Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

XYZ Sports Company-Customer Segmentation

Group 8

Gonçalo Ferreira, number: 20230492

Pedro Adonis, number: 20230994

Sebastião Oliveira, number: 20220558

January 2024

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

INDEX

1. Introduction

This report refers to the Data Mining project regarding a dataset containing information of a fitness facility called "XYZ Sports Company", with customer related data, between June 1st, 2014, and October 31st, 2019. Our project is to create marketing segmentation of the company's clients, so that the company can understand better its clients, deliver more personalized services, and optimize customer satisfaction. This is achieved through clustering techniques that were taught throughout this course. We are aiming for a not very high number of clusters and for these to have a certain degree of balance, as marketing segmentation is only profitable for the company if the clusters are well defined and have a reasonable number of customers but are also not too generic. According to our interpretation of the guidelines given, we should create customer segmentation and then give insights into the different activities that the clients prefer to participate in based on them and the rest of the data. In the end, we will combine these clusters with the different activities, to provide more relevant insights. Our work will help the company understand better the characteristics of its clients and what segmentations they belong to.

2. Data Exploration and Preprocessing.

2.1. Exploratory Data Analysis

The first step was to import the provided dataset into a dataframe. The dataset initially had 14942 observations and 31 variables. We looked for duplicates and saw that we had a situation where 2 observations had the exact same attributes, except for the ID, where the latter was 18942 and 22873, so we decided to drop one of them.

Since one of the columns was called "ID", and after checking that there are no remaining duplicates, we decided to set this variable as the index. This information can be checked in Figure 1. The following set was to see the existing Nans. As it is shown in Figure 2, you can check the existing missing values. We had these in the following variables: *Income, AthleticsActivities, WaterActivities, FitnessActivities, DanceActivities, TeamActivities,RacketActivities,CombatActivities,NatureActivities, SpecialActivities, OtherActivities, NumberOfFrequencies, AllowedWeeklyVisitsBySLA* and *HasReferences.*

The next step consisted in separating metric from non-metric features. This was achieved by creating a list, on which we inserted manually the names of metric features. The non-metric features were obtained by taking out the names of metric columns from the dataframe's columns. The separation can be seen in Figures 3 and 4.

We supported our data exploration with histograms and boxplots. These are meant only for metric features but gave us good first insights and a general idea regarding data distribution and presence of outliers. Our metric variables seem to display an irregular distribution and the boxplots reveal to us that the data contains a lot of outliers.

Regarding the non-metric features, we plotted bar charts. From the graph's visualization, we can conclude that we have more women than men in our dataset, as well as some imbalanced variables,

as we can see in Figure 5. We can also conclude that most activities' variables have clients that did not participate in them, meaning that the 0 is more frequent. We have 2 variables only containing values in one category, which are *DanceActivities* and *NatureActivities.* In our dataset, no client participated in these activities.

Additionally, we made some graphs combining different variables. We started by making a bar chart with the average income of the clients that participated in each activity, as we can see in Figure 6. The activities with richer clients on average are *SpecialActivities* and *OtherActivities.* The activities with poorer clients on average are *WaterActivities* and *TeamActivities*.

Lastly, we checked for the distribution of men and women attending the different types of activities, as can be found in Figure 7. *AthleticActivities* and *CombatActivities* are more frequented by men, whereas *WaterActivities and FitnessActivities* are more frequented by women. *TeamActivities*, *RacketActivities* and *SpecialActivities* are the most balanced activities. *OtherActivities* are only practiced by women, but it must be considered the fact that there are only 23 clients that match this case, so for further analysis and conclusions more data would be needed.

2.2. Data Cleaning and Outliers Removal

To fill in the missing values, we assumed that a missing value in *AthleticsActivities*, *WaterActivities*, *FitnessActivities*, *DanceActivities*, *TeamActivities, RacketActivities, CombatActivities, NatureActivities, SpecialActivities, OtherActivities* meant that the person did not participate in them. Regarding *NumberOfFrequencies*, we believe that the missing values here are because the person did not come yet to the gym. As such, all the variables mentioned in this paragraph will be filled with 0, because it stands for not participating in the activities variables and means the absence of frequency in *NumberOfFrequencies*. We still need to address the missing values existing in Income, *AllowedWeeklyVisitsBySLA* and *HasReferences*.

We dropped *DanceActivities* and *NatureActivities* from our dataset, as these always displayed the value 0, meaning that no one took part in these two activities. Given this, they will provide us with no relevant information, and can be dropped.

To create meaningful information, we decided to create a new variable called *TotalActivities*, which corresponds to the sum of all remaining 8 activities, to check how is the distribution of clients regarding the frequency of different possible activities. The value counts of this new variable can be found below in figure 8. As we can see, even though we concluded that the absence of participation in each activity was the most common value (and indeed, it is), one could conclude that most people do not participate in the activities. However, with the value counts we can see that this line of thinking is wrong. Although the participation is rather small in most cases, most people attend at least one type of activity, which is an important insight.

When dealing with the missing values in Income, we wanted to be thorough, as just filling with the mode or median could bias our work. Apart from the missing values, we also had values that were 0. These are plausible values, meaning that the person simply has no income. In the hope of better understanding the existing 0s, we decided to explore this variable according to the age, and we could verify that the highest age, when a client has 0 as reported income, is 16. This makes total sense, as

most 16 years olds do not work. However, not all clients under 16 years of age have 0 as income, just 84%.

Out of the 495 missing values in Income, 363 are from clients with less than 16 years of age. As the most plausible situation is for people these ages to not be working, we will fill these missing values with "0". After this, there are still 132 missing values. Median imputation was chosen for the remaining missing values, preserving the integrity of the income distribution.

Regarding the variable *AllowedWeeklyVisitsBySLA's* missing values, we decided to fill the missing values with 0, meaning that the person doesn't have a subscription service, instead this person is allowed to frequent some classes.

Lastly, the only remaining feature with missing values is *HasReferences*. We interpreted these as possible human errors, and the employee simply forgot to input in the system that this person had no references. Also, 0 is the majority class, so these missing values were filled with "0".

We also created a variable called *Days_between*. This variable displays the number of days a person was enrolled in the gym. It was computed through the difference between *EnrollmentFinish* and *EnrollmentStart*.

Now the next step was to deal with outliers, as these have a very strong impact on the analysis of a variable. We first attempted to remove outliers based on the traditional interquartile range, with 1.5. However, as seen previously, we have a high quantity of outliers in our dataset. Removing the outliers based on the IQR, would remove 50,17 % of the dataset, which is an unbearable value. We tried adjusting the 1.5, into other values like 3 or 6, but the percentage of outliers was still very high. As such, we decided to not use the traditional method and instead go with manual filtering.

After several experiments, we removed values where the Age was higher than 78, Income higher than 8000 units, *DaysWithoutFrequency* higher than 900, *LifetimeValue* higher than 4200 and *NumberOfFrequencies* higher than 400. With this approach, we were able to maintain 98,18% of our dataset, a more reasonable value.

