Group Project Data Mining 2023/2024

XYZ Sports Company: Customer Analysis and Business Strategies

Masters in Data Science and Advanced Analytics

Data Mining

Afonso Gorjão (20230575), Diogo Almeida (20230737), Pedro Carvalho (20230554)

February 7, 2024

Abstract

If companies could effectively segment their customers, they would likely see significant financial gains and higher customer satisfaction. Such segmentation enables businesses to make more customer-centric decisions, potentially leading them to peak performance. To initiate this analytical process for XYZ Sports Company, we utilized its customer database, sourced from the company's ERP system, covering the period from June 1st, 2014, to October 31st, 2019. This project is designed to examine customer data and split them into distinct segments, revealing unique customer characteristics. By applying data mining techniques, including data preprocessing and clustering, our aim is to reveal the varied interests and behaviors of different customer groups and determine the most effective marketing strategies for each segment. The implementation is expected to not only enhance service offerings but also refine its marketing tactics, leading to more tailored and impactful customer interactions. This strategic endeavor positions XYZ Sports Company to strengthen its market position. In the end, we uncover two different clustering solutions. Both were made using the k-means algorithm.

Contents

I. Introduction

In an increasingly data-driven society, the need to study and predict markets, trends and consumer behavior is growing. Various studies and forecasts are carried out to give companies an edge in an increasingly competitive market.[1] Like contemporary society, the world of fitness and sports is dynamic and rapidly evolving due to changing customer preferences and the need for personalized service offerings.

XYZ Sports Company, a fitness facility, recognizes the imperative of adapting its marketing and customer engagement strategies to stay ahead in the competitive landscape. This project represents a strategic initiative by the company that has as its primary objective to develop a nuanced understanding of its diverse customer base through comprehensive data mining and segmentation techniques.

The project is anchored in the concept of market segmentation, which involves categorizing customers into distinct groups based on various characteristics such as demographic profiles, behavioral patterns, economic insights, and favorite activities. By dissecting the customer base into segments, the company aims to gain knowledge of the preferences and behaviors of different customer groups. This understanding is crucial for tailoring marketing efforts, optimizing service delivery, and enhancing overall customer satisfaction.

This report will provide an in-depth explanation of the reasoning and strategies employed throughout the development of customer segmentation, as well as an explanation of each cluster's patterns and insights for business improvement at many levels.

II. Exploratory Data Analysis and Data Preprocessing

In data mining, EDA is crucial in understanding the initial insights, uncovering the underlying structure, identifying anomalies, and extracting meaningful patterns from large volumes of data [3]. This particular dataset is complex and has many features, some difficult to fully comprehend, meaning some data explorations need to be performed. In addition, this is a dataset simulating real-world data, sometimes incomplete and inconsistent, lacking in certain behaviors or trends, and containing some errors. A strong data preprocessing could be key to getting the best out of data for further analysis.

2.1 Feature Analysis

We initiated our data analysis by categorizing the features of our dataset into distinct types. This categorization involved separating the features into numerical (Figure 3) and categorical groups (Figure 4), we utilized histograms and pie charts to gain insights into the distribution of data across them. Within the categorical features, we further converted them into boolean and date features for more practical handling during the analysis process. *Gender* was dummies-encoded to become a boolean feature, meaning all our categorical features became either boolean or date type.

As observed, *DanceActivities* and *NatureActivities* features had no values (Figure 4), and *OtherActivities* had only 0.2% of True, so we decided to drop these features from our dataset.

We also tried to perform some analysis of the pattern evolution over time. We concluded that the percentage of customers indulged in *WaterActivities* is increasing, with lots of customers only indulging in *FitnessActivities* having dropped out. This shows that the clients and their habits are changing and will be approached later in this report.

2.2 Coherence checking

Finding and dealing with inconsistencies or contradictions in the data can significantly impact the quality of the data mining process and the accuracy of its results[3].

The most staggering cases of data inconsistency come from date features. Many customers have the same *EnrollmentStart* as *EnrollmentFinish*. Those customers have *Dropout* equal to 0, meaning they are still enrolled in the gym. Other customers still registered had *EnrollmentFinish* equal to the ”today” *fixed_date*, the date of the data collection. So, we needed to establish that, for equal Start and Finish dates, *EnrollmentFinish* would be updated to the *fixed_date*.

After solving issues related to Enrollment dates, we focused on Last Period dates. The Last Period is supposed to feature, at least partially, inside the Enrollment. Cases where the Last Period is fully comprehended before *EnrollementStart* or after *EnrollementFinish* are incoherent and cannot be interpreted into a viable scenario, having to be dropped from the dataset.

Another problem with inconsistent data lies in the comparison between *NumberOfFrequencies* and *RealNumberOfVisits*. The first should always be higher or equal because it belongs to the entire period of affiliation. This is not the case in many instances, too many in percentage to just drop. We will therefore split them into two cases:

- the customers whose *LastPeriodStart* is before their enrollment might have visits counted previously by any reason (a trial, a visit just to enroll, etc.): the *RealNumberOfVisits* will be updated to *NumberOfFrequencies* because we don’t want to consider the visits before the *EnrollmentStart*;
- the customers whose *LastPeriodStart* is after their enrollment do not have any logical reason for the extra visits in the Last Period, they will be, therefore, dropped.

After removing inconsistent instances, we will now focus on the logic for the Last Period. This period is supposed to give information about the customer’s most recent time in the gym. However, before and after enrollment, not only is the customer not going to the gym, but he is not paying anything, he is no longer affiliated with the gym. With that in mind, we decided to shorten the Last Period so that it is not outside Enrollment.

After implementing these changes, we found that only a minimal portion of the dataset, less than 0.3%, was removed, meaning we considered the majority of the dataset suitable for analysis. With this preprocessing stage, we improved consistency in the date-related features.

2.3 Missing Values Imputation

We plotted the missing values for every feature (Figure 5). Having no means to infer if the missing values for Activities and *HasReferences* are related to people not having references or not indulging in any activity, we will consider them as Missing Completely at Random. Same for the *NumberOfFrequencies* feature. As for *Income*, those values seem to be Missing at Random, as they are highly correlated with underage people. The *AllowedWeeklyVisitsBySLA* missing values can be slightly correlated with small *AllowedNumberOfVisitsBySLA*, but we’ll impute them without accounting for that ourselves.

The missing values for categorical features were imputed with the mode. As for numerical features, we opted for KNNImputer, as it preserves the relationships between data and adjusts to variations in the underlying structure across different sections of the data, which is good in our extensive dataset. As KNN is sensitive to unscaled data we scaled to impute and then immediately unscaled again (for feature engineering).

III. Feature Engineering

After unscaling our data, we will try to modify and create features that are more representative and better at describing our customers to improve the outcomes of our data analysis [3].

The main thing that caught our attention was the non-normalization of the data. Any feature related to the count of events or money will be highly correlated with the others because they all depend on the time spent in the gym. Using these raw features would mean that the length of the enrollment period would be accounted for many times and the features would not cover every possible aspect and characteristic of the customer.

Within the same line of thinking, the amount of time between *LastPeriodStart* and *LastPeriodFinish* was not the same for every instance. Being the Last Period is just a fictional period that could be used to measure the habits of customers in their last days of enrollment, this means that the count of *RealNumberOfVisits* does not carry any meaning as it stands, because it is measured for different time intervals.

With the need for normalization in mind, we created many engineering features:

- *EnrollmentDuration*: number of days between *EnrollmentStart* and *EnrollmentFinish*;
- *LastPeriodDuration*: number of days between *LastPeriodStart* and *LastPeriodFinish*;
- *LastNumberOfVisits_norm*: number of visits in the Last Period normalized, dividing *RealNumberOfVisits* by *LastPeriodDuration*;
- *NumberOfVisits_norm*: number of visits in all Enrollment normalized, dividing *NumberOfFrequencies* by *EnrollmentDuration*;
- *AllowedVisitsRate_LastPeriod*: converting *LastNumberOfVisits_norm* in visits per week multiplying by seven and comparing with *AllowedWeeklyVisitsBySLA*, obtaining the exploitation rate of the service by the customer;
- *AllowedVisitsRate_Enrollment*: same logic but using *NumberOfVisits_norm* (the *AllowedWeeklyVisitsBySLA* is given for the last two months, but we are assuming that the customer had the same service hired for the whole enrollment);
- *LP_Visits_DecayRatio*: ratio between *LastNumberOfVisits_norm* and *NumberOfVisits_norm*, trying to observe a ”disappearance” pattern in the last days of enrollment;
- *ValueDuringEnroll*: value per day on enrollment, dividing *LifetimeValue* by *EnrollmentDuration*;
- *ValuePerAttendance*: value per visit to the gym during enrollment, dividing *LifetimeValue* by *NumberOfFrequencies*;
- *NumberOfRenewals_norm*: number of renewals in all Enrollment normalized, dividing *NumberOfFrequencies* by *EnrollmentDuration*;
- *AttendedClasses_norm*: number of attended classes in all Enrollment normalized, dividing *AttendedClasses* by *EnrollmentDuration*;
- *ClassChoice_Frequency*: number of attended classes by visits to the gym in all Enrollment, dividing *AttendedClasses* by *NumberOfFrequencies*, allowing to check if the customer frequently attends classes or usually goes to the gym to workout by himself;
- *HowLongSinceLastVisit*: number of days between *DateLastVisit* and ”today” (*fixed_date*), allowing us to understand how long ago the customer stopped going to the gym to try associating usual patterns and behaviors with time evolution;
- *MultipleActivities*: boolean feature to check if the customer indulged in more than one activity.

IV. Feature Selection and Outlier Treatment

Outliers significantly affect the results of data analysis. We need to identify data points that deviate so much from other observations that will disrupt both our scaling and our clustering process [3].

4.1 Manual Outlier Removal

First, we tried to identify outliers and remove them based on the interquartile range. However this removed too much data and we resorted to manually truncating the numerical features, based on the box plots, making sure to not discard too much data. It was an iterative process since we only took outliers from the features we used to cluster. In this step, we removed approximately 5% of the data, which is high but still much better than removing based on the interquartile range. Later we are going to reintegrate the outliers in the clustering solution so we saved them in a dedicated data frame.

4.2 Feature Selection

Plotting the correlation heatmap for all features (after the feature engineering), it became apparent that some of these features might be correlated with each other. Consequently, we faced the decision of determining which features to retain or drop, guided by the principles of relevancy and redundancy. (Figure 6)

We decided for this project to try two different clustering approaches. We will try customer profiling with and without segmentation of features. For that, we need to select features for both the non-segmented clustering and for the two (our best result was obtained with two segments) segments. The three sets of features and their correlation can be seen in Figure 7, Figure 8, and Figure 9.

For the non-segmented approach, the features selected are barely correlated and cover every aspect of the data, with *Age*, a demographic feature; *NumberOfVisits_norm*, that shows the frequency rate; *ValueDuringEnroll*, the monetary aspect; and *ClassChoice_Frequency*, that represents the personal preference of the customer for the type of workout. This is, therefore, a balanced subset of features.

For the segmented approach, the group tried many reasonings and different ideas. Having many features to choose from, some ideas for segments did not seem to lead to a good clustering. For instance, the Last Period behavior of a customer did not lead to interpretable and satisfactory clusters. After a long iterative process considering both metrics (inertia and R2 score, for example) and the interpretability of groups, we decided on our final segmentation. We have a first segment for more customer-behavior-related features: *Age*, *AllowedVisitRate_Enrollment*, *ClassChoice_Frequency*; this segment covers demographic, service utilization rate, and workout preference. Our second segment is based on the gym's profit, a very important aspect of any business. Its features are both value-related: *ValueDuringEnroll* and *ValuePerAttendance*.

Before proceeding, we need to scale our data. We opted for MinMax after several iterations because not only did the clustering come out better, but also because MinMax outperforms standardization in terms of interpretability. Giving a common range for all features allows for a clearer comparison and an easier reading of the scaled values [4].

4.3 DBScan

With the features selected, we are able, once again, to try and find more outliers with the DBScan algorithm. In principle, the outliers found by DBscan are more pertinent since it finds these outliers in multidimensional planes (contrasting with our 1-dimensional manual outlier removal). This step can only be made after feature selection since DBScan is a clustering algorithm and hence prone to the curse of dimensionality.

We performed outlier removal for the 3 groups of selected features (features for non-segmented solution, features for segment 1, and features for segment 2). We chose the radius of the DBScan applying the elbow method on the graph of the sorted distribution of minimum distances since we want the radius to be small enough to catch some outliers (too small and every point is an outlier) but not big enough to not catch outliers at all (or very few). The min_samples hyper-parameter was taken to be 2x the number of selected features in each case (rule of thumb). In this step, we removed 1.97% more data. The outliers were appended to the outlier dedicated data frame.

