NOVA
IMS
Information Management School

Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Customer Segmentation of XYZ Sports Company

Group 50

Catarina Silva, number: 20230368

Francisco Castro, number: 20230992

Pedro Catarro, number: 20230463

January, 2024

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

Abstract

This project is based on a question that all companies face at a certain stage of their life cycle, "Who are our customers?". Understanding who their customers are is crucial for companies because, by answering this question, they can determine how to satisfy these customers, keep them loyal to the company, and also attract new ones. Our goal with this project is to segment customers and, based on these segments, understand their value, demographic data, and the types of sports these customers prefer to participate in.

To create the clusters, we began by exploring the data and understanding our variables. Next, we preprocessed our data, by doing some tasks such as searching for duplicates, treating missing values and outliers, performed feature engineering, and other important steps. Following that, we initiated the clustering process by deciding on the perspectives we want to use. We then calculated the optimal number of clusters for each perspective and determined the most suitable clustering methods. Later, we merged the clusters from all perspectives. In the end, this study showed us that we will end up with a final number of clusters of four.

We will conclude the project with an analysis of each cluster and understand the clients they represent. Then we provide suggestions for the business and marketing strategies, such as promotions and the increase the number of classes of a particular activity. Finally, we will end with a brief conclusion.

Key words: Customer; Cluster;

INDEX

1. Introduction

Knowing who their customers are is crucial for companies. This helps them align with the needs of existing customers and acquire new ones. Market segmentation helps identify these customers geographically, demographically, psychologically, and behaviorally. Understanding the differences between these segments allows organizations to make better strategic choices about opportunities, product definition, positioning, promotions, pricing, and target marketing.

In this project, we were asked to divide the customers of XYZ Sports Company, a well-established fitness facility serving the community for several years, into segments. The goal is to gain insights into the customers, draw conclusions on how to better serve them, and understand the types of sports they prefer to participate in.

For the execution of this project, we were provided with a dataset containing information about customers collected between June 1st, 2014, and October 31st, 2019.

Throughout this report, we will explain all the steps taken to reach the final result, the segmentation of customers into what we believe are the most accurate. We will also suggest business strategies and marketing approaches based on our findings.

2. Data Exploration

To begin our data mining project, we first set clear goals and figured out the best ways to achieve them. Once we had a plan, our next big task was to really get to know our data. This means we looked at the dataset's size, checked what each piece of information was called, and made sure our data was organized in a way that makes it easy to work with. Understanding our data well is super important for making good decisions throughout the project. To commence this exploration, we began by importing the requisite libraries, ensuring that we had the necessary tools at our disposal for robust data analysis. Following this, we proceeded to import the dataset, laying the foundation for the subsequent stages of our project.

One of the primary tasks in our data exploration journey was to visualize the dataset's fundamental characteristics. We initiated this process by assessing the dataset's scale, delving into the number of observations it encapsulated. Understanding the size of our dataset is foundational, providing insights into the volume of data that forms the basis of our analyses.

Moving forward, we delved into the specifics of the dataset by scrutinizing the variable names and their associated data types. This examination unveiled the structure and composition of the dataset, enlightening the types of information captured by each variable. Additionally, recognizing the importance of a unique identifier, we strategically set the "ID" column as the index. This not only streamlined the data but also facilitated more efficient referencing and retrieval of information during subsequent analyses.

Our approach to data exploration is driven by the principle of extracting meaningful insights while minimizing unnecessary complexity. By visualizing the dataset's size, understanding variable characteristics, and optimizing the dataset structure, we laid a robust foundation for the subsequent stages of our data mining project. This meticulous exploration sets the stage for more in-depth analyses, ensuring that our insights are grounded in a comprehensive understanding of the underlying data.

3. Data preprocessing

3.1. Duplicates

In our data mining analysis, we performed a check for duplicate values in the dataset using the code '**data.duplicated().sum()'**. The resulting output indicated a count of 0, signifying that there are no duplicate values present in the dataset. This absence of duplicates ensures the integrity and uniqueness of our data, laying a solid foundation for reliable analyses and meaningful insights throughout the project.

3.2. Missing values

In managing missing data within our dataset, we implemented various strategies for imputation and removal. (Annex 1)

Addressing missing values in the 'Income' column, we chose the K-Nearest Neighbors (KNN) imputation technique. Utilizing scikit-learn's KNNImputer with 'n_neighbors=5' and 'weights="uniform,"' we imputed missing values based on the values of their five nearest neighbors. This approach ensures a data-driven imputation, leveraging the surrounding context to estimate missing values more accurately.

For categorical columns such as 'HasReferences' and 'AllowedWeeklyVisitsBySLA' a straightforward strategy was employed. Missing entries in the 'HasReferences' column were filled with the string 'False,' and missing values in the 'AllowedWeeklyVisitsBySLA' column were replaced with the string '0.'

