Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

CUSTOMER SEGMENTATION

XYZ SPORTS COMPANY

Group 32

Diogo Barros, 20230555

Diogo Roiçado, 20230557

Duarte Mendes, 20230494

January, 2024

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

INDEX

Contents

1. Introduction

In the dynamic world of sports and fitness, understanding and catering to customer needs is not just an advantage, but a necessity. XYZ Sports Company has established itself as a pivotal member of the fitness community, continually seeking to enhance its services and deepen customer relationships. The cornerstone of achieving this lies in a profound understanding of the existing customer base. By learning from the rich data of current customers, organisations can pave the way for innovative strategies and improved customer targeting.

This project, "Customer Segmentation Strategy for XYZ Sports Company", aims to dissect and analyse the wealth of customer-related data available through the company's ERP system. Covering data collected from June 1st, 2014, to October 31st, 2019, our objective is to segment the customer base into distinct groups characterised by their unique attributes and behaviours. This segmentation will enable the company to deliver more personalised services, sharpen its marketing strategies, and ultimately foster a deeper engagement with its community.

By employing a variety of data mining techniques and embracing a multidimensional view of customer attributes — including demographic and behavioural factors — we aim to reveal insightful customer segments. These segments will serve as a blueprint for tailoring services and unveiling new opportunities in the sports and fitness industry. Our journey through data exploration, customer understanding, and strategic implementation will lay down the foundation for a data-driven approach to customer relationship management at XYZ Sports Company.

2. Data Exploration

For this project, we were given a detailed dataset that included both encounter-level data, such as a unique customer ID and the personal information of each customer, that was provided at the beginning of the project.

The starting point of our work was to check the data type of the features, concluding that most of them were numerical, with a total of 31 features and 14942 records. As the unique identifier was *ID,* we defined it as the index of the dataset and then proceeded to search for duplicate entries before any transformation: one was found, which was immediately removed. We knew that we should keep searching for further inconsistencies and aspects to deal with in the data. We continued to explore the dataset by plotting the proportion of missing values per feature (Figure 1). We also noticed that most of the features with missing values were either metric or binary features, which was taken in mind for the next phases of our work.

In line with this previous conclusion, our procedure was to divide and define the existing features into three categories: metric features, non-metric features and binary features. After this division, we decided to explore both metric and binary features using the *describe()* method, in order to keep track of potential outliers. We were able to see that some variables showed a very high percentage of '0' values, and decided to handle them in a latter step, due to their need of a deeper analysis and interpretation. It also came to our attention that *AllowedNumberOfVisitsBySLA* shouldn't be float as it is, but integer (by our interpretation of the variable), so we rounded it down (e.g. 8.9 becomes 8). For example, someone can come to the facilities 8 times in a specific week if they are allowed to do so 8.9 times. The main goal of using the *describe()* method was to find differences between the mean and median, and also if the maximum or minimum value in any feature was too high or too low, respectively. As so, plotting the histograms of binary and metric features (Figure 2), and the boxplots for metric ones (Figure 3), became relevant to confirm and strengthen our understanding of the data.

Having completed the initial data exploration phase, it was of great importance to look for inconsistencies or incoherences within the data. First, we noticed that there were children in the dataset, which is normal since everyone can do certain physical activities, but in what refers to the *Income* feature, we further noticed that some children had income associated. This can happen in some rare cases when a person is 16 or 17 years old, but not entirely for 4 or 5 years old, so we decided to have a further look. We then figured out that some parents filled the enrolment form of their children with their own income, leading to 200 incoherent *Income* records. Having this in mind, our decision was to assume that any children with less than 18 years old should have 0 income. We also noticed that 2422 of the records had the *EnrollmentStart* and *EnrollmentFinish* in the same day, which was strange at the beginning. An in-depth analysis was carried out to analyse those values, which led us to notice that those records were the ones that also have *Dropout* has 0. This finding would mean that these are the still-enrolled customers, making sense with the fact that there is no *EnrollmentFinish* record for them. We also realised that there were three users that visited the facilities during their enrolment but didn't pay anything. We figured, since there was no correlation among those, that they own something like a giveaway. Continuing with the exploration, there were also 481 records of clients that made more visits to the facility during the period between *LastPeriodStart* and *LastPeriodFinish*, than during the period between *EnrollmentStart* and *EnrollmentFinish*. This shouldn't be possible because clients are just allowed to visit the facilities during their enrolment period, we took it into consideration for the rest of the project.

Another incoherence found was that there were 48 clients that made more visits to the facilities between *LastPeriodStart* and *LastPeriodFinish* than they were allowed to in *AllowedNumberOfVisitsBySLA*. There were a lot of inconsistencies regarding visits to the space because there were 208 occurrences of *DateLastVisit* being after the *LastPeriodFinish* which isn't supposed to be possible. We were also able to locate records of customers with *HasReferences* = 1 and *NumberOfReferences* = 0: since there is no sense in this, we decided to transform *HasReferences* to 0 regarding *NumberOfReferences*, since it is the second that determines the first variable.

Having gathered all the inconsistencies present in the data, we opted by filtering the data to only remove records where two or more inconsistencies were detected, for the sake of keeping most of the provided data. This meant that a total of 18 rows (corresponding to a removal of 0.11% of data) were detected and so removed, due to the lack of truthfulness compared to the rest of the records.

At the end of our process of data exploration, a correlation matrix was created with the purpose of understanding which features could possibly be dropped because of high correlation (Figure 4 - *Age* with *Income*, *HasReferences* with *NumberOfReferences*, …). However, we would need more insights from the data before making such decisions.

3. Data Preprocessing and Preparation

After further exploration of the dataset, we emerged to the preparation and cleaning of the provided data, to assess meaningful insights that can be relevant to the subsequent steps of our project. The Data Preprocessing step became a crucial phase of our project, for its huge relevance in the correct analysis and output extraction of the dataset.

3.1 Missing Values

The correct identification and handling of all the missing records within our dataset (missing value imputation) is vital in the Data Preprocessing step, and an adequate imputation method needed to be performed after an analysis of each feature, to not lose the information and relevancy of it. Firstly, we saw that *Age* and *Income* were highly correlated, so we decided to fill the missing values in Income using Linear Regression for customers that have an *Age* higher than 18 (Figure 5), and for the remaining ones we filled with zeros (it was not supposed to kids to have income). The same Linear Regression method was applied for features such as *AllowedNumberOfVisitsBySLA* and *AllowedWeeklyVisitsBySLA*, as their correlation was 0.67. *NumberOfFrequencie*s revealed a skewed right distribution, which led us to replace the missing values with the median. From our analysis of missing values of the *Activities* features, it became very clear that missing values here meant that the customer was not enrolled in the activity, and probably forgot to fill that space. According to this interpretation, we replaced all the missing values in activities with 0. In the case of *HasReferences,* we noticed that there were 12 missing values related to a *NumberOfReferences > 0*, in which we knew that the respective *HasReferences* had to be 1.

3.2 Visual Exploration and Outliers

Following the visualisation performed previously, we were able to identify both possible univariate outliers/uncommon values. By this time, we had already noticed that there were a lot of outliers from the boxplots in some features like *LifetimeValue* or *Days WithoutFrequency* (Figure 6). For this reason, we decided to perform a deeper analysis to understand which of them were actual outliers, and which ones corresponded to the famous quote “The Bill Gates Effect” (not common but real values). All the values, despite some seeming unlikely to happen, were possible. To shrink the space, we decided to manually remove outliers using a threshold for the features in which putting limits could make sense. After applying it to *DaysWithoutFrequency, LifetimeValue, NumberOfFrequencies, AttendedClasses, AllowedNumberOfVisitsBySLA* and *RealNumberOfVisits* we were able to remove 0.8% of the data. Before the decision of handling them by hand, we also tried the IQR method, but since it was removing 45% of the data (which was too much), we decided not to follow this approach. After this, counting together with the 18 rows removed before, at this point we were left with 0.9909% of the initial data, which was a good amount since we were still going to look for bivariate outliers.

Following this analysis, we decided to pairplot all the numerical variables (Figure 8), without *AllowedWeeklyVisitsBySLA*, *NumberOfRenewal*s and *NumberOfReferences*. As we suspected, there were some bivariate outliers possible to visualise and some high-related features, such as *Age* and *Income,* that we had seen before. Our decision was not to handle these outliers by this time, since we could possibly take care of them using DBSCAN later, and because we could also be losing valuable information for clustering.

3.3 Feature Engineering

During the data exploration and preprocessing steps, we gained a deep understanding of our dataset. We opted to move on to feature engineering and feature creation.

Firstly, we proceeded with applying *get_dummies* in *Gender*, transforming it into *Gender_Male*, a binary variable. To also have information about the activities, we created the variable *TotalActivities*, representing the total number of activities in which each customer was enrolled in.

As it made a lot of sense to the future analyses, we would do we created a variable that shows the entire enrolment period of each customer in days, *EnrollmentTime* using the pandas function *pd.to_datetime* applied in *EnrollmentFinish* subtracted by *EnrollmentStart*. Using the same logic, we also created *LastPeriodTime* from *LastPeriodFinish* and *LastPeriodStart.* Further exploring this last feature, we concluded that all the *LastPeriodTime* records would all represent semesters, assuming the form of either one, two or three semesters. So, with the help of *LastPeriodTime*, we created *ContractPeriod,* which represented the total of semesters for which each person was enrolled in on its last contract. The big assumption here was that once a customer enrolled themselves into the company, he never had a semester where he was never enrolled. Additionally, as we thought it would make sense for the segmentation process, we created a feature that could represent the number of monthly visits to the facilities by the customer, related to the entire enrolment period as *EnrollmentMonthlyVisits.* It consists in multiplying by 30 (days) the total number of visits of each customer between their *EnrollmentStart* and *EnrollmentFinish*, and then dividing it by their *EnrollmentTime* (in case the customer didn't have a contract anymore), or by the number of days between their *EnrollmentStart* and the last day of collected data - 31/10/2019 (in case the customer is still enrolled). By using this approach, we were able to solve the underlying issue with instances that have the same dates on both *EnrollmentStart* and *EnrollmentFinish*.

We also wanted to have a feature that could represent the number of customers still enrolled in XYZ Sports at the end of the analysed period, which led us to *StillEnrolled* (a better version of *Dropout*): it was made basically from the customers with *EnrollmentStart = EnrollmentFinish* and *Dropout = 0* that got the value 1, 0 otherwise.

4. Feature Selection

It is relevant for the business to know how well a specific topic has developed overtime, and XYZ SPORTS COMPANY is not an exception. For us to understand when the months were where more customers used to join the company and specific activities, we decided to plot that respective graphic over time. This plot became especially relevant to understand the peak times, that are, most of the time for almost every feature, in September - October and March - April (Figure 8). The decrease in the number of people enrolled in each activity was also noticeable in the graphs, with some cases of activities that stopped to have people enrolled, like *RacketActivities* and *OtherActivities* (Figure 9). This conclusion could have induced us to think about deleting these features, but we also thought that it could be possible, as a marketing strategy, to take them back, and for this reason we proceeded with their analysis.

As a result of the engineering done before, we decided to drop some of the variables that were already transformed into new ones. Having this in mind, we dropped *EnrollmentStart, EnrollmentFinish, LastPeriodStart, LastPeriodFinish*.

Filter methods also became relevant to our feature selection analysis, and by putting it in practice it allowed us to gain some traction on the extraction of meaningful data insights from the variables, while selecting the most relevant features to perform our clustering approach. Our decision was to first explore relevant statistics from the structure of our data, using variable variance analysis for that purpose. We were able to verify which variables did not have such an informative character. *NatureActivities* and *DanceActivites* had a variance of 0 and *AthleticsActivities, RacketActivities* and *OtherActivities* 0.0843, 0.1515 and 0.0427, respectively. As they contain insignificant information because they're predominantly populated with zeros, we could confidently remove them. There were other features with very low variance but based on all the information we got since here, we decided to only remove those.

After the creation of the new variables, it was important to understand if we created any redundant or irrelevant features, so we displayed again the new correlation matrix (Figure 10). As it was expected, *EnrollmentTime, TotalSemesters, LifetimeValue* and *NumberOfFrequencies* were highly related with more than two other features, which was normal based on the definition and on normal customer behaviour. As a result of the previous imputation in *HasReferences* based on *NumberOfReferences* and a previously high correlation between them we figured that *HasReferences* was giving redundancy to the data, so we dropped it. The same happened with the *TotalSemester* previously created as it was well represented by *EnrollmentTime. StillEnrolled* was created to fix *Dropout*, which was dropped too. Regarding the feature *AllowedWeeklyVisitsBySLA,* more than just being high correlated with *AllowedNumberOfVisitsBySLA,* we understood, by the interpretation of both, that they can introduce a bit of redundancy in the data since the first one represents the same as the second one but weekly. So, we dropped *AllowedWeeklyVisitsBySLA.* Finally, as it was mentioned before (FIGURE 5) both *Age* and *Income* were highly correlated. Even though *Income* has a bigger domain as it represents a bigger amount of data, we figured that because of the demographic interpretation of *Age* being more relevant for marketing purposes than *Income,* we should drop *Income.*

4.1 Data Standardization

Before applying any cluster techniques or algorithms, the processes of scaling, standardising or normalising the data and respective metric features arises as of utmost importance. Since our dataset had features on different scales, those with larger ones would dominate the distance metric and potentially lead to misleading results. Our decision fell on applying *StandardScaler* on the metric features.

5. Clustering Application

In the next phase, our approach involved employing a range of clustering techniques to capture the multifaceted nature of customer data. These techniques helped us to identify homogeneous groups within the heterogeneous market, enabling XYZ Sports Company to tailor its services and marketing approaches more effectively. Each customer segment was uniquely analysed to understand each value proposition and demographic makeup, as well as the sports activities that most resonate with them.

We decided to initially apply clustering to the entire dataset to gain a comprehensive overview of customer segments. This approach allows us to capture broad patterns and potentially uncover unexpected segments without bias. After establishing these initial clusters, we proceed to a more focused segmentation, refining our strategy based on the insights gained.

5.1 DBSCAN

The first clustering method applied was DBSCAN: our purpose was to use a method that could handle and identify possible noise within the data. We concluded, using 4 clusters, as it was predicted using the bandwidth value (Figure 11) that there were 698 noisy points identified. The respective R^2 score was of 0.3313, which means that approximately 32% of the data's variability was explained by the clustering model. This value can be considered low, suggesting that while there is some structure of the data captured by the clusters, still a significant portion of variability remains unexplained. The identified noise points did not provide that much insight about their characteristics but, after the application of UMAP visualisation (Figure 12), the visualisation showed us one main cluster, suggesting a high degree of similarity among the data points. This clustering method provided us with an understanding of how the noise points are distributed. However, since we wanted to try other different clustering techniques, we followed with our clustering analysis.

5.2 K-means

Our next clustering approach consisted of applying the k-means algorithm, considering a range of 1 to 10 clusters. The results were evaluated based on the R^2 metric (Figure 13). The k-means did not present a clear elbow in the corresponding plot. This ambiguity prompted further investigation into cluster validity. We turned our attention to silhouette analysis, plotting the silhouette graph and scores (Figure 14). This process suggested that a two-cluster solution had the highest score but, as we can see in the plot (Figure 15) there are negative values on cluster 1. This indicates that some data points are, on average, closer to members of the cluster 0. Based on that visualisation, we could conclude that this might not be the best representation of the underlying data structure. We also tried to analyse the inertia plot (Figure 16), and even after adjusting the random state parameter a few times, a distinct elbow was still not perceptible.

5.2.1 K-Medoids

To address the limitations of K-Means's sensitivity to outliers and non-metric features. We decided to use K-Medoids, robust to noise and applicable to various types of data. It chooses actual data points as clusters centres, offering meaningful representations. Using the Inertia plot, we could easily figure an elbow (Figure 17). As we could tell by the UMAP associated with a 4-cluster decision, the clusters were not very well defined (Figure 18) and since there wasn't much to hyper tune, we kept going.

5.2.2 K-Prototypes

To analyse the impact of the binary features we employed K-prototypes. By leveraging this algorithm, we aimed to uncover meaningful patterns and relationships, potentially revealing how specific combinations of binary features characterise different groups. We applied the method with init as *Cao,* an initialization method specifically designed for the K-Prototypes algorithms to handle categorical data effectively. In the average silhouette plot (Figure 19) we could see that 3 and 4 clusters would be a good choice to analyse, so we decided to study both. After applying the method, we added the labels column to the dataset and group the data frame by the cluster labels and then calculated the average of all the other features for each cluster. We could see some correlation between the binary features of the cluster 0 and 2 and for 1 and 2, which could indicate that the behaviour of the people associated to those clusters could be very similar in terms of those features (Figure 20 and Figure 21), for example for *FitnessActivities* that had values of 0.72 and 0.63 for clusters 0 and 2, respectively. While for cluster one it was 0.07. This indicates that it is an important feature for predicting customer behaviour, cluster 0 and 2 might be approached similarly by a marketing campaign that leverages customer *FitnessActivities*. The same analysis was done for 4 clusters and the same interpretation was taken (Figure 22). In practical terms, such insights could be leveraged for more targeted strategies, as we kept in mind for the next phase.

5.3 Self-Organizing Maps

Self-Organizing Maps (SOMs) are useful for exploring the structure of the data. Our idea was to apply SOM as a pre-processed step before applying clustering algorithms. The reduction made by SOM can improve the efficiency and effectiveness of the algorithms.

With this in mind, we built the SOM with the weights assigned as random and with a grid of 50x50 as we figured it would be a good approach since it can capture a lot of detail in the data. We will use this grid to do the evaluation on the following steps.

5.3.1 K-Means on top of SOM

Employing K-Means on top of Self-Organizing Maps, merges the strengths of both techniques. SOMs organise complex data into a structured grid. Applying K-Means to SOM units refines clustering, leveraging SOM's structure preservation and K-Means' precision. To decide the best number of features we've not only observed the inertia of the SOM, but we've plotted the structure of the SOMs for the different range of clusters and visualised what number of clusters would have been a good choice.

We decided to proceed with 5 clusters, because visually there is a good separation between clusters and the inertia plot revealed an elbow on that point. There were two major clusters (performing almost 72% of records) but that doesn't seem like a bad solution when compared to the remaining ones.

To generalise this analysis to the full dataset we needed to check for the best matching units and cross them with all data. The final R^2 for this solution is 0.38. Some of the relevant interpretations we took from here were related with the adolescents being the ones who attend more classes, the recent customers (low renewals) intend to show no interest into attending classes that are only related to *FitnessActivities* and a similar behaviour among the customers that didn't came for a long time.

5.3.2 Hierarchical clustering on top of SOM

To perform the SOM with the Hierarchical Clustering we will need to test what is the optimal dimension of the clustering, so we will get the best R^2 for the multiple methods applied on the SOM grid. The ward linkage metric was the best method to proceed (Figure 23). The plot of the clustering differs a bit when compared with the K-Means algorithm since we always have one major cluster. According to both methods we were uncertain about the number to choose, the doubts were between 3 and 5 (Figure 24). We've tested both approaches and none of them was any better than the clustering performed by the K-Means on top of SOM. So we decided to not move any further with this analysis.

5.4. Customer Segmentation

Customer segmentation is vital for business, aiming to understand diverse customer needs effectively. By dividing the customer base into clusters sharing similar traits, companies can tailor their strategies and campaigns more precisely. An RFM (Recency, Frequency, Monetary Value) approach was taken to understand the frequency and the reactions of customers to the provided classes. The objective is to make our company closer to its clients, adapting to a better satisfaction.

Our first approach was to split our dataset into demographic, behavioural and product features. For the product we had included the activities offered from our facilities, but we weren't able to do any good clustering, so we decided to merge this approach with the behavioural.

5.4.1 Demographic analysis

For the demographic analysis, at this point we only can include *Age* and *LifetimeValue*, we also added the *UseByTime* and *Gender* variables but they didn't influence the segmentation so we will not consider them for further analysis.

The measure chosen for the optimal number of clusters was once again the R^2 metric. The number of clusters that outlined our model was 6, using the K-Means approach (Figure 25). Even though 3 seemed the best answer, after some exploration and from the insights we had from before, we wanted to separate the best possible relation between *Age* and *LifetimeValue*. Using some visualisation (Figure 26), we were able to see that there is a linear separation of the clusters: two that divide the customers with a high *LifetimeValue* from all ages, and 4 more dividing the *Age* between medium and low *LifetimeValue*.

5.4.2. Product Behavioural analysis

We began by integrating features related to the frequency of customer visits to the facilities, their relationship (on time) with the company and what were the type of activity that the customer preferred to attend (not only the offered activities but if they preferred to be on its own way), we chose features like: *AttendedClasses*, *DaysWithoutFrequency*, *NumberOfFrequencies*, ….

Again, the methods of evaluation of the optimal clusters were applied and the method that ended up being chosen was K-Means (Figure 27). At this time, we weren't sure if we needed to choose 5 or 7 clusters. So, we started by applying K-Means with 5 clusters as it could generalise better to new data. It was observed that roughly 70% of the data was characterised by the average customer, which is normal since it represents the population well. But we had the intention to distribute in a better way for the merging purpose so we could capture finer variations and subgroups within the data, allowing for a more detailed analysis and the opportunity to combine similar clusters into larger, more generalised groups later. As we could see by the plots (Figure 28) there were some clusters characterised by their singularity. Cluster 3 was defined by the customers who attended the most classes, cluster 4 by the customers that dropped out but had a lot of days before doing it without visiting the facilities. Cluster 5 was characterised by the customers who have been enrolled for a short period of time.

5.4.3. Merging perspectives

To have a more robust understanding of our customers we decided to merge these perspectives. While demographic data provides insights into who customers are, product behavioural data reveals how they interact with the services of the facilities. It can show which products a certain age group prefers and if the different genders can influence the type of service. This combined understanding empowers business to tailor marketing strategies and develop targeted products and services.

We started by doing a crosstab table to understand how our dataset was distributed by each label in each perspective (Figure 29). A data Frame of the centroid of each cluster was computed by calculating the distances of each datapoint to the centroid. After applying Hierarchical Clustering, to that data frame containing the centroids, every label was merged to the one that was assigned by the method, as we ended by choosing to stop when we have the total of 6 (Figure 30), as it was visible and in conformity with the previous analysis. A TSNE visualisation was made (Figure 31). To understand the factors impacting this analysis the most, we constructed a decision tree. This tree predicted the clustering performance. Also, we identified the variables with the highest R^2. According to this metric, it's evident that behavioural variables had the greatest impact. Additionally, we observed that *Age* and *Gender* do not significantly influence the segmentation of our enrolled clients. Although *Age* didn't have a high R^2, based on the business knowledge we thought that by doing a segmentation with this variable we could have better insights on how the customers are distributed. The following results were based on (Figure 32).

Our cluster analysis has painted a vivid picture of our customer base. Starting with Cluster 0, these are our long-term members who tend to be older and are the most frequent visitors, a clear sign of their loyalty and the habit of regular visits they've developed over time. Moving to the youthful side, Cluster 3 represents our youngest members who not only participate actively in numerous classes but also contribute significantly to our revenue, likely because they're eager to explore and take advantage of what the company offers.

Then there's Cluster 1, where members engage with a variety of activities, but the lower *LifetimeValue* and visitation rates suggest these offerings might be losing their charm. Cluster 4 stands out as well, with members who have had long memberships but show big gaps in attendance, which eventually leads to them dropping out. It's a reminder for us to investigate re-engaging strategies.

Cluster 5 shows a burst of activity with members who join for short but intensive periods, possibly with specific goals that align with short-term offerings of the company. And lastly, Cluster 2 where customer almost didn't attend any class, they paid almost no money while enrolled, and were also the ones who didn't visit the facilities, displaying a disengagement like Cluster 4.

By understanding these patterns, we can tailor our services to better fit our customers' needs, whether it's nurturing the loyalty of our long-term members or sparking the interest of the less engaged. This detailed view of our customer segments is crucial for shaping a more personalized approach to customer service.

6. Business Strategies and Marketing Approaches

- *School in Shape*: If the customers are still studying in a school, when showing their school card, they can get a 15% discount on the enrol of some activity. They are also allowed to bring two friends that are still studying for a class for free. There is an opportunity to also apply to be an ambassador of XYZ Sports where their social media presence will be use in a different level in exchange of some advantages. Improve their *AllowedNumberOfVisitsBySLA*, since these customers are the ones who attend more classes. – Cluster 3

- *Loyalty Stars*: As a reward for long-term customers, it will be offered a long-term contract of 3 years with a 20% total discount. An invite for the annual dinner of the company. And personalized health plans to keep them engaged and for them to feel appreciated. This would be applied for customers that were enrolled for at least 3years. – Cluster 0

- *Activities Lovers:* For the customers who attend *FitnessAcitivites* a discount of 25% if they also enrol in other activity at the same time. At the same time the company could open a free day of just activities in the entire facilities for all the customers to try out. – Cluster 1

- *2nd Chance*: As we've had numerous customers who previously dropped out, but have recently shown renewed interest in our offerings, now is an opportune moment to re-engage them. A compelling campaign would involve applying a semester-long discount for returning customers or involving an offer designed just for them based on previous insights– Cluster 4

- *Goal-Oriented*: Offer short-term programs geared towards common goals that the customer might have. To keep them interested, the company could provide tools and support towards their goals. – Cluster 5

7. Conclusion

In concluding this project on customer segmentation for XYZ Company, we've discovered important insights about the company's customers. Our careful work in looking at the data, preparing it, and dividing the customers into groups has given us useful segments that show different customer needs and behaviours.

We carefully dealt with any issues in the data, making sure our groups were accurate and understood that customer data can be complex. We chose specific features in the data because we wanted our findings to be useful and practical for XYZ Company. We used different methods to group the customers, which helped us understand the many aspects of the data. By looking at both who the customers are and what they buy, we got a complete view of the customer base. This helps XYZ Company make better marketing plans. Our findings offer specific ideas for marketing, like focusing on younger customers through school programs or using the popularity of fitness activities to build a community. These plans are made for the different groups we found, making XYZ Company's marketing more effective. In summary, this project has given XYZ Company a better understanding of its customers and has provided valuable information that can be used to improve how the company connects with customers and grows its business.

8. References

- *Developer Interface — Kprototypes 0.1.2 Documentation*. [online] Available at: https://kprototypes.readthedocs.io/en/latest/api.html
- Shiledarbaxi, N. (2021). *Comprehensive Guide To K-Medoids Clustering Algorithm*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/comprehensive-guide-to-k-medoids-clustering-algorithm/
- Coenen, Andy, and Adam Pearce. "Understanding UMAP." *Pair-Code.github.io*, https://pair-code.github.io/understanding-umap/
- Tjoonk, Niels. "What Does a T-SNE Plot Show?" *Single Cell Discoveries*, 11 Jan. 2023, https://www.scdiscoveries.com/blog/knowledge/what-does-a-t-sne-plot-show/

Appendix

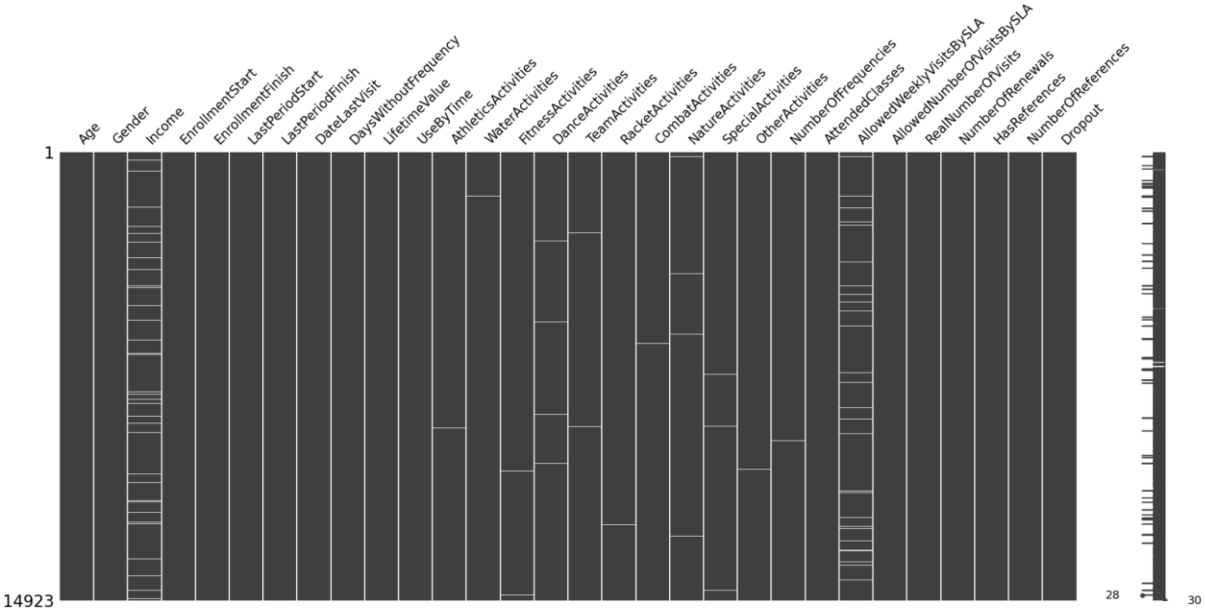


Figure 1 - Missing values on each feature

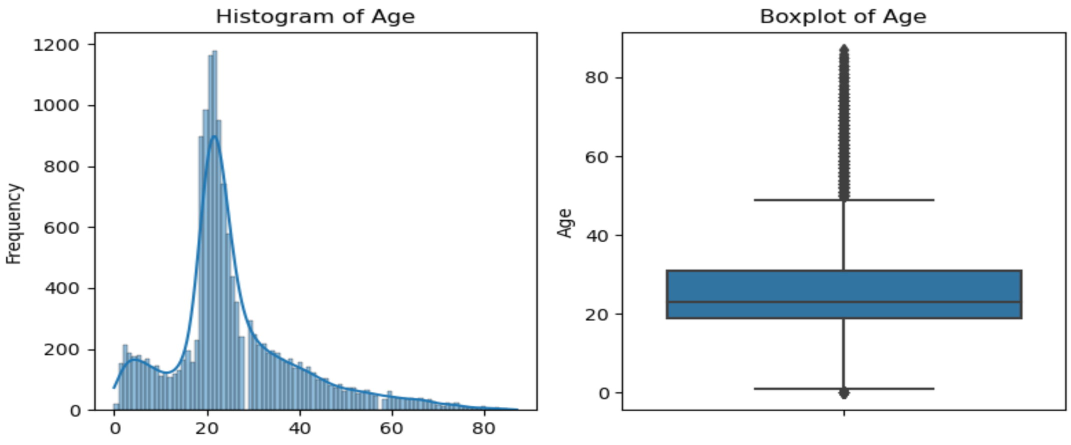


Figure 2 - Example of Histogram and Box-Plot of metric features

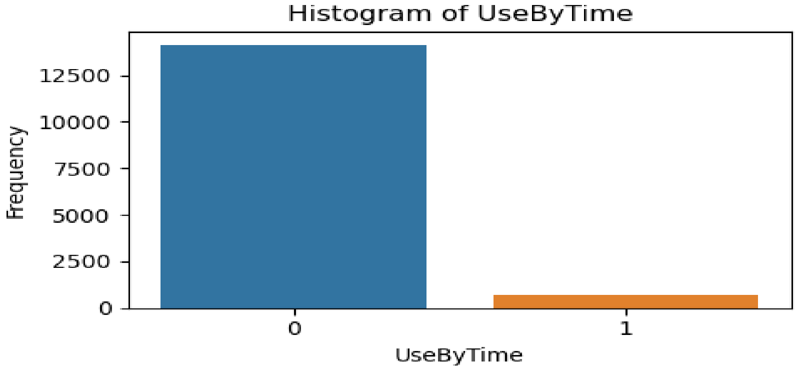


Figure 3 - Histogram Example of a Binary Feature

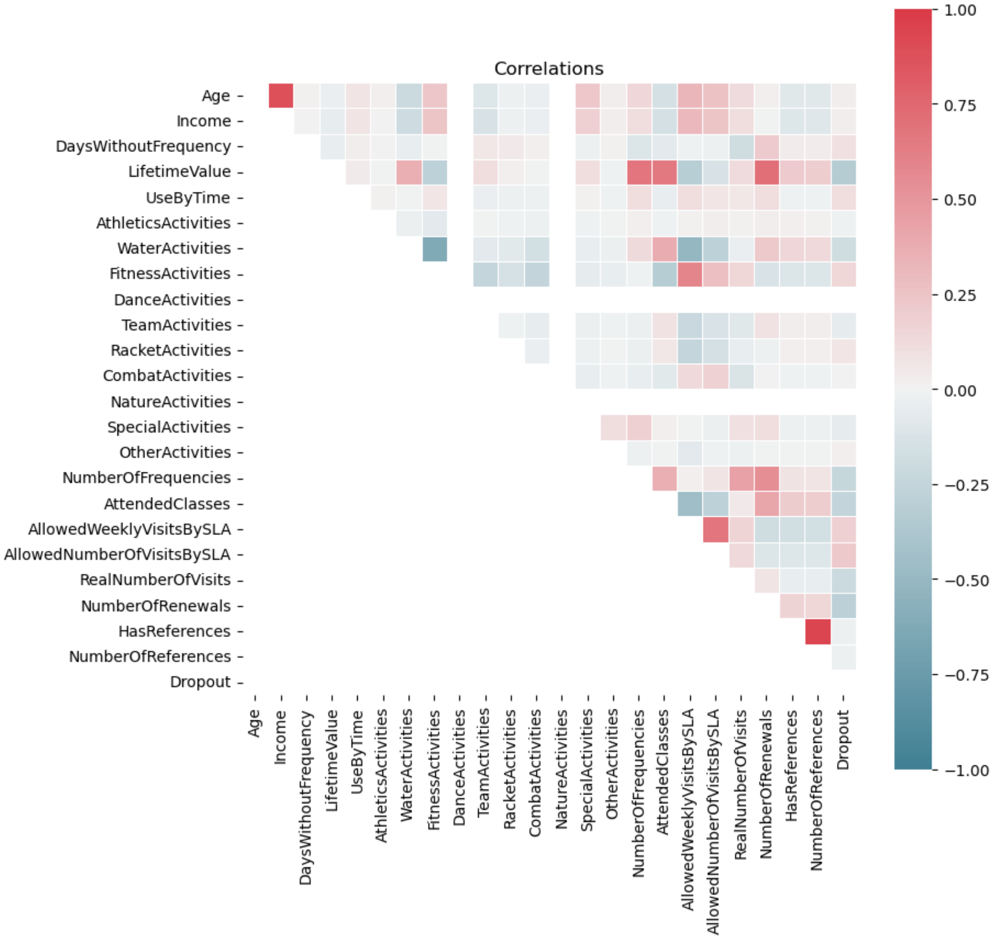


Figure 4 - Correlation Heatmap

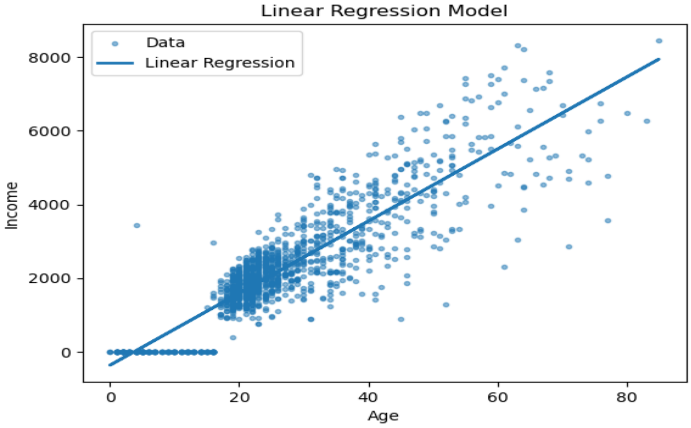


Figure 5 - Plot of Age and Income with a linear regression applied

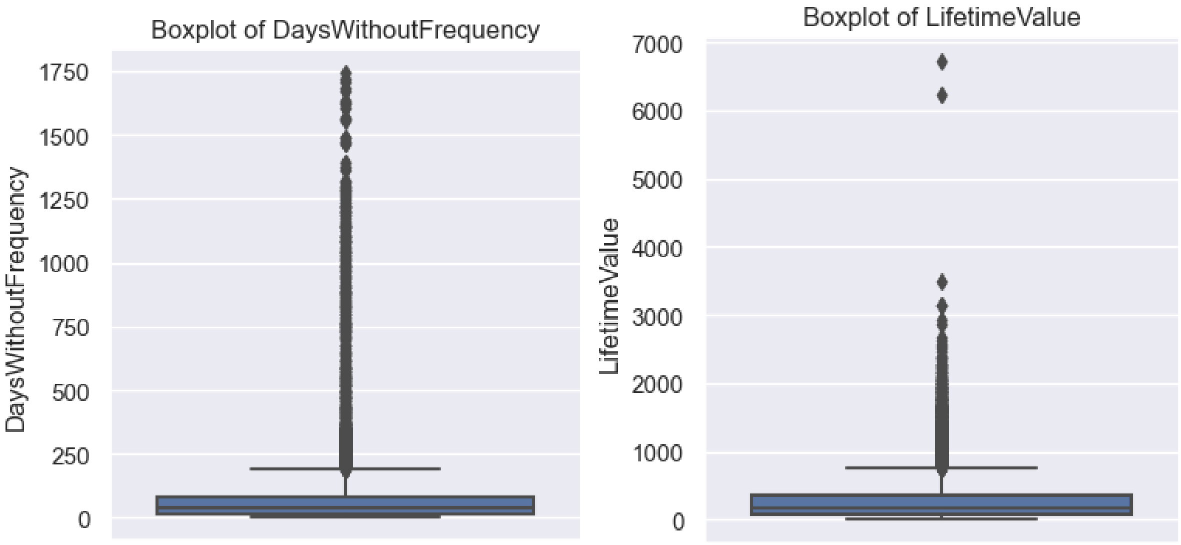


Figure 6 - Box plot of DaysWithoutFrequency and Lifetime Value

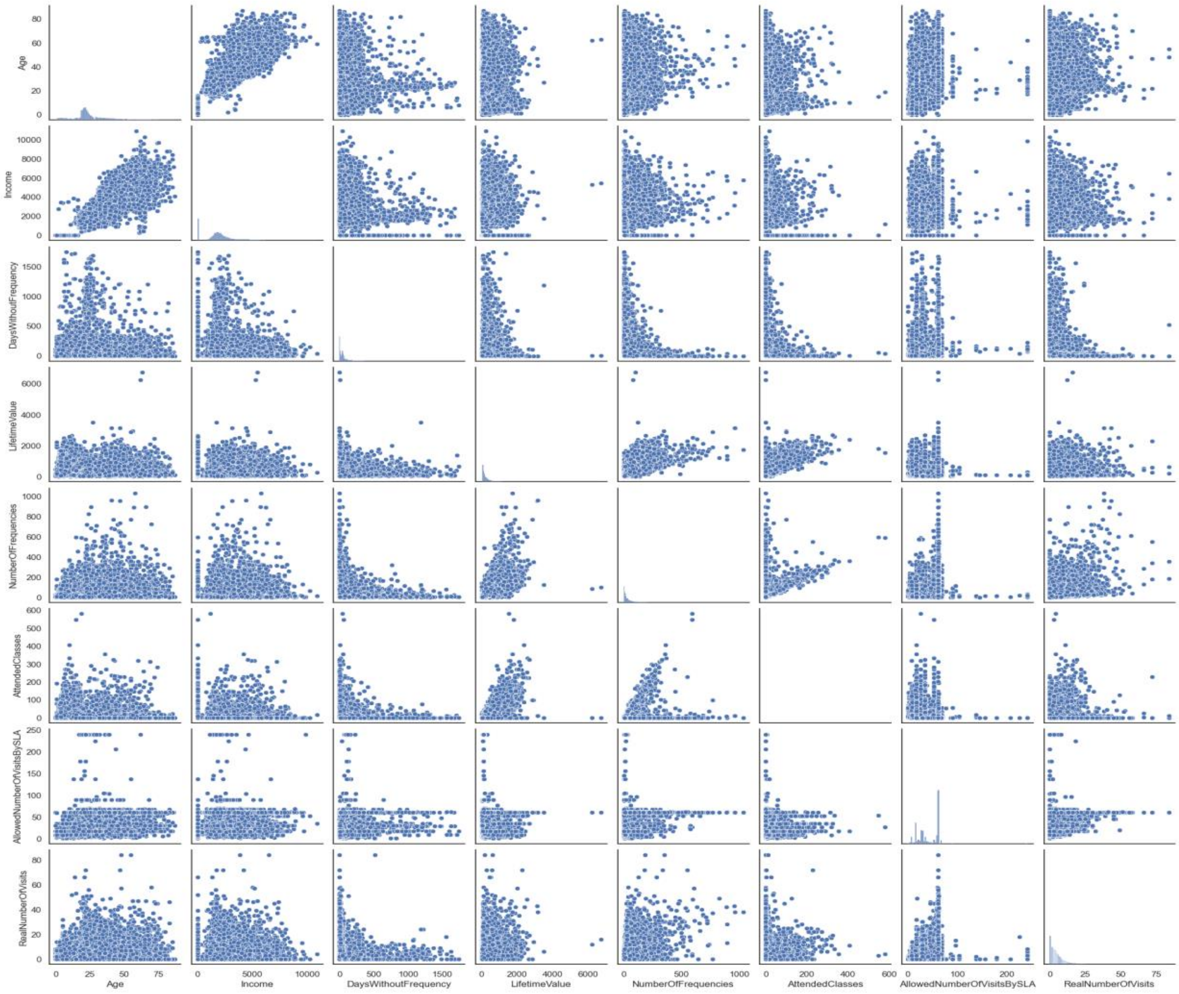


Figure 7 - Pair- Plots of the Numerical Features

Note: As it is hard to visualise the names, the respective features are *Age, Income, DaysWithoutFrequency, LifetimeValue, NumberOfFrequencies, AttendedClasses, AllowedNumberOfVisitsBySLA* and *RealNumberOfVisits*

