Data Mining I - Project report

5th January 2018

Carolina Marques M2016126@novaims.unl.pt
Madalena Rodrigues M2016315@novaims.unl.pt
Maria Inês Silva D20170088@novaims.unl.pt

1. Problem and data description

For this project, we received data from a fictional insurer in Portugal, with reference to 2016. We were asked to develop a customer segmentation in such a way that it will be possible for the Marketing Department to better understand all the different customers' profiles.

To achieve this goal, we used techniques from the Descriptive Modelling field, also known as Unsupervised Learning. In this field, the aim is to describe and summarize large unlabeled data sets and it includes three main types of analysis - clustering, association rules and visualization. Since the Marketing Department wished to have a customer segmentation, the main focus of our project was the clustering analysis. With this type of analysis, individual customers were grouped together based on their similarity (the clusters) and each group's representatives (the centroids) were extracted and analysed. Using fictional representatives of groups of real customers does lead to a loss of detail, however, in return, we gained interpretability and understanding of the customers, which enabled more suited and efficient marketing decisions.

The data set received is comprised of 10296 customers and 13 features. In particular, for each customer, the following variables were available:

Variables	Description	SAS Encoding
ID	Customers' ID	Intervalar
FirstPolicy	Year of the customer's first policy	Intervalar
Birthday	Customer's birthday year	Intervalar
Education	Academic degree	Ordinal
Salary	Gross monthly salary in euros	Intervalar
Area	Living area	Nominal
Children	Binary variable (1 = has children)	Binary
CMV	Customer monetary value	Intervalar
Claims	Claims rate in last 2 years	Intervalar
Motor	Premiums in LOB Motor, in euros	Intervalar
Household	Premiums in LOB Household, in euros	Intervalar
Health	Premiums in LOB Health, in euros	Intervalar
Life	Premiums in LOB Life, in euros	Intervalar
Work	Premiums in LOB Work compensations, in euros	Intervalar

***Table 1:** Description of variables and SAS encoding*

Given the data, we decided to build two separate segmentations:

- Product segmentation - The product segmentation considered the columns related with the annual premiums per Line of Business (LOB) and aimed to group together customers with similar insurance policies.
- Value segmentation - The value segmentation considered the remaining features and aimed to characterize the customers in relation with their value to the company.

2. Data exploration

Before building the customer segmentation, we explored the raw data in order to check for missing values and other issues (such as outliers) that may undermine the construction of insightful segmentations.

2.1 Categorical features

We explored the categorical features, which represent labels or classes. The ordered labels, such as Education, are called ordinal, while unordered labels, such as Area and Children, are called nominal. Figure 1 illustrates the distribution of the 3 categorical features present in the raw data set. We concluded that the number of missing values is not significant (38 customers in total) and that there is no class with such a low number of observations that justifies ignoring it.

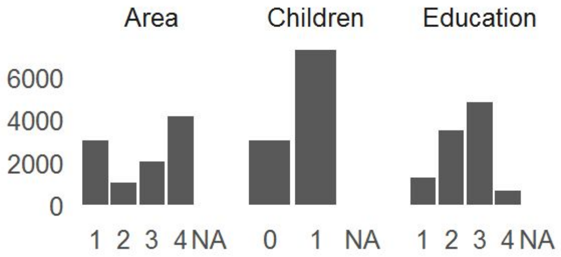


Figure 1: *Distribution of raw categorical features*

2.2 Numeric features

For the numerical features, we computed some summary statistics, which are presented in tables 2 and 3. We concluded that missing values are not significant and that almost all features have strange values that could be outliers.

Variable	Mean	Standard Deviation	Median	% of Missing Values	Minimum	Maximum
Birthday	1968	20	1968	0.2%	1028	2001
Claims	0.7	2.9	0.7	0.0%	0	256.2
CMV	177.9	1945.8	186.9	0.0%	-165680.4	11875.9
FirstPolicy	1991	511	1986	0.3%	1974	53784
Salary	€2,506.70	€1,157.40	€2,501.50	0.3%	€333.00	€55,215.00

Table 2: *Summary statistics of numerical features*

Variable	Mean	Standard Deviation	Median	% of Missing Values	Minimum	Maximum
Health	€171.60	€296.40	€162.80	0.4%	-€2.10	€28,272.00
Household	€210.40	€352.60	€132.80	0.0%	-€75.00	€25,048.80
Life	€41.90	€47.50	€25.60	1.0%	-€7.00	€398.30
Motor	€300.50	€211.90	€298.60	0.3%	-€4.10	€11,604.40
Work	€41.30	€51.50	€25.70	0.8%	-€12.00	€1,988.70

Table 3: *Summary statistics of numerical features*

The following plots further illustrates the conclusions presented before.

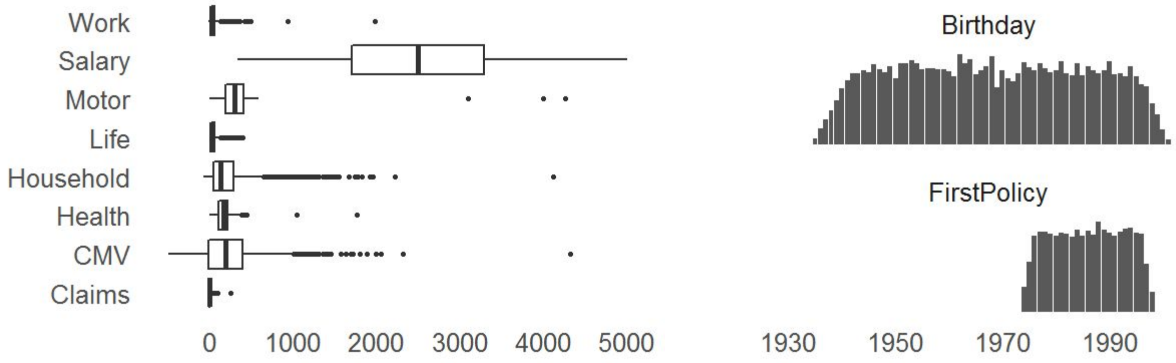


Figure 2: *Distribution of raw numerical features*
(box-plots for interval features and bar-plots for discrete features)

3. Data preparation

Based on the exploration performed on the raw data, we identified some transformations that needed to be done in the data preparation phase. This phase is essential in any Data Mining task as it increases the quality and signal of the data. Since all clustering algorithms are data driven, the quality of the resulting clusters are heavily dependent on the quality of the data used to build them.

3.1 Feature transformation

a) Treatment of missing values

Missing values indicate the inexistence of information stored in the data set. There are two main strategies to deal with missing values - remove observations/features or input new values. In this particular case, categorical features were filled with the most frequent class and numerical features were filled with the median. The idea here is to avoid excluding customers and/or important features. For this purpose, we used the median to control to the outliers' bias.

b) Treatment of reservals

In this data set, negative premiums represent reversals occurred the 2016 and paid in 2015. In other words, they represent customers that cancelled their insurance policy in that specific LOB before the end of policy and that are no longer customers in that LOB. Since in the product segmentation we were only interested in knowing how much the customers spend on each LOB, we changed all negative premiums to zero. We also created a new ordinal variable to indicate how many reversals each customer experienced in 2016 as we thought it might be an interesting feature for the value segmentation.

c) Treatment of dates

The features FirstPolicy and Birthday correspond to calendar years and thus do not belong to the interval variable type[1]. To transform them into an interval type and to help interpret the clusters, we substituted them by the new features Fidelity and Age, respectively. (Fidelity = 2016 - FirstPolicy & Age = 2016 - Birthday).

d) Treatment of outliers

Another necessary step was the extraction of outliers. They consist of unusual or atypical observations which are too far from the rest of the observations or are inconsistent with the process that generates the data. As such, the common practice is to isolate outliers in the training phase (i.e., the computation of the clusters) in order to avoid their biased effect on the clusters and to include them later in the analysis by attributing a relevant cluster to them.