Recognizing the importance of balancing data completeness and robustness, we implemented a threshold-based approach to remove rows with missing values. Rows containing missing values in columns representing less than 5% of the total data were dropped. This decision aimed to preserve a significant portion of the dataset while minimizing the impact of missing values on the overall dataset.

3.3. Feature engineering

In the process of preparing our dataset for analysis, we incorporated several feature engineering techniques to derive meaningful insights and simplify the data. One significant aspect involved the calculation of the 'EnrollmentTime' feature, which quantified the duration of enrollment for each individual by computing the time difference in days between the 'EnrollmentStart' and 'EnrollmentFinish' timestamps. This information was then converted to an integer, and unnecessary columns related to enrollment timestamps were dropped to streamline the dataset.

Another feature, 'DaysSinceLastVisit,' was engineered to measure the time elapsed since the last visit until a fixed reference date. Utilizing the 'DateLastVisit' column, the code computed the number of days between 'DateLastVisit' and a predefined reference date ('questionary_date'). The resulting data was stored in a new column, and the original 'DateLastVisit' column was subsequently dropped for simplicity.

The feature engineering process also addressed various activities, such as athletics, water sports, fitness, team sports, racket sports, combat, special activities, and others. The 'TotalActivities' feature was introduced to capture the cumulative count of activities for each individual. Additionally, the code calculated the count of each activity type, although the results were not stored or printed for further analysis.

Further refinement included the generation of the 'NActivities' feature by assigning the 'TotalActivities' column from the activity-related dataframe to the main dataset (Annex 2). Additionally, the 'AvgLifeTimeValue' feature was engineered by dividing the 'LifetimeValue' by the duration of enrollment in years, offering insights into the average monetary value generated by individuals over their enrollment period. To handle potential division by zero cases, any instances of 'np.inf' (infinity) resulting from the calculation were replaced with 0. Then to get a better presentation of our features, we have renamed all of them.

Overall, these feature engineering steps contributed to a more insightful and streamlined dataset, setting the stage for effective analysis.

3.4. Feature selection

The code focuses on preparing the dataset for analysis through various feature selection methods. It starts by visualizing activity counts based on gender (Annex 3), then provides an overall view of activity participation. Unnecessary columns are removed to simplify the dataset. Categorical distribution plots offer insights into participant distribution across different activity categories. Correlation analysis explores relationships between enrollment, income, and dropout rates. Age group analysis considers age as a categorical variable. Correlation matrix heatmaps identify strong correlations influencing feature selection (Annex 4). Redundant or less significant columns are strategically removed, aiming to improve efficiency. A final correlation matrix heatmap ensures retention of the most relevant features for subsequent analysis. These steps collectively contribute to refining the dataset and enhancing the effectiveness of future analyses.

3.5. Outliers

Concerning outliers, we created different plots to get a clearer picture of them and figure out which ones to eliminate (Annex 5). In this context, the code employs various visualizations, including categorical distribution plots and numeric variable analyses, to detect and handle outliers. By employing box plots and histograms, the code allows for a comprehensive exploration of the data's distribution, enabling the manual removal of outliers based on predefined criteria. We successfully retained approximately 98% of the data after removing outliers manually, preserving nearly all the dataset while eliminating the most extreme outliers.

We opted against further outlier elimination using DBSCAN because we were concerned that it might result in the loss of important information, especially considering that we had already addressed the most extreme outliers.

3.6. Data Normalization

For data normalization, we thoroughly assessed three distinct scaling methods—Robust Scaler, MinMax Scaler, and Standard Scaler—applied across a range of clustering algorithms. The primary goal was to understand the influence of various scaling approaches on both the performance of the algorithms and the resultant clusters.

For each scaling method, the analyses were conducted in separate notebooks, systematically evaluating the clustering algorithms and the clusters they generated. After an exhaustive examination, we determined that the Standard Scaler surpassed the other scaling methods for our dataset. This decision was grounded in the observation that the Standard Scaler produced clusters that were not only more diverse but also showcased improved characteristics compared to clusters obtained through alternative scaling methods.

The preference for the Standard Scaler stemmed from the intention to improve precision and distinctiveness of the clusters derived from the clustering algorithms. This choice aimed to align more effectively with the specific characteristics of our data, ensuring an optimized performance in the clustering analysis.

4. Segmentation and Clustering

4.1. Algorithm Testing for Clustering Evaluation

In the process of obtaining clusters for customer segmentation, we approached the task by dividing the features into three perspectives: demographic and value features, behavior features, and preference features. The demographic and value features included attributes such as 'Age,' 'AvgLifeTimeValue,' 'LifetimeValue,' and 'EnrollmentTime.' Behavior features encompassed metrics like 'DaysWithoutFrequency' 'NumberOfFrequencies' 'AttendedClasses', 'AllowedWeeklyVisits', 'AllowedNVisits', 'NVisits', 'NRenewals', 'NReferences', 'DaysSinceLastVisit' and 'NActivities' , while preference features represented the customer's activities, such as 'Athletics,' 'Water,' 'Fitness,' and ‘Other’.

