Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Marketing Segmentation for XYZ Sports Company

Diogo Reis, 20230481

Marta Jesus, 20230464

Pedro Cerejeira, 20230442

January, 2024

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Index

1. Introduction

XYZ Sports Company is a fitness facility. It counts with more than nine different activities for all ages that have been contributing to and supporting the community for some years now.

The main goal of every company, just like XYZ Sports Company, is to expand the business and gain new customers. The first thing that must be done is to understand the actual customers by their behaviour, including the activities each attends and their aggregated values and demographics. Therefore, it will be easier to segment client's preferences and improve the services offered. The dataset used is from June 1st, 2014 until October 31st, 2019, and was provided by the ERP system of the company. The project aims, through the use of clustering algorithms, to provide a data-driven marketing strategy based on customer segmentations to improve engagement, and gain insights about which activities are preferred and hence personalized services.

2. Data Preprocessing

The main goal of data pre-processing is to transform features into a representation that is more worthwhile for the goal of the project [1].

In this section, the dataset is going to be analyzed in order to understand the data, which is essential to obtain good performance. Initially, it was concerned about what kind of information each feature could provide, if there were inconsistencies, and any other problems related to duplicated or missing values (which were treated using different approaches) and data types. Before data preparation, which consists of transforming current variables and creating new ones (what is done in the feature engineering subsection), visualizations were plotted, and outliers were treated. The dataset used had 14942 observations and 30 features, with the index defined to be *ID* as each customer has a unique number.

2.1 Duplicated Values

To avoid treating irrelevant data, the first step was checking for possible duplicated values and dropping the one that was found. Throughout the project, multiple copies of the previous datasets were created, allowing making changes without affecting them and avoiding having to run the entire notebook.

2.2 Descriptive Analysis

To get some insights into the data, descriptive analyses of each variable were performed. It was observed possible outliers as the maximum values and the 75-percentile had significant discrepancies. More into it, grouping some features by *Gender* was acquired that the men would spend more money in the gym facilities than women. Moreover, both *DanceActivities* and *NatureActivities* had zero records, so they ended up being dropped from the dataset *(fig. 1).*

2.3 Checking Data Types

Ensuring the correct data types for every feature is crucial. Binary features like *UseByTime*, *AthleticsActivities*, *WaterActivities*, *FitnessActivities*, *TeamActivities*, *RacketActivities*, *CombatActivities*, *SpecialActivities*, and *OtherActivities*, and features with float-type, *NumberOfFrequencies*, *AllowedWeeklyVisitsBySLA*, *HasReferences,* were handled as they would become more convenient as integers.

2.4 Data Visualization

Visualizations help understand the distribution of the data in a more intuitive way. Variables were categorized as numerical, *int64,* and *float64,* except for binary features as they only have 2 values, followed by plots of boxplots, histograms, and pair plots. These provided insights revealing that the majority of the company's clients are under 25 years old, the dropout usually occurs within the first year and the number of visits to the facilities is less than 25 (fig. 2). Regarding activities, it was clear that kids predominantly frequent *WaterActivities,* while adults with more than 30 years are more common in *OtherActivities*. *TeamActivities* typically involve individuals with an age range between 10 and 20 years and *SpecialActivities* embrace a range of age bigger than other activities (fig 3). Each activity was also checked the gender and it was observed that *OtherActivities* were mostly frequented by women with more than 20 years. Men frequent more *TeamActivities* and *CombatActivities,* while women go more for the *RacketActivities* (fig 4).

2.5 Missing Values

To maintain data consistency, it is important to address missing values. *Income* had 3.3% of missing values and due to its high correlation with *Age* (87.5%) the Nans were addressed using *Age* as an estimator. Since only 0.86% of individuals under 16 years old had reported income, and considering it as residual, further in the project they are going to be treated as outliers. The Nan values of those under 16 were filled with 0. For individuals aged between 16 and 64 a Linear Regression was used to estimate the incomes as the ones above 64 are likely retired. For the ones over 65, the Nan were filled with the average income of this group, presuming that they are in the retirement age and won't increase their income.

For the *NumberOfFrequencies* and the activities with Nan values, considering the low count, they were filled with the mode. The last feature with missing values, *AllowedWeeklyVisitsBySLA*, was addressed using the median since it didn't have any high correlation with other features, and it is more resistant to outliers.

2.6 Outliers Treatment

There are multiple possible approaches to deal with outliers. In this project two options were considered, manual and IQR methods. Going back to the *Age/Income* relationship to deal with the individuals under 16 with reported income, it was noted that individuals aged 2 had income, which was almost impossible. Assuming that age groups with less than 2% reporting income were classified as outliers, and consequently dropped. For the remaining numerical features, boxplots were plotted revealing that many of them presented quite substantial inconsistencies. Thresholds for 8 features were established and the percentage of data kept after removing the outliers was 98.782%, less than 3%, and considered acceptable. The IRQ method, commonly used for the normally distributed variables, was also applied. However, after performing this method the removed data was way bigger than 3%. An alternative approach involved combining both filters (manual and IRQ) resulting in a removal of data less than 1%. The method used in this project was the manual filters and it showed being more consistent and impactful for the outcome.

2.7 Data Preparation

Data preparation is the process where the goal is to extract the maximum insights possible from the current data by transforming features.

As *Gender* only assumed two values, Female and Male, it was transformed into 0's and 1's as they are easier to analyze. The dates of *EnrollementStart, EnrollementFinish, LastPeriodStart, LastPeriodFinish,* and

DateLastVisit were converted to *datatime* values. For the references, it was checked if there were some incoherences. *HasReferences* it´s a True or False variable so, if the *NumberOfReferences* was equal to 0 and *HasReferences* equal to 1, this last column should be updated to 0. Also, if *NumberOfReferences* was bigger than 0 and *HasReferences* equal to 0, it would be updated to 1.

It was also checked if the facility visits would have some incoherence. The value counts of both *AllowedWeeklyVisitsBySLA* and *AllowedNumberOfVisitsBySLA* showed that this last variable was a float-type variable which is impossible. To correct this, *AllowedWeeklyVisitsBySLA* was multiplied by 26 as it the number of weeks half a year has, and *AllowedNumberOfVisitsBySLA* was updated. It was checked if *RealNumberOfVisits was* bigger than the *AllowedNumberOfVisitsBySLA* to confirm that no individual would go to the gym facilities more times than the ones they were allowed to. The last variable checked was *LifeTimeValue* and it was found that three individuals had 0 and so, they were dropped.

2.8 Feature Engineering

Some new variables were created to either simplify existing ones or to add overall value to the dataset. Therefore, *TimeEnrolled* was developed being the difference between *EnrollementFinish* and *EnrollementStart* so, the number of days an individual was enrolled in the gym was clear. However, by doing this, it was discovered that they would have signed up and dropped out on the same day, having 0 days enrolled. To address this, it was assumed that these individuals could use the facilities on that day. Consequently, a day was added to *TimeEnrolled*, setting 1 as the minimum number of days enrolled in the facilities. Subsequently, *TimeEnrolled_Months* was created by dividing *TimeEnrolled* by 30, representing approximately the number of months a person was enrolled, which will prove useful further in the project.

Adding to this, *PeriodEnrolled* was defined, following a similar logic as before, calculated as the difference between *LastPeriodFinish* and *LastPeriodStart.* This was also converted to days and since no individuals had 0 days, no adjustments were needed.

