Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

CUSTOMER SEGMENTATION FOR XYZ SPORTS COMPANY

Group 21

Rodrigo Silva, number: 20230536

Nicolau Dulea, number: 20230544

Joana Gonçalves, number: 20230977

January 2024

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

Abstract

XYZ Sports Company, a well-established fitness facility, seeks to enhance its marketing strategies and customer engagement by developing a comprehensive customer segmentation strategy. Leveraging data from its ERP system, the project aims to create a data-driven strategy for understanding customers, delivering personalized services, and optimizing marketing efforts. The process involved data exploration, analysis, preprocessing, preparation, feature engineering, and selection. Clustering methods, including hierarchical clustering, K-Means, DBSCAN, and Mean Shift, were employed to delineate customer segments based on diverse perspectives.

Following cluster analysis, our dataset revealed seven distinct customer segments: three for children, three for adults, and one for seniors. Key characteristics of these groups include: Kids_1, the youngest engaging in water activities with the highest lifetime value; Kids_2, older children with a focus on combat, team, and fitness activities, experiencing a higher dropout rate despite high lifetime value; Kids_3, an intermediate group with the lowest lifetime value but more references; Adults_1, consistent fitness-focused customers with higher renewals and longer usage times; Adults_2, sharing similarities with Adults_1 but with the least lifetime value, no references, and higher dropout; Adults_3, the oldest among adults, engaging in water, team, and racket activities; and Seniors, a distinct group with the highest income and frequency, active in fitness and water activities, participating in more special activities compared to other age groups.

This segmentation lays the foundation for targeted service customization and optimized marketing strategies. The culmination of this project provides invaluable insights into each customer segment's unique value, demographics, and sports activity preferences. This understanding forms the basis for customizing services and optimizing marketing strategies to align with distinct demographics and preferences. The discussion offers general insights and proposes campaigns, emphasizing the potential for business improvement, customer retention strategies, and the implementation of customer feedback mechanisms. The concluding remarks on the strategic importance of leveraging data analytics for customer centered innovation, maintaining a competitive edge, and making informed decisions. Further recommendations suggest studying patterns related to drop-out features and developing predictive models to understand factors contributing to dropouts.

Index

1. Introduction

Marketing strategies have traditionally been based on market research, demographics, or statistical methods of data analysis. With the emergence of data mining processes, this task can be improved. A popular application of clustering, a data mining technique to create homogeneous groups, is market segmentation. Clusters formed for a business purpose are usually called “segments''. The identification of customers’ groups makes use of diverse characteristics and perspectives on customers. This predictive modeling allows the understanding of the differences and rules that characterize the various segments, leading to better strategic choices (Berry & Linoff, 1997; Jain et al., 1999).

An example of that is XYZ Sports Company, a well-established fitness facility that aims to develop a comprehensive customer segmentation strategy. To enhance its marketing strategies, improve customer engagement, and tailor its services, XYZ would like to understand the value and demographics of each customer segment, as well as gain insights into the different sports activities that customers prefer to participate in. This project will leverage the dataset provided by the company’s ERP system, which includes customer-related data, collected between June 1st 2014 and October 31st 2019. The company’s goal is to create a data-driven strategy that will enable it to better understand its customers, deliver more personalized services, and optimize marketing efforts.

2. Data exploration and analysis

The metadata provided for this project is described in Appendix 1. This dataset is composed of 14942 instances and 26 features. On the whole dataset, there was one duplicate value, which was removed.

After an initial exploration of our data, we decided to do a segmentation of the dataset based on the customer’s age. Therefore, we have three initial groups: one for children, with ages ranging between 0 and 15 years; adults, from 16 years of age to 64; and elders, aged 65 and over. Young customers are inherently different from adults, who in turn are also different from the elderly, having distinct patterns of consumption. This implies that the marketing approaches will not be the same, making sense to analyze them separately.

The elders’ group is constituted by 345 individuals, being a cluster by itself. Therefore, some of the typical pre-processing methodologies were not necessary, nor clustering algorithms. For adults and kids, most of the steps for pre-processing and clustering described in this report are identical for both sets of customers. Where different techniques or values are employed, it will be mentioned.

2.1. Data preprocessing

2.1.1. Definition of type features and correction of data types

To prepare the data, the features were divided according to their data types:

- Metric features (continuous): *Age, Income, DaysWithoutFrequency, LifetimeValue, NumberOf Frequencies, AttendedClasses, AllowedWeeklyVisitsBySLA, AllowedNumberOfVisitsBySLA, RealNumberOfVisits, NumberOfRenewals* and *NumberOfReferences*.

- Nonmetric features (booleans): *Gender, UseByTime, AthleticsActivities, NatureActivities, Dance Activities, WaterActivities, FitnessActivities, TeamActivities, RacketActivities, CombatActivities, SpecialActivities, OtherActivities, HasReferences* and *Dropout.*
- Date features: *EnrollmentStart, EnrollmentFinish, LastPeriodStart, LastPeriodFinish, DateLastVisit*

The date features were converted from object (string) type to date type. The nonmetric features (except Gender) are described in our metadata as being Booleans. However, they are encoded as real numbers. Since True and False are translated into 1 and 0, respectively, we will not change the data types of these features.

2.1.2. Coherence checking

When analyzing our data, some incoherences were found, among them:

- *AllowedWeeklyVisitsBySLA* > *AllowedNumberOfVisitsBySLA*: It doesn't make sense that a customer has more weekly visits than total visits. One approach could be to delete these variables, but due to the small number of instances and the possibility of using these variables for feature engineering, we decided to exclude these instances. Later we can reintroduce and classify them.
- *HasReferences* > 0 but *NumberOfReferences* = 0: It is not possible that a customer is stated to have references, but then has no number of references. Since we have few instances where this occurs, it was assumed that we have an error on *HasReferences* and convert it to 1.
- Same *EnrollmentStart* as *EnrollmentFinish*: The date of start and finish of enrolment cannot be the same, especially when the date of the last visit to the facilities is greater than the finish date. We can see that none of these customers is a dropout, and the date of the last visit is in the last semester of 2019. This means that these customers are still enrolled i.e., they don't have a date for *EnrollmentFinish*. For the purpose of this analysis, we will consider their *EnrollmentFinish* as October 31st, 2019, the final date when data was collected.

2.1.3. Data visualization

To gain some insights of our data, some visualizations were created: histograms and boxplots for metric features and bar charts for non-metric features. It was during this visualization that we noticed that some features could benefit from a transformation, due to their right-skewed distribution. To achieve that, the Box-Cox transformation (Box & Cox, 1964) was employed. It was also possible, through bar charts, to assess that there were no incoherences of values for the non-metric features. Consequently, the features transformed were:

- For the children dataset: *DaysWithoutFrequency, LifetimeValue, NumberOfFrequencies, AttendedClasses, AllowedNumberOfVisitsBySLA, RealNumberOfVisit, NumberOfRenewals* and *NumberOfReferences.*
- For the adult and elder dataset: *Age, Income, DaysWithoutFrequency, LifetimeValue, NumberOf Frequencies, AttendedClasses, AllowedNumberOfVisitsBySLA, RealNumberOfVisits* and *NumberOf Renewals.*

2.1.4. Outlier removal

We decided to initiate our data preparation with outliers' removal, so extreme values would not accentuate their effect during missing values imputation. Different methods were tested to understand which were more effective: Z-score, Interquartile-Range (IQR) and visual inspection based on histograms and boxplots of the variables. Visual inspection was the method that preserved more of our children data set (99.53%) and adults data set (99.41%).

