Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Customer Segmentation Strategy for XYZ Sports Company

Group 36

André Filipe Silva, 20230972

Catarina Reis, 20230981

João Gonçalves, 20230560

January, 2023

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

INDEX

Table of Contents

[1] Service Level Agreement

1. Introduction

In the fitness industry, understanding customer behavior is key. Clustering segmentation helps identify unique engagement patterns, allowing management to segment the market, a pivotal step in any marketing plan. It is our hope that applying clustering techniques to XYZ Sports Company's customer data will reveal distinct segments, that can be worked upon with marketing plans for increased value to the company.

Our analysis aims to provide XYZ Sports with insights to dynamically adapt to diverse customer demands, refining and personalizing offerings for an engaging fitness experience. This not only enhances customer satisfaction but also strategically positions the company in the competitive fitness market, aligning business decisions with the unique needs of each segment.

2. Data Exploration

The dataset used for our research is from the company's ERP system, having a rich set of customer-related data collected between June 1st, 2014, and October 31st, 2019. This dataset spans 14942 rows and comprises 30 features.

One of our initial steps was to eliminate the *ID* feature, as its sole purpose is customer identification. We opted to use the default index, which already provides a unique identification for each member.

We started by conducting a simple examination of the data *[Table 1]* and *[Figure 1]*. Notably, we observed that most customers are young adults, with an average age of around 26 years, having the oldest member in our dataset 87 years old. There are also more females in the facility center, and the most popular activities are water and fitness activities.

Additionally, a significant portion of members goes more than 40 days without returning to the facility, indicating a potential area for improvement. Notably, external references rarely influence members to start their sports journey at this facility. We also identified missing values, with *Income* and *AllowedWeeklyVisitsBySLA* being the most affected. Curiously, we found members under one year old, suggesting that this facility caters to a diverse audience, including newborns.

Also, the observations include the presence of zero income values, likely representing children being financially dependent on another adult.

Another relevant point to mention is that two of the activities, *DanceActivities* and *NatureActivities* were not frequented by any members present in our dataset *[Figure 1]*. Thus, we are going to assume that the fitness center had this type of service in the past, which was discontinued. This way, we decided to remove these features.

These are preliminary insights, and now we are ready to dive deeper into our data.

2.1. Types conversion

Data types significantly impact clustering success. Recognizing their significance, we meticulously converted features to ensure a prevalence of numeric types, aligning with the requirements of most clustering methods.

There were a lot of features with data types not prone for analysis. We will not go into the technical details of this, as it is out of the scope of a report such as this (details can be consulted in the notebook). *HasReferences, AllowedWeeklyVisitsBySLA, EnrollmentStart, EnrollmentFinish* underwent transformation.

2.2. Duplicate records

The dataset revealed one duplicate record. Given how uncommon it is to have the same information for two or more individuals across all features, we decided to remove the duplicated record, assuming it represents the same individual.

2.3. Data coherence

To ensure a reliable analysis of our data, we made a structured coherence analysis based on the assumption that we were dealing with a Portuguese company. This way, we could justify some situations, without the need to remove and lose relevant information.

We started by looking at dates, making sure their values made sense. Having in consideration the fact that our data was collected between June 1st, 2014, and October 31st, 2019, we check for incoherences in across the following features *EnrollmentStart*, *EnrollmentFinish*, *LastPeriodStart*, *LastPeriodFinish* and *DateLastVisit*. Upon analysis, we found one issue – it was observed that 768 records within the dataset display instances where the *DateLastVisit* falls outside the designated enrolment period. This discrepancy raises concerns about data coherence. In response, we removed the *LastPeriodStart* and *LastPeriodFinish* features. This decision was based on the recognition that *DateLastVisit* already gives us insights about the user's last visit, including the semester in which the last visit occurred.

We also checked if anyone visited the fitness center after their membership period ended. We did find 208 cases where it happened and the way we decided to deal with them was by setting the date of last visit to the last period finish.

Another thing we checked was if anyone went over the allowed number of visits. Since we found 48 cases where the members did exceed the number of allowed visits, we assumed it was a system error, thus we set up their number of visits to the maximum number officially allowed.

For the *Income* variable, we made sure that individuals under 16 had their income set to 0, following the Portuguese legal working age. This adjustment affected 17 records, making our income data more accurate.

Also, we noticed a curious situation where some people had *HasReferences* as true, but *NumberOfReferences* was 0. To make sense of this, we checked the ages of these members. Given that they were all above 8 years old — a reasonable age to receive recommendations (from friends, for example)— we assumed that a *NumberOfReferences* value of 0 likely indicated a potential error or missing information. Our treatment for this inconsistency was to set to False if the *NumberOfReferences was 0.*

Finally, we notice that there were customers who didn't pay anything to the fitness center between their enrolment dates, having the *LifeTimeValue* equal to 0. Since we only had 3 cases in this situation, we agreed in removing those records because they didn't make sense.

2.4. Data Visualization

Effective data visualization in the initial exploration phase is crucial for uncovering patterns, understanding variable relationships, and identifying key features.

Looking at *[Figure 1]*, we can start by pointing out that most members have a full-week SLA[1] to the facility center. The second most prevalent SLA is for two visits per week.

Observing the histograms in *[Figure 2]*, it's apparent that most features exhibit leftward skewness, indicating a notable presence of outliers. Dealing with these outliers is vital for precise clustering, particularly in techniques relying on distances. Further details on addressing outliers will be discussed in the subsequent section titled "Outliers Removal" (3.1).

While examining Spearman correlations *[Figure 3]*, we observed that most of the variables showed moderate correlations. Two pairs of variables stood out: *Income* and *Age*, with a positive correlation of 0.87, and *AllowedWeeklyVisitsBySLA* and *AttendedClasses*, with a negative correlation of -0.85. The features *NumberOfFrequencies* and *LifetimeValue* also exhibited a positive correlation of 0.74, similar to *AllowedNumberOfVisitsBySLA* and *AllowedWeeklyVisitsBySLA*, which had a positive correlation of 0.69. From this analysis, it would become obvious to drop either Age or Income. However, as we wanted both features to stay somehow, we later transformed Age into a categorical variable and left Income as is. A more thorough justification for this decision will be provided in the upcoming section on "Feature Selection."

The analysis of age-related activity preferences *[Figure 4]* reveals distinct patterns. Water activities dominate among children, spanning from newborns to 13 years old, aligning with known benefits for children, especially newborns. Combat Activities show a small but notable presence until age 13, increasing in adolescence along with a rise in Team Activities.

Teenagers exhibit a balanced interest in Combat and Team Activities. Young adults prioritize fitness activities, that persists beyond age 30 and continues through older age brackets. Notably, as individuals age, there is a gradual introduction of special activities. This also makes sense, as people begin to deal with certain limitations and disabilities as they get older.

In *[Figure 5]*, one can clearly see the dominance of fitness activities among other activities, followed by water activities, which also have a significant presence but with around half of the overall participations compared to fitness. Thus, fitness and water activities emerge as the highest-participation categories overall. Even though fitness stands out as more popular among both genders, it has a bigger emphasis on female engagement. Finally, athletics, water, and racket sports show a balanced participation between both genders.

In financial terms, the 7-13 age group is the standout contributor, peaking at around 690 in *LifetimeValue* *[Figure 6]*. Even the 0-6 age range is significant, hitting approximately 560. For the teen and adult categories (14-18, 31-50, 51-100), the revenue is steady but more moderate, ranging from around 320 to 430. However, the 19-30 age group shows a dip, with a *LifetimeValue* around 185, indicating a need for strategic enhancements.

In essence, our revenue drivers are the 7-13 age group, closely followed by the 0-6 age range. Teens and adults contribute consistently, while the 19-30 group requires focused strategies for increased financial impact.

3. Data Preprocessing

3.1. Outlier Removal

Taking in consideration the boxplots and histograms analyzed in Data Visualization (2.4.), we already know that there were some cases that stood out, having a concerning presence of outliers. To solve this issue and search for more outliers that were not already detected, we started by applying two methods to deal with the outlier removal, including: **IQR Method** and **Manual Filters,** remaining with approximately 79.9% and 98.15% of the data, respectively.

As one of the requirements for **Z-Score** removal assumes normal distributions and we can see in the histograms that we have a lot of skewed data, we did not use this method.

After the manual outlier removal, we re-evaluated the boxplots, which did not exhibit a significant change, aligning with our expectations based on the nature of the data:

- *Age*: Since the fitness center has customers of all ages, it makes sense to consider all ages for our clustering purpose .
- *DaysWithoutFrequency*, *AttendedClasses*, *NumberofFrequencies*, *LifetimeValue* and *RealNumberOfVisits*: The diversity of values present in these features are necessary to represent the different engagement levels of each customer.

3.2. Feature engineering

Our original dataset had a lot of features in formats that were difficult to work with, so in order to simplify our analysis and extract more valuable information from our data, we created the following features:

- ***AgeGroup***: Based on the numerical feature *Age*, we created a categorical feature to categorize individuals into distinct age groups: Children (0-12), Teenagers (13-19), Young Adults (20-39), Adults (40-59) and Seniors (above 60).
- ***MembershipDuration:*** Calculated from *EnrollmentStart* to *EnrollmentFinish* inclusively, *MembershipDuration* succinctly represents the length of membership in days.
- ***MonthlyVisits***: Indicates the average monthly visits per customer.
- ***MonthlySpending***: Indicates the average monthly spending per customer.
- ***TotalActivities***: Counts the total number of activities in which the customer participated.

3.3. Transforming skewed data

By looking at the distribution of the metric features *[Figure 8]*, we can notice that we are essentially dealing with left-skewed data. Since some clustering algorithms may be sensitive to the skewed nature of the data, we tried to reduce the skewness. For this, we assessed the skewness measure for each feature, with values exceeding 1 or falling below -1, indicating a highly right or left-skewed distribution, respectively. To align features more closely with a normal distribution, we applied square and cube root transformations.

Despite our best efforts to make our data follow a normal distribution, even after applying the transformations, our data continues to exhibit skewness, although we did achieve very good improvements regarding some variables, such as *DaysWithoutFrequency, LifetimeValue, NumberOfFrequencies*, and *RealNumberOfVisits*. *[Figure 9]*

3.4. Scaling

Before using K-Nearest Neighbors (KNN), we'll make sure all numerical values are on the same scale. Since our data was originally left-skewed and after transformation it still exhibits skewness, it's important to choose a scaling method that is robust to such non-normal distributions. This way, we used the MinMaxScaler.

We thought about using the StandardScaler but this method assumes that the data follows a Gaussian distribution which is not our case.

3.5. Missing values

At this point, after removing and creating features, we searched for missing values. Specifically, we observed that income had 0.15% missing values, and each of the activity-related features had less than 0.07% missing values. For Income we used the KNNImputer to fill in the missing values.

For the various types of Activities offered by the facility, we assumed that missing values were equivalent to non-participation – so we filled those missing values with 0. For *NumberOfFrequencies* and *MonthlyVisits*, after visualizing their distributions, we decided to fill the missing values with the median.

