Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance (A2Z)

Group N

Luca Sousa: 20220601

Mariana Lavinha number: 20220565

Martim Santos number: 20220540

janeiro, 2023

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

INDEX

1. Introduction

This report was developed with the purpose of helping A2Z Insurance (A2Z) better understand their customers by designing a market segmentation study that takes as input the ABT (Analytic Based Table) provided prior to this analysis. Market segmentation divides a market into smaller segments. A market segment consists of a group of customers who share a similar set of needs and wants (Kotler & Keller, 2012).

Taking this in mind, this project is going to focus on the characteristics and further definition of each group of customers as well as some final recommendations of approaches A2Z should take to create better customer value and stronger long-term loyalty relationships.

2. Data Exploration

2.1. Non-graphical analysis

During this initial phase, the company's ABT structure was analyzed. The dataset initially has 10296 observations and 14 variables. In case of any doubts, the metadata regarding the dataset can be seen in appendix 1. After looking at its shape and which types of data were being recorded in each column some initial actions were taken.

The columns *GeoLivArea*, *Children* and *EducDeg* had their type of data transformed to categories due to them being of categorical nature. Since the first two were non-ordinal, it was just passed the *astype()* method while for the last one it had to be defined the order of each element prior to the transformation. This action helps with memory usage and performance.

As for EducDeg, we decided to bin the education levels. After analyzing the data, we determined that it would be useful to divide the population into two groups based on whether they had a PhD or not. This decision was based on the finding that the average *Customer Monetary Value* and *Claims Rate* was significantly different for those with a PhD, making it a useful indicator for our bins. By dividing the population in this way, we can more easily compare and analyze the data for these two groups.

We also wanted to reduce the number of categories and make it easier to visualize the data. By grouping the education levels into two categories, you can more easily see the distribution of education levels within the population.

CustID was defined as the index and 3 duplicate values were dropped from the table.

2.2. Coherence Check

There are several issues with the data that need to be addressed. First, the value for *FirstPolYear* cannot be greater than 2016. Second, the value for *BirthYear* of 1028 does not make sense, as it is 100 years or below. Third, the value for *FirstPolYear* must be greater than or equal to the *BirthYear* as individuals cannot have insurance before they are born. Fourth, individuals cannot spend more than

they earn, so it is important to ensure that the data reflects this. Finally, we should verify that the age of individuals is consistent with their salary, as it would not be realistic for a 16-year-old to have a very high salary.

2.3. Descriptive statistics & Graphical analysis

Appendices 2 and 3 show descriptive statistics for metric and non-metric variables, respectively. From looking at the tables it can be noticed some extreme values for several of the variables, something that needs to be accounted for. Appendix 5 shows the first visual exploration done: bar-charts. These bar-charts give information regarding class frequencies among customers. There's a high discrepancy between customers with and without children as well as customers with PhD or only having basic education versus the rest. Less than 10% of our customers live in the area categorized as 2. This could be due to different reasons such as the zone not being so much of a residential area and more of an industrial one or maybe the insurance services aren't well promoted through local channel distribution campaigns. Whichever, this should be something to be discussed with A2Z stakeholders later on.

3. Data Preprocessing

3.1 Outliers

In this data preprocessing step, we first examined histograms and boxplots for each variable in the dataset (Appendix 6 and 7). Based on these visualizations, we identified potential outliers for several variables.

For the variable "MonthSal," which represents gross monthly salary, we decided to set a limit wage of 5021. This decision was based on the distribution of values in the variable, as values above this threshold appeared to be extreme.

For the variable "PremMotor," we determined that values above 600 were extreme and should be removed. Similarly, for the variable "PremHousehold," we determined that values above 2300 were extreme and should be removed. For the variable "PremHealth," we found that values above 450 were extreme and should be removed.

We decided not to eliminate any values for the variable "PremLife," as all values seemed to be relatively close together and did not appear to be extreme.

Overall, these data preprocessing steps will help us to better understand the distribution of values within our dataset and ensure that our analyses are based on reliable and accurate data.

3.2 Missing Values

Appendix 4 shows the missing values that the dataset holds. A replacement transformation was done based on the median, for most of the metric variables and on the mode for the categorical variables.

The premium variables went through a different approach based on the following assumption: You don't need to have an insurance service for every option. Therefore, the missing values were replaced by zero meaning the customer in question doesn't have a contract containing that specific insurance service.

3.3 Feature Engineering

We first created a variable called "Age" (2016 - BirthYear) to represent the length of time that a customer has been with the company. Next, we created a variable called "Fidelity" (2016 - FirstPolYear) to track how long a customer has been with the company. We also created the variable "Prem_Total" which represents the sum of all total positive premiums. Finally, we removed customers who did not have any premiums purchased in the current year (2016) from the dataset, as these customers were not relevant to the analysis. Appendix 8 shows the Boxplots after the outliers treatment and feature engineering.

3.4 Correlations

The Spearman method was used to identify correlations in the data, as it is often used in cases where the variables are not necessarily normally distributed or the relationship between the variables is not necessarily linear. A correlation heatmap was generated to visualize the relationships between the variables. The results showed that there is a strong negative correlation (-0.98) between Claims_Rate and CMV. The next highest correlations were between Motor and the other premiums, ranging from -0.64 to -0.7. Most of the other correlations in the heatmap were weak, as indicated by the pale colors in the graphic matrix. (Appendix 9)

3.1. 3.3 Variable Selection

To improve the performance of our models in the future, it is important to conduct an exploratory analysis of the relationships between our variables, in order to identify the most effective set of variables to use when applying clustering algorithms.

At this step we chose to drop the Age variable since it has inconsistent values and a strong correlation with the MonthSal variable (0.93). As for the Fidelity variable, after analyzing the correlation matrix, we can conclude that it does not show a strong correlation with any variable. So we decided to remove this variable.

4. Cluster algorithms

4.1. Demographic segmentation

The variables *EducDeg_phd_0*, *Children*, and *MonthSal* were chosen for the demographic segment because they provided relevant information about the characteristics of our customers. The 'EducDeg_phd_0' variable allowed us to differentiate between customers with a PhD and those without, which can be useful for understanding their level of education and potentially their income or occupation. We splitted the original variable due to the symmetry found in the other points, when compared to key features, such as Customer Monetary Value and Monthly Salary.

The 'Children' variable was included to capture information about the family size of our customers, which can be useful for understanding their household composition and potentially their lifestyle and needs. The 'MonthSal' variable was included to capture information about the monthly income of our customers, which can be useful for understanding their financial situation and potentially their purchasing power.

