NOVA IMS
Information Management School

Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

XYZ Sports Company:

Customer Segmentation

Group 79
Pedro Miguel Silva Carvalho, number: 20230487

Maria Carlota Fortunato, number: 20230382

January, 2024

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

INDEX

1. Introduction

The aim of this project is to cluster different segments of customers/users of a fitness facility, known as XYZ Sports Company. Clustering is important to understand each type of customer we are dealing with in our business, how can we target them more efficiently and how can we answer their specific needs. In the fitness/sports business, it is crucial to develop a specialized segmentation as there are many different needs, so more than improving our connection with our customers, we can develop programs that are most beneficial to them.

For the business side this is an example of an efficient approach where, by knowing our specific target, we can develop marketing campaigns directly to them. On the other hand, we get the advantage of saving resources and efforts by not selling products that have no value to certain groups, and instead directing them only to those who use them.

Having the above in consideration, as well as the requests for the deliveries, we will do a data driven cluster approach that will group users under three perspectives, which, in turn, will allow us to understand them under the value and demographics of each segment, as well as gain insights into the different sports activities that customers prefer to participate in. Furthermore, we will develop specific marketing campaigns for each unique cluster. We will use historical data from June 1st 2014 to October 31st 2019 from XYZ Sports Company.

2. Data Exploration

The first step when performing any project of this kind is to get an understanding of which type of variables the data set contains and how is the dataset actually structured. Therefore, we started by importing the given records from the gym. The data set is composed of 14942 rows and 30 columns. ID was set as an index because it is a unique identifier of each user. We immediately identified a duplicate entry and decided to delete it since it was only one.

A key step to perform when analyzing a dataset is to verify if the data types are stored according to the information provided from the metadata. In fact, many changes had to be made. Starting with the dates, for further manipulation we decided to use the method 'to_datetime' that stores them in the precise date format and not as a string. Then, regarding binary variables, we decided to transform all that applied to 0 and 1. We used Int64 in order to preserve Not a Number (Nan) untouched by now. Fourteen columns were affected in this step. We also stored 'NumberOfFrequences' as an Int64 to facilitate its manipulation.

With these steps, the data we got became more consistent which allows us to move on to pre-process data.

3. Data Pre-processing

3.1 Feature Split

The initial data contained, as it commonly happens, a mix of features, some of which numerical and others categorical or binary. Therefore, we initiated the data preprocessing part by categorizing the dataset features into two groups: non*_metric_features* and *metric_features*. The first one includes

variables such as Gender and binary variables that indicate if the user is enrolled in specific activities, the second contains numerical attributes such as age and income.
This split allows us to, in the course of our project, apply preprocessing techniques specific to each category, ensuring that we handle them in the best way.

3.2 Coherence Checking

Amongst all the volume of data we have, there might be some inconsistencies on how the data is stored or how some variables may have some missing records that can be easily imputed, after checking its coherence. We identified two situations of that kind. The first one concerns the columns 'HasReferences' and 'NumberOfReferences'. The former had twelve rows with NaN values. However, after checking the number of references all of them had at least one reference. As a result, we changed these records to a binary value of 1.

Regarding the columns 'Age' and 'Income', we checked if there was the inconsistency of a minor earning any money. In fact, there were situations like that. At this stage, we made our first assumption: children cannot earn money until they are 16, only at that age they can earn some sort of allowance from parents and that is the value declared as income. Prior to that age, the parents imputed their own income, which is not valid. As so, we gathered all observations that fulfilled this criterion (20 rows) and set the income value to 0. Further value inconsistencies that we did not uncover immediately will be dealt with in the next sections.

3.3 Feature Engineering

In a data mining project such as this one, this section is of high value. After some brainstorming and trying to understand the business, our aim was to extract meaningful information from variables that were not of great use the way they were presented, and transform them. We also tried to build different ratios that helped with the analysis of consumer behavior.

Overall, we built three date variables with method *dt.days* to have countable lengths. In the new variable 'EnrollmentLenght' there was a problem with people being enrolled for zero days. We assumed that this was a technical problem by the gym system and we imputed the date present in this user's 'LastPeriodFinish'. Two ratios and one cumulative variable were also made. **Figure 1** in the annex provides information on all these new variables. After this step, we dropped all date variables as they provided no longer relevant information for the clustering task. To understand some relationships we plotted every metric feature relationship with each other (**Figure 2** in the annex), which will be also useful in other sections.

3.4 Outliers Removal

Taking care of outliers is a relevant step because outliers can influence clusters results and give us false targeting groups for our marketing strategy. As a common practice, and in order to avoid losing relevant information, we opted by doing a manual filtration at this stage. For that, we built initial visualizations: boxplots for metric features and count plots for non-metric features (**Figures 3** and **4**, respectively). Our main filtration was based on observation of the boxplots. **Figure 5** in the annex shows the cut-off values we used. At this stage, we did not consider non-metric features to have outliers; however, we decided to drop univariate activities: 'NatureActivities' and 'DanceActivities' as they will provide no splitting criteria for clustering. This was based on the assumption that these are new activities in the gym so at the data collection date no user was enrolled yet.

After this filtration, we keep 99.54% of the original data. We decided to take care of outliers at this stage to avoid performing this step multiple times due to the introduction of new variables in the previous section and to allow the usage of the imputation method in section 3.6, which is sensitive to outliers and noise.

3.5 Metric Features Normalization

In data analysis it's very common to have features that use different scales, which can cause issues when we use clustering algorithms that depend on distances between data points. The normalization process is used as a way to solve this problem, and in this case our group opted for the Min-Max scaling technique. This method transforms the metric features to a common scale, usually between 0 and 1. It starts by finding the minimum and maximum values observed and then rescaling them, by doing this proportionality is ensured (the smallest value becomes 0 and the largest becomes 1, everything else in between is proportional).

3.6 Missing Values

As it is known, missing values represent an absence of information, meaning that there is no value stored for a certain variable or participant. These values need to be correctly handled or there's a risk that there will be inaccurate results. In our dataset four metric features had missing values and all the activities had missing values.

For the metric features we chose to use K-Nearest Neighbors (KNN) Imputation, a technique that fills in the missing values by using their k-nearest neighbors. It starts by computing the distances between the data points that have missing values and the other data points in the dataset, then it selects the k data points that are closest to it (using the Euclidean distance) and, after that, it uses the values found to fill in the missing values. For the non metric features we opted to fill in the missing values using the mode, most frequently occurring value of each respective feature. Since missing values were at maximum 44 we chose this simple method that is not based on distances and is appropriate for categorical features.

3.7 Outliers Dataset pre-processing & DBSCAN

Since we will later introduce outliers again, to provide clusters to these specific customers and potential new ones, we also need to take care of this dataset. For that, we followed the same steps described in the previous two points. We normalized with MinMax scaler and we took care of missing values the same two ways we did before.

After this, we applied DBSCAN density based technique, a robust approach to identify new potential outliers that we might not have considered relevant before and remove noise. With it, points that are neither border points nor core points have a label -1 and are considered outliers. To get the threshold we use KNN algorithm to understand distance between neighbors, only considering metric features. Then, we plot it and use the elbow method to know the right eps. We used an epsilon of 0.45. We identified extra 347 outliers with DBSCAN. This plot is present in **Figure 6** in the annex. We did this at this stage to ensure that no rows had missing values when computing the distances.

4. **Feature Selection**

Our group used a correlation matrix to identify potential redundancy among the metric features, setting a threshold of 0.7. We identified several pairs that might have variables considered for removal. This was important information to take into account when deciding how to divide our variables for clustering. **Figure 7** in the annex shows the correlation matrix for our metric features.

