Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Customer Segmentation

Group AC

Alex Santander number: 20220658

Daila Alexandre number: r20191182

Diogo Silva number: 20221393

January, 2023

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

INDEX

1. Introduction

The goal of this Data Mining project is to segment the customers of a Portuguese insurance company (A2Z) based on their characteristics, features, and patterns found in the ABT that was provided.

It is intended to create clusters and assess customer profiles, providing the marketing department with valuable information to better understand their customers and create targeted marketing strategies. This will help the company optimize its resources, reduce costs, and maximize customer satisfaction by personalizing its approach to each segment of clients and improve overall results.

The project was developed in Python using a Jupyter notebook, applying content taught in Data Mining classes and learned in self-study.

2. Data Exploration and Treatment

2.1. General Information

This section has the objective to check initial information about the dataset.

There are 10296 rows and 14 attributes, one of them being the ID. These attributes are mainly divided in three groups: a customer's sociodemographic profile, such as Birth Year and Education Degree; the value that a customer spends in each Premium and the money that was not spent due to the cancellation of the package by the client, and customer value for the company such as Customer Monthly Value. The data type is initially float for all variables except *EducDeg* which is object. The table 1 in the annex has more information about each attribute.

Upon reviewing the summary statistics (table 2), it was noticed that some of the metric variables may contain outliers, as indicated by the wide range between the minimum and maximum values and the 1st and 4th quartile. Inconsistencies were also observed, such as clients with birth years that fall outside of the expected range. In addition, some of the categorical variables were found to have missing values. These issues will be addressed later when handling outliers, missing values, and ensuring the data is coherent.

2.2. Transformation of Variables & Data Types

The variables *FirstPolYear*, *Age*, *GeoLivArea*, and *Children* are naturally integers but it is necessary to verify if there are any decimal values, if there are, these values should be handled appropriately before the data type is changed. In case, there are no decimal values, the features will be converted into after treating missing values.

The index *CustID* data type was changed to integer, the variable *EducDeg* was transformed to ranks to make the data easier to understand by removing unnecessary information and make it usable in the clustering methods and finally, *BirthYear* was converted to *Age* in order to increase its interpretability.

2.3. Duplicates

Checking for duplicates is a key step when preparing the data for segmentation. When doing this, the group realized there were three duplicate rows. After carefully examining them, it was decided to remove the duplicates from the dataset, because even though they have different IDs, these observations showed the same values for all variables and have a low possibility of being a different person.

2.4. Features Distributions

A visual representation of each feature was created using histograms. For categorical features, the absolute frequencies were plotted (Figure 1). For metric data, the data was discretized and plotted, with the axis limited, excluding outliers, in order to improve the visualization of the distributions (Figure 2).
The variables *PremHousehold*, *PremLifem*, *PremWork*, and *CustMonVal* all exhibit a right-skewed distribution. On the other hand, *MonthSal*, *PremMotor*, and *PremHealth* show almost normal distributions.

2.5. Incoherences

There are several constraints that should be applied to the data in order to ensure its quality and accuracy. These include:

1. The *Age* variable should be between 0 and 110, as people can't live up to 110 years.
2. A person under the age of 16 should not have a value for the *EducDeg* variable indicating that they have a BSc/MSc or PhD degree.
3. A person under the age of 16 should not have a value for the *MonthSal* variable.
4. The variables *FirstPolYear*, *BirthYear*, *MonthSal*, and *ClaimsRate* should not have negative values.
5. The value of the *PremHousehold*, *PremLife*, *PremWork*, and *CustMonVal* variables should not be larger than the annual salary, which is calculated as the *MonthSal* variable multiplied by 12.
6. The *MonthSal* variable should be at least 530, which is the minimum wage in Portugal in 2016.
7. It is generally assumed that parents are over the age of 18, so individuals with *Age* below 18 should have 0 *Children.*
8. People who have *MotorPrem* are assumed to drive, which are usually with *Age* over 18.
9. People who don't have *MonthSal* shouldn't have *PremWork* (if they don't have a salary, they don't have a job and don't have an insurance)
10. The *FirstPolYear* variable should be between the value of the *BirthYear* variable and 2016.

These are some general assumptions that could be made about the data based on common patterns or trends that are often observed in real-world situations. However, there may be exceptions or unusual circumstances that do not follow these assumptions. For each one of these constraints, a careful analysis was made as follows:

1. An individual has 998 years of *Age* which is clearly an error and so was be removed.
2. There is no incoherences.
3. There are 12 individuals under 16 with a salary which is impossible by Portuguese standards, however, there are international clients and so, they will remain the same.
4. There are no incoherences.
5. There is one person whose annual value in insurance exceeds the annuity salary, however, other people help in the payment, such as parents, so it may not be incorrect.
6. People between 16 and 18 can only work part-time, thus may receive less than the minimum wage, so this constraint does not apply for these individuals. It is also usual for people with ages between 18 and 25 to work part-time, so thew will not be considered as incoherences.
7. Although the group couldn't find the teen pregnancy rates in Portugal, in America 'Teen birth rates' continued to decline from 17.4 per 1,000 females in 2018 to 16.7 per 1,000 females in 2019' - this is from Ages 13 to 19. Considering that Portugal should have a

value similar to these, the value seems reasonable. Also, keep in mind the clients are not just portuguese.

8. There are 210 individuals under 18 with Motor Premium. It is a low number that could be explained by minors driving 50 cc vehicles or foreign clients.
9. There are no incoherences.
10. There are 1997 people with the First Policy year before the Birth Year. This is a significant number and so a deeper analysis was made. The group analyzed *FirstPolYear* and *Age* correlations with other variables for: original dataset without missing values, dataset where the values of these variables for these observations were switched and dataset without the incorrect observations. Despite the changes made, the correlation of the two variables with the remaining appears to have decreased. The *FirstPolYear* variable is suspected to be incorrect due to its low correlations with every variable, as it should at least have a mild/high correlation with *CustMonVal*, as it is expected that clients who have been with the company for a longer period of time would have brought more value to the company. As a result, the *FirstPolYear* variable will be removed.

2.6. Outliers

Outliers' detection is the process of identifying values or unusual observations in a specific dataset that are different from the rest of the values. This step is important because it has a significant impact on the results of the statistical analysis, such as standard deviation, mean and median and in some of the clustering methods.

At first glance, there are potential outliers when looking into the statistical description of the dataset (Table 2) as mentioned in the General Information section. The distance between the 75% quartile and the maximum value in most features is large. There is only one feature with a minimum value very far away from its 25% quartile (*CustMonVal*).

For a better understanding of the study data, visual representations have been created: Boxplots and Histograms. Box Plots (Figure 3) have shown the presence of outliers in all features, except 'Age'. Histograms (Figure 4) also show the same outlier presence and highly skewed data.

The first insight that can be withdrawn is that the data contains a great amount of outliers. Most of the variables have outliers, as seen by comparing the boxplot and histogram for each feature. Some of these outliers, such as those in *MonthSal* and *PremHealth*, are considered 'true' outliers as they are far from the rest of the distribution. However, the outliers in *PremLife*, *PremHousehold*, and *PremWork* are simply the skewed portion of the distribution.

To handle these outliers, IQR was the first method tried. The IQR is a robust measure of dispersion that is less affected by outliers than other measures, such as the standard deviation. However, using 25 and 75 quartiles within a 1.5 threshold led to a loss of 14% of the data, which is a substantial quantity of data to be eliminated. As an alternative, a filter method was the best choice and the threshold for each variable was carefully chosen.

At the end of the outlier's elimination process, 310 observations were deleted keeping 96.65% of the dataset rows.

2.7. Missing Values

The first step to detect missing values was to see if the data contained special characters that represented the missing values. So, before finding NaN values, those special characters were transformed into NaN values as well.

There are 8 columns with missing values, 324 empty cells in total, distributed in 275 rows. None of the features has so many missing values that makes it unusable.These quantities represent 2.76% of the total rows in the dataset (table 3).

It is assumed that the lack of data for the Premium variables means that the individuals do not have that insurance and therefore the real value is 0 (as they do not spend any money in that premium). As seen in the incoherences section, *Age* and *MonthSal* have a very high correlation, and so KNN imputation is used to fill each-other's missing values with k=5. The missing values of *EducDeg*, *Children*, and *GeoLivArea* are deleted as there are so few values missing. Making the deleted observations 3.497% of the dataset.

2.8. Visualization

Before clustering, it is important to visualize the data to see beforehand if there are any pattern trends or relationships and to identify any potential issues besides the ones already treated. Histograms and boxplots of all variables were already assess in the Outliers section.
The first visualization is the Pairwise relationships between numerical features. This view refers to the relationships between pairs of variables in a dataset (figure 5).
As a result of this view, there were two principal statements:

- *Age* and *MonthSal* present a strong positive relationship.
- *ClaimsRate* and *CustMonVal* present a atrong negative correlation.

The second visualization is Discriminant Power Boxplots. This boxplot summarizes the distribution of the discriminant power scores of the categorical features using a set of summary statistics, such as the median, upper and lower quartiles, and minimum and maximum values. These are the conclusions for each categorical feature:

- *GeoLivArea*: No significant discriminating power (Figure 6).
- *Children*: Has some discriminating power in features *Age* and *MonthSal* (Figure 7).
- *EducDeg*: Has strong discriminating power (Figure 8).

2.9. Feature Engineering

2.9.1. New Variables

By creating new variables that are derived from existing ones, there is the possibility of uncovering patterns and relationships that may not be immediately apparent. Also, the accuracy of your data mining model improves by providing additional information or context.
New variables created are the following:

1. **Total_Balance**: Represent the sum of all premiums for each client, considering the cancelation of packages as negative values.
2. Each client has a different mix of premiums purchased, and not all of them have the 5 premiums the company is offering. Some of them still have the possibility to buy more premiums and others canceled the premiums they had. For that reason, three features have been created:
    1. **Number_Prems**: This feature reflects the number of premiums each customer had purchased, and is important because there are a 30% of unsubscripted packages among all customers.
    2. **Num_cancaled_subscriptions**: Represents the number subscriptions canceled for each customer.
    3. **Money_from_cancelations**: Is related to Canceled Subscriptions features. In this case, is the monetary value lost suffered by the company due to the client's decision to cancel one of theirs premiums.
3. **Money_spent_by_client**: Is money corresponding only to the active premiums.
4. **Money_spent_over_salary:** Ratio of what the customer spends in relation of their annual salary.

5. Ratios of each premium devided by the total balance were created: **Health_Ratio, Life_Ratio, Motor_Ratio, Household_Ratio and Work_Ratio**

2.9.2. Scaling

Scaling refers to the process of transforming variables to a common scale. This is a necessary process because variables may be measured on different scales, affecting the accuracy and interpretability of models.

Robust scaling was chosen for normalization because many of the variables are skewed and a large number of outliers were retained. This method is suitable for handling such data. Robust scaling is also known as robust normalization or robust standardization, is a method of scaling variables that is less sensitive to the presence of outliers in the data. Unlike standardization or min-max scaling, which are based on the mean and standard deviation or range of the data, robust scaling uses the median and interquartile range (IQR) to transform the data.

2.10. Correlations

After creating the new variables, it was checked if any of the numerical variables were univariate (with variance equal to 0) in order to drop them if existent, which was not the case.

A correlation matrix is a table showing the correlation between multiple variables. It is a useful tool for identifying the relationships between different variables and understanding how they may be related.

The dataset had close to normal and skewed distributions. Therefore, using only Pearson distribution for skewed features may not be appropriate. In this case, Kendall's tau correlation was aslo incorporated into the analysis; the correlations results can be seen in table 4.

After visualizing the correlations matrix, the following was observed:

- *MonthSal* and *Age* have a very high correlation and so, *MonthSal* is chosen as the potential to buy packages is more important than how old the clients are;
- *CustMonVal* and *ClaimsRate* have a very high correlation, but *ClaimsRate* is decided to be more important as it is based on the last 2 years and not since they became clients, which maintains the relevance of the more recent clients;
- *TotalBalance* and *Profit* have an extremely high correlation with *PremHousehold* and to each-other.
- Ratio variables have a high correlation with the variable they are based on, so only the ratio variable or the original variable can be used. - After trying both, the original variables were chosen as it led to better results.

2.11. Perspectives

Following the project description suggestions, three perspectives were created using the selected features: the **Sociodemographic** perspective with *Children*, *MonthSal* and *EducDeg*; the **Product** perspective with *PremHealth*, *PremHousehold*, *PremLife*, *PremMotor* and PremWork; and the **Value** perspective with *ClaimsRate*, *Money_spent_by_client*, *Money_from_cancelations*, *Money_spent_over_salary* and *Num_Prems*.

Once these perspectives contain new variables, the discriminatory power of the new categorical features (*Num_Prems* and *Num_canceled_subscriptions*) was assessed. (Figures 8 and 9)

2.12. PCA

Principal Components Analysis (PCA) was employed to see if a smaller input space could be created and better results achieved. This method is designed to measure variation and is more effective when used numeric features rather than categorical ones. Therefore, PCA was not applied to the **Social demographic** perspective, as it only has one numeric variable. Additionally, the company wants to understand which types of insurance are most interesting to customers, so PCA was not performed on the **Product** perspective in order to analyze them separately. As a result, PCA was only carried out on the **Value** perspective.

The Scree Plot and the Cumulative variance (figure 11) indicate that only one principal component should be retained, which accounts for nearly 100% of the variance. However, upon examining its histograms (figure 12), no clear clusters are visible. The results are not very interpretable or meaningful, and clustering methods applied to the original features produce better results. As a result, PCA is not used for dimensionality reduction for the rest of the project.

3. Clustering

In this section, several clustering methods were applied, taking into account the details of each of the perspectives.

In advance, it was determined that Gaussian Mixture would not be used for the **Product** and **Value** perspectives due to the fact that only two out of the three selected numeric variables showed a distribution close to normal, as determined by the normal Quantile-Quantile test (figure 13). Both DBScan and Mean Shift were also not used for the same two perspectives because they create clusters with a small number of observations, indicating that the data may be sparse or that the clusters have different densities. However, DBScan was still used to identify and remove 50 outliers from the dataset, making the percentage of deleted data 4%.

Furthermore, Hierarchical clustering and Hierarchical clustering on top of SOM for the same perspectives (**Value** and **Product**) provided poor results - as it also led to created clusters with a very small number of observations.

To determine the optimal number of clusters using K-means, the elbow method was utilized, silhouette score, and dendrogram analysis. The dendrogram was generated by applying hierarchical clustering to a larger number of clusters created through K-means clustering.

