshold for zero income at 16 instead of 18, but it was reasoned that individuals aged 16 to 18 are typically engaged in schooling and might not be working full time.

Post-project, we discussed the concept of using linear regression to fill missing income values based on age, given the strong correlation observed between age and income. This correlation suggests that as individuals get older, their income tends to increase.

- Activities

We opted for a uniform approach in handling missing values across all activities. Our rationale was based on the observation that each activity typically has a small number of attendees and consequently we chose to fill the missing values with zeros, presuming that these absences indicate the individual did not engage in the respective activity.

- Number of frequencies

We chose to remove these rows because the 'NumberOfFrequencies' values were so different from one another that it was impossible to make any reliable predictions about them. Since they represented only a tiny part of the whole dataset, getting rid of them shouldn't affect the clustering much, and it seemed better than using potentially inaccurate or misleading information in our analysis.

- Has References

Upon observation, we noticed that the missing values (NaNs) in this specific variable corresponded to entries where the number of references was greater than 0. Therefore, we made the decision to replace these NaNs with the value 1.

- Allowed Weekly Visits by SLA

We initially attempted to predict the number of visits based on the 'AllowedWeeklyVisits' variable. However, our analysis indicated that this approach was ineffective due to the limited and specific values within the 'AllowedWeeklyVisits' column. Even employing a linear regression yielded unsatisfactory results. Upon further investigation, we found that around 3.58% of the rows had missing values (NaNs) in this 'AllowedWeeklyVisits' variable. Considering this proportion is relatively small in the dataset and aiming to preserve the integrity of the stipulated outcomes outlined in the service level agreement, we made the decision to drop these rows. This action ensures a more accurate analysis by maintaining the reliability of the specified visit allowances.

Feature Engineering

Gender

Initially, we explored employing the one-hot encoding technique for the 'Gender' feature. However, upon evaluation, we found that these binary variables were essentially inversely related, introducing redundant information without enhancing the analysis.

Subsequently, we opted for label encoding as our final solution for handling the 'Gender' feature. This approach assigned 'Male' as 1 and 'Female' as 0, transforming the categorical data into numerical format, while avoiding the redundancy introduced by one-hot-encoding.

Enrollment Duration

We began by computing the 'EnrollmentDuration_days' by subtracting the enrollment start date from the enrollment finish date, capturing the duration in days. Realizing the need for a more extensive timeframe, we transformed this duration into months ('EnrollmentDuration_months'). As well as semesters 'EnrollmentDuration_semesters'. We intentionally focused on semesters as it aligns well with our upcoming analysis techniques, ensuring a contextually relevant examination of enrollment lengths within the company's framework.

Last period duration

We calculated the days between 'LastPeriodStart' and 'LastPeriodFinish anticipating its potential relevance for other SLA related variables. This mirrors our previous approach, determining the duration of the last period in days, similar to the previous feature creations. However, we ultimately didn't give much use to this feature in our subsequent analyses or investigation.

These following features offer valuable insights into customer behavior, specifically focusing on their engagement recency, frequency and monetary contributions throughout their enrollment. They play a pivotal role in RFM analysis, aiding in customer segmentation based on their engagement patterns and overall value.

Recency

We introduced the 'Recency' column to calculate the duration in days between a customer's 'Date of Last Visit' and the fixed date '2019-10-31', serving as our dataset's final reference point. This measure provides us with the number of days since the customer's last interaction, potentially indicating whether they remain active or have disengaged from our business.

Frequency

This feature computes the interaction frequency a customer has had with the sports facility by dividing the 'Number of Frequencies' by the 'Enrollment Duration in Semesters'. It aims to quantify the customer's engagement frequency within each semester of their enrollment. Higher values indicate increased interaction frequency, reflecting a more actively engaged customer.

Monetary:

The 'Monetary' feature evaluates a customer's 'Lifetime Value' over their 'Enrollment Duration in Semesters'. It calculates the average monetary contribution per semester by dividing the customer's 'Lifetime Value' (or spending) by their enrollment duration in semesters. Elevated values may suggest customers who consistently make monetary contributions throughout their enrollment, signifying their sustained financial impact on the business.

Data Visualization and Analysis

This stage was pivotal for understanding the dataset's distribution. Initially, a comprehensive descriptive analysis unveiled the mean values for each variable, providing a foundational understanding of expected values within the dataset. Subsequently, variance calculations were performed across all variables to gauge their informational significance. Variables exhibiting a variance below the threshold of 0.01—specifically athletics activities, nature activities, dance activities, and other activities—ended up being removed.

Following this, a correlation heatmap (fig. 1) was constructed to detect relationships between variables. While numerous variables exhibited high correlation, a threshold of 0.95 correlation was applied to eliminate highly correlated features. Consequently, only Enrollment Duration days and months were removed, retaining Enrollment Duration Semesters due to their perfect correlation of 1, as they represented the same variable divided by a constant.

The next step involved dividing features into numeric and non-numeric types to streamline the ongoing analysis. Initially, histograms (fig. 2) were plotted for numeric features to visualize their distribution and identify potential outliers. However, upon reassessment, it was decided that box plots (fig. 3) would be more appropriate for outlier identification and subsequent treatment.

Additionally, a profile report was generated at the beginning of the project to explore potential variables for clustering, offering insights and ideas for subsequent clustering analysis.

