Universidade NOVA de Lisboa
NOVA Information Management School

Ivan Jure Parać
Nuno Melo Bourbon
Stuart Gallina Ottersen

A2Z INSURANCE

Lisbon, 2021

Universidade NOVA de Lisboa

NOVA Information Management School

L I S B O N

Ivan Jure Parać
Nuno Melo Bourbon
Stuart Gallina Ottersen

Programme: Data Science
Course: Data Mining

A2Z INSURANCE

Mentors:

Fernando Bação
Joao Fonseca
David Silva

Lisbon, December 2021.

Ivan Jure Parać
Nuno Melo Bourbon
Stuart Gallina Ottersen

Statement of Authenticity

Hereby We state that this document, Our project, is authentic, authored by Us, and that, for the purposes of writing it, We have not used any sources other than those allowed for this project. Ethically adequate and acceptable methods and techniques were used while preparing and writing this report.

Abstract

This report was prepared for A2Z Insurance in an effort to improve customer retention and acquisition by finding similarities between existing customers and building customer segments that can be targeted using directed marketing approaches. The original data, consisting of customer personal and business-related information, was first explored before undergoing several preprocessing steps to be used for clustering. Multiple clustering algorithms were tested and evaluated to create clusters based on sociodemographic, value, and product features. Finally, the clusters obtained with these three approaches were merged together to arrive at the four final customer segments. For each segment, possible marketing approaches were recommended.

Keywords: data mining; python; big data; preprocessing; data analysis; clustering; unsupervised learning

Table of Contents

1. Introduction

The present report was prepared for the A2Z insurance company with the goal of defining customer segments based on data collected for more than 10,000 clients. It describes the steps, approaches, and decisions taken to produce the final segments, as well as the rationale behind them. It is divided into four main sections, namely data exploration, data preprocessing, clustering approaches, and major insights from the customer segments obtained. The conclusions of this report are expected to guide the strategies taken by the Marketing department and ultimately boost the company's revenue.

2. Data Exploration

"*Data exploration is about efficiently extracting knowledge from data even if we do not know exactly what we are looking for.*" [1]

The dataset provided consisted of 10,296 observations, each representing a customer. It contained 14 features, of which only one was not in numeric format, including customer's personal information, metrics of their value to the company, and premium paid per type of insurance policy (***Table 1***). An early assessment of the dataset revealed a total of 372 missing values across 9 features and the presence of extreme values representing possible outliers. Before addressing any major issues, we set *customer ID* as an index and searched for duplicated entries. A total of 3 duplicates were found, which were dropped, thus producing a dataset with 10,293 observations and 13 features. A regex expression was used to retrieve only the numbers from the *EducDeg* column, which was then converted to *float* so that all data was numeric.

A major incoherence was found when looking at the *policy* and *birth years*, which was that ~19% of customers signed their first policy before they were born. Such a high number of inconsistencies suggests a systematic error, which we assumed to be the insertion of the years into the wrong fields. In these cases, the policy and birth years were swapped. We also found it strange that no customers were acquired after 2001. This suggests either the database stopped being updated or a lack of effort in customer acquisition. In fact, the number of new customers steadily declined from 1997 to 2001 (**Figure 1**).

Further exploration revealed that the average customer is expected to be 50 years old and earn a salary of 2,507€. More than half (53%) of the customers have, at least, a Bachelor's degree, and over 70% have children (**Figure 2**). The *Children* feature was assumed to be referring to under-aged children, as the proportion of customers with children decreases as they get older. Not much is known about where customers live since no information other than a label-encoded geographical location was provided.

Almost 22% of all customers were found to have at least one *negative premium*, which could indicate mere price adjustments or early policy cancellation. If interpreted as the latter, expenditure

on reacquiring these customers should be limited. The expected spending habits of these customers were also found to be different, spending considerably more on *motor* and less on all other premiums. Further, ~27% of customers had a *negative monetary value*, meaning that the company loses money with approximately 1 out of 4 customers.

3. Data Preprocessing

3.1. Univariate Outlier Detection

When looking at the box-plots of metric features we noticed two obvious outliers in *FirstPolYear* (53784) and *BirthYear* (1028) (**Figure 3**). We removed the entire row for the former but kept the latter and replaced it with 1928, assuming it arose from a typing mistake due to the proximity between 0 and 9 in a typical *qwerty* keyboard.

For the remaining features, deciding which observations constituted outliers was not as simple because, despite their extremeness, we had no way of knowing whether they were real or incorrect. Given our limited domain knowledge, we decided to take a conservative approach, and manually removed only 44 observations (0.43% of the data), corresponding to the most extreme values, to prevent them from biasing the clusters produced (**Figure 3**). The outliers were saved in a separate dataset to be labeled after defining the clusters.

3.2. Dealing With Missing Values

Depending on the feature, rows with missing values were either dropped or imputed, as shown in **Table 2**. The same missing value treatment was applied to the outliers removed from the main dataset.

Before removing rows with missing birth years we attempted to impute them, but in at least one row the estimated birth year occurred after the first policy. Since the number of missing values was fairly low our final decision was to simply discard them. In total, 58 rows were dropped due to missing values: 12 due to incompleteness, 44 due to missing policy and/or birth years, and 2 due to missing information regarding education.

Missing salary values were imputed using simple linear regression. *BirthYear* was chosen as the regressor based on an almost perfect (-0.9) negative correlation with salary. *K-Neighbors* regressor produced worse results, and while a multiple linear regression led to a slight improvement, it was too marginal to justify the increased complexity. To impute missing values in *Children*, logistic regression was applied. Again, *BirthYear* was used as a regressor, as indicated by a high correlation with *Children*, *Recursive Feature Elimination*, and *LASSO* regression coefficients **(Figure 4)**.

3.3. Feature Engineering

A total of 12 new features were created in an attempt to obtain new ones that would aid the clustering efforts (**Table 3**). A few of these, like *Age*, are "quality-of-life" features, in the sense that they were created because they were easier to interpret than the original ones. As the premium values provided are annual, monthly salary was converted to yearly salary.

A total of 27 outliers were removed from the newly created features following the approach previously described. These were again stored in a separate dataset.

3.3.1. Feature Transformation

Some of the metric features were heavily skewed, presenting a long right tail. This can negatively impact the clustering process [2]. In an effort to normalize these distributions, both square root, and log_{10} transformations were tested, after ensuring no negative and only positive values, respectively. In the end, square root transformation was chosen for producing smoother distributions than the alternative. The upper values of *PremTotal* and *EffortRatio* were also clipped to decrease skewness. Final results are shown in **Figure 5**.