3.1. Sociodemographic

3.1.1. K-Prototypes

Two thirds of this perspective are categorical variables and only one numerical, causing most clustering methods to be discarded. K-prototypes, which is a variant of K-Means that can handle categorical data, was performed.

Plotting the cost function (WCSS) (figure 14), the elbow curve does not clearly indicate the number of clusters to use as it does not have a defined elbow, but it curves in 3 clusters. So, in order to achieve the optimum number of clusters, a hybrid hierarchical clustering was also performed using the ward method to calculate the distance between clusters. Observing its dendrogram (figure 15), the biggest jump is achieved when 3 clusters are formed.

Therefore, 3 clusters were formed using K-Prototypes and upon profiling them, it was concluded that they provided good results. By analyzing the variables' distributions (figure 16), it was possible to conclude that the first cluster consists only of individuals with children, while the third cluster consists only of individuals without children, these two present similar distribution in Education and similar range for Month Salary, however its distribution shows that people without children tend to have higher salary. Most people in the second cluster have children, but they differ from the rest by their lower salary and slightly lower education. The number of individuals seems well distributed across the 3 clusters (figure 17).

3.2. Value

3.2.1. K-Means

After looking at the outputs of the previously methods (elbow methods and dendogram), the number of clusters first chosen was 4, but, as it led to 2 clusters with very few observations.Therefore, K-Means was performed with 2 clusters. Profiling the clusters, even though they seemed to be clearly separated (figure 18), one cluster only had a few individuals while the other contained most of the individuals in the data set (figure 19). Leading to the conclusion that the cluster with few elements may be outliers. This clustering was not selected.

3.2.2. Self-Organizing Maps

In this perspective, the group also applied a SOM algorithm as a visualization tool. The grid used was a 50x50 neurons. Learning rate and neighborhood size hyperparameters were used in its default values.
Silhouette plot and mean by cluster was also taken into consideration.
From the component planes visualization (figure 20), the feature *Money_from_cancelations* shows very clear a big cluster and another small in the corner. *Money_spent_over_salary* has two clusters and *ClaimsRate, Money_spent_by_client* seems to have three but not very explicit.
A U-Matrix (figure 21) was also constructed in order to spot clusters. By its shape it is visible that there are three clusters but a decision can not be made based only in the visualization of the heatmap.

3.2.2.1. K-Means on top of SOM

Som is a powerful algorithm, but the interpretation of the heatmap is subjective. Applying K-means on top of the SOM output, helps to see and determine the number of cluster for this perspective.
After analysing the profile of the clusters it was concluded that there was a high value cluster that did not have any cancelations of packages, has the highest amount of money spent has the highest ratio of money spent over salary. There is also a low value cluster as it has all the clients that cancelled packages, and has the lowest amount of money spent. The third cluster is of medium value as all the metrics are in between the metrics of the previous clusters.
In this section the visualizations are analysed:

- Silhouette Score analysis: cluster 0 has a strong representation, while cluster 2 exhibits some miss represented values and cluster 1 has a small proportion of misclassification. Overall, the results are still satisfactory. (figure 22)
- PCA 2 Dimensions: two clusters show the same values for the 2nd Principal Component and differ in the 1st PC where one shows higher values. The remaining cluster rounds around the 1st PC where the two other meet and dispersed along the 2nd PC (figure 23)
- t-SNE 2D: Two clusters are located in the middle and are separated by an almost linear boundary, while the remaining cluster is situated around them, forming a wrapping shape (figure 24)
- Analyze UMAP 2D - similar to t-SNE but the cluster that is evolving the two others is further apart (figure 25)

These visualizations lead the group to conclude that although the three clusters are different, there is one further away from the other two.

3.3. Product

3.3.1. K-Means

After looking at the output of the previously mentioned methods the number of clusters decided was 4 based on the dendrogram and Elbow Method (figures 26 and 27). In the analysis of the profile of each cluster, these seem to be well segmented, with clear differences as seen in figure 28 and different ranges of each variable (figure 29).

3.3.2. Self-Organizing Maps

For the final perspective, the group applied the same approach used in value perspective. From the component planes visualization (figure 30), it can be estimated that both *PremHealth* and *PremMotor* seem to have two similar clusters (one located in the upper right corner and the other at the bottom, more towards the left). The rest of the features seem to have between two and three clusters. From the U-Matrix (figure 31) it is also possible to identify three clusters. However, similarly to the value perspectives, these visualizations are not so clear and so other approaches are taken into account.

3.3.2.1. K-Means on top of SOM

In the same manner to what was done for the value perspective, K-Means was applied on top of SOM to help in visualization and cluster association.
Through the analysis of the profiles it is clear that there is a clear predominance of different prems in different clusters.One cluster has a predominance of the PremHealth, other has a predominance of PremMoto and the last one has a predominance of PremHousehold, PremLife, PremWork.
In this section The group presents the following visualizations:

- Silhouette Score: The silhouette scores for cluster zero show good representation, while clusters one and two have some misclassified values. However, overall the silhouette scores are still satisfactory (figure 32).
- PCA 2 Dimensions: The clusters form a pyramid shape, with three layers. The elements in the bottom layer are more sparse compared to those in the top layer, indicating that individuals in one cluster are more similar to each other than those in the other cluster (figure 33).
- t-SNE 2 Dimensions: a cluster seem to start where the other ends, cluster 1 and 2 are further apart and cluster 0 stands in the middle (figure 34).
- UMAP 2 Dimensions: Umap show the same figure as t-SNE but inverted with cluster 0 in the center, however this one shows a few outliers leaving the aggregate that is made up of the 3 clusters (figure 35).

This clustering method was chosen over ‘simple’ K-means since it has a better separation of the relevant prem packages of each cluster.

4. Final Clusters

The clustering methods that were merged were K-prototypes with 3 clusters for the **Sociodemographic** perspective, K-means on top of Som with 3 clusters for the **Value** Perspective and K-means on top of Som with 3 clusters for the **Product** Perspective.
Before merging them, the clusters were analyzed and the numbers they were assigned to were switched to descriptions of the cluster, in order to facilitate interpretability.

The clusters in the **Product** perspective were changed to the corresponding Prem package that had the highest value. The clusters in the **Sociodemographic** perspective were changed to small descriptions (such as ‘0_high_sal_No_children__higher_educ’). The clusters in the **Value** perspective were changed to ‘High_value’, ‘Medium value’ and ‘Low value’.

After grouping the dataframe with the selected features and the 3 perspectives’ clusters, the grouping arrived at the table 4 where the names of the clusters were displayed.

From that dataframe, the dendrogram was created in order to decide the number of final clusters that would be chosen. As seen in figure 36, the selected number of clusters was 8, which led to 3 clusters with barely any observations, and, through trial and error, 5 final clusters were chosen.

From table 5 and the profiling it (figure 37) is visible that cluster 0, which contains 566 clients, is characterized by having a high income, few canceled packages, high PremHealth, a medium value of all Prems;

Cluster **1** contains 2033 clients and is characterized by having low salary, very high PremHousehold and PremWork, High money lost to cancellations and High money spent by client over salary;

Cluster **2** contains only 7 clients, and could possibly be ignored, but the group decided not to. This cluster has a very high salary, Very low ratio of money spent over salary, very high claims rate;

Cluster **3**, containing 1180 clients, has as a common denominator a low salary, very low money lost from cancellations, very high PremMotor and very low PremHealth;

Cluster **4** contains 6094, being the largest cluster, and is characterized by having high money lost from cancellations.

The visualization graphs of the final clusters are good, as there is a noticeable separation of the clusters. Although there is some overlap this is to be expected as the graphs are a simplification of a space with 11 dimensions (figure 39, 29 and 40).

5. Marketing Strategy

As the clients in cluster **0** are stable customers, with high income, and medium spending in Prems. As they seem to be loyal to the company it might be best to try and persuade them to spend more, ir order to so the company could try offering small discounts for upgrading their premiums or rewards for their loyalty.

The clients in cluster **1** are low-income individuals that spent a lot of their salary on the insurances so they were more prone to canceling the packages. The strategy suggested is to offer a discount in PremMotor if the client pays for PremHousehold and PremWork.

The clients in cluster **2** spend a lot on insurances, except in motor, which means there is potential to increase the sales. The proposed strategies are if the client has two packages,offer 10% discount in PremMotor; When they have more than 2 packages and one is PremMotor make it possible for them to earn a reward, like a raffle;

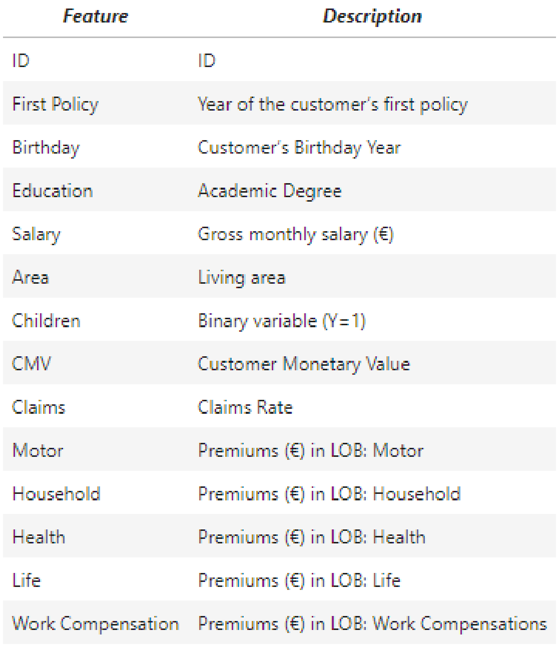
The clients in cluster **3** seem to be people that only have insurance for the vehicles with the company, not forgetting that it is mandatory by law to have this insurance. A possible approach could be that if they have PremMotor, a discount of 10% or 20% in PremHealth in the next 2 years.

The clients in cluster **4** seem to be mostly people that canceled premiums, and are possibly moving to another insurance company, in order to bring them back it is proposed to give a 30% discount in another premium of their choice for 1 year (with fidelity).

6. Conclusion

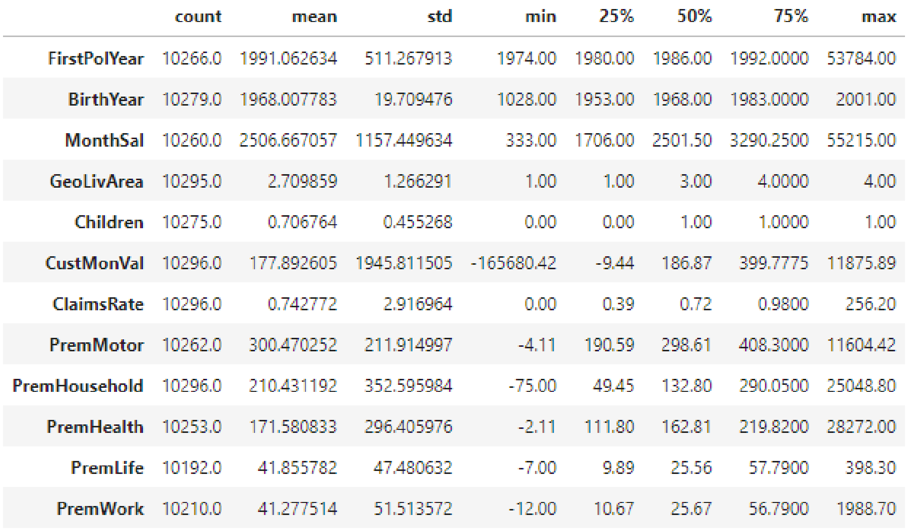
The group has successfully identified the relevant features, identified relevant customer segments based on which a marketing campaign could be stratized on. The number of final clusters chosen seem to be appropriate to the analysis.
Since the density based clusterings lead to bad results, it might suggest that the data is either sparse or has varying density. Since K-means clustering has the best performance, and it doesn't perform as well when the data has different densities, it suggests that the data is only sparse. Since K-means clustering is very sensitive to outliers, the good performance should mean that the group performed a good treatment of the outliers. Due to the nature of the data SOM is a good alternative. SOM is relatively tolerant of missing or noisy data, and still capable to find patterns in the data. Using K-means and SOM combined helps improve cluster accuracy, enhancing interpretability, and the ability to handle large datasets.
The proposed marketing strategies are not very specified as there is no more detailed information about them, so the company should carefully analyze them having in mind their budget focusing on loyal and potential clients such as in cluster 0.

7. Appendix



Feature	*Description*
ID	ID
First Policy	Year of the customer's first policy
Birthday	Customer's Birthday Year
Education	Academic Degree
Salary	Gross monthly salary (€)
Area	Living area
Children	Binary variable (Y=1)
CMV	Customer Monetary Value
Claims	Claims Rate
Motor	Premiums (€) in LOB: Motor
Household	Premiums (€) in LOB: Household
Health	Premiums (€) in LOB: Health
Life	Premiums (€) in LOB: Life
Work Compensation	Premiums (€) in LOB: Work Compensations

Table 1 – Features description



	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10266.0	1991.062634	511.267913	1974.00	1980.00	1986.00	1992.0000	53784.00
BirthYear	10279.0	1968.007783	19.709476	1028.00	1953.00	1968.00	1983.0000	2001.00
MonthSal	10260.0	2506.667057	1157.449634	333.00	1706.00	2501.50	3290.2500	55215.00
GeoLivArea	10295.0	2.709859	1.266291	1.00	1.00	3.00	4.0000	4.00
Children	10275.0	0.706764	0.455268	0.00	0.00	1.00	1.0000	1.00
CustMonVal	10296.0	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	0.742772	2.916964	0.00	0.39	0.72	0.9800	256.20
PremMotor	10262.0	300.470252	211.914997	-4.11	190.59	298.61	408.3000	11604.42
PremHousehold	10296.0	210.431192	352.595984	-75.00	49.45	132.80	290.0500	25048.80
PremHealth	10253.0	171.580833	296.405976	-2.11	111.80	162.81	219.8200	28272.00
PremLife	10192.0	41.855782	47.480632	-7.00	9.89	25.56	57.7900	398.30
PremWork	10210.0	41.277514	51.513572	-12.00	10.67	25.67	56.7900	1988.70