Handling Outliers

Initially, we attempted to utilize the IQR (Interquartile Range) method to address outliers, but this approach resulted in the removal of a considerable portion of the dataset. Consequently, we opted for a manual approach, leveraging box plots from our data visualizations to establish specific thresholds for outlier removal. The targeted features subjected to outlier removal were continuous and numeric, distinguishing them from discrete numerical features, which bear categorical characteristics. Thus, it was deemed inappropriate to remove values from the latter. Post-filtering, we retained around 99.21% of the original dataset, indicating the removal of a minor portion. This process aimed to ensure a more balanced dataset by eliminating extreme values, thereby enhancing the precision of subsequent analyses.

Customer Segmentation

RFM Analysis

The RFM model allows to segment XYZ sports company customers based on their past transaction behavior. It analyses customer behaviour based on three essential factors: recency, frequency and monetary value.

The first step to perform RFM analysis was to gather the necessary information from the dataset. For that we created the features tailored to this analysis “Recency”, “Frequency” and “Monetary”.

Each customer was then assigned a score for each of the variables. The score ranges from 5, the highest (most recent, most frequent, with the most amount spent), to 1, the lowest possible score.

With the summatory of the individual scores, we derive the overall RFM Score of each customer, providing insights into their value to the business. This final score allows us to categorize clients into "High Value," "Medium Value," and "Low Value" from a business standpoint.

The business takeaways from the RFM Analysis are the following:

- Overall, XYZ sports company gym has historically targeted “Medium Value” clients. However, when looking at the current customers, the majority are “High Value" clients (fig. 4).
- Those under the age of 18 constitute 8.5% of the overall "High Value" customers and 6% of the "Medium Value" customers. This demographic is particularly crucial to target, given their potential for long-term customer relationships, owing to their youth.
- There is no discernible direct correlation (fig. 5) observed among the RFM metrics. Customer behavior in terms of recency, frequency, and monetary spending may not move uniformly, and adjustments in marketing strategies might need to be tailored to the specific dynamics of each metric independently.

Clustering analysis

During our project, we explored various approaches to determine the variables suitable for clustering. Initially, we constructed a cluster solely based on customer activities, aiming to pinpoint the most prevalent activity among our customer base. Employing k-means and U-map for visualization due to the high dimensionality, we visualized this cluster. However, upon reconsideration, we realized that this approach essentially resembled a frequency graph of all activities, albeit more complex to interpret visually. Consequently, we decided to omit this approach from our analysis.

Exploring further, we attempted to establish alternative behavioral clusters distinct from the activity-based one. Specifically, we delved into a demographic cluster using variables such as Lifetime Value, Dropout, NumberOfRenewals, HasReferences, and NumberOfReferences. The objective was to create a cluster that could shed light on customer value, loyalty, and satisfaction. However, we encountered a challenge: although these features are categorized as numeric in their data type, some possess a limited range of outcomes, such as NumberOfReferences, while others are binary, like HasReferences. Considering these limitations, we opted to discard this clustering perspective and redirect our focus towards the next two approaches.

- Demographic Perspective

As a way to detect potential tendencies in the demographics of the customer base, our first approach was to attempt to find clusters using the features Age, Gender and Income.

The Demographic Perspective underwent significant refinements during the project's evolution. Initially, it encompassed gender as a feature, aiming to discern income and age disparities by gender. However, we opted to remove this feature due to its contribution to increased graphical complexity—a challenge stemming from its addition resulting in higher dimensionality (three dimensions). This adjustment streamlined our analysis, enhancing its comprehensibility.

This exploration uncovered a notable inconsistency: some children exhibited remarkably high incomes. Subsequently, this discovery prompted us to revisit the preprocessing stage, implementing adjustments to rectify these inconsistencies.

In the actual clustering phase, we employed the elbow method. This method involves computing the Within-Cluster-Sum of Squared Errors (WSS) across varying numbers of clusters (k) and selecting the k at which the WSS begins to decrease. The result showed 3 clusters as the optimal solution (fig.7). Subsequently, visualizing these clusters provided us with the following output:

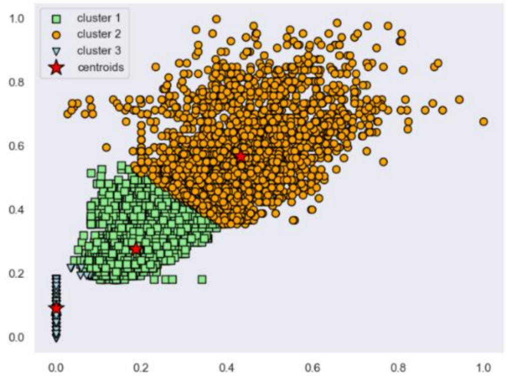


Figure 6 – Demographic Clusters

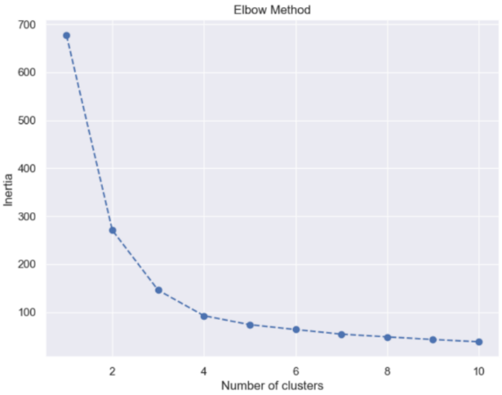


Figure 7 – Elbow Method for Demographic Clusters

Through this visualization (fig. 6) we have distinctly delineated three clusters. After this we created tables featuring the mean values of each, characterizing them:

Cluster 0: This cluster comprises predominantly younger individuals, averaging around 24 years old, with moderate income and relatively limited engagement in activities. Their spending habits and class participation are lower, resulting in higher dropout rates compared to the other clusters.

Despite a relatively high engagement recency, they possess fewer references, a finding that intrigued us. This cluster stood as our most substantial group, comprising a total of 9383 individuals.