V. Clustering

After preparing our dataset, we can start our clustering phase. We will, as previously explained, perform two different clustering approaches. We then need to evaluate the metrics for both.

5.1 Inertia and Silhouette with K-means

To help with the decision on the number of clusters utilized, Inertia and Silhouette are two metrics that need to be observed. We used the K-Means clustering algorithm for these metrics, given it's a strong and simple algorithm and these metrics alone did not define our decision on the number of clusters.

For the non-segmented solution, the inertia elbow plot was not very useful, pointing to ranges between 3 to 6 clusters, as seen in Figure 10. Analyzing the Silhouette plots and coefficient values for a range of clusters of 2 to 9 and analyzing the clustering solutions for different numbers of clusters, we opted for 5. The silhouette plot can be seen in Figure 11.

As for the segmented solution, we opted to proceed with 3 clusters for each segment. Both silhouette plots (Figure 13 and Figure 15) and coefficients were pretty satisfactory. Only the inertia plot for segment 1 (Figure 12) seemed to indicate 4 clusters, but after iterating, our solution led to better cluster profiles, agreeing with silhouettes and inertia plot for segment 2 (Figure 14).

5.2 Clustering algorithms

Besides attempting the clustering with different numbers of clusters, the group tried applying several clustering algorithms. After implementing methods like K-Means and Hierarchical Clustering for every reasonable number of clusters, applying K-Means to SOM (Self-Organizing Map), and trying MeanShift Clustering, we concluded that the best ones were K-Means and Hierarchical Clustering. After applying both segmented and non-segmented solutions, the results that we thought were more representative of our dataset agreed with R2 score results; K-means had better solutions for all clusterings. K-means is a straightforward algorithm and produces easily interpretable clusters, making it easier to understand the data structure and distribution than in Hierarchical Clustering [2].

Our best R2 scores were: for the non-segmented clustering, 0.77 with K-Means and 0.75 with HC (both with 5 clusters); for the segmented clustering in *seg1*, 0.833 with K-Means and 0.826 with HC (both with 3 clusters), and in *seg2*, 0.63 with K-Means and 0.61 with HC (both with 3 clusters). Even in the crosstab and merging process, K-Means had better results so we decided to proceed with every clustering done with K-Means.

5.3 Segmentation Crosstab and Merging

In this step, we look at the Crosstab (*seg1*,*seg2*) which is the segmented solution. Our objective is to merge low-frequency clusters and decrease the total number of clusters in this solution. The merging is done by seeing which centroid the cluster we want to merge is closer to. After the closest cluster is identified (it was done with Minkowski distance) we merge the two clusters. We started with 9 clusters (3x3) and ended with 5.

5.4 Cluster Model Performance and Outlier Classification

5.4.1 Not Segmented Solution

After settling on a solution we tried to reintegrate the outliers with two methods and kept the one that gave a better silhouette score. The first was training a Decision Tree on the data giving the cluster labels as the target variable. We split the data into training data and test data and the decision tree correctly predicted 94% of the test labels. This is a very good result and shows that the clusters have distinct characteristics. Using the trained model we predicted the labels for the outliers. The silhouette score before the outlier reintroduction was 0.39 and after reintroduction 0.37. It is expected a decrease in silhouette scores since the outliers will likely be far away from their respective centroid.

The second solution was to simply see which centroid each outlier was closer to and label the outliers accordingly. The silhouette score before the outlier reintroduction was 0.39 and after reintroduction 0.31. Since the Decision Tree decreased less the silhouette score we labeled the outliers according to the Decision tree and reintroduced them in our solution.

5.4.2 Segmented Solution

For the segmented solution, we tried to reintegrate the outliers with the same two methods used for the not-segmented solution. The decision tree correctly predicted 93% of the test labels. Once again this is a very good result and shows the distinct characteristics of the clusters. For outlier classification by the decision tree the silhouette score before reintroduction was 0.38 and after reintroduction 0.33. For the outlier classification by nearest centroid the silhouette score before reintroduction was 0.38 and after reintroduction 0.26. Once again since the decision tree decreased less the silhouette score we labeled the outliers according to the Decision tree and reintroduced them in our solution.

VI. Final Solutions, Profiles and Marketing Strategy

6.1 Not Segmented Solution

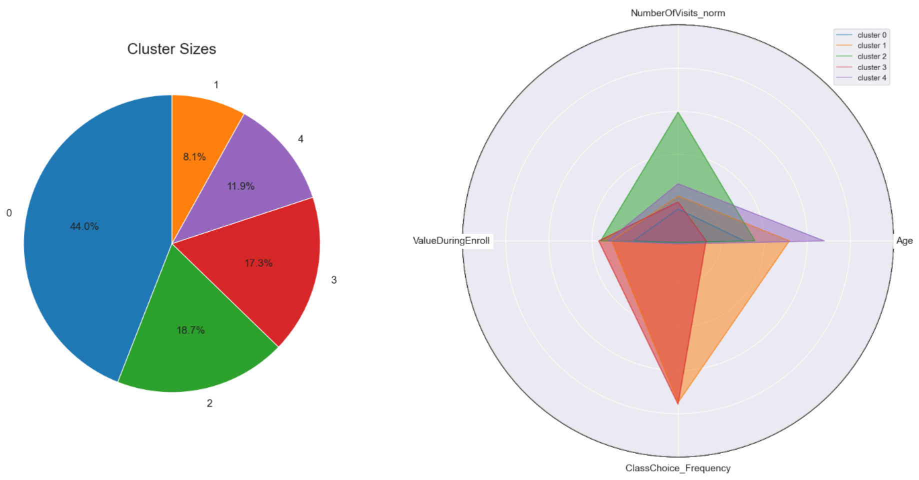


Figure 1: Cluster Profiles for non-segmented clustering

In the Figure above, we can see the data distribution between clusters and the radar pointing out the cluster's characteristics regarding the selected features for clustering. The data distribution, not ideal, is strongly conditioned by a great number of customers who don't seem to stand out in any aspect useful to the company.

We will now identify and characterize each cluster with the use of their individual profiles and our Centroids Table (Figure 21).

- Cluster 0 - Figure 16: The most general and hardest to classify cluster. Composed mainly of customers who indulged only in fitness activities and did not attend any classes. These customers, as previously mentioned tend to have a higher dropout rate, and looking at the centroid, are the oldest in terms of gym age (biggest mean for *HowLongSinceLastVisit*). Also from the centroids, this is the one with smaller *NumberOfFrequencies* (also normalized) and *LifetimeValue*, and also smaller *ValueDuringEnroll*. This means these customers have or had a cheap service, but don't take much advantage of it. The gym did not profit much from them and they were the ones that took less advantage of the service hired.

  From a marketing and business perspective, the company should try one of two approaches. XYZ can focus on improving its fitness facilities; a possible upgrade could result in fewer dropouts from fitness users, which has been seen a lot over the years, and in a higher frequency rate that can lead to the subscription of more expensive services. The alternative is to try luring these customers into attending classes, which not only leads to fewer dropouts but also to subscriptions to more expensive services. The creation of fitness-related classes could be a way of doing it.

  It has to be noted that the apparently ”cheap” services hired might be related to inflation or changes in prices over the years, as the average customer in this cluster left the gym approximately 2.5 years ago.

- Cluster 1 - Figure 17: The grown-up class attendees. These customers are adults who attend classes. They have a lower dropout rate and most of them don't indulge in fitness activities. They have a considerable *EnrollmentDuration* and *LifetimeValue*, but their *ValueDuringEnroll* and *ValuePerAttendance* could be better. They take advantage of the service hired in terms of weekly frequency (biggest *AllowedVisitsRate_Enrollment* mean), which is related to attending classes, that tend to have a predetermined number of occurrences per week (they are, on average, allowed to go 2.8 times to the gym per week).

  Regarding the marketing and business perspective, these customers tend to have a high income and can maybe pay for higher service rates. Giving them trials to new classes, and trying to make them come more times to the gym, is a possible solution. People often love gym classes for the socializing component, so by convincing a small group, others might join.

- Cluster 2 - Figure 18: The gym freaks. The great majority of customers are keen on fitness activities. They don't go to classes and many of them are not enrolled for more than one year. They have a higher *NumberOfVisits_norm* and their weekly usage rate is similar to the class attendees, despite being authorized to go for 7 days a week on average. Their *ValuePerAttendance* is the lowest, meaning they are the ones that have less profit from the gym facility.

  Similarly to Cluster 0, the approach here could be to improve fitness facilities, as it is unlikely that these customers want to start attending classes. They are determined while they are enrolled, so XYZ can only try to make them stay longer in the gym facility, although these are the less-profiting clients, so the focus should not be here.

- Cluster 3 - Figure 19: The kids. With an average *Age* of less than 10 years and a really low rate of *FitnessActivies*, there is no doubt that these are young people that go on average once every two weeks to either swim or practice Team, Racket or Combat activities. They have the lower *Dropout* rate, the longer *EnrollmentDuration*, and the bigger *LifetimeValue* and *ValuePerAttendance*. These statistics show that there is either a bigger satisfaction rate for these customers (or their families) or partnerships between schools.

From a business point of view, this cluster seems good, but we do not have access to the expenses associated with these young people, either in instructors' or cleaning staff's salaries. If there is profit, XYZ is on a good path and can only try to establish more (if there already are) school partnerships or attract families into enrolling children in the facility.

- Cluster 4 - Figure 20: The elder people. These customers' main characteristic is being older. They also have a lower service utilization rate and a higher *ValuePerAttendance*, only surpassed by those enrolled in classes.

  Although age alone might not look like a decisive factor for marketing and business strategies, the customers' average income certainly is. Although most of these customers seem to be already paying more than what they get from XYZ, their very high income when compared to other customers indicates that the company's profit can increase. Sending promotions and information about more expensive and better packages might be a good approach, as well as strong marketing campaigns for classes. These customers tend to have more money to spend, so a good marketing investment might be enough to convince them to try new packages or experiences and end up spending more money on the company.

It is clear that, although the distribution could be better, as well as some aspects of the used features could improve (like the differentiation of clusters regarding *ValueDuringEnroll*), the clusters' interpretability is rather acceptable. Each group of customers has unique and distinguishable characteristics and several approaches could be taken with this clustering solution.

The t-SNE (t-distributed stochastic neighbor embedding) of this clustering solution can be found in Figure 28. It is a way to visualize the different clusters. It uses statistical methods to visualize high-dimensional data.

6.2 Segmented Solution

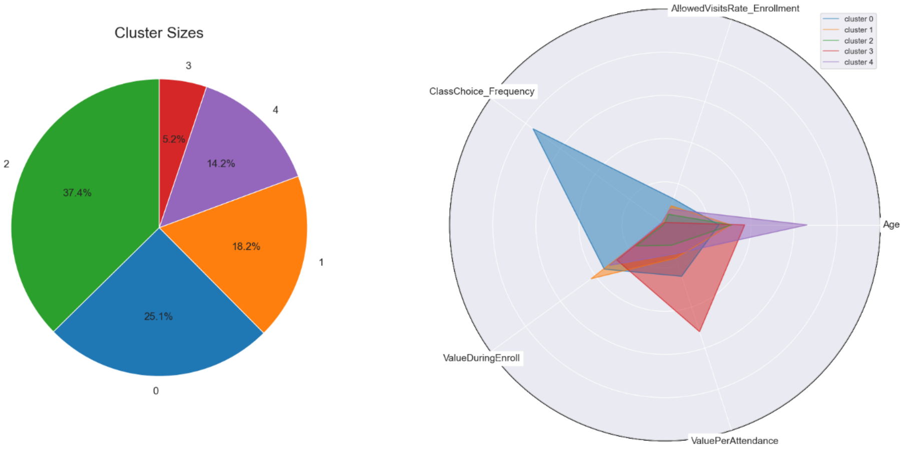


Figure 2: Cluster Profiles for segmented clustering

From the Figure above, we can conclude that the data distribution through the clusters is slightly better. The Centroids Table for this clustering solution can be seen in Figure 27.

- Cluster 0 - Figure 22: This cluster is a mix of clusters 1 and 3 from the previous clustering solution. Rather than separating class attendees into children and adults, this clustering solution groups them in the same cluster. They have the highest weekly utilization rate for their package and their *EnrollmentDuration* is on average the longest. As previously mentioned for children, if the gym is making a profit even after deducting the expenses associated, XYZ should try to lure more customers like this and try to make them stay, as they are reliable.

