NOVA IMS Information Management School

MAA

Master Program in Advanced Analytics

Customer Clustering in the Insurance Sector by Means of Unsupervised Machine Learning

An internship report

Thies Bücker

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

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

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

CUSTOMER CLUSTERING IN THE INSURANCE SECTOR BY MEANS OF UNSUPERVISED MACHINE LEARNING

by

Thies Bücker

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics.

Advisor: Mauro Casteli

April 2016

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Mauro Casteli, of Nova Information Management School of Universidade Nova de Lisboa for all helpful advices and guidance. Also I want to thank the team of Accenture who gave me the chance to work on a real problem and provided me with all necessary resources that were needed to conduct the report. Without their passionate support I would not have been able to conduct the report.

ABSTRACT

Clustering is one of the most frequently applied techniques in machine learning. An overview of the most comon algorithms, problems and solutions is provided in this report. In modern times, customer information is a curtail success factor in the insurance industry. This work describes a way how customer data can be leveraged in order to provide useful insights that help driving business in a more profitable way. It is shown that the available data can serve as a base for customer segmentation on which further models can be built upon. The customer is investigated in three dimensions (demographic, behavior, and value) that are crossed to gain precise information about customer segments.

KEYWORDS

Customer segmentation, clustering, k-means, unsupervised learning, segmentation

INDEX

1. Introduction 1
2. Motivation 2
3. Clustering 3
   - 3.1. Definition and Application 3
     - 3.1.1. Definition 3
     - 3.1.2. Fields of Research 4
     - 3.1.3. Clustering Approaches 4
   - 3.2. Hierarchical Clustering 5
     - 3.2.1. Dendrogram 6
     - 3.2.2. Similarity Measures 7
     - 3.2.3. Metrics 8
     - 3.2.4. Divisive Algorithms 9
   - 3.3. Partitional Clustering 10
     - 3.3.1. K-Means Algorithm 10
     - 3.3.2. Advantages and Disadvantages of K-Means 10
     - 3.3.3. Outlier Detection 12
     - 3.3.4. Initializations for K-Means 15
     - 3.3.5. Number of Clusters 20
   - 3.4. Cluster validation & Similarity Measures 21
     - 3.4.1. External Validation Techniques 22
     - 3.4.2. Internal Validation Techniques 24
4. Customer Clustering For an insurance 29
   - 4.1. Data Sources 29
   - 4.2. Segmentation 31
     - 4.2.1. Demographic Segmentation 32
     - 4.2.2. Behavior Segmentation 37
     - 4.2.3. Value Segmentation 45
   - 4.3. Final Deliverable 54
5. Results and Limitations 56
6. Bibliography 58
7. Appendix 63
   - 7.1. List of variables in CAR 63

7.2. Correlation Matrix for Demographic Segmentation 97
7.2.1. Section 1 of the demographic correlation matrix 99
7.2.2. Section 2 of the demographic correlation matrix 100
7.2.3. Section 3 of the demographic correlation matrix 101
7.2.4. Section 4 of the demographic correlation matrix 101
7.2.5. Section 5 of the demographic correlation matrix 102
7.2.6. Section 6 of the demographic correlation matrix 102
7.3. Profiling of Demographic segments 102
7.4. Profiling of Behavior segments 106
7.5. Profiling of Value segments.....109

LIST OF FIGURES

Figure 1 - Dendrogram example 6
Figure 2 - Similarity measures 7
Figure 3 - Creation of the CAR 30
Figure 4 - Number of customers in demographic segments 35
Figure 5 - Age distribution of demographic segments 35
Figure 6 - Scatter plot of the first two principal components for the demographic segmentation 37
Figure 7 - Number of customers in behavior segments 42
Figure 8 - Number of years since last purchase 42
Figure 9 - Number of new policies (12 month) vs. number of total policies 43
Figure 10 - Number of contracts as an insured person 44
Figure 11 - Scatter plot of the first two principal components for the behavior segmentation 45
Figure 12 - Number of clients in value segments 50
Figure 13 - Average current value of financial products per client 51
Figure 14 - Average technical margin by value segment 51
Figure 15 - Scatter plot of first two principal components for value segmentation 53

LIST OF TABLES

Table 1 - Definitions of clustering 3
Table 2 - Classical outlier definitions 13
Table 3 - Translation from Event/Non-Event to True/False Positive/Negative. Correctly classified are marked in green, while incorrectly classified observations are marked in orange. 22
Table 4 - Categories and key words for topic extraction from phone call notes 31
Table 5 - Fraction of missing values for demographic segmentation 33
Table 6 - Fraction of missing values for behavior variables 39
Table 7 - Correlations among behavior variable 40
Table 8 - Fraction of missing values among value variables 47
Table 9 - Correlation among value variables 49
Table 10 - Overview of the demographic correlation matrix 98
Table 11 - First section of the demographic correlation matrix 99
Table 12 - Second section of the demographic correlation matrix 100
Table 13 - Third section of the demographic correlation matrix 101
Table 14 - Fourth section of the demographic correlation matrix 101
Table 15 - Fifth section of the demographic correlation matrix 102
Table 16 - Sixth section of the demographic correlation matrix 102

1. INTRODUCTION

The Master's degree in Advanced Analytics contains a praxis phase, which can be done in form of an internship with a company. For this report Accenture has been chosen to be the partnering company for the internship. Because of the partnership between Accenture and Nova IMS it is ensured that the content of the internship fits the topics of the university course. Due to that reason the internship has been held in the Advanced Analytics department of Accenture in Lisbon between the 26.10.15 and 22.04.16. The main reason for the internship is to show the student relevant topics in the area of advanced analytics while improving his practical hard skills and soft skills.

This report consists of a theoretical and a practical part. Both parts throw light on the topics of the internship. This is mainly clustering, but also other areas, like data preprocessing are touched in the report. In the theoretical part typical approaches of clustering are explained, and problems are pointed out. Later a collection of possible solutions for these problems is explained. The theoretical part and the practical part serve different purposes. While the former answers mainly general mathematical questions and describes properties of various concepts, the latter describes the practical steps that have been undertaken for a solution with a real data set.

2. MOTIVATION

In the past years the environment for business has changed and is changing on an accelerating velocity. Due to globalization new competitors emerge and due to modern communication technology and media, prices and services are comparable easily, cost-efficiently and fast. Enterprises that ran their business model successfully for years, might face new challenges and obstacles because of quickly changing customer behavior or technologies that are applied by a competitor. These changes increase the pressure on companies of all industries. One of the industries that is effected most by these changes is the insurance industry. In this particular industry the number of clients of a company has is one of the main success drivers and has a major impact on the profitability of the company. And thus, customer retention becomes one of the key concerns of insurance businesses in our days. The cost of attracting new customers is much higher than the cost of retaining the clients that are already related to the company. For this reason it is crucial to the enterprises to keep their clients active. In order to do so, clients must be managed in a customized way that suits their needs and makes them profitable to the enterprise. For an insurance it is indispensable to precisely know its clients. Clients cannot all be treated in the same manner, but must be handled in a tailored way. In pursuance of that it is necessary for a company to know its customers. In many businesses information about the client is rarely used, even though it is available. One way of getting to know the customer base and identifying possible management strategies is clustering. Clustering provides the applicant with a sense of what a typical client is. Further it equips the researcher with a solid understanding of different customer segments. Based on these segments specific strategies can be worked out to target the customer in a more specific, and so more successful and efficient way.

This internship report demonstrates how such a segmentation is being accomplished for an insurance company. Further it illustrates a variety of different approaches that can be used to apply clustering for different problems in various fields.

3. CLUSTERING

Because clustering has been discussed widely in literature over the past decades, a variety of definitions can be found while researching about clustering. That is why it should be started by giving an overview of definitions. This goes together with a brief description of the fields in which clustering is applied. Later on different clustering branches will be described, compared and common problems will be explained.

3.1. DEFINITION AND APPLICATION

Clustering is not one specific algorithm, but rather a group of algorithms with similar goals. This section should give conclusion about the matter of clustering, provide an overview of different clustering types, and should give a brief summary of the field in which clustering is applied broadly.

3.1.1. Definition

The exact definition of clustering varies depending on the researcher and the respective research field and topic. In general clustering is considered a form of unsupervised machine learning (A. Jain, Murty, & Flynn, 1999). In contrast to supervised machine learning, it does not try to approach a target value. However it might be helpful to compare the outcome of clustering algorithm to the real clusters, if available. Specific techniques are described in section External Validation Techniques.

The following table provides an overview of different definitions.

Author(s)	Definition
(Jain, Murty, & Flynn, 1999, p.264)	"Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)"
(Liu, Li, Xiong, Gao, & Wu, 2010, p.911)	"dividing a set of objects into clusters such that objects within the same cluster are similar while objects in different clusters are distinct"
(Rendón et al., 2008, p.241)	"determine the intrinsic grouping in a set of unlabeled data, where the objects in each group are indistinguishable under some criterion of similarity"
(A. K. Jain, 2010, p.651)	"Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity"
(Likas, Vlassis, & Verbeek, 2003, p.12)	"defined as a region in which the density of objects is locally higher than in other regions"

Table 1 - Definitions of clustering

In essence clustering pursues the objective of identifying groups of data that are similar to data of the same group and less similar to data of other groups. In this work the terms groups, segments, clusters, and classes are used synonymously.

3.1.2. Fields of Research

Such properties, as described above, are desirable in a wide range of fields. The following list should provide an overview of some of the most relevant fields in clustering:

- **Customer Segmentation:** To provide companies with potentially interesting insights about their customer, and help them manage different customer segments in different ways.
- **Human Genetic Clustering:** Applies clustering to human genetic data in order to draw conclusions about population structure (de Hoon, Imoto, Nolan, & Miyano, 2004).
- **Medical Imaging:** Uses clustering to make similarly behaving structures insight the body visible and interpretable on a monitor.
- **Image Segmentation:** Clustering is used to detect similar regions in a visual image. This technique is applied in research topics like object recognition, video analysis for self-driving cars, and map creation from satellite images.

3.1.3. Clustering Approaches

Due to its appeal to many research areas, a variety of different techniques have emerged over time. This section will distinguish some of the main groups of clustering algorithms.

Bottom-up clustering under the use of a similarity measure is called hierarchical agglomerative clustering. Bottom-up means, that each observation starts in a cluster on its own (singleton) and is then merged with the most similar cluster, or observation, to form a bigger cluster. This process ends at the latest when all observations are merged together into one single cluster. The similarity measure is a parameter of the algorithm, and can be chosen, for example, from the list provided in Similarity Measures (Duda, Hart, and Stork 2001).

One advantage of the agglomerative method is, that the distances at which the clusters have been merged can provide useful information. This information can be useful when deciding the number of clusters. Details about how to use these distances are provided in section Dendrogram.

Cimiano, Hotho, & Staab (2004) point out that the difference between agglomerative and divisive approaches is the bottom-up approach (agglomerative) in contrast to the top-down method (divisive). In divisive clustering algorithms all observations start in a single cluster, and are then subdivided into sub clusters, until each observation ends up in a cluster on its own.

The differentiation between monothetic and polythetic offers information about the sequential or simultaneous usage of variables. Anderberg (1973) applied a monothetic algorithm that considers one variable after another. The first variable divides the population into two groups, and the second variable subdivides these groups into two sub groups each. This process is carried out until all variables are used. The number of resulting clusters is determined as two by the power of the number of used variables. A. Jain et al. (1999) point out, that many applications have a

dimensionality (number of used variables) that is greater than 100, and would result in a number of clusters that "is so large that the data set is divided into uninterestingly small and fragmented clusters" (p.275). This behavior is the reason why polythetic methods find greater response in research. All further described algorithms use the polythetic approach.

Hard clustering is distinct from fuzzy clustering in terms of the membership of an observation to a cluster. While in hard clustering each observation is assigned to one single cluster, it can be assigned to multiple clusters in fuzzy clustering. In fuzzy clustering each observation has a degree of membership to each cluster, so that the association with the segments is relative. Observations that seem to fit better to one cluster have a greater degree of membership for that cluster, than observations that are only weakly associated to a cluster (A. Jain et al., 1999).

One advantage of fuzzy clustering is, that it can be translated to hard clustering by assigning an observation to the cluster with the highest degree of membership. A translating from hard clustering to fuzzy clustering is not possible without further calculations.

Another dimension that differentiates clustering approaches is how the applied algorithm finds its optimum. If it includes a form of random search (stochastic), or if it approaches the optimal value using a step by step deterministic method (A. Jain et al., 1999).

Clustering techniques can also be divided into incremental and non-incremental clustering. Charikar, Chekuri, Feder, and Motwani (2004) define incremental clustering as follows:

> for an update sequence of n points in M, maintain a collection of k clusters such that as each input point is presented, either it is assigned to one of the current k clusters or it starts off a new cluster while two existing clusters are merged into one (p.1419)

The condition to split a cluster, and which and when clusters are joined together is subject to the respective implementation. Incremental clustering is used when data sets become so large, that they are too big to be processed at once, and therefore need to be processed observation by observation. In this case it is important to minimize the number of scans through the data set (A. Jain et al., 1999).

Non-incremental clustering, in contrast, deals with the whole data set at once. The algorithms that will be presented are non-incremental functions.

3.2. Hierarchical Clustering

Two of the major branches in clustering are hierarchical clustering and partitional clustering. This chapter will give an overview of the most important topics in hierarchical clustering, while the following chapter will do so for the partitional clustering algorithms.

"Hierarchical clustering algorithms produce a nested series of partitions based on a criterion to merging or splitting clusters based on similarity" (Jain et al., 1999, p.268). Hierarchical algorithms can be split into two main categories, agglomerative and divisive. In which agglomerative stands for a bottom up approach that starts with all observations in a single cluster, and merges the observations step by step according to a similarity measure, until the entire population is merged into the same cluster. In a subsequent step the dendrogram is used to define the optimal number of clusters. The dendrogram is explained in the section Dendrogram.

Of a divisive approach is spoken, when a top down method is used. In contrast to the agglomerative approach, all observations start in a single cluster and are split based on a similarity measure until clusters cannot be split anymore, because they only consist of a single observation (Cimiano et al., 2004).

3.2.1. Dendrogram

The procedure agglomerative methods use, merges the most similar observations in an iterative manner. The dendrogram is a chart that gives conclusion about the distance at which two observations/clusters have been joined. It can be used as an indicator for the number of clusters. An example of a dendrogram for a hypothetical clustering task can be seen in Figure 1. The y-axis illustrates the distance at which two observations have been merged. It can be observed, that the two clusters that are joined at last, have a distance greater than 200. In this case the researcher decides to have three final clusters, which have a distance of greater than 75. The next two clusters that would be joined together would be the green and the blue ones.

Note that the dendrogram does not give conclusion about the used similarity measure and metric.

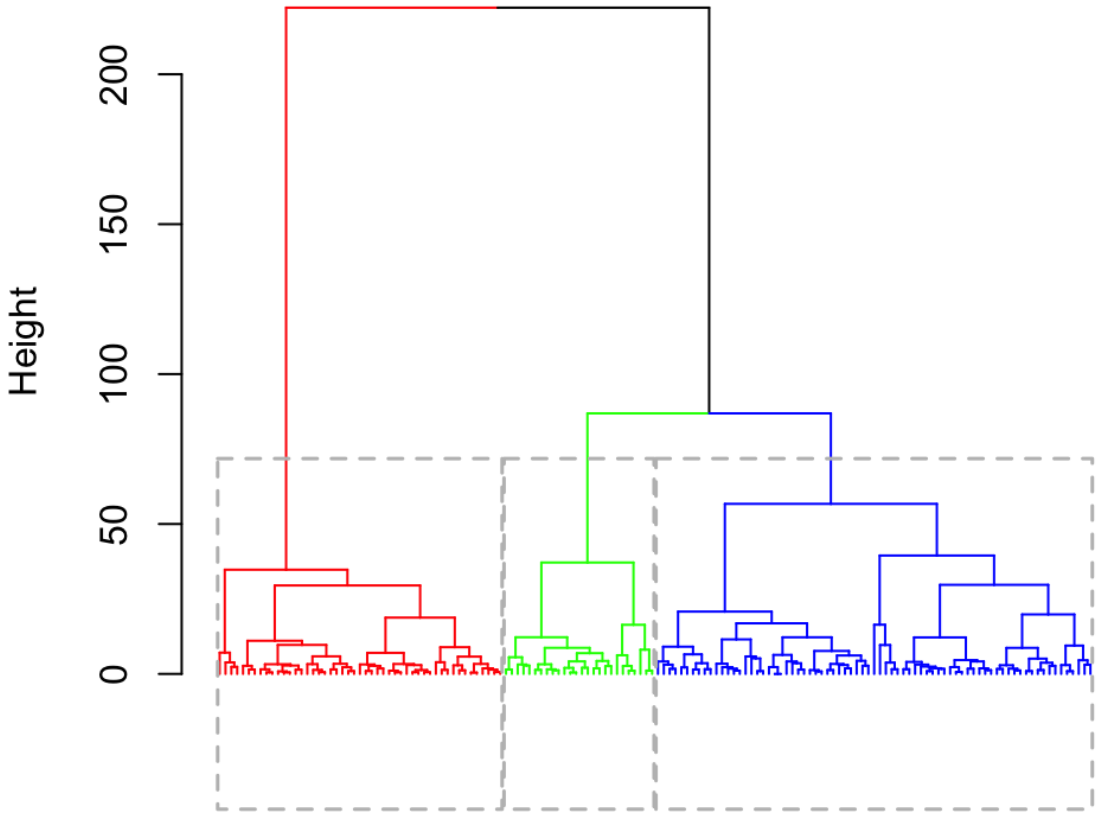


Figure 1 - Dendrogram example

3.2.2. Similarity Measures

The criteria on which the distances between clusters are evaluated are called similarity measures. These techniques determine how distances are considered and which clusters, or observations, are merged at next. A variety of different similarity measures exist. The most important ones are explained in this section. Figure 2 illustrates how selected similarity measures evaluate the distance between two hypothetical segments.

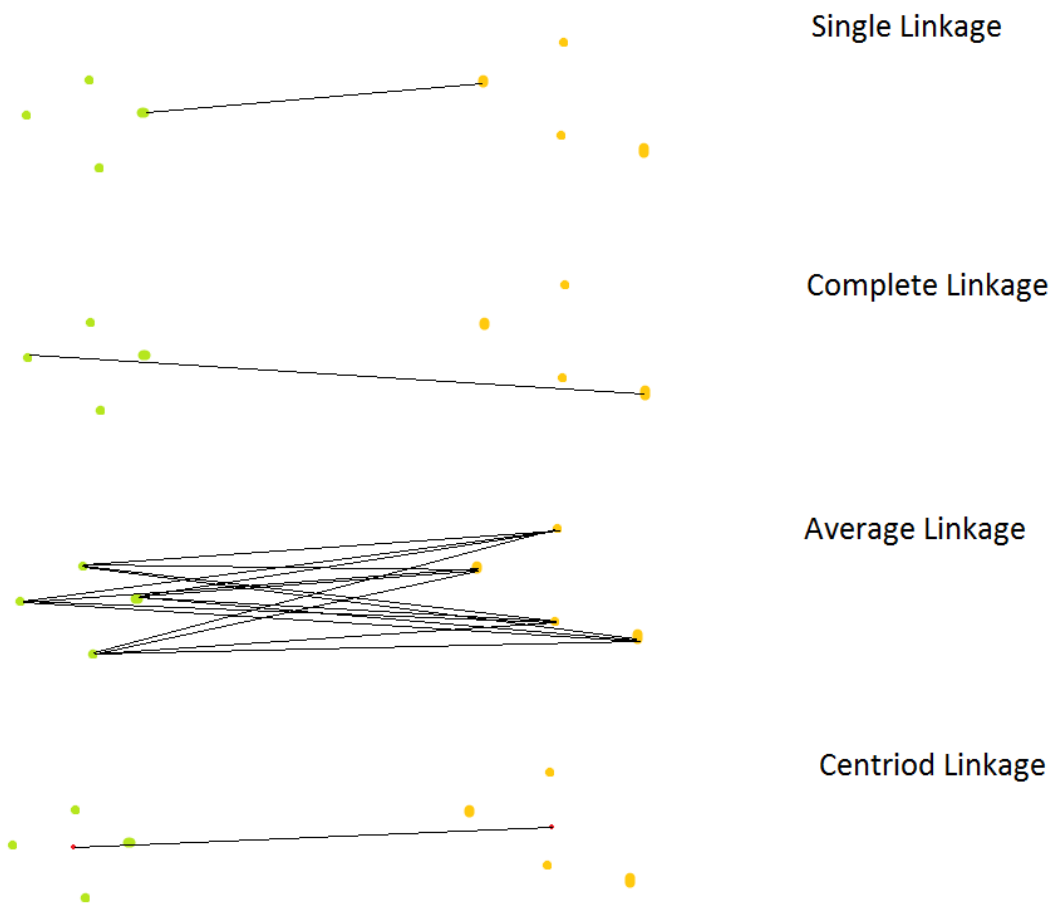


Figure 2 - Similarity measures

According to (Sibson, 1973) single linkage is one of the oldest methods to measure similarity. It was first proposed by Sneath (1957). In agglomerative clustering methods, single linkage groups the observations/clusters into new clusters that are the most similar. Similarity is defined as the minimal distance between two observations/clusters. Where the distance between two segments is defined as the minimal distance of their observations. This is the distance between the two most similar observations from different clusters. The property of only requiring one link for a combination of groups is giving the method its name (Gower & Ross, 1969). An exemplary illustration of single linkage can be viewed in Figure 2 - Similarity measures.

Complete linkage was proposed by King (1967). It describes an approach that is similar to the single linkage method, but uses the farthest observations of two clusters, instead of the nearest. The

advantage over single linkage is, that it avoids the so called chaining. Cho, Lee, and Lee (2009) describe this phenomenon as the occurrence of a chain of elements that can be extended for long distances without regard of the overall shape of the emerging cluster. An exemplary illustration of complete linkage can be viewed in Figure 2 - Similarity measures.

The average linkage, or UPGMA (Unweighted Pair Group Method with Arithmetic Mean), is another implementation of a similarity measure for hierarchical clustering (Dawyndt, De Meyer, & De Baets, 2006). Instead of the nearest, or furthest observations from two clusters, it uses the average pairwise distance between all observations of two clusters to define their distance. An exemplary illustration of average linkage can be viewed in Figure 2 - Similarity measures.

"The centroid linkage distance reflects the distance between the centers of two clusters" (Bolshakova & Azuaje, 2003, p.828). As the average linkage, also the centroid linkage does not define the distance between two clusters as the distance between only one pair of observations. It consists of a two level procedure, in which, at the first level, the centroids of each cluster are computed. And in the second step the distances between those centroids are computed. The two closest clusters are joined accordingly. An exemplary illustration of centroid linkage can be viewed in Figure 2 - Similarity measures.

As defined by Gowda and Krishna (1977), the mutual neighbor distance takes into account the environment of the observations. The distance is defined as $NN(x_i,x_j)+NN(x_j,x_i)$ where NN means neighbor number. If all observations are ordered according to their distance to x_i in ascending order. The neighbor number is defined as the rank of x_j. So if x_j is the nearest neighbor of x_i the neighbor number equals to one. Note that, if x_i is set remotely, x_j might be its nearest neighbor, but not vice-versa. In order to include information about the neighbor number of x_i to x_j and of x_j to x_i, the sum of both ranks is taken and expressed as the mutual neighbor distance.

Ward's Method, proposed by Ward (1963), focuses on the variation of the clusters. Merges are not performed based on the distance between observations of two clusters, or their centroids. In this case the chosen indicator is the overall variance from observations around the centroid they have been assigned to. The method compares the total variance of all possible merges, and choses the one that increases the total variance of the population the least. Note that the total variance is always increasing when two clusters are joined together. This happens because clusters become sparser, until all observations are included in a single cluster.

3.2.3. Metrics

Of the presented similarity measures, the single linkage, the complete linkage, the average linkage, the centroid linkage, and the mutual neighbor distance use metrics to evaluate the similarity of observations/clusters. These metrics are a parameter of the similarity measurement. A metric needs to fulfill four conditions (Arkhangel'skii & Pontryagin, 1990)*:*

1. Non-Negativity: A distance cannot be negative, and is always expressed by a positive number.

2. Identity of indiscernibles: When two points are equal, their distance is zero.

3. Symmetry: The distance from a point X to another point Y is the same, as the distance form point Y to point X.

4. Subadditivity, or triangle inequality: The direct distance between two points X and Y must be smaller or equal to the distance from X over Z to Y.

All metrics need to fulfil these conditions. The following section provides an overview of four popular metrics (Deza & Deza, 2009):

1. Euclidean metric: $(\sum_{i=1}^{k}(x_i - y_i)^2)^{\frac{1}{2}}$ Where X and Y are two k-dimensional observations.

2. City-block metric: $\sum_{i=1}^{k}|x_i - y_i|$ It is the sum over the positive differences between X and Y in each dimension.

3. Minkowski metric: $(\sum_{i=1}^{k}(x_i - y_i)^p)^{\frac{1}{p}}$ It is a generalized form of the Euclidean metric, with p=2. The city block metric is a special case of the Minkowski metric with p=1, that assumes all differences as positive.

4. Mahalanobis distance: $((\det A)^{\frac{1}{n}}(x-y)A^{-1}(x-y)^T)^{\frac{1}{2}}$ A denotes the correlation matrix of the data set, and X and Y are the row vectors that should be compared to one another. The Mahalanobis distance is the sum of the dimension wise difference of the observations X and Y. This difference is calculated in a space, which transforms the right angled axis of the Euclidean space to a Mahalanobis space. In the Mahalanobis space the angle between two axis is anti-proportional to their correlation, obtained by their correlation matrix (Davidson, 2002).

3.2.4. Divisive Algorithms

Since the previous sections dealt mainly with agglomerative approaches, thies section is dedicated to diviceve methods. Divisive algorithms start with all observations in one single cluster and then divide the population into the most distinct sub groups.

In this section two examples of divisive algorithms should be explained. The first one is the minimal spanning tree.

As Kou, Markowsky, & Berman (1981) explain, a spanning tree is a construct of edges (links) that connect all observations of a data set. These links are associated with a distance measure. The tree with the minimum overall distance (minimal span) of all possibilities, while connecting all observations, is called the minimal spanning tree (MST). A discussion on how to find the minimal spanning tree is not part of this work.

Zahn (1970) uses these minimal spanning trees for a divisive algorithm. The MST can be seen as the first cluster that includes all observations. If one of the links is taken out, the result contains two separate trees. The algorithm proposed by Zahn (1970) eliminates the longest link at each iteration. Consequently an additional cluster is created in each iteration until none of the observations are connected anymore. The researcher can now decide on the optimal number of clusters by comparing the solutions of different iterations.

The second divisive algorithm that should be presented is bisecting k-means. The bisecting k-means clustering algorithm tries to combine the time efficiency of the k-means algorithm with the high quality results from hierarchical clustering (Steinbach, Karypis, & Kumar, 2000). For now it is enough to know that the k-means algorithm divides a chosen cluster into two sub clusters. A detailed discussion of the k-means algorithm is given in K-Means Algorithm. The algorithm starts with all observations in a single cluster. This cluster is split into two sub clusters by using the standard k-means algorithm. In order to obtain the best split of a chosen cluster, the k-means algorithm is initialized a number of times. The iteration that results with the lowest average pairwise distance of within the sub clusters is chosen to perform the split. This procedure is repeated until the predetermined number of clusters is reached. Steinbach et al. (2000) note, that there is only little difference between the possible methods of selecting a cluster to split in each iteration. These methods can be based on the techniques that are introduced in Internal Validation Techniques.

3.3. Partitional Clustering

Partitional clustering, is the counterpart to hierarchical clustering. The name signifies that the observations are not grouped according to a hierarchy (for example a hierarchy of distances). In Partitional clustering observations are grouped all at the same time. The most popular representative of this class is the k-means algorithm. Due to its popularity the focus of this chapter will be on the k-means algorithm.

3.3.1. K-Means Algorithm

This chapter should give an overview of the working method of the k-means algorithm. Important properties and problems are explained in the following sections.

The algorithm, as defined by MacQueen (1967) is an iterative process that starts with a predefined number of initial cluster centers (seeds), that corresponds to the number of clusters that should be defined. After the initialization of the seeds, observations are assigned to their closest seed (taking into account all dimensions). Then the mean of the observations that have been assigned to the same group is computed. This mean is called centroid. In each subsequent iteration the centroids of the previous iteration are treated as seeds. This procedure iterates until one of the stopping criteria is met. The stopping criteria are:

- The maximum number of iterations has been reached
- The centroids converge: The difference in the position of all centroids of the current iteration is similar enough to their respective position in the previous iteration. So the distance between observations and centroids can be considered minimized.

All observations that are assigned to the same centroid in the final iteration belong to one cluster.

3.3.2. Advantages and Disadvantages of K-Means

Due to its wide range of application fields, it is necessary to obtain a more detailed overview of the capabilities of k-means. This section studies the main strength and weaknesses of the k-means algorithm.

(MacQueen, 1967) points out that k-means is a particularly easy algorithm, which makes it preferable over other methods. This can also be an advantage when explaining the method to non-specialists.

Its greatest strength is probably its low computational cost in comparison to hierarchical clustering. Steinbach et al. (2000) emphasize the linear relation of time and complexity of data set. This computational efficiency makes it the technique of choice for many problems with large data sets.

Although these properties are desirable, there are a number of drawbacks of the k-means algorithm. One of which is, that the number of clusters (k) needs to be pre-defined. Especially in comparison with hierarchical methods this is a disadvantage. When there is no prior knowledge about the number of clusters, a number of heuristics can be applied (see Number of Clusters). Nevertheless it is necessary to note, that these methods are only heuristics, and cannot be seen as definitive solution for this problem. In conclusion it is always necessary to run k-means with different k, in order to find the best suitable model. When doing so there are two major problems. The first is, that it is difficult to compare models with a different number of clusters. Even though some of the measures in Cluster validation & Similarity Measures penalize models that are using a higher number of clusters, because they naturally lead to a lower sum of square error, it is not said that the penalty fits each problem and number of clusters in the same way. The second problem is, that even if k-means is initialized with different k, there cannot be certainty that the optimal number of clusters has been found within these iterations.

Another problem with rerunning k-means is, that it minimizes the distance between observations and centroids depending on an initial solution, and so converges towards a local optimum. The relation between the initialization, and the result is deterministic. Once it has found a local optimum (smallest distance between observations and centroids), a termination condition is met. To further explore the search space, multiple initializations are necessary (P.S. Bradley, U.M. Fayyad, 1998). A number of different initialization methods, that help to explore the search space to the largest extend, is described in Initializations for K-Means. The solutions of the different initializations with the same number of clusters can be compared using the measures provided in Cluster validation & Similarity Measures.

The inability of separating overlapping clusters correctly is another drawback of many clustering algorithms, including k-means. In the case of overlapping true clusters, k-means will allocate each observation to its closest centroid, and so, create an artificial threshold by which observations are grouped with either of the clusters (Cleuziou, 2008).

Another drawback of k-means is, that, in its original implementation, it belongs to the class of hard clustering algorithms (Hard vs. Fuzzy) and as such cannot estimate the quality of the membership of an observation to a cluster (Davidson, 2002). This means that, on one hand, observations that could have been assigned to two clusters with almost the same likelihood, are only assigned to one (as described above). On the other hand this means that outliers, which don't belong to any cluster, are assigned to the closest centroid. Both of these properties bias the centroids, which are calculated as the mean of their allocated observations. Outliers bias the centroids by pulling them in their direction, and observations that are located in in the overlapping area of two true clusters bias them to a direction further apart from each other. One can imagine this behavior by picturing two slightly overlapping Gaussian distributions. Observations that are located in the overlapping area will be assigned to the closest centroid (group-mean), not to the mean of their original distribution. This will

lower the estimated mean of the distribution with the smaller original group-mean, and raise the estimated mean of the distribution with the higher original group-mean.

K-means takes in to account the overall Euclidean distances between observations and centroids. This weights each variable equally and makes k-means sensitive to the scale of the variable (Davidson, 2002). For example, if the size of humans should be clustered into two groups, using the features height (in meters) and weight (in kilograms). It is very likely, that the effect of weight will dominate the outcome. This is, because the height typically only varies on a decimal (0.1s) level, while the weight might change on a decamal (10s) level. In this specific case it might be a solution to use centimeters instead of meters for the height measure. In general, this drawback is oftentimes counteracted with transforming all values into a range (e.g. between 0 and 1), or normalizing them to z-scores. Z-scores are the deviation from the mean of a variable, expressed in standard deviations. However, (Davidson, 2002) notes, that this transformation adds more weight to variables with a low standard deviation.

K-means considers all variables regardless of their relation. If, for example, two variables are characteristics of the same underlying factor, both will be taken into consideration, and thus, the underlying factor will be weighted with two variables. Davidson (2002) suggests two ways of dealing with that problem. The first is to perform Factor Analysis[1] in order to extract the underlying factors of all variables, and then perform k-means on the obtained factors. A disadvantage of this procedure might be the difficult interpretability of the factors (Brown, 2015). The other suggested way of tackling this issue is, to use the correlations among the variables and compute the Mahalanobis distance (see Metrics). This distance reduces the effect of highly correlated variables in the algorithm, by taking into account correlations when computing the distances between observations. In the Mahalanobis space, axis (variables) that are correlated are located closer together (angle of less than 90 degree). This reduces the effect those axes have, when distance between observations is measured.

Another problem of k-means, notes Davidson (2002), is that, in the original form, clusters are only described by their centroids. These centroids only provide the means of a cluster. They are unable to provide a full view of the cluster, since they are hiding information like the standard deviation and the number of observations assigned to the cluster.

As stated above, clustering is applied in a variety of different fields (Fields of Research). Therefore the applications are numerous and the type of data varies. Another downside of k-means is, that in its original implementation it is only able to detect spherical (or hyper-spherical) shaped clusters. (Fred & Jain, 2002) identify this behavior as one of the major weaknesses of the k-means algorithm.

3.3.3. Outlier Detection

As stated in the section above (Advantages and Disadvantages of K-Means) outlier are a main concern when using k-means. When detected, outlier can be deleted from the data set or, preferably, transformed into a usable value, so that they do not bias the result of the k-means

[1] Factor Analysis is a dimensional reduction technique. It describes underlying factors by investigating the correlations among variables. Further it shows the relation of those factors and the original variables to provide a description of the factors. A discussion of factor analysis is not subject of this work.

algorithm. For a start the two classical definitions of outlier are given in Table 2 - Classical outlier definitions.

Author(s)	Definition
(Hawkins, 1980, p.1)	"An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism"
(Barnett, & Lewis, 1994, p.175)	"an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"

Table 2 - Classical outlier definitions

In this section techniques to detect outliers are presented. In general these techniques can be divided according to different features. In this work two of these features are described. The first feature identifies methods with underlying assumptions like, for example, that the data was generated by a certain type of distribution. These methods are called parametric. Their counterparts, the non-parametric methods, do not assume a specific type of underlying distribution.

The second feature by with outlier detection methods can be distinguished is by how many variables are considered when detecting an outlier. Univariate methods only consider the value of a single variable, while multivariate methods consider multiple dimensions to identify an outlier (Ben-gal, 2005).

The first method that is presented is called the percentile method. Using this method, all observations are ordered according to the dimension that should be scanned for outliers. Then a value above which points are considered outliers, and a value below which points are considered outliers are defined. Depending on the variable it can be appropriate to set zero as the lower limit. Usually both values are defined using percentiles of the dimension of interest. If an observation is spotted as an outlier its value can simply be set to the respective boundary. This method does not assume a specific kind of distribution, and thus is considered non-parametric.

The main advantage of this method is its low computational cost.

Disadvantages of this method are that the threshold above (or below) which an observation is an outlier must be determined manually. Also this is a univariate technique, and so only considers one dimension at a time. Another drawback is, that outliers are only determined by their rank of a specific variable, comparing with the rank of other observations. This method rigorously determines a chosen percentage of a sample as outliers, without caring about their localization comparing with the rest of the set.

The second method presented uses the standard deviation of a variable. This method takes into consideration the relative location of an observation to the mean of a variable. It is similar to the percentile method in the sense that observations are ordered according to their value of only one variable (univariate), and the top (or/and bottom) observations are identified as outliers. In contrast to the method described above, the threshold, based on which an observation is determined an outlier, is given as a multiple of the standard deviation of the respective variable. Using this

procedure the number of outliers is not provided manually, but identified by the characteristic of the data.

A drawback is that this method assumes the variable to be normal distributed, which in reality is not always the case (Acuna & Rodriguez, 2004). This property makes it part of the parametric approaches.

Bay and Schwabacher (2003) proposed the so called simple distance based outlier detection. This method aims to identify the predefined number of outliers in a specific subset of the population. The procedure computes the average distance for all observations in the subset to their nearest neighbors. The size of the neighborhood, and the size of the subset are parameters of the algorithm. The neighborhood of an example can be chosen from the whole population. The observations with the highest average distance (the metric is also a parameter of the algorithm) to their nearest neighbors are considered outliers.

Bay & Schwabacher (2003) note that they do not provide any guideline for the right choice of the parameters, which also includes the size of the investigated subset. Desirable properties of this method are that it is a multivariate technique that does not make any assumption about the underlying distribution. On the other side it only identifies a pre-defined number of outliers, which might be difficult to guess.

The last presented method for detecting outliers is the local outlier factor (LOF). This method, proposed by Breunig, Kriegel, Ng, & Sander (2000) uses the concept of neighborhood. An integer k defines the number of neighbors. All observations at the same distance from the respective observation as the furthest neighbor are also called neighbors of this observation. As an effect of this the number of neighbors can vary if there are multiple neighbors at the same distance. The distance of the furthest neighbor is called k-distance. In the next step the so called reachability-distance is calculated. The reachability-distance between two objects is their distance, according to a metric, but at least the k-distance. So all neighbors are assumed to be at the k-distance. Note that this definition does not qualify to be a metric, because it is not symmetric. The distance from A to B can be different than the distance from B to A, if B is a neighbor of A, but not vice versa. This property is important when calculating the density of an observation. The so called local reachability density of a point is calculated as the inverse average reachability-distance from all its neighbors to the point.

The local outlier factor is defined as the average ratio of the local reachability densities of the neighbors and the point of interest. A value around one signifies that the investigated observation is approximately as dense as its neighbors, while a value that is much greater than one indicates that the investigated observation is a lot less dense than its neighbors. The latter is an indicator for an outlier.

In contrast to other methods, the proposed procedure does not strictly tell if an observations is an outlier, it rather assigns a degree of being an outlier to each observation, based on which the researcher can decide.

Further advantages of this method are, that multiple dimension are considered at the same time to compute the local outlier factor, and that it defines an outlier based on the properties of the local environment (neighbors), not on the properties of the whole population. The latter improves outlier

detection of different distributions within the same data set. While doing so it does not make specific assumptions about the underlying distribution, and thus is part of the non-parametric methods. Also the masking-effect[2] can be tuned with the size of the neighborhood. If a greater neighborhood is used, outliers are less likely to mask each other.

3.3.4. Initializations for K-Means

As discussed above (Advantages and Disadvantages of K-Means) the k-means algorithm converges towards a local optimum, and the relation between initialization and result is deterministic. Therefore the initial centroids (seeds) have a major impact on the solution that will be found. This section will give an overview of different initialization techniques, which can be used in order to explore the search space in the most effective way possible.

The initialization methods are grouped by their relation of time performance to the data set's complexity. Complexity is synonymous with the number of observations in the data set. At first the initialization techniques that show a linear relation of complexity and performance are presented.

A (asymptotical) linear relation between complexity of the data set and the performance of the initialization means, that if the (already high) complexity of the data set increases by a certain proportion, the time it takes to perform the seed initialization increases by a similar proportion. This is, when the highest order of the time-complexity-relation is one.

- Forgy (1965) proposes a method, that assigns all observations to one of the k clusters at random, and then calculates the mean of each cluster. These means are used as initial seeds.

- The method proposed by Jancey (1966) is to randomly generate seeds. Celebi, Kingravi, & Vela (2013) note that, if the space is not filled with observations, this procedure might lead to empty clusters, when the generated values are too far apart from the data points.

- MacQueen (1967) suggested two methods of initialization. The first one assumes the first k observations to be the initial cluster centers, which has the drawback of being sensitive to the sorting of the data set. To overcome this disadvantage, the second method selects k randomly chosen observations to be the initial solutions. (Celebi et al., 2013) point out, that this allocates a higher probability for the initial cluster centers to be in an area of dense population, however there is no mechanism that prevents the chosen seeds to be outliers.

- Ball & Hall (1967) take the overall centroid as a first seed, and the select arbitrary observations, that have a minimal distance T to all previously chosen seed, until k seeds are defined. In many cases it is difficult to define a reasonable distance T, however it helps to provide a well spread initial solution. Tou and Gonzales (1974) use a similar method. The only difference is, that they use the first observation as the first seed, not the overall centroid.

[2] "It is said that an outlier masks a second one close by if the latter can be considered outlier by himself, but not any more if it is considered along with the first one. Equivalently after the deletion of one outlier other instance may emerge as an outlier. Masking occurs when a group of outlying points skews the mean and covariance estimates toward it, and the resulting distance of the outlying point from the mean is small" (Acuna & Rodriguez, 2004, p.7)

- Gonzalez (1985) proposes an initialization method, which spreads out the cluster centers. It starts with one arbitrarily chosen observation, called the head of the cluster. Then one of the observations that is the furthest from this head is defined to be the head of the next cluster. Observations are assigned to one of the clusters based on which head is closest to them. Until the number of clusters is reached, the observation, that is the furthest from its head is defined to be the head of the next cluster, and all observations that are closer to the new head, than to their actual cluster heads, are assigned to the new cluster.

- Al-Daoud (1996) proposes a density based approach. In a first step the space is divided into M uniform hypercubes, which do not overlap. The total number of centroids is assigned to the hypercubes, proportionally to their respective number of observations. Each hypercube's centroids are chose at random. Celebi et al. (2013) identify a disadvantage in the difficult choice of the number of hypercubes.