Cluster 1: Composed of middle-aged individuals with an average age of 49, this cluster demonstrates higher incomes and moderate engagement across various activities. They tend to exhibit greater spending, higher class attendance, and a slightly lower dropout rate. While their engagement recency is lower than Cluster 0, they also display a comparatively low number of references. Within this cluster, the count of individuals was smaller, specifically totalling 2606.

Cluster 2: This cluster predominantly comprises children, averaging around 8 years old, belonging to lower income brackets. Despite their young age, they actively engage in activities, showcasing higher spending habits and relatively lower dropout rates. While their engagement recency is lower than Clusters 0 and 1, they exhibit the highest number of references among the three groups. This cluster represented our smallest group, encompassing a total of 2389 individuals.

Following the initial clustering process, we attempted to visualize the clusters by excluding individuals without income. The aim was to observe whether segmenting among individuals with income would result in more refined clusters. Employing the same procedure as previously done, we noticed that the third cluster formed after this refinement contained a notably smaller number of individuals. This observation led us to conclude that pursuing this approach wouldn't provide meaningful segmentation, so we just kept it as it was.

- <u>Purchase Behavior Perspective</u>

Following the RFM analysis, we thought it would be interesting to cluster the customer into groups using those same features. This perspective could give us insight into customer segments with similar purchasing behaviors.

The clusters were found using the K-Means algorithm. To find the optimal K number of clusters, we used two methods:

1) The Elbow method

The objective is to identify an "elbow" point in the curve, signifying the optimal number of clusters for clustering. In this case, the optimal number of clusters is between 2 and 3 (fig.8).

2) The Silhouette Coefficient

The silhouette coefficient is the average value of all individual silhouette coefficients calculated for points in the dataset. This metric indicates the extent to which each point aligns with its assigned cluster, providing insight into the overall quality of the clustering.

The graph (fig. 9) shows that the best score is achieved with K=3, which ended being the final

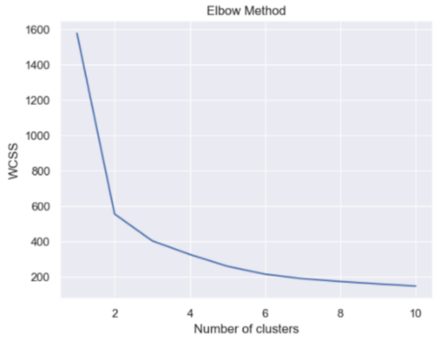


Figure 8 – Elbow Method for Purchase Behavior Clusters

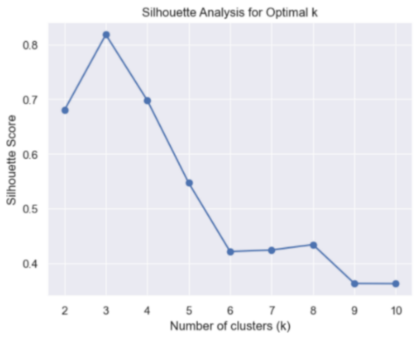


Figure 9- Silhouette Analysis for Purchase Behavior Clusters

decision for the number of clusters.

Resulting Clusters:

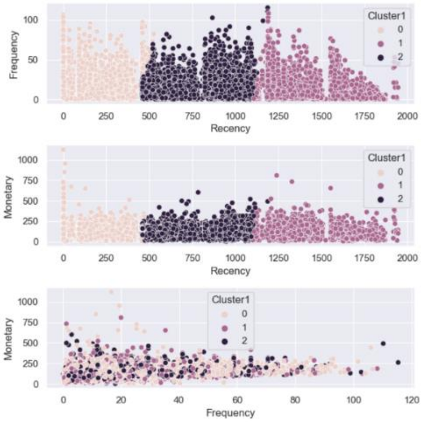


Figure 11 – Purchase Behavior Clusters

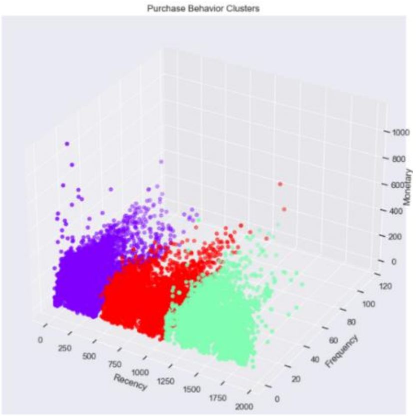


Figure 10 – Purchase Behavior Clusters

The primary variation among the three clusters is predominantly observed in the recency factor. The frequency and monetary aspects show relatively consistent patterns across the clusters (fig 10 & 11).

This aligns with the nature of the business. The gym operates mostly in a subscription-based business model, resulting in a relatively uniform spending pattern per semester for most customers. Also, the distribution of frequency, or how often customers visit the gym in a semester, appears to be relatively consistent among the clusters regardless of the recency.

Cluster 0 comprises the most recent customers, including 96% of the current customer base. Cluster 1 represents customers with a medium recency, and Cluster 2 consists of the least recent customers.

From these clusters we could draw the following segments:

- Recent Customer Engagement: Target customers in Cluster 0 (most recent) for specific engagement initiatives, promotions, or loyalty programs.
- Medium-Recency Engagement: Implement strategies tailored to customers in Cluster 1 (medium recency) to maintain their engagement and potentially encourage more frequent interactions.
- Reactivation Strategies: Develop reactivation campaigns for customers in Cluster 2 (least recent) to encourage them to re-engage with the business. Consider offering special incentives or promotions to win them back.