The next three new variables are ratios, as they will help see the kind of engagement individuals have within a month. Firstly, *RatioOfMonthlyFrequency,* created by dividing the *NumberOfFrequencies* by *TimeEnrolled_Months,* indicates approximately how many times an individual uses the gym facilities in a month. The next one was created to understand how many renewals the individual would make during a month, named *RatioOfRenewals,* calculated by dividing *NumberOfRenewals* with *TimeEnrolled_Months.* The last variable defined was *RatioOfMonthlyAttendedClasses* to check the commitment individuals have to activities. However, there was an issue during their making, as they are divided by *TimeEnrolled_Months,* which has 2348 individuals with 0 as a value, and so *inf* were defined as values in every ratio variable. To solve this problem, it was necessary to replace the *inf* values with Nan and then replace them with 0, assuming these individuals would have a 0 ratio frequency, renewals, and attended classes as a value.

To simplify the dataset and since the activities were binary, *ActivityDiversityIndex* was created as the sum of all the activities an individual could attend. Additionally, a feature to analyze the membership plan each individual had been introduced. *MembershipPlan* will help them understand how much money they would spend every month they are enrolled in the gym. It is defined as the *LifeTimeValue* of each individual divided by *TimeEnrolled_Months.*

3. Variable Selection

To measure which features were the least redundant and consequently have the most importance to the segmentation analysis a variable selection was performed. Initially, it was decided to drop all *datetime64 [ns]* variables, since they are not useful to any clustering method, and the information needed from those variables was already extracted on feature engineering.

Moreover, a correlation matrix was plotted to evaluate the correlations between all numerical variables. Every pair of variables with a correlation value equal to or superior to 0.7 was considered redundant, so to address this, one of those two variables was selected and dropped (fig 5).

In the end, *HasReferences* was dropped not only due to its higher correlation with other variables that were removed but also because it is categorical, and the same information was already captured by *NumberOfReferences. Income* was also dropped, given its high correlation with *Age* which was already considered. The last one, *NumberOfRenewals* was also dropped as it is correlated with *LifeTimeValue* which is an important feature for the segmentation further on the project.

4. Data Normalization

In this section, the first thing done was dividing the features into metric and non-metric, depending on their data type. The goal was to provide all the metric features with a common scale, giving the attributes the same weight *[2, p. 113].* The normalization technique chosen was Min-Max-Scaler, as it is less disruptive and avoids distortion. With it, the data is scaled in a [0,1] range.

5. Segmentations

To enable better performance of the clustering methods and after many different approaches tested, the best way found to split the features into smaller subsets was: Sociodemographic = [*Age*, *LifetimeValue*, *ActivityDiversityIndex*], in this it is expected to capture the different age groups, the value each spend and how many activities those individuals practice; Engagement = [*NumberOfReferences*, *RatioOfRenewals*, *RatioOfMontlhyFrequency*, *RatioOfMonthlyAttendedClasses*, *MembershipPlan*], aiming to represent the groups by their ratio. All the variables in these subsets are numerical, meaning that they are available to be used in clustering algorithms.

6. Clustering Models

Four different methods for each segmentation were performed, applying the same techniques for both. The evaluation of the results was done by analyzing the mean value of each feature and the cluster distributions.

6.1 Hierarchical + Partition Methods

K-means is one distance partition-based algorithm that partitions the data into *k* clusters. Initially, one random partition is created and then starts an iterative process in which every object is reassigned to a similar cluster, where the objects are more related. The algorithm stops once it finds the optimal clusters. *[2, p. 452]*

The hierarchical method can be classified as agglomerative or divisive. In this project was only used the agglomerative method, which uses a bottom-up approach. It starts with every object forming an isolated cluster and then iteratively merges all clusters until they are all into one or a certain condition is reached

[2, p.459]. There are four aggregation rules: *single*, *complete*, *average*, and *ward* which determine the distance metric to be used between all pairs of points merging those that minimize the rule *[3]*. The most common and efficient one is the *ward* whose objective is to minimize the variance by minimizing the sum of squared differences between all clusters *[3]*. When selecting the *ward* method, the affinity parameter, which also selects the distance metric to be used based on pair-wise rules, it's mandatory to use of Euclidean distance *[3*]. The *distance_threshold* sets the maximum distance in which the clusters stop merging. *[3]*

For both segmentations, the first thing done was to evaluate which method reached the highest R² value through various numbers of clusters. The R² represents the proportion of the variability in the y-variable that can be accounted for by the regression. The highest value occurs when the regression perfectly aligns with the dataset. For this purpose, the R² was applied to compare between *K-means,* initialized without specifying the number of clusters, and agglomerative clustering with various linkage methods *[4, p.182]*.

For the socio-demographic segmentation, *K-means* got the highest R² result, followed, as expected, by the *ward* linkage method. In this case, the R² graph also suggested that the optimum number of clusters should be between 3 and 4 (fig. 6). To confirm which was the optimal number of clusters, the *silhouette score* was also applied, a test that compares the cohesion for all observations: how similar this point is regarding the points assigned to the same cluster compared to other clusters. The values of the silhouette coefficient vary between [-1,1] being the best result the nearest to 1 as possible. This one suggested that the ideal number of clusters would be 2, decreasing the coefficient till increasing again with 5 clusters. That being said, 3 and 4 clusters were tested for this segmentation (fig. 7)

For the engagement segmentation, the R² graph showed once again that *K-means* was the one that reached the higher score, with 3 or 4 clusters. The *silhouette score* showed the highest value for 4 clusters so, this was the final choice to test (fig. 7)**.**

6.2 Self-Organizing-Maps

Self-Organizing-Maps, also called *SOM* is an unsupervised technique related to neural networks. Their goal is to adjust its units to match the input data, ensuring that the network represents the data as closely as possible. *Hierarchical* and *K-means* clustering were applied on top of *SOM*. The model was trained with a grid of 10x10, for both segmentations, giving results good enough to observe the cluster distributions.

For the *K-means*, an inertia graph was plotted, showing evidence that the optimal number of clusters to use would be 3, after using the elbow method, for both socio-demographic and engagement segmentation. For the *Hierarchical*, the initial R² graph confirmed that *ward* method linkage is the most efficient when compared to the others, it was the one used. Then, after plotting the dendrogram and defining a threshold of 12, the chosen number of clusters to use was 3, for both segmentations.

However, when those results were transferred into all the dataset, the cluster assignments weren't as accurate as they previously showed. The cluster distribution changed negatively as it assigned most of the observations into a single cluster making it non-well distributed.

6.3 Density-Based-Clustering

The last method applied was density-based clustering, which as the name shows is based on the distance notion. The main idea is to grow each cluster until the density (number of data points) surpasses some threshold. These kinds of methods are also useful to find outliers and decrease the noise *[2, p. 449].*

DBSCAN was performed, an algorithm that separates the data between high and low density. It only has two parameters: *min_samples* and *eps*. The first one defines how the algorithm stands for noise, being appropriate to increase this parameter when a dataset contains many outliers. The *eps* control the local neighbourhood of the points which means that if it's too small some of the data cannot be clustered and be classified as noise and if it's too large can create bigger clusters that don't give relevant information. Simplifying, higher *min_samples* or lower *eps* means that it is necessary to have a higher density between data points to form clusters. *[5]*

On both segmentations, evaluating the *eps* parameters by plotting the *k-distance* graph, was the first thing done. In the socio-demographic segmentation, the graph indicated an *eps* equal to 0.2 while in the engagement segmentation, the *eps* was roughly 0.35. Then, a function was used to evaluate the values of the *min_samples,* with the selected *eps* value fixed, giving a view of the variation of the number of clusters and their respective noise. For the socio-demographic, was selected min_samples = 6, resulting in 4 clusters and 9 observations classified as noise. On the other hand, for the engagement segmentation, the min_samples selected were 2 and it returned 2 clusters with a noise of 1.