Next, we scaled our metric variables. Scaling is important, as it brings all metric variables into the same scale and allows for better comparisons. We went with Min Max Scaler, as it preserves the original distribution of the data, within the new range. Standard Scaler was not the most appropriate one in this scenario, as our data set is highly irregular and so, not normal nor normalike.

Furthermore, the distributions of metric variables were analyzed using descriptive statistics and visualizations such as boxplots, showing a better understanding of variation, central tendency, and outliers in figure 9.

2.3. Feature selection

For the feature selection, we used Spearman's correlation for metric variables, as this correlation method doesn't require necessarily linear relationships. It detects patterns among non-linear relationships. For better interpretation of correlation's results, we used a heatmap of the correlation matrix in figure 10, aiding in understanding the interdependence of the variables. With that in mind, the most correlated pairs were *Age* and *Income* (0.9), *NumberOfFrequencies* and *LifetimeValue* (0.7),

NumberOfRenewals and *LifetimeValue* (0.7), *AllowedNumberOfVisitsBySLA* and *AllowedWeeklyVisitsBySLA* (0.7) and lastly, *AttendedClasses* and *AllowedWeeklyVisitsBySLA* (-0.8).

Based on the correlation's results, we decided to exclude *Income*, *LifetimeValue* and *AllowedWeeklyVisitsBySLA* from our metric features, as highly correlated variables bring the same info into our model, while contributing to the curse of dimensionality. Age was preferred over Income as it was deemed a more significant factor in the context of a gym environment. We believe that the differences in a gym with clients of different ages are bigger than the ones with different incomes. Normally the monthly fee is fixed or somewhat similar, so the income will not be relevant. We also think that customers of different ages engage in different activities rather than customers with different incomes. We decided to drop *LifetimeValue* and *AllowedWeeklyVisitsBySLA*, as we believe that the pairs with which these variables are correlated present more valuable information for our future segmentation.

2.3.1. Principal component analysis (PCA)

The Principal Component Analysis (PCA) is a technique intended to reduce input space, by transforming the data set into uncorrelated variables, to ensure that we only have unique information in our data set. These new variables are called Principal Components. This can only be applied to metric features. Initially, all metric features were copied to a new dataframe (df_pca), and PCA was applied. This step involved fitting the PCA model to our metric features and then transforming these features into principal components. By doing that, we reduced the complexity of our data and ensured no redundancies.

In Figure 11 the scree plot and the cumulative variance plot were used to determine the number of components to retain. These plots helped us identify the 'elbow point'. After that, it was decided to retain six principal components, allowing us to focus on the most relevant aspects of our data. This interpretation was able to be better presented in Figure 12 within Color-coded Significance.

However, PCA must be used and interpreted with caution, as it comes with some challenges, such as hard interpretability of results and the fact that it assumes linearity among the variables. As such, despite our attempt, we did not rely on PCA for dimensionality reduction.

2.3.2. Non-Metric Features

Regarding non-metric features, the feature selection is more challenging, as we can't perform a set of techniques. We decided to remove the temporal variables, *EnrollmentStart*, *EnrollmentFinish*, *LastPeriodStart*, *LastPeriodFinish* and *DateLastVisit,* as temporal variables cannot be transformed using encoding. With that in mind, the OneHotEncoder transformed the categorical features into a format good for modeling, merging with the rest of the dataset.

3. Clustering Algorithms

3.1. Using DBSCAN to remove outliers.

DBSCAN can also be a way to identify and remove outliers in the dataset. The graph from Figure 13 was plotted to determine the appropriate epsilon value, and accordingly, we went with an epsilon of 0.25. After fitting DBSCAN to the metric features, it found 4 clusters. The outliers identified by

DBSCAN, the cluster with values equal to –1, were removed from the dataset that was then used in clustering. After the final solution, these outliers will be reincluded.

3.2. Customer Segmentation

We decided to create 2 different perspectives, to optimize clustering results and allow for a better analysis. We created a perspective called *Demographics,* composed by *Age, NumberOfRenewals, NumberOfReferences* and *TotalActivities*. We believe this is an interesting perspective because these variables are more related to the client itself.

The 2nd perspective is called *Usage,* and contains *DaysWithoutFrequency,NumberOfFrequencies,RealNumberOfVisits, AllowedNumberOfVisitsBySLA* and *AttendedClasses.* This perspective refers directly to a client's frequency in the gym, engagement in classes.

We will be attempting several techniques of clustering for both our perspectives, decide which one to use and later apply them to the final clustering solution for each perspective.

3.3. Demographic

3.3.1. K-Means and Hierarchical Clustering

In this method, the K-Means is normally used with a higher K (higher number of clusters) and then the number of clusters is reduced with the Hierarchical. Firstly, in K-Means, it was chosen 5 as the number of clusters and then we got 0.99 as the R2 score. Then, in the Hierarchical clustering, it was chosen ward for the method and 5 as number of clusters. It was then obtained the same score since we decided to keep all the clusters that were chosen in the K-Means phase.

3.3.2. Mean-Shift

Mean-Shift is a centroid-based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. For this algorithm, bandwidth needs to be estimated and it can be done by changing the parameter called quantile. But why bother doing all of this? This is done because depending on the value of the bandwidth, the number of clusters used in the algorithm will change. After several experiments, it was found 3 as the best K. The r2 score obtained was 0.89, which is worse than the previous method but also a decent result.

3.3.3. DBSCAN

DBSCAN is also known as Density-Based Spatial Clustering of Applications. It is used to find core samples of high density and expands clusters from them. It is normally good when the data contains clusters with similar density. Firstly, it is needed to find the best epsilon which is the maximum distance between two samples for one to be considered as in the neighborhood of the other. This is the most important DBSCAN parameter to choose appropriately for a data set and distance function.

To do that, it was used the Nearest Neighbors with twenty as number of neighbors and decided 0.2 as result. With that value, DBSCAN was performed, and 4 clusters were obtained. However, the R^2 score obtained, which was 0.42, suggests that the model explains a smaller portion of the variance compared to other clustering methods used in the analysis. This could indicate that DBSCAN, while

useful for identifying outliers, may not capture the dataset's structure as effectively as hierarchical methods or mean-shift clustering in this scenario.

3.3.4. Self-Organizing Maps

Lastly, the Self-Organizing Map method, also known as SOM, was performed. The SOM works by adjusting the units to the data in the input space, so that the network is, as best as possible, representative of the training dataset. The SOM was built using a map size of 50x50 dimensions, a random initialization, a 'gaussian' neighborhood function, a batch training and hexagonal topology/lattice. Then, it was trained with an unfolding and fine-tuning phase of 100 iterations. These two phases are used to make the training of the SOM more effective. The unfolding phase spreads the units in the region of the input space where the data patterns are located, and the fine-tuning phase is where small adjustments are made to reduce the quantization error and center the units in the areas where the density of patterns is highest.