- For children, the parameters used to filter data were: *LifetimeValue, AllowedNumberOfVisits BySLA, DaysWithoutFrequency* and *AttendedClasses.*
- Whereas for adults, the parameters used to filter data were: *LifetimeValue, AllowedNumberOf Visits BySLA, AllowedNumberOfVisitsBySLA* and *RealNumberOfVisits.*

2.1.5. Missing values

For both the children and seniors' datasets, there were not many missing values, except for *AllowedWeeklyVisitsBySLA* and *NumberOfFrequencies*, as can be seen on Appendix 2 (figures 8 and 10, respectively). Also, for these features, KNN Imputer was applied after normalization, using Robust Scaler. This scaler was chosen due to the presence of outliers in the dataset. In the adults' dataset (Appendix 2, figure 9), KNN Imputer was applied for *AllowedWeeklyVisitsBySLA* and *Income*, following the same normalization.

To ensure a client is enrolled in at least one activity, missing values were set to 1 if the person was not enrolled in any other activity. Due to the low quantity of missing values in *HasReferences* and the activities features, and since they are binary, the mode was used as the method of imputation.

2.2. Feature engineering

Several features were created based on the existing ones:

- *Recency*: How many days have passed since the last visit (2019-10-31 - *DateLastVisit*).
- *EnrollmentTime*: Days enrolled (*EnrollmentFinish* - *EnrollmentStart*).
- *AverageSpent*: Average money spent per visit (*LifetimeValue* / *NumberOfFrequencies*).
- *HasRenewals*: If a client made a renewal or not.
- *ServiceVisitsRatio*: Ratio between the allowed and the real number of visits in the service time (*RealNumberOfVisits* / *AllowedNumberOfVisitsBySLA*).
- *AttendedClassesPerVisit*: (*AttendedClasses* / *NumberOfFrequencies*).

After the engineering of these new features, the date variables were removed from the dataset. They would not be useful for clustering and relevant information was extracted from them through the new features created. Non-metric features were updated, adding *HasRenewals*, while the rest of the new variables were added to metric features.

2.3. Second round of data preparation

After new features were created, the dataset was visually reassessed for possible transformations and outliers. The previously transformed features were once again transformed along with some new ones using the Box-Cox transformation, those being: *EnrollmentTime*, *AverageSpent*, *ServiceVisitsRatio*, *AttendedClassesPerVisit* and *EnrollmentTimeByRenewal* (Appendix 3).

For outliers the same methods as in point 2.1.4. were tested. For both children and adults, visual inspection proved to be the method that kept more data (98.8% and 99.23 %, respectively).

2.4. Feature selection

To avoid the curse of dimensionality, the number of features were reduced based on redundancy and relevancy. The former was assessed through analysis of the correlation matrix for the metric features.

For the children dataset (Appendix 4, figure 12) certain feature selection choices were made: the feature *DaysWithout Frequency* showed high correlation with *Recency* and *ServiceVisitsRatio*. Both *Recency* and *ServiceVisitsRatio* seemed essential and were retained, while *DaysWithoutFrequency* was taken out.

NumberofFrequencies and *AttendedClasses* showed similarity with *LifetimeValue* and were correlated with *AttendedClassesPerVisit*. In the interest of avoiding redundancy, *NumberofFrequencies* and *Attended Classes* were excluded, while *LifetimeValue* and *AttendedClassesPerVisit* were maintained.

For the features *AllowedWeeklyOfVisitsBySLA* and *AllowedNumberOfVisitsBySLA*, a decision was made to exclude one. Due to a higher prevalence of missing values in *AllowedWeeklyOfVisitsBySLA*, it was removed. *RealNumberOfVisits* demonstrated similarity and correlation with *ServiceVisitsRatio* and for this reason was excluded. *EnrollmentTime* showed high correlation and similarity with features that were retained previously, *LifeTimeValue* and *EnrollmentTimeByRenewal*. To avoid redundancy, *EnrollmentTime* was omitted in favor of the other correlated features.

For the adults' datasets (Appendix 4, figure 13), we followed the same feature selection choices as in the children's dataset, with one modification being the exclusion of *AllowedNumberOfVisitsBySLA*. After analysis, it was found that keeping this feature in the model for adults had a negative effect on the model's performance, leading to lower scores. Therefore, *AllowedNumberOfVisitsBySLA* was also excluded for the adults' datasets to enhance overall model performance.

Based on this, the features used for the children were: *Age*, *LifetimeValue*, *AllowedNumber OfVisitsBySLA*, *NumberOfRenewals*, *NumberOfReferences*, *Recency*, *AverageSpent*, *ServiceVisitsRatio*, *AttendedClassesPerVisit* and *EnrollmentTimeByRenewal* (figure 1).

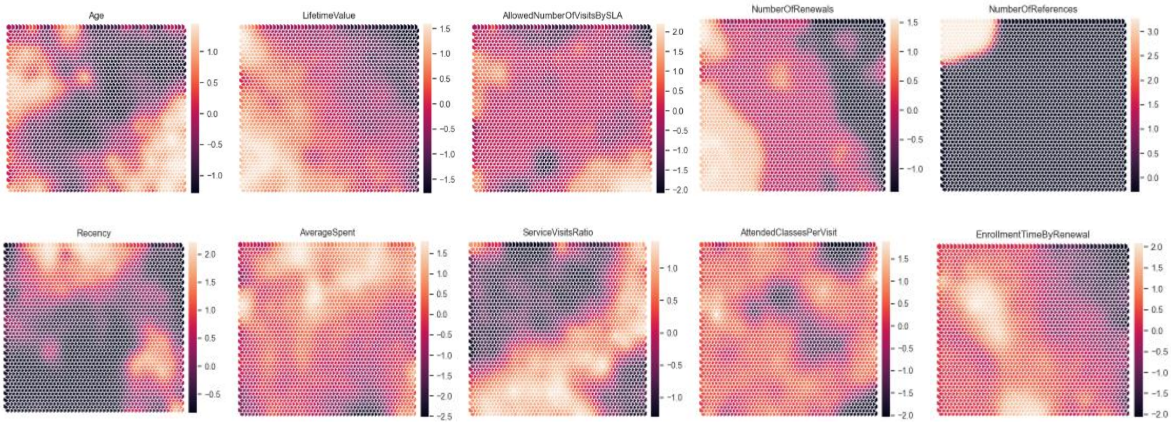

Figure 1 Component planes of the features selected for children clustering.

And for the adults set: *Age*, *LifetimeValue*, *NumberOfRenewals*, *NumberOfReferences*, *Recency*, *AverageSpent*, *ServiceVisitsRatio*, *AttendedClassesPerVisit* and *EnrollmentTimeByRenewal* (figure 2).

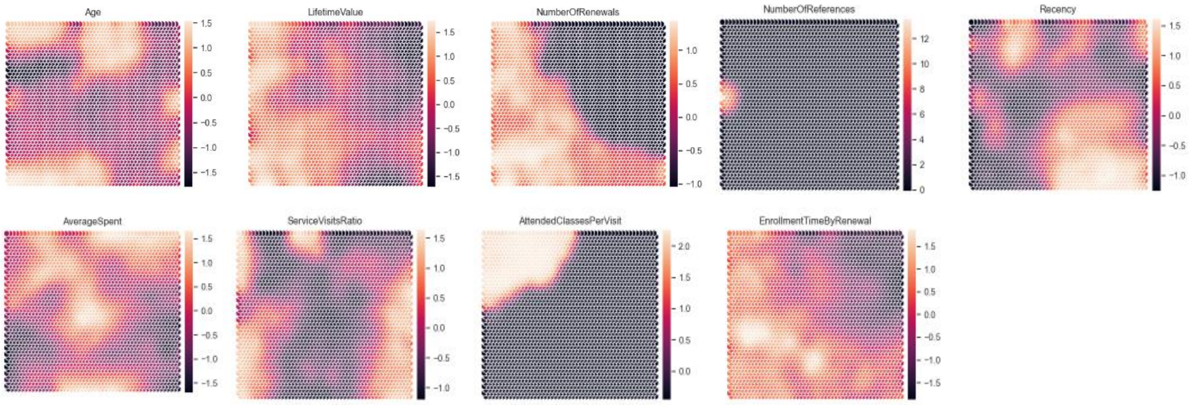

Figure 2 Component planes of the features selected for adults clustering.