For the identification of numerical features' outliers we applied the Gaussian assumption, i.e., any observations lying outside the interval defined by the mean +/- 3 times the standard deviation were considered outliers. These thresholds can be checked in annex 6.1. Note, however, that the feature CMV was treated differently. After applying the Gaussian assumption and plotting its empirical density function, we verified that this feature still had some outliers and thus we decided to apply a fixed threshold of +/-15000. Finally, since we didn't consider the Claims in the training phase (more details on this decision will follow), we didn't filter the outliers related to it.

In the end, we isolated 425 customers due to outliers, which corresponds to 4.1% of the initial data set. We recognize that not all features have a Gaussian distribution, however, for simplicity, we used this criteria.

After applying these 4 transformations, we plotted the new features' distributions (empirical densities for numerical features and bar-plots for categorical features), which can be found in annex 6.2.

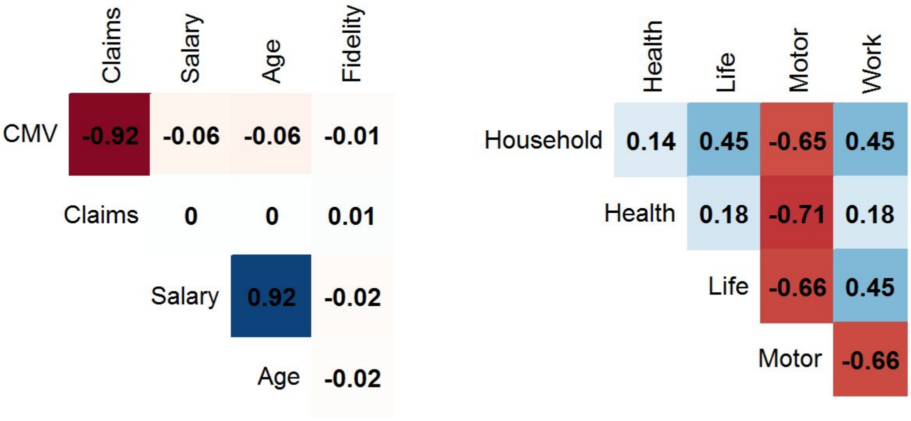
3.2 Feature selection

Firstly, it's important to note that since the usual clustering algorithms rely on distances, they don't work well with categorical features. Therefore, the features Education, Area, Children and Reversals were excluded from the training phase and were only used in the profiling phase.

The second point to note is the redundancy. In other words, we should only consider features in the training phase that bring new information, which implies that we should avoid including features

[1] Note that in SAS Miner we encoded them as interval to simplify the features' transformations, however, this does not influence the analysis.

with high correlation (either positive or negative). Figure 4 presents the correlation matrices of numerical features used in the value and product segmentations, separately. In the value segmentation, the pairs of features Salary-Age and CMV-Claims show strong correlations and thus for the training phase we excluded the Age and the Claims. On the other hand, there are no strong correlations between the features used on the product segmentation and thus we kept them all for the training phase.



	Claims	Salary	Age	Fidelity
CMV	-0.92	-0.06	-0.06	-0.01
Claims		0	0	0.01
Salary			0.92	-0.02
Age				-0.02

	Health	Life	Motor	Work
Household	0.14	0.45	-0.65	0.45
Health		0.18	-0.71	0.18
Life			-0.66	0.45
Motor				-0.66

Figure 4: *Correlation matrices for both segmentations, value on the left and product on the right*

4. Clustering

As mentioned in the first section, clustering algorithms reduce the complexity and increase the understandability of large data sets by separating the data into a small number of clusters based on their similarity. Since observations belonging to the same cluster are considered similar, we can reduce the size and complexity of the initial data set by considering the representatives of each clusters, the centroids.

For this project we chose a widely used clustering algorithm, the k-means. Its main advantages are its simplicity (k-means is a very simple algorithm that can be intuitively understood) and its efficiency (if we consider a reasonable number of clusters and iterations, it is a linear-time algorithm). On the other hand, the algorithm has 4 main disadvantages:

- It is very sensitive to outliers. To mitigate this issue, as shown in the previous section, we isolated all outliers from the training phase.
- It is also sensitive to the initial seeds and may terminate in a local optimum. To mitigate this issue, we used the SAS Miner initiation called Princomp. In this option, the initial seeds are set in the plane defined by the first two principal components, in an evenly spaced manner. The idea is to disperse the initial seeds in the two directions of maximum variance of the initial data set.
- The user needs to set from the start the desired number of clusters, k, which can be subjective. In SAS Miner, we run the algorithm for a variety of k's and saved the resulting within-cluster standard deviation to build an elbow-plot and decide on the optimal number of clusters.
- The k-means does not work well with non-spherical clusters. Plotting a pairwise scatter plot over all features used in each segmentation did not reveal any non-spherical shape among the data.

4.1 Analysis of the optimal number of clusters

Figure 5 shows the elbow plots for both segmentations separately. It includes the within-cluster standard deviation for clustering solutions with 2 to 9 clusters.

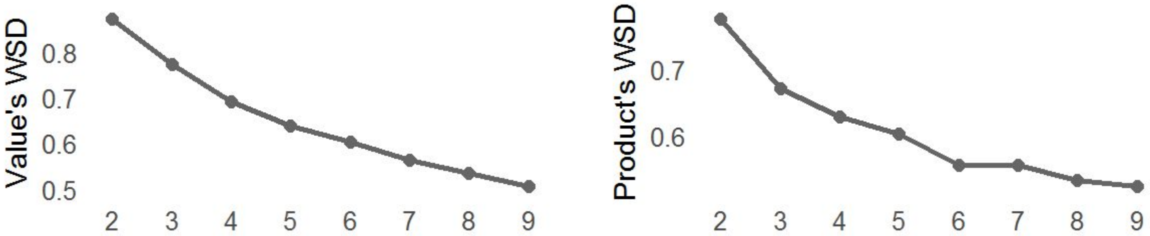


***Figure 5:** Elbow plots for both segmentations*

Based on the elbow criteria, we decided to further investigate 3 different clustering solutions for each segmentation by analysing the following plots for each solution (which are presented in annex 6.3 and 6.4):

- A pie chart with the distribution of customers over all the clusters.
- The input mean plot, which shows the 0-1 standardized centroids.
- Distributions of all relevant features (empirical density for numerical features and frequency bar-plots for categorical features) for each cluster and excluding the outliers.

For the value segmentation, we analysed solutions with 3, 4 and 5 clusters. We ended up choosing the solution with 4 clusters since the solution with 3 clusters does not differentiate the feature Fidelity and the solution with 5 clusters does not bring enough new differentiation when compared with the solution with 4 clusters.