For cluster evaluation, we decided to exclude three binary (non-metric) features ('UseByTime,' 'Dropout,' 'Gender_binary'). The decision to exclude these three binary features during cluster evaluation was driven by the nature of the analysis. These features may not have provided meaningful insights into customer segmentation based on their inherent binary nature, as their values are restricted to two categories and would lead to compatibility issues with some clustering algorithms. Excluding them allowed for a more focused analysis on metric features that could better capture the nuances of customer behavior and value, thereby enhancing the effectiveness of the clustering algorithms.

On the other hand, in the preference perspective, other binary features were retained as they played a crucial role in understanding customer interests in various activities. Unlike 'UseByTime,' 'Dropout,' and 'Gender_binary,' the binary features in the preference perspective were relevant to the analysis, as they directly contributed to identifying the specific activities customers were more inclined towards. Retaining these binary features enriched the analysis by providing insights into customer preferences and behaviors related to their chosen activities, contributing to a more nuanced and targeted segmentation approach.

After defining the three perspectives for analysis, we proceeded to test clustering algorithms. We commenced our analysis by directing our attention to the categorical features, specifically those pertaining to preferences within the dataset. In addressing this specific set of features, we opted for the utilization of the K-modes algorithm. K-modes is well-suited for categorical data and was employed to partition the data into three distinct clusters.

The decision to exclusively use the K-modes algorithm for the preference features was based on the favorable results it provided. In contrast, when we experimented with the DBSCAN algorithm using the Jaccard's distance metric, the results were less satisfactory. The DBSCAN algorithm, in this context, produced only one cluster. Due to this discrepancy and the inadequate performance of DBSCAN for our specific dataset and features, we decided to discard its results for further analysis.

The remaining two perspectives, consisting of metric features only, underwent testing with various clustering algorithms, including k-means, hierarchical clustering, self-organizing maps (SOM) with k-means, SOM with hierarchical, Meanshift, DBSCAN.

Following an exhaustive examination of various clustering algorithms applied to both perspectives, a meticulous assessment of each algorithm's performance in terms of cluster distribution was conducted. Through a detailed analysis, we identified and subsequently selected the algorithms that demonstrated optimal results for each respective perspective. This selection process was characterized by a thorough consideration of the unique characteristics and requirements inherent to each perspective, ensuring that the chosen algorithms were well-suited for the nuances of the data in question.

Following this comprehensive assessment of various clustering algorithms, it was determined that the most effective clustering algorithm for both the demographic and value perspective, as well as the behavior perspective, was a combination of the self-organizing map (SOM) and k-means. This hybrid approach demonstrated superior performance, yielding clusters that were not only well-defined but also reflective of the nuanced characteristics within each perspective.

With the optimal clustering algorithms identified for each perspective, the subsequent step involved the amalgamation of labels obtained from the chosen algorithms across the three perspectives. This process aimed to integrate the information derived from diverse clustering techniques, enhancing the overall accuracy and richness of the final segmentation. To achieve this integration, hierarchical clustering was employed, leveraging its ability to systematically merge labels from the demographic and value perspective, the behavior perspective, and the preference perspective.

In our pursuit of identifying the most suitable clustering approach, we explored alternative methodologies to potentially yield distinct clusters. Despite experimenting with various methods, including attempts to merge only demographic and value perspectives with behavioral perspectives separately (without the preference perspective), the outcomes demonstrated striking similarities. Consequently, we determined that amalgamating all three perspectives—demographic and value, behavioral and preference—proved to be the optimal strategy for achieving a thorough and comprehensive understanding of consumer behavior. This comprehensive integration of perspectives is expected to offer a complete understanding of consumer behavior, enhancing the depth of our analysis for better-informed insights.

In conclusion, the final outcome yielded four clusters as the ultimate result, each characterized by the following respective counts: Cluster 1 with 3058 instances, Cluster 3 with 7747 instances, Cluster 2 with 1987 instances, and Cluster 0 with 1474 instances.

4.2. Visual representation of clusters (T-SNE) and feature importance

Following the consolidation of labels from the merged perspectives, we undertook a visualization using t-SNE (t-distributed stochastic neighbor embedding)(fig1). This visualization technique was employed to gain a more intuitive and interpretable understanding of the spatial relationships between data points within the four clusters.