3. Clustering

To do clustering of our datasets, several methods were tested. We also normalized the features selected using RobustScaler. This scaler was chosen due to its robustness to outliers, since we know that even after transformations and outlier removal, some features still have extreme values. To evaluate the clustering performance and help deciding the number of clusters, different metrics were used: R^2, Silhouette Coefficient (Rousseeuw, 1987), Calinski- Harabasz Index (Caliński & Harabasz, 1974) and Davies-Bouldin Index (Davies & Bouldin, 1979). The resulting scores for the several algorithms can be found in Appendix 5 (table 2 for children, table 3 for adults).

3.1. Hierarchical clustering

For hierarchical clustering, we started by analyzing the best linkage to be used on the AgglomerativeClustering algorithm, between "ward", "complete", "average" and "single". “Ward” proved to be the linkage that provided the best R^2, so the distance used was the Euclidean. By visual inspection of the resulting dendrograms and different scores, we decided that 3 clusters were the ideal both for children and for adults (Appendix 6, figures 14 and 15, respectively), using this clustering method.

3.2. K-Means

Knowing the disadvantages of the centroid selection process in k-means, we decided to use k-means++ (Arthur & Vassilvitskii, 2007) (since it generally initializes the centroids more distant from each other) and 300 initializations to maximize the global optimum. Besides the previously mentioned metrics, the inertia values were considered too. We also based ourselves on the prior analysis of hierarchical clusters to assess the clusters, due to their visual representation of distances. By inspection of the resulting scores, we maintained the decision of 3 clusters for both datasets.

3.3. Mean Shift

This technique (Comaniciu & Meer, 2002) for cluster density has the advantage of not requiring a priori specification of the number of clusters, relying on the parameter bandwidth (size of region to search). The bandwidth was determined separately using the estimate_bandwidth function. However, since the quantile parameter for this function needs to be manually set, it is still a subjective algorithm. Using this method to assign the data points by the estimated density did not provide great results for any of the datasets.

3.4. DBSCAN

To take advantage of the irregular shape of our clusters, we tested DBSCAN (Ester et al., 1996) to create them based on the density of our data. We determined the maximum distance between two samples with the help of the elbow method and used the number of dimensions plus one as a rule of thumb to determine the minimum number of samples to use. The result was not as informative as we had hoped for in both sets of data.

3.5. SOM

In our research, we utilized self-organizing maps (SOM) (Kohonen, 1995) to visualize component planes for the selected features, and during hierarchical and k-means clustering, using hit maps. This allowed us to visualize the spatial distribution of similarities and differences for both these methods. U-matrix analysis (Appendix 7) was used to gain insight into the data's underlying structure. However, the results were not very informative.

As k-means clustering provided the best results, it was decided to combine it with SOM. The k-means SOM uses the learned structure of a self-organizing map to initialize the centroids for a subsequent k-means clustering step. This integration overcomes some of the limitations associated with the traditional k-means algorithm, but for this data and cluster analysis, the result was unsatisfactory.

Final decision

For both children and adults, we decided to use K-Means clustering with three clusters. This number of clusters was also supported by hierarchical clustering' scores. The choice of using K-Means is based on its overall favorable scores (Appendix 5) and performance results, as evidenced in figure 3. This decision is guided by our intuitive understanding and is in line with our business logic, making it the most sensible and practical option for our specific context.

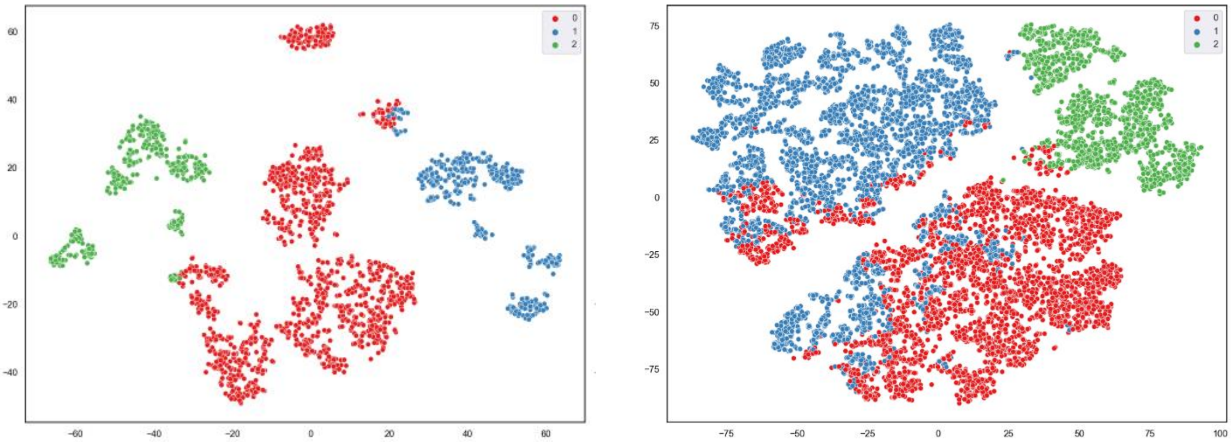

Figure 3 t-SNE for the children dataset (left) and adults' dataset (right), using K-Means, with 3 clusters each.

4. Profiling

After the different clustering algorithms were tested, our dataset was divided into 7 segments: 3 for children, 3 for adults and 1 for seniors. On figure 4, we can see the distribution of customers between the groups. Although the whole dataset appears to be uneven, between age groups the clusters are more balanced. On Appendix 8 it is possible to have an overview of the differences between the clusters' means for the dataset features.

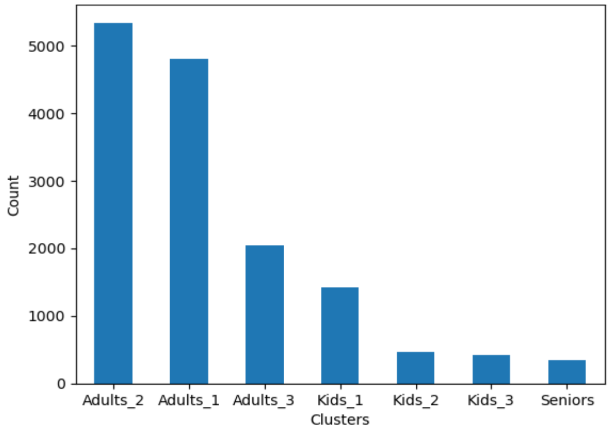

Figure 4 Number of instances per cluster.

4.1. Characterization of groups

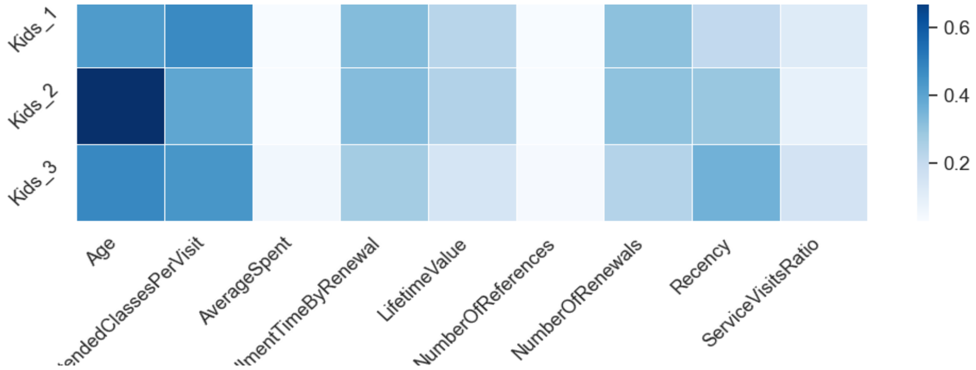

Figure 5 Heatmap evidencing differences between the kids' clusters.