Finally, for *AllowedWeeklyVisitsBySLA and AllowedNumberofVisitsBySLA*, which are categorical variables we inputted the mode.

Post-missing value imputation, a check on the variable histograms shows consistent results, confirm that the distributions of the variables remain mostly intact, which is our main goal when performing this process.

3.6. Reassessing Outliers using DBScan

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that groups together data points that are close to each other based on a density criterion. The most important parameter to decide on when using DBSCAN is 'epsilon' – the acceptable maximum distance for two points to be considered in the vicinity of one another. There is an auxiliary plot often performed to help choose epsilon – in this case we decided for an epsilon of 0.4.

DBSCAN ended up removing 0.93% of our data, leaving us with 99.07%.

This was our final chosen method to remove outliers, disregarding Manual or IQR.

4. Features Selection

Feature selection is a very important step in a task like this. It is the features that are chosen that give (or not) predictive power towards solving our problem. We tackle the selection of numerical and categorical features separately.

4.1. Numerical features

For numerical feature selection, we chose based on a combination of two methods: the Spearman Correlation Matrix and the Component Planes using Self-Organized Maps, after all the preprocessing was done. Looking at the results and correlations, we decided to keep the following metric variables: Income, LifetimeValue, DaysWithoutFrequency, NumberOfFrequencies, RealNumberOfVisits, NumberOfRenewals, MembershipDuration. These features appeared to strike a good balance of importance to our solution and no multicollinearity. There is a note to be made here: several features would be kept according to the Spearman Correlation Matrix, but the Component Planes showed their poor differentiating nature, and as such a decision was made not to keep them.

4.2. Categorical features

Regarding feature selection on categorical feature, our thought process is much simpler. All except one of our categorical features are binary. If we want to differentiate customers, we need one thing: variability. As such, we are only interested in categorical features that have a significant amount of customers in both categories. Going back to *[Figure 1]*, one can easily visually inspect this. As such, the categorical features to keep were Gender, WaterActivities, FitnessActivities, AllowedWeeklyVisitsBySLA and Dropout. In a first analysis, we still included CombatActivities, but later figured out it badly skewed our results due to lack of variability.

Segmentation Perspectives

Based on our features, we decided to perform customer segmentation on three perspectives, so that we could gather the best patterns from our data. Those segments were Socio-Demographic, Value, and Product. On the Socio-Demographic front, we relied on the features *AgeGroup*, *Income*, *Gender* and *AllowedWeeklyVisitsBySLA* to characterize our customers. For Value, we had a number of useful metric features – *LifetimeValue*, *DaysWithoutFrequency*, *NumberOfFrequencies*, *RealNumberOfVisits, NumberOfRenewals* – and a categorical feature – *Dropout.* Finally, for our Product segmentation, we relied on the variability presented by the features *WaterActivities, FitnessActivities, CombatActivities, MembershipDuration*.

5. Clustering

Before we begin describing our clustering techniques and results, it is essential to say that, due to the fact that many (if not the majority) of the features employed in our clustering process were categorical, we did not use Euclidean Distances as a way to measure distances between points, but instead went with Gower distance[1], which can account for categorical variables as well as metric variables.

Also, as a matter of consistency, and since all our segments have a mix of categorical and metric features, we decided to go with the three same clustering techniques for the data: K-Prototypes, Hierarchical Clustering, and K-Medoids.

[1] Tuerhong, G., & Kim, S. B. (2014). Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert systems with applications*, *41*(4), 1701-1707.

K-Prototypes is a clustering algorithm that works by partitioning the data into K clusters, similar to K-means. However, it works well in the presence of categorical variables, unlike K-means. Due to problems with its implementation though, we mostly used K-Prototypes to start getting a sense of how many cluster our data would present us with based on the elbow method. As a matter of precaution, we would like to note that with this method we are only getting an initial suggestion for the number of possible clusters in our data and not really a formal solution.

Hierarchical Clustering is one of the classical clustering algorithms used in the data science industry. The algorithm starts by treating each data point as a separate cluster and then combines two data points by a measure of distance. It involves a lot of hyperparameter tuning, such as 'linkage' or 'metric'.

Finally, K-medoids is a partitioning technique of clustering that splits the data set of n objects into k clusters, where the number k of clusters is assumed to be know a priori (which is why we leave this method to the end, so we have a sense of how many clusters we have in our data beforehand). The algorithm works by first choosing k number of random points from the data and assigning these k points to k number of clusters. These are the initial medoids. For all the remaining data points, the distance from each medoid is calculated and the data point is assigned to the cluster with the nearest medoid. The algorithm continues to iterate until the medoids no longer change. We used the method 'fasterpam'.

Our rationale for clustering is as follows: We plot the Cost Curve for K-Prototypes and get an initial suggestion for the number of clusters. We take that suggestion and perform Hierarchical Clustering. With those two methods together, we get a pretty good sense of what our solution should be. In the end, we perform K-Medoids and compared the results with Hierarchical Clustering.

5.1. Socio-Demographic Clustering

Starting with Socio-Demographic Clustering, using the K-Prototypes technique pointed us towards 4 clusters in our data *[Figure 10]*. We took that forward to Hierarchical Clustering, where we computed the Gower Distances, and silhouette scores for every linkage model (single, complete, average) *[Figure 11]*. We got our best results with the 'complete' linkage model, which in turn provided us with an answer that 4 clusters would be the best way to go when analyzing the dendrogram. Since it did not go against what we had found previously in the first method we decided to go forward with 4 clusters to the K-Medoids method *[Figure 12]*. Doing our computations, we achieved a silhouette score of 0.7582 with the 4 clusters using the K-Medoids method.

5.2. Value Clustering

Looking at our K-Prototypes Cost Curve, we are inclined towards choosing 3 clusters *[Figure 13]*. Again, moving on to the Hierarchical Clustering, computing scores for a range of different clusters and all linkage models, we find a sweet spot at linkage model 'complete' with 3 clusters *[Figure 14]*.Taking it to K-Medoids *[Figure 15]*, for the 3 clusters we obtain a silhouette score of 0.6580.

5.3. Product Clustering

Looking at the K-Prototypes Cost curve proved to be a bit more difficult *[Figure 16]*. Several numbers could have been reasonably chosen. We attempted to strike a balance between performance and simplicity and chose 4 clusters.

Computing the Gower Distances once more in Hierarchical Clustering, and the silhouette scores, we chose 'average' linkage and 4 clusters, with a silhouette score here of 0.7 *[Figure 17]*. Going into K-Medoids, the results were similar with a silhouette score of 0.7045 for 4 clusters *[Figure 18]*.

6. Cluster Results Analysis

In our work, K-Medoids proved more reliable than Hierarchical Clustering when it comes to the final results – perhaps due to our abundance of categorical variables – and as such, K-medoids was our preferred method when choosing the outcome for clustering.

With this said, we ran all the code for the cluster segments again using K-medoids, this time going a bit further – we looked into how many values each cluster had, if they were balanced and whatnot. This is when we started to get some hints that some of our clusters might need manual merging.

Looking at **Socio-Demographic**, clusters 3 and 4 were small and sparse (1741 and 1297 observations respectively, compared with the much larger siblings standing at 7126 and 4639 observations). After taking a glance at the metric variable that is a part of this cluster, *Income,* we also notice that the mean values for the clusters are not too far apart from each other. Based on all of this, we decided to merge these clusters with the bigger ones, ending with a final solution of two clusters for the Socio-Demographic segmentation. Doing a deep dive into this segment, we can see the clusters continue to not differ much on income. However, the differences between the remaining variables are stark. One of the clusters is older on average. This cluster is also predominantly female (60-40 ratio) and has fewer contracted visits per week compared to the other cluster. The second cluster is composed of younger individuals, predominantly male (65-35 ratio) but with a higher contract of weekly visits *[Figure 19]*.

Passing onto **Value**, although we see clusters with a lot of parallel tendencies, there are visible differences amongst them. The Lifetime Value of Cluster 1 is higher, meaning they provide higher value to the business, and this is probably due to the fact that they spend less days without frequenting our fitness facility. This is reflected by the considerably higher number of frequencies and real number of visits, which obviously reflects on the higher number of contract renewals. This is all in comparison to cluster 0 *[Figure 20]*.

Finally, the **Product** segment. On this segment, we can identify that the clients with the shortest membership duration, are the ones who are the most into fitness activities. On the other hand (and cluster), the clients who tend to stay with the business the longest are the ones most invested in water related activities *[Figure 22]*.

7. Final Solution

In the end, after employing a variety of methods and analysis techniques, and attempting to analyze the data from a series of different perspectives we must merge to find one final solution. By merging, we have a total of 12 different clusters. To identify our final solution, we resorted to hierarchical clustering (as used in class) *[Figure 22]* and ended up with 3 different clusters. One thing is visible in these clusters – Income is similar between the three, meaning that it is not a dividing factor when adhering to a fitness facility (which is backed-up by real-life data – there are gyms for every wallet out there). Now, let us take a deeper look at the clusters: *[Figure 23][Figure 24]*

Cluster 0 – This is the cluster that brings most value to our business - jointly with Cluster 1 they have the biggest *LifetimeValue* in the dataset. These clients are the ones that frequent the gym more often. They are predominantly female, and they seem to prefer Water Activities to Fitness Activities. One concerning characteristic of this cluster is that their Membership Duration is extremely low, compared to the other clusters. Given the fact that these clients seem to enjoy the gym, there could be a special follow-up with them to understand why they are leaving and try to get information from there that would be useful to get them to stay. Maybe the gym is overcrowded? Maybe it doesn't have the most suitable gear? This seems like the sort of issue that needs client feedback in order to get better.

Cluster 1 – Along with Cluster 0, this cluster brings us a lot of value to the business. They do not come as consistently as clients in Cluster 0, but they are also predominantly female. In this case they do prefer Fitness Activities to Water Activities. Unlike Cluster 0, they rarely quit the gym. They stay for a long time. It could be useful to get their feedback on what makes them stay so long and compare it to the feedback gotten from the Cluster 0 clients. Also, the best way to improve the value obtained from this cluster would probably be to provide them with discounts with "Refer-A-Friend" type of campaign, so they can bring friends and family to the gym. Going to the gym alone can often be quite a chore at the start, but if a new person has the company of someone that seems to enjoy gym as much as our cluster 1 clients (and we're basing this "love" on the *NumberOfRenewals* and *MembershipDuration*), they are more likely to stick around, and so it seems like a great way to gather new clients and expand.

Cluster 2 – We would classify this cluster as the more "unmotivated" type. They go to the gym the lowest number of times out of the three clusters. As for every cluster, they are predominantly female, but these clients in particular seem to love Fitness Activities. What strikes us as odd is that, even though their motivation does not seem very high, their membership duration is decent. Even so, they don't renew their contract often. Perhaps these clients could be given discounts towards their renewals so that they don't drop so often. Also, targeted campaigns to incentivize them by the gym coaches could help them get their motivation back on track and hopefully boost their confidence and get them to go to the gym more.