Table 2 – Summary Statistics

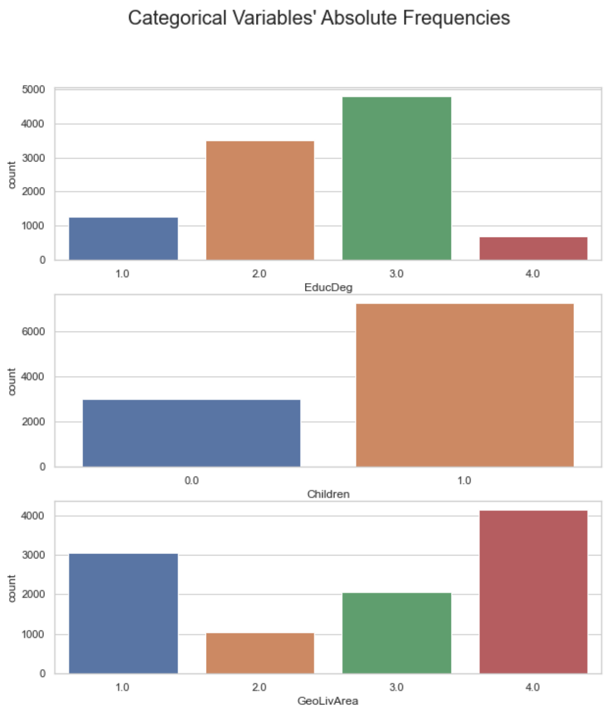


Figure 1 – Categorical Variables' Absolute Frequencies

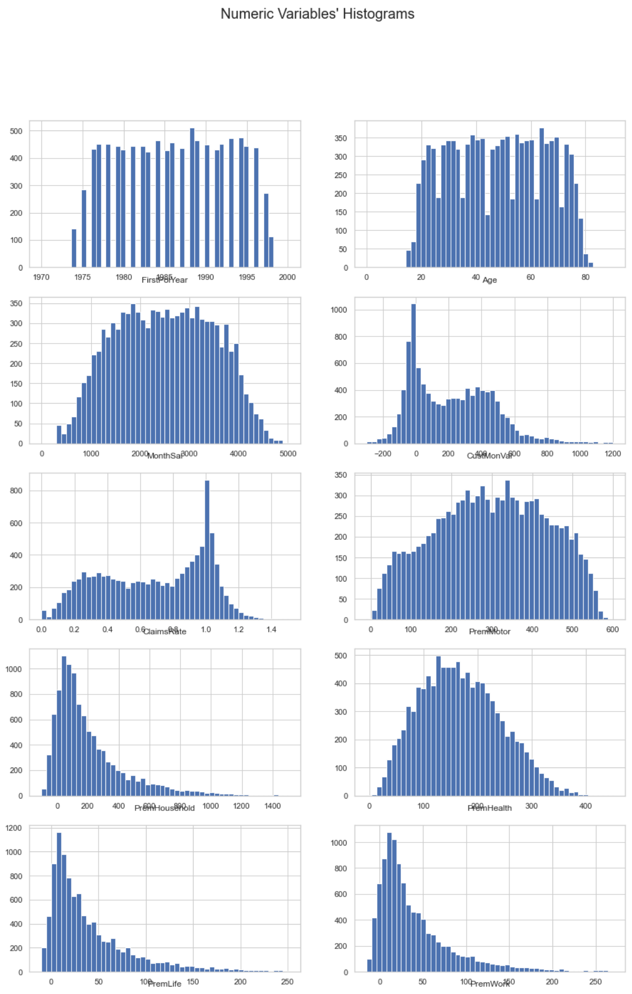


Figure 2 – Numeric Variables' Histograms

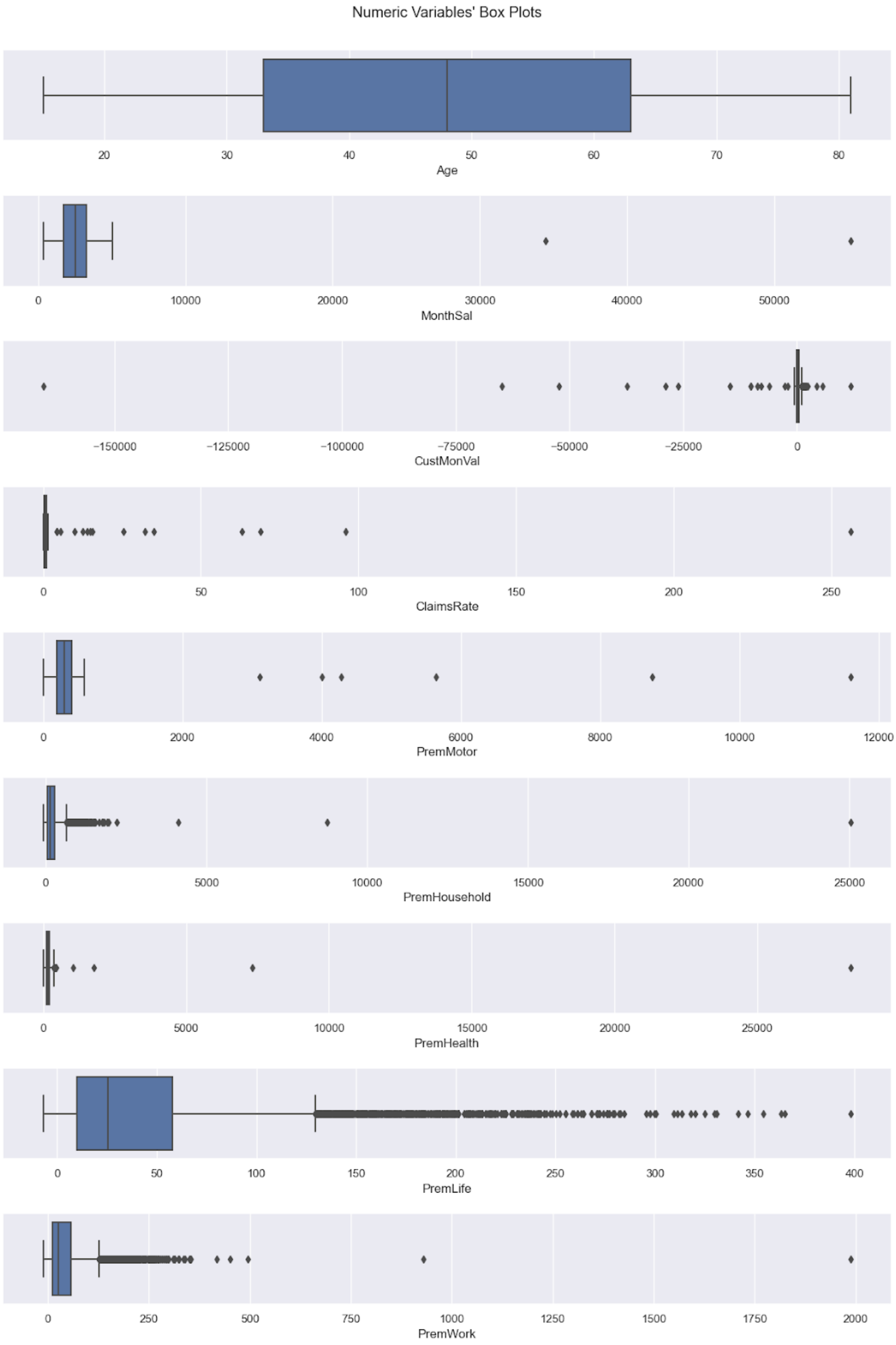


Figure 3 – Numeric Variables' Boxplot

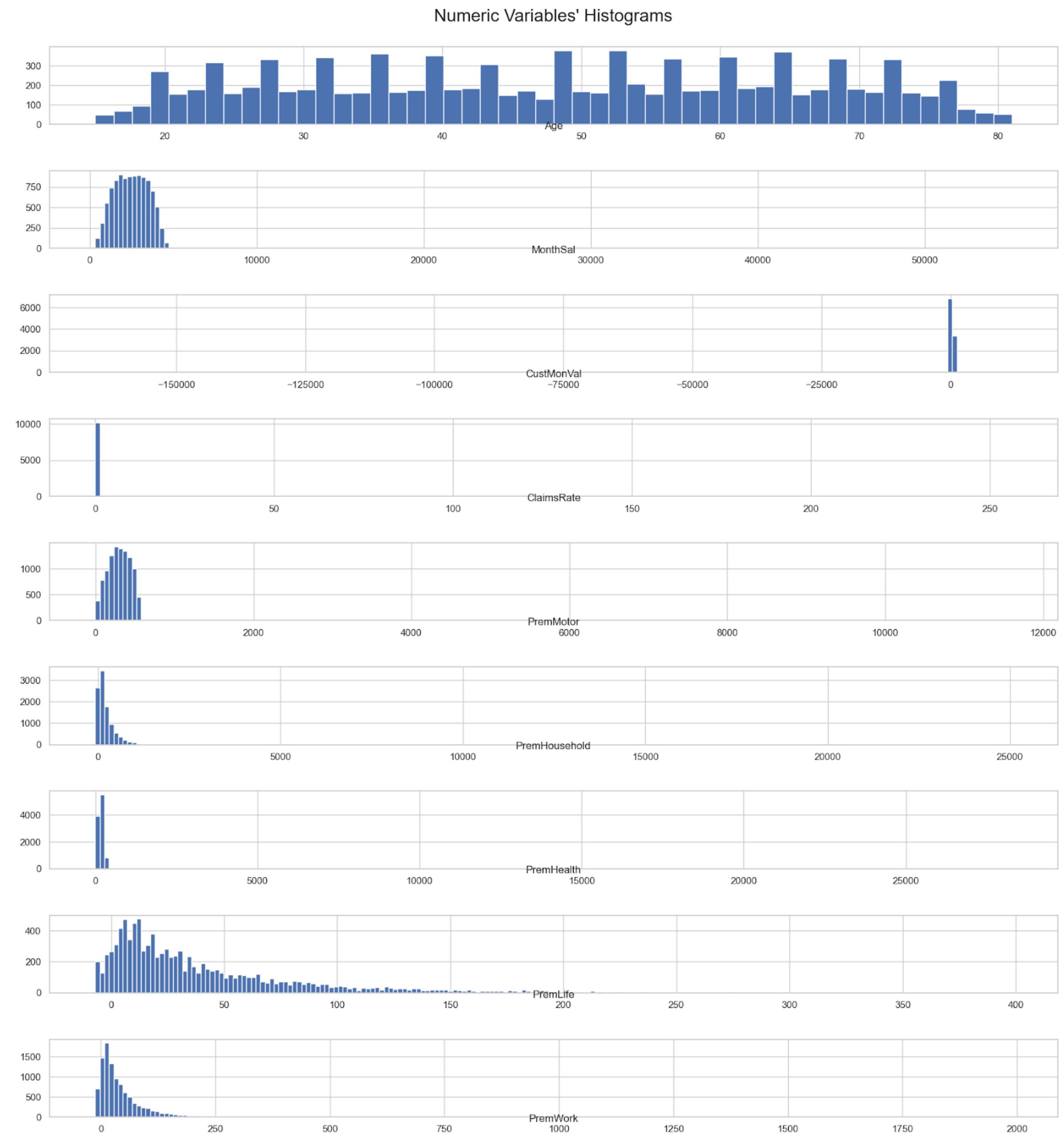
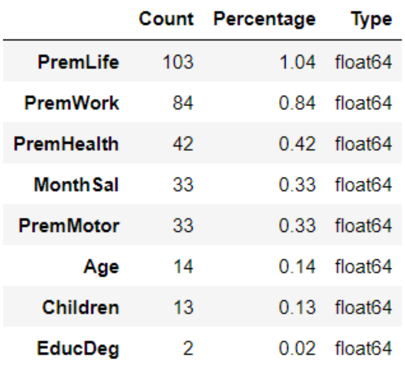


Figure 4 - Numeric Variables' Histograms



	Count	Percentage	Type
PremLife	103	1.04	float64
PremWork	84	0.84	float64
PremHealth	42	0.42	float64
MonthSal	33	0.33	float64
PremMotor	33	0.33	float64
Age	14	0.14	float64
Children	13	0.13	float64
EducDeg	2	0.02	float64

