Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Customer Segmentation and Data-Driven Strategy Enhancement for XYZ Sports Company

Group 94

Andriani Kakoulli, number: 20230484

Eugénia Rosário, number: 20220598

Matheus Felisberto, number: 20230585

January 2024

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

Table of Contents

1. Introduction

This project aims to produce a comprehensive customer segmentation strategy for XYZ Sports Company fitness facility. It is based on an extended dataset, sourced from the company's Enterprise Resource Planning (ERP) system, encompassing customer-related data spanning from June 1st, 2014, to October 31st, 2019.

The dataset is a detailed compilation of customer data, including unique identifiers (ID), demographic information, enrollment details, activity participation, and usage metrics. It also covers financial aspects, engagement levels, and relational data. Based on the data provided, an Exploratory Analysis was made to better understand the relationship between features and better fitting for the feature engineering process, in order to develop an appropriate segmentation model./ identify the variables that should be used to segment customers.

After the exploratory analysis process, and choosing the variables that better suit the segmentation model, we resorted to different Data Mining algorithms in order to segment those clients into 3 main groups/clusters to better understand its customers, deliver more personalized services, and optimize marketing efforts.

2. Data Profiling

As a first approach towards the dataset, a general data exploration had to be done in order to understand our data. By this exploration, we were able to absorb information about the customers regarding demographics (e.g. age, income), preferable sports activities (e.g. *AthleticActivities*, *WaterActvities*) and information from a value perspective or in other words, the customer's behavior in relation to the sport facility (e.g. *DaysWithoutFrequency*, *EnrollmentStart*, *EnrollmentFinish*). These three perspectives towards the sports company's customers reveal important details which will be proven to be useful for the customers segmentation.

3. Data Pre-processing

The work of data preprocessing involves handling or imputing the missing values, scaling of data and handling outliers. Overall, we tried to clean and transform the data in ways that would be beneficial for future use in clustering.

3.1 Initial Transformations

First of all, the unique identifier (index) was set to be the *ID* variable, an already existing feature of the dataset, after it was transformed from the range of 1000 until 24941 to a sequence of consecutive numbers from 0 to 14941. For the duplicates there was no action to be taken since all customers were unique.

Other noticeable aspects of interest was the zero number of enrollments in *DanceActivities* and *NatureActivities*, so it was decided to drop these two variables. Moreover, *AthleticActivities* and *OtherActivities* had very small numbers of enrollments of 0.7% and 0.2% respectively, though

it was decided to keep them for future use in feature engineering. Our attention was drawn to variables with their data type mistaken, so we needed to transform them in their proper types. Variables of *Gender, HasReferences, UseByTime* and all variables concerning activities had to be converted to type boolean, while *EnrollmentStart, EnrollmentFinish, LastPeriodStart, LastPeriodFinish* and *DateLastVisit* had to be converted from type object to type date, in order to be able to use them as dates in future actions.

An action was taken regarding *Age* in relation with *Income*[Figure1]. Assuming that the data were collected in Portugal and noting that the legal working age in Portugal is above 16 years old, the few values of variable *Income* of users with *Age* below 16 were firstly set as NaN and were later imputed with zero.

3.2 Missing Values

In this dataset, there were various features that contained missing values but only had a small percentage of them, except variables of *Income* and *AllowedWeeklyVisitsBySLA*[Figure2]. The following imputation strategies were crucial for handling missing data and ensured the dataset's integrity and reliability for subsequent analyses.

Each feature was handled with a different approach. It was shown that all *activities* had a small population of missing values, but instead of dropping the corresponding entities, we followed the assumption that NaN-values in these fields were customers that did not enroll and should be imputed with zero. Missing values of variable *Income* were imputed with zero as well, but the assumption behind this action will be explained later. We filled the NaN-values of the variable *HasReferences* based on the values of the existing variable *NumberOfReferences* given the logical relation that these two variables have.

Regarding the feature of *AllowedWeeklyVisitsBySLA*, the number of missing values was large enough to decide not to drop them. Instead, an observation was made regarding a pattern that occurred concerning this variable and the variable of *AllowedNumberOfVisitsBySLA*. We created another column in the dataset named *Pattern* in favor of the plenty entities with the observation that: the allowed number of visits per week *is* the number of visits that the customer is allowed to make (according to the service he hired in the last two months), multiplied by 7 seven (as there are seven days in a week and the week concerns us), divided by 61 (assuming that two months equal 30 + 31 days). [Example: for better understanding of this pattern, observing the row *ID*=1 which did not have missing values in the two relative cells, the value of *AllowedNumberOfVisitsBySLA* was 17.42, and multiplied by 7/61 gave the number 1.99, which rounds to 2 and it was equal to the corresponding value of *AllowedWeeklyVisitsBySLA* of *ID*=1.] Sometimes, this pattern was valid for the number of visits made in one month rather than two, but no obvious reason/pattern was revealed, so we continued with this assumption and more specifically in two months.

A similar approach was attempted but we could not find a way to calculate the number of visits to the sports facility between the dates indicated in *EnrollmentStart* and *EnrollmentFinish*. Since only a percentage of 0.17% would be deleted, leaving us with a dataset of 99.83% of the initial, we chose to drop the NaN-values of this variable.

3.3 Outlier Detection

Through meticulous analysis of box plots(figure3), we discerned that the dataset's features can be effectively bifurcated into two distinct categories based on the extremity of their values. The first category encompasses features exhibiting pronouncedly extreme values, such as *AttendedClasses*, *AllowedWeeklyVisitsBySLA*, *AllowedNumberOfVisitsBySLA*, *NumberOfReferences.* These features are characterized by significant variations, indicating potential outliers or unusual data points. In contrast, the second category comprises features with moderately extreme values such as *Age, Income, RealNumberOfVisits, NumberOfRenewals.*

In the applied outlier detection and removal methodology, two distinct filtering criteria based on the Interquartile Range (IQR) method were employed to preprocess data for analysis for each of the created sets. Firstly, the set of extreme values was selected. The 10th and 95th percentiles of these features were calculated and the IQR was determined as the difference between these percentiles. Outlier thresholds were set using 1.5 times the IQR above the 95th percentile and below the 10th percentile. Each feature was then filtered to identify values within these thresholds, and these filters were combined. This process was then repeated with the set of features with moderate values, using the 25th and 75th percentiles. Finally, the two filters were combined, this combination meant that a data point was retained if it was not considered an outlier by either filter.