After the model was trained, the components planes were used to observe the distribution of the data and they can give other insights such as feature importance, feature correlation and outlier detection. As visualizations tools, U-Maps and Hit Maps were used to understand better the cluster solution.

On top of the SOM model, it was tested K-Means and Hierarchical. With the K-Means, firstly the number of clusters to use was decided by performing the Elbow Method and the Silhouette Method, and 3 was the number chosen. The results were not good since the division of clients among the clusters was very imbalanced and most of the clients were in one of the clusters. For the Hierarchical, 'ward' was found to be the best method and then with the help of a Dendrogram, it was chosen 5 as the number of clusters. The results were better than the ones with K-Means since the distribution of the clients by the clusters had a better balance.

To conclude, the final SOM clustering solution was performed where 4 clusters were used, that had a high imbalance, and it obtained a r2 score of 0.3 which was unsatisfactory.

3.4. Usage

3.4.1. K-Means & Hierarchical Clustering

In this method, the K-Means is normally used with a higher K (higher number of clusters) and then the number of clusters is reduced with the Hierarchical. It was decided 6 as number of clusters for the K-Means and a r2 score of 0.99 was obtained. Regarding the Hierarchical, it was used 'ward' because it was the best method and then a Dendrogram was done to decide the number of clusters, which ended up being 4. The r2 score obtained was 0.96.

3.4.2. Mean-Shift

Here we did the same as explained in the Mean-Shift part of the *Demographic* perspective. After several experiments, it was found 3 as the best K. Even with this low number chosen, the dispersion of the clients was not balanced at all and so the results were already expected not to be good. The r2 score obtained was 0.37, which is not good as expected.

3.4.3. DBSCAN

Like in the other perspective here it was used the Nearest Neighbors with twenty as number of neighbors and decided 0.07 the epsilon value. With that value, DBSCAN was performed, and 5 clusters were obtained. As r2 score it was obtained 0.56.

3.4.4. SOM

Lastly, the SOM clustering algorithm was performed. It was built and trained in the same way that was done in the Demographic Perspective. Component Planed, U-Maps and Hit Maps can also be seen. Like in the previous perspective, on top of SOM, K-Means and Hierarchical were used with the same methods and in the same way. On K-Means 3 was the number of clusters used. On Hierarchical it was used ward as the method and 4 as number of clusters. For the final SOM solution, 3 clusters were used and a r2 score of 0.14 was obtained, so it also was unsatisfactory like in the demographic perspective. This suggests and the complexity of this algorithm suggests that there could be future improvements on the results of this method.

3.5. Final Clustering Solution per Perspective

After applying different algorithms and seeing the results, the chosen method for the final cluster solution incorporated K-Means and Hierarchical algorithms. After several experiments to identify the best number of clusters for each perspective, it was decided that 3 was the best value for both perspectives.

Regarding the Demographic perspective, Cluster 0 includes the youngest clients with the highest number of renewals and highest number of references. Cluster 1 has the oldest clients and the medium value of number of renewals and the lowest value in the number of references. Cluster 2 consists also of young clients with the lowest number of renewals. The variable *TotalActivities* does not present significant differences between the three existing clusters.

For the usage clustering Cluster 0 has the highest *RealNumberOfVisits* and number of frequencies lowest number of days without frequency, which includes highly active clients with the most class attendance, suggesting high engagement. Cluster 1 has the lowest *RealNumberOfVisits* and number of frequencies, and the highest days without frequency, representing the clients that have the lowest levels of attendance. In Cluster 2, which consists of clients with highest number of attended classes and low *RealNumberOfVisits*, number of frequencies and number of days without frequency, signaling an opportunity for targeted engagement and retention strategies. The variable *AllowedNumberOfVisitsBySLA* does not present significant differences between the three existing clusters. The clustering profiles of these perspectives' clusters are in figure 14.

Finally, we arrive at the merging of perspective part, here we have two problems to address. The first one is that there are too many clusters (9 clusters), so a reduction of this number is needed. The second one is about the size of the clusters, there are clusters with too few clients compared to total number of existing clients, an easy example is a cluster that has only 257 clients. So, to address both these problems, first we merged the lowest frequency clusters into closest clusters, we decided to make a threshold with 500 clients as minimum and then, we used Hierarchical Clustering and with that produced a Dendrogram that helped in defining the final number of clusters. With that Dendrogram that you can see in figure 15, it was concluded that 3 was the best number of clusters.

The merged cluster analysis shows three key customers in the gym, as you can see in figure 16. Cluster 0 has 3701 clients and suggests a younger age type of client with higher gym frequency, class attendance and renewals, indicating a highly active and possibly well-suited segment. Cluster 1 has 2175 clients and consists of older clients with moderate gym usage, showing a moderate frequent attendance and a moderate number of renewals. Lastly, Cluster 2 is the biggest cluster with 8513 clients, shows a younger age group with the least gym involvement and renewal number that may require additional engagement and tailored marketing strategies to increase activity levels.

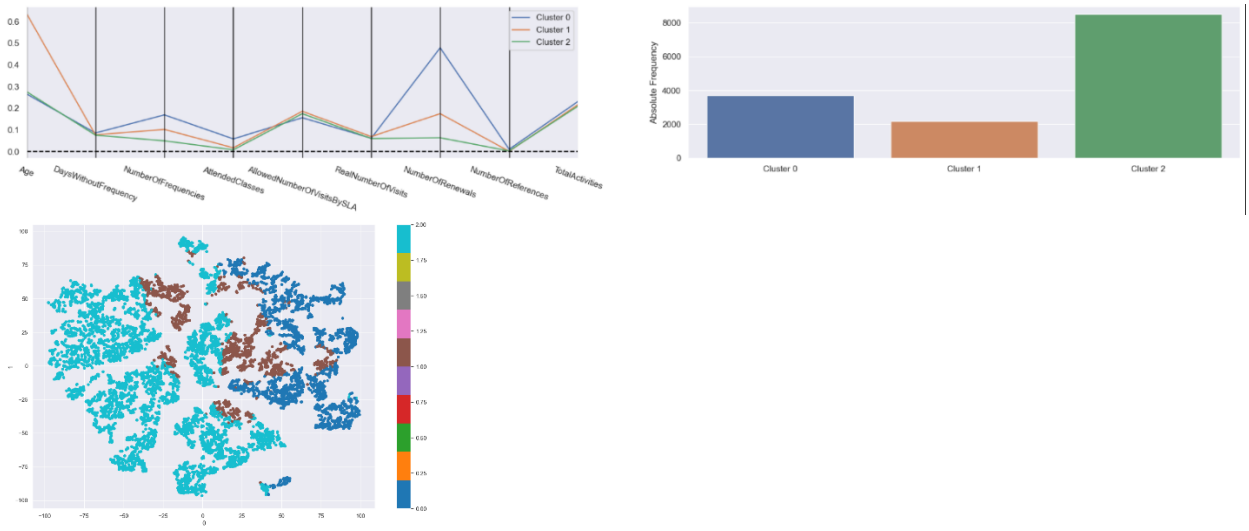


Figure 17. Final clustering profiling with t-sne

Those three clusters presented similar values of days without frequency, allowed number of visits and total number of visits. Another similarity is on attended classes between Cluster 1 and Cluster 2.