- **Age and Income:** Correlation = 0.88
- **Lifetime Value and Number of Renewals:** Correlation = 0.71
- **Lifetime Value and Enrollment Length:** Correlation = 0.79
- **Allowed Number of Visits and Allowed Weekly Visits:** Correlation = 0.71
- **Average Classes per Visit and Allowed Weekly Visits:** Correlation = -0.86
- **Visits Ratio and Real Number of Visits:** Correlation = 0.72
- **Enrollment Length and Number of Renewals:** Correlation = 0.91

We opted to use three clustering methods: the first one based on value, the second based on demographics/social aspects and the last one based on activities.

For the clustering related to the customer value we considered the variables: *LifetimeValue*; *NumberOfFrequencies*; *NumberofRenewals*; *EnrollmentLenght*; *VisitsRatio*; *AverageClassesperVisit*; *AttendedClasses*. As mentioned in our notebook, we found that all of these variables provided insights into the customer's long-term relationship with the service, making them fit to be grouped together. Due to the high correlation that *EnrollmentLenght* had with two other variables, we opted to remove it in order to avoid issues such as multicollinearity and redundancy.

In the clustering based on demographics, we aimed to get insights into individuals with similar characteristics and who also share similar recent engagement patterns. For that, we considered the variables: *Age*; *Income*; *Gender*; *Recency*. After noticing the high correlation between the variables Age and Income, we opted to remove age from the selected features for this clustering, as we attempted both ways and Income was providing greater information in the final solution.

Finally, in the clustering for activities, we took into account the variables that represented the participation of customers in the different sports activities, aiming to identify groups with similar preferences. We used the variables: *AthleticsActivities*; *WaterActivities*; *FitnessActivities*; *TeamActivities*; *RacketActivities*; *CombatActivities*; *SpecialActivities*; *NumberofActivities*. The variable *OtherActivities* was not included due to lack of variance it had, which we believed could introduce noise to the clustering.

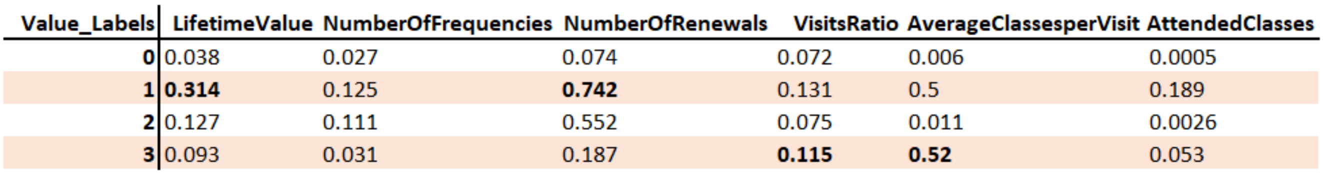
This way with these specific feature selection methods we go over the curse of multidimensionality. Adding to that we decided to not use Principal Component Analysis since, even though it reduces the input space, it provides no value to explain characteristics of customers, the factor we are after with this project.

5. Individual Clustering

5.1 Value Perspective- K-means Algorithm

For this first approach we used the K-means algorithm, with it we could only choose metric features, as it is based on distances and binary variables would result in a too great impact since they can only be 0 and 1. What we did first was to create an independent data set with only the value variables we chose. Then, we ran an initial instance of k-means to get an initial prediction. The initiation 'k-means++' was used to ensure that the initial centroids were not close to each other, providing a better final solution.

To understand the correct number of clusters for this scenario we run this algorithm eleven times to cluster from a range of one to eleven different groups. With this, we stored information about inertia and plotted it. As seen in **Figure 8** in the annex, the number of clusters before the inertia slope decline slows down is four, being that the number we went with. We also computed the silhouette score, which gave the same information. After running the final algorithm, we got 4 clusters. Cluster 0 is the one with the lower value customers as they barely come to the facilities, do not engage in many activities and do not provide a large lifetime value to us. Cluster 1 represents our high value customers. They have a large lifetime value, they are the ones renewing their contract many times, and they are the most engaging customers. We assumed them as our most loyal users. Clusters 2 and 3 lie in the middle with the nuance that cluster 3 have low-value but high engaging customers. As such, they might come fewer times to the facility but when they come, they tend to participate in the classes.



Value_Labels	LifetimeValue	NumberOfFrequencies	NumberOfRenewals	VisitsRatio	AverageClassesperVisit	AttendedClasses
0	0.038	0.027	0.074	0.072	0.006	0.0005
1	**0.314**	0.125	**0.742**	0.131	0.5	0.189
2	0.127	0.111	0.552	0.075	0.011	0.0026
3	0.093	0.031	0.187	**0.115**	**0.52**	0.053

Table 1- Value Perspective Cluster Centroids

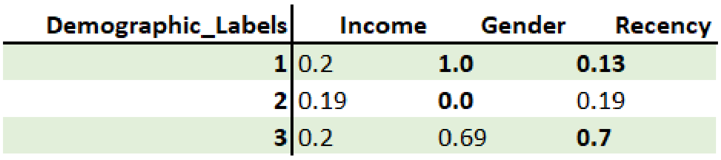
To ensure the quality of our final solution we computed R-squared which gave a result of approx. 72.4% of the variance of our clusters being explained by these variables, which is considered a good value. We also tested the SOM algorithm, however, besides not providing such clearer division, the R-squared was only about 32% being those the reasons we opt for K-means.

5.2 Demographic Perspective- K-prototypes Algorithm

Our dataset is composed of many binary features so it would be of value to find an algorithm that enables us to use them when clustering. As our self-studies task we found K-prototypes. This algorithm combines k-means properties for metric features and k-modes for non-metric features. The former uses the squared Euclidean distance as a dissimilarity measure whilst the latter uses the number of mismatches in categories. Because k-prototypes were built having K-means as a basis it keeps its efficiency (Huang, 1998).

Firstly, we defined the features to use and then we had to specify which column indexes are the categorical ones, which in the demographic perspective was just the 'Gender' column. Then we built the "elbow" plot after iterating for a range of 1 to 11 possible clusters. **Figure 9** in the annex shows this plot where we identified three as the optimal number. To get the final solution we ran the algorithm with this cluster number, added the labels to the data frame and grouped the values by each label.

Cluster 0 is composed of only women, with middle-low income. Cluster 1 has only male users that do not stay much time without coming to the facilities. Cluster 3 is the more discrepant in terms of recency where users here have stayed a long period without coming to the gym, they are mainly women and have middle-low income as well.



Demographic_Labels	Income	Gender	Recency
1	0.2	**1.0**	**0.13**
2	0.19	**0.0**	0.19
3	0.2	0.69	**0.7**

Table 2- Demographic Perspective Cluster Centroids

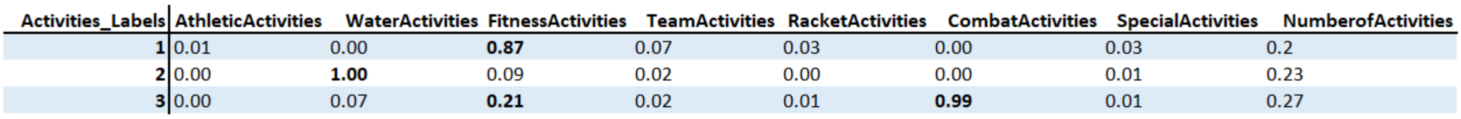
Cluster 2 is the one with higher number of observations, 5901, which is an indicator that a lot of people in the gym stayed a long period of time without using the facilities again, being why the target marketing is so important for each one of these groups.

5.3 Activities Perspective- K-prototypes Algorithm

The third perspective that made sense to explore was the Activities clustering. Although we managed to create one new numeric variable with them, it made sense to include them in clustering as well. Because they are all binary, we had to use, again, the K-prototypes algorithm. Since these variables do not have a high variance, we expected from the beginning that only certain activities would have high cardinality in the clusters. This is no issue since it can signal in which activities the facility can focus more since they attract the most customers.