- **Kids_1**: it is the group with the youngest customers. Although it is the group with the least average money spent per visit, it is the one that has the highest lifetime value, among children. The main activities that this group practices are water activities.
- **Kids_2**: group with the oldest children. In this group the dropout is bigger, even though the lifetime value and frequency of the facilities is high. This group practices more combat, team, and fitness activities, compared with other children.
- **Kids_3**: this is an intermediate group in terms of age and activities practiced, however, it is the group with the least lifetime value. Although the dropout is not as big as the previous one, this

group has a bigger recency value, frequenting the facilities less than the others. Among children, this is the group that has more references.

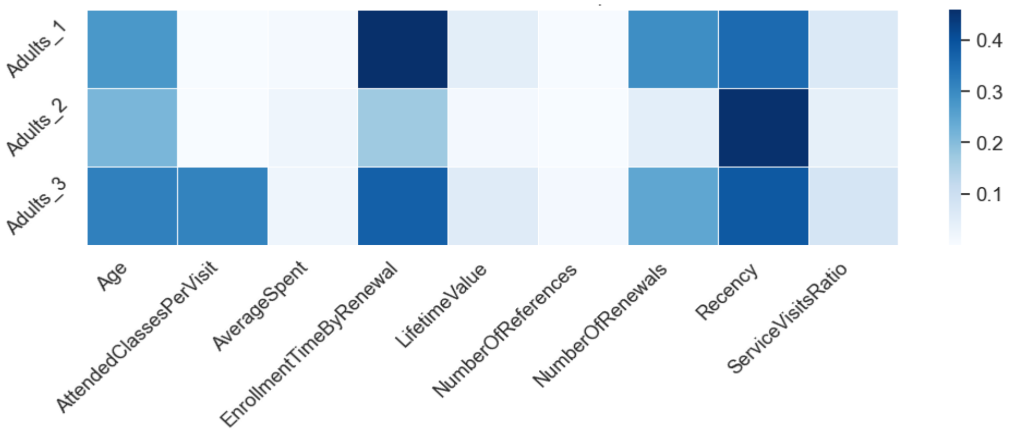

Figure 6 Heatmap evidencing differences between adults' clusters.

- **Adults_1:** this group represents more constant customers, having more renewals and a higher number of visits and use by time. However, they present the least average money spent per visit. This group practices mainly fitness activities.
- **Adults_2:** very similar to Adults_1 in terms of age and activities practiced. It is the only adult group that does not practice any special activities. This is the group with the least lifetime value, has no references and the dropout and recency values are higher.
- **Adults_3:** this group is the oldest among adults, having greater income and more references. For the rest it is more intermediate between the other adult groups, despite being the one with less allowed visits. This group is the one that practices more water, team and racket activities.
- **Seniors:** constitute a group by themselves, unlike other age groups. They present the highest income among all groups and are the most frequent ones. This group practices more fitness and water activities and are the users that engage in more special activities.

4.2. Characterization of activities

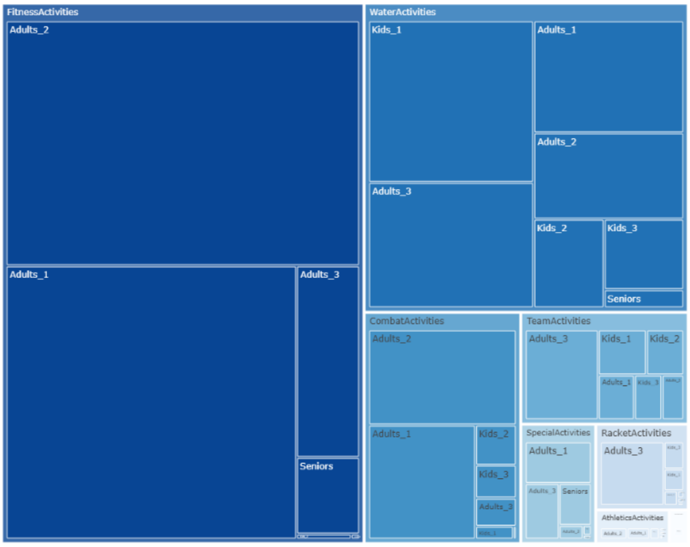

Figure 7 Tree map for the several activities practiced by the different clusters.

We can observe (figure 7) that the main activities practiced on XYZ facilities are fitness and water activities, followed by combat and team activities. The predominant activities for the age groups are fitness for adults and seniors and water activities for children. We can also note that *OtherActivities* are almost irrelevant, while nature and dance activities are not practiced at all.

4.3. Reclassifying outliers

Finally, our previously excluded outliers need to be classified into one of the clusters previously defined. For that, a Decision Tree Classifier was trained with a dataset composed of the different clusters. Since we did not have many outliers (only 122 instances), we do not expect the cluster values to differ much. Thus, 28 instances were added to children clusters (4 to Kids_1, 6 to Kids_2 and 18 to Kids_3) and 94 to adults (46 to Adults_1, 34 to Adults_2 and 14 to Adults_3). Since the datasets were divided based on age, it was expected that no outliers would be reassigned to the seniors' cluster.

5. Discussion: General insights and possible campaigns

XYZ would like to understand the value and demographics of each customer segment, as well as gain insights into the different sports activities that customers prefer to participate in. While going through the business information, we discovered useful information for the format and management of the business.

With a balanced number of men and women and a considerable proportion of affiliates in relation to the possibility of using the plan, the company can improve its business by making better use of the spaces it has and improving strategies to retain customers. One of our suggestions would be to reduce or eliminate dance and nature activities, since there are no customers enrolled in them. On the other hand, those with very little participation, such as "Other activities", where a total of 28 people participated, can be converted to occasional workshops.

The clustering work done so far has also given rise to ideas about marketing and how the company can use clusters in the business context. There are a significant number of dissatisfied customers, particularly in the Kids_3 and Adults_2 clusters, for whom the company could promote a discount to encourage them to return. Using the income information will also be valuable for suggesting more products and services, especially for the Seniors and Adults_3 groups. It would also be interesting to pay attention to Kids_2, who are the group of children who most often drop out of activities, so offering packages of classes or discounts on different activities would be a way to keep them in the company at the time of transition from child to teenager, when tastes tend to change.

It should be noted that a questionnaire could be created for each client to rate the facility, so that information can be gathered about the equipment, leisure areas and friendliness of the staff. The company could use this data to make improvements, which would not only result in more customers, but also in the loyalty of current ones.

6. Conclusion

With the delineation of customer segments, we now have invaluable knowledge about the unique value each segment brings to our business. This newfound knowledge is the basis for tailoring our services and fine-tuning our marketing efforts to meet specific demographics and preferences. As we navigate the dynamic fitness industry, this project positions XYZ not only to meet current customer expectations, but also to proactively adapt to emerging trends.

From our perspective, this academic project underscores the fundamental role of data-driven decision making in defining our strategic initiatives. The success of this initiative is consistent with our commitment to using analytics to make informed decisions, engage customers, and maintain a competitive edge. As XYZ continues to evolve in the fitness industry, this research is a testament to the enduring value of leveraging data analytics for strategic planning, customer-centric innovation, and the continued success of our unique perspective.

For further research, we recommend studying the patterns over drop out features. A good approach would be to develop a predictive model to gain knowledge from the people who tend to drop the facility. This could be an interesting approach to understand what the clients value the most.

7. References

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics.

Berry, M. J. A., & Linoff, G. (1997). Data Mining Techniques for Marketing, Sales, and Customer Support. (2nd ed., 2004). John Wiley & Sons.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society B, 26, 211-252.

Caliński, T., & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. Communications in Statistics - Theory and Methods, 3, 1-27.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2), 224-227.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR (pp. 226–231). AAAI Press.

Jain, A. K., Murthy, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. ACM Computing Reviews.

Kohonen, T. (1995). Self-Organizing Maps. Springer, Berlin, Heidelberg.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics, 20*, 53–65

8. Appendix

8.1. Appendix 1 - Metadata

Table 1 Metadata provided by XYZ's ERP system.

