Data Mining I project - Draft report

#### ***Maria Ines Silva***

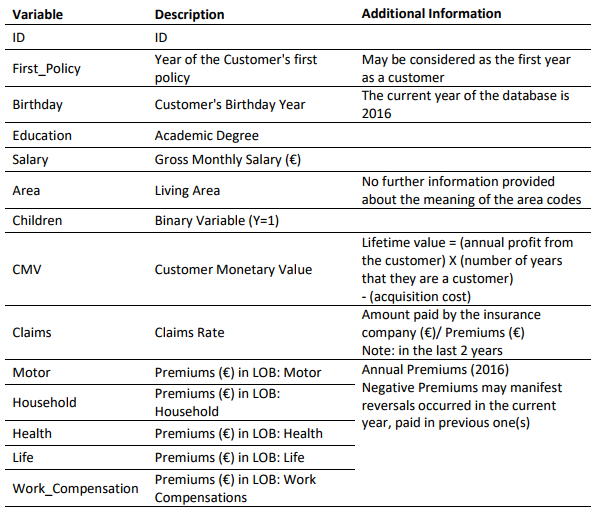
#### ***4th January 2018***

* [Problem and data description](#gjdgxs)
* [Data exploration (EXPLORE)](#30j0zll)
  + [Categorical features](#1fob9te)
  + [Numeric features](#3znysh7)
* [Data preparation (MODIFY)](#2et92p0)
  + [Feature transformation](#tyjcwt)
  + [Feature selection - Value segmentation](#3dy6vkm)
  + [Feature selection - product segmentation](#1t3h5sf)
* [Clustering (MODEL)](#4d34og8)
  + [Value segmentation](#2s8eyo1)
  + [Product segmentation](#17dp8vu)
* [Profiling (ASSESS)](#3rdcrjn)
  + [Value segmentation](#26in1rg)
  + [Product segmentation](#lnxbz9)
  + [Consolidated segmentation](#35nkun2)

Problem and data description

For this project, we received data from a fictional insurer in Portugal. 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.

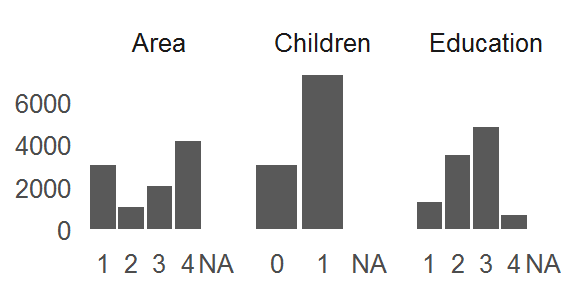
For each customer, the following variables are available:



Data exploration (EXPLORE)

In the data set received there are 10296 customers and 13 features. In particular, there are 3 categorical features and 10 numeric features.

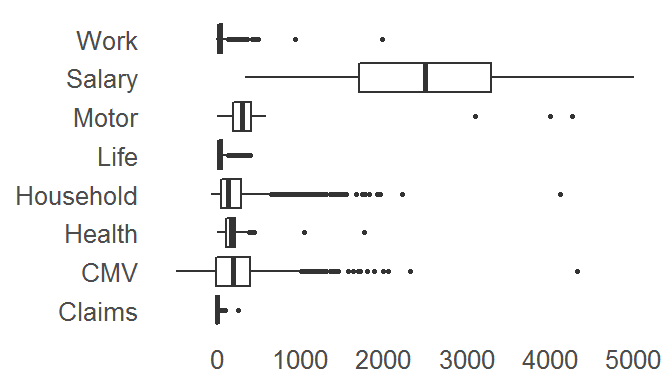
Categorical features

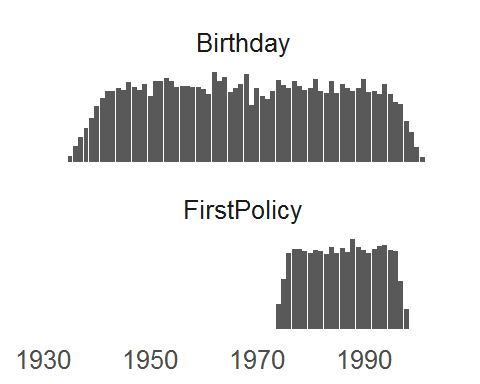


For all 3 features, there are a total 38 observations with missing values.

Numeric features

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Product | Mean | Standar deviation | Median | % missing values | Minimum | Maximum | left threshold | right threshold |
| Birthday | 1968.0 | 19.7 | 1968.0 | 0.002 | 1028.0 | 2001.0 | 1909 | 2027 |
| Claims | 0.7 | 2.9 | 0.7 | 0.000 | 0.0 | 256.2 | -8 | 9 |
| CMV | 177.9 | 1945.8 | 186.9 | 0.000 | -165680.4 | 11875.9 | -5660 | 6015 |
| FirstPolicy | 1991.1 | 511.3 | 1986.0 | 0.003 | 1974.0 | 53784.0 | 457 | 3525 |
| Health | 171.6 | 296.4 | 162.8 | 0.004 | -2.1 | 28272.0 | -718 | 1061 |
| Household | 210.4 | 352.6 | 132.8 | 0.000 | -75.0 | 25048.8 | -847 | 1268 |
| Life | 41.9 | 47.5 | 25.6 | 0.010 | -7.0 | 398.3 | -101 | 184 |
| Motor | 300.5 | 211.9 | 298.6 | 0.003 | -4.1 | 11604.4 | -335 | 936 |
| Salary | 2506.7 | 1157.4 | 2501.5 | 0.003 | 333.0 | 55215.0 | -966 | 5979 |
| Work | 41.3 | 51.5 | 25.7 | 0.008 | -12.0 | 1988.7 | -113 | 196 |