3.3.2. Scaling

Different combinations of scalers (*MinMaxScaler*, *StandardScaler* and *RobustScaler*) and clustering methods were tested [3]. In general, *RobustScaler* did not perform as well as the other two (as evaluated by intrinsic methods and visual inspection of the clusters formed), probably due to it being based on interquartile ranges, and only be applied after removing the most extreme observations.[4] Overall, *MinMaxScaler* performed better, and was the chosen scaler for the sociodemographic clustering, where the features used had no outliers.[5] For value and product clustering, we chose *StandardScaler* due to its higher robustness to outliers. It standardizes features by removing the mean and scaling to unit variance. A normal variance of data is a requirement for clustering algorithms like K-Means, without it they will behave badly and give out incorrect results. [6]

3.4. Feature Selection

We found we could subdivide the features available into three big categories: *Sociodemographic*, *Value*, and *Product*. To make the problem easier to tackle, and to more easily understand customer profiles, clustering was done for each of these segments. For each segment, features were selected based on their relevancy and redundancy, as determined by their correlation (Pearson) and continuous evaluation of the clusters obtained.

3.4.1. Sociodemographic Clusters

Sociodemographic features available included *Generation*, *Age*, *YearSal*, *EducDeg*, *Children*, and *GeoLivArea*. Clusters obtained with these features can help identify trends in customer behavior

based on their personal information. Only yearly salary, education degree, and children were used during the clustering process. *Age* and *EducDeg* were dropped due to their high correlation (and likely redundancy) with *YearSal*, which also suffered less extensive preprocessing seeing as 19% of its values were not swapped. *GeoLivArea* was discarded because of its ambiguity and low discriminant power.

3.4.2. Value Clusters

Value clustering was carried out using features related to customer-business relationship and profitability, including *FirstPolAge*, *CustYears* (which replaced *FirstPolYear*), *CMV*, *ClaimsRate*, *PremTotal*, and *EffortRatio*. These features contain information about the duration of the customer relationship, the total premium paid, the value of the customer, and spending behavior. The goal was to identify the most valuable customer segments, retention rates, and pinpoint any customer segments where money was being lost.

3.4.3. Product Clusters

Product clustering used features related to the premiums paid by the customer for each insurance type. It is expected to help identify customer spending habits. When creating a marketing approach, product features are typically combined with sociodemographic features such as *Age*, *YearSal*, or *Children* so that specific groups of people can be targeted.

3.5. Multivariate Outlier Detection

Unlike univariate outlier detection, which checks if each individual value is an outlier, multivariate outlier detection checks if the combination of features results in an outlier. While centroid-based clustering algorithms, such as K-Means, can be used to detect outliers, this should be avoided as they require prior information regarding the number of clusters and their size, and assume a flat geometry. Because this was not known, we opted for a density-based approach, namely DBSCAN.

Density-based algorithms can be used to detect outliers because they create clusters based on the density of observations. Because outliers will, by definition, be placed outside of high-density regions, they are left out of the clusters formed. Based on the results obtained with *DBSCAN*, we removed a total of 92 observations, which were included in the separate dataset for the outliers [7].

By the end of preprocessing, a total of 221 observations were removed, representing 2.15% of the original dataset. The remaining 10,072 observations were used to create clusters.

4. Clustering

4.1. Sociodemographic

Sociodemographic clustering was performed based on three features, namely *YearSal*, *EducDeg*, and *Children*. Despite being categorical, the ordinal nature of the level of education allowed it to be used as a metric feature in distance-based algorithms without having to encode it as dummy variables, even if the low cardinality over-weighed its importance. This issue of cardinality was not so easily ignored for *Children*, due to its binary nature, which makes the application of distance-based clustering algorithms less than ideal. For this reason, when testing most clustering methods, we did not include children among the selected features. DBSCAN and BIRCH, being density-based, clearly highlighted the issues of using low cardinality features, by separating customers according to their education and mostly ignoring their salaries **(Figure 6)**.

To use the information provided by *EducDeg* and *Children* features while avoiding issues caused by low cardinality, we decided to use *K-prototypes* as our final solution for sociodemographic clustering. *K-prototypes* still uses a distance measure but adds a similarity measure for categorical features (in our case, *EducDeg* and *Children*) that uses the number of matching categories [8]. A total of 4 clusters were created by this approach **(Figure 7)**.

4.1.1. K-prototypes: Theoretical Explanation

K-prototypes is a variation on *K-means* and *K-modes* that allows it to handle mixed data types. *K-means* uses Euclidean distance which is not fitting when working with categorical values, while *K-modes* uses dissimilarities between data points as a measure of distance. All three algorithms are based on the same concept where K initial means/modes/prototypes (called centroids from here on) are selected, each observation is then assigned to the closest of these K centroids based on a dissimilarity measure. After all observations have been assigned the position of the centroid is recalculated based on all the points assigned to it, the points are then retested and possibly reassigned to a new centroid. After a number of iterations, no change is seen and we have arrived at our final clusters. What differentiates *K-prototypes* from the other algorithms is its dissimilarity measure [8].

$$d2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \tag{4.1}$$

The first term in this dissimilarity measure represents the Euclidean distance of the metric attributes, while the second term is the matching dissimilarity in categorical attributes. The γ is used to avoid favoring either type of data [8].

4.2. Value

Testing clustering for value was quite extensive and many of the attempted methods produced similar results. Some methods were easy to exclude, such as *DBSCAN* and *AffinityPropagation*, which created either too many clusters or very uneven clusters, regardless of the hyperparameters used. We selected the two main contenders from among the remaining algorithms based on R^2 values and the evenness of the distribution of observations by the different clusters. As such, our chosen approach for value clustering was to use *SOM* followed by *K-means* **(Figure 8)**. We initially used five features, but eventually narrowed it down to only 3 (*FirstPolAge*, *PremTotal*, *EffortRatio*, based on their relative importance to defining the final clusters after merging the different approaches. A total of 6 clusters were created from these three features (**Figure 9**). It was interesting to find that neither *CMV* or *ClaimsRate* were relevant to differentiate the final clusters.

4.3. Product

For product clustering, two different feature combinations were tested. The first one used the values of the premiums (Motor, House, Health, Life, and Work), while the second used the premium ratios created by dividing each of the premiums by the total premium (MotorRatio, HouseRatio, HealthRatio, LifeRatio, and WorkRatio). After testing multiple clustering methods, we concluded that using premium ratios did not improve the clustering solutions obtained, and also made the results harder to interpret. Therefore, we ended up choosing the original premium features rather than the ratios. Most clustering algorithms produced fairly similar results, with no one solution standing out as the best. In the end, we based our decision on the algorithm that captured the most variance, as assessed by the R^2, which was the combination of *SOM* with *K-means* **(Figure 10)**. The characteristics of the 5 final clusters obtained are shown in **Figure 11** [9].

4.4. Merging the clusters

The labels obtained with the three clustering approaches were combined to produce a total of 108 clusters of various sizes, each containing 1 to 800 observations (**Figure 12**). To have a manageable number of clusters, we chose 300 observations as the threshold for the minimum number of observations a cluster needed to have to be considered as such. This resulted in the creation of 8 clusters and 6,029 unlabelled observations.