- Bradley & Fayyad (1998) also suggest a more stage procedure. At first the data set is partitioned into a number of randomly selected subsets. For each of those subsets k-means with the wanted number of clusters is applied. It should be initialized, using randomly chosen observations of the subset as seeds (second method proposed by (MacQueen, 1967), as described above). The resulting centroids from each subset are joined together into one superset. This set of centroids is then clustered by k-means a number of times. Each time, the centroids of another subset are chosen to be the seeds for the k-means algorithm in the superset. The seeds of the iteration with the smallest sum of square error are chosen to be the seeds for the k-means run on the whole population.

- Su & Dy (2007) suggest an initialization method that is using principal component analysis (PCA)[3]. The data set is divided according to a divisive algorithm. It uses the first principal component of the covariance matrix that passes through the centroid of the data set. An orthogonal hyperplane is laid through the principal component, intersecting at the centroid. This plane divides the data set into two subsets. This procedure is repeated for the cluster with the greatest sum of square error, until the predefined number of clusters has been reached. The centroids of the resulting clusters serve as seeds for the k-means algorithm.

- Another initialization method that uses principal component analysis is widely used. It calculates the first two principal componenty, and initializes k equally distant seeds on the grid of the plane spanned by the first two principal components.

- Onoda, Sakai, & Yamada (2012) propose an initialization method based on Independent Component Analysis (ICA)[4]. An observations is chosen to be a seed if it has the minimum cosine distance[5] from one of the independent components.

[3] Principal Component Analysis (PCA) is a technique that rotates the axis of a data set, so that a great proportion if the intrinsic information can be expressed with less dimensions. PCA projects the observations on the principal components of their covariance matrix. A principal component is an eigenvector multiplied with its eigenvalue. The principal components are ordered in descending order by the amount of variance they account for. In this algorithm, only the first principal component (accounting for the greatest proportion of variance) is considered. A discussion of Principal Component Analysis is not part of this work.

[4] Independent Component Analysis (ICA) is a technique that calculates new axis from the observations in a data set. The property that these new axis should have is, that they are independent. While the new axis are

At next initialization methods are presented, whose execution cost is loglinear to the complexity of the data set. Loglinear means, that the cost of execution is linear to a logarithm of the number of observations of the data set.

- Hartigan & Wong (1979) propose a method that first sorts all observations according to their distance from the overall centroid. This sorting procedure is the main cost driver for the initialization. Observations are chosen to be seeds by applying the formula (1 + (L – 1) * (M/K)) to their rank in the sorted data set. Where L is the number of the cluster (between 1 and the number of clusters), M is the total number of observations in the data set, and K is the total number of clusters. (Hartigan & Wong, 1979) point out that this initialization method guarantees, that none of the initial clusters will be empty.

- Al-Daoud (2005) presents a procedure that only takes one dimension in to consideration when initializing the seeds. This dimension is said to be the one with the greatest variance. All observations are ordered according to their values in that dimension, and then divided into k groups. The means of each group are chosen to be the initial centroids for the k-means algorithm.

- Redmond & Heneghan (2007) suggest a method that uses the kd-tree[6] procedure. The threshold to create leaf buckets is set to the number of observations divided by ten times the number of clusters wanted, so that each cluster has approximately ten buckets. For each leaf bucket, its density is computed as the division of number of observations by its volume. The volume can be simply obtained by multiplication of the length of all dimensions, where the length is the difference between the respective minimum and the respective maximum of a dimension in the leaf bucket. The density is then associated with the mean of the leaf bucket. The means are the potential seeds of the k-means algorithm, but only k should be chosen. Seeds should have two properties. They should represent a densely populated leaf bucket, and they should be well spread in the space. In order to optimize both constraints Redmond & Heneghan (2007) compute the distance between the leaf bucket means as the Euclidian distances weighted by their density. The first seed is the mean of the most dense leaf bucket. The second seed is chosen by the above described distance in descending order. The third and all subsequent seeds, until k is reached, are chosen using the distance to the closest seed (that has already been chosen), weighted by the density of the leaf bucket they represent.

---

independent of each other, they should still be correlated with the original data, in order to provide maximum interpretability.

[5] The cosine distance is a distance measure for two vectors. It is defined as the cosine of the angle between two vectors. If the observations of the two vectors are given the cosine distance can be computed as $\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$ where A_i and B_i are observations of the vectors A and B.

[6] A kd-tree (k-dimensional-tree) is a top-down hierarchical scheme of partitioning data. It splits a data set alongside its longest dimension. The longest dimension is defined as the maximal difference between minimum and maximum of a dimension across all observations in a set (or subset). The split is performed at the median of the longest dimension. In order to save computational cost, the mean might be used instead of the median. This partition will be performed on each resulting bucket, until a leaf bucket is created. A leave bucket is a bucket, which contains a number of observations that is lower than a predefined threshold. The kd-tree is created when all partitions are leaf buckets (Redmond & Heneghan, 2007).

Two problems are noted by Redmond & Heneghan (2007). First, the density for less dense leaf buckets is underestimated. So the distance computation is dominated by leaf buckets with a higher density. This drawback can be overcome by using the inverted rank density, instead of the absolute amount. (It needs to be inverted, so that the highest density is represented by the highest rank).

The second identified problem is, that outliers have impact on the density and the distance of a leaf bucket. In general outliers are an issue for k-means (Advantages and Disadvantages of K-Means). According to Redmond & Heneghan (2007) the effect of outliers on the distance causes the algorithm to choose observations from leaf buckets containing outliers as seeds, and the effect of the outliers on the mean of the leaf bucket would bias the true seed. They overcome this issue by assuming that outliers are mainly found in less dense leaf buckets. So they only consider 80% of the densest leaf bucket means to be potential seeds.

- Hasan, Chaoji, Salem, & Zaki (2009) propose a method that aims to avoid choosing outliers for seeding. They define six criteria that should be satisfied by their initialization method:

    1. It should be computationally inexpensive
    2. It should be parameter free, and if it uses parameters, they should not have a major influence on the seeds
    3. It should be deterministic and not sensitive to the order of the data set
    4. It should be not sensitive to outliers
    5. It should be robust to clusters of different scale and density
    6. It should be robust to different cluster sizes and skewness

  The approach is based on distances between points, and the local outlier factor, which has been introduced in Local Outlier Factor. The initialization works as follows: At first a randomly chosen observation is defined as seed number zero. All observations are ordered in descending order according to their distance to that observation. The furthest observation is the next potential seed, but it is only considered if it is not an outlier (LOF close to one). If the furthest observation is identified as an outlier, the next observation in descending order is considered and so on. In order to save computational recourses the LOF is only calculated for the next potential seed, not for the whole data set. This procedure will finally find a second seed. Once it has been found, all remaining observations are sorted in descending order according to their minimum distance to their closest of the already defined seeds. This procedure repeats until k + 1 seeds have been found. In a final step the initial seed zero is dropped from the set of seeds, because it was chosen randomly. Hasan et al. (2009) point out that the computation of the distances to the already chosen seeds is not an additional cost of computation, because these distances would be calculated by the k-means algorithm in the first iteration anyway, and now can be used while executing the algorithm. Hasan et al. (2009) highlight that this initialization algorithm satisfies all criteria that are listed above.

The last part of this chapter presents initialization methods whose performance quadratically decreases with increasing complexity of the data set. Comparing with the other initialization methods, the ones that are presented here have a worse performance when the size of the data set increases.

- Lance (1967) proposes the usage of hierarchical algorithms to choose the location of seeds for k-means and leave the choice of similarity measure and metric to the user, because the right choice depends on the properties of the data set. The performance is determined by the performance of the applied hierarchical clustering algorithm. As stated in section Advantages and Disadvantages of K-Means, these algorithms generally take a large amount of time to process complex data sets. So the major drawback of this method is that it takes away k-mean's advantage over hierarchical approaches, the fast execution.

- Astrahan (1970) defines his method based on two distance parameters. The first distance parameter defines the radius of an observation, which includes all its neighbors. The density of an observation is defined as the number of neighbors within that radius. All observations are sorted descendingly by their density. The densest observation is chosen to be the first potential seed. All remaining observations are defined as potential seeds in descending order of density, if they have at least at distance equal to the second parameter from their closest potential seed. This procedure is repeated until no further potential seed can be found. In case the number of potential seeds exceeds the number of wanted clusters, a hierarchical clustering algorithm is performed on all potential seeds. This hierarchical clustering algorithm is performed until the exact number of wanted seeds is reflected by the number of groups created by the algorithm. The centers of the resulting groups serve as initial seeds for k-means. Celebi et al. (2013) note that the major drawback of this method are the two distance parameters that have major influence on the resulting seeds and the standard implementation does not provide a heuristic for an educated guess to set them.

- Kaufman, & Rousseeuw (1990) define another method with a quadratic time-complexity relation. The algorithm choses the population mean as the first seed, and then the observation that reduces the sum of square error the most for each iteration. At each iteration the distances with the potentially new cluster seed have to be computed. This makes the procedure having a not desirable performance.

- Cao, Liang, & Jiang (2009) create a method that uses cohesion[7] and coupling[8]. These two concepts serve different objectives. Cohesion is used as an indicator of centrally of an observation in its neighborhood, while coupling indicates the distance between two neighborhoods. Both concepts depend on the neighborhood of an observation. In order to define an appropriate neighborhood Cao et al. (2009) normalize all variables into a range

[7] The cohesion of an observation is defied by Cao et al. (2009) as the number of observations who's neighborhood is a sub set of the neighborhood of the investigated observation. The number of observations for whom that applies is divided by the number of observations that share at least one element of their neighborhood with the neighborhood of the observation of interest. The higher the cohesion of an observation is, the more centrally it is located in its neighborhood.

[8] Cao et al. (2009) define as coupling of two observations the number of elements that are included in both observation's neighborhoods, divided by the union of their neighborhoods. The higher coupling is, the more similar are the two observations, because they share a larger fraction of their neighborhood.

between zero and one and compute the average distance between all observations (where the metric is subject to the researcher). They define this average distance as the neighborhood radius. All observations within that radius of an observation are called neighbors of the observation. Based on these neighbors, cohesion and coupling are calculated. The first cluster center is the observation with the maximal cohesion. All following initial centers are chosen in descending order by their cohesion. The second criteria they must satisfy is a coupling with all previously selected seeds, which is lower than the average distance between all observations. This procedure is stopped when the wanted number of seeds is reached.

3.3.5. Number of Clusters

As mentioned above (Advantages and Disadvantages of K-Means), one of the major drawbacks of the k-means algorithm against hierarchical approaches is that the number of clusters needs to be pre-defined in k-means. This chapter provides some heuristics that help to choose a reasonable number of clusters.

As a first guideline Pham, Dimov, & Nguyen (2005) state that the number of clusters should be large enough to reflect the characteristics of the data set, but at the same time significantly smaller than the number of observations. The choice of the best number of clusters is similar to the question of whether the result of clustering is good or not. Therefore some of the techniques described in Internal Validation Techniques can be applied to determine reasonable k, as well. Oftentimes the number of clusters is defined by the business case, or by knowledge of the data source. All other cases need to use a heuristics. This chapter explains procedures that are customized for the purpose of evaluating the best number of clusters.

The first technique that is presented is X-Means. This method provides the applicant with the freedom of supplying a range of possible k, instead of a fix value. The idea of Pelleg & Moore (2000) is that a regular k-means algorithm is run with the lower boundary of the number of clusters. When the algorithm converges all final centroids (parents) are split into two children centroids and are moved apart from each other (proportional to the size of their cluster). In the next step local k-means are run within each cluster, using the two generated centroids as seeds. When this run has finished clusters are evaluated locally using the BID index (see: BIC-Index). If the solution with only the parent centroid scores a higher BIC-value than the children, it is kept. Otherwise the children are preferred over their parents. Then a new iteration of k-means is run on the whole population. This time it uses the suggested number of seeds from the previous run. This procedure stops when either the upper boundary of clusters has been reached, of the BIC-score does not improve anymore when adding new seeds.

Ray & Turi (1999) propose another procedure that uses a validity measure. This validity measure is computed as the ratio of the intra-cluster distance to the inter-cluster distance. Low values of this measure indicate good solutions, because both, a low intra-cluster distance and a high inter-cluster distance are desirable properties of a good clustering solution.

Initially all observations start in the same cluster in the validity is computed. In each iteration the cluster with the greatest variance is re-clustered using k-means with two seeds. In each iteration the resulting validity is obtained. The algorithm stops when a maximum number of k has been reached.

All observed values for the validity measure are compered, and the iteration that gave the lowest score determines the number of clusters.

Note that the validity measure can also serve as an indicator to choose better solutions over worse, when both options have the same number of clusters.

The cubic clustering criterion (CCC) is another technique of determining the number of clusters of a data set. It was proposed by Sarle (1983). It is based on the idea that good clustering solutions are able to explain a greater proportion of the overall variance, than a classifier that assumes that the data is equally distributed in the space. It is calculated using the formula:

$$\ln\left(\frac{1-E(R^2)}{1-R^2}\right)*K$$

$E(R^2)$ is the expected value of R^2 under the assumption that the data is distributed equally in space It is obtained by the R^2 of hypercubes in which the space is divided. The number of hypercubes corresponds to the number of clusters to be checked for validity. In the formula R^2 denotes the actually observed R^2 by the obtained clusters, and K is a factor stabilize the variance Sarle (1983). K will not be further explained in this work since the main outcome of the CCC is determined by the previous term. Greater values of the cubic clustering criterion indicate a better grouping, because the centroids of the clusters seem to be more descriptive for the population.

However it should be noted that R^2 naturally raises with increasing number of clusters, and, therefore the cubic clustering criterion cannot be interpreted on its own. It is necessary to evaluate the context by computing the CCC for different numbers of clusters. The CCC will increase with the number of clusters, but the effect of an additional cluster will asymptotically decrease. It is subject to the specific problem to determine a rational number of clusters, which can be supported by the CCC.

3.4. Cluster validation & Similarity Measures

After different clustering algorithms have been introduced, and a number of initialization methods and distance measures have been explained. Now the clustering can be performed. Typically, in stochastic techniques, for each clustering problem many iterations are necessary until a satisfying solution has been found. This section gives conclusion about how a good solution can be distinguished from a bad solution.
As stated above (Definition), clustering seeks to identify groups in data where observations are similar insight the group, and different when compared across groups. This raises the question of how similarity should be measured. Although researchers have proposed a number of different similarity measures, a single best method cannot be defined. “Evaluating the quality of clustering results is still a challenge in recent research.” (Färber et al., 2010, p.1). In general the clusters should have a high intra-cluster similarity (low intra-cluster distances) and low inter-cluster similarity (high inter-cluster distances). So one can say, the lower the intra-cluster-distance, the higher the inter-cluster-distance, the better is the validity of the segmentation.
The choice of a similarity measure/metric depends on the research question and the underlying type of data.
One of the most commonly used similarity measures is the standard deviation. Here the intra-cluster standard deviation and inter-cluster standard deviation are compared. A result is considered good, if

it shows a low intra-cluster standard deviation and a high inter-cluster standard deviation. This concept is picked up by some of the techniques below.

The two main branches of cluster validation are internal and external validation (Rendón, Abundez, Arizmendi, & Quiroz, 2011). Internal validation is based on information intrinsic to the data, while external information considers previous knowledge about the data (for example the true classification or distribution where observations were generated from).

The following sections will give an overview of popular cluster validation techniques.

3.4.1. External Validation Techniques

For the external validation four indexes will be introduced. These indexes are the F-measure, entropy, purity, and the NMI-Measure. According to Rendón et al. (2011) these are the most frequently referenced in literature.

Following Powers (2011) the F-measure is based on two measures, Recall and Precision. In order to explain these measures, it is necessary to describe the concept of true positive, false positive, true negative, and false negative observations. These concepts are based on whether an observation is an event, or a non-event, and whether it was classified as an event, or a non-event. An event can be understood as the observation fulfilling some sort of criteria, like for example: Client owns a red car is an event in contrast to client does not own a red car would be a non-event. The following table translates the event/non-events to true/false positive/negative:

	Event	Non-Event
Classified as Event	True Positive (TP)	False Positive (FP)
Classified as Non Event	False Negative (FN)	True Negative (TN)

Table 3 - Translation from Event/Non-Event to True/False Positive/Negative. Correctly classified are marked in green, while incorrectly classified observations are marked in orange.

The words “true” and “false” indicate whether an observation has been classified correctly, while the words “positive” and “negative” indicate whether an observation was classifies as an event or a non-event.

Having introduced these concepts, recall and precision will be defined consequently:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Recall is the True Positive Rate, which indicates the proportion of TP over all events. It gives conclusion about what fraction of all events have been classified as events.

Precision, self-explanatorily, gives conclusion about how precise the classification is. It does so by calculating the proportion of TP over all as event classified observations.

Powers (2011) notes that neither of the measures captures information about how well the model handles negative cases. If information about negative cases should be gained, Inverse Recall and Inverse Precision can be calculated. These follow the same formulas as Precision and Recall, but exchange events with non-events. (For example: Client does not own a red car would be the event, while client owns a red car would be a non-event).

A classification is considered "good" when both measures are close to one. This is the case when there is only a minimal proportion of wrongly classified observations (orange fields in Table 3 - Translation from Event/Non-Event to True/False Positive/Negative. Correctly classified are marked in green, while incorrectly classified observations are marked in orange.). The F-measure is an attempt to combine Recall and Precision to a single indicator. Larsen & Aone (1999) calculate the F-measure as:

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The F-measure increases when Precision or Recall increase and has a maximum of one, which is reached when both, Precision and Recall are equal to one. Therefore a higher F-measeure indicates a better fit of the model.

At next the concept of purity will be introduced. As the names says, it measures how pure a cluster is. The cluster purity is obtained, by taking the most frequent class of each cluster and divide it by the cluster size. So the cluster purity is the fraction of the most frequent class of a cluster. The overall purity is calculated by the weighted average of all cluster purities. The cluster purities are weighted by the respective cluster size divided by the total number of observations.

A model is considered "good" if it is very pure (if the model reflects the labels). In general great values (close to one) indicate good solutions. Nevertheless Strehl & Ghosh (2002) note, that purity is biased and favors smaller clusters. Clusters that only consists of one observation are scored as perfectly pure. They might be perfectly pure, but a model that defines every observation in a single cluster does not provide the needed information.

A measure that is similar to purity is entropy. In clustering entropy is defined as:

$$E_j = -\sum_i p_{ij} \log_2 p_{ij}$$

Originally Shannon (1948) introduced this measure without the index j. Here it is used to clarify that Entropy is calculated for each cluster.

In each cluster the negative sum of proportions p_{ij} of each class i is needed to define the entropy. It evaluates to zero if a cluster only consists of objects of the same class, and is greater when a cluster is less pure.

In order to obtain the entropy of the whole data set, a weighted average of the entropies over all clusters is calculated. It is weighted by the number of observations in the respective cluster. Entropy displays the same behavior as purity and assumes a perfect score (zero) if a cluster consists of only one observation.

Strehl, Ghosh, & Mooney (2000) note that entropy should be preferred over purity, because it considers information about the whole distribution, and not only about the most frequent class.

A different approach is the NMI-measure. NMI stands for normalized mutual information. The motivation for this measure is that oftentimes validity indexes depend on the chosen metric, and therefore lacking the opportunity to compare results from algorithms with different metric. Strehl et al. (2000) define the mutual information as:

$$\Lambda(K,\lambda) = \frac{1}{n}\sum_{l=1}^{k}\sum_{h=1}^{g} n_l^{(h)} \frac{\log\left(\frac{n_l^{(h)} n}{\sum_{i=1}^{k} n_i^{(h)} \sum_{i=1}^{g} n_l^{(i)}}\right)}{\log(k * g)}$$

Where K is the true labeling and λ is the result from the algorithm. Suppose that k denotes the number of true clusters, and g is the number of clusters that resulted from the algorithm. Consequently k * g is the number of possible combinations of k and g. Further, n denotes the total number of observations in the data set, and $n_l^{(h)}$ stands for the number of observations that were assigned to cluster h of the algorithm and have l as a true label. The sum $\sum_{i=1}^{k} n_i^{(h)}$ stands for the number of observations that were assigned to cluster h by the algorithm, and the sum $\sum_{i=1}^{g} n_l^{(i)}$ signifies the number of observations that are truly labeled as l. Note that the base of the logarithm is not a parameter of the measure, because it does not have any effect on the result (it is reduced by the second algorithm of the same base). The term in the numerator of the log describes the ratio of observations that are in a specific combination of clusters (real label l and cluster h) over the expected number of observations under the assumption of independence. If this ratio is greater than one, it means, that more observations than expected are clustered as l and h. One signifies a match with the expected value, and a value between zero and one stands for a lower occurrence than expected. In a final step the weighted average over the log to the base k * g of these values is calculated. It is weighted by the fraction of observations with the respective combination of clusters.

For the evaluation of the NMI-measure it can be said, that a higher value indicates a better fit for the actual labeling. It is worth noting that a positive property of the NMI-measure is, that, it punishes solutions that use many clusters, by dividing by the log of number of possible combinations of clusters.

3.4.2. Internal Validation Techniques

Internal validation techniques distinguish from the external validation techniques in the need of information. By contrast, internal validation techniques do not require prior knowledge about the data, like for example a label, which indicates the true cluster of an observation (Larsen & Aone, 1999).

For the internal validation five indexes will be introduced in the following section these are the BIC, CH, DB, SIL, Dunn and NIVA indexes. It will be started with the BIC-index.

The Bayesian Information Criterion was proposed by Raftery (1986) and is calculated as

$$BIC = -2\ ln(L) + s * ln(n)$$

Where L is the likelihood that the chosen parameters generate the data that were used in the segmentation. In other words, it is the probability that the observed values occur under the assumption of the chosen model. S is the number of free parameters that are estimated. And N stands for the number of observations.

The greater the likelihood of the actual observations under the assumption that the model is correct (L), the smaller becomes the BIC-Index. It can be said that a small BIC-Index indicates a better solution than a higher BIC-Index.

The second term (s*ln(n)) increases the BIC-Index, if more parameters are taken into account or more observations are considered, this penalizes a model for complexity.

Another measure is the CH-Index. The CH-Index was proposed by Calinski and Harabasz (1974) and computes the same value as the Pseudo-F statistic. The index is calculated as

*CH = (SSB / SSW) * (N - K) / (K - 1)*

SSB denotes the overall between cluster variance, calculated as the squared Euclidian distance from the respective cluster mean to the overall mean, multiplied by the number of observations in the respective cluster and summed over all clusters. Note that it is not divided by the overall sample size.

SSW is the counterpart of SSB and stands for overall within cluster variance. It is simply calculated as squared Euclidian distance of each observation to its respective cluster mean. These values are summed up over all observations (and so over all clusters).

N stands for the number of observations in the population and K is the number of clusters.

A higher SSB value increases the CH-Index, while SSW decreases its value. In conclusion a high CH value indicates a good fit for the data, while a small CH value does the opposite. According to this interpretation the number of sample observations raises the fit of the model, without changing the model. When comparing different models for the same sample this factor becomes irrelevant.

K, the number of identified clusters, decreases the CH-Index and so, indicates a worse result. While SSB / SSW tends to become greater when the number of clusters increases, and one could be confused if the CH-Index is only raising because of a higher number of clusters, this penalty against usage of many clusters is another good property of the index.

At next the DB-Index will be introduced. The DB-Index was developed by Davies & Bouldin (1979). It is calculated as:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

σ_i and σ_j describe the average distances of all observations within the clusters i and j to their respective centroid. The number of clusters is given by n, and $d(c_i, c_j)$ is the distance between the cluster centroids of the clusters i and j. Good segmentations should have a small value in the numerator and a great value in the denominator. The maximum function selects the most similar cluster j for every chosen cluster i (indicated by close centroids and clusters that are scattered out).

The index is the global average of the similarity measures of each cluster with its most similar cluster. Since similarities across clusters should be avoided, smaller values of the DB-Index indicate a better result.

Another concept for cluster validation is the silhouette-index. The silhouette technique is described by Rousseeuw (1987). It computes the similarity of each observation to its assigned cluster and compares it to the similarity to the other clusters. The formula of the Silhouette Index is:

$$s(i) = \frac{(b(i) - a(i))}{Max\{a(i), b(i)\}}$$

In this case i denotes an observation. Every observation has an assigned value of the Silhouette Index. This value describes how certain an observation is a member of the cluster it has been assigned to. In order to calculate this, b(i) is computed as the minimum of the average distances the observation has to all observations of another cluster, compared across all clusters, but the cluster i has been assigned to. In other words, i has an average distance to each of the other clusters. The minimum of those average distances defines b(i). a(i), however, is the average distance to the observations in the same cluster as i.

Given the formula above, s(i) is defined as smaller than one and greater than negative one. A Silhouette Index that is close to one indicates a well matched observation, because the average distance to the closest cluster is much greater than the average distance to observations within the same cluster. If the Silhouette Index is close to negative one, the observation seems to be not a good fit for the cluster (closer to observations from other cluster, than to observations of the same cluster), and might actually belong to another cluster. If the Silhouette Index is close to zero, which means that the average distance to observations of the neighboring cluster is the same as the average distance to observations of the same cluster, the observation is at the border between those two clusters.

One of the advantages of this method is that a value is calculated for every observation of the entire data set. Using this information, the researcher is able to gain information about groups of well clustered and groups of badly clustered observation. In order to obtain a single measure for the whole population, the average of Silhouette Indexes over all observations can be calculated.

Another characteristic of this method is that it can be interpreted graphically. The observations within each cluster are sorted in descending order by their Silhouette Index values and displayed as a bar chart. A cluster that only displays long bars can be considered a strong cluster, whereas a cluster of smaller or negative bars indicates a weak cluster.

One of the most popular indexes is the Dunn Index. It was presented in Dunn (1973). It is calculated as:

$$Dunn = \frac{min_{1 \le i \le j \le n} d(i,j)}{max_{1 \le k \le n} d'(k)}$$

The term in the numerator gives conclusion about the minimal inter-cluster distance across all clusters. The term in the denominator calculates intra-cluster distance for the cluster with the

greatest intra-cluster distance. The metric is a parameter of the index, and can, among others, be taken from the list of metrics provided in Metrics.

In essence the Dunn-Index compares the minimal distance between the two clusters to the maximal distance within a cluster. Better results score higher because they tend to have denser cluster that are more spares.

As a last example of cluster validity measures the NIVA-Index will be introduced. NIVA-Index was proposed by Rendón et al. (2008) and consists of a two stage procedure.

In the first stage all observations of a cluster c are subdivided into l subgroups. Rendón et al. (2008) used the OSB clustering algorithm[9] to subdivide the clusters, but other algorithms can be applied as well.

In the second stage the NIVA-Index is calculated for each of the natural clusters. Its formula is:

$$NIVA(C) = \frac{Compac(C)}{SpexG(C)}$$

Where the Compac(c) measure describes the average compactness over all clusters. And SpexG(C) describes the separability between the clusters.

The compactness (compac(c)) is defined as the average over all clusters of the product of the compactness (Esp(c_i)) and the seperability Sepx(c_i) between the clusters. N is the total number of clusters. The applied formula is:

$$Compac(C) = \frac{1}{N}\sum_{i=1}^{N} Esp(c_i) * SpexS(c_i)$$

This raises the question of how its elements, Esp(c_i) and SpexS(c_i), are defined. At first Esp(c_i) will be explained here. Its value can be calculated using the following formula:

$$Esp(c_i) = \frac{1}{l_i}\sum_{k=1}^{l_i}[\frac{1}{n_i^k}\sum_{j=1}^{n_i^k}\left(d(x_j, x_{j+1})\right)]$$

$d(x_j, x_{j+1})$ describes the distance between an element of sub cluster l_i and its nearest neighbor. This distance is averaged over all n_i^k elements of the respective sub cluster. In a subsequent step, this resulting sub cluster average distance between each observation and its nearest neighbor, is averaged over all sub clusters in cluster c_i. It can be understood as a two stage average of the distance of an observation to its nearest neighbor within the same sub cluster. The first stage

[9] "The clustering algorithm uses the clustering criteria in order to detect connected components, for which an object x_j belongs to a cluster k, if only there is an object $x_{j,h}$ (nearest neighbor from x_j) such that the Euclidean distance between the two objects is greater than a calculated threshold, called similarity threshold. In order to calculate the similarity threshold, we use a heuristic, which consists of calculating an average of distances between x_j and $x_{j,h}$ each time an object $x_{j,h}$ enters cluster k." (Rendón et al., 2008, p.243)

aggregates over observations in the same sub cluster and the second stage aggregates over all sub clusters within on cluster.

In order to fully understand Compac(C), its second element, Spex(c_i)will be introduced next. It uses the formula:

$$SpexS(c_i) = \frac{1}{l_i}\sum_{k=1}^{l_i}\{\max\{d(sv_k, sv_j)\}\}$$

$\max\{d(sv_k, sv_j)\}$ describes the distance between a respective sub cluster centroid and the sub cluster centroid that is the furthest apart within the same cluster. $SpexS(c_i)$ can be interpreted as the average distance over all sub clusters within the one cluster from a sub cluster centroid to the respective furthest sub cluster centroid.

The only element that is left to define in order to build the NIVA-Index is SpexG(C). It is defined as:

$$SpexG(C) = \frac{1}{N}\sum_{i=1}^{N}\{\min_{\substack{j\in C \\ i\neq j}}\{d(v_i, v_j)\}\}$$

Where v_i and v_j are centroids of clusters. SpexG(C) can be thought of, as the average distance of a cluster's centroid to its nearest neighbor centroid. SpexG(C) is called average separability of C groups.

The NIVA-Index depends the three values that are explained above. In essence the index decreases when sub clusters are dense (low distance from observation to its nearest neighbor, given by Esp(c_i)), and sub clusters are located close together (low maximum distance between sub cluster centroids, measured by SpexS(c_i)). On the other side, it increases when the clusters are not well separated (low minimum distance between cluster centroids, indicated by SpexG(C)). Considering that, it can be said, that segmentations with a low NIVA-Index tend to have preferable properties over segmentations with a higher NIVA-index.

4. CUSTOMER CLUSTERING FOR AN INSURANCE

Now, that a variety of theoretical concepts are introduced, this section will focus on the analysis that was carried out throughout the internship.

The project of the internship is to cluster the clients of an insurance. The goal of the project is to group customers in order to identify and characterize similarities and differences. The objective of the project is not to create the best solution from a mathematical point of view, but to create subsets of customers that are interesting from a business perspective. The gained insights should be used to manage customers in a more profitable way. Simply partitioning the population into groups by applying fixed rules would also be a valid procedure to achieve this goal. But it is decided, that a machine learning technique (k-means clustering) should play a role when defining the segments, too.

The project starts out as a segmentation for the health department of the insurance, but during the course of the project other LOBs[10] are included in the analysis to provide a full outlook of the client. The following sections will describe the work in the first project of the internship.

4.1. DATA SOURCES

The first time of the already running project is dedicated to getting to know the data sources, the team and the tools that are provided. The data science team consists of two data scientists and me as an intern. SAS Enterprise Miner and SAS Enterprise Guide are the tools that were provided by the insurance. The main data sources are:

- BICP: A view on a customer level including demographic information and aggregated information from the policy level.
- RCOM (Radar Comercial): A view on a policy level. It includes unaggregated information for each policy.
- HLTH (tables containing information about health clients and policies): This is a compilation of data sets containing detailed information, like, for example claims, payments, and insurance types, for health insurances.
- INE (Instituto Nacianal de Estatística, engl: national statistics institute): These publicly available statistics are used as an external data source to characterize the demographic status of a client.
- Contact Center: The contact center is a department of the insurance that is responsible for the customer contact. The notes from phone calls will be evaluated.

The first mile stone of the project is to create the CAR (Customer Analytic Record). It should contain a 360 degree view of the customer. Therefore it includes information from all described data sources and serves as a base for the segmentation. It will be also used to profile the resulting segments. An illustration of how CAR is merged can be viewed in Figure 3.

[10] LOB = Line of Business. Six LOBs were identified: Health, Auto, Home, Life Risk, Life Other, and Other. Each policy and all its properties are assigned to one LOB.

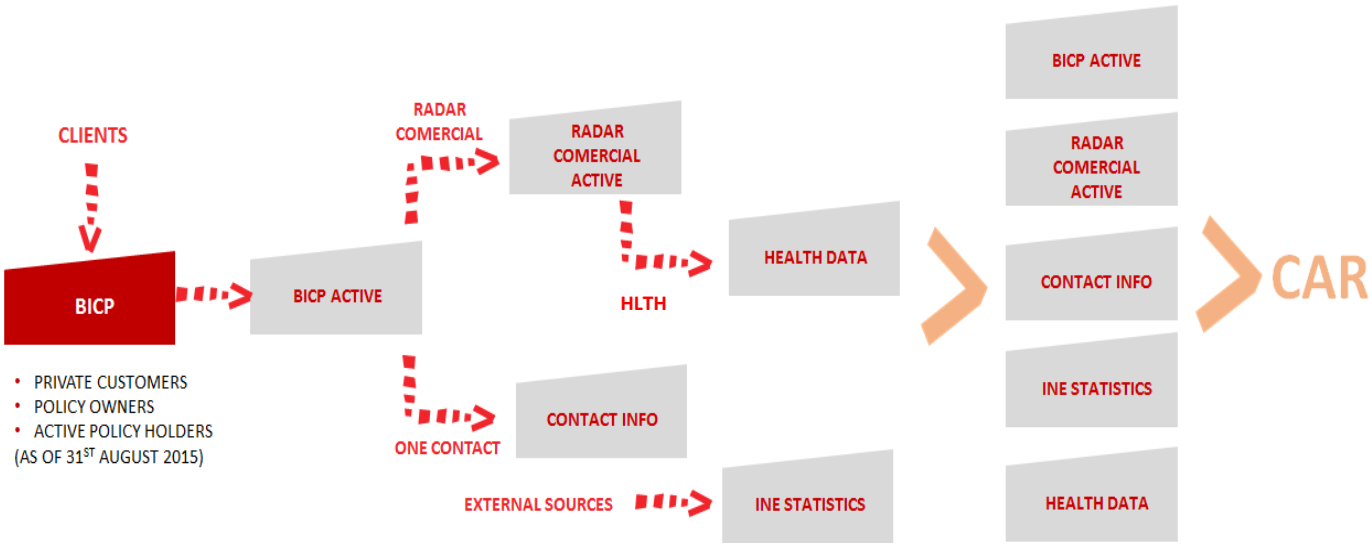


Figure 3 - Creation of the CAR

It can be seen, that the start point to select customers that are included in the analysis is BICP. The cutoff date for the analysis is 31.08.2015. Whenever clients are referenced as having specific properties, they are meant to apply by that date, if not specified differently. Included in CAR are clients that are active policy holders that own at least one policy. Companies are excluded from the analysis, because they are likely to show a different behavior and are typically acting on a different scale than individual persons. Also excluded are duplicate records and clients that faded away. The population of CAR is completed by health clients from HLTH that cannot be found in BICP. The same filter is applied to the health clients. In the end, CAR consists of 2.275.178 clients that are characterized by a total of 2.380 variables. A full list of variables can be viewed in the appendix (List of variables in CAR).

It needs to be said that the information form the contact center are extracted from the notes of phone calls. These notes suffer from bad data quality and the technique that is used to extraction does probably not capture all contained information. It is agreed on using key words in order to assign a call to a category. A list of categories and used key words is shown in Table 4. The categories should conclude about the line of business that is subject of the talk, and about reason of the contact. Prior to assigning categories to the notes, all notes have been stripped. This means that all accents and special Portuguese characters have been removed in order to be compliant to the software. Also key words are typically the stem of a word so that the capture all forms and conjugations of the respective key word.

Category	**Key Words**
Life	vida, acident, morte, viag, funeral, cacador, familia
Auto	matricula, veiculo, viatura, auto, automovel, carro, peritag, condutor, conduc, para&choque, moto, embarcacao

Health	saude, doenca, cancro, hospital, medic, ambulancia, terapia, terapeuta, ferimento, lessao, clinico, deficiente, mulher
Home	duch, quarto, corredor, banho, levatorio, cozinha, sala, seguro casa, prejuízo&agua, raio, tapete, computador, televis, fogo, telhado, habitacao, bicicleta, condominio, domestica
Opening	abertura, assinar, fazer seguro, simulacao
Cancellation	cancel, anul
Call back	call&back, cham&volta, liga&volta
Not reached	mailbox, atende&telef, nao&respond
Inclusion	inclu
Extension	prorroga, renova
Payment	paga, troca&cheque, troca&chq
Alteration	atualiza, altera, substitu, retific, corrig, aperfeico, simulacao, aumenta, mudanca
Claim	carta&declinacao, relama, dano, queixar

Table 4 - Categories and key words for topic extraction from phone call notes

4.2. SEGMENTATION

In order to give the insurance the chance to gain knowledge about the customers from different perspectives, it is decided to create three different dimensions that should be segmented. The identified dimensions are demographic, behavior, and value. The method that is used for clustering follows the same schema in each dimension. The variables that will be used for the segmentation have been selected according to a variety of criteria.

In the first place all variables that have a fraction of missing values of greater than 30% are not allowed to enter the segmentation process. The same applies to categorical variables with more than 20 different values. These would extend the space in a way that would slow down the clustering algorithm too much. The remaining variables are allocated to one of the dimensions. The client ID (variables chave and tipo_chave in the appendix List of variables in CAR) is allocated to all of the dimensions. Due to the amount of variables it is necessary to investigate the correlation among variables of the same dimension. Uncorrelated variables that had a proportion of less than 30% of missing values were proposed to be used as input for the segmentation. Because the named restrictions are excluding some variables that are considered interesting from a business perspective, business experts are allowed to override the suggestions made by the 30% rule and correlations.

Because of its scalability (discussed in Advantages and Disadvantages of K-Means) the k-means algorithm was identified to be the best clustering algorithm for a data set for this size (about 16 GB). Before applying the algorithm to the input variables, some preprocessing needs to be performed.

Apart from data completeness and correlation investigation, the presence of outliers is taken care of by using the Percentile method. All values above the 99th percentile are set to the value of the 99th percentile. And all values below the 1st percentile were set to the 1st percentile.

After the outlier treatment, missing values are imputed. Categorical variables are anticipated to assume the most frequent value, and the mean is imputed to numeric variables. Some variables are treated in a special way. These special treatments are explained in the respective chapters below.

Also data standardization is used to fit the data to the algorithm. In all dimensions dividing the difference from the mean by the standard deviation leads to better results than using a range between 0 and 1. Better, in this case, means that the results could be interpreted by business in a more intuitive way. The standardization is applied after the removal of outliers and after the treatment of missing values. Categorical variables are encoded as dummy variables for each of their levels.

At first the segmentation process is carried out on a sample of 10% of the data. Once an acceptable solution is found, the settings are applied to the whole population. During the course of clustering the input variables change constantly. These changes also consider suggestions from business experts so that it is ensured that the result would fit the business needs.

In order to evaluate if a cluster is stable all initialization methods that are available in SAS Enterprise Miner are applied. In the end, the initialization method using principal components, as explained in Linear Relation of Complexity and Performance is chosen to create the most interesting clusters.

All in all about 1000 iterations are performed, where, among others, also principal component analysis for dimensional reduction is applied. The solutions using principal component analysis are rejected because of their difficult interpretability and will not be part of this work.

Details for of the final segmentations for each dimension will be presented in the following sections.

4.2.1. Demographic Segmentation

The demographic segmentation aims to provide information about groups of clients in a similar demographic situation. Its outcome is explained in this chapter.

For a better understanding of the outcome, it is necessary to understand the variables that are used to create the segments. The used variables are explained below.

1. Gender: The gender of a client is captured by two binary variables. One indicating men, the other one women. The perfect collinearity of these variables considers this feature twice for the clustering algorithm. Below will be shown why this is considered a good behavior.

2. Age: The age of the client in years.

3. Contactablity: This is a numeric variable that assumes a value of three if the client provided his email address, a value of two if the clients did not provide his email address, but his

mobile phone number, a value of one if the client did not provide his email address or mobile phone number, but his landline number, and zero if none of the above are provided. It is based on the variables Indicador_Email, Indicador_Telefone_Fixo, and Indicador_Telemovel.

4. Civil Status: Civil Status gives conclusion about the marital status of a client. It is a categorical variable that can assume the values married, divorced, separated, single, and widow.

5. Preferred Channel: This categorical variable gives conclusion about the preferred contact channel of the client. Possible values are banc, broker, agent, and others.

6. District: This categorical variable provides information about the localization of the customer. It is an aggregated version of the variable distrito_ctt_dem. The aggregated version that is used in the segmentation can assume the values Aveiro, Açores, Beja, Braga, Bragança, Castelo Branco, Coimbra, Faro, Guarda, Leiria, Lisboa, Madeira, Porto Alegre, Porto, Santarém, Setúbal, Viana do Castelo, Vila Real, Viseu, and Évora.

7. Tenure: This is a numeric variable indicating for how many years the client owns his oldest policy.

The proportion of missing values for each variable can be seen in Table 5

Label	Variable	Fraction Missing
Tenure	ANTIGUIDADE	10%
Age	IDADE	2%
Civil Status	ESTADO_CIVIL	27%
Preferred Channel	CANAL_PREFERENCIAL	8%
District	distrito_agg_ctt_dem	0%
Gender	SEXO	2%
Contactablity	Contactablity	8%

Table 5 - Fraction of missing values for demographic segmentation

For the demographic clustering the imputation of missing values was changed for some variables. Missing values of civil status and gender were imputed in a stratified way. So missing values did not automatically count as members of the most frequent class, but rather were assigned to one of all possible classes based on their relative frequency.

The only two numerical variables in the demographic segmentation are tenure and age. They have a correlation coefficient of 0.3, and therefor are considered mainly uncorrelated. Appart from that the only correlations with an argument of greater than 0.3 are between civil status married and single (-0.86) and gender male and female (-1).

