

**Data Mining Project**

**MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

**XYZ Sports Company: Customer Segmentation**

Group 12

Beatriz Santos, number: 20230521

Emília Santos, number: 20230446

Ricardo Kayseller, number: 20230450

January, 2024

INDEX

[1. Introduction iii](#_Toc155560954)

[2.1. Duplicates iii](#_Toc155560955)

[2.2. Visualizations iv](#_Toc155560956)

[3.1. Missing Values iv](#_Toc155560957)

[3.2. Outliers iv](#_Toc155560958)

[3.3. Data Types v](#_Toc155560959)

[3.4. Feature Engineering v](#_Toc155560960)

[3.5. Data Normalization vi](#_Toc155560961)

[3.6. Encoding vi](#_Toc155560962)

[4.1. K-Means (All metric features) vi](#_Toc155560963)

[4.2. K-Modes (All categorical features) vi](#_Toc155560964)

[4.3. Hierarchical Clustering (K-Means and K-Modes merge) vii](#_Toc155560965)

[4.4. Self-Organizing Maps – MiniSOM vii](#_Toc155560966)

[4.5. Density Based Clustering viii](#_Toc155560967)

[4.5.1. DBSCAN viii](#_Toc155560968)

[4.6. Clustering by Perspectives viii](#_Toc155560969)

[4.6.1. K-Prototypes ix](#_Toc155560970)

[4.6.2. Testing on Hierarchical and K-Means x](#_Toc155560971)

[4.6.3. Merging the Perspectives x](#_Toc155560972)

[4.6.4. Final Clusters Analysis xi](#_Toc155560973)

[7. References xiv](#_Toc155560974)

[8. Appendix xv](#_Toc155560975)

# Introduction

XYZ Sports Company, an established fitness facility with a longstanding commitment to community wellness, aims to redefine its customer engagement and marketing strategies through advanced data mining clustering techniques. This project is centered around the development of a strategic customer segmentation framework, utilizing an extensive dataset from the company's ERP system, covering the period from June 1st, 2014, to October 31st, 2019. The objective is to dissect the customer base into distinct segments by analyzing a spectrum of data encompassing demographic details, behavioral patterns, and sports activity preferences. This segmentation is intended to empower XYZ Sports Company with deeper insights into its customer base, enabling the delivery of more tailored services and the formulation of targeted marketing campaigns.

1. **Data Exploration**

To better explore our data, an understanding of the business problem must be done. Afterwards, we can analyse the dataset and its variables, conduct the descriptive statistics, and look for possible errors, duplicate, check data types and did some visualizations.

Firstly, we started by loading the CSV file ‘XYZ\_sports\_dataset’ that was assigned to this project. This file is composed of 14 942 rows and 31 columns, that correspond to our variables, from which 16 of the variables are considered floats, 9 of them as integers and 6 as objects (Annex 1). We have this information by using the method ‘.info()’, which also reveals the features that have missing values: *AllowedWeeklyVisitsBySLA, Income, AthleticsActivities, WaterActivities, FitnessActivities, DanceActivities, TeamActivities, RacketActivities, CombatActivities, NatureActivities, SpecialActivities, OtherActivities, NumberOfFrequencies* and *HasReferences*.

Going through the descriptive statistics relative to the numeric variables (Annex 2), we observed that for some variables (*Age, DaysWithoutFrequency, Income, NumberOfFrequencies, LifetimeValue, AttendedClasses, AllowedNumberOfVisitsBySLA, RealNumberOfVisits*) the range between the third quartile (75%) and the maximum value is too big, which indicates the possible existence of outliers. For instance, the values of the third quartile and the max value for the variable *NumberOfFrequencies* are, respectively, ‘45.000’ and ‘1031.00’. The group also noticed that many variables have the minimum value as ‘0’ and maximum as ‘1’, possibly meaning that these are binary, boolean or considered categorical features. Regarding the descriptive statistics relative to the categorical (object) features (Annex 3), we can observe the number of frequencies, how many unique values the features have, the value that appears the most in the dataset and the respective frequency. For example, *Gender* has 2 unique values and the most common one is ‘Female’, appearing 8931 times.

Checking the count of each value for all variables, we came across some aspects that are worth analyzing, such as the fact that the features *DanceActivities* and *NatureActivities* only have one value.

## Duplicates

After exploring the dataset that was given to us, we checked for duplicates and one was found, so we dropped it, ending up with a dataset with no duplicates.

## Visualizations

To better understand our data, we produced some visualizations. From those, we concluded that our dataset has over 8000 ‘Female’ individuals and about 6000 males (Annex 4). Looking at the numeric variables’ histograms, we observed that most of the instances lie between the 20 and 30 years old (Annex 5) and that his most frequent number of visits to the sports facility between the EnrollmentStart date and EnrollmentFinish date is between 0 and 200 (Annex 6). Regarding the distribution of the monthly *Income,* we stated that most of the individuals earn less than 3000 (monetary units) (Annex 7). Analyzing the numeric variables box plots, we saw that various features, such *Age* and *Income* had outliers, some of which we removed (Annex 8). We did a few comparisons between variables, for instance, between *Age* and *AllowedWeeklyVisitsBySLA*, concluding that most of the users with less than 60 years can go as many times as they want, weekly, to the sports facilities, according to the service hired (Annex 9). The other comparison made was relative to *Age* and *Income*, stating that few individuals with less than 20 receive more than 2000 and that the older the person is, the higher the income (Annex 10). Relatively to the Correlation Matrix, the variables *Age* and *Income* are the most correlated ones (0.88), followed by *NumberOfRenewals* and *LifetimeValue* (0.71), and *LifetimeValue* and *NumberOfFrequencies* (0.67) (Annex 11). Lastly, we conducted non-metric variables’ absolute frequencies, from which we identified high imbalanced variables such as *HasReferences*, *OtherActivities* and *AthleticsActivities* (Annex 12).