For the segmentation, we decided to use K prototypes. K-means, Self-Organizing Maps (SOM), DBSCAN, and Mean-Shift are all clustering algorithms that are used to group similar data points together. Most of these algorithms are not well-suited for categorical data because they rely on the distance between data points to determine similarity. Categorical data consists of discrete values and does not have a natural ordering or a notion of distance. As a result, using these algorithms on categorical data can lead to suboptimal or even meaningless clusters.

Additionally, these algorithms are designed to work with continuous numerical data and may not be able to handle the high dimensionality that is often present in categorical data. K-prototypes is a clustering algorithm that is specifically designed to handle mixed data, which includes both numerical and categorical data (Nguyen et al., 2011). It is an extension of the K-means algorithm that has been modified to handle categorical data by using a mode-based distance measure rather than a traditional Euclidean distance measure.

After applying the K prototypes algorithm, the demographic segment was divided into 3 clusters, which were evenly distributed with almost one third of observations in each. This balanced distribution can give us a good representation of the characteristics of our customers within each cluster and can provide useful insights when merged with the product segment.

4.2. Value (Product & Customer Value) segmentation.

For the second segmentation the following variables were chosen: *CustMonVal*; *ClaimsRate*; *PremMotor*; *PremHousehold*; *PremHealth*; *PremLife*; *PremWork*. Before applying different algorithms, it was decided the best feature scaling technique to use and some background research was made. Since the data presented wasn't normally distributed and had a strong amount of outliers, specially in the Premiums variables, it was decided to apply the Robust Scaler technique instead of others such as MinMax, MaxAbs or Standard Scaler. The first two due to their sensitiveness to outliers and the last one due to lack of normally distributed data. Robust Scaler is calculated by using the interquartile range (Kumar, 2022).

First, a Hierarchical Clustering method was applied to try to watch some initial results but as it can be seen in Appendix 10, even trying different cluster numbers and changing the linkage type, by itself is not a good enough algorithm to solve our segmentation problem.

K-means clustering was also used to try to define some value segments. Inertia can be recognized as a measure of how internally coherent clusters are (2.3. Clustering, n.d.). By plotting different inertia values depending on the number of clusters it can be estimated a possible optimal range for K-means. From it, a possible optimal solution was observed in a range between 2 to 6 clusters (Appendix 11). To shorten this range, a silhouette score was also made. Using this metric, it could be concluded that the best number of clusters to use with K-means was 2 with a silhouette score of 0.36 (Appendix 12). Using the optimal solution, K-means got a 0.32 R-Squared score.

We also used DBSCAN and Mean-Shift. With DBSCAN we obtained a R-Squared score of 0.06 with a 2 cluster solution (one major cluster and one “residual” cluster) while with Mean-Shift, we obtained a R-Squared score of 0.33 with a 6 cluster solution.

Last but not least SOM was used. After its creation, we implemented Hierarchical Clustering on top of it and calculated R-Squared scores for different parameter values and defined a dendrogram (Appendix 13 and 14). We obtained a final 4 cluster solution with a R-Squared of 0.44. In Appendix 15 it can be seen a HitMap showing the distribution of the 4 clusters in the SOM. Due to its high R-Squared value and good performance on high dimensional data, we decided to use SOM as our clustering algorithm for Value segmentation.

5. Final clusters

In the end, the analysis resulted in 3 demographic clusters and 4 product clusters. We used hierarchical clustering to join the segments together. This method allows us to merge clusters that are similar to each other based on their distance from one another, by using the concatenated cluster centroids as our basis for comparison. After merging the clusters, we created a contingency table for each cluster to better understand the characteristics of each group. In total, we ended up with 4 merged clusters, which we are able to visualize using t-SNE in the appendix 16.

Analyzing the merged clusters (Appendix 18) it is possible to see a decreasing frequency pattern. Therefore, cluster 0 has the highest number of customers and the opposite for cluster 3. Cluster 3 is the one with the most distinct values from the other clusters. These customers have, on average, higher premiums, lower salaries and have the lowest average age. However this cluster contains the customers who have the highest average of Customer Monetary Value. After our analysis and taking into account that cluster 3 presents the highest values of the CustMonVal and Prem variables, we decided to classify this group as our "loyal customers". Cluster 2 seems to be the most similar to cluster 4, regarding the variables "MonthSal" and "Age". The main difference can be seen in premiums, with lower values. Thus, we classified this cluster as our “money saver customers”. In cluster 1 we can observe the customers with the highest average age, as well as salary. However, regarding the CustMonVal variable, this cluster does not present such a positive result compared to cluster 3. We defined this cluster as “older customers”. Finally, in cluster 0, that we defined as our “generic customers”, we can conclude that, despite being the one with the highest number of customers, it presents the values of the variables with the smallest variations.

6. Marketing Approaches & Conclusions

From our analysis, we created different marketing approaches and advices for A2Z to take into account.

Our older customers represent a strong potential due to their high income and overall good engagement with A2Z services. Our goal is not to increase engagement but to get these customers to purchase higher value insurances.

- ➔ Create new positional insurance services with higher premium values that focus on benefits looked by older generations.

Generally, customers that receive less are younger. Due to that, a portion of younger customers who aren't loyal customers are the ones that tend to be more cautious and money savers. This is our "money saver" customers. Even though that's the case we can still try some approaches to promote engagement and interest from them doing the follow:

- ➔ Try a cross-selling approach by creating a global insurance pack that will include all the different types of insurance A2Z offers. Make a global promotion value for people younger than 30 called "A2Z GLOBAL Y".
- ➔ Since this segment focuses more on health and life insurances, create a long-term approach pack for the two of them called "A2Z take care" that gives you different discounts and vouchers for spa, health and fitness related businesses when the two are bought together for a four year fidelity period.

For our loyal customers, a pack of marketing approaches can be defined. Because loyalty should be rewarded, it should be evident to our customers how valued they are. Our loyal customers tend to be young yet with family and work ideals in their mindset. Because of that, and since they don't have high salaries we recommend the follow:

- ➔ Create a loyalty program of five years and for each insurance type subscribed, the user gets a 10%-15% discount in premium values. If two or more insurance types are subscribed at the same time, the user gets a 5% extra discount on top of it.

The ABT provided by A2Z was in need of a deep cleaning and restructure. Either way, we were able to achieve good results with what was provided to us, taking the time to analyze the overall dataset and getting a notion of our customers. As it can be seen in appendix 17, using a classification tree, we were able to represent the different stages of partition in our clustering. It could be seen that Monthly Salary has been the most decisive variable in our partition and that customers spend overall more money in HouseHold insurance than the rest. Our Classification Tree estimates that on average, we are able to predict 95.62% of the customers correctly. We hope that A2Z takes on consideration the various insights given in this report to support a stronger understanding of their customers and a stronger long-term relationship with them.