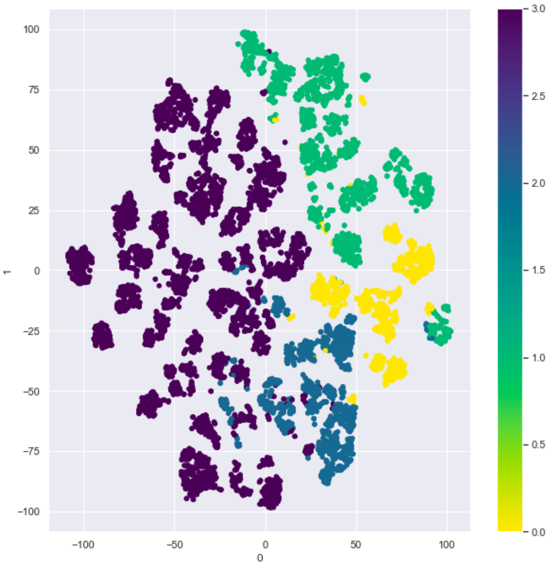


Fig 1 : T-SNE representation of the final clusters

Furthermore, in order to uncover the factors influencing cluster formation and comprehend the pivotal features shaping segmentation, we conducted an analysis of feature importance across all perspectives. This encompassed evaluating the influence of each feature on the overall clustering outcome. The outcome of this analysis revealed 'AllowedWeeklyVisits' 'EnrollmentTime' and 'AllowedNVisits' as the paramount features influencing cluster formation, with the respective values of 0.739, 0.597 and 0.394. The determination of feature importance underscored these specific attributes as noteworthy contributors to the discernible patterns and distinctions observed among the identified clusters.

Finally, following this comprehensive process, we employed a Decision Tree Classifier on the merged dataset (with the merged labels of the tree perspectives) to evaluate the predictive accuracy of our segmentation approach and confirm the effectiveness of the clusters derived from the intricate clustering process.

The results demonstrated an impressive predictive accuracy of 91.99% when the Decision Tree Classifier was applied to the merged dataset. This percentage signifies the estimated proportion of accurately classified customers by the model, underscoring the efficacy of our segmentation approach.

The high accuracy level reinforces the reliability and utility of the obtained clusters in capturing meaningful patterns within the customer data.

In this manner, the utilization of the Decision Tree Classifier served as a crucial validation step, providing quantitative evidence of the practical utility of the segmentation results derived from the earlier clustering methodology employed in the analysis.

5. Understand our clusters and suggestions of business applications for the findings and marketing approaches for each cluster.

Analyzing these 4 clusters (Clusters 0, Clusters 1, Clusters 2, and Clusters 3), we can extract various insights about the customers (Fig.2. and Fig.3.).

Cluster 0 is the smallest cluster but has the highest lifetime value. These customers visit the facilities with great frequently and, based on their contracted services, can access the installations more often. They are more likely to renew and have enrolled in more different activities, with a focus on water, fitness, and combat sports.

Cluster 1 is the second-largest cluster, comprehending individuals with an average age lower than the others. Customers in this cluster, on average, have fewer visits, which could be explained by the fact that they have fewer visits allowed to XYZ Sports Company facilities based on the services they've contracted.

Cluster 2, representing the majority of XYZ's customers (more than half). This majority represents many customers who did not renew despite not having a highest dropout rate. The graph indicates that customers in this cluster enroll in fewer diverse activities, with fitness, water, and combat being the most enrolled. Given that this cluster constitutes a significant portion of XYZ's customer base, efforts should be made to please and attract them.

Cluster 3 includes customers who have been enrolled for a longer time, but he also represented a lot of people who dropout. It is observed that customers in this cluster predominantly subscribed water and fitness activities.

Considering that customers in Cluster 2 is the cluster that represent most of our costumers we should have special attention to it. The cluster 2 costumers renew less on average, so the company could offer promotions after a period of enrollment ends to incentivize renewals. To attract more customers fitting into this cluster, the company could invest in advertising that promotes the sports most practiced by them, such as water activities, fitness activities and combat.

Finally, XYZ Sports should increase the variety of fitness and water activities to please their customers and serve them all while maintaining good relationships. Additionally, the company could offer trial classes for other activities to attract customers to try them. Addressing the issue of dropouts, the company could adopt progressive discounts and long-term benefits to encourage customers not to quit.