1. **Data Preprocessing**

Firstly, and taking into account our data exploration, we decided to remove the variables *NatureActivities* and *DanceActivities*, since they only had ‘1’’s and ‘0’’s as values, giving us no insights. We also dropped *HasReferences*, because the information that it gave us was basically the same as *NumberOfReferences* (value ‘0’ in this feature means that the individual has no references). The feature *OtherActivities* was the other variable that was removed, it was very imbalanced, i.e., 98% of its values wer ‘0’’s.

## Missing Values

In our dataset, it was possible to identify missing values in the following features: *Income* (3.31%), *AthleticsActivities* (0.24%), *WaterActivities* (0.25%), *FitnessActivities* (0.23%), *TeamActivities* (0.23%), *RacketActivities* (0.25%), *CombatActivities* (0.22%), *SpecialActivities* (0.29%), *OtherActivities* (0.23%), *NumberOfFrequencies* (0.17%) and *AllowedWeeklyVisitsBySLA* (3.58%).

So, to replace these missing values, we decided to use the KNN Imputer for the metric features (*Income*, *NumberOfFrequencies, AllowedWeeklyVisitsBySLA*) and mode to the non-metric ones (*AthleticsActivities*, *WaterActivities*, *FitnessActivities*, *TeamActivities*, *RacketActivities*, *CombatActivities*, *SpecialActivities, OtherActivities).* We choose to do the mode for the categorical variables, since the KNN Imputer, generally, involves distance measures, such as the Euclidean distance, which makes this method more effective in numeric features.

## Outliers

Having the numeric variables’ box plots (Annex 8) to observe the outliers, we chose to remove the ones on the variables that had the most amount: *Age, Income, DaysWithoutFrequency, NumberOfFrequencies, AttendedClasses, AllowedNumberOfVisitsBySLA, RealNumberOfVisits* and *NumberOfRenewals.* To deal with these we resort to the manual method of defining thresholds, for instance, considering the feature *NumberOfRenewals*, we decided to consider the numbers above 6 as outliers.

## Data Types

In this phase, we took into consideration some variables with float as data type (*Income, LifetimeValue, NumberOfFrequencies, AllowedWeeklyVisitsBySLA, AllowedNumberOfVisitsBySLA*) and all the variables that had only two unique values (*UseByTime, AthleticsActivities, WaterActivities FitnessActivities DanceActivities TeamActivities, RacketActivities, CombatActivities, SpecialActivities, Dropout*), these being ‘0’ and ‘1’, and converted them to integers. Additionally, the variables regarding dates (*EnrollmentStart, EnrollmentFinish, LastPeriodStart, LastPeriodFinish, DateLastVisit*) were converted to datetime. These conversions were made to easily operate with the variables.

## Feature Engineering

In the feature engineering phase of our data analysis, we strategically expanded our dataset with the creation of additional variables derived from existing data. This approach aimed to get better insights of costumers’ behaviors within the fitness center’s ecosystem. By crafting these new variables, we intended to extract and quantify subtle patterns of engagement, attendance, and overall facility utilization.

* *LastPeriod\_Duration:* To gain insights into recent customer activity, we've devised a metric that quantifies the duration of a customer's most recent interaction with the gym. This metric is determined by calculating the time difference in days between two important dates: *LastPeriodStart* and *LastPeriodFinish*. These dates originally come in a date format but require some feature engineering to transform them into numerical data. This transformation is essential as it provides a more relevant and interpretable form of data for our analysis. Essentially, this feature plays a crucial role in assessing current customer engagement levels and understanding their recent activity patterns.
* *Enrollment\_Duration*: Similarly, we converted the *EnrollmentStart* and *EnrollmentFinish* dates into a continuous variable. This feature represents the total length of a customer's enrollment period, calculated from the *EnrollmentStart* to the *EnrollmentFinish* date. It provides a window into the long-term engagement of a customer, offering insights into loyalty and potentially indicating the lifetime value of the customer to the gym.
* *AverageVisitFrequency*: This feature was developed to assess the regularity of a customer’s visits, consisting in the ratio of the total number of real visits (*RealNumberOfVisits*) to the entire enrollment duration (*Enrollment*\_Duration).
* *ClassAttendanceRatio*: Calculated as the ratio of attended classes (*AttendedClasses*) to the total number of gym visits (*NumberOfFrequencies*), this feature aims to quantify the customer’s preference for structured classes over other normal gym activities. It highlights the importance of group fitness classes in the customer's gym usage pattern.

## Data Normalization

To normalize our data, the method that we used was the Standard Scaler. Since we still had a considerable number of outliers, even after the outlier removal phase, this was go to choice. We also tested with MinMaxScaler but the results that we got using Standard Scaler were the best ones, in general.

