Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

<Group 44>

<Peter Falterbaum>, number: <20230956>

<Stepan Kuznetsov>, number: <20231002>

<Ugochukwu Onyeri>, number: <20230817>

<January>, <2024>

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

INDEX

1. Introduction

In the dynamic realm of fitness and wellness, comprehending and catering to the diverse preferences of customers is paramount to success. XYZ Sports Company, a fitness establishment, acknowledges this imperative and embarks on a data-driven odyssey to refine its customer segmentation strategy. This endeavor delves into the company's meticulously curated ERP system, leveraging a wealth of customer-centric data spanning six years (2014 – 2019), to uncover distinct customer segments based on their unique attributes and behaviors.

The ultimate aim is to cultivate a tailored approach that empowers XYZ Sports Company to deliver personalized services, optimize marketing endeavors, and foster profound customer engagement. By deciphering the value and demographics of each segment, coupled with insights into their favored sporting pursuits, XYZ can effectively tailor its offerings and marketing messages to resonate with each customer cohort.

This project encompasses exploring the data, identifying germane variables for segmentation, employing clustering techniques to discern distinct customer groups, and conducting a comprehensive analysis of the characteristics of each segment. The chosen number of clusters will be rigorously justified based on both data considerations and business objectives. The gleaned insights from this analysis will inform the development of targeted marketing strategies and personalized service offerings that cater to the specific requirements and preferences of each customer segment.

The pinnacle of XYZ Sports Company's success lies in its ability to comprehend, engage, and retain its diverse customer base. By embracing a data-driven approach to customer segmentation, XYZ unlocks a treasure trove of knowledge that empowers it to deliver exceptional customer experiences and achieve sustainable growth in the competitive fitness industry.

2. Data overview

The dataset for XYZ Sports Company encompasses a wide range of customer information, crucial for segmentation analysis. Spanning from June 1, 2014, to October 31, 2019, presents a comprehensive view of customer demographics and behaviors. The dataset contains 14,942 entries and 31 variables. This rich dataset includes traditional demographic information like age and gender, complemented by detailed financial, engagement, and membership features. Moreover, it features behavioral data such as types of activities customers are enrolled in, frequency of visits to the facility, and dropout rates. This diverse array of data points allows for a nuanced understanding of customer profiles, laying the foundation for effective segmentation. By analyzing these varied aspects, we can identify distinct patterns and trends, essential for crafting targeted marketing strategies and improving customer engagement.

Financial Features: The financial features, including 'Income', 'LifetimeValue', and 'NumberOfRenewals', offer insights into the economic aspects of customer relationships. These parameters are crucial for understanding customer value and their long-term commitment to the company's services.

Engagement Features: Engagement features such as 'DaysWithoutFrequency', 'NumberOfFrequencies', 'RealNumberOfVisits', 'NumberOfReferences', and 'AttendedClasses' provide a window into customer activity levels. These metrics are instrumental in gauging the intensity and frequency of customer interaction with the company's services.

Membership Features: Membership features including 'EnrollmentDuration', 'EnrollmentStart', 'DateLastVisit', and 'AllowedWeeklyVisitsBySLA', reveal patterns in membership longevity and usage. They help in understanding customer retention and their adherence to service level agreements (SLAs).

In addition to the listed features, the dataset also includes categorical features such as 'Gender', 'Dropout', along with various activity types like 'AthleticsActivities', 'WaterActivities', etc. These categorical variables, although not used in the initial clustering, provide context to customer preferences and behaviors. The decision to exclude these categorical features from the initial clustering phase is rooted in their qualitative nature, which differs fundamentally from the quantitative nature of the other features and requires different analytical approaches.

The temporal scope of the dataset, spanning over five years, is a valuable asset for longitudinal analysis, though our current focus excludes seasonal variances. This omission is a deliberate choice to concentrate on customer comparison rather than temporal fluctuations, which could be an avenue for future research. Regarding data integrity, the dataset is largely complete, with notable exceptions in 'AllowedWeeklyVisitsBySLA' and 'Income', each missing approximately 3% of values.

Annex 1 highlights skewed distributions in some features, a critical consideration for clustering algorithms like K-Means that require normalization. Therefore, scaling techniques are employed to balance the varying scales of features like income and attended classes, ensuring equal weightage in distance calculations.

Overall, this multifaceted dataset sets the stage for a robust segmentation analysis, allowing us to unravel complex customer patterns and inform targeted strategies for XYZ Sports Company.

3. Data Preprocessing

The efficacy of a data-driven project is fundamentally linked to the caliber and preparedness of the data it utilizes. This chapter meticulously outlines the comprehensive preprocessing steps executed to uphold the dataset's integrity and enhance its suitability for sophisticated analysis. In the initial phase of data preparation, we removed the ID column, recognizing its redundancy for our analytical objectives.

Additionally, we employed one-hot encoding for the 'Gender' feature. This transformation was vital in converting categorical data into a numerical format, seamlessly integrating it into our models. Such encoding not only preserves the intrinsic information within the categorical variable but also renders it compatible with the algorithms that require numerical input.

A significant transformation was also applied to the date columns. By converting these into integer values, representing the number of days elapsed since the earliest enrollment date in the dataset, we were able to engage with time-based data more efficiently. This quantification of time intervals is not just a mere conversion; it's crucial to capture the temporal dynamics of customer engagement in a format amenable to our analytical tools.

3.1 Missing Values Imputation

While the dataset does not exhibit a high ratio of missing values, addressing these gaps is crucial for maintaining data integrity. To tackle this, we utilized a KNN Imputer with 10 neighbors, a method chosen for its effectiveness in preserving the dataset's underlying structure. This approach ensures the continuity and reliability of our data, which is essential for accurate segmentation analysis. A comprehensive overview of the initially missing values and their respective ratios is provided in Annex 2.

3.2 Data Exploration

As highlighted in the previous chapter, our dataset exhibits a range of skewed features along with varying scales. To comprehensively understand these aspects, we employed boxplots as a key analytical tool. This manual analysis, focused on identifying and evaluating outliers (for instance shown in appendix 4), revealed significant findings.

We observed a notable presence of outliers in features like “AttendedClasses”, “DaysWithoutFrequency”, “NumberOfFrequencies”, and “RealNumberOfVisits”. These outliers are not merely statistical anomalies; they potentially represent unique or extreme customer behaviors that could impact our analysis. For instance, high values in “AttendedClasses” might indicate exceptionally engaged customers, while extreme values in “DaysWithoutFrequency” could reflect periods of inactivity.

These findings are critical as they influence decisions on data treatment and the robustness of our clustering algorithms. Understanding the nature and extent of these outliers helps in determining whether to adjust, normalize, or retain these data points in their original form, ensuring that our customer segmentation model is both accurate and representative.

For a detailed visual representation of these outliers and their distribution across the dataset, refer to the boxplots provided in Annex 3. These visualizations offer a clear perspective on the spread and extremities in our data, serving as a valuable reference for our preprocessing strategy.

3.3 Feature Generation

In our approach to enhancing the analysis of customer data, we have introduced new features aimed at providing more specific insights. 'EnrollmentDuration' has been calculated to measure the length of a customer's membership in days. This feature is crucial for understanding the period of active engagement a customer has with our facilities.

Additionally, we have created the 'VisitsPerDay' metric. This is calculated by dividing the total number of visits, represented by 'NumberOfFrequencies', by the 'EnrollmentDuration'. This metric is important as it gives a clear indication of the frequency of facility usage on a daily basis, offering a quantifiable measure of customer engagement.