In the end, and by the number of features per subset, it was known à prior, that the optimum number of clusters should be between 3 and 5.

6.4 Final Clusters for Each Segmentations

The final choice was made based mostly on the cluster distributions, as these, were the ones that showed a better balance for the data division (fig. 8).

The clustering method chosen for each segmentation was *K-means,* with 3 and 4 clusters for the socio-demographic and engagement segmentation, respectively.

For socio-demographics, cluster 0 (composed by around 65.99% of individuals) has a medium age of 23.8 years old, a 175.62€ of *LifeTimeValue* and participates, on average, in 1 activity. For cluster 1, the medium age is 49.3 years old, the average LifeTimeValue is 410.56€ and each individual participates, more or less, in 1 activity. In cluster 2, was found that the average age is 8.21 years old, *LifeTimeValue* is 651.65€ and each individual joins 1 activity.

For engagement, all clusters have almost the same *NumberOfReferences* which is less than 1. Clusters 0 and 3 have both less than 2 frequencies per month on average. Cluster 1 has 2.46 visits and cluster 2, the one with more monthly frequencies, has 7.6. Related to monthly attended classes, cluster 0 is the one with the lowest average value: 0.23. Cluster 1 is the one with the most attended classes per month with a value of 0.56. Clusters 2 and 3 have, respectively, 0.53 and 0.46 monthly visits. About *MembershipPlan*, cluster 2 is the one with the highest amount spent per month (34.9€), followed by cluster 1 (with 22.1€), cluster 0 (with 19.83€) and cluster 3 (with 17.1€). On *RatioOfRenewals*, during all the time enrolled, it's possible to understand that cluster 0 is the one with the lowest value, cluster 1 is the one with the highest value (0.20). Cluster 2 and 3 have, respectively, 0.02 and 0.1 *RatioOfRenewals*.

7. Clustering Analysis

7.1 Merging the perspectives

Initially, the two clustering methods chosen were merged by using *Hierarchical* clustering (fig. 9). The result was of 4 clusters, the bigger one with 52,2% of the gym customers, the second one with 31,84%, and the other two clusters, with the lowest distributions, with 9.02% and 6,94%, respectively (fig. 10).

In the final clustering merged table, it can be observed that clusters 1 and 2 are the regular customers of the gym as they have the lowest *DaysWithoutFrequency*. It is highly important to maintain them so we can improve their experience and promote return visits, as they are the ones with fewer *TimeEnrolled_Months.* The ones in cluster 1 spend, per month, more in the gym facilities. On the other side, clusters 0 and 3 come less to the facilities however compared to the previous ones, these have a higher *TimeEnrolled_Months,* but a *Dropout* close to, or 1. This indicates that the strategies should pass by attracting as well as retaining them.

For each cluster, a description and the characteristics of them will be explained as well as the marketing strategies for them.

Firstly, cluster 0, is composed of individuals with an average age of 9.25 years old. In this cluster, we have the kids that enrolled more in *WaterActivities* (fig. 12), and that have a 1.59 average of *RatioOfMonthAttendendClasses,* demonstrating that they are enrolled in the gym for the activities. These are the ones that have a higher monthly membership, assuming that *WaterActivities* are the most expensive.

For this cluster, since they have a high rate of *Dropout* even though they are the ones with more time enrolled, the company can develop a marketing strategy based on this. These customers are important for the gym as they are the ones that spend more money monthly, so a plan could be offered. This may be a promotion based on a flexible membership plan as they are kids and probably during the summer months, they don´t attend the gym facilities. In these months, they could pay half of the price they usually pay, guaranteeing a reduction of the dropout rate (it is close to 1), as they will be enrolled at least for the next scholar year.

Cluster 1 is composed of individuals with an average of 29.3 years old, which are young adults. The main activity practised by these customers is *FitnessActivities*, existing also a relevant proportion of individuals that practice *WaterActivities* and *CombatActivities (fig. 12)*. It's a group with low *TimeEnrolled_Months* and on average 50.3 *DaysWithoutFrequency,* but with the high *RatioOfMonthlyFrequency*, meaning that these are the customers who come more often to the facilities. This group of individuals spends an average of 24.7€ per month. It's also the group with lower *NumberOfReferences*, the lowest *RatioOfAttendedClasses* and a moderate *Dropout* (fig. 11).

The marketing strategies implemented for this group could be: since they spend a moderate monthly value and are older, at least compared with cluster 0 (which is the only one with a high *MembershipPlan*), these individuals probably are also interested in health campaigns which could mean implementing nutritionists, physiotherapists and other healthcare professionals to make regular monitoring. To increase the *NumberOfReferences*, a good option should be to create a reference system, rewarding the clients that bring new customers to the company. To get extra loyalty and increase the *TimeEnrolled_Months* should be interesting to create different levels of clients according to the duration they are enrolled giving extra offers to the older clients.

Cluster 2 is, on average, the younger. Since it´s the group with the lowest *TimeEnrolled_Months* but also with the lowest *Dropout*, they are the most recent customers of the company. Their preferred activity is *WaterActivities (fig. 12)* showing the same pattern between age and favourite activity as cluster 0, which is sustained because the *RatioOfMontlhyAttendedClasses* is bigger than the *RatioOfFrequency*. On average, they spend monthly 14.8€, (fig.11).

To ensure that the company will retain these and that will acquire new clients, a good approach is creating packs for the new clients and creating group discounts for new enrolments, for example to schools and other entities that work with children.

The last cluster is number 3, it shows the individuals with an average age of 29.27 years old and has the lowest *MembershipPlan.* We can tell that even though they are the ones that pay less and have a reasonable *TimeEnrolled_Months,* about 18 months, their *Dropout* is 1, showing that they are not satisfied in the gym, and they will eventually drop out. These individuals frequent *FitnessActivities* (fig. 12) the most, and so the marketing strategy will focus on this. Firstly, it is highly important to understand why they are not pleased, and then promote workshops and new classes focused on fitness (fig.11).

For this last cluster and combined with cluster number 2, it is notable that the kids from cluster 2 go few times to the gym and don´t have a high rate of dropout, contrasting the adults in cluster 3 who also don't go very often but have a *Dropout* rate of 1. To try to promote their activity in the gym, it could make sense to apply a family package, where 2 adults plus a child could be part of it. As the gym also wants to make money, and these clusters are the ones that pay less, putting a specific price on the package could help the gym earn more. This way, they would stop paying for the times they go, that are few, and have a fixed price every month.

8. Feature importance

In the final part of this project, the focus was directed toward assessing the importance of each feature in the clusters. To determine the proportion of the total sum of squares (SS) explained between the clusters, the R^2 was used. The output of the function showed that *Age*, *LifeTimeValue,* and *RatioOfRenewals* are the features that had a higher R^2, establishing them as being the most influential features.