A 2D visualization of our final solution using t-SNE can be found in *[Figure 25]*.

It is important to discuss the limitations of our work. We worked with a lot of categorical variables and not many metric variables. Performing clustering with categorical variables proved to be a more complex task and the results did not turn out as good as we wanted them to. A lot of numerical features were just not useful to our task, so they were dropped. We had a lot of datetime features that we did not know what to do to begin with, and maybe some more exploration could have been done in that sense. All in all, we don't feel our solution is bad, but we feel like it could have benefitted from more information.

8. References

- Ziafat, H., & Shakeri, M. (2014). Using Data Mining Techniques in Customer Segmentation. Hasan Ziafat Int. Journal of Engineering Research and Applications.
- Tuerhong, G., & Kim, S. B. (2014). Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert systems with applications*, *41*(4), 1701-1707.

9. Appendix

Figures and Tables

	count	mean	std	min	25%	50%	75%	max
Age	14942.0	26.015794	14.156582	0.00	19.00	23.00	31.000	87.00
Income	14447.0	2230.816086	1566.527734	0.00	1470.00	1990.00	2790.000	10890.00
DaysWithoutFrequency	14942.0	81.224936	144.199576	0.00	13.00	41.00	83.750	1745.00
LifetimeValue	14942.0	302.561871	364.319566	0.00	83.60	166.20	355.075	6727.80
NumberOfFrequencies	14916.0	40.120542	65.466459	1.00	7.00	18.00	45.000	1031.00
AttendedClasses	14942.0	10.152456	29.154202	0.00	0.00	0.00	3.000	581.00
AllowedNumberOfVisitsBySLA	14942.0	41.636299	21.066166	0.56	25.72	38.99	60.970	240.03
RealNumberOfVisits	14942.0	5.320707	6.332958	0.00	1.00	4.00	7.000	84.00
NumberOfRenewals	14942.0	1.205260	1.381305	0.00	0.00	1.00	2.000	6.00
NumberOfReferences	14942.0	0.022286	0.166777	0.00	0.00	0.00	0.000	3.00