## Encoding

In this data preprocessing phase, we focused on encoding categorical and boolean features. For the boolean columns, such as *UseByTime*, *DropOut* and activity features, we converted the ‘True’ and ‘False’ values to numerical equivalents, with ‘True’ being replaced by ‘1’ and ‘False’ by ‘0’. This transformation simplifies the interpretation of these variables for subsequent analysis. Additionally, we addressed the categorical *Gender* column by creating a mapping dictionary that assigns 'Male' to 0 and 'Female' to 1, enabling us to represent gender as a numerical feature.

1. **Clustering Algorithms**

## K-Means (All metric features)

In the initial phase of clustering, the K-Means algorithm was chosen to evaluate the performance of the metric features of our dataset. K-Means algorithm is one of the most used algorithms in Clustering Analysis and works by iteratively assigning data points to the nearest cluster centroid and updating the centroids to minimize the sum of squared distances between data points and their assigned centroids. Our exploration began with an examination of the inertia values (Annex 13)**,** which quantifies the sum of squared distances between data points and their respective cluster centers. Visualizing these values in an "Inertia plot" helped us identify an optimal number of clusters by locating the 'elbow' point where inertia starts to stabilize. Subsequently, we evaluated the quality of our clustering solutions using the average silhouette score, which measures cluster cohesion and separation. By analyzing these scores and employing silhouette plots, we determined that the most meaningful and well-separated clusters could be achieved with a count of 5 (Annex 14). In our analysis, it's worth noting that although the silhouette score indicated a relatively higher value for 2 clusters, we ultimately determined that 5 clusters was the optimal choice. The final clustering solution utilized the K-Means algorithm with 'k-means++' initialization.

Our cluster analysis uncovered five distinct member segments within the dataset, each with its own characteristics. Cluster 0 represents moderately aged, mostly female members with moderate income, engagement in fitness activities, but lower visit frequency. Cluster 1 comprises older, high-income members actively participating in fitness activities with high visit frequency. Cluster 2 consists of older, loyal, and highly engaged members with low income. Cluster 3 includes younger, low-income members with moderate engagement and lower visit frequency. Cluster 4 represents younger, mostly female, low-income members with good engagement and high visit frequency.

## K-Modes (All categorical features)

To effectively manage the categorical attributes in our dataset, we utilized the K-Modes algorithm. Unlike the traditional K-Means, which is optimal for numerical data due to its dependency on distance metrics, K-Modes is adept with categorical data. It operates by creating clusters based on the frequency of category matches among data points, using modes (the most frequent category) instead of means. This method uses a simple matching dissimilarity measure to handle categorical objects, substituting the mean of a cluster with its mode, the most frequent category. We applied K-Modes using three distinct initialization methods: "Huang", "Random", and "Cao". These methods differ in how they initially position the cluster centroids. To discern the most effective approach, we used the elbow method to identify the optimal cluster count, coupled with a comparison of silhouette scores for each initialization.

The silhouette score (a metric that quantifies the similarity of an object to its own cluster relative to other clusters), served as our criterion for cluster quality. The "Huang" initialization emerged as the leading method, achieving the highest silhouette score (0.50). The silhouette score of 0.5 attained in our clustering suggests that while the customer segments are not perfectly delineated, they are sufficiently distinct to enable the identification of unique groups within the XYZ Sports Company's customer base. The Elbow method (Annex 15)suggested three clusters as the optimal number for this initialization. In this segmentation, Cluster 0 is characterized by a predominantly male demographic with a strong preference for fitness and combat activities, but it faces a high dropout rate of 83%. Cluster 1, primarily female, shows a significant inclination towards water activities with a slightly better retention rate at 72%. Lastly, Cluster 2, exclusively male, is entirely focused on fitness activities but has the highest dropout rate at 86%, indicating a critical engagement challenge.

## Hierarchical Clustering (K-Means and K-Modes merge)

To enhance our customer segmentation, we integrated the results of K-Modes and K-Means clustering. This combination approach leverages the strengths of both methods to create a more comprehensive view of our customer base. We chose five clusters to strike a balance between detail and practicality. This number allows for meaningful differentiation without overly complicating the segmentation, aligning with our marketing strategies and operational capabilities. This decision, rooted in domain knowledge, ensures that the segmentation is both insightful and manageable for targeted customer engagement. The five combined clusters reveal distinct customer profiles: Cluster 1 balances gender with a slight fitness preference and average income but shows lower engagement and a higher dropout rate. Cluster 2 indicates a similar gender mix with slightly higher income and a preference for water activities, along with a higher retention rate. Cluster 3 has a balanced gender distribution, average income, and shows a mixed interest in activities with a moderate dropout rate. Cluster 4, while similar in income and gender mix, exhibits a higher dropout rate and a varied activity preference. Lastly, Cluster 5 stands out with the highest income levels, a balanced gender mix, and a strong inclination towards fitness activities, but also the highest dropout rate, suggesting a need for targeted retention strategies.

## Self-Organizing Maps – MiniSOM

SOM's training began with an initial Quantization Error (QE) of 1.2284, indicating the average dissimilarity between the input data and their respective best matching units (BMUs) within the map. Upon the completion of the training phase, the QE was reduced to 1.177. The reduction in QE demonstrates the SOM's adaptation to the underlying data structure, yielding a more accurate and refined representation of the input space. This improvement signifies the neural network's learning progression, culminating in a map where the neuron's weight vectors more closely mirror the input vectors they represent.