Due to its size, a full correlation matrix including all dummy variables of the demographic dimension can be found only in the appendix (Correlation Matrix for Demographic Segmentation).

In coordination with the insurance it was agreed on assigning weights to the chosen variables. According to which both gender variables and the age of the customer should count four times. The weights were assigned by first standardizing all variables to display their deviation from the mean as a multiple of their standard deviation, and then the weighted variables were multiplied by the respective weight. This procedure increases the relative variability of the variables of interest and so increases their impact on the distance of an observation to its centroid. This procedure is uncommon in statistical focused solutions of clustering. In this case, however, the purpose was not to create mathematically optimal clusters, but rather to create segments, that help the business differentiating groups of customers. Gender and age of the client were assumed to have a major impact on the on selection of a target strategy for the customers, so it was decided to assign the weights.

After various iterations the algorithm resulted with six clusters. A description of the clusters can be done using the weighted variables:

1. Young Women: This cluster consist mainly of female customers in the young adult age.
2. Young Men: This cluster consist mainly of male customers in the young adult age.
3. Adult Women: This cluster consist mainly of female customers in the adult age.
4. Adult Men: This cluster consist mainly of male customers in the adult age.
5. Senior Women: This cluster consist mainly of female customers in the senior age.
6. Senior Men: This cluster consist mainly of male customers in the senior age.

A distribution of clients among the demographic segments is illustrated in Figure 4. It can be observed that the segments young men, adult women, and senior women all have an approximate size of 390.000 customers, while adult men form the greatest cluster of 563.228 clients. Young women and senior women form the smallest cluster with a rough size of 260.000 customers. Naturally the male cluster are greater, because 60% of the population consists of male.

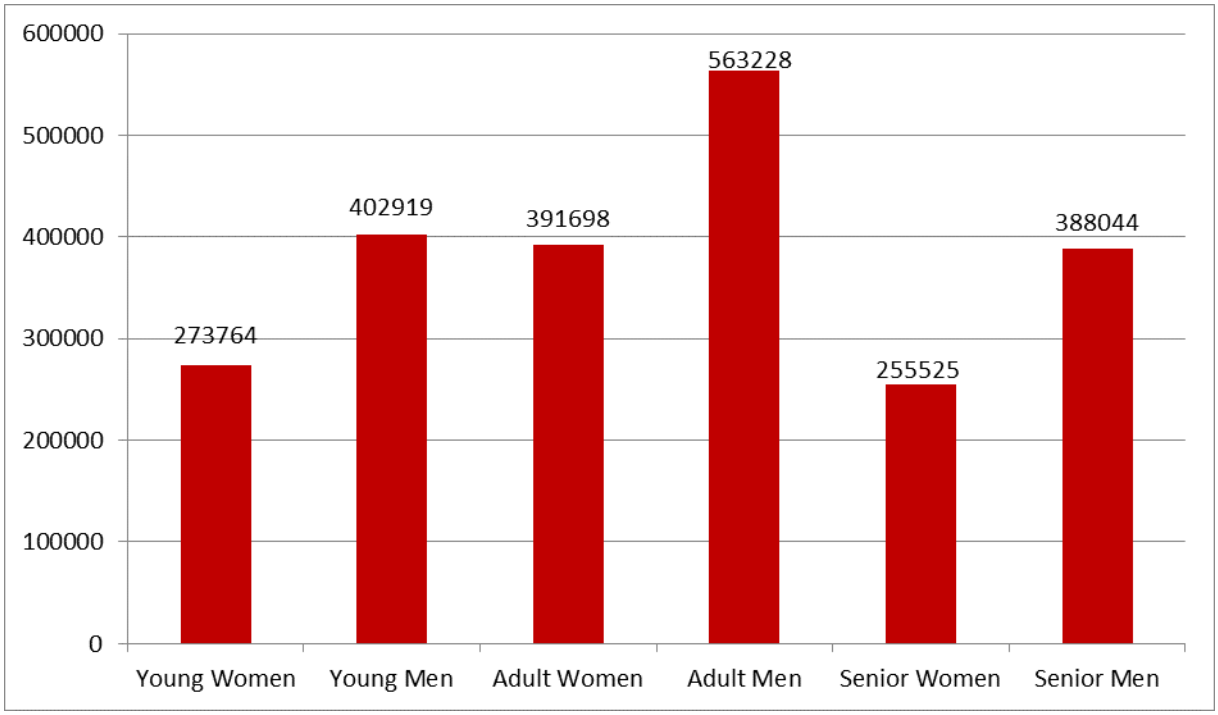


Figure 4 - Number of customers in demographic segments

Figure 5 displays the age distribution of the clients among demographic segments. While the male cluster always tend to be slightly older than their female counterparts, the gap between young, adult, and senior segments is clearly visible.

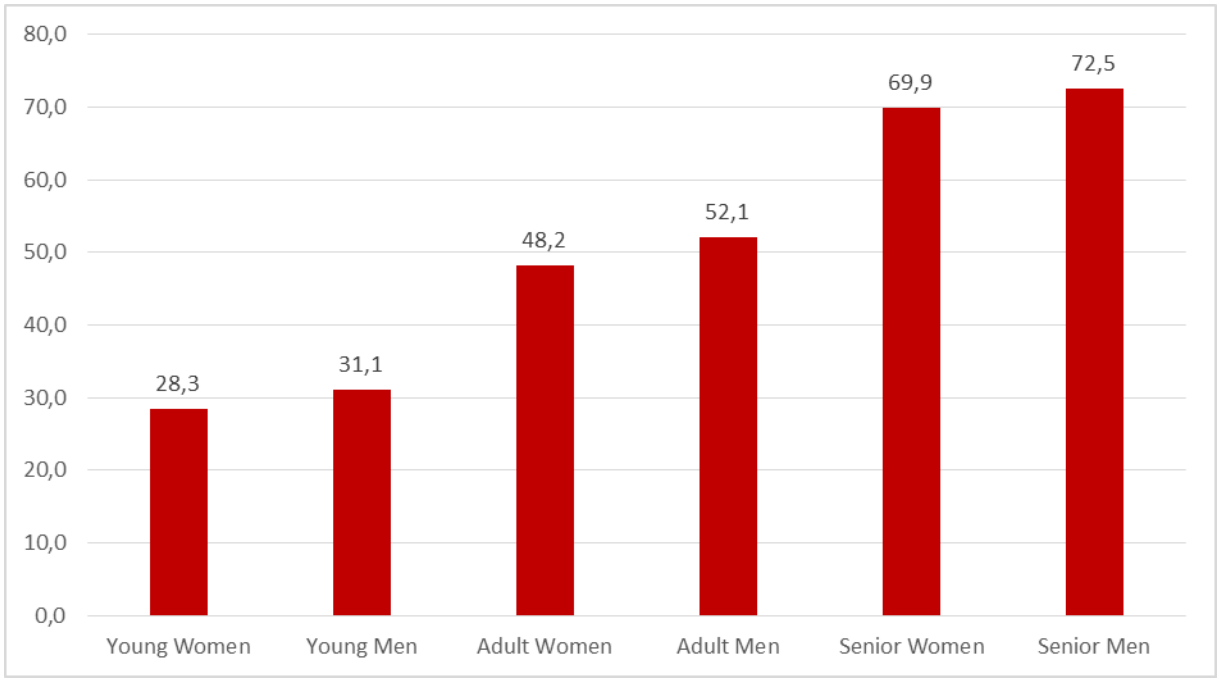


Figure 5 - Age distribution of demographic segments

The male segments consist of 100% male and the female segments consist of 100% female. From the gender perspective the assignment to either male, or female cluster is a clear cut. For the age however it is not. All segments also contain few examples of young and senior customers. This is due to the influence of the other variables. The distribution of all variables among the demographic segments is given in the appendix (Profiling of Demographic segments).

Although the cluster selection was based on qualitative criteria, quantitative cluster validity measures will be presented. Two of the Internal Validation Techniques are applied. Due to the size of the data set, they are only applied to a 5% fraction of the overall population.

The first measure that is computed is the CH-index. In the 5% sample it reaches an SSB value of 2,387,421, and a SSW value of 736,601.1. Further it considered 113,758 observations with 36 dimensions. Applying the formula of the CH-index to those values, a level of 10,531.08 is reached.

By comparing SSB to TSS (total sum of squares, 3,124,022), it can be observed, that the segmentation accounts for 23.58% of the overall variation. For a segmentation using six centroids to describe 113,758 observations, this is a reasonable result. The segments serve well as an indicator for the client's demographic situation.

To provide a second measure, the Dunn-index is computed. In the demographic segmentation the maximum intra-cluster distance is 10.37. It is measured as the maximum distance from an observation to its centroid. The minimal Euclidean inter-cluster distance is measured between the clusters of young women and adult women, as 4.61. It is examined using the centroids method. Calculating the Dunn-index from these two distances, a ratio of 0.4445 results. This result signifies that there is at least one observation that is further apart from its centroid than the distance between the two closest centroids.

Finally Figure 6 - Scatter plot of the first two principal components for the demographic segmentation displays a visual distribution of all observations in the 5% sample. The two dimensional plot is obtained from the first two principal components of the correlation matrix of the variables that are used in the demographic segmentation. An interpretation of the principal components is not part of this work. These principal components account for 8.24% of the total variance. The percentage of explained variance is low, because of the dimensionality of the used data set (36 variables). In the first place the overall dispersion of the data points is examined. The positioning of the point suggests a four cluster solution with one cluster in each of the quarters of the plot. The usage of categorical variables explains the occurrence of visible gaps in the graphic. Before concluding that a four cluster solution would have been a better fit, it needs to be considered, that the plot only accounts for 8.24% of the total variation. This is also an explanation why the chosen clusters are distributed in the space. As mentioned above, the chosen segmentation accounts for 23.58% of the total variance. Despite the dispersion, a system of cluster membership can be obtained. Each of the four parts of the visualization starts with the young women (red) on the left side, followed by young men (brown). The areas of the adult women (green) and adult men (turquoise) can be found close to them. Observations of senior women (blue) are typically located below, and senior men (purple) can be found on the right side.

Figure 6 - Scatter plot of the first two principal components for the demographic segmentation

4.2.2. Behavior Segmentation

At next the behavior of the clients is segmented. The behavior dimension should result in groups of clients that share a common characteristic behavior. This behavior should identify clients that can be managed in the same way, but should be handled differently from other segments. This section will explain the used variables and offer an overview of the main characteristics of each cluster.

In alignment with business experts 18 numeric variables have been identified to enter the clustering process. These variables are:

1. Number of annulled policies in the last 36 month
2. Number of month since last purchase
3. Number of premiums paid after the end date in the last 12 month
4. Number of active policies
5. Number of active product families
6. Ratio of active policies to total policies

7. Number of contracts as insured person
8. Average of months from the beginning to the end of the policy
9. Sum of the number of renewed policies in the last 12 month
10. Sum of the number of new policies in the last 12 month
11. Yearly increase in active product families. Calculated as the difference between PPFAMPROD_ACT (number of active product families) and FAMPROD_12M (number of active product families 12 month ago)
12. Sum of the number of active policies that began in the last 24 month
13. Sum of policies which came from substitutions and are active
14. Number of distinct channels
15. Number of simulations in the last 12 month
16. Number of non-financial claims
17. Number of month since the last closure of a non-financial claim
18. Frequency of annulment. Calculated as the quotient of number of annulled policies (HSAPANU_TOT) and tenure of the customer (ANTIGUIDADE)

The fraction of missing values per variable can be obtained from Table 6. The table clarifies that all variables but the last four satisfy the condition of having at the most 30% of missing values. Frequency of annulment, number of annulled policies in the last 36 month, number of non-financial claims, and number of month since the last closure of a non-financial claim have fractions of missing values that are above the threshold of 30%. Never the less these variables are taken into account because after discussing advantages and disadvantages with the insurance, it was decided that they carry important information to form useful clusters. The high percentage of missing values in the number of financial claims can be explained by the calculation of the variable that only assigns a value if there was a financial claim. So a large proportion of the 72% missing values can be explained by the non-existence of a financial claim. For this reason it was valid to change the missing value treatment for this variable to impute zeros instead of the variable mean.

Label	Variable	Fraction Missing
Number of month since the last purchase	NMESES_ULT_COMPRA	9%
Number of distinct channels	NUM_CANAIS	8%
Number of simulations in the last 12 month	NUM_SIMUL_12M	8%
Number of premiums paid after the end date in the last 12 month	PAG_RECCOBRDT_12M	14%
Number of active policies	PPAPOLVIG_ACT	9%
Number of active product families	PPFAMPROD_ACT	9%
Ratio of active policies to total policies	PPRAC_APOL_VIG	9%
Number of contracts as insured person	PPTOTTPREL_PS	9%

Sum of the number of active policies that began in the last 24 month	sum_policy_vig_b24M_rcom	13%
Sum of the number of renewed policies in the last 12 month	cou_pol_ren_12M_rcom	13%
Yearly increase in active product families	slope_PPFAMPROD	9%
Average of months from the beginning to the end of the policy	avg_maturity_mon_rcom	13%
Sum of policies which came from substitutions and are active	sum_sub_vig_rcom	13%
Sum of the number of new policies in the last 12 month	count_pol_new_12M_rcom	13%
Frequency of annulment	freq_HSAPANU_TOT	61%
Number of annulled policies in the last 36 month	ANU_TOTAPO_ULT36M	61%
Number of non-financial claims	SinNFinAbert	72%
Number of month since the last closure of a non-financial claim	SinNFinMesUltFech	80%

Table 6 - Fraction of missing values for behavior variables

The correlations for these variables are provided in Table 7. It can be concluded that most of the variables are uncorrelated. Although there are some variables scoring high correlations. For example the average of the month from the beginning of the policy to the end of the policy is correlated with the number of month since the last purchase (0.84). Also the number of renewed policies in the last 12 month shares a correlation coefficient of 0.79 with the number of active policies. The latter is also correlated with the number of active product families with a correlation of 0.79. Despite these correlations it was decided to keep all variables in the segmentation in order to capture all intrinsic information. This procedure is justified by the business purpose of the project. The insurance is interested in all characteristics and wants their clients to be clustered according to all of them.

Label	Variable	NMESES_ULT_COMPRA	NUM_CANAIS	NUM_SIMUL_12M	PAG_RECCOBRDT_12M	PPAPOLVIG_ACT	PPFAMPROD_ACT	PPRAC_APOL_VIG	PPTOTTPREL_PS	avg_maturity_mon_rcom	sum_sub_vig_rcom	count_pol_new_12M_rcom	sum_policy_vig_b24M_rcom	cou_pol_ren_12M_rcom	slope_PPFAMPROD	freq_HSAPANU_TOT	ANU_TOTAPO_ULT36M	SinNFinAbert	SinNFinMesUltFech
Number of month since the last purchase	NMESES_ULT_COMPRA	1,00	-0,12	-0,15	-0,09	-0,19	-0,19	0,23	-0,10	0,84	-0,17	-0,27	-0,35	-0,11	-0,19	-0,26	-0,23	-0,05	0,12
Number of distinct channels	NUM_CANAIS	-0,12	1,00	0,05	0,15	0,34	0,39	0,02	0,20	0,04	0,09	0,06	0,10	0,27	0,18	-0,11	0,06	0,04	-0,02
Number of simulations in the last 12 month	NUM_SIMUL_12M	-0,15	0,05	1,00	0,06	0,12	0,10	-0,05	-0,02	-0,12	0,09	0,38	0,32	0,05	0,10	0,15	0,12	0,06	-0,07
Number of premiums paid after the end date in the last 12 month	PAG_RECCOBRDT_12M	-0,09	0,15	0,06	1,00	0,29	0,32	0,00	0,18	-0,06	0,07	0,05	0,11	0,23	0,07	-0,04	0,05	0,35	-0,06
Number of active policies	PPAPOLVIG_ACT	-0,19	0,34	0,12	0,29	1,00	0,79	0,03	0,53	-0,03	0,21	0,29	0,40	0,79	0,34	-0,12	0,23	0,08	-0,07
Number of active product families	PPFAMPROD_ACT	-0,19	0,39	0,10	0,32	0,79	1,00	0,05	0,51	-0,02	0,17	0,14	0,24	0,60	0,47	-0,20	0,12	0,10	-0,06
Ratio of active policies to total policies	PPRAC_APOL_VIG	0,23	0,02	-0,05	0,00	0,03	0,05	1,00	-0,12	0,19	-0,29	-0,08	-0,09	-0,09	-0,01	-0,38	-0,40	0,13	-0,09
Number of contacts as insured person	PPTOTTPREL_PS	-0,10	0,20	-0,02	0,18	0,53	0,51	-0,12	1,00	-0,02	0,01	-0,02	-0,02	0,11	0,61	-0,16	-0,01	0,03	-0,01
Average of months from the beginning to the end of the policy	avg_maturity_mon_rcom	0,84	0,04	-0,12	-0,06	-0,03	-0,02	0,19	-0,02	1,00	-0,06	-0,25	-0,30	0,05	-0,16	-0,27	-0,11	-0,07	0,14
Sum of policies which came from substitutions and are active	sum_sub_vig_rcom	-0,17	0,09	0,09	0,07	0,21	0,17	-0,29	0,01	-0,06	1,00	0,03	0,26	0,32	-0,04	0,15	0,26	-0,02	-0,01
Sum of the number of new policies in the last 12 month	count_pol_new_12M_rcom	-0,27	0,06	0,38	0,05	0,29	0,14	-0,08	-0,02	-0,25	0,03	1,00	0,72	0,04	0,35	0,23	0,23	0,04	-0,08
Sum of the number of active policies that began in the last 24 month	sum_policy_vig_b24M_rcom	-0,35	0,10	0,32	0,11	0,40	0,24	-0,09	-0,02	-0,30	0,26	0,72	1,00	0,26	0,23	0,26	0,27	0,06	-0,13
Sum of the number of renewed policies in the last 12 month	cou_pol_ren_12M_rcom	-0,11	0,27	0,05	0,23	0,79	0,60	-0,09	0,11	0,05	0,32	0,04	0,26	1,00	-0,10	-0,03	0,39	0,03	-0,06
Yearly increase in active product families	slope_PPFAMPROD	-0,19	0,18	0,10	0,07	0,34	0,47	-0,01	0,61	-0,16	-0,04	0,35	0,23	-0,10	1,00	-0,09	-0,09	0,06	-0,02
Frequency of annulment	freq_HSAPANU_TOT	-0,26	-0,11	0,15	-0,04	-0,12	-0,20	-0,38	-0,16	-0,27	0,15	0,23	0,26	-0,03	-0,09	1,00	0,52	-0,05	-0,01
Number of annulled policies in the last 36 month	ANU_TOTAPO_ULT36M	-0,23	0,06	0,12	0,05	0,23	0,12	-0,40	-0,01	-0,11	0,26	0,23	0,27	0,39	-0,09	0,52	1,00	-0,04	-0,05
Number of non financial claims	SinNFinAbert	-0,05	0,04	0,06	0,35	0,08	0,10	0,13	0,03	-0,07	-0,02	0,04	0,06	0,03	0,06	-0,05	-0,04	1,00	-0,51
Number of month since the last closure of a non financial claim	SinNFinMesUltFech	0,12	-0,02	-0,07	-0,06	-0,07	-0,06	-0,09	-0,01	0,14	-0,01	-0,08	-0,13	-0,06	-0,02	-0,01	-0,05	-0,51	1,00

Table 7 - Correlations among behavior variable

In this dimension all variables enter the k-means algorithm unweighted. After several iterations, a five cluster solution seems to contain the most interesting insights for business. These five segments are:

1. Loosing Client: This segment mainly consists of clients that are inactive for a long time. Due to their low contact rate with the insurance they might be more likely to leave the insurance.

2. Recent Purchaser: Customers in this segment are more likely to have purchased a policy within the last year. This segment does not consist exclusively of new customers, but contains a large proportion of new clients.

3. Often Insured: The often insured client consists of a greater than average proportion of clients that are insured in other clients policies. This could be an interesting target group to cross sell other products, because these clients are already in touch with the insurance, but do not have a large portfolio.

4. Big Client: These clients excel in the number of policies they own. They tend to own significantly more policies than average.

5. Health User[11]: This is a segment that only contains customer who's only contact with the insurance is that they are insured in the health policy of another customer. They tend to be younger and form a good target for upselling and cross selling when they reach a certain age or career level. Health user are sorted into a special cluster because they are especially interesting to the health department of the insurance.

The number of clients per segment can be obtained from the graph in Figure 7. It can be observed that the size of the clusters is decreasing from the left to the right. Loosing clients seem to be the biggest fraction of the population (932.229 clients). This might be due to the high average age of the population (50 years). Clients that buy new policies tend to be young, so, many customers are identified as loosing clients. The second largest segment are the resent purchaser (601.915 clients). By further investigating this segment, the insurance can create specific strategies to make its recent purchasers profitable. In the population there are 366.366 clients segmented as often insured customers. The two smallest segments are big clients, with 195.241 customers, and health users with 179.427 clients.

[11] Health User are a special segment. In the original segmentation they were not included, because they are not part of BICP. After the original segmentation has been performed, the insurance desired to add these clients to the analysis. In the demographic dimension they are sorted in correctly, because all needed variables were available. In the behavior dimension they are included as an additional cluster, because they all share the common property of not owning a single policy. They are only insured in the health policy of another client. In the value dimension they are sorted according to their annually payments as either very profitable, or value loosing.

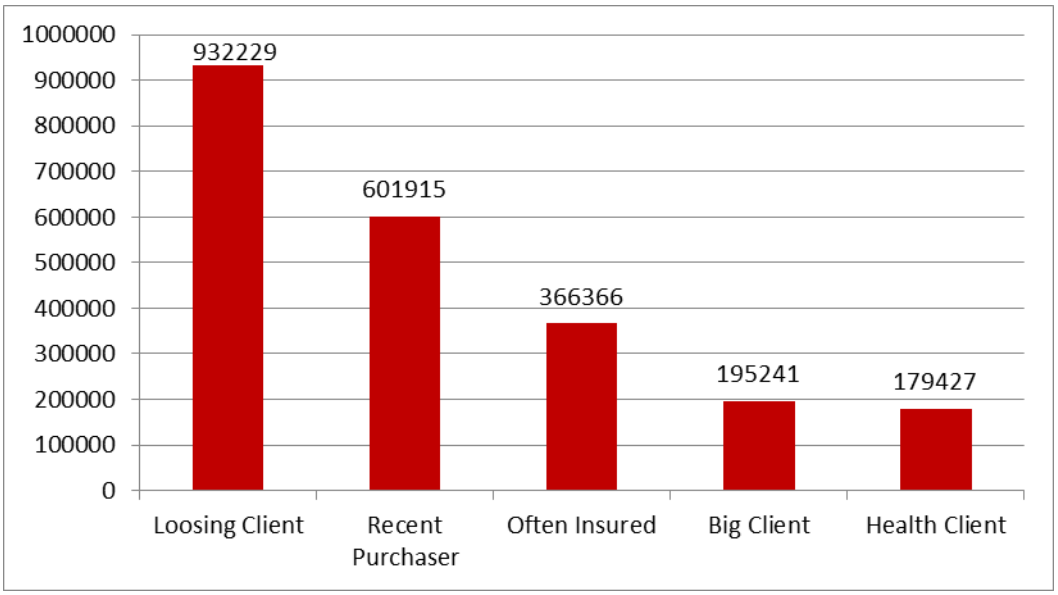


Figure 7 - Number of customers in behavior segments

Figure 8 - Number of years since last purchase demonstrates that loosing clients, on average, have the longest time since the last purchase (8.24 years). This is the value that describes this segment best. Also often insured customers have not bought a new policy for 3.95 years. But this is related to their behavior as being more likely to be insured in someone else's policy. This graph only concludes about policies that are owned by a customer. This is also the reason why health user do not appear in any of the following figures. They do not own a single policy, and therefore do not qualify for such kind of analysis. The value is the lowest for recent purchaser (1.27 years) and big clients (1.3 years) which is not surprising, because big clients are assumed to buy new policies more often than the other segments, and recent purchaser, by definition, have bought more recently on average.

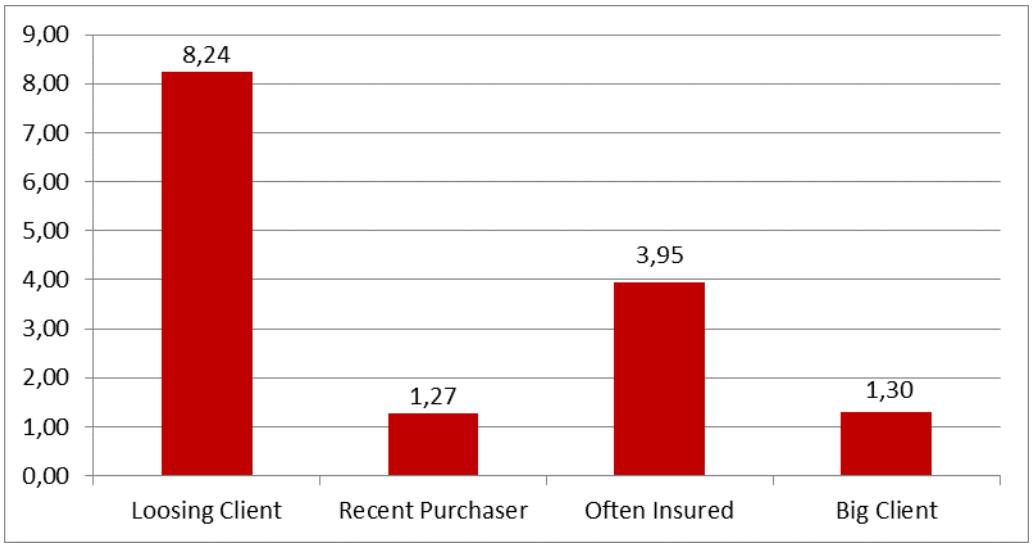


Figure 8 - Number of years since last purchase

Figure 9 - Number of new policies (12 month) vs. number of total policies compares the total number of policies a customer owns and the number of policies that a customer has bought in a specific segment within the last 12 month. This graphic illustrates the main difference between recent

purchaser and big clients. While both segments on average purchased 0.65 policies within the last year, the total number of policies differs considerably. As explained above, for big clients it is a normal behavior to purchase new policies, while recent purchasers seem to buy new products less frequently. The next step for the insurance could be to investigate the reasons why recent purchaser bought a new policy and implement their findings in marketing campaigns. On the other side the further development of recent purchaser needs to be traced and investigated in order to manage them in the most profitable way.

Also the inactivity of losing clients is displayed by the chart. On average they did not buy any policy in the last year. On the other side they are customers with a rather small average portfolio size of only 1.42 policies per client. This is the same average number of total policies as the recent purchaser. But for the latter, the number of bought policies in the last year indicates growth, not shrinkage, for the portfolio size. The second biggest clients can be found in the often insured segment. With 3.19 policies they are also larger than average customers. An average client owns 2.12 policies. The health users are not listed, because they do not own any policy, and so do not qualify for the analysis.

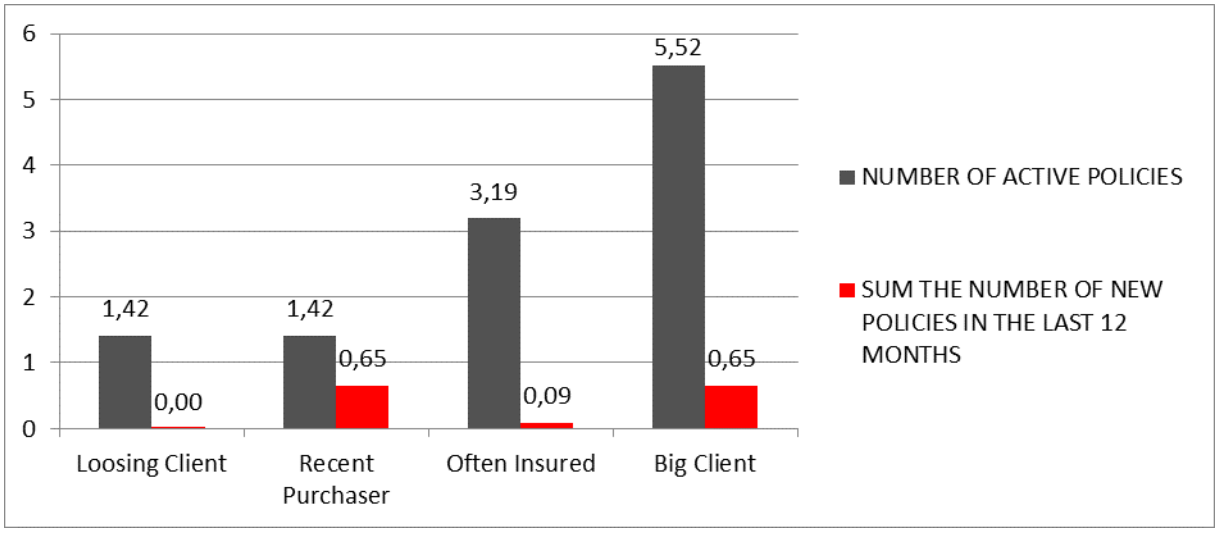


Figure 9 - Number of new policies (12 month) vs. number of total policies

Finally the number of policies as an insured person is provided by the graphic in Figure 10. As expected, often insured clients score the highest (on average 1.64 policies as an insured person). The second highest score, 0.76, is reached by big clients. As stated above, this probably has its reason in the high total number of policies. It should also be noted, that recent purchaser, on average, have four times more contracts than losing clients (0.16 compared to 0.04). But it should also be said that both values are low when comparing them to the remaining two listed segments. Health user are not listed in the figure, because the graph only includes clients that own at least one active policy. This is not the case for any of the health user, and so, they do not qualify for the analysis.

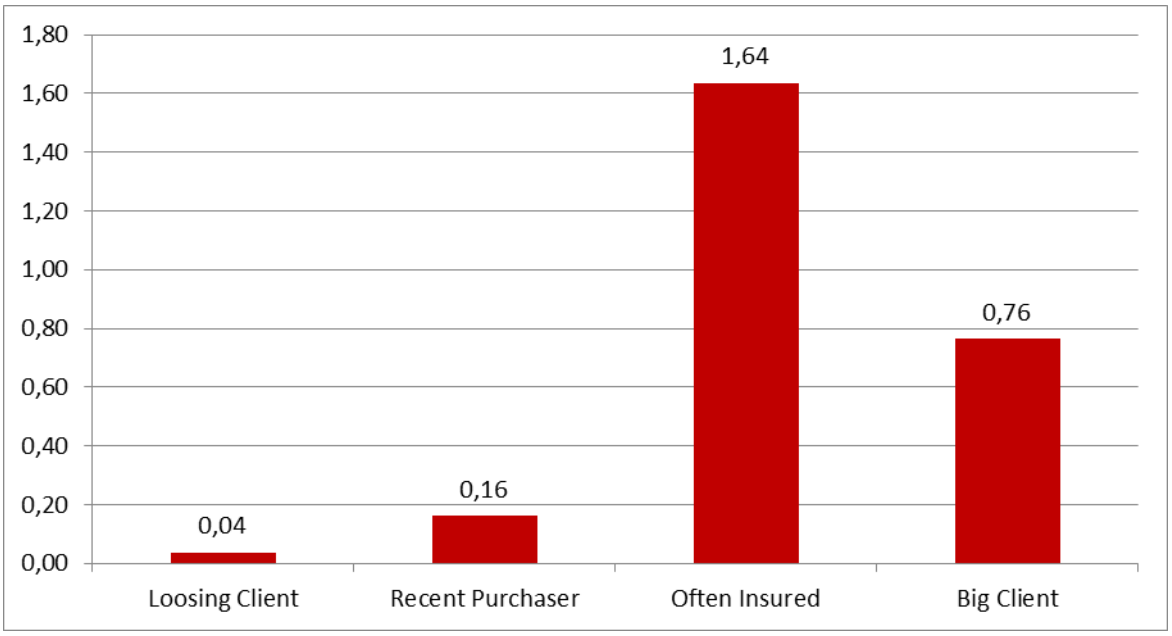


Figure 10 - Number of contracts as an insured person

Above discussed are the features that best describe the segments. A full list of the averages for each variable and each segment can be obtained in the appendix (Profiling of Behavior segments).

Although the cluster selection was based on qualitative criteria, quantitative cluster validity measures should be presented. Two of the Internal Validation Techniques are applied. Due to the size of the data set, they are only applied to a 5% fraction of the overall population. It is the same randomly chosen fraction that has been used for the quantitative validation measures of the demographic segmentation.

The first measure that is computed is the CH-index. In the 5% sample it reaches a SSB value of 532,512, and a SSW value of 1,515,104. Further it considered 113,758 observations with 18 dimensions. Applying the formula of the CH-index to those values, a level of 2,315.537 has been reached.

By comparing SSB to TSS (2,047,616), it can be observed that the segmentation accounts for 26% of the overall variation. For a segmentation using only five clusters to describe 113,758 observations, this is a reasonable result. The segments serve well as an indicator for the client's behavior.

To provide a second measure, the Dunn-index is computed. In the behavior segmentation the maximum intra-cluster distance is 15.28. It is measured as the maximum distance from an observation to its centroid. The minimal inter-cluster distance is measured between loosing clients and health user, as 1.63. To compute this value the centroid method from Metrics has been applied. The ratio of the minimal inter-cluster distance to the maximal intra-cluster distance is 0.1067. This is a result that shows the dispersion of observations in the space. The two closest cluster are almost ten times closer than the furthest observation from its centroid.

Finally, an attempt to visualize the distribution of the observations and their cluster membership is provided in Figure 11 - Scatter plot of the first two principal components for the behavior segmentation. The plotted scores of the first two principal components of the correlation matrix of

the behavior variables account for 22.09% of the overall variation in the 18 variables. Given the dense population of data points, true, separated clusters do not seem to exist. However, the segmentation produces such clear results that the segments are easily identifiable in the plot. Four clusters are clearly separable in the graphic. The loosing clients (red) in the bottom right corner, the recent purchaser (army green) on top, the often ensured clients (green) at the bottom, and the big clients (blue) on the left hand side. The latter seem to scatter out more than the other clusters. Health users are barely visible in this plot. This is because of their origin. They have been assigned to their own cluster after the other observations have been clustered. The health user segment consists of clients that are only insured in someone else's health policy and do not own any other policy with the insurance. They are in their own cluster because the used variables are not available for these customers. Due to that they have been imputed with the respective variable mean and all principal component evaluate to zero for all health clients. They are represented by a single pink dot with the coordinates zero/zero.

Figure 11 - Scatter plot of the first two principal components for the behavior segmentation

4.2.3. Value Segmentation

The third, and last, dimension of the segmentation is value. This dimension gives conclusion about the profitability of the customers. It should help to indicate whether a client is likely to be profitable, and therefore should be managed in a different manner than clients that make losses. This section will explain the resulting segments.

The variables have not only been chosen based on their properties, like correlation and percentage of missing values, but also they have been aligned with business experts to ensure that all desired information is captured in the segmentation.

14 variables have been selected to serve as input variables for the clustering algorithm. These variables are:

1. Average of all PCA[12] from active policies
2. A client's value in the current month
3. Current value of financial product holding
4. Total PCA of policies in force
5. Average of total processed commissions for active policies
6. Average of the PCA of the homologous year
7. Average PCA of all policies renewed in the last 12 month (by either completing a year or substitution)
8. Sum of current year processed commissions divided by the sum of processed premiums
9. Average of current month PCA minus the average of homologous month PCA
10. Sum of claims value year to date
11. Sum of policies that were substituted (same LOB purchased within a 7 day window of another policy ending) for a policy with a smaller PCA
12. Sum of policies that were substituted (same LOB purchased within a 7 day window of another policy ending) for a policy with a higher PCA
13. Sum of current month's technical margin minus the sum of homologous month's technical margin
14. Value of the client in the homologous month of the past year

During the iterations it became clear, that the business line Life Other causes extremely large values comparing with the other LOBs. This was caused by financial products that were included in Life Other. In coordination with business matter experts it was decided to remove financial values from

[12] PCA = Premio Anualizado Comercial; engl: yearly comercial premiums are the payments a customer has to make for a whole year.

the variables 1,2,4,5,6,7,8,9,10,13, and 14. The financial values are summarized in variable 3 (current value of financial product holding).

The fraction of missing values in the value dimension can be obtained from Table 8. Nearly all of the variables satisfy the criteria that a maximum of 30% of the variable can be missing. Variables that do not satisfy this, are the average PCA of all policies renewed in the last 12 month (by either completing a year or substitution), the sum of current year's processed commissions divided by the sum of processed premiums, and the sum of claims value year to date. The latter is available only for 8% of the data. This is due to the nature of the variable. If a client does not have a claim this field is missing. Due to this behavior the missing value treatment for this variables was changed from inserting the mean to inserting a zero, and it is valid to consider this variable for the segmentation. The other variables with a missing fraction above 30%, the average PCA of renewed policies in the last 12 month, and the ratio of commissions to premiums, were included in the clustering algorithm because of business demands. The fraction of missing values for these variables do not exceed the 30% limit by too much. They are all below 40%.

Label	Variable	Fraction Missing
Total PCA of policies in force	PPPCANAOFIN	9%
Current value of financial product holding	PPPROVM_ATUAL	9%
Average of total processed commissions for active policies	avg_com_tot_vig_nf_rcom	27%
Average of the PCA of the homologous year	avg_pca_hom_nf_rcom	30%
Average PCA of all policies renewed in the last 12 month (by either completing a year or substitution)	avg_pca_ren_12M_nf_rcom	34%
Average of all PCA from active policies	avg_vig_pca_nf_rcom	27%
Sum of current year processed commissions divided by the sum of processed premiums	rat_com_prems_nf_rcom	38%
Average of current month PCA minus the average of homologous month PCA	slope_avg_pca_nf_rcom	30%
Sum of claims value year to date	sum_claim_ytd_nf_rcom	92%
Sum of policies that were substituted (same LOB purchased within a 7 day window of another policy ending) for a policy with a smaller PCA	sum_sub_decrease_rcom	13%
Sum of policies that were substituted (same LOB purchased within a 7 day window of another policy ending) for a policy with a higher PCA	sum_sub_increase_rcom	13%
Sum of current month's technical margin minus the sum of homologous month's technical margin	slope_tech_margin_nf	8%
A client's value in the current month	tech_margin_nf_a	8%
Value of the client in the homologous month of the past year	tech_margin_nf_h	8%

Table 8 - Fraction of missing values among value variables

The correlations among these variables are provided by Table 9. Even though most variables are not correlated and have correlation coefficients around zero, there are some variables that are highly correlated. In previous versions of the segmentation they were not included, however it was decided to consider them in the segmentation in order to weight some characteristics of the client higher and to establish a full view on the customer. The sum of claims value year to date, difference in the

technical margins of the current month and the homologous month of the previous year, and the client's value of the current month are highly correlated (bottom, right hand corner of the correlation matrix in Table 9). The sum of claims value year to date and the current value of the client are negatively correlated. In previous versions of the segmentation only the current value of the client was included in the segmentation. Due to advice from business experts the sum of claims value year to date was added to the segmentation space.

The high correlation between the average PCA of the homologous year, the average PCA of all policies renewed in the last 12 month, and the average of all PCA from active policies can be interpreted as a higher weight on these variables, which also adds slightly different information.

Label	Variable	PPPCANAOFIN	PPPROVM_ATUAL	avg_com_tot_vig_nf_rcom	avg_pca_hom_nf_rcom	avg_pca_ren_12M_nf_rcom	avg_vig_pca_nf_rcom	rat_com_prems_nf_rcom	sum_sub_decrease_rcom	sum_sub_increase_rcom	slope_avg_pca_nf_rcom	sum_claim_ytd_nf_rcom	slope_tech_margin_nf	tech_margin_nf_a	tech_margin_nf_h
Total PCA of policies in force	PPPCANAOFIN	1,00	-0,02	0,23	0,30	0,34	0,35	0,00	0,06	0,05	-0,01	0,00	0,00	0,02	0,01
Current value of financial product holding	PPPROVM_ATUAL	-0,02	1,00	0,01	0,01	0,01	0,01	0,00	-0,01	0,07	0,00	-0,01	0,00	0,00	0,00
Average of total processed commissions for active policies	avg_com_tot_vig_nf_rcom	0,23	0,01	1,00	0,56	0,63	0,65	0,01	0,01	0,04	-0,01	0,01	0,00	0,01	0,00
Average of the PCA of the homologous year	avg_pca_hom_nf_rcom	0,30	0,01	0,56	1,00	0,89	0,87	0,00	0,03	0,03	-0,06	0,01	0,00	0,01	0,01
Average PCA of all policies renewed in the last 12 month (by either completing a year or substitution)	avg_pca_ren_12M_nf_rcom	0,34	0,01	0,63	0,89	1,00	0,97	0,00	0,01	0,06	-0,04	0,01	0,00	0,01	0,01
Average of all PCA from active policies	avg_vig_pca_nf_rcom	0,35	0,01	0,65	0,87	0,97	1,00	0,00	0,00	0,06	-0,02	0,01	0,00	0,02	0,01
Sum of current year processed commissions divided by the sum of processed premiums	rat_com_prems_nf_rcom	0,00	0,00	0,01	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Sum of policies that were substituted (same LOB purchased within a 7 day window of another policy ending) for a policy with a smaller PCA	sum_sub_decrease_rcom	0,06	-0,01	0,01	0,03	0,01	0,00	0,00	1,00	0,06	0,01	0,00	0,00	0,01	0,00
Sum of policies that were substituted (same LOB purchased within a 7 day window of another policy ending) for a policy with a higher PCA	sum_sub_increase_rcom	0,05	0,07	0,04	0,03	0,06	0,06	0,00	0,06	1,00	-0,01	0,00	0,01	0,00	-0,01
Average of current month PCA minus the average of homologous month PCA	slope_avg_pca_nf_rcom	-0,01	0,00	-0,01	-0,06	-0,04	-0,02	0,00	0,01	-0,01	1,00	0,01	0,00	0,00	0,00
Sum of claims value year to date	sum_claim_ytd_nf_rcom	0,00	-0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,01	1,00	-0,89	-0,90	0,03
Sum of current month's technical margin minus the sum of homologous month's technical margin	slope_tech_margin_nf	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	-0,89	1,00	0,58	-0,82
A client's value in the current month	tech_margin_nf_a	0,02	0,00	0,01	0,01	0,01	0,02	0,00	0,01	0,00	0,00	-0,90	0,58	1,00	-0,01
Value of the client in the homologous month of the past year	tech_margin_nf_h	0,01	0,00	0,00	0,01	0,01	0,01	0,00	0,00	-0,01	0,00	0,03	-0,82	-0,01	1,00

Table 9 - Correlation among value variables

In order to meet the business expectations weights were assigned to different variables. To assign weights all variables were standardized by subtracting the mean and dividing them by the standard deviation. Then the variables of a client's current value of a client's financial product holding and the current value of a client are multiplied by four, and the client's average PCA of all active policies is multiplied by two. The latter is only multiplied by two, because its correlation with other variables already increases its effect on the segmentation. As explained in Variables for the Demographic Segmentation this amplifies the variance of the weighted variables, and k-means will react accordingly.