Feature name	Description
ID	Unique identifier of the record
Age	Age of the user on October 31st 2019 if it is not a dropout, or age of the user at date specified in attribute *EnrollmentFinish* if it is a dropout
Gender	Gender of the user
Income	Monthly salary of user
EnrollmentStart	Date of first enrollment
EnrollmentFinish	Finish date of last enrollment
LastPeriodStart	Start date of the last activity or the last two months if less
LastPeriodFinish	End date of last activity or last two months if less
DateLastVisit	Date and time of the user's last visit to the sport facility
DaysWithoutFrequency	Number of days the user did not visit the facility before being considered a dropout
LifetimeValue	Total amount paid by the customer during the period in which he was enrolled (between *EnrollmentStart* and *EnrollmentFinish*)
UseByTime	Indicates whether the user was enrolled in this form of use
AthleticsActivities	Indicates if the user was ever enrolled in athletics activities during the period between *EnrollmentStart* and *EnrollmentFinish*
WaterActivities	Indicates if the user was ever enrolled in water activities during the period between *EnrollmentStart* and *EnrollmentFinish*
FitnessActivities	Indicates if the user was ever enrolled in fitness activities during the period between *EnrollmentStart* and *EnrollmentFinish*
DanceActivities	Indicates if the user was ever enrolled in dance activities during the period between *EnrollmentStart* and *EnrollmentFinish*
TeamActivities	Indicates if the user was ever enrolled in team activities during the period between *EnrollmentStart* and *EnrollmentFinish*
RacketActivities	Indicates if the user was ever enrolled in racket activities during the period between *EnrollmentStart* and *EnrollmentFinish*
CombatActivities	Indicates if the user was ever enrolled in combat sports activities during the period between *EnrollmentStart* and *EnrollmentFinish*
NatureActivities	Indicates if the user was ever enrolled in nature activities during the period between *EnrollmentStart* and *EnrollmentFinish*
SpecialActivities	Indicates if the user was enrolled in sports for disabled people

	during the period between *EnrollmentStart* and *EnrollmentFinish*
OtherActivities	Indicates if the user was ever enrolled in other activities that do not fall into the other categories during the period between *EnrollmentStart* and *EnrollmentFinish*
NumberOfFrequencies	Number of visits to the sports facility since the date indicated in *EnrollmentStart* and the date indicated in *EnrollmentFinish*
AttendedClasses	Number of classes the user attended between *EnrollmentStart* and *EnrollmentFinish*
AllowedWeeklyVisitsBySLA	Indicates the number of weekly visits that the user can make to the facilities according to the service he had hired in the last 2 months of his registration
AllowedNumberOfVisitsBySLA	Indicates the total number of visits that the user can make to the facilities according to the service he had hired in the last 2 months of his registration
RealNumberOfVisits	Indicates the actual number of visits that the user made to the facilities in the last period his registration
NumberOfRenewals	Number of renewals during the registration period
NumberOfReferences	Number of people with which the user is related by family relationship or friendship
HasReferences	This field contains the value True if *NumberOfReferences*> 0, or False otherwise
Dropout	Represents the user's enrollment status

8.2. Appendix 2 - Missing values

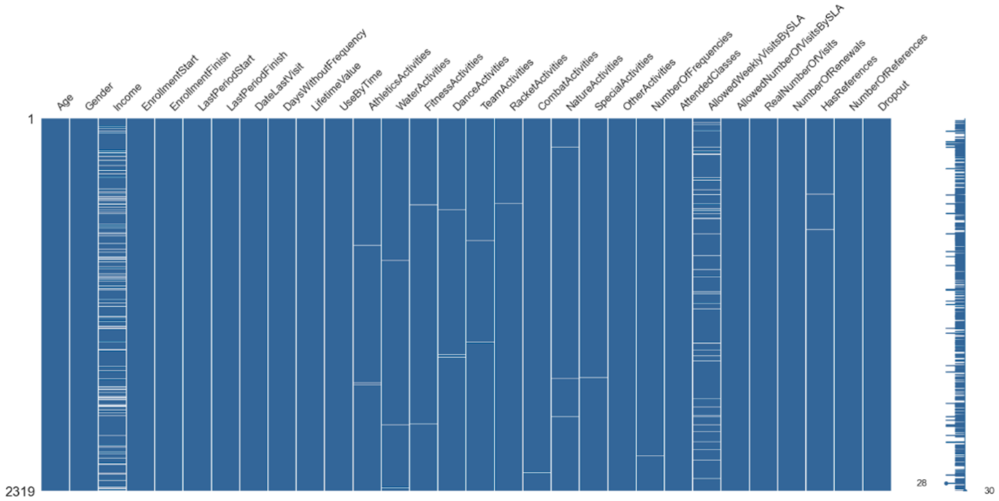

Figure 8 Missing values for the children dataset.

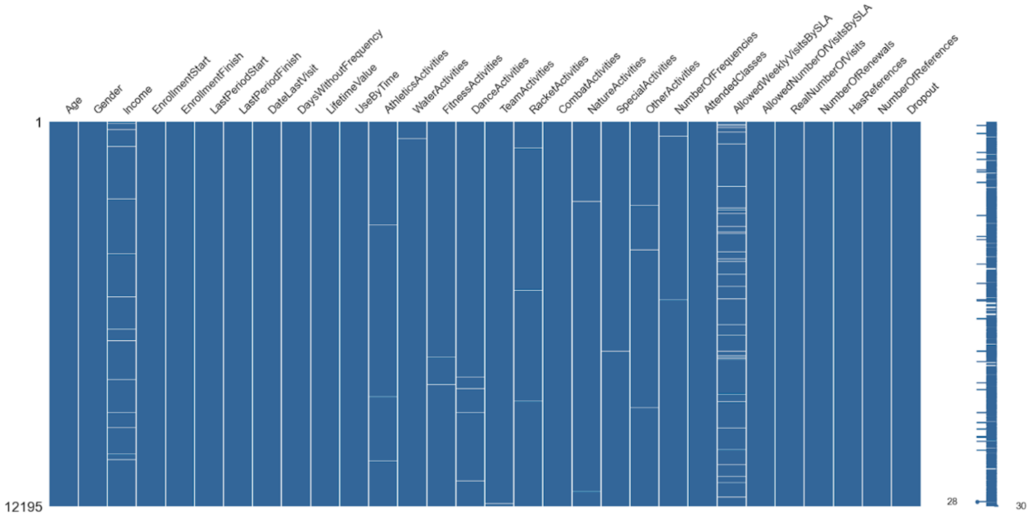

Figure 9 Missing values for the adults dataset.

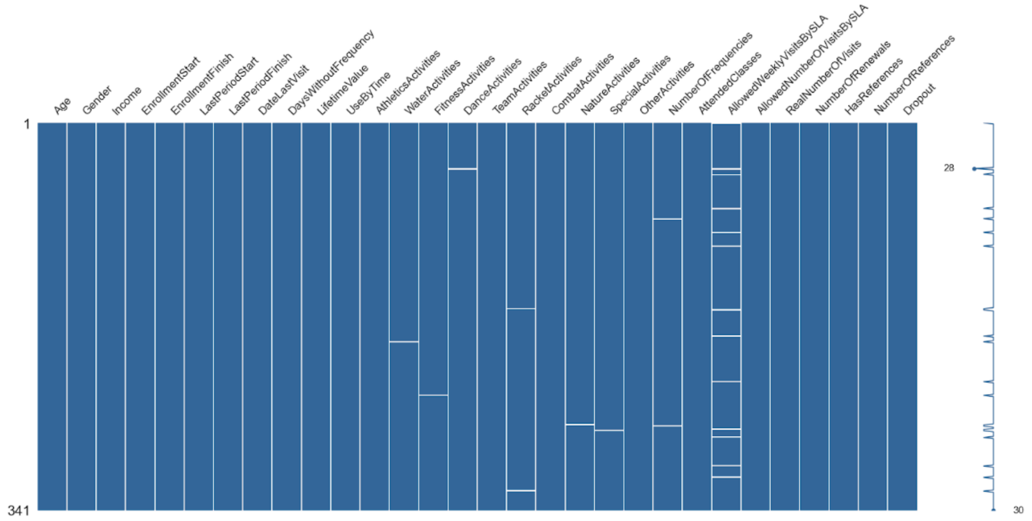

Figure 10 Missing values for the seniors dataset.