Table 1: Descriptive Statistics of numerical features

Figure 1: Categorical Feature Frequency Distributions

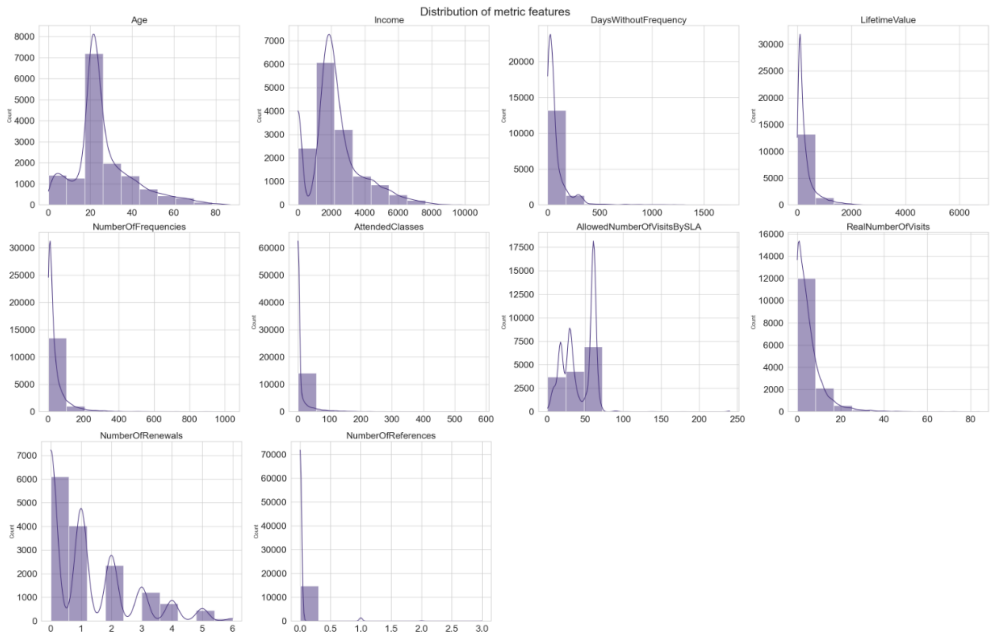

Figure 2: Distribution of numerical features

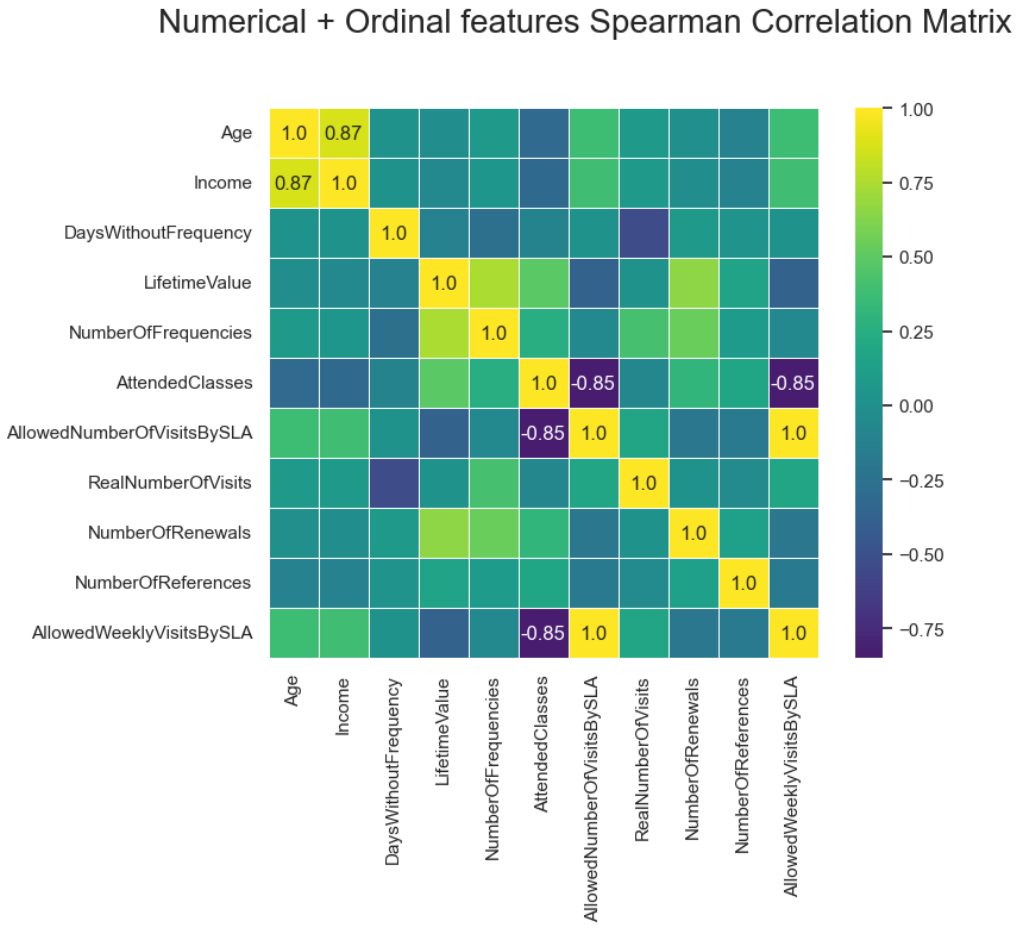

Figure 3: Spearman Correlation Matrix – Numerical and Ordinal features

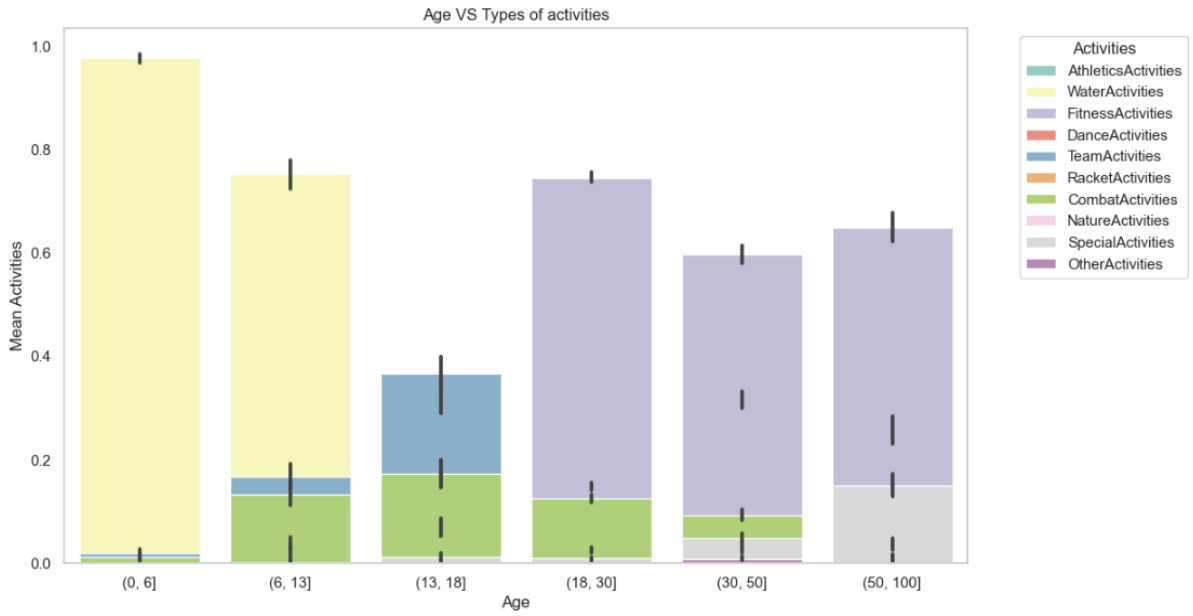

Figure 4: Age VS Types of Activities

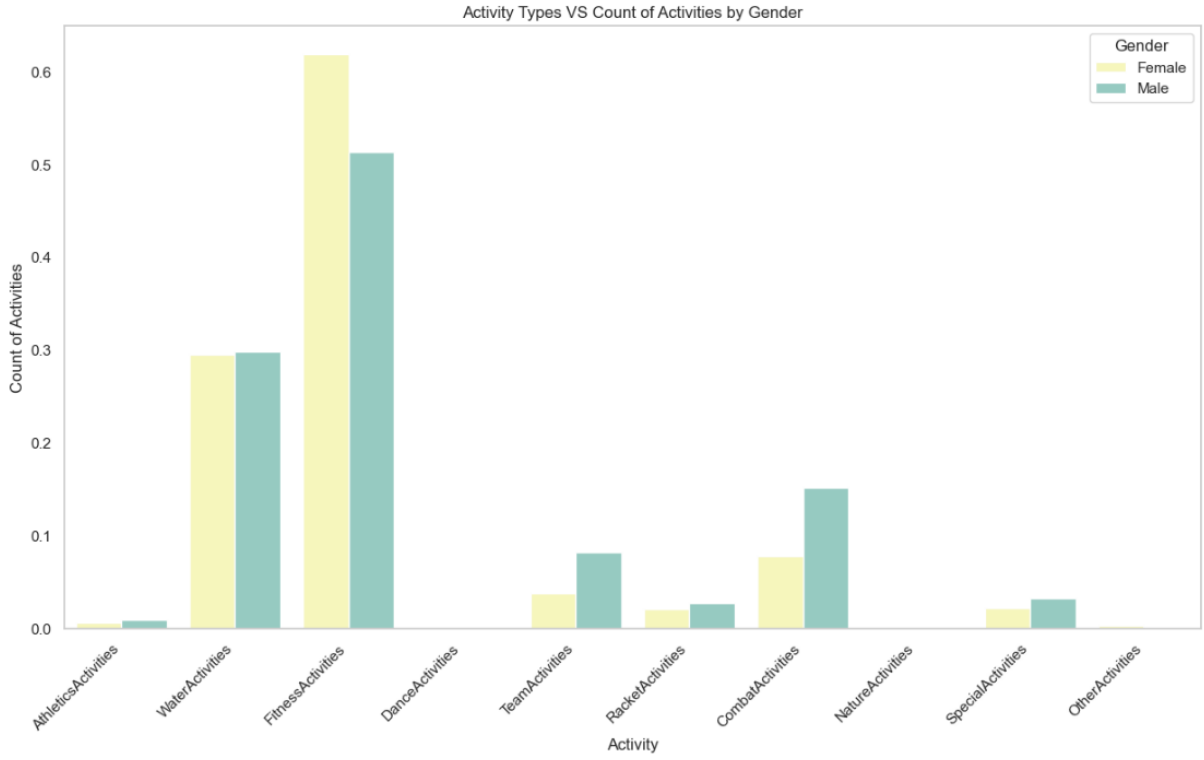