Related to these clusters, there are also some characteristics of each cluster to be obtained when compared to the different activities that exist. Cluster 0 has the highest water activities and team activities practice. The higher value of water activities is expected since they are the younger group and normally the gym's swimming pools have a lot of young people that are learning how to swim or improving their capabilities. Cluster 1 has the highest special activities and other activities practice and the lowest combat and team activities practice. Cluster 2 has the lowest special activities practice and paired with Cluster 1 they have the highest values in fitness activities and similar values water activities also. Cluster 0 and Cluster 1 have similar values Combat Activities.

3.5.1. Assess feature importance and reclassifying outliers.

Here the objective is to understand better the results of the clustering solution and talk about the reclassification of the outliers that were removed earlier with the DBSCAN.

Using the R2 it could be seen what proportion of each variables total variance in the data is explained between the clusters. As you can see in Figure 18, age and number of renewals are the 2 variables better explained followed by number of frequencies and attended classes. The rest of the variables are less explained between the clusters.

Then a Decision Tree Classifier was used with the Gini importance criterion to also evaluate the clusters. It is estimated that on average, the model can predict 97.64% of the customers correctly. In

Figure 19 the fully grown decision tree can be visualized, and with that see what are the separations made and the variables and thresholds used to do them.

4. Marketing Strategies/Business Applications

Cluster 0: This cluster has the youngest clients, most attended classes, the one that goes more times per week to the gym and by far the highest number of renewals. With that in mind, a satisfactory marketing approach would be to launch fitness challenges and online competitions on social media platforms for younger audiences that are relevant in this clustering and use influencers to promote those events to increase brand visibility and appeal to this age group's digital habits. This cluster also displays the highest number of people attending *WaterActivities* and *TeamActivities.* With this in mind, we recommend the gym to make sure that the swimming and *TeamActivities* materials are always in good state, as these will suffer more wear, since are more used by the people of this cluster. To keep these clients satisfied, the gym should always invest in good equipment.

Cluster 1: This cluster is the one with the highest age and both second largest number of frequencies and renewals. A relevant marketing strategy for this group could be create wellness programs focused on senior age people. By doing this the academy could develop programs like low-impact aerobics classes or yoga, which can cater to their specific health needs, interests and be a well-suited place for social interaction. This is the cluster where people attend the most *SpecialActivities.* Given this fact, the gym should make sure that it has the right disability inclusive infrastructure and workers prepared to help people with disabilities.

Cluster 2: As this is the cluster with the least number of renewals and frequencies, and presents a high percentage of young people, a possible marketing strategy could be decreasing renewals costs for these clients, such as contractual costs or maybe creating a new monthly payment category for clients of younger ages, to ensure that these continue linked to the company. Since they show up the least in the gym, and don't take part in most of the classes provided by the gym, the gym could create a new activity, to attract them. A good example of a new type of activity could be padel. Even though the gym already offers *RacketActivities,* but if so far, this has not been enough to attract young people, maybe it should redirect its focus towards a trending sport nowadays, like Padel. Also, by looking at the frequency of these cluster's clients in the different activities, we see that the 3rd most frequent activity is *CombatActivities*. A possible approach could be increasing the gym's offer in terms of this activity, because there are a lot of different types of combat activities and diversifying it could help captivating its customers.

5. Conclusion

In conclusion, the project's goals have been successfully achieved by applying the data mining techniques learned in class to a real-world project scenario, thereby creating relevant client segmentation.

Even though we managed to reach our conclusions, it is our belief that we faced some challenges throughout this project. Firstly, the dataset was composed of a very high percentage of outliers, which complicated our analysis. To bypass this struggle, manual filtering was used as it was unreasonable to remove outliers with the IQR, since the percentage of outliers was very high.

Another obstacle was reducing the input space. It was very hard and at the same time essential for our project since dimensionality complicates the process of creating clusters. We based our feature selection mostly on common knowledge and relevance for possible cluster computing.

A technique that proved very useful in starting the process of clustering was dividing features into different perspectives, in our case, *demographic* and *usage. It* was hard to create the perspectives since we interpreted our dataset as homogeneous in a certain way and several combinations could make sense here. Nevertheless, we believe that the approach we adopted was the best one.

After this, we made several attempts with different types of clusters algorithms, different parameters, and different possible combinations of numbers of clusters. This proved also to be very challenging, as very seldom did the clusters display a balanced number of clients and well-shaped clusters.

For our final clustering solution, we decided to go with K-Means and Hierarchical algorithms. After several attempts, we believe that this was the best solution that we were able to achieve. This created 3 clusters for each perspective and 3 clusters after merging them.

Lastly, as requested, we were able to characterize our clusters, and based on their characteristics, we provided possible marketing strategies to the XYZ Sports Company. We strongly believe that the suggestions we presented will optimize the company's customers' satisfaction, by delivering personalized services and focusing on new possible relevant services.

For future improvements, more testing and experiments could be done with more time. Specifically in the SOM algorithm, more fine tuning could probably improve the results exponentially since it's a complex clustering method.

6. References

Mukaka M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal : the journal of Medical Association of Malawi, 24(3), 69–71.

https://www.kaggle.com/code/yesilyurttemre/customer-segmentation-in-the-marketing-campaign

Sklearn's documentation (2023), Clustering: Plot hierarchical clustering dendrogram Available at :

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

Sklearn's documentation (2023), Clustering: sklearn.tree.DecisionTreeClassifier Available at :

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Sklearn's documentation (2023), Clustering: sklearn.cluster Available at :

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster

7. Apendices

```
<class 'pandas.core.frame.DataFrame'>
Index: 14942 entries, 10000 to 24941
Data columns (total 30 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         14942 non-null  int64
 1   Gender                      14942 non-null  object
 2   Income                      14447 non-null  float64
 3   EnrollmentStart             14942 non-null  object
 4   EnrollmentFinish            14942 non-null  object
 5   LastPeriodStart             14942 non-null  object
 6   LastPeriodFinish            14942 non-null  object
 7   DateLastVisit               14942 non-null  object
 8   DaysWithoutFrequency        14942 non-null  int64
 9   LifetimeValue               14942 non-null  float64
 10  UseByTime                   14942 non-null  int64
 11  AthleticsActivities         14906 non-null  float64
 12  WaterActivities             14905 non-null  float64
 13  FitnessActivities           14907 non-null  float64
 14  DanceActivities             14906 non-null  float64
 15  TeamActivities              14907 non-null  float64
 16  RacketActivities            14905 non-null  float64
 17  CombatActivities            14909 non-null  float64
 18  NatureActivities            14895 non-null  float64
 19  SpecialActivities           14898 non-null  float64
 20  OtherActivities             14907 non-null  float64
 21  NumberOfFrequencies         14916 non-null  float64
 22  AttendedClasses             14942 non-null  int64
 23  AllowedWeeklyVisitsBySLA    14407 non-null  float64
 24  AllowedNumberOfVisitsBySLA  14942 non-null  float64
 25  RealNumberOfVisits          14942 non-null  int64
 26  NumberOfRenewals            14942 non-null  int64
 27  HasReferences               14930 non-null  float64
 28  NumberOfReferences          14942 non-null  int64
 29  Dropout                     14942 non-null  int64
dtypes: float64(16), int64(8), object(6)
memory usage: 3.5+ MB
```