The integration of these new features is designed to enhance our segmentation analysis. By including 'EnrollmentDuration' and 'VisitsPerDay', we are better equipped to group customers in a manner that reflects their engagement patterns more accurately. This improved segmentation is expected to aid in developing more focused marketing strategies and in enhancing the customer experience.

3.4 Data Scaling

Data scaling was identified as a necessary step in our preprocessing workflow, given the non-normal distribution and presence of outliers in our dataset. To address this, we selected the RobustScaler for its reduced sensitivity to outliers. This choice ensures a more stable foundation for our clustering algorithms, as it minimizes the distortion caused by extreme values.

By applying the RobustScaler, we have standardized the scale of our data, a critical step for the clustering phases that ensue. This standardization is pivotal for the accuracy of our customer segmentation. It ensures that each variable contributes equally to the analysis, irrespective of its original scale or variability. This equal weighting is essential for a fair and accurate representation of each customer in the clustering process.

3.5 Feature Reduction & Correlations

In our quest to understand the intricate relationships within our dataset, we conducted a comprehensive correlation analysis between various features (see appendix 4). This involved two key approaches:

- **Overall Correlation Matrix Heatmap**: Initially, we plotted a complete correlation matrix heatmap, providing a broad overview of all inter-feature relationships. This visual representation was instrumental in identifying both strong and weak correlations across the dataset.
- **Filtered Correlation Heatmap**: To focus more sharply on the most influential relationships, we employed a filtered correlation heatmap. By setting a threshold of ±0.3, we were able to isolate and focus on correlations that exhibit significant strength. This selective approach was crucial in distinguishing the most impactful feature interactions.

Based on these insights, we made strategic decisions to refine our feature set. This included removing features such as 'LastPeriodStart', 'LastPeriodFinish', 'Age', 'EnrollmentFinish', and 'HasReferences'. The rationale behind this was to eliminate redundant data and mitigate potential multicollinearity, ensuring a more streamlined and effective analysis.

Furthermore, we simplified our feature space by excluding activities like 'DanceActivities' and 'NatureActivities'. This aggregation not only reduced complexity but also enhanced the interpretability of our model. For a detailed view of the correlation analysis and the decisions it informed, please refer to the correlation matrix provided in Appendix 3.

4. Segmentation Analysis

After presenting the preprocessing phase, we now move on to the clustering stage of our analysis. In our project with XYZ Sports Company, clustering will enable us to segment the customer base into distinct groups based on their characteristics and behaviors. We are expecting to gain insights for targeted marketing strategies and personalized customer service, improving business performance and customer satisfaction. Our preparatory work in feature selection, data cleaning, and dimensionality reduction has set a solid foundation for implementing effective clustering.

4.1 DBSCAN

DBSCAN, a density-based clustering algorithm, is well-suited for identifying groups of observations based on the concept of distance and a minimum number of points. Unlike K-Means, DBSCAN is advantageous in detecting clusters of arbitrary shapes, not limited to spherical forms. However, it faces challenges with clusters of varying densities due to its reliance on a uniform set of parameters – epsilon (eps) and minimum samples – for all clusters.

In our analysis, we employed the K-Distance Graph – which is shown in Figure 2 – to determine an optimal eps value, identifying a pronounced elbow at a distance suggesting an eps value of 3.5. We set min_samples to 15. This configuration, chosen after experimenting with various parameter combinations, yielded the most favorable silhouette value, an indicator of cluster quality. A silhouette score close to 1 suggests well-separated and distinct clusters, and our score of approximately 0.844 indicates a successful clustering outcome.

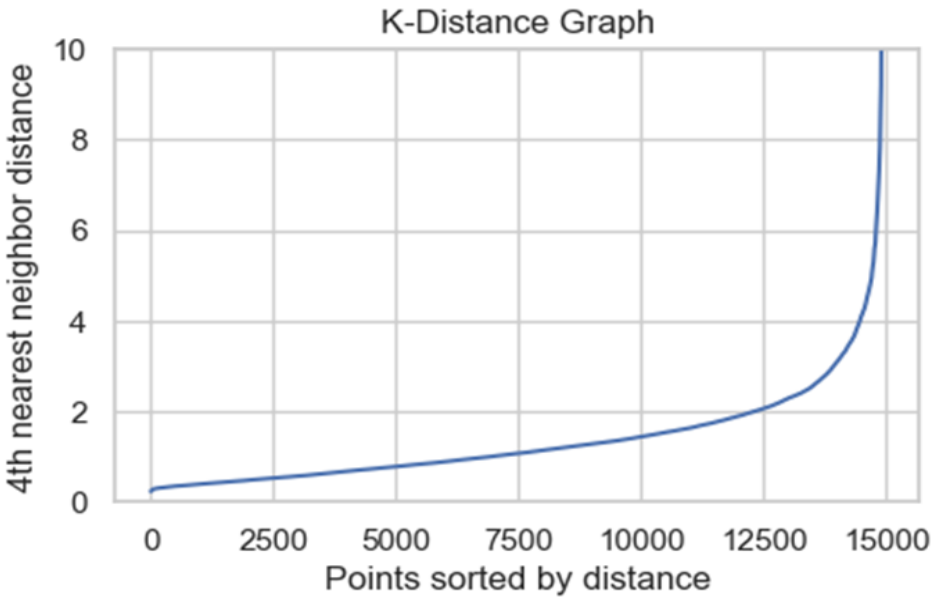


Fig. 2 K-Distance graph to determine optimal eps value for DBSCAN clustering.

The application of DBSCAN resulted in the identification of three clusters: -1, 0, and 1. Notably, cluster -1, comprising 337 entries, was designated for outlier observations. Cluster 1 was exceptionally small, containing only 14 data points out of the total 14,942. The remaining majority were classified into cluster 0. The identified outliers in cluster -1 were subsequently removed from the dataset to ensure a more robust and representative analysis in subsequent clustering approaches.

The DBSCAN process effectively segregated the data, with the majority of data points forming a coherent cluster and outliers clearly identified. This preliminary clustering step was crucial in refining the dataset, setting a solid foundation for more targeted clustering techniques to follow. The identification of a tiny cluster 1 can be analyzed in further steps but might not be sufficient for interpretation due to its size.

4.2 Clustering algorithms

The data was analyzed using different perspectives on the data and therefore applying several cluster analyses independently. The overall goal is to converge the found clusters into a final segmentation which is primarily evaluated on the business use assuming technical reliability. Therefore, three main categories of features were defined:

1) **Financial Features**: which reflect the financial side of the client's relationship with the fitness center, such as the client's income, the total amount spent at the center, and the number of membership renewals.
2) **Engagement Features**: These variables help to assess how actively clients visit the fitness center, their frequency of activities, and their overall level of engagement.
3) **Membership Features**: These characteristics are related to the duration and terms of clients' memberships at the fitness center.

KMeans clustering was applied to segment the XYZ Sports Company's customer base from three angles—financial, engagement, and membership—using subsets of data specifically selected for each perspective. Three clusters were created for financial and engagement segments, while four were chosen for the membership segment. Furthermore, hierarchical clustering was performed on the centroids of the KMeans clusters to explore the data structure at a deeper level.