For the product segmentation, we looked at the solutions with 2, 3 and 4 clusters. In the end, we chose the solution with 2 clusters because it gives a good enough differentiation for the features Health and Motor and increasing the number of clusters does not increase the differentiation for the other features.

It is also important to note that, for both segmentations, the pie charts show an homogenous distribution of customers among all clusters and thus all clusters have a meaningful expression in the data set.

5. Profiling and marketing strategies

The profiling phase was composed of two main steps. Firstly, we profiled the two segmentations separately. Then, we consolidated the segmentations to have a global view of the customers and decide on the most appropriate marketing strategy, both in terms of the value to the company and the type of products used.

5.1 Value segmentation

Figure 6 contains the input mean plots for the value segmentation, which allowed as to define and profile its 4 clusters. Figure 7 further illustrates the properties of each cluster properties by presenting their distribution over all variables relevant for the value segmentation.

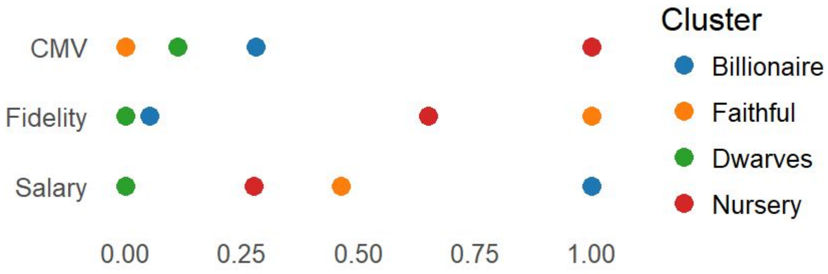


Figure 6: *Input mean plot for the value segmentation*

The first cluster is composed of the so-called **Billionaires**, which are distinguished from the other clusters by showing an above average salary. However, the high level of income does not lead to higher levels of monetary value. They are older individuals, with a low probability of having children and slightly higher level of education than other clusters, who are very recent customers.

The second cluster contains the **Faithfuls**. As the name implies, it's the cluster with the most loyal customers. They also have the lowest monetary value, which is not ideal for the insurer.

The next cluster is the **Dwarves** since they fall short on all variables expressing value as clients. They are young adults with a high probability of having children and lower levels of education.

The fourth cluster was designated as **Nursery**. Since these customers have the highest level of monetary value, they should be given more attention to than other customers. Unfortunately, they show high levels of reversals, which means the insurer may be losing them as customers.

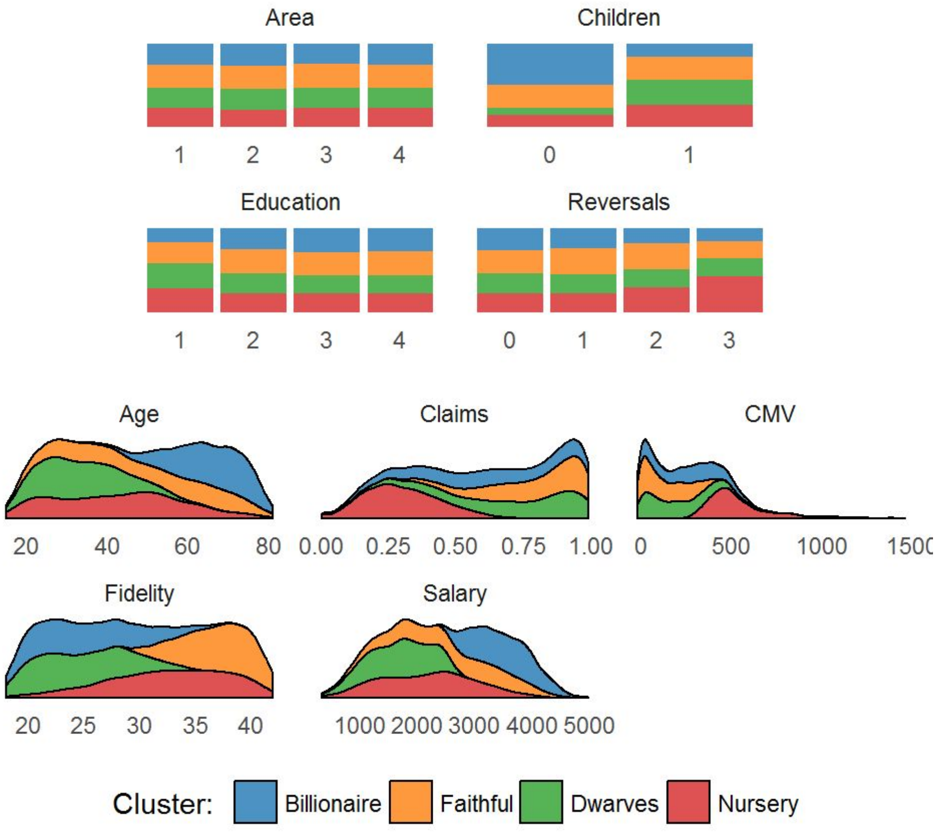


Figure 7: *Distributions of clusters over all variables concerning the value segmentation*

5.2 Product segmentation

Figures 8 and 9 contain the input mean plot and the empirical density functions related with the product segmentation, which were used to profile the 2 product clusters.

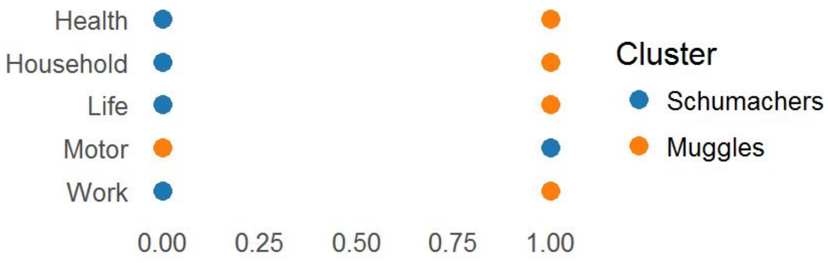


***Figure 8:** Input mean plot for the product segmentation*

We named the first cluster of **Schumachers**. They are characterized by high premiums in car insurance and low to zero premiums in the remaining products. This means that they are individuals with expensive cars and/or higher covers than the mandatory third party liability, who either spend below average on the remaining products or bought those products with competitors.

The second and final cluster, the **Muggles**, represent the average people as they don't show any special spending on any product. However, in comparison with the Schumachers, they have lower premiums on their car insurance cover and slightly higher premiums on their health insurance cover.