Table 3 – Missing Values

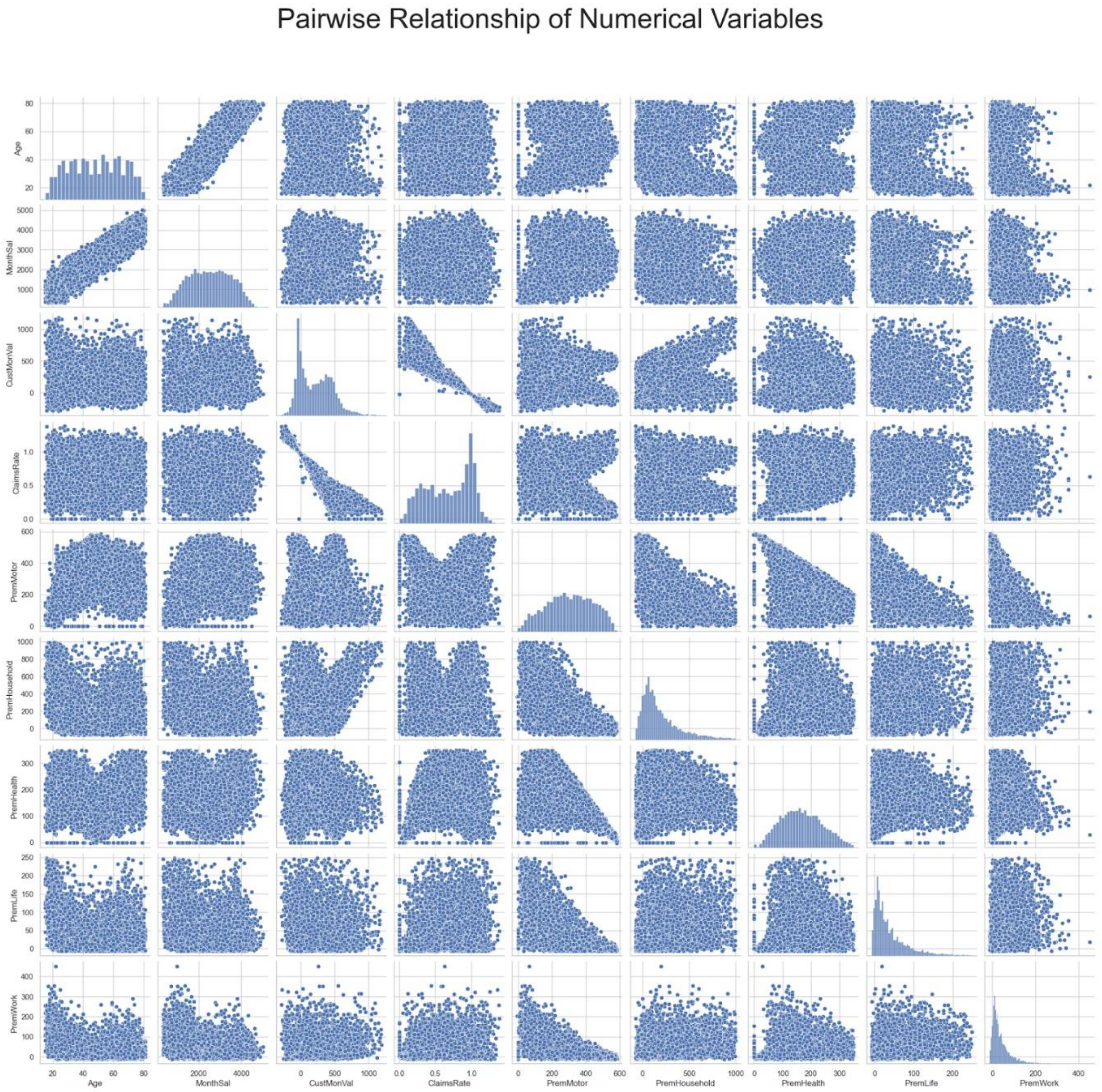


Figure 5 – Pairwise Relationship of Numerical Variables

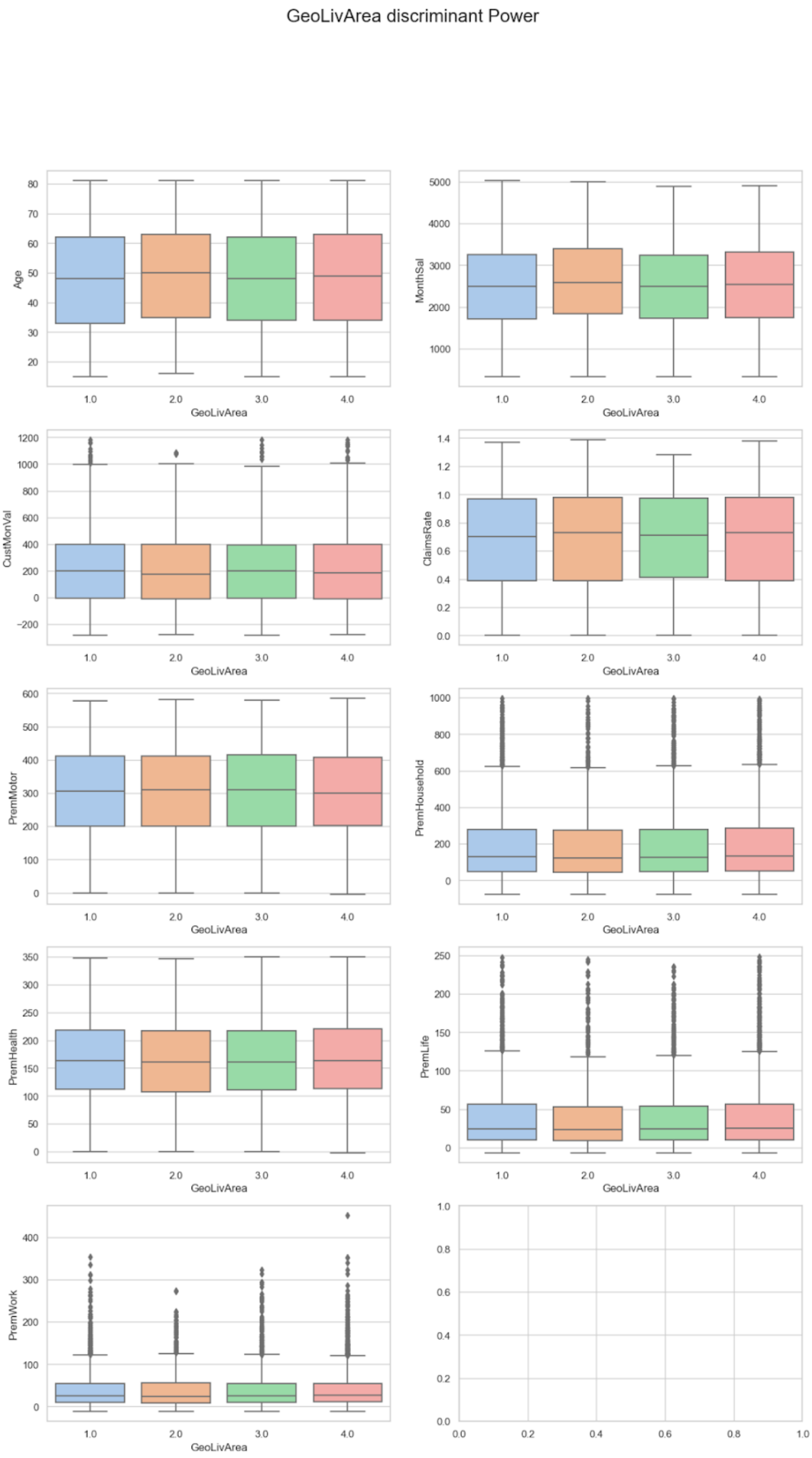


Figure 6 – GeoLiv Discriminant Power

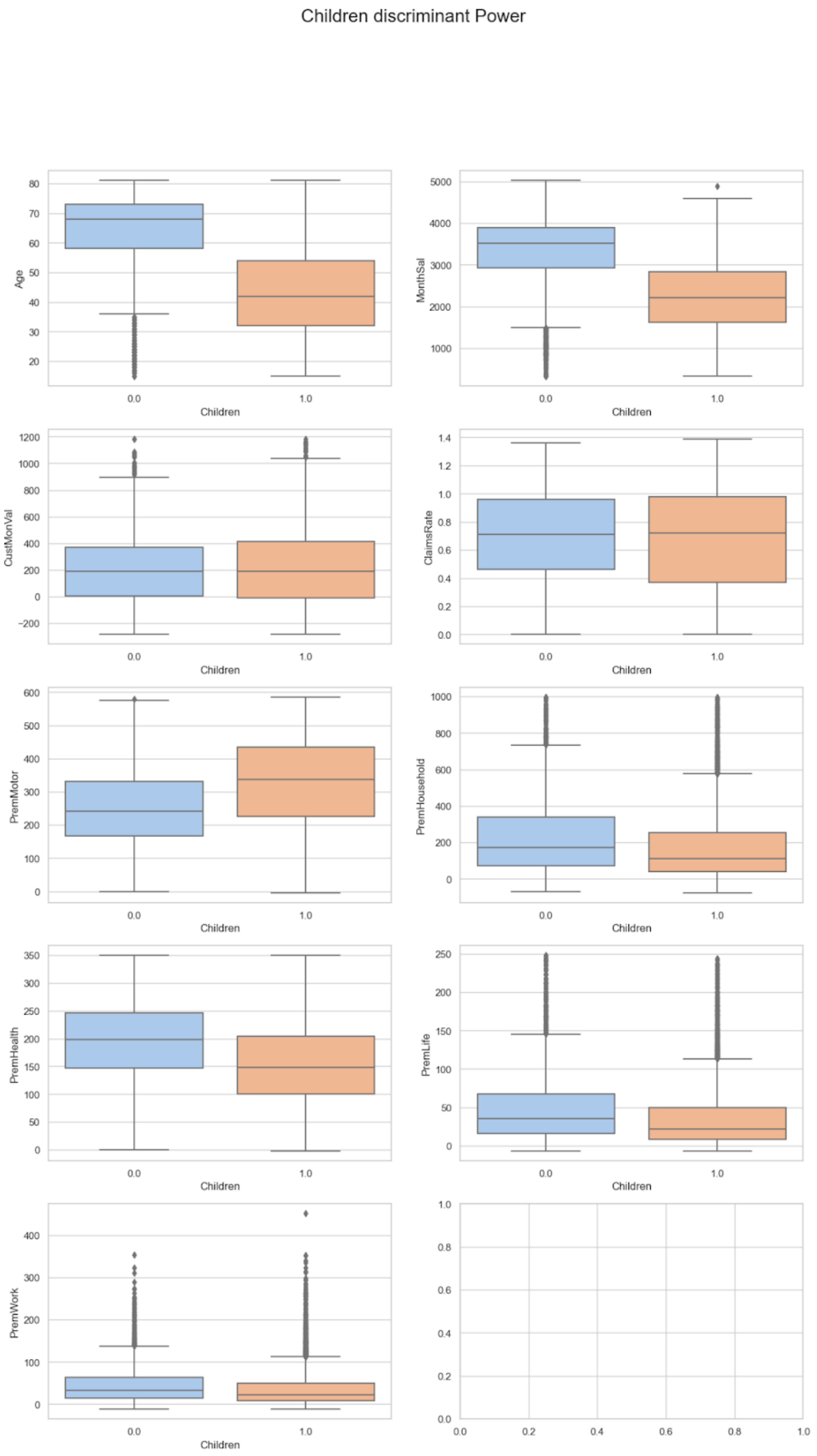


Figure 7 – Children Discriminant Power

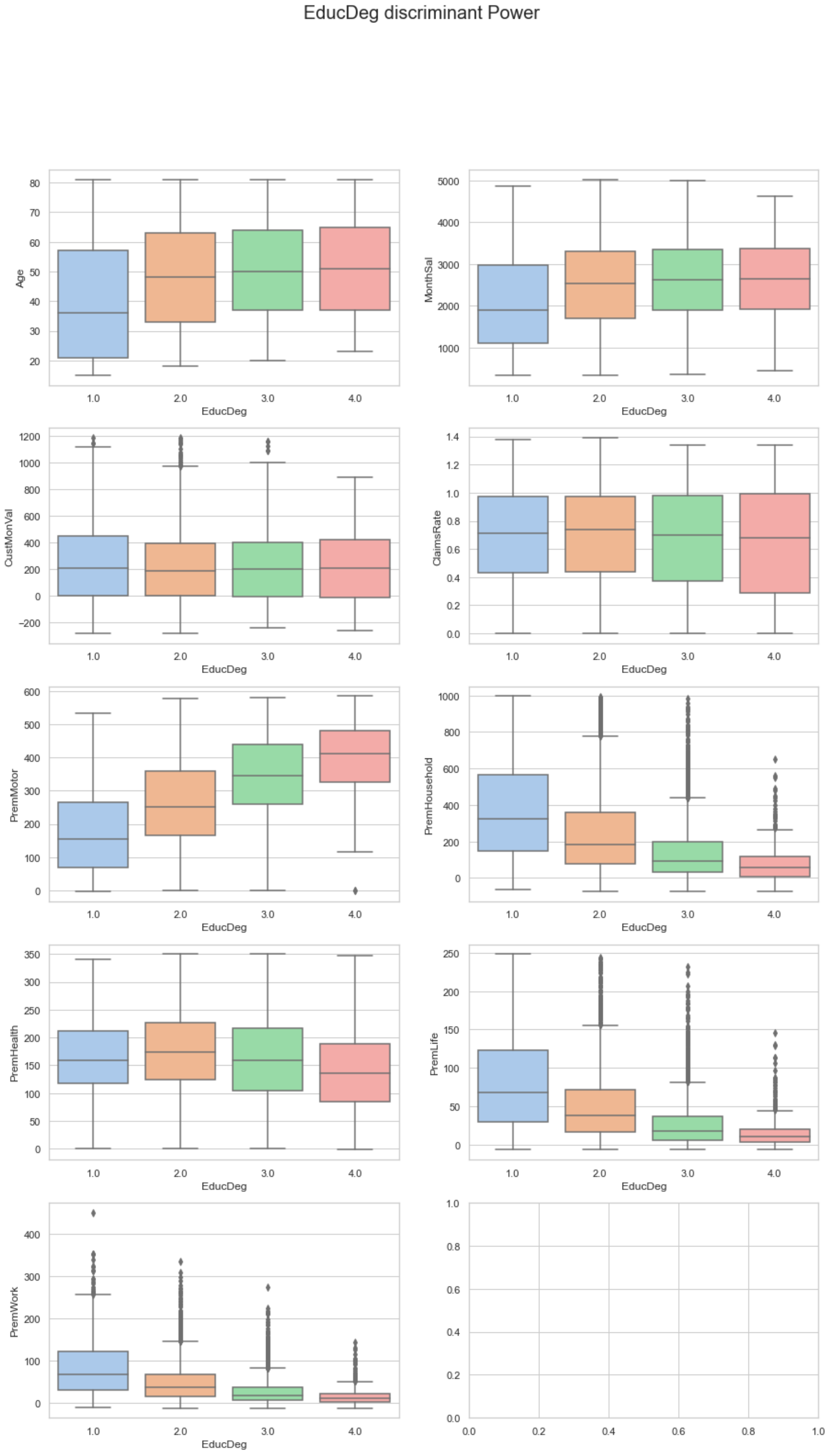


Figure 8 - EducDeg Discriminant Power

Feature 1	Feature 2	Pearson	Kendall
Age	MonthSal	0.925	0.766
ClaimsRate	CustMonVal	-0.936	-0.886
PremHouseHold	Total Balance	0.982	0.922
Health Ratio	PremHealth	0.89	0.738
Life Ratio	PremLife	0.951	0.905
Motor Ratio	PremMotor	0.968	0.86
Motor Ratio	PremHouseHold	-0.741	-0.613
Motor Ratio	Total Balance	-0.708	-0.592
HouseHold Ratio	PremHouseHold	0.969	0.983
HouseHold Ratio	Total Balance	0.95	0.905
HouseHold Ratio	Motor Ratio	-0.793	-0.616
Work Ratio	PremWork	0.953	0.904
Canceled subscription	Num Prems	-1	-1
Lost money	Num Prems	0.597	0.886
Lost money	Canceled subscription	-0.597	-0.886