A new DataFrame was created, and approximately 97.57% of the original data was retained. This filtered DataFrame represented a significant retention of the original data, ensuring that the filtering criteria were effective yet not overly restrictive.

3.4 Scaling

The goal of scaling is to standardize or normalize the feature values, ensuring their comparability and reducing the impact of scale variations. We chose a subset of numerical features that included *Age, Income, DaysWithoutFrequency, LifetimeValue, NumberOfFrequencies, NumberOfRenewals, NumberOfReferences*, and applied various scaling techniques on that subset. These techniques included StandardScaler, MinMaxScaler, RobustScaler and MaxAbsScaler. The decision of the scaling method was based on illustrating the distributions before and after scaling, using boxplots(figure4) revealing the spread of values and histograms showing how scaling alters the frequency distribution, and among the aforementioned techniques, we decided to apply StandardScaler since it seemed to worked better in clustering.

3.5 Coherence Checking

Before proceeding with further analysis, we conducted an assessment of data coherence. During this process, we identified several inconsistencies within the dataset. The most notable discrepancy involved the presence of 200 records as being under 18 years old and 20 records under 16 years old, having associated income values. This inconsistency was particularly evident in cases where the age was listed as 0. This issue brings us to another point concerning the Age feature, which included very young children registered at the gym. Additionally, we noted a concern with the ages of customers who had a dropout status (Dropout equals 1). In these instances, the age data reflected the customer's age at the time of dropout, making it outdated. To

address this, we introduced a new feature aimed at updating the age for such cases (CalculatedAge).

In our analysis, we also discovered 2422 instances where the *DateLastVisit* was more recent than the date of *EnrollmentFinish*. Initially, this might suggest an error or typo. However, upon further consideration, we recognized that these occurrences might not necessarily indicate a mistake. It's plausible that these users visited the gym for reasons other than regular workouts or sessions. For example, they might have dropped in to inquire about better membership conditions, or for other interactions, during which their visit was recorded. This understanding led us to treat these instances with a different perspective, considering them as potential customer engagements rather than outright errors.

Furthermore, 48 instances were identified where the *RealNumberOfVisits* of the customer exceeded the *AllowedNumberOfVisitsBySLA*, indicating inconsistencies in visit counts that need to be addressed.

Additionally, we also observed discrepancies between the *HasReferences* and *NumberOfReferences* fields. Specifically, there were instances where *HasReferences* was marked as True, yet *NumberOfReferences*was recorded as zero. Conversely, we also found cases where *HasReferences* was False, but *NumberOfReferences*was shown a non-zero count. These inconsistencies have been flagged for correction.

4. Feature Engineering

Leveraging the data that were eventually clear and complete, we delved into the process of feature engineering tailored for clustering analysis, a pivotal step uncovering inherent patterns and groupings within the XYZ Sports Company dataset. We introduced a set of diverse features based on previous remarks, designed to capture key aspects of user behavior and characteristics.

Building upon what was observed in the coherence checking regarding age consistency, we needed to introduce the feature of *CalculatedAge*. This feature computed the difference in EnrollmentStart date and the last day the data were collected (31-10-2019). The values of the new feature were then compared to the values of the customers' age, and if the given age was smaller than the calculated age, we led to the action of replacing the 'wrong' value of given age with the one that was consistent with our data.

Afterwards, we introduced the feature of *MembershipDuration,* which was used in following steps, indicating the number of days between the start and finish of the user's enrollment. In like manner, the feature of *LastPeriodDuration* was created to address the duration between LastPeriodStart and LastPeriodFinish for future use. The latter also showed low cardinality, and with respect to the metadata, we figured it might be beneficial to group its values using a map indicating the period between LastPeriodStart and LastPeriodFinish[half year(182 days), one year(365 days) and one and a half year(546 days)]. Based on this grouping, we added three variables to the dataset which were the dummy variables of the groups of *LastPeriodDuration.*

Other membership characteristics created were the *VisitFrequency,* which is the average number of visits per month by the user to the facility, and *AvgMonthlySpending,* which gives the average amount spent by the user per month during their membership. By the last mentioned features, we were able to make comparisons concerning monthly behavior and visits-spending comparisons.

Some features regarding the customer were added. Grouping the age in a variable named *AgeGroup* and the income in the variable *IncomeBracket* was beneficial for visualizing related information. The variable of *EngagementScore* gave a composite (average) score on visit frequency, class attendance ratio and activity diversity, which all contributed to the overall behavior of the customer.

The engagement of the customer to the sports facility was explored by four features. Starting with a feature very important to the company, the *DropoutRiskScore* was used to represent the likelihood of the user dropping out based on their *DaysWithoutFrequency.* The *ClassAttendanceRatio* new feature gave the ratio of the number of classes attended, to the total number of facility visits, while the *UtilizationRatio* gave the ratio of the actual number of visits, to the allowed number of visits (as per the SLA). Since the data were collected until the 31st of October 2019, the *LastActivityRecency* gave the number of days of the user's last visit to the facility in respect to that date.

Variables indicating dates were proven to be useful for clustering the customer's behavior in reference to the sports facility, so based on some of those, we created new features of temporal information of the user: *EnrollmentStartMonth* and *EnrollmentStartYear* gave information about the month and the year the customer was enrolled, respectively. Likewise, the features of *DateLastVisitMonth* and *DateLastVisitYear* were formed, indicating the month and year when the user last visited the facility. The *StartSeason* and *FinishSeason* features were created to display the season of the year that the user was enrolled, as long as the season that the user finished their enrollment.

Lastly, in terms of revenue, the feature of *RevenuePerVisit* was established, implying the average revenue generated by the user at each visit.