Figure 5: Activity Types VS Count of Activities by Gender

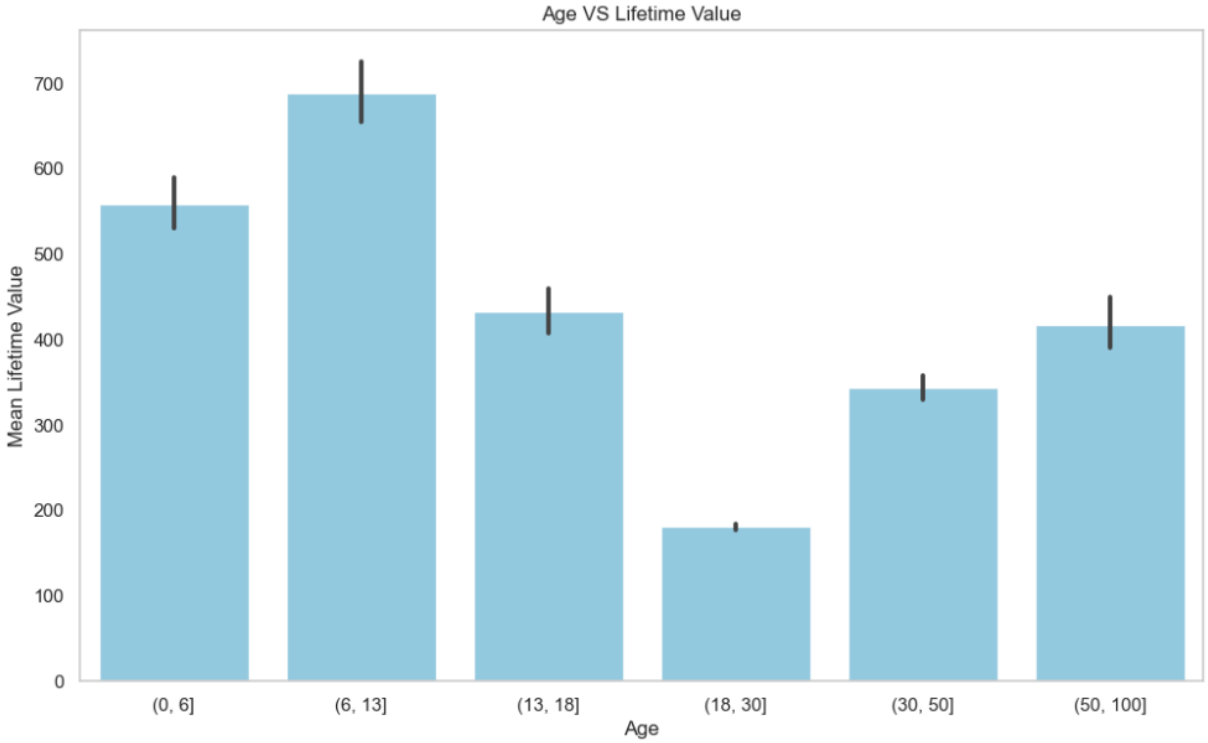

Figure 6: Age VS Lifetime Value

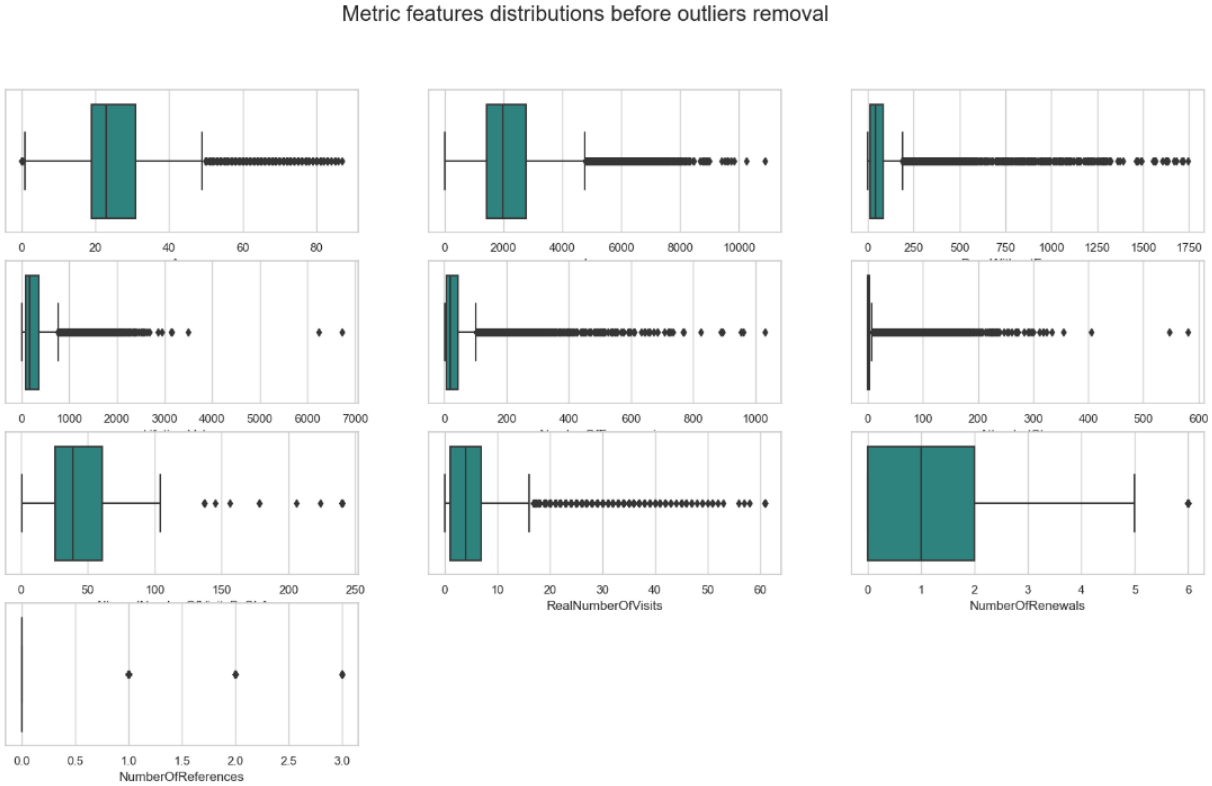

Figure 7

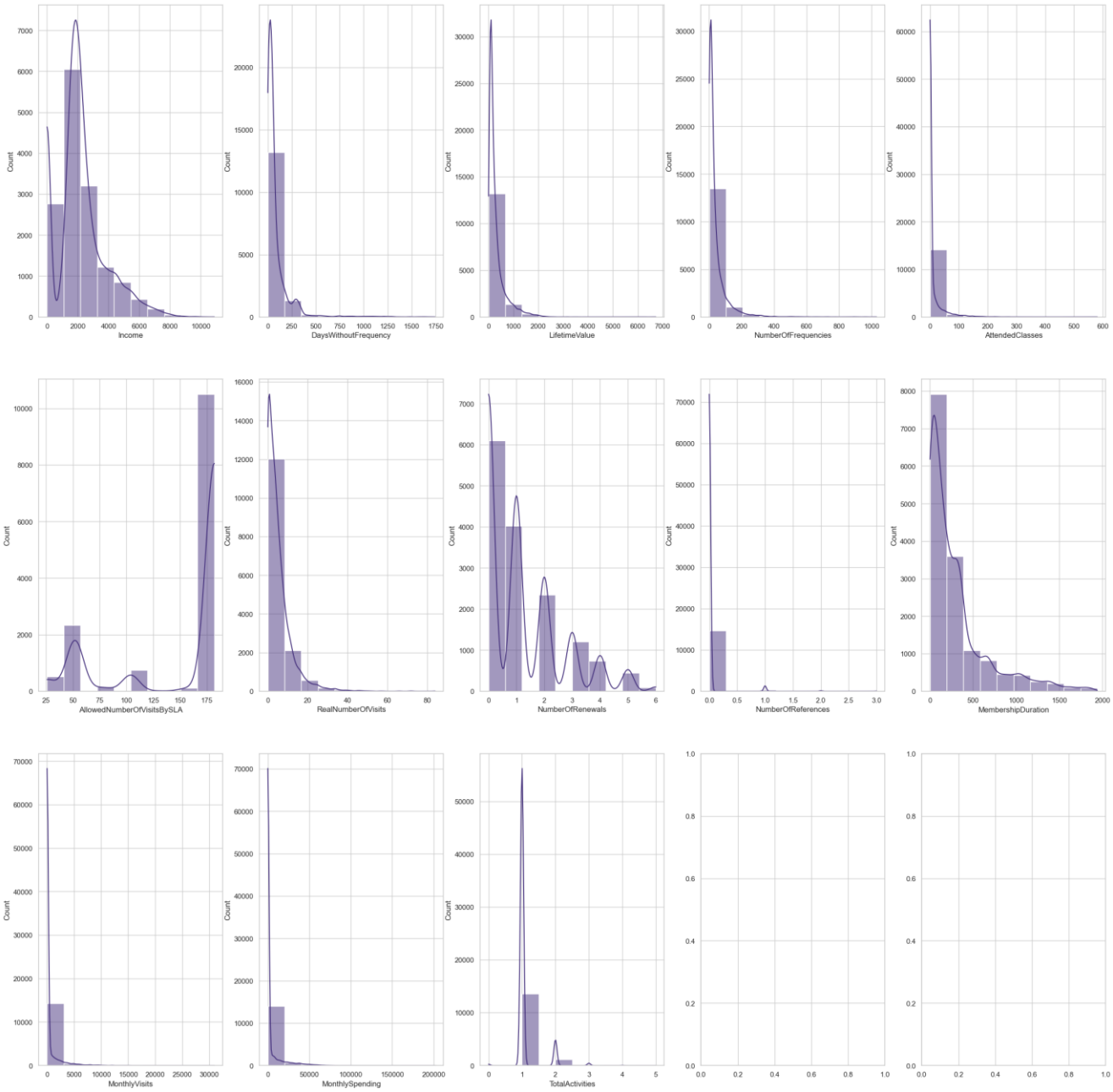

Figure 8: Distribution of metric features before transformations

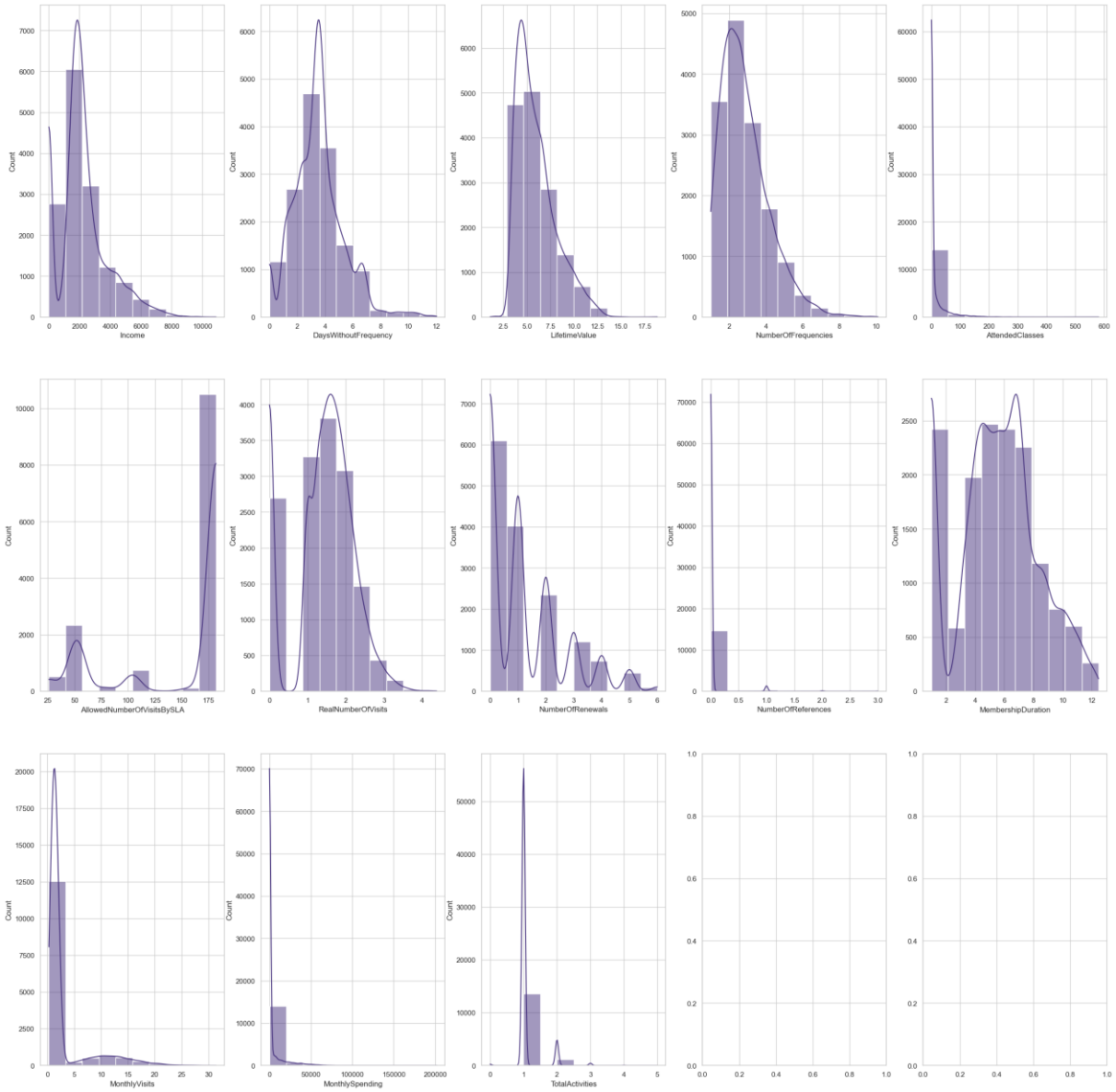

Figure 9: Distribution of metric features after transformations

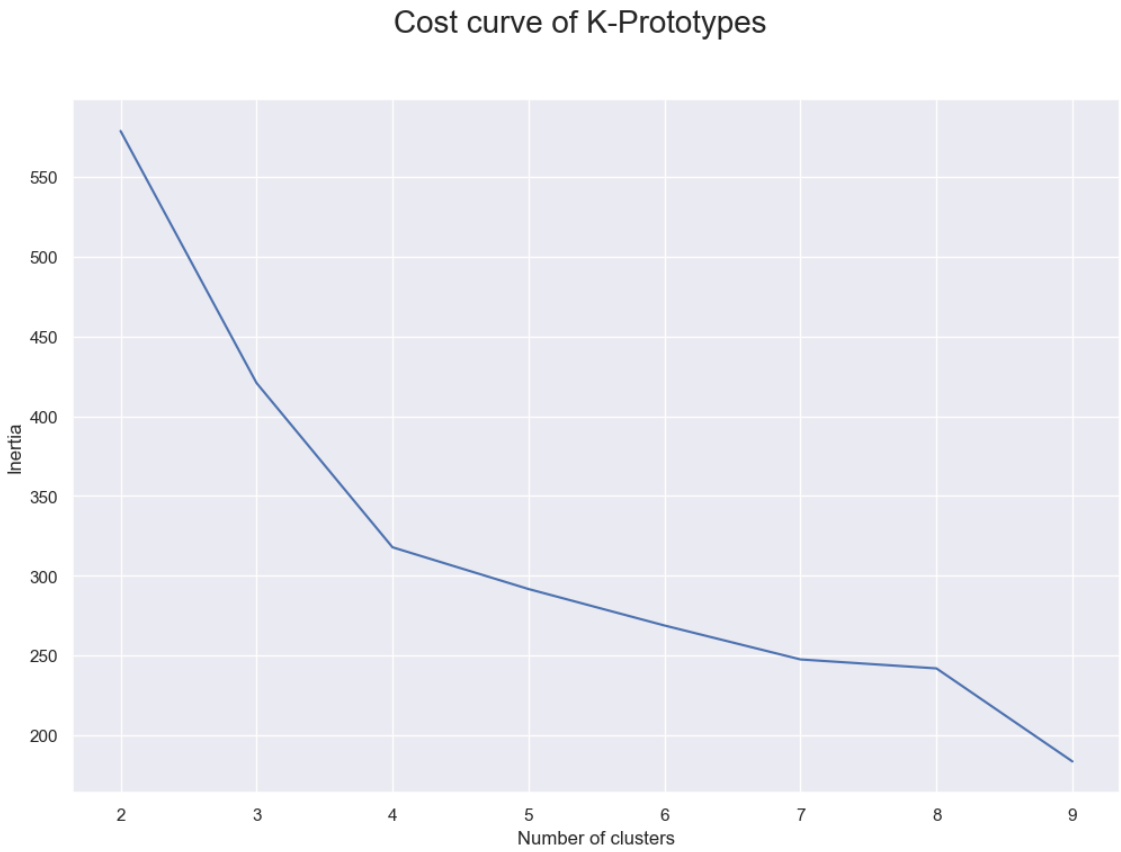