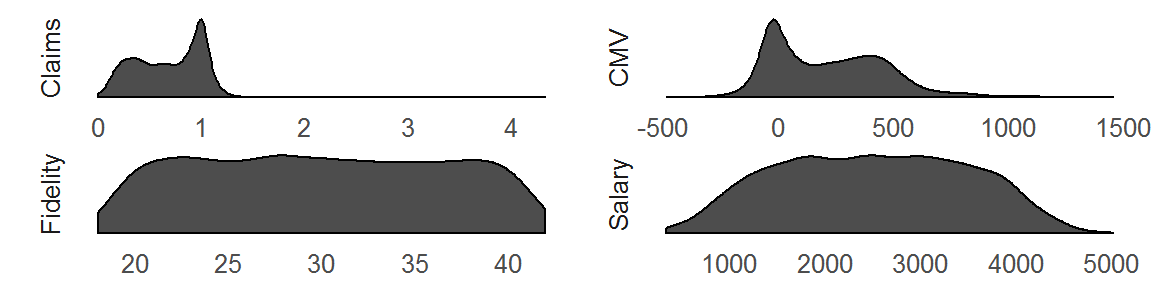
Data preparation (MODIFY)

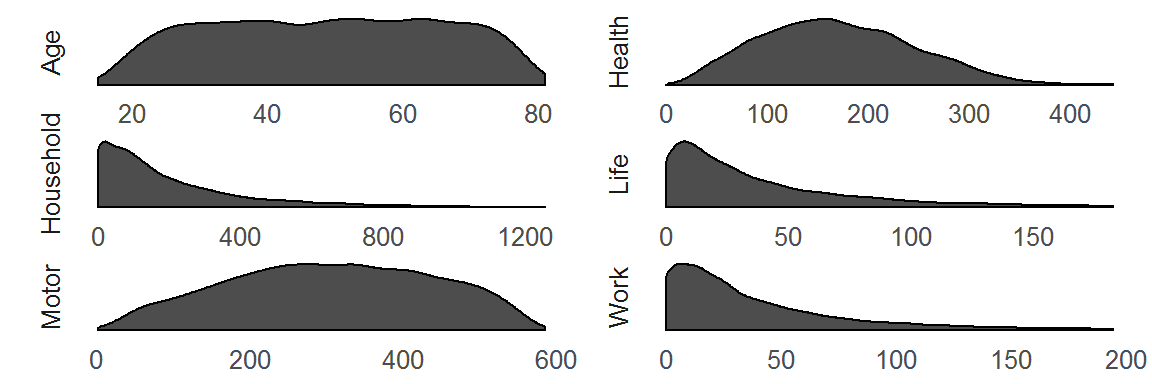
Feature transformation

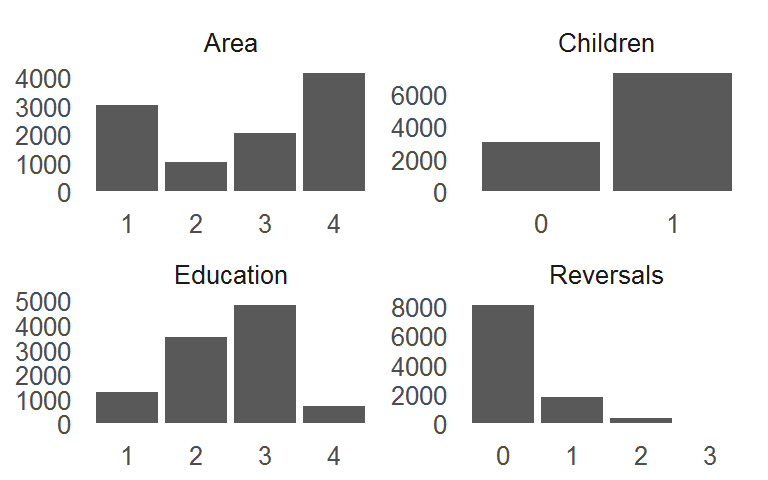
The fist step in our data preparation was to clean the data and make some meaningful transformations. In particular, we preformed the following transformations:

1. **Teatment of missing values** - categoricals were filled with the most frequent class and numericals were filled with the median. The idea here is to avoid excluding customers and /or important information. For this purpose, we used the median to control to the outliers’ bias.
2. **Treatment of Reversals** - since negative premiums represent reversals, we’ll change all negative premiums to zero and we’ll add a new ordinal variable to indicate how many reversals which costumer experienced in 2016.
3. **Treatment of Dates** - the features FirstPolicy and Birthday correspond to calendar years and thus do not belong to the interval variable type. To transform them into a interval type and to help interpret the clusters, these features will be substituted by the Fidelity and Age, respectively. (Fidelity = 2016 - FirstPolicy & Age = 2016 - Birthday).
4. **Treatment of outliers** - numeric features’ outliers will be identified via the Gaussian assumption, i.e., any observations lying outside the interval defined by the mean +/- 3 times the standard deviation will be considered outliers. Note that in the entire sample, there are 425 outliers, which corresponds to 4.1% of the initial set of customers. (We recognize that not all features have a Gaussian distribution, however, for simplicity, we’ll use this criteria)

After these transformations, the features showed the following distributions:







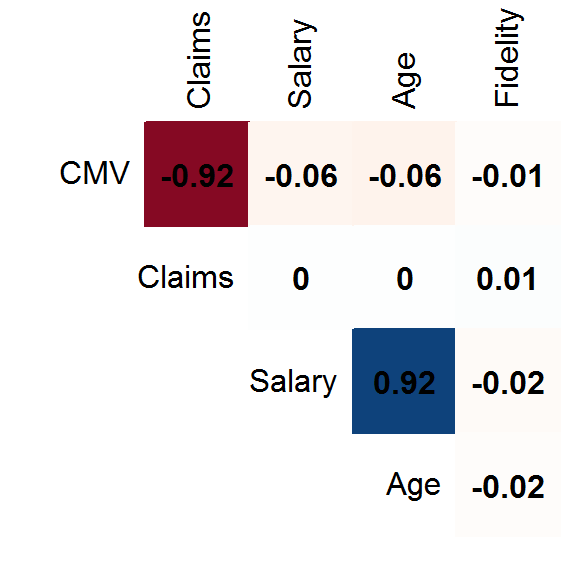
Given the data, our proposal is to build two segmentations - value and product. The product segmentation will consider the columns related with the annual premiums per type of product and will aim to group together customers with similar insurance policies. The value segmentation will considered the remaining features and will aim to characterize the costumers in relation with their value to the company.