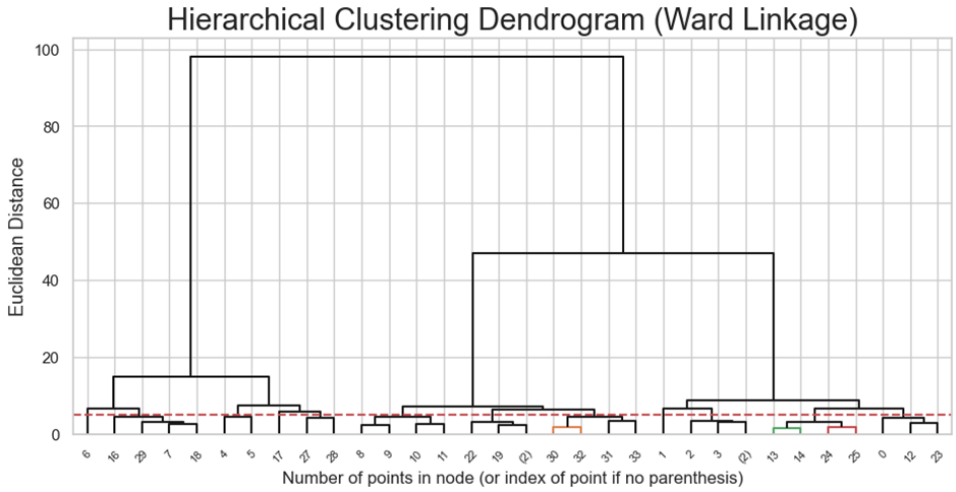


Fig. 3 Hierarchical clustering dendogram

4.3 Clusters obtained

When analyzing the mean values of the clusters and perspectives, we found out that “AttendedClasses” being scaled still has much higher value ranges than the other features. Therefore, we decided to obtain

charts but with a rescaled axis (by leaving out “Attended Classes”) to better visualize and understand the variation of feature means across different perspectives.

For the **financial** perspective (showed in figure 4) Cluster 0 has a consistent profile, with no extreme deviations, suggesting a segment with uniform financial behavior. However, there is a noticeable peak in "Number of Emergency Visits," which could indicate a higher frequency of financial distress or unexpected financial needs in this group compared to others. Cluster 1 has the most members, as shown in the bar chart on the right. Cluster 2 has the lowest mean values in nearly all the metrics, especially in "Days Without Frequency" and "Lifetime Value," which may suggest that this cluster consists of individuals with lower financial activity or engagement.

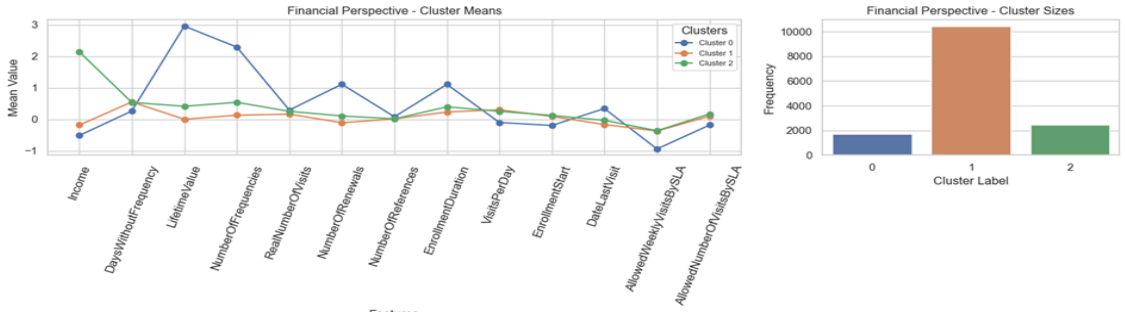


Fig. 4 Cluster Means – Financial Perspective

The **engagement** perspective (presented in figure 5) revealed clusters with varying levels of facility use, from frequent visitors to those less regular, highlighting opportunities for tailored engagement strategies. Cluster 0 has a significant peak in the "Lifetime Value" metric, indicating that individuals in this group may have a higher overall value from an engagement. This could reflect greater customer loyalty or a higher level of interaction with services/products over time. Cluster 1 shows elevated mean values in "Number of Emergency Visits" and "Enrollment Duration," which could indicate a pattern of urgent engagement and longer-term involvement with services/products. Cluster 2 is the smallest and has lower mean values in "Income," "Lifetime Value," and "AllowedWeeklyVisitsBySLA", which suggest less engagement and lower income levels.

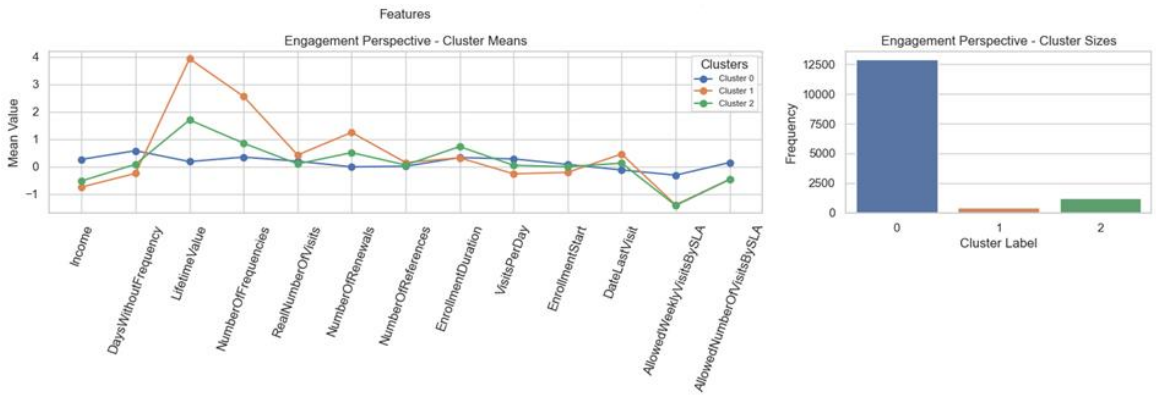


Fig. 5 Cluster Means – Engagement Perspective

Membership clustering (figure 6) showed differences in the duration and frequency of visits, indicating varying levels of commitment to the fitness facility. Cluster 0 has moderate mean values across most metrics without any significant peaks. This reflects to a group with average membership engagement and no extreme behaviors in terms of the metrics measured. Cluster Has a big peak in "Days Without Frequency," which may suggest periods of inactivity or less frequent engagement in membership-

related activities. Cluster 2 is characterized by a significant peak in "Enrollment Duration," indicating members in this cluster tend to stay enrolled for longer periods, which can be a sign of loyalty or sustained interest. This cluster also has a slightly elevated mean in "Lifetime Value," meaning that these members may be more valuable over the long term. Cluster 3 has lower mean values across all metrics, with the most pronounced dip in "Lifetime Value." This could reflect a group with the least engagement and potentially the lowest contribution to membership value.

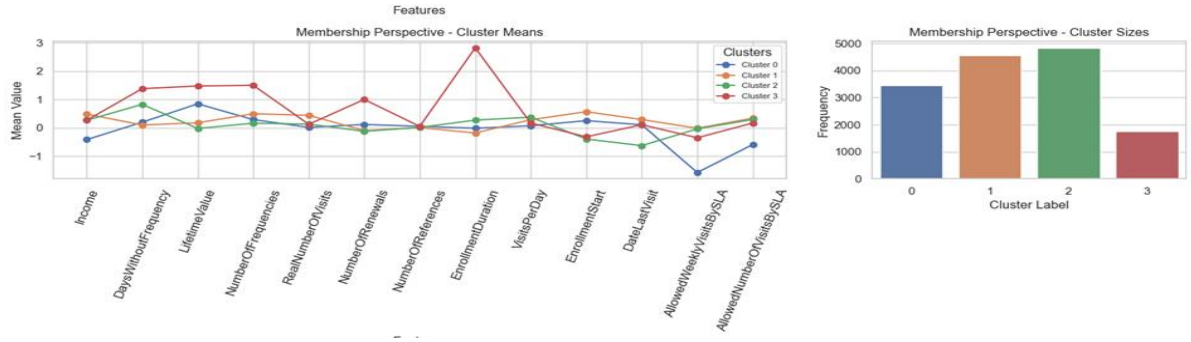


Fig. 6 Cluster Means – membership Perspective

Income Distribution:

Insights: Clusters with higher median incomes, especially cluster 3, may represent groups with higher income levels. The presence of outliers in clusters 0, 1, and 3 shows that individual values significantly exceeding the general income distribution within it. A large interquartile range in clusters 0, 1, and 3 can suggest a significant income variability.