At next the clusters that resulted from the k-means algorithm will be explained.

After several iterations the following four clusters emerged:

1. Financial Client: Customers in this segment are characterized by their premiums for financial products that clearly differentiates them from customers in other segments.
2. Value Loosing Clients: on average these clients cause losses.
3. Standard Client: The value performance of these clients is close to the global average.
4. Very Profitable Clients: This segment is the most valuable. Clients tend to cause great profits.

Figure 12 gives information about the number of customers in each of the value segments. It shows that value loosing clients (1.003.246) and standard clients (944.498) are the biggest segments, whereas financial clients (81.878) and very profitable clients (245.556) are smaller segments.

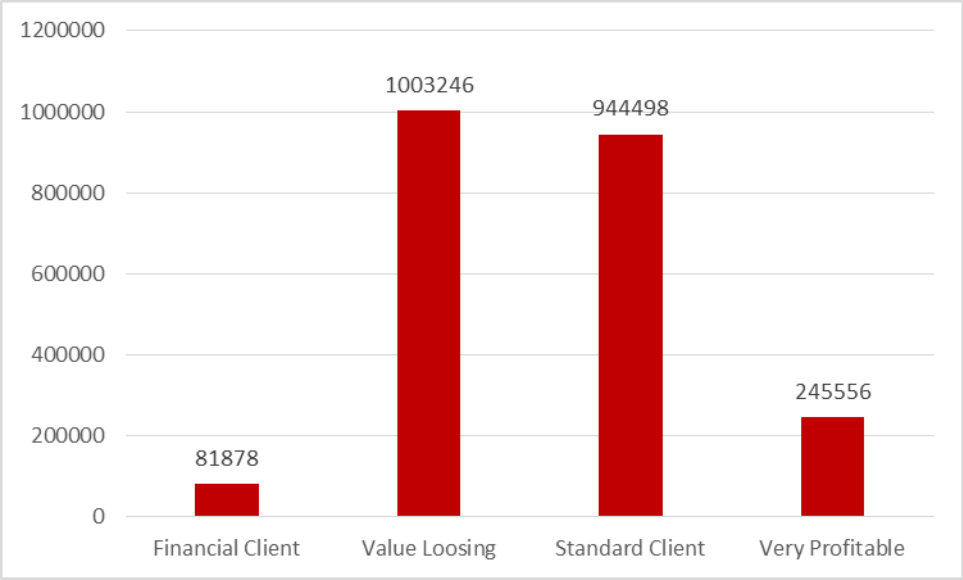


Figure 12 - Number of clients in value segments

Figure 13 displays the average current value of financial products for each segment. This graphic is most capable of characterizing financial clients. Their financial product's value (73.233.10 €) is a multiple of the average value of the other segments. This indicates that financial clients have different interests than the other clients which should be reflected by the strategy these customers are targeted by.

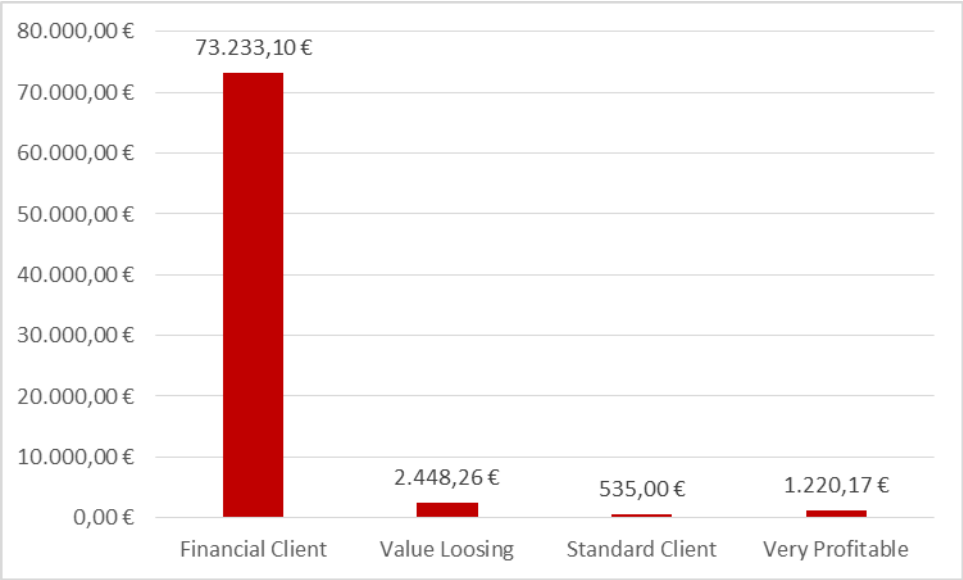


Figure 13 - Average current value of financial products per client

Figure 14 displays the average profit/loss a client causes by segment. The average profit of the population is 53.01€. Very profitable clients cause about six times the average profit (623.04€), while value loosing clients cause a loss of 139.27€. Financial Clients have a value that is slightly lower than average (39.22€). It can be positively noted, that standard clients have a solid above average profit of 146.47€.

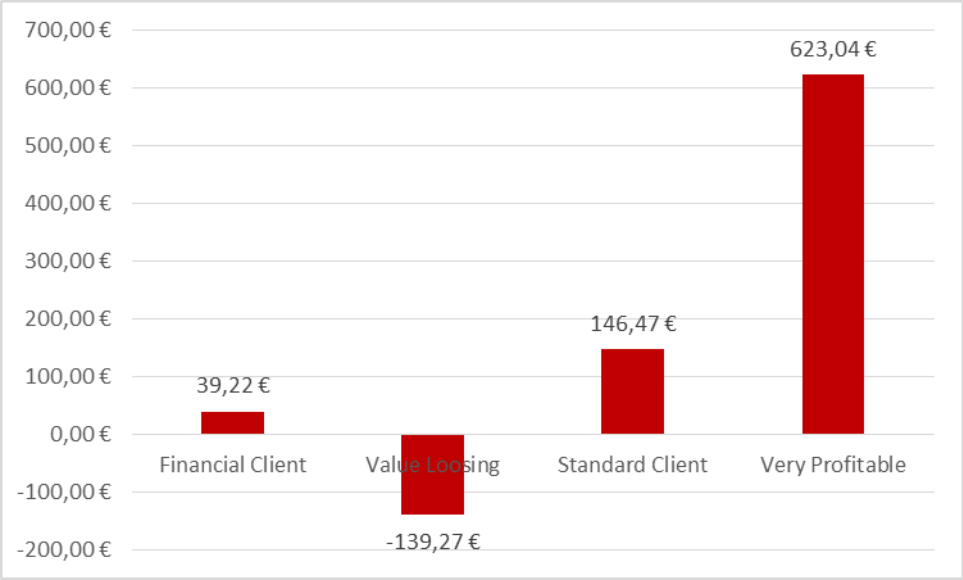


Figure 14 - Average technical margin by value segment

The variables that are discussed above serve best to describe the characteristics of each segment. However a complete list of averages for all segments from all used features is given in the appendix (Profiling of Value segments).

Although the cluster selection was based on qualitative criteria, quantitative cluster validity measures should be presented. Two of the Internal Validation Techniques are applied. Due to the size of the data set, they are only applied to a 5% randomly chosen fraction of the overall population.

The first measure that is computed is the CH-index. In the 5% sample it reaches a SSB value of 2,159,960, and a SSW value of 3,186,619. Further it considered 113,758 observations with 14 dimensions. Applying the formula of the CH-index to those values, a level of 5,930.63 has been reached.

By comparing SSB (2,159,960) to TSS (5,346,579), it can be observed, that the segmentation accounts for 40.4% of the overall variation. For a segmentation using only four clusters for 113,758 observations in 14 dimensions, this is a good result. Even when it is considered that the partly high correlation among the input variables causes the proportion of explained variance to increase, this is a reasonable result. The segments serve well as an indicator for the client's value.

To provide a second measure, the Dunn-index is computed. In the value segmentation the maximum intra-cluster distance is 36.92. It is measured as the maximum distance from an observation to its centroid. The minimal inter-cluster distance is measured between value loosing clients and standard clients. Its amount is 3.5. The Dunn-index evaluates to 0.0948, which signifies that there is at least one observation that are far more distant from its centroid than the distance between the two closest clusters.

In order to provide the reader with a graphical overview of the segmentation a two dimensional scatterplot for the 5% sample is presented. The two dimensions are the two first principal components. Due to the partly high correlation of the input variables (see Table 9 - Correlation among value variables), these first two principal components are able to account for 27.44% of the total variation in the 14 variables.

The first observation that can be made from Figure 15 is that in the two dimensional plot no real clusters are visible. This is astonishing given the fact, that 40.4% of the overall variation is expressed by the cluster centroids. Nevertheless, the absence of real clusters is an obstacle for the segmentation and leads to difficulties when clustering the observations. Figure 15 also provides a visual impression of how the clusters can be separated. Areas of the clusters of value loosing clients (green), standard clients (blue), and very profitable clients (purple) are visually separated. Financial clients (red), however, are spread across the whole plot. This is probably due to its reliance on variables that are not captured well by the first two principal components. A discussion of the resulting principal components is not part of this work. They only serve to visualize the cluster separation and overall distribution of data points.

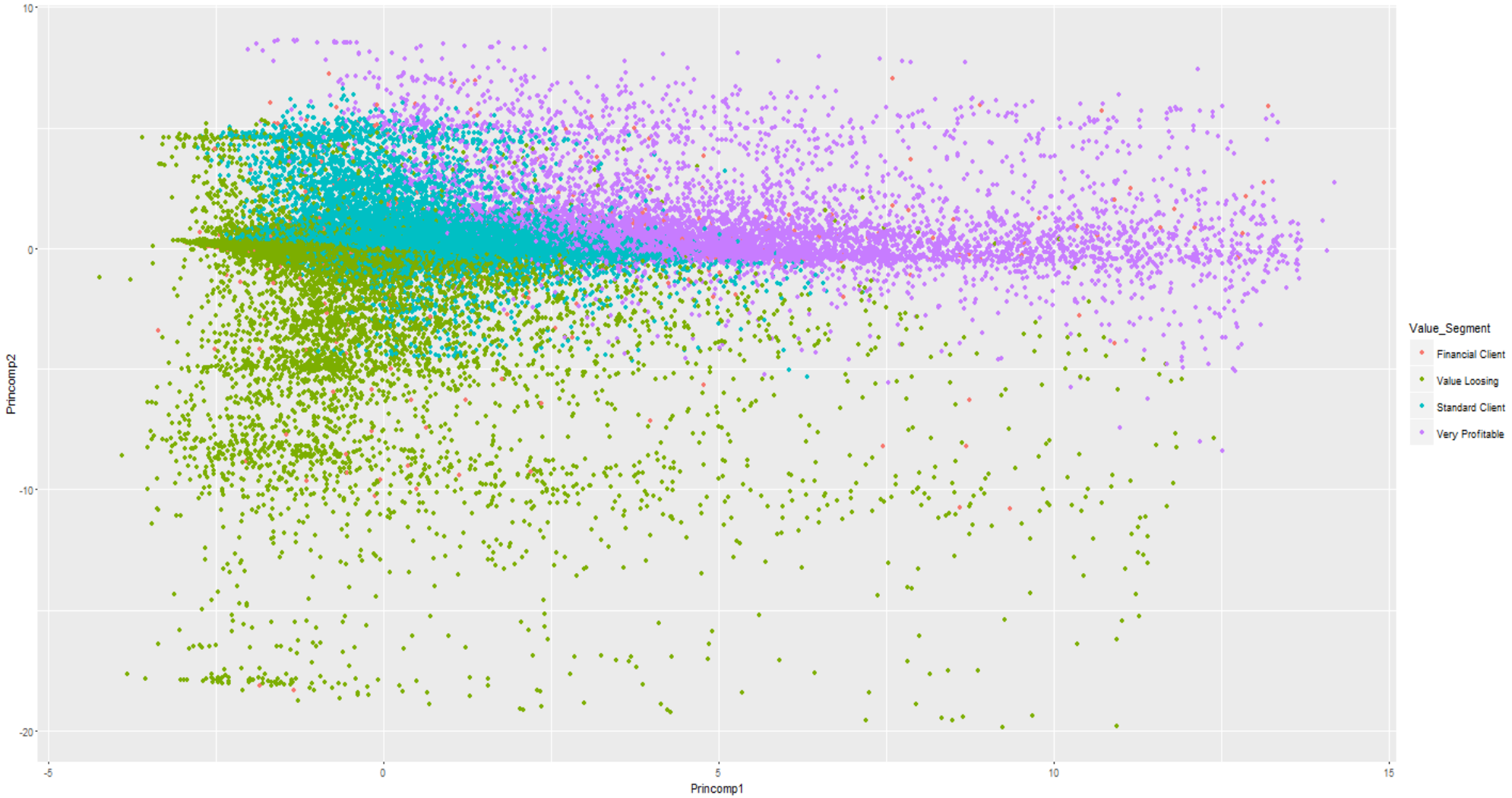


Figure 15 - Scatter plot of first two principal components for value segmentation

4.3. FINAL DELIVERABLE

After the completion of the segmentation for all dimensions, the profiling for all segments was performed. This profiling did not only include the created segments (also referred to as macro-segments), but also micro-segments that resulted when macro-segments are crossed. When crossing the six demographic segments with the five behavior segments and the four value segments, a total of 199 micro- and macro-segments emerged. This includes also microsegments, which resulted when crossing only two dimensions, and the total population. In order to give a full profile of all microsegments two dashboard have been created. One dealing with indicators characterizing the segments in the environment of the whole enterprise, and one that characterizes the segments according to their health indicators.

The enterprise wide dashboard provides the following information:

- Sizes of the population, the segment, number of individual person, and number of ENI
- Client status (new, recent, old), marital status, tenure, gender
- Age distribution
- LOB that has been used for the purchases of the first product (Health, Home, Auto, Life Risk, Life Other, Other)
- Policy distribution (Health, Home, Auto, Life Risk, Life Other, Other)
- Contact channel distribution (Banc, Agent, Broker, Other)
- Financial Summary with total values (PCA per LOB and technical margin)
- Financial Summary with average values (PCA per LOB and technical margin)
- Portfolio size (average number of active policies, average number of active product families)

The analysis should focus on health variables. Due to that a second dashboard was created, providing the following information for health clients:

- Segment Size, percentages of ENI, individual person, health policy owner that also own a policy in another LOB, health title owner that also own a policy in another LOB, clients that are insured in a health policy and also own a policy in another LOB, clients that are insure in a health policy but are not title owner of a health policy, clients that are insure in a health policy and are also title owner of a health policy, clients that are insured in a health policy of an individual person, clients that are insured in a health policy of a corporate
- Average tenure of policies, average tenure of health policies
- Client status (new, recent, old), marital status, gender
- Age distribution
- The insurance issues different kinds of cards for their health clients. Cards basically signify the quality of the health insurance. The distribution these cards is given in the dashboard

- For each card type the percentage of clients that did not use it in the past three years
- Every health coverage has a limit up to with a claim is covered. The proportion of the limit that was actually used by the client was investigated for each quarter after the renovation of a policy and for every coverage type. All clients have the option of using a special network of hospitals that offers special conditions for clients of the insurance. The usage within and outside the network is displayed separately.
- For usage of the network the insurance has special discounts for their clients. Graphics that give conclusion about the discounted and undiscounted value of payments to hospitals inside and outside the network are implemented in the dashboard. As well as the top three coverages that cause the highest discounted and undiscounted payments for each segment are available.
- A compilation of the top five coverages that are owned by the greatest proportion of health clients is given
- A drilldown for the top five owned coverages that also displays the card type (as explained above) is implemented in the dashboard
- A comparison of the time of cancellation, relative to the beginning of a contract, is performed between health policies and policies of other LOBs
- All claims are associated to one coverage type, the relative frequency for each coverage is given in the dashboard
- For each claim a specialty (more detailed description of the claim) is recorded. The dashboard contains the top five ambulant specialties and top three in-patient specialties are projected on the dashboard
- Another statistic gives conclusion about the most used coverages for clients that have a specific kind of card

Using these two dashboards a variety of different characteristics can be investigated and compared across micro-segments. Examples of findings are not part of this work since they are confidential.

5. RESULTS AND LIMITATIONS

In the end the goal of the project was achieved. The customer base has been segmented in a way that satisfies the insurance. New insights could be gained based on the delivered dashboards and a base for future investigation has been established with the segmentation. Further, business strategies can now be aligned with the special needs of a specific customer segments.

Although the solution was successful and the client agreed with the major outcomes of the project, there are some recommendations for possible improvements in the future. In the presented case most of the external statistics that were obtained were specific for a geographical region (or post code area), and thus did not have much influence on a customer level, but rather on a regional level. To differentiate clients that live in the same area, they did not provide additional information.

One more thing that might improve the segmentation is the outlier treatment. The percentile method classifies outlier too rigorously. It is possible that the readjustment of the top and bottom percentile destroys characteristics of the population. A case by case decision to detect outliers based on for example the LOF might lead to better results.

Apart from that, the outcome of the segmentation could be improved by improving the quality of data. The highest potential improvement is in the notes from phone calls. If the company focuses on a better interpretability of the notes, the behavior of the clients can be anticipated more precisely and so the notes inherit more valuable insights.

It has been pointed out that k-means was the algorithm of choice because it was the most applicable one to the size of the data set. If more computational resources are provided other algorithms might be applied. For example it can be expected that the results of a hierarchical algorithm form stronger cluster in the sense of a smaller intra-cluster distances. However in order to apply hierarchical algorithms on a data set of this size computational power must increase significantly. The time complexity of hierarchical algorithms is quadratic and in test runs a five percent sample was considered too big to apply hierarchical clustering.

It should also be noted that this project could be carried out in a different manner when the purpose was to create a more descriptive segmentation. In this case strict boundaries for the correlation of used variables and percentage of missing values would be installed. However for the sake of this project the interpretability of the segments had top priority, and therefor variables needed to be weighted, and also variables could enter the segmentation that did not satisfy the installed criteria.

Also the segmentation can only be seen as a first step to gain insights. Due to the nature of the used variables that capture past information, it does not have predictive power. Predictive models need to be implemented in order to validate the findings from the segmentation, and to assign a numerical probability for, for example, churn and cross-sell opportunities. The performed segmentation already created the ground on which such a predictive model can be grown. Different models can be calculated for different segments, based on the specific needs.

Another thing that is necessary to do is to establish a clear scope of the project in the cooperation with a client company. This is necessary because the quality of the solution is suffering when redefining the environment and goal during the course of the project.

One more thing that is important to note is, that the undertaken project considers values as off 31.08.2015. Thus the created model is time-invariant and will not change over time. However the underlying customer base will chenage. This means than after a certain time it is recommended to create a new version of the analysis to capture new insights and see if new segments emerge or old segments disappear.

6. BIBLIOGRAPHY

Acuna, E., & Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *International IPSI 2004 Conference, Symposium on Challenges in Internet and Interdisciplinary Research*.

Al-Daoud, A., (2005). *A New Algorithm for Cluster Initialization*, Proc. of the 2nd World Enformatika Conf., 2005, 74–76.

Anderberg, M. R., (1973). *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY.

Astrahan, M. M. (1970). *Speech analysis by clustering, or the hyperphoneme method* (No. AIM-124). Stanford Univ Calif Dept of Computer Science.

Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral science*, *12*(2), 153-155.

Barnett, V., & Lewis, T. (1994). Outliers in statistical data. *International Journal of Forecasting,* 12(1), 175-176.

Bay, S., & Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 29–38.

Ben-Gal, I. (2005). *Outlier detection*. In Data mining and knowledge discovery handbook (pp. 131 146). Springer US.

Bolshakova, N., & Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing*, 83(4), 825–833.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *Proceedings of the 2000 Acm Sigmod International Conference on Management of Data*, 1–12.

Brown, T. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications, 1-38.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Cao, F., Liang, J., & Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers and Mathematics with Applications*, 58(3), 474–483.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1), 200 -10.

Charikar, M., Chekuri, C., Feder, T., & Motwani, R. (2004). Incremental Clustering and Dynamic Information Retrieval. *SIAM Journal on Computing*, 33(6), 1417–1440.

Cho, M., Lee, J., & Lee, K. M. (2009). Feature correspondence and deformable object matching via agglomerative correspondence clustering. *Proceedings of the IEEE International Conference on Computer Vision*, (Iccv), 1280–1287.

Cimiano, P., Hotho, A., & Staab, S. (2004). Comparing Conceptual , Divisive and Agglomerative Clustering for Learning Taxonomies from Text. *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, Including Prestigious Applicants of Intelligent Systems*, 16, 435–439.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. *19th International Conference on Pattern Recognition*.

Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *Computer Science Department of State University of New York (SUNY): Technical Report 02-2*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.

Dawyndt, P., De Meyer, H., & De Baets, B. (2006). UPGMA clustering revisited: A weight-driven approach to transitive approximation. *International Journal of Approximate Reasoning*, 42(3), 174–191.

de Hoon, M. J. L., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9), 1453–1454.

Deza, M. M., & Deza, E. (2009). *Encyclopedia of distances*. Springer Berlin Heidelberg, (pp. 1-583).

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well Separated Clusters, *Journal of Cybernetics,* **3** (3), 32–57.

Färber, I., Günnemann, S., Kriegel, H.-P., Kröger, P., Müller, E., Schubert, E., … Zimek, A. (2010). On Using Class-Labels in Evaluation of Clusterings. *Proc. 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust 2010) in Conjunction with 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*, Washington, DC, USA, 9.

Fred, A. L. N., & Jain, a. K. (2002). Data clustering using evidence accumulation. *Object Recognition Supported by User Interaction for Service Robots*, 4, 276–280.

Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, *21*(3), 768-769.

Gonzalez, F. (1985). Clustering To Minimize Intercluster Distance. *Theoretical Computer Science*, 38, 293–306.

Gowda, K. C., & Krishna, G. (1977). Nonparametric clustering using the concept of mutual nearest neighbourhood. *Report No. EE*, *43*.

Gower, J. C., & Ross, G. J. S. (1969). Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1), 54-64.

Hartigan, J. a., & Wong, M. a. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1), 100–108.

Hasan, M. Al, Chaoji, V., Salem, S., & Zaki, M. J. (2009). Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, 30(11), 994–1002.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 1(3), 264–323.

Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany*, *14*(1), 127-130.

Kaufman, L., Rousseeuw, P., (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience.

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, *62*(317), 86-101.

Kou, L., Markowsky, G., & Berman, L. (1981). A fast algorithm for Steiner trees. *Acta Informatica*, 15(2), 141–145.

Lance, G. N. (1967). A general theory of classificatory sorting strategies: II. Clustering systems. *The Computer Journal*, 10(3), 271–277.

Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 99*, 5, 16–22.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of intenal clustering validation measures. *IEEE Internatinal Conference on Data Mining*, 911–916.

MacQueen, J. B. (1967). Kmeans Some Methods for classification and Analysis of Multivariate Observations. *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*, 1(233), 281–297.

Moh'd B, A. D., & Roberts, S. A. (1996). New methods for the initialisation of clusters. *Pattern Recognition Letters*, *17*(5), 451-455.

Onoda, T., Sakai, M., & Yamada, S. (2012). Careful seeding method based on independent components analysis for k-means clustering. *Journal of Emerging Technologies in Web Intelligence*, 4(1), 51–59.

P.S. Bradley, U.M. Fayyad, U. (1998). Refining Initial Points for K-Means Clustering. *ICML*, 91-99.

Pelleg, D., Pelleg, D., Moore, A. W., & Moore, A. W. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International Conference on Machine Learning Table of Contents*, 727–734.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1), 103–119.

Pontriagin, L. S., & Arhangel'skii, A. V. (1990). *General topology I: Basic concepts and constructions dimension theory*, Springer US.

Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Raftery, A. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society*, Series B, 48, 249-350.

Ray, S., & Turi, R. H. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, 137–143.

Redmond, S. J., & Heneghan, C. (2007). A method for initialising the K-means clustering algorithm using kd-trees. *Pattern Recognition Letters*, 28(8), 965–973.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus External cluster validation indexes. *International Journal of Computers and Communications*, 5(1), 27-34.

Rendón, E., Garcia, R., Abundez, I., Gutierrez, C., Gasca, E., Del Razo, F., & Gonzalez, A. (2008). NIVA: A robust cluster validity. *Proceedings of the 12th WSEAS International Conference on Communications*, 209–213.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C), 53–65.

Sarle, W. (1983). Cubic Clustering Criterion. SAS Institute. *Cary, NC*.

Shannon C. E., (1948). A Mathematical Theory of Communication, *The Bell System Technical Journal*, 379 – 423.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1), 30-34.

Sneath, P. H. A. (1957). The Application of Computers to Taxonomy. *Journal of General Microbiology*, 17(1), 201–226.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. *6th ACM SIGKDD, World Text Mining Conference*, 1–2.

Strehl, A., & Ghosh, J. (2002). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3, 583–617.

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of Similarity Measures on Web-page Clustering. *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 58 – 64.

Su, T., & Dy, J. (2007). In search of deterministic methods for initializing K-means and Gaussian mixture clustering. *Intelligent Data Analysis*, 1–42.

Tou, J. T., & Gonzalez, R. C. (1974). Pattern recognition. *Applied Mathematics and Computation, Addison-Wesley, MA*, 377 – 399.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.

Zahn. (1970). Graph-theoretical methods for detecting and describing gestalt clusters. *Journal of Chemical Information and Modeling*, 53(9), 1689–1699.

7. APPENDIX

7.1. LIST OF VARIABLES IN CAR

Note that several variables contain asterisks (*). Asterisks indicate a group of variables that have been calculated in the same way for different filters. Eg: ind_plan_* pinpoints all variables that indicate that the client owns a policy of type plan with a specific coverage that is indicated by the *.

Also note that whenever the name of the insurance appeared in a variable, it was changed to "INSURANCE" in order to be compliant with the confidentiality agreement with the insurance.

No	Variable	Label	Origin
1	ANT_AP	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "AC PESSOAIS"	BICP
2	ANT_ATCO	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "AT CONTA OUTREM"	BICP
3	ANT_ATCP	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "AT CONTA PROPRIA"	BICP
4	ANT_ATED	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "AT EMP DOM"	BICP
5	ANT_AUTO	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "AUTO"	BICP
6	ANT_CACA	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "CACA"	BICP
7	ANT_CAP	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "CAPITALIZACAO"	BICP
8	ANT_EMB	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "EMBARCACAO"	BICP
9	ANT_MRH	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "MRH"	BICP
10	ANT_MRN	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "MRN"	BICP
11	ANT_NVDO	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "NAO VIDA OUTROS"	BICP
12	ANT_PPR	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "PPR"	BICP
13	ANT_RCF	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "RC FAMILIAR"	BICP
14	ANT_RCG	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "RC GERAL"	BICP
15	ANT_SD	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "SAUDE"	BICP
16	ANT_VDO	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "VIDA OUTROS"	BICP
17	ANT_VDR	Antiguidade(anos),da apolice mais antiga em vigor, da familia de produto "VIDA RISCO"	BICP
18	ANTIGUID_CC_COND_HAB1	Antiguidade(anos),da apolice mais antiga em vigor, da	BICP

		família “CC_COND_HAB_1”	
19	ANTIGUIDADE	Antiguidade do Cliente	BICP
20	ANTIGUIDADE_OBJETO	Antiguidade do Objeto	BICP
21	ANU_APOFAM_AP	# de apolices anuladas p/Fam.Prod-Ac.Pessoais	BICP
22	ANU_APOFAM_ATCO	# de apolices anuladas p/Fam.Prod-AT Conta Outrem	BICP
23	ANU_APOFAM_ATCP	# de apolices anuladas p/Fam.Prod-AT Conta Própria	BICP
24	ANU_APOFAM_ATED	# de apolices anuladas p/Fam.Prod-AT EMP DOM	BICP
25	ANU_APOFAM_AUTO	# de apolices anuladas p/Fam.Prod-Auto	BICP
26	ANU_APOFAM_CACA	# de apolices anuladas p/Fam.Prod-Caça	BICP
27	ANU_APOFAM_CAP	# de apolices anuladas p/Fam.Prod-Capitalização	BICP
28	ANU_APOFAM_EMB	# de apolices anuladas p/Fam.Prod-Embarcação	BICP
29	ANU_APOFAM_MRH	# de apolices anuladas p/Fam.Prod-MRH	BICP
30	ANU_APOFAM_MRN	# de apolices anuladas p/Fam.Prod-MRN	BICP
31	ANU_APOFAM_NVDO	# de apolices anuladas p/Fam.Prod-Não Vida Outros	BICP
32	ANU_APOFAM_PPR	# de apolices anuladas p/Fam.Prod-PPR	BICP
33	ANU_APOFAM_RCF	# de apolices anuladas p/Fam.Prod-RC Familiar	BICP
34	ANU_APOFAM_RCG	# de apolices anuladas p/Fam.Prod-RC Geral	BICP
35	ANU_APOFAM_SD	# de apolices anuladas p/Fam.Prod-Saúde	BICP
36	ANU_APOFAM_VDO	# de apolices anuladas p/Fam.Prod-Vida Outros	BICP
37	ANU_APOFAM_VDR	# de apolices anuladas p/Fam.Prod-Vida Risco	BICP
38	ANU_EFEC_APOFAM_AP	# de apolices anuladas efect.p/Fam.Prod-Ac.Pessoais	BICP
39	ANU_EFEC_APOFAM_ATCO	# de apolices anuladas efect.p/Fam.Prod-AT Conta Outrem	BICP
40	ANU_EFEC_APOFAM_ATCP	# de apolices anuladas efect.p/Fam.Prod-AT Conta Própria	BICP
41	ANU_EFEC_APOFAM_ATED	# de apolices anuladas efect.p/Fam.Prod-AT EMP DOM	BICP
42	ANU_EFEC_APOFAM_AUTO	# de apolices anuladas efect.p/Fam.Prod-Auto	BICP
43	ANU_EFEC_APOFAM_CACA	# de apolices anuladas efect.p/Fam.Prod-Caça	BICP
44	ANU_EFEC_APOFAM_CAP	# de apolices anuladas efect.p/Fam.Prod-Capitalização	BICP
45	ANU_EFEC_APOFAM_EMB	# de apolices anuladas efect.p/Fam.Prod-Embarcação	BICP
46	ANU_EFEC_APOFAM_MRH	# de apolices anuladas efect.p/Fam.Prod-MRH	BICP
47	ANU_EFEC_APOFAM_MRN	# de apolices anuladas efect.p/Fam.Prod-MRN	BICP
48	ANU_EFEC_APOFAM_NVDO	# de apolices anuladas efect.p/Fam.Prod-Não Vida Outros	BICP
49	ANU_EFEC_APOFAM_PPR	# de apolices anuladas efect.p/Fam.Prod-PPR	BICP
50	ANU_EFEC_APOFAM_RCF	# de apolices anuladas efect.p/Fam.Prod-RC Familiar	BICP
51	ANU_EFEC_APOFAM_RCG	# de apolices anuladas efect.p/Fam.Prod-RC Geral	BICP
52	ANU_EFEC_APOFAM_SD	# de apolices anuladas efect.p/Fam.Prod-Saúde	BICP
53	ANU_EFEC_APOFAM_VDO	# de apolices anuladas efect.p/Fam.Prod-Vida Outros	BICP
54	ANU_EFEC_APOFAM_VDR	# de apolices anuladas efect.p/Fam.Prod-Vida Risco	BICP
55	ANU_EFEC_FATPAG_ULT12M	# de apolices anuladas por falta pagamento ult.12M	BICP
56	ANU_EFEC_FATPAG_ULT1M	# de apolices anuladas por falta pagamento ult.1M	BICP
57	ANU_EFEC_FATPAG_ULT36M	# de apolices anuladas por falta pagamento ult.36M	BICP
58	ANU_EFEC_N_MESES_2ANUL	Nº meses decorridos entre duas últimas anulações	BICP
59	ANU_EFEC_NMESES_FAM_ULTANUL_AP	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Ac.Pessoais	BICP

60	ANU_EFEC_NMESES_FAM_ULTANUL_ATCO	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-AT Conta Outrem	BICP
61	ANU_EFEC_NMESES_FAM_ULTANUL_ATCP	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-AT Conta Própria	BICP
62	ANU_EFEC_NMESES_FAM_ULTANUL_ATED	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-AT EMP DOM	BICP
63	ANU_EFEC_NMESES_FAM_ULTANUL_AUTO	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Auto	BICP
64	ANU_EFEC_NMESES_FAM_ULTANUL_CACA	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Caça	BICP
65	ANU_EFEC_NMESES_FAM_ULTANUL_CAP	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Capitalização	BICP
66	ANU_EFEC_NMESES_FAM_ULTANUL_EMB	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Embarcação	BICP
67	ANU_EFEC_NMESES_FAM_ULTANUL_MRH	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-MRH	BICP
68	ANU_EFEC_NMESES_FAM_ULTANUL_MRN	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-MRN	BICP
69	ANU_EFEC_NMESES_FAM_ULTANUL_NVDO	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Não Vida Outros	BICP
70	ANU_EFEC_NMESES_FAM_ULTANUL_PPR	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-PPR	BICP
71	ANU_EFEC_NMESES_FAM_ULTANUL_RCF	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-RC Familiar	BICP
72	ANU_EFEC_NMESES_FAM_ULTANUL_RCG	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-RC Geral	BICP
73	ANU_EFEC_NMESES_FAM_ULTANUL_SD	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Saúde	BICP
74	ANU_EFEC_NMESES_FAM_ULTANUL_VDO	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Vida Outros	BICP
75	ANU_EFEC_NMESES_FAM_ULTANUL_VDR	Nº meses decorridos desde ult.Anul.efect.p/Fam.Prod-Vida Risco	BICP
76	ANU_EFEC_NMESES_ULTANUL	Nº meses decorridos desde ult.Anul.efectiva	BICP
77	ANU_EFEC_PCAFAM_AP	Valor PCA das apol.anuladas efect.p/Fam.Prod-Ac.Pessoais	BICP
78	ANU_EFEC_PCAFAM_ATCO	Valor PCA das apol.anuladas efect.p/Fam.Prod-AT Conta Outrem	BICP
79	ANU_EFEC_PCAFAM_ATCP	Valor PCA das apol.anuladas efect.p/Fam.Prod-AT Conta Própria	BICP
80	ANU_EFEC_PCAFAM_ATED	Valor PCA das apol.anuladas efect.p/Fam.Prod-AT EMP DOM	BICP
81	ANU_EFEC_PCAFAM_AUTO	Valor PCA das apol.anuladas efect.p/Fam.Prod-Auto	BICP
82	ANU_EFEC_PCAFAM_CACA	Valor PCA das apol.anuladas efect.p/Fam.Prod-Caça	BICP
83	ANU_EFEC_PCAFAM_EMB	Valor PCA das apol.anuladas efect.p/Fam.Prod-Embarcação	BICP
84	ANU_EFEC_PCAFAM_MRH	Valor PCA das apol.anuladas efect.p/Fam.Prod-MRH	BICP
85	ANU_EFEC_PCAFAM_MRN	Valor PCA das apol.anuladas efect.p/Fam.Prod-MRN	BICP
86	ANU_EFEC_PCAFAM_NVDO	Valor PCA das apol.anuladas efect.p/Fam.Prod-Não Vida Outros	BICP
87	ANU_EFEC_PCAFAM_RCF	Valor PCA das apol.anuladas efect.p/Fam.Prod-RC Familiar	BICP
88	ANU_EFEC_PCAFAM_RCG	Valor PCA das apol.anuladas efect.p/Fam.Prod-RC Geral	BICP

89	ANU_EFEC_PCAFAM_SD	Valor PCA das apol.anuladas efect.p/Fam.Prod-Saúde	BICP
90	ANU_EFEC_PCAFAM_VDO	Valor PCA das apol.anuladas efect.p/Fam.Prod-Vida Outros	BICP
91	ANU_EFEC_PCAFAM_VDR	Valor PCA das apol.anuladas efect.p/Fam.Prod-Vida Risco	BICP
92	ANU_EFEC_PCATOT_ULT12M	Valor PCA das apolices anuladas efect.Ult.12M	BICP
93	ANU_EFEC_PCATOT_ULT1M	Valor PCA das apolices anuladas efect.Ult.1M	BICP
94	ANU_EFEC_PCATOT_ULT36M	Valor PCA das apolices anuladas efect.Ult.36M	BICP
95	ANU_EFEC_TOTAPO_ULT12M	# de apolices anuladas efect.ult.12M	BICP
96	ANU_EFEC_TOTAPO_ULT1M	# de apolices anuladas efect.ult.1M	BICP
97	ANU_EFEC_TOTAPO_ULT36M	# de apolices anuladas efect.ult.36M	BICP
98	ANU_NMESES_FAMPROD_AP	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Ac.Pessoais	BICP
99	ANU_NMESES_FAMPROD_ATCO	Nº meses decorridos desde ult.Anul. p/Fam.Prod-AT Conta Outrem	BICP
100	ANU_NMESES_FAMPROD_ATCP	Nº meses decorridos desde ult.Anul. p/Fam.Prod-AT Conta Própria	BICP
101	ANU_NMESES_FAMPROD_ATED	Nº meses decorridos desde ult.Anul. p/Fam.Prod-AT EMP DOM	BICP
102	ANU_NMESES_FAMPROD_AUTO	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Auto	BICP
103	ANU_NMESES_FAMPROD_CACA	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Caça	BICP
104	ANU_NMESES_FAMPROD_CAP	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Capitalização	BICP
105	ANU_NMESES_FAMPROD_EMB	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Embarcação	BICP
106	ANU_NMESES_FAMPROD_MRH	Nº meses decorridos desde ult.Anul. p/Fam.Prod-MRH	BICP
107	ANU_NMESES_FAMPROD_MRN	Nº meses decorridos desde ult.Anul. p/Fam.Prod-MRN	BICP
108	ANU_NMESES_FAMPROD_NVDO	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Não Vida Outros	BICP
109	ANU_NMESES_FAMPROD_PPR	Nº meses decorridos desde ult.Anul. p/Fam.Prod-PPR	BICP
110	ANU_NMESES_FAMPROD_RCF	Nº meses decorridos desde ult.Anul. p/Fam.Prod-RC Familiar	BICP
111	ANU_NMESES_FAMPROD_RCG	Nº meses decorridos desde ult.Anul. p/Fam.Prod-RC Geral	BICP
112	ANU_NMESES_FAMPROD_SD	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Saúde	BICP
113	ANU_NMESES_FAMPROD_VDO	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Vida Outros	BICP
114	ANU_NMESES_FAMPROD_VDR	Nº meses decorridos desde ult.Anul. p/Fam.Prod-Vida Risco	BICP
115	ANU_NMESES_ULTANUL	Nº meses decorridos desde ult.anulação	BICP
116	ANU_PCAFAM_AP	Valor PCA das apol.anuladas p/Fam.Prod-Ac.Pessoais	BICP
117	ANU_PCAFAM_ATCO	Valor PCA das apol.anuladas p/Fam.Prod-AT Conta Outrem	BICP