- Value-based perspective

In this approach, customers are segmented not only by demographic characteristics (Age and Income) but also by their potential profitability to the business, merging the two previous perspectives. The RFM Score is used to integrate the profitability element into the clustering process.

The algorithm employed was K-Means with the same methods to determine the optimal number of clusters:

1) Elbow method: showed an optimal k between 3 and 4, without a clear elbow though.

2) Silhouette Score: the best silhouette score is clearly obtained when the number of clusters is 3, which was our final decision.

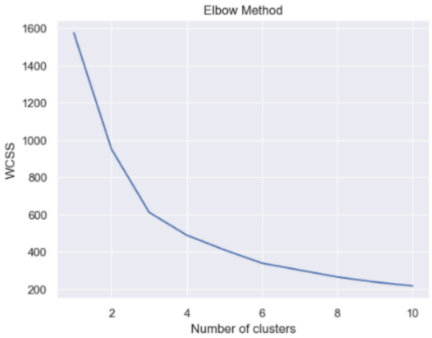


Figure 12 – Elbow Method for Value Based Clusters

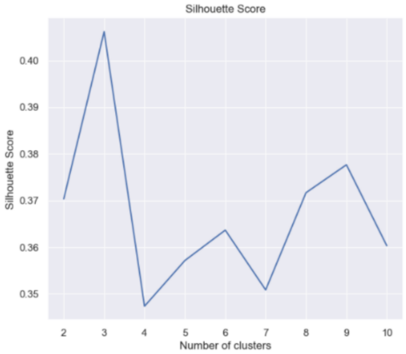


Figure 13 -silhouette Score for Value Based Clusters

Resulting clusters:

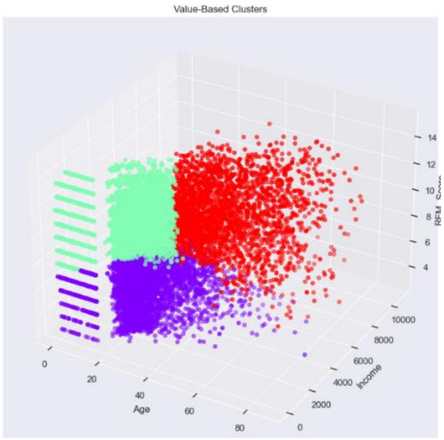


Figure 24 - Value Based Clusters

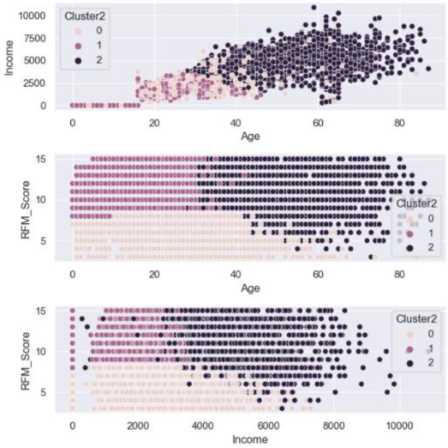


Figure 15 - Value Based Clusters

Cluster 0 is defined by individuals with modest to moderate incomes and a low RFM score, suggesting lower engagement and spending behavior. This group likely comprises customers with relatively restrained involvement in expenditures.

In contrast, Cluster 1 includes individuals with low incomes but notable high RFM scores. This cluster is characterized by younger customers who, despite their lower income levels, showcase heightened engagement and spending patterns. These customers demonstrate a noteworthy commitment to interactions with the business.

Cluster 2 consists of individuals with the highest income, predominantly from an older demographic. This cluster represents customers with substantial financial resources and potentially long-standing relationships with the business, as evidenced by a high RFM score. The members of this cluster are likely to be valuable patrons who have established enduring connections with the company.

Based on these clusters, we can perform the following segmentation:

- Low-Engagement Segment: Target customers in Cluster 0 (low/medium income, low RFM) for engagement initiatives. Focus on increasing their frequency and monetary contributions to improve their overall value.

- High-Value, Younger Segment: Concentrate on customers in Cluster 1 (low income, high RFM) who exhibit both high engagement and younger demographics. Tailor marketing efforts to maintain and enhance their loyalty.
- High-Value, Established Segment: Prioritize customers in Cluster 2 (highest income, older) who have a high RFM score. These customers represent a valuable, established segment, and efforts can be directed toward maintaining satisfaction and encouraging continued loyalty.

It's noteworthy that the majority of the current gym members fall into Cluster 1 (approximately 58%), emphasizing the importance of customizing marketing strategies for the younger demographic. However, Cluster 2 accounts for 26% of the existing clientele and should be a primary business focus for growth, given their higher RFM Score, indicating greater potential profitability.

Conclusion

From merging the insights from both demographic analysis and RFM segmentation, we are going to provide some recommendations for the sports facility strategic plan.

To optimize operational efficiency and customer satisfaction, we propose the preservation and potential expansion of activities that are preferred by customers, namely Water Activities, Fitness Activities, Combat Activities, and Team Activities (fig. 16). Simultaneously, a critical examination of less popular offerings may allow for cost savings. Moreover, the demographic lens reveals a particular emphasis on tailoring activities to age groups, with a notable focus on the younger demographic, identified as medium to high value customers. This underscores the importance of aligning offerings with the preferences and lifestyles of this segment.

Also, recognizing the diverse income levels within our customer base, we advocate for the implementation of subscription plans tailored to different economic brackets. Premium plans are suggested to attract and retain older and wealthier clientele, while the introduction of affordable options aims to capture and engage the younger and less affluent demographic. This nuanced approach to service and pricing aligns with the identified clusters and ensures a targeted and

effective strategy for sustained growth and customer retention in the dynamic fitness industry.