Lifetime Value Distribution:

Insights: Cluster 3 stands out for its high median lifetime value, meaning that customers in this group tend to be more valuable on average. Clusters 0 and 1 also exhibit relatively high median lifetime values but with a significant number of outliers. Clusters 4 through 7 show lower median lifetime values, with cluster 7 having the lowest median, possibly reflecting a more homogeneous and lower customer value in this group.

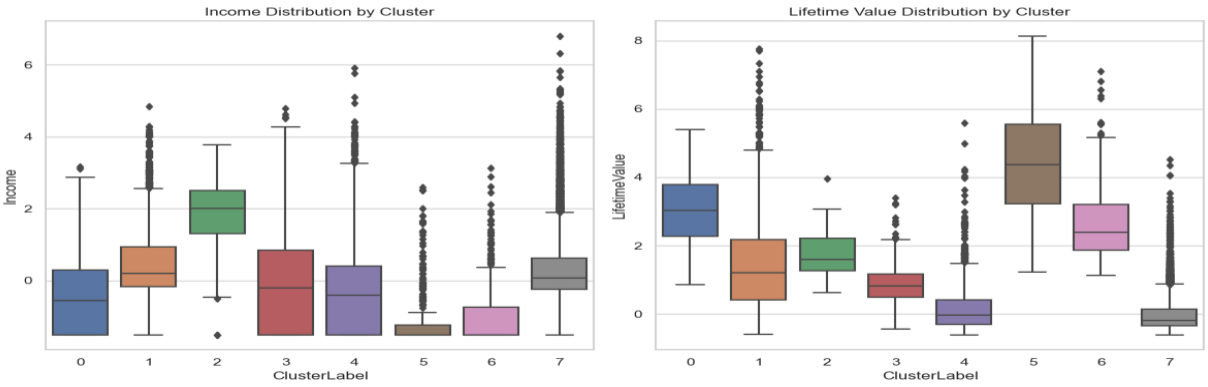


Fig. 7 Distribution of income and lifetime value across all clusters.

4.4 Analysis of Activities

4.4.1 Fitness activities totals compared for each cluster.

From annex 8, fitness activities are the most popular type of activity. Water activities, combat activities, racket activities, and special activities are all slightly less popular, with each having below 10% of participants. Special activities are the least popular type of activity of participants. It is worthy to note on cluster 7, fitness activities means are more present compared to others. More emphasis should be on this cluster. Overall, fitness activities are a popular choice for customers of this sports company. However, there is also a significant number of customers who prefer other types of activities. This suggests that the company should offer a variety of activities to appeal to a wider range of customers.

4.4.2 Fitness activities mean compared for each cluster.

As you can see from Annex 9, the mean of water activities is significantly higher than the mean of the other activities except for fitness activities in cluster 1 and 7. This means that customers are more likely to participate in water activities than other types of activities. Here are some potential reasons for this:

- Water activities are often seen as being more beneficial for health and fitness.
- Water activities can be done at a variety of levels, so they are accessible to people of all ages and fitness levels.
- Water activities can be done alone or with others, so they can be adapted to individual preferences.

The higher mean for water activities suggests that this is an area where the sports company could focus their marketing efforts. They could promote the benefits of water activities, offer discounts or incentives for participating in fitness activities, or create new fitness classes or programs.

4.4.3 TSNE visualization of the combined clusters

XYZ Sports Company can leverage t-SNE (t-distributed Stochastic Neighbor Embedding) for an insightful visualization of the customer segmentation results derived from clustering algorithms. By applying t-SNE to the multidimensional data of customer activities, the company can effectively reduce the complexity to a two-dimensional plane. This visual representation would allow for a clearer understanding of how customers are grouped based on their activity preferences.

4.5 Recommendations to the business

Targeted Promotions: Clusters with high lifetime value but low frequency of visits could be targeted with loyalty programs and promotions to increase their visitation rates.

Personalized Communication: For clusters that show high engagement but lower financial spend, personalized communication highlighting premium membership benefits could encourage upgrades.

Referral Programs: Clusters that have high numbers of renewals and references could be leveraged for referral programs, encouraging them to bring new members in exchange for benefits.

Retention Strategies: For any clusters exhibiting declining engagement or membership duration, retention strategies such as feedback surveys and re-engagement campaigns can be introduced.

Personalized workout plans: For fitness enthusiasts, suggest specific workout plans based on their fitness level, goals, and preferences.

Exclusive Membership Perks and Offers: Offer discounts, early access to events, or exclusive merchandise for specific customer segments.

Targeted Wellness and Nutrition Guidance: Provide personalized nutrition advice and meal plans for customers aiming to improve their diet.

5. Conclusion

The project provided valuable insights into exploring the data, identifying germane variables for segmentation, employing clustering techniques to discern distinct customer groups, and conducting a comprehensive analysis of the characteristics of each segment. The chosen number of clusters were rigorously justified based on both data considerations and business objectives. The insights from this analysis will inform the development of targeted marketing strategies and personalized service offerings that cater to the specific requirements and preferences of each customer segment.

The segmentation approach was methodically divided into several phases. Initial steps involved thorough data exploration and preprocessing, where we focused on scaling the data, pruning irrelevant features, excising misleading outliers, and imputing missing values. This meticulous preparation was essential to ensure the accuracy and relevance of our subsequent analyses.

With the data thus refined, we embarked on a layered analysis of customer segments. We employed three distinct perspectives to delineate domain-related clusters, later synthesizing these findings for a comprehensive view. To ascertain the most effective segmentation, we experimented with various clustering algorithms, evaluating their performance using the silhouette score. Additionally, we calculated the inertia metric to gauge the internal coherence of clusters, though its application was limited to the K-Means algorithm. This dual metric approach – leveraging both silhouette scores and inertia – validated our decision to utilize K-Means clustering across all perspectives. K-Means consistently demonstrated the most stable performance across varying cluster numbers.

In the financial perspective, we identified three distinct clusters. The largest cluster exhibited average feature values, while the other two displayed significant differentiation. One cluster represented customers with markedly higher incomes. In contrast, the third cluster, with the lowest average income, showed peaks in “LifeTimeValue”, “NumberOfFrequencies”, “NumberOfRenewals”, and “EnrollmentDuration”.