The Self-Organizing Map's U-Matrix provides a compelling visual representation of the high-dimensional data's structure, capturing the similarity or dissimilarity among the data points. In the given U-Matrix, the presence of cells with values of 0.91 and 1.00, which are likely represented by warmer colors, stands out significantly. These values suggest that the corresponding neurons have feature vectors that are quite distinct from those of their immediate neighbors in the map, indicating potential outliers or unique data points that do not cluster closely with others.

Furthermore, cells with values of 0.7, 0.71, and 0.73 also show a relatively high degree of dissimilarity but to a lesser extent than the most extreme values. These could represent transitional areas in the data space, where clusters may begin to differentiate from one another. (

On the other hand, the predominance of cool, blue colors across much of the matrix suggests that most neurons have similar weight vectors to their neighbors, indicating clusters of closely related data points. These cooler shades denote shorter distances between neurons, revealing areas where data points cluster more densely. This prevalence of similarity points to the presence of groups or patterns within the dataset that the SOM has successfully learned to identify and cluster together. Please see the graph for observation in(Annex 16).

## Density Based Clustering

### DBSCAN

Upon applying DBSCAN to our dataset, which includes the metric features, the algorithm identified a total of seven distinct clusters. This estimation is critical as it informs us of the density and distribution of our data points in the multi-dimensional feature space, reflecting the underlying structure of our dataset.

In the clustering analysis using DBSCAN, the data was segmented into distinct clusters, excluding noise identified by the label '-1'. This noise label indicates data points that didn't fit well with any cluster. The remaining labels, (0, 1, 2, 3, 4, 5), represent well-defined clusters. An R-squared value of 0.5895 was achieved, indicating that the model's clusters explain about 59% of the variance within the data. This level of explanation highlights a significant separation and consistency within the clusters, suggesting that the DBSCAN model has effectively captured the underlying structure of the data.

## Clustering by Perspectives

To comprehensively understand the customer base of XYZ Sports Company, we embarked on a segmentation strategy that examines the data from three distinct perspectives: demographic, behavioral, and sports preferences.

**Demographic perspective**: The demographic data includes A*ge*, *Income*, and *Gender*, offer insights into the socioeconomic status and life stages of the customers. This data is relatively stable over time and can be easily measured, making it a reliable base for segmentation. Segmenting customers by age, gender, and income enables a nuanced approach to market differentiation, where age-related segmentation informs service customization to cater to life stage-specific fitness needs, a gender differentiation allows for targeted offerings that resonate with the unique preferences of each gender and income segmentation ensures that pricing and membership options are attuned to the financial realities of diverse economic groups, ensuring the sports facility's services are accessible and appealing to all segments of the community.

**Behavioral perspective**: Behavioral segmentation is grounded in the analysis of customers' interactions and their engagement patterns with XYZ Sports Company. It encompasses a range of features that reflect how often and how deeply customers engage with the facility and its services. This includes: *DaysWithoutFrequency, LifetimeValue, NumberOfFrequencies, AttendedClasses, AllowedWeeklyVisitsBySLA, AllowedNumberOfVisitsBySLA, RealNumberOfVisits, NumberOfRenewals, NumberOfReferences, LastPeriod\_Duration, Enrollment\_Duration, AverageVisitFrequency, ClassAttendanceRatio, UseByTime* and *Dropout*.

**Sports\_preferences**: The segmentation based on sports activities features preferences—namely, water activities, fitness activities, team activities, racket activities, combat activities, and special activities, provides insight into the customers' interests and participatory patterns at XYZ Sports Company.

### K-Prototypes

For the customer segmentation of XYZ Sports Company, we employed the K-Prototypes clustering algorithm, which is designed to handle mixed data types. This method is particularly advantageous for our dataset as it contains both numerical and categorical features., by combining the K-Means clustering algorithm approach to numerical data with the K-Modes approach to categorical data. To determine the optimal number of clusters, we utilized the Elbow Method, to determine the best number of clusters. The Elbow plot showed in (Annex 17), suggests a bend at the five-cluster mark.

The K-Prototypes clustering applied to our dataset has been quantitatively assessed using multiple metrics. The Silhouette Score of 0.2012, points to a lack of clear demarcation between clusters, implying overlap, as also depicted by the t-SNE visualization where clusters intermingle. This overlap might partly arise from the way K-Prototypes computes distances for categorical data, which may not always yield distinctly separated clusters. The Calinski-Harabasz Score, also known as the Variance Ratio Criterion, is defined as the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters, measuring how well-separated the clusters are and how compact they are. Higher value of CH index means the clusters are dense and well separated. Its Score of 2394.63, however, indicates a reasonable level of cluster density and separation, although there's room for improvement. The Davies-Bouldin Index (DBI) is an internal evaluation scheme for clustering algorithms where a lower DBI indicates better clustering. The Davies-Bouldin Index measures the quality of clustering models. It finds the average similarity of each cluster to its most similar cluster, where similarity is the ratio of distances between clusters and within clusters. A perfect clustering solution would have a DBI of 0, or some score close to 0, that would indicate the clusters were well compacted and separated. The score of 1.6615 we obtained is moderate, suggesting that clusters are neither too dispersed nor too compact, but with some room for improvement in separation. The Adjusted Rand Index (ARI) is a measure of the similarity between two data clustering’s. It is an adjustment of the Rand Index that accounts for the chance grouping of elements, which makes it more reliable for comparing the agreement of different cluster assignments of the same dataset. The Adjusted Rand Score of 0.7319 reflects a strong agreement and consistency in the clustering process, signaling that the clusters formed are stable and not random. For visual cluster analysis, we utilized t-SNE (Annex 18)reducing the dimensions to a two-component space with a perplexity setting of 30 and 300 iterations. This visualization underscores some cluster overlap, reinforcing the moderate Silhouette Score. Clusters show differentiation, yet they are not crisply defined, with some extending into others' space.