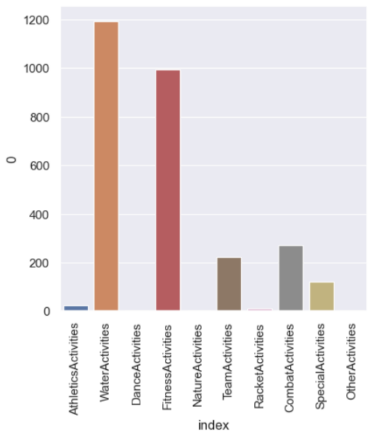


Figure 16 – Activity Frequency of Current Customers

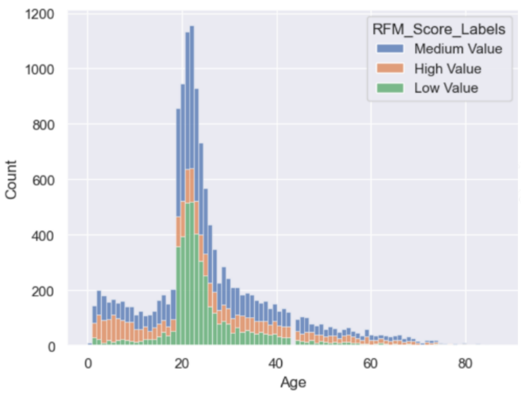


Figure 17 – Merged Age Distribution/ RFM Score

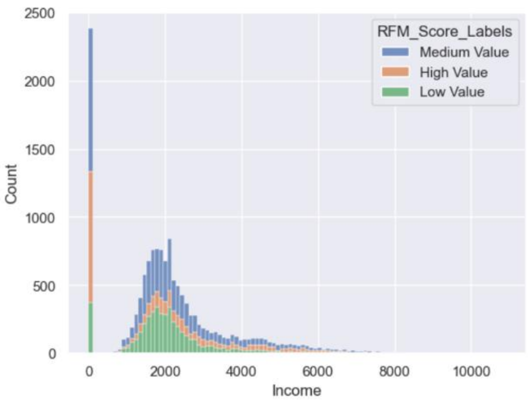


Figure 18- Merged Income Distribution/ RFM Score

Annex

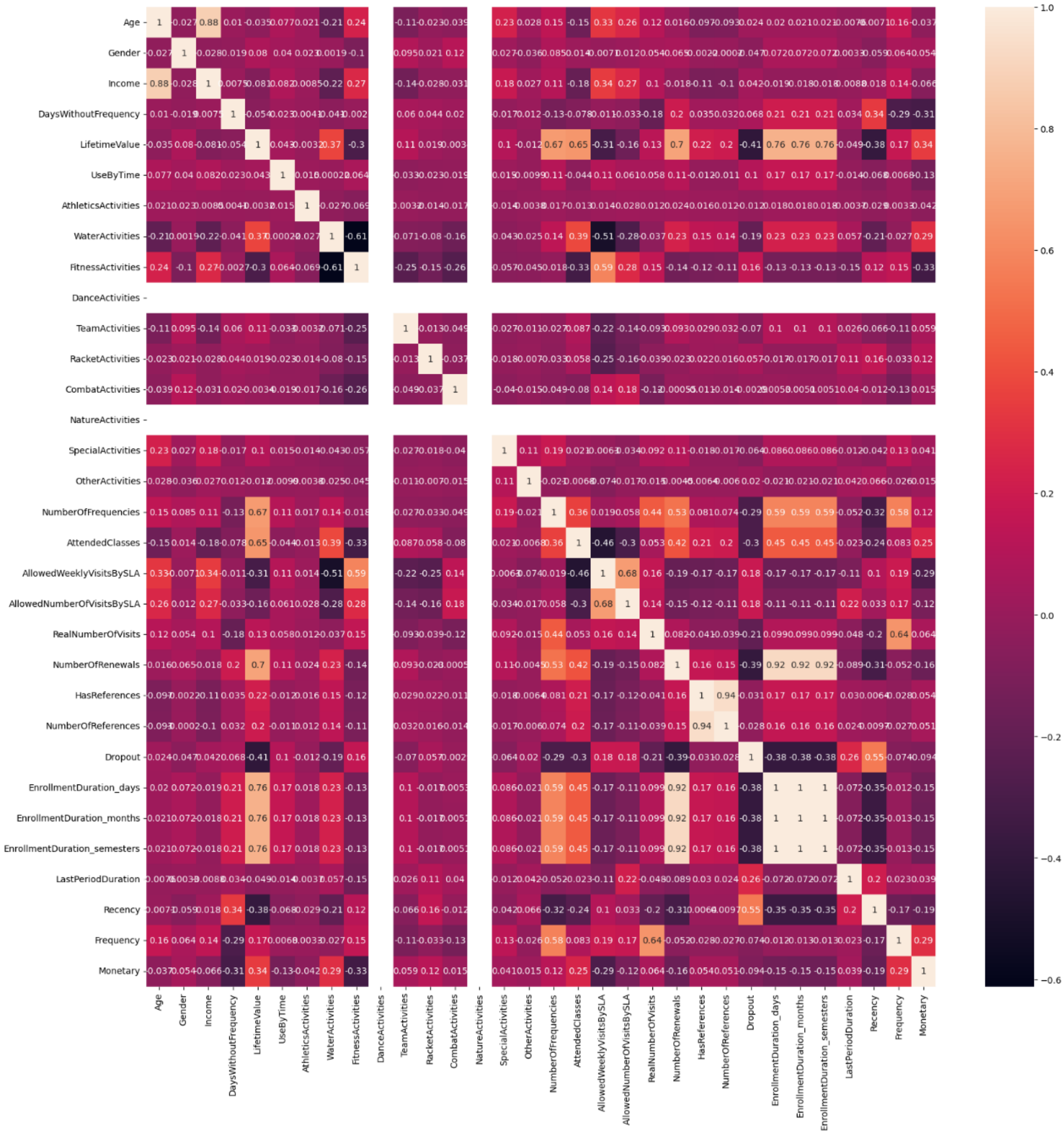
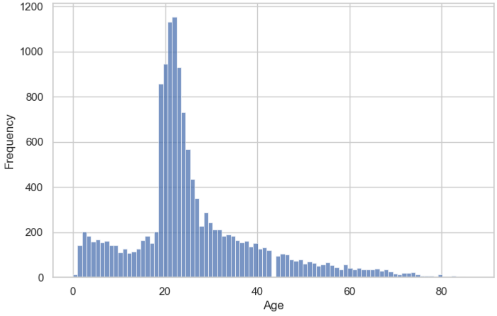
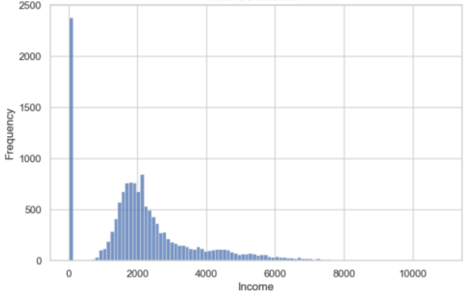