The graph of t-Distributed Stochastic Neighbour Embedding also known as *t-SNE* was plotted, to observe the distribution of the clusters. It is normally used for exploratory data analyzes and clustering as it represents the similarities between neighbours by creating a probability distribution *[6]*. In the plotted t-SNE (fig.13) it is clear the cluster distribution.

Another method used to understand feature importance was training a *Decision Tree* with *max_depth* = 3 using the final merged labels as the target variable. The goal is to differentiate between classes to obtain the purest leaves. Firstly, the data was split using the *train_test_split* method and right after fitted into the *DecisionTreeClassifier [7].* The results indicated that 95.15% of the customers were predicted correctly, on average. The Decision Tree was plotted, to better understand, the relationships between clusters and features and can be seen in (fig. 14)**.** After that, the feature importance was seen being *Age* and *RatioOfRenewals* the ones with the better results.

9. Conclusion

The effectiveness of something is directly influenced by the effort invested in it, and that is all the project is about. Unfortunately, data is not always clean or even balanced and in this case was neither of those. Throughout the project, various pre-processing techniques and feature engineer methods were explored and implemented to treat, clean and transform variables to gain some valuable knowledge about the company clients. After those segmentations - which suffered different tests - were created and different clustering approaches were applied to them to achieve the main goal of the project. At the end feature importance were assessed through R^2 and using a *Decision Tree*.

The expectations for this project were to help the XYZ Sports Company gain insights into their customer's behaviour and preferences by partitioning them into clusters and creating marketing approaches, not only to increase the satisfaction and engagement, but also to retain and gain new clients by improving the services offered. The main approaches suggested were, some promotions to encourage loyalty, healthcare services, familiar packs to try to involve families more as well as group discounts, to bring schools and Free-Time-Activities. Were also suggested to implement specific workshops and a reference system to reward the customers that bring new ones to the gym.

Yet, it was challenging as there were various date time datatypes and binary features, that could not be handled as numerical since the distance to each other calculated in the clusters would be wrong and would give biased results. There are also several ways of doing the segmentations and the chosen features for each one are the ones that would provide the best insights for customer value and behaviour that guided our marketing plan.

10. References

[1] *6.3. Preprocessing data — scikit-learn 1.3.2 documentation*. (n.d.). Scikit-learn. Retrieved January 5, 2024, from https://scikit-learn.org/stable/modules/preprocessing.html

[2] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition*. Elsevier Science.

[3] *sklearn.cluster.AgglomerativeClustering — scikit-learn 1.3.2 documentation*. (n.d.). Scikit-learn. Retrieved January 2, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

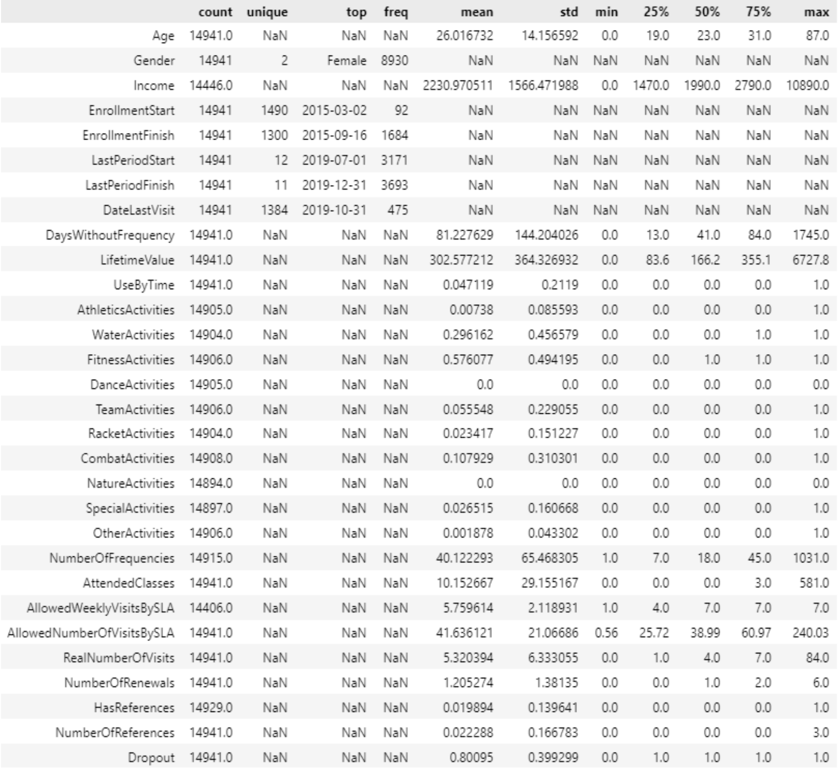
[4] Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics*. Wiley.

[5] *2.3. Clustering — scikit-learn 1.3.2 documentation*. (n.d.). Scikit-learn. Retrieved January 2, 2024, from https://scikit-learn.org/stable/modules/clustering.html

[6] Erdem (burnpiro), K. (2020, April 22). t-SNE clearly explained. Medium. https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

[7] *sklearn.tree.DecisionTreeClassifier — scikit-learn 1.3.2 documentation*. (n.d.). Scikit-learn. Retrieved January 5, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

11. Appendix



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	14941.0	NaN	NaN	NaN	26.016732	14.156592	0.0	19.0	23.0	31.0	87.0
Gender	14941	2	Female	8930	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Income	14446.0	NaN	NaN	NaN	2230.970511	1566.471988	0.0	1470.0	1990.0	2790.0	10890.0
EnrollmentStart	14941	1490	2015-03-02	92	NaN	NaN	NaN	NaN	NaN	NaN	NaN
EnrollmentFinish	14941	1300	2015-09-16	1684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LastPeriodStart	14941	12	2019-07-01	3171	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LastPeriodFinish	14941	11	2019-12-31	3693	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DateLastVisit	14941	1384	2019-10-31	475	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DaysWithoutFrequency	14941.0	NaN	NaN	NaN	81.227629	144.204026	0.0	13.0	41.0	84.0	1745.0
LifetimeValue	14941.0	NaN	NaN	NaN	302.577212	364.326932	0.0	83.6	166.2	355.1	6727.8
UseByTime	14941.0	NaN	NaN	NaN	0.047119	0.2119	0.0	0.0	0.0	0.0	1.0
AthleticsActivities	14905.0	NaN	NaN	NaN	0.00738	0.085593	0.0	0.0	0.0	0.0	1.0
WaterActivities	14904.0	NaN	NaN	NaN	0.296162	0.456579	0.0	0.0	0.0	1.0	1.0
FitnessActivities	14906.0	NaN	NaN	NaN	0.576077	0.494195	0.0	0.0	1.0	1.0	1.0
DanceActivities	14905.0	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TeamActivities	14906.0	NaN	NaN	NaN	0.055548	0.229055	0.0	0.0	0.0	0.0	1.0
RacketActivities	14904.0	NaN	NaN	NaN	0.023417	0.151227	0.0	0.0	0.0	0.0	1.0
CombatActivities	14908.0	NaN	NaN	NaN	0.107929	0.310301	0.0	0.0	0.0	0.0	1.0
NatureActivities	14894.0	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SpecialActivities	14897.0	NaN	NaN	NaN	0.026515	0.160668	0.0	0.0	0.0	0.0	1.0
OtherActivities	14906.0	NaN	NaN	NaN	0.001878	0.043302	0.0	0.0	0.0	0.0	1.0
NumberOfFrequencies	14915.0	NaN	NaN	NaN	40.122293	65.468305	1.0	7.0	18.0	45.0	1031.0
AttendedClasses	14941.0	NaN	NaN	NaN	10.152667	29.155167	0.0	0.0	0.0	3.0	581.0
AllowedWeeklyVisitsBySLA	14406.0	NaN	NaN	NaN	5.759614	2.118931	1.0	4.0	7.0	7.0	7.0
AllowedNumberOfVisitsBySLA	14941.0	NaN	NaN	NaN	41.636121	21.06686	0.56	25.72	38.99	60.97	240.03
RealNumberOfVisits	14941.0	NaN	NaN	NaN	5.320394	6.333055	0.0	1.0	4.0	7.0	84.0
NumberOfRenewals	14941.0	NaN	NaN	NaN	1.205274	1.38135	0.0	0.0	1.0	2.0	6.0
HasReferences	14929.0	NaN	NaN	NaN	0.019894	0.139641	0.0	0.0	0.0	0.0	1.0
NumberOfReferences	14941.0	NaN	NaN	NaN	0.022288	0.166783	0.0	0.0	0.0	0.0	3.0
Dropout	14941.0	NaN	NaN	NaN	0.80095	0.399299	0.0	1.0	1.0	1.0	1.0