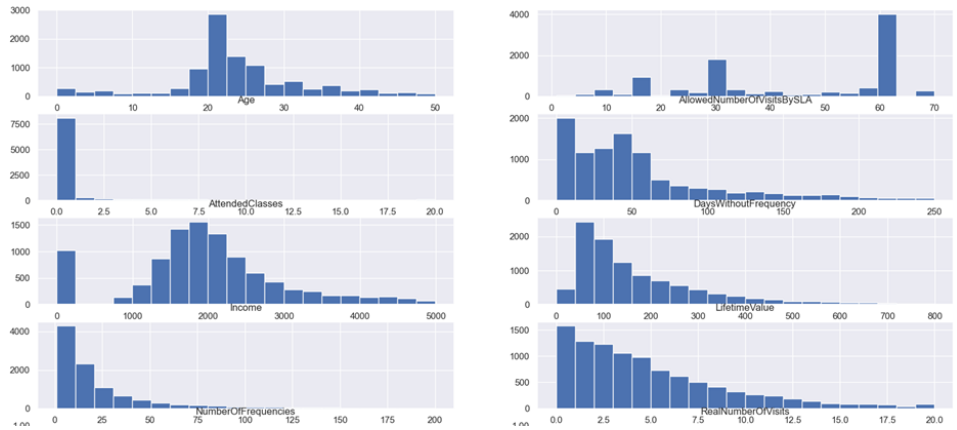
The engagement perspective revealed three separate clusters: the largest encompassing over 86% of customers, a second cluster with approximately 10%, and a third, comprising around 4%. This segmentation particularly highlighted a concentration of high “LifeTimeValue” and “NumberOfFrequencies” within the smallest cluster.

From the membership perspective, four clusters emerged as most insightful. Three of these were approximately equal in size, with the fourth being about half as large. The smallest cluster distinctly diverged from the others, notably in “EnrollmentDuration” and exhibiting the highest mean values in “DaysWithoutFrequency”, “LifeTimeValue”, “NumberOfFrequencies”, and “NumberOfRenewals”.

Regarding the data and the opportunities to gain insights, we decided to exclude the analysis of seasonal effects. Further research could combine results found in this study and gain more business relevant insights, for example seasonal marketing. We found some limitations analyzing thoroughly outliers that we detected using dbscan and the cluster perspectives. It could be highly beneficial to investigate more into these outliers to find optimization solutions.

Appendix

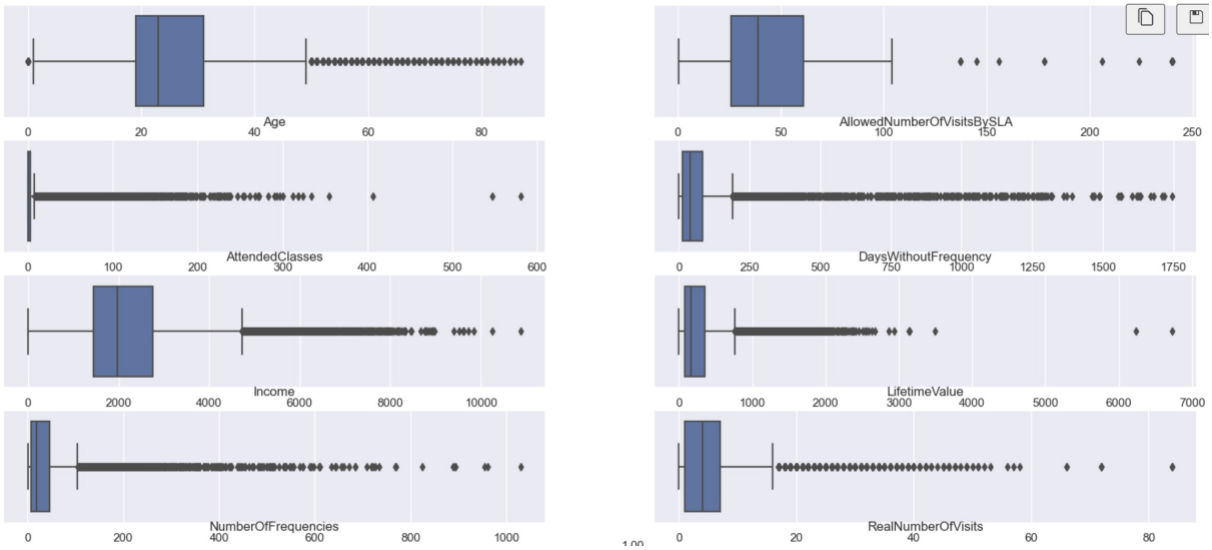
Annex 1 – Histogram of key features.



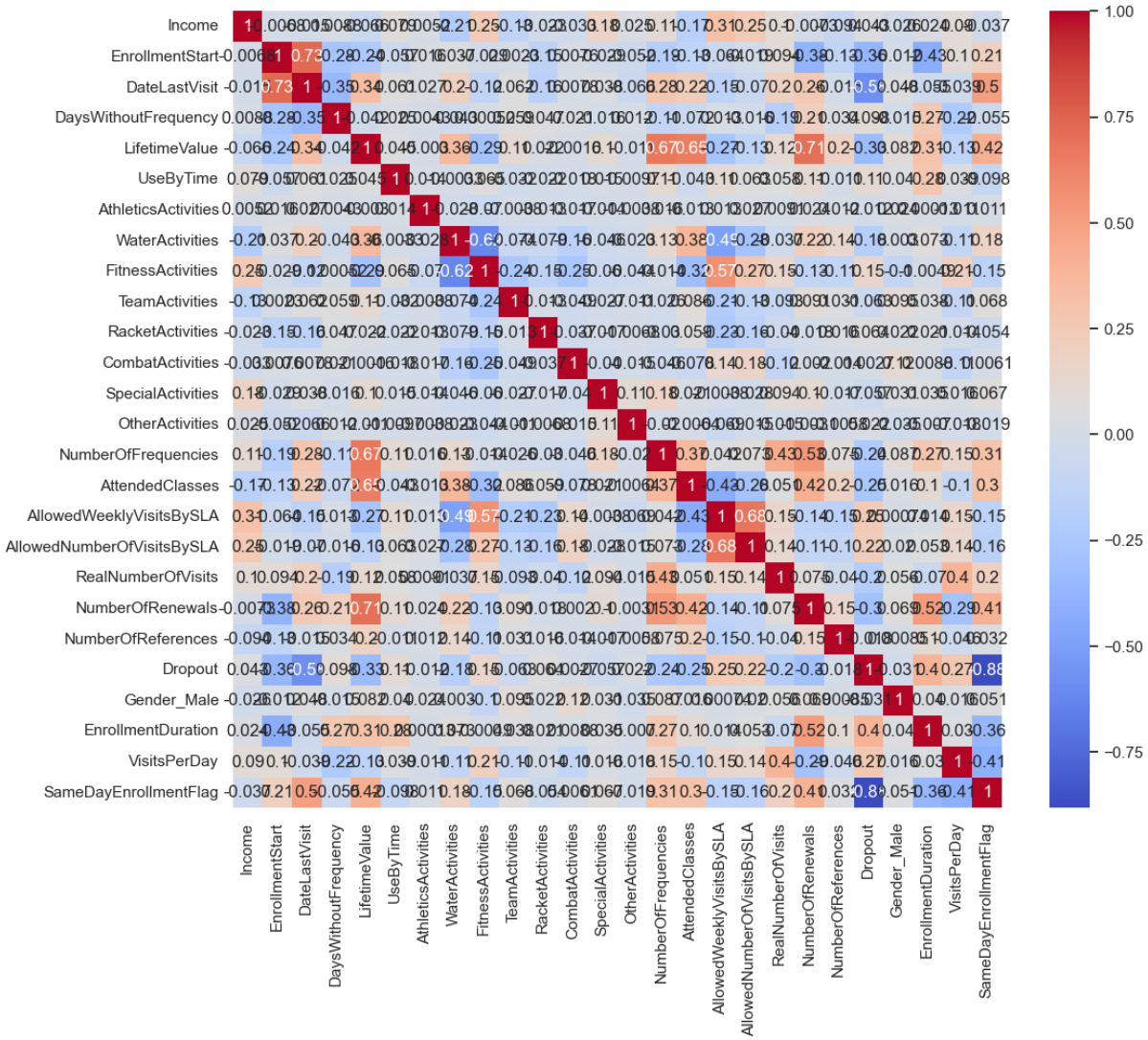
Annex 2 – Initially missing values.