- Cluster 1 - Figure 23: The early dropouts. These customers have a very low *EnrollmentDuration*. Their *ValueDuringEnroll* is very high (they are correlated as it is easier to maintain a high attendance rate with a smaller enrollment period) but so is their utilization rate, having the higher *NumberOfVisits_norm*.

  XYZ should try to convince these users to stay for longer because in the short period they are enrolled (average of 6 months) they are determined, go to the gym, and have a high value. They have a high rate of *FitnessActivities* following the trend of more dropouts in this type of users and suggesting that the gym should invest more here if they want to keep these customers.

- Cluster 2 - Figure 24: The most frequent users and the lower *ValuePerAttendance*. These customers are like a mix of previous clusters 0 and 2. As previously mentioned, they are not very profitable to the gym but are assiduous, so a marketing and promoting effort should be considered, although not a priority.

- Cluster 3 - Figure 25: The profit ones. These customers have a huge *ValuePerAttendance*. They also have a short *EnrollmentDuration*, but in the time they are enrolled, they spend a lot for what they get in return.

  These are the customers who enroll but give up at the start, so an effort to make them keep trying should be made. Either trying to convince them to try classes or reminding them in any way to go to the gym, the goal is to delay their service unsubscription.

- Cluster 4 - Figure 26: Cluster very similar to previous cluster 4. Both characterization and business analysis would be similar.

Like the previous one, the clustering solution using segmentation has its flaws but is quite well-interpretable and useful for business-related analysis. The t-SNE (t-distributed stochastic neighbor embedding) of this clustering solution can be found in Figure 29.

VII. Conclusion

The data we were faced with lacked consistency. Despite its remarkably low missing value counts it left a lot to be desired in terms of timekeeping. Most features referring to dates were inconsistent and required some thought. Since most of the other features need to be normalized this is a very unfortunate characteristic of the database. However, we think that both clustering solutions presented are not only satisfactory but complement each other.

The choice between the two solutions would depend not only on technical aspects but also on the company's intentions and desired business strategy. Neither can fully grasp the characteristics of the customer base but together they paint a good general picture and give valuable information for possible directed marketing campaigns. Further exploring other clustering techniques could improve the results of this project, which contains a challenging data set that is often noisy and difficult to understand, like much real-world data.

References

[1] Kashwan, K. R. and Velu, C. 2013. Customer Segmentation Using Clustering and Data Mining Techniques. International Journal of Computer Theory and Engineering, 856-861

[2] Zhang, W., & Yang, Z. (2014). A comparison of k-means clustering and hierarchical clustering for document clustering. International Journal of Information Technology & Decision Making, 13(02), 289-304

[3] Han, J., Kamber, M. 2006, Data Mining – Concepts and Techniques, Morgan Kaufmann, Elsevier Inc

[4] Zou, C., Zhang, J., Li, Q., & Hu, J. (2018). A comparative study of feature normalization methods for interpretable machine learning. Pattern Recognition Letters, 111, 133-141.

Annex

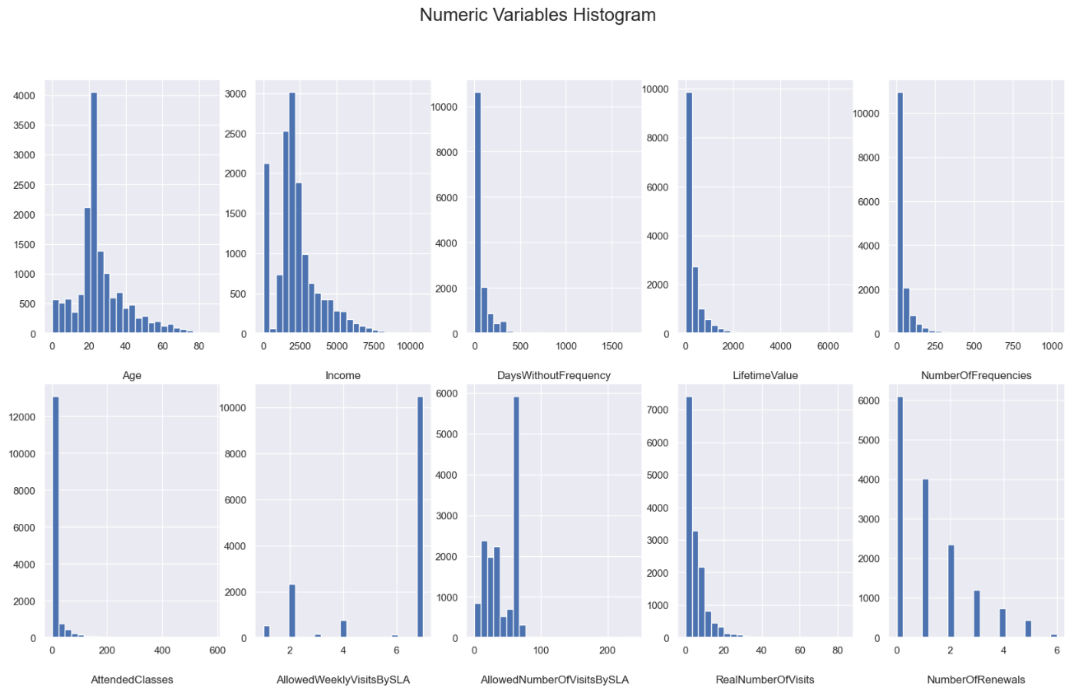


Figure 3: Histograms of the numerical features

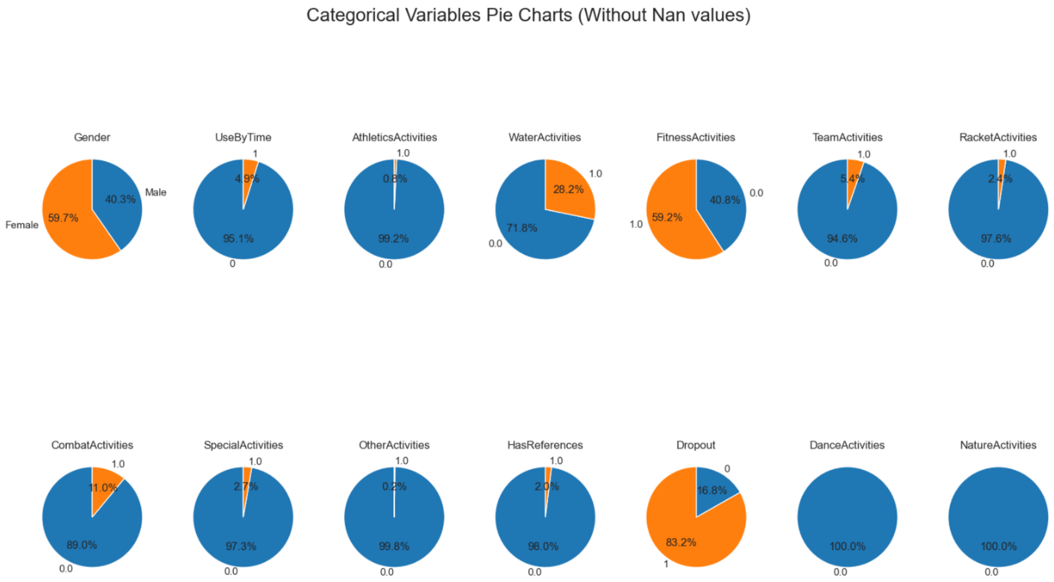


Figure 4: Pie Charts of the categorical features

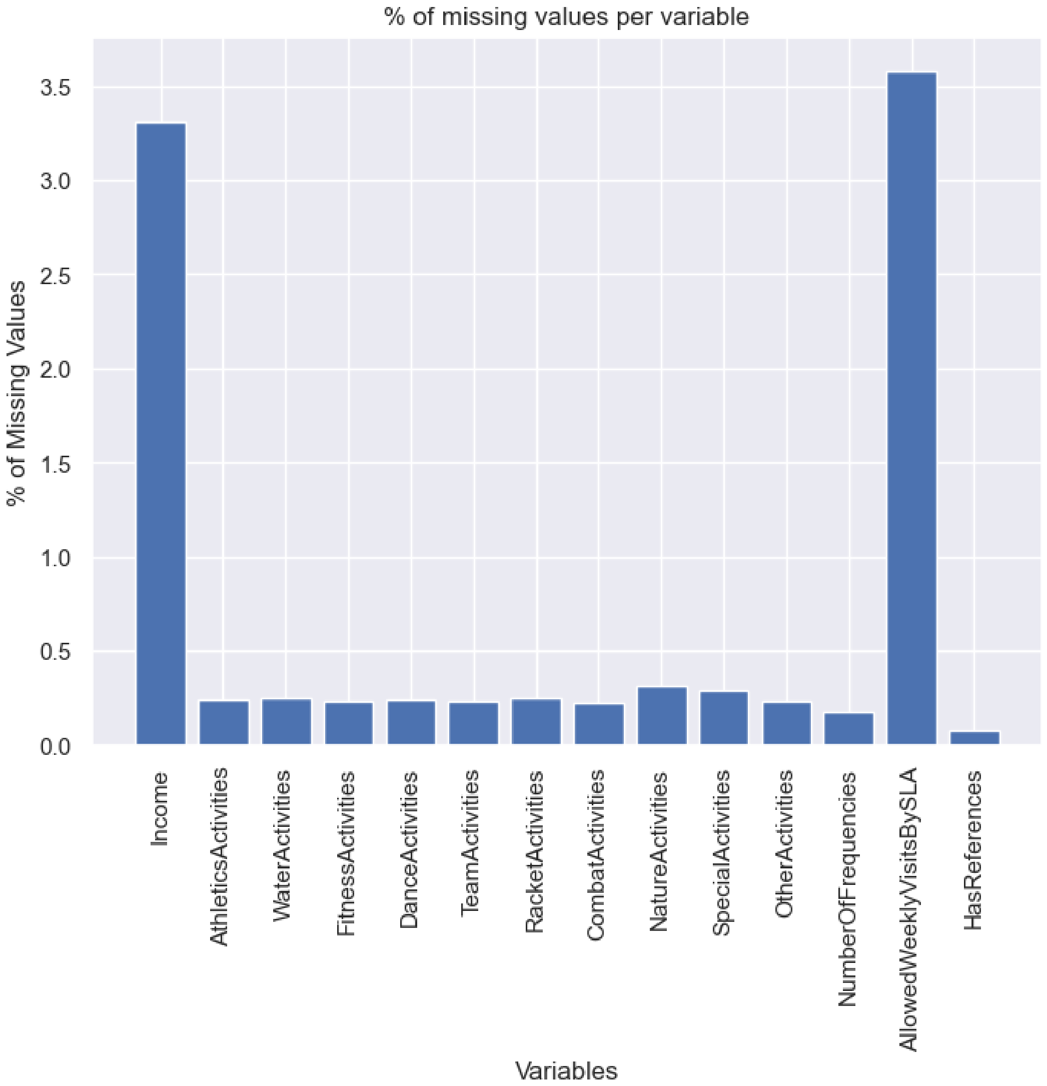


Figure 5: Percentage of missing values in the features

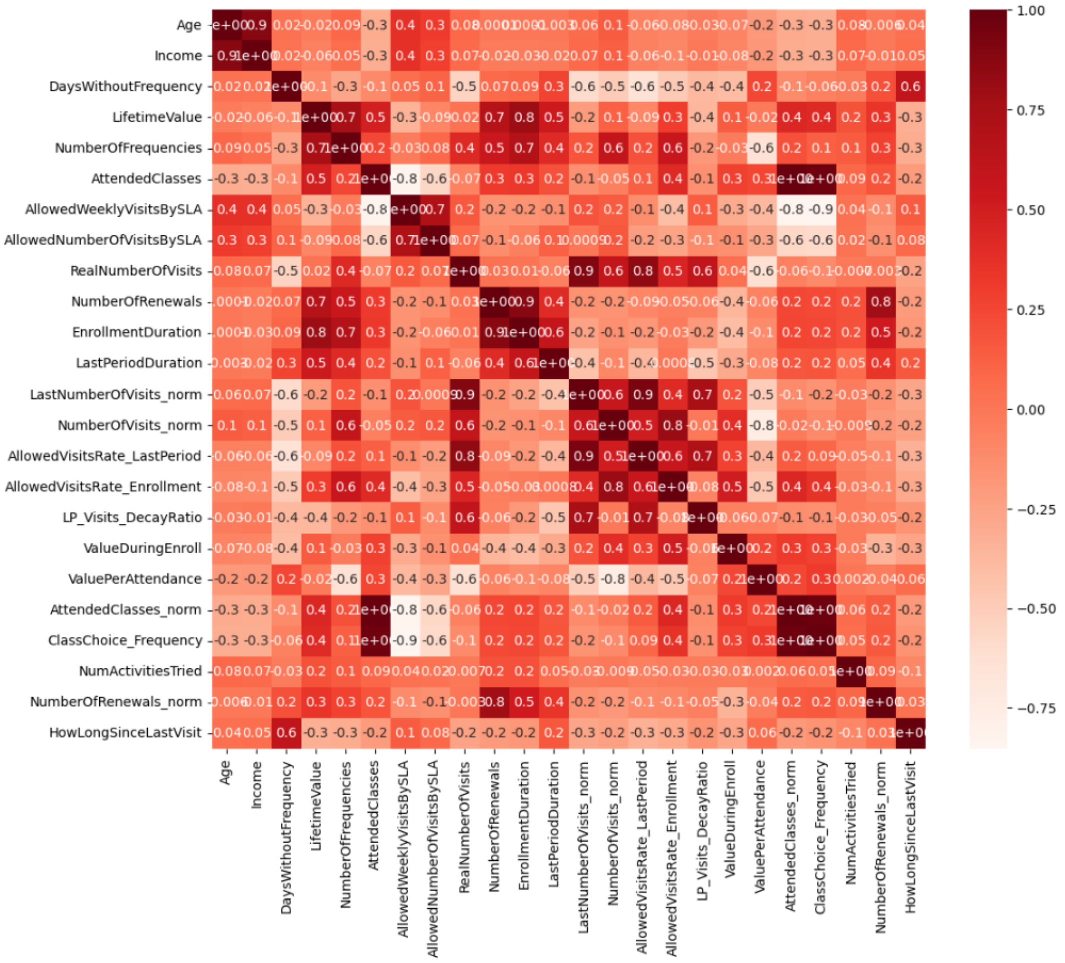


Figure 6: Correlation between all of the features (including engineered features)