4.1 Inspection of *MembershipDuration*

An issue was observed regarding some values of *MembershipDuration* being equal to zero[figure5]. More inspection was done to check the hypothesis if the customers with this anomaly in the data were active members of the sports company. This assumption was first rejected by the financial aspect, as it was shown that neither of these customers had *LifetimeValue* equal to zero, meaning that they had paid a certain amount to the company. It was also rejected since the *LastPeriodDuration* had non-zero values for all customers, and the assumption of another form of enrollment was not applicable as *UseByTime* feature did not indicate that the customers were enrolled in this form of use.

Considering the fact that none of the aforementioned assumptions was able to explain this issue, we followed the action of replacing the dates regarding enrollment with the dates indicated in LastPeriodStart or LastPeriodFinish respectively.

5. Clustering

In order to understand the customers of XYZ Sports Company, we are introducing a comprehensive cluster strategy that encloses three key perspectives: value, demographics, and sports activities. This multifaceted approach aims to segment the customer base into distinct groups, each characterized by unique attributes and behaviors.

Our methodology incorporates a series of sophisticated analytical techniques, such as Principal Component Analysis, the Elbow Method, KMeans clustering and DBSCAN, to name a few. Our metric to evaluate our clusters was the silhouette score, which measures the quality based on cohesion and separation within the clusters.

5.1 Demographic

This aspect delves into the demographic characteristics of our customers, such as age, number of references, which in this context can mean family size, and income levels. Important to mention that we did not use gender as a feature because we cannot measure distance from binary features, and our dataset contains gender as binary.

With demographic clustering, we were able to understand the customer base in a way that future communication and product offerings can suit the company's objectives, targeting the right audience.

5.1.2 Dimensionality reduction

We have applied the PCA within our selected demographic features, hence the high-dimensional data was transformed into a lower-dimensional space, making it easier to analyze even considering the non significant information loss during this process. This process was fundamental to avoid the curse of dimensionality, even though from a demographic perspective we do not have as many features, reducing noise, and uncovering hidden patterns.

5.1.3 KMeans

In order to determine the number of clusters, which is a required hyperparameter for KMeans, we have used the Elbow Method. In the demographic clustering, our optimal number of clusters according to the elbow was 4. After the ensemble, the silhouette score for the cluster was 0.51. We decided then to ensemble another clustering algorithm.

5.1.4 DBSCAN

This clustering algorithm does not require establishing the number of clusters beforehand, but it has some hyperparameters that can be tuned in order to best fit the data. The most important is the eps, and we used the default 0.5. The silhouette score was 0.67, a great improvement from the KMeans.

5.1.5 Clusters

After the clusterization there is one of the most important steps in the journey, the cluster analysis. To have a great explainability from the model we analyzed again the data to identify the patterns that DBSCAN identified. KMeans did not perform well as compared to DBSCAN, since it

usually performs better for circle shaped clusters, while DBSCAN can handle outliers (see Figure 7 and 11 in Appendix for a detailed view of the clusters).

- **Seniors:** Identified as outliers, confirmed by the age boxplot, their age average is 44, they hold the biggest income average, and also spent more with the business throughout the time.
- **Adults:** The adult group usually does not have references (family or friends), but they have the second highest income average, and the lowest lifetime value.
- **Kids:** Averaging 8 years, having no income, having references and the second higher lifetime value.
- **Youngs:** 19 years on average, the lowest income on average, having references and the third lifetime value on average.

5.1.6. Business solutions and marketing strategies for each cluster

Different business applications and marketing approaches can be suggested for each identified cluster:

- **Seniors:** With the highest income and significant spending with the business, tailored high-value products and services can be offered to this group. Exclusive memberships or premium offerings might be particularly appealing.
- **Adults**: This group, characterized by a good income but lower engagement (lifetime value), might benefit from targeted promotions that increase their engagement and loyalty. Offering referral benefits could also be effective, considering their lower number of references.
- **Kids**: Despite their lack of income, they have a high lifetime value and references, indicating influence through family or friends. Marketing strategies could focus on family-oriented products and services, or partnerships with educational or children-centric brands.
- **Youngs**: With the lowest income but a fair amount of references and lifetime value, this group might be receptive to budget-friendly options, student discounts, or social media marketing campaigns that leverage their connectivity and potential as brand ambassadors.

5.2 Value

Our methodology to cluster XYZ Sports Company from the value perspective follows the same from the demographic. We first reduced the dimensionality, then applied KMeans, DBSCAN, and Gaussian Mixture Models to identify the optimal clusterization. Followed by the evaluation using the silhouette score.

To improve even more our cluster quality, we have used the UMAP as dimensionality reduction instead of PCA, focusing on more robust results. As hyperparameters we have chosen 100 numbers of neighbors, and a minimum distance of 0.1. Also, we selected the euclidean distance as the metric. The same methodology was applied, the Elbow Method gave us the optimum number of

clusters and we created the KMeans cluster. The silhouette score was 0.56. The closer we get to 1, the better our clustering solution(see Figure 8 and 12 in Appendix for a detailed view of the clusters). After analyzing the results, we came to the conclusion of the main characteristics of each customer segmentation.

- **Olders**: customers in this segment hold the longest membership duration, although their monthly spending and lifetime value are not the highest. In contrast, their class attendance and engagement score are the highest between the three categories.
- **Golden**: the key characteristics include the highest lifetime value, consequently the highest average monthly spending along with the utilization ratio. Although, their days without frequency are the highest, and the recency is the lowest.
- **Common**: customers in this category go to the facility mostly to take classes, since they hold the lowest attendance ratio amongst all categories. Also, they are on bottom regarding lifetime value and monthly average spending, and utilization ratio. However, they have the lowest days without frequency, meaning they are loyal to their physical activities

5.3 Sports

Clustering customers by sports perspective revealed interesting patterns regarding their preferences, and brought insights on how to increase the number of enrollments by creating plans for exclusive activities, for instance.

As previously mentioned, our methodology includes the dimensionality reduction using UMAP, then the clustering using KMeans and DBSCAN. The UMAP counts on 20 as the number of neighbors to be considered a cluster and a minimum distance of 0.1. The KMeans silhouette score was slightly better than DBSCAN and with fewer clusters, hence we decided to go further with it (see Figure 9 and 13 in Appendix for a detailed view of the clusters).