8.3. Appendix 3 - Metric features’ distributions

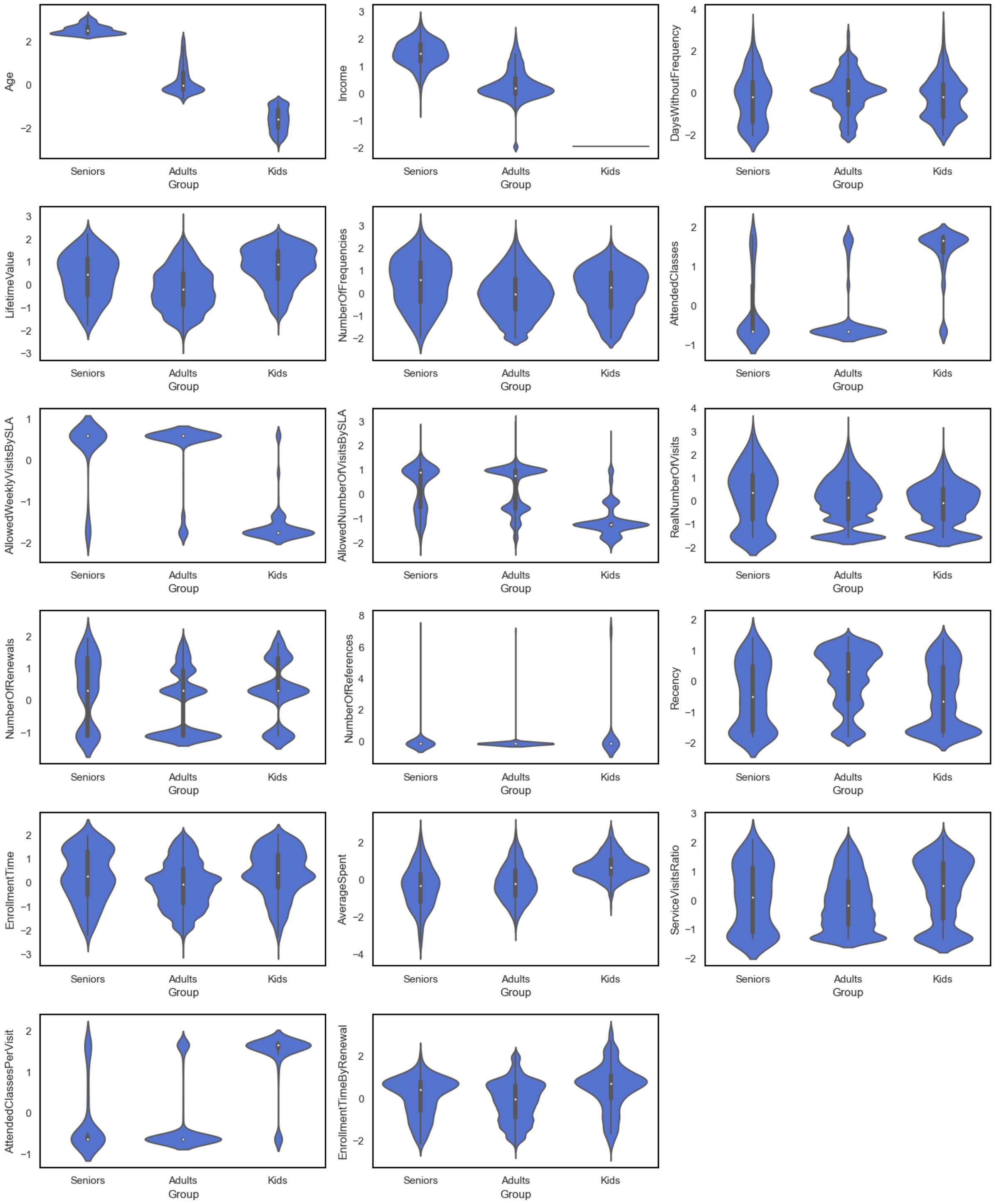

Figure 11 Metric features' violin plots for the three age groups.

8.4. Appendix 4 – Correlation matrices

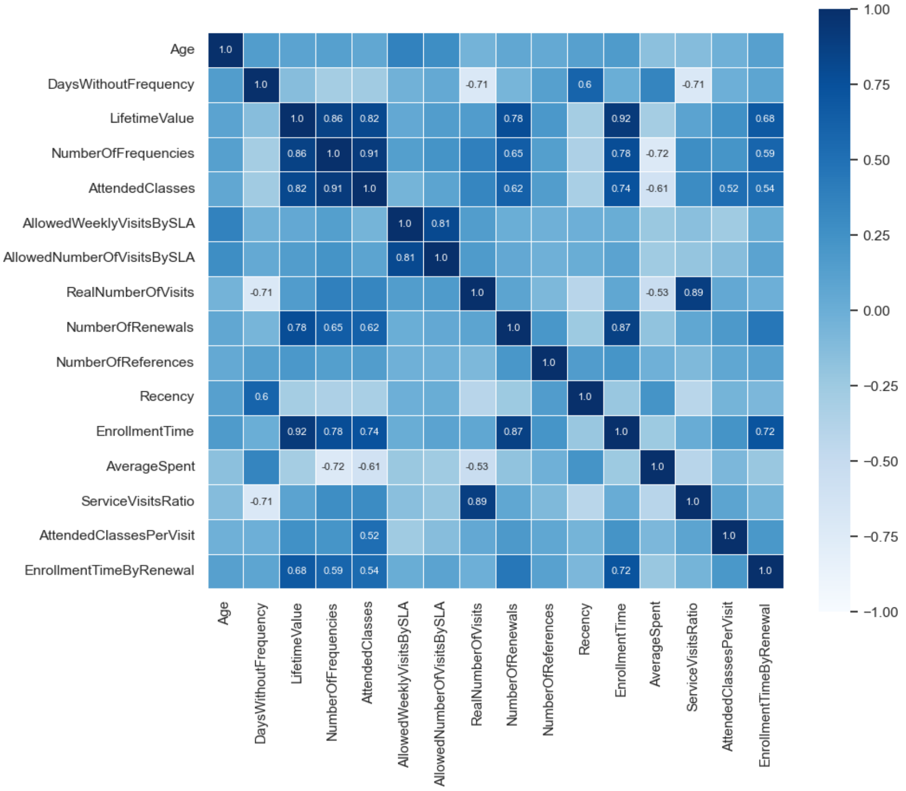

Figure 12 Correlation matrix for children metric features.