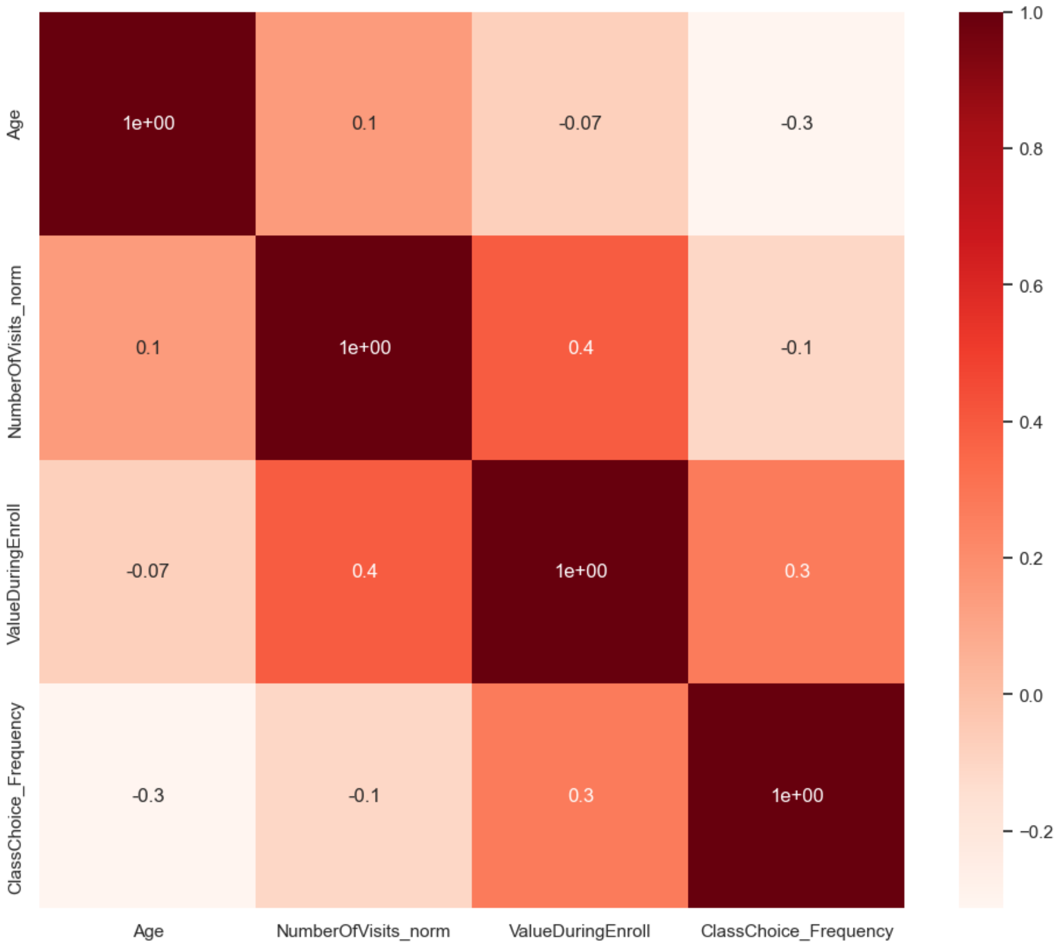


Figure 7: Correlation between the features selected for the not segmented solution