118	ANU_PCAFAM_ATCP	Valor PCA das apol.anuladas p/Fam.Prod-AT Conta Própria	BICP
119	ANU_PCAFAM_ATED	Valor PCA das apol.anuladas p/Fam.Prod-AT EMP DOM	BICP
120	ANU_PCAFAM_AUTO	Valor PCA das apol.anuladas p/Fam.Prod-Auto	BICP
121	ANU_PCAFAM_CACA	Valor PCA das apol.anuladas p/Fam.Prod-Caça	BICP
122	ANU_PCAFAM_EMB	Valor PCA das apol.anuladas p/Fam.Prod-Embarcação	BICP
123	ANU_PCAFAM_MRH	Valor PCA das apol.anuladas p/Fam.Prod-MRH	BICP
124	ANU_PCAFAM_MRN	Valor PCA das apol.anuladas p/Fam.Prod-MRN	BICP
125	ANU_PCAFAM_NVDO	Valor PCA das apol.anuladas p/Fam.Prod-Não Vida Outros	BICP
126	ANU_PCAFAM_RCF	Valor PCA das apol.anuladas p/Fam.Prod-RC Familiar	BICP
127	ANU_PCAFAM_RCG	Valor PCA das apol.anuladas p/Fam.Prod-RC Geral	BICP
128	ANU_PCAFAM_SD	Valor PCA das apol.anuladas p/Fam.Prod-Saúde	BICP
129	ANU_PCAFAM_VDO	Valor PCA das apol.anuladas p/Fam.Prod-Vida Outros	BICP
130	ANU_PCAFAM_VDR	Valor PCA das apol.anuladas p/Fam.Prod-Vida Risco	BICP
131	ANU_PCATOT_ULT12M	Valor PCA das apolices anuladas Ult.12M	BICP
132	ANU_PCATOT_ULT1M	Valor PCA das apolices anuladas Ult.1M	BICP
133	ANU_PCATOT_ULT36M	Valor PCA das apolices anuladas Ult.36M	BICP
134	ANU_TOTAPO_ULT12M	# de apolices anuladas ult.12M	BICP
135	ANU_TOTAPO_ULT1M	# de apolices anuladas ult.1M	BICP
136	ANU_TOTAPO_ULT36M	# de apolices anuladas ult.36M	BICP
137	apol_novas_12meses_AP	Nº de Apolice Novas nos ultimos 12 meses da familia de produto AP	BICP
138	apol_novas_12meses_ATCO	Nº de Apolice Novas nos ultimos 12 meses da familia de produto ATCO	BICP
139	apol_novas_12meses_ATCP	Nº de Apolice Novas nos ultimos 12 meses da familia de produto ATCP	BICP
140	apol_novas_12meses_AUTO	Nº de Apolice Novas nos ultimos 12 meses da familia de produto AUTO	BICP
141	apol_novas_12meses_CACA	Nº de Apolice Novas nos ultimos 12 meses da familia de produto CACA	BICP
142	apol_novas_12meses_CAP	Nº de Apolice Novas nos ultimos 12 meses da familia de produto CAP	BICP
143	apol_novas_12meses_EMB	Nº de Apolice Novas nos ultimos 12 meses da familia de produto EMB	BICP
144	apol_novas_12meses_MRH	Nº de Apolice Novas nos ultimos 12 meses da familia de produto MRH	BICP
145	apol_novas_12meses_MRN	Nº de Apolice Novas nos ultimos 12 meses da familia de produto MRN	BICP
146	apol_novas_12meses_NVDO	Nº de Apolice Novas nos ultimos 12 meses da familia de produto NVDO	BICP
147	apol_novas_12meses_PPR	Nº de Apolice Novas nos ultimos 12 meses da familia de produto PPR	BICP
148	apol_novas_12meses_RCF	Nº de Apolice Novas nos ultimos 12 meses da familia de produto RCF	BICP
149	apol_novas_12meses_RCG	Nº de Apolice Novas nos ultimos 12 meses da familia de produto RCG	BICP
150	apol_novas_12meses_SD	Nº de Apolice Novas nos ultimos 12 meses da familia de produto SD	BICP
151	apol_novas_12meses_VDO	Nº de Apolice Novas nos ultimos 12 meses da familia de	BICP

		produto VDO	
152	apol_novas_12meses_VDR	Nº de Apolice Novas nos ultimos 12 meses da familia de produto VDR	BICP
153	apolices_novas_12meses	Nº Apolices Novas nos ultimos 12 meses	BICP
154	apolices_novas_1mes	Nº Apolices Novas no ultimo mês	BICP
155	apolices_novas_36meses	Nº Apolices Novas nos ultimos 36 meses	BICP
156	AU_FLAG_SEM_DANOS_PP	Flag sem cobertura: danos próprios	BICP
157	AU_FLAG_SEM_VAL_AQUISC33	Indicator of aquesition without additional value	BICP
158	AU_nOUTROS_TIPOS_VEICULOS	Number of cehicles of type “others”	BICP
159	AU_nREBOQUES	Number of times client has been towed	BICP
160	AU_nSEGUROS_CARTA	Number of auto insurances of type “card”	BICP
161	AU_nVEICULOS_COMERCIAIS	Number of vehicles for commercial purposes	BICP
162	AU_nVEICULOS_LIGEIROS	Number of light vehicles	BICP
163	AU_nVEICULOS_MOTOCICLOS	Number of motorcycles owned	BICP
164	AU_nVEICULOS_PESADOS	Number of heavy vehicles	BICP
165	AU_OBJ_MAIS_RECENT	Most recent insured object	BICP
166	CANAL_GEST_CLIENTE	Canal de Gestor de Cliente	BICP
167	CANAL_MED_ULT_COMP	Canal Mediação- Ultima compra	BICP
168	CANAL_MEDNIV90_ULT_COMP	Canal Mediação NIV90 - Ultima compra	BICP
169	CANAL_PREFERENCIAL	Corresponde ao canal definido pela tabela Regras Leads	BICP
170	CHAVE	Chave do Registo	BICP
171	COD_CONCELHO	Código Concelho	BICP
172	CONCELHO_MORADA	Concelho da Morada	BICP
173	COUNT_UR_MR	# de objetos MRH em vigor (total)	BICP
174	CP4	Codigo postal 4 digitos	BICP
175	CP7	Codigo Postal	BICP
176	Decil_Fin	Prémio Comercial Anualizado	BICP
177	DENS_POP_CONC	Densidade populacional (N.º/ km²) por Concelho de residência	BICP
178	dtinform	Data de Informacao	BICP
179	EMIS_PCA_12M	Indicador com o somatório dos PCA de produtos subscritos em campanhas nos últimos 12 meses	BICP
180	EMIS_PCA_1M	Indicador com o somatório dos PCA de produtos subscritos em campanhas no último ultimo mês	BICP
181	EMIS_PCA_36M	Indicador com o somatório dos PCA de produtos subscritos em campanhas nos últimos 36 meses	BICP
182	ENQ_GEST_CLIENT	Enquadramento de Gestor de Cliente	BICP
183	ENQ_PREFERENCIAL	Corresponde ao enquadramento definido pela tabela Regras Leads	BICP
184	ENQ_ULT_COMP	Canal Mediação NIV90 Enquadramento - Ultima compra	BICP
185	ENT_PREFERENCIAL	Corresponde à entidade definida pela tabela Regras Leads	BICP
186	ENT_ULT_COMP	Numero da Entidade de Última Compra	BICP

187	ESTADO_CIVIL	Estado Civil da Entidade	BICP
188	ESTADO_CLIENTE	De acordo com a existência de apólices em vigor ou data de anulação/término da última apólice em vigor,como Tomador	BICP
189	FAIXA_ETARIA	Faixa Etaria de Entidade	BICP
190	FAM_PRIM_COMP	Familia de primeiro produto subscrito (primeira compra)	BICP
191	FAM_ULT_COMP	Família de último produto subscrito (compra mais recente)	BICP
192	FINC_CCENTER_NET	VALOR PROD CCENTER_NET	BICP
193	FINC_CGD	VALOR PROD CGD	BICP
194	FINC_CLOSE	VALOR PROD CLOSE	BICP
195	FINC_Conta_Entg_Cresc	Nr de Entgregas Prog Crescentes	BICP
196	FINC_CTT	VALOR PROD CTT	BICP
197	FINC_DC_AGENCIAS	VALOR PROD DC_AGENCIAS	BICP
198	FINC_DC_MEDIACAO	VALOR PROD DC_MEDIACAO	BICP
199	FINC_DCB	VALOR PROD DCB	BICP
200	FINC_DCC	VALOR PROD DCC	BICP
201	FINC_NrApol_Entg_Prog	Nr apolices com entregas programadas	BICP
202	FINC_NrApol_Entg_Unicas	Numero de apolices com entregas unicas	BICP
203	FINC_OPEN	VALOR PROD OPEN	BICP
204	FINC_TP_PROD	Tipo Produto Finc	BICP
205	FINC_Valor_Entg_Prog	Valor das entregas prog ultimos 12 meses	BICP
206	FLAG_AlvoCamp	Flag Alvo de Campanha	BICP
207	FLAG_ASS_VIAG_EXTRA		BICP
208	FLAG_CONT_DIR_OUT	Flag Contacto Direto Outbound	BICP
209	FLAG_CONTEUDO	Flag Conteudo	BICP
210	Flag_EmisCamp	Flag de Emissão em Campanha	BICP
211	FLAG_INCIDENCIA	Indicador da existencia de Incidência	BICP
212	FLAG_INIBICAO_CONTACTO	Indicador de Inibicao de Contacto Comercial	BICP
213	FLAG_OBJ_NOVO	Flag Objeto Novo C3.3	BICP
214	FLAG_OBJ_VALIOSOS	Flag Objetos Valiosos	BICP
215	FLAG_SEGURADO	Se tem pelo menos uma apolice como segurado/pessoa segura	BICP
216	FLAG_SEM_DIV_SEGURA	Flag sem cobertura: Dívida Segura	BICP
217	FLAG_SEM_FEN_SISMIC	Flag sem cobertura: Fenomenos Sismicos	BICP
218	FLAG_SEM_FURTO_ROUBO	Flag sem cobertura: Furto ou Roubo	BICP
219	FLAG_SEM_PERDA_TOTAL	Flag sem cobertura: Perda Total	BICP
220	FLAG_SEM_QIV	Flag sem cobertura: Assistência em Viagem Extra	BICP
221	FLAG_SEM_VEIC_SUBST	Flag sem cobertura: veiculo substituição	BICP
222	HSAPANU_APOFAM_AP	# apólices por Família Prod. Anuladas - Ac.Pessoais	BICP
223	HSAPANU_APOFAM_ATCO	# apólices por Família Prod. Anuladas - AT Conta Outrem	BICP
224	HSAPANU_APOFAM_ATCP	# apólices por Família Prod. Anuladas - AT Conta Própria	BICP
225	HSAPANU_APOFAM_ATED	# apólices por Família Prod. Anuladas - AT EMP DOM	BICP
226	HSAPANU_APOFAM_AUTO	# apólices por Família Prod. Anuladas - Auto	BICP
227	HSAPANU_APOFAM_CACA	# apólices por Família Prod. Anuladas - Caça	BICP
228	HSAPANU_APOFAM_CAP	# apólices por Família Prod. Anuladas - Capitalização	BICP
229	HSAPANU_APOFAM_EMB	# apólices por Família Prod. Anuladas - Embarcação	BICP

230	HSAPANU_APOFAM_MRH	# apólices por Família Prod. Anuladas - MRH	BICP
231	HSAPANU_APOFAM_MRN	# apólices por Família Prod. Anuladas - MRN	BICP
232	HSAPANU_APOFAM_NVDO	# apólices por Família Prod. Anuladas - Não Vida Outros	BICP
233	HSAPANU_APOFAM_PPR	# apólices por Família Prod. Anuladas - PPR	BICP
234	HSAPANU_APOFAM_RCF	# apólices por Família Prod. Anuladas - RC Familiar	BICP
235	HSAPANU_APOFAM_RCG	# apólices por Família Prod. Anuladas - RC Geral	BICP
236	HSAPANU_APOFAM_SD	# apólices por Família Prod. Anuladas - Saúde	BICP
237	HSAPANU_APOFAM_VDO	# apólices por Família Prod. Anuladas - Vida Outros	BICP
238	HSAPANU_APOFAM_VDR	# apólices por Família Prod. Anuladas - Vida Risco	BICP
239	HSAPANU_PCAAPOFAM_AP	Valor PCA das Apól.Anuladas p/Fam.Prod-Ac.Pessoais	BICP
240	HSAPANU_PCAAPOFAM_ATCO	Valor PCA das Apól.Anuladas p/Fam.Prod-AT Conta Outrem	BICP
241	HSAPANU_PCAAPOFAM_ATCP	Valor PCA das Apól.Anuladas p/Fam.Prod-AT Conta Própria	BICP
242	HSAPANU_PCAAPOFAM_ATED	Valor PCA das Apól.Anuladas p/Fam.Prod-AT EMP DOM	BICP
243	HSAPANU_PCAAPOFAM_AUTO	Valor PCA das Apól.Anuladas p/Fam.Prod-Auto	BICP
244	HSAPANU_PCAAPOFAM_CACA	Valor PCA das Apól.Anuladas p/Fam.Prod-Caça	BICP
245	HSAPANU_PCAAPOFAM_EMB	Valor PCA das Apól.Anuladas p/Fam.Prod-Embarcação	BICP
246	HSAPANU_PCAAPOFAM_MRH	Valor PCA das Apól.Anuladas p/Fam.Prod-MRH	BICP
247	HSAPANU_PCAAPOFAM_MRN	Valor PCA das Apól.Anuladas p/Fam.Prod-MRN	BICP
248	HSAPANU_PCAAPOFAM_NVDO	Valor PCA das Apól.Anuladas p/Fam.Prod-Não Vida Outros	BICP
249	HSAPANU_PCAAPOFAM_RCF	Valor PCA das Apól.Anuladas p/Fam.Prod-RC Familiar	BICP
250	HSAPANU_PCAAPOFAM_RCG	Valor PCA das Apól.Anuladas p/Fam.Prod-RC Geral	BICP
251	HSAPANU_PCAAPOFAM_SD	Valor PCA das Apól.Anuladas p/Fam.Prod-Saúde	BICP
252	HSAPANU_PCAAPOFAM_VDO	Valor PCA das Apól.Anuladas p/Fam.Prod-Vida Outros	BICP
253	HSAPANU_PCAAPOFAM_VDR	Valor PCA das Apól.Anuladas p/Fam.Prod-Vida Risco	BICP
254	HSAPANU_PCATOT	Valor PCA das Apólices Anuladas (Não Fin)	BICP
255	HSAPANU_TOT	# Apolices Anuladas	BICP
256	HSAPANU_TOTFAM	# famílias de produtos Anuladas	BICP
257	HSAPCAD_APOFAM_AP	# apólices por Família Prod.Caducadas - Ac.Pessoais	BICP
258	HSAPCAD_APOFAM_ATCO	# apólices por Família Prod.Caducadas - AT Conta Outrem	BICP
259	HSAPCAD_APOFAM_ATCP	# apólices por Família Prod.Caducadas - AT Conta Própria	BICP
260	HSAPCAD_APOFAM_AUTO	# apólices por Família Prod.Caducadas - Auto	BICP
261	HSAPCAD_APOFAM_CACA	# apólices por Família Prod.Caducadas - Caça	BICP
262	HSAPCAD_APOFAM_MRH	# apólices por Família Prod.Caducadas - MRH	BICP
263	HSAPCAD_APOFAM_NVDO	# apólices por Família Prod.Caducadas - Não Vida Outros	BICP
264	HSAPCAD_APOFAM_RCF	# apólices por Família Prod.Caducadas - RC Familiar	BICP

265	HSAPCAD_APOFAM_RCG	# apólices por Família Prod.Caducadas - RC Geral	BICP
266	HSAPCAD_PCATOT	Valor PCA das Apólices Caducadas(Não Fin)	BICP
267	HSAPCAD_TOT	# Apolices Caducadas	BICP
268	HSAPCAD_TOTFAM	# famílias de produtos Caducadas	BICP
269	IDADE	Idade	BICP
270	IND_APL_GRP	Indicadores de apolices grupo	BICP
271	Indicador_Email	Indicador de Email	BICP
272	Indicador_Telefone_Fixo	Indicador de Telefone Fixo	BICP
273	Indicador_Telemovel	Indicador de Telemovel	BICP
274	INDICE_RENDIMENTO	INDICE_RENDIMENTO	BICP
275	LR_SEM_CONTEUDO	# de locais de risco MRH com objetos tipo fracção e imóvel e sem objetos tipo conteudo (todos em vigor)	BICP
276	LR_SEM_PAREDES	# de locais de risco MRH com objetos tipo conteudo e sem objetos tipo fracção ou imovel (todos em vigor)	BICP
277	MAX_NRLUGAR	# Max Lugares	BICP
278	MAX_VFIN_NC	Max de valor de aplicação em simulações seg Financeiro não concretizado	BICP
279	meses_2_ult_compras	Nº meses decorridos entre as 2 últimas compras da entidade	BICP
280	meses_apol_nova_ult_compra_AP	Nº meses decorridos desde a ultima compra da familia de produto AP	BICP
281	meses_apol_nova_ult_compra_ATCO	Nº meses decorridos desde a ultima compra da familia de produto ATCO	BICP
282	meses_apol_nova_ult_compra_ATCP	Nº meses decorridos desde a ultima compra da familia de produto ATCP	BICP
283	meses_apol_nova_ult_compra_AUTO	Nº meses decorridos desde a ultima compra da familia de produto AUTO	BICP
284	meses_apol_nova_ult_compra_CACA	Nº meses decorridos desde a ultima compra da familia de produto CACA	BICP
285	meses_apol_nova_ult_compra_CAP	Nº meses decorridos desde a ultima compra da familia de produto CAP	BICP
286	meses_apol_nova_ult_compra_EMB	Nº meses decorridos desde a ultima compra da familia de produto EMB	BICP
287	meses_apol_nova_ult_compra_MRH	Nº meses decorridos desde a ultima compra da familia de produto MRH	BICP
288	meses_apol_nova_ult_compra_MRN	Nº meses decorridos desde a ultima compra da familia de produto MRN	BICP
289	meses_apol_nova_ult_compra_NVDO	Nº meses decorridos desde a ultima compra da familia de produto NVDO	BICP
290	meses_apol_nova_ult_compra_PPR	Nº meses decorridos desde a ultima compra da familia de produto PPR	BICP
291	meses_apol_nova_ult_compra_RCF	Nº meses decorridos desde a ultima compra da familia de produto RCF	BICP
292	meses_apol_nova_ult_compra_RCG	Nº meses decorridos desde a ultima compra da familia de produto RCG	BICP
293	meses_apol_nova_ult_compra_SD	Nº meses decorridos desde a ultima compra da familia de produto SD	BICP
294	meses_apol_nova_ult_compra_VDO	Nº meses decorridos desde a ultima compra da familia de produto VDO	BICP

295	meses_apol_nova_ult_compra_VDR	Nº meses decorridos desde a ultima compra da familia de produto VDR	BICP
296	MOV_VEMIT_ENT_PROG_12M	Montante subscrito de entregas programadas financeiros, últimos 12 meses	BICP
297	MOV_VEMIT_ENT_PROG_1M	Montante subscrito de entregas programadas financeiros, último mês	BICP
298	MOV_VEMIT_ENT_PROG_36M	Montante subscrito de entregas programadas financeiros, últimos 36 meses	BICP
299	MRH_CAP_HAB_PRINC	Capital da casa dada como habitação principal	BICP
300	MRH_CAPIT_CASA_MAIS_VALIOSA	Capital da casa mais valiosa	BICP
301	MRH_FLAG_COB_FEN_SISM	Flag Fenomenos sismicos	BICP
302	MRH_NR_LOCRISCO	# de locais de risco MRH distintos em vigor	BICP
303	MRH_SUM_CAPIT_LRISCO	Somatorio do capital seguro dos locais de risco distintos com MRH em vigor	BICP
304	MRH_SUM_CONTEUDO	Values of contrats in force that type “conteudo”	BICP
305	MRH_SUM_OUTROS	Values of contrats in force that type “outros”	BICP
306	MRH_SUM_PAREDES	Values of contrats in force that type “paredes”	BICP
307	NACIONALIDADE	Nacionalidade	BICP
308	NIV90_GEST_CLIENTE	Codigo NIV90 de Gestor de Cliente	BICP
309	NIV90_PREFERENCIAL	Correspondente ao canal definido pela tabela Regras Leads	BICP
310	NMESES_ULT_ANULACAO	Número de meses decorridos desde última anulação	BICP
311	NMESES_ULT_COMPRA	Número de meses decorridos desde última compra	BICP
312	NMESES_US_NC	Número de meses decorridos desde última simulação não convertida	BICP
313	NMESES_US_NC_AP	Número de meses decorridos desde última simulação não convertida AP	BICP
314	NMESES_US_NC_AT	Número de meses decorridos desde última simulação não convertida AT	BICP
315	NMESES_US_NC_AUTO	Número de meses decorridos desde última simulação não convertida AUTO	BICP
316	NMESES_US_NC_HAH	Número de meses decorridos desde última simulação não convertida HAH	BICP
317	NMESES_US_NC_MHS	Número de meses decorridos desde última simulação não convertida MHS	BICP
318	NMESES_US_NC_MR	Número de meses decorridos desde última simulação não convertida MR	BICP
319	NMESES_US_NC_VD	Número de meses decorridos desde última simulação não convertida VD	BICP
320	NR_LR_COM_PARED_CONTEUD	# de locais de risco MRH com objetos tipo fracção e imóvel e com objetos tipo conteudo (todos em vigor)	BICP
321	NR_OBJ_DETIDOS	Number of objects in dept	BICP
322	Num_AlvoCamp_12M	Número de campanhas de que foi alvo nos últimos 12 meses	BICP
323	Num_AlvoCamp_1M	Número de campanhas de que foi alvo no último mês	BICP
324	Num_AlvoCamp_36M	Número de campanhas de que foi alvo nos últimos 36 meses	BICP
325	NUM_CANAIS	Número de canais distintos	BICP
326	NUM_CONTAS_N80	Número de contas grupo NIV80	BICP
327	NUM_CONTAS_N90	Número de contas grupo NIV90	BICP

328	Num_EmisCamp_12M	Número de emissões em campanha nos últimos 12 meses	BICP
329	Num_EmisCamp_1M	Número de emissões em campanha no último mês	BICP
330	Num_EmisCamp_36M	Número de emissões em campanha nos últimos 36 meses	BICP
331	NUM_ENT_CLIENT	Numero de Entidades do Cliente	BICP
332	NUM_SIMUL_12M	Número de Simulações, últimos 12 meses	BICP
333	NUM_SIMUL_1M	Número de Simulações, último mês	BICP
334	NUM_SIMUL_24M	Número de Simulações, últimos 24 meses	BICP
335	NUM_SIMUL_AP_12M	Número de Simulações AP, últimos 12 meses	BICP
336	NUM_SIMUL_AT_12M	Número de Simulações AT, últimos 12 meses	BICP
337	NUM_SIMUL_AUTO_12M	Número de Simulações AUTO, últimos 12 meses	BICP
338	NUM_SIMUL_CA_12M	Número de simulações convertidas ajustadas, últimos 12meses	BICP
339	NUM_SIMUL_CA_1M	Número de simulações convertidas ajustadas, último mês	BICP
340	NUM_SIMUL_CA_24M	Número de simulações convertidas ajustadas, últimos 24meses	BICP
341	NUM_SIMUL_CA_AP_12M	Número de simulações convertidas ajustadas AP, últimos 12 meses	BICP
342	NUM_SIMUL_CA_AT_12M	Número de simulações convertidas ajustadas AT, últimos 12 meses	BICP
343	NUM_SIMUL_CA_AU_12M	Número de simulações convertidas ajustadas AUTO, últimos 12 meses	BICP
344	NUM_SIMUL_CA_HAH_12M	Número de simulações convertidas ajustadas HAH, últimos 12 meses	BICP
345	NUM_SIMUL_CA_MHS_12M	Número de simulações convertidas ajustadas MHS, últimos 12 meses	BICP
346	NUM_SIMUL_CA_MR_12M	Número de simulações convertidas ajustadas MR, últimos 12 meses	BICP
347	NUM_SIMUL_CA_VD_12M	Número de simulações convertidas ajustadas VD, últimos 12 meses	BICP
348	NUM_SIMUL_CONV_12M	Número de simulações convertidas, nos últimos 12 meses	BICP
349	NUM_SIMUL_HAH_12M	Número de Simulações HAH, últimos 12 meses	BICP
350	NUM_SIMUL_MHS_12M	Número de Simulações MHS, últimos 12 meses	BICP
351	NUM_SIMUL_MR_12M	Número de Simulações MR, últimos 12 meses	BICP
352	NUM_SIMUL_VD_12M	Número de Simulações VD, últimos 12 meses	BICP
353	NUMERO_VEICULOS	Nº de veiculos	BICP
354	NUMERO_VEICULOS_COM_DP	Numero de veiculos com DP	BICP
355	OBJ_COMP_CR_HIPOT	# objetos comprados com credor hipotecário	BICP
356	OBJETO_VAL_ACT	Objeto - Valor Actual - COGEN3.3	BICP
357	PAG_NRMDPAG	Nº medio dias pagamento recibo premio continuado após emissão Mês	BICP
358	PAG_NRMESUPAG	Nº meses desde último pagamento premio e cancelamento premio Mês	BICP
359	PAG_PERC_RECCBTAG_DD_1M	% Recibos de premio pagos por meio de pagamento (TaG) DD, mês	BICP
360	PAG_PERC_RECCBTAG_DIRETOS_1M	% Recibos de premio pagos por meio de pagamento (TaG) Diretos Mês	BICP
361	PAG_PERC_RECCBTAG_MEDI	% Recibos de premio pagos por meio de pagamento (TaG)	BICP

	ADOR_1M	MEDIADOR Mês	
362	PAG_RecCanEsEmit_12M	Nº recibos estornos e cancelamento de estornos emitidos, últimos 12 meses	BICP
363	PAG_RecCanEsEmit_1M	Nº recibos estornos e cancelamento de estornos emitidos, último mês	BICP
364	PAG_RecCanEsEmit_36M	Nº recibos estornos e cancelamento de estornos emitidos, últimos 36 meses	BICP
365	PAG_RecCanPrCob_12M	Nº recibos de prémios e cancelamento de prémios pagos, últimos 12 meses	BICP
366	PAG_RecCanPrCob_1M	Nº recibos de prémios e cancelamento de prémios pagos, último mês	BICP
367	PAG_RecCanPrCob_36M	Nº recibos de prémios e cancelamento de prémios pagos, últimos 36 meses	BICP
368	PAG_RecCanPrEmit_12M	Nº recibo de prémios e cancelamento de prémios emitidos, últimos 12 meses	BICP
369	PAG_RecCanPrEmit_1M	Nº recibo de prémios e cancelamento de prémios emitidos, último mês	BICP
370	PAG_RecCanPrEmit_36M	Nº recibo de prémios e cancelamento de prémios emitidos, últimos 36 meses	BICP
371	PAG_RECCBTAG_DD_12M	Nº Recibos de premio pagos por meio de pagamento (TaG) DD, últimos 12 meses	BICP
372	PAG_RECCBTAG_DD_1M	Nº Recibos de premio pagos por meio de pagamento (TaG) DD, últimos mês	BICP
373	PAG_RECCBTAG_DD_36M	Nº Recibos de premio pagos por meio de pagamento (TaG) DD, últimos 36 meses	BICP
374	PAG_RECCBTAG_DIRETOS_12M	Nº Recibos de premio pagos por meio de pagamento (TaG) Diretos, últimos 12 meses	BICP
375	PAG_RECCBTAG_DIRETOS_1M	Nº Recibos de premio pagos por meio de pagamento (TaG) Diretos, último mês	BICP
376	PAG_RECCBTAG_DIRETOS_36M	Nº Recibos de premio pagos por meio de pagamento (TaG) Diretos, últimos 36 meses	BICP
377	PAG_RECCBTAG_MEDIADOR_12M	Nº Recibos de premio pagos por meio de pagamento (TaG) Mediador, últimos 12 meses	BICP
378	PAG_RECCBTAG_MEDIADOR_1M	Nº Recibos de premio pagos por meio de pagamento (TaG) Mediador, último mês	BICP
379	PAG_RECCBTAG_MEDIADOR_36M	Nº Recibos de premio pagos por meio de pagamento (TaG) Mediador, últimos 36 meses	BICP
380	PAG_RECCOBRDT_12M	Nº Recibos de prémios pagos após data de vencimento, últimos 12 meses	BICP
381	PAG_RECCOBRDT_1M	Nº Recibos de pémios pagos após data de vencimento, último mês	BICP
382	PAG_RECCOBRDT_36M	Nº Recibos de prémios pagos após data de vencimento, últimos 36 meses	BICP
383	PAG_VARECCBR_12M	Valor total de Recibos Cobrados, últimos 12 meses	BICP
384	PAG_VARECCBR_1M	Valor total de Recibos Cobrados, último mês	BICP
385	PAG_VARECCBR_36M	Valor total de Recibos Cobrados, últimos 36 meses	BICP
386	PAG_VARECCBRDT_12M	Valor total recibos de prémio pagos após vencimento, últimos 12 meses	BICP
387	PAG_VARECCBRDT_1M	Valor total recibos de prémio pagos após vencimento, último	BICP

		mês	
388	PAG_VARECCBRDT_36M	Valor total recibos de prémio pagos após vencimento, últimos 36 meses	BICP
389	PAG_VAREEMIT_12M	Valor Total Recibos Emitidos, últimos 12 meses	BICP
390	PAG_VAREEMIT_1M	Valor Total Recibos Emitidos, último mês	BICP
391	PAG_VAREEMIT_36M	Valor Total Recibos Emitidos, últimos 36 meses	BICP
392	PAG_VRECPCBR_12M	Valor de Recibos por Cobrar, último 12 meses	BICP
393	PAG_VRECPCBR_1M	Valor de Recibos por Cobrar, último mês	BICP
394	PAG_VRECPCBR_36M	Valor de Recibos por Cobrar, último 36 meses	BICP
395	pca_apol_novas_12meses	PCA de apólices novas dos ultimos 12 meses	BICP
396	pca_apol_novas_12meses_AP	PCA de Apolice Novas nos ultimos 12 meses da familia de produto AP	BICP
397	pca_apol_novas_12meses_ATCO	PCA de Apolice Novas nos ultimos 12 meses da familia de produto ATCO	BICP
398	pca_apol_novas_12meses_ATCP	PCA de Apolice Novas nos ultimos 12 meses da familia de produto ATCP	BICP
399	pca_apol_novas_12meses_AUTO	PCA de Apolice Novas nos ultimos 12 meses da familia de produto AUTO	BICP
400	pca_apol_novas_12meses_CACA	PCA de Apolice Novas nos ultimos 12 meses da familia de produto CACA	BICP
401	pca_apol_novas_12meses_CAP	PCA de Apolice Novas nos ultimos 12 meses da familia de produto CAp	BICP
402	pca_apol_novas_12meses_EMB	PCA de Apolice Novas nos ultimos 12 meses da familia de produto EMB	BICP
403	pca_apol_novas_12meses_MRH	PCA de Apolice Novas nos ultimos 12 meses da familia de produto MRH	BICP
404	pca_apol_novas_12meses_MRN	PCA de Apolice Novas nos ultimos 12 meses da familia de produto MRN	BICP
405	pca_apol_novas_12meses_NVDO	PCA de Apolice Novas nos ultimos 12 meses da familia de produto NVDO	BICP
406	pca_apol_novas_12meses_PPR	PCA de Apolice Novas nos ultimos 12 meses da familia de produto PPR	BICP
407	pca_apol_novas_12meses_RCF	PCA de Apolice Novas nos ultimos 12 meses da familia de produto RCF	BICP
408	pca_apol_novas_12meses_RCG	PCA de Apolice Novas nos ultimos 12 meses da familia de produto RCG	BICP
409	pca_apol_novas_12meses_SD	PCA de Apolice Novas nos ultimos 12 meses da familia de produto SD	BICP
410	pca_apol_novas_12meses_VDO	PCA de Apolice Novas nos ultimos 12 meses da familia de produto VDO	BICP
411	pca_apol_novas_12meses_VDR	PCA de Apolice Novas nos ultimos 12 meses da familia de produto VDR	BICP
412	pca_apol_novas_1mes	PCA de apólices novas do ultimo mes	BICP
413	pca_apol_novas_36meses	PCA de apólices novas dos ultimos 36 meses	BICP
414	PERC_QUADROS_CONC	Proporção de profissionais socialmente mais valorizados (%) por Concelho de residência	BICP
415	POP_RESID_CONC	População residente (N.º) por Concelho de residência	BICP
416	POTENCIA	Veículo - Potência	BICP
417	PP_VC_A	Valor do Cliente, mês atual	BICP

418	PP_VC_A_AP	Valor do Cliente, mês atual, familia de prod AP	BICP
419	PP_VC_A_ATCO	Valor do Cliente, mês atual, familia de prod ATCO	BICP
420	PP_VC_A_ATCP	Valor do Cliente, mês atual, familia de prod ATCP	BICP
421	PP_VC_A_ATED	Valor do Cliente, mês atual, familia de prod ATED	BICP
422	PP_VC_A_AUTO	Valor do Cliente, mês atual, familia de prod AUTO	BICP
423	PP_VC_A_CACA	Valor do Cliente, mês atual, familia de prod CACA	BICP
424	PP_VC_A_EMB	Valor do Cliente, mês atual, familia de prod EMB	BICP
425	PP_VC_A_MRH	Valor do Cliente, mês atual, familia de prod MRH	BICP
426	PP_VC_A_MRN	Valor do Cliente, mês atual, familia de prod MRN	BICP
427	PP_VC_A_NVDO	Valor do Cliente, mês atual, familia de prod NVDO	BICP
428	PP_VC_A_RCF	Valor do Cliente, mês atual, familia de prod RCF	BICP
429	PP_VC_A_RCG	Valor do Cliente, mês atual, familia de prod RCG	BICP
430	PP_VC_A_SD	Valor do Cliente, mês atual, familia de prod SD	BICP
431	PP_VC_H	Valor do Cliente, mês homólogo	BICP
432	PP_VC_H_AP	Valor do Cliente, mês homólogo, familia de prod AP	BICP
433	PP_VC_H_ATCO	Valor do Cliente, mês homólogo, familia de prod ATCO	BICP
434	PP_VC_H_ATCP	Valor do Cliente, mês homólogo, familia de prod ATCP	BICP
435	PP_VC_H_ATED	Valor do Cliente, mês homólogo, familia de prod ATED	BICP
436	PP_VC_H_AUTO	Valor do Cliente, mês homólogo, familia de prod AUTO	BICP
437	PP_VC_H_CACA	Valor do Cliente, mês homólogo, familia de prod CACA	BICP
438	PP_VC_H_EMB	Valor do Cliente, mês homólogo, familia de prod EMB	BICP
439	PP_VC_H_MRH	Valor do Cliente, mês homólogo, familia de prod MRH	BICP
440	PP_VC_H_MRN	Valor do Cliente, mês homólogo, familia de prod MRN	BICP
441	PP_VC_H_NVDO	Valor do Cliente, mês homólogo, familia de prod NVDO	BICP
442	PP_VC_H_RCF	Valor do Cliente, mês homólogo, familia de prod RCF	BICP
443	PP_VC_H_RCG	Valor do Cliente, mês homólogo, familia de prod RCG	BICP
444	PP_VC_H_SD	Valor do Cliente, mês homólogo, familia de prod SD	BICP
445	PP_VPREMIOS_12M	Prémios Processados 12 Meses	BICP
446	PP_VPREMIOS_1M	Prémios Processados Mês	BICP
447	PP_VPREMIOS_36M	Prémios Processados 36 Meses	BICP
448	PPAPOLVIG_12M	# Apolices Vigor - 12M	BICP
449	PPAPOLVIG_1M	# Apolices Vigor - 1M	BICP
450	PPAPOLVIG_ACT	# Apolices Vigor - Hoje	BICP
451	PPCABAZ	Cabaz	BICP
452	PPFAMPROD_12M	# Familias Produtos com Apolices Vigor - 12M	BICP
453	PPFAMPROD_1M	# Familias Produtos com Apolices Vigor - 1M	BICP
454	PPFAMPROD_ACT	# Familias Produtos com Apolices Vigor - Hoje	BICP
455	PPINDPSFP_AP	Indic.Pessoa Segura por Familia Prod - Ac.Pessoais	BICP
456	PPINDPSFP_SD	Indic.Pessoa Segura por Familia Prod - Saúde	BICP
457	PPINDPSFP_VDR	Indic.Pessoa Segura por Familia Prod - Vida Risco	BICP
458	PPPCA_FAMPROD_AP	Tot.PCA Apolices Vigor por Familia Prod.- Ac.Pessoais	BICP
459	PPPCA_FAMPROD_ATCO	Tot.PCA Apolices Vigor por Familia Prod.- AT Conta Outrem	BICP
460	PPPCA_FAMPROD_ATCP	Tot.PCA Apolices Vigor por Familia Prod.- AT Conta Própria	BICP
461	PPPCA_FAMPROD_ATED	Tot.PCA Apolices Vigor por Familia Prod.- AT EMP DOM	BICP
462	PPPCA_FAMPROD_AUTO	Tot.PCA Apolices Vigor por Familia Prod.- Auto	BICP

463	PPPCA_FAMPROD_CACA	Tot.PCA Apolices Vigor por Familia Prod.- Caça	BICP
464	PPPCA_FAMPROD_EMB	Tot.PCA Apolices Vigor por Familia Prod.- Embarcação	BICP
465	PPPCA_FAMPROD_MRH	Tot.PCA Apolices Vigor por Familia Prod.- MRH	BICP
466	PPPCA_FAMPROD_MRN	Tot.PCA Apolices Vigor por Familia Prod.- MRN	BICP
467	PPPCA_FAMPROD_NVDO	Tot.PCA Apolices Vigor por Familia Prod.- Não Vida Outros	BICP
468	PPPCA_FAMPROD_RCF	Tot.PCA Apolices Vigor por Familia Prod.- RC Familiar	BICP
469	PPPCA_FAMPROD_RCG	Tot.PCA Apolices Vigor por Familia Prod.- RC Geral	BICP
470	PPPCA_FAMPROD_SD	Tot.PCA Apolices Vigor por Familia Prod.- Saúde	BICP
471	PPPCA_FAMPROD_VDO	Tot.PCA Apolices Vigor por Familia Prod.- Vida Outros	BICP
472	PPPCA_FAMPROD_VDR	Tot.PCA Apolices Vigor por Familia Prod.- Vida Risco	BICP
473	PPPCA_PS	PCA das Apolices por tipo relação - pessoa segura	BICP
474	PPPCA_TM	PCA das Apolices por tipo relação - tomador	BICP
475	PPPCANAOFIN	Total de PCA das Apolices em Vigor(não financeiros)	BICP
476	PPPROVM_ANO_ANT	Valor Total Produtos Financeiros - 12M	BICP
477	PPPROVM_ATUAL	Valor Total Produtos Financeiros - Hoje	BICP
478	PPPROVM_MES_ANT	Valor Total Produtos Financeiros - 1M	BICP
479	PPRAC_APOL_VIG	Rácio Apólices em vigor	BICP
480	PPRAC_FP_AP	Rácio Apólices em vigor por Famlia Prod - Ac.Pessoais	BICP
481	PPRAC_FP_ATCO	Rácio Apólices em vigor por Famlia Prod - AT Conta Outrem	BICP
482	PPRAC_FP_ATCP	Rácio Apólices em vigor por Famlia Prod - AT Conta Própria	BICP
483	PPRAC_FP_ATED	Rácio Apólices em vigor por Famlia Prod - AT EMP DOM	BICP
484	PPRAC_FP_AUTO	Rácio Apólices em vigor por Famlia Prod - Auto	BICP
485	PPRAC_FP_CACA	Rácio Apólices em vigor por Famlia Prod - Caça	BICP
486	PPRAC_FP_CAP	Rácio Apólices em vigor por Famlia Prod - Capitalização	BICP
487	PPRAC_FP_EMB	Rácio Apólices em vigor por Famlia Prod - Embalagem	BICP
488	PPRAC_FP_MRH	Rácio Apólices em vigor por Famlia Prod - MRH	BICP
489	PPRAC_FP_MRN	Rácio Apólices em vigor por Famlia Prod - MRN	BICP
490	PPRAC_FP_NVDO	Rácio Apólices em vigor por Famlia Prod - Não Vida Outros	BICP
491	PPRAC_FP_PPR	Rácio Apólices em vigor por Famlia Prod - PPR	BICP
492	PPRAC_FP_RCF	Rácio Apólices em vigor por Famlia Prod - RC Família	BICP
493	PPRAC_FP_RCG	Rácio Apólices em vigor por Famlia Prod - RC Geral	BICP
494	PPRAC_FP_SD	Rácio Apólices em vigor por Famlia Prod - Saúde	BICP
495	PPRAC_FP_VDO	Rácio Apólices em vigor por Famlia Prod - Vida Outros	BICP
496	PPRAC_FP_VDR	Rácio Apólices em vigor por Famlia Prod - Vida Risco	BICP
497	PPRAC_PRODFIN	Rácio de Apólices Produtos Financ. em Vigor	BICP
498	PPTOTFAMPROD_AP	# Apolices Vigor por Familia Prod.- Ac.Pessoais	BICP
499	PPTOTFAMPROD_ATCO	# Apolices Vigor por Familia Prod.- AT Conta Outrem	BICP
500	PPTOTFAMPROD_ATCP	# Apolices Vigor por Familia Prod.- AT Conta Própria	BICP
501	PPTOTFAMPROD_ATED	# Apolices Vigor por Familia Prod.- AT EMP DOM	BICP
502	PPTOTFAMPROD_AUTO	# Apolices Vigor por Familia Prod.- Auto	BICP
503	PPTOTFAMPROD_CACA	# Apolices Vigor por Familia Prod.- Caça	BICP
504	PPTOTFAMPROD_CAP	# Apolices Vigor por Familia Prod.- Capitalização	BICP
505	PPTOTFAMPROD_EMB	# Apolices Vigor por Familia Prod.- Embarcação	BICP
506	PPTOTFAMPROD_MRH	# Apolices Vigor por Familia Prod.- MRH	BICP
507	PPTOTFAMPROD_MRN	# Apolices Vigor por Familia Prod.- MRN	BICP