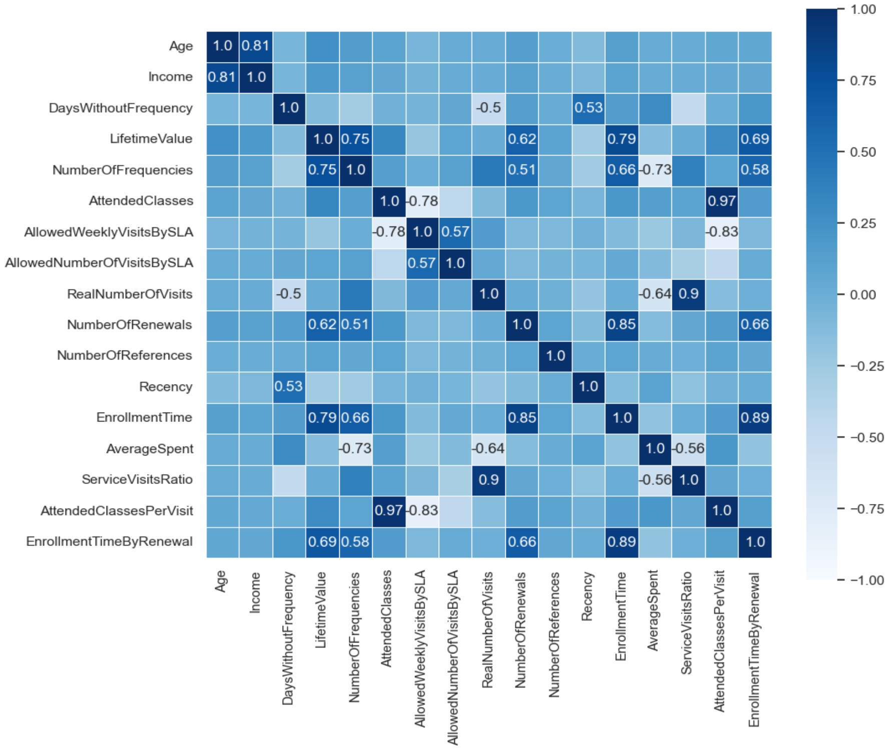

Figure 13 Correlation matrix for adults metric features.

8.5. Appendix 5 – Clustering scores

Table 2 Scores obtained for the different algorithms tested for the children dataset.

Clustering method	Number of clusters	R^2	Silhouette	Calinski and Harabasz	Davies-Bouldin
Hierarchical	2	0.5771	0.6056	3138.38	0.5859
	3	0.8519	0.6813	6610.06	0.4628
	4	0.9008	0.6651	6959.40	0.5113
KMeans	3	0.8630	0.6934	7243.14	0.4424
	4	0.90574	0.6809	7360.56	0.4991
	5	0.93015	0.6655	7646.50	0.5996
Mean Shift	3	0.9048	0.6802	7283.13	0.5026
DBSCAN	3	0.0203	0.1913	23.78	7.2545
SOM	3	0.4341	0.0651	587.68	3.4132

Table 3 Scores obtained for the different algorithms tested for the adults dataset.

Clustering method	Number of clusters	R^2	Silhouette	Calinski and Harabasz	Davies-Bouldin
Hierarchical	2	0.2839	0.3815	4826.13	1.1970
	3	0.3874	0.2069	3848.30	1.4652
	4	0.4548	0.1632	3383.08	1.6558
KMeans	3	0.4305	0.2400	4599.23	1.4818
	4	0.4965	0.2113	4000.59	1.4780
	5	0.5405	0.2006	3577.97	1.4326
Mean Shift	4	0.2863	0.3177	1627.01	1.3939
DBSCAN	3	0.0132	0.0316	81.63	2.3992
SOM	3	0.4015	0.2107	4082.21	1.6495

8.6. Appendix 6 - Dendrograms

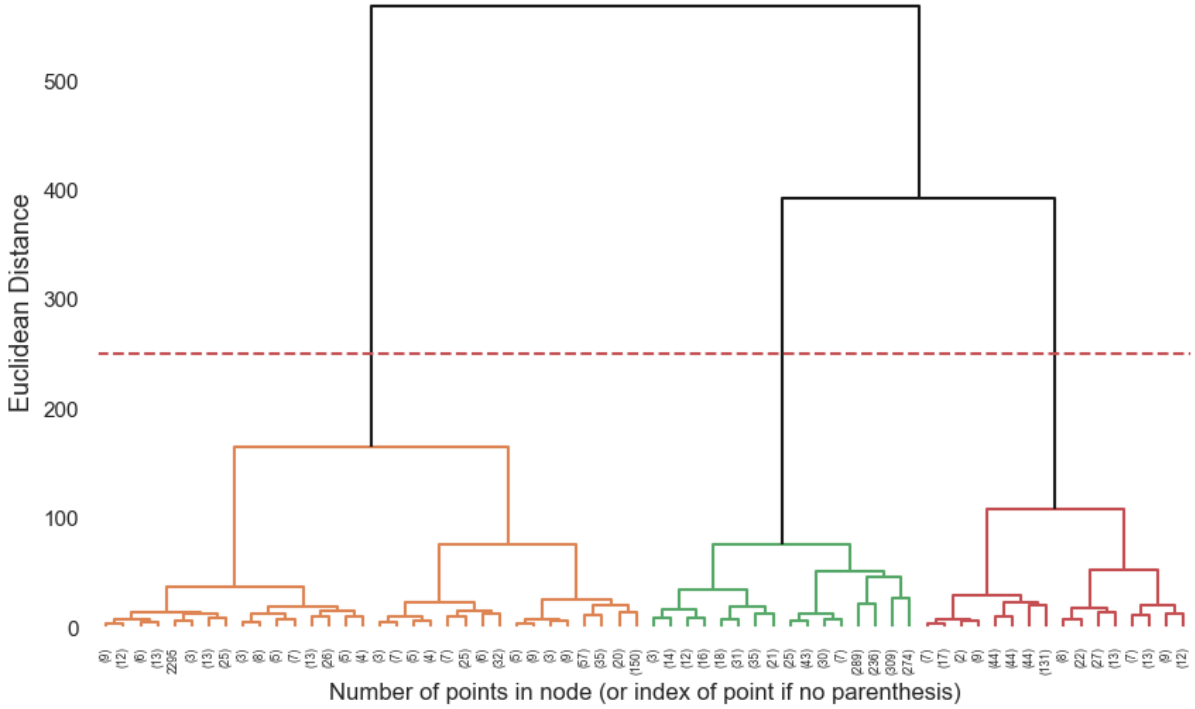

Figure 14 Hierarchical clustering for the children dataset.

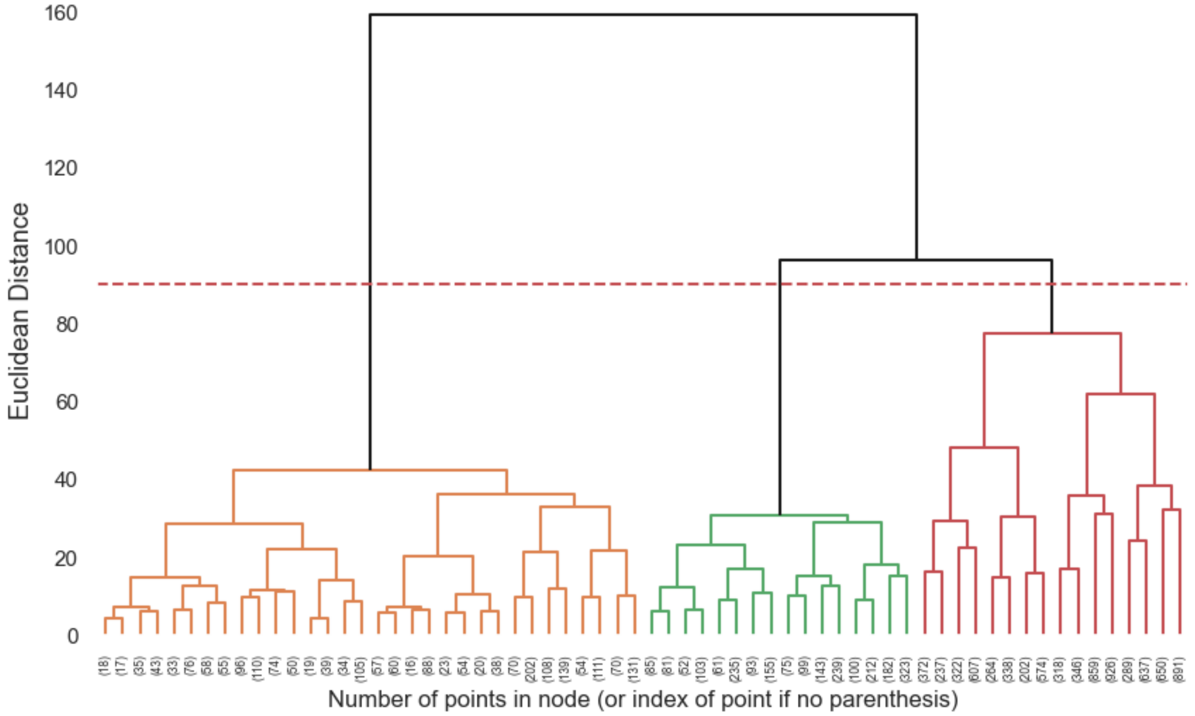

Figure 15 Hierarchical clustering for the adults' dataset.