	Missing Values	Percentage
AllowedWeeklyVisitsBySLA	535	3.580511
Income	495	3.312810
NatureActivities	47	0.314550
SpecialActivities	44	0.294472
WaterActivities	37	0.247624
RacketActivities	37	0.247624
AthleticsActivities	36	0.240932
DanceActivities	36	0.240932
FitnessActivities	35	0.234239
TeamActivities	35	0.234239
OtherActivities	35	0.234239
CombatActivities	33	0.220854
NumberOfFrequencies	26	0.174006
HasReferences	12	0.080311

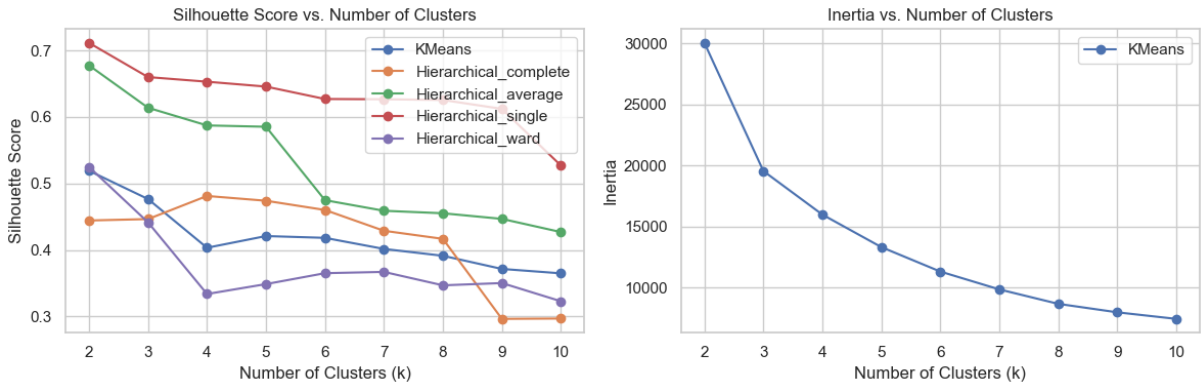
Annex 3 – Boxplots of a sample of features.



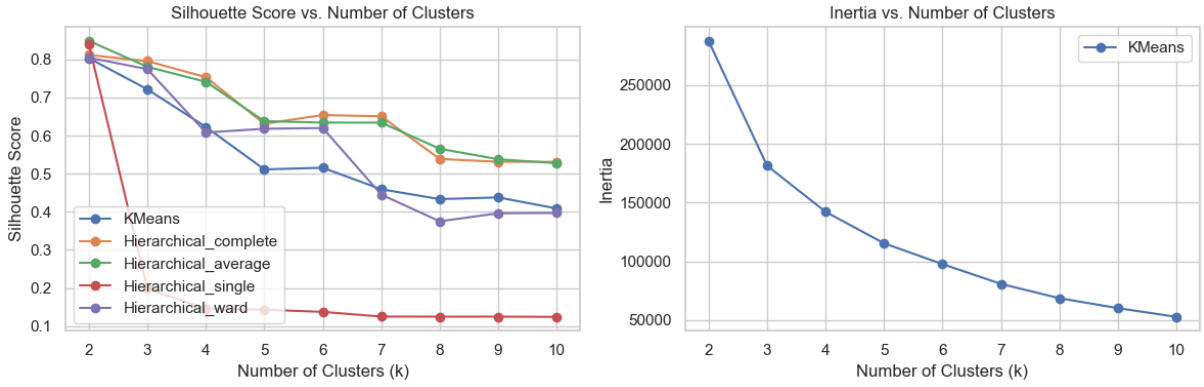
Annex 4 – Correlation matrix.



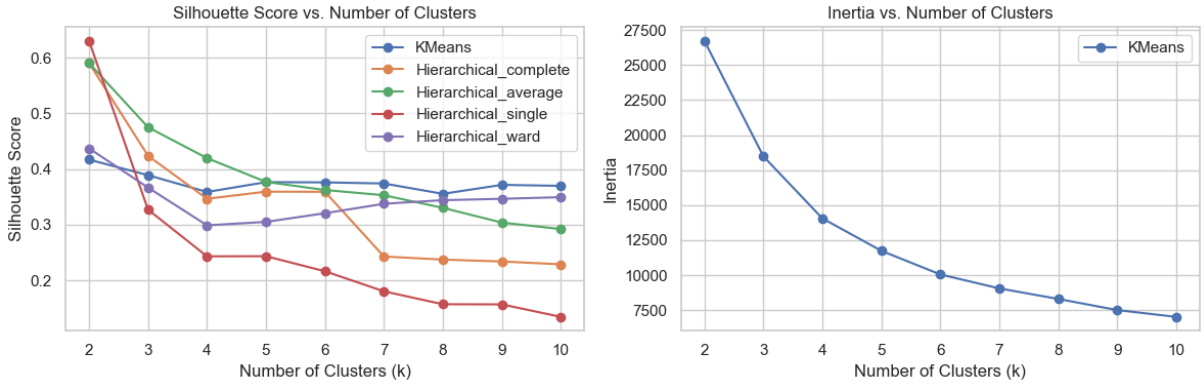
Annex 5 – **financial** perspective: comparing silhouette scores of different clustering algorithms and inertia for different k's using kmeans.



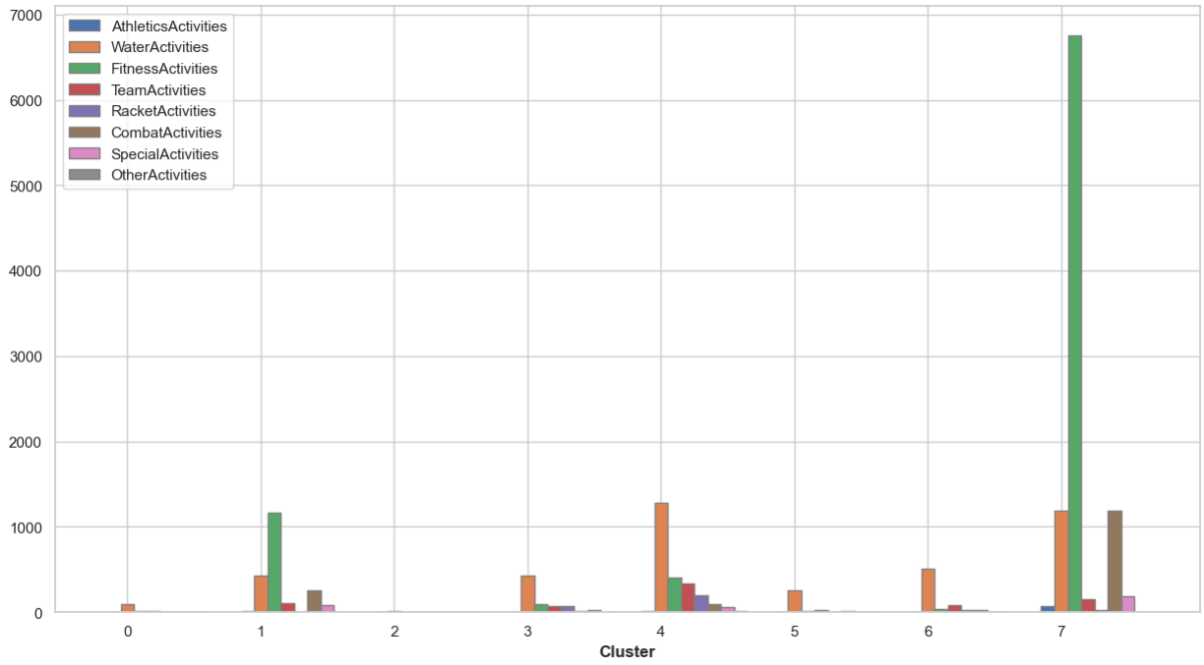
Annex 6 – **engagement** perspective: comparing silhouette scores of different clustering algorithms and inertia for different k's using kmeans..