The next step is to select which variables should be considered in each segmentation.

Feature selection - Value segmentation

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, i.e.e, be the computation of the clusters. However, they can and were used in the profiling phase.

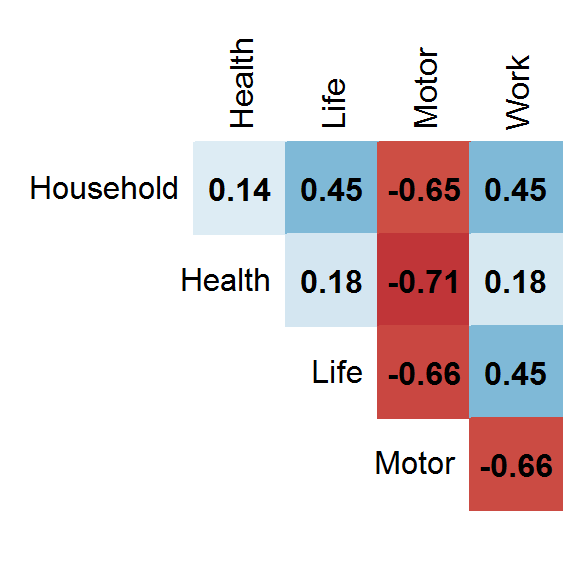
The second point to note is the study of redundancy. In other words, we should only consider variables in the training phase that bring new information, which implies that we should avoid including variables with high correlation (either positive or negative). The following plot illustrates the correlation of the numeric variables related with the value segmentation:



The pairs of features Salary-Age and CMV-Claims show string correlations and thus for the training phase we excluded the Age and the Claims.

Feature selection - product segmentation

Considering now the variables related with the product segmentation, the following plot shows their correlation matrix:

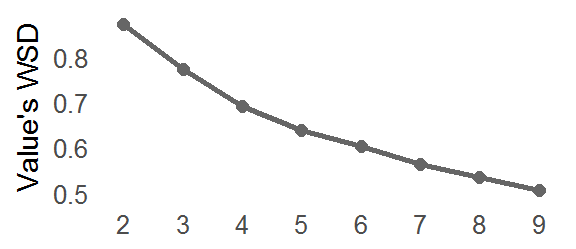


There are no strong correlations between the features and thus we kept them all for the training phase.

Clustering (MODEL)

For both value and product segmentations, the model used is was the k-means. It is a widely used and easy to understand clustering algorithm. **[Main disadvanges? Why can we still use it? Options chosen in SAS?]**

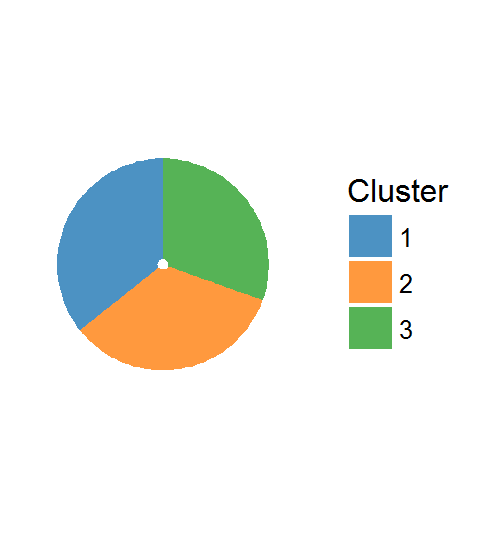
Value segmentation

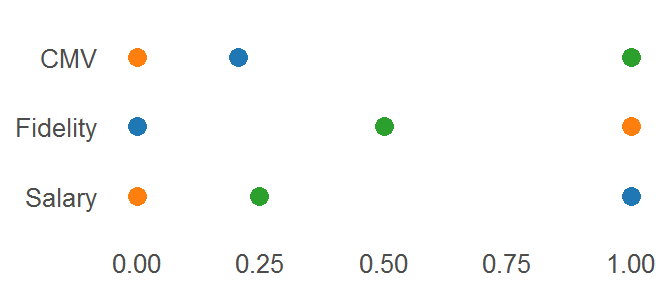


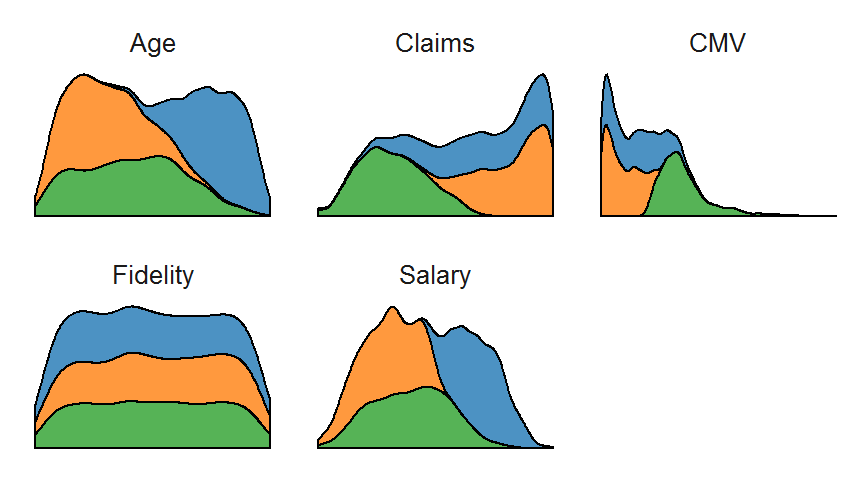
Based on the elbow criteria, we decided to further investigate solutions with 3, 4 and 5 clusters for the value segmentation. Next, for each solution, we’ll present 3 types of plots:

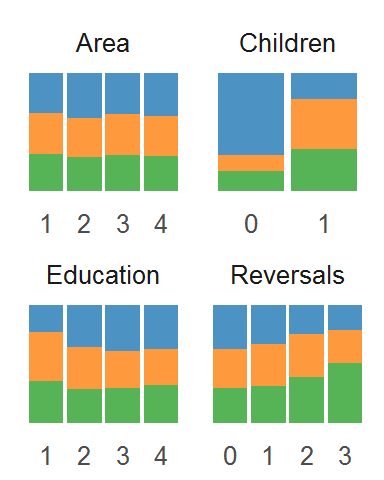
* A pie chart with the distribution of all customers between the different clusters;
* The input mean plot, which shows, for each variable used in the training phase, ….
* Distributions of all relevant features (density for numericals and frequency for categoricals) among each cluster, excluding the outliers.