To assign the unlabelled observations to the existing clusters, we attempted hierarchical clustering, semi-supervised learning, and decision tree classification. While *AgglomerativeClustering* and *LabelSpreading* produced results similar to those obtained with a decision tree-based approach, we decided to use the latter as it is robust to the presence of outliers and handles mixed data types fairly well. Specifically, we used a *RandomForestClassifier*, which was more accurate in predicting the originally assigned labels than a simple *DecisionTree*. The outliers were assigned to existing clusters in a similar fashion. The 8 clusters were then manually merged based on similarity and marketability, which further decreased their number down to 4 (**Figure 13**), which were considered to be a more manageable number of clusters from a marketing point of view .

While we used Uniform Manifold Approximation and Projection (UMAP) to visualise the clustering results before and after manually merging them, there is one important caveat, which is that education degree and children were not treated as metric features, but as categorical, by K-prototypes. These two features were left out of UMAP and, as such, we were cautious about making inferences regarding the quality of the clustering approach chosen based on this method. Nevertheless, the clusters produced seemed adequate as there were clear patterns in how the observations were distributed in the UMAP, both before and after manually merging them (**Figure 14**). In the final solution, two (orange and green) of the four clusters have a higher degree of overlap but were not merged due to differences in education, as well as the amount paid for house and motor premiums.

5. Customer Profiles

Based on the four final clusters obtained we suggest four profiles that are broad generalizations of the type of customers each segment might contain. Customer acquisition strategies, as well as ways to improve profitability and improve relations with existing customers, are proposed below.

5.1. The Home Buyers with High-risk Jobs

Adults in this segment are expected to be younger and have a lower level of education. Their salary is also the lowest of the four segments. Despite this, they have the highest CMV, which can be attributed to higher expenses (compared to the remaining segments) with life, work, and house insurances. Life and work premiums are higher for this segment likely due to the tendency of people with lower education have to assume higher-risk jobs, which prompts the insurance company to increase premium rates. On the other hand, high expenditure on house insurance premiums could be related to the acquisition of the first house.

As a customer acquisition strategy, we propose working with real estate companies to incorporate A2Z's house insurance products in the services these companies provide to young homeowners. Special partnerships with labor unions and syndicates could also help young, uneducated people have access to information regarding advantageous life and work insurances which they might otherwise not hear about.

On the other hand for customers in this segment or those with similar characteristics, a directed advertisement campaign can be effective. Here the existing relationship with the customer improves the probability of purchase.

5.2. Expensive Car Owners

Recognizable by a high level of education, and expected age of 50 years old, and a high probability of having children, this segment spends a significant sum on motor insurance policies. It likely includes people who have paid their housing mortgages and are comfortable enough in life that

they can now afford to spend more money on certain luxuries, such as cars and respective insurances.

Similarly to the home buyers, we suggest collaborations with car dealerships to acquire customers with similar profiles. We believe advertising the insurance as part of the car price can help find new customers, and that once a customer relationship has been established other products are more easily sold.

For the existing customer, we again suggest doing directed ad campaigns to up-sell an improved, more expensive version of the current policy. Seeing as this is the segment that spends the least on all other types of policies, it may be worth developing some cross-selling strategies, such as additional motor insurance perks with the subscription to one of the other policies.

5.3. The Retirees

This segment consists of elderly people who spend moderate amounts of money across all insurance types. They are expected to have the highest salary which means that even though they have high spending, it is still a small percentage of their income.

To obtain more customers in this segment we suggest creating "ease of mind" packages, as the cost is less of an issue. These would be broad policies that include most insurance types at a higher premium but which would also provide access to higher quality services. These high-end services can be advertised combined with an "ease of use" aspect that retirees will likely find appealing.

Similarly targeted advertisements can be applied to existing customers. For instance, this customer segment is the one that spends the most money on health insurance. On one hand, this may be because of higher prices imposed by the insurance company, given that elderly people are generally more at risk of having health issues. However, it may also indicate that this segment is particularly worried about their health. Therefore, they could be interested in more expensive policies covering health expenses that their current policies do not.

5.4. Family Package

This segment has people with the highest education and a very high percentage of people with children. They have a very low average age for their first policy, which might indicate that they are children of previous or current customers. They are one of the least valuable customers, but not by a large margin. The spending is predominantly on car premiums, the house being on the lower together with health. Being quite similar demographically to the cluster of house buyers, we think it might contain people that have not bought a house but rather rent a home but have disposable income to spend on cars. This might also be reflected in the effort ratio which is around half of the home buyers.

To acquire more customers we suggest advertising "family" packages. The packages could include bundles for car, house, and/or health, and give out free or very cheap insurance for children, which would, in turn, make it more likely for those children to later stay at A2Z Insurance. After being

a customer from childhood it is that much harder to leave so this approach makes it not profitable short term, but long term it can produce really good results. The reason why we recommend this segment for the family package is that it had the highest number of children if we do not count the 50-year-old segment which has more children but those children are going to be older on average, so it is better to target younger parents with younger children.

One more viable strategy would be an attempt at "converting" an existing customer to a higher profit segment which would either be done by contacting the customer with new deals, bundles, or options, or "forced" conversion when the ongoing plan is canceled in next couple month, which forces customers to switch their plans. The first approach is soft and much more likely to not aggravate the customer, while the second approach is the hard one, which risks customers leaving for another insurance company.

6. Conclusion

We believe that A2Z insurance can create directed marketing campaigns to the aforementioned segments to improve customer acquisition, which seems to have been neglected since 2001. At the same time, the segments created may be used to improve relations with existing customers, as they make it easier to recommend new and improve existing services. The strategies proposed in this report were designed to have both of these goals in mind.

Acquiring new customers has to be a priority since as time passes there are going to be fewer customers who are still with the A2Z insurance agency. Before that number gets too low and the company starts hemorrhaging money, it is paramount to invest in marketing campaigns and get some "new blood". A good place to start would be to target future customers with children, who would provide longevity to the company. Since this approach would make money in the future but lose money at present, other customer profiles that pay more at the moment should also be targeted so they minimize the risk of overspending.

At the same time, existing customers can be marketed towards with phone calls or e-mails with new bundles that would give them more options, but that would cost more which is beneficial to the company.

Bibliography

[1] Stratos Idreos, Olga Papaemmanouil, Surajit Chaudhur, "Overview of data exploration techniques,"

[2] Keke Chen and Ling Liu, "Cluster rendering of skewed datasets via visualization," 2003.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] *Sklearn.preprocessing.robustscaler*, Available at `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html`.

[5] *Sklearn.preprocessing.sklearn.preprocessing.minmaxscaler*, Available at `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`.

[6] *Sklearn.preprocessing.standardscaler*, Available at `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`.

[7] *Sklearn.cluster.dbscan*, Available at `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`.

[8] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery volume*, vol. 2, no. 2, pp. 283–304, 1998.

[9] K. Pang, *Self-organizing maps*, Available at `https://www.cs.hmc.edu/~kpang/nn/som.html`, 2003.