Figure 1. Initial Features

```
Age                             0
Gender                          0
Income                        495
EnrollmentStart                 0
EnrollmentFinish                0
LastPeriodStart                 0
LastPeriodFinish                0
DateLastVisit                   0
DaysWithoutFrequency            0
LifetimeValue                   0
UseByTime                       0
AthleticsActivities            36
WaterActivities                37
FitnessActivities              35
DanceActivities                36
TeamActivities                 35
RacketActivities               37
CombatActivities               33
NatureActivities               47
SpecialActivities              44
OtherActivities                35
NumberOfFrequencies            26
AttendedClasses                 0
AllowedWeeklyVisitsBySLA      535
AllowedNumberOfVisitsBySLA      0
RealNumberOfVisits              0
NumberOfRenewals                0
HasReferences                  12
NumberOfReferences              0
Dropout                         0
dtype: int64
```

Figure 2. Inital Existing Nans

```
: df[metric_features].columns

: Index(['Age', 'Income', 'DaysWithoutFrequency', 'LifetimeValue',
         'NumberOfFrequencies', 'AttendedClasses', 'AllowedWeeklyVisitsBySLA',
         'AllowedNumberOfVisitsBySLA', 'RealNumberOfVisits', 'NumberOfRenewals',
         'NumberOfReferences'],
        dtype='object')
```

Figure 3. Metric Features

```
df[non_metric_features].columns

Index(['Gender', 'EnrollmentStart', 'EnrollmentFinish', 'LastPeriodStart',
       'LastPeriodFinish', 'DateLastVisit', 'UseByTime', 'AthleticsActivities',
       'WaterActivities', 'FitnessActivities', 'DanceActivities',
       'TeamActivities', 'RacketActivities', 'CombatActivities',
       'NatureActivities', 'SpecialActivities', 'OtherActivities',
       'HasReferences', 'Dropout'],
      dtype='object')
```