Figure 3 - Correlation Heatmap

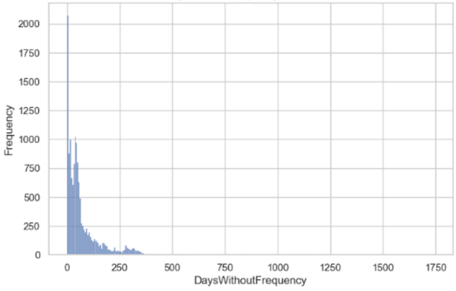
Age Distribution



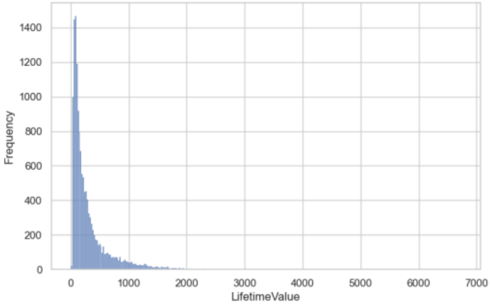
Income Distribution



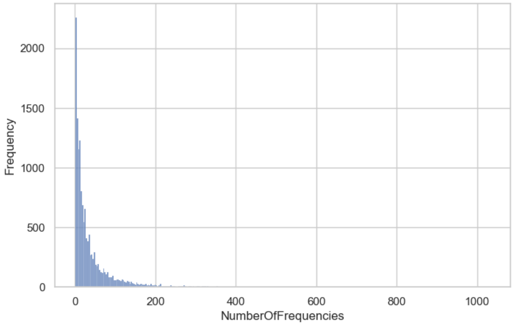
DaysWithoutFrequency Distribution



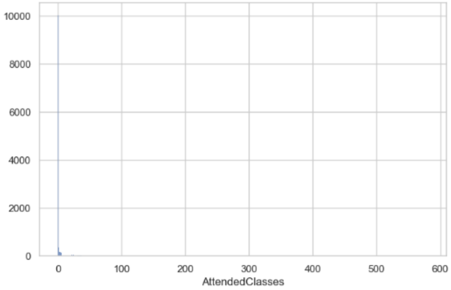
LifetimeValue Distribution



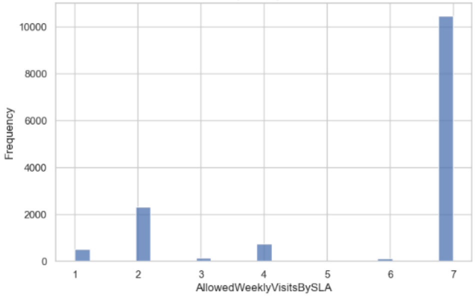
NumberOfFrequencies Distribution



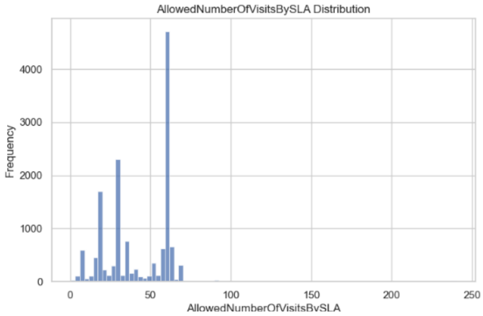
AttendedClasses Distribution



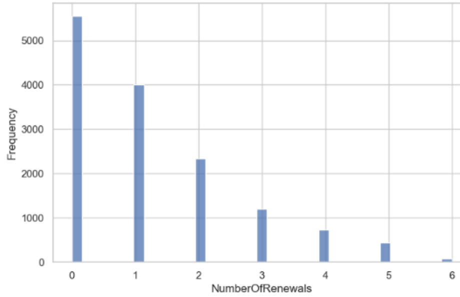
AllowedWeeklyVisitsBySLA Distribution



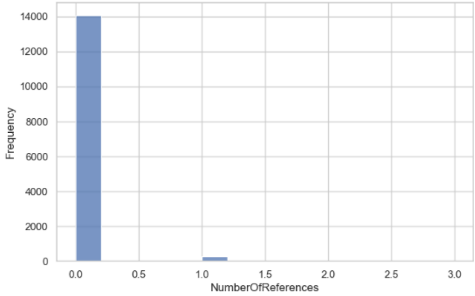
AllowedNumberOfVisitsBySLA Distribution



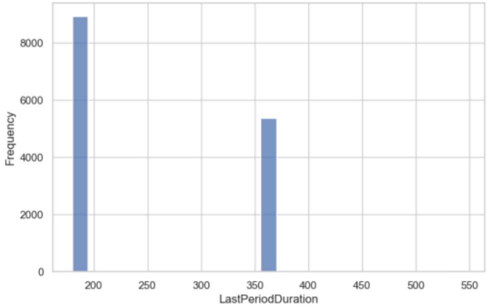
NumberOfRenewals Distribution



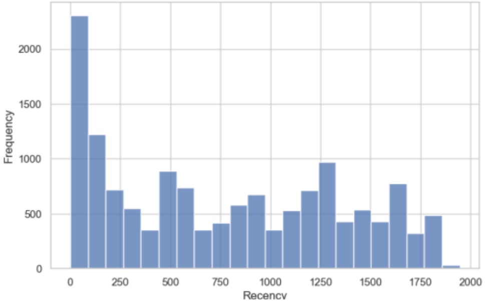
NumberOfReferences Distribution



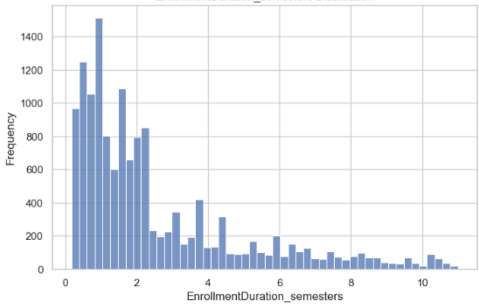
LastPeriodDuration Distribution



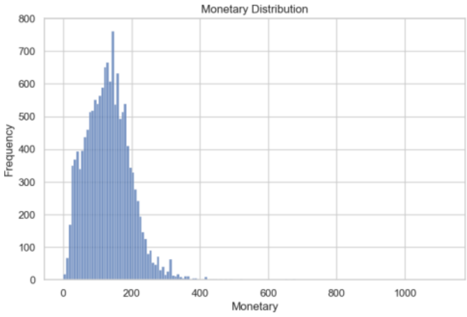
Recency Distribution



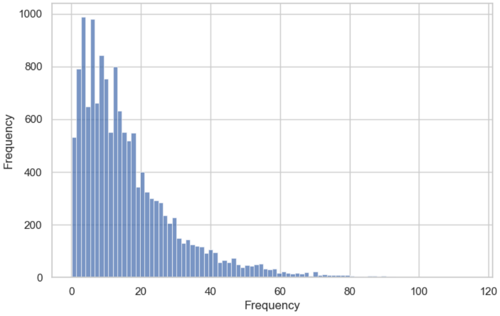
EnrollmentDuration_semesters Distribution