3 Clusters

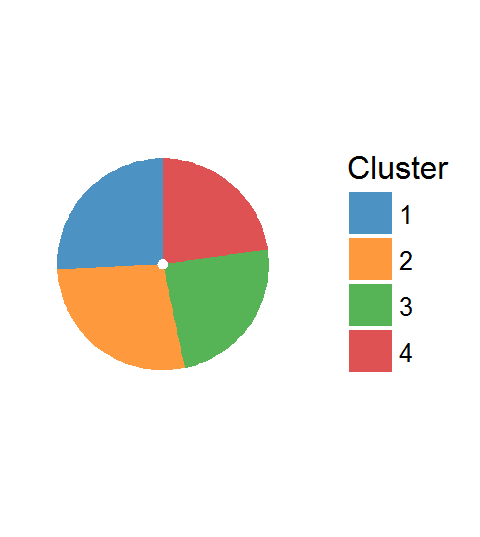


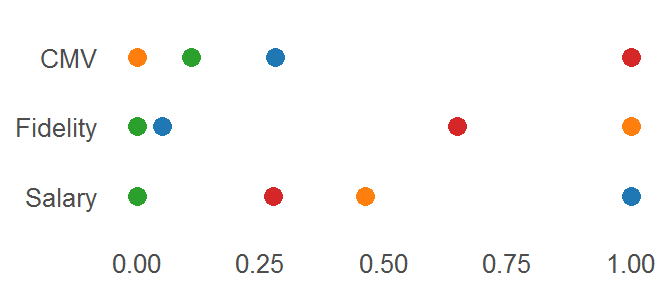


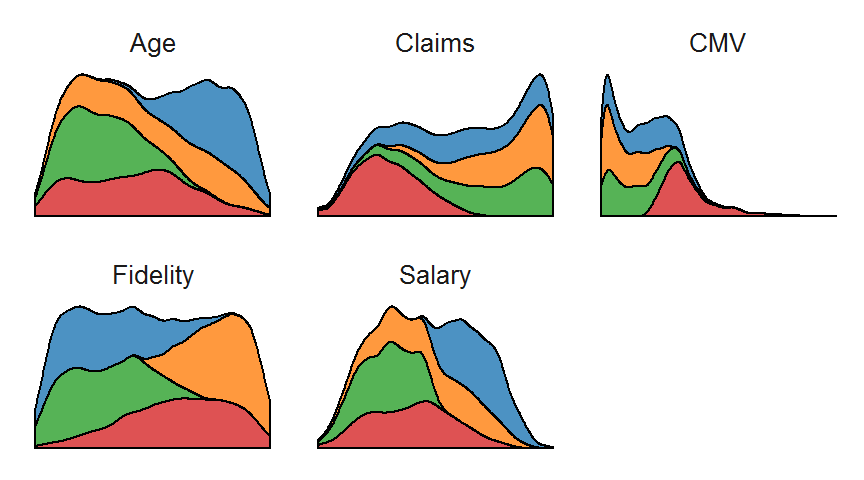


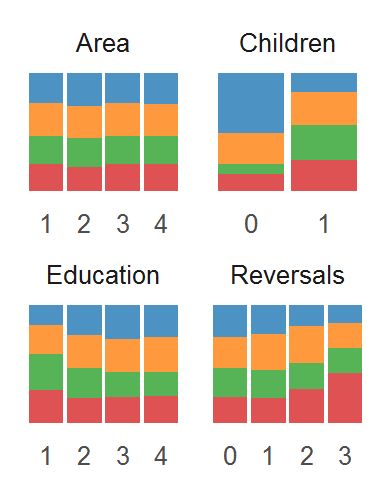


4 Clusters

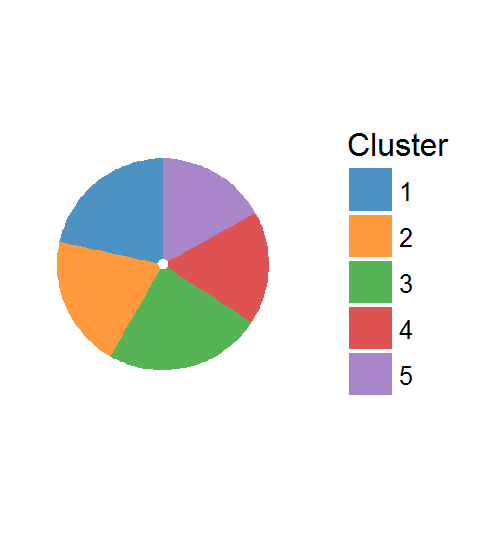


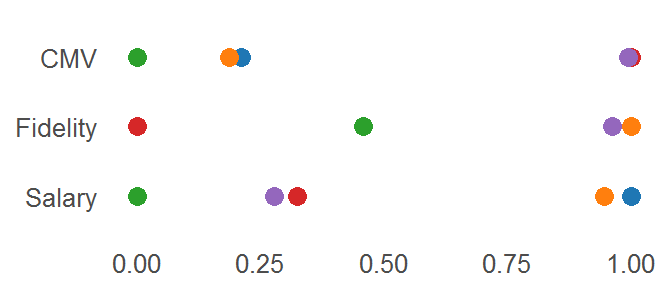


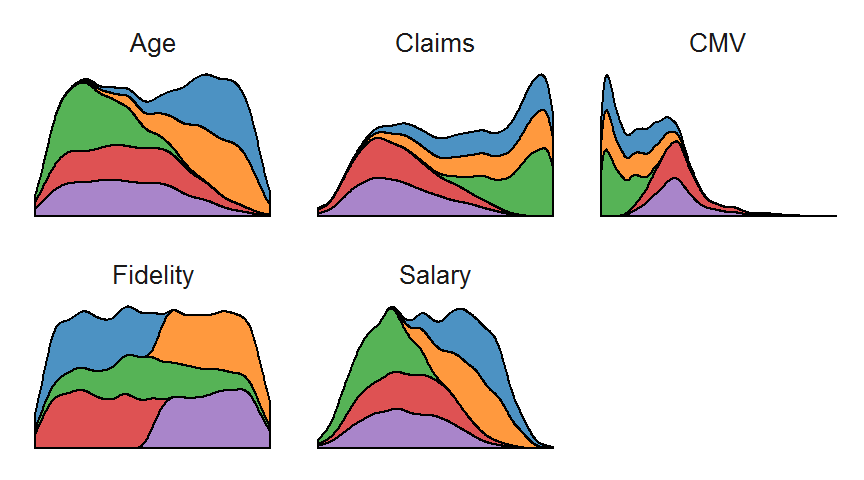


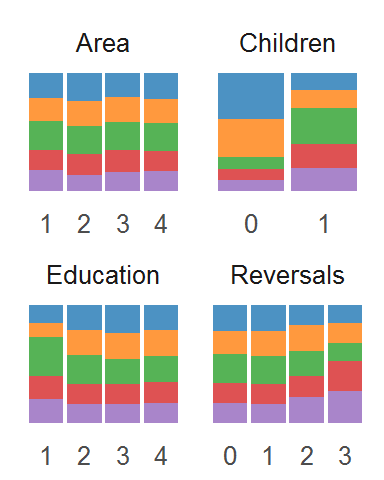


5 Clusters



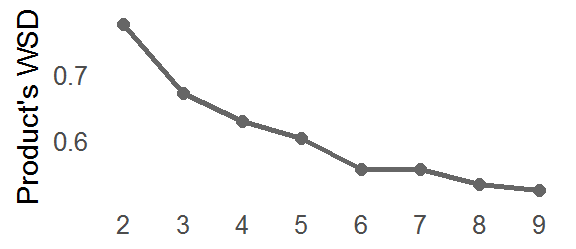