7. References

2.3. Clustering. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/clustering.html

Kotler, P., & Keller, K. L. (2012). *Marketing Management*. Prentice Hall.

Kumar, V. (2022, February 6). *Sklearn Feature Scaling with StandardScaler, MinMaxScaler, RobustScaler and MaxAbsScaler*. MLK - Machine Learning Knowledge. https://machinelearningknowledge.ai/sklearn-feature-scaling-with-standardscaler-minmaxscaler-robustscaler-and-maxabsscaler/

Nguyen, X.-M., Debnath, B., & Fu, Y. (2011). K-prototypes: A clustering algorithm for mixed numeric and categorical data. Knowledge and Information Systems, 29(2), 283-312.

8. Appendix

Variable	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 (Children=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

Figura 1: Metadata

	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10293.0	1991.050131	510.596893	1974.00	1980.00	1986.00	1992.00	53784.00
BirthYear	10293.0	1968.006898	19.694456	1028.00	1953.00	1968.00	1983.00	2001.00
MonthSal	10293.0	2506.602545	1155.492229	333.00	1707.00	2502.00	3289.00	55215.00
CustMonVal	10293.0	177.929963	1946.091554	-165680.42	-9.44	187.03	399.86	11875.89
ClaimsRate	10293.0	0.742728	2.917385	0.00	0.39	0.72	0.98	256.20
PremMotor	10293.0	299.508928	212.288828	-4.11	190.26	298.39	407.52	11604.42
PremHousehold	10293.0	210.419863	352.635041	-75.00	49.45	132.80	290.05	25048.80
PremHealth	10293.0	170.836503	296.031364	-2.11	110.91	162.03	218.93	28272.00
PremLife	10293.0	41.425887	47.428782	-7.00	9.89	25.45	57.01	398.30
PremWork	10293.0	40.938779	51.440777	-12.00	10.00	25.56	56.01	1988.70

Figura 2 - Descriptive Statistics - Metric Features

	count	unique	top	freq
EducDeg	10293	4	3 - BSc/MSc	4816
GeoLivArea	10293.0	4.0	4.0	4143.0
Children	10293.0	2.0	1.0	7281.0

Figura 3: Descriptive Statistics - Non-metric features

```
FirstPolYear      30
BirthYear         17
EducDeg           17
MonthSal          36
GeoLivArea         1
Children          21
CustMonVal         0
ClaimsRate         0
PremMotor         34
PremHousehold      0
PremHealth        43
PremLife         104
PremWork          86
dtype: int64
```