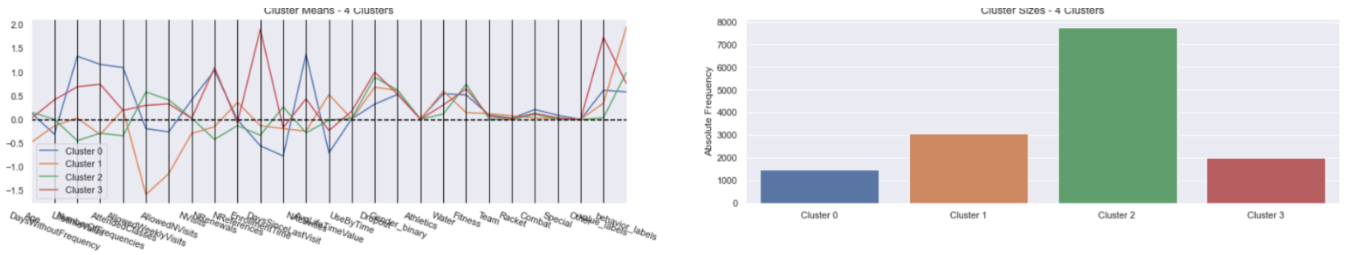


Fig 2 : Cluster profiling of the final clusters

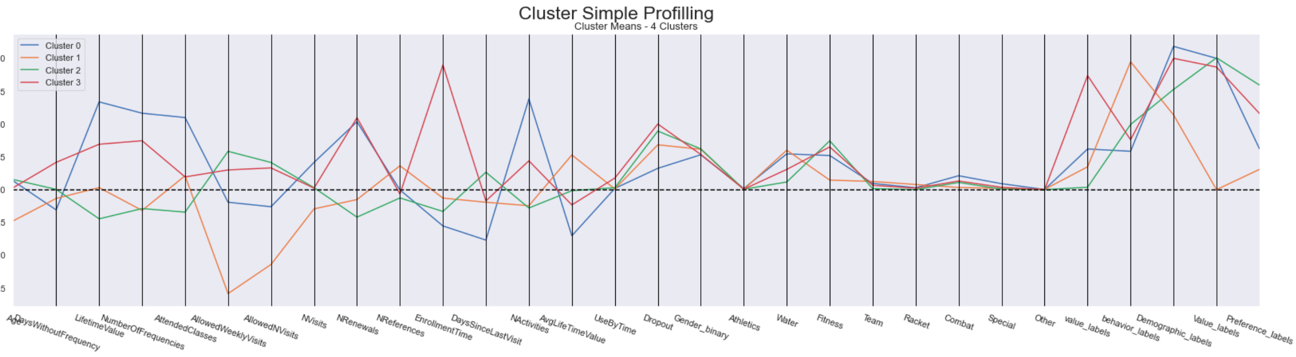


Fig 3 : Cluster profiling for final clusters and all features

6. Conclusion

We started this project by exploring the data and understanding our features. Next, we preprocessed our data, by searching for duplicates, treating missing values, doing features engineering, selecting the best features to our problem, treating outliers and doing the normalization of our data. Some steps of our data preprocessing phase were not carried out in a specific order, but as the needs were identified. This made our work more fluid and better to identify errors.

To cluster we choose 3 perspectives that represent demographic, behaviour of our costumer and the activities they prefer to participate. After testing other clustering algorithms, we conclude that K-modes was the best for the third perspective, and a combination of the self-organizing map (SOM) and k-means for the other two. Finally, we used hierarchical clustering to merge the clusters that we obtained, and we end up with 4 clusters.

It's important to note that this was the best approach that we tested, but with more time we could test more approach in other to find better cluster.

To conclude, we made a brief analysis of our cluster and found the cluster that represent most of our customers (Cluster 2). We discovered that these cluster customers don't renew their enrollment that much, and we suggest offering promotions after the end of the enrollment period. Then, we identified some trends as all cluster have a lot of people participating in fitness and water activities and suggest increasing the number of classes of that activates. Finally, we propose that the XYZ Sports company adopt progressive discounts and long-term benefits to reduce the number of dropouts.

7. References

[1] sklearn.preprocessing.StandardScaler. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

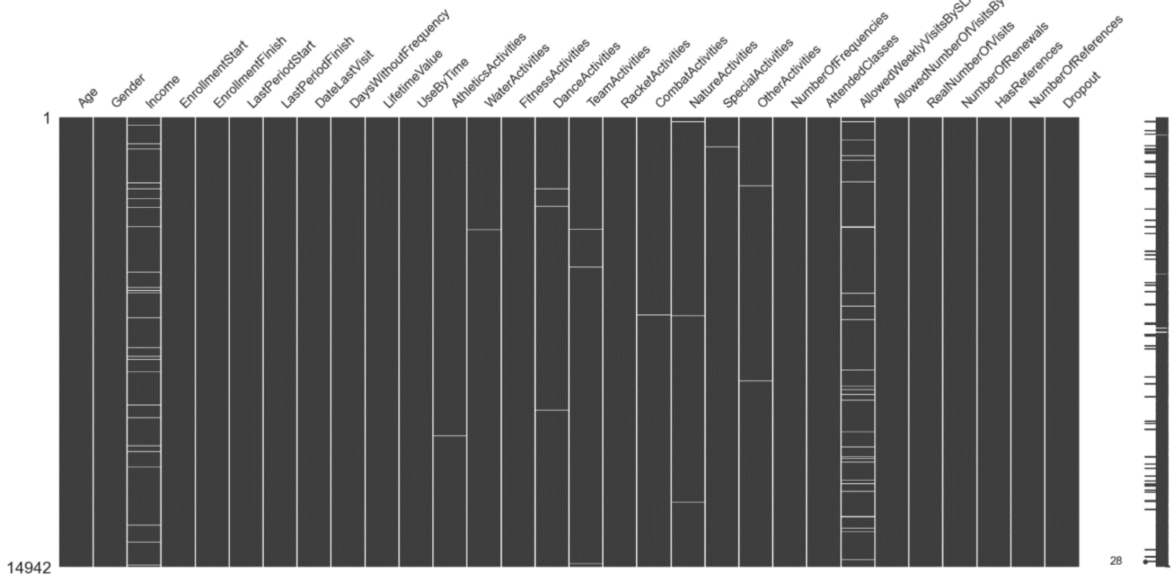
[2] sklearn.preprocessing.RobustScaler. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