508	PPTOTFAMPROD_NVDO	# Apolices Vigor por Familia Prod.- Não Vida Outros	BICP
509	PPTOTFAMPROD_PPR	# Apolices Vigor por Familia Prod.- PPR	BICP
510	PPTOTFAMPROD_RCF	# Apolices Vigor por Familia Prod.- RC Familiar	BICP
511	PPTOTFAMPROD_RCG	# Apolices Vigor por Familia Prod.- RC Geral	BICP
512	PPTOTFAMPROD_SD	# Apolices Vigor por Familia Prod.- Saúde	BICP
513	PPTOTFAMPROD_VDO	# Apolices Vigor por Familia Prod.- Vida Outros	BICP
514	PPTOTFAMPROD_VDR	# Apolices Vigor por Familia Prod.- Vida Risco	BICP
515	PPTOTPS_APOVIG	# Pessoas seguras em Apólices em Vigor	BICP
516	PPTOTTPFRAC_ANUAL	# Apolices por tipo Fracionamento - Anual	BICP
517	PPTOTTPFRAC_MENSAL	# Apolices por tipo Fracionamento - Mensal	BICP
518	PPTOTTPFRAC_SEMESTRAL	# Apolices por tipo Fracionamento - Semestral	BICP
519	PPTOTTPFRAC_TRIMESTRAL	# Apolices por tipo Fracionamento - Trimestral	BICP
520	PPTOTTPFRAC_UNICA	# Apolices por tipo Fracionamento - Única	BICP
521	PPTOTTPPAG_DD	# Apolices por tipo Pagamento - DD	BICP
522	PPTOTTPPAG_DIRETA	# Apolices por tipo Pagamento - Direta	BICP
523	PPTOTTPREL_PS	Numero de contractos como pessoa segura/aderente	BICP
524	PPTOTTPREL_TM	Número Contractos como Tomador	BICP
525	PreVendas_ap_AP	# apólices AP com subscrição (não em vigor)	BICP
526	PreVendas_ap_ATCO	# apólices ATCO com subscrição (não em vigor)	BICP
527	PreVendas_ap_ATCP	# apólices ATCP com subscrição (não em vigor)	BICP
528	PreVendas_ap_AUTO	# apólices AUTO com subscrição (não em vigor)	BICP
529	PreVendas_ap_CACA	# apólices CACA com subscrição (não em vigor)	BICP
530	PreVendas_ap_CAP	# apólices CAP com subscrição (não em vigor)	BICP
531	PreVendas_ap_MRH	# apólices MRH com subscrição (não em vigor)	BICP
532	PreVendas_ap_MRN	# apólices MRN com subscrição (não em vigor)	BICP
533	PreVendas_ap_NVDO	# apólices NVDO com subscrição (não em vigor)	BICP
534	PreVendas_ap_PPR	# apólices PPR com subscrição (não em vigor)	BICP
535	PreVendas_ap_RCF	# apólices RCF com subscrição (não em vigor)	BICP
536	PreVendas_ap_RCG	# apólices RCG com subscrição (não em vigor)	BICP
537	PreVendas_ap_SD	# apólices SD com subscrição (não em vigor)	BICP
538	PreVendas_ap_VDR	# apólices VDR com subscrição (não em vigor)	BICP
539	PreVendas_apolices	# de apólices com subscrição (não em vigor)	BICP
540	PreVendas_fm_prod	# familias de produtos com subscrição (não em vigor)	BICP
541	PreVendas_pca	Somatório de pca subscritos (não em vigor)	BICP
542	PreVendas_pca_AP	Somatório de pca AP subscritos (não em vigor)	BICP
543	PreVendas_pca_ATCO	Somatório de pca ATCO subscritos (não em vigor)	BICP
544	PreVendas_pca_ATCP	Somatório de pca ATCP subscritos (não em vigor)	BICP
545	PreVendas_pca_AUTO	Somatório de pca AUTO subscritos (não em vigor)	BICP
546	PreVendas_pca_CACA	Somatório de pca CACA subscritos (não em vigor)	BICP
547	PreVendas_pca_CAP	Somatório de pca CAP subscritos (não em vigor)	BICP
548	PreVendas_pca_MRH	Somatório de pca MRH subscritos (não em vigor)	BICP
549	PreVendas_pca_MRN	Somatório de pca MRN subscritos (não em vigor)	BICP
550	PreVendas_pca_NVDO	Somatório de pca NVDO subscritos (não em vigor)	BICP
551	PreVendas_pca_RCF	Somatório de pca RCF subscritos (não em vigor)	BICP
552	PreVendas_pca_RCG	Somatório de pca RCG subscritos (não em vigor)	BICP

553	PreVendas_pca_SD	Somatório de pca SD subscritos (não em vigor)	BICP
554	PreVendas_pca_VDR	Somatório de pca VDR subscritos (não em vigor)	BICP
555	PROFISSAO	Profissão	BICP
556	PROP_AUT_CONC	Proporção de utilização do automóvel nas deslocações (%) por Concelho de residência	BICP
557	RACIO_CAV_OBJETO_NV_VALIOSO	Valuable object of type “none life”	BICP
558	RANK_MRH_CAP_CASA_MAIS_VALIOSA	Decil do montante da casa mais valiosa	BICP
559	RANK_VA_OBJ_MAIS_VALIOSO	Decil do valor (campo "Valor do objeto mais valioso com seg em vigor") atual da UR mais valiosa em vigor	BICP
560	RECHEIO_CASA_MAIS_VALIOSA	Capital do recheio da casa mais valiosa	BICP
561	REGIAO	Litoral ou Interior	BICP
562	ROT_MED_OBJ	rotação média de objetos	BICP
563	SALES_INDEX	SALES_INDEX	BICP
564	SEMAFORO	Classificação do Cliente (Abordagem Comercial)	BICP
565	SEXO	Sexo da Entidade	BICP
566	SinAbert	#aberturas de sinistros	BICP
567	SinFinAbert	#aberturas sinistros financeiros	BICP
568	SinFinMorte	montantes pagos p tipo sinistro financeiro Morte	BICP
569	SinFinResgParc	montantes pagos p tipo sinistro financeiro Resgate Parcial	BICP
570	SinFinResgTot	montantes pagos p tipo sinistro financeiro Resgate Total	BICP
571	SinFinTranfer	montantes pagos p tipo sinistro financeiro Transferência	BICP
572	SinFinVencim	montantes pagos p tipo sinistro financeiro Vencimento	BICP
573	SinNFinAbert	#aberturas sinistros não financeiros	BICP
574	SinNFinAbertAP	#sinistros abertos p família produto não financeiro AP	BICP
575	SinNFinAbertATCO	#sinistros abertos p família produto não financeiro ATCO	BICP
576	SinNFinAbertATCP	#sinistros abertos p família produto não financeiro ATCP	BICP
577	SinNFinAbertATED	#sinistros abertos p família produto não financeiro ATED	BICP
578	SinNFinAbertAUTO	#sinistros abertos p família produto não financeiro AUTO	BICP
579	SinNFinAbertCACA	#sinistros abertos p família produto não financeiro CACA	BICP
580	SinNFinAbertEMB	#sinistros abertos p família produto não financeiro EMB	BICP
581	SinNFinAbertMRH	#sinistros abertos p família produto não financeiro MRH	BICP
582	SinNFinAbertMRN	#sinistros abertos p família produto não financeiro MRN	BICP
583	SinNFinAbertNVDO	#sinistros abertos p família produto não financeiro NVDO	BICP
584	SinNFinAbertRCF	#sinistros abertos p família produto não financeiro RCF	BICP
585	SinNFinAbertRCG	#sinistros abertos p família produto não financeiro RCG	BICP
586	SinNFinAbertSD	#sinistros abertos p família produto não financeiro SD	BICP
587	SinNFinAbertVDO	#sinistros abertos p família produto não financeiro VDO	BICP
588	SinNFinAbertVDR	#sinistros abertos p família produto não financeiro VDR	BICP
589	SinNFinCstAbert	custo sinistros abertos não financeiros	BICP
590	SinNFinCstEmAbert	custo sinistro exercicio não financeiros em aberto	BICP
591	SinNFinEmAbert	#sinistros não financeiros em aberto	BICP
592	SinNFinFechCstAP	custo sinistro exercicio fechados no ano p familia produto não financeiro AP	BICP

593	SinNFinFechCstATCO	custo sinistro exercicio fechados no ano p familia produto não financeiro ATCO	BICP
594	SinNFinFechCstATCP	custo sinistro exercicio fechados no ano p familia produto não financeiro ATCP	BICP
595	SinNFinFechCstATED	custo sinistro exercicio fechados no ano p familia produto não financeiro ATED	BICP
596	SinNFinFechCstAUTO	custo sinistro exercicio fechados no ano p familia produto não financeiro AUTO	BICP
597	SinNFinFechCstCACA	custo sinistro exercicio fechados no ano p familia produto não financeiro CACA	BICP
598	SinNFinFechCstEMB	custo sinistro exercicio fechados no ano p familia produto não financeiro EMB	BICP
599	SinNFinFechCstMRH	custo sinistro exercicio fechados no ano p familia produto não financeiro MRH	BICP
600	SinNFinFechCstMRN	custo sinistro exercicio fechados no ano p familia produto não financeiro MRN	BICP
601	SinNFinFechCstNVDO	custo sinistro exercicio fechados no ano p familia produto não financeiro NVDO	BICP
602	SinNFinFechCstRCF	custo sinistro exercicio fechados no ano p familia produto não financeiro RCF	BICP
603	SinNFinFechCstRCG	custo sinistro exercicio fechados no ano p familia produto não financeiro RCG	BICP
604	SinNFinFechCstSD	custo sinistro exercicio fechados no ano p familia produto não financeiro SD	BICP
605	SinNFinFechCstVDO	custo sinistro exercicio fechados no ano p familia produto não financeiro VDO	BICP
606	SinNFinFechCstVDR	custo sinistro exercicio fechados no ano p familia produto não financeiro VDR	BICP
607	SinNFinMesUltAbert	#meses desde última abertura de sinistro não financeiro	BICP
608	SinNFinMesUltFech	#meses desde encerramento de último sinistro não financeiro	BICP
609	SUM_AU_PLANO_AUTOESTIMA	# apólices por plano	BICP
610	SUM_AU_PLANO_L	# apólices por plano	BICP
611	SUM_AU_PLANO_XL	# apólices por plano	BICP
612	SUM_AU_PLANO_XS	# apólices por plano	BICP
613	SUM_CAPIT_CONTEUDO	Somatório do capital seguro do recheio dos MRH em vigor	BICP
614	SUM_MRH_P_BASE	Nr de apolices em vigor por plano MRH usado	BICP
615	SUM_MRH_P_EXTRA	Nr de apolices em vigor por plano MRH usado	BICP
616	SUM_MRH_P_TOTAL	Nr de apolices em vigor por plano MRH usado	BICP
617	SUM_OF_OBJ_NOVO	# objetos comprados novos	BICP
618	SUM_OF_VEIC_VAL_ACT	Somatório do valor actual dos objetos com seg em vigor	BICP
619	TAXA_ACT_CONC	Taxa de actividade (%) da população residente por Concelho de residência	BICP
620	TIPO_CHAVE	Tipo de Registo Chave (NIF ou CR)	BICP
621	TIPO_CLIENTE	Tipo de Entidade	BICP
622	TIPO_PROD_SAUD	Tipo produto saude	BICP
623	TIPO_PRODUTO_AU	Tipo de produto auto detido	BICP
624	TIPO_PRODUTO_MR	Tipo de produto MRH detido	BICP

625	VA_OBJETO_MAIS_VALIOSO	Valor atual do objeto mais valioso com seg em vigor	BICP
626	VALOR_MAX_AMBULATORIO	Valor max de coberturas em ambulatorio	BICP
627	VALOR_MAX_ESTOMATOLOGIA	Valor max de coberturas em ESTOMATOLOGIA	BICP
628	VALOR_MAX_INTERNAMENTO	Valor max de coberturas em internamento	BICP
629	VEIC_MVALIOSO_NC	Valor automóvel mais valioso em simulações Auto não convertida	BICP
630	VEIC_VAL_NOVO	Veículo - Valor em Novo - COGEN3.3	BICP
631	ZONA_RESIDENCIA	A zona de residência é considerada Urbana, Semi-Urbana ou Rural de acordo com sua a densidade populacional (hab/km2)	BICP
632	profissao_type	Aggregated professions by the table BICP_PROF	BICP
633	ind_first_*	Indicate the * product line is in the list of first products in BICP's FAM_PRIM_COMP	BICP
634	cou_policy_rcom	Count all policies of the client	RCOM
635	sum_policy_new_rcom	Sum all policies of the client which are new	RCOM
636	sum_vig_rcom	Sum all policies of the client which are active	RCOM
637	sum_anul_rcom	Sum all policies of the client which are annulled	RCOM
638	ind_policy_new_rcom	Indicate client has a new policy	RCOM
639	ind_vig_rcom	Indicate client has has a policy in active	RCOM
640	ind_anul_rcom	Indicate client has as annulled policy	RCOM
641	sum_pca_rcom	Sum of all pca	RCOM
642	avg_pca_rcom	Average of all pca	RCOM
643	max_pca_rcom	Max of all pca	RCOM
644	min_pca_rcom	Min of all pca	RCOM
645	sum_vig_pca_rcom	Sum of all pca from active policies	RCOM
646	avg_vig_pca_rcom	Average of all pca from active policies	RCOM
647	max_vig_pca_rcom	Max of all pca from active policies	RCOM
648	min_vig_pca_rcom	Min of all pca from active policies	RCOM
649	sum_anul_pca_rcom	Sum of all pca from annulled policies	RCOM
650	sum_claim_ytd_rcom	Sum of cssinexe	RCOM
651	avg_claim_ytd_rcom	Average of cssinexe	RCOM
652	sum_com_tot_rcom	Sum of comtoapl	RCOM
653	avg_com_tot_rcom	Average of comtoapl	RCOM
654	min_mon_end_rcom	Min of months to end	RCOM
655	avg_maturity_mon_rcom	Average of months to end	RCOM
656	min_day_end_rcom	Min of days to end	RCOM
657	avg_maturity_day_rcom	Average of days to end	RCOM
658	sum_sub_rcom	Sum of policies which came from substituition	RCOM
659	sum_ren_rcom	Sum of policies which came from substituition or older than 1 year (renewal definition)	RCOM
660	sum_sub_vig_rcom	Sum of policies which came from substituition and are active	RCOM
661	sum_ren_vig_rcom	Sum of policies (which came from substituition or older than 1 year (renewal definition)) and are active	RCOM

662	min_months from beginning_rcom	Min of months from beginning	RCOM
663	min_mon_canc_rcom	Min of mon_canc	RCOM
664	max_mon_canc_rcom	Max of mon_canc	RCOM
665	avg_mon_canc_rcom	Avg of mon_canc	RCOM
666	mon_canc_0_2_rcom	Sum of cases when mon_canc between Jan and March and was not substituted	RCOM
667	mon_canc_3_5_rcom	Sum of cases when mon_canc between April and Jun and was not substituted	RCOM
668	mon_canc_6_8_rcom	Sum of cases when mon_canc between Jul and Sept and was not substituted	RCOM
669	mon_canc_9_11_rcom	Sum of cases when mon_canc between Oct and Dec and was not substituted	RCOM
670	mon_canc_0_2_1yr_rcom	Sum of cases when mon_canc between Jan and March of last year and was not substituted	RCOM
671	mon_canc_3_5_1yr_rcom	Sum of cases when mon_canc between April and Jun of last year and was not substituted	RCOM
672	mon_canc_6_8_1yr_rcom	Sum of cases when mon_canc between Jul and Sept of last year and was not substituted	RCOM
673	mon_canc_9_11_1yr_rcom	Sum of cases when mon_canc between Oct and Dec of last year and was not substituted	RCOM
674	cou_niv90_rcom	Number of agents that sold policies	RCOM
675	cou_niv80_rcom	Count distinct niv80	RCOM
676	cou_niv70_rcom	Count distinct niv70	RCOM
677	cou_niv60_rcom	Count distinct niv60	RCOM
678	cou_niv50_rcom	Count distinct niv50	RCOM
679	cou_niv40_rcom	Count distinct niv40	RCOM
680	cou_niv90_vig_rcom	Count distinct niv90	RCOM
681	cou_niv80_vig_rcom	Count distinct niv80	RCOM
682	cou_niv70_vig_rcom	Count distinct niv70	RCOM
683	cou_niv60_vig_rcom	Count distinct niv60	RCOM
684	cou_niv50_vig_rcom	Count distinct niv50	RCOM
685	cou_niv40_vig_rcom	Count distinct niv40	RCOM
686	sum_agnt_life_rcom	Number of agents that sold life policies	RCOM
687	sum_agnt_non_life_rcom	Number of agents that sold non-life policies	RCOM
688	sum_agnt_auto_rcom	Number of agents that sold auto policies	RCOM
689	sum_agnt_health_rcom	Number of agents that sold health policies	RCOM
690	sum_agnt_life_vig_rcom	Number of agents that sold life policies which are active	RCOM
691	sum_agnt_non_life_vig_rcom	Number of agents that sold non-life policies which are active	RCOM
692	sum_agnt_auto_vig_rcom	Number of agents that sold auto policies which are active	RCOM
693	sum_agnt_health_vig_rcom	Number of agents that sold health policies which are active	RCOM
694	frac_*_pol_rcom	Fraction of * LoB policies to the total number of policies	RCOM
695	frac_*_pol_vig_rcom	Fraction of *LoB active policies to the total number of active policies	RCOM
696	sum_BON_rcom	Sum of policies from Bonança corporac	RCOM

697	sum_FMS_rcom	Sum of policies from the insurance corporac	RCOM
698	sum_BON_vig_rcom	Sum of active policies from Bonança corporac	RCOM
699	sum_FMS_vig_rcom	Sum of active policies from the insurance corporac	RCOM
700	sum_*_frac_rcom	Number of policies with payments of *fractions	RCOM
701	sum_vig_*_frac_rcom	Number of active policies with payments of *fractions	RCOM
702	max_sit_prof_rcom	Max of the professional situation defined in the policy	RCOM
703	max_vig_gold_rcom	Indicates an active policy has the client markes as a gold client	RCOM
704	max_anul_gold_rcom	Indicates an annulled policy has the client markes as a gold client	RCOM
705	avg_anul_pca_rcom	Average annulled PCA	RCOM
706	max_anul_pca_rcom	Max annulled PCA	RCOM
707	min_anul_pca_rcom	Min annulled PCA	RCOM
708	max_claim_ytd_rcom	Max of policy claim value from the beginning of the year	RCOM
709	min_claim_ytd_rcom	Min of policy claim value from the beginning of the year	RCOM
710	sum_claim_vig_ytd_rcom	Sum of claims from the active policies from the beginning of the year	RCOM
711	sum_claim_hom_rcom	Sum of claims from the active policies from the beginning of the homologous year	RCOM
712	avg_claim_hom_rcom	Average of claims from the active policies from the beginning of the homologous year	RCOM
713	sum_pca_hom_rcom	Sum of the PCA of the homologous year	RCOM
714	avg_pca_hom_rcom	Average of the PCA of the homologous year	RCOM
715	sum_prems_rcom	Sum of the premiums since the begining of the year	RCOM
716	avg_prems_rcom	Average of the processed premiums since the begining of the year	RCOM
717	max_prems_rcom	Max of the processed premiums since the begining of the year	RCOM
718	min_prems_rcom	Min of the processed premiums since the begining of the year	RCOM
719	sum_prems_vig_rcom	Sum of the processed premiums since the begining of the year for the active policies	RCOM
720	max_com_tot_rcom	Max of total processed comissions	RCOM
721	min_com_tot_rcom	Min of total processed comissions	RCOM
722	sum_com_tot_vig_rcom	Sum of total processed comissions for active policies	RCOM
723	avg_com_tot_vig_rcom	Average of total processed comissions for active policies	RCOM
724	max_com_tot_vig_rcom	Max of total processed comissions for active policies	RCOM
725	min_com_tot_vig_rcom	Min of total processed comissions for active policies	RCOM
726	sum_tenure_vig_eq_0_rcom	Sum of policies active policies with less than 1 year	RCOM
727	sum_tenure_vig_eq_1_rcom	Sum of policies active policies with more than 1 year and less than 2	RCOM
728	sum_tenure_vig_eq_2_rcom	Sum of policies active policies with more than 2 years and less than 3	RCOM
729	sum_tenure_vig_0_2_rcom	Sum of policies active policies with more than 0 year and less than 3	RCOM
730	sum_tenure_vig_gt_2_rcom	Sum of policies active policies with more than 3 years	RCOM

731	avg_cost_vig_policy_rcom	Sum of total processed comissions for active policies divided by number of active policies	RCOM
732	avg_agnt_cost_policy_rcom	Sum of total processed comissions for active policies divided by number of distinct agents (cod_niv90)	RCOM
733	sum_expense_ytd_rcom	Sum of claims plus comissions	RCOM
734	sum_revenue_ytd_rcom	Sum of premiums minus sum expenses	RCOM
735	rat_revenue_pca_rcom	Sum of revenue divided by the PCA	RCOM
736	sum_expense_vig_rcom	Sum of claims plus comissions for active policies	RCOM
737	sum_revenue_vig_ytd_rcom	Sum of premiums minus sum expenses for active policies	RCOM
738	rat_revenue_pca_vig_rcom	Sum of revenue divided by the PCA for active policies	RCOM
739	policy_frequency_rcom	Total number of policies divided by the number of year the client is in the insurance	RCOM
740	sum_backtarded_rcom	Sum of policies with a begin date before the emission date	RCOM
741	slope_sum_pca_rcom	Sum of current year PCA minus the sum of homologous month PCA	RCOM
742	slope_avg_pca_rcom	Average of current year PCA minus the average of homologous month PCA	RCOM
743	sum_premios_rcom	Sum of current year premiums	RCOM
744	avg_premios_rcom	Average of current year premiums	RCOM
745	slope_sum_premios_rcom	Sum of current year premiums minus the sum of homologous month premiums	RCOM
746	slope_avg_premios_rcom	Average of current year premiums minus the average of homologous month premiums	RCOM
747	slope_sum_claim_rcom	Sum of current year claim value minus the sum of homologous month claim value	RCOM
748	slope_avg_claim_rcom	Average of current year claim value minus the average of homologous month claim value	RCOM
749	rat_claim_prems_rcom	Sum of current year claim value divided by the sum of processed premiums	RCOM
750	rat_com_prems_rcom	Sum of current year processed comissions divided by the sum of processed premiums	RCOM
751	avg_mon_sub_rcom	Average number of months since a substituition (same LoB purchased within a 7 day window of another policy ending)	RCOM
752	min_mon_sub_rcom	Min number of months since a substituition (same LoB purchased within a 7 day window of another policy ending)	RCOM
753	max_mon_sub_rcom	Max number of months since a substituition (same LoB purchased within a 7 day window of another policy ending)	RCOM
754	avg_mon_end_rcom	Average number of months for a policy to end	RCOM
755	max_mon_end_rcom	Max number of months for a policy to end	RCOM
756	min_maturity_mon_rcom	Youngest policy in months	RCOM
757	max_maturity_mon_rcom	Oldest policy in months	RCOM
758	mat_rollover_rcom	When a policy was substituited (same LoB purchased within a 7 day window of another policy ending) then 1, else when policy has ended then 2, when a policy will end in the next 3 months then 3 else 4	RCOM
759	sum_sub_increase_rcom	Sum of policies that were substituted (same LoB purchased within a 7 day window of another policy ending) for a policy	RCOM

		with a higher PCA	
760	sum_sub_decrease_rcom	Sum of policies that were substituted (same LoB purchased within a 7 day window of another policy ending) for a policy with a smaller PCA	RCOM
761	sum_canal_banca_postal_rcom	Sum of policies from a postal bank channel (cdniv20 equals 'DCPB')	RCOM
762	sum_canal_corretor_rcom	Sum of policies from a broker channel (cdniv20 equals 'IDCC')	RCOM
763	sum_canal_agencia_rcom	Sum of policies from an agency channel (cdniv20 equals 'IDCM')	RCOM
764	sum_canal_outro_rcom	Sum of policies from another channel (cdniv20 not equal to 'IDCM', 'IDCC', 'DCBP')	RCOM
765	sum_vida_cp_rcom	Sum of life capitalization policies (ramosp equals 'CP')	RCOM
766	sum_vida_ms_rcom	Sum of life mixed policies (ramosp equals 'MS')	RCOM
767	sum_vida_pp_rcom	Sum of life retirement savings policies (ramosp equals 'PP')	RCOM
768	sum_vida_pu_rcom	Sum of life capitalization/unit linked policies (ramosp equals 'PU')	RCOM
769	sum_vida_rn_rcom	Sum of life rent policies (ramosp equals 'RN')	RCOM
770	sum_vida_rs_rcom	Sum of life risk policies (ramosp equals 'RS')	RCOM
771	sum_vida_ul_rcom	Sum of life unit linked policies (ramosp equals 'UL')	RCOM
772	sum_vida_ul_cap_rcom	Sum of life capitalization plus unit linked policies (ramosp is equal to 'CP', 'PU', 'UL')	RCOM
773	sum_pca_vida_cp_rcom	Sum of the PCA from life capitalization policies (ramosp equals 'CP')	RCOM
774	sum_pca_vida_ms_rcom	Sum of the PCA from life mixed policies (ramosp equals 'MS')	RCOM
775	sum_pca_vida_pp_rcom	Sum of the PCA from life retirement savings policies (ramosp equals 'PP')	RCOM
776	sum_pca_vida_pu_rcom	Sum of the PCA from life capitalization/unit linked policies (ramosp equals 'PU')	RCOM
777	sum_pca_vida_rn_rcom	Sum of the PCA from life rent policies (ramosp equals 'RN')	RCOM
778	sum_pca_vida_rs_rcom	Sum of the PCA from life risk policies (ramosp equals 'RS')	RCOM
779	sum_pca_vida_ul_rcom	Sum of the PCA from life unit linked policies (ramosp equals 'UL')	RCOM
780	sum_pca_vida_ul_cap_rcom	Sum of the PCA from life capitalization plus unit linked policies (ramosp is equal to 'CP', 'PU', 'UL')	RCOM
781	avg_pca_vida_cp_rcom	Average of the PCA from life capitalization policies (ramosp equals 'CP')	RCOM
782	avg_pca_vida_ms_rcom	Average of the PCA from life mixed policies (ramosp equals 'MS')	RCOM
783	avg_pca_vida_pp_rcom	Average of the PCA from life retirement savings policies (ramosp equals 'PP')	RCOM
784	avg_pca_vida_pu_rcom	Average of the PCA from life capitalization/unit linked policies (ramosp equals 'PU')	RCOM
785	avg_pca_vida_rn_rcom	Average of the PCA from life rent policies (ramosp equals 'RN')	RCOM
786	avg_pca_vida_rs_rcom	Average of the PCA from life risk policies (ramosp equals 'RS')	RCOM

787	avg_pca_vida_ul_rcom	Average of the PCA from life unit linked policies (ramosp equals 'UL')	RCOM
788	avg_pca_vida_ul_cap_rcom	Average of the PCA from life capitalization plus unit linked policies (ramosp is equal to 'CP', 'PU', 'UL')	RCOM
789	sum_policy_anul_b*M_rcom	Sum of the number of cancelled policies that began in the last *months	RCOM
790	sum_policy_vig_b*M_rcom	Sum of the number of active policies that began in the last *months	RCOM
791	sum_policy_new_b*M_rcom	Sum of the number of new policies that began in the last *months	RCOM
792	sum_pca_b*M_rcom	Sum of the PCA for the policies that began in the last *months	RCOM
793	sum_pca_act_*M_rcom	Sum of the PCA for the active policies that began in the last *months	RCOM
794	avg_pca_act_*M_rcom	Average of the PCA for the active policies that began in the last *months	RCOM
795	count_pol_new_*M_rcom	Number of policies, excluding policies that came from substituitions, that began in the last *months	RCOM
796	sum_pca_new_*M_rcom	Sum of the PCA from policies, excluding policies that came from substituitions, that began in the last *months	RCOM
797	avg_pca_new_*M_rcom	Average of the PCA from policies, excluding policies that came from substituitions, that began in the last *months	RCOM
798	mon_canc_*M_rcom	Number of policies, excluding policies that came from substituitions, that were cancelled in the last *months	RCOM
799	sum_pca_canc_*M_rcom	Sum of the PCA from policies, excluding policies that came from substituitions, that were cancelled in the last *months	RCOM
800	avg_pca_canc_*M_rcom	Average of the PCA from policies, excluding policies that came from substituitions, that were cancelled in the last *months	RCOM
801	mon_exp_*M_rcom	Number of policies, excluding policies that came from substituitions, that expired in the last *months	RCOM
802	sum_pca_exp_*M_rcom	Sum of the PCA from policies, excluding policies that came from substituitions, that expired in the last *months	RCOM
803	avg_pca_exp_*M_rcom	Average of the PCA from policies, excluding policies that came from substituitions, that expired in the last *months	RCOM
804	count_pol_sub_*M_rcom	Number of policies that were substituted (same LoB purchased within a 7 day window of another policy ending) in the last *months	RCOM
805	sum_pca_sub_*M_rcom	Sum of the PCA from policies that were substituted (same LoB purchased within a 7 day window of another policy ending) in the last *months	RCOM
806	avg_pca_sub_*M_rcom	Average of the PCA from policies that were substituted (same LoB purchased within a 7 day window of another policy ending) in the last *months	RCOM
807	sum_diff_pca_sub_1M_rcom	Sum of the difference (new minu old) of PCA from policies that were substituted (same LoB purchased within a 7 day window of another policy ending) in the last *months	RCOM

808	avg_diff_pca_sub_1M_rcom	Average of the difference (new minu old) of PCA from policies that were substituted (same LoB purchased within a 7 day window of another policy ending) in the last *months	RCOM
809	cou_act*M_rcom	Sum the number of active policies that were active * months ago	RCOM
810	cou_policy_rcom	Count all policies of the client	RCOM by LOB
811	sum_policy_new_rcom	Sum all policies of the client which are new	RCOM by LOB
812	sum_vig_rcom	Sum all policies of the client which are active	RCOM by LOB
813	sum_anul_rcom	Sum all policies of the client which are annulled	RCOM by LOB
814	ind_policy_new_rcom	Indicate client has a new policy	RCOM by LOB
815	ind_vig_rcom	Indicate client has has a policy in active	RCOM by LOB
816	ind_anul_rcom	Indicate client has as annulled policy	RCOM by LOB
817	sum_pca_rcom	Sum of all pca	RCOM by LOB
818	avg_pca_rcom	Average of all pca	RCOM by LOB
819	max_pca_rcom	Max of all pca	RCOM by LOB
820	min_pca_rcom	Min of all pca	RCOM by LOB
821	sum_vig_pca_rcom	Sum of all pca from active policies	RCOM by LOB
822	avg_vig_pca_rcom	Average of all pca from active policies	RCOM by LOB
823	max_vig_pca_rcom	Max of all pca from active policies	RCOM by LOB
824	min_vig_pca_rcom	Min of all pca from active policies	RCOM by LOB
825	sum_anul_pca_rcom	Sum of all pca from annulled policies	RCOM by LOB
826	sum_claim_ytd_rcom	Sum of cssinexe	RCOM by LOB
827	avg_claim_ytd_rcom	Average of cssinexe	RCOM by LOB
828	sum_com_tot_rcom	Sum of comtoapl	RCOM by LOB
829	avg_com_tot_rcom	Average of comtoapl	RCOM by LOB
830	min_mon_end_rcom	Min of months to end	RCOM by LOB
831	avg_maturity_mon_rcom	Average of months to end	RCOM by LOB
832	min_day_end_rcom	Min of days to end	RCOM

			by LOB
833	avg_maturity_day_rcom	Average of days to end	RCOM by LOB
834	sum_sub_rcom	Sum of policies which came from substituition	RCOM by LOB
835	sum_ren_rcom	Sum of policies which came from substituition or older than 1 year (renewal definition)	RCOM by LOB
836	sum_sub_vig_rcom	Sum of policies which came from substituition and are active	RCOM by LOB
837	sum_ren_vig_rcom	Sum of policies (which came from substituition or older than 1 year (renewal definition)) and are active	RCOM by LOB
838	min_months from beginning_rcom	Min of months from beginning	RCOM by LOB
839	min_mon_canc_rcom	Min of mon_canc	RCOM by LOB
840	max_mon_canc_rcom	Max of mon_canc	RCOM by LOB
841	avg_mon_canc_rcom	Avg of mon_canc	RCOM by LOB
842	mon_canc_0_2_rcom	Sum of cases when mon_canc between Jan and March and was not substituted	RCOM by LOB
843	mon_canc_3_5_rcom	Sum of cases when mon_canc between April and Jun and was not substituted	RCOM by LOB
844	mon_canc_6_8_rcom	Sum of cases when mon_canc between Jul and Sept and was not substituted	RCOM by LOB
845	mon_canc_9_11_rcom	Sum of cases when mon_canc between Oct and Dec and was not substituted	RCOM by LOB
846	mon_canc_0_2_1yr_rcom	Sum of cases when mon_canc between Jan and March of last year and was not substituted	RCOM by LOB
847	mon_canc_3_5_1yr_rcom	Sum of cases when mon_canc between April and Jun of last year and was not substituted	RCOM by LOB
848	mon_canc_6_8_1yr_rcom	Sum of cases when mon_canc between Jul and Sept of last year and was not substituted	RCOM by LOB
849	mon_canc_9_11_1yr_rcom	Sum of cases when mon_canc between Oct and Dec of last year and was not substituted	RCOM by LOB
850	cou_niv90_rcom	Count distinct niv90	RCOM by LOB
851	cou_niv80_rcom	Count distinct niv80	RCOM by LOB
852	cou_niv70_rcom	Count distinct niv70	RCOM by LOB
853	cou_niv60_rcom	Count distinct niv60	RCOM by LOB
854	cou_niv50_rcom	Count distinct niv50	RCOM by LOB
855	cou_niv40_rcom	Count distinct niv40	RCOM by LOB
856	cou_niv90_vig_rcom	Count distinct niv90	RCOM by LOB

857	cou_niv80_vig_rcom	Count distinct niv80	RCOM by LOB
858	cou_niv70_vig_rcom	Count distinct niv70	RCOM by LOB
859	cou_niv60_vig_rcom	Count distinct niv60	RCOM by LOB
860	cou_niv50_vig_rcom	Count distinct niv50	RCOM by LOB
861	cou_niv40_vig_rcom	Count distinct niv40	RCOM by LOB
862	cou_act*M_rcom	Sum the number of active policies that were active * months ago	RCOM by LOB
863	TAXA_ACT_CONC_ine_dem	Taxa de actividade (%) da população residente por Concelho de residência	INE
864	RENSA_CONC_ine_dem	Valor médio mensal das rendas (euros) dos alojamentos familiares clássicos arrendados por Concelho	INE
865	PERC_QUADROS_CONC_ine_dem	Proporção de profissionais socialmente mais valorizados (%) por Concelho de residência	INE
866	PROP_AUT_CONC_ine_dem	Proporção de utilização do automóvel nas deslocações (%) por Concelho de residência	INE
867	POP_RESID_CONC_ine_dem	População residente (N.º) por Concelho de residência	INE
868	DENS_POP_CONC_ine_dem	Densidade populacional (N.º/ km²) por Concelho de residência	INE
869	REGIAO_ine_dem	Litoral ou Interior	INE
870	INDICE_RENDIMENTO_ine_dem	INDICE_RENDIMENTO	INE
871	SALES_INDEX_ine_dem	SALES_INDEX	INE
872	distrito_ctt_dem	Designação Distrito	INE
873	Sum_cont_cont	Sum all contacts	CC
874	sum_cont_in_cont	Sum contacts with interaction inbound	CC
875	sum_cont_out_cont	Sum contacts with interaction outbound	CC
876	sum_email_in_cont	Sum email contacts with interaction inbound	CC
877	sum_email_out_cont	Sum email contacts with interaction outbound	CC
878	sum_pres_in_cont	Sum presential contacts with interaction inbound	CC
879	sum_voz_in_cont	Sum voice contacts with interaction inbound	CC
880	sum_voz_out_cont	Sum voice contacts with interaction outbound	CC
881	sum_agcm_in_cont	Sum all agency interactions inbound	CC
882	sum_cc_in_cont	Sum all agency interactions inbound	CC
883	min_mon_in_cont	Months to the earlierst contact inbout	CC
884	min_mon_out_cont	Months to the earlierst contact outbout	CC
885	sum_cont_Life_cont	Number of contact with a word suggesting life	CC
886	sum_cont_Auto_cont	Number of contact with a word suggesting auto	CC
887	sum_cont_Health_cont	Number of contact with a word suggesting health	CC
888	sum_cont_Home_cont	Number of contact with a word suggesting home	CC
889	sum_cont_Claim_cont	Number of contact with a word suggesting claim	CC
890	sum_cont_Opening_cont	Number of contact with a word suggesting opening	CC
891	sum_cont_Cancellation_cont	Number of contact with a word suggesting cancellation	CC

892	sum_cont_Call_back_cont	Number of contact with a word suggesting call back	CC
893	sum_cont_Not_reached_cont	Number of contact with a word suggesting not reached	CC
894	sum_cont_Inclusion_cont	Number of contact with the word inclusion	CC
895	sum_cont_Extension_cont	Number of contact with a word suggesting extension	CC
896	sum_cont_Payment_cont	Number of contact with a word suggesting payment	CC
897	sum_cont_Alteracao_cont	Number of contact with a word suggesting change	CC
898	num_coverages	total number of coverages in policies own	HLTH
899	ncov_critical_illness	total number of critical illness coverages	HLTH
900	ncov_disability	total number of disability income coverages	HLTH
901	ncov_premio	total number of premium coverages	HLTH
902	ncov_life_protection	total number of life (protection) coverages	HLTH
903	ncov_medical	total number of medical coverages	HLTH
904	perc_critical_illness	total number of critical illness coverages / total number of inforce coverages	HLTH
905	perc_disability	total number of disability income coverages / total number of inforce coverages	HLTH
906	perc_premio	total number of premium coverages / total number of inforce coverages	HLTH
907	perc_life_protection	total number of life (protection) coverages / total number of inforce coverages	HLTH
908	perc_medical	total number of medical coverages / total number of inforce coverages	HLTH
909	flag_critical_illness	flag for critical illness coverage owned	HLTH
910	flag_disability	flag for disability coverage owned	HLTH
911	flag_premio	flag for premium coverage owned	HLTH
912	flag_live_protection	flag for live protection coverage owned	HLTH
913	flag_medical	flag for medical coverage owned	HLTH
914	num_cartoes	number of policies of the client of the type 'cartão'	HLTH
915	num_simplecare	number of policies of the client of the type 'simplecare'	HLTH
916	perc_cartoes	percentage of policies of the client of the type ‘cartão’ over the total number of policies he owns	HLTH
917	perc_simplecare	percentage of policies of the client of the type ‘simplecare’ over the total number of policies he owns	HLTH
918	num_cartoes_maternal	number of policies of the client of type ‘cartão-maternal’	HLTH
919	num_cartoes_geral	number of policies of the client of type ‘cartão-geral’	HLTH
920	num_cartoes_vital	number of policies of the client of type ‘cartão-vital’	HLTH
921	num_cartoes_dental	number of policies of the client of type ‘cartão-dental’	HLTH
922	perc_cartoes_maternal	percentage of policies of the client of the type ‘cartão-maternal’ over the total number of policies he owns	HLTH
923	perc_cartoes_geral	percentage of policies of the client of the type ‘cartão-geral’ over the total number of policies he owns	HLTH
924	perc_cartoes_vital	percentage of policies of the client of the type ‘cartão-vital’ over the total number of policies he owns	HLTH
925	perc_cartoes_dental	percentage of policies of the client of the type ‘cartão-dental’ over the total number of policies he owns	HLTH

926	sinister_premium	relationship between the amount spent on sinisters and the total premium paid in health by the client	HLTH
927	max_limit	maximum limit of coverage	HLTH
928	sum_limit_inter	sum of values in every police a client has of the 'internamento' coverage limits	HLTH
929	sum_limit_amb	sum of values in every police a client has of the 'ambulatorio' coverage limits	HLTH
930	months_limit	number of months since the last change in a limit in an owned coverage	HLTH
931	usage	percentage of usage since the last change of a limit	HLTH
932	remaining	percentage of remaining to the limit since the last change of a limit	HLTH
933	nqt1	percentage of usage of the client in the first quarter after renovation for the last 15 months	HLTH
934	nqt2	percentage of usage of the client in the second quarter after renovation for the last 15 months	HLTH
935	nqt3	percentage of usage of the client in the third quarter after renovation for the last 15 months	HLTH
936	nqt4	percentage of usage of the client in the fourth quarter after renovation for the last 15 months	HLTH
937	nq1	percentage of network usage of the client in the first quarter after renovation for the last 15 months	HLTH
938	nq2	percentage of network usage of the client in the second quarter after renovation for the last 15 months	HLTH
939	nq3	percentage of network usage of the client in the third quarter after renovation for the last 15 months	HLTH
940	nq4	percentage of network usage of the client in the fourth quarter after renovation for the last 15 months	HLTH
941	nq1_ris	percentage of usage in risk coverages of the client in the first quarter after renovation for the last 15 months	HLTH
942	nq2_ris	percentage of usage in risk coverages of the client in the second quarter after renovation for the last 15 months	HLTH
943	nq3_ris	percentage of usage in risk coverages of the client in the third quarter after renovation for the last 15 months	HLTH
944	nq4_ris	percentage of usage in risk coverages of the client in the fourth quarter after renovation for the last 15 months	HLTH
945	rq1	average of money reimbursed to the client during the first quarter after renovation for the last 15 months	HLTH
946	rq2	average of money reimbursed to the client during the second quarter after renovation for the last 15 months	HLTH
947	rq3	average of money reimbursed to the client during the third quarter after renovation for the last 15 months	HLTH
948	rq4	average of money reimbursed to the client during the fourth quarter after renovation for the last 15 months	HLTH
949	flag_children	whether insured is insurer's children	HLTH
950	flag_ascending	whether insured is insurer's ascending	HLTH