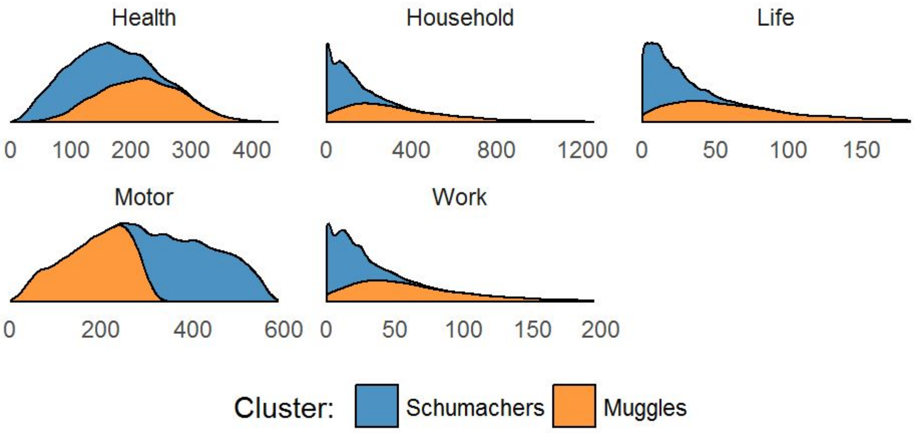


***Figure 9:** Distributions of clusters over all variables concerning the product segmentation*

5.3 Consolidated segmentation

When we consolidated the two segmentations, we got 8 clusters. Their centroids and frequency are presented in table 4. Although customers are almost uniformly distributed among all clusters, we considered that 8 is too large a number for creating different marketing strategies. Therefore, we removed the 2 clusters with the lowest number of customers by joining them in the clusters that shares the same cluster of the value segmentation. In particular, Nursery Muggles is integrated into Nursery Schumachers to form the cluster Nursery and Dwarves Schumachers is integrated into Dwarves Muggles to form the cluster Dwarves.

Cluster		Frequency	Centroids							
Value	Product		CMV	Salary	Fidelity	Household	Health	Life	Motor	Work
Nursery	Muggles	9.9%	540	2182	33	330	215	62	185	62
Dwarves	Schumachers	11.0%	114	1675	25	88	132	17	406	17
Billionaire	Muggles	11.9%	195	3513	25	330	215	62	185	62
Faithful	Muggles	12.7%	61	2526	37	330	215	62	185	62
Dwarves	Muggles	12.9%	114	1675	25	330	215	62	185	62
Nursery	Schumachers	12.9%	540	2182	33	88	132	17	406	17
Billionaire	Schumachers	13.8%	195	3513	25	88	132	17	406	17
Faithful	Schumachers	14.9%	61	2526	37	88	132	17	406	17

***Table 4:** Centroids and frequency of initial consolidated segmentation*

Figure 10 illustrates the distribution of all customers over the final 6 consolidated clusters. As expected, all clusters have a relevant proportion of customers.

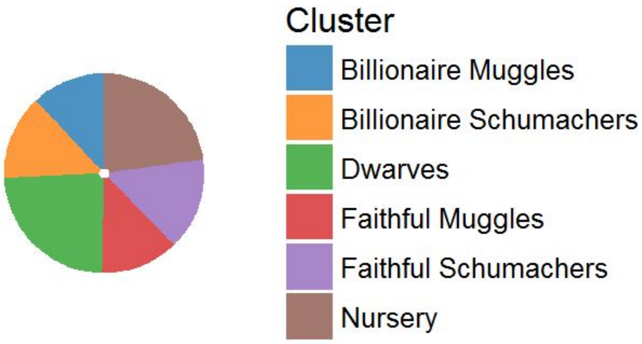


***Figure 10:** Pie chart of the distribution of customers over the final consolidated clusters*

In addition, we plotted the distribution of all relevant variables, which can be found in annex 6.5, to check if the conclusions obtained from the profiling of each separate segmentation changed with the consolidation. Tables 5 and 6 summarize the main characteristics of each cluster and the possible marketing approaches for those customers.

Cluster	Main characteristics	Strategies
Billionaire Muggles	- High salary - Recent customers - Average monetary value - Older individuals - Low probability of having children - Almost no reversals - Low spending on car insurance - Average spending on other products	These are good new customers with some value to the insurer. The goal is to maintain/improve their satisfaction level and to make them future loyal clients. Since they are older individuals with high incomes, they demand a quality service. Thus, the strategy for this cluster is to provide the best service and to offer special perks tailored for them.

***Table 5:** Final clusters properties and marketing strategies*

Cluster	Main characteristics	Strategies
Billionaire Schumachers	- High salary - Recent customers - Average monetary value - Older individuals - Low probability of having children - Slightly higher education levels - High spending on car insurance - Low to zero spending on other products	These are new high potential customers. They are recent customers with high salaries that spend mainly on car insurance. Thus, they can be targeted for selling other products in a package with their car insurance policy and, at the same time, make an effort to provide a good services and perks in order to retain them.
Dwarves	- Low salary - Recent customers - Low to average monetary value - Young adults - High probability of having children	The Dwarves are recent customers with children and low incomes and thus low prices should be an important factor to them. In addition, their monetary value is not as high as desirable. Therefore, the strategy here is twofold: - Reduce as much as possible the operational costs by simplifying the services provided in order to offer cheaper policies - Offer new deals that extend to their families to attract them to our insurance company in the future, such life and health covers for the rest of the family
Faithfull Muggles	- Older customers (more loyal) - Low monetary value - Slightly lower education levels - Almost no reversals - Low spending on car insurance - Average spending on other products	Although these customers have a low monetary value, their claim rate is not higher than the average. This means that the reason for their low value lies in their operational costs. Also, they are very loyal customers and thus we should strive to make them profitable for the insurer. The strategy here is to reduce the operational costs by simplifying the services provided and retain them.
Faithfull Schumachers	- Older customers (more loyal) - Very low monetary value and high claims' rate - Slightly higher education levels - High spending on car insurance - Low to zero spending on other products	These are the worst customers. High claim rates and low monetary value means that the pricing of their policies is not adequate. Thus, pricing rules needs to be revised, even if it means losing some customers.
Nursery	- High monetary value and low claim's rate - Slightly lower investment in health insurance - Expressive number of reversals	These are the best customers. The strategy for this group should be a quality service and tailored perks. We want to increase their satisfaction and make them life loyal.

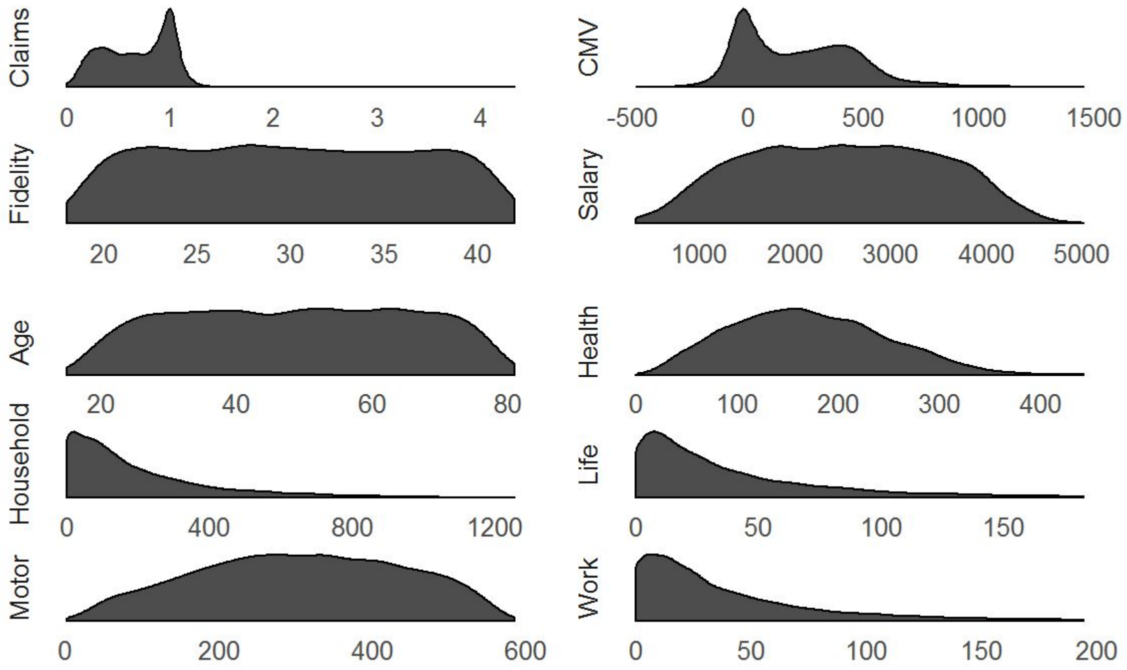
Table 6: *Final clusters properties and marketing strategies*