List of Figures

List of Tables

Appendix

1. Tables

Feature	***Description***
CustID	Unique customer ID
FirstPolYear	Year of the customer's first policy
BirthYear	Customer birth year
EducDeg	Education degree (Basic, High School, BSc/MSc, and PhD)
MonthSal	Customers gross monthly salary (€)
GeoLivArea	Label-encoded customers geographical location
Children	Binary values representing does customer have children or not
CustMonVal	Lifetime value = (annual profit from the customer) X (number of years that they are a customer) - (acquisition cost)
ClaimsRate	Amount paid by the insurance company (€)/ Premiums (€) in the last 2 years
PremMotor	Customers spending on premiums in regards to *motors*
PremHousehold	Customers spending on premiums in regards to *household*
PremHealth	Customers spending on premiums in regards to *health*
PremLife	Customers spending on premiums in regards to *life*
PremWork	Customers spending on premiums in regards to *work*

Table 1: Original Features

Case	***Approach***	***Dropped/imputed***
FirstPolYear, *BirthYear*, *EducDeg* and rows missing more than 30% of values	Dropped rows	58 removed
Premiums	Assumed missing values meant that no premium was paid, imputed with zero	215 imputed
Monthly salary	Imputed with linear regression using *BirthYear* as regressor	33 imputed
Children	Imputed using logistic regression with *BirthYear* as regressor	13 imputed

Table 2: Approaches used to handle missing values.

Feature	***Description***
Age	Customer age as of 2016
FirstPolAge	Customer age at the time of the first policy creation
CustYears	Number of years a customer has been with the company
Generation	Generation (Millennial, Gen X,...) a customer belongs to
YearSal	Yearly salary
PremTotal	Sum of all premiums paid by the customer
EffortRatio	Proportion of yearly salary spent on the company
PremRatios (5 features)	Proportion of PremTotal spent on each insurance type

Table 3: Features engineered from the ones originally provided.

2. Figures

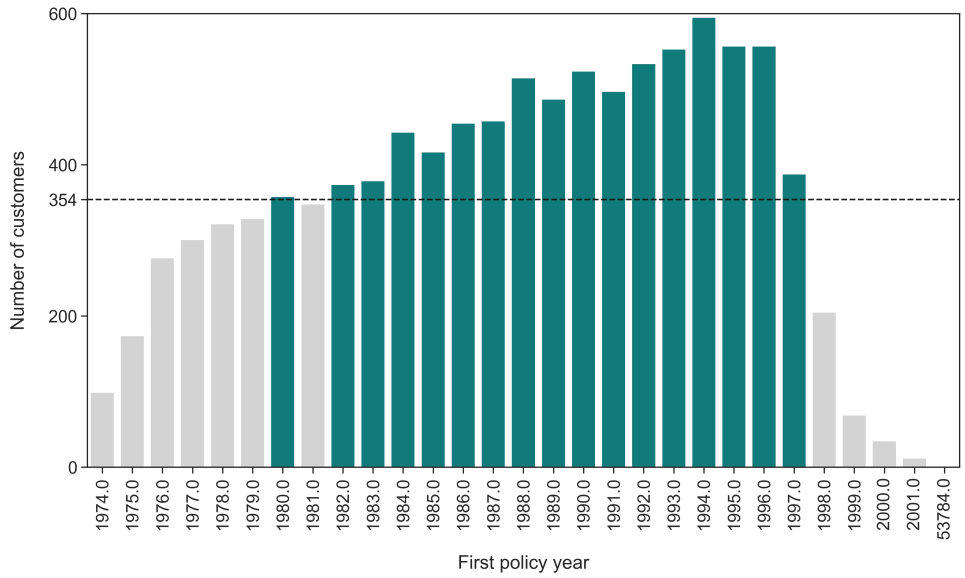

Figure 1: New customers acquired per year.

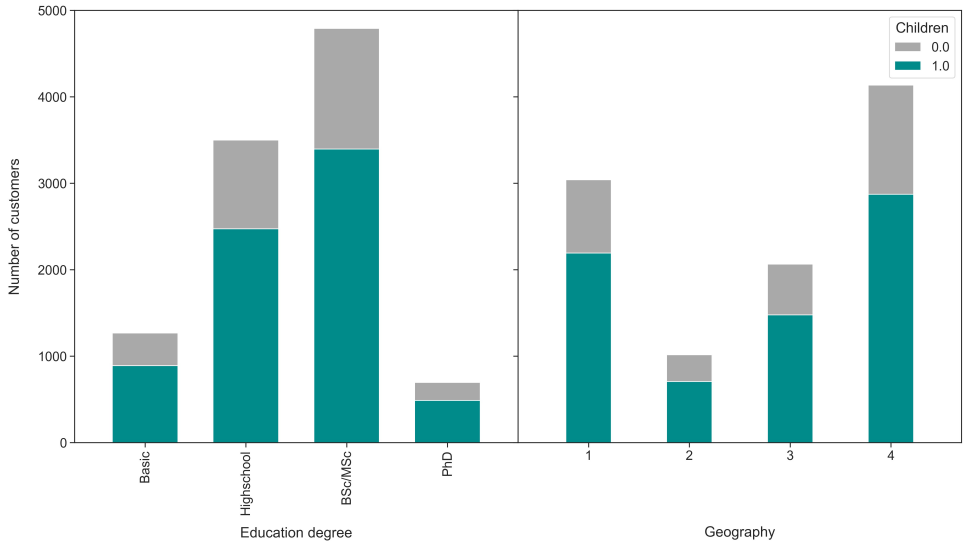

Figure 2: Initial exploration of customer sociodemographic data.