951	flag_spouses	whether insured is insurer's spouse	HLTH
952	flag_elder_spouses	whether insured is insurer's elder spouse	HLTH
953	flag_adult_children	whether insured is insurer's adult children	HLTH
954	flag_grandchildren	whether insured is insurer's grandchildren	HLTH
955	flag_others	whether insured is insurer's other relatives	HLTH
956	flag_parents	whether insured is insurer's parents	HLTH
957	flag_titular	whether insured is insurer	HLTH
958	flag_relatives	whether insured is insurer's relatives	HLTH
959	n_relatives_ins	number of family members insured	HLTH
960	n_childre_ins	number of children insured	HLTH
961	n_ascending_ins	number of ascending insured	HLTH
962	n_spouses_ins	number of spouses insured	HLTH
963	n_elder_spouses_ins	number of elder spouses insured	HLTH
964	n_adult_children_ins	number of adult children insured	HLTH
965	n_grandchildren_ins	number of grandchildren insured	HLTH
966	n_others_ins	number of other relatives insured	HLTH
967	n_parents_ins	number of parents insured	HLTH
968	n_titular_ins	number of children insured	HLTH
969	n_policies	total number of polices of the client	HLTH
970	perc_nao_titular	percentage of his policies in which the client is just tomador and not titular	HLTH
971	ratio_children	ratio of children insured in the policies of the client	HLTH
972	ratio_ascending	ratio of ancestors insured in the policies of the client	HLTH
973	ratio_spouses	ratio of spouses insured in the policies of the client	HLTH
974	ratio_elder_spouses	ratio of elder spouses insured in the policies of the client	HLTH
975	ratio_adult_children	ratio of adult children insured in the policies of the client	HLTH
976	ratio_grandchildren	ratio of grandchildren insured in the policies of the client	HLTH
977	ratio_parents	ratio of parents insured in the policies of the client	HLTH
978	ratio_relatives	ratio of relatives insured in the policies of the client	HLTH
979	ratio_others	ratio of other relationships insured in the policies of the client	HLTH
980	clientes_ INSURANCE	Indicator for clients that own at least one policy of INSURANCE	HLTH
981	pessoa_seg_saude	Indicator for clinets that do not own a single policy of the insurance	HLTH
982	clientes_ INSURANCE _ENI	Indicator for clients that are clientes_ INSURANCE and ENI	HLTH
983	clientes_ INSURANCE _PS	Indicator for clients that are clientes_ INSURANCE and individual person	HLTH
984	clientes_ INSURANCE_tomador_sd	Indicator for clients that are clientes_ INSURANCE and own a health policy	HLTH
985	clientes_ INSURANCE _title_sd	Indicator for clients that are clientes_ INSURANCE and have a health policy where they are titular	HLTH
986	clientes_ INSURANCE _insured_sd	Indicator for clients that are clientes_ INSURANCE and have a health policy where they are insured person	HLTH
987	pessoa_seg_saude_group	Indicator for clients that are pessoa_seg_saude and are insured in a corporate policy	HLTH
988	pessoa_seg_saude_indiv	Indicator for clients that are pessoa_seg_saude and are	HLTH

		insured in an individual policy	
989	pessoa_seg_saude_title	Indicator for clients that are pessoa_seg_saude and have a title	HLTH
990	pess_seg_sd_insured_no_title	Indicator for clients that are pessoa_seg_saude and do not have a title	HLTH
991	nqt1_*	percentage of usage of the client in the first quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
992	nqt2_*	percentage of usage of the client in the second quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
993	nqt3_*	percentage of usage of the client in the third quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
994	nqt4_*	percentage of usage of the client in the fourth quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH

995	nq1_*	percentage of network usage of the client in the first quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
996	nq2_*	percentage of network usage of the client in the second quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
997	nq3_*	percentage of network usage of the client in the third quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
998	nq4_*	percentage of network usage of the client in the fourth quarter after renovation for the last 15 months for a specific coverage. Specific coverages are indicated by the * that stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR	HLTH
999	*_MTC	Sum of health clients with a specific client status (* stands for abandonado, conhecide, de_saida, perdido, recente, recuperado, NA). Uses RCOM information from the first	HLTH

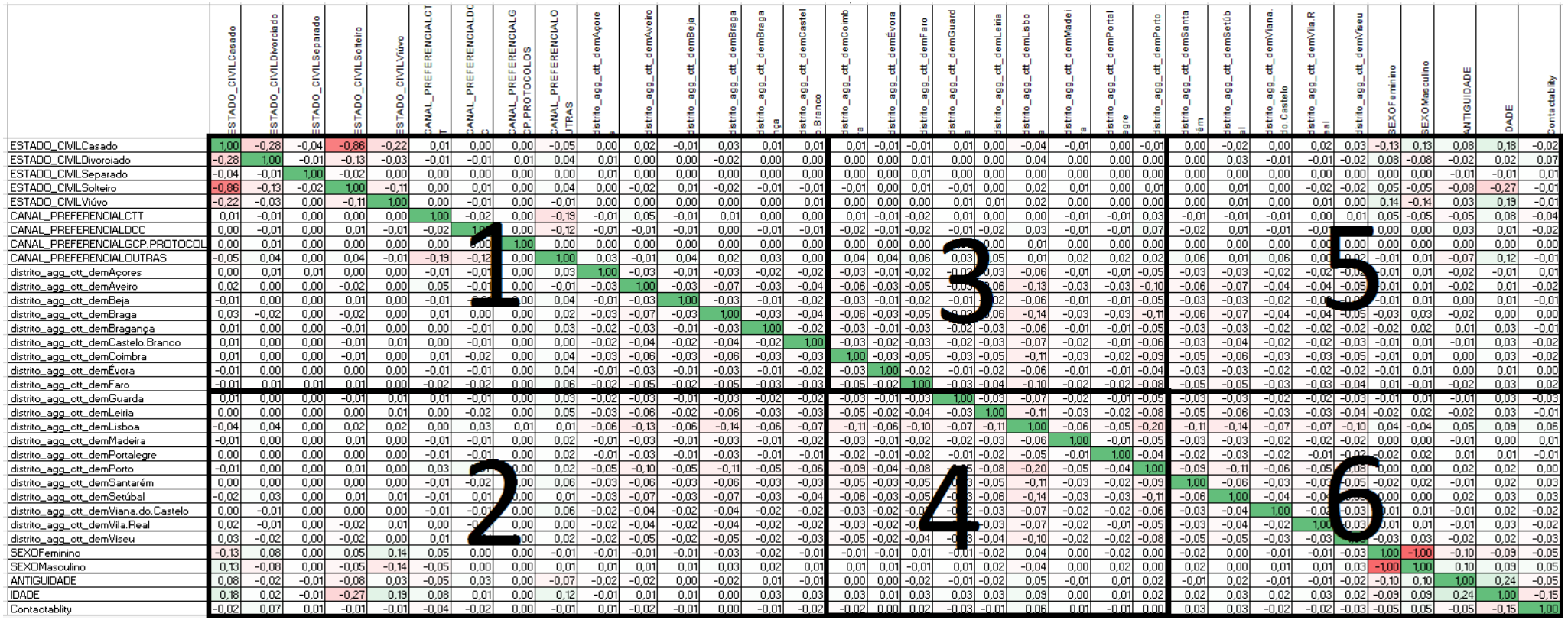
		version of the car	
1000	ENI	Indicator for ENI clients	HLTH
1001	PS	Indicator for individual person	HLTH
1002	*_amb	Indicator for a client with a claim in a specific ambulant specialty in the last 36 month. (* stands for Anatomia_Patologia, Anestesiologia, Biopsias, Cardiologia, Cardiologia_Pediatria, Cirurgia_Cardio_Toracica, Cirurgia_Geral, Cirurgia_Geral_Bloco_PF, Cirurgia_Maxilo_Facial, Cirurgia_Pediatrica, Cirurgia_Plastica_Reconstr, Cirurgia_Vascular, Dermatologia, Doenças_Infecciosas, Endocrinologia, Enfermagem, Estomatologia, Gastrenterologia, Gastrenterologia_Pediátrica, Genética, Ginecologia, Ginecologia_Bloco_PF, Hematologia_Clinica, Imuno_Alergologia, Imuno_Hemoterapia, KIT, MFR, MGF, Medicina_Interna, Medicina_Nuclear, NA, Nefrologia, Neuro_Radiologia, Neurocirurgia, Neurologia, Oftalmologia, Oncologia_Médica, Ortopedia, Ortopedia_Bloco_PF, Otorrino, Patologia_Clinica, Pediatria, Pneumologia, Psiquiatria, Radiologia_Imagiologia, Radioncologia, Reumatologia, TNC, Terapia_Fala, Urgência, Urologia, Urologia_Bloco_PF)	HLTH
1003	*_internamento	Indicator for a client with a claim in a specific in-patient category in the last 36 month. (* stands for Atos_Cirurgicos, Tecnicas_Invasivas, Medicina, Oncologia, NA, Parto, Terap_Med_Gerais)	HLTH
1004	ind_claim_*	Indicates whether a client had a claim with a specific coverage in the last 36 month. (*stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR)	HLTH
1005	shown_val_plan_*	Sum of value of the claims, that has been shown to the client. For clients that own a plan and had claims in the last 36 month. These values are separated by the coverage that covered the claim (*stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR)	HLTH

1006	paid_val_plan_*	Sum of paid of the claims, that has been paid to the client. For clients that own a plan and claims in the last 36 month. These values are separated by the coverage that covered the claim (*stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR)	HLTH
1007	reimburse_payment	Sum of reimburse payment	HLTH
1008	net_payment	Sum of network payment	HLTH
1009	net_presented_val	Sum of network shown value	HLTH
1010	reimburse_presented_val	Sum of reimburse shown value	HLTH
1011	AMBULATORIO	Indicates whether a client had an "ambulantorio" claim in MTC_REDE_REEMBOLSO in the last 36 month	HLTH
1012	INTERNAMENTO_PARTO	Indicates whether a client had an "internamento" or "parto" claim in MTC_REDE_REEMBOLSO in the last 36 month	HLTH
1013	ESTOMATOLOGIA	Indicates whether a client had an "estomatologia" claim in MTC_REDE_REEMBOLSO in the last 36 month	HLTH
1014	OTHERS	Indicates whether a client had a claim that was not "ambulatorio", "internamento", "parto" or "estomatologia" in MTC_REDE_REEMBOLSO in the last 36 month	HLTH
1015	ind_*	Indicates whether a client owns a speial coverage (*stands for INTERNAMENTO, AMBULATORIO, ESTOMATOLOGIA, PARTO, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR, or a combination of INTERNAMENTO and PARTO)	HLTH
1016	ind_plan_*	Indicates whether a client owns a plan with speial coverage (*stands for AMBULATORIO, ESTOMATOLOGIA, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR, or a combination of INTERNAMENTO and PARTO)	HLTH

1017	ind_card_*	Indicates whether a client owns a card with speial coverage (*stands for AMBULATORIO, ESTOMATOLOGIA, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR, or a combination of INTERNAMENTO and PARTO)	HLTH
1018	ind_simplc_*	Indicates whether a client owns simplecare with speial coverage (*stands for AMBULATORIO, ESTOMATOLOGIA, SUBSIDIO_DIARIO, CHECK_UP, MEDICAMENTOS, PROTESES, ASSIST_HOSPITALAR, CHECK_UP_2, EXT_ESTRANGEIRO, ORTOTESES, COBERTURA_ESPECIAL, TERAPEUTICAS, DOENCAS_GRAVES, 2nd_OPINIAO, DOENCAS_GRAVES_2, REGISTO_ENTRADA, APOIO_DOMICILIARIO, BEST_DOCTORS, ESTOMATOLOGIA_WDA, PROTECCAO_PAGAMENTO, ONCOLOGIA, PREMIO, OUTROS, A_DEFINIR, or a combination of INTERNAMENTO and PARTO)	HLTH
1019	cou_pol_ren_12M_rcom	Sum of the number of renewed policies in the last 12 month	RCOM
1020	avg_com_tot_vig_nf_rcom	Average of total processed comissions for non-financial active policies	RCOM
1021	avg_pca_ren_12M_nf_rcom	Average PCA of renewed non-financial policies in the last 12 month	RCOM
1022	avg_vig_pca_nf_rcom	Average PCA of active non-financial policies in the last 12 month	RCOM
1023	rat_com_prems_nf_rcom	Sum of current year processed commissions divided by the sum of processed premiums	RCOM
1024	slope_avg_pca_nf_rcom	Average of current month non-financial PCA minus the average of homologous month non-financial PCA	RCOM
1025	sum_claim_ytd_nf_rcom	Sum of claims value of non-financial policies year to date	RCOM
1026	slope_tech_margin_nf	Sum of current month's non-financial technical margin minus the sum of homologous month's non-financial technical margin	RCOM
1027	tech_margin_nf_a	A client's value of non-financial policies in the current month	RCOM
1028	tech_margin_nf_h	Value of non-financial policies of the client in the homologous month of the past year	RCOM

7.2. Correlation Matrix for Demographic Segmentation

The demographic segmentation includes categorical variables that lead to a variety of dummy variables for each of their levels. This increases the size of the correlation matrix in a manner that makes it necessary to split it into six pieces and show them separately. The following table gives an overview of the localization of sub section in the total correlation matrix.



	ESTADO_CIVILCasado	ESTADO_CIVILDivorciado	ESTADO_CIVILSeparado	ESTADO_CIVILSolteiro	ESTADO_CIVILViúvo	CANAL_PREFERENCIALCTT	CANAL_PREFERENCIALDCC	CANAL_PREFERENCIALGCP.PROTOCOLOS	CANAL_PREFERENCIALOUTRAS	distrito_agg_ctt_demAçore	distrito_agg_ctt_demAveiro	distrito_agg_ctt_demBeja	distrito_agg_ctt_demBraga	distrito_agg_ctt_demBraga	distrito_agg_ctt_demCastel o.Branco	distrito_agg_ctt_demCoimb ra	distrito_agg_ctt_demÉvora	distrito_agg_ctt_demFaro	distrito_agg_ctt_demGuard a	distrito_agg_ctt_demLeiria	distrito_agg_ctt_demLisbo a	distrito_agg_ctt_demMadei a	distrito_agg_ctt_demPortal egre	distrito_agg_ctt_demPorto	distrito_agg_ctt_demSanta ém	distrito_agg_ctt_demSetub al	distrito_agg_ctt_demViana. do.Castelo	distrito_agg_ctt_demVila.R eal	distrito_agg_ctt_demViseu	SEXOFeminino	SEXOMasculino	ANTIGUIDADE	IDADE	Contactablity
ESTADO_CIVILCasado	1,00	-0,28	-0,04	-0,86	-0,22	0,01	0,00	0,00	-0,05	0,00	0,02	-0,01	0,03	0,01	0,01	0,01	-0,01	-0,01	0,01	0,00	-0,04	-0,01	0,00	-0,01	0,00	-0,02	0,00	0,02	0,03	-0,13	0,13	0,08	0,18	-0,02
ESTADO_CIVILDivorciado	-0,28	1,00	-0,01	-0,13	-0,03	-0,01	-0,01	0,01	0,04	0,01	0,00	0,00	-0,02	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,03	-0,01	-0,01	-0,02	0,08	-0,08	-0,02	0,02	0,07
ESTADO_CIVILSeparado	-0,04	-0,01	1,00	-0,02	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,01
ESTADO_CIVILSolteiro	-0,86	-0,13	-0,02	1,00	-0,11	0,00	0,01	0,00	0,04	0,00	-0,02	0,01	-0,02	-0,01	-0,01	-0,01	0,00	0,01	-0,01	0,00	0,02	0,01	0,00	0,01	0,00	0,01	0,00	-0,02	-0,02	0,05	-0,05	-0,08	-0,27	-0,01
ESTADO_CIVILViúvo	-0,22	-0,03	0,00	-0,11	1,00	0,00	-0,01	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,01	0,02	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,14	-0,14	0,03	0,19	-0,01
CANAL_PREFERENCIALCTT	0,01	-0,01	0,00	0,00	0,00	1,00	-0,02	0,00	-0,19	-0,01	0,05	-0,01	0,01	0,00	0,00	0,01	-0,01	-0,02	0,01	0,00	0,00	-0,01	-0,01	0,03	-0,01	-0,01	-0,01	0,00	0,01	0,05	-0,05	-0,05	0,08	-0,04
CANAL_PREFERENCIALDCC	0,00	-0,01	0,00	0,01	-0,01	-0,02	1,00	0,00	-0,12	-0,01	-0,01	-0,01	0,00	-0,01	-0,01	-0,02	-0,01	-0,02	-0,01	-0,02	0,03	-0,01	-0,01	0,07	-0,02	0,01	-0,01	-0,01	0,00	0,00	0,00	0,03	0,01	-0,02
CANAL_PREFERENCIALGCP.PROTOCOL	0,00	0,01	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CANAL_PREFERENCIALOUTRAS	-0,05	0,04	0,00	0,04	-0,01	-0,19	-0,12	0,00	1,00	0,03	-0,01	0,04	0,02	0,03	0,00	0,04	0,04	0,06	0,03	0,05	0,01	0,02	0,02	0,02	0,06	0,01	0,06	0,00	0,02	-0,01	0,01	-0,07	0,12	-0,01
distrito_agg_ctt_demAçores	0,00	0,01	0,01	0,00	0,00	-0,01	-0,01	0,00	0,03	1,00	-0,03	-0,01	-0,03	-0,02	-0,02	-0,03	-0,01	-0,02	-0,02	-0,03	-0,06	-0,01	-0,01	-0,05	-0,03	-0,03	-0,02	-0,02	-0,01	0,01	0,01	-0,02	-0,01	0,01
distrito_agg_ctt_demAveiro	0,02	0,00	0,00	-0,02	0,00	0,05	-0,01	0,00	-0,01	-0,03	1,00	-0,03	-0,07	-0,03	-0,04	-0,06	-0,03	-0,05	-0,03	-0,06	-0,13	-0,03	-0,03	-0,10	-0,06	-0,07	-0,04	-0,04	-0,05	0,01	0,01	-0,02	0,01	-0,02
distrito_agg_ctt_demBeja	-0,01	0,00	0,00	0,01	0,00	-0,01	-0,01	0,00	0,04	-0,01	-0,03	1,00	-0,03	-0,01	-0,02	-0,03	-0,01	-0,02	-0,01	-0,02	-0,06	-0,01	-0,01	-0,05	-0,03	-0,03	-0,02	-0,02	-0,01	-0,01	0,01	0,00	0,01	-0,01
distrito_agg_ctt_demBraga	0,03	-0,02	0,00	-0,02	0,00	0,01	0,00	0,00	0,02	-0,03	-0,07	-0,03	1,00	-0,03	-0,04	-0,06	-0,03	-0,05	-0,03	-0,06	-0,14	-0,03	-0,03	-0,11	-0,06	-0,07	-0,04	-0,04	-0,05	-0,03	0,03	-0,02	0,00	0,00
distrito_agg_ctt_demBragança	0,01	0,00	0,00	-0,01	0,00	0,00	-0,01	0,00	0,03	-0,02	-0,03	-0,01	-0,03	1,00	-0,02	-0,03	-0,01	-0,03	-0,02	-0,03	-0,06	-0,01	-0,01	-0,05	-0,03	-0,03	-0,02	-0,02	-0,02	-0,02	0,02	0,01	0,03	-0,01
distrito_agg_ctt_demCastelo.Branco	0,01	0,00	0,00	-0,01	0,01	0,00	-0,01	0,00	0,00	-0,02	-0,04	-0,02	-0,04	-0,02	1,00	-0,03	-0,02	-0,03	-0,02	-0,03	-0,07	-0,02	-0,01	-0,06	-0,03	-0,04	-0,02	-0,02	-0,03	-0,01	0,01	-0,01	0,03	-0,02
distrito_agg_ctt_demCoimbra	0,01	0,00	0,00	-0,01	0,00	0,01	-0,02	0,00	0,04	-0,03	-0,06	-0,03	-0,06	-0,03	-0,03	1,00	-0,03	-0,05	-0,03	-0,05	-0,11	-0,03	-0,02	-0,09	-0,05	-0,06	-0,03	-0,03	-0,05	-0,01	0,01	0,00	0,03	-0,02
distrito_agg_ctt_demÉvora	-0,01	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,04	-0,01	-0,03	-0,01	-0,03	-0,01	-0,02	-0,03	1,00	-0,02	-0,01	-0,02	-0,06	-0,01	-0,01	-0,04	-0,03	-0,03	-0,02	-0,02	-0,02	-0,01	0,01	0,00	0,01	0,00
distrito_agg_ctt_demFaro	-0,01	0,01	0,01	0,01	0,00	-0,02	-0,02	0,00	0,06	-0,02	-0,05	-0,02	-0,05	-0,03	-0,03	-0,05	-0,02	1,00	-0,03	-0,04	-0,10	-0,02	-0,02	-0,08	-0,05	-0,05	-0,03	-0,03	-0,04	0,01	-0,01	-0,02	0,03	0,02
distrito_agg_ctt_demGuarda	0,01	0,00	0,00	-0,01	0,01	0,01	-0,01	0,00	0,03	-0,02	-0,03	-0,01	-0,03	-0,02	-0,02	-0,03	-0,01	-0,03	1,00	-0,03	-0,07	-0,02	-0,01	-0,05	-0,03	-0,03	-0,02	-0,02	-0,03	-0,01	0,01	-0,01	0,03	-0,03
distrito_agg_ctt_demLeiria	0,00	0,00	0,00	0,00	0,01	0,00	-0,02	0,00	0,05	-0,03	-0,06	-0,02	-0,06	-0,03	-0,03	-0,05	-0,02	-0,04	-0,03	1,00	-0,11	-0,03	-0,02	-0,08	-0,05	-0,06	-0,03	-0,03	-0,04	-0,02	0,02	-0,02	0,03	-0,01
distrito_agg_ctt_demLisboa	-0,04	0,04	0,00	0,02	0,02	0,00	0,03	0,01	0,01	-0,06	-0,13	-0,06	-0,14	-0,06	-0,07	-0,11	-0,06	-0,10	-0,07	-0,11	1,00	-0,06	-0,05	-0,20	-0,11	-0,14	-0,07	-0,07	-0,10	0,04	-0,04	0,05	0,09	0,06
distrito_agg_ctt_demMadeira	-0,01	0,00	0,00	0,01	0,00	-0,01	-0,01	0,00	0,02	-0,01	-0,03	-0,01	-0,03	-0,01	-0,02	-0,03	-0,01	-0,02	-0,02	-0,03	-0,06	1,00	-0,01	-0,05	-0,03	-0,03	-0,02	-0,02	-0,02	0,00	0,00	-0,01	0,00	0,01
distrito_agg_ctt_demPortalegre	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,02	-0,01	-0,03	-0,01	-0,03	-0,01	-0,01	-0,02	-0,01	-0,02	-0,01	-0,02	-0,05	-0,01	1,00	-0,04	-0,02	-0,03	-0,02	-0,01	-0,02	-0,02	0,02	0,01	0,01	-0,01
distrito_agg_ctt_demPorto	-0,01	0,00	0,00	0,01	0,00	0,03	-0,01	0,00	0,02	-0,05	-0,10	-0,05	-0,11	-0,05	-0,06	-0,09	-0,04	-0,08	-0,05	-0,08	-0,20	-0,05	-0,04	1,00	-0,09	-0,11	-0,06	-0,05	-0,08	0,00	0,00	0,02	0,02	0,00
distrito_agg_ctt_demSantarém	0,00	0,00	0,00	0,00	0,00	-0,01	-0,02	0,00	0,06	-0,03	-0,06	-0,03	-0,06	-0,03	-0,03	-0,05	-0,03	-0,05	-0,03	-0,05	-0,11	-0,03	-0,02	-0,09	1,00	-0,06	-0,03	-0,03	-0,05	-0,02	0,02	-0,01	0,02	0,03
distrito_agg_ctt_demSetúbal	-0,02	0,03	0,00	0,01	0,01	-0,01	0,01	0,00	0,01	-0,03	-0,07	-0,03	-0,07	-0,03	-0,04	-0,06	-0,03	-0,05	-0,03	-0,06	-0,14	-0,03	-0,03	-0,11	-0,06	1,00	-0,04	-0,04	-0,06	0,00	0,00	0,02	0,03	0,03
distrito_agg_ctt_demViana.do.Castelo	0,00	-0,01	0,00	0,00	0,00	-0,01	-0,01	0,00	0,06	-0,02	-0,04	-0,02	-0,04	-0,02	-0,02	-0,03	-0,02	-0,04	-0,02	-0,03	-0,07	-0,02	-0,02	-0,06	-0,03	-0,04	1,00	-0,03	-0,03	0,01	0,01	-0,01	0,02	-0,02
distrito_agg_ctt_demVila.Real	0,02	-0,01	0,00	-0,02	0,01	0,00	-0,01	0,00	0,00	-0,02	-0,04	-0,02	-0,04	-0,02	-0,02	-0,03	-0,02	-0,04	-0,02	-0,03	-0,07	-0,02	-0,01	-0,05	-0,03	-0,04	-0,02	1,00	-0,03	0,01	0,01	-0,01	0,03	-0,02
distrito_agg_ctt_demViseu	0,03	-0,02	0,00	-0,02	0,00	0,01	-0,01	0,00	0,02	-0,02	-0,05	-0,02	-0,05	-0,02	-0,03	-0,05	-0,02	-0,04	-0,03	-0,04	-0,10	-0,02	-0,02	-0,08	-0,05	-0,05	-0,03	-0,03	1,00	-0,03	0,03	-0,02	0,02	-0,03
SEXOFeminino	-0,13	0,08	0,00	0,05	0,14	0,05	0,00	0,00	-0,01	-0,01	-0,01	-0,01	-0,03	-0,02	-0,01	-0,01	-0,01	0,01	-0,01	-0,02	0,04	0,00	-0,02	0,00	-0,02	0,00	-0,01	-0,01	-0,03	1,00	-1,00	-0,10	-0,09	-0,05
SEXOMasculino	0,13	-0,08	0,00	-0,05	-0,14	-0,05	0,00	0,00	0,01	0,01	0,01	0,01	0,03	0,02	0,01	0,01	0,01	0,01	0,01	0,02	-0,04	0,00	0,02	0,00	0,02	0,00	0,01	0,01	0,03	-1,00	1,00	0,10	0,09	0,05
ANTIGUIDADE	0,08	-0,02	-0,01	-0,08	0,03	-0,05	0,03	0,00	-0,07	-0,02	-0,02	0,00	-0,02	0,01	-0,01	0,00	0,00	-0,02	-0,01	-0,02	0,05	-0,01	0,01	0,02	-0,01	0,02	-0,01	-0,01	-0,02	-0,10	0,10	1,00	0,24	-0,05
IDADE	0,18	0,02	-0,01	-0,27	0,19	0,08	0,01	0,00	0,12	-0,01	0,01	0,01	0,00	0,03	0,03	0,03	0,01	0,03	0,03	0,03	0,09	0,00	0,01	0,02	0,02	0,03	0,02	0,03	0,02	-0,09	0,09	0,24	1,00	-0,15
Contactablity	-0,02	0,07	0,01	-0,01	-0,01	-0,04	-0,02	0,00	-0,01	0,01	-0,02	-0,01	0,00	-0,01	-0,02	-0,02	-0,02	0,00	0,02	-0,03	-0,01	0,06	0,01	-0,01	0,00	0,03	0,03	-0,02	-0,02	-0,03	0,05	-0,05	-0,15	1,00

Table 10 - Overview of the demographic correlation matrix

7.2.1. Section 1 of the demographic correlation matrix

	ESTADO_CIVILCasado	ESTADO_CIVILDivorciado	ESTADO_CIVILSeparado	ESTADO_CIVILSolteiro	ESTADO_CIVILViúvo	CANAL_PREFERENCIALCTT	CANAL_PREFERENCIALDCC	CANAL_PREFERENCIALGCP.PROTOCOLOS	CANAL_PREFERENCIALOUTRAS	distrito_agg_ctt_demAçores	distrito_agg_ctt_demAveiro	distrito_agg_ctt_demBeja	distrito_agg_ctt_demBraga	distrito_agg_ctt_demBragança	distrito_agg_ctt_demCastelo.Branco
ESTADO_CIVILCasado	1,00	-0,28	-0,04	-0,86	-0,22	0,01	0,00	0,00	-0,05	0,00	0,02	-0,01	0,03	0,01	0,01
ESTADO_CIVILDivorciado	-0,28	1,00	-0,01	-0,13	-0,03	-0,01	-0,01	0,01	0,04	0,01	0,00	0,00	-0,02	0,00	0,00
ESTADO_CIVILSeparado	-0,04	-0,01	1,00	-0,02	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00
ESTADO_CIVILSolteiro	-0,86	-0,13	-0,02	1,00	-0,11	0,00	0,01	0,00	0,04	0,00	-0,02	0,01	-0,02	-0,01	-0,01
ESTADO_CIVILViúvo	-0,22	-0,03	0,00	-0,11	1,00	0,00	-0,01	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	0,01
CANAL_PREFERENCIALCTT	0,01	-0,01	0,00	0,00	0,00	1,00	-0,02	0,00	-0,19	-0,01	0,05	-0,01	0,01	0,00	0,00
CANAL_PREFERENCIALDCC	0,00	-0,01	0,00	0,01	-0,01	-0,02	1,00	0,00	-0,12	-0,01	-0,01	-0,01	0,00	-0,01	-0,01
CANAL_PREFERENCIALGCP.PROTOCOLOS	0,00	0,01	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CANAL_PREFERENCIALOUTRAS	-0,05	0,04	0,00	0,04	-0,01	-0,19	-0,12	0,00	1,00	0,03	-0,01	0,04	0,02	0,03	0,00
distrito_agg_ctt_demAçores	0,00	0,01	0,01	0,00	0,00	-0,01	-0,01	0,00	0,03	1,00	-0,03	-0,01	-0,03	-0,02	-0,02
distrito_agg_ctt_demAveiro	0,02	0,00	0,00	-0,02	0,00	0,05	-0,01	0,00	-0,01	-0,03	1,00	-0,03	-0,07	-0,03	-0,04
distrito_agg_ctt_demBeja	-0,01	0,00	0,00	0,01	0,00	-0,01	-0,01	0,00	0,04	-0,01	-0,03	1,00	-0,03	-0,01	-0,02
distrito_agg_ctt_demBraga	0,03	-0,02	0,00	-0,02	0,00	0,01	0,00	0,00	0,02	-0,03	-0,07	-0,03	1,00	-0,03	-0,04
distrito_agg_ctt_demBragança	0,01	0,00	0,00	-0,01	0,00	0,00	-0,01	0,00	0,03	-0,02	-0,03	-0,01	-0,03	1,00	-0,02
distrito_agg_ctt_demCastelo.Branco	0,01	0,00	0,00	-0,01	0,01	0,00	-0,01	0,00	0,00	-0,02	-0,04	-0,02	-0,04	-0,02	1,00
distrito_agg_ctt_demCoimbra	0,01	0,00	0,00	-0,01	0,00	0,01	-0,02	0,00	0,04	-0,03	-0,06	-0,03	-0,06	-0,03	-0,03
distrito_agg_ctt_demÉvora	-0,01	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,04	-0,01	-0,03	-0,01	-0,03	-0,01	-0,02
distrito_agg_ctt_demFaro	-0,01	0,01	0,01	0,01	0,00	-0,02	-0,02	0,00	0,06	-0,02	-0,05	-0,02	-0,05	-0,03	-0,03

Table 11 - First section of the demographic correlation matrix

7.2.2. Section 2 of the demographic correlation matrix

	ESTADO_CIVILCasado	ESTADO_CIVILDivorciado	ESTADO_CIVILSeparado	ESTADO_CIVILSolteiro	ESTADO_CIVILViúvo	CANAL_PREFERENCIALCTT	CANAL_PREFERENCIALDCC	CANAL_PREFERENCIALGCP.PROTOCOLOS	CANAL_PREFERENCIALOUTRAS	distrito_agg_ctt_demAçores	distrito_agg_ctt_demAveiro	distrito_agg_ctt_demBeja	distrito_agg_ctt_demBraga	distrito_agg_ctt_demBragança	distrito_agg_ctt_demCastelo.Branco
distrito_agg_ctt_demGuarda	0,01	0,00	0,00	-0,01	0,01	0,01	-0,01	0,00	0,03	-0,02	-0,03	-0,01	-0,03	-0,02	-0,02
distrito_agg_ctt_demLeiria	0,00	0,00	0,00	0,00	0,01	0,00	-0,02	0,00	0,05	-0,03	-0,06	-0,02	-0,06	-0,03	-0,03
distrito_agg_ctt_demLisboa	-0,04	0,04	0,00	0,02	0,02	0,00	0,03	0,01	0,01	-0,06	-0,13	-0,06	-0,14	-0,06	-0,07
distrito_agg_ctt_demMadeira	-0,01	0,00	0,00	0,01	0,00	-0,01	-0,01	0,00	0,02	-0,01	-0,03	-0,01	-0,03	-0,01	-0,02
distrito_agg_ctt_demPortalegre	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,02	-0,01	-0,03	-0,01	-0,03	-0,01	-0,01
distrito_agg_ctt_demPorto	-0,01	0,00	0,00	0,01	0,00	0,03	0,07	0,00	0,02	-0,05	-0,10	-0,05	-0,11	-0,05	-0,06
distrito_agg_ctt_demSantarém	0,00	0,00	0,00	0,00	0,00	-0,01	-0,02	0,00	0,06	-0,03	-0,06	-0,03	-0,06	-0,03	-0,03
distrito_agg_ctt_demSetúbal	-0,02	0,03	0,00	0,01	0,01	-0,01	0,01	0,00	0,01	-0,03	-0,07	-0,03	-0,07	-0,03	-0,04
distrito_agg_ctt_demViana.do.Castelo	0,00	-0,01	0,00	0,00	0,00	-0,01	-0,01	0,00	0,06	-0,02	-0,04	-0,02	-0,04	-0,02	-0,02
distrito_agg_ctt_demVila.Real	0,02	-0,01	0,00	-0,02	0,01	0,00	-0,01	0,00	0,00	-0,02	-0,04	-0,02	-0,04	-0,02	-0,02
distrito_agg_ctt_demViseu	0,03	-0,02	0,00	-0,02	0,00	0,01	-0,01	0,00	0,02	-0,02	-0,05	-0,02	-0,05	-0,02	-0,03
SEXOFeminino	-0,13	0,08	0,00	0,05	0,14	0,05	0,00	0,00	-0,01	-0,01	-0,01	-0,01	-0,03	-0,02	-0,01
SEXOMasculino	0,13	-0,08	0,00	-0,05	-0,14	-0,05	0,00	0,00	0,01	0,01	0,01	0,01	0,03	0,02	0,01
ANTIGUIDADE	0,08	-0,02	-0,01	-0,08	0,03	-0,05	0,03	0,00	-0,07	-0,02	-0,02	0,00	-0,02	0,01	-0,01
IDADE	0,18	0,02	-0,01	-0,27	0,19	0,08	0,01	0,00	0,12	-0,01	0,01	0,01	0,00	0,03	0,03
Contactablity	-0,02	0,07	0,01	-0,01	-0,01	-0,04	-0,02	0,00	-0,01	0,01	-0,02	-0,01	0,00	-0,01	-0,02

Table 12 - Second section of the demographic correlation matrix

7.2.3. Section 3 of the demographic correlation matrix

	distrito_agg_ctt_demCoimbra	distrito_agg_ctt_demÉvora	distrito_agg_ctt_demFaro	distrito_agg_ctt_demGuarda	distrito_agg_ctt_demLeiria	distrito_agg_ctt_demLisboa	distrito_agg_ctt_demMadeira	distrito_agg_ctt_demPortalegre	distrito_agg_ctt_demPorto
ESTADO_CIVILCasado	0,01	-0,01	-0,01	0,01	0,00	-0,04	-0,01	0,00	-0,01
ESTADO_CIVILDivorciado	0,00	0,00	0,01	0,00	0,00	0,04	0,00	0,00	0,00
ESTADO_CIVILSeparado	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
ESTADO_CIVILSolteiro	-0,01	0,00	0,01	-0,01	0,00	0,02	0,01	0,00	0,01
ESTADO_CIVILViúvo	0,00	0,00	0,00	0,01	0,01	0,02	0,00	0,00	0,00
CANAL_PREFERENCIALCTT	0,01	-0,01	-0,02	0,01	0,00	0,00	-0,01	-0,01	0,03
CANAL_PREFERENCIALDCC	-0,02	-0,01	-0,02	-0,01	-0,02	0,03	-0,01	-0,01	0,07
CANAL_PREFERENCIALGCP.PROTOCOLOS	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00
CANAL_PREFERENCIALOUTRAS	0,04	0,04	0,06	0,03	0,05	0,01	0,02	0,02	0,02
distrito_agg_ctt_demAçores	-0,03	-0,01	-0,02	-0,02	-0,03	-0,06	-0,01	-0,01	-0,05
distrito_agg_ctt_demAveiro	-0,06	-0,03	-0,05	-0,03	-0,06	-0,13	-0,03	-0,03	-0,10
distrito_agg_ctt_demBeja	-0,03	-0,01	-0,02	-0,01	-0,02	-0,06	-0,01	-0,01	-0,05
distrito_agg_ctt_demBraga	-0,06	-0,03	-0,05	-0,03	-0,06	-0,14	-0,03	-0,03	-0,11
distrito_agg_ctt_demBragança	-0,03	-0,01	-0,03	-0,02	-0,03	-0,06	-0,01	-0,01	-0,05
distrito_agg_ctt_demCastelo.Branco	-0,03	-0,02	-0,03	-0,02	-0,03	-0,07	-0,02	-0,01	-0,06
distrito_agg_ctt_demCoimbra	1,00	-0,03	-0,05	-0,03	-0,05	-0,11	-0,03	-0,02	-0,09
distrito_agg_ctt_demÉvora	-0,03	1,00	-0,02	-0,01	-0,02	-0,06	-0,01	-0,01	-0,04
distrito_agg_ctt_demFaro	-0,05	-0,02	1,00	-0,03	-0,04	-0,10	-0,02	-0,02	-0,08

Table 13 - Third section of the demographic correlation matrix

7.2.4. Section 4 of the demographic correlation matrix

	distrito_agg_ctt_demCoimbra	distrito_agg_ctt_demÉvora	distrito_agg_ctt_demFaro	distrito_agg_ctt_demGuarda	distrito_agg_ctt_demLeiria	distrito_agg_ctt_demLisboa	distrito_agg_ctt_demMadeira	distrito_agg_ctt_demPortalegre	distrito_agg_ctt_demPorto
distrito_agg_ctt_demGuarda	-0,03	-0,01	-0,03	1,00	-0,03	-0,07	-0,02	-0,01	-0,05
distrito_agg_ctt_demLeiria	-0,05	-0,02	-0,04	-0,03	1,00	-0,11	-0,03	-0,02	-0,08
distrito_agg_ctt_demLisboa	-0,11	-0,06	-0,10	-0,07	-0,11	1,00	-0,06	-0,05	-0,20
distrito_agg_ctt_demMadeira	-0,03	-0,01	-0,02	-0,02	-0,03	-0,06	1,00	-0,01	-0,05
distrito_agg_ctt_demPortalegre	-0,02	-0,01	-0,02	-0,01	-0,02	-0,05	-0,01	1,00	-0,04
distrito_agg_ctt_demPorto	-0,09	-0,04	-0,08	-0,05	-0,08	-0,20	-0,05	-0,04	1,00
distrito_agg_ctt_demSantarém	-0,05	-0,03	-0,05	-0,03	-0,05	-0,11	-0,03	-0,02	-0,09
distrito_agg_ctt_demSetúbal	-0,06	-0,03	-0,05	-0,03	-0,06	-0,14	-0,03	-0,03	-0,11
distrito_agg_ctt_demViana.do.Castelo	-0,03	-0,02	-0,03	-0,02	-0,03	-0,07	-0,02	-0,02	-0,06
distrito_agg_ctt_demVila.Real	-0,03	-0,02	-0,03	-0,02	-0,03	-0,07	-0,02	-0,01	-0,05
distrito_agg_ctt_demViseu	-0,05	-0,02	-0,04	-0,03	-0,04	-0,10	-0,02	-0,02	-0,08
SEXOFeminino	-0,01	-0,01	0,01	-0,01	-0,02	0,04	0,00	-0,02	0,00
SEXOMasculino	0,01	0,01	-0,01	0,01	0,02	-0,04	0,00	0,02	0,00
ANTIGUIDADE	0,00	0,00	-0,02	-0,01	-0,02	0,05	-0,01	0,01	0,02
IDADE	0,03	0,01	0,03	0,03	0,03	0,09	0,00	0,01	0,02
Contactablity	-0,02	0,00	0,02	-0,03	-0,01	0,06	0,01	-0,01	0,00

Table 14 - Fourth section of the demographic correlation matrix

7.2.5. Section 5 of the demographic correlation matrix

	distrito_agg_ctt_demSantarém	distrito_agg_ctt_demSetúbal	distrito_agg_ctt_demViana.do.Castelo	distrito_agg_ctt_demVila.Real	distrito_agg_ctt_demViseu	SEXOFeminino	SEXOMasculino	ANTIGUIDADE	IDADE	Contactablity
ESTADO_CIVILCasado	0,00	-0,02	0,00	0,02	0,03	-0,13	0,13	0,08	0,18	-0,02
ESTADO_CIVILDivorciado	0,00	0,03	-0,01	-0,01	-0,02	0,08	-0,08	-0,02	0,02	0,07
ESTADO_CIVILSeparado	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,01
ESTADO_CIVILSolteiro	0,00	0,01	0,00	-0,02	-0,02	0,05	-0,05	-0,08	-0,27	-0,01
ESTADO_CIVILViúvo	0,00	0,01	0,00	0,01	0,00	0,14	-0,14	0,03	0,19	-0,01
CANAL_PREFERENCIALCTT	-0,01	-0,01	-0,01	0,00	0,01	0,05	-0,05	-0,05	0,08	-0,04
CANAL_PREFERENCIALDCC	-0,02	0,01	-0,01	-0,01	-0,01	0,00	0,00	0,03	0,01	-0,02
CANAL_PREFERENCIALGCP.PROTOCOLOS	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CANAL_PREFERENCIALOUTRAS	0,06	0,01	0,06	0,00	0,02	-0,01	0,01	-0,07	0,12	-0,01
distrito_agg_ctt_demAçores	-0,03	-0,03	-0,02	-0,02	-0,02	-0,01	0,01	-0,02	-0,01	0,01
distrito_agg_ctt_demAveiro	-0,06	-0,07	-0,04	-0,04	-0,05	-0,01	0,01	-0,02	0,01	-0,02
distrito_agg_ctt_demBeja	-0,03	-0,03	-0,02	-0,02	-0,02	-0,01	0,01	0,00	0,01	-0,01
distrito_agg_ctt_demBraga	-0,06	-0,07	-0,04	-0,04	-0,05	-0,03	0,03	-0,02	0,00	0,00
distrito_agg_ctt_demBragança	-0,03	-0,03	-0,02	-0,02	-0,02	-0,02	0,02	0,01	0,03	-0,01
distrito_agg_ctt_demCastelo.Branco	-0,03	-0,04	-0,02	-0,02	-0,03	-0,01	0,01	-0,01	0,03	-0,02
distrito_agg_ctt_demCoimbra	-0,05	-0,06	-0,03	-0,03	-0,05	-0,01	0,01	0,00	0,03	-0,02
distrito_agg_ctt_demÉvora	-0,03	-0,03	-0,02	-0,02	-0,02	-0,01	0,01	0,00	0,01	0,00
distrito_agg_ctt_demFaro	-0,05	-0,05	-0,03	-0,03	-0,04	0,01	-0,01	-0,02	0,03	0,02