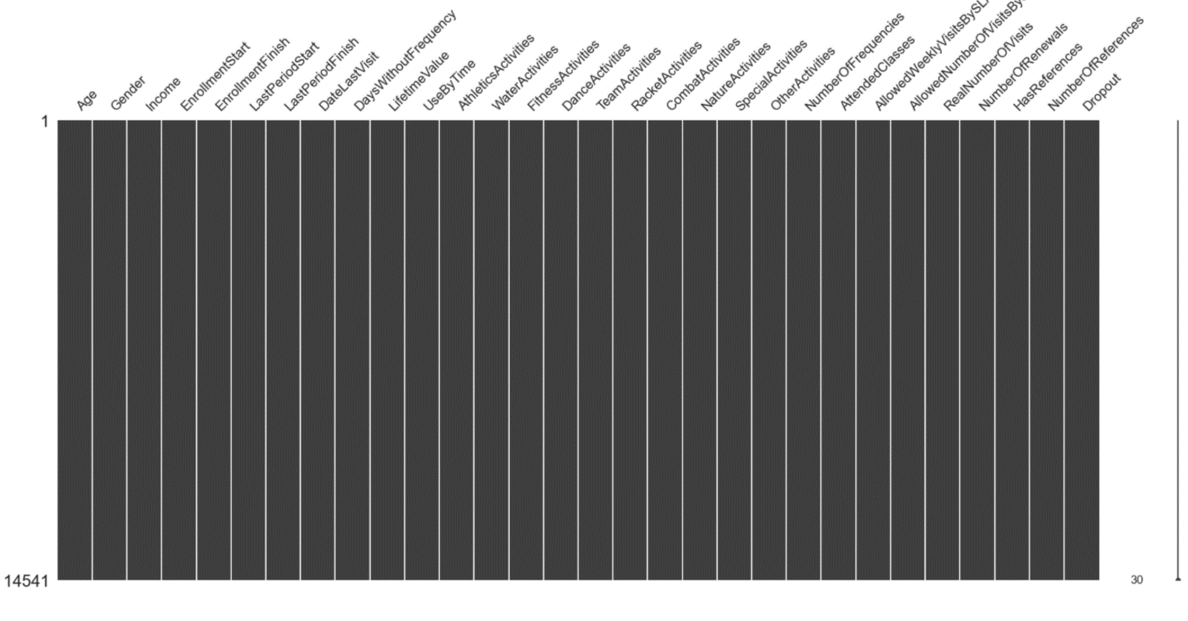
[4] 2.3. Clustering. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/clustering.html

[5] Sklearn.cluster.KMeans. (n.d.-b). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