By analyzing the clusters, we can conclude that Cluster 0 represents younger or less affluent individuals who are highly engaged in fitness activities but less so in team and racket activities. Cluster 1 comprises the least affluent and possibly older members who are interested in water activities, indicating a need for age-friendly classes. Cluster 2 includes below-average income and age members who enjoy social aspects of exercise, making group classes and team-based activities appealing. Cluster 3 consists of slightly below-average income and age individuals who could benefit from efficient, high-intensity workouts due to their busy schedules. Lastly, Cluster 4 comprises the most affluent and possibly older customers who focus on health maintenance, suggesting the potential for personalized fitness plans and luxury wellness services.

### Testing on Hierarchical and K-Means

Taking into consideration only the metric variables, on this case, present on the demographic and behavioral perspectives.

The resulting data presented (Annex 19) and (Annex 20) a comparison of R-squared values across different numbers of clusters (k) for various clustering methods: K-Means, complete, average, single, and Ward’s method. K-Means achieved slightly higher R-squared values than Ward's method, in both (*demographic and behavioral*) indicating a small but notable advantage in explaining the variance within the clusters, but the decision was made to proceed with Ward's method for several reasons. Ward's method is recognized for its ability to create clusters by minimizing the total within-cluster variance. Hence, each step of the algorithm aims to merge the pair of clusters that will result in the least increase in total within-cluster variance after the merge. This approach is particularly beneficial when the dataset contains naturally hierarchical structures, allowing for a more nuanced grouping that can reflect the complexity of the data.

Furthermore, Ward's method tends to create more balanced cluster sizes, which is advantageous when analyzing demographic data to avoid overemphasizing small clusters that may not be as statistically significant. This method is also generally more robust to noise and outliers compared to K-Means, leading to more reliable and interpretable clusters, especially in datasets where outliers or noise can significantly skew the clustering.

### Merging the Perspectives

In the process of consolidating the perspectives, we performed a merge of the contingency table values to adjust for low counts, ensuring the robustness of the statistical analysis. The original contingency table provided a cross-tabulation of '*behavior\_labels'* and *'demographic\_labels'*, revealing the frequency distribution across various clusters identified by the two distinct labeling processes.

The merge strategy involved combining the less populous clusters with similar ones to maintain statistical significance. This was necessary because clusters with very few members can be less reliable and may not represent meaningful patterns in the data. By consolidating these clusters, we enhanced the interpretability of the results and reduced the potential for statistical anomalies caused by small sample sizes.

After the merge, the contingency table was updated to reflect the new cluster counts. The low-frequency labels, specifically those with counts of *67, 145, 177, 124, 207*, and *183*, were merged with their adjacent clusters, leading to a new distribution of counts across the clusters. (Annex 21) and (Annex 22)

With the Hierarchical merging, we obtained a total of 5 clusters, after analyzing the corresponding dendrogram and using ‘ward’ as the linkage method. It states that the silhouette score in k-means, was in agreement with the results from the hierarchical dendrogram decision.

### Final Clusters Analysis

Looking at the final merged clusters (Annex 23), which resulted in 5 clusters, and to the profiling with the categorical features (corresponds to the *sports\_preferences* perspective) (Annex 24 and 25)*,* we analysed the following:

**Cluster 0:** This cluster indicates a younger demographic with less disposable income. These group of individuals can be characterized as the ones who have a class attendance ration very high, despite having a low engagement metrics like real number of visits and number of frequencies. However, they have one of the lowest numbers of allowed weekly visits by SLA. The most attended activities are the ones on water, followed by the team activities. This could suggest that they may have busy schedules, prioritizing classes that are specific to their interests or fitness goals.

**Cluster 1:** This is the oldest group and the one that has the highest income, has also a high number of allowed visits, weekly and general, by SLA, which means displays above-average engagement. Looking at the number of renewals, these individuals have a low amount of them. The fitness activities are the go-to activities on these cluster. They could also represent a segment that is willing to spend more on services.

**Cluster 2:** Contrarily to the cluster 1, these persons have the lowest income and age, which my indicate that these people may be studying and still have no income. Relatively to the Lifetime Value, these cluster is the one with the highest level, as well as on the variable attended classes, i.e., it is the cluster were the persons attend to most classes, despite having the lowest number of allowed weekly visits by SLA. It’s also the group with more renewals. Their engagement level is very high.

**Cluster 3:** These group lies in the middle of the others in most of the variables, except in the number of allowed visits, weekly and general, which is very high. The last period duration correspondent to these cluster is the highest and it has the lowest number of renewals. The most practiced activity is fitness, followed by combat and water activities.

**Cluster 4:** The Lifetime Value is the lowest in this cluster and the same happens with the class attendance ratio. They don’t attend very much to the classes, even with a high allowed number of weekly visits, when compared to others. The duration of the last period is also very low in these group. Regarding the activities, as observed in cluster 1 and 3, these individuals are most interested in fitness activities.