- **Fitness**: customers who majoritarily go to practice fitness activities.
- **Athletic and Water**: they only take athletic and water activities.
- **Racket and Combat**: their preferred activities include mostly racket and combat, but also a few sessions on water and athletics.
- **Ghosts**: in the sports perspectives, they are not present. It can be the case of other activities practitioners or inactive users, like parents who pay for a family plan but only the kids visit the facility. This is a hypothesis that can be confirmed in the later section on the contingency matrix.
- **Water and Fitness**: they also practice a little racket and athletic activities besides their main ones water and fitness.
- **Explorers**: users whose main characteristic is practicing the whole available activities.

5.4 Discoveries and insights

After analyzing the contingency matrix (see Figure 10 in Appendix), we could observe interesting patterns that can help the XYZ Sports Company to reach their goals with this report, meaning they can understand the value and demographics, and yet have insights regarding users activities preferences.

Adults show a preference for Fitness activities, with 6233 in the Common category. That may suggest that fitness activities in particular are popular among adults with a "normal" contribution to the finances. However, Older Adults have a inclination to be Explorers with 1484 in the Common category.

Among the Kids category, very few participate in sports categories, that may imply limited interest or availability of sports for kids in the facility. Creating an environment kids-friendly can be an interesting initiative to bring more customers in. Youngs and Seniors also do not seem to prefer participating in all sports activities, that also can mean less opportunities or even interest.

The Golden category brings an insight regarding Adults and Olders, showing a higher participation in Athletic and Water and Explorers categories, respectively. This may suggest a higher value or even premium preference in these groups for these specific activities.

A niche group identified can be a target for specialized marketing or programs, the Seniors in Water and Fitness sports category.

There are two sports categories that have generally low numbers across all demographic groups, Racket and Combat and Water and Fitness. This suggests that either a niche appeal or less overall interest.

6. Conclusions

This report outlines a structured approach adopted to segment customers from XYZ Sports Company. Initially, we explored the data to gain insights and identify potential challenges that could impact the quality of our segmentation. To address these challenges and enhance the clustering process, we strategically imputed missing values based on their relevance to clustering objectives and normalized the data to ensure uniform significance across all features, an essential step for distance-based algorithms.

In line with the report's objectives, we have created three different perspectives based on demographics, value and sports activities. Given the amount of features at our disposal we have used techniques such as Principal Component Analysis and UMAP to reduce the dimensionality of the data and improve our clusterization solution quality. The algorithms chosen for the clusterization were KMeans, DBSCAN and Gaussian Mixture Model, and we have used the silhouette score to evaluate their performances, a reliable metric for evaluating cluster performance.

With this report the XYZ Sports Company can have a multi-faced understanding of their customers, engage, target marketing campaigns, create new monthly-plans, develop new business opportunities among other insights.

References

- Taborda, J. T (2024). Portuguese labor law and employment contracts. *Expatica.*

Appendix

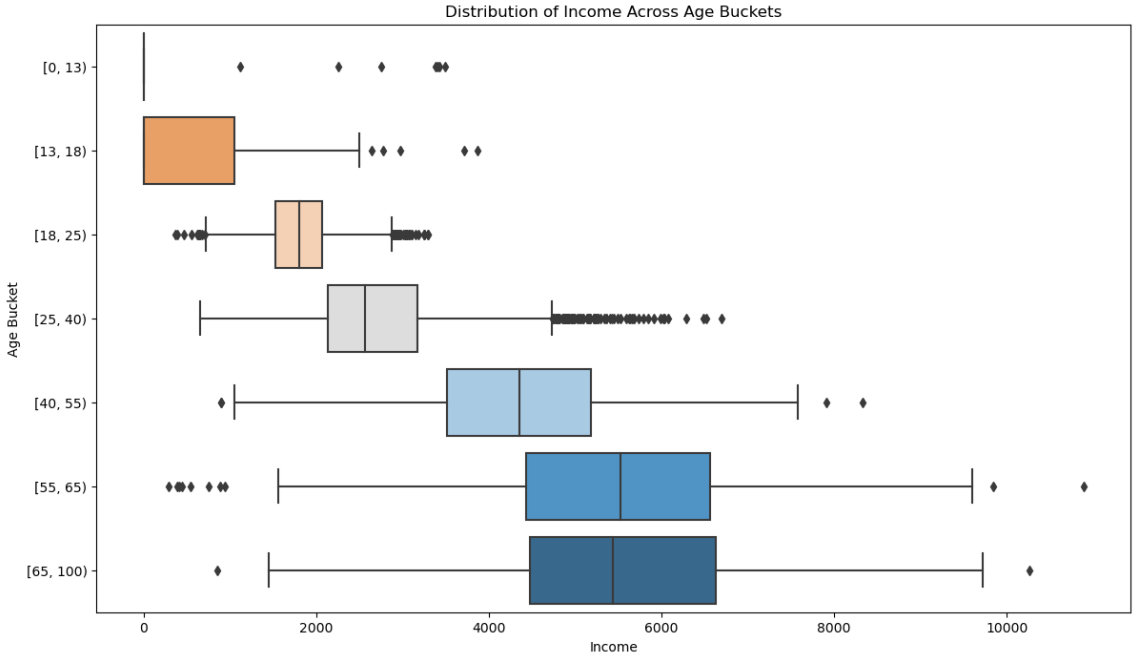

Figure 1: Income Distribution Across Ages

	Missing Count	Missing Percentage
Income	512	3.43%
AthleticsActivities	36	0.24%
WaterActivities	37	0.25%
FitnessActivities	35	0.23%
TeamActivities	35	0.23%
RacketActivities	37	0.25%
CombatActivities	33	0.22%
SpecialActivities	44	0.29%
OtherActivities	35	0.23%
NumberOfFrequencies	26	0.17%
AllowedWeeklyVisitsBySLA	535	3.58%
HasReferences	12	0.08%

Figure 2: Dataframe showing features' missing value counts and corresponding percentages