Figure 1- Descriptive Analyses of Features

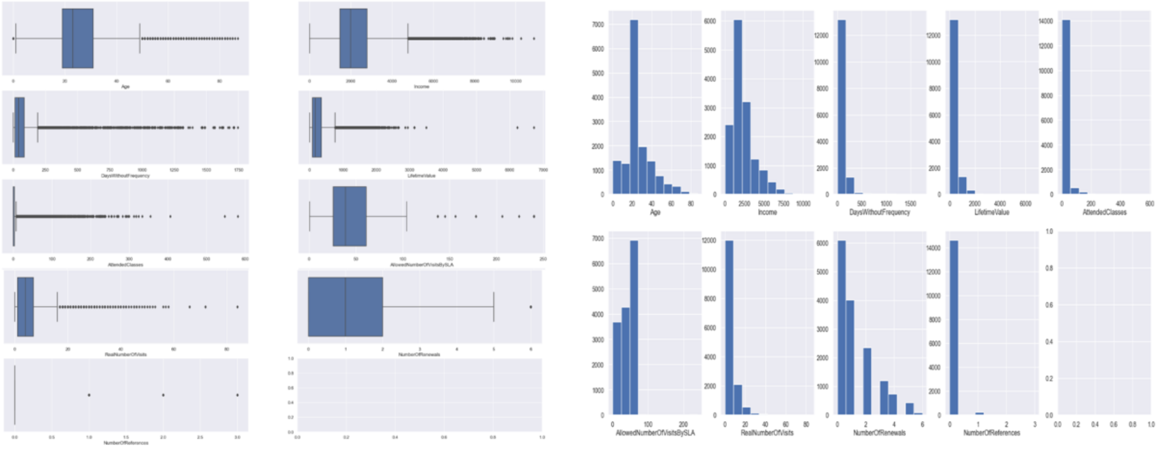


Figure 2- Features Visualizations

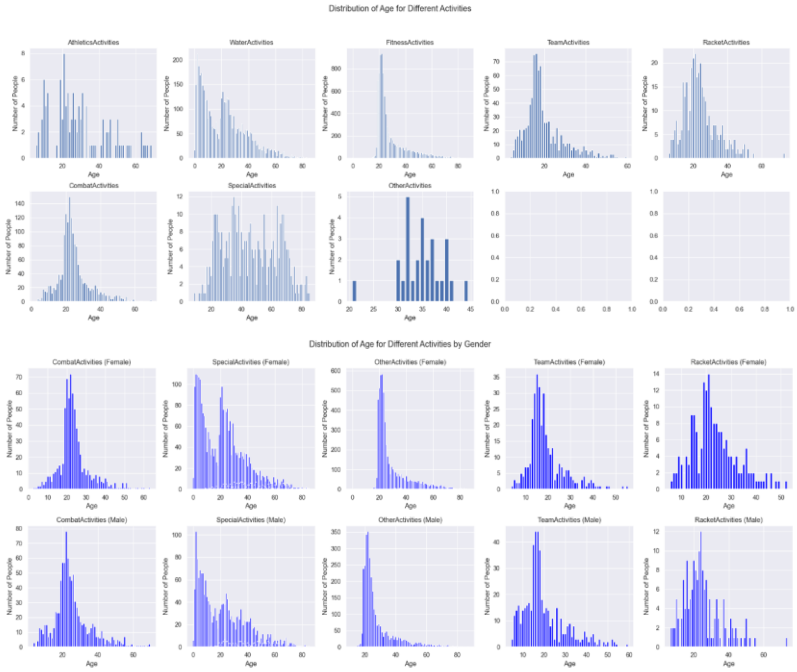


Figure 3- Histograms Activities vs Age and Gender

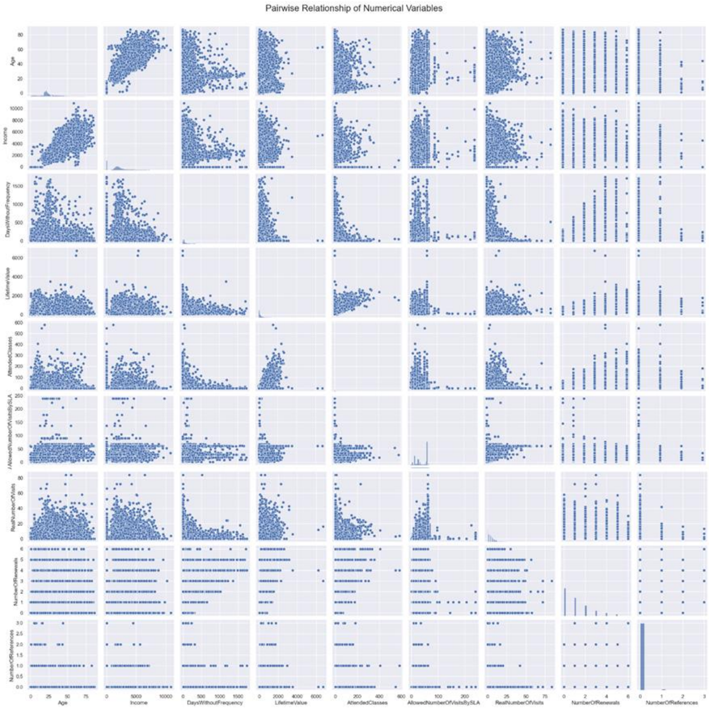


Figure 4- Pairwise Numerical Variables