1. **Business applications and marketing approaches**

A white text with black lines

Description automatically generatedTaking into account the final 5 clusters, we are able to consider the following business and marketing approaches for each one:

1. **Conclusion**

In conclusion, the customer segmentation project for XYZ Sports Company has culminated in the identification of distinct customer clusters, each embodying unique characteristics and engagement patterns within the fitness facility. Throughout the project, we engaged in comprehensive data exploration, preprocessing, and the application of multiple clustering techniques, which provided valuable insights into customer demographics, behaviors, and activity preferences. The final clusters, although differentiated with discernible profiles, may not exhibit perfect separation according to the clustering validation scores. This is a common occurrence in real-world data mining projects due to the inherent complexity and variability of customer data. Despite this, the clusters have offered a strategic foundation for crafting tailored marketing strategies and business applications designed to enhance customer engagement and retention. Our journey through data mining revealed the intrinsic complexities of working with mixed data types, particularly the challenge of clustering with categorical features. Many clustering algorithms are predominantly distance-based and, as such, are less effective with non-numeric data. Addressing this challenge was crucial in our project, as our goal was to achieve meaningful customer segmentation that could translate into actionable business strategies. By Using advanced techniques and the thoughtful application of algorithms like K-Prototypes and Hierarchical Clustering, we were able to navigate these complexities and identify seven distinct customer clusters. The final clusters, while offering valuable insights, are not without the potential for further refinement. The segmentation process, though guided by robust methodologies, yielded clusters that weren't perfectly distinct according to the silhouette scores and other validation metrics. Opportunities for enhancement remain open, potentially through the incorporation of additional data, further refinement of clustering techniques, or by introducing new variables to more accurately reflect the evolving landscape of fitness trends and customer behavior. This adaptive approach ensures that the segmentation framework remains dynamic and responsive to the changing needs of the customer base, underlining XYZ Sports Company's commitment to data-driven decision-making and excellence in customer service.

# References

Aprilliant, A. (2022, March 31). The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical). Medium. <https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb>

Brownlee, J. (2020, June 23). kNN Imputation for Missing Values in Machine Learning. Machine Learning Mastery. <https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/>

Davies-Bouldin Index. (2023, November 5). GeeksforGeeks. https://www.geeksforgeeks.org/davies-bouldin-index/

Hierarchical Clustering in Machine Learning. (2019, May 8). GeeksforGeeks. <https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/?ref=lbp>

Jaiswal, S. (2020, July 10). K-ModesClustering. Medium. https://medium.com/@shailja.nitp2013/k-modesclustering-ef6d9ef06449

K-Mode Clustering in Python. (2022, December 28). GeeksforGeeks. <https://www.geeksforgeeks.org/k-mode-clustering-in-python/>

Medium. (n.d.). Medium. Retrieved January 7, 2024, from <https://engineering.teknasyon.com/how-to-normalize-your-unsupervised-data-for-clustering-methods-9389298d20d>

Rohitkeswani. (2021, October 13). K-Prototypes Clustering Algorithm. Medium. https://medium.com/@keswani-rohitkumar/k-prototypes-clustering-algorithm-f5d8e09a0104#:~:text=K%2Dprototype%20algorithm%20is%20an

Ruberts, A. (2020, May 16). K-Prototypes - Customer Clustering with Mixed Data Types. Well Enough. https://antonsruberts.github.io/kproto-audience/

Soliya, S. (2021, July 12). Customer Segmentation using k-prototypes algorithm in Python. Analytics Vidhya. <https://medium.com/analytics-vidhya/customer-segmentation-using-k-prototypes-algorithm-in-python-aad4acbaaede>