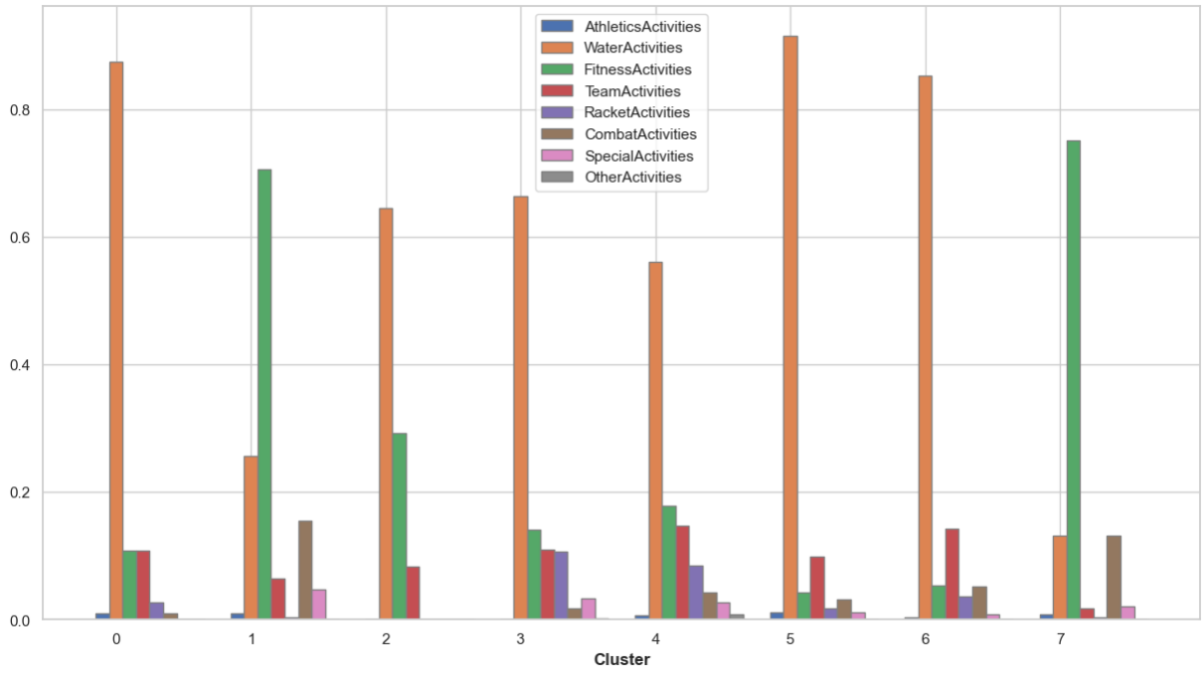
Annex 7 – **membership** perspective: comparing silhouette scores of different clustering algorithms and inertia for different k's using kmeans..



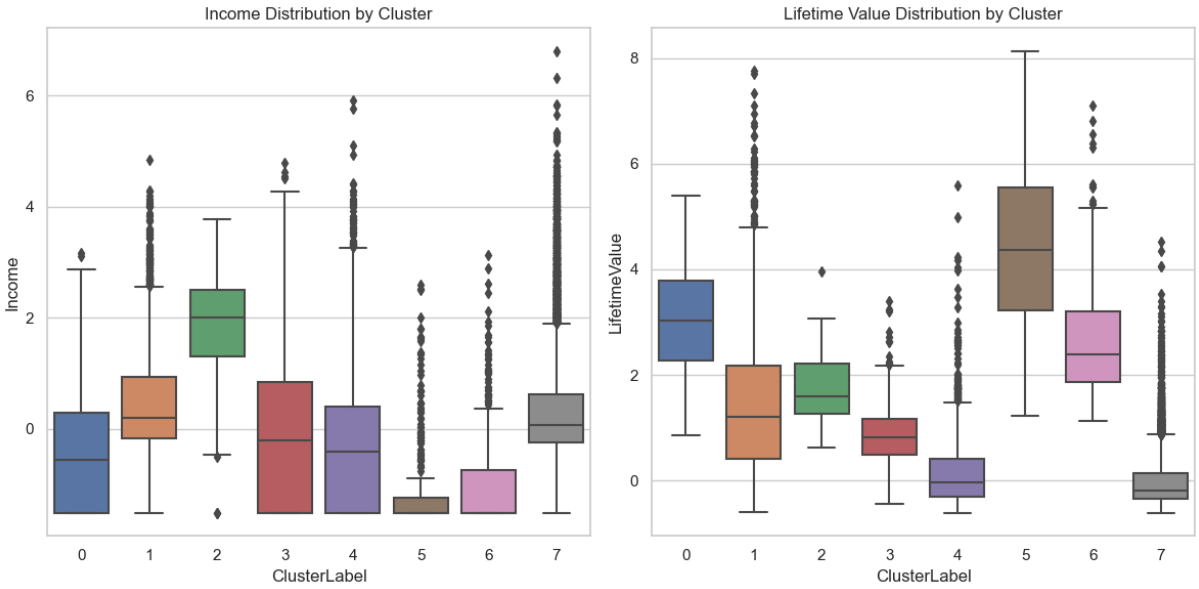
Annex 8 – fitness activities **totals** compared for each cluster.