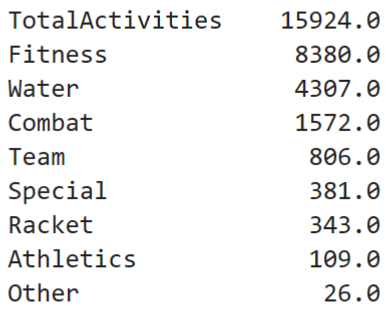
8. Appendix

Annex 1 – Treatment of missing values



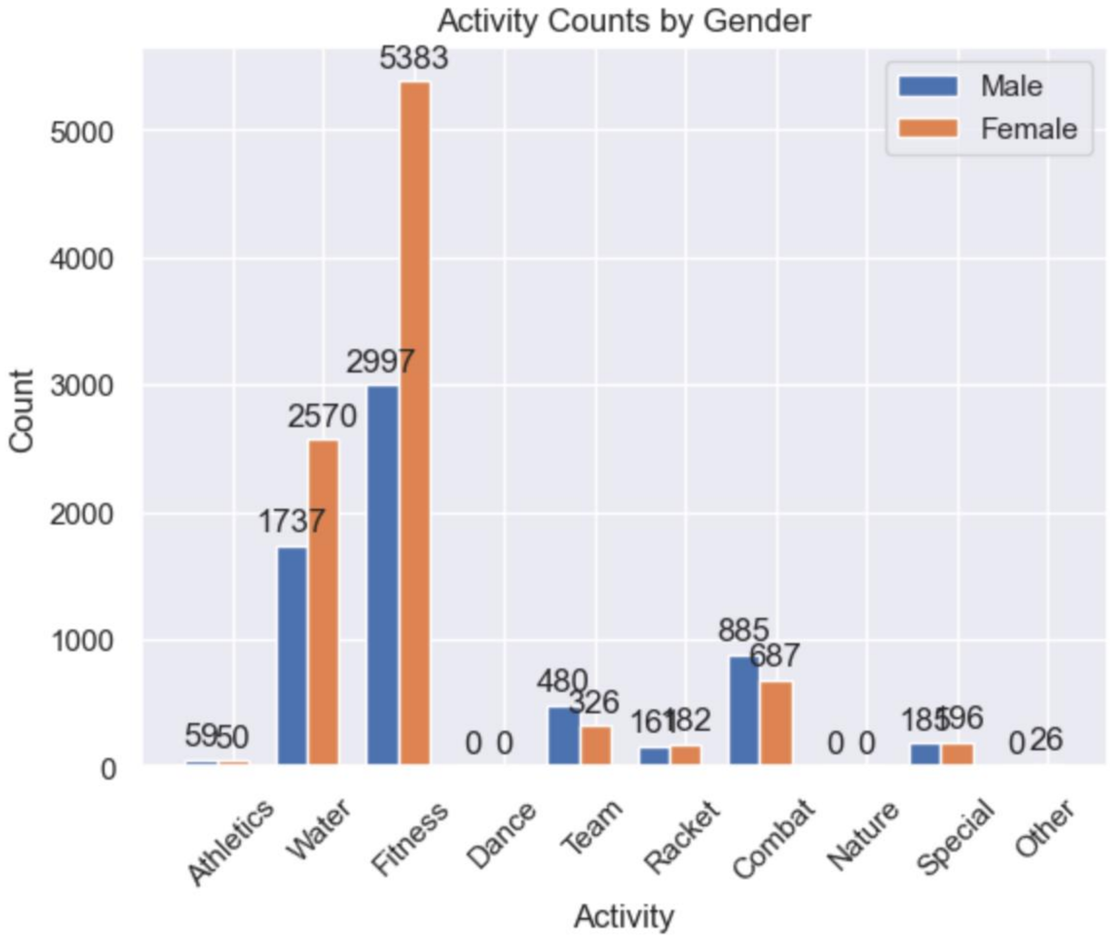


Annex 2 – New variable TotalActivities that later became NActivities

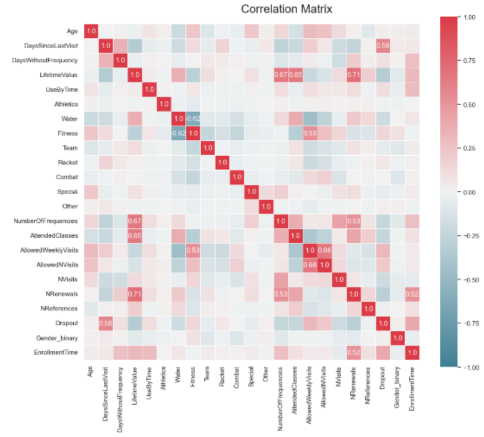


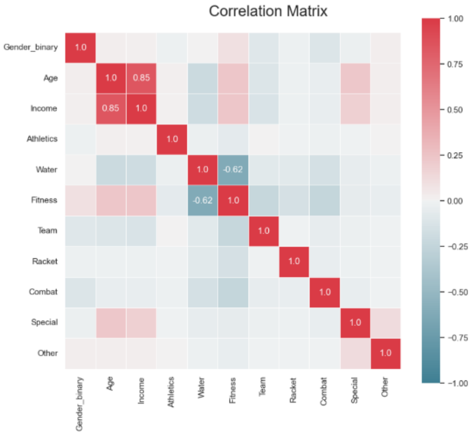
```
TotalActivities    15924.0
Fitness             8380.0
Water               4307.0
Combat              1572.0
Team                 806.0
Special              381.0
Racket               343.0
Athletics            109.0
Other                 26.0
```

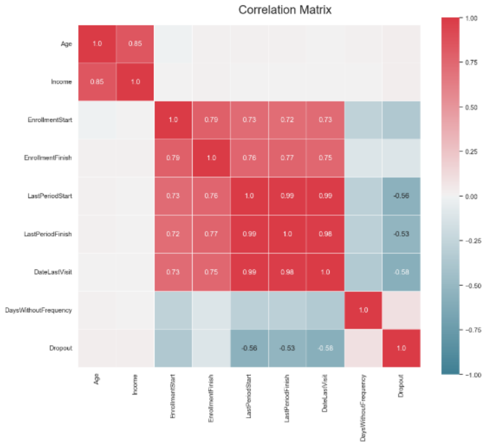
Annex 3 – Activity counts based on gender