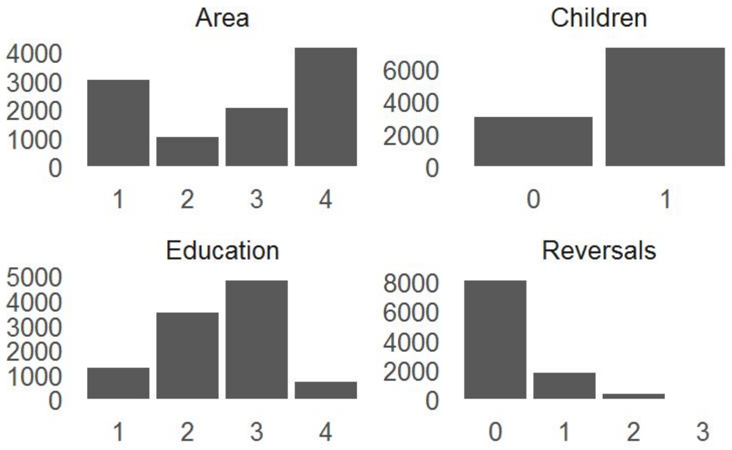
6. Annex

6.1 Thresholds for identifying outliers in numerical features

Threshold	Birthday	Claims	CMV	FirstPolicy	Salary	Health	Household	Life	Motor	Work
Left	1909	-8	-5660	457	-€966	-€718	-€847	-€101	-€335	-€113
Right	2027	9	6015	3525	€5,979	€1,061	€1,268	€184	€936	€196

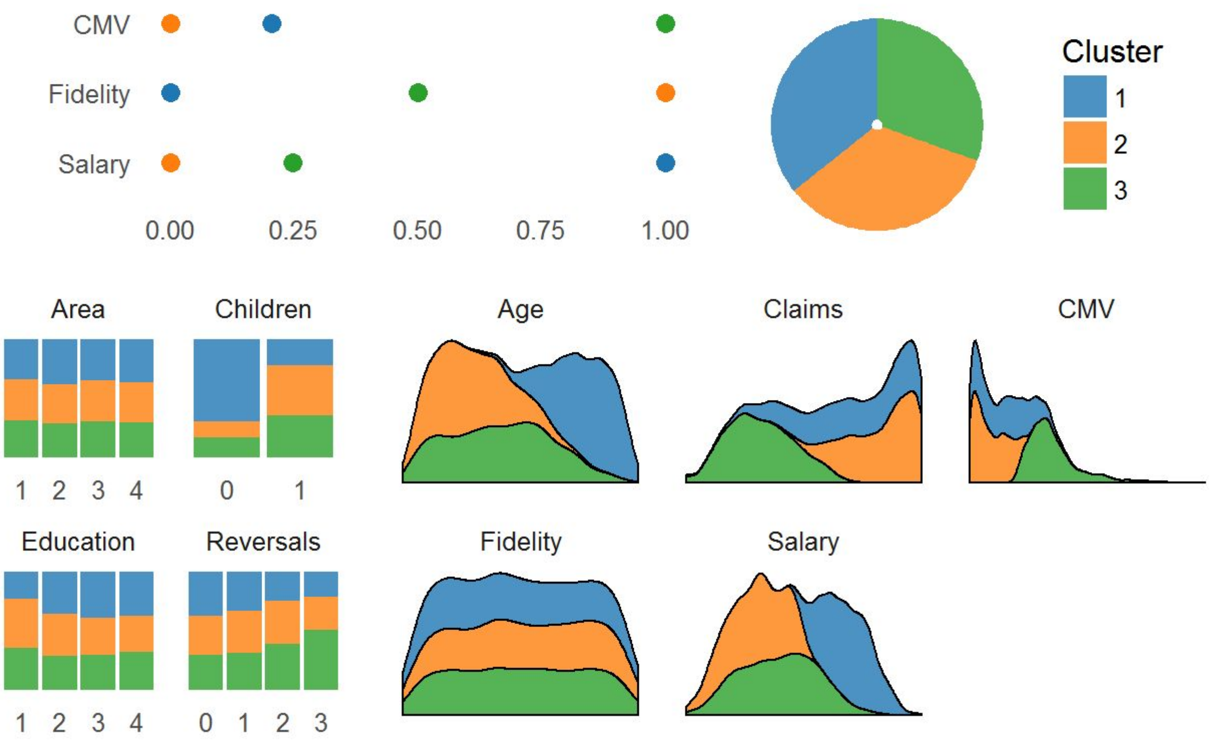
6.2 Distribution of final “cleaned” features