Table 4 – Pearson and Kendall tau correlation coefficients

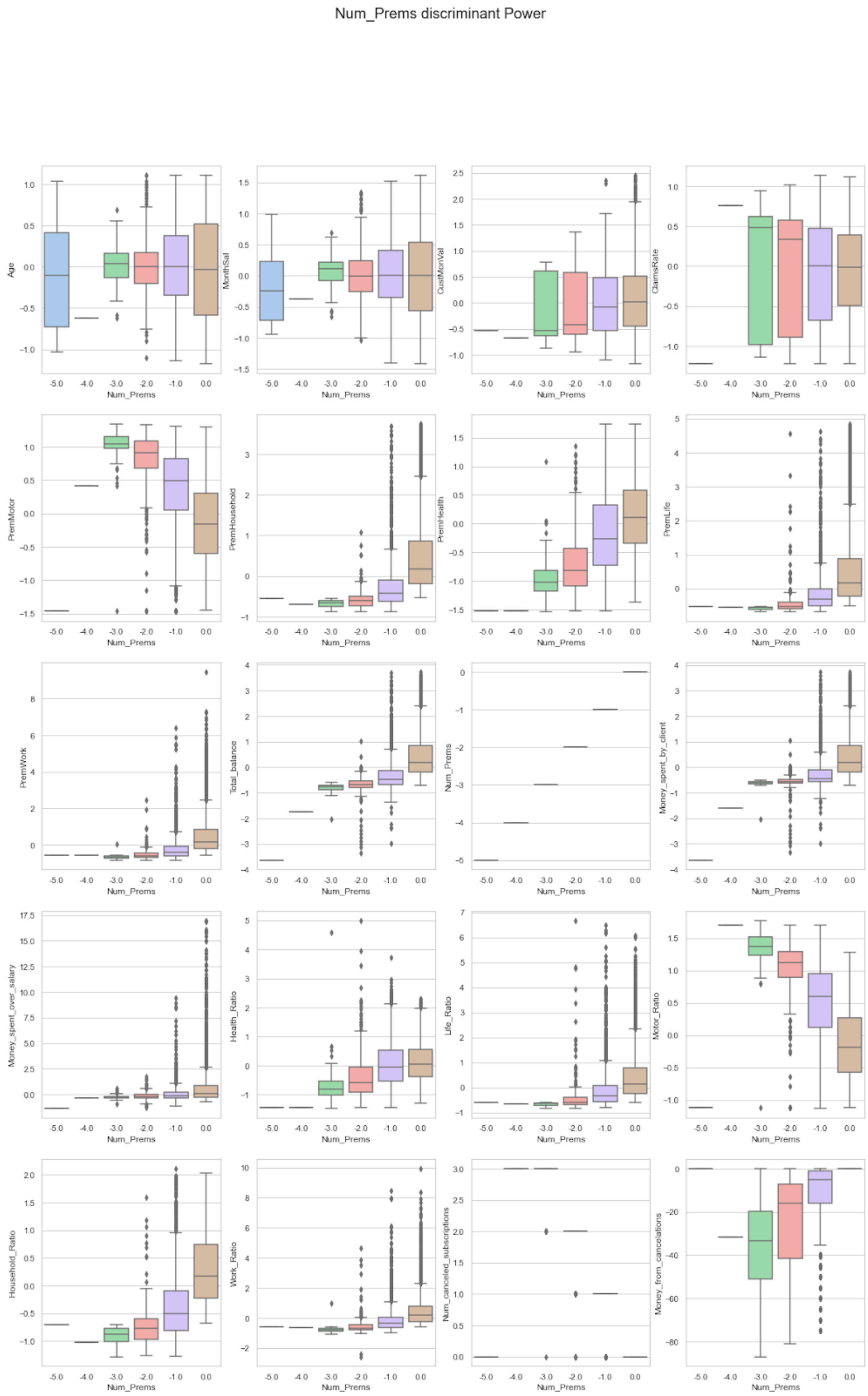


Figura 9 – Num_Prems discriminant Power

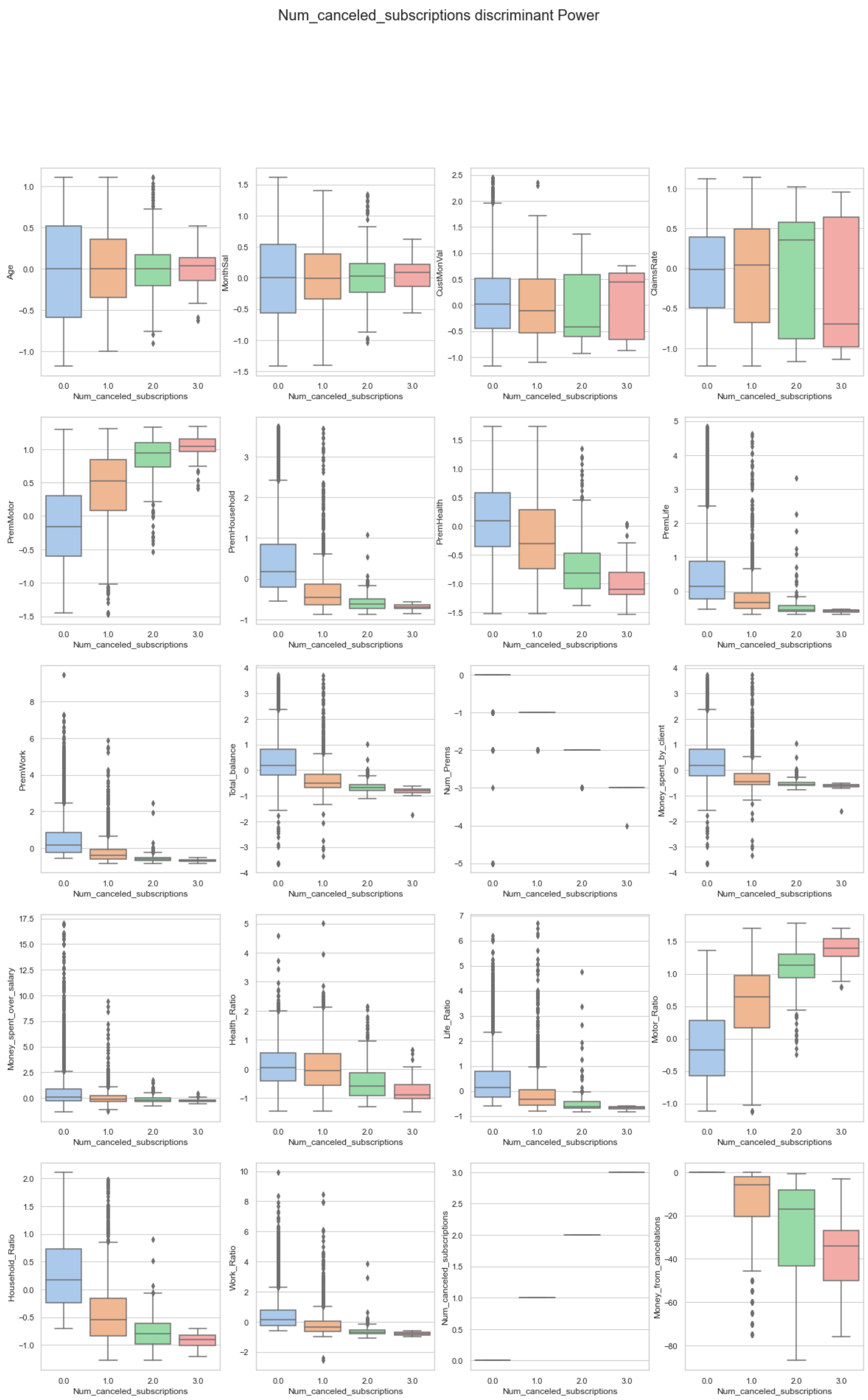


Figure 10 – Num_canceled_subscriptions discriminant Power

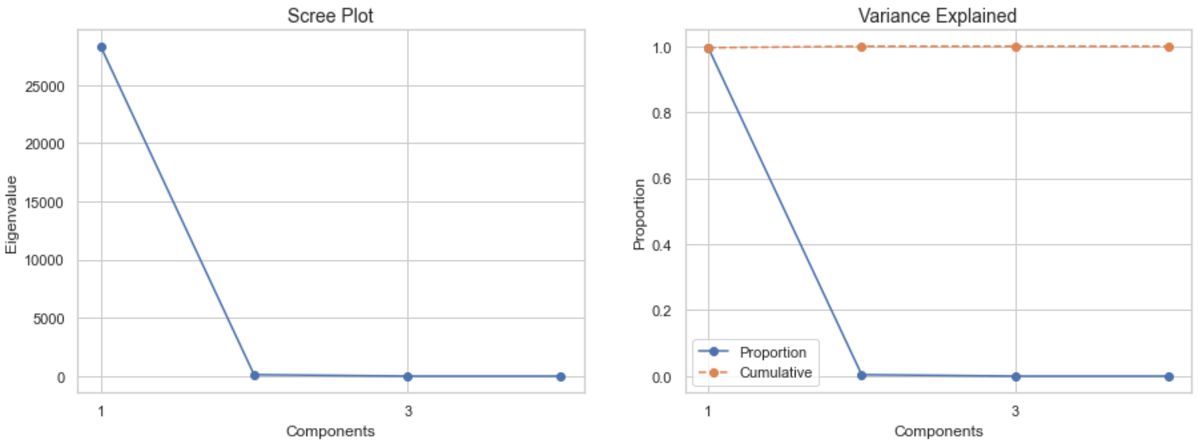


Figure 11 – Eigenvalue and Variance (PCA for value perspective)

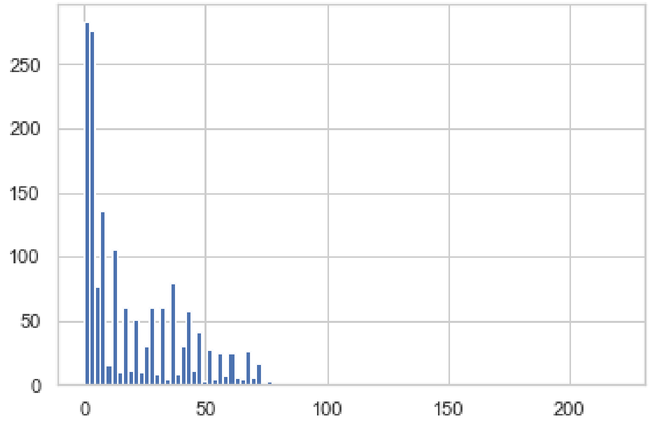


Figure 12 – PC0 distribution (PCA for value perspective)

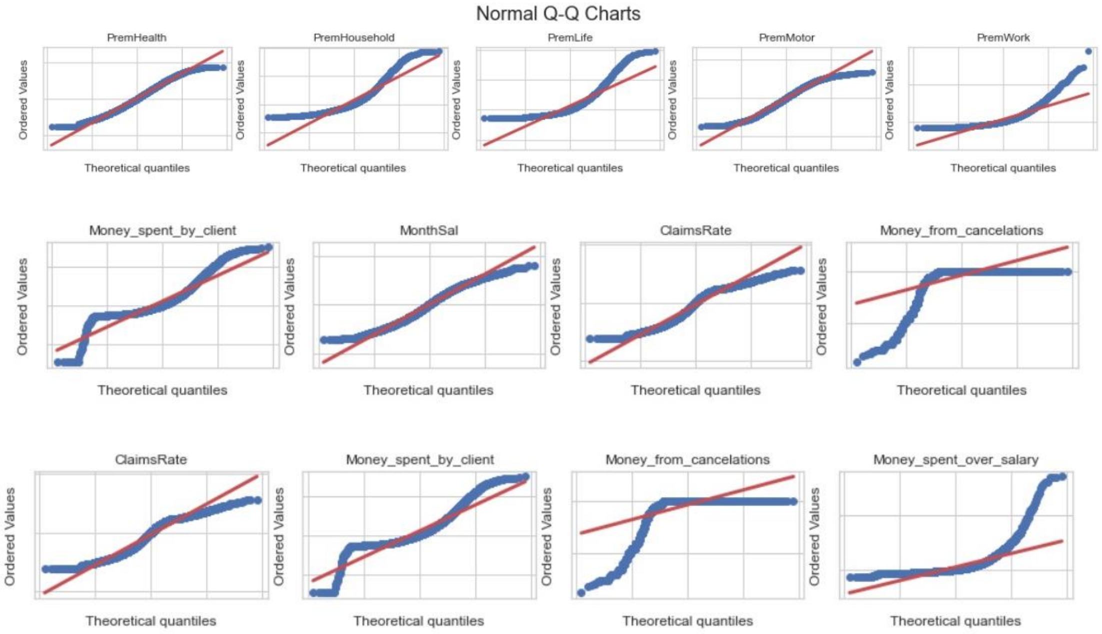


Figure 13 – Normal Quantile-Quantile Plots

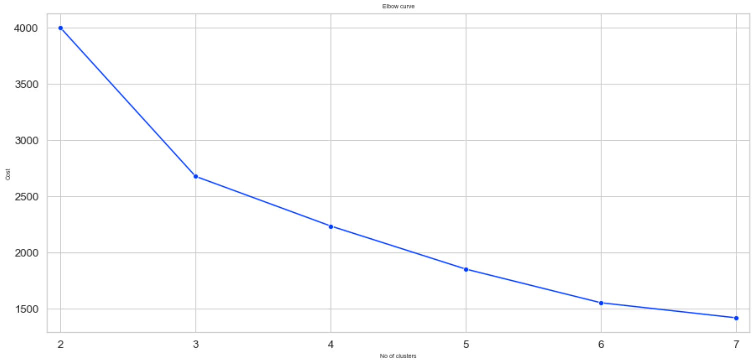


Figure 14 – Cost function for K-Prototypes in Sociodemographic perspective

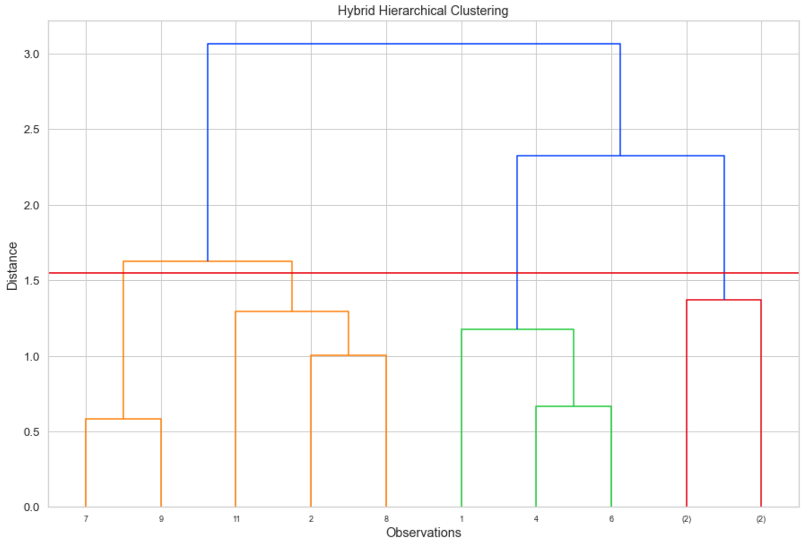


Figure 15 – Dendogram for Sociodemographic perspective

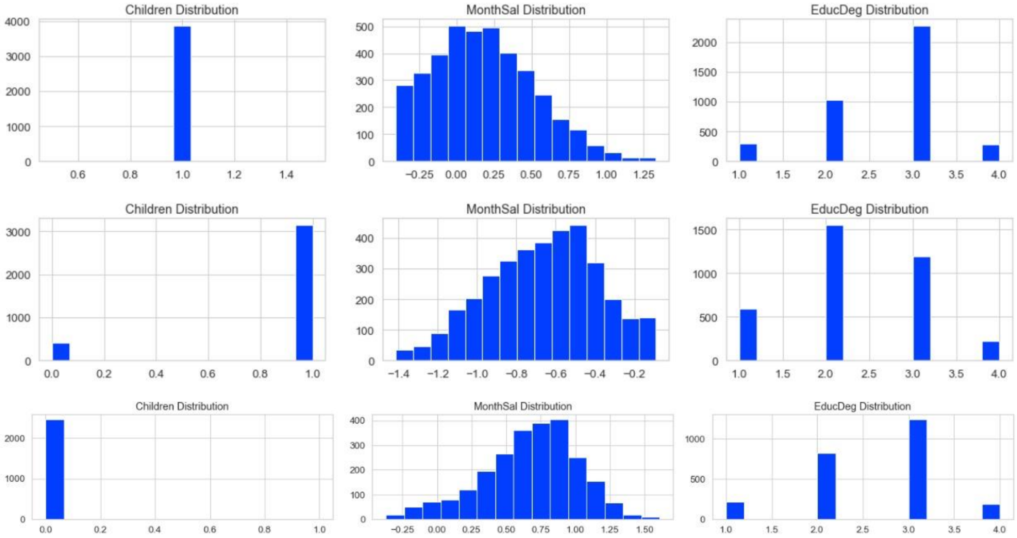


Figure 16 – Variables' distributions for each sociodemographic prespectives clusters