Figura 4: Missing Values

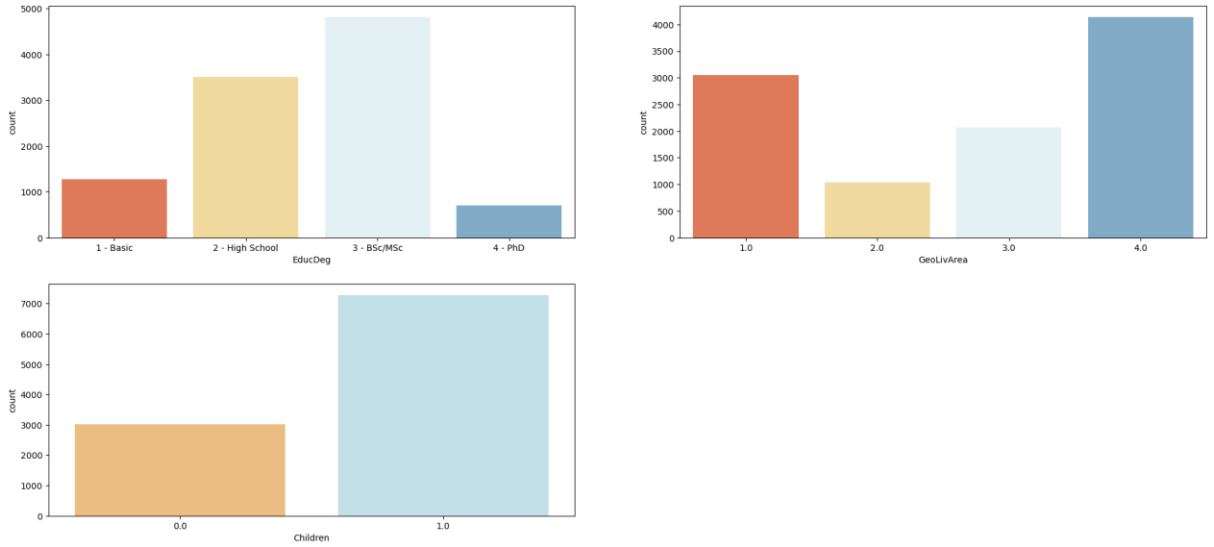

Figura 5: Bar Charts

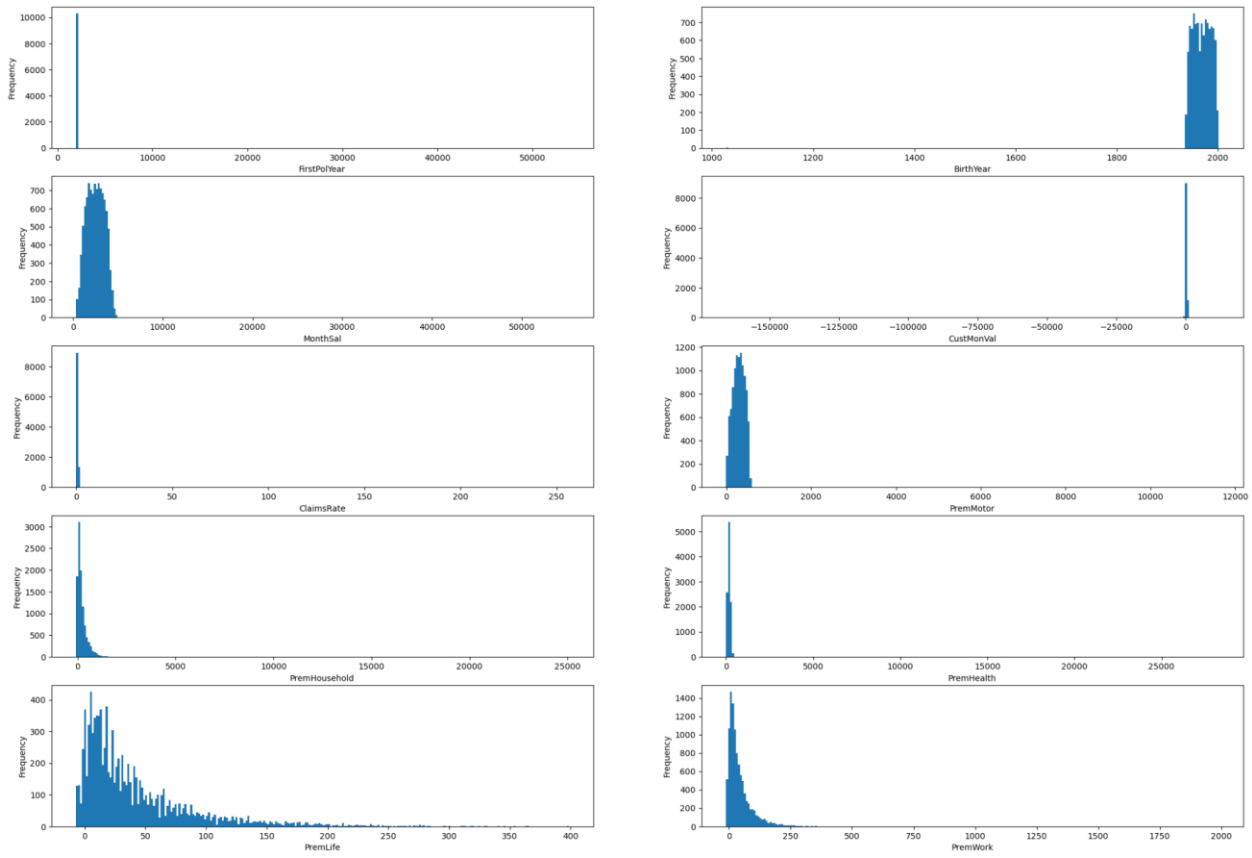

Figura 6: Histograms

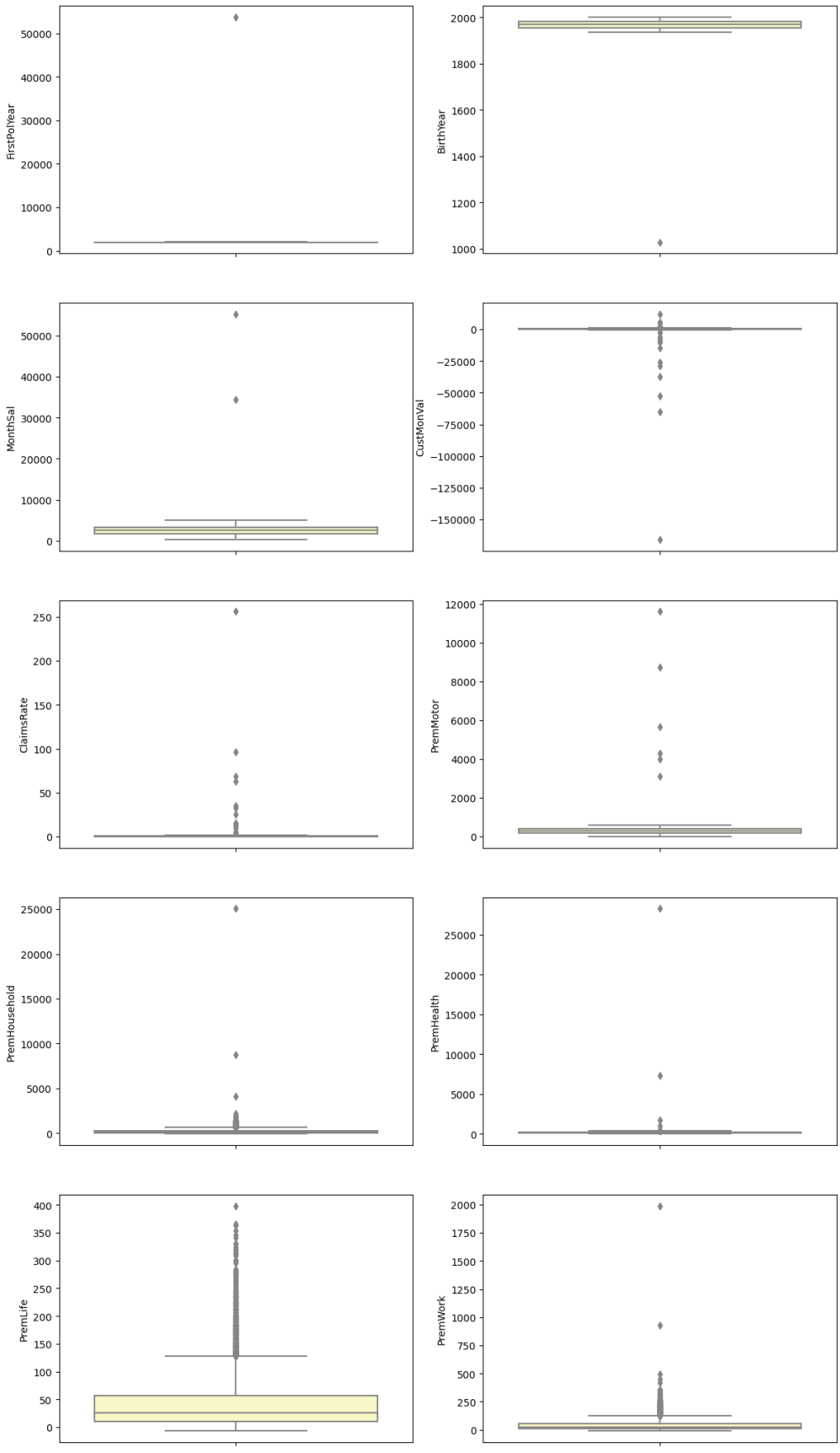

Figura 7: Initial Boxplots