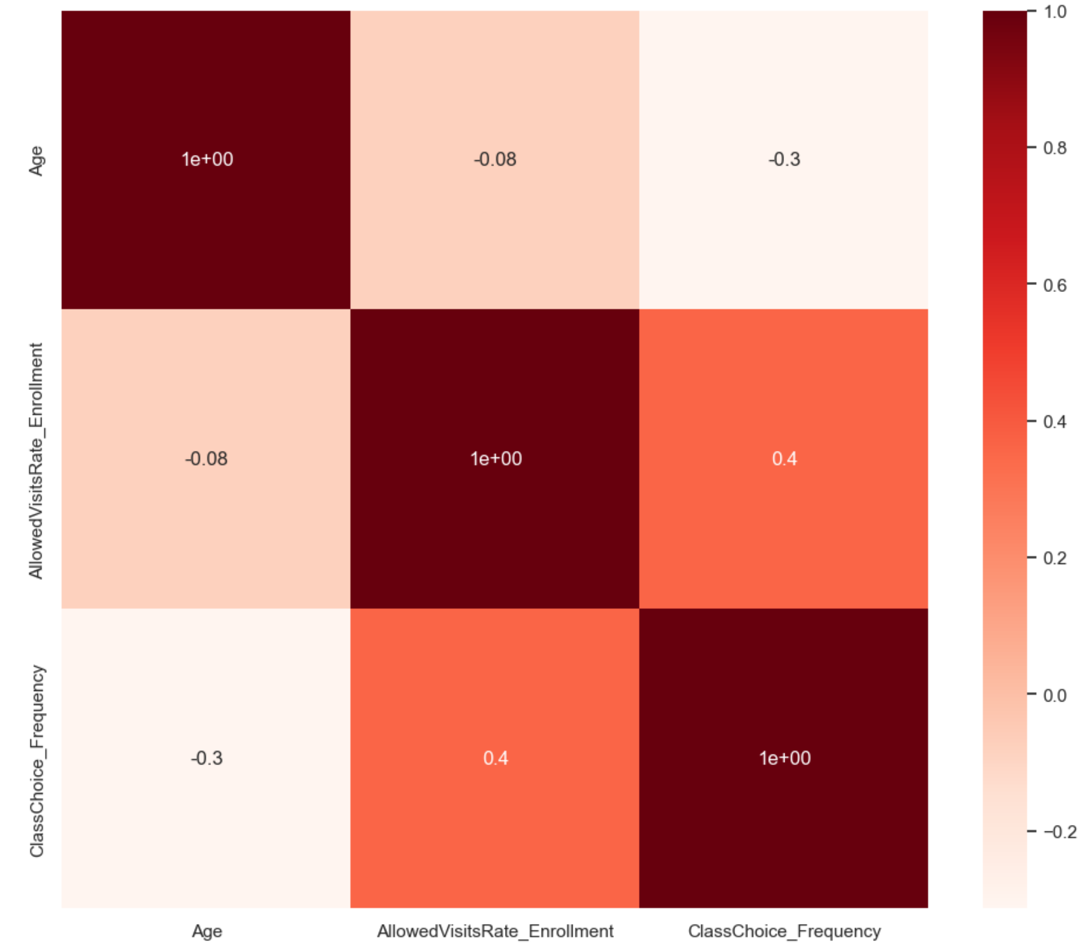


Figure 8: Correlation between the features selected for *seg1*

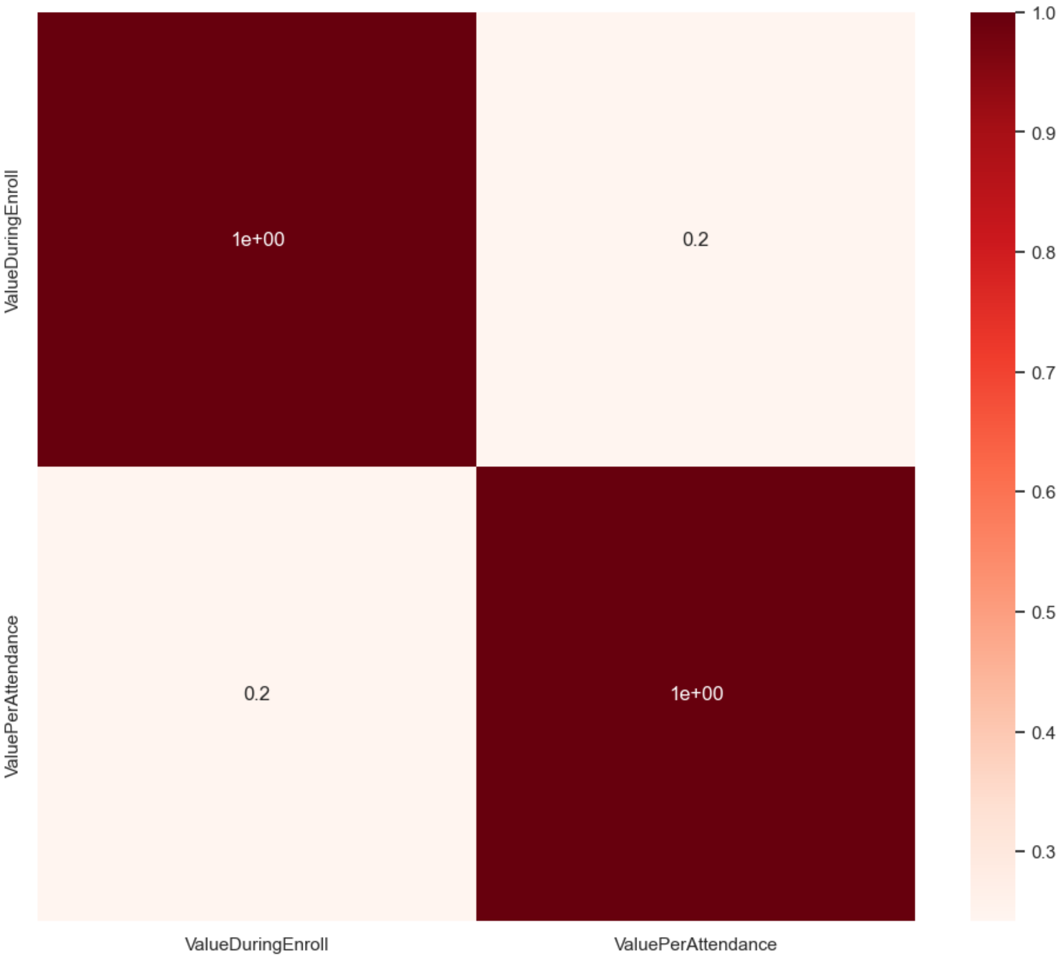


Figure 9: Correlation between the features selected for *seg2*

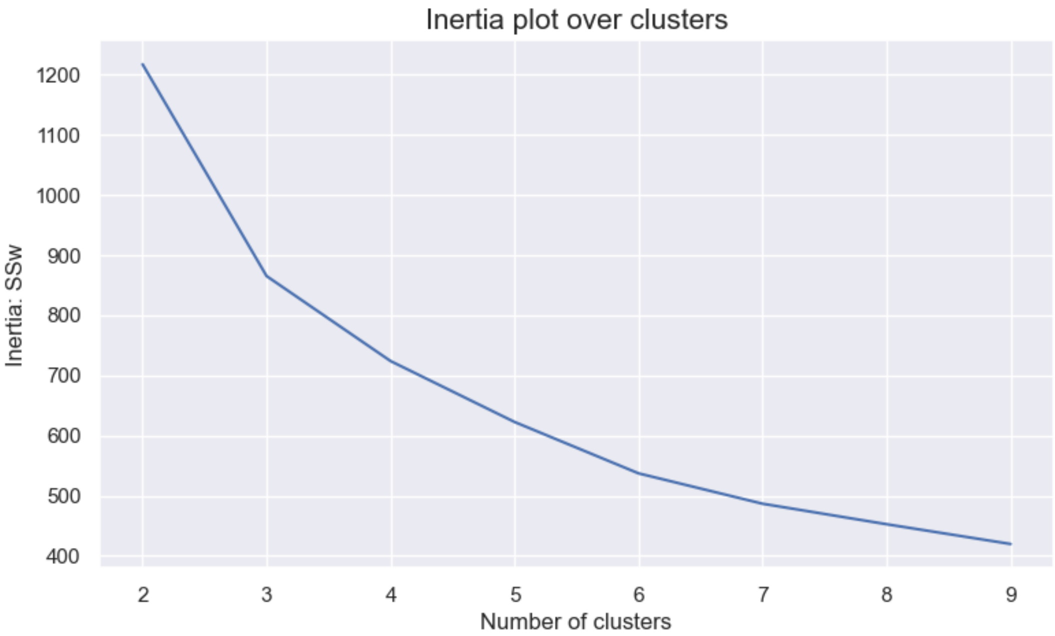


Figure 10: Inertia plot for the not segmented solution

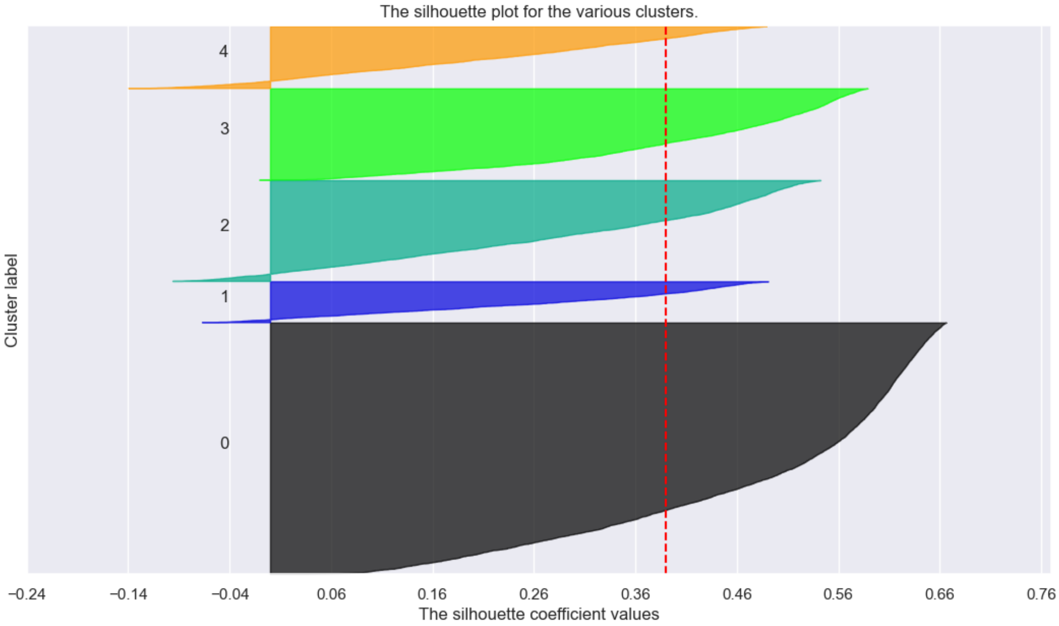


Figure 11: Silhouette for the not segmented solution. The red line is the average silhouette