Figure 10: Cost curve of K-Prototypes for Socio-Demo perspective

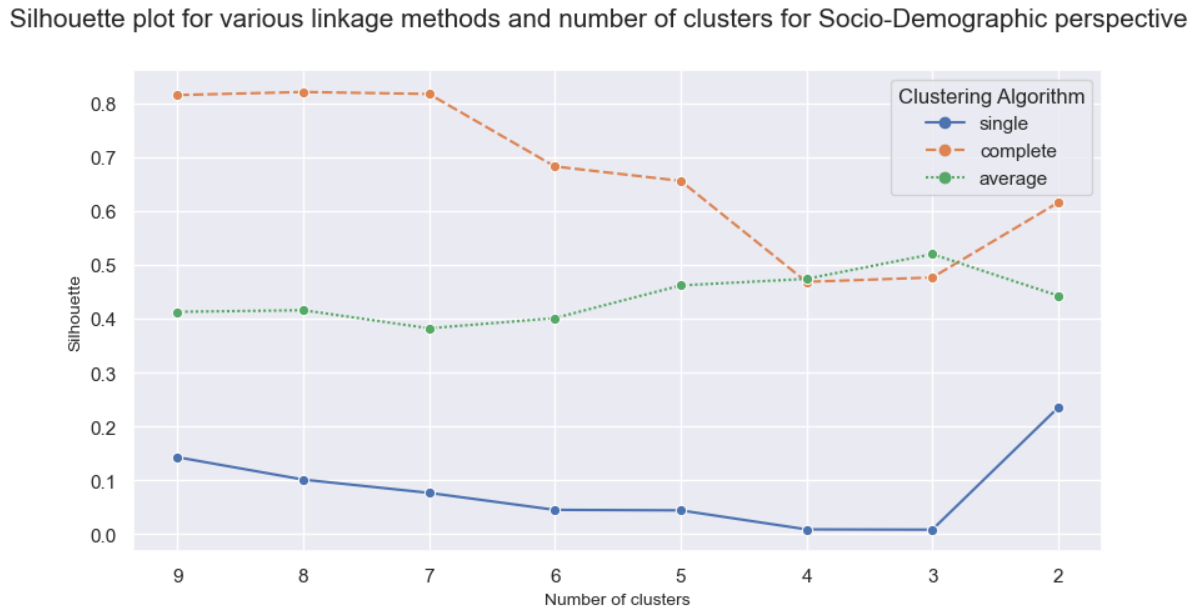

Figure 11

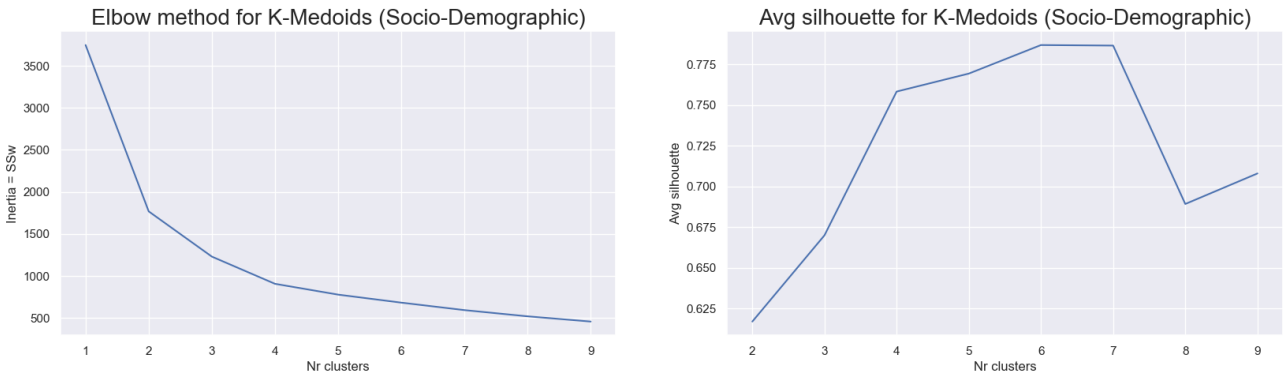

Figure 12

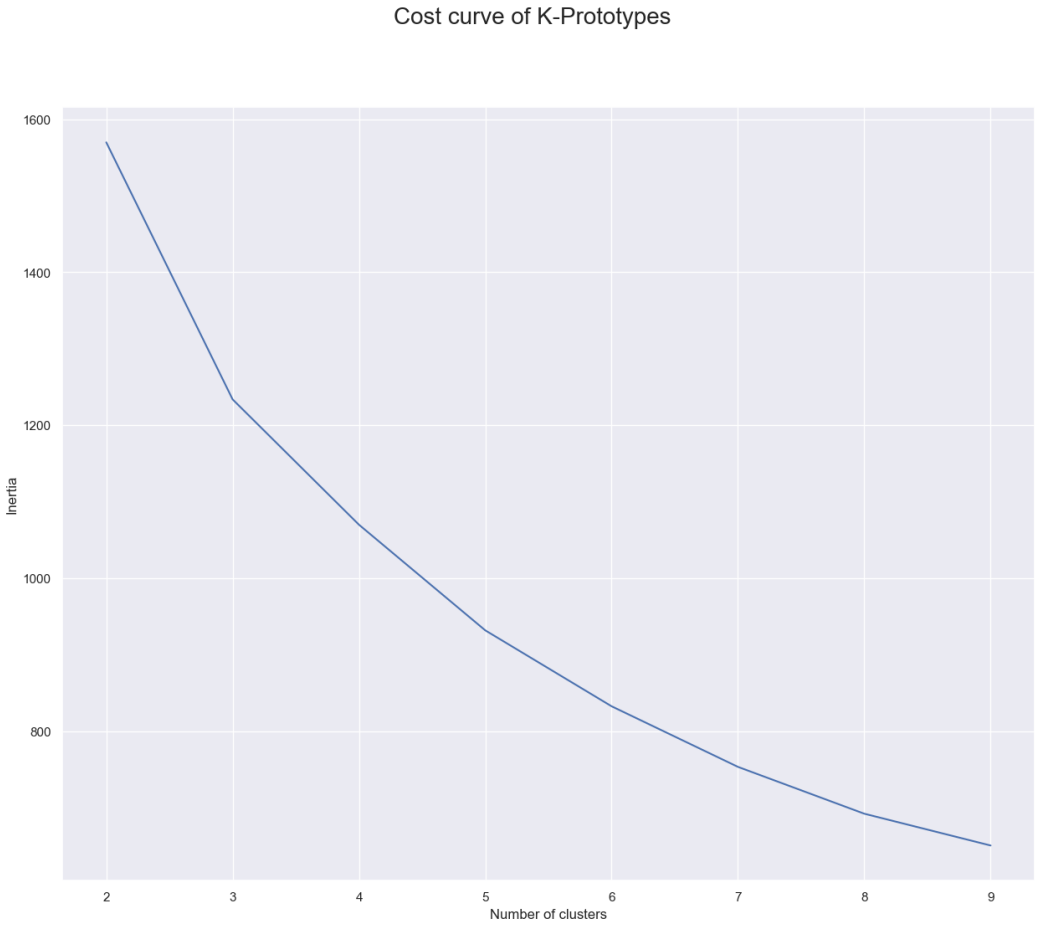

Figure 13: Cost curve of K-Prototypes for Value perspective

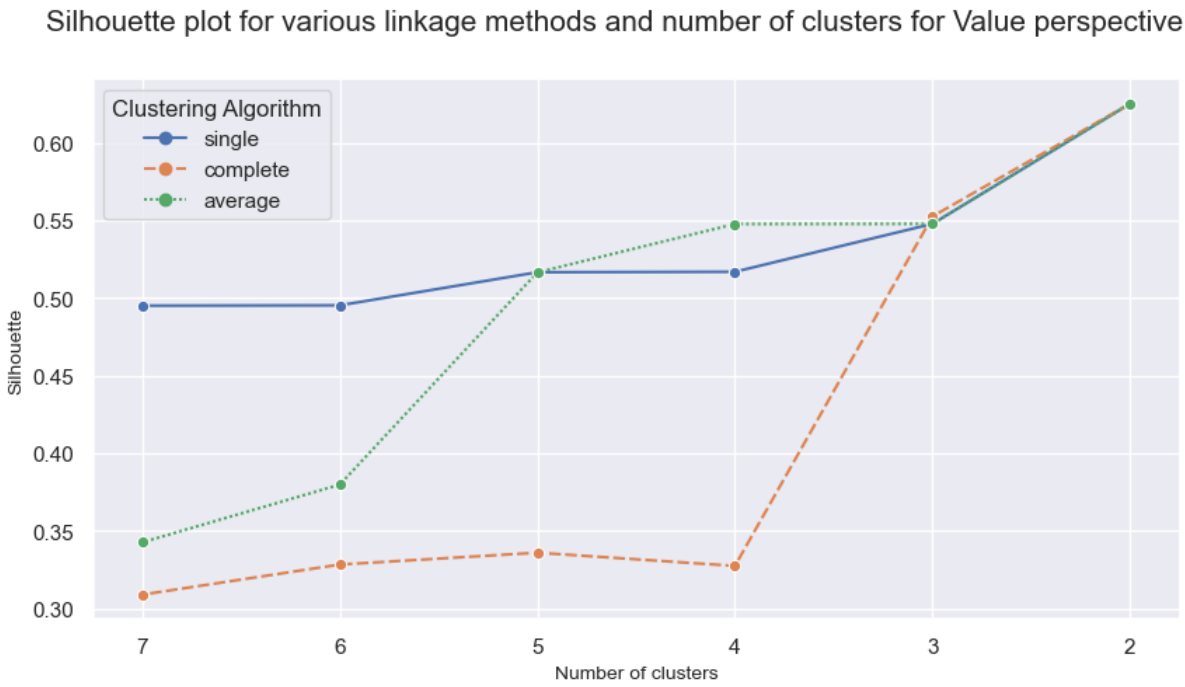

Figure 14

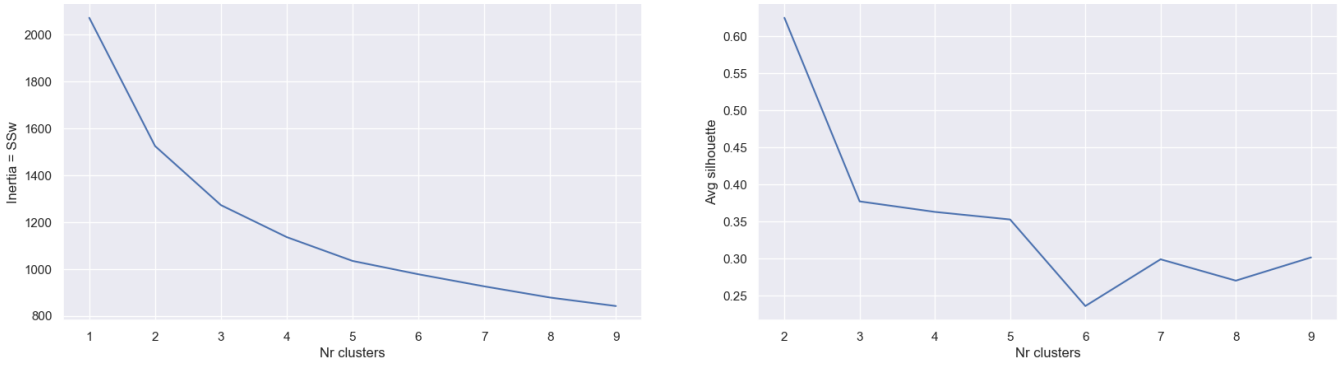

Figure 15: elbow method for K-Medoids(left) value perspective; Avg silhouette for K-Medoids(right) value perspective