Figure 4. Non Metric Features

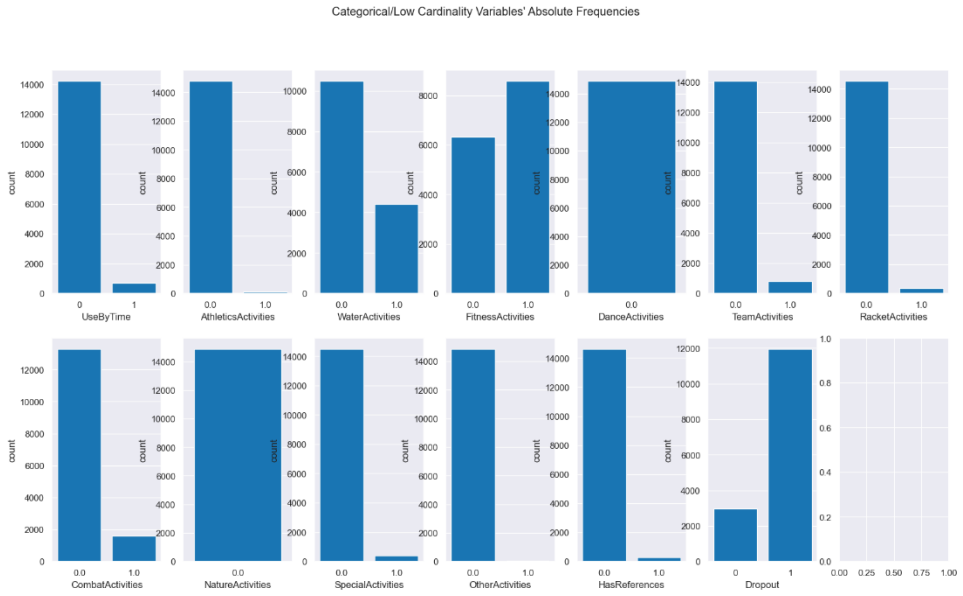


Figure 5. Categorical Variables' Bar Charts

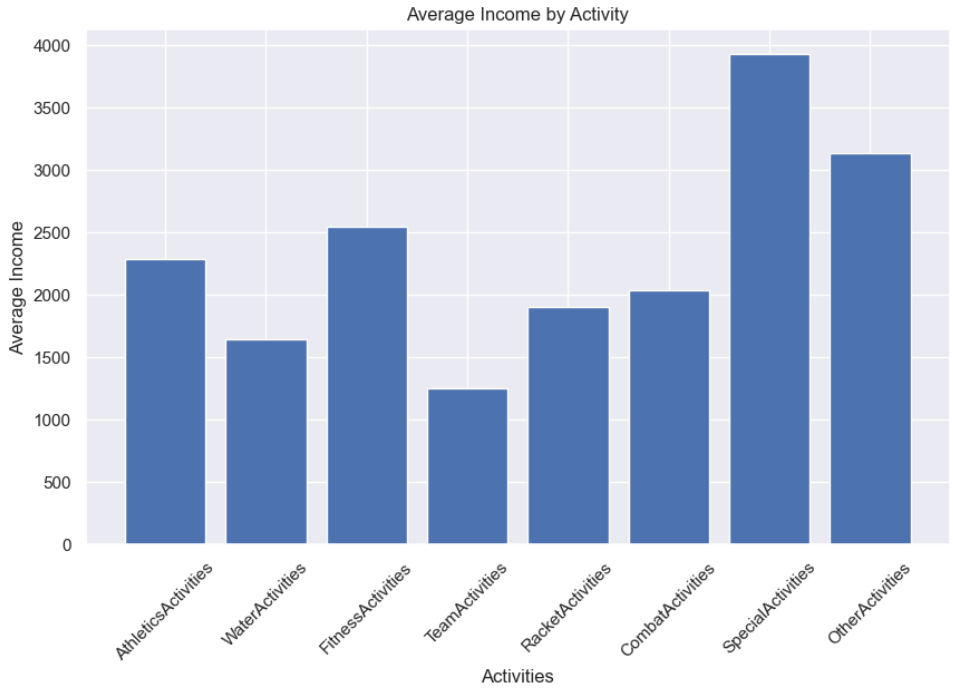


Figure 6. Average Income by Activity

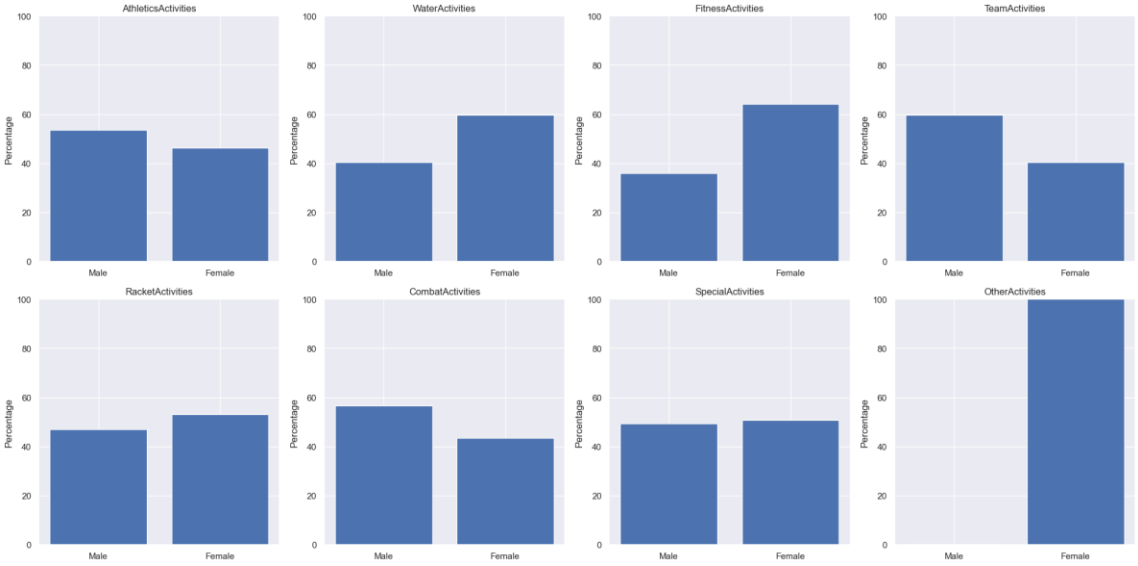


Figure 7. Distribution of men and women by the different activities

```
TotalActivities
1.0    13606
2.0     1187
3.0       98
0.0       39
4.0        9
5.0        2
```