8.7. Appendix 7 – U-matrices

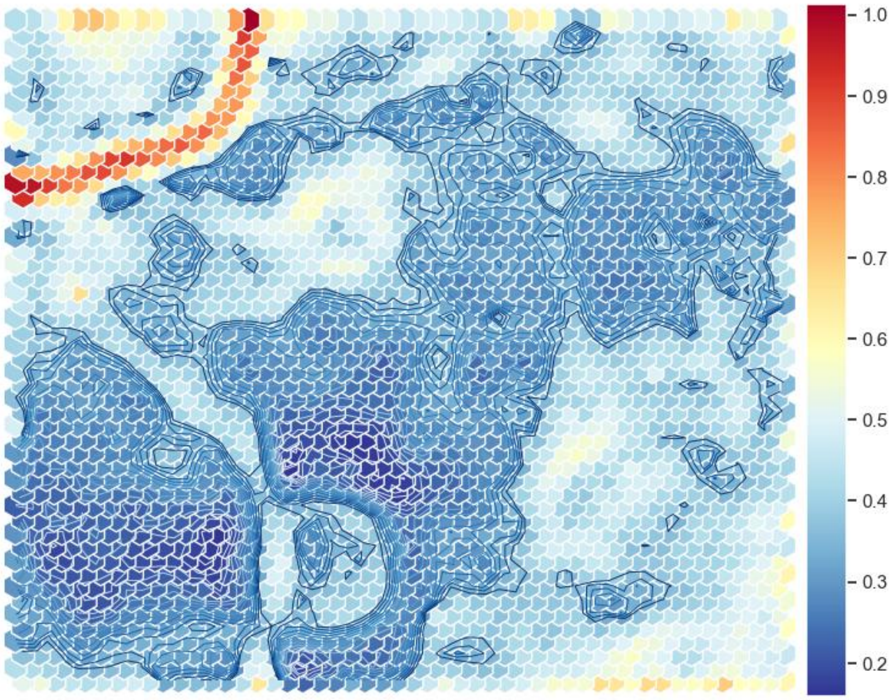

Figure 16 U-matrix for the children dataset.

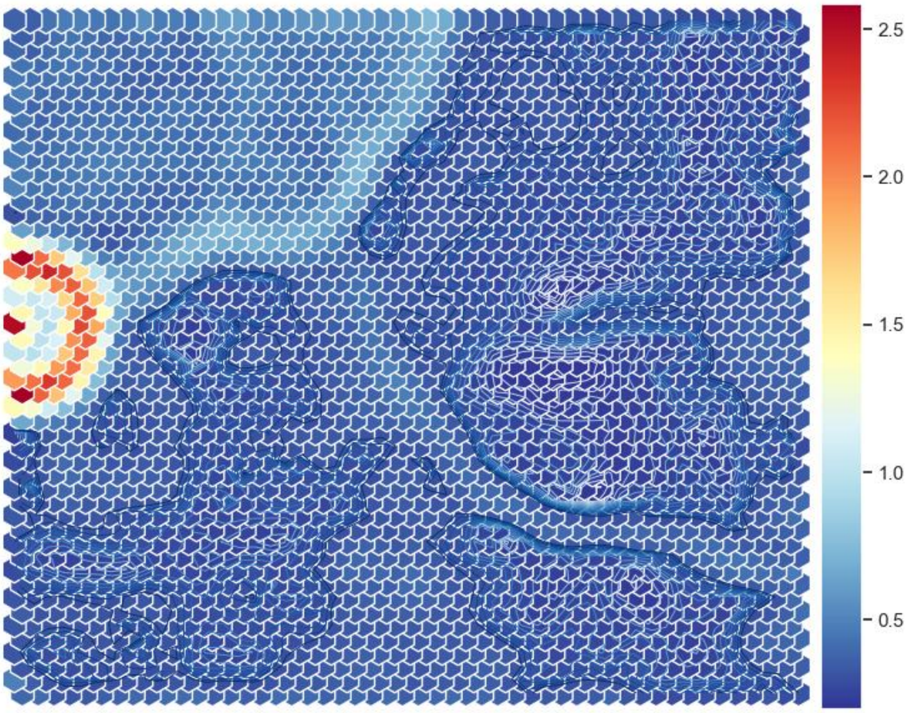

Figure 17 U-matrix for the adults' dataset.

8.8. Appendix 8 – Clusters' means

Table 4 Features' means for the different clusters.

	Adults_1	Adults_2	Adults_3	Kids_1	Kids_2	Kids_3	Seniors
Age	29.22	26.20	31.34	6.40	10.62	7.18	71.05
DaysWithoutFrequency	87.76	69.38	84.69	59.09	72.78	95.51	75.27
LifetimeValue	345.12	89.96	402.66	636.12	667.62	424.51	417.66
UseByTime	0.10	0.02	0.04	0.00	0.01	0.00	0.03
AthleticsActivities	0.01	0.01	0.00	0.01	0.00	0.01	0.01
WaterActivities	0.17	0.12	0.49	0.91	0.65	0.64	0.20
FitnessActivities	0.77	0.76	0.28	0.00	0.03	0.00	0.66
DanceActivities	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TeamActivities	0.02	0.01	0.19	0.09	0.21	0.17	0.00
RacketActivities	0.00	0.00	0.13	0.02	0.00	0.08	0.00
CombatActivities	0.13	0.14	0.03	0.02	0.18	0.16	0.01
NatureActivities	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SpecialActivities	0.04	0.00	0.06	0.00	0.00	0.01	0.24
OtherActivities	0.00	0.00	0.01	0.00	0.00	0.00	0.00
NumberOfFrequencies	69.50	9.32	42.91	41.46	60.76	25.23	78.58
AttendedClasses	0.18	0.00	31.32	36.75	48.36	20.28	8.20
AllowedWeeklyVisits	6.96	6.93	3.34	2.11	4.81	1.45	6.34
AllowedNumberOfVisits	49.82	48.04	26.32	17.40	41.65	7.95	46.76
RealNumberOfVisits	7.57	4.21	4.11	3.73	6.05	2.21	7.85
NumberOfRenewals	1.76	0.27	1.48	1.55	1.54	1.19	1.68
HasReferences	0.01	0.00	0.02	0.08	0.08	0.10	0.01
NumberOfReferences	0.01	0.00	0.03	0.09	0.09	0.12	0.01
Dropout	0.79	0.92	0.79	0.57	0.67	0.55	0.61
Recency	695.39	889.66	753.22	388.43	546.18	672.30	501.43
EnrollmentTime	612.37	135.72	514.69	636.56	624.05	491.05	585.77
AverageSpent	8.73	19.14	21.49	25.93	26.28	42.41	13.33
HasRenewals	0.87	0.22	0.69	0.77	0.78	0.68	0.70
ServiceVisitsRatio	0.16	0.10	0.20	0.22	0.16	0.28	0.18
AttendedClassesPerVisit	0.00	0.00	0.79	0.82	0.68	0.76	0.15
EnrollmentTimeByRenewal	325.87	126.02	266.67	355.95	352.64	297.50	257.77
Male	0.44	0.36	0.37	0.44	0.49	0.46	0.37
Income	2601.45	2292.86	2749.64	0.00	0.00	0.00	5526.52