Following the steps described in the previous section we arrived at the elbow plot (**Figure 10** in the annex) with an optimal number of three clusters. To get the final solution we ran the algorithm with this cluster number, added the labels to the data frame and grouped the values by each label.

As expected, only a few activities are predominant in each cluster. In cluster 1, the majority of people go to Fitness Activities. In cluster 2, everyone goes to the Water Activities whilst in cluster 3 everyone goes to Combat Activities.



Activities_Labels	AthleticActivities	WaterActivities	FitnessActivities	TeamActivities	RacketActivities	CombatActivities	SpecialActivities	NumberofActivities
1	0.01	0.00	**0.87**	0.07	0.03	0.00	0.03	0.2
2	0.00	**1.00**	0.09	0.02	0.00	0.00	0.01	0.23
3	0.00	0.07	**0.21**	0.02	0.01	**0.99**	0.01	0.27

Table 3- Activities Perspective Cluster Centroids

Regarding density, clusters 1 and 2 have a high density. However, cluster 3 only has 1586 observations, which means that when merging the data set we might disregard this focus group to prevent us from having too many marketing needs. We also noticed that these cluster 3 have some frequency in Fitness Activities so we might join them to cluster 1.

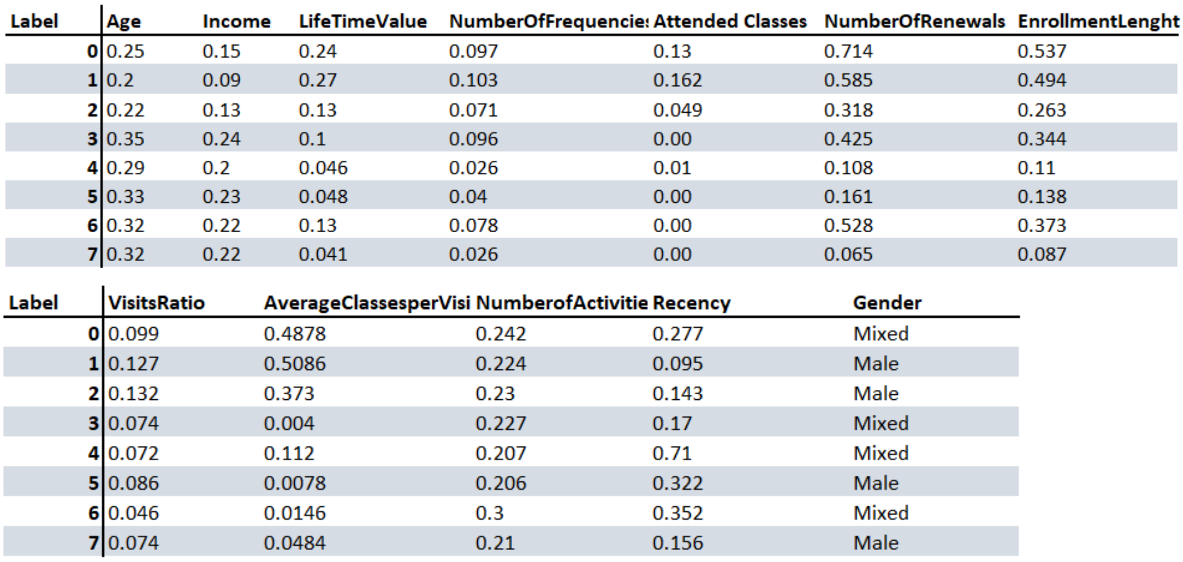
6. Combining Clusters Solution

This section aims at combining all perspectives and taking value out of the final solution to help the facility in getting the best knowledge of their customers and to provide useful solutions.

6.1 Merging Perspectives

To merge the perspectives we used hierarchical clustering. Firstly, we collected all labels from each category. Then, we built a data frame with them and only the metric features used. The reason for this previous step is that hierarchical clustering uses distances from points to centroids and so we could not include the binary variables in them. Since we have at least one metric feature in all the three perspectives we decided to go forward with this approach.

After visualizing the dendrogram (**Figure 11** in the annex), we could understand that more than 10 clusters would not be needed. Therefore, after some iterations to find the best match between cluster density and information provided we came to the final value of **eight** clusters. We decided to go with this number even though some clusters do not have that high cardinality because we wanted to preserve some focus clusters for our marketing approaches, which, otherwise, would be lost in a bigger cluster. The final centroids are as follows:



Label	Age	Income	LifeTimeValue	NumberOfFrequencie	Attended Classes	NumberOfRenewals	EnrollmentLenght
0	0.25	0.15	0.24	0.097	0.13	0.714	0.537
1	0.2	0.09	0.27	0.103	0.162	0.585	0.494
2	0.22	0.13	0.13	0.071	0.049	0.318	0.263
3	0.35	0.24	0.1	0.096	0.00	0.425	0.344
4	0.29	0.2	0.046	0.026	0.01	0.108	0.11
5	0.33	0.23	0.048	0.04	0.00	0.161	0.138
6	0.32	0.22	0.13	0.078	0.00	0.528	0.373
7	0.32	0.22	0.041	0.026	0.00	0.065	0.087

Label	VisitsRatio	AverageClassesperVisi	NumberofActivitie	Recency	Gender
0	0.099	0.4878	0.242	0.277	Mixed
1	0.127	0.5086	0.224	0.095	Male
2	0.132	0.373	0.23	0.143	Male
3	0.074	0.004	0.227	0.17	Mixed
4	0.072	0.112	0.207	0.71	Mixed
5	0.086	0.0078	0.206	0.322	Male
6	0.046	0.0146	0.3	0.352	Mixed
7	0.074	0.0484	0.21	0.156	Male

Figure 12 in the annex presents the number of users per cluster, whilst **Figure 13** in the annex presents a summary on the main level per feature of each cluster and a first attempt to name these clusters. The analysis performed in the next section as well as the previous figure were done with the help of **Figure 14** in the Annex.

6.2 Profiling- Marketing Strategies

This section aims at providing specific marketing approaches for each customer cluster we identified before.

If the sports facility wants to keep clients engaged in their classes, they could target Clusters 0 and 6. Since **Cluster 0** is composed of low income high engagement clients, to keep this rate, the center could offer early access to classes' enrollment where these customers would profit from a premium experience due to their loyalty to the facilities. This way we promote commitment to us, which translates in an efficient way to retain these customers' value. In turn, since **Cluster 6** individuals have higher income and less engagement we suggest that to increase their participation in the classes, the facility could offer some discount per extra class the user enrolled in, this way we can take advantage of their higher income availability and increase both our revenues and their frequencies.

If the sports center wants to reduce customers' recency as a way to increase their lifetime value they could target **Cluster 4**. These individuals gave up too early without trying any classes. We suggest that, as a way to make them return more often, the gym should offer the renewal fee and create a point system where after 15 frequencies the user would earn a free class. This way we make the clients have an objective, which motivates them to come and might make them enroll in a class they enjoy, which improves their value.

To target Cluster 1 and 2 the facility could invest in social aspects of the clients. Regarding **Cluster 1**, as they are considered the best customers and are linked to water activities, the gym could offer them some water equipment as a reflection of their preference for this activity. Besides that, they could open a friends' day where these clients could bring a friend to try the water class and, since they enjoy it, they would be good advocates and could make their friend stay. For the low-income customers in **Cluster 2**, since they come regularly we just want to alleviate their financial burden a bit and offer them a discount per each friend that they bring to the facility, attributing the discount to the friends as well. Therefore, if client Y of this cluster brings a friend Z in December, in January both have a 45% discount. In February if no friend is introduced the price comes to normal.

For the remaining three clusters, we suggest the facility takes a different approach using a Personal Trainer perspective. For **Cluster 7**, since they are mainly focused on fitness but lack a sense of commitment, we suggest the facility offers a free training experience. This experience would consist of a personal trainer spending a single session with the user and drawing up a training plan specific for the fitness activity with a predefined schedule. This could make the user goals achieved more quickly, which will motivate them to continue to come and make more renewals, which, in turn, increases our value.