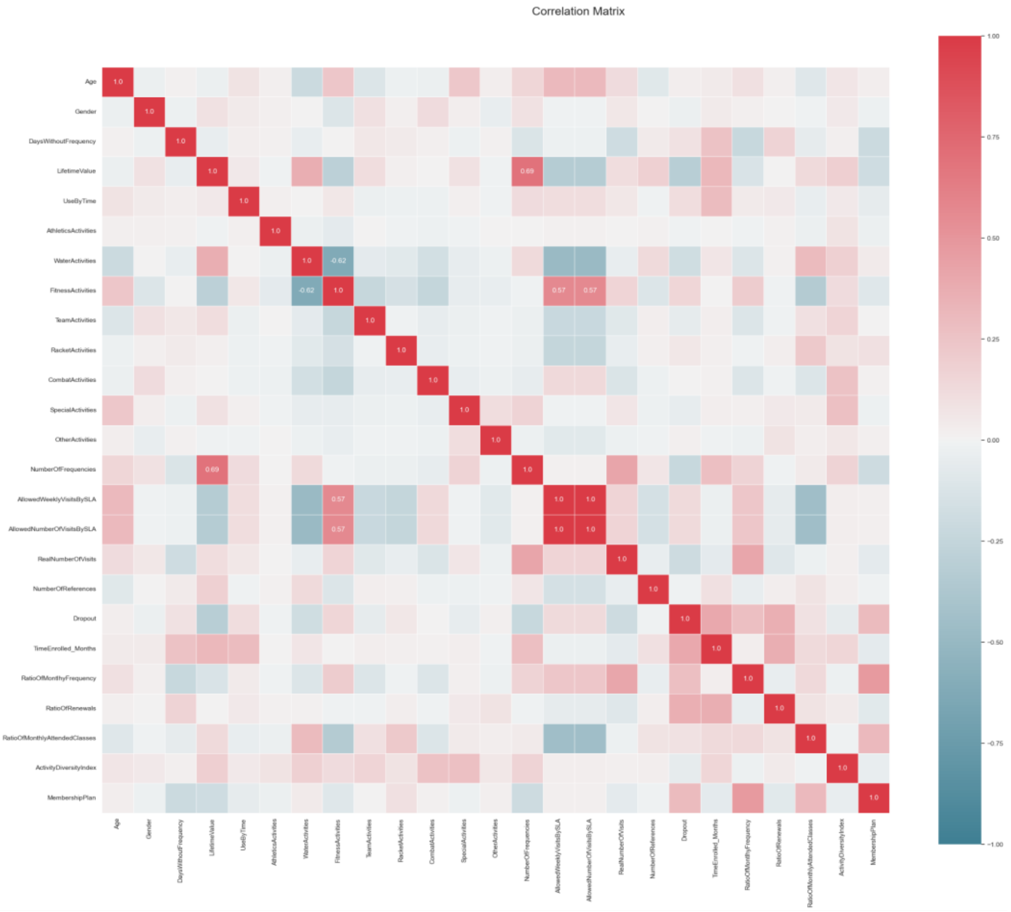


Figure 5- Final Correlation Matrix

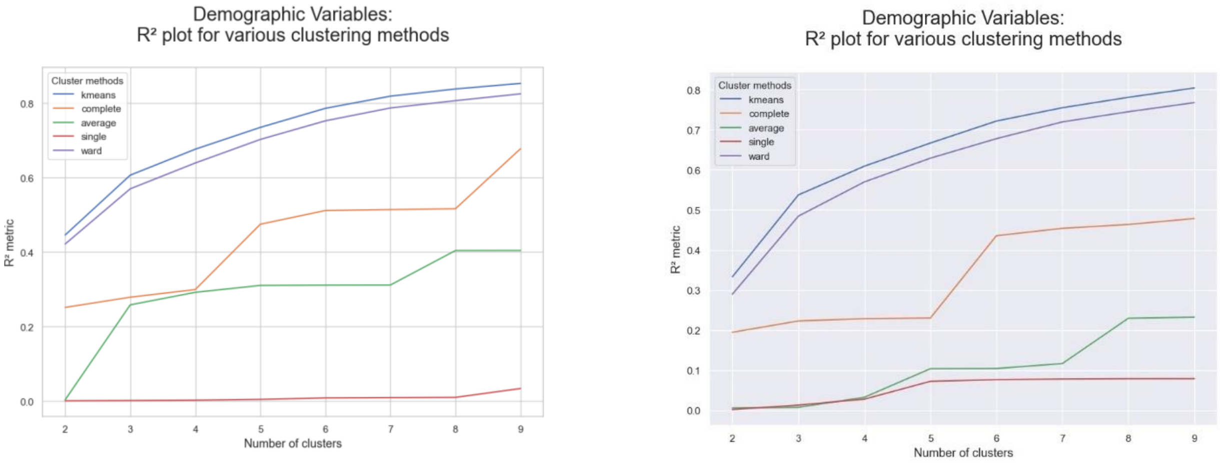


Figure 6- Socio-demographic and Engagement, respectively

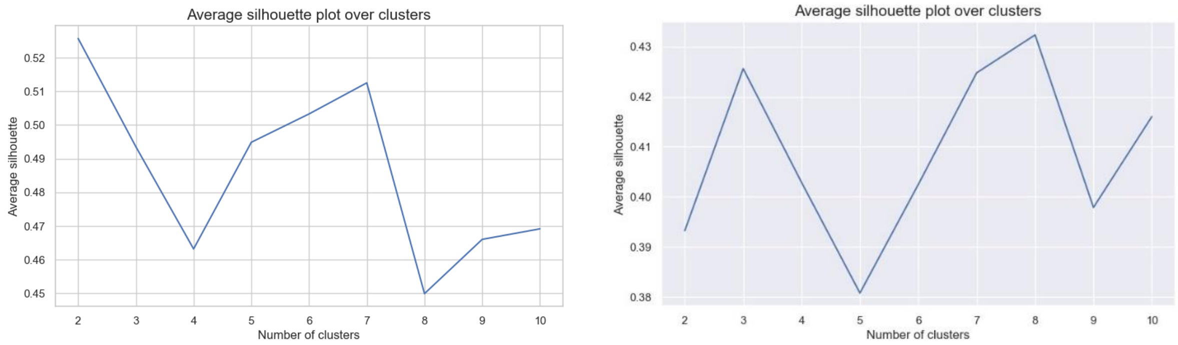


Figure 7- Silhouette for Socio-demographic and Engagement, respectively