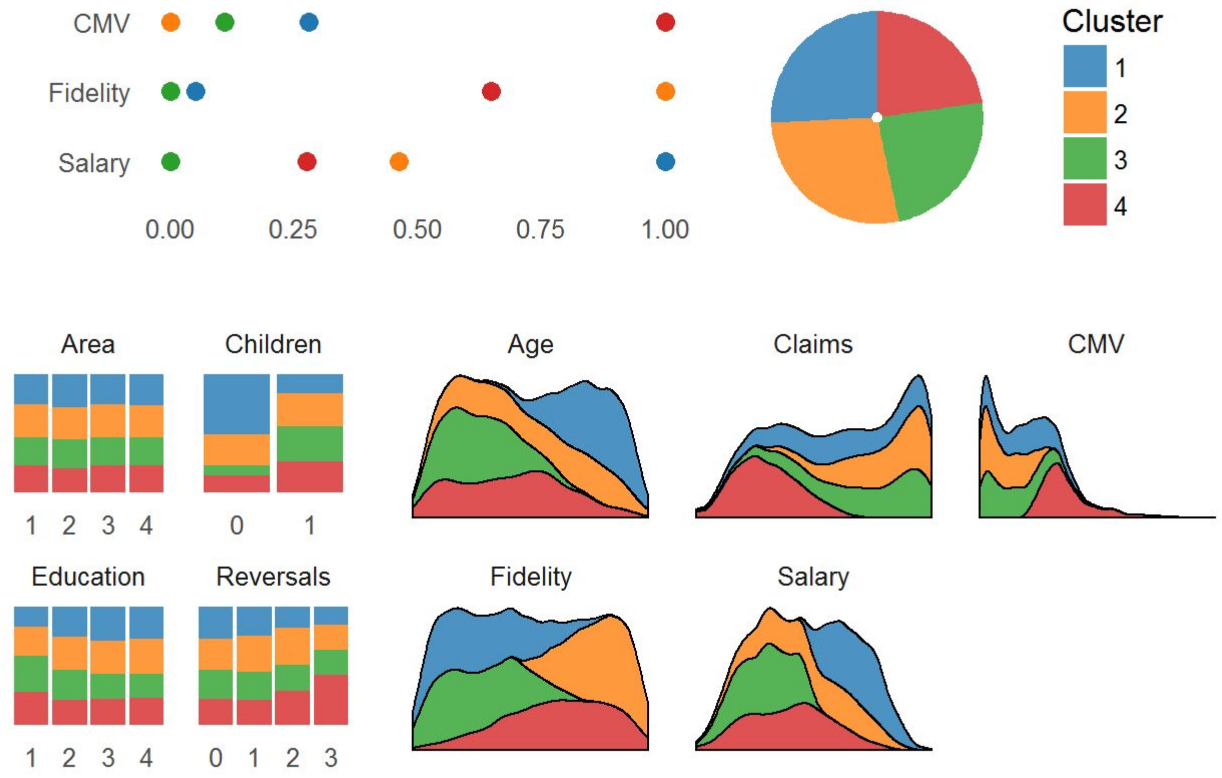


6.3 Results of the analysis on 3 different solutions for the value segmentation

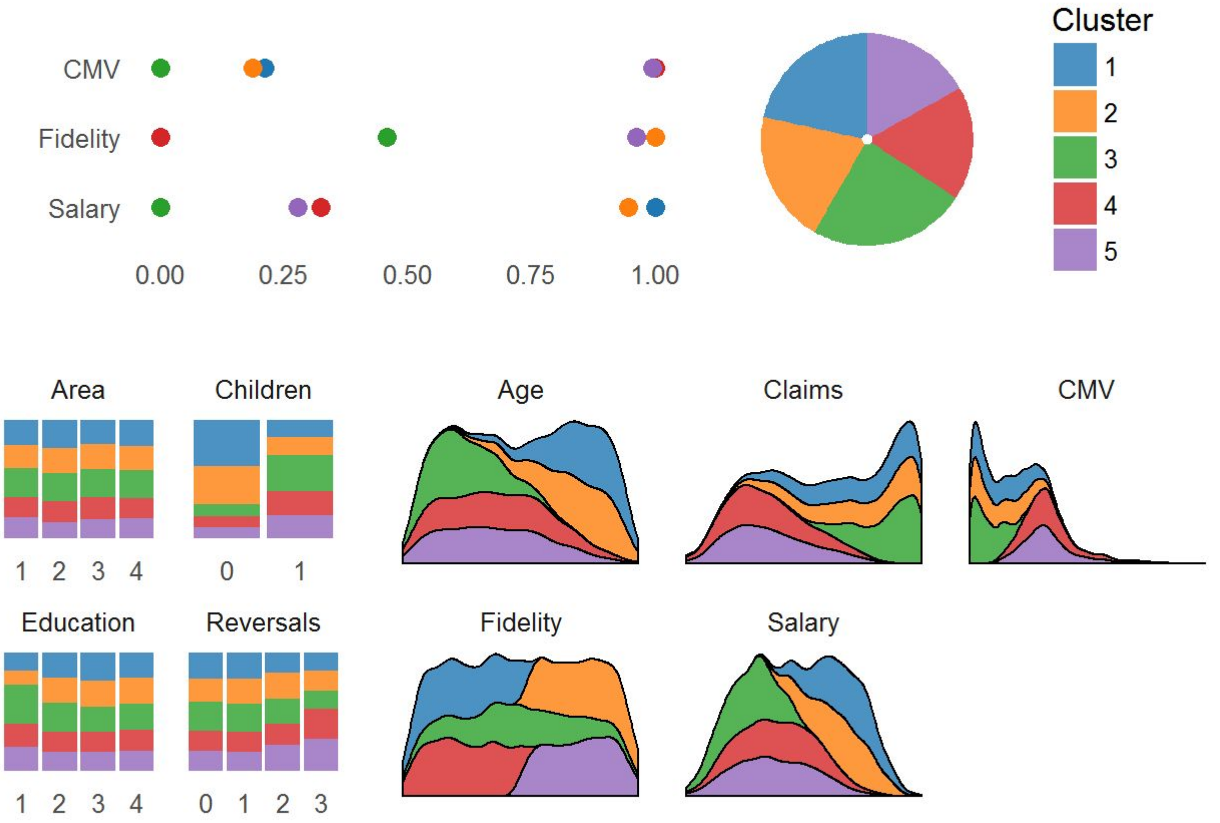
Solution with 3 clusters



Solution with 4 clusters

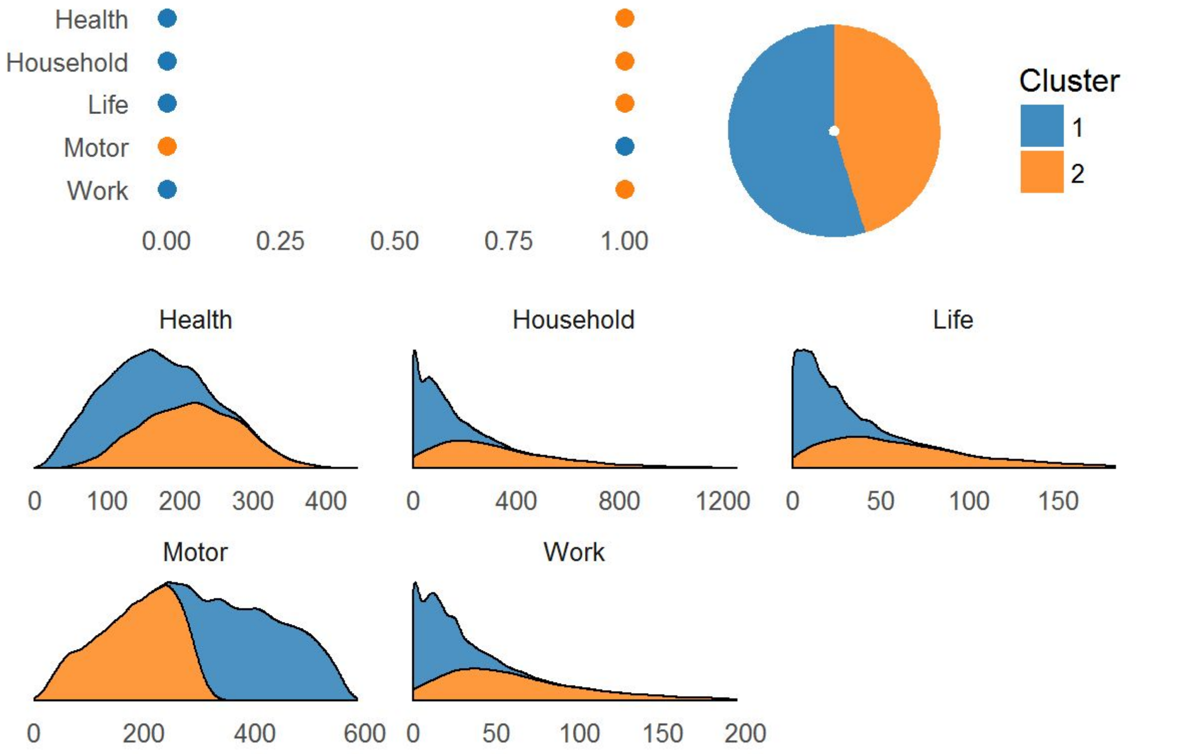


Solution with 5 clusters

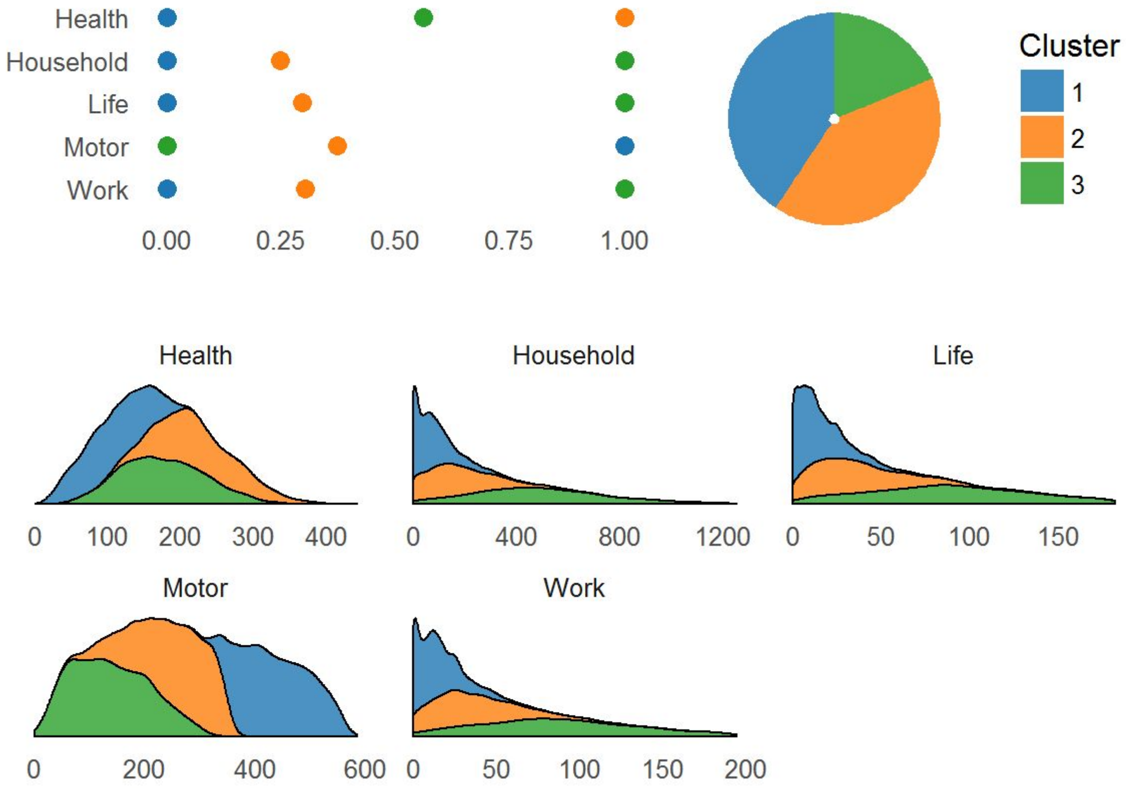


6.4 Results of the analysis on different solutions for the product segmentation