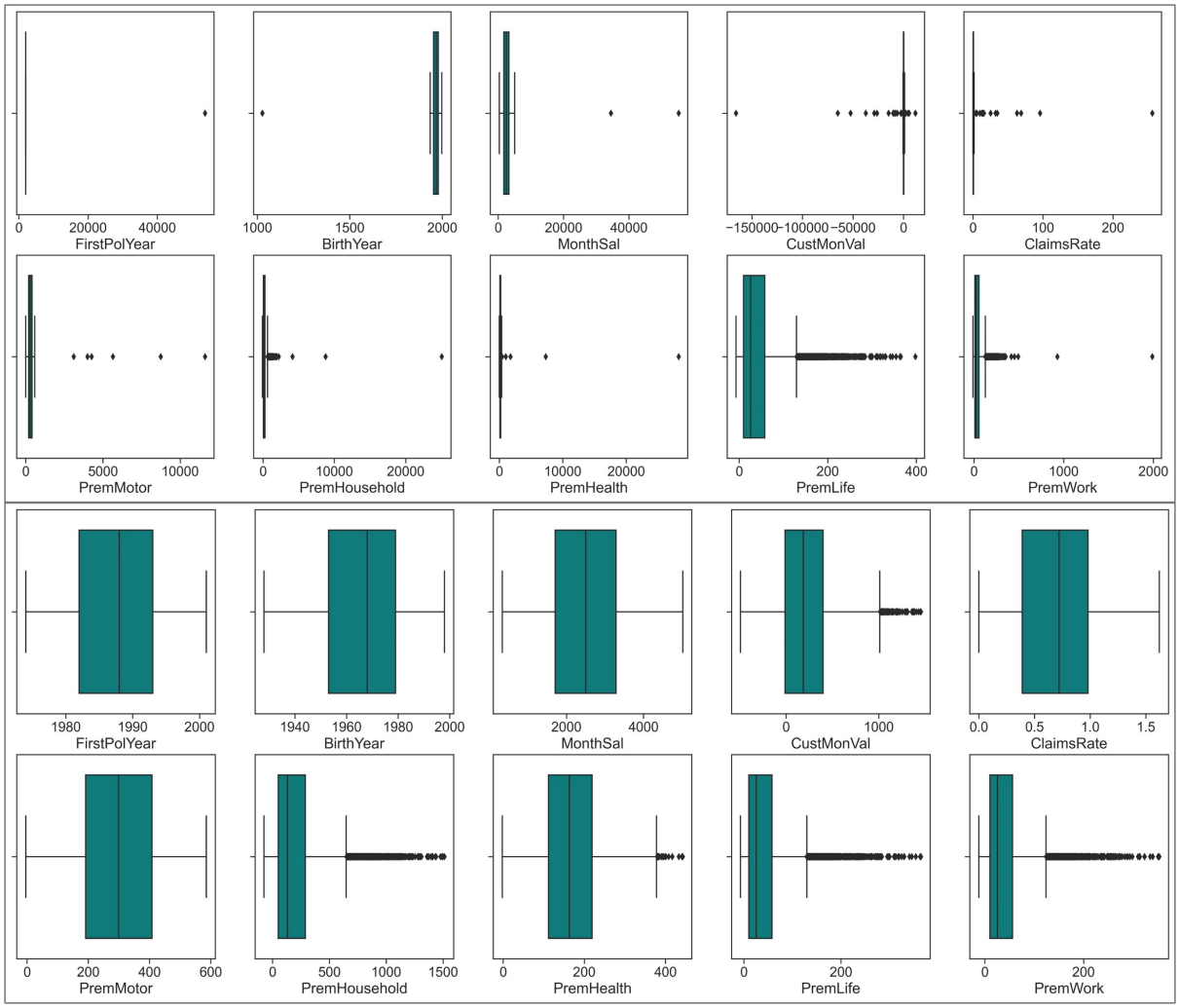

Figure 3: Box-plots of metric features before (top panel) and after (bottom panel) outlier removal.

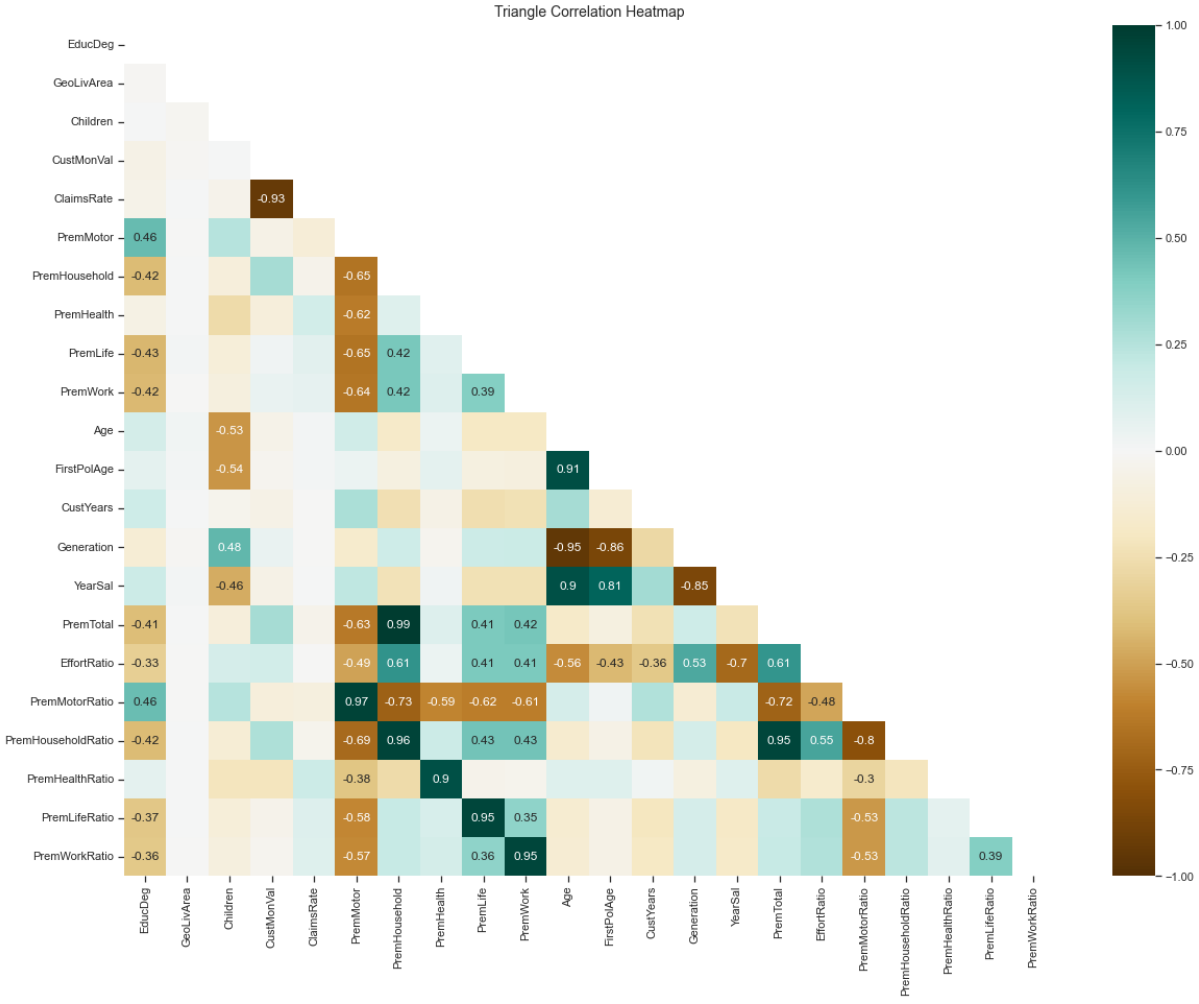

Figure 4: Correlation heatmap of Pearson correlations between features.

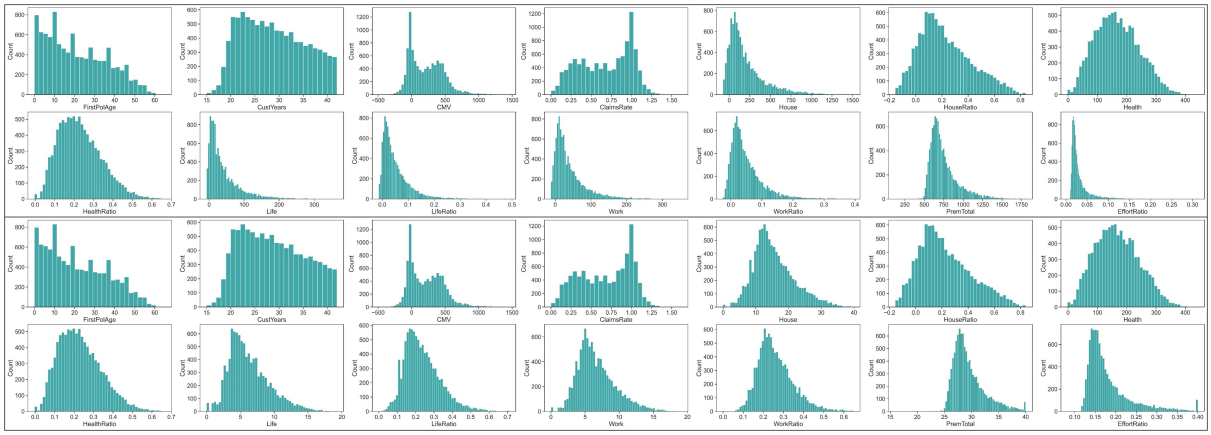

Figure 5: Histograms of metric features before (top panel) and after (bottom panel) square root transformation of skewed features and clipping of *PremTotal* and *EffortRatio*.