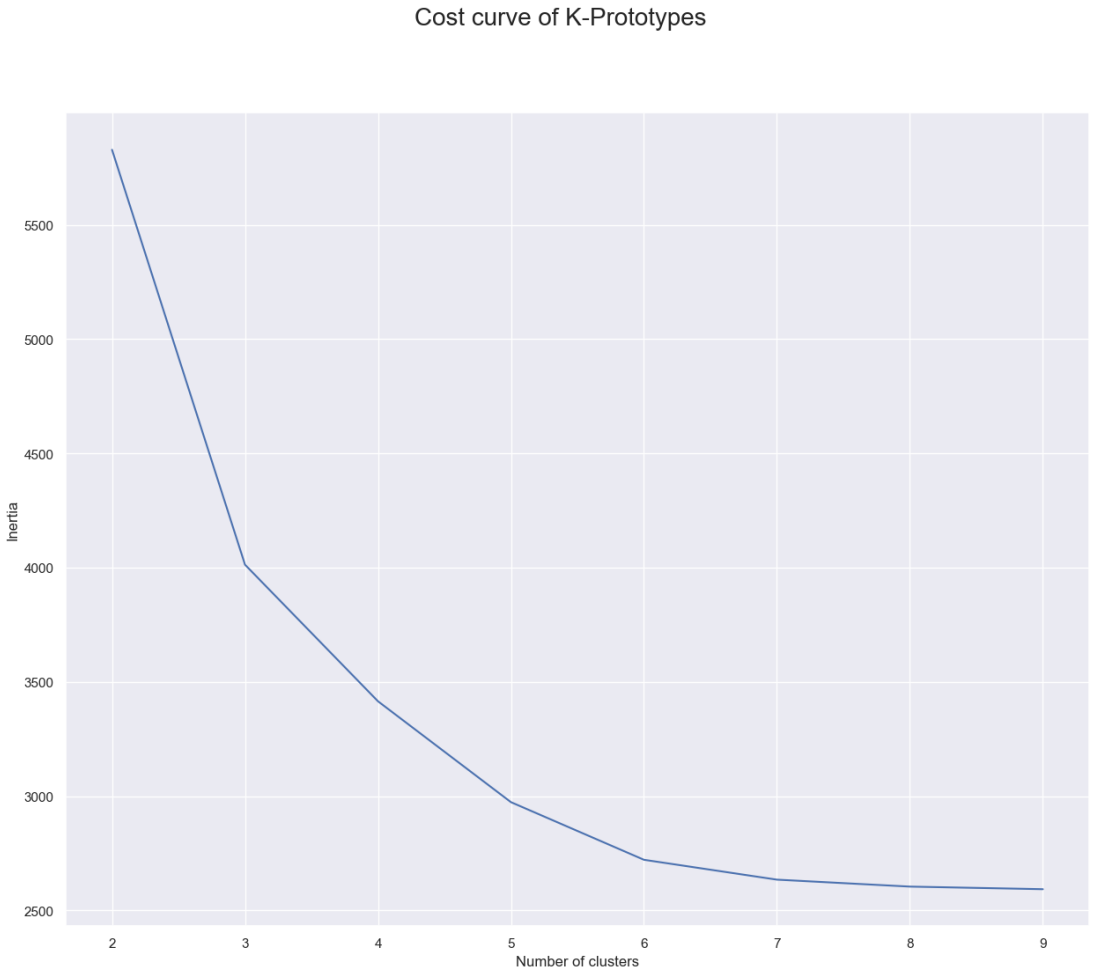

Figure 16: Cost curve of K-Prototypes for Product perspective

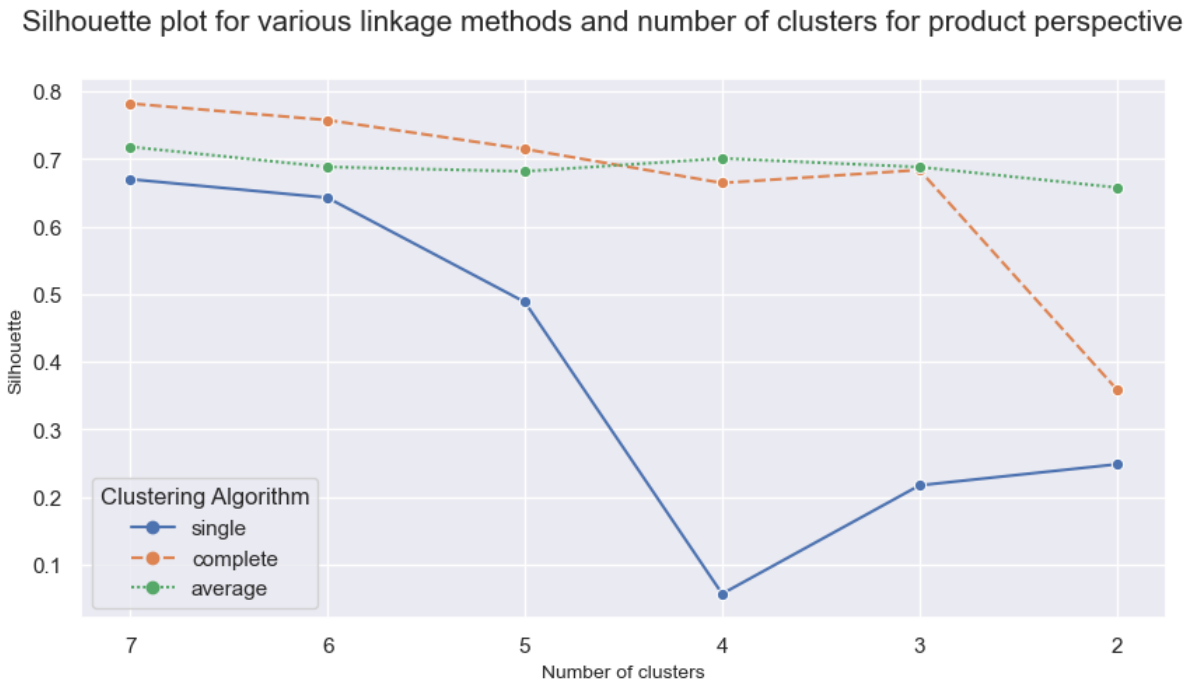

Figure 17

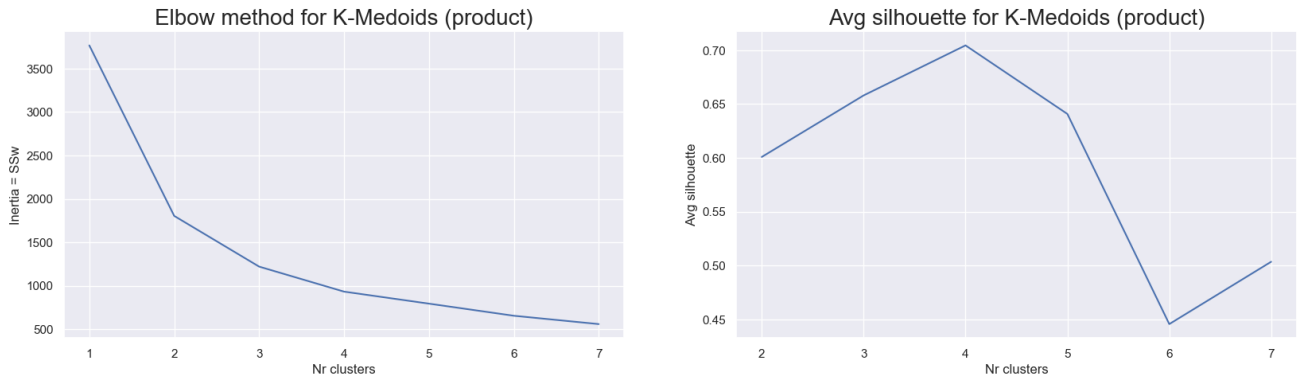

Figure 18

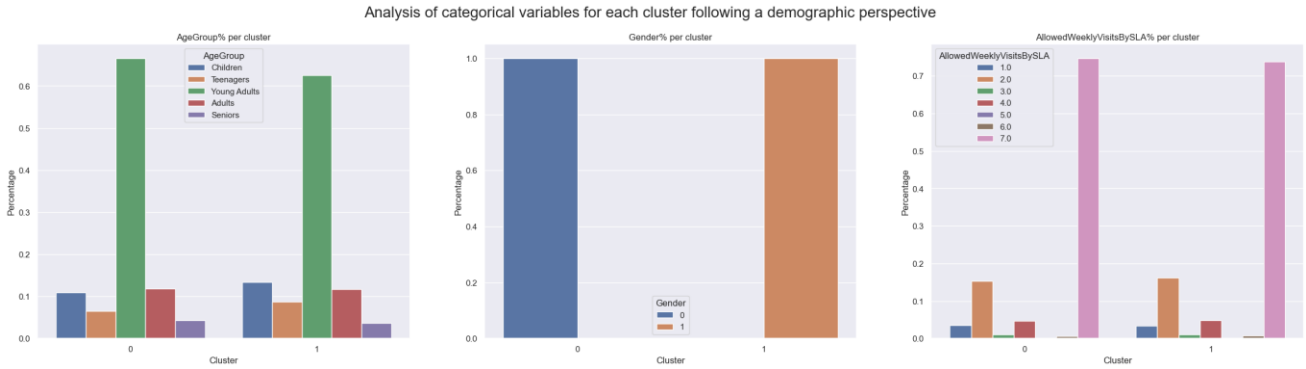

Figure 19

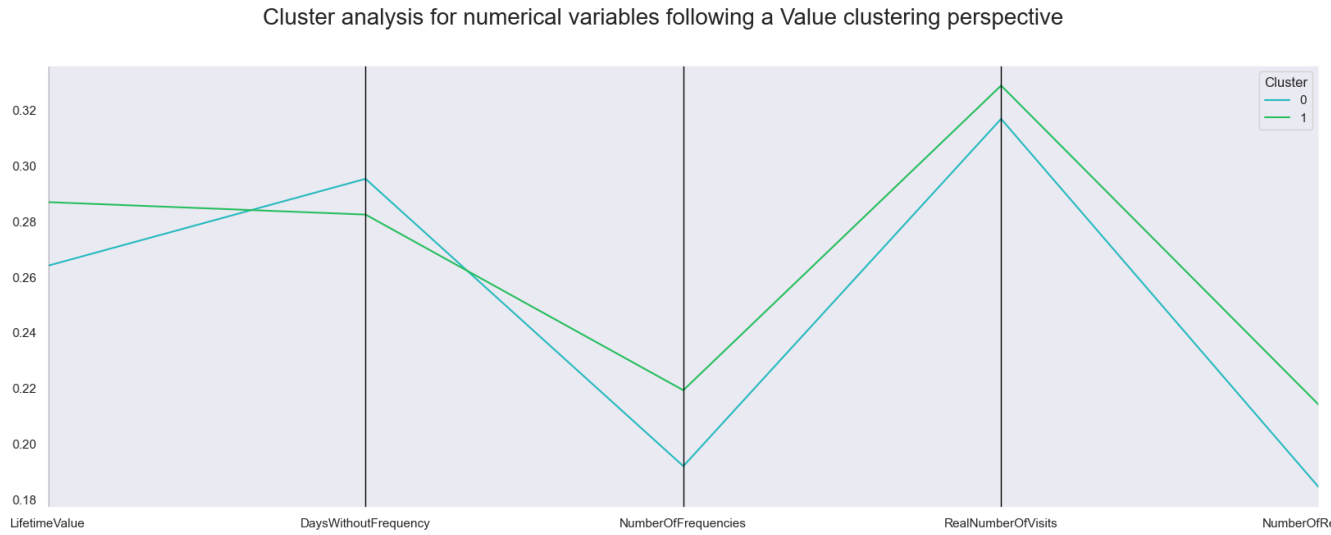

Figure 20