Monetary Distribution



Frequency Distribution



RealNumberOfVisits Distribution

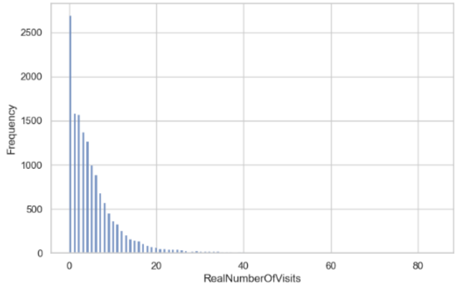


Figure 4 - Features Distributions

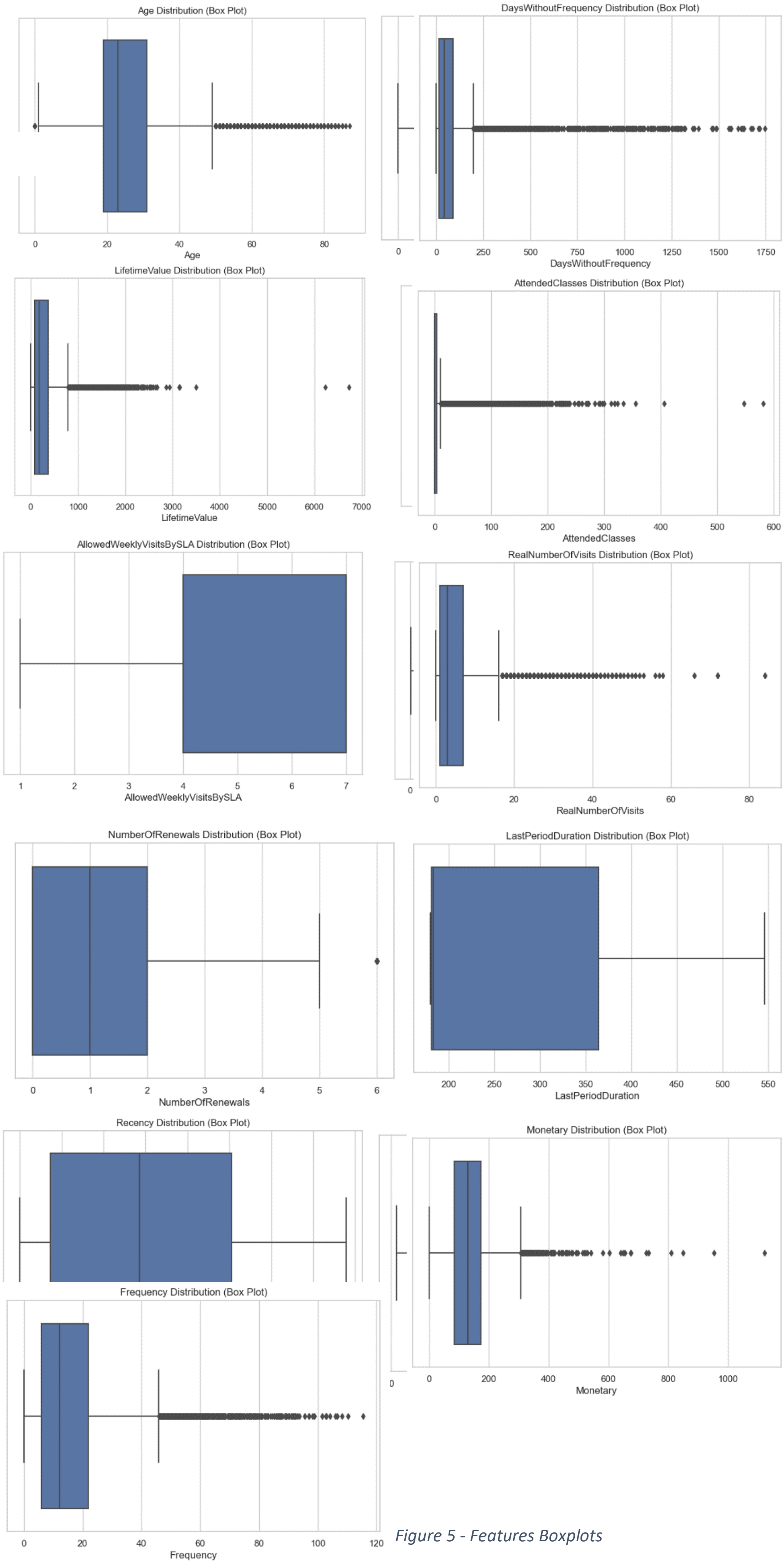


Figure 5 - Features Boxplots