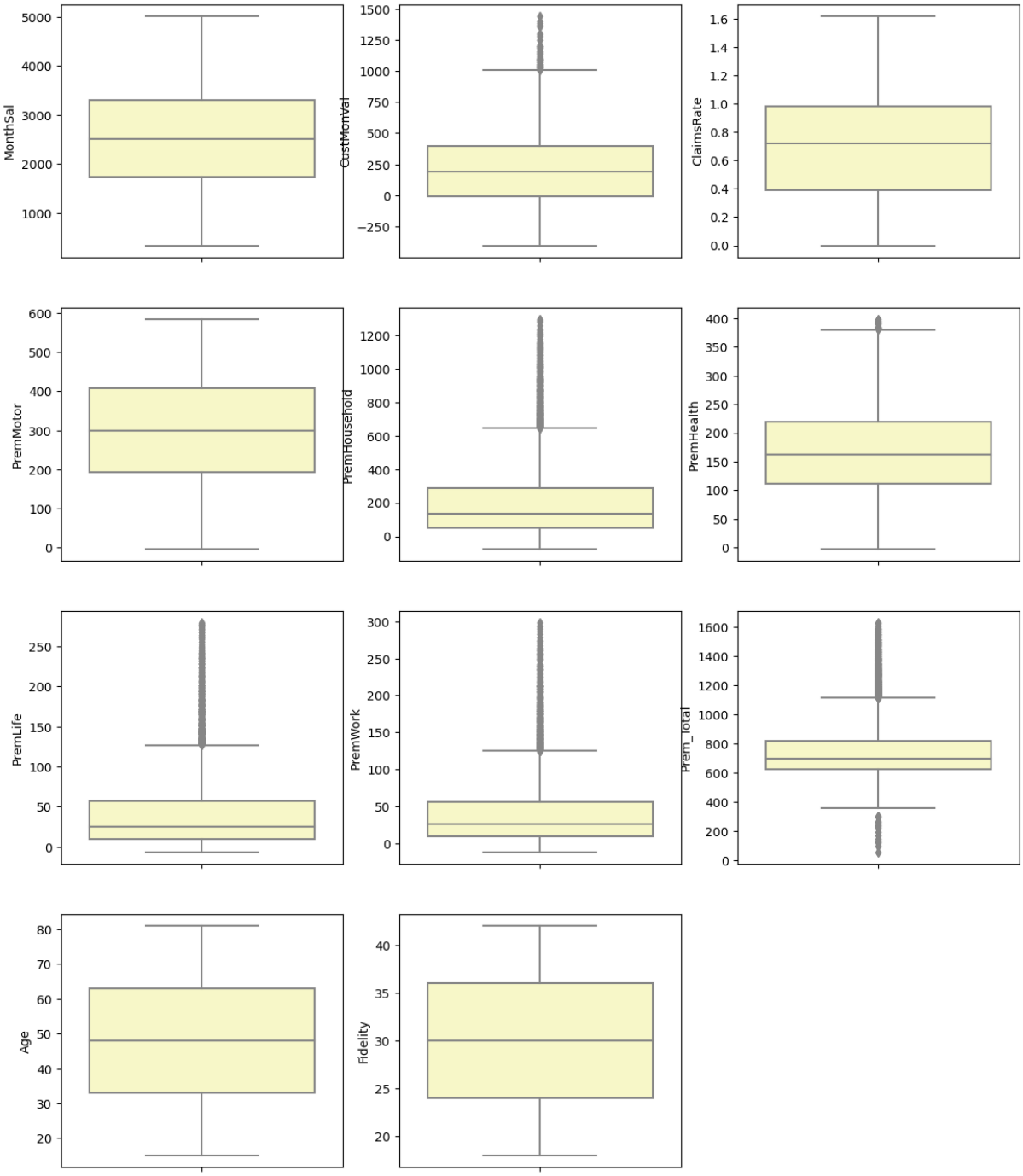

Figura 8: Final Boxplots

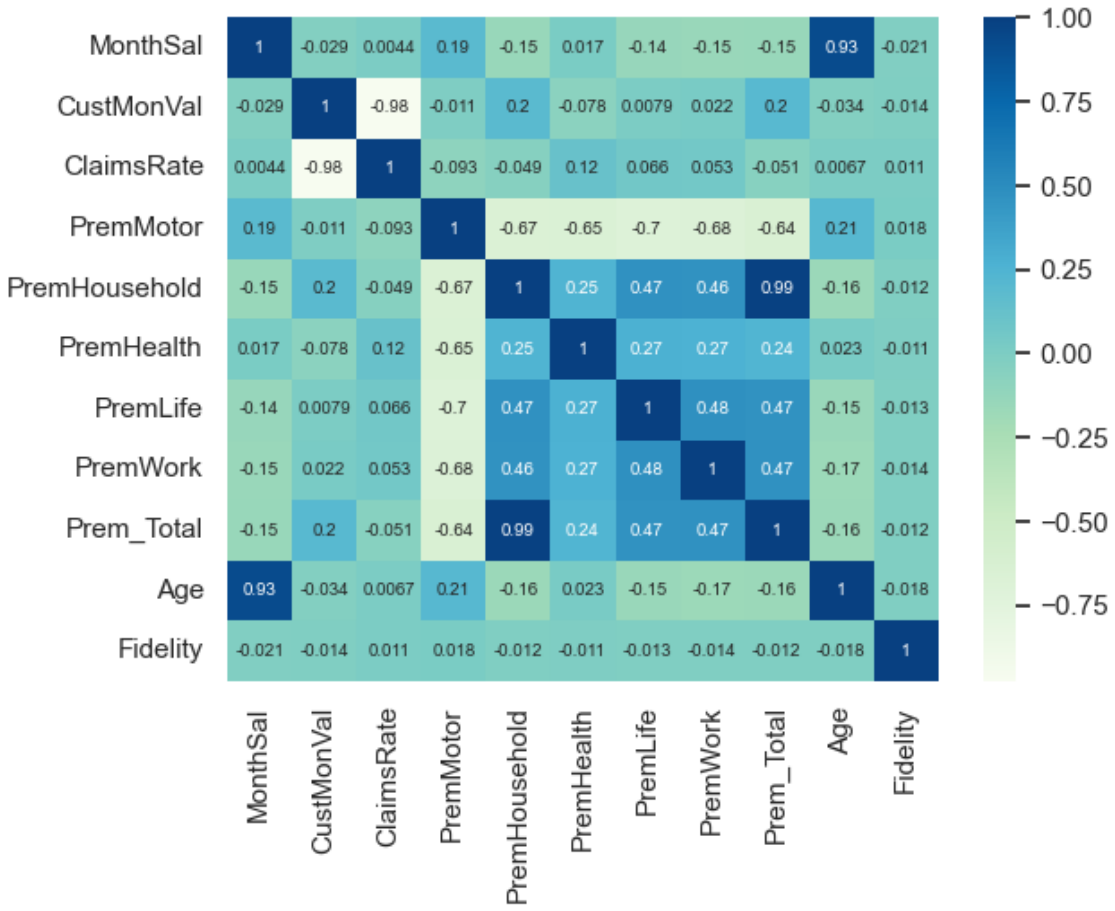

Figura 9: Correlation Matrix

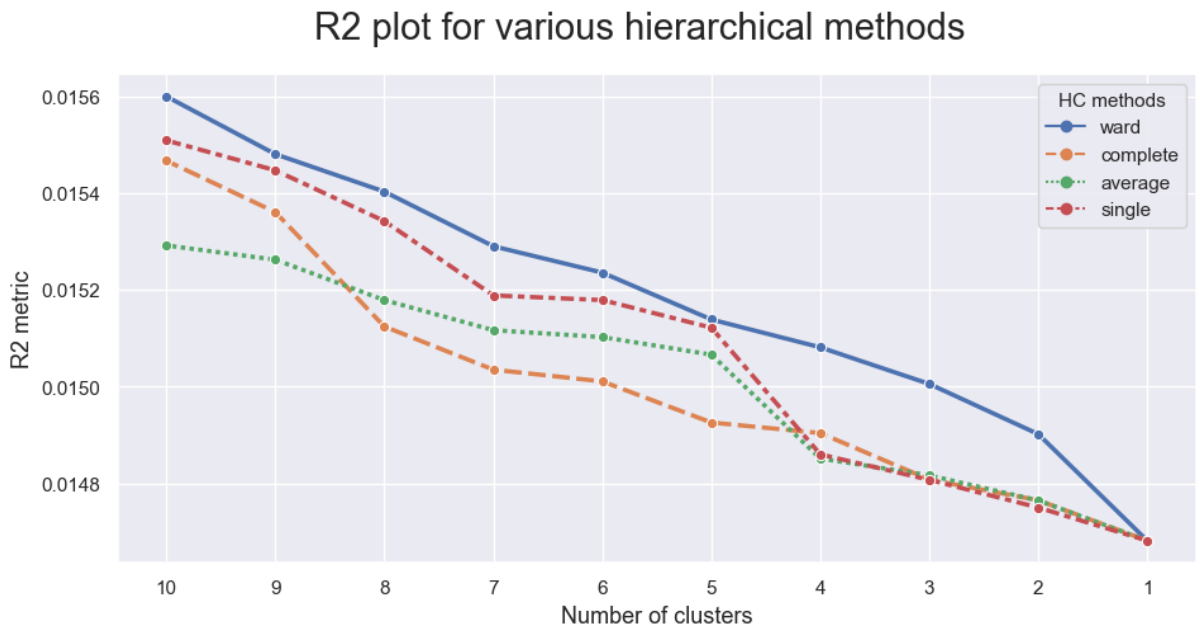

Figura 10: Hierarchical value segmentation