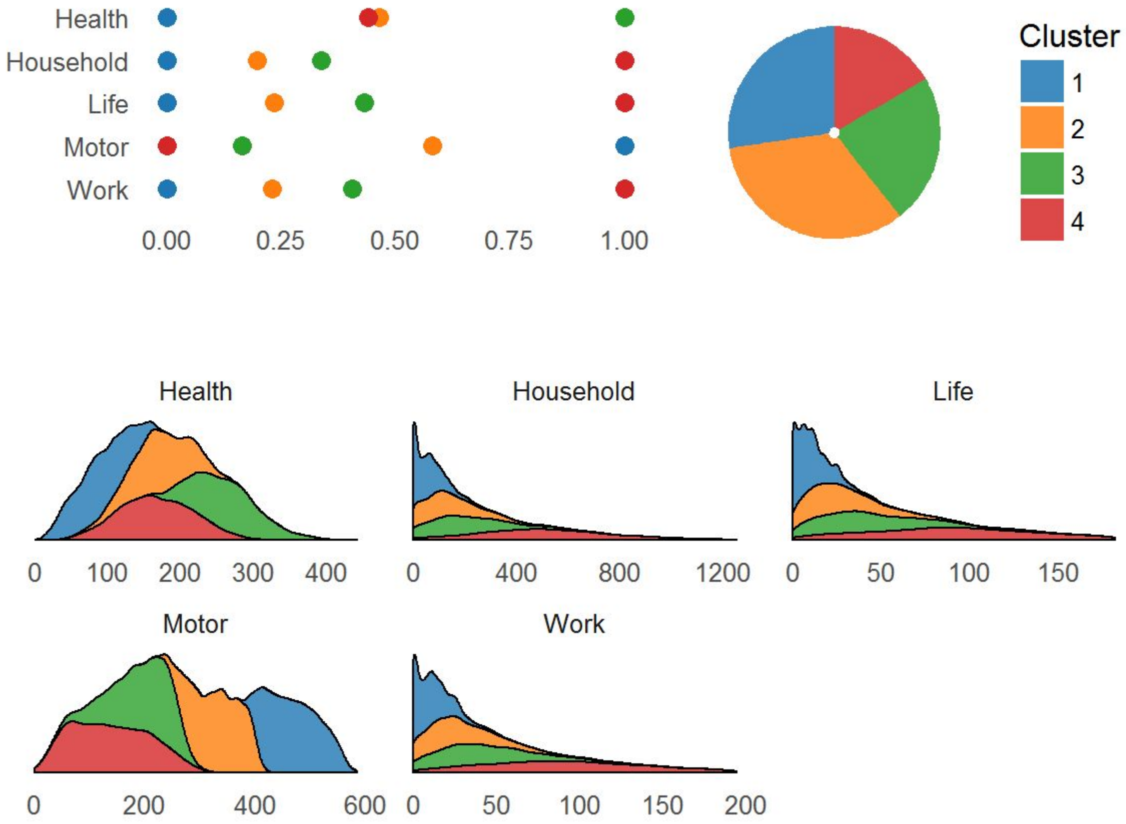
Solution with 2 clusters



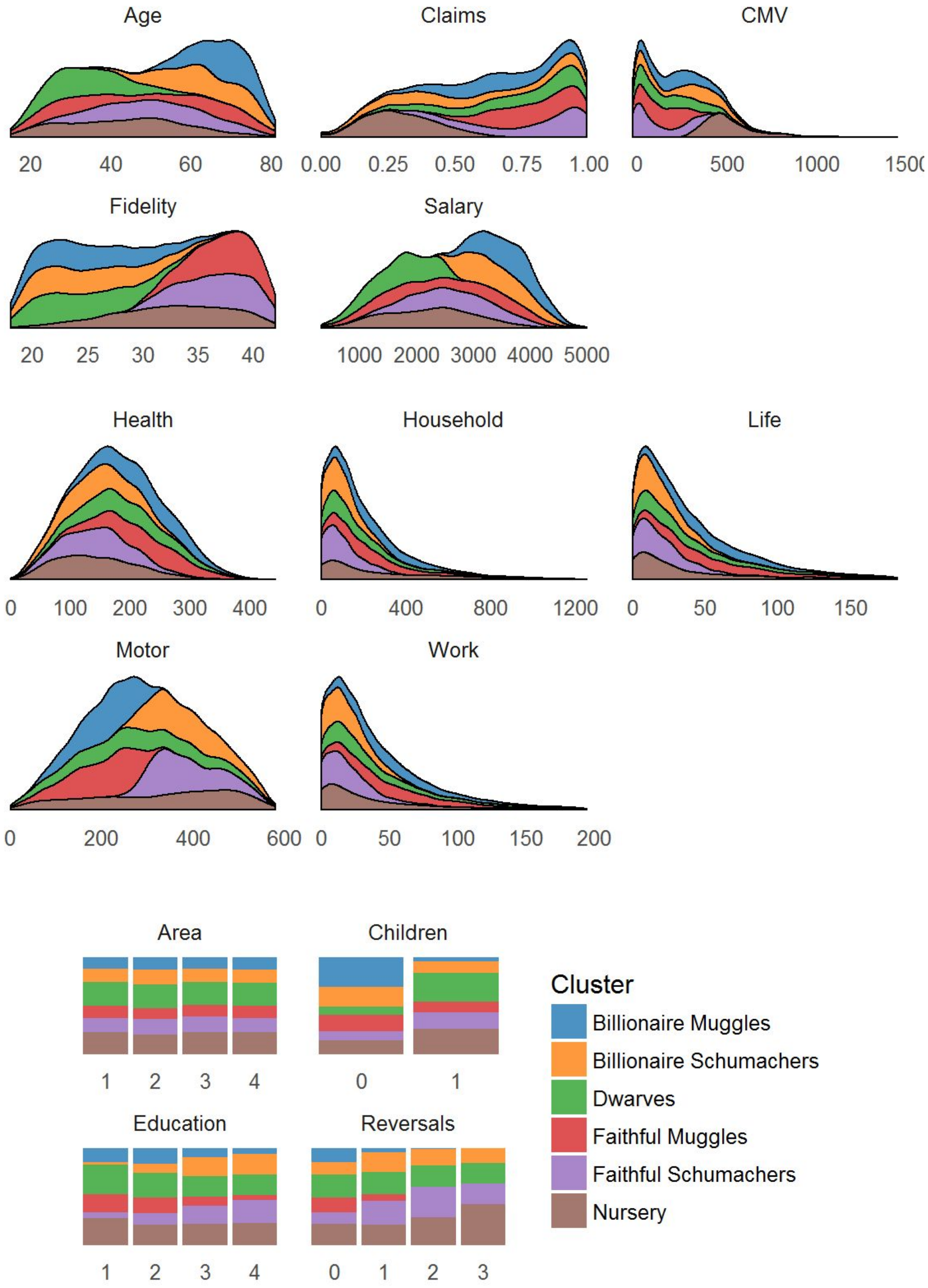
Solution with 3 clusters



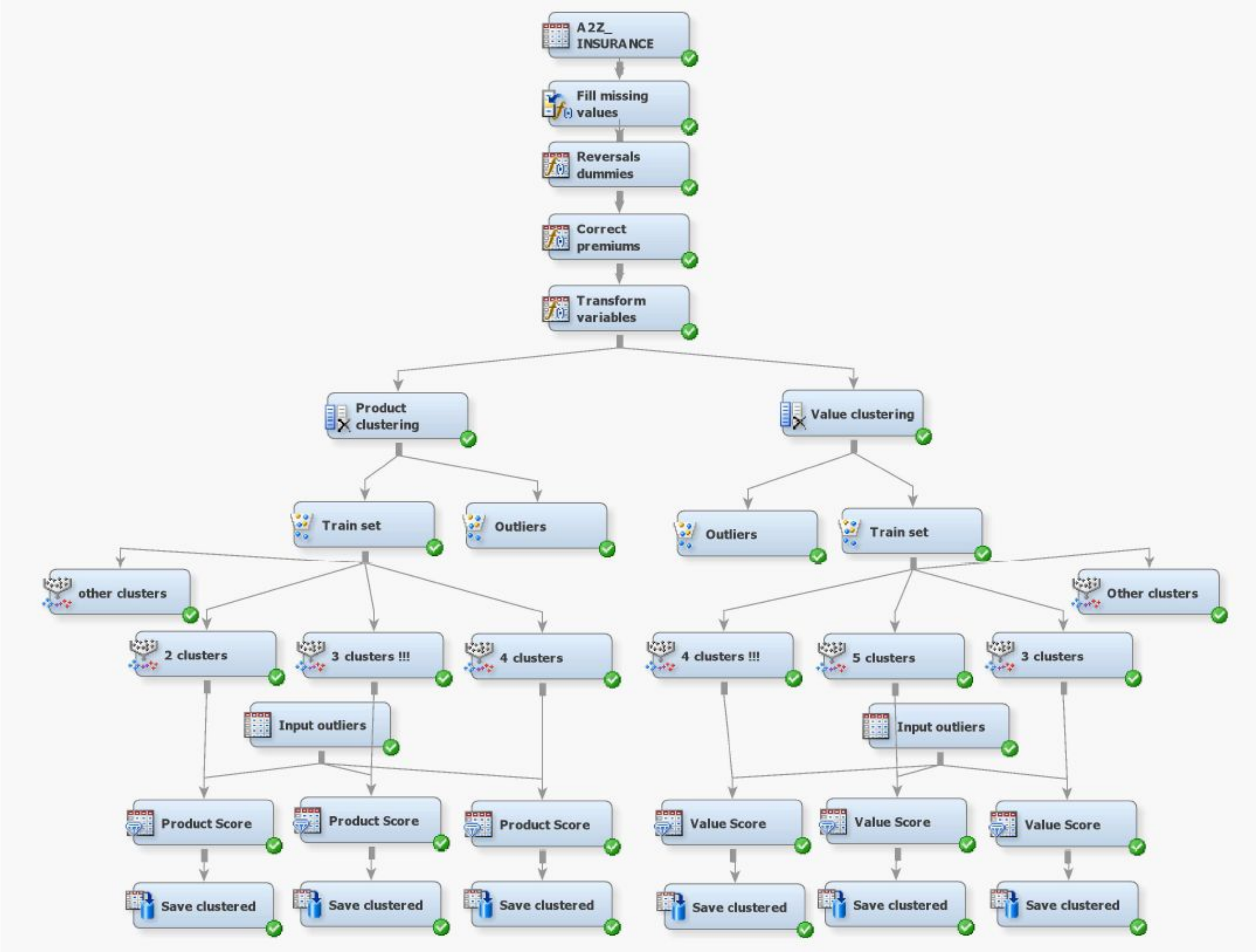
Solution with 4 clusters



6.5 Distribution of final consolidated clusters over all relevant variables



6.6 Project diagram from SAS Miner


A2Z_ INSURANCE
Fill missing values
Reversals dummies
Correct premiums
Transform variables
Product clustering
Value clustering
Train set
Outliers
Outliers
Train set
other clusters
Other clusters
2 clusters
3 clusters !!!
4 clusters
4 clusters !!!
5 clusters
3 clusters
Input outliers
Input outliers
Product Score
Product Score
Product Score
Value Score
Value Score
Value Score
Save clustered
Save clustered
Save clustered
Save clustered
Save clustered
Save clustered