Regarding **Clusters 3** and **5**, we are dealing with high income clients. For the former group we would create a premium experience. Since they do not want to engage in classes but tend to come, we would create a personalized program where they could pay a certain amount for a personal trainer for all the visits, nutritionist appointments and extra hours for practices. This way these high-income users would feel more valued and would provide more value to the fitness center. The latter has the difference that they are only male and mainly go to fitness activities. The premium experience could be translated into the creation of a hub where the facility would bring advanced bodybuilders and experts in the field to share tips, practice knowledge and training exchange experiences to keep them motivated and engaged in their target activity.

With all these strategies, the facility can easily target their customers and make them enjoy and enrich their experience in the gym, making them valuable to us and us to them.

6.3 Multidimensionality Visualization

One useful tool to ease the understanding of the distribution of clusters in a high dimension problem is to use visualizations specifically built for that purpose. We choose to use T-SNE. This method breaks down all points to a two-dimension space where clusters can be observed. As seen in **Figure 15** in the annex, our T-SNE visualization does not provide a great division in clusters. We believe that even though our strategy is cohesive during the work we did there is still some noise present that is influencing the points distribution. However, some bigger groups can be easily identified in the visualization.

7. Outliers Reintroduction

As it was explained before, we initially identified and removed potential outliers from the dataset. Here, as a way to allow those customers to also have a marketing campaign assigned we decided to reintroduce them.

To predict cluster labels for these reintroduced outliers, our group used a decision tree classifier, using merged_labels as the defining target. The dataset was split into training and testing sets, using the function train_test_split, with 20% of data being used for testing. The algorithm achieved an accuracy of 72.63%, on average, when attributing customers to the right cluster, based on the test data. After that, we were able to use this trained model to predict cluster labels for the outliers within the remaining dataset.

8. Conclusion

The project we developed gives relevance to the whole data mining process and allows us to understand how it is done and what we can take from it. A good preprocessing line is a key element in generating good solutions, as so we did an extensive yet cohesive approach that generated a clean dataset for the clustering. With it, we applied three clustering perspectives, value, demographic and activity based, using different clustering methods.

As the real world is complex, the approach taken on these perspectives gave more insights for our final solution. With the multiple combinations we tried, we understood the complexity and importance of choosing relevant features. After combining everything, the final solution of eight clusters fell in line with our expectation of having focus groups and broader groups as it is what we are used to dealing with in our daily lives. With these specific groups, we were able to identify what characterized them, which is an important step for a company if it wants to know their customers. Only with these differences we were able to come up with specific marketing approaches for each group that better reflected their habits and needs. Our eight clusters are different from each other, some more specific, others more general but each one has a strong independent campaign to get to the customers in them, ranging from premium experiences to single class discounts.

This is a clear example of how data mining practices are of major relevance to businesses. With these tasks a company, like the fitness facility we worked with, can save many resources in only targeting the specific groups and not lose time in marketing certain promotions to people who will never be moved by them. The visualizations we generated, especially Figure 14, were a great support to our decisions and to increase the degree of confidence in our solution.

We now understand that this practice is not mainly logical and we need to have some judgment towards decisions we have to make. As so, another key element for us was to try to understand the business itself and how it works. Only after this, we felt ready to come up with relevant strategies.

Having all the above in consideration, we agree that we did a good job and provided quality information to XYZ sports center, which will help them to target their clients more efficiently and, as a consequence, increase their value.

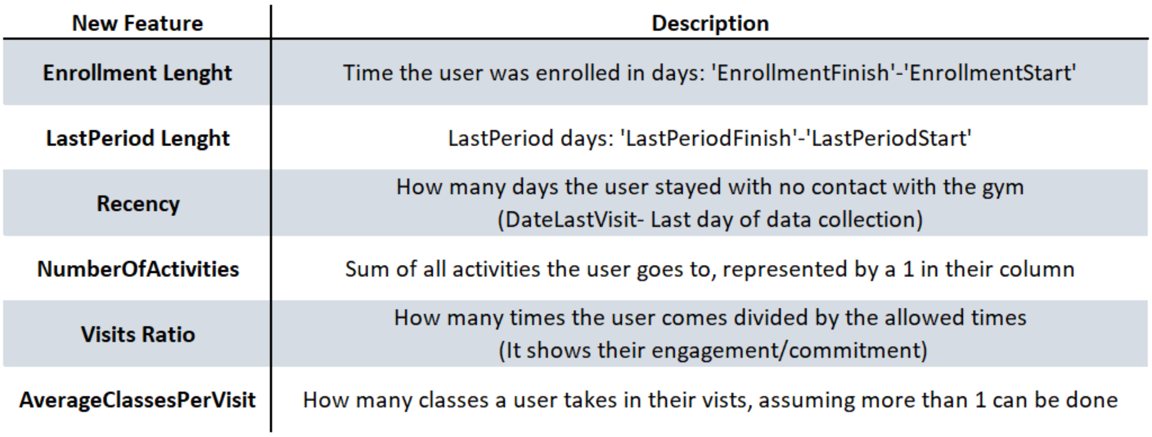
9. References

Anon., 2023. [Online]
Available at: https://medium.com/@balajicena1995/dbscan-clustering-2a577d384e61
[Accessed November 2023].

Huang, Z., 1998. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,* s.l.: Kluwer Academic Publishers.

Notebooks from classes on course GitHub: https://github.com/fpontejos/Data-Mining-23-24/

10. Appendix



New Feature	Description
Enrollment Lenght	Time the user was enrolled in days: 'EnrollmentFinish'-'EnrollmentStart'
LastPeriod Lenght	LastPeriod days: 'LastPeriodFinish'-'LastPeriodStart'
Recency	How many days the user stayed with no contact with the gym (DateLastVisit- Last day of data collection)
NumberOfActivities	Sum of all activities the user goes to, represented by a 1 in their column
Visits Ratio	How many times the user comes divided by the allowed times (It shows their engagement/commitment)
AveragéClassesPerVisit	How many classes a user takes in their vists, assuming more than 1 can be done