Figura 11: Inertia K-means value segmentation

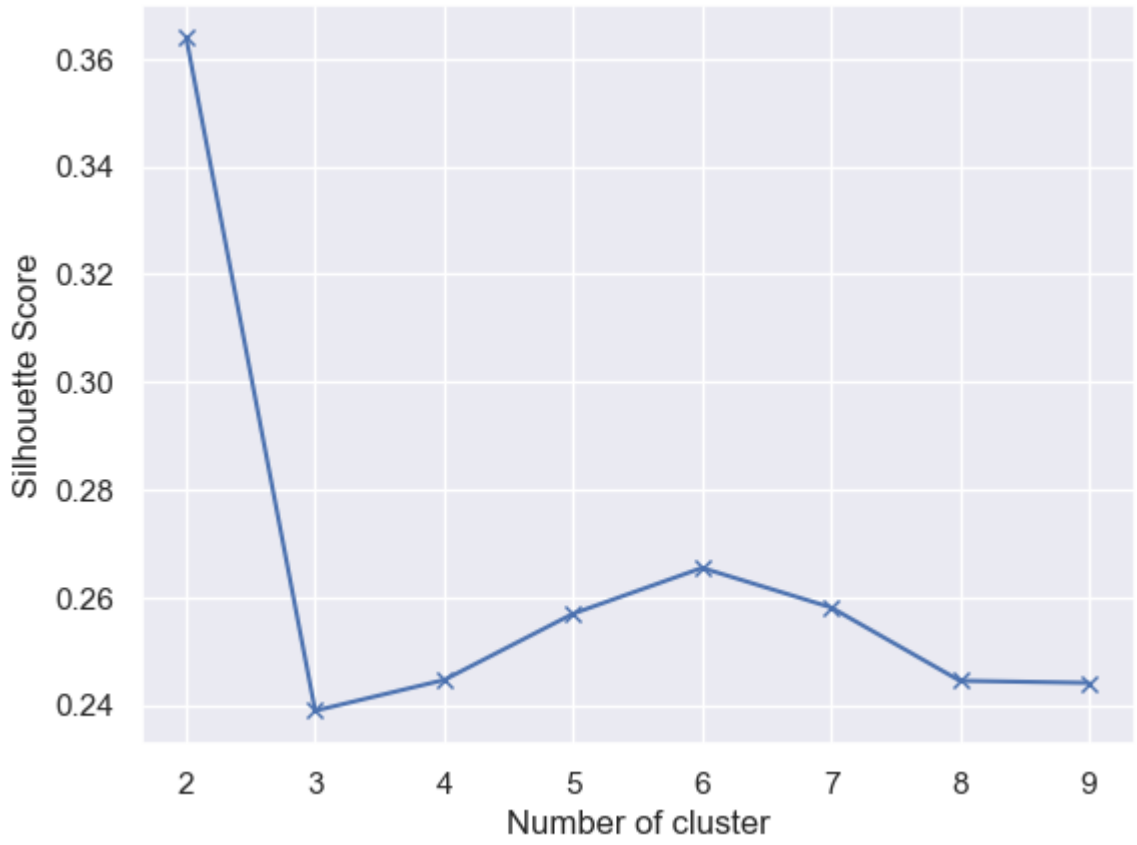

Figura 12: Silhouette score Kmeans

Figura 13: SOM Value

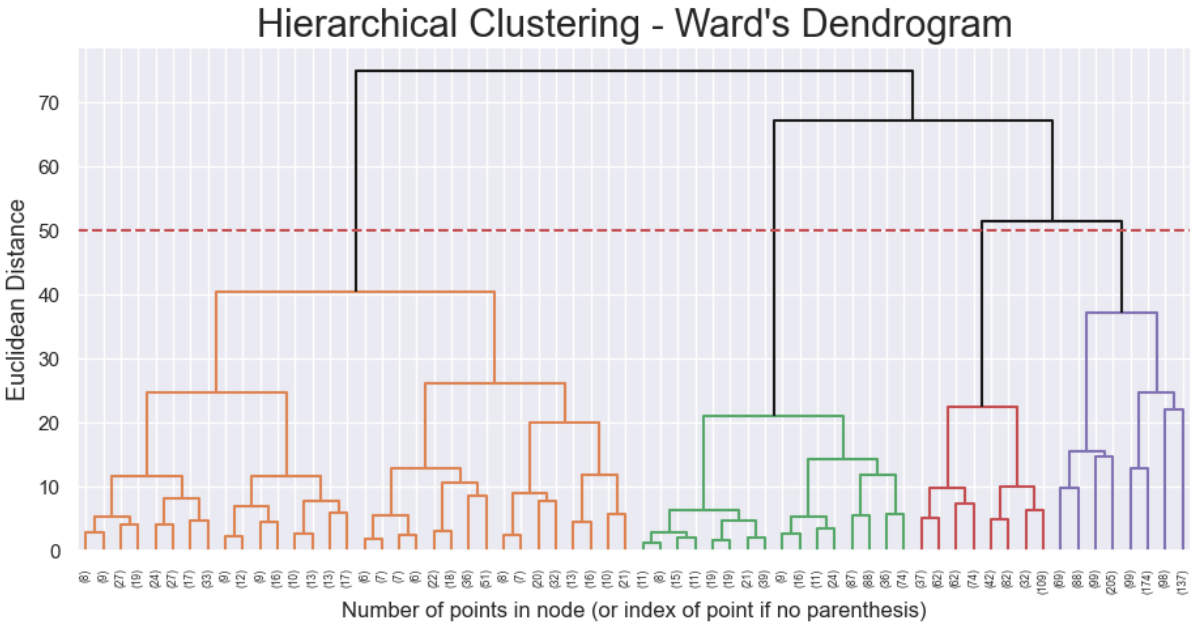