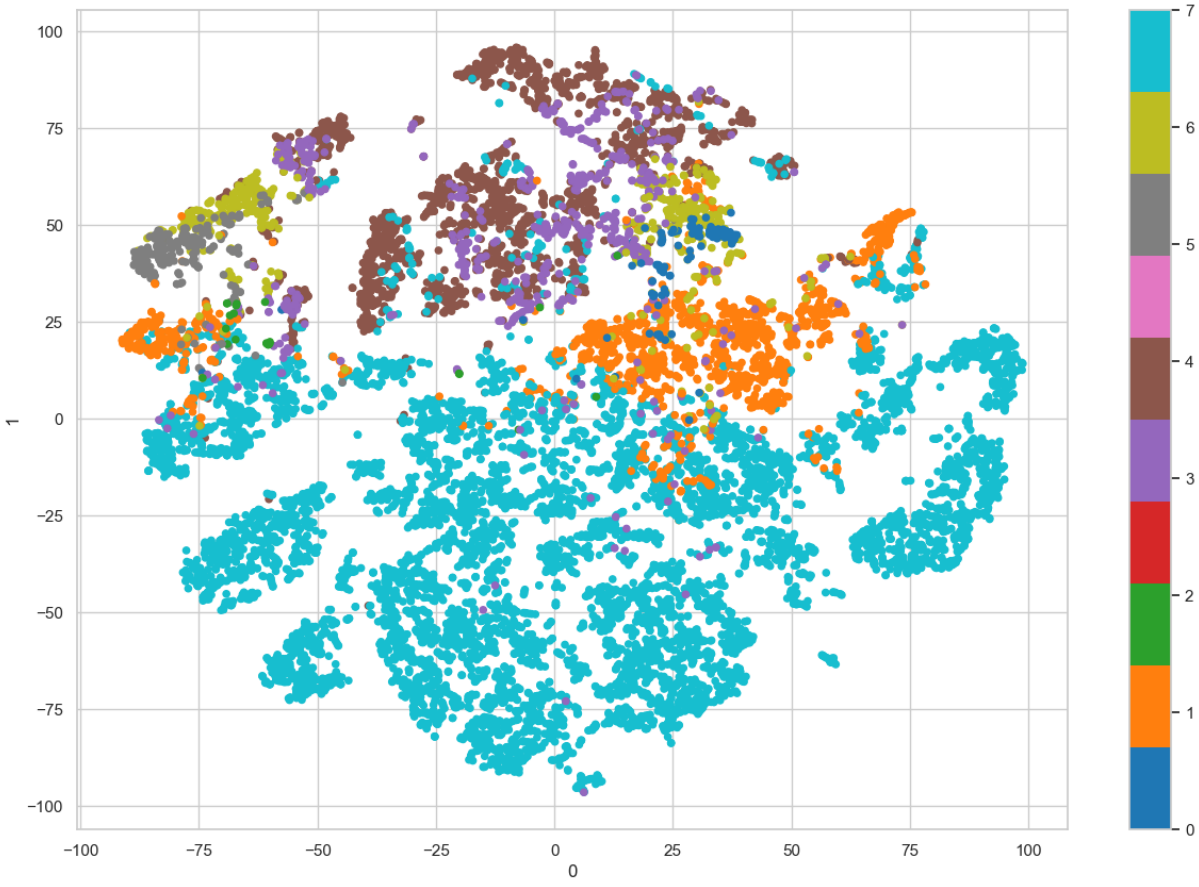
Annex 9 – fitness activities **mean** compared for each cluster.



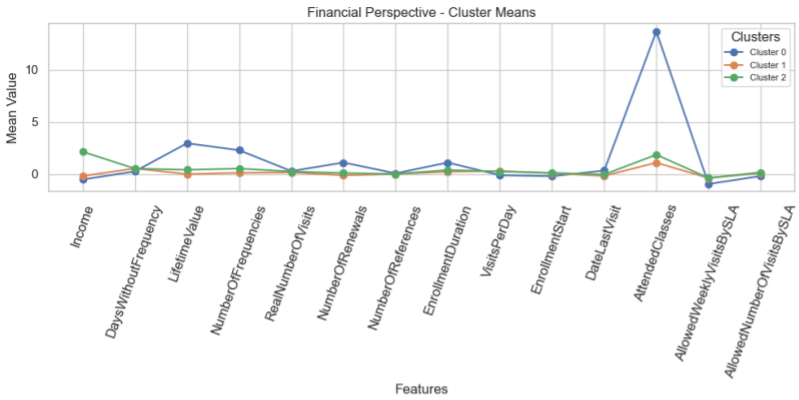
Annex 10 – Income and Lifetime value distribution compared for each cluster.

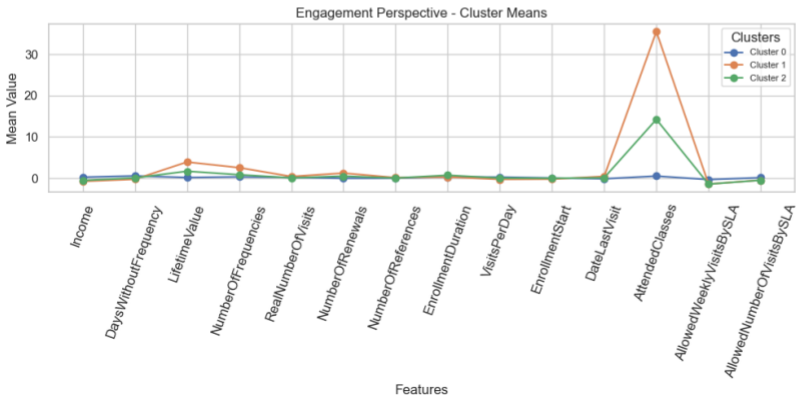


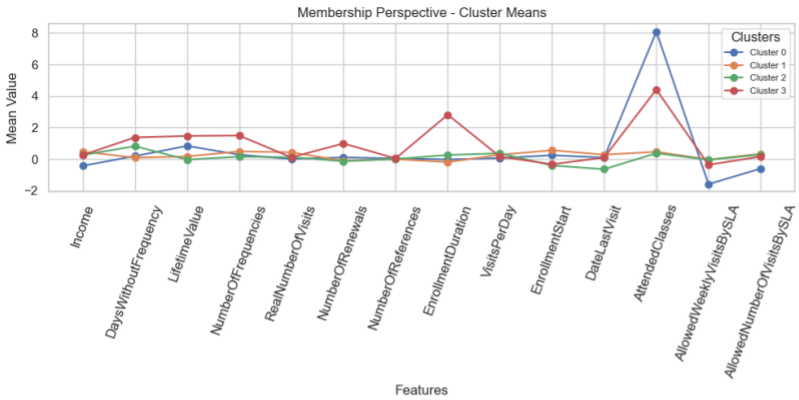
Annex 11 – TSNE visualization of the combined clusters.

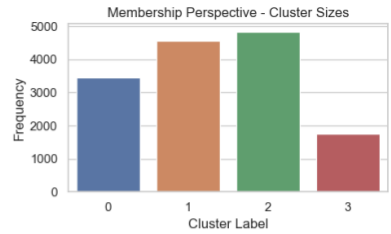


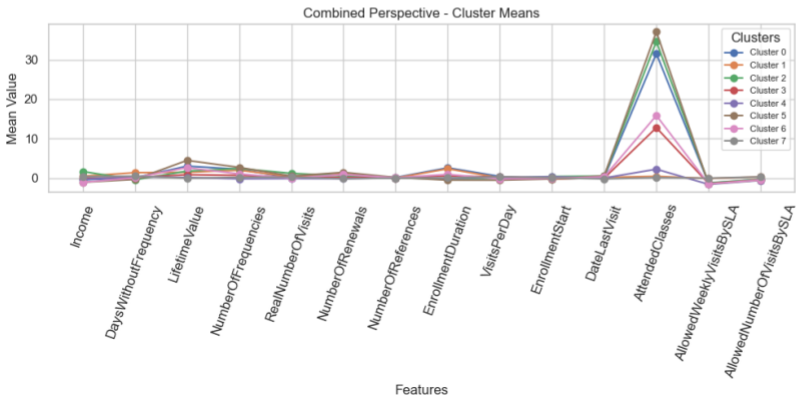
Annex 12 – means of features per cluster – with “AttendedClasses”.

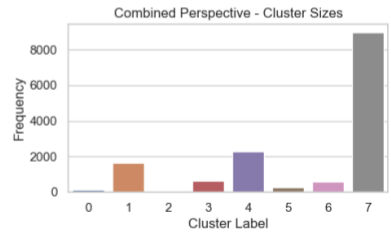












Annex 13 – means of features per cluster – **without** “AttendedClasses”.