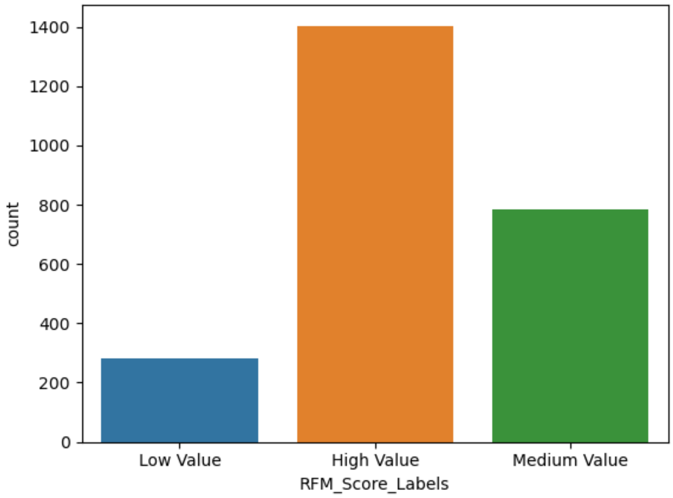


Figure 6 - RFM Labels of Current Customers

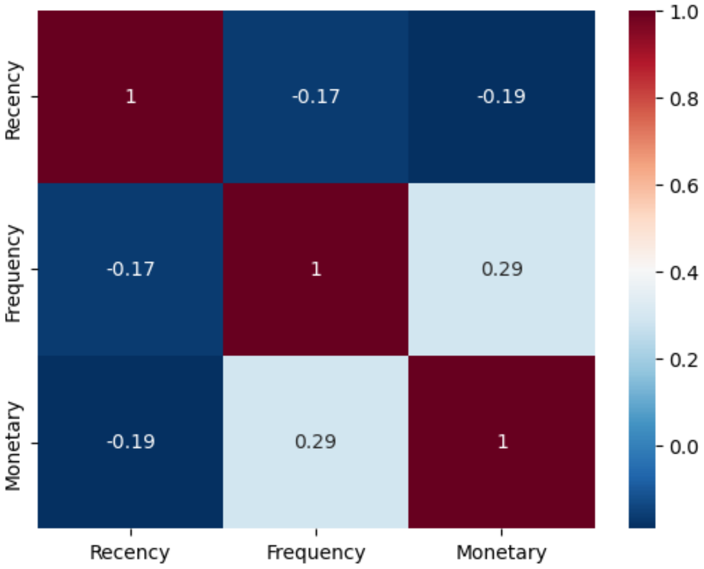


Figure 7- RFM Heatmap

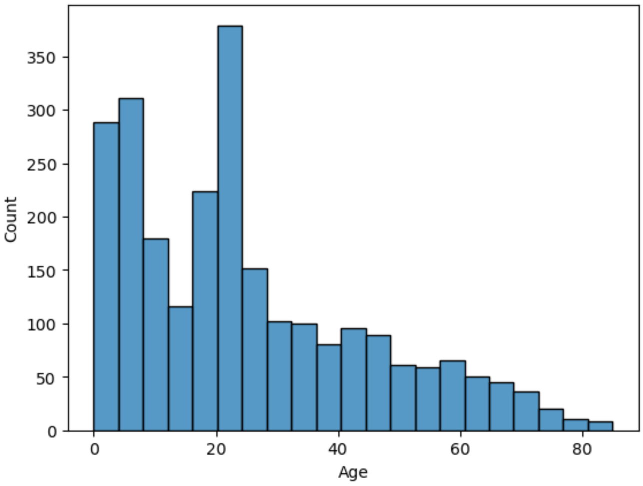


Figure 8 - Age Distribution of Current Customers