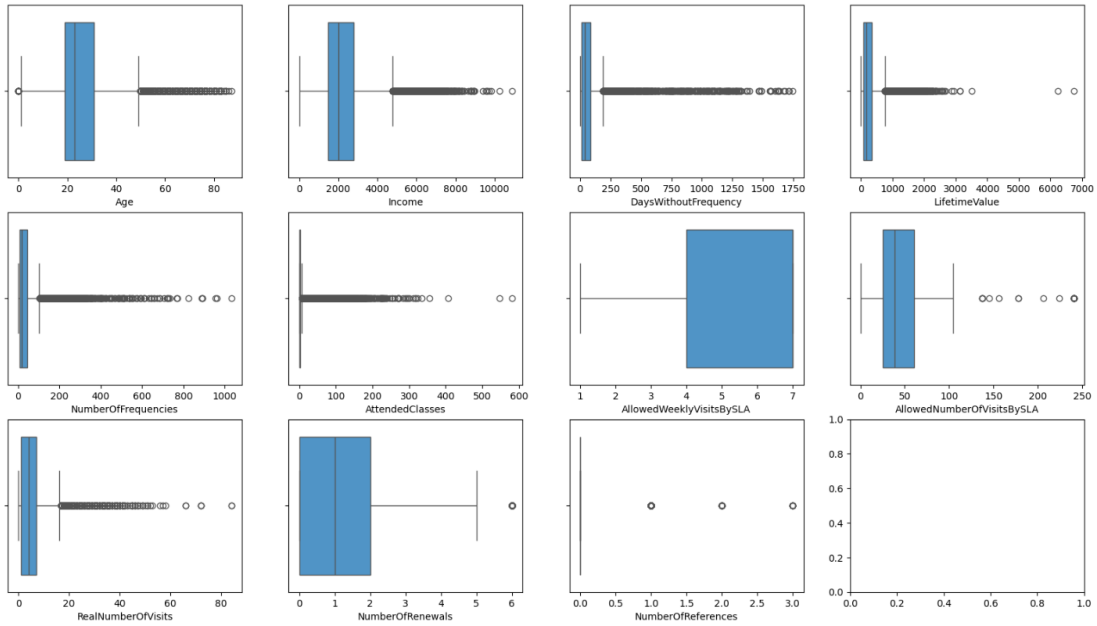

Figure 3 : Boxplots of numerical features before handling the outliers

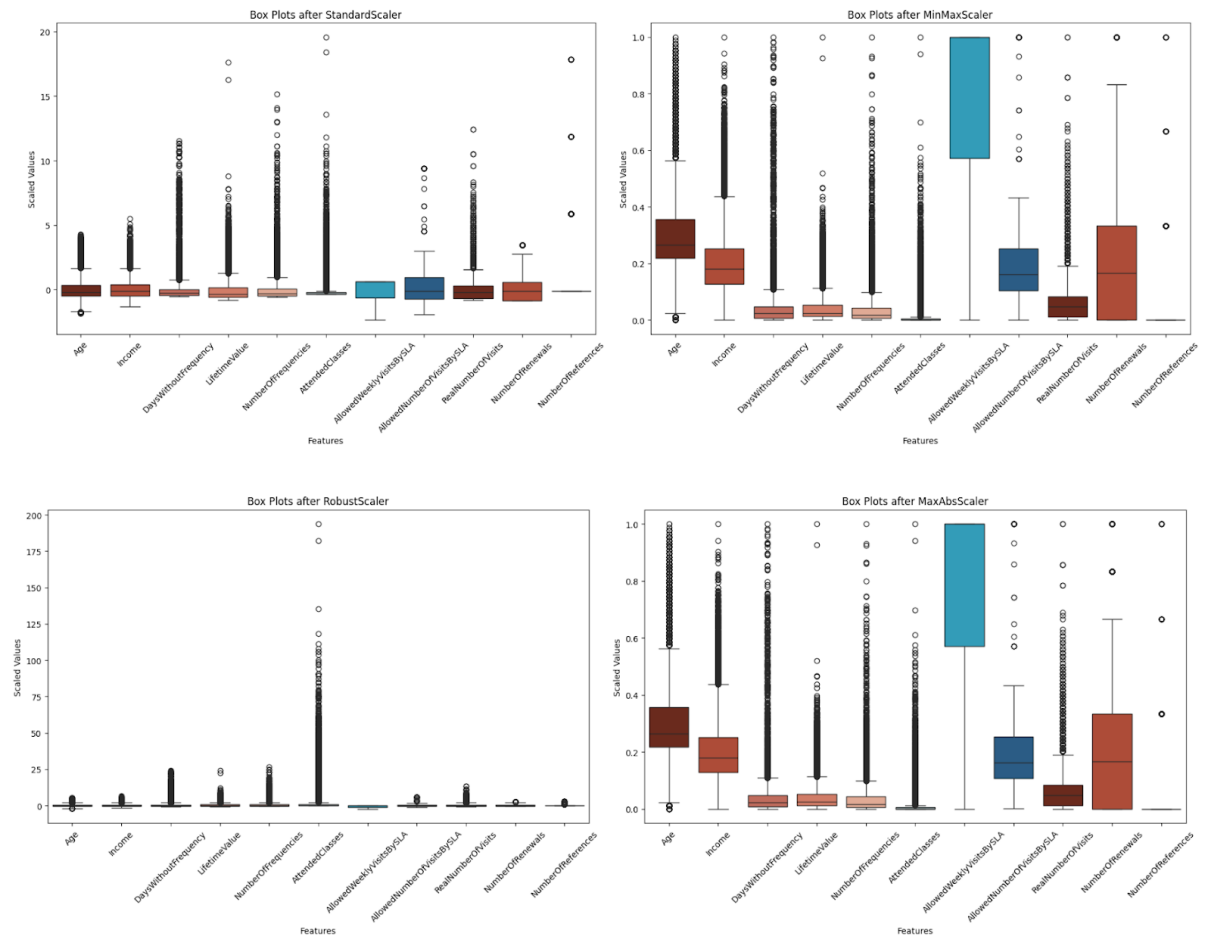

Figure 4 : Boxplots of features before scaling and after using different methods of scaling