# Appendix

Annex 1 – Data Information

A screenshot of a computer

Description automatically generated

Annex 2 – Descriptive Statistics for Metric Features

A screenshot of a computer

Description automatically generated

Annex 3 – Descriptive Statistics for Non-Metric Features

A screenshot of a number

Description automatically generated

Annex 4 – *Gender* Distribution

A graph of distribution of gender

Description automatically generated

Annex 5 – *Age* Distribution

A graph of age distribution

Description automatically generated

Annex 6 – Distribution of *NumberOfFrequencies*

A graph with numbers and a bar

Description automatically generated

Annex 7 – Distribution of *Income*

A graph of income distribution

Description automatically generated

Annex 8 – Numeric Variables Box Plots

A screenshot of a graph

Description automatically generated

Annex 9 – *Age* Vs *AllowedWeeklyVisitsBySLA*

A graph of a number of blue dots

Description automatically generated

Annex 10 – *Age* Vs *Income*

A graph of a number of dots

Description automatically generated with medium confidence

Annex 11 – Correlation Matrix

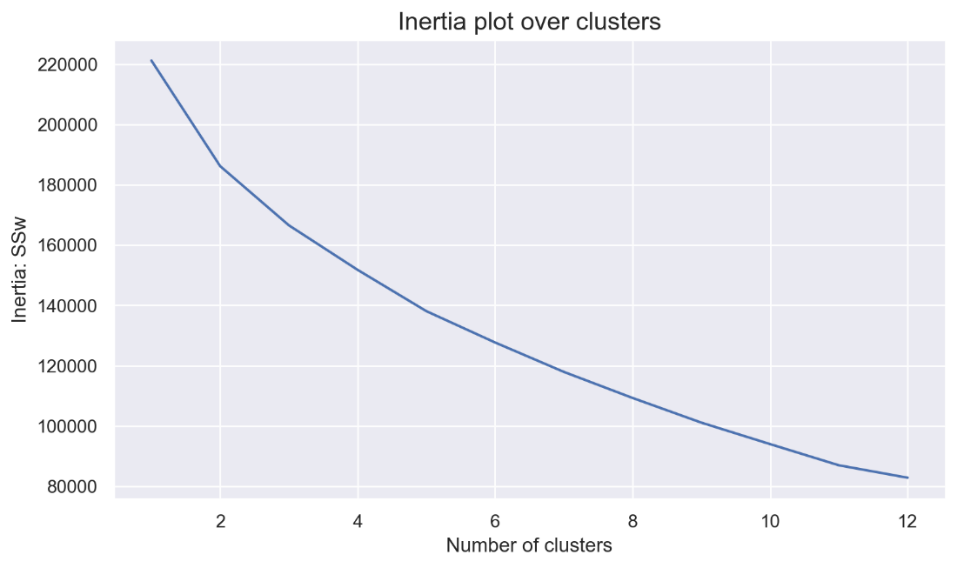
A graph of a number of people

Description automatically generated with medium confidence

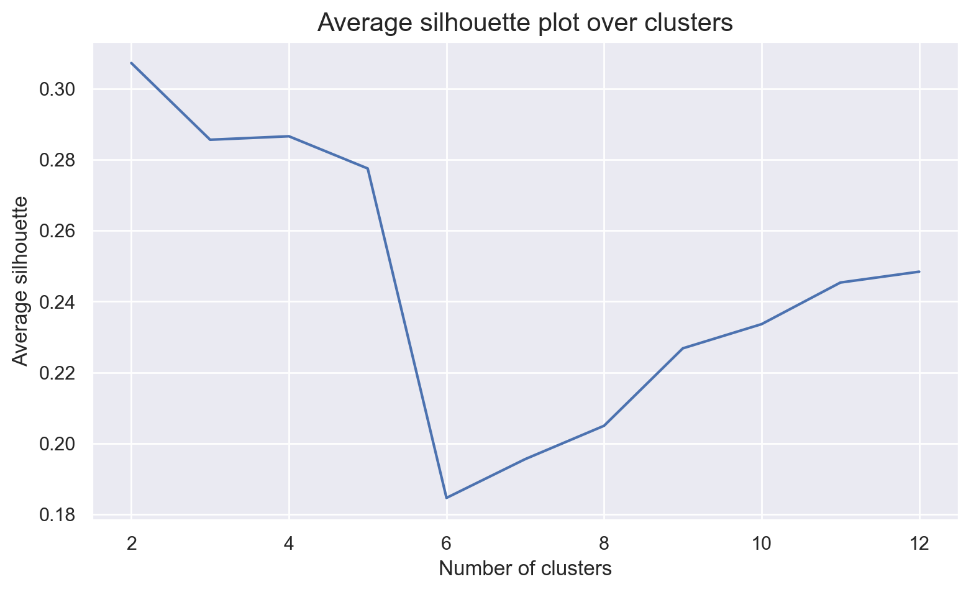
Annex 12 - Non-Metric Variables' Absolute FrequenciesA close-up of a graph

Description automatically generated

Annex 13 - Inertia plot over clusters for K means graphic



Annex 14 - Average silhouette plot over clusters for k means graphic

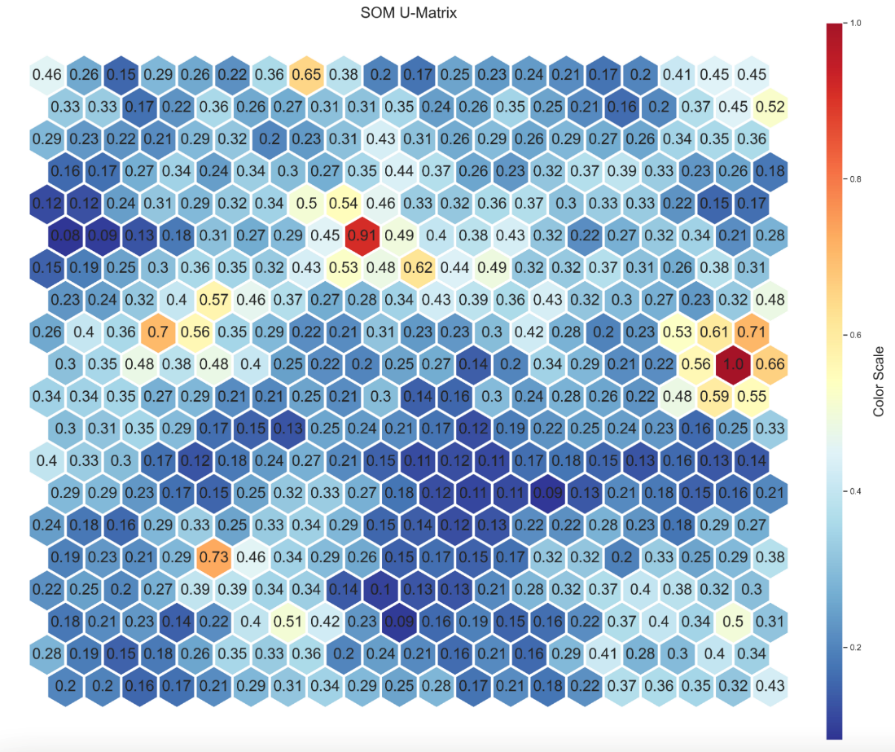


Annex 15 - Elbow Curve for K-Modes (“Huang”)