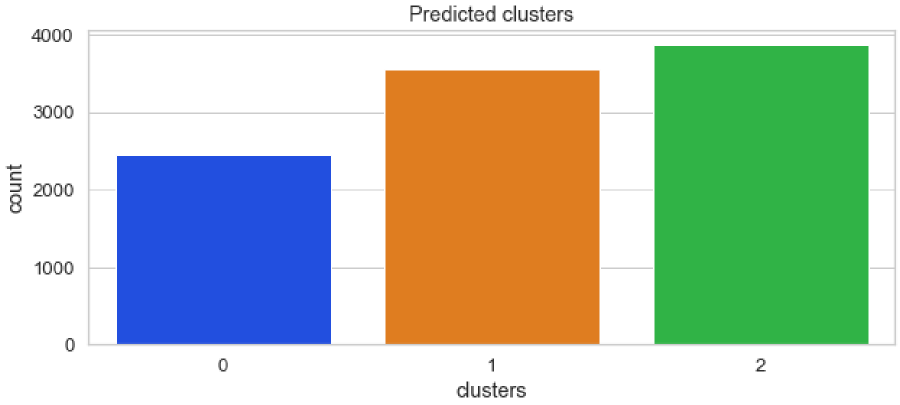


Figure 17 – Count of elements of Sociodemographic clusters

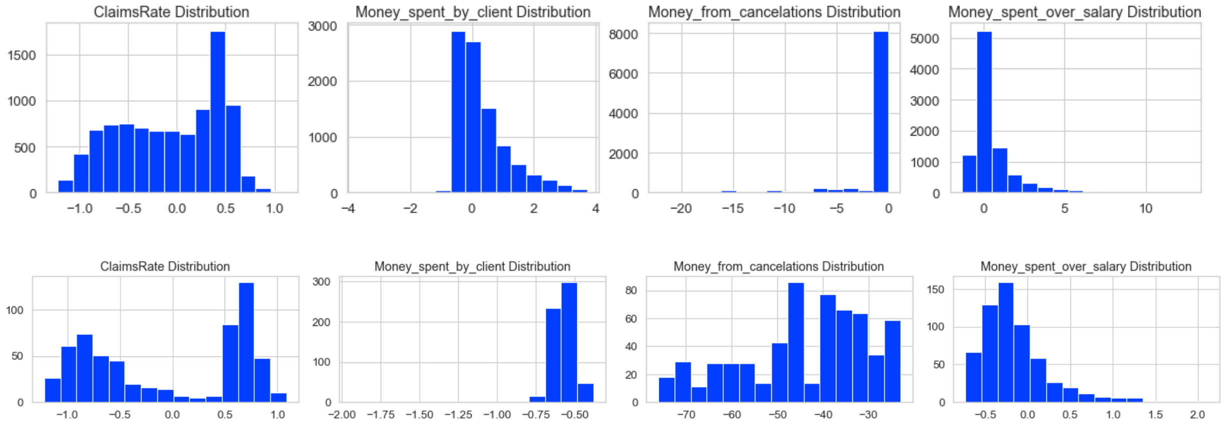


Figure 18 - Variables' distributions for each Value prespectives clusters

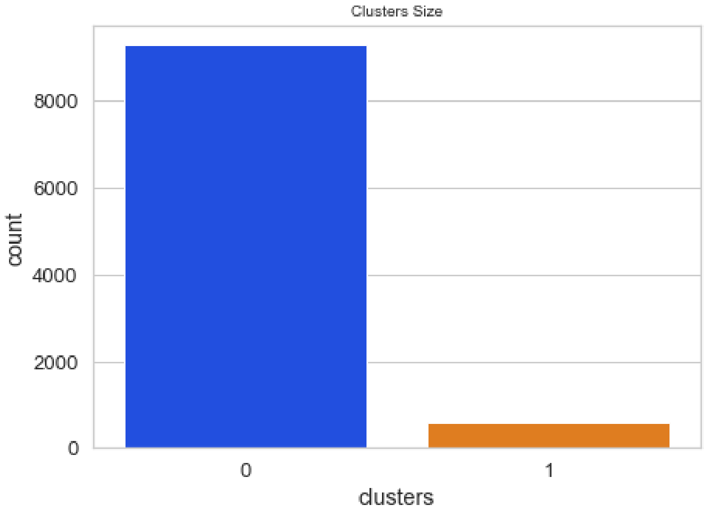


Figure 19 - Count of elements of Value clusters

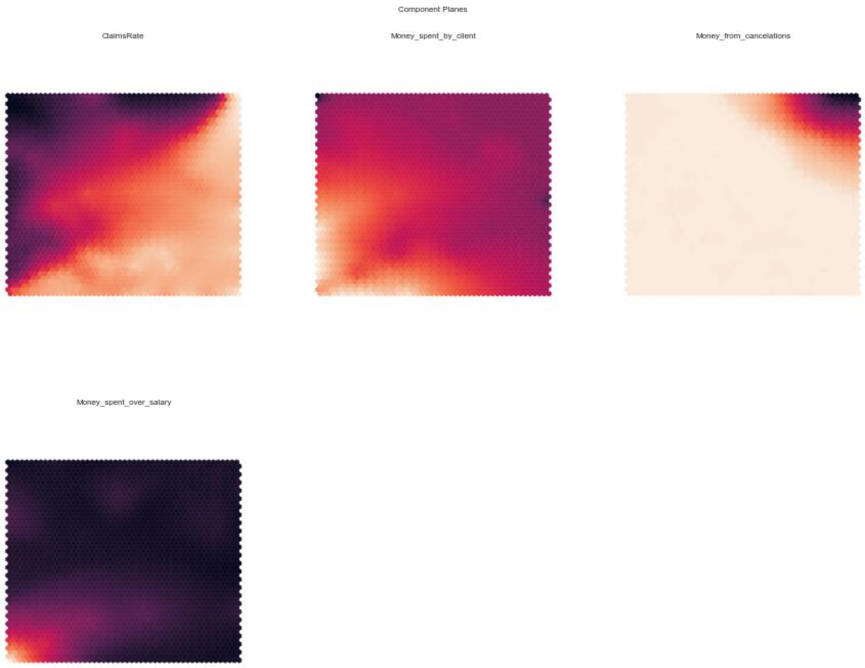


Figure 20 – Component Planes SOM for Value prespective

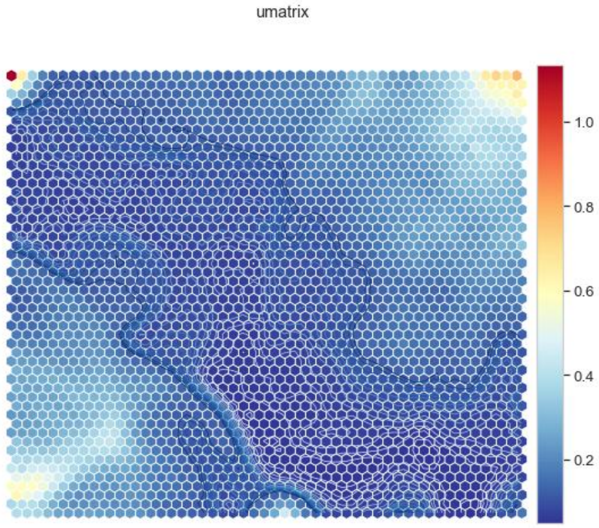


Figura 21 – U-Matrix SOM for Value prespective

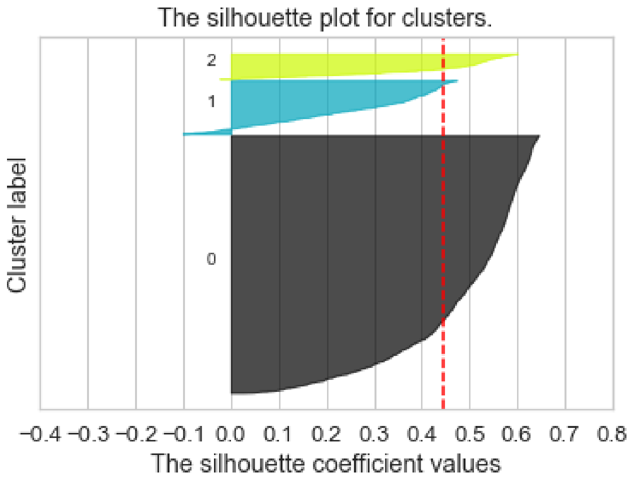


Figure 22 – Silhouette coeficiente for clusters K-Means on top of SOM for the Value prespective

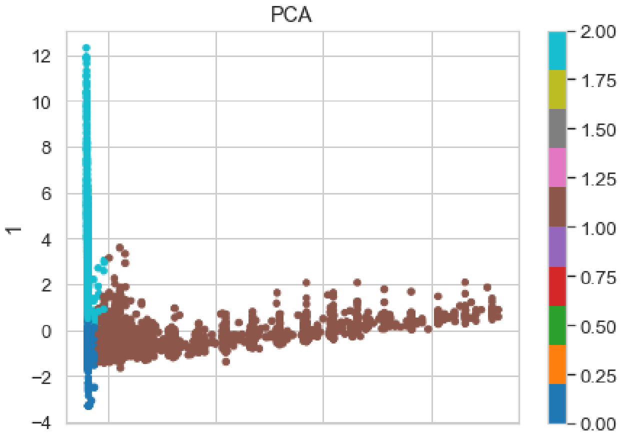


Figure 23 – 2 Dimensional PCA for clusters K-Means on top of SOM for the Value prespective

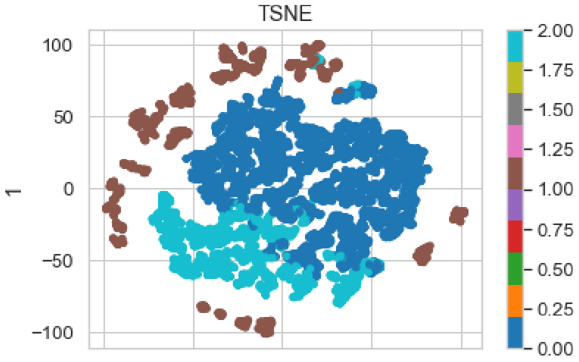


Figura 24 – 2 Dimensional t-SNE for clusters K-Means on top of SOM for the Value prespective

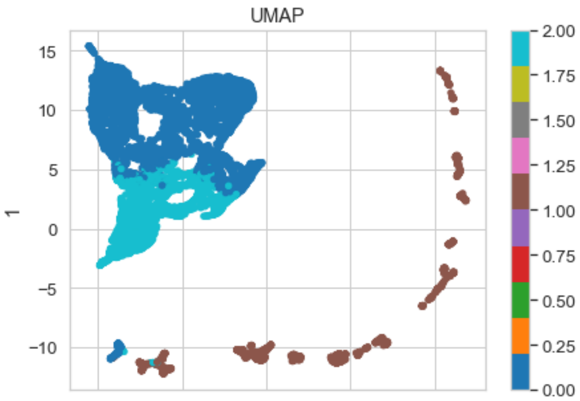


Figura 25 - 2 Dimensional Umap for clusters K-Means on top of SOM for the Value prespective