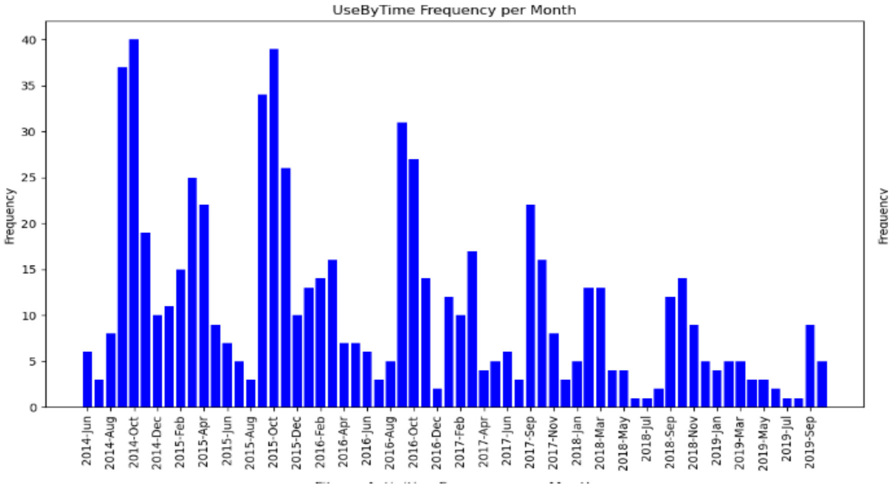


Figure 8 - UseByTime frequency per month

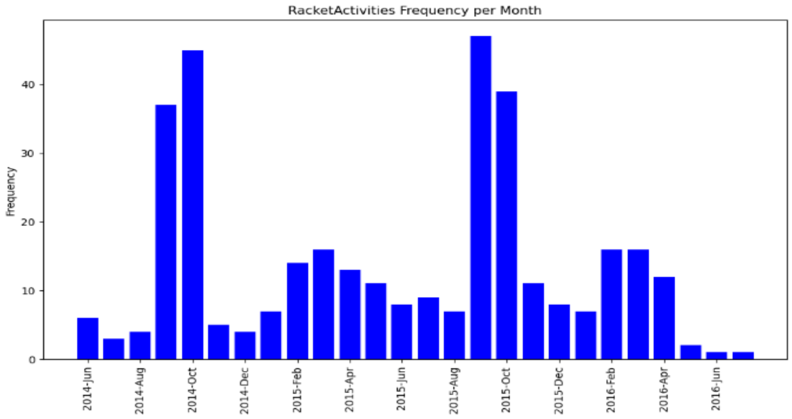


Figure 9 - RacketActivities frequency per month

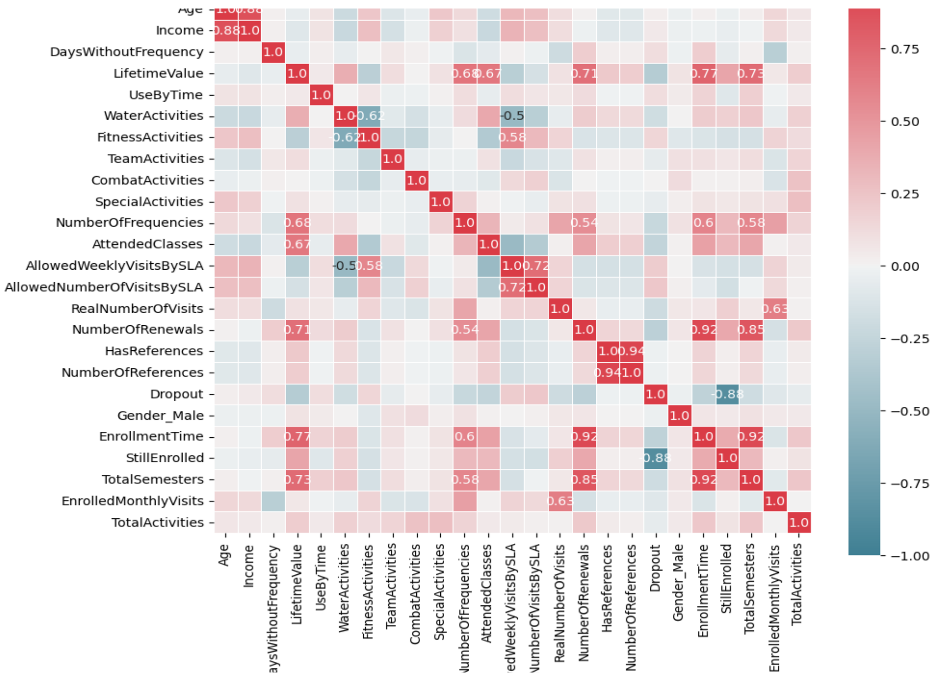


Figure 10 - Correlation Matrix after feature engineering

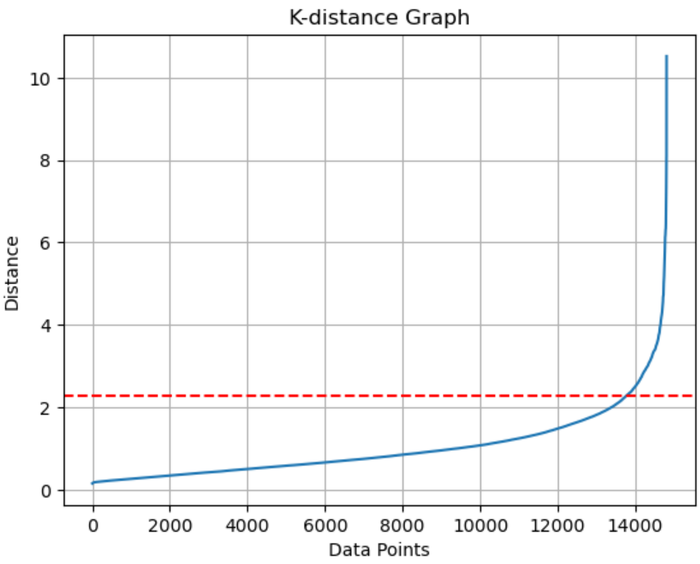


Figure 11 – Optimal value of eps for DBSCAN

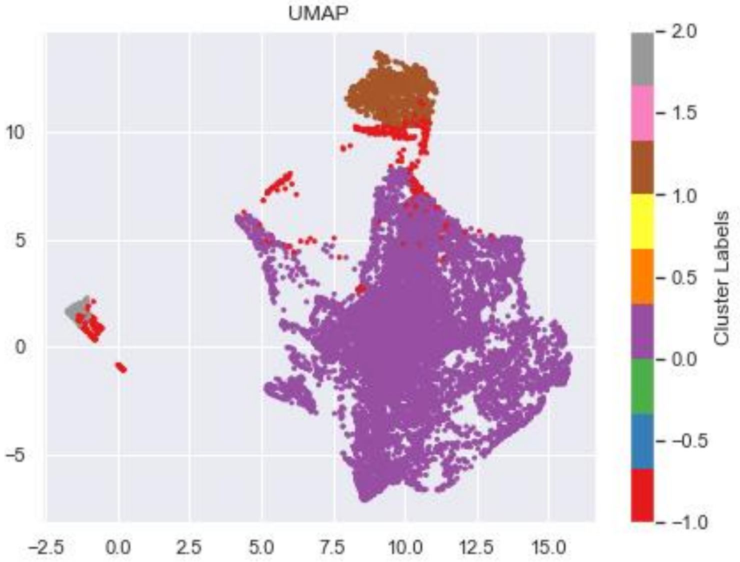


Figure 12 - UMAP visualisation for DBSCAN

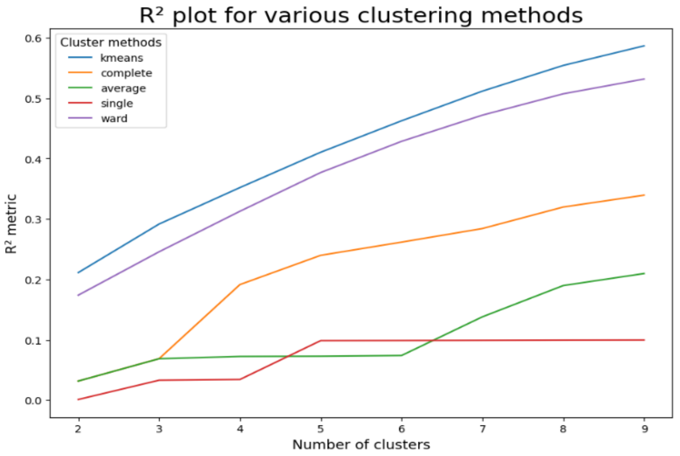


Figure 13 - R^2 plot over k-means and hierarchical clustering algorithms

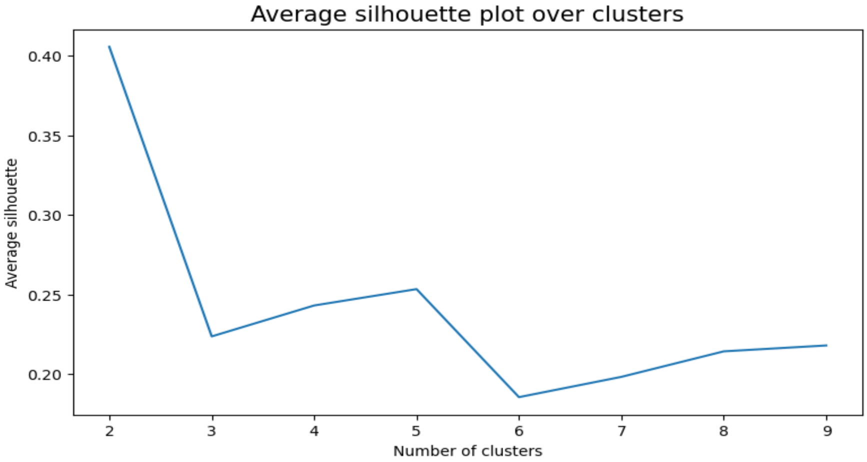


Figure 14 - Average Silhouette plot over clusters

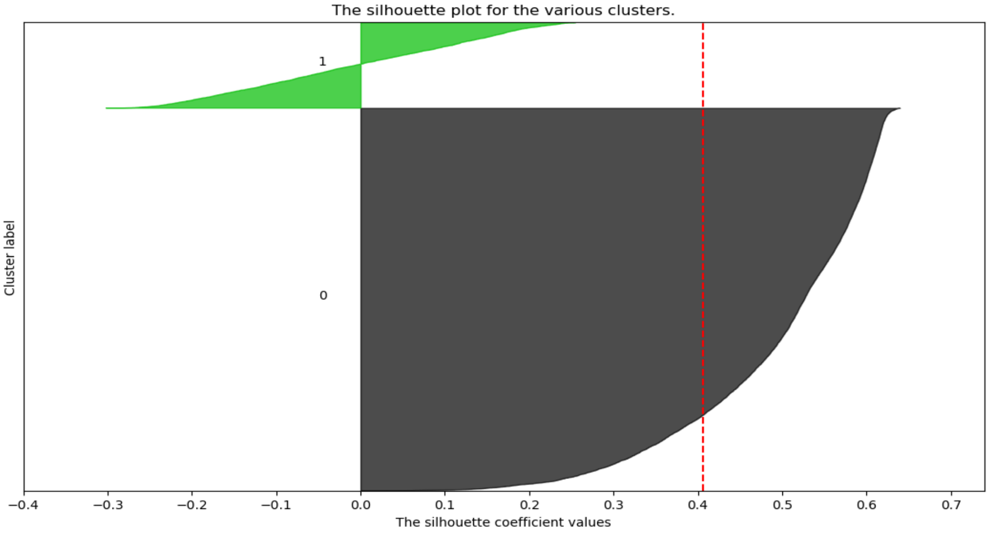


Figure 15 - Silhouette plot for 2 clusters