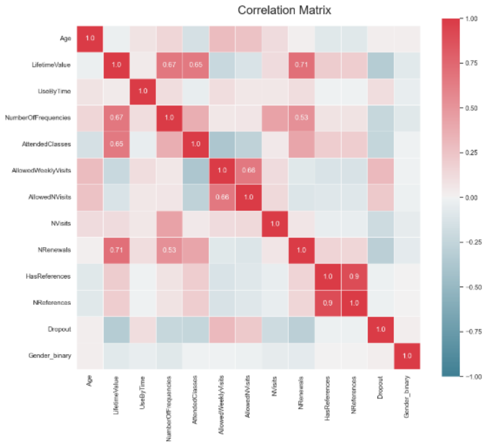


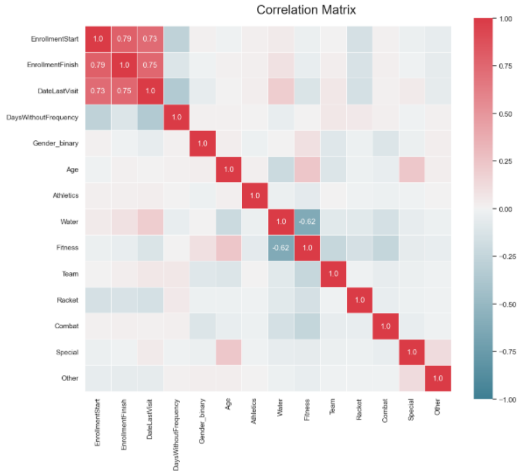
Annex 4 – Correlation matrix heatmaps



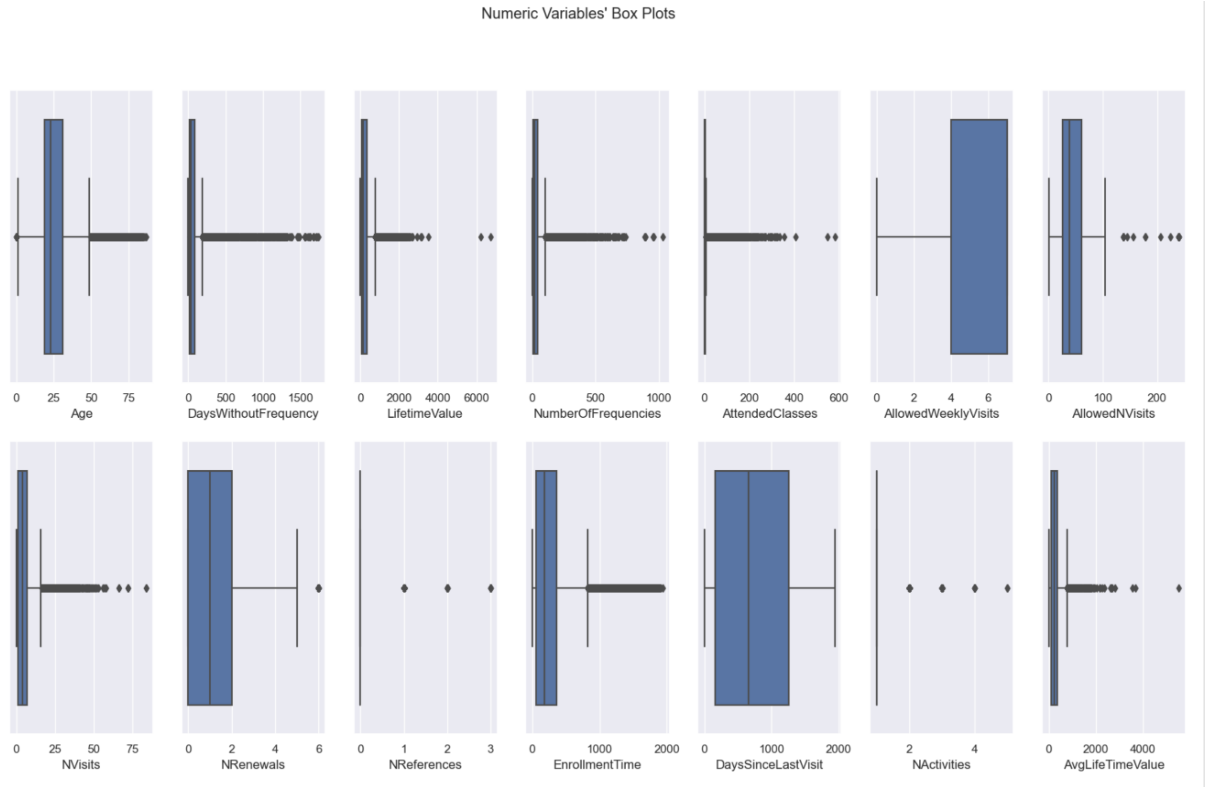








Annex 5 – Box Plots so see the outliers


Numeric Variables' Box Plots
Age
DaysWithoutFrequency
LifetimeValue
NumberOfFrequencies
AttendedClasses
AllowedWeeklyVisits
AllowedNVisits
NVisits
NRenewals
NReferences
EnrollmentTime
DaysSinceLastVisit
NActivities
AvgLifeTimeValue