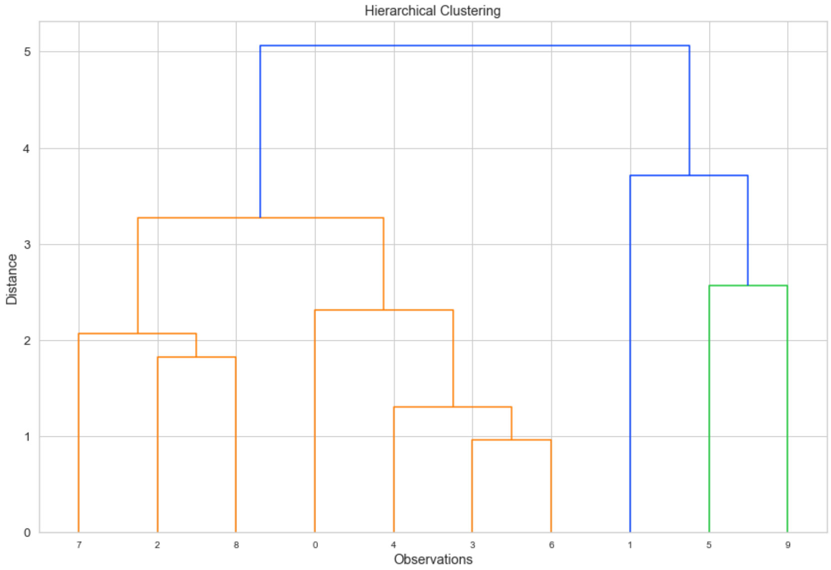


Figure 26 - Dendogram for Product perspective

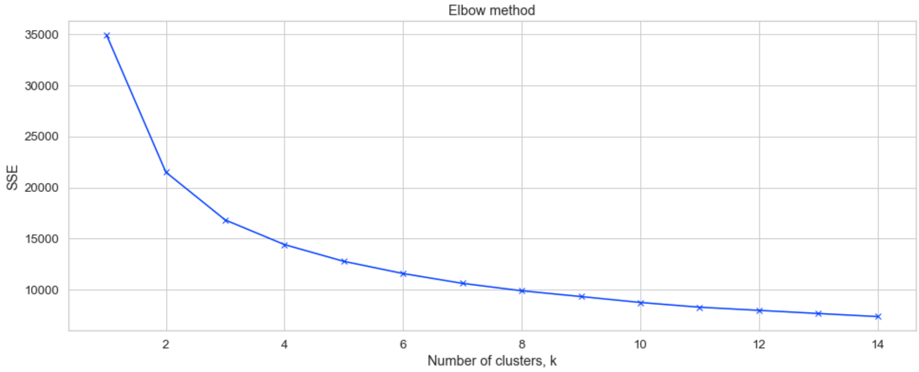
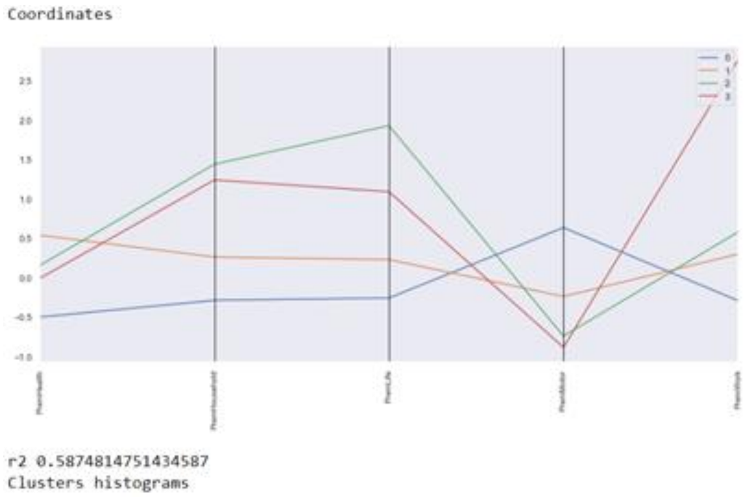


Figure 27 - Cost function for K-Means in Product perspective



```
Coordinates
```

```
r2 0.5874814751434587
Clusters histograms
```

Figura 28 – Parallel plot for K-Means clusters in Product perspective

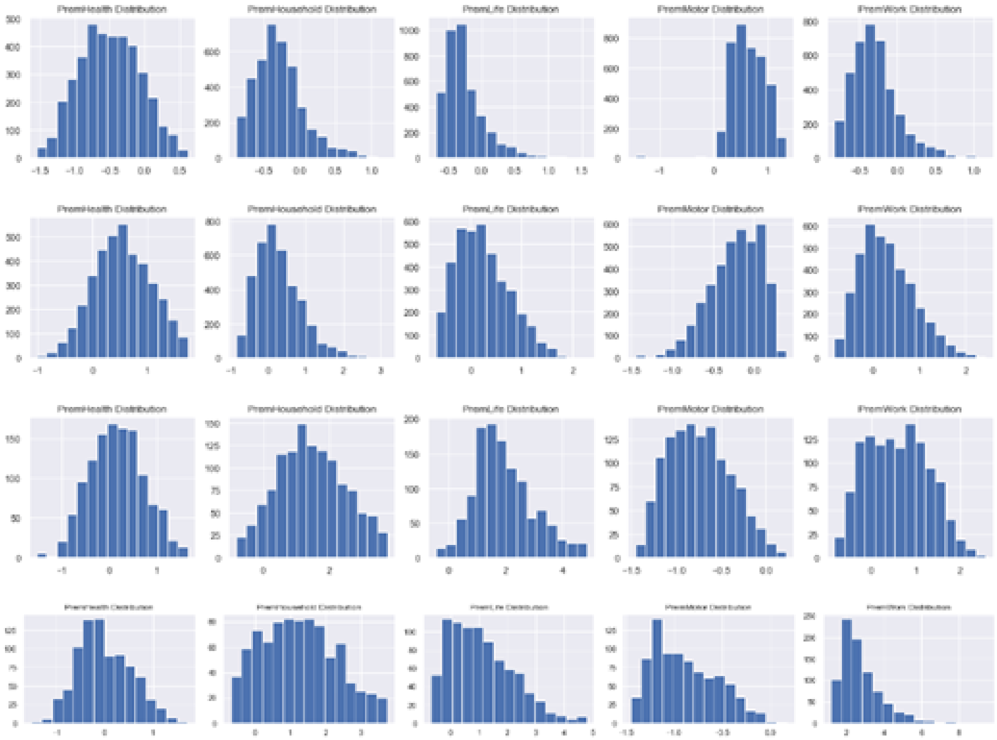


Figure 29 - Variables' distributions using K-Mean for each Product prespective

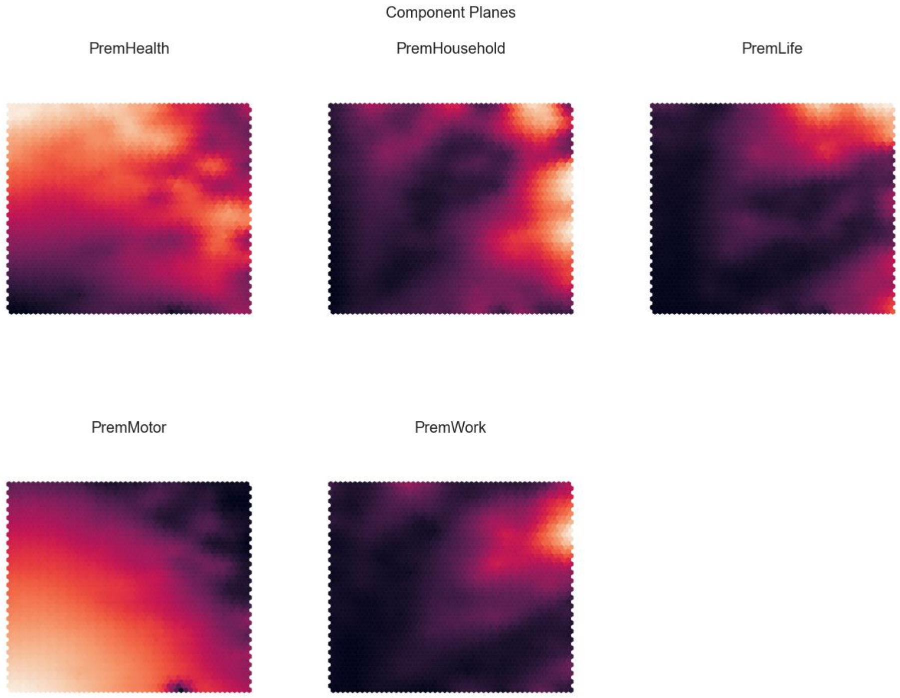


Figure 30 - *Component Planes SOM for Product prespective*

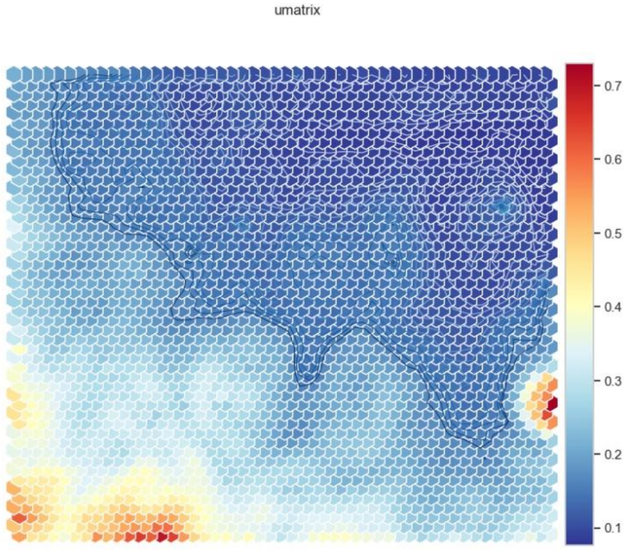


Figure 31 – U-Matrix SOM for Product prespective

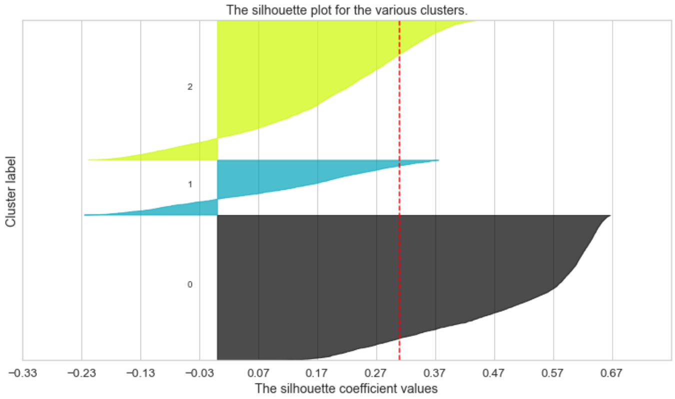


Figure 32 - Silhouette coeficiente for clusters K-Means on top of SOM for the Product prespective

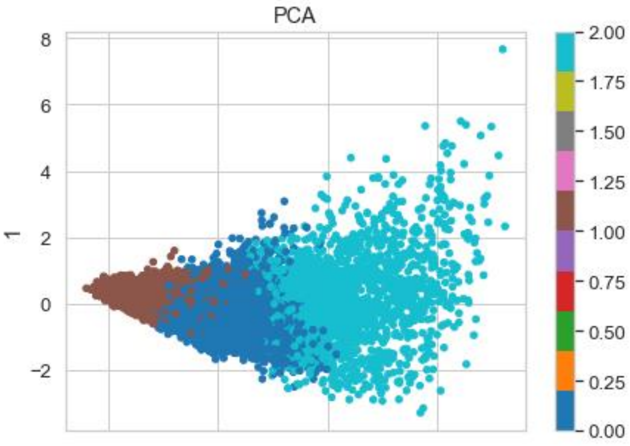


Figure 33 - 2 Dimensional PCA for clusters K-Means on top of SOM for the Product prespective

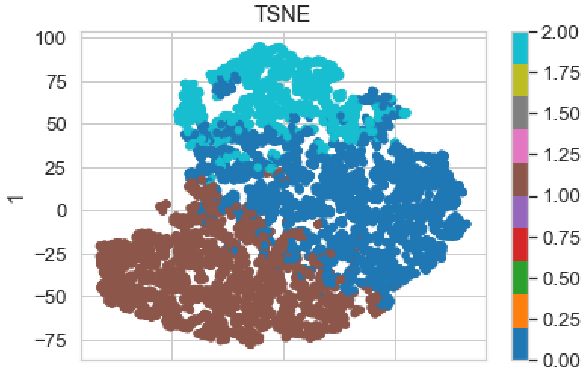


Figure 34 - 2 Dimensional t-SNE for clusters K-Means on top of SOM for the Prooduct prespective

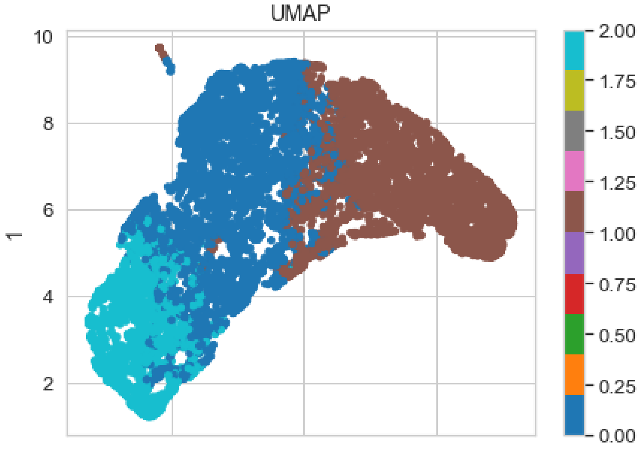
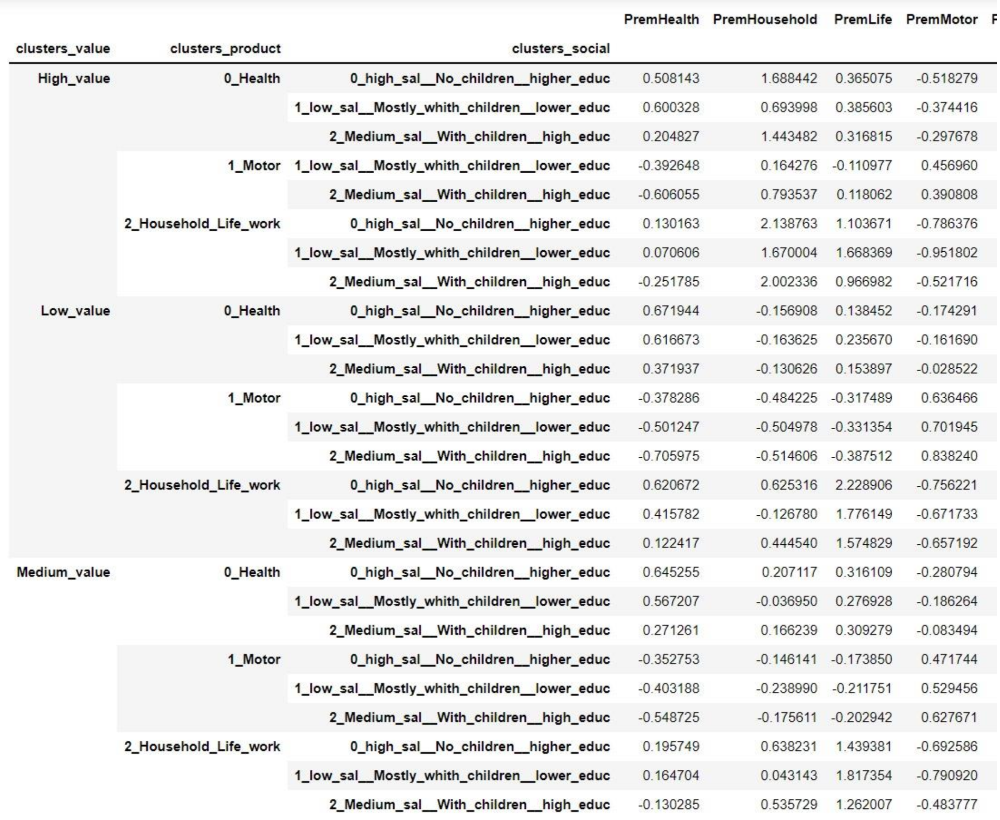