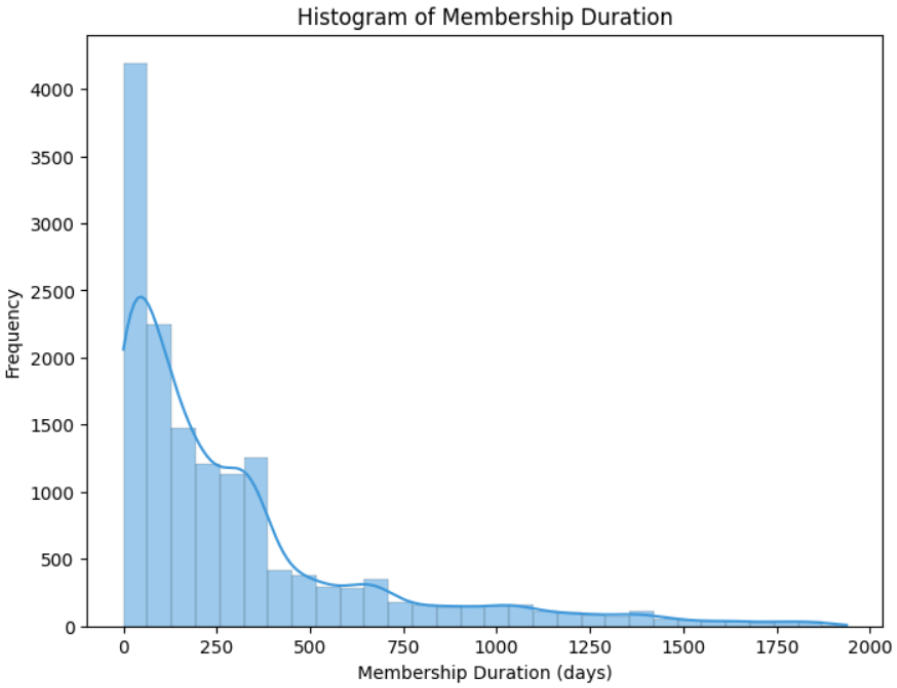

Figure 5 : Distribution of MembershipDuration before feature engineering

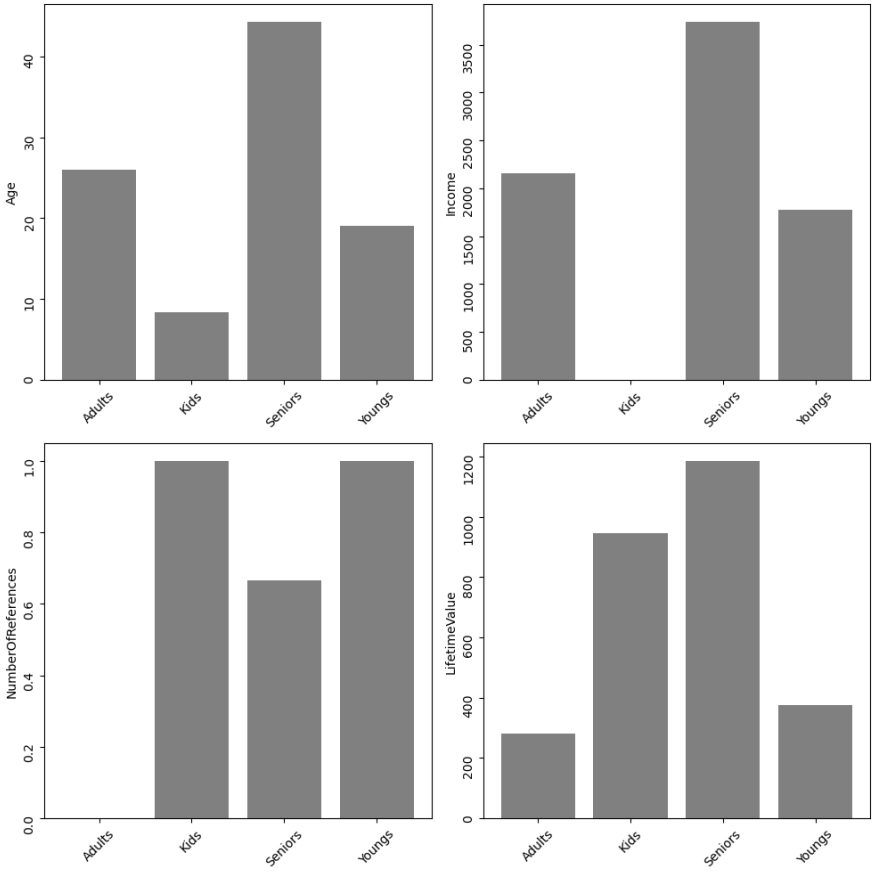

Figure 7: Plot showing demographic clusterization

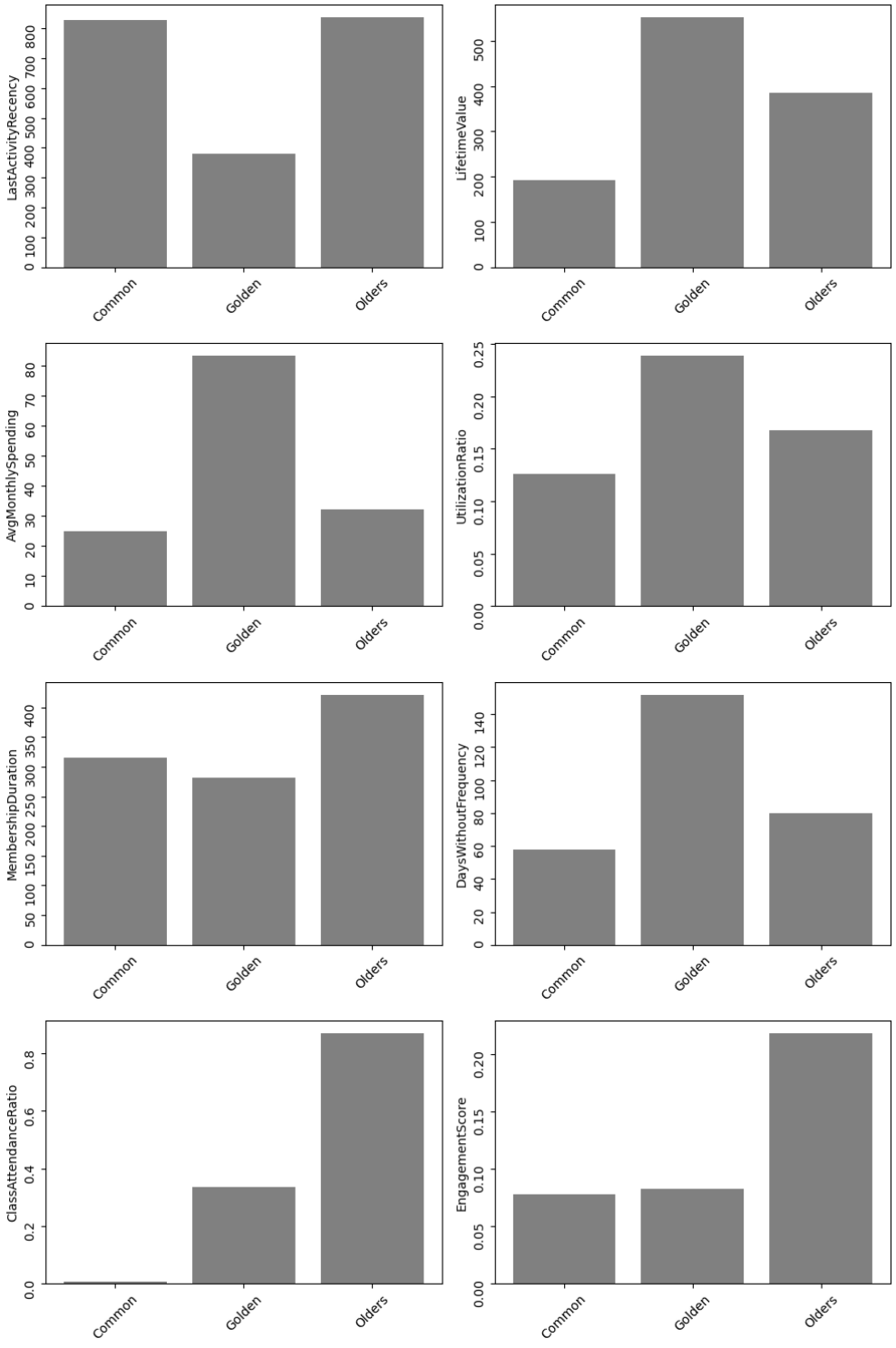

Figure 8: Plot showing value clusterization

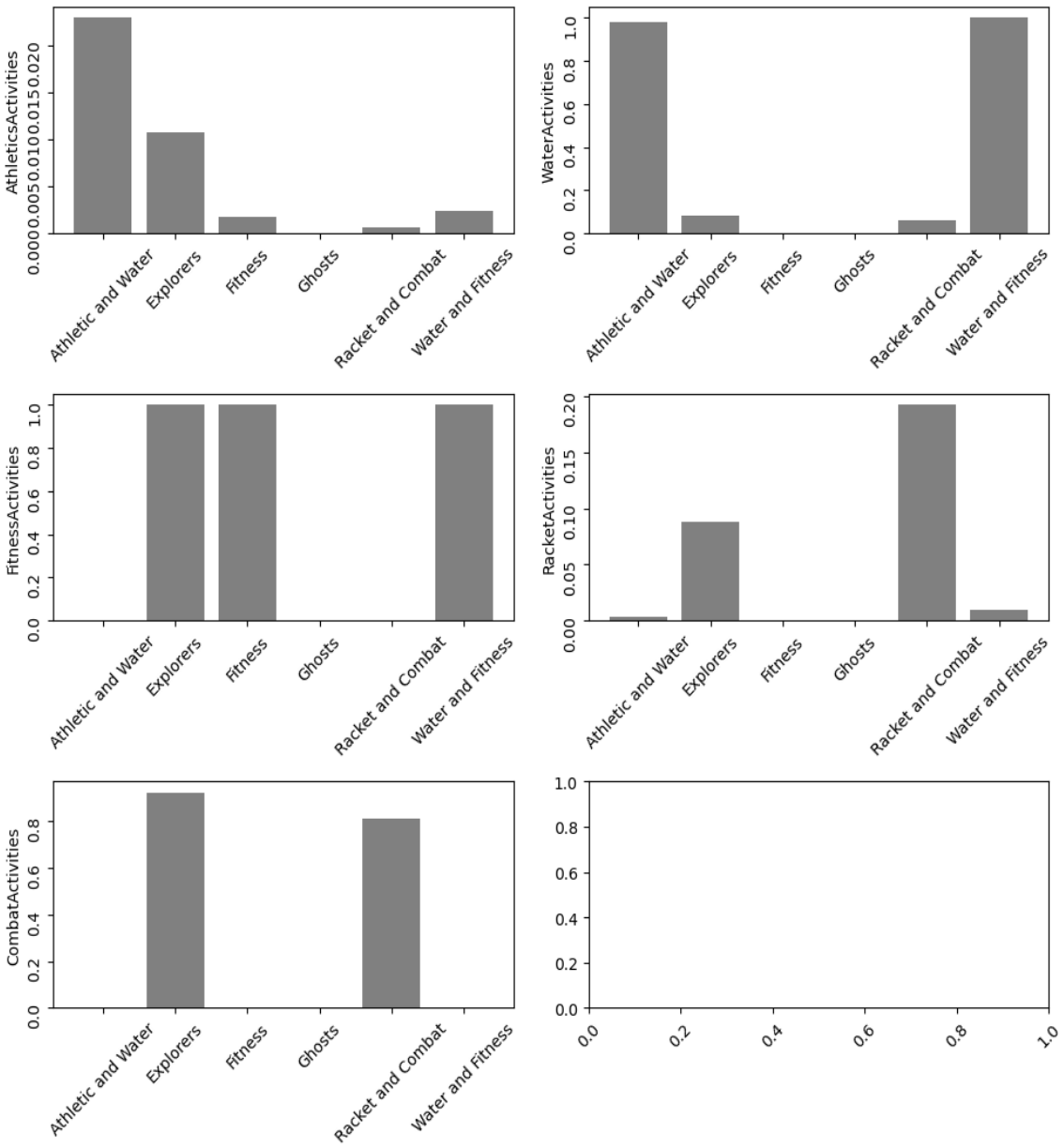

Figure 9: Plot showing sports clusterization

Demographic/DBSCAN	Sports/KMeans Value/KMeans	Athletic and Water	Explorers	Fitness	Ghosts	Racket and Combat	Water and Fitness
Adults	**Common**	1172	254	6233	160	930	249
	Golden	1051	99	1204	226	251	85
	Olders	1484	18	249	456	331	65
Kids	**Common**	2	0	4	0	3	0
	Golden	51	0	0	0	6	0
	Olders	101	0	0	2	17	1
Seniors	**Common**	10	2	32	0	4	7
	Golden	24	1	33	11	3	14
	Olders	28	0	8	9	2	6
Youngs	**Common**	0	0	4	0	3	2
	Golden	0	0	2	0	0	0
	Olders	3	0	0	2	1	0

Figure 10: Contingency matrix of clusters

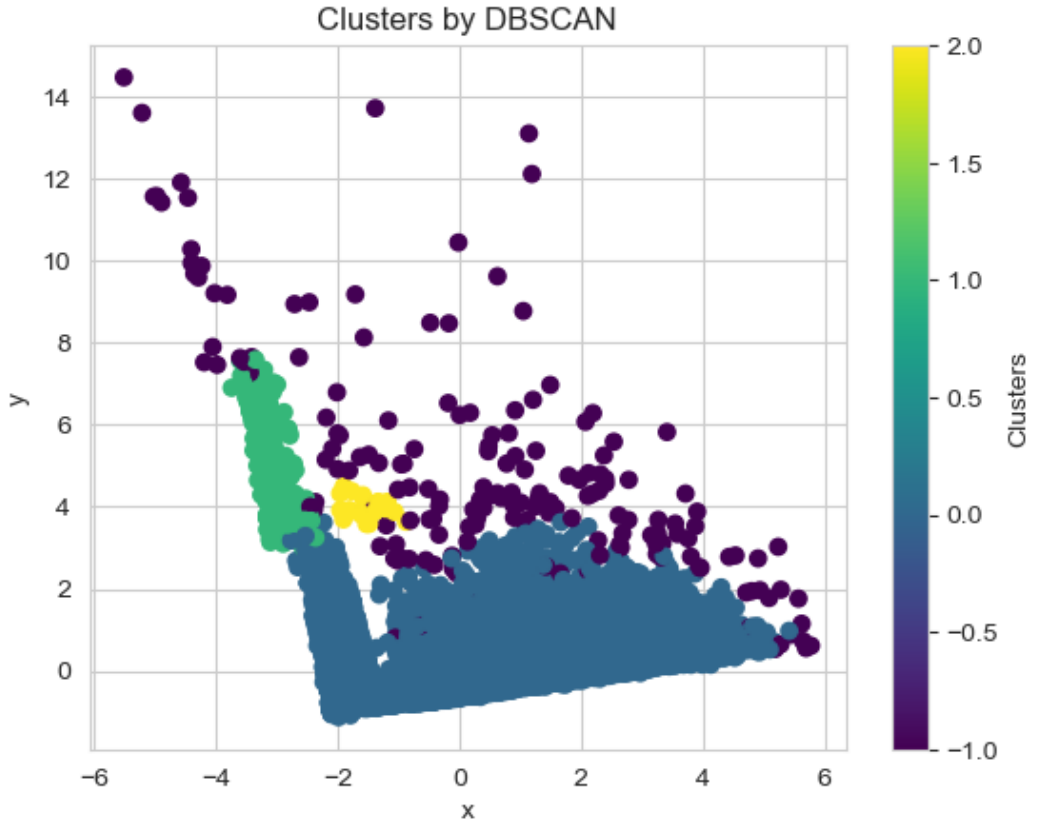

Figure 11: Final demographic cluster

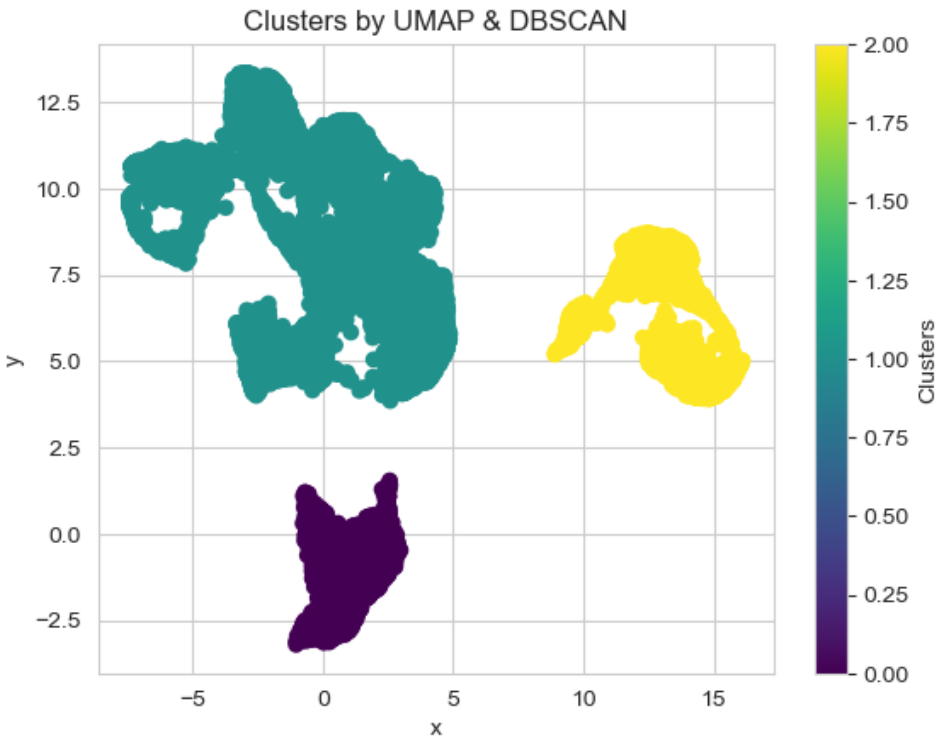

Figure 12: Final value cluster

Figure 13: Final sports cluster