Figura 14: Dendrogram

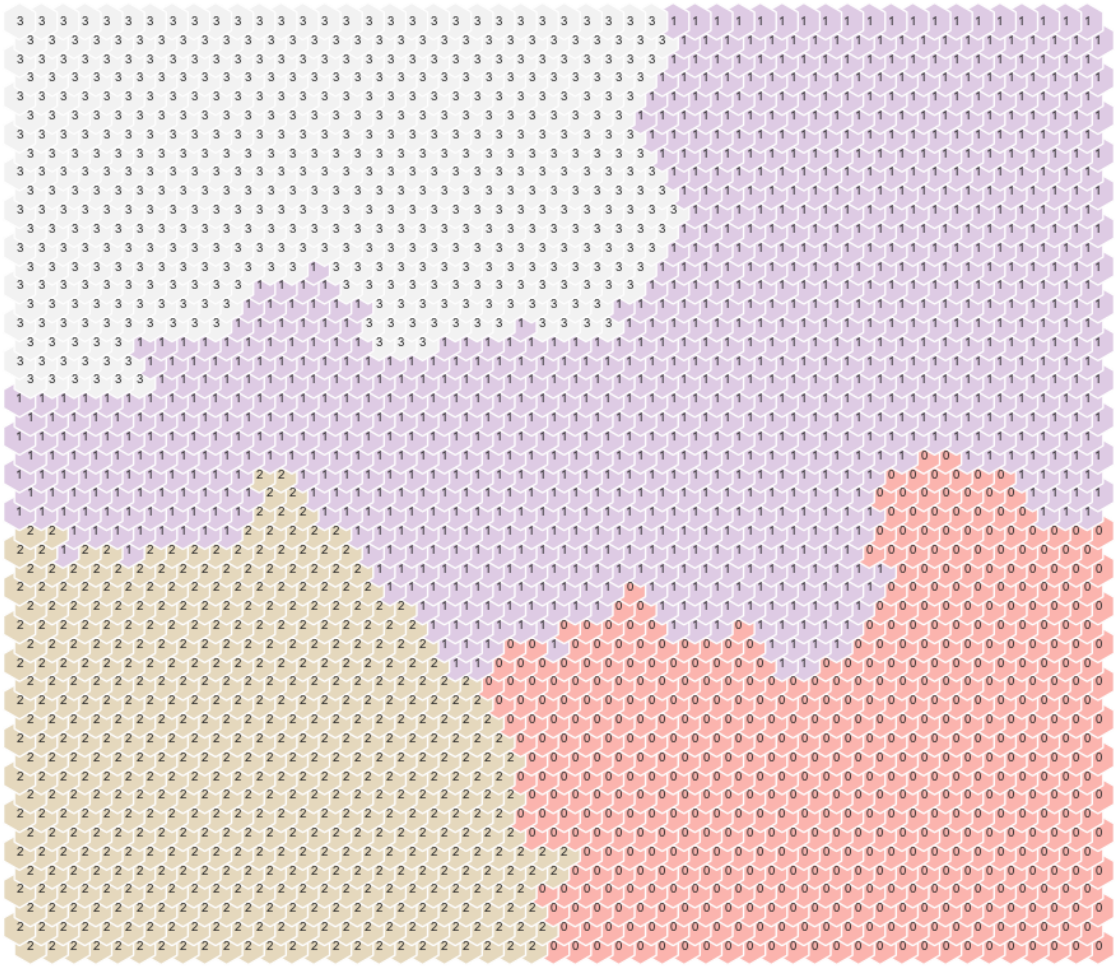

Figura 15: Clustering

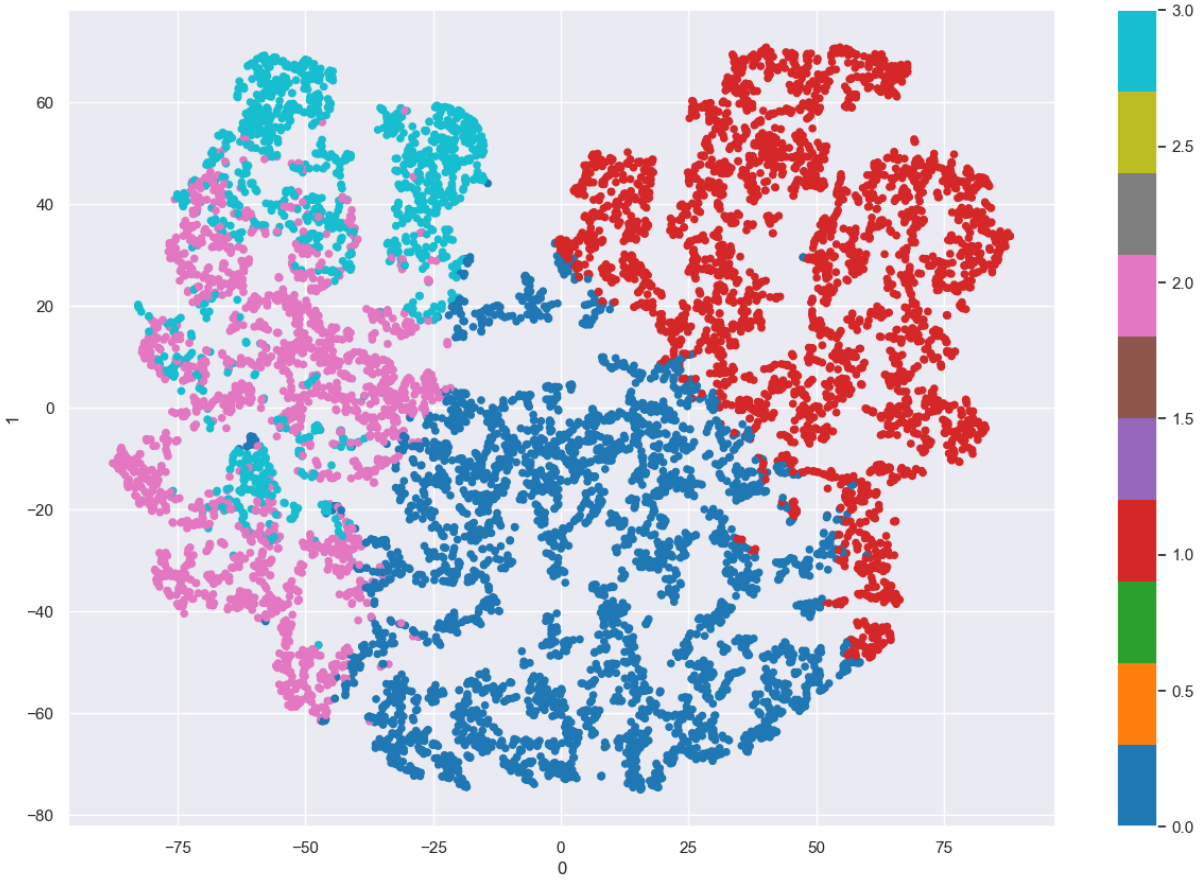

Figura 16: t-SNE 2D representation of the obtained clusters.

Image X.4

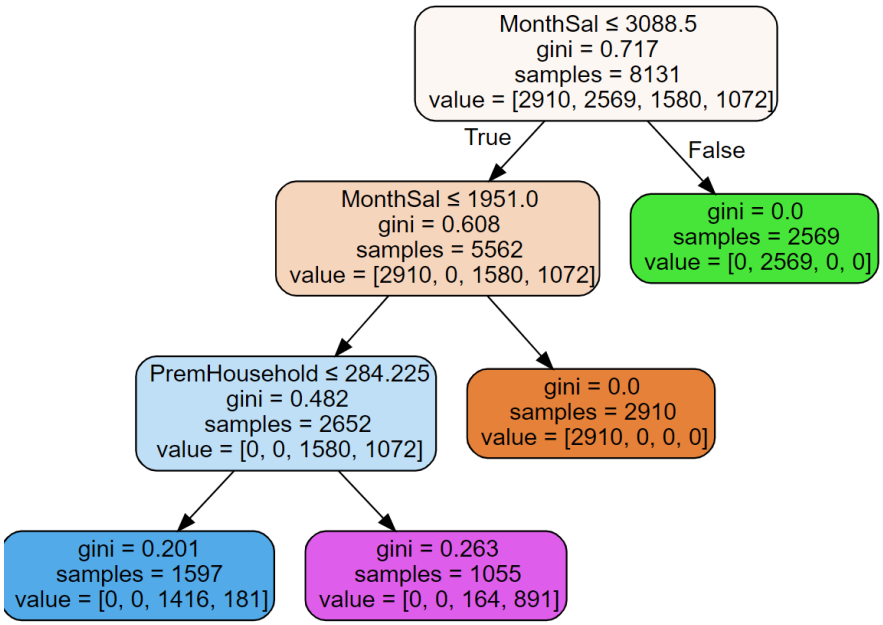