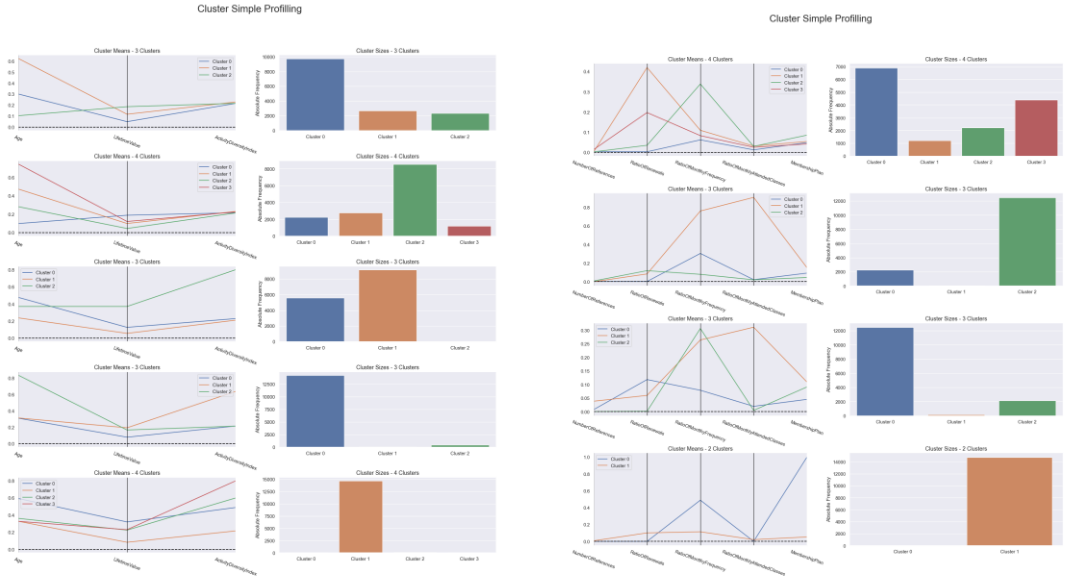


Figure 8- Socio-demographics and Engagement final clusters, respectively

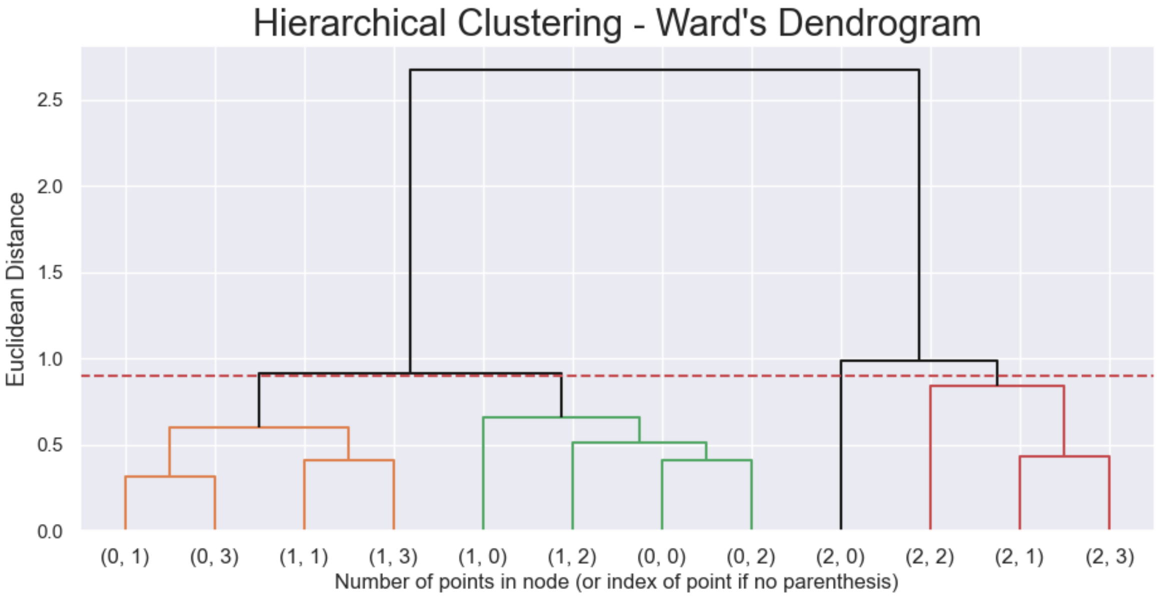


Figure 9- Dendrogram for Merging Cluster

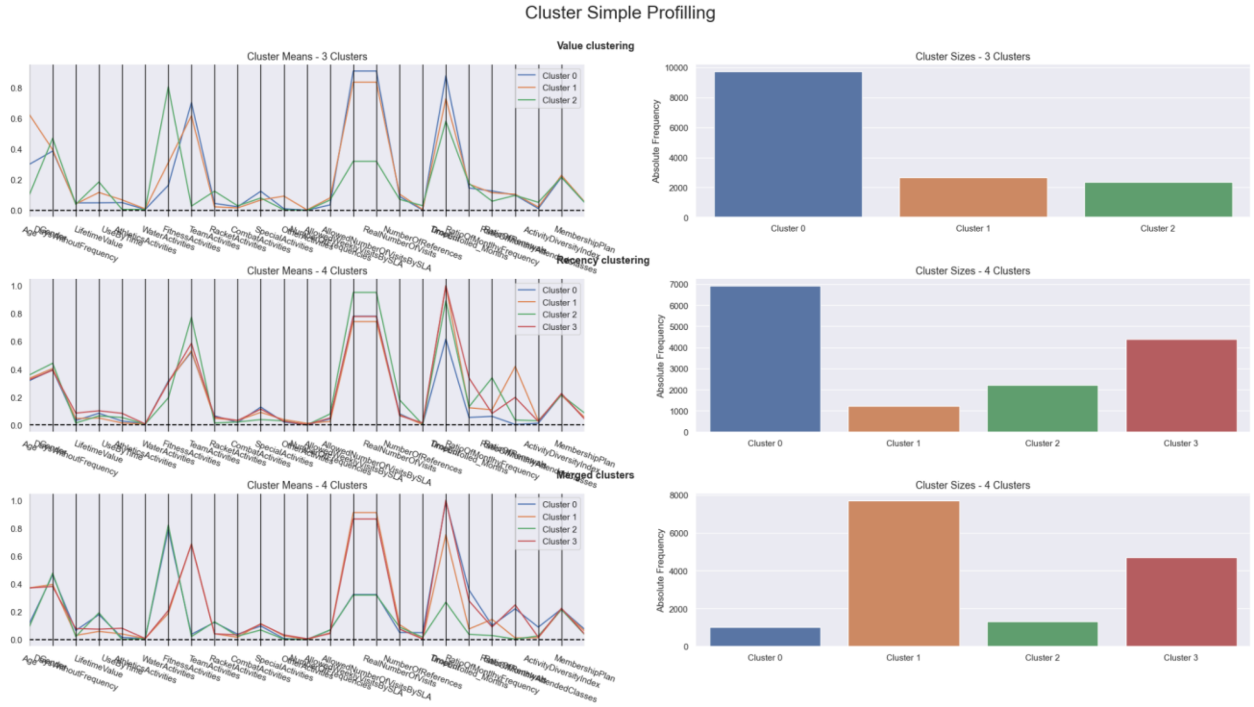
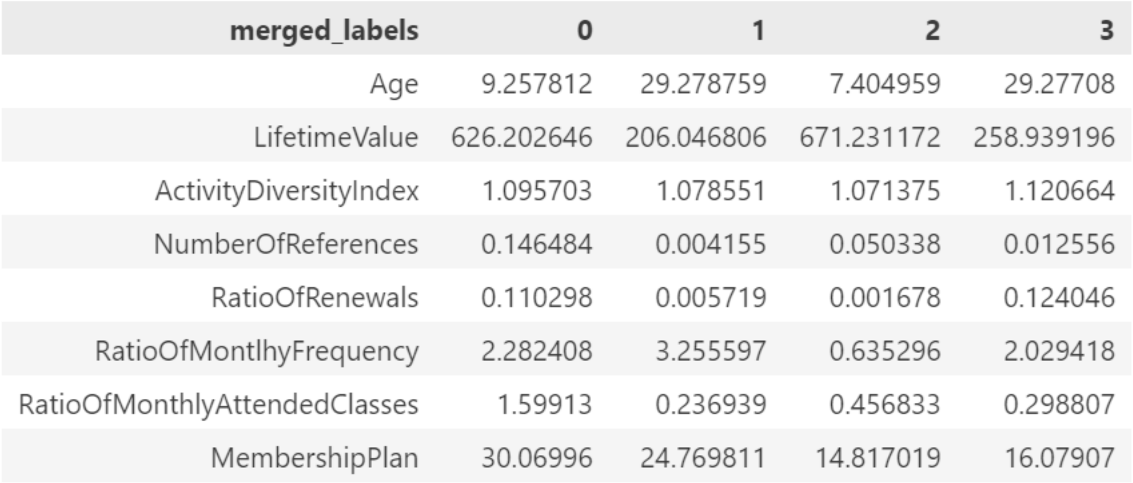