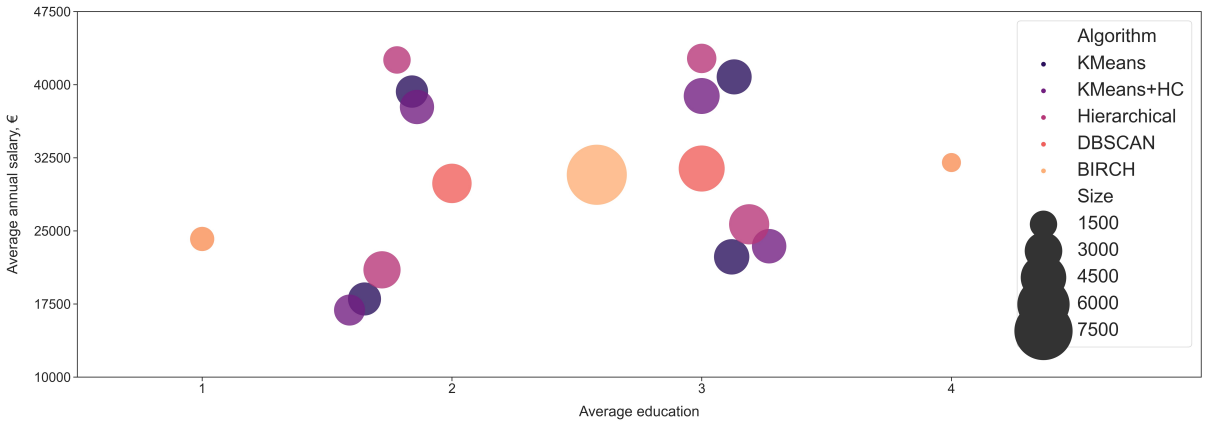

Figure 6: Sociodemographic clustering - comparison of clustering algorithms and cluster sizes.

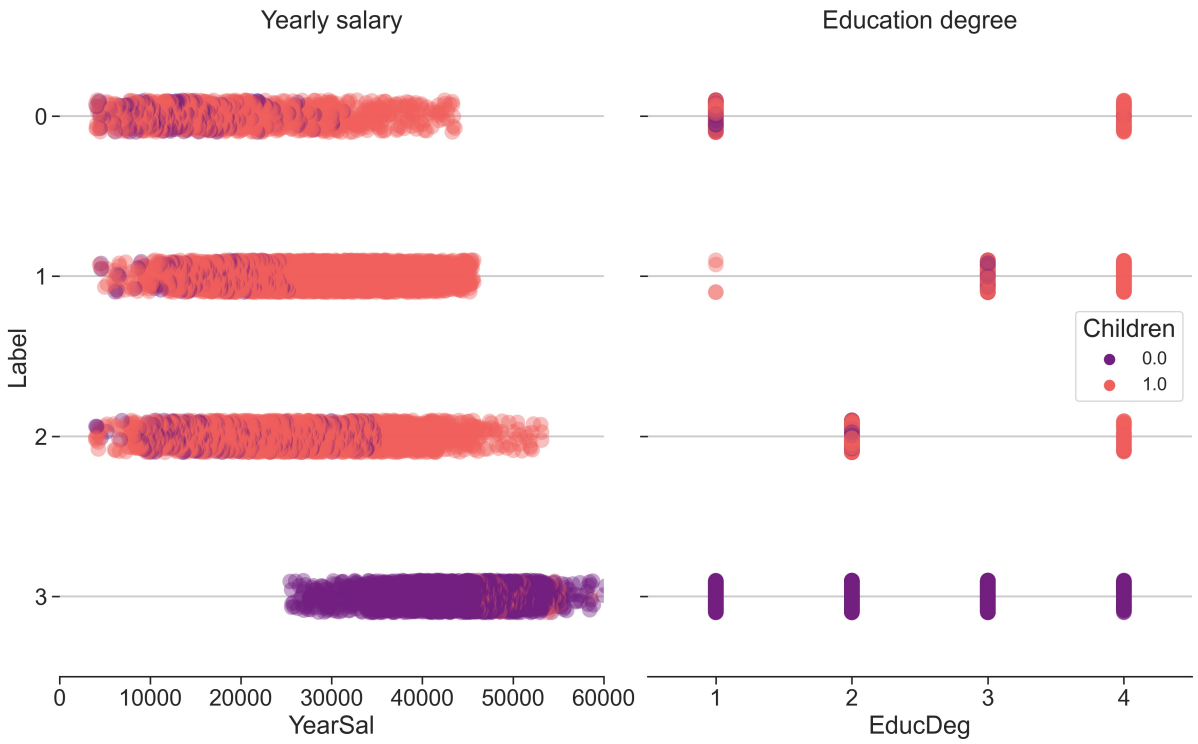

Figure 7: Characteristics of the final sociodemographic clusters obtained.