Figura 17: Classification Tree, representing the stages of partition in our clustering. Monthly Salary has been the most decisive.

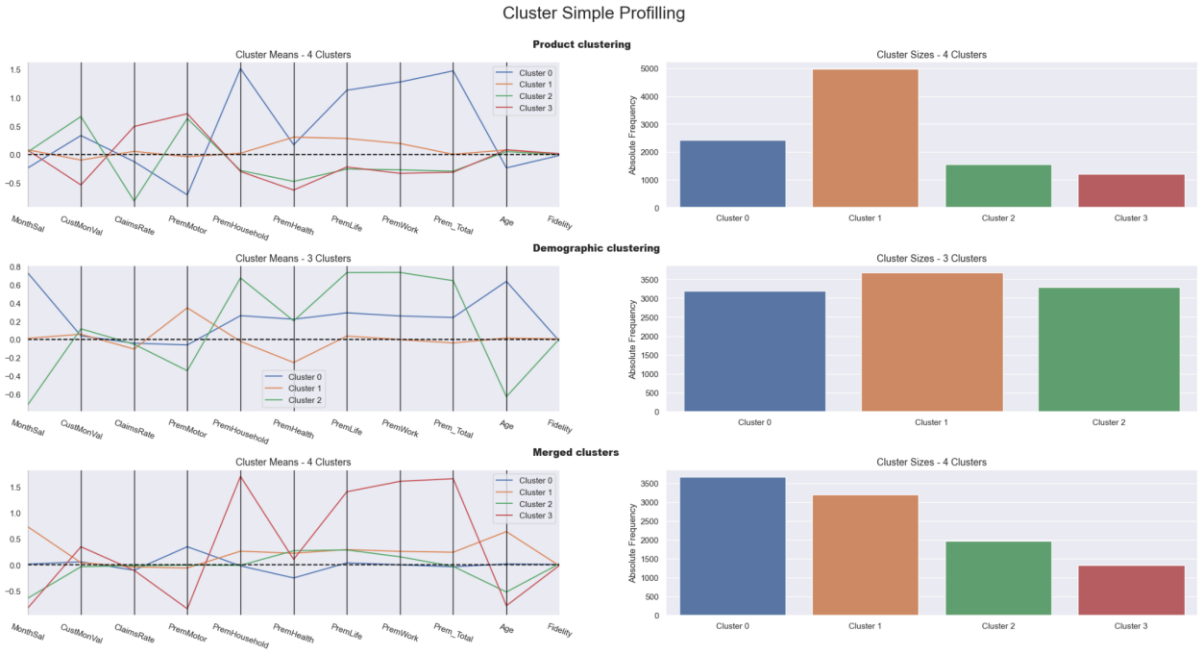

Figura 18: Profilling