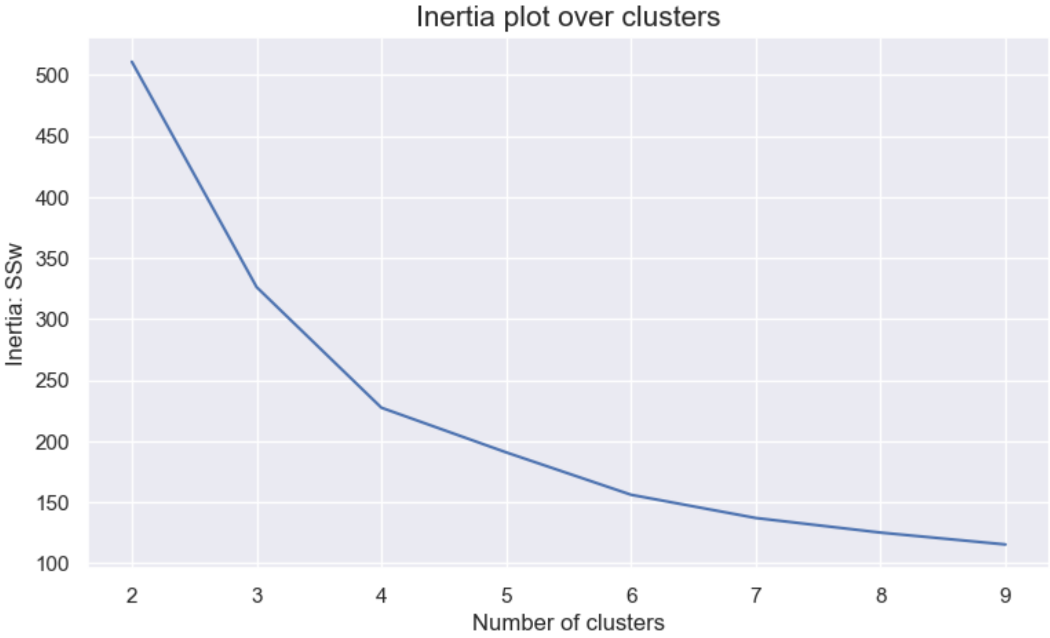


Figure 12: Inertia plot for *seg1*

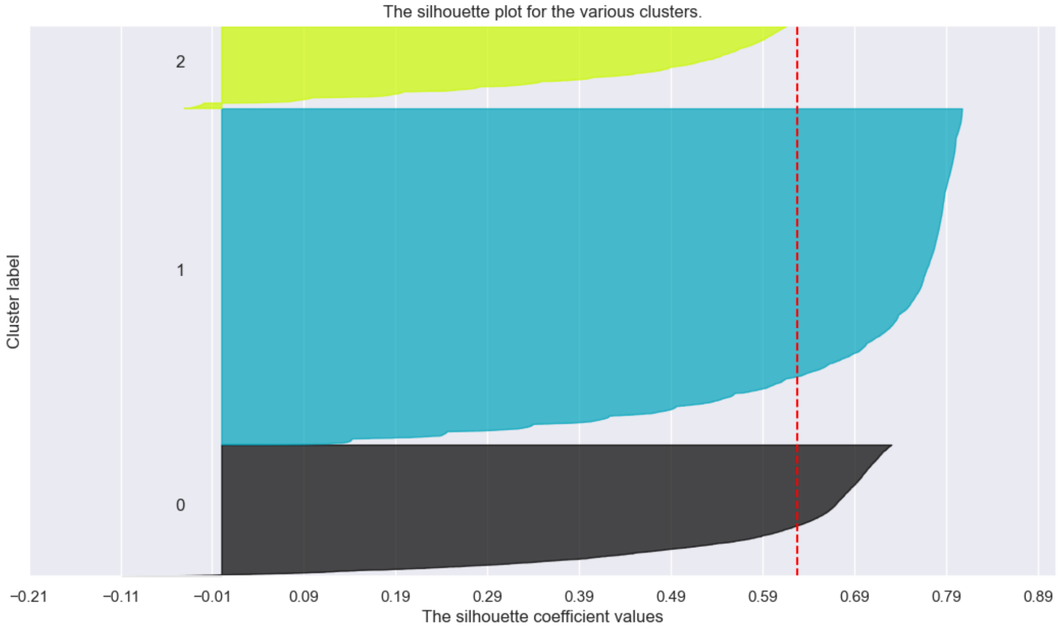


Figure 13: Silhouette for *seg1*. The red line is the average silhouette

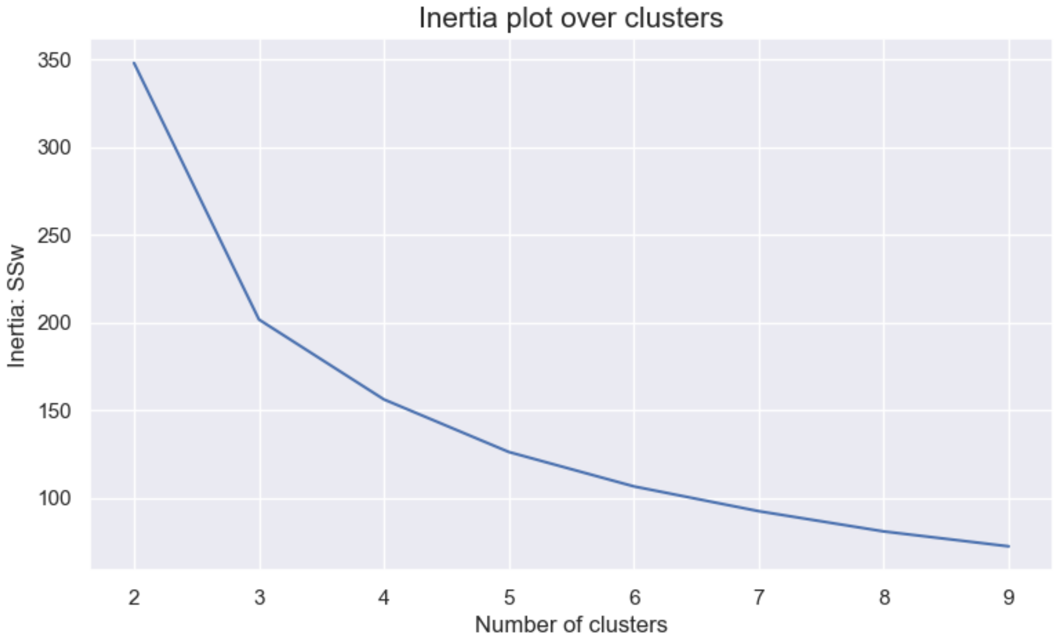


Figure 14: Inertia plot for *seg2*

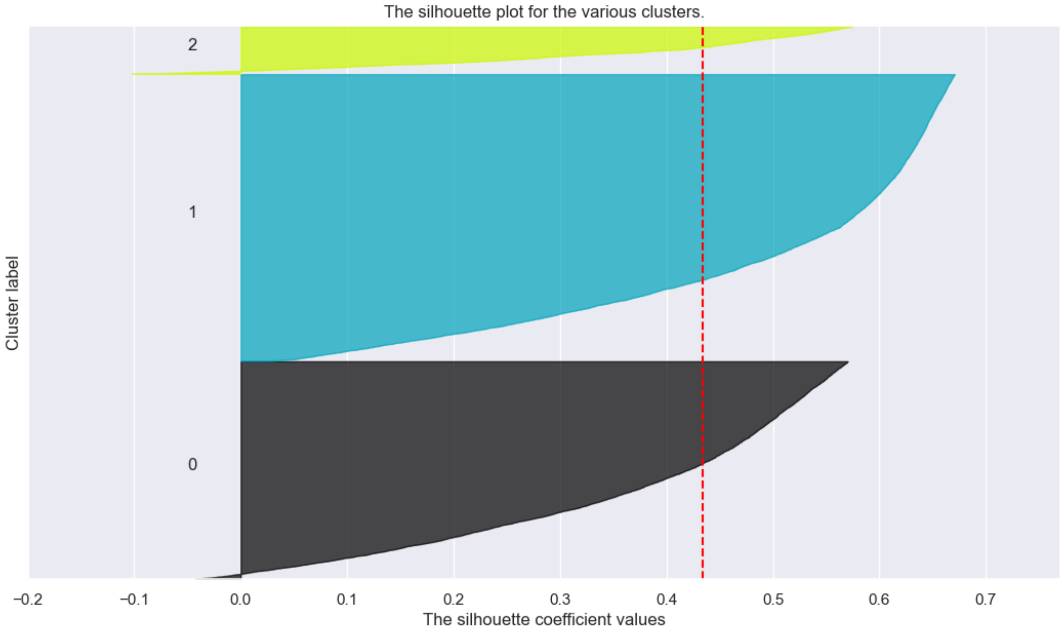


Figure 15: Silhouette for *seg2*. The red line is the average silhouette

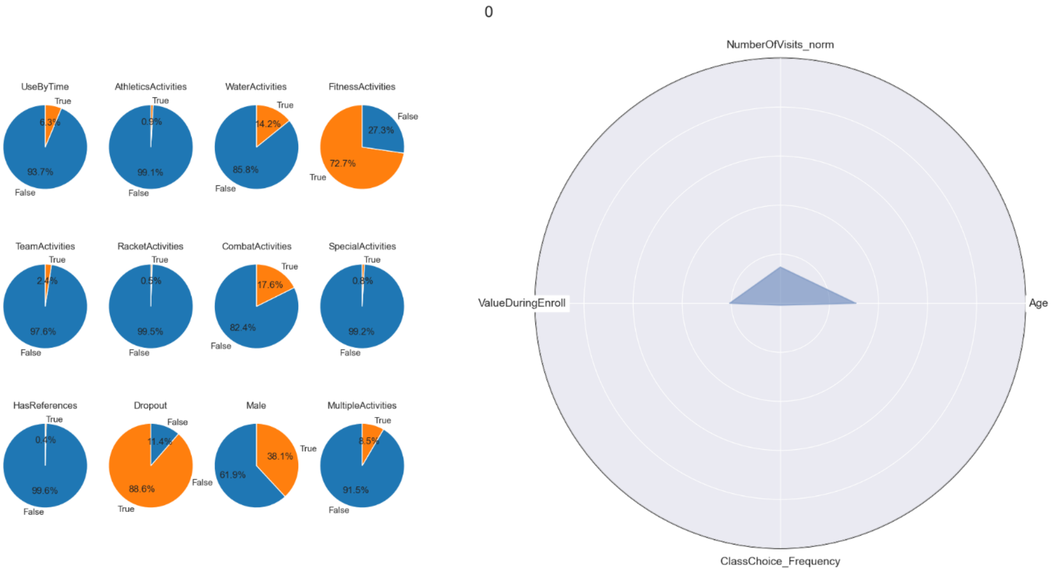


Figure 16: Not segmented solution - cluster 0 profile

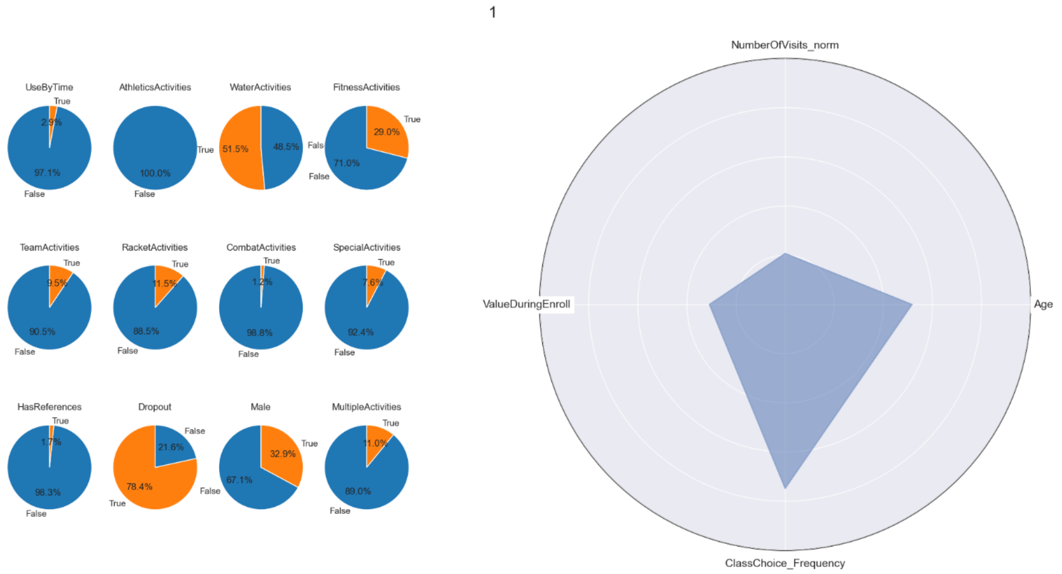


Figure 17: Not segmented solution - cluster 1 profile

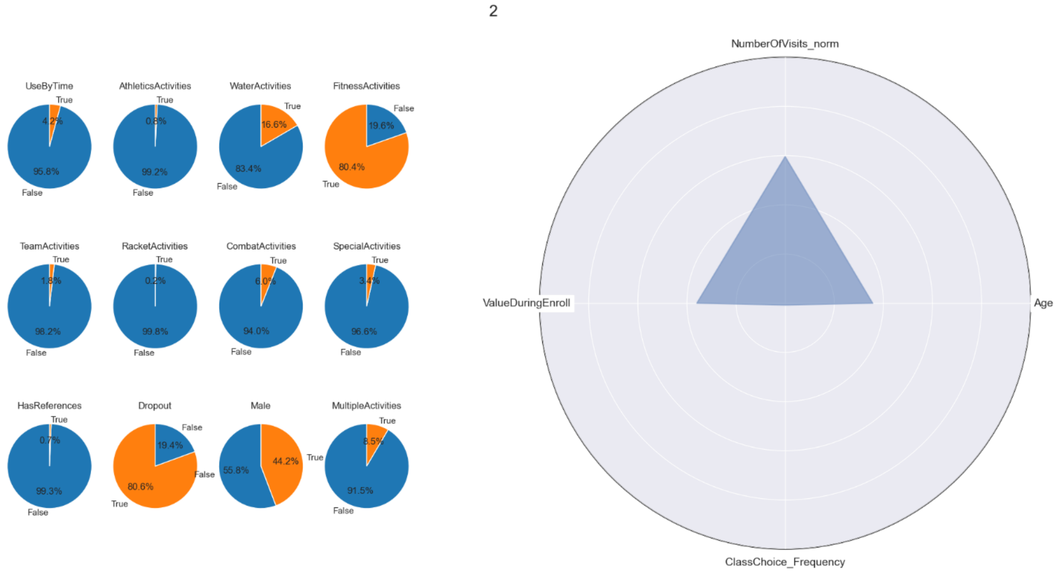


Figure 18: Not segmented solution - cluster 2 profile

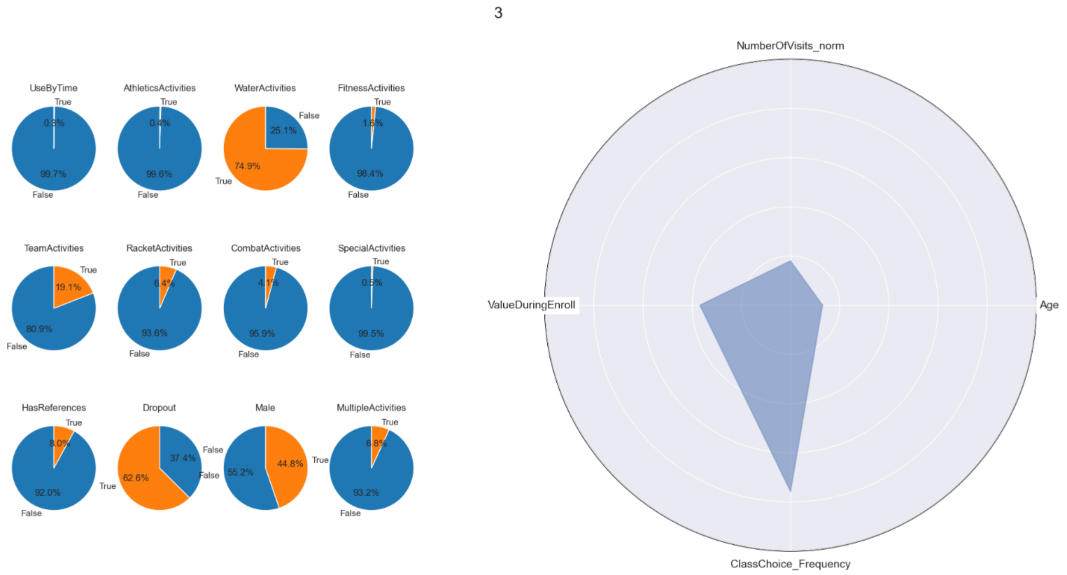


Figure 19: Not segmented solution - cluster 3 profile

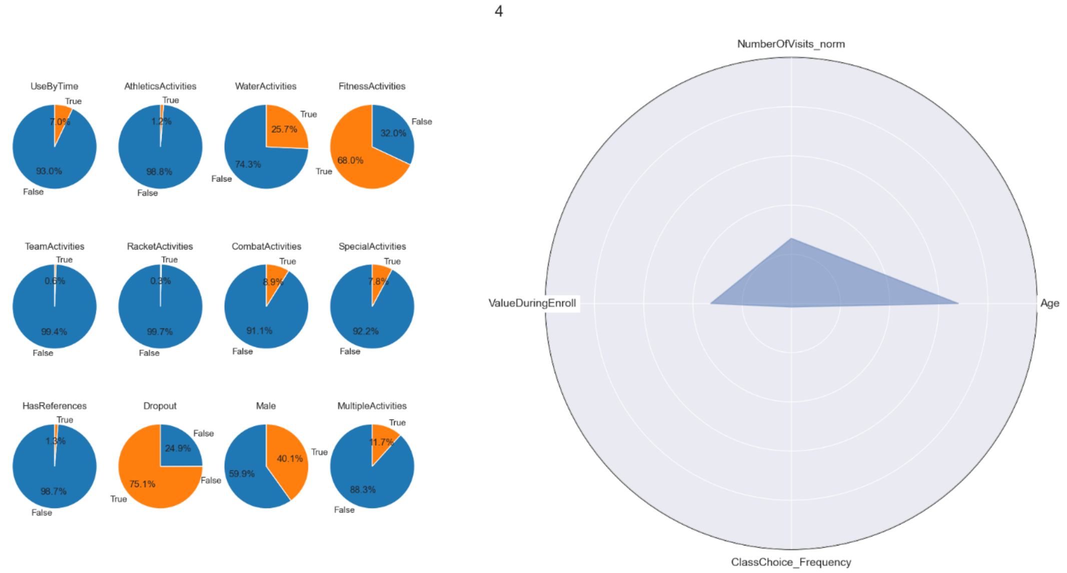
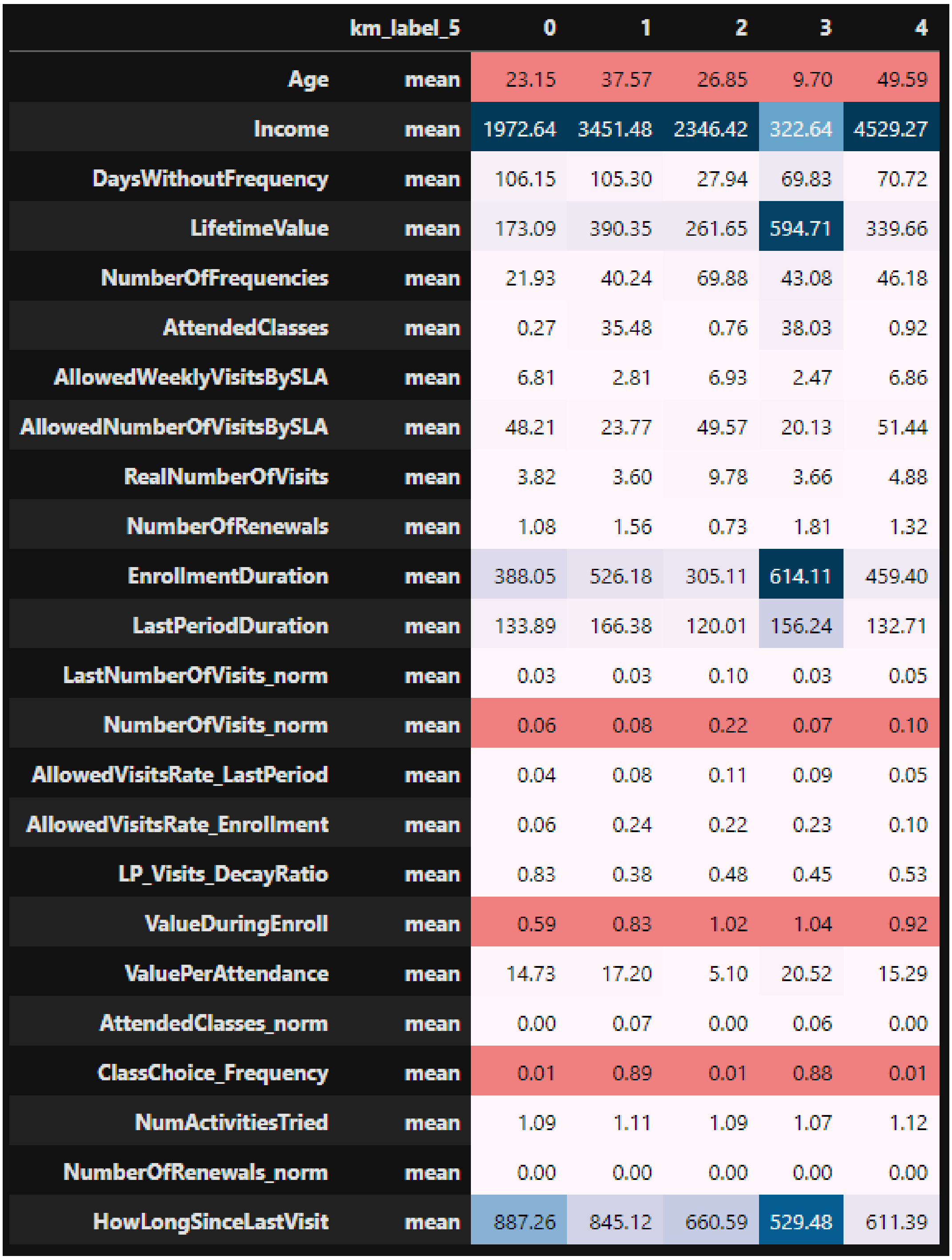