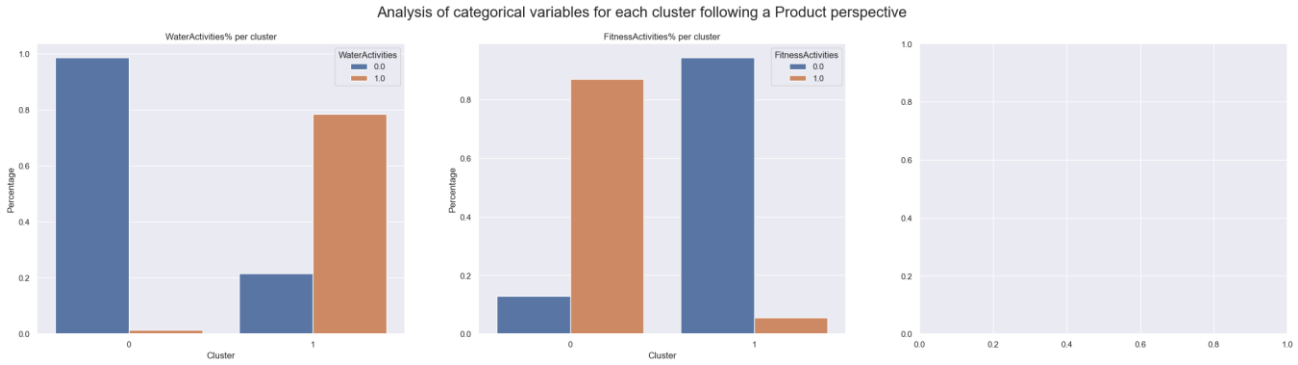

Figure 21

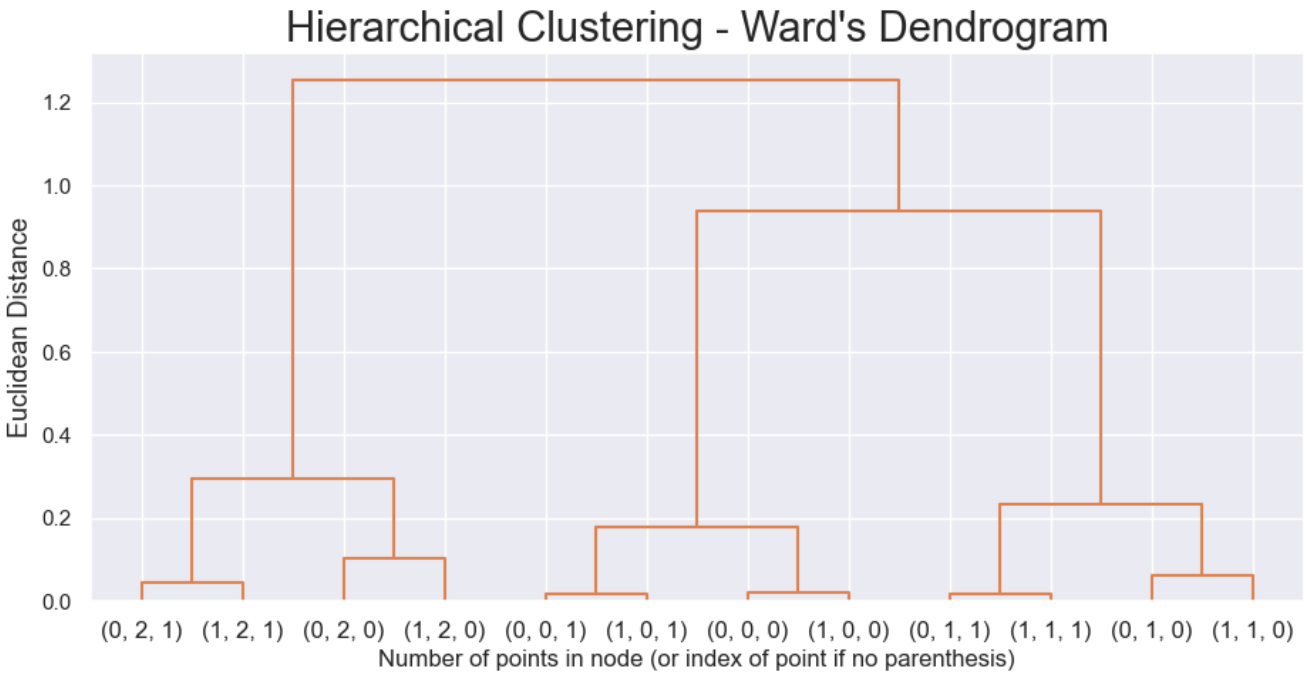

Figure 22: Merging Cluster Perspectives

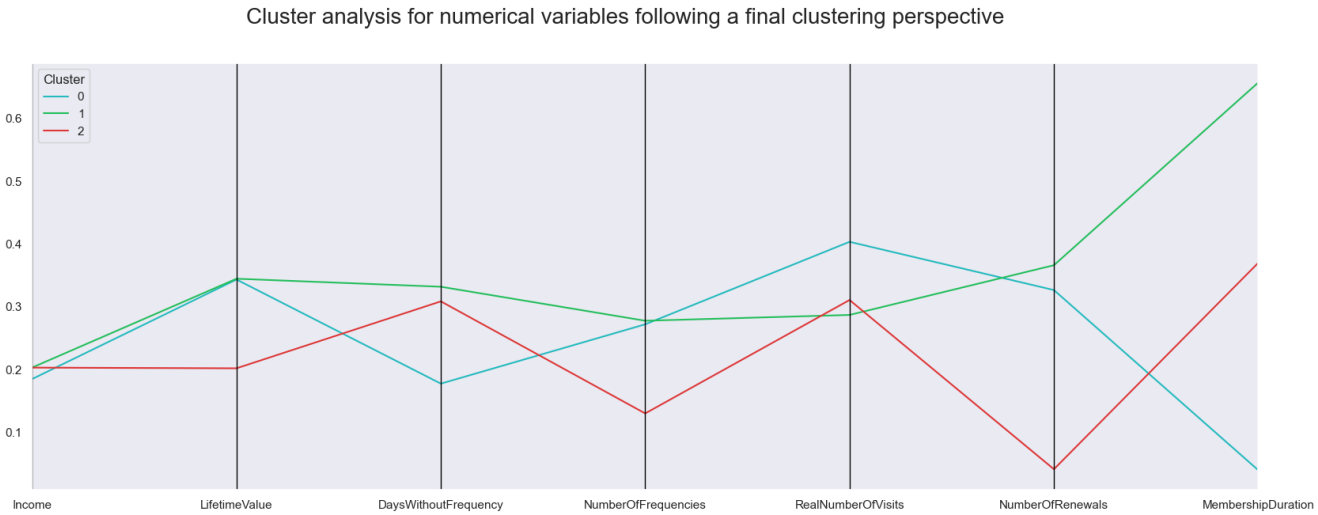

Figure 23

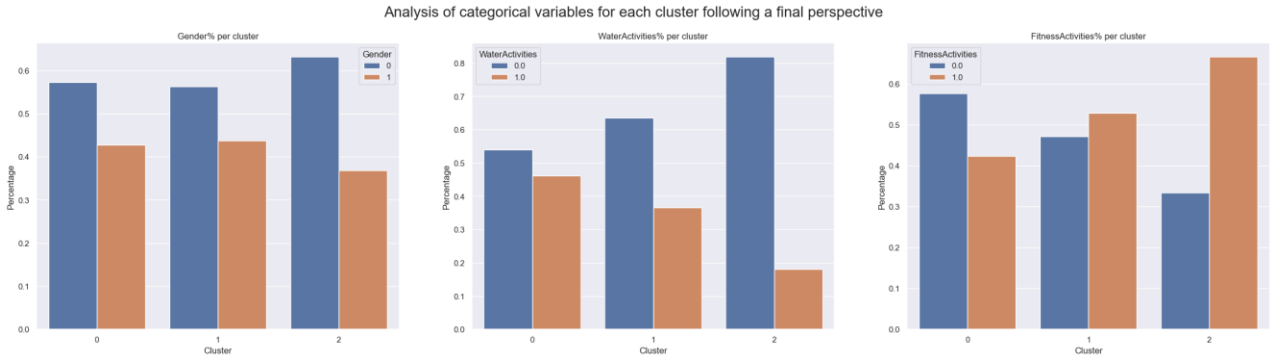

Figure 24

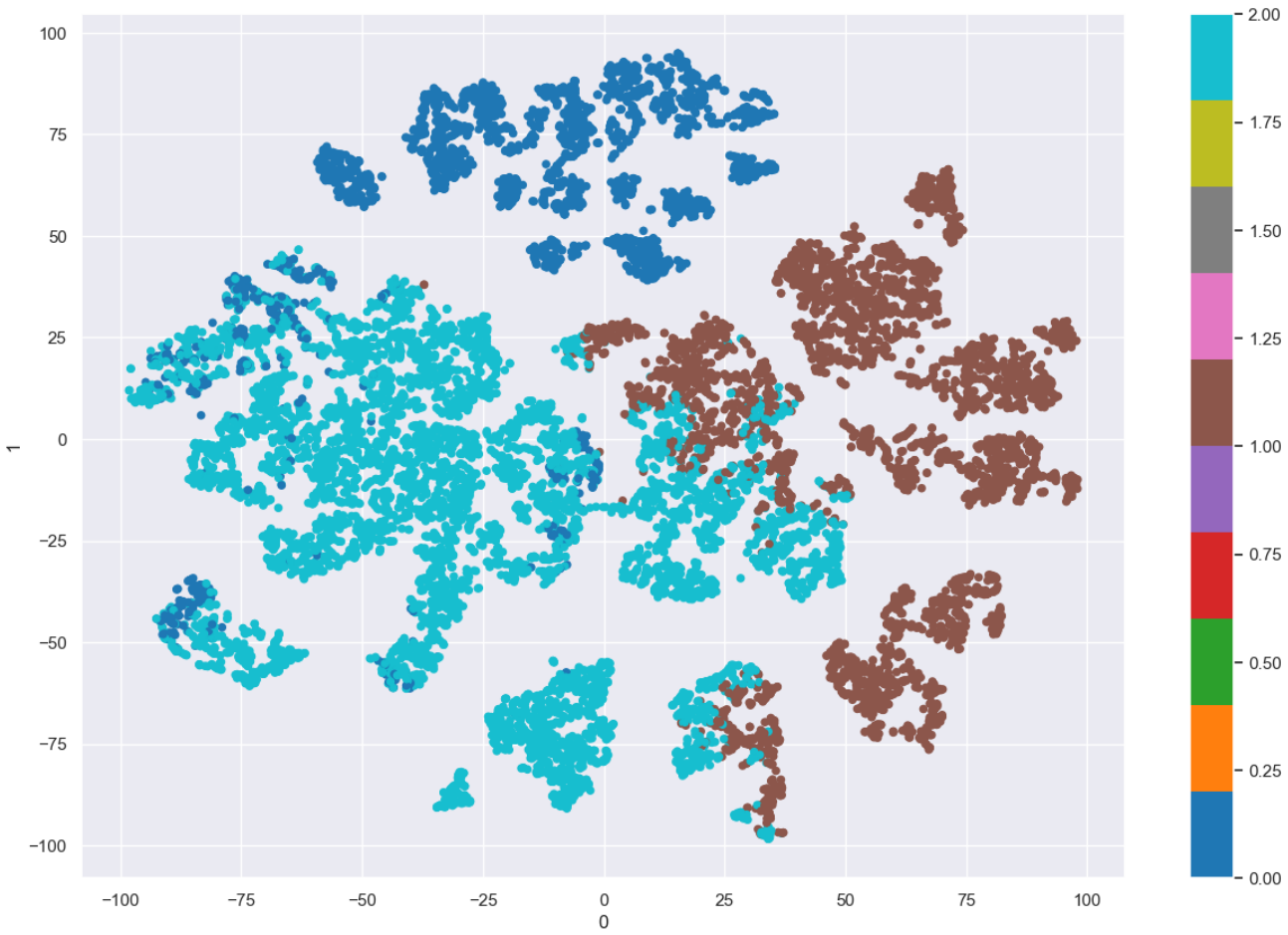

Figure 25: Cluster visualization using t-SNE