In the end we chose the solution with 4 clusters. If we chose 3 clusters, then the feature Fidelity would have the same distribution on all the clusters.

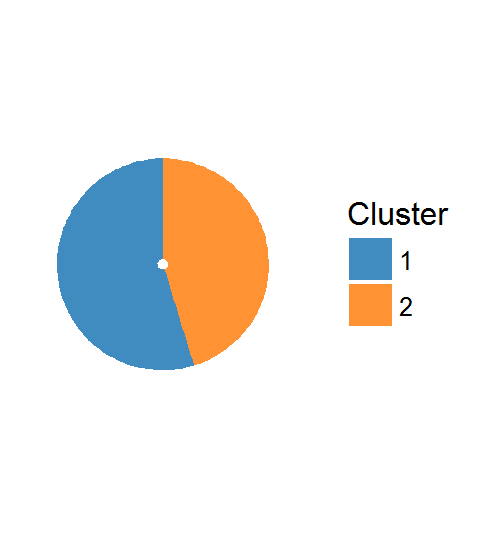
Product segmentation

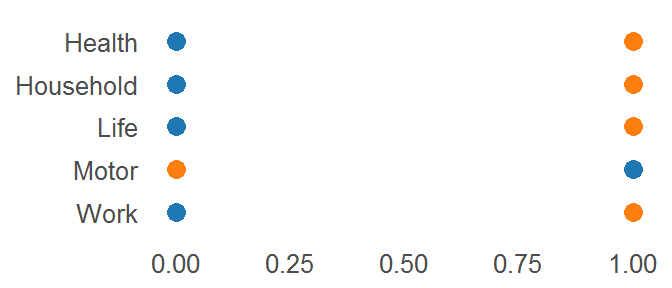


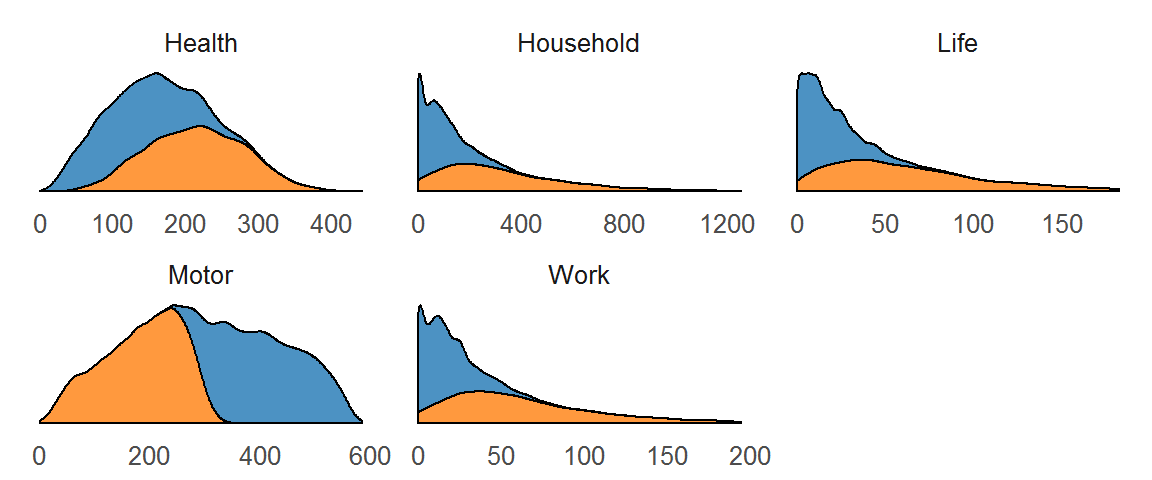
Based on the elbow criteria, we decided to further investigate solutions with 2, 3 and 4 clusters for the product segmentation.