Figure 1- New Features Description

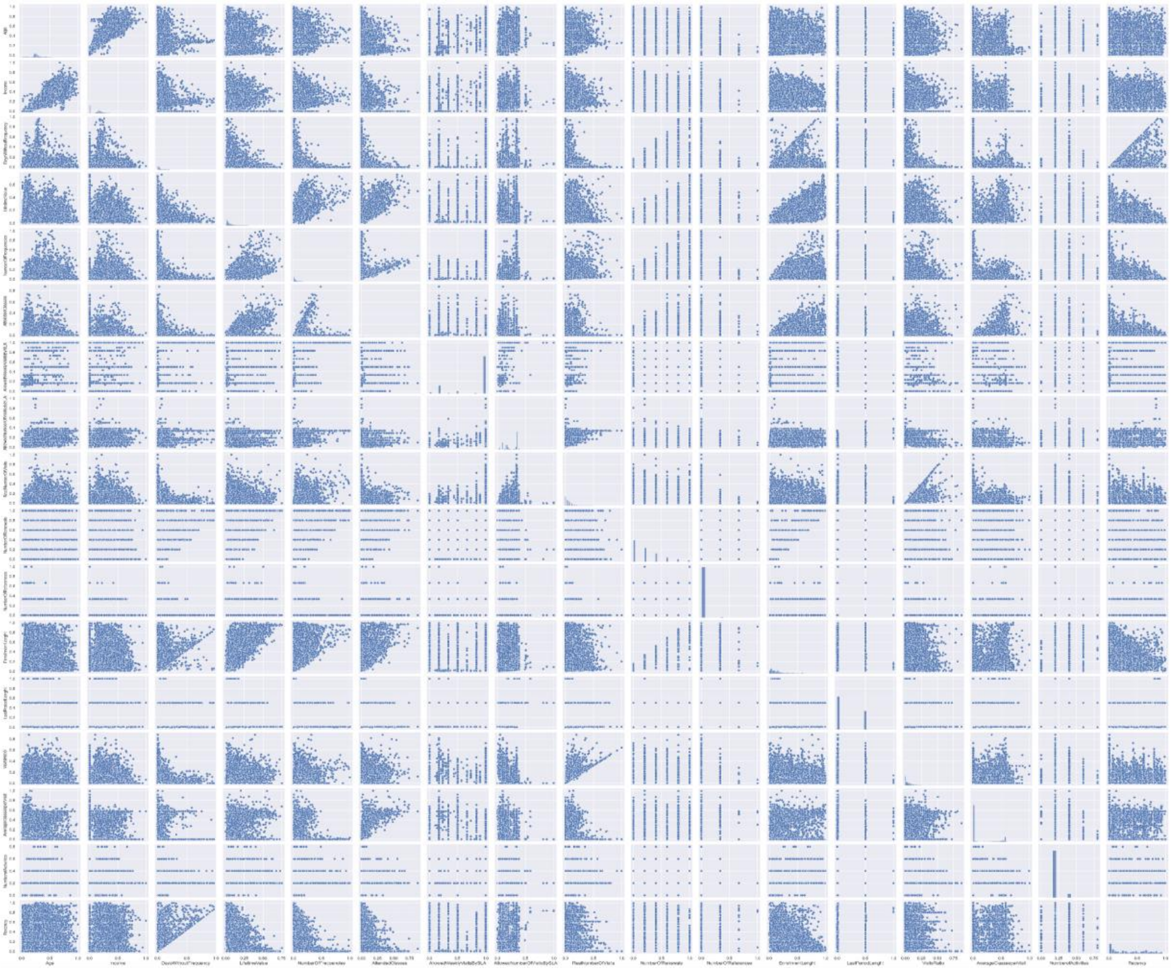


Figure 2- Metric Features Pairwise Relationship

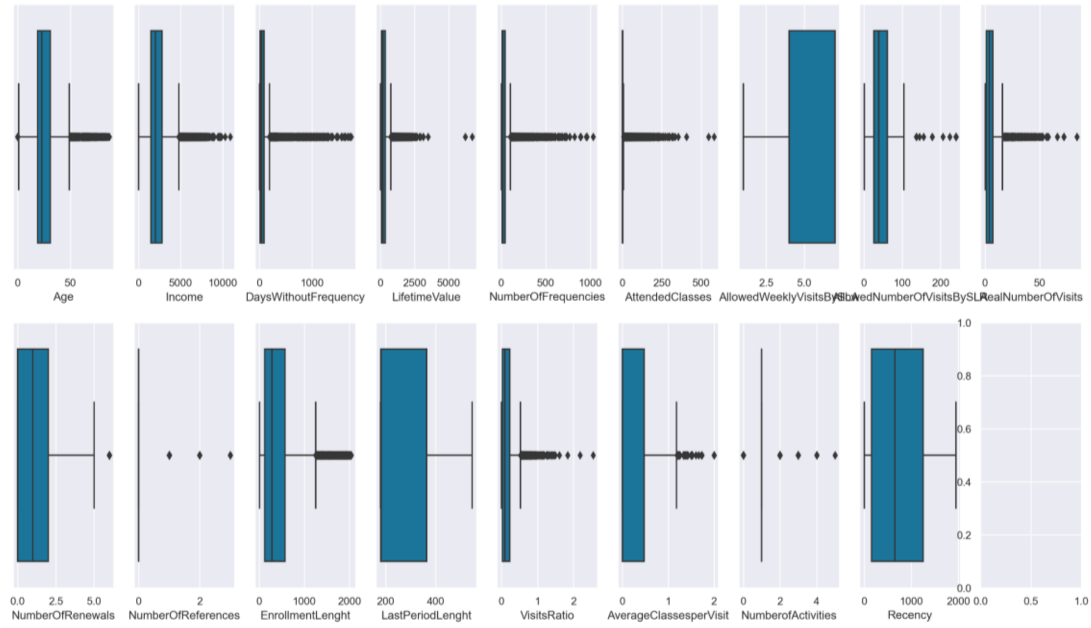


Figure 3- Metric Features Boxplot

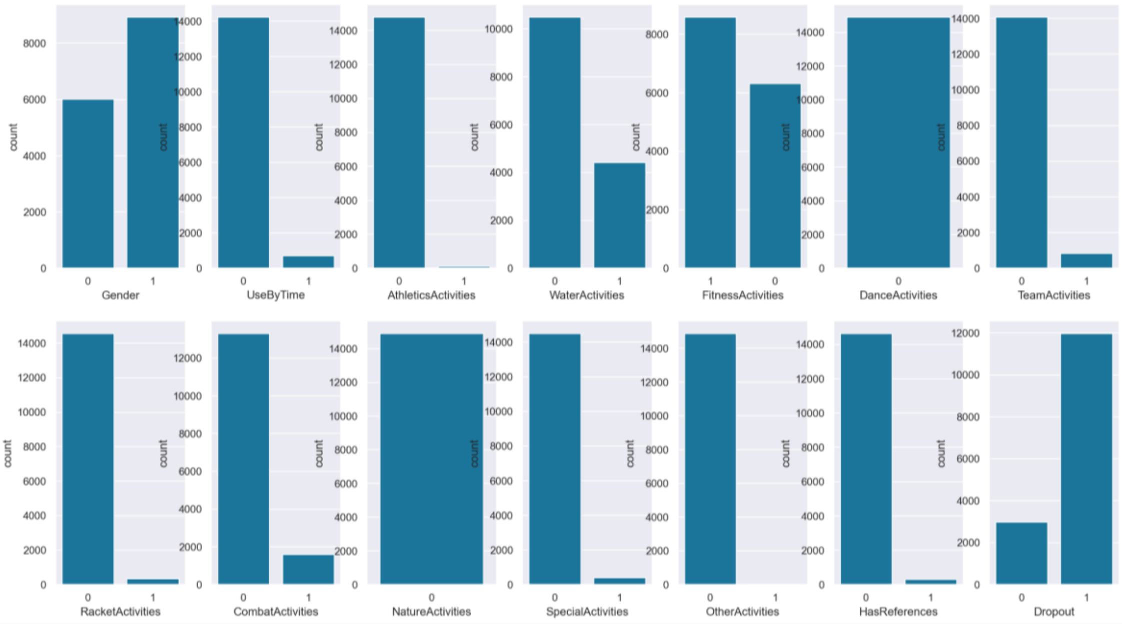


Figure 4- Non-metric Features Count Plots

Variable	Cut-off values
LifeTimeValue	5000
NumberOfFrequencies	750
AllowedWeeklyVisits	200
AttendedClasses	500
RealNumberOfVisits	80
VisitsRatio	2
AverageClassesperVisit	2
NumberOfRenewals	If >5 then is 5

Figure 5- Outliers Cut-off Values

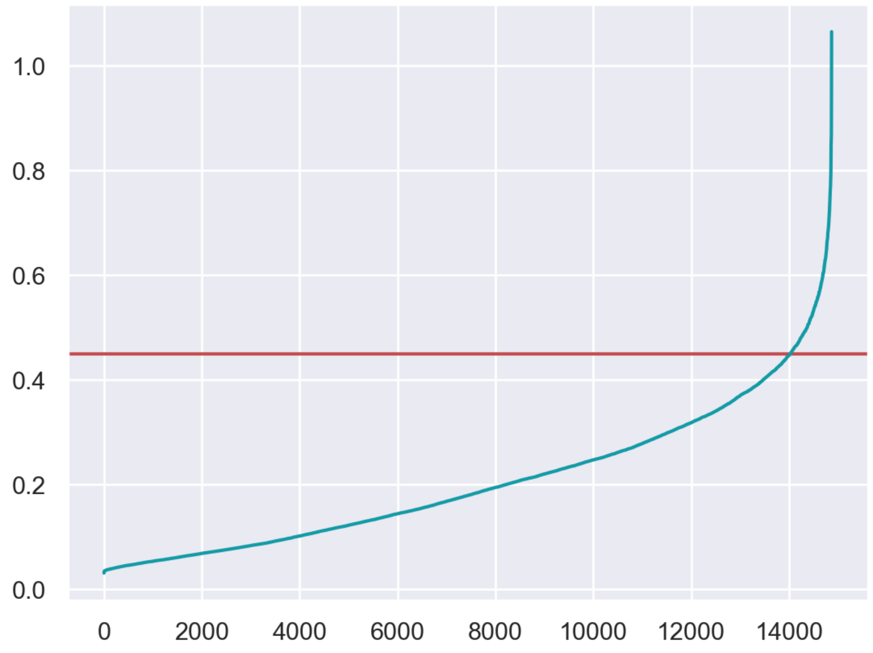


Figure 6- DBSCAN Epsilon Plot

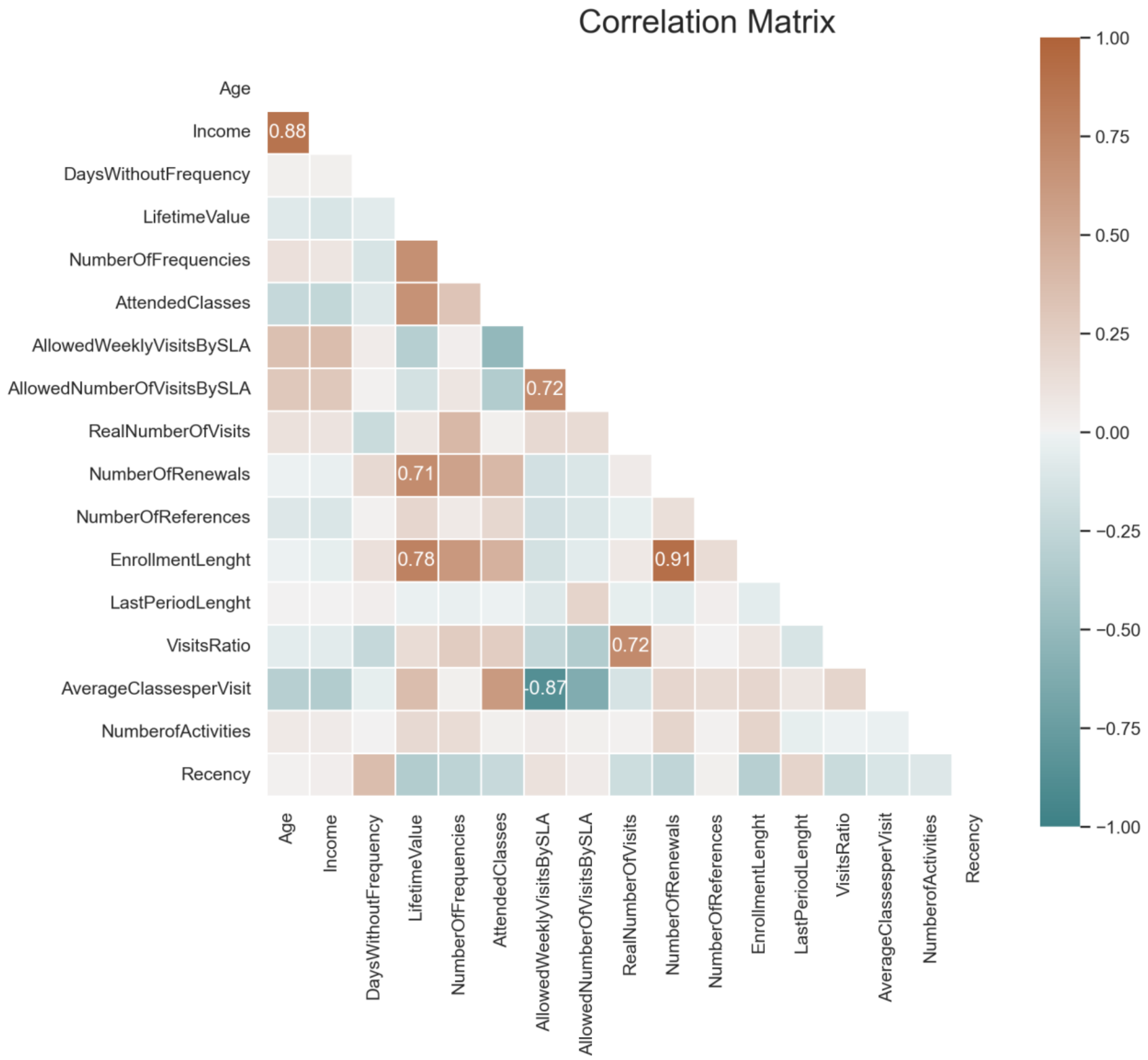


Figure 7- Metric Features Correlation Matrix

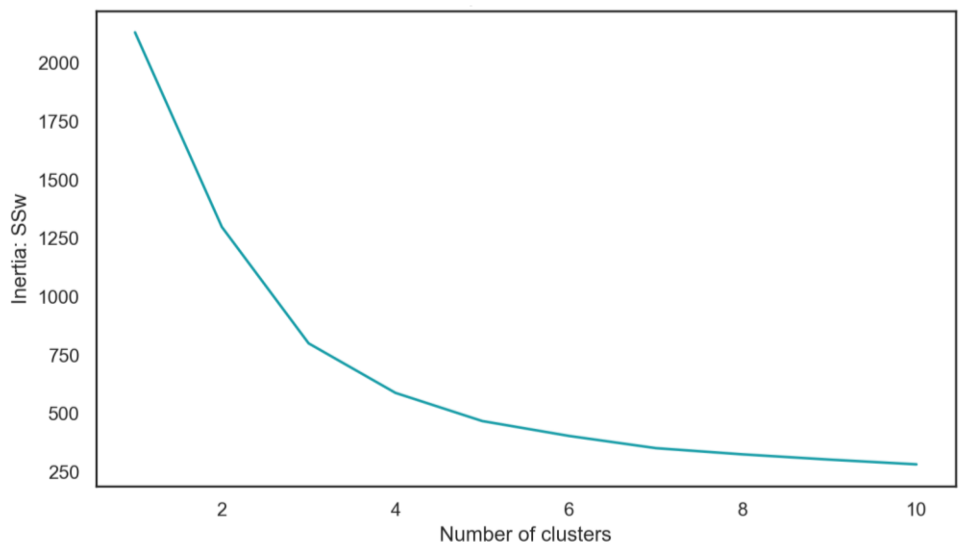


Figure 8- Inertia plot over the clusters Value Perspective

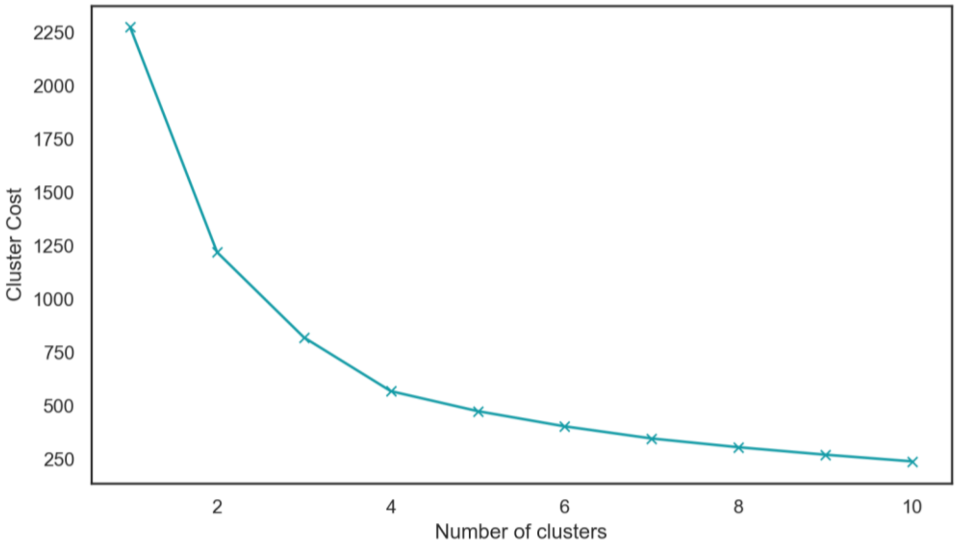


Figure 9- Elbow Method Plot Demographic Perspective

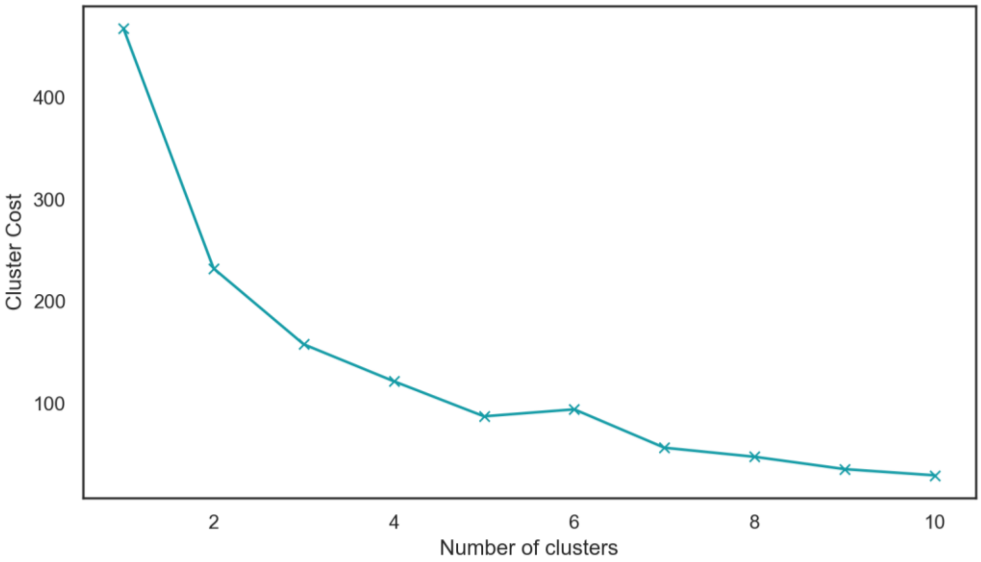


Figure 10- Elbow Method Plot Activities Perspective

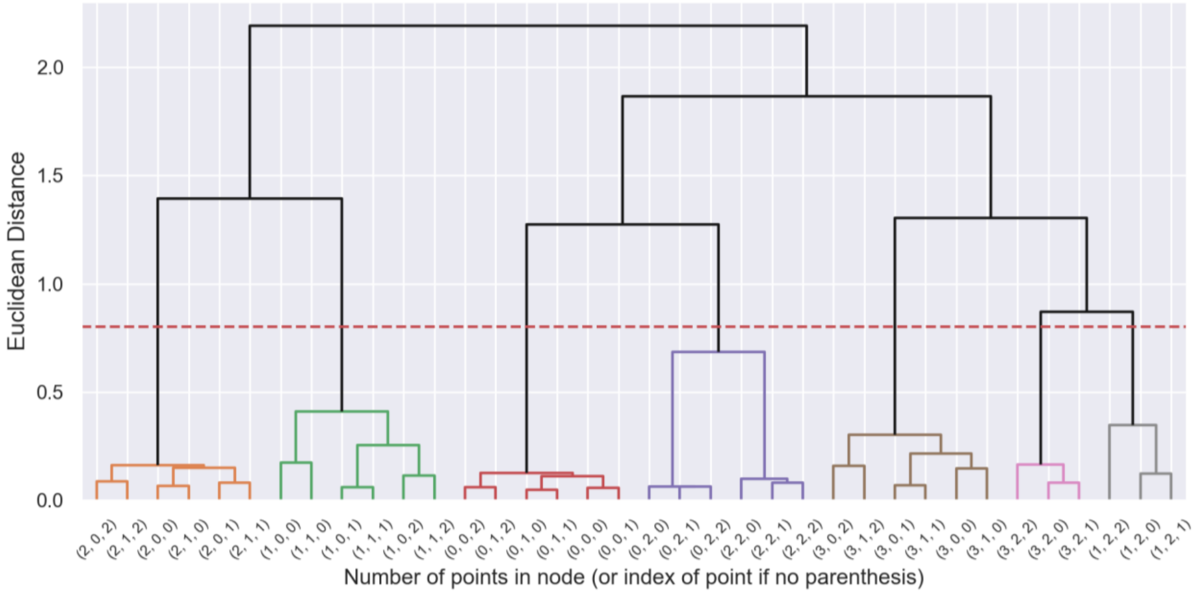
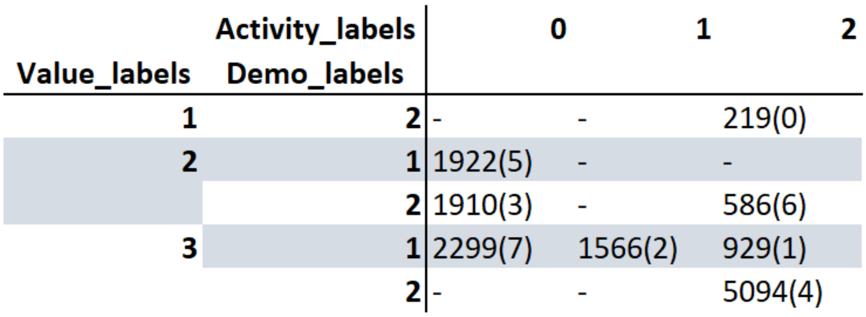
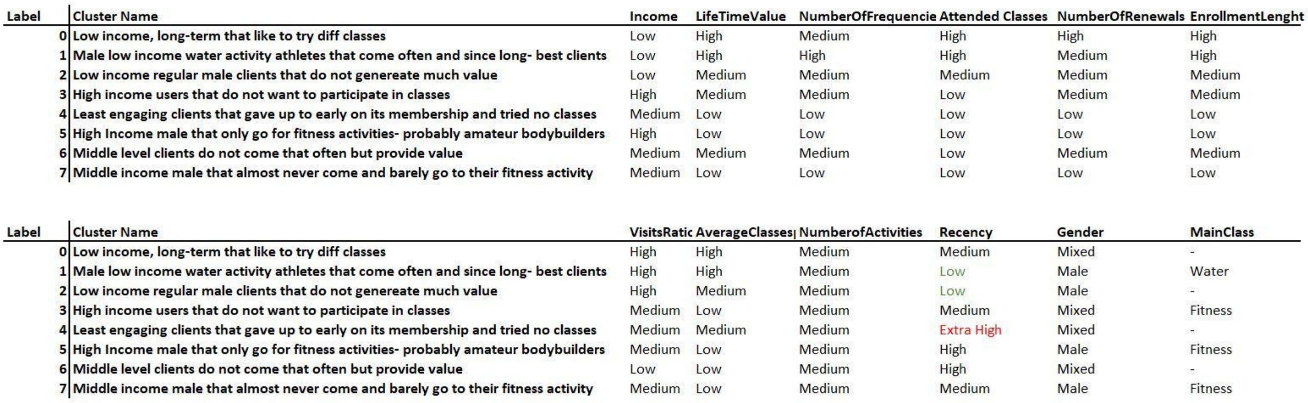


Figure 11- Dendrogram for Hierarchical Clustering



	Activity_labels	0	1	2
Value_labels	Demo_labels			
1	2	-	-	219(0)
2	1	1922(5)	-	-
	2	1910(3)	-	586(6)
3	1	2299(7)	1566(2)	929(1)
	2	-	-	5094(4)

Figure 12- Cluster Value Counts



Label	Cluster Name	Income	LifeTimeValue	NumberOfFrequencie	Attended Classes	NumberOfRenewals	EnrollmentLenght
0	Low income, long-term that like to try diff classes	Low	High	Medium	High	High	High
1	Male low income water activity athletes that come often and since long- best clients	Low	High	High	High	Medium	High
2	Low income regular male clients that do not genereate much value	Low	Medium	Medium	Medium	Medium	Medium
3	High income users that do not want to participate in classes	High	Medium	Medium	Low	Medium	Medium
4	Least engaging clients that gave up to early on its membership and tried no classes	Medium	Low	Low	Low	Low	Low
5	High Income male that only go for fitness activities- probably amateur bodybuilders	High	Low	Low	Low	Low	Low
6	Middle level clients do not come that often but provide value	Medium	Medium	Medium	Low	Medium	Medium
7	Middle income male that almost never come and barely go to their fitness activity	Medium	Low	Low	Low	Low	Low

Label	Cluster Name	VisitsRatic	AverageClasses|	NumberofActivities	Recency	Gender	MainClass
0	Low income, long-term that like to try diff classes	High	High	Medium	Medium	Mixed	-
1	Male low income water activity athletes that come often and since long- best clients	High	High	Medium	Low	Male	Water
2	Low income regular male clients that do not genereate much value	High	Medium	Medium	Low	Male	-
3	High income users that do not want to participate in classes	Medium	Low	Medium	Medium	Mixed	Fitness
4	Least engaging clients that gave up to early on its membership and tried no classes	Medium	Medium	Medium	Extra High	Mixed	-
5	High Income male that only go for fitness activities- probably amateur bodybuilders	Medium	Low	Medium	High	Male	Fitness
6	Middle level clients do not come that often but provide value	Low	Low	Medium	High	Mixed	-
7	Middle income male that almost never come and barely go to their fitness activity	Medium	Low	Medium	Medium	Male	Fitness

Figure 13- Final Cluster Interpretation

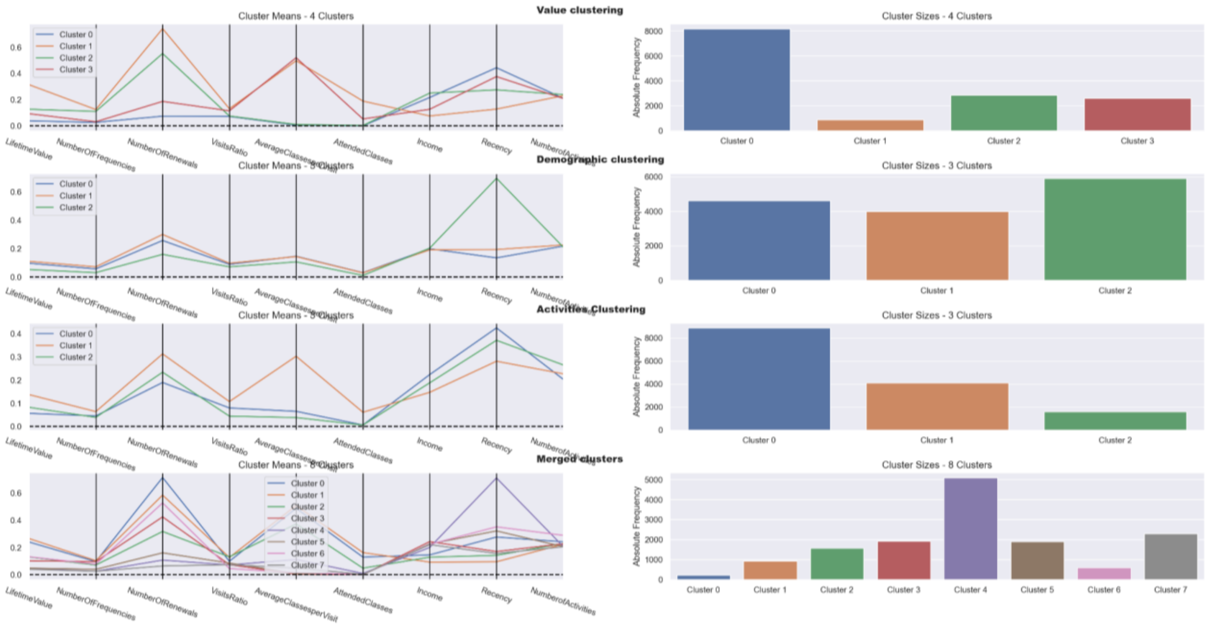


Figure 14- Final Cluster Profiling

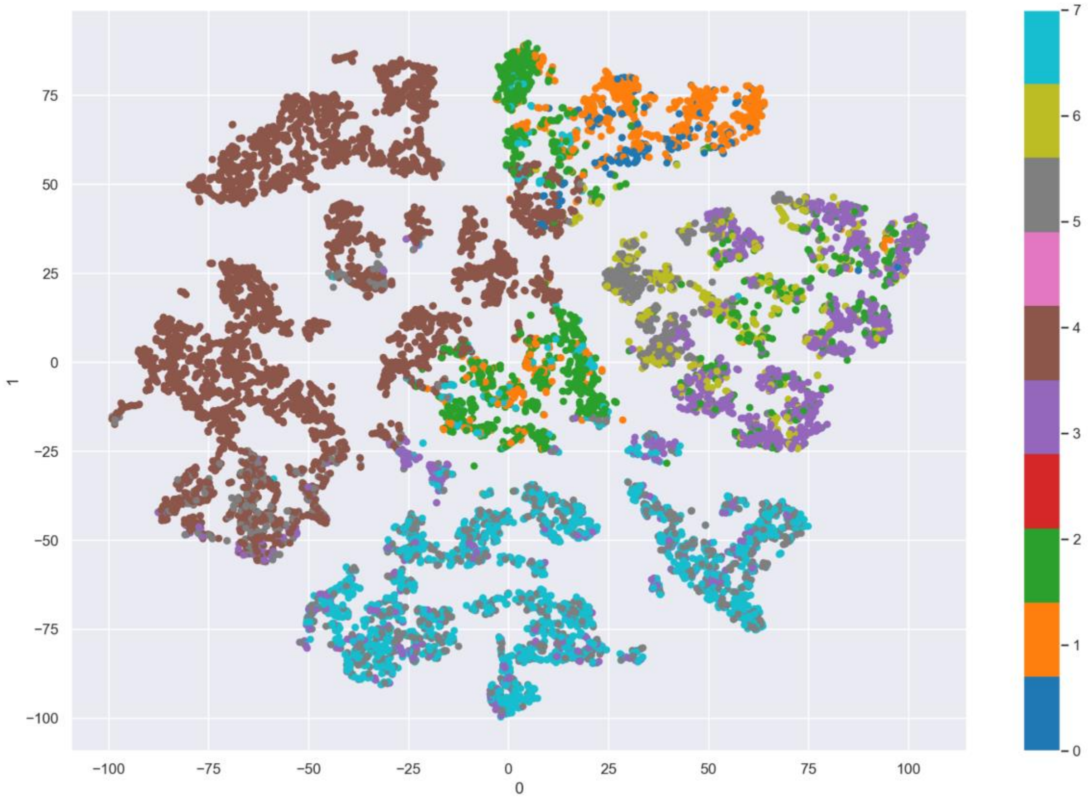


Figure 15- T-SNE Visualization