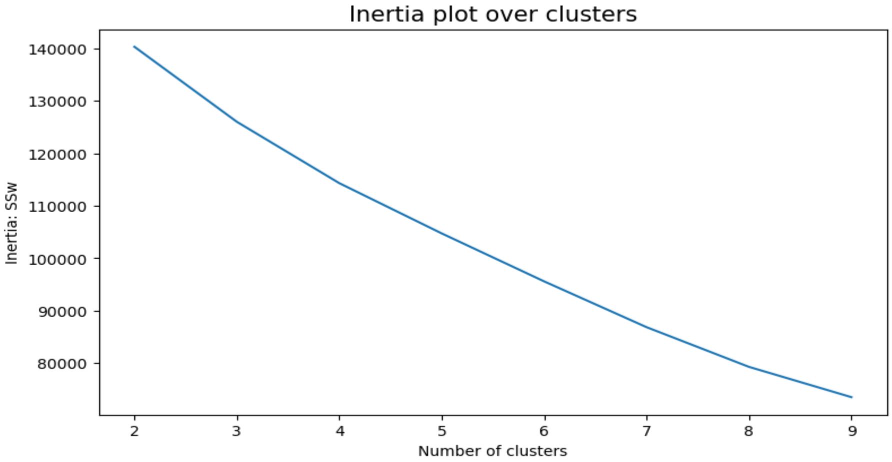


Figure 16 - Inertia plot over clusters

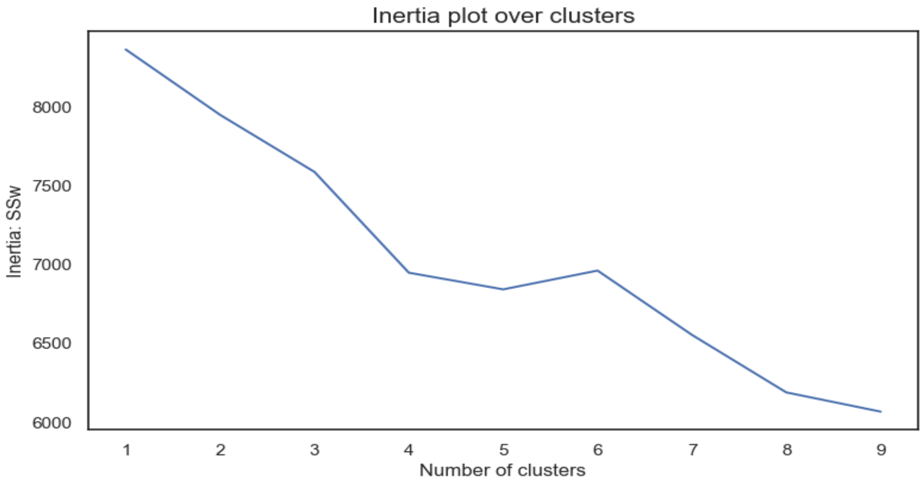


Figure 17 - Inertia plot for K-Medoids

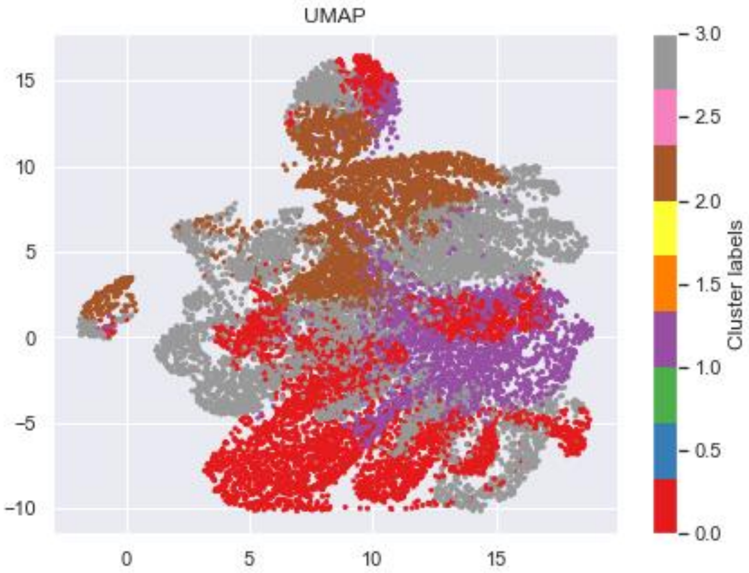


Figure 18 - UMAP for K-Medoids

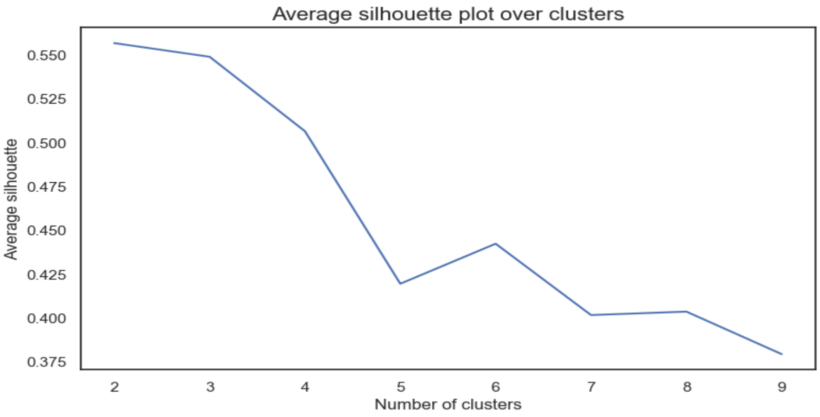


Figure 19 - Average silhouette plot over clusters for K-Prototypes

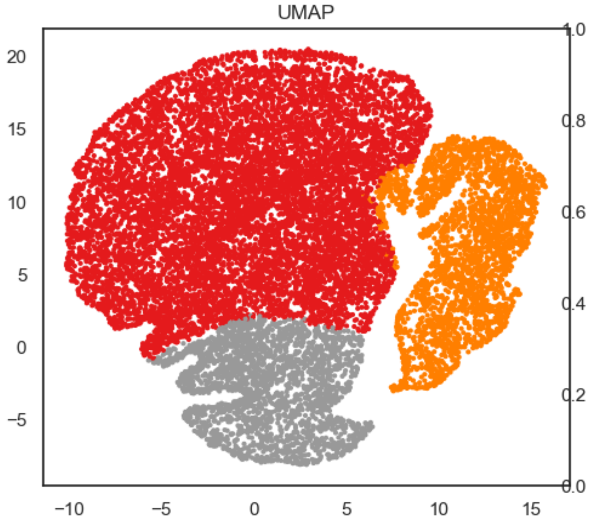


Figure 20 - UMAP with 3 clusters for K-Prototypes

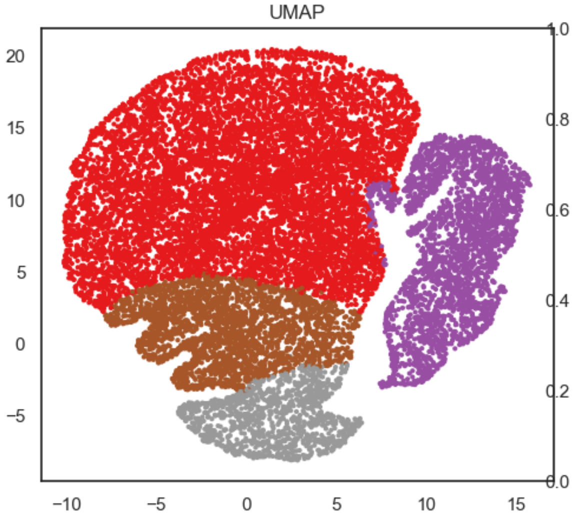


Figure 21- UMAP with 4 clusters for K-Prototypes

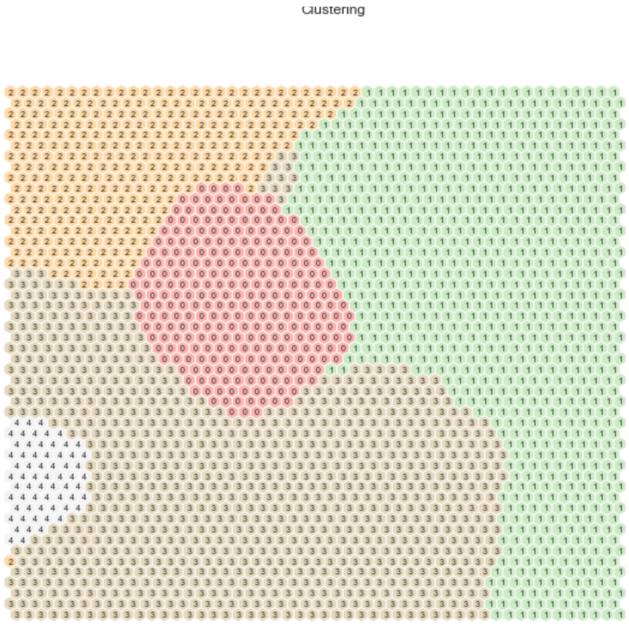


Figure 22 - 5 cluster visualisation for K-means on top of SOM

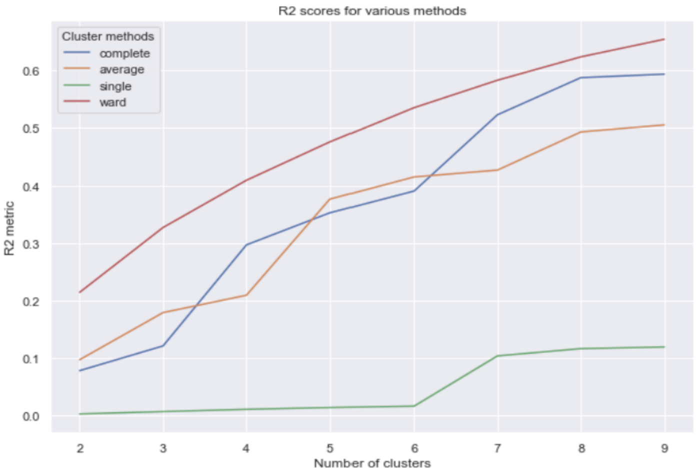


Figure 23 - R^2 scores for several Hierarchical methods, on Hierarchical Clustering on top of SOM

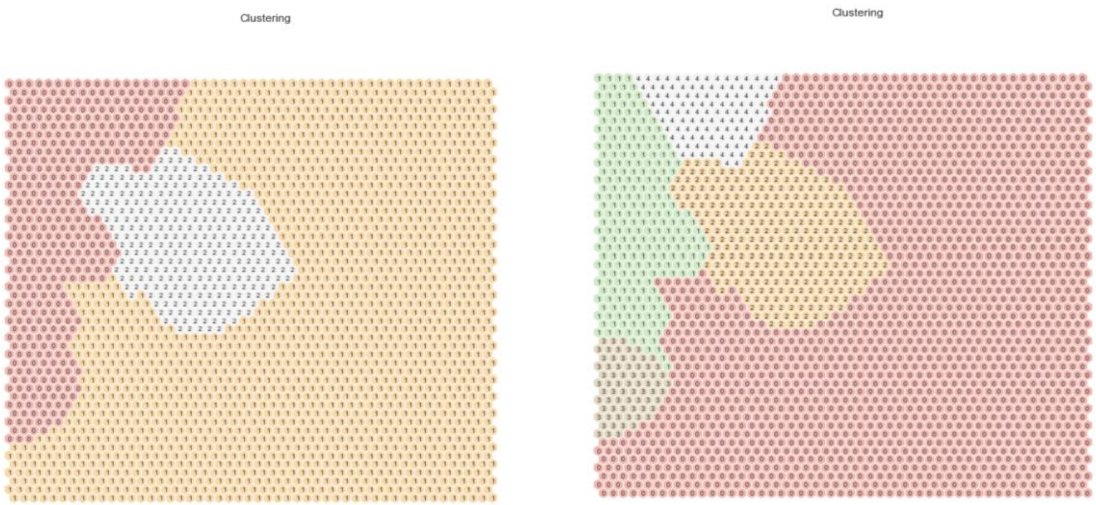


Figure 24 - 3 and 5 cluster visualisation for Hierarchical Clustering on top of SOM

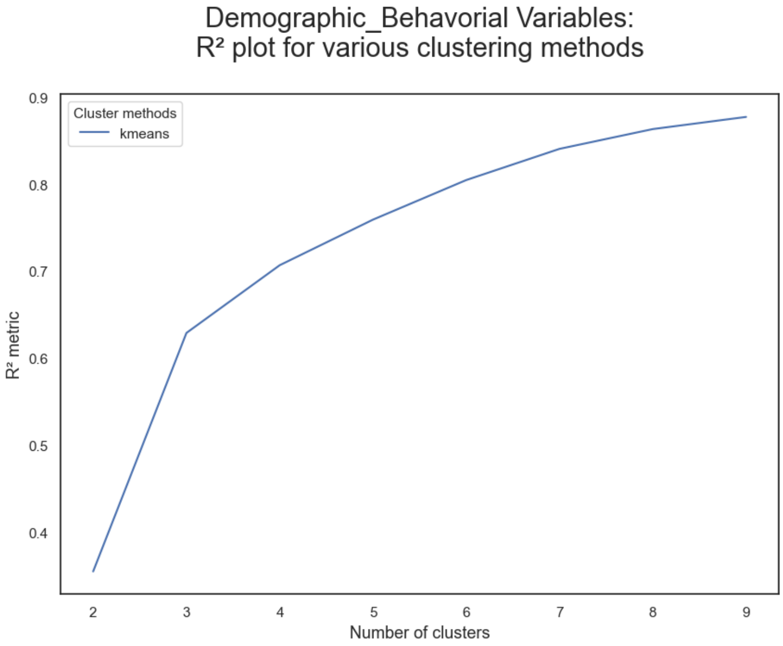


Figure 25 - R^2 plot for k-means on demographic behavior variables

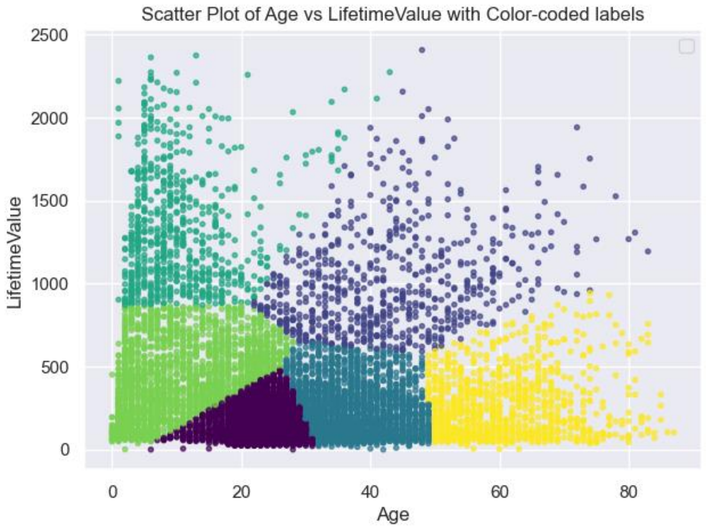


Figure 26 - Scatter plot of Age vs LifetimeValue with color-coded labels

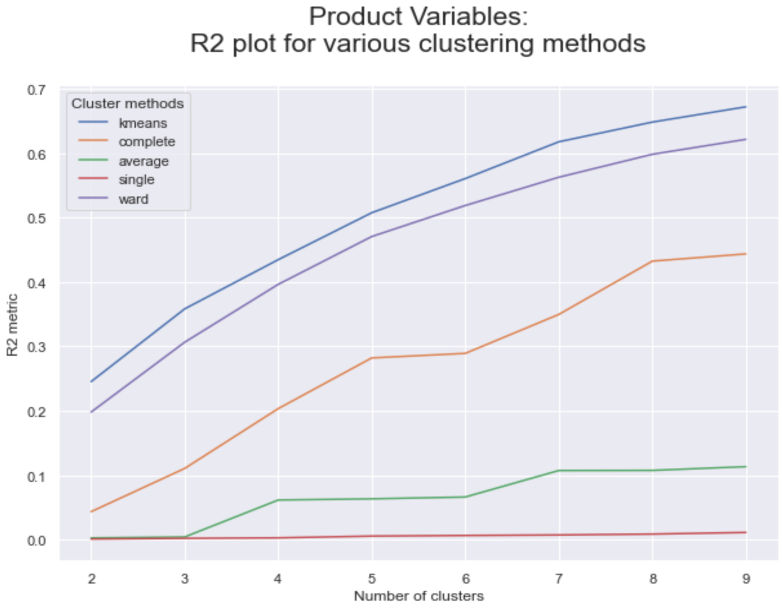
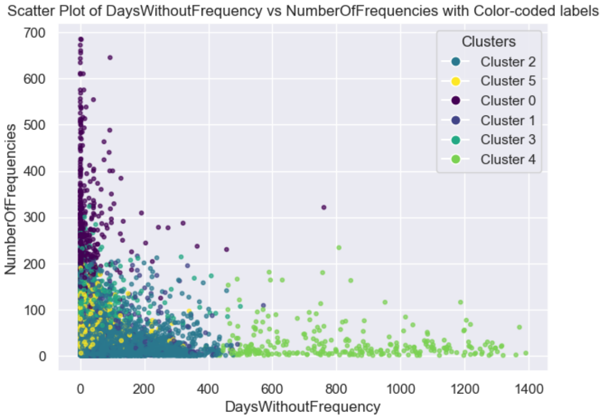
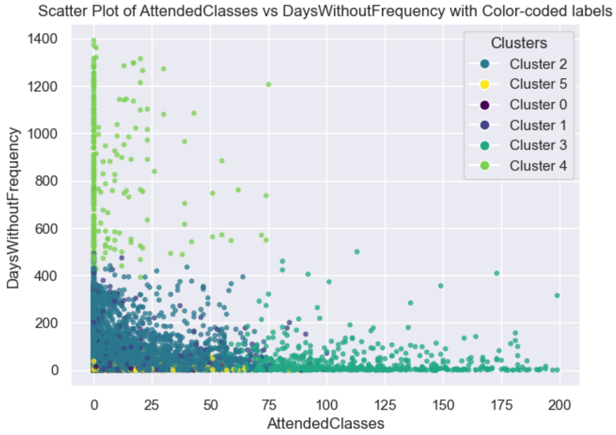


Figure 27 - R^2 plot for k-means for the behaviour analysis





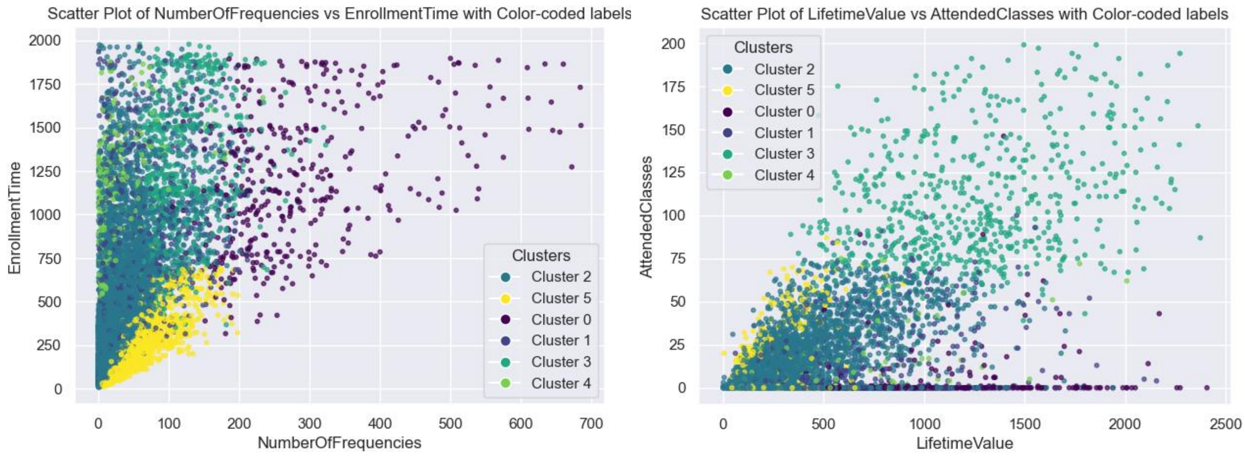
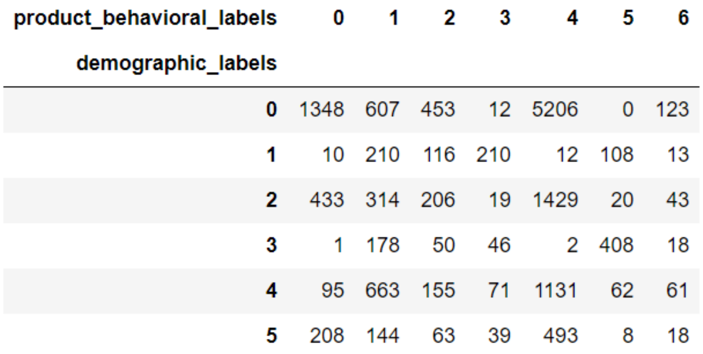


Figure 28 - Distribution of each cluster among relation between features (Examples)



product_behavioral_labels	0	1	2	3	4	5	6
demographic_labels							
0	1348	607	453	12	5206	0	123
1	10	210	116	210	12	108	13
2	433	314	206	19	1429	20	43
3	1	178	50	46	2	408	18
4	95	663	155	71	1131	62	61
5	208	144	63	39	493	8	18

Figure 29 - Table of labels for segmentation

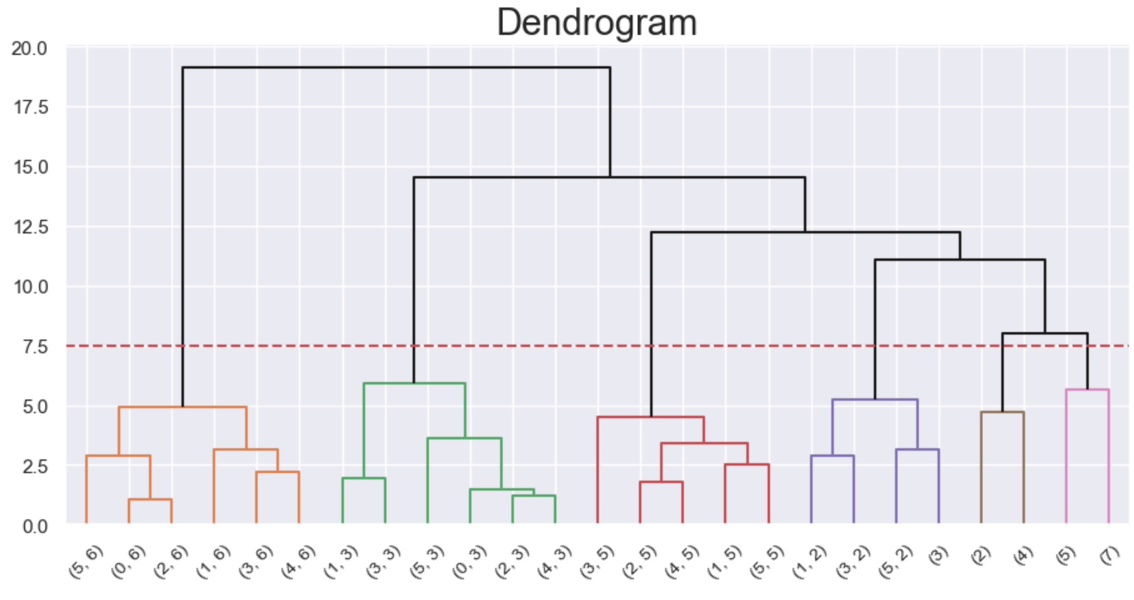


Figure 30 - Dendogram of the clusters and threshold applied

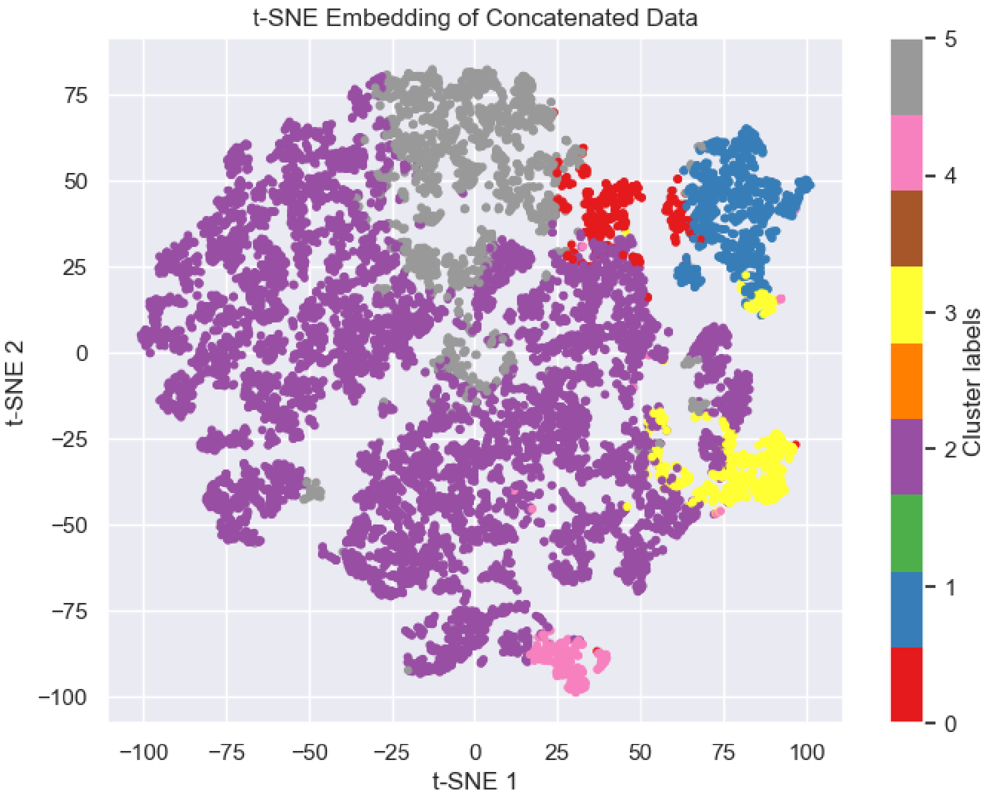


Figure 31 - TSNE representation of the final clusters

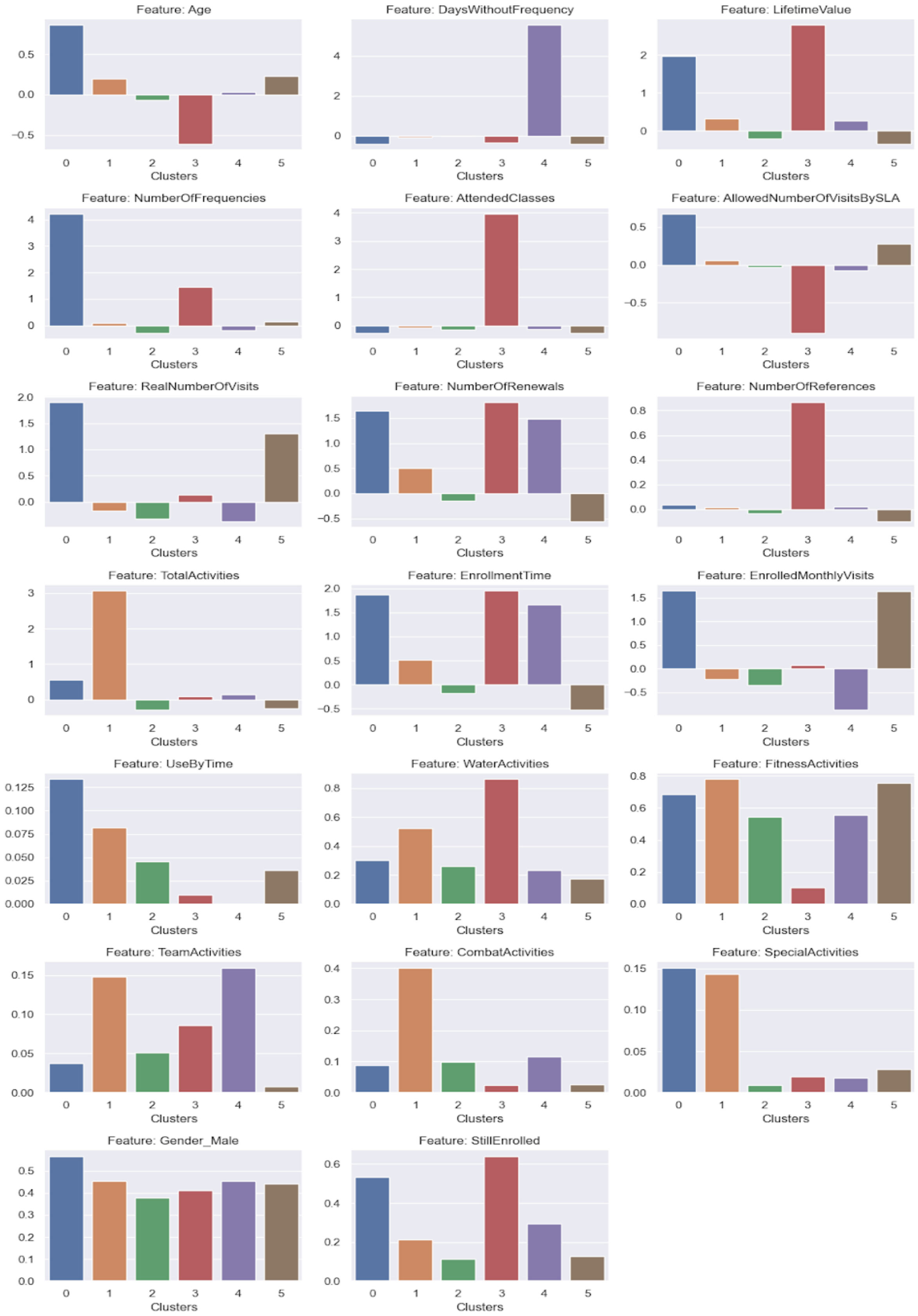


Figure 32 - Cluster profiling