The same plots showed for the value segmentation are now presented for the 3 solutions of the product segmentation.

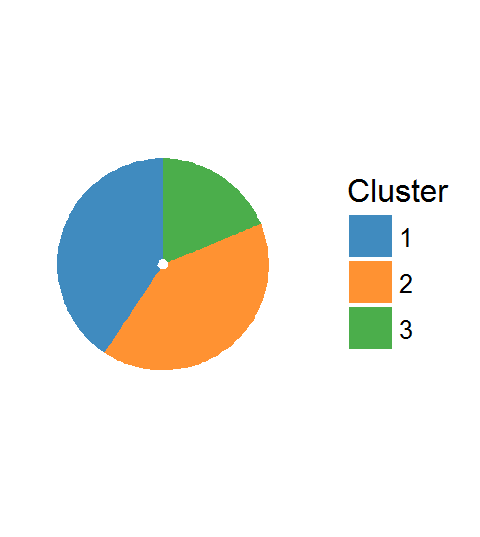
2 Clusters

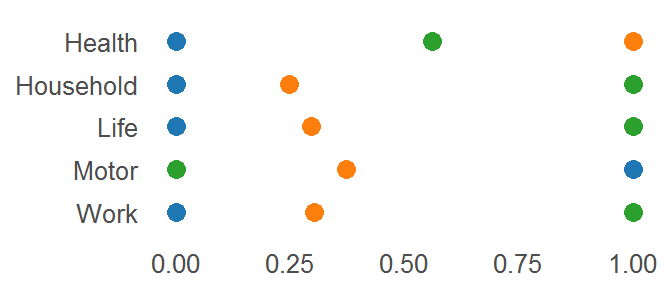


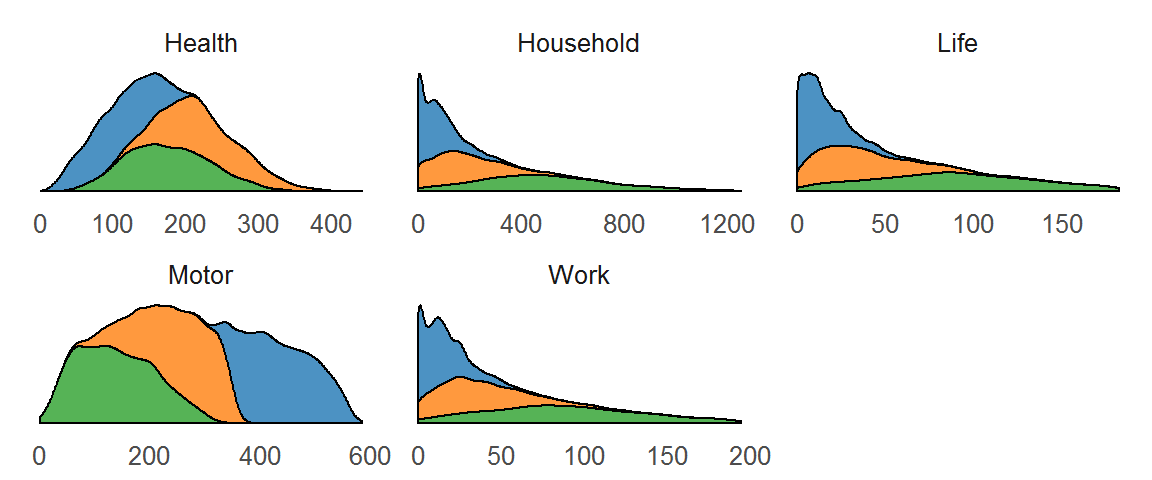




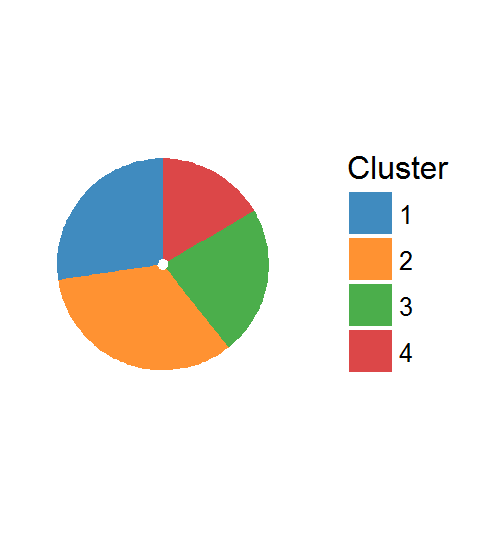
3 Clusters

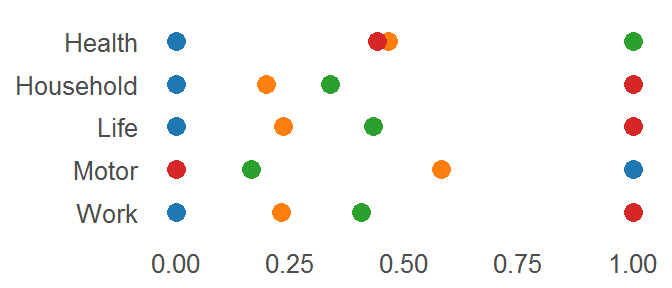


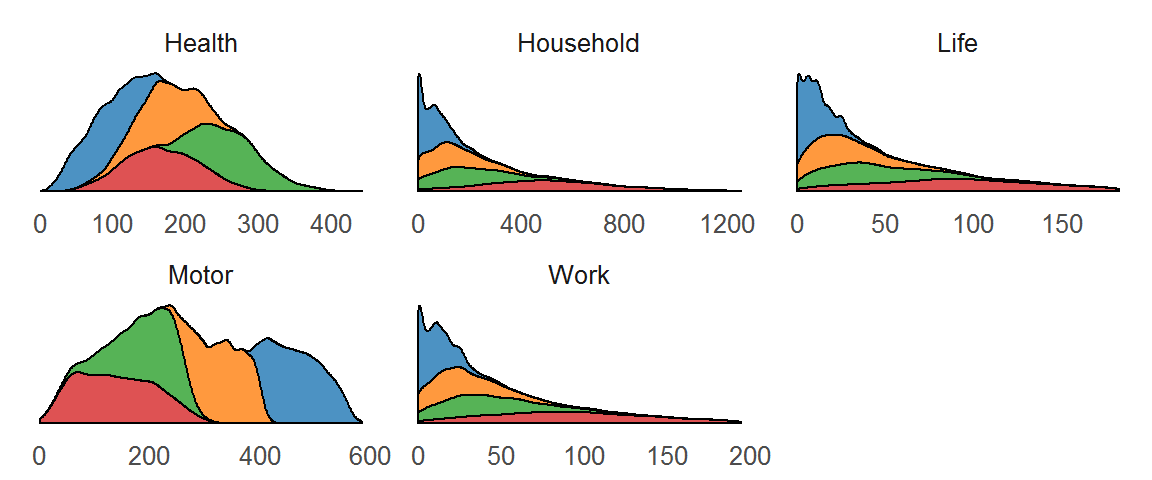




4 Clusters





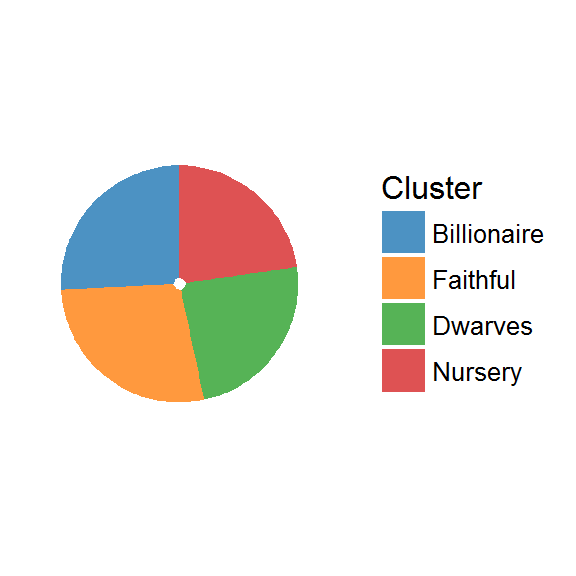


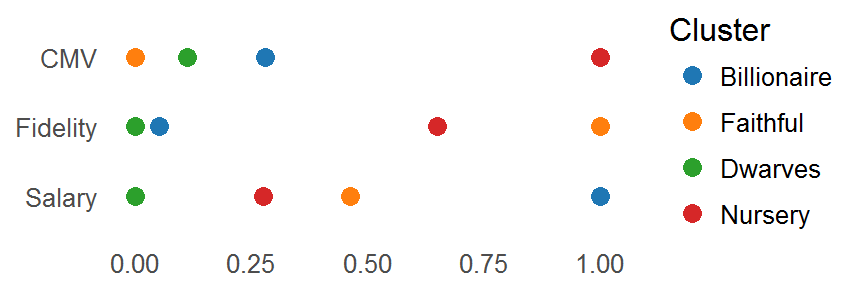