Figure 10- Final Clusters Analyses



merged_labels	0	1	2	3
Age	9.257812	29.278759	7.404959	29.27708
LifetimeValue	626.202646	206.046806	671.231172	258.939196
ActivityDiversityIndex	1.095703	1.078551	1.071375	1.120664
NumberOfReferences	0.146484	0.004155	0.050338	0.012556
RatioOfRenewals	0.110298	0.005719	0.001678	0.124046
RatioOfMontlhyFrequency	2.282408	3.255597	0.635296	2.029418
RatioOfMonthlyAttendedClasses	1.59913	0.236939	0.456833	0.298807
MembershipPlan	30.06996	24.769811	14.817019	16.07907

Figure 11- Clusters Table for Analyses

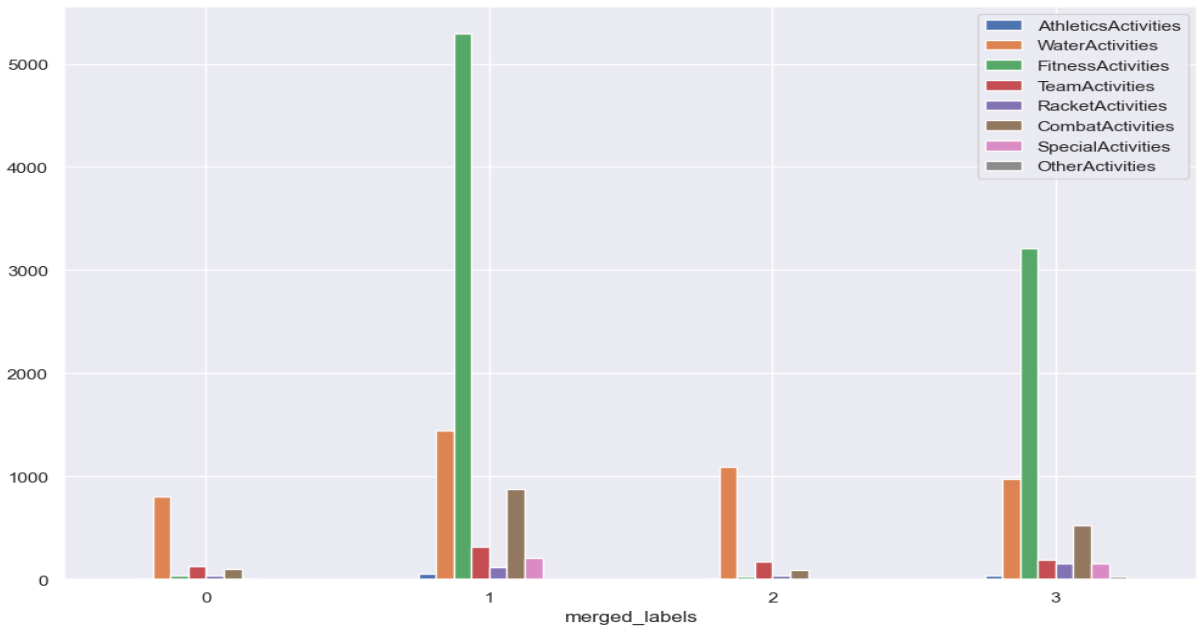


Figure 12- Activities by cluster

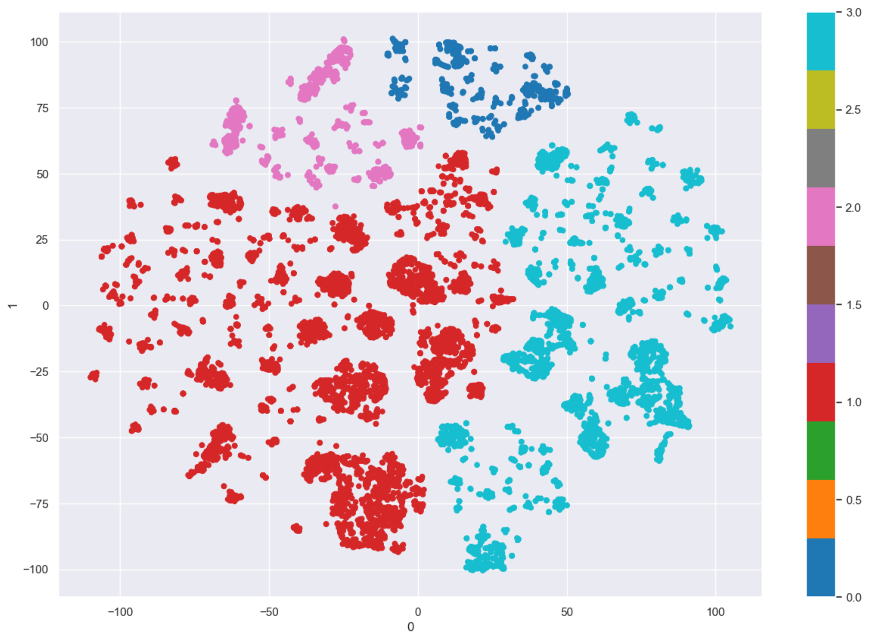


Figure 13- T-SNE for the final clusters

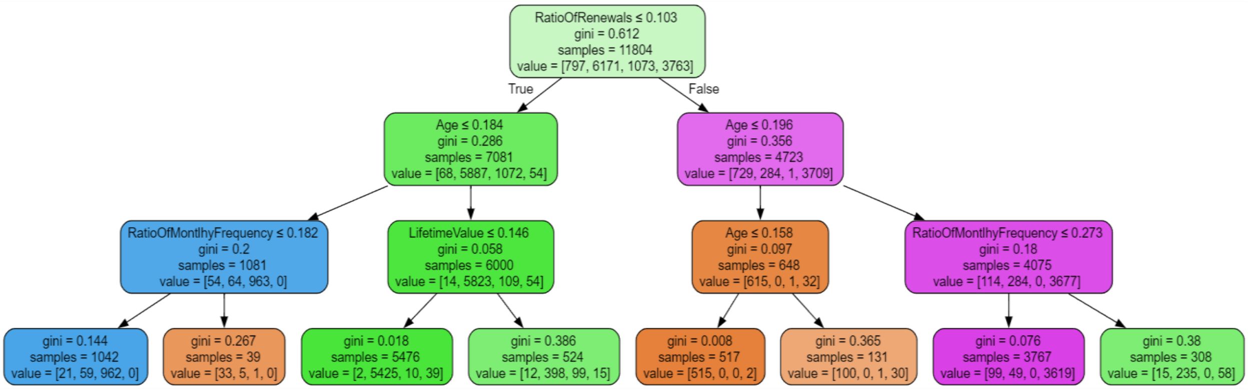


Figure 14- Decision Tree for Feature Importance