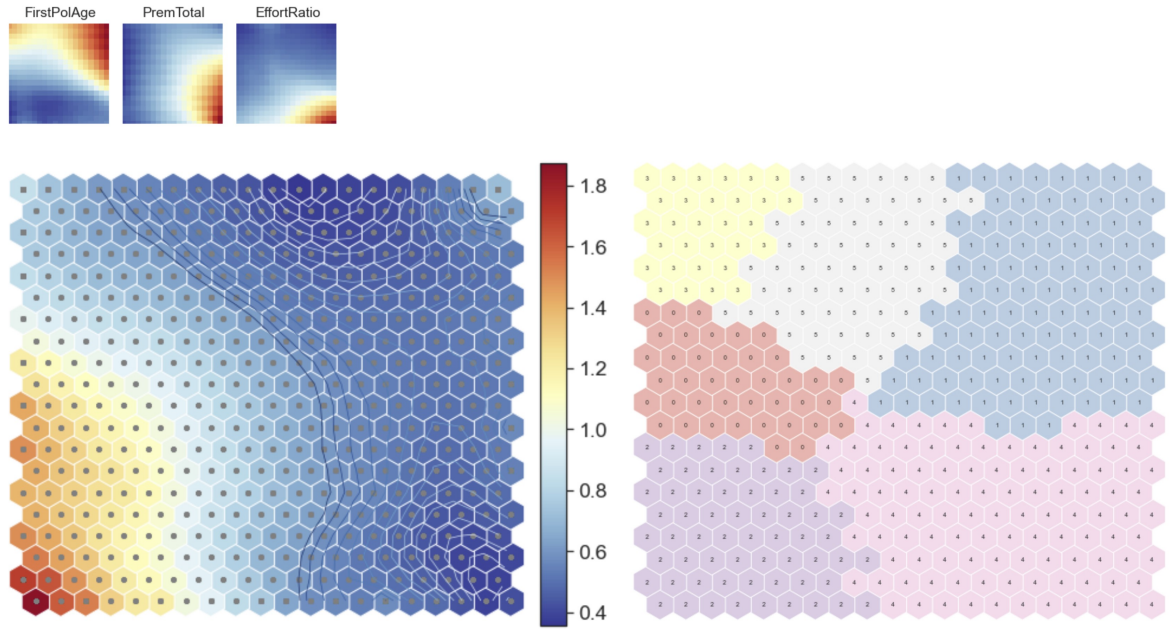

Figure 8: Value clustering with Self-organising maps (SOM) and K-means: features used (top left), U-matrix obtained (left), and clustering results (right).

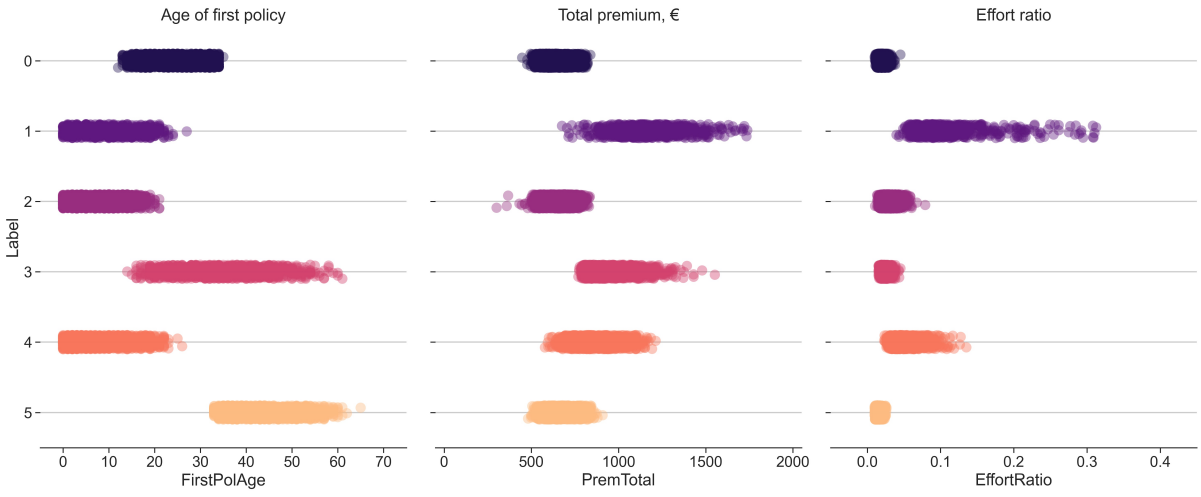

Figure 9: Characteristics of the final value clusters obtained.

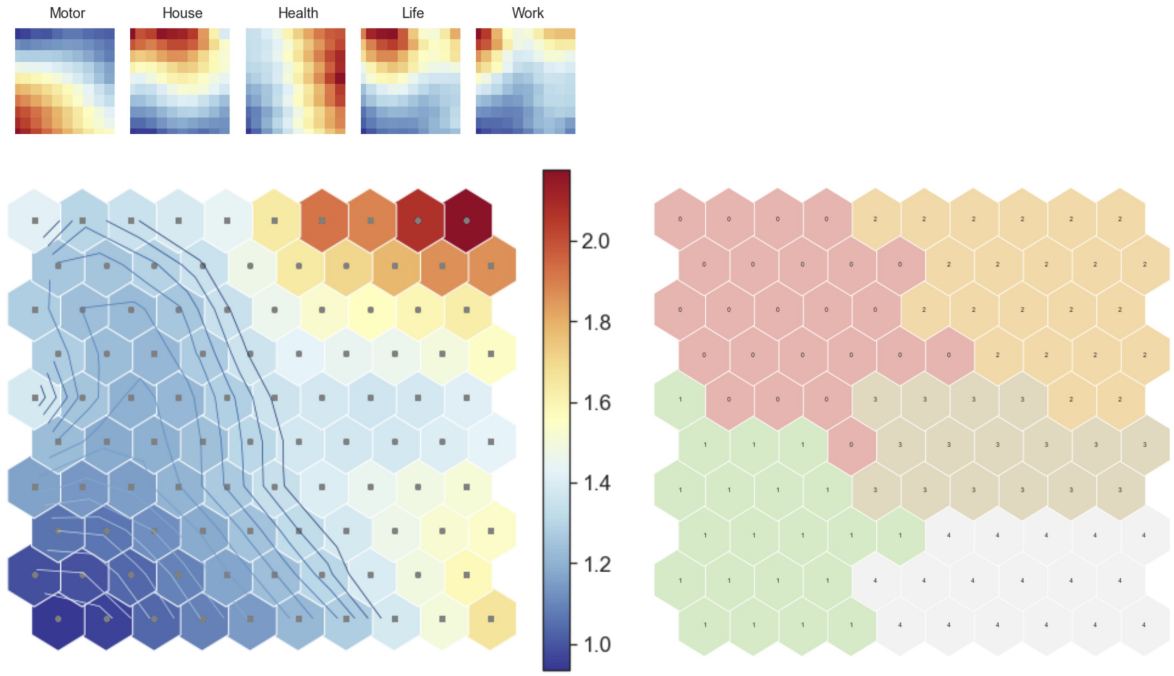

Figure 10: Product clustering with Self-organising maps (SOM) and K-means: features used (top left), U-matrix obtained (left), and clustering results (right).

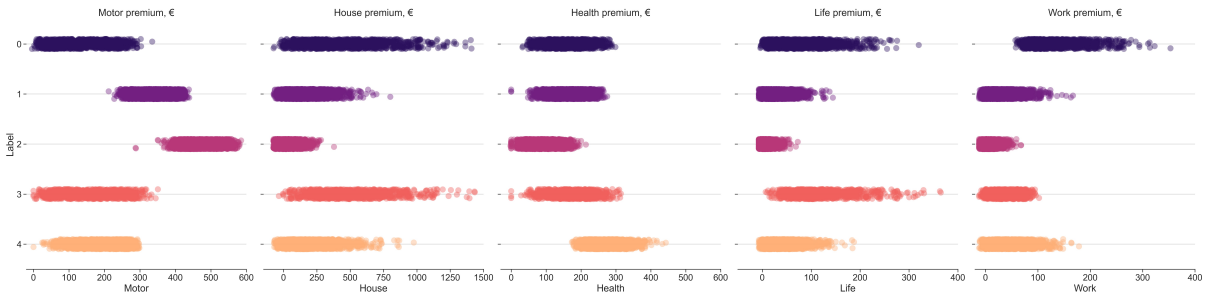

Figure 11: Characteristics of the final product clusters obtained.

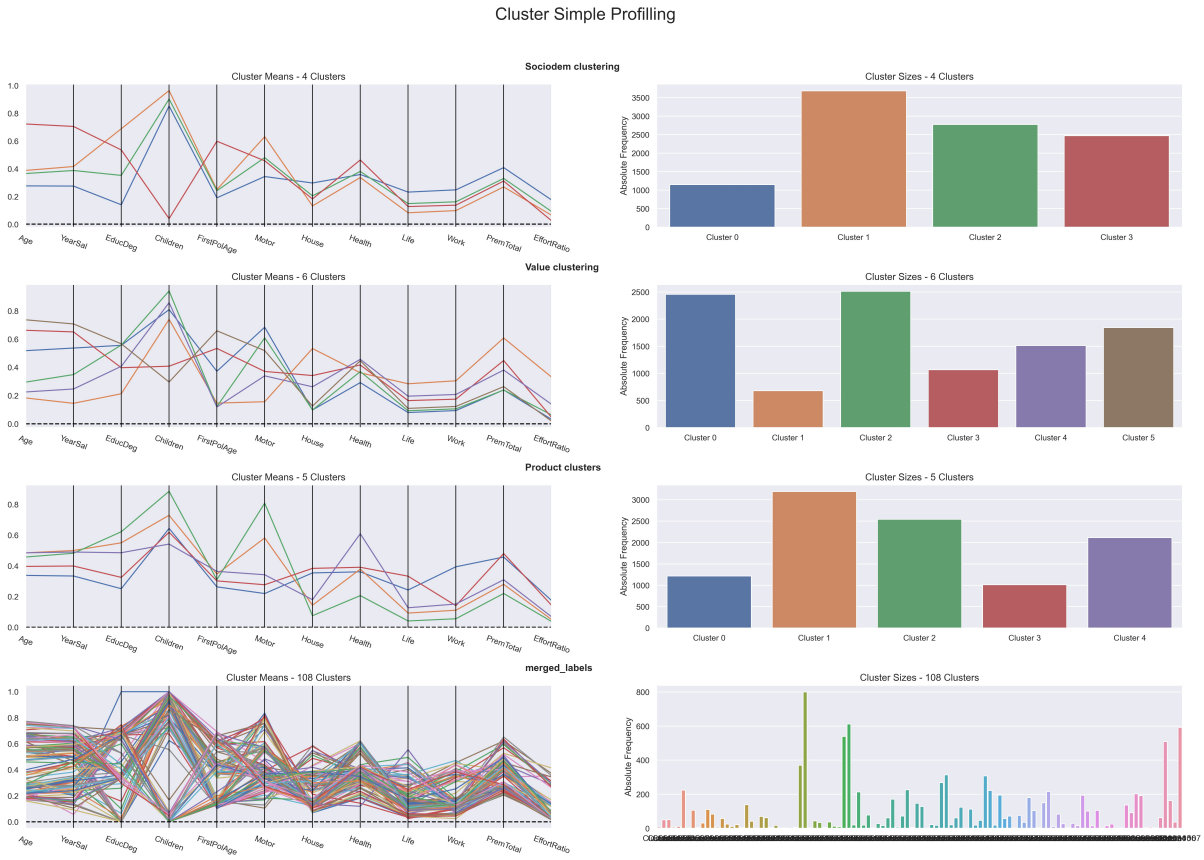

Figure 12: Summary of the clusters obtained by the sociodemographic, value and product approaches, as well as of the 108 clusters obtained by merging the three approaches together.

Cluster	Age	YearSal	EducDeg	Children	CMV	PremTotal	FirstPolAge	EffortRatio	Motor	House	Health	Life	Work
1	37.731228	20784.74834	1.727368	0.864912	235.87877	857.790712	11.006316	0.053479	205.092239	336.199386	179.697123	69.628519	67.173446
2	37.012124	21607.96879	3.10238	0.911989	211.50982	704.529358	7.424338	0.037104	334.374787	145.451684	169.541329	27.50304	27.658518
3	54.109844	33992.11936	2.790875	0.915927	215.1183	654.648576	24.874102	0.019847	425.681555	83.801267	109.663688	17.523118	17.978948
4	69.035388	43696.6903	2.519787	0.178082	201.48704	754.852907	40.628615	0.017559	256.989699	208.163756	206.85809	42.174722	40.66664

Figure 13: Detailed description of the final four clusters obtained by merging the results from the individual sociodemographic, value, and product clustering approaches.

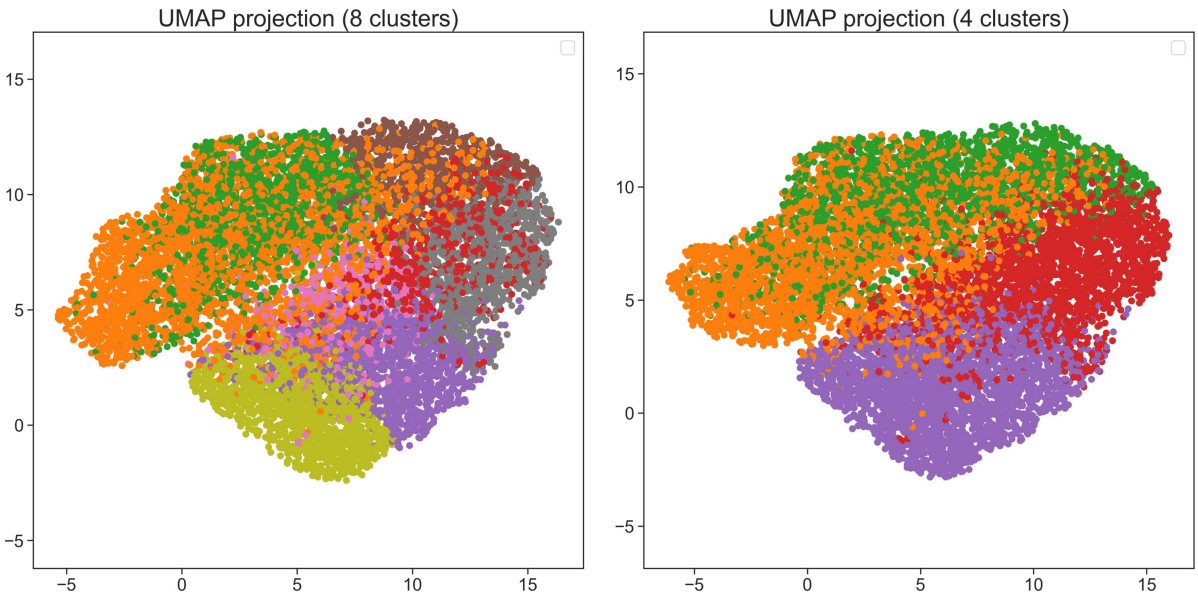

Figure 14: UMAP representation before and after manually merging clusters based on similarity.