In the end, we chose the solution with 2 clusters because it gives a good enough differentiation for the features Health and Motor. Note that increasing the number of clusters does not increase the differentiation for the other variables.

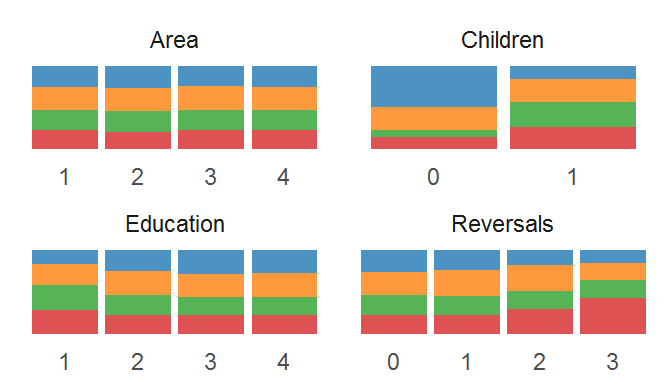
Profiling (ASSESS)

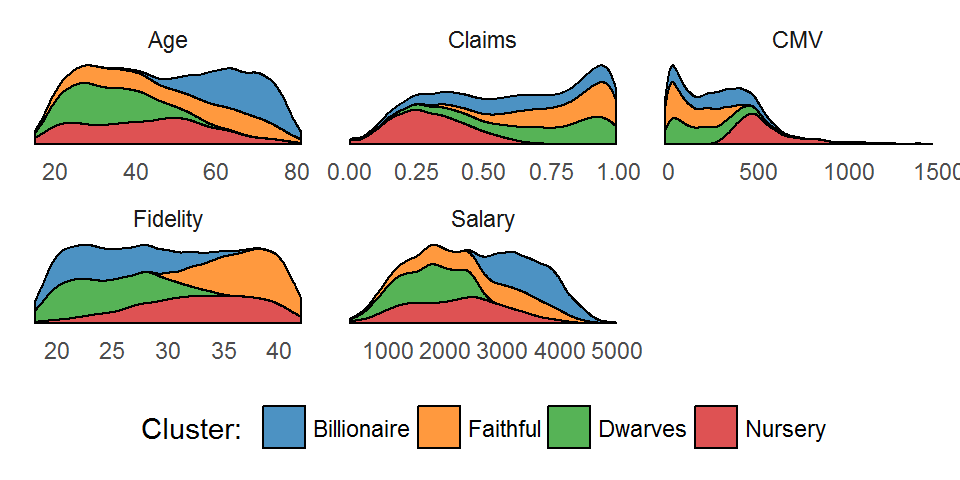
The profiling will encompass two main steps. Firstly, we’ll profile the two segmentations separately. Secondly, we’ll consolidate both segmentations to have a global view of the customers and decide on the Marketing strategy.