Figure 20: Not segmented solution - cluster 4 profile



	km_label_5	0	1	2	3	4
Age	mean	23.15	37.57	26.85	9.70	49.59
Income	mean	1972.64	3451.48	2346.42	322.64	4529.27
DaysWithoutFrequency	mean	106.15	105.30	27.94	69.83	70.72
LifetimeValue	mean	173.09	390.35	261.65	594.71	339.66
NumberOfFrequencies	mean	21.93	40.24	69.88	43.08	46.18
AttendedClasses	mean	0.27	35.48	0.76	38.03	0.92
AllowedWeeklyVisitsBySLA	mean	6.81	2.81	6.93	2.47	6.86
AllowedNumberOfVisitsBySLA	mean	48.21	23.77	49.57	20.13	51.44
RealNumberOfVisits	mean	3.82	3.60	9.78	3.66	4.88
NumberOfRenewals	mean	1.08	1.56	0.73	1.81	1.32
EnrollmentDuration	mean	388.05	526.18	305.11	614.11	459.40
LastPeriodDuration	mean	133.89	166.38	120.01	156.24	132.71
LastNumberOfVisits_norm	mean	0.03	0.03	0.10	0.03	0.05
NumberOfVisits_norm	mean	0.06	0.08	0.22	0.07	0.10
AllowedVisitsRate_LastPeriod	mean	0.04	0.08	0.11	0.09	0.05
AllowedVisitsRate_Enrollment	mean	0.06	0.24	0.22	0.23	0.10
LP_Visits_DecayRatio	mean	0.83	0.38	0.48	0.45	0.53
ValueDuringEnroll	mean	0.59	0.83	1.02	1.04	0.92
ValuePerAttendance	mean	14.73	17.20	5.10	20.52	15.29
AttendedClasses_norm	mean	0.00	0.07	0.00	0.06	0.00
ClassChoice_Frequency	mean	0.01	0.89	0.01	0.88	0.01
NumActivitiesTried	mean	1.09	1.11	1.09	1.07	1.12
NumberOfRenewals_norm	mean	0.00	0.00	0.00	0.00	0.00
HowLongSinceLastVisit	mean	887.26	845.12	660.59	529.48	611.39