Figure 8. TotalActivities' value counts

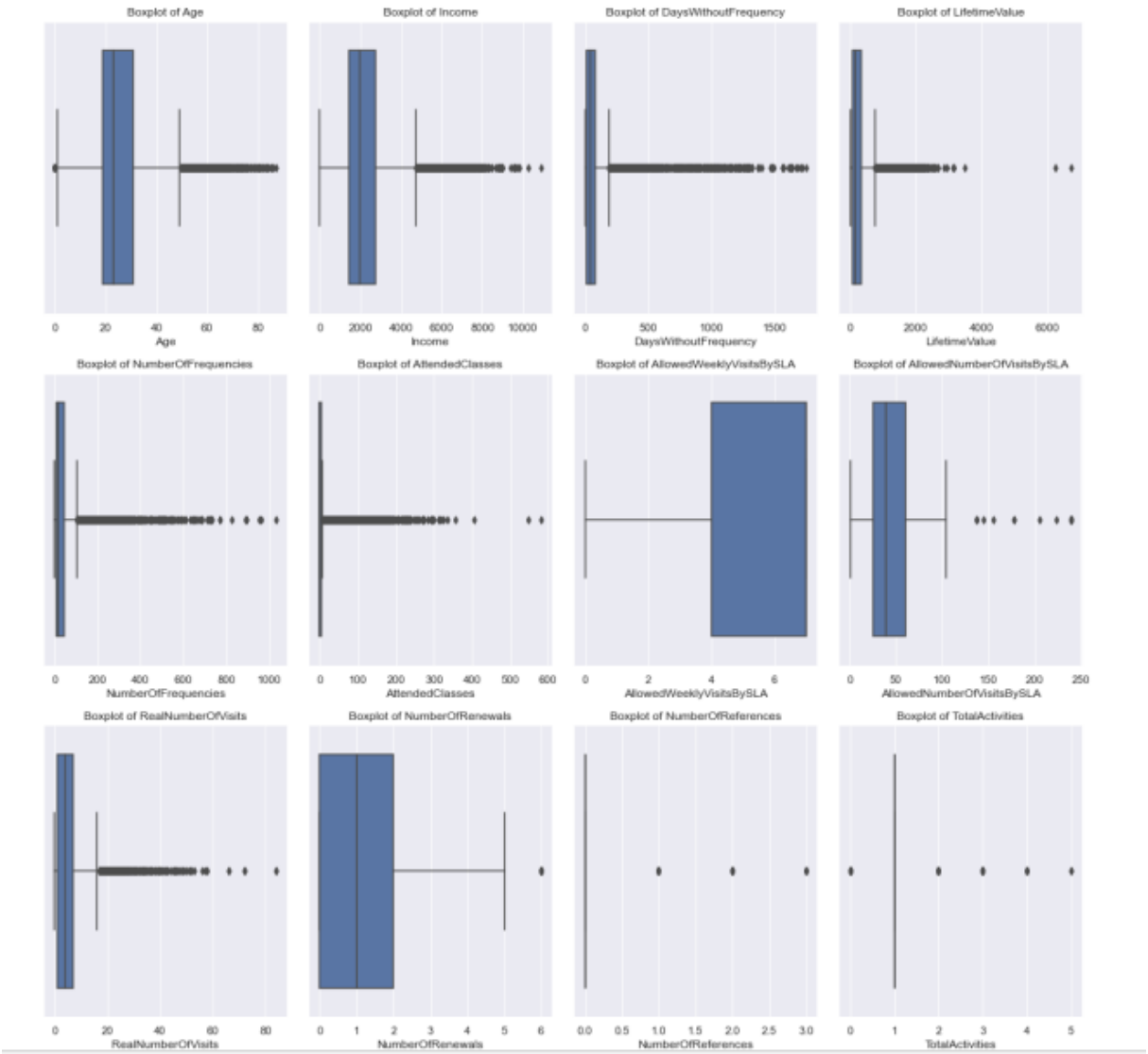


Figure 9. Metric-features' Boxplots

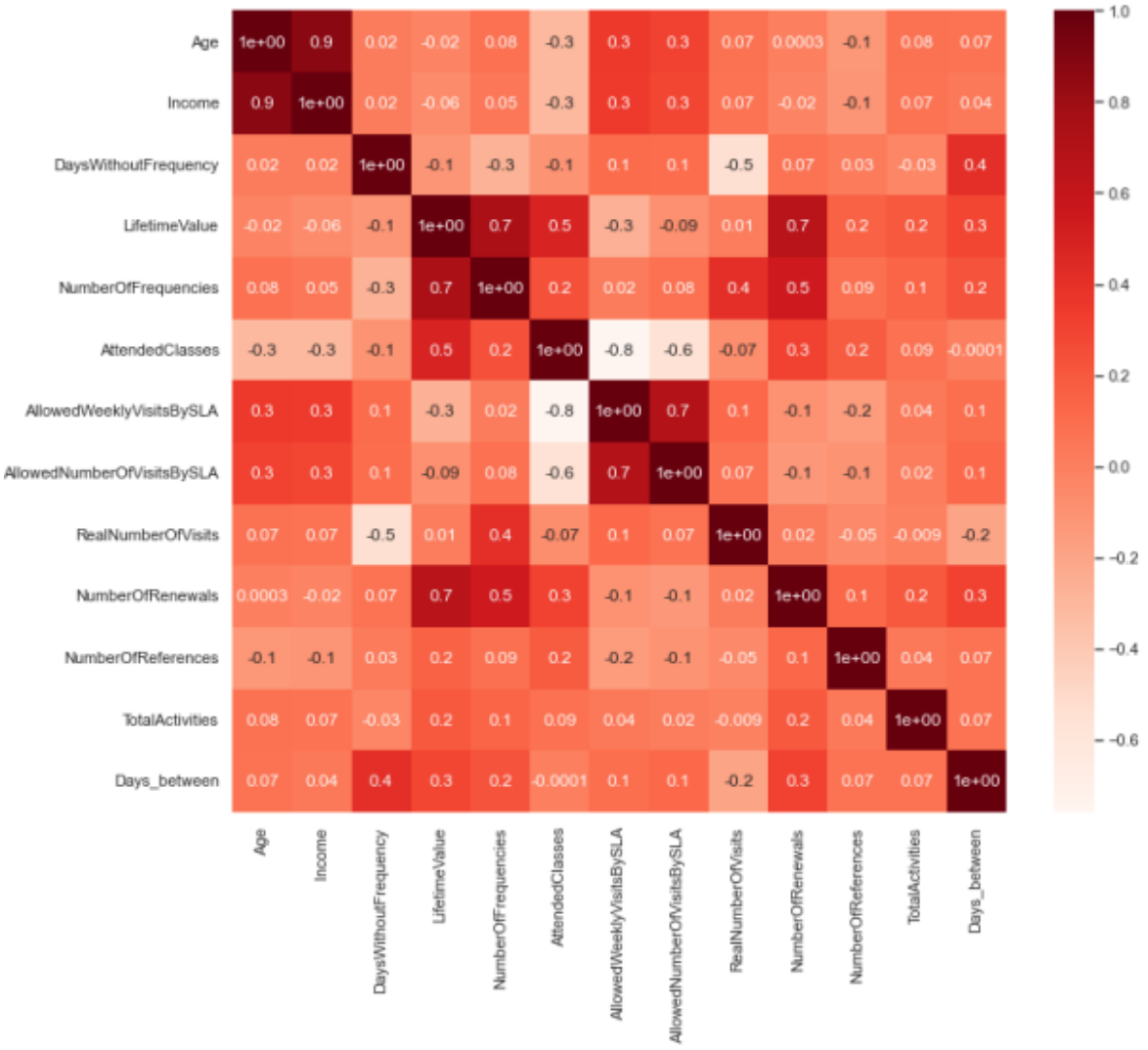


Figure 10. Spearman's Correlation Heatmap

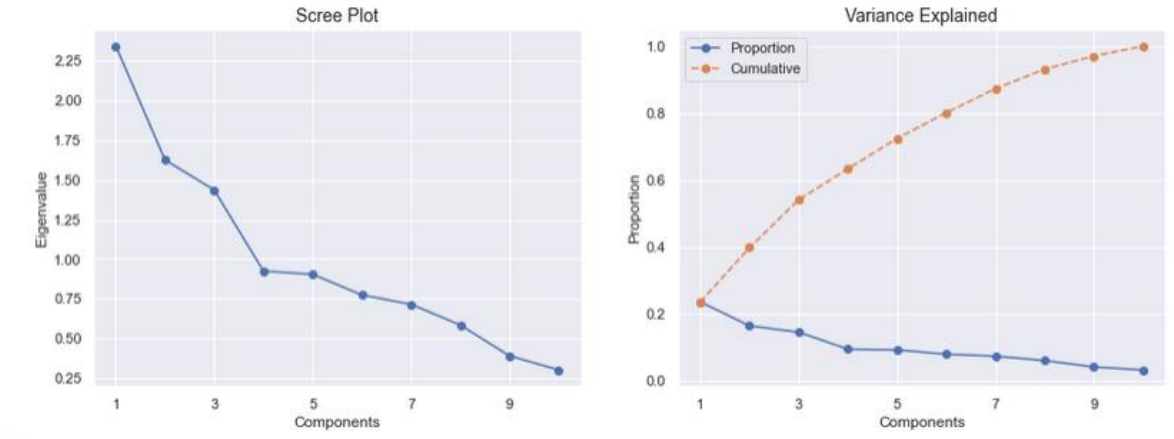


Figure 11. Scree and Variance Plot

	PC0	PC1	PC2	PC3	PC4	PC5
Age	0.054066	0.988075	0.021468	0.100967	-0.084022	-0.053147
DaysWithoutFrequency	0.168502	-0.005738	-0.519967	0.640772	0.519997	-0.138344
NumberOfFrequencies	0.651986	0.148294	0.451969	-0.367347	0.382917	-0.192285
AttendedClasses	0.429687	-0.218845	0.384522	-0.006361	-0.020545	-0.363264
AllowedNumberOfVisitsBySLA	-0.064746	0.364193	-0.136470	-0.254665	0.398661	0.749923
RealNumberOfVisits	0.066967	0.165757	0.387742	-0.370016	0.323887	-0.000569
NumberOfRenewals	0.949526	-0.074161	0.188137	0.198600	-0.077807	0.104876
NumberOfReferences	0.164679	-0.122163	0.017893	0.021558	-0.021965	-0.147611
TotalActivities	0.242725	0.061370	0.029381	-0.033024	-0.014809	0.047173
Days_between	0.734437	0.035799	-0.602400	-0.298519	-0.059197	-0.056613