Value segmentation

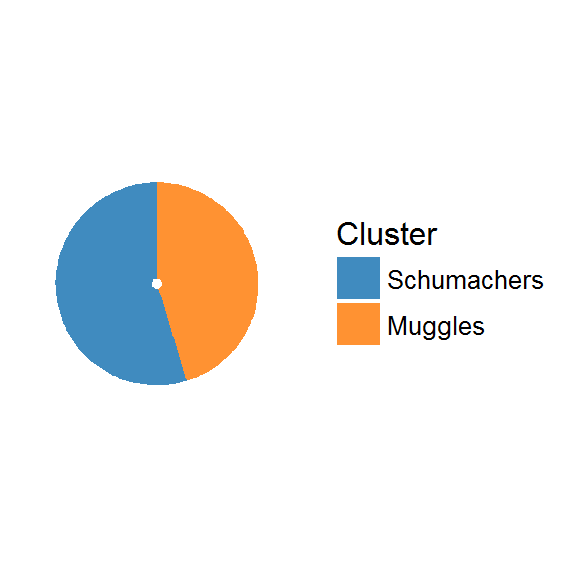


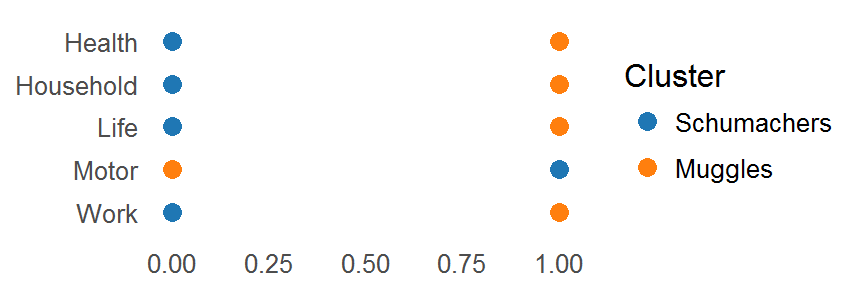


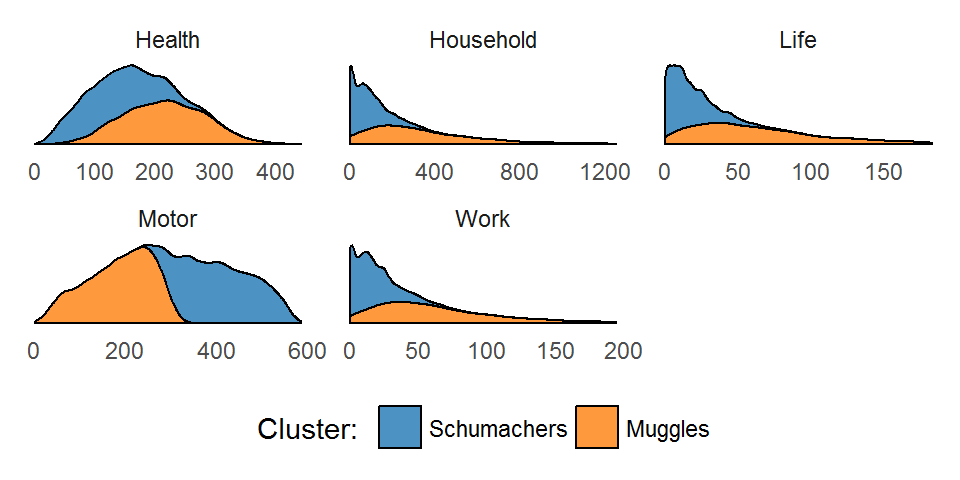




Product segmentation







Consolidated segmentation

When we consolidate the two segmentation, we get 8 clusters:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| value\_cluster | product\_cluster | number | freq | full\_cluster | CMV | Salary | Fidelity | Household | Health | Life | Motor | Work |
| Nursery | Muggles | 1015 | 0.0985820 | Nursery Muggles | 539.95743 | 2182.164 | 32.53171 | 329.63775 | 214.8857 | 61.72158 | 185.1436 | 61.83922 |
| Dwarves | Schumachers | 1130 | 0.1097514 | Dwarves Schumachers | 114.05033 | 1675.434 | 24.79577 | 88.39709 | 132.2176 | 17.16481 | 405.8858 | 17.28489 |
| Billionaire | Muggles | 1229 | 0.1193667 | Billionaire Muggles | 195.19015 | 3513.246 | 25.40324 | 329.63775 | 214.8857 | 61.72158 | 185.1436 | 61.83922 |
| Faithful | Muggles | 1303 | 0.1265540 | Faithful Muggles | 60.79571 | 2526.169 | 36.71732 | 329.63775 | 214.8857 | 61.72158 | 185.1436 | 61.83922 |
| Dwarves | Muggles | 1330 | 0.1291764 | Dwarves Muggles | 114.05033 | 1675.434 | 24.79577 | 329.63775 | 214.8857 | 61.72158 | 185.1436 | 61.83922 |
| Nursery | Schumachers | 1331 | 0.1292735 | Nursery Schumachers | 539.95743 | 2182.164 | 32.53171 | 88.39709 | 132.2176 | 17.16481 | 405.8858 | 17.28489 |
| Billionaire | Schumachers | 1425 | 0.1384033 | Billionaire Schumachers | 195.19015 | 3513.246 | 25.40324 | 88.39709 | 132.2176 | 17.16481 | 405.8858 | 17.28489 |
| Faithful | Schumachers | 1533 | 0.1488928 | Faithful Schumachers | 60.79571 | 2526.169 | 36.71732 | 88.39709 | 132.2176 | 17.16481 | 405.8858 | 17.28489 |

Although customers are almost uniformly distributed among all clusters, we considered that 8 is too large a number of clusters for creating different Marketing strategies. Therefore, we will aggregate the 2 clusters with the lowest number of customers in order to have only 6 clusters.

To do it, we chose the value sgmentation as the main one and thus we change the product segmentation when we need to reduce the number of cluster.

After applying these changes, the final clusters have the following ditributions and pie chart:

