Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Customer Segmentation Report

Group J

Inês Ribeiro, number: m20210595

José Dias, number: m20211009

January, 2022

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

Table of Contents

Table of Figures

1. Introduction

This project is about an insurance company that operates in Portugal.

The goal is to develop a Customer Segmentation based on a database of 10.290 customers.

With this Customer Segmentation, the company will be able to improve their performance. The segmentation allows us, not only to choose which kinds of customers to target (maybe there are some groups of customers that are not worth doing any marketing campaign because they do not resonate with the company message/values) but also to actually do different marketing campaigns for each different cluster/group of customers. This could potentially increase the company profits.

During this project, the data was worked in a way that would be beneficial for the cluster analysis and later, the customer segmentation.

The project can be found in a GitHub repository, with the following link: https://github.com/josedias97/Data-Mining-Project.git .

2. Data Pre-processing

In order to have a good quality data so that a proper and reliable clustering can take place, the data needs to be treated and cleaned.

The first thing that needed to be done was to check and remove any duplicated observations: there were 3 observations removed at this step.

After observing the descriptive statistics table of the data, it was noticeable that there were some missing values, incoherencies and some possible outliers that needed to be looked at in more detail.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
CustID	10296.0	NaN	NaN	NaN	5148.5	2972.34352	1.0	2574.75	5148.5	7722.25	10296.0
FirstPolYear	10266.0	NaN	NaN	NaN	1991.062634	511.267913	1974.0	1980.0	1986.0	1992.0	53784.0
BirthYear	10279.0	NaN	NaN	NaN	1968.007783	19.709476	1028.0	1953.0	1968.0	1983.0	2001.0
EducDeg	10279	4	b'3 - BSc/MSc'	4799	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthSal	10260.0	NaN	NaN	NaN	2506.667057	1157.449634	333.0	1706.0	2501.5	3290.25	55215.0
GeoLivArea	10295.0	NaN	NaN	NaN	2.709859	1.266291	1.0	1.0	3.0	4.0	4.0
Children	10275.0	NaN	NaN	NaN	0.706764	0.455268	0.0	0.0	1.0	1.0	1.0
CustMonVal	10296.0	NaN	NaN	NaN	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	NaN	NaN	NaN	0.742772	2.916964	0.0	0.39	0.72	0.98	256.2
PremMotor	10262.0	NaN	NaN	NaN	300.470252	211.914997	-4.11	190.59	298.61	408.3	11604.42
PremHousehold	10296.0	NaN	NaN	NaN	210.431192	352.595984	-75.0	49.45	132.8	290.05	25048.8
PremHealth	10253.0	NaN	NaN	NaN	171.580833	296.405976	-2.11	111.8	162.81	219.82	28272.0
PremLife	10192.0	NaN	NaN	NaN	41.855782	47.480632	-7.0	9.89	25.56	57.79	398.3
PremWork	10210.0	NaN	NaN	NaN	41.277514	51.513572	-12.0	10.67	25.67	56.79	1988.7

Figure 1: Descriptive statistics of the dataset

2.1. Coherence Checks

There were performed the following coherence checks:

- **FirstPolYear**: from the descriptive statistics table, there does not seem to be any problem with the minimum value of this variable. However, and since this database is from 2016, the maximum value of 53784 does not make sense. Any observations with values of FirstPolYear greater than 2016 were removed;
- **BirthYear**: the minimum value of this feature is 1028, which cannot be correct, therefore it was decided to check a threshold of a maximum of 120 years (record of the oldest person alive ever) and the observations that did not comply with this rule were removed;
- **FirstPolYear** and **BirthYear**: it also does not make sense to do an insurance policy before being born, so any observation that has a value of BirthYear greater than the value of FirstPolYear would be removed. However, there were 1997 observations in this situation, corresponding to almost 20% of the data. The best approach was to remove one of the variables causing this incoherence. As the BirthYear is more susceptible to mistakes, since it is an information provided by the customer and not by the company, this was the feature removed;
- **MonthSal:** it was considered incoherent for customers to spend more money in insurances than they earn. The incoherent observations were removed.

Based on the descriptive statistics table, the remaining metric variables did not seem to have any incoherence. It is also important to point out that there was no need to perform coherence checks in the categorical features since their categories could be observed in their barplots and everything seemed to be coherent.

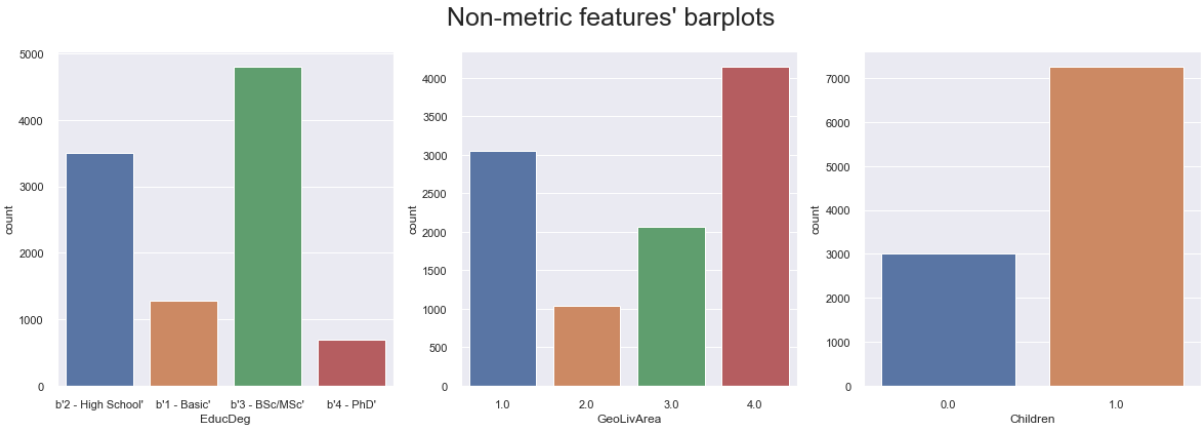

Figure 2: Categorical features barplots

After the coherence checks there were 10290 customers in the dataset, meaning that there were 3 observations removed in this step.

2.2. Missing Values

Before proceeding to the outlier treatment, it was checked if there was any variable with more than 20% of observations with missing values. If so, this feature would have been removed. In this case, the feature with the greatest percentage of missing values was PremLife (1.01%) and therefore, no feature was removed.

2.3. Outliers

The treatment of outliers could not be done with the dataset as it was because of the missing values. The missing values were temporarily filled with the median of the corresponding variable in order to be able to deal with the outliers.

To detect them, the function LocalOutlierFactor from sklearn was used. It considered 84 observations to be outliers. Those 84 observations, approximately 0.82% of the data, were removed, leaving the dataset with 10206 customers.

2.4. Missing Values (continuation)

To perform the missing values' treatment, the approach used was to fill the missing values with the KNN imputer. One popular technique for imputation is a K-Nearest Neighbor model. A new sample is imputed by finding the samples in the training set "closest" to it and averages these nearby points to fill in the value.

This method was used for all categorical and numerical variables except the premiums.

The categorical variable EducDeg had string values, but KNN imputer only works with numerical data, so in order to be able to fill in the missing values with this method, it was necessary to transform the strings into their numerical representation.

The premium variables were analyzed differently. The missing values in a premium variable were simply replaced by 0, this was done under the assumption that the customer did not spend anything on that specific insurance.

2.5. Feature Engineering

At this point, it was decided to create new features that could be more useful and improve the interpretability of our models.

The first variable to be created was Client_Years, it represents the number of years that passed since the customer became a client of this company. This is much more interpretable than FirstPolYear.

The second variable to be created was Yearly_Salary, which is the remuneration of the customer per year instead of per month (MonthSal). As the premiums are annual, this variable seems to make more sense than MonthSal.

The third variable that was created was Total_Premiums, it corresponds to the sum of all the premium variables (PremMotor, PremHousehold, PremHealth, PremLife and PremWork).

The fourth variable created was Profit, this is a binary variable that takes the value of 1 if the client represents a profit for the company, and 0 otherwise. This was calculated with the already existent variable ClaimsRate (if ClaimsRate is below 1, the client represents Profit to the company).

At this point, the observations that had a value of 0 in the variable Total_Premiums were deleted as it was assumed that these customers did not have any purchases in the year of 2016.

The fifth variable created was Effort_Rate, it represents the proportion of salary the clients spend in premiums.

Then, there were 5 more features created that correspond to the proportion of each premium in the variable Total_Premiums (Motor_Ratio, Household_Ratio, Health_Ratio, Life_Ratio and Work_Ratio). When the value of any of these ratios was negative, that value was replaced that by 0, because it was assumed that the customer did not spend money on the related premium.

The last feature created was Cancelled, it is also a binary variable which takes the value of 1 if the clients cancelled at least one premium, and takes the value of 0 otherwise.

2.6. Feature Selection

After feature engineering, the process of feature selection started to choose the most adequate variables for clustering.

2.6.1. Metric Features

The Pearson correlations between metric features were analyzed and although some features are highly correlated with each other, there was not any feature removed since they could be useful for clustering. However, highly correlated features should not be together on the selected features for each cluster.

2.6.2. Categorical Features

For the categorical features there were used multiple methods to check if the features had any importance for the analysis.