Figure 12. Color Coded Significance

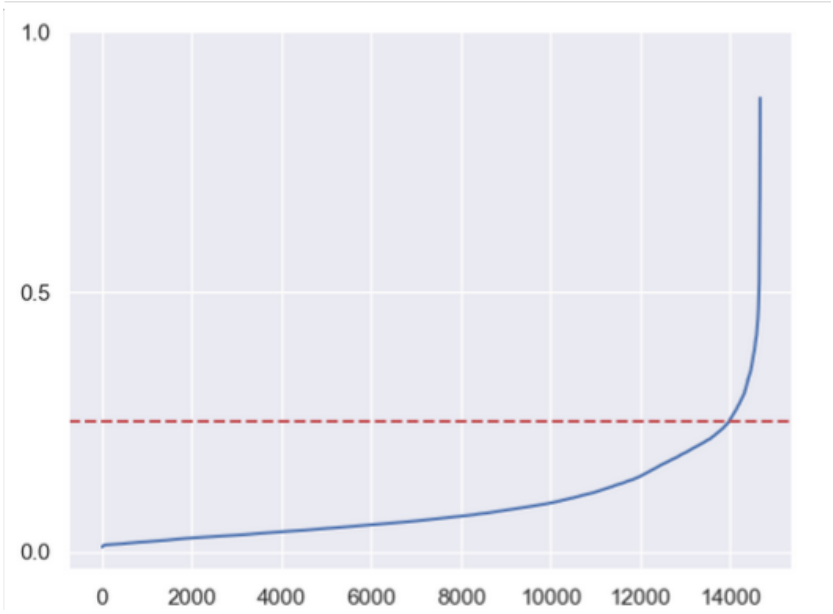
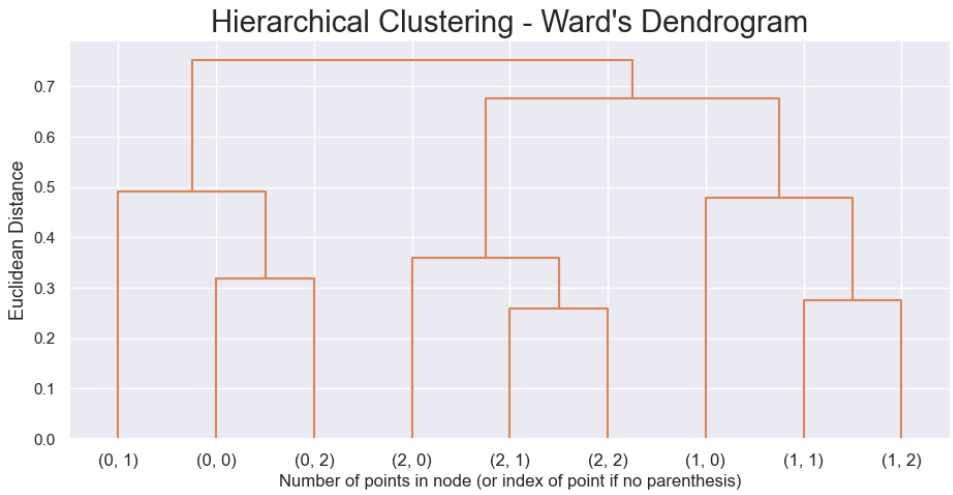


Figure 13. K-distance graph to find out the right eps value



13

Figure 15: Ward's dendrogram, illustrating the hierarchical clustering of data based on Euclidean distances

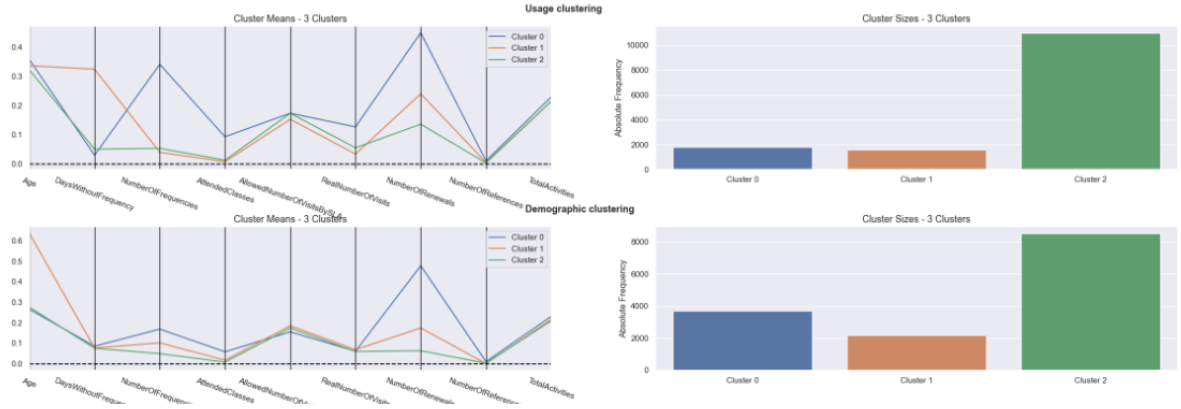


Figure 16: Demographic's and Usage's perspective profile report.

Age	0.560853
DaysWithoutFrequency	0.001543
NumberOfFrequencies	0.184564
AttendedClasses	0.126883
AllowedNumberOfVisitsBySLA	0.012797
RealNumberOfVisits	0.002037
NumberOfRenewals	0.658405
NumberOfReferences	0.008774
TotalActivities	0.023637

Figure 18: Feature importance through R2

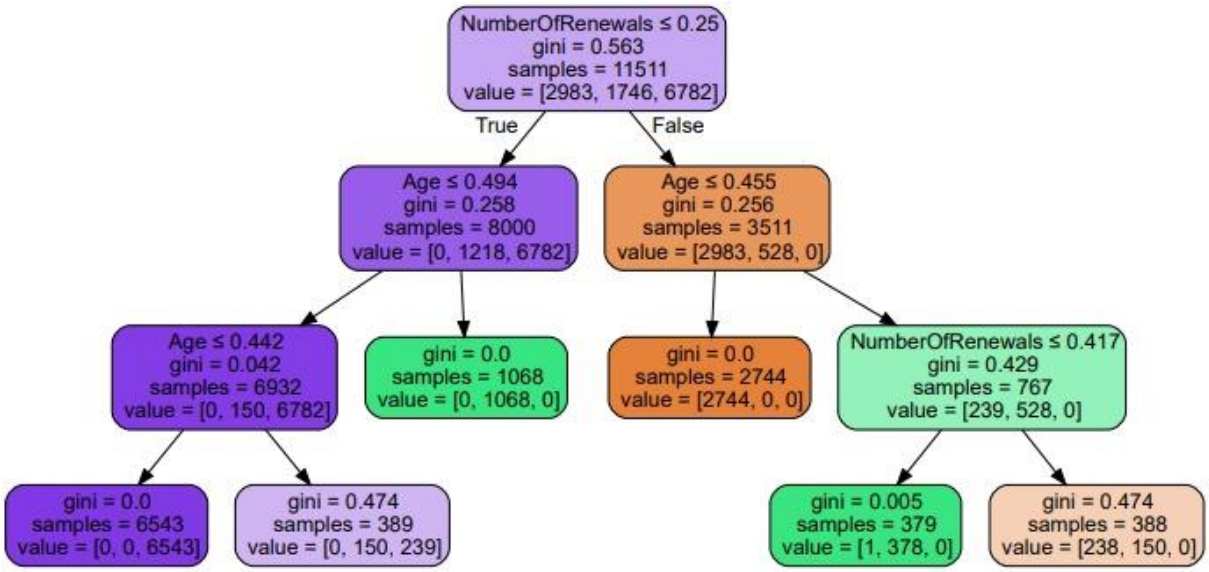


Figure 19: Decision Tree