Uma imagem com texto, captura de ecrã, file, Gráfico

Descrição gerada automaticamente

Annex 16 – SOM U-Matrix



Annex 17 – Elbow Curve for K-Prototypes

Uma imagem com texto, captura de ecrã, file, Gráfico

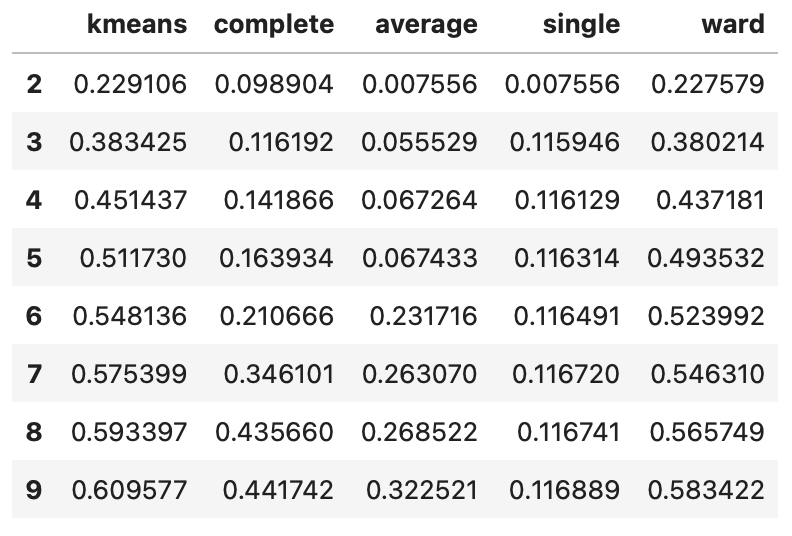
Descrição gerada automaticamente

Annex 18 – T-SNE visualization for K prototypes

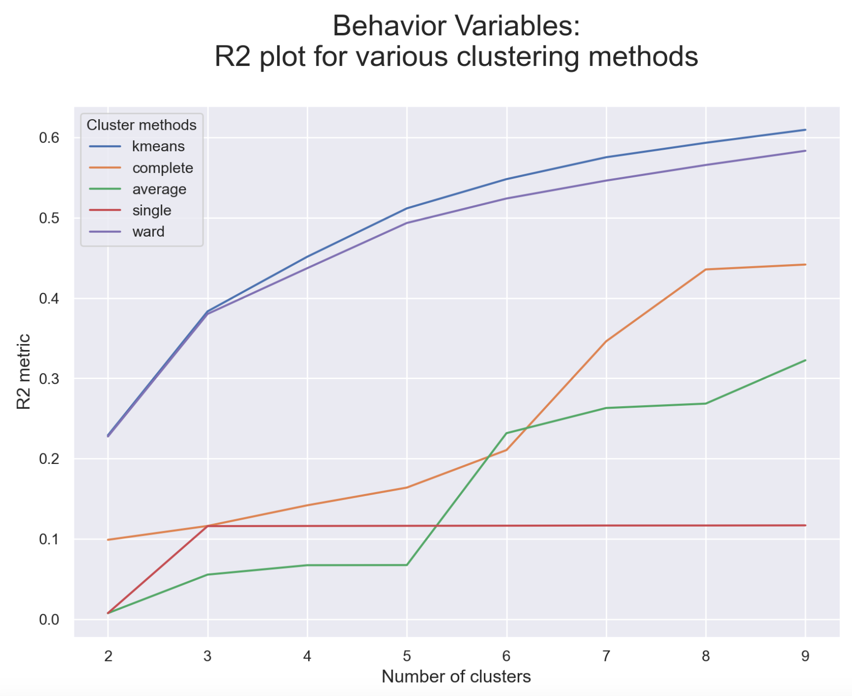
Uma imagem com texto, captura de ecrã, mapa

Descrição gerada automaticamente

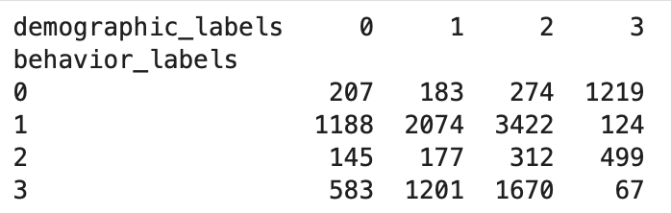
Annex 19 - R2 optimal clusterer on demographic variables



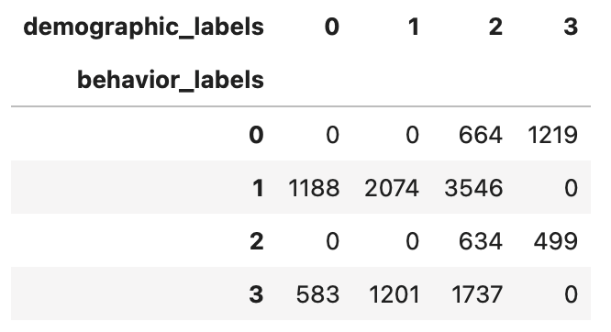
Annex 20 – R2 plot on various clustering methods on behaviour variables



Annex 21 - Contingency table 1



Annex 22 - Contingency table 2



Annex 23 – Hierarchical Clustering - Ward`s Dendrogram

A diagram of a clustering structure

Description automatically generated

Annex 24 - Cluster analysis

A close-up of a graph

Description automatically generated

Annex 25 - Profiling with unused/categorical features

A graph of different colored bars

Description automatically generated