Firstly, it was checked if there was any variable that for one category had more than 90% of the observations, because that would mean that the other categories had very few observations, making the variable not much relevant.

The opposite was also checked. If a feature had too many categories, it would have few observations in each one (less than 10% of the observations) making it also not relevant, hence it would not make sense to keep that variable.

In both cases, there were not any features to remove.

To finish, it was checked if the different categories of a certain categorical feature had a different effect on the metric features. As it can be seen in the boxplots (these can be found in the appendix section) there are no significant differences between categories of GeoLivArea in the metric variables. Consequently, this variable was concluded not important and was removed.

Contrary to what happens with GeoLivArea, the different categories of the variables EducDeg, Children, Cancelled and Profit have impact on (at least some of) the metric variables. Thus, these variables were kept.

2.7. Data Normalization

To finish the data preprocessing, it was created a scaled dataframe. This is important because for most clustering methods, distances are analyzed, meaning that if the data isn't scaled, some features take more importance in the clustering algorithms than others.

3. Clustering

3.1. Clustering perspectives

After the pre-processing was finished, it needed to be decided what kind of segmentation was going to be done (what perspectives were going to be analyzed).

There were 2 different and interesting perspectives to study:

- Value Perspective: includes the features that are related with value (features: CustMonVal, Effort_Rate, Total_Premiums, Cancelled, Profit)
- Consumption perspective: includes all the premium ratios (Motor_Ratio, Household_Ratio, Health_Ratio, Life_Ratio, Work_Ratio)

It was tried as much as possible to not include correlated features in the same cluster perspective.

Both perspectives were used in the clustering methods, with the exception of K-Prototypes, that was not used for Consumption because it does not have any categorical features (and the algorithm needs both metric and categorical features).

3.2. DBSCAN

The first clustering algorithm used was DBSCAN. This was the first algorithm because it detects the outliers automatically, thus, these were removed to run the other clustering methods.

The DBSCAN is an algorithm that creates the clusters by analyzing the densities. The biggest advantages of this method are that it is not necessary to predefine the number of clusters, like it is in most algorithms, and the clusters can have an arbitrary shape. It is only needed to predefine 2 parameters: epsilon (the radius defining the neighborhood) and MinPts (minimum number of points in the ε-neighborhood). The disadvantages of this method are that we need to define these parameters, and that it does not perform well when the clusters are of varying density.

Based on the density of an area of data points, DBSCAN can

The results of DBSCAN were not great especially because the R squared was low for both perspectives and it yielded a one cluster solution (considering that the other cluster is considered to be noise or outliers).

3.3. Hierarchical Clustering

Since the Hierarchical Clustering algorithm was used in almost every method, it is important to make a brief introduction of this method. To get the first cluster, the two points with the smallest distance are found in the distance matrix, then the next two points, and so on, until there is just one cluster (in the end). In the beginning of the process, each point is a cluster by itself, and in the end of the process all points are in the same cluster; it needs to be decided how many clusters to keep, because ending up with just one cluster is not be useful. This method is not very good in isolation (does not yield great results), but it can definitely be used in conjunction with other methods to find out how many clusters to use.

3.4. K-Prototypes

In the dataset, there were some categorical variables that seemed to be quite important, therefore, it would be of high importance to find a clustering algorithm that could include both numerical and categorical variables. The chosen algorithm was the K-Prototypes algorithm. The K-Prototypes algorithm allows the usage of categorical variables, and it is a mixture between the K-Means and K-modes algorithms.

The only clustering perspective used with K-Prototypes was the value perspective, because the consumption perspective did not have any categorical features.

To start the process, 10 clusters were chosen with 10 initializations, and it seemed to be a big enough number of initial clusters to give a general idea of the structure of the data. Later, the elbow method was used to decide the final number of clusters to be used with K-Prototypes, and converge to a final solution.

The final clusters of K-Prototypes were composed of 3 clusters, and the results were quite satisfactory.

3.5. K-Means & Hierarchical

Another clustering method that was used and was in fact the chosen one for the final clusters, for both perspectives, was K-Means and Hierarchical Clustering.

The K-Means algorithm is very susceptible to outliers so, the observations that were considered by DBSCAN to be noise or outliers were removed (these observations were not dealt with as outliers in the pre-processing). By doing this, it is possible to use K-Means and there is no need to use K-Medoids.

Not all variables from the value perspective were used as some of them were categorical and K-Means can only work with numeric data. This means that the variables ‘CustMonVal’, ‘Effort_Rate’ and ‘Total_Premiums’ were

used and the variables 'Cancelled' and 'Profit' were not. This was not a problem in the consumption perspective as its variables (premium ratios) are all numerical.

For both value and consumption perspectives, the K-Means algorithm was first run with 20 clusters, to be able to represent the structure of the data. Then, it was applied the Hierarchical Clustering algorithm and with the help of a dendrogram it was determined that the appropriate number of clusters should be 4 for the value perspective and 6 for the consumption perspective.

After this, K-Means was run again with the number of clusters found by the Hierarchical Clustering algorithm. To have an idea on how the different features vary from cluster to cluster, the variables' histograms for the different clusters were plotted, and once again, this was done for both perspectives.

3.6. SOM & Hierarchical

The SOM algorithm is a specific type of a neural network, and it can be used both for clustering and for high dimensional visualization.

With the SOM algorithm, it is important to check the U-Matrix (distances between the units of the SOM), Component Planes (average value in each variable for each one of the units in the SOM) and Hit Maps (representing the size of hexagon based on the number of individuals classified in that unit; the hit map represents the number of individuals classified in each one of the units of the map).

To better decide the number of units, it was applied Hierarchical Clustering on top of SOM using the units obtained as observations. After that, the dendrograms for the two perspectives were plotted and analysed. For value segmentation the number of clusters chosen was 5 and for consumption segmentation it was 6.

The results of these clusters were quite good, there was a good variation between the different clusters. However, it was decided not to use this method because by analysing the histograms of the K-Means' clusters and the histograms of the SOM's clusters, the ones that were obtained with the K-Means algorithm seemed to be more promising.

3.7. Best Clusterer

The best clusterer, for both value and consumption perspectives, was the K-Means combined with the Hierarchical Clustering method.

The DBSCAN method was by far the worst one.

The SOM with Hierarchical Clustering and the K-Prototypes methods yielded very good variability in the features between the different clusters. Despite all that, the K-Means was considered to be the best as the differences between clusters were more visible.

3.8. Cluster Visualization

3.8.1. t-SNE (t-distributed Stochastic Neighbour Embedding)

The objective is to project an n dimensional space into a 2-dimensional space. It models it in a way that the neighbours in a high dimensional space continue to be neighbours in a low dimensional space. Although this method is adequate to visualize high-dimensional data, it is also highly recommended to use another dimensionality reduction method, such as PCA.

With this in mind, it was created a data frame with PCA to improve the quality of the visualization of the clusters. With the usage of the elbow method, 3 principal components were retained.

This resulted in 4 compact, cohesive and very well distanced clusters.

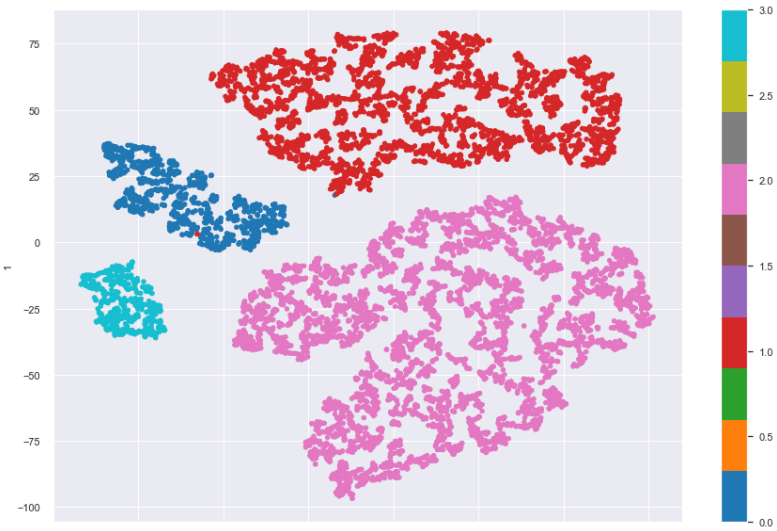

Figure 3: t-SNE cluster visualization

3.8.2. UMAP

Even though the results gotten with t-SNE were quite good, it could be beneficial to try the Uniform Manifold Approximation and Projection (UMAP).

Similarly, to t-SNE, PCA´s were used to do the visualization with UMAP (to reduce dimensionality).

The results obtained with UMAP were not great, and so the visualization of the clusters obtained with t-SNE was much better.

3.9. Cluster Profiling

In the cluster profiling part, it can be seen how different the four final clusters are from each other.

Here is a short description of the differences between the clusters:

- **Cluster 0**: this cluster has higher values than the others for the features CustMonVal, Effort_Rate, Total_Premiums, PremHousehold and PremLife and has lower values than all other clusters for PremMotor and MonthSal. These customers are the ones who spend more on Premiums, have a higher percentage of their salary on insurances, and they only spend a high amount on Life and Household insurances. That means that these customers do not have the biggest buying power (as it can be seen by the MonthSal), they do not receive the highest wages, therefore they will only spend money on the insurances that are strictly necessary. As expected, the level of education of the individuals in this clusters is also lower, on average. This cluster is the second cluster with less individuals;
- **Cluster 1**: this cluster has much higher values than the others for the features PremHealth and MonthSal, and the remaining features have, on average, a similar value to the other clusters. This cluster is the second biggest cluster in terms of number of individuals, and it has an average value for most variables, that means that this is a “typical customer”. They earn an above average salary, but looks like they are also more conservative in their buying patterns. They also have an above average level of education, and almost half of the individuals do not have children;
- **Cluster 2**: this cluster has much higher values than the others for the feature PremMotor and MonthSal, and has lower values than all other clusters for Effort_Rate, Total_Premiums, and all other premium with the exception of PremMotor. This is the biggest cluster when it comes to the number of individuals.

The effort rate is quite low, but the total premiums are also low, so with these two variables is very difficult of know if these customers have a higher buying power. However, it can be seen that they spend much more on Motor insurance, which is related to premium goods. That could mean that in fact these customers might have a higher buying power (as it can be seen by the MonthSal). They also have a high level of education;

- **Cluster 3**: this cluster has higher values than the others for the feature PremWork and has lower values than all other clusters for CustMonVal. They earn an average Monthly Salary. This is the smallest cluster. They have the highest Work_Ratio. This could mean that these customers are new customers of the company (because of the low CustMonVal, while subscribing mainly to the Work insurance), and they could represent a high value cluster for marketing campaigns (they can definitely increase the number of insurances they buy);

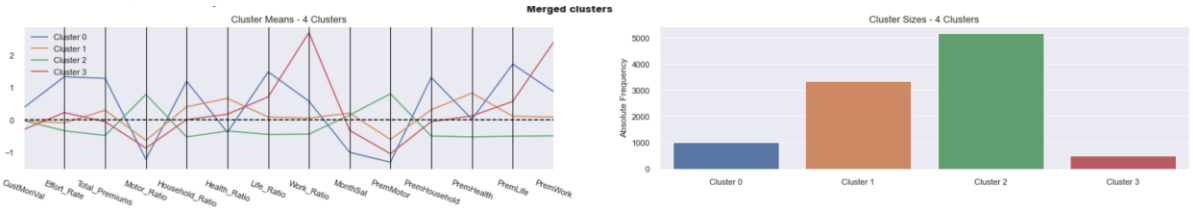

Figure 4: Merged clusters profiles

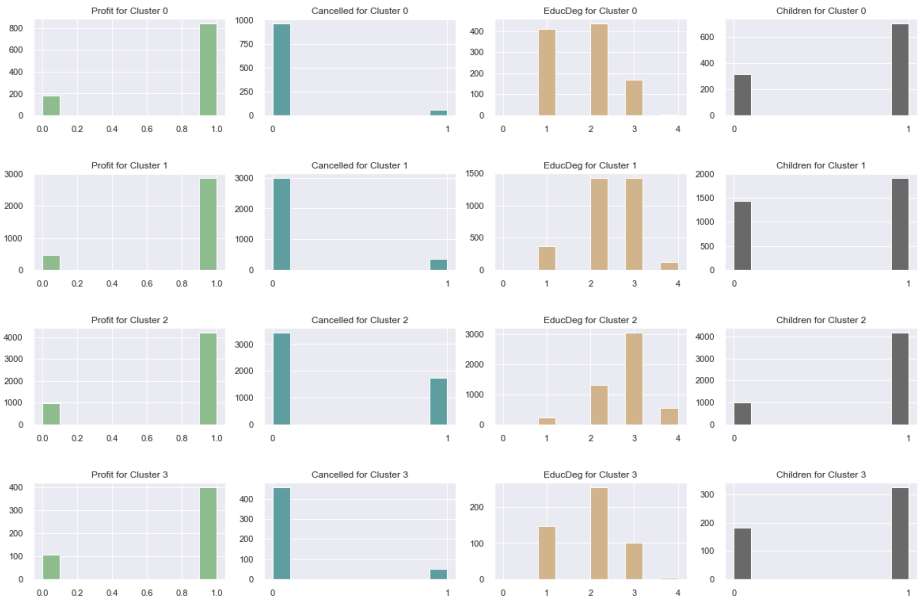

Figure 5: Final clusters categorical features histograms

3.10. Decision Tree

To be able to interpret the cluster solutions, it is very beneficial to visualize a decision tree, and check the rules that the algorithm used to label each cluster. It can also be used to predict the cluster of new observations.

With a decision tree with a max depth of 3, the model was able to predict the test observations with an accuracy of around 88%. The rules to classify each cluster are the following:

- **Cluster 0:** if PremMotor <= 0.008 and Effort_Rate > 1.118 and Work_Ratio <= 2.109, then a customer belong to this cluster with a probability of around 84%
- **Cluster 1:** if PremMotor <= 0.008 and Effort_Rate <= 1.118 and Work_Ratio <= 1.441, then a customer belongs to this cluster with a probability of around 79%

- **Cluster 2:** if PremMotor > 0.008 and Health_Ratio <= 1.046 and Household_Ratio <= 0.703, then a customer belongs to this cluster with a probability of around 98%
- **Cluster 3:** if PremMotor <= 0.008 and Effort_Rate <= 1.118 and Work_Ratio > 1.441, then a customer belongs to this cluster with a probability of around 71%

4. Marketing Approach

Given that after merging the two perspectives, there were four clusters, the marketing team should definitely target the customer of each cluster differently.

The customers of cluster 0 have low salaries and they will not buy anything that is not an essential good for them. They already invest a lot in Life insurance and House insurance (the more essential ones). The marketing team should keep targeting these two, and try to increase the purchases of Health insurance in this cluster. The campaign should be used to inform them that Health insurance is also an essential one. A good approach would be to give them a discount if they buy the three together.

The customers of cluster 1 are very conservative, they are people with a high education level, do not spend money on unnecessary things, and probably live a very simple life without much luxury. This cluster has a high potential because the customers have a high monetary availability. However, they are not going to want to spend a lot of money. They already spend a lot on the Health insurance, but there are two more insurances considered essential and should be looked at: Life and House insurances. Contrary to what was done with the customer from cluster 1, to these customers, a discount should be offered if they buy one more insurance together with the Health insurance. Once again, here a pack of two should be offered because they are very conservative and will not feel comfortable spending money in two more products.

The customer from cluster 2 are the opposite of cluster 1 in terms of buying behaviour. They just spend money in a more superfluous insurance (Car insurance), and do not spend much in the essential ones. They earn high salaries and have high education levels (similarly to cluster 1). There is a big potential in this cluster, because with the exception of Car insurance, they do not spend money in other insurances. Since they have the financial availability to spend money in more insurances, and they are not conservative, they probably would not be more likely to buy even if we offered them a discount. Given that, the best approach for the marketing team is to spend more of the budget to contact/advertise them in different ways and more frequently (do more marketing campaigns targeting this customer specifically). Given that this is the biggest cluster, they should be exposed more frequently to the marketing campaigns. It is expected that the House insurance grows more in sales than any other, because it is more related with a material good.

The customer from cluster 3 spend much more in Work insurance. This is probably because they have a dangerous job (it is the smallest cluster, which makes sense because most people do not have a dangerous job). They earn an average salary, so they probably can afford one more insurance. They already spend some capital on the Life insurance, so it is recommended that the marking team focuses on increasing the sales for Health insurance. They are probably interested in that, since it is related to the insurances that they already buy (for primary needs). Since not all the customer buy the Life insurance, the marketing team could launch a campaign where they offer a discount if they buy two of the insurances together (one of them being Work insurance). It is better to offer this discount only with two products (instead of three) because the customers of this cluster do not have a high financial availability.

5. Pain Points

The first (and biggest) pain point of this project is related with the process of merging perspectives. The Dendrogram was quite difficult of analyse. The initial analysis results in choosing either 6 or 9 (5 would also be acceptable).

However, after seeing the visualization with t-SNE, none of these seemed to be a good number of clusters, this is because there were very large clusters but there were also very small clusters really close to each other.

Given that the outliers were removed twice, and that the number of observations would be too big to be considered outliers, the chosen number of clusters for the final solution was 4. That made much more sense in the t-SNE visualization. The number of clusters chosen might not have been the most optimal one when looking at the dendrogram.

The second pain point that it is important to mention is that the results of K-Prototypes, SOM and K-Means (the last 2 in conjunction with Hierarchical Clustering algorithm) were very similar. In the final solution, the K-Means method was used for the two perspectives. But maybe another combination could be better and give different clusters. The main issue is that the method was chosen just looking at the histograms, and since the histograms were very similar, it could be difficult to choose the best one.

The third pain point, that is also related with the second one is that the results of the R squared of all the models were quite high, and it is probably not very accurate. However, it was not used to choose the best method.

Having that said, the cluster solution resulted in 4 very different clusters, and it could benefit the campaigns of the marketing team, so it is unknown if fixing these pain points would change the solution in a significant/better way, but are still important to mention.

6. Conclusion

The goal of this project was to develop a customer segmentation of an insurance company with a given dataset.

In the beginning, the data was not ready to be analysed and there was some data treatment to be done due to the presence of outliers and noise, missing values and incoherencies.

Even after this, there was still some work to be done in terms of creating new variables and deciding if every variable was in fact important or not for the analysis that was going to be done.

There were two perspectives considered in this customer segmentation analysis, and for each one, different methods of clustering were tested in order to try to obtain the best results possible.

The final cluster solution, with the two perspectives merged, was composed of 4 different clusters of customers. A marketing campaign/gameplan was suggested to try to get the most out of the campaigns that the company has.

To finish, the company should definitely prioritise the campaigns of clusters 1 and 2, because those give more immediate results given that the customers have more budget to spend.

7. Bibliography

https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

https://www.geeksforgeeks.org/ml-principal-component-analysispca/

https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.cluster.hierarchy.dendrogram.html

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

https://scikit-learn.org/stable/modules/clustering.html#clustering

Practical Classes Notebooks

8. Appendix

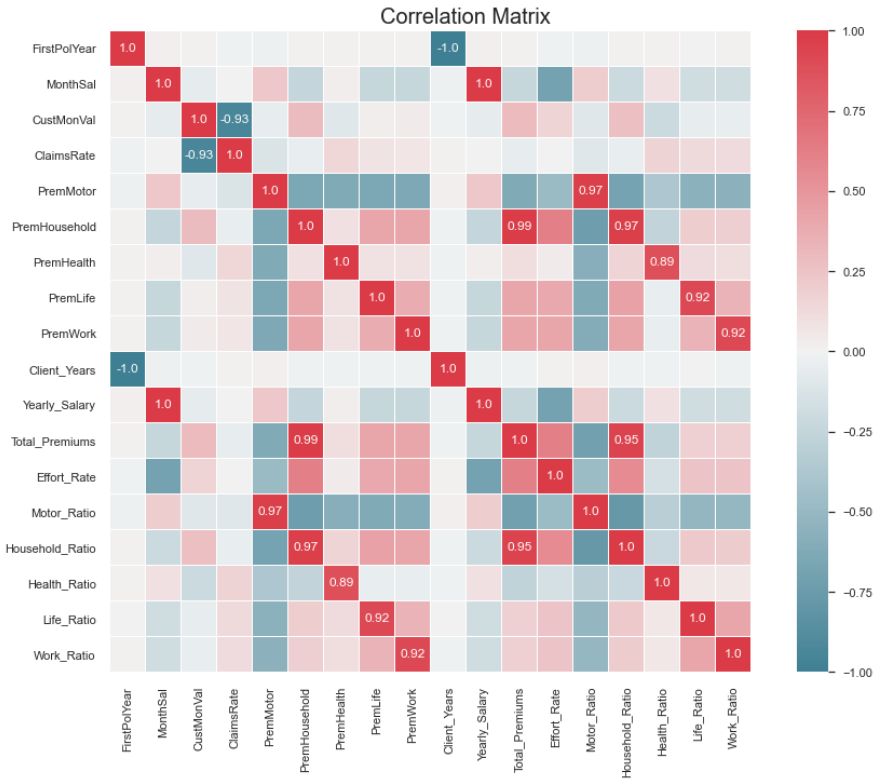

Figure 6: Correlation Matrix

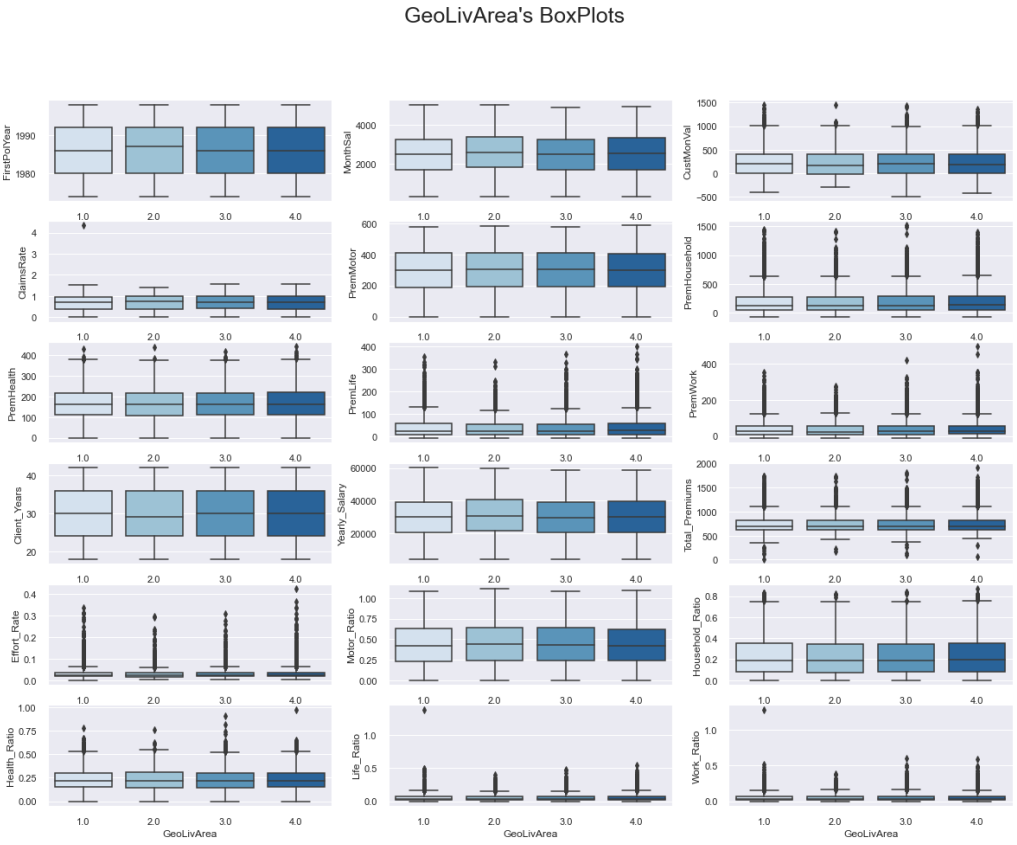

Figure 7: GeoLivArea´s BoxPlots

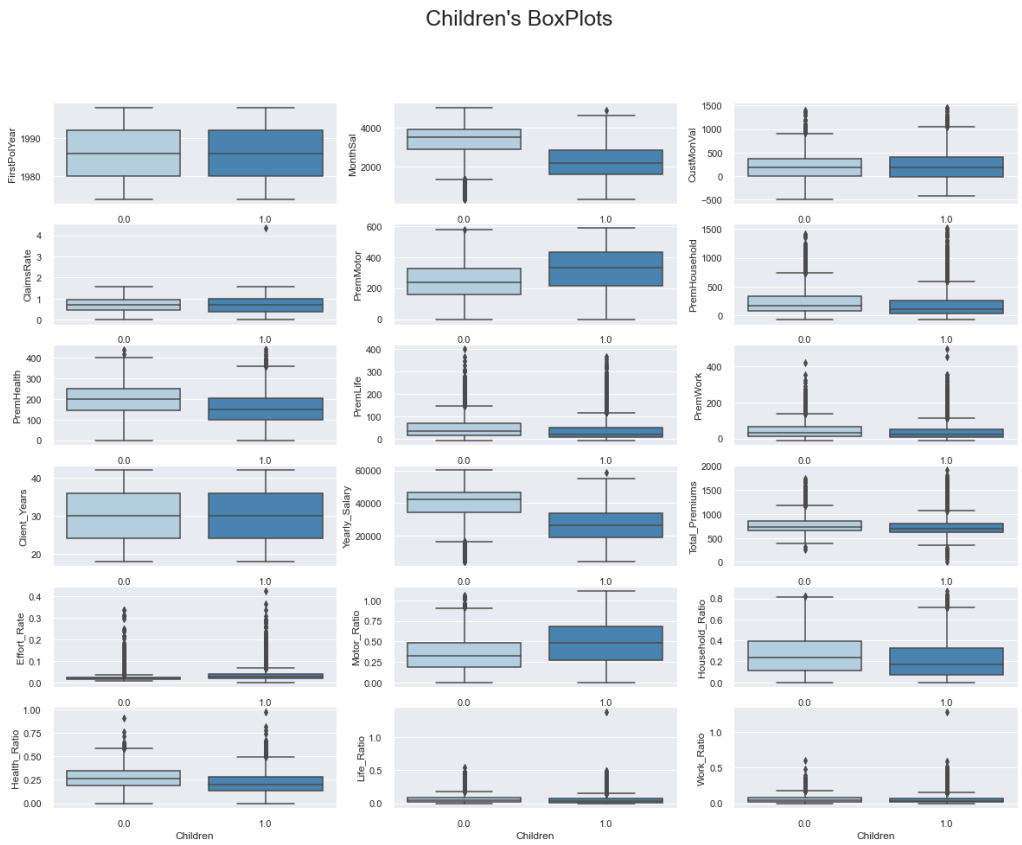

Figure 8: Children´s BoxPlots

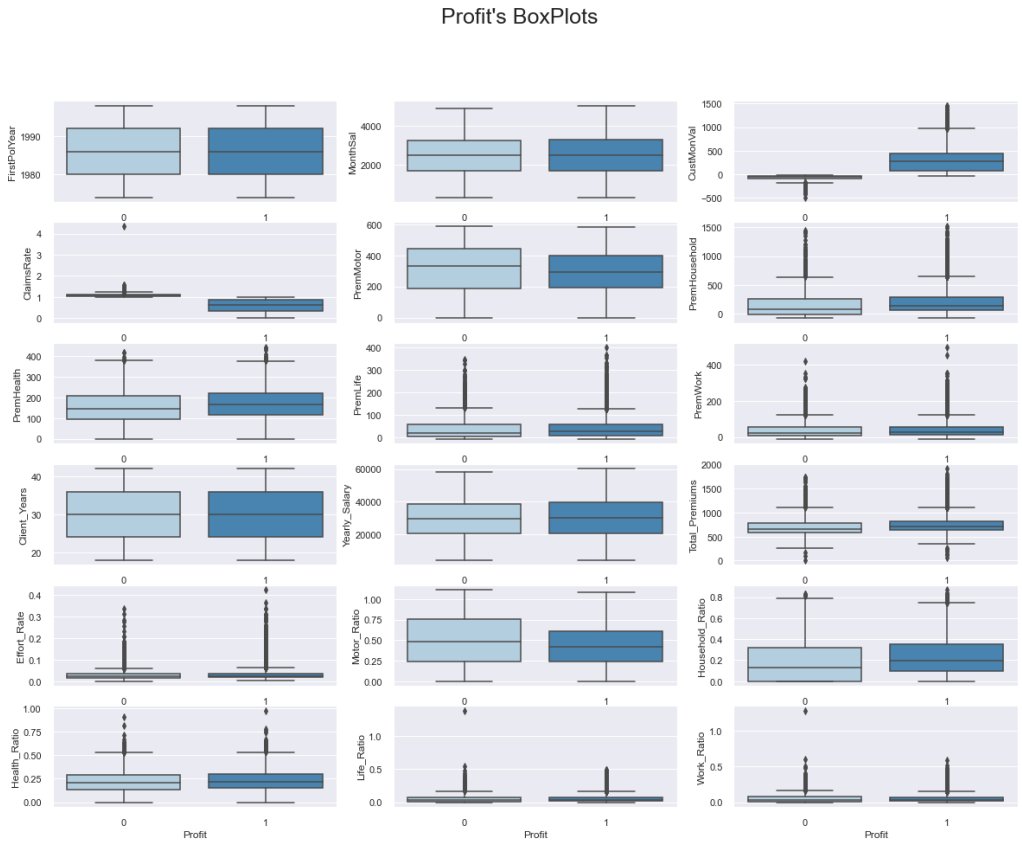

Figure 9: Profit´s BoxPlots

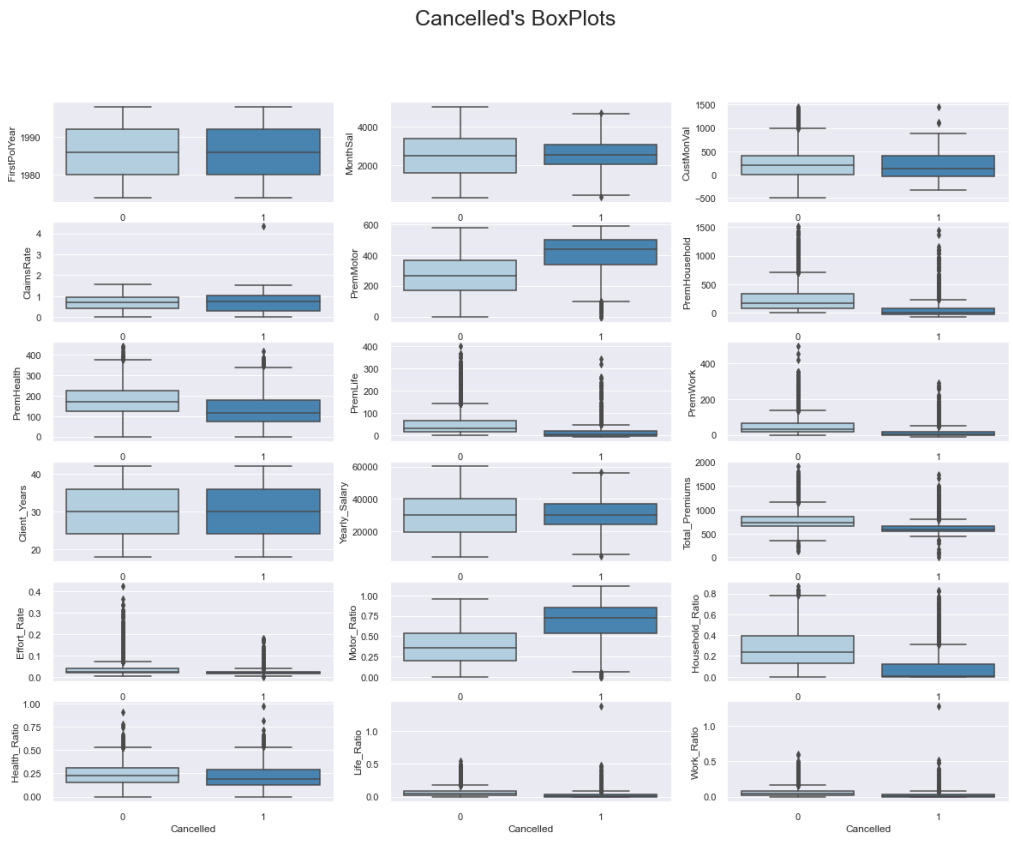

Figure 10: Cancelled´s BoxPlots

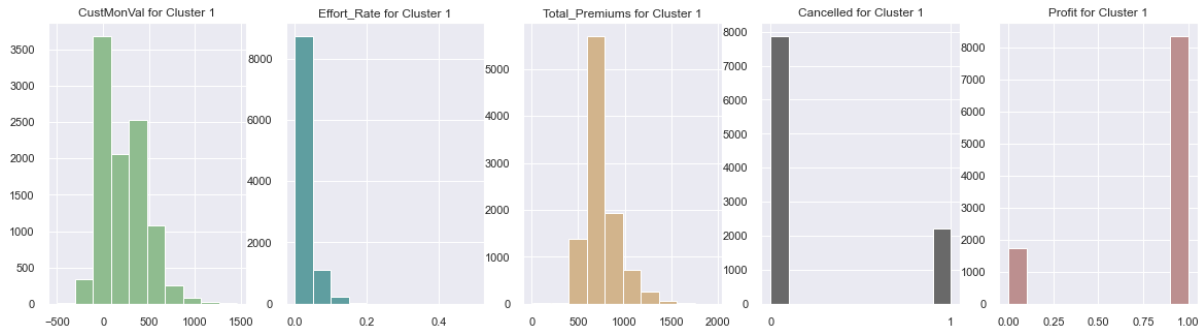

Figure 11: DBSCAN value perspective histogram

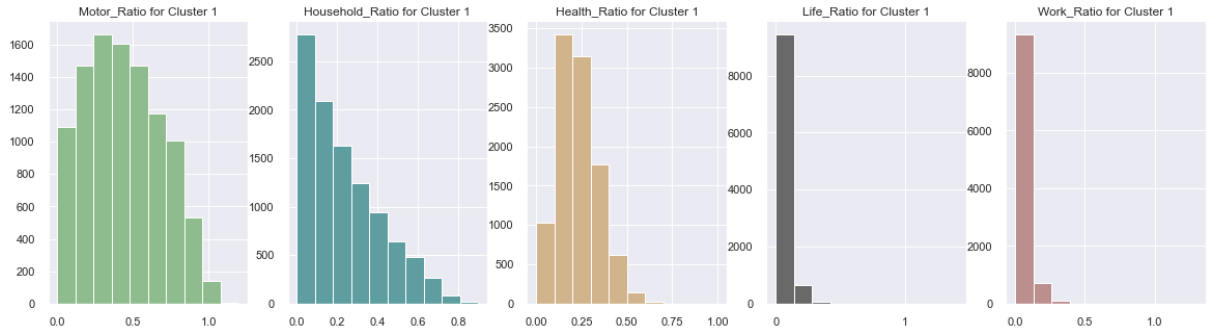

Figure 12: DBSCAN consumption perspective histogram

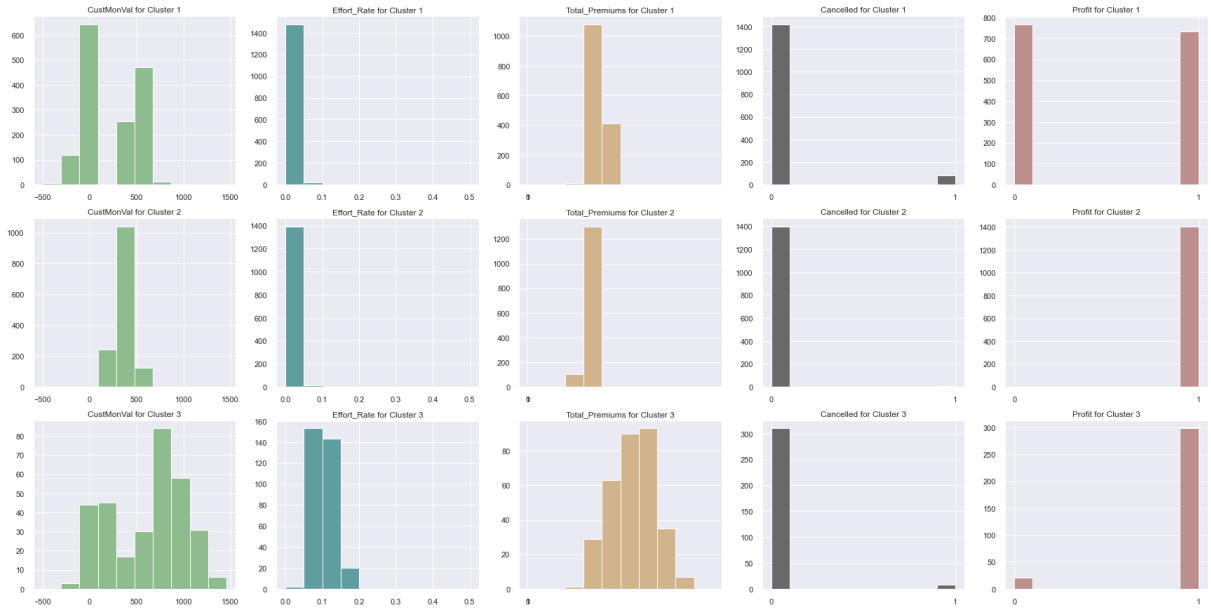

Figure 13: K-Prototypes Value perspective histogram

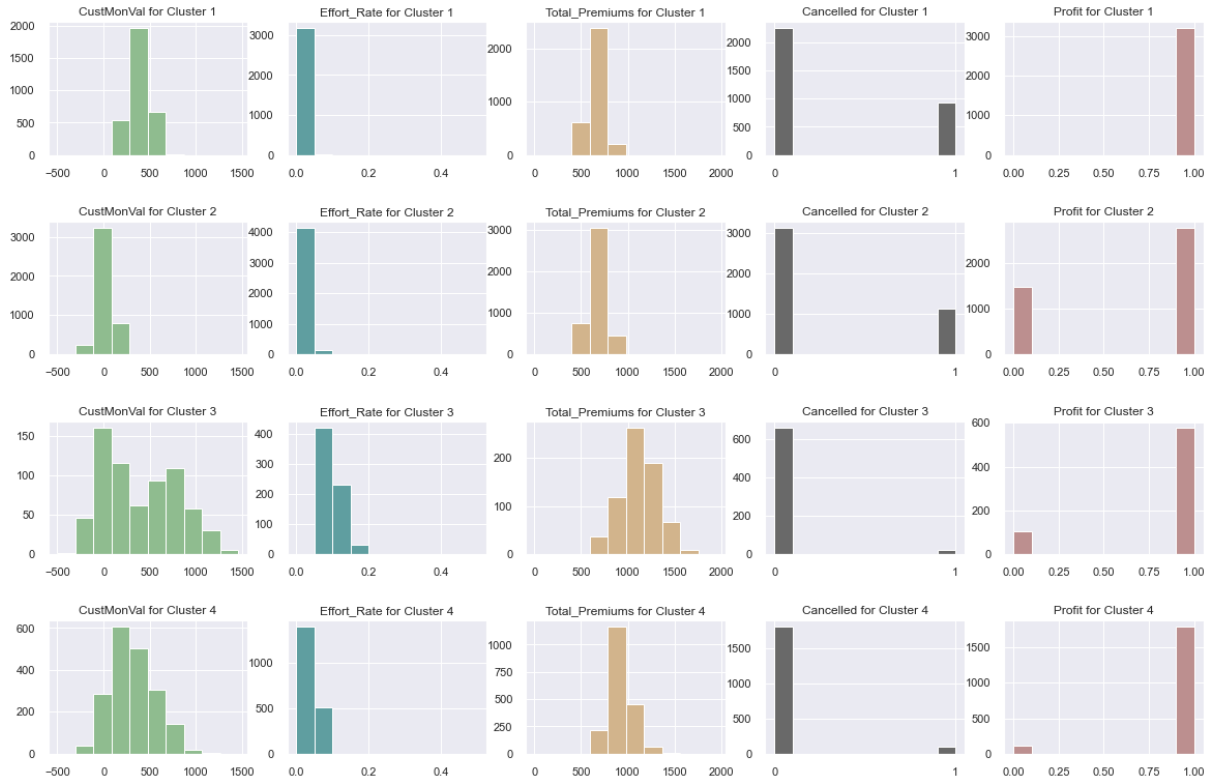

Figure 14: K-Means + Hierarchical Value perspective histogram

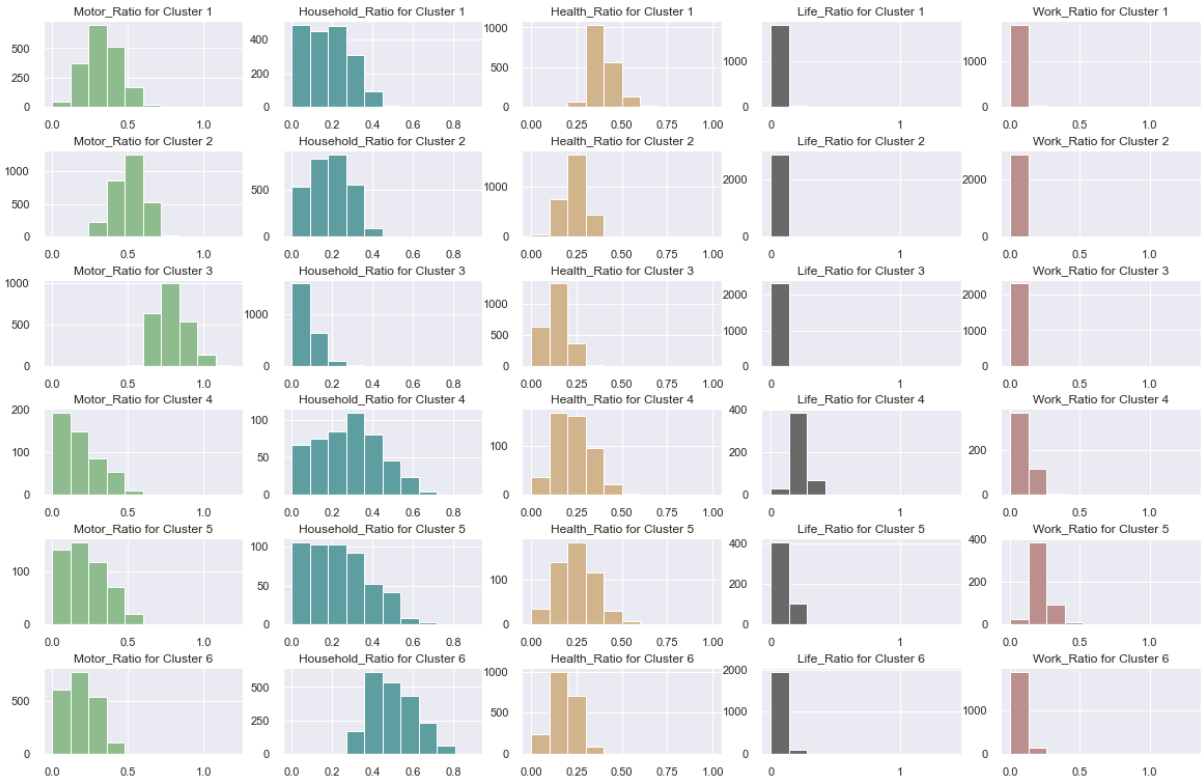

Figure 15: K-Means + Hierarchical Consumption perspective histogram

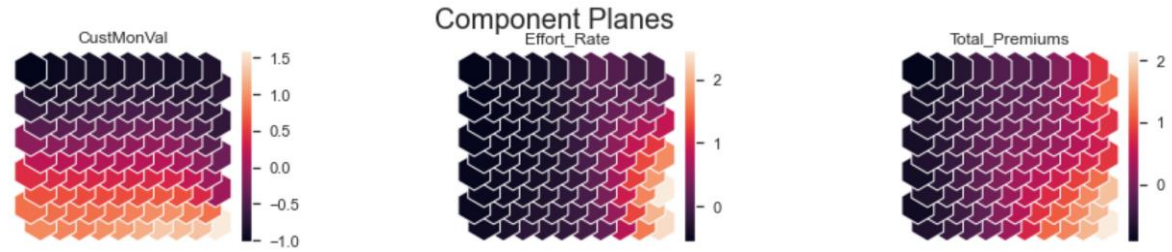

Figure 16: Component Planes Value Perspective

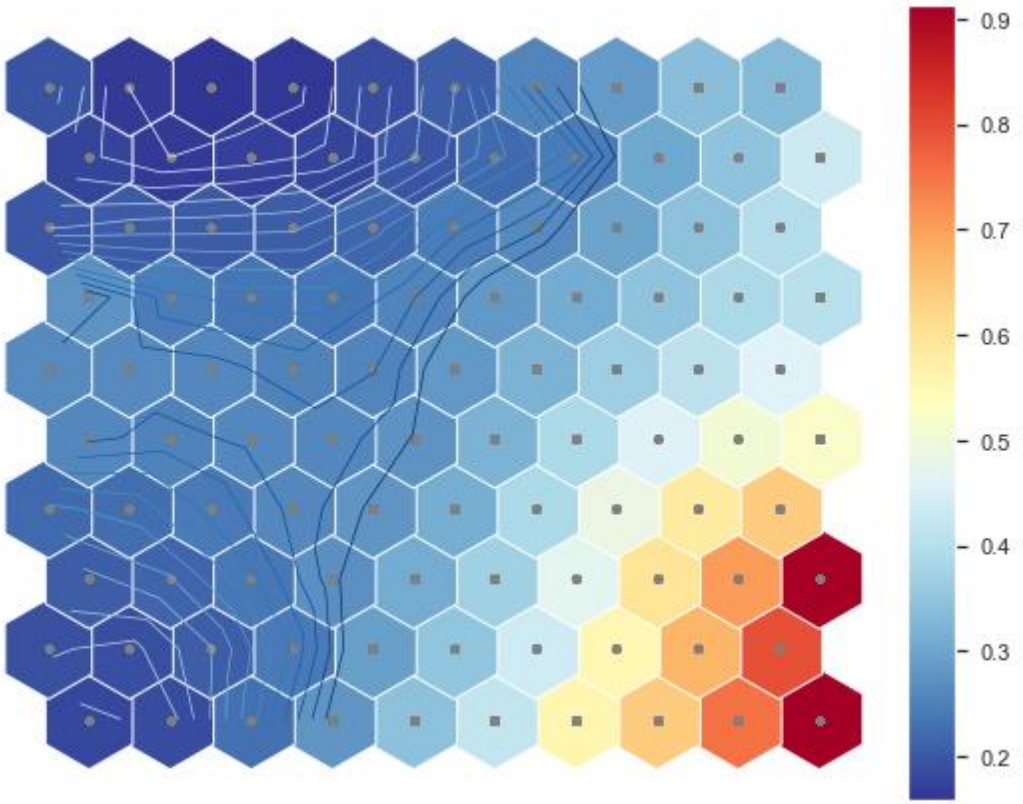

Figure 17: U-Matrix Value Perspective

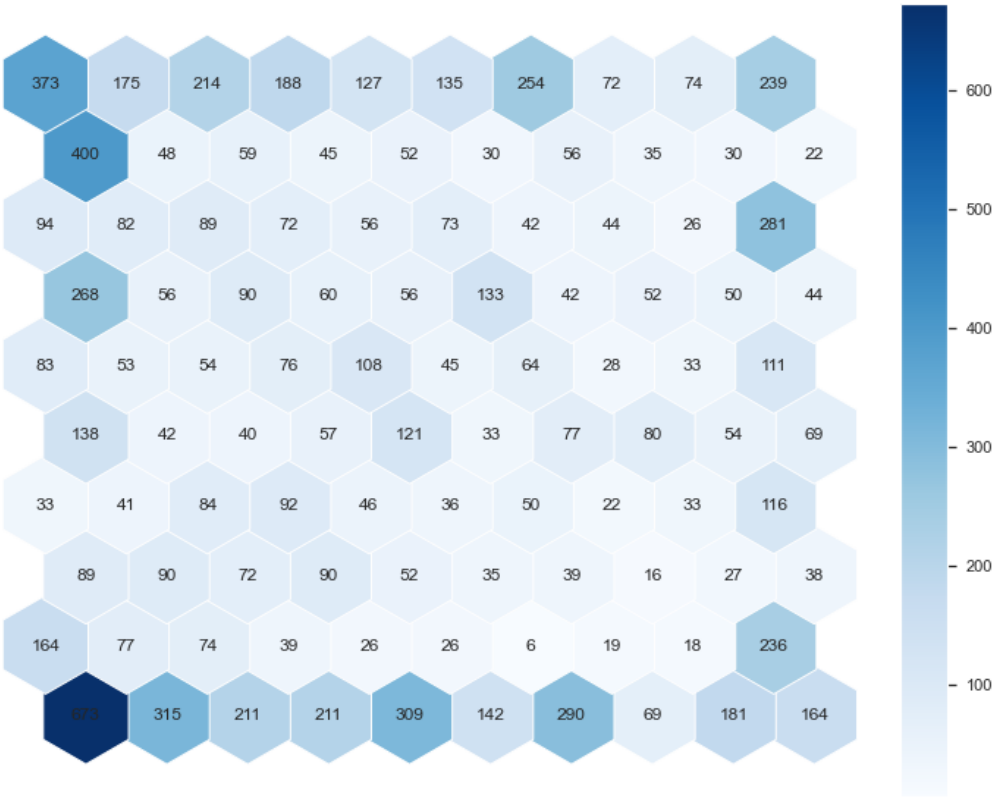

Figure 18: Hit Map Value Perspective

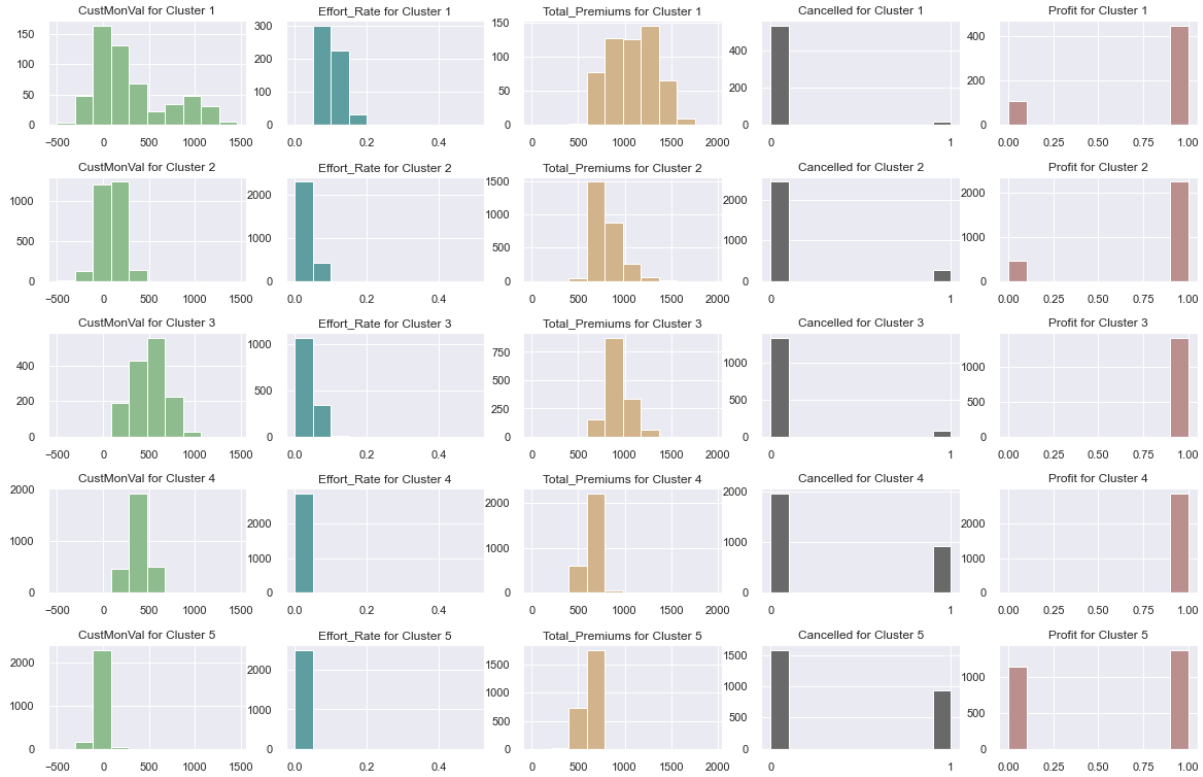

Figure 19: SOM + Hierarchical Value perspective histogram

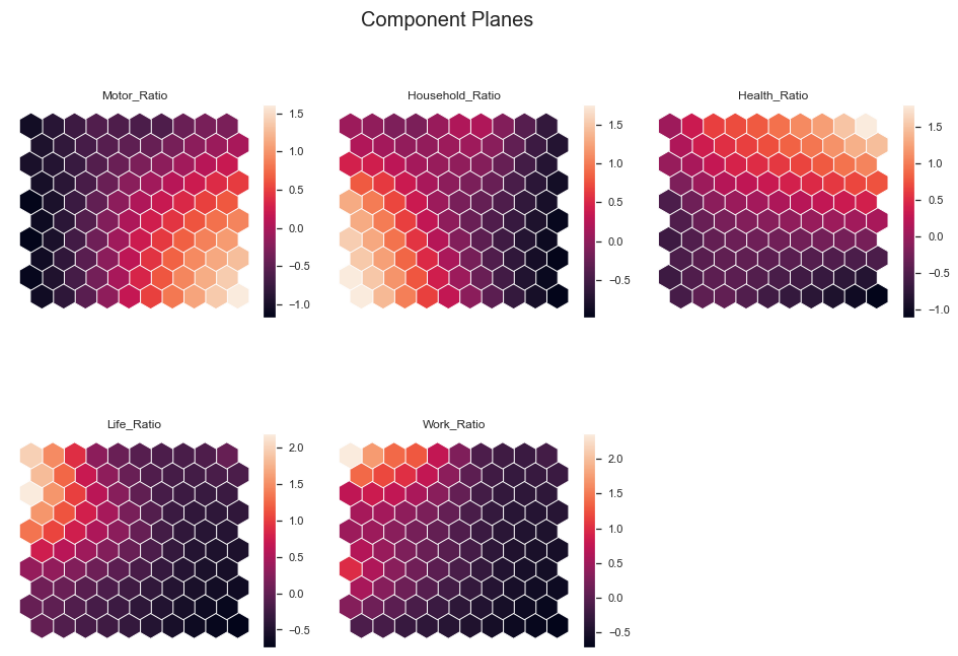

Figure 20: Component Planes Consumption Perspective

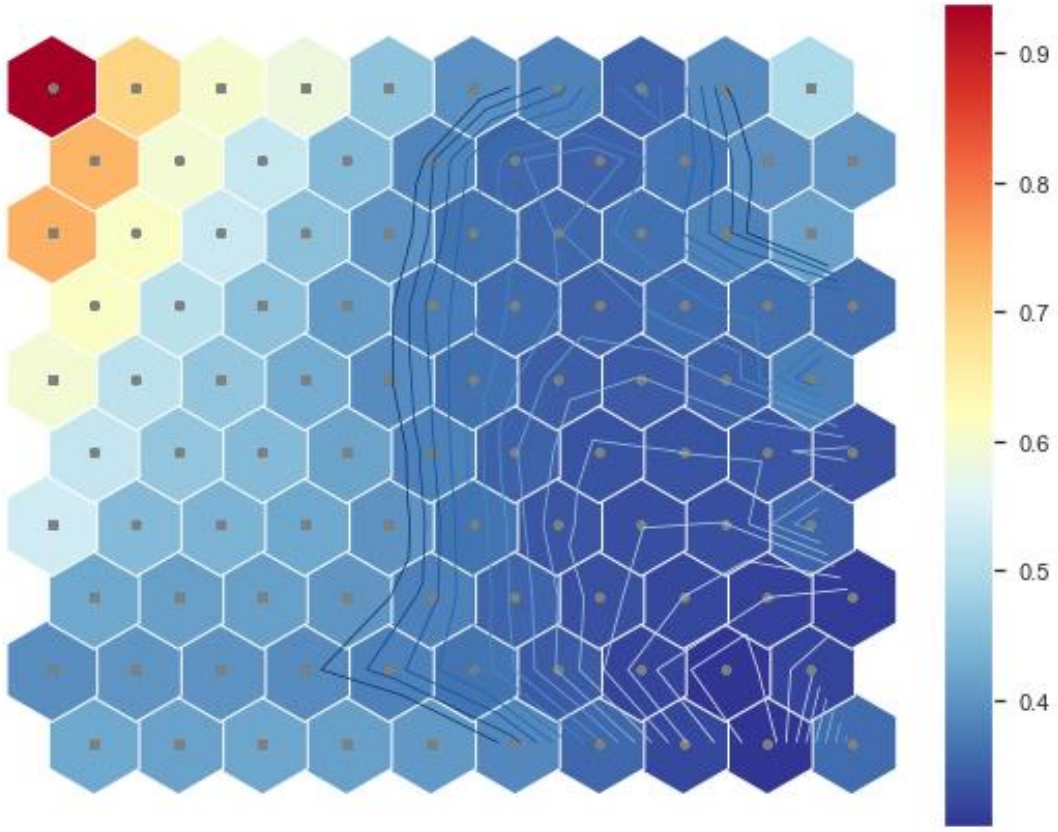

Figure 21: U-Matrix Consumption Perspective

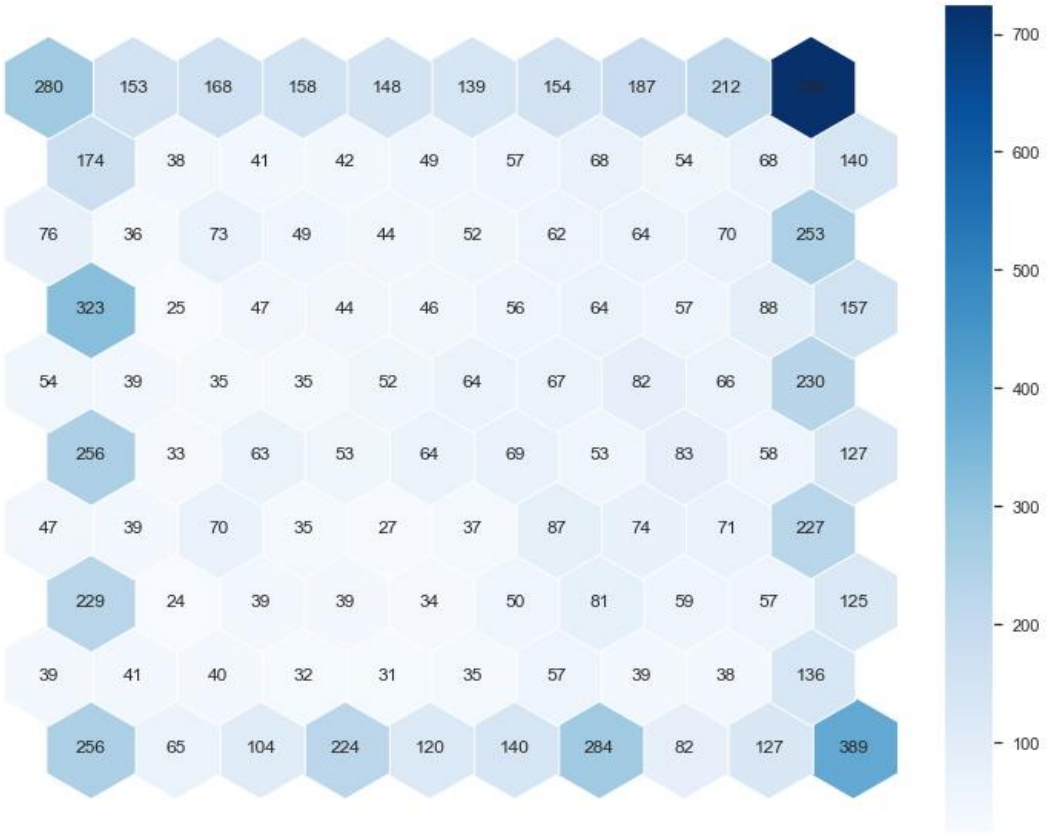

Figure 22: Hit Map Consumption Perspective

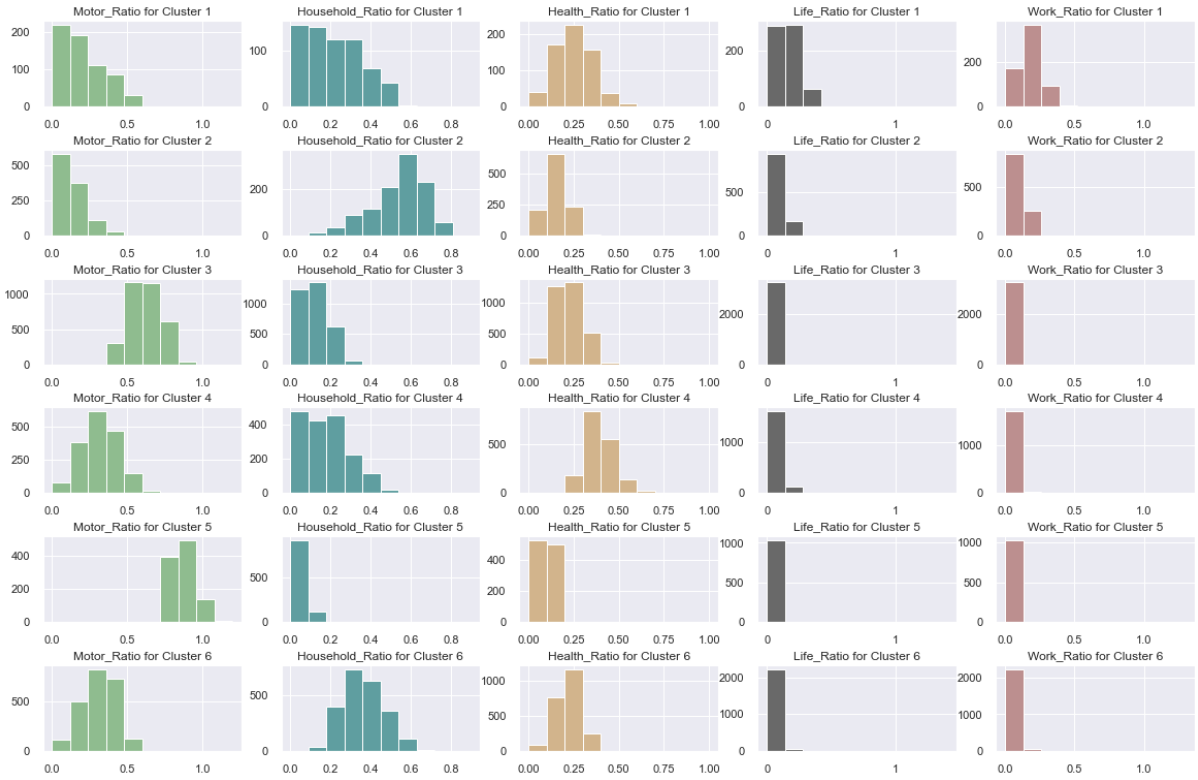

Figure 23: SOM + Hierarchical Consumption perspective histogram

Figure 24: Decision Tree