Figure 35 - Umap for clusters K-Means on top of SOM for the Product prespective



			PremHealth	PremHousehold	PremLife	PremMotor	F
clusters_value	**clusters_product**	**clusters_social**					
High_value	**0_Health**	**0_high_sal__No_children__higher_educ**	0.508143	1.688442	0.365075	-0.518279	
		1_low_sal__Mostly_whith_children__lower_educ	0.600328	0.693998	0.385603	-0.374416	
		2_Medium_sal__With_children__high_educ	0.204827	1.443482	0.316815	-0.297678	
	1_Motor	**1_low_sal__Mostly_whith_children__lower_educ**	-0.392648	0.164276	-0.110977	0.456960	
		2_Medium_sal__With_children__high_educ	-0.606055	0.793537	0.118062	0.390808	
	2_Household_Life_work	**0_high_sal__No_children__higher_educ**	0.130163	2.138763	1.103671	-0.786376	
		1_low_sal__Mostly_whith_children__lower_educ	0.070606	1.670004	1.668369	-0.951802	
		2_Medium_sal__With_children__high_educ	-0.251785	2.002336	0.966982	-0.521716	
Low_value	**0_Health**	**0_high_sal__No_children__higher_educ**	0.671944	-0.156908	0.138452	-0.174291	
		1_low_sal__Mostly_whith_children__lower_educ	0.616673	-0.163625	0.235670	-0.161690	
		2_Medium_sal__With_children__high_educ	0.371937	-0.130626	0.153897	-0.028522	
	1_Motor	**0_high_sal__No_children__higher_educ**	-0.378286	-0.484225	-0.317489	0.636466	
		1_low_sal__Mostly_whith_children__lower_educ	-0.501247	-0.504978	-0.331354	0.701945	
		2_Medium_sal__With_children__high_educ	-0.705975	-0.514606	-0.387512	0.838240	
	2_Household_Life_work	**0_high_sal__No_children__higher_educ**	0.620672	0.625316	2.228906	-0.756221	
		1_low_sal__Mostly_whith_children__lower_educ	0.415782	-0.126780	1.776149	-0.671733	
		2_Medium_sal__With_children__high_educ	0.122417	0.444540	1.574829	-0.657192	
Medium_value	**0_Health**	**0_high_sal__No_children__higher_educ**	0.645255	0.207117	0.316109	-0.280794	
		1_low_sal__Mostly_whith_children__lower_educ	0.567207	-0.036950	0.276928	-0.186264	
		2_Medium_sal__With_children__high_educ	0.271261	0.166239	0.309279	-0.083494	
	1_Motor	**0_high_sal__No_children__higher_educ**	-0.352753	-0.146141	-0.173850	0.471744	
		1_low_sal__Mostly_whith_children__lower_educ	-0.403188	-0.238990	-0.211751	0.529456	
		2_Medium_sal__With_children__high_educ	-0.548725	-0.175611	-0.202942	0.627671	
	2_Household_Life_work	**0_high_sal__No_children__higher_educ**	0.195749	0.638231	1.439381	-0.692586	
		1_low_sal__Mostly_whith_children__lower_educ	0.164704	0.043143	1.817354	-0.790920	
		2_Medium_sal__With_children__high_educ	-0.130285	0.535729	1.262007	-0.483777	

Table 4

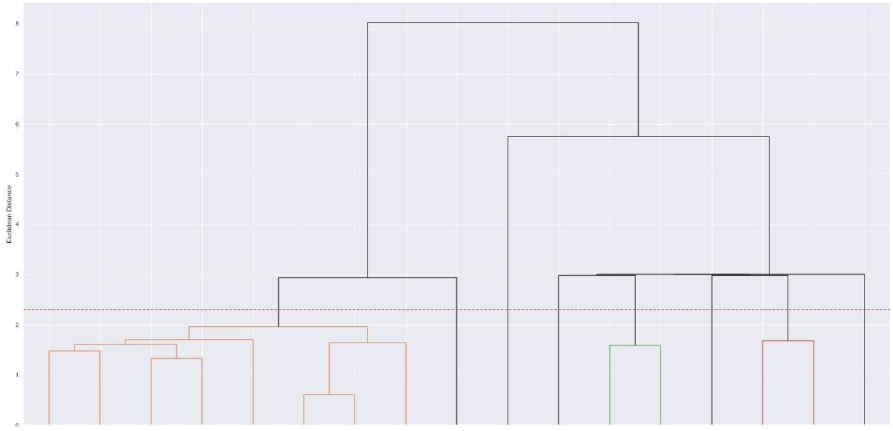
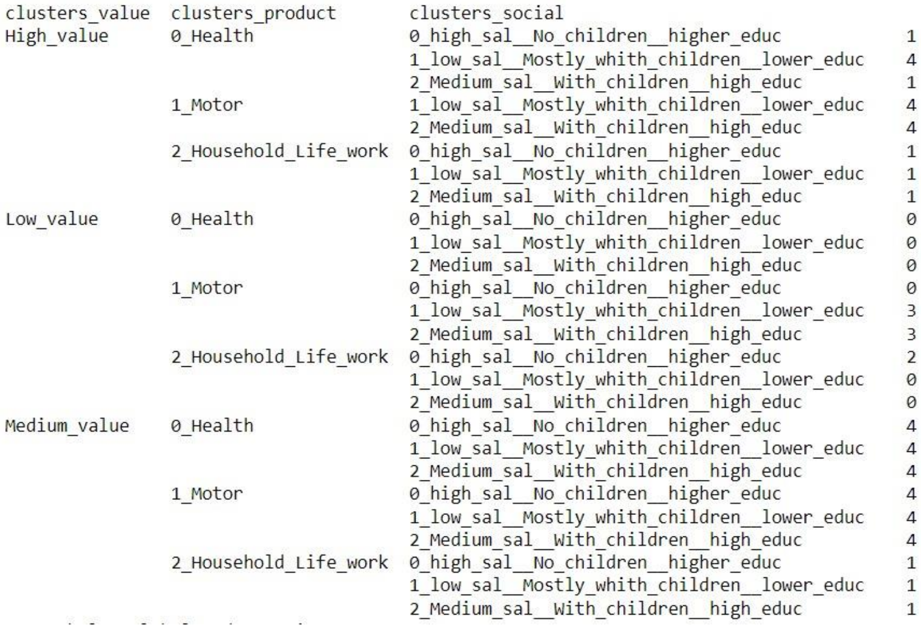
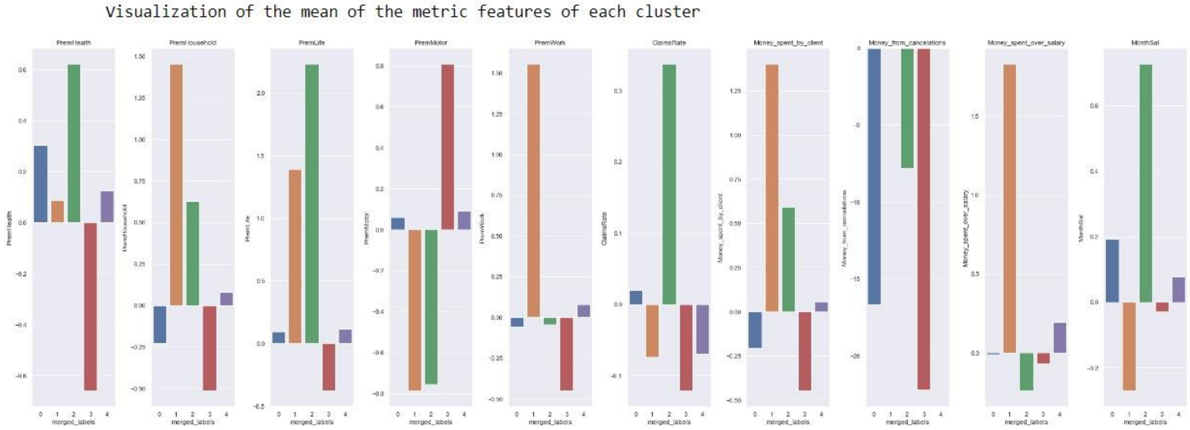


Figure 36 – Dendogram



```
clusters_value  clusters_product     clusters_social
High_value      0_Health             0_high_sal__No_children__higher_educ        1
                                     1_low_sal__Mostly_whith_children__lower_educ   4
                                     2_Medium_sal__With_children__high_educ      1
                1_Motor              1_low_sal__Mostly_whith_children__lower_educ   4
                                     2_Medium_sal__With_children__high_educ      4
                2_Household_Life_work 0_high_sal__No_children__higher_educ       1
                                     1_low_sal__Mostly_whith_children__lower_educ   1
                                     2_Medium_sal__With_children__high_educ      1
Low_value       0_Health             0_high_sal__No_children__higher_educ        0
                                     1_low_sal__Mostly_whith_children__lower_educ   0
                                     2_Medium_sal__With_children__high_educ      0
                1_Motor              0_high_sal__No_children__higher_educ        0
                                     1_low_sal__Mostly_whith_children__lower_educ   3
                                     2_Medium_sal__With_children__high_educ      3
                2_Household_Life_work 0_high_sal__No_children__higher_educ       2
                                     1_low_sal__Mostly_whith_children__lower_educ   0
                                     2_Medium_sal__With_children__high_educ      0
Medium_value    0_Health             0_high_sal__No_children__higher_educ        4
                                     1_low_sal__Mostly_whith_children__lower_educ   4
                                     2_Medium_sal__With_children__high_educ      4
                1_Motor              0_high_sal__No_children__higher_educ        4
                                     1_low_sal__Mostly_whith_children__lower_educ   4
                                     2_Medium_sal__With_children__high_educ      4
                2_Household_Life_work 0_high_sal__No_children__higher_educ       1
                                     1_low_sal__Mostly_whith_children__lower_educ   1
                                     2_Medium_sal__With_children__high_educ      1
```

Table 5



Visualization of the mean of the metric features of each cluster

Figure 37 – Final Clusters Profiles

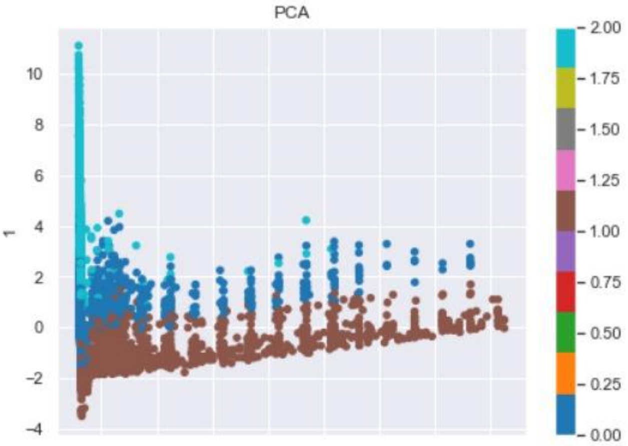


Figure 38 – 2 Dimensional PCA for clusters K-Means on top of SOM for the Value prespective

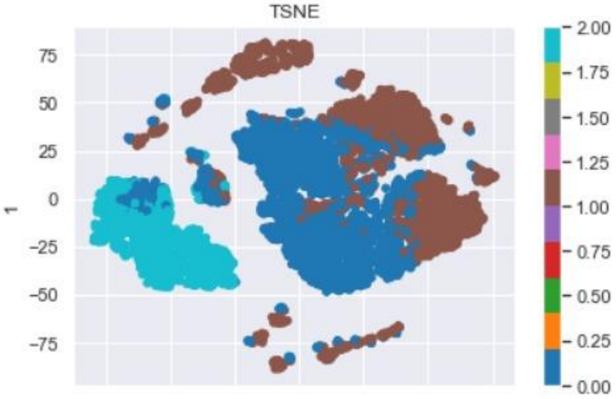


Figure 39 – 2 Dimensional t-SNE for clusters K-Means on top of SOM for the Value prespective

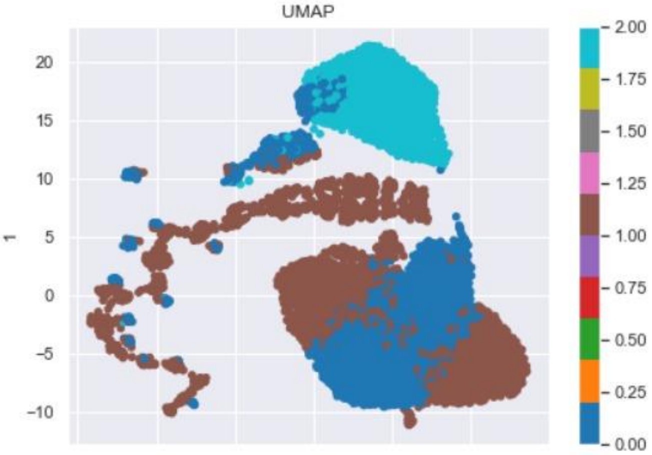


Figure 40 – Umap for clusters K-Means on top of SOM for the Value prespective

Cluster	Sociodemographic	Value	Product	Evaluation	Marketing Campaigns
0 (566)	Low H: Income	L: Money from cancelations	H: PremHealth M: All Prems	Stable costumers with high income, but medium in all insurances, loyal so try to get then spend more	- Offer a gift (sound speaker, watch) to show appreciation for being loyal -Offer small discount to upgrade insurances they alrady have
1 (2033)	Medium VL: salary	VH: Money spent by cliente & over salary H: Money from cancelations	VH: PremHousehold & PremWork L: PremMotor	Low-income individuals that spent a lot on insurances so probably that is why they cancel them.	-If they have PremHousehold and PremWork, offer discount in PremMotor
2 (7)	Low VH: salary	VL: Money spent over salary VH: Claims Rate	VH: PremHealth & PremLife VL: PremMotor	Special costumers: spent a lot in insurances, except motor (maybe they use táxi/uber). Potential to spend more	-If has two packages, 10% discount in PremMotor. -When they have more than 2 packages and one is PremMotor, wins na object -if they have 2 packages and subscribe to basic PremMotor, upgrad to the best PremMotor
3 (1180)	Low L: salary	VL: Money from cancelations	VH: PremMotor VL: PremHealth	People who just have insurance for the veichle (because is mandatory and cant afford more insurances)	- if they have PremMotor, discount of 10% or 20% in PremHealth in the next 2 years
4 (6094)	Medium	H: Money from cancelations	Medium	Regular people that canceled premiums. Maybe changed company – bring them back	- 30% discount in other premium of their choice for 1 year (with fidelity)
VL = Very Low, L = Low, M = Medium, H = High, VH = Very High					

Table 6 – Marketing strategies