Figure 21: Centroids of not segmented solution

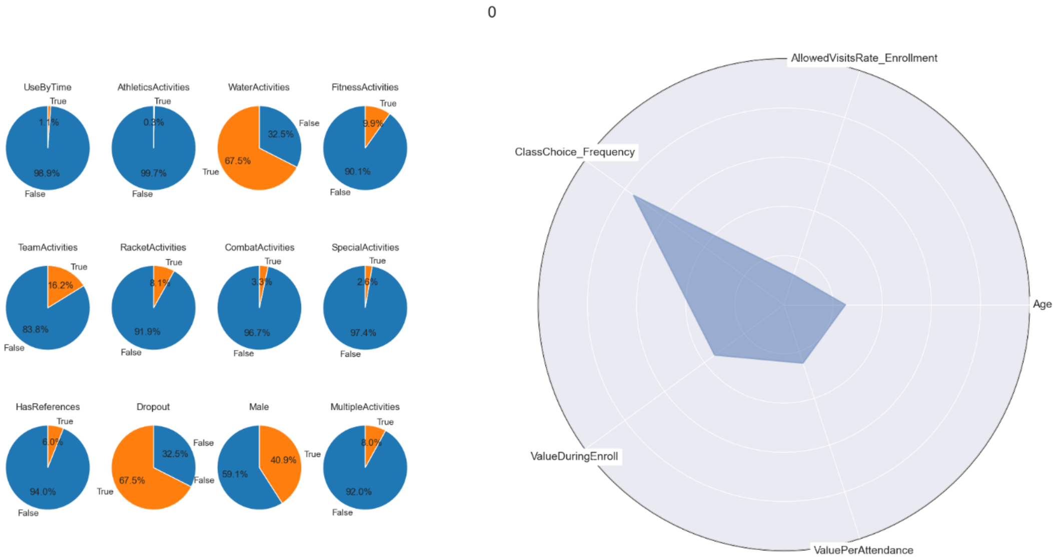


Figure 22: Segmented solution - cluster 0 profile

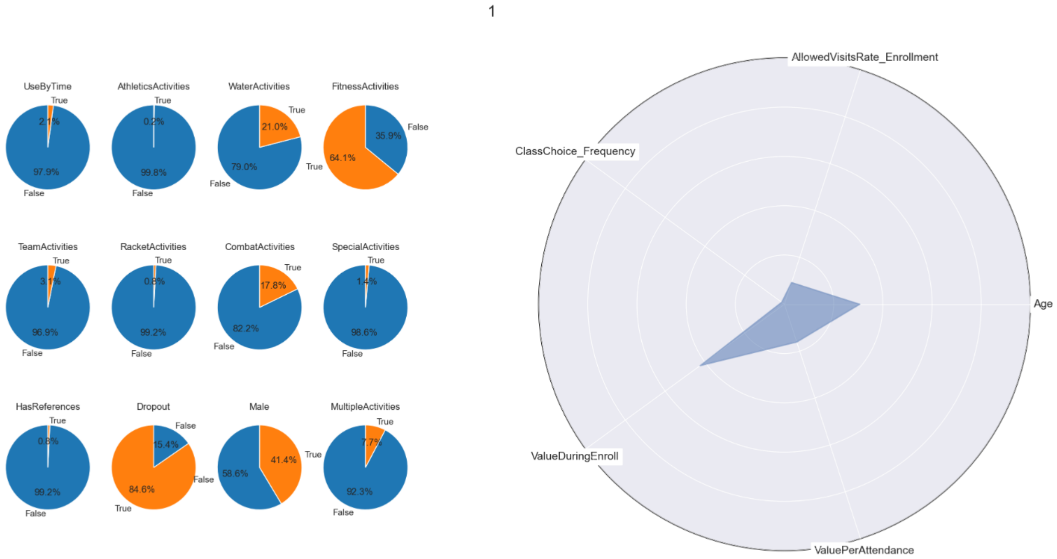


Figure 23: Segmented solution - cluster 1 profile

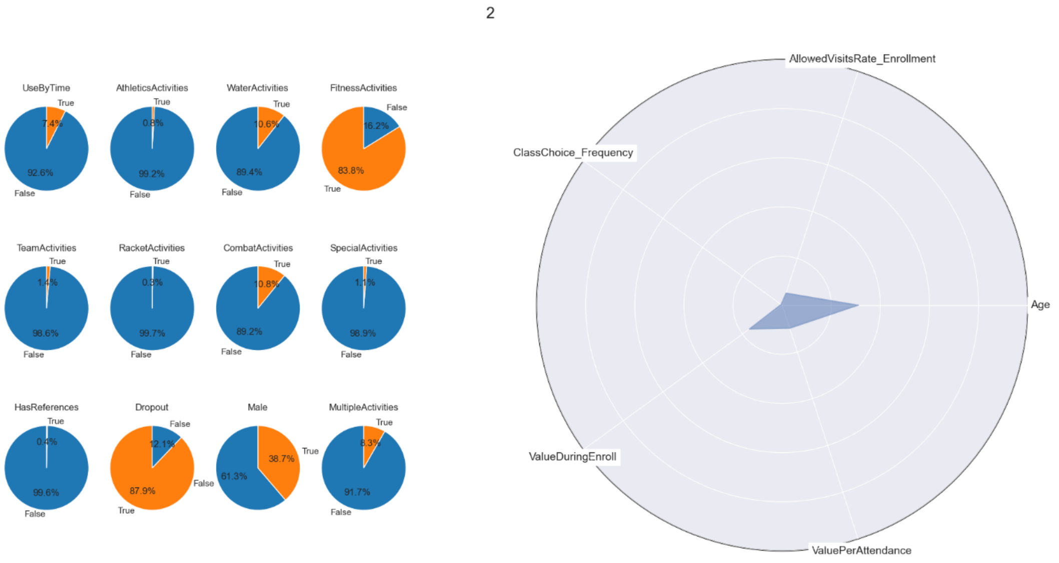


Figure 24: Segmented solution - cluster 2 profile

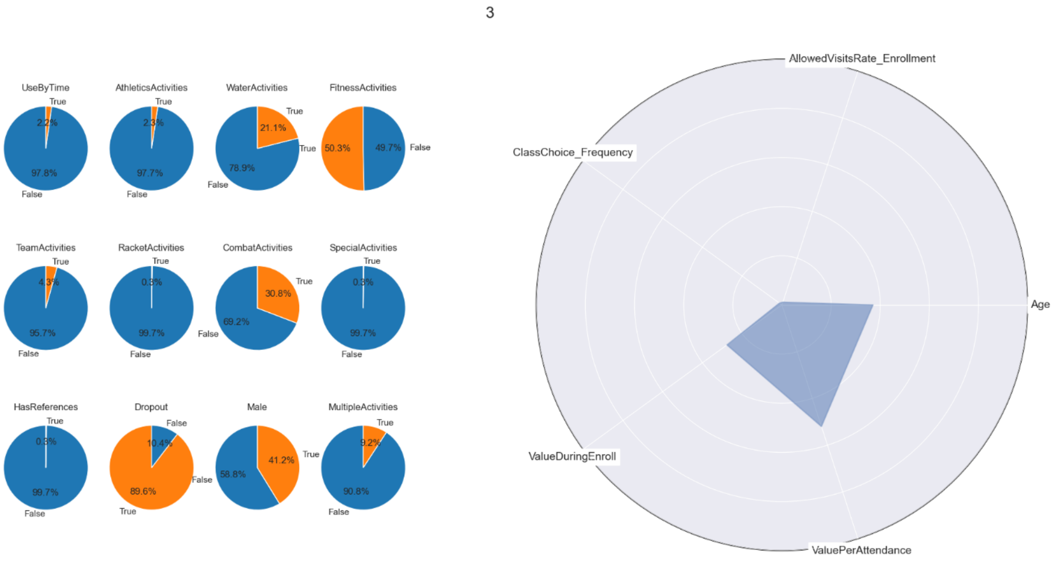


Figure 25: Segmented solution - cluster 3 profile

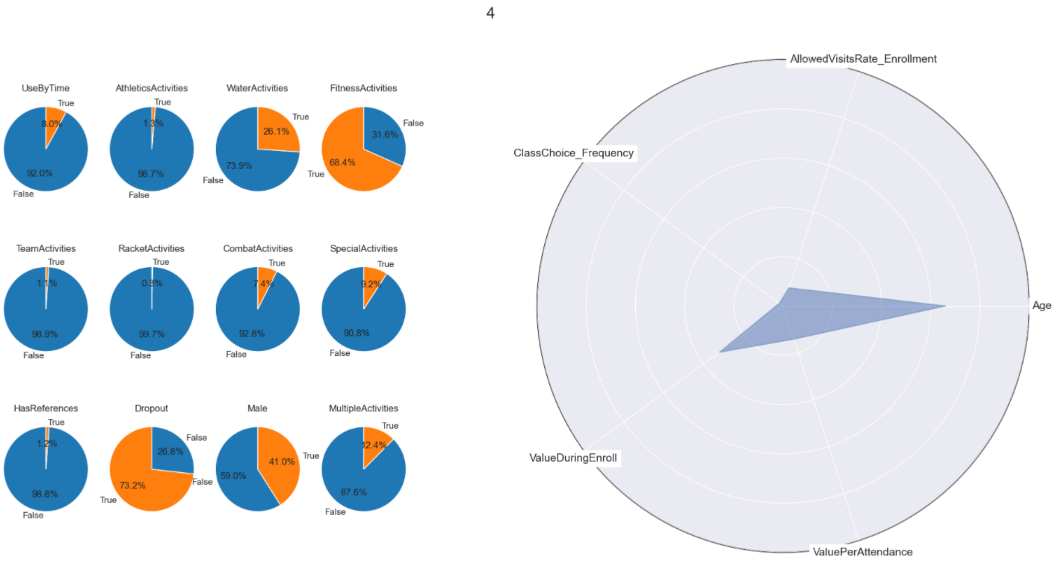
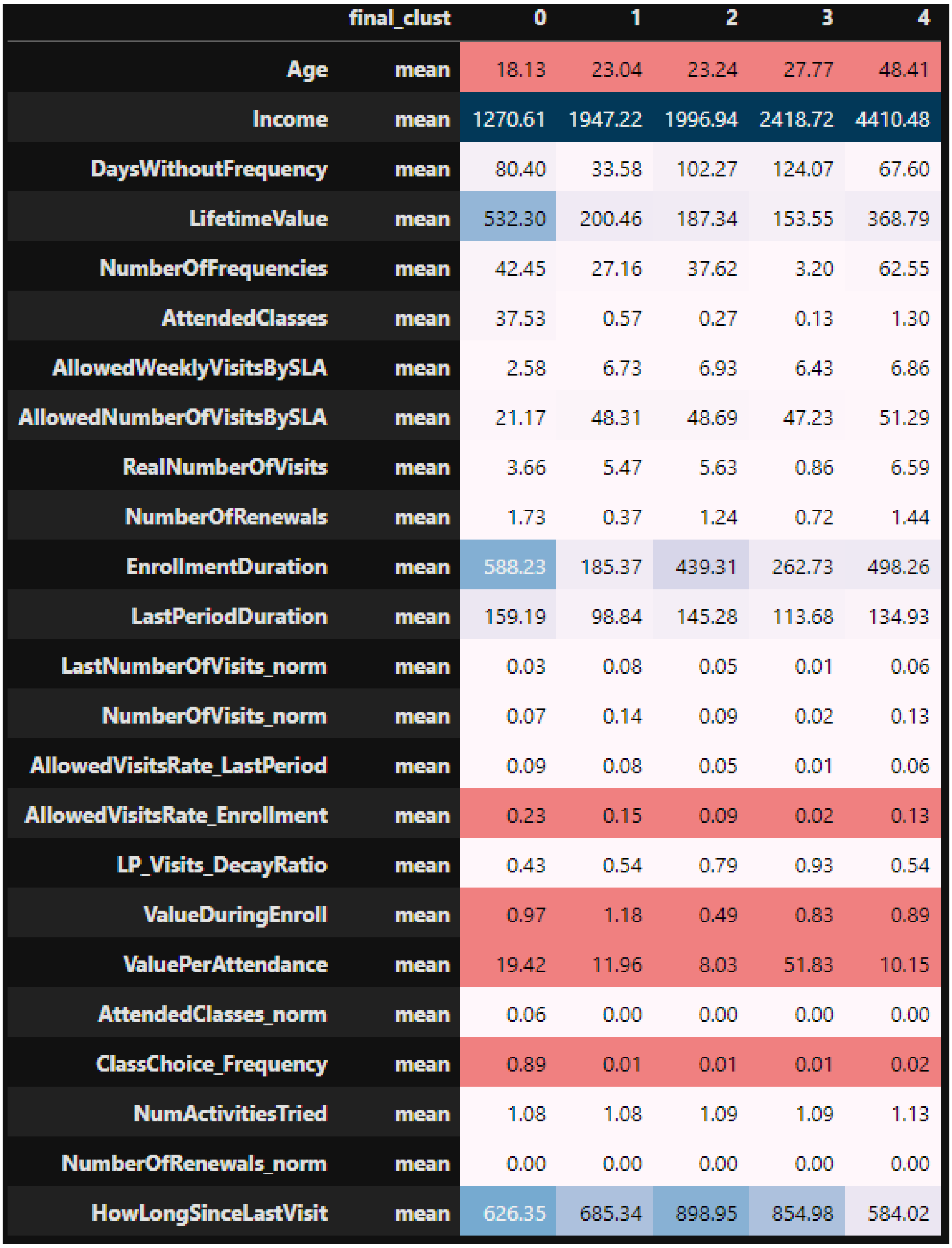


Figure 26: Segmented solution - cluster 4 profile



	final_clust	0	1	2	3	4
Age	mean	18.13	23.04	23.24	27.77	48.41
Income	mean	1270.61	1947.22	1996.94	2418.72	4410.48
DaysWithoutFrequency	mean	80.40	33.58	102.27	124.07	67.60
LifetimeValue	mean	532.30	200.46	187.34	153.55	368.79
NumberOfFrequencies	mean	42.45	27.16	37.62	3.20	62.55
AttendedClasses	mean	37.53	0.57	0.27	0.13	1.30
AllowedWeeklyVisitsBySLA	mean	2.58	6.73	6.93	6.43	6.86
AllowedNumberOfVisitsBySLA	mean	21.17	48.31	48.69	47.23	51.29
RealNumberOfVisits	mean	3.66	5.47	5.63	0.86	6.59
NumberOfRenewals	mean	1.73	0.37	1.24	0.72	1.44
EnrollmentDuration	mean	588.23	185.37	439.31	262.73	498.26
LastPeriodDuration	mean	159.19	98.84	145.28	113.68	134.93
LastNumberOfVisits_norm	mean	0.03	0.08	0.05	0.01	0.06
NumberOfVisits_norm	mean	0.07	0.14	0.09	0.02	0.13
AllowedVisitsRate_LastPeriod	mean	0.09	0.08	0.05	0.01	0.06
AllowedVisitsRate_Enrollment	mean	0.23	0.15	0.09	0.02	0.13
LP_Visits_DecayRatio	mean	0.43	0.54	0.79	0.93	0.54
ValueDuringEnroll	mean	0.97	1.18	0.49	0.83	0.89
ValuePerAttendance	mean	19.42	11.96	8.03	51.83	10.15
AttendedClasses_norm	mean	0.06	0.00	0.00	0.00	0.00
ClassChoice_Frequency	mean	0.89	0.01	0.01	0.01	0.02
NumActivitiesTried	mean	1.08	1.08	1.09	1.09	1.13
NumberOfRenewals_norm	mean	0.00	0.00	0.00	0.00	0.00
HowLongSinceLastVisit	mean	626.35	685.34	898.95	854.98	584.02

Figure 27: Centroids of segmented solution

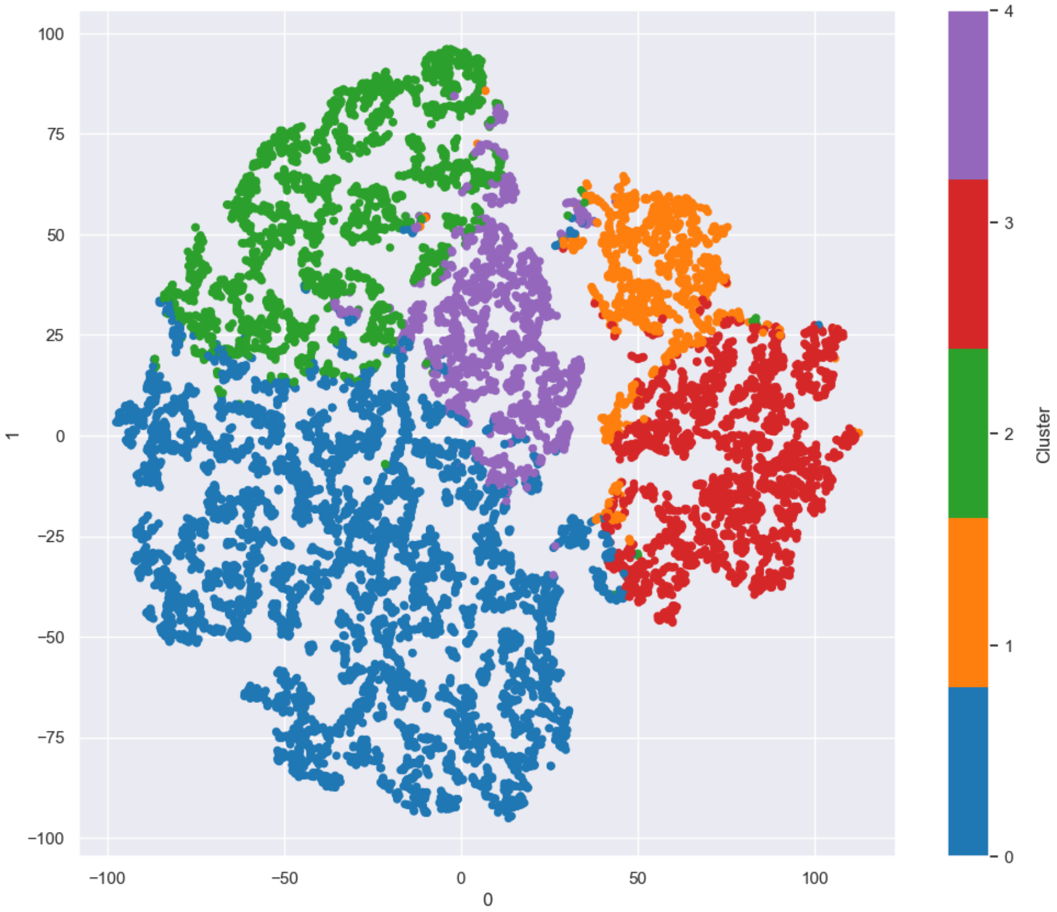


Figure 28: t-SNE graph for the not segmented solution

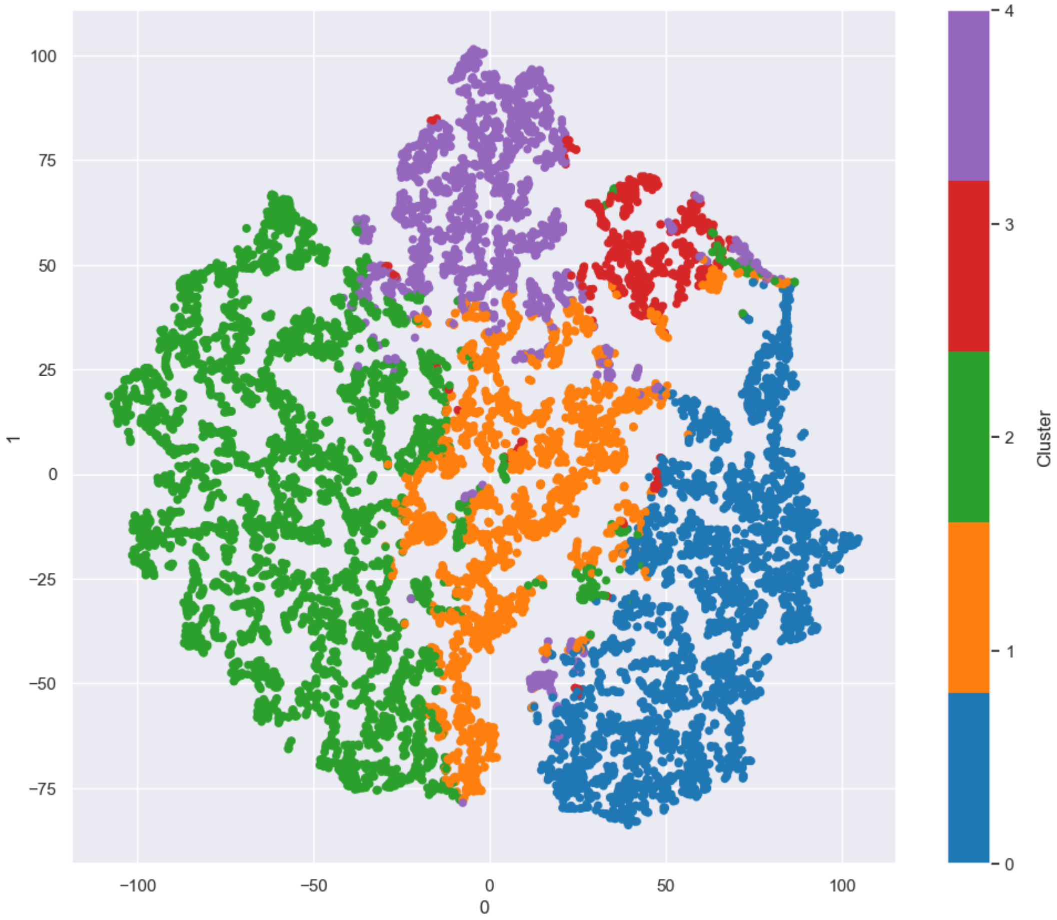


Figure 29: t-SNE graph for the segmented solution