Table 15 - Fifth section of the demographic correlation matrix

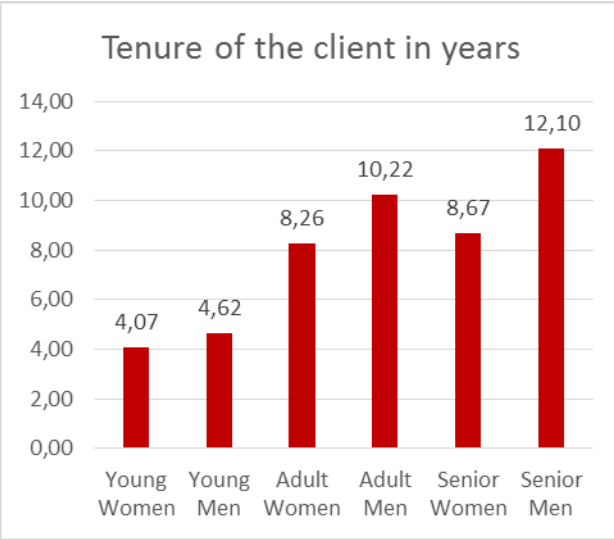
7.2.6. Section 6 of the demographic correlation matrix

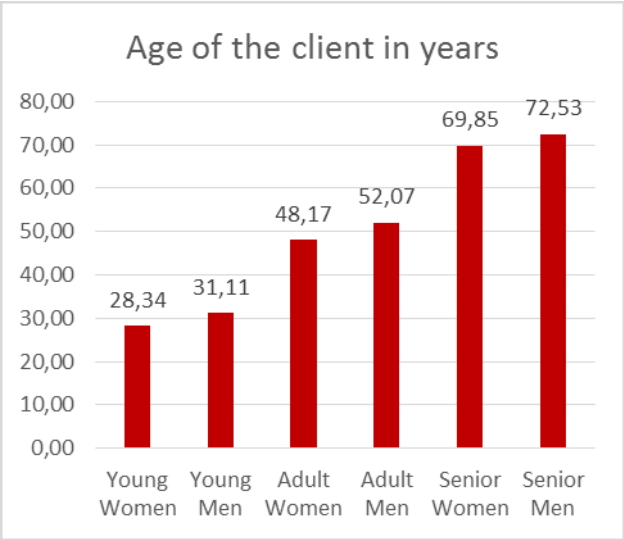
	distrito_agg_ctt_demSantarém	distrito_agg_ctt_demSetúbal	distrito_agg_ctt_demViana.do.Castelo	distrito_agg_ctt_demVila.Real	distrito_agg_ctt_demViseu	SEXOFeminino	SEXOMasculino	ANTIGUIDADE	IDADE	Contactablity
distrito_agg_ctt_demGuarda	-0,03	-0,03	-0,02	-0,02	-0,03	-0,01	0,01	-0,01	0,03	-0,03
distrito_agg_ctt_demLeiria	-0,05	-0,06	-0,03	-0,03	-0,04	-0,02	0,02	-0,02	0,03	-0,01
distrito_agg_ctt_demLisboa	-0,11	-0,14	-0,07	-0,07	-0,10	0,04	-0,04	0,05	0,09	0,06
distrito_agg_ctt_demMadeira	-0,03	-0,03	-0,02	-0,02	-0,02	0,00	0,00	-0,01	0,00	0,01
distrito_agg_ctt_demPortalegre	-0,02	-0,03	-0,02	-0,01	-0,02	-0,02	0,02	0,01	0,01	-0,01
distrito_agg_ctt_demPorto	-0,09	-0,11	-0,06	-0,05	-0,08	0,00	0,00	0,02	0,02	0,00
distrito_agg_ctt_demSantarém	1,00	-0,06	-0,03	-0,03	-0,05	-0,02	0,02	-0,01	0,02	0,03
distrito_agg_ctt_demSetúbal	-0,06	1,00	-0,04	-0,04	-0,05	0,00	0,00	0,02	0,03	0,03
distrito_agg_ctt_demViana.do.Castelo	-0,03	-0,04	1,00	-0,02	-0,03	-0,01	0,01	-0,01	0,02	-0,02
distrito_agg_ctt_demVila.Real	-0,03	-0,04	-0,02	1,00	-0,03	-0,01	0,01	-0,01	0,03	-0,02
distrito_agg_ctt_demViseu	-0,05	-0,05	-0,03	-0,03	1,00	-0,03	0,03	-0,02	0,02	-0,03
SEXOFeminino	-0,02	0,00	-0,01	-0,01	-0,03	1,00	-1,00	-0,10	-0,09	-0,05
SEXOMasculino	0,02	0,00	0,01	0,01	0,03	-1,00	1,00	0,10	0,09	0,05
ANTIGUIDADE	-0,01	0,02	-0,01	-0,01	-0,02	-0,10	0,10	1,00	0,24	-0,05
IDADE	0,02	0,03	0,02	0,03	0,02	-0,09	0,09	0,24	1,00	-0,15
Contactablity	0,03	0,03	-0,02	-0,02	-0,03	-0,05	0,05	-0,05	-0,15	1,00

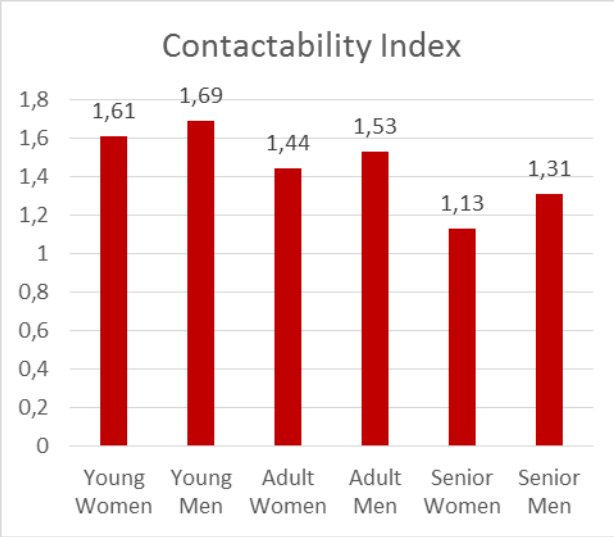
Table 16 - Sixth section of the demographic correlation matrix

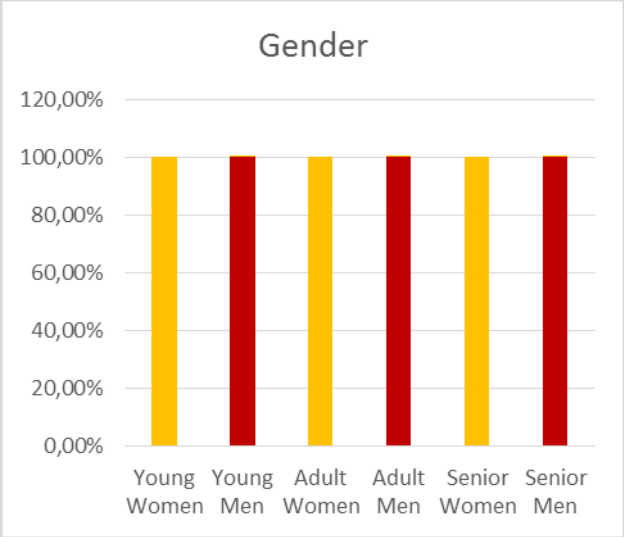
7.3. PROFILING OF DEMOGRAPHIC SEGMENTS

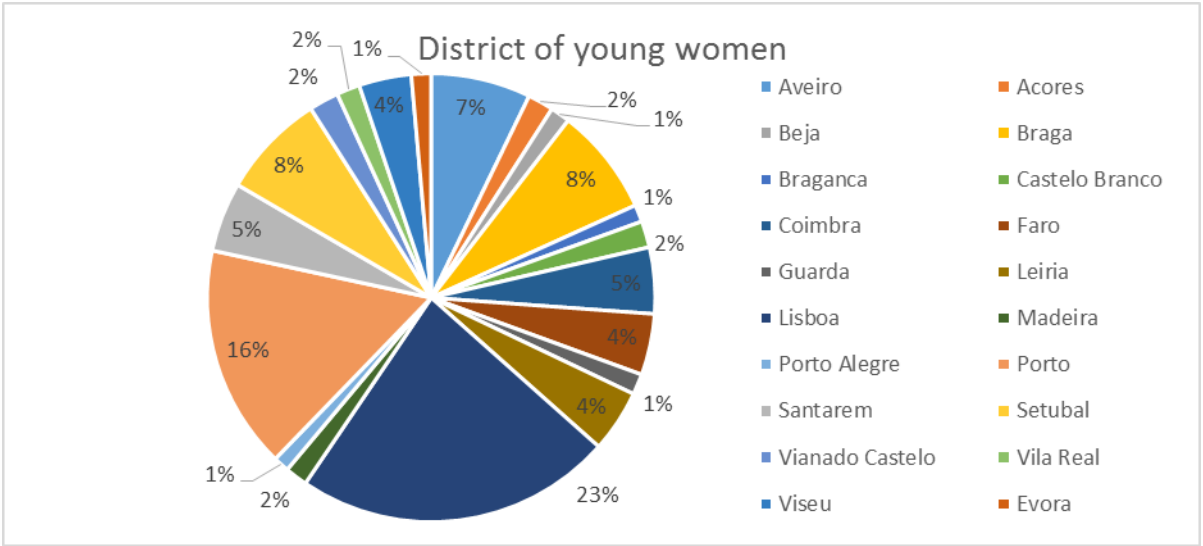
This section displays the means of all used variables for the demographic segmentation.

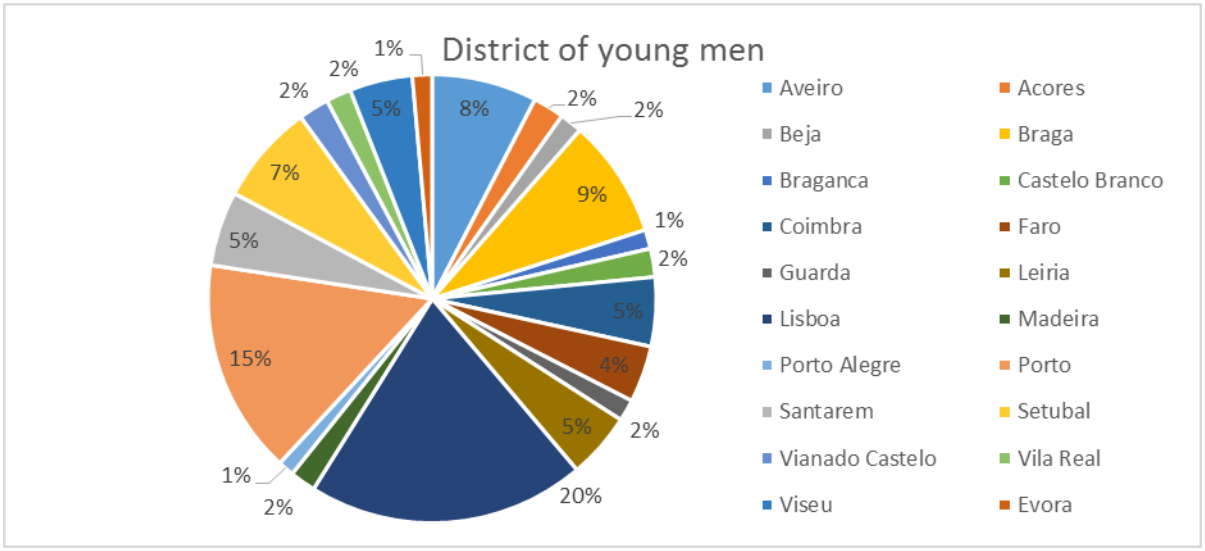

Tenure of the client in years
14,00
12,00
10,00
8,00
6,00
4,00
2,00
0,00
4,07
4,62
8,26
10,22
8,67
12,10
Young Women
Young Men
Adult Women
Adult Men
Senior Women
Senior Men

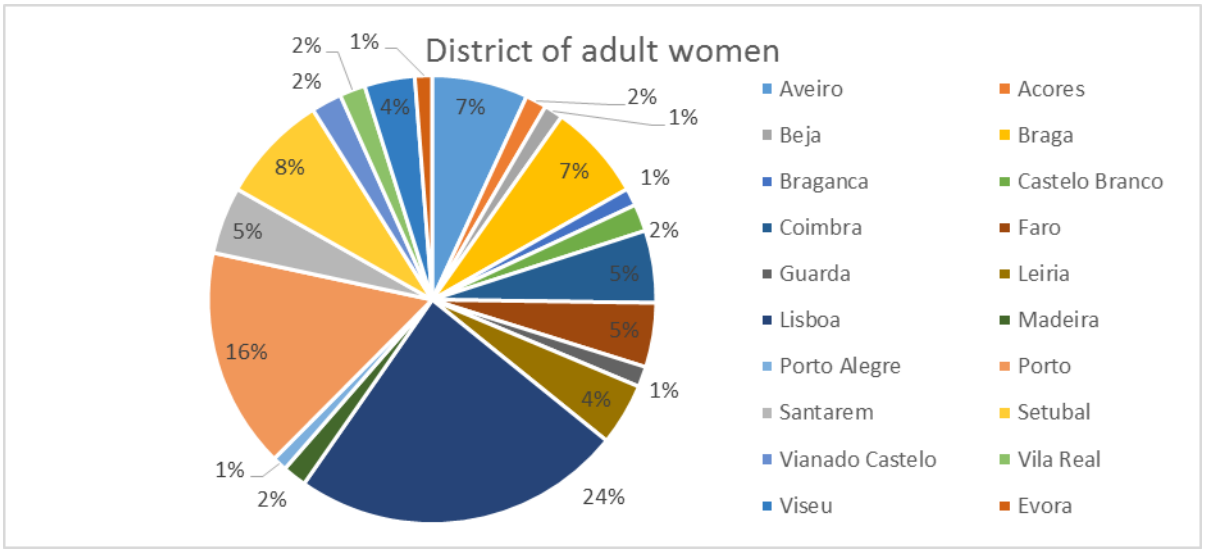

Age of the client in years
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
28,34
31,11
48,17
52,07
69,85
72,53
Young Women
Young Men
Adult Women
Adult Men
Senior Women
Senior Men

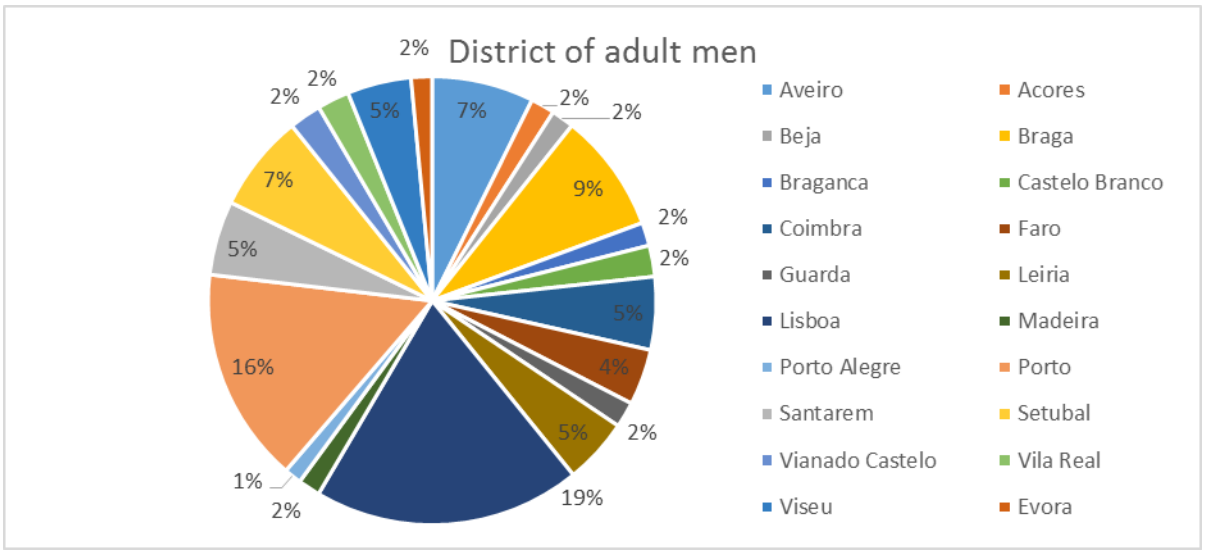

Contactability Index
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
1,61
1,69
1,44
1,53
1,13
1,31
Young Women
Young Men
Adult Women
Adult Men
Senior Women
Senior Men

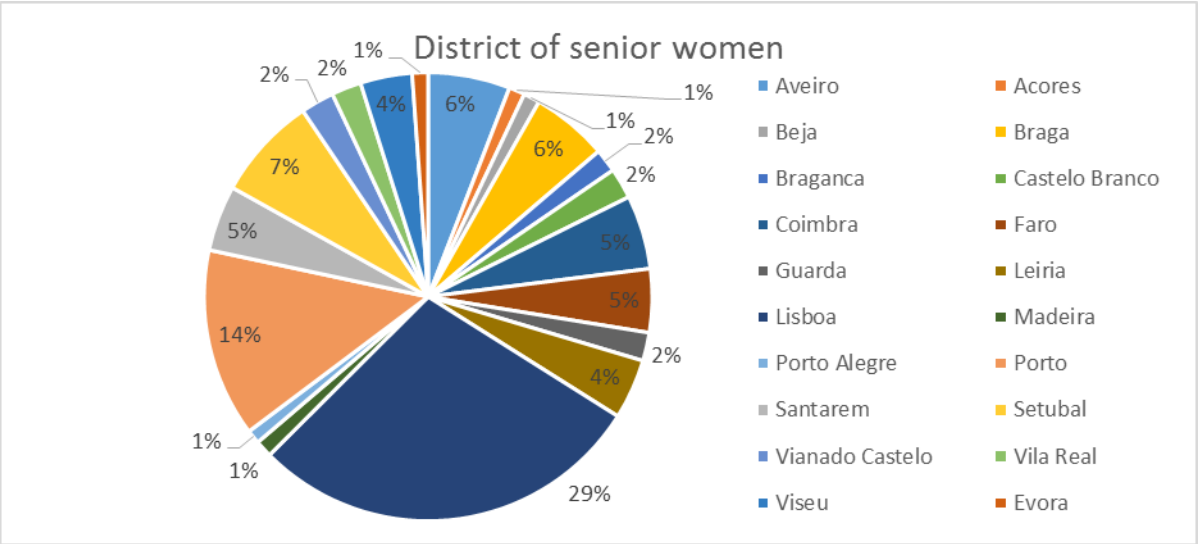

Gender
120,00%
100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
Young Women
Young Men
Adult Women
Adult Men
Senior Women
Senior Men

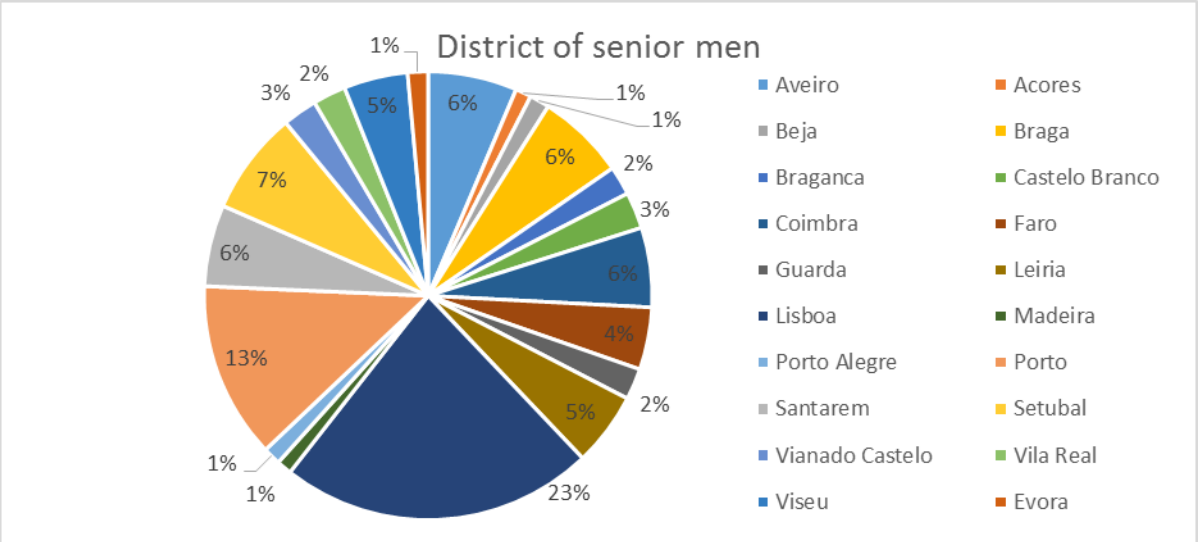

District of young women
7%
2%
1%
8%
1%
2%
5%
4%
1%
23%
2%
1%
16%
5%
8%
2%
2%
1%
4%
Aveiro
Acores
Beja
Braga
Braganca
Castelo Branco
Coimbra
Faro
Guarda
Leiria
Lisboa
Madeira
Porto Alegre
Porto
Santarem
Setubal
Vianado Castelo
Vila Real
Viseu
Evora

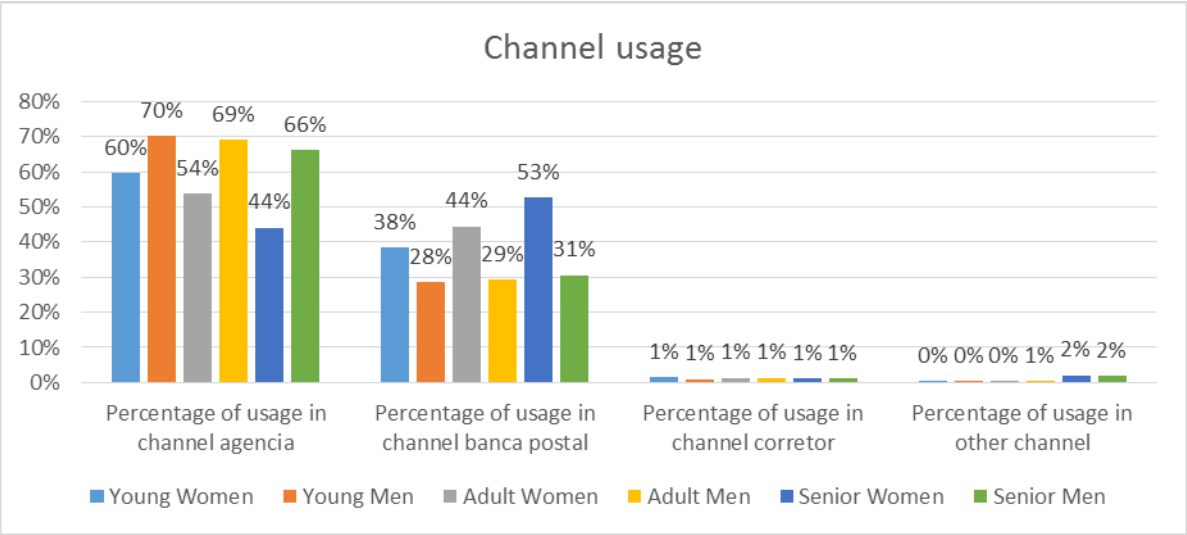

District of young men
8%
2%
2%
9%
1%
2%
5%
4%
2%
5%
20%
2%
1%
15%
5%
7%
2%
2%
5%
1%
Aveiro
Acores
Beja
Braga
Braganca
Castelo Branco
Coimbra
Faro
Guarda
Leiria
Lisboa
Madeira
Porto Alegre
Porto
Santarem
Setubal
Vianado Castelo
Vila Real
Viseu
Evora

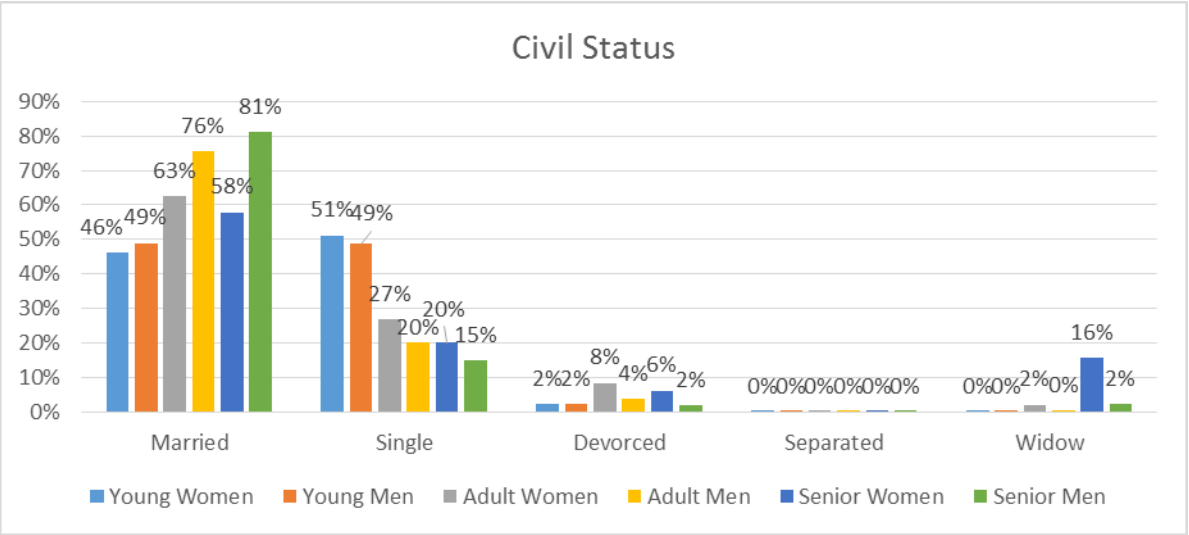

District of adult women
7%
2%
1%
7%
1%
2%
5%
5%
1%
4%
24%
2%
1%
16%
5%
8%
2%
2%
4%
1%
Aveiro
Acores
Beja
Braga
Braganca
Castelo Branco
Coimbra
Faro
Guarda
Leiria
Lisboa
Madeira
Porto Alegre
Porto
Santarem
Setubal
Vianado Castelo
Vila Real
Viseu
Evora


District of adult men
7%
2%
2%
9%
2%
2%
5%
4%
2%
5%
19%
2%
1%
16%
5%
7%
2%
2%
5%
2%
Aveiro
Acores
Beja
Braga
Braganca
Castelo Branco
Coimbra
Faro
Guarda
Leiria
Lisboa
Madeira
Porto Alegre
Porto
Santarem
Setubal
Vianado Castelo
Vila Real
Viseu
Evora


District of senior women
6%
1%
1%
2%
2%
5%
5%
2%
4%
29%
1%
1%
14%
5%
7%
2%
2%
4%
Aveiro
Acores
Beja
Braga
Braganca
Castelo Branco
Coimbra
Faro
Guarda
Leiria
Lisboa
Madeira
Porto Alegre
Porto
Santarem
Setubal
Vianado Castelo
Vila Real
Viseu
Evora

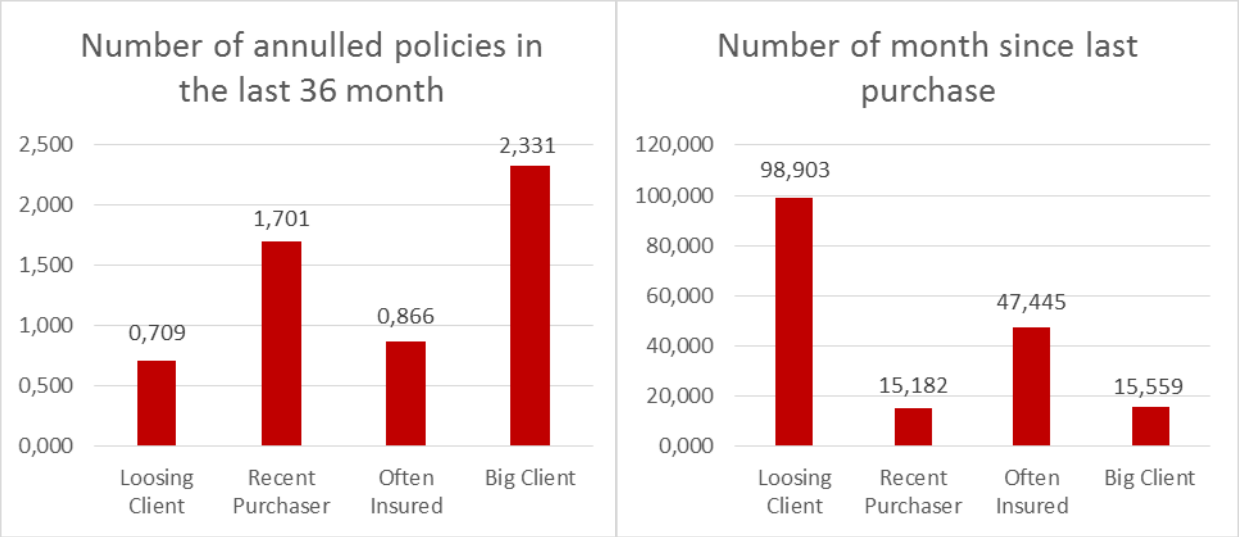

District of senior men
6%
1%
1%
2%
6%
3%
6%
4%
2%
5%
23%
1%
1%
13%
6%
7%
3%
2%
5%
Aveiro
Acores
Beja
Braga
Braganca
Castelo Branco
Coimbra
Faro
Guarda
Leiria
Lisboa
Madeira
Porto Alegre
Porto
Santarem
Setubal
Vianado Castelo
Vila Real
Viseu
Evora

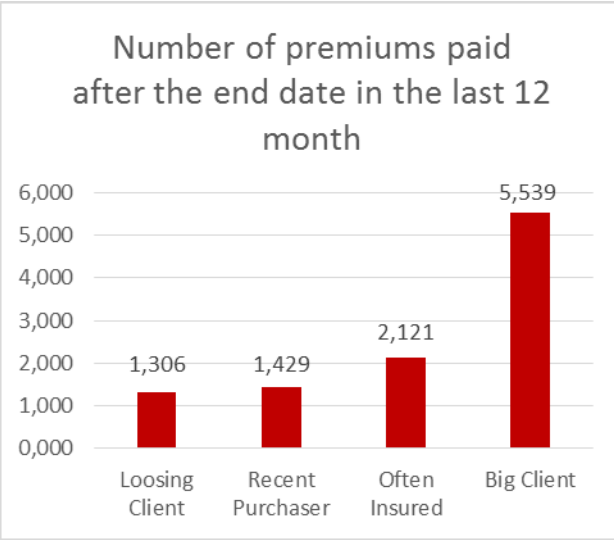

Channel usage
80%
70%
60%
50%
40%
30%
20%
10%
0%
60%
70%
54%
69%
44%
66%
38%
28%
44%
29%
53%
31%
1% 1% 1% 1% 1% 1%
0% 0% 0% 1% 2% 2%
Percentage of usage in channel agencia
Percentage of usage in channel banca postal
Percentage of usage in channel corretor
Percentage of usage in other channel
Young Women
Young Men
Adult Women
Adult Men
Senior Women
Senior Men

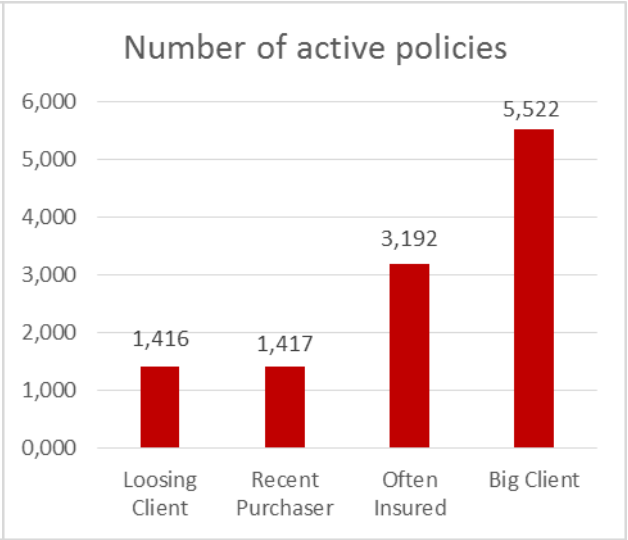


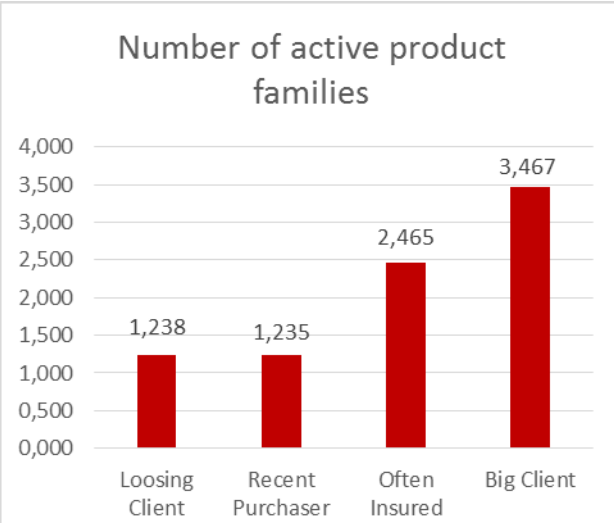
7.4. Profiling of Behavior segments

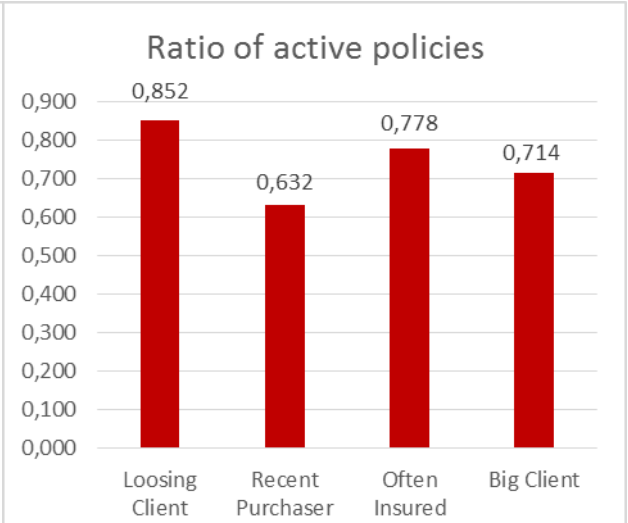
This section gives an overview of the means of all behavior segments in the used variables of the behavior segmentation. Health user are excluded from the graphics because they do not own any policy and are only insured in some else's health policy, and thus do not qualify for the analysis. They are segmented into one cluster because they have in common that they are insured in a health policy but do not own any policy.

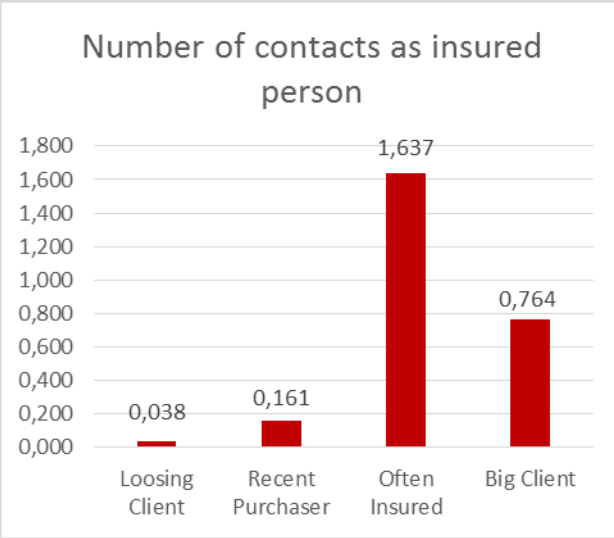


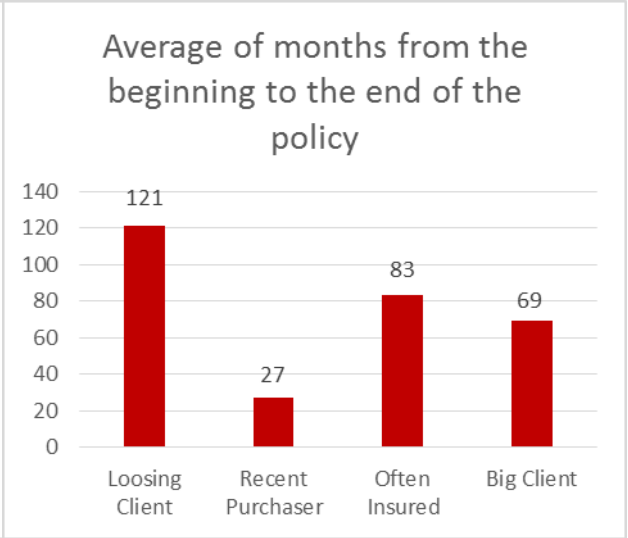

Number of premiums paid after the end date in the last 12 month
6,000
5,000
4,000
3,000
2,000
1,000
0,000
1,306
1,429
2,121
5,539
Loosing Client
Recent Purchaser
Often Insured
Big Client

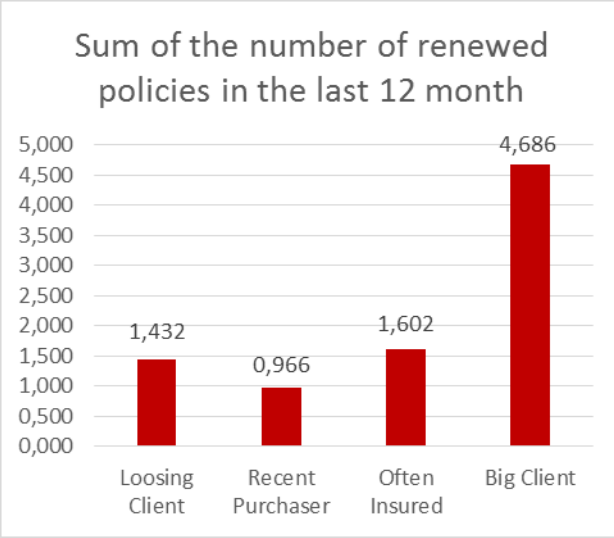

Number of active policies
6,000
5,000
4,000
3,000
2,000
1,000
0,000
1,416
1,417
3,192
5,522
Loosing Client
Recent Purchaser
Often Insured
Big Client

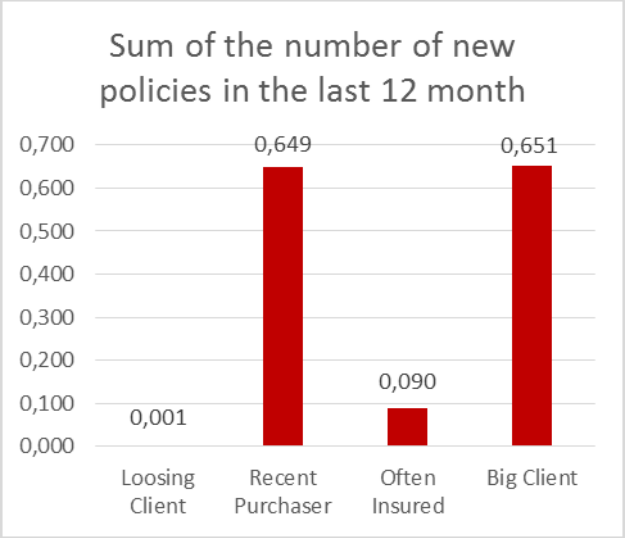

Number of active product families
4,000
3,500
3,000
2,500
2,000
1,500
1,000
0,500
0,000
1,238
1,235
2,465
3,467
Loosing Client
Recent Purchaser
Often Insured
Big Client

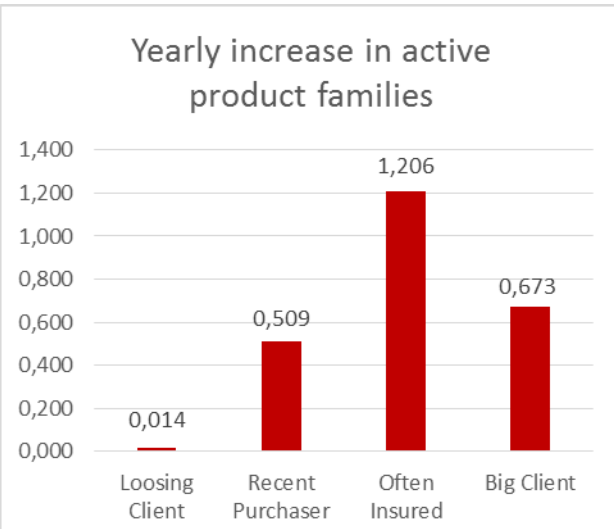

Ratio of active policies
0,900
0,800
0,700
0,600
0,500
0,400
0,300
0,200
0,100
0,000
0,852
0,632
0,778
0,714
Loosing Client
Recent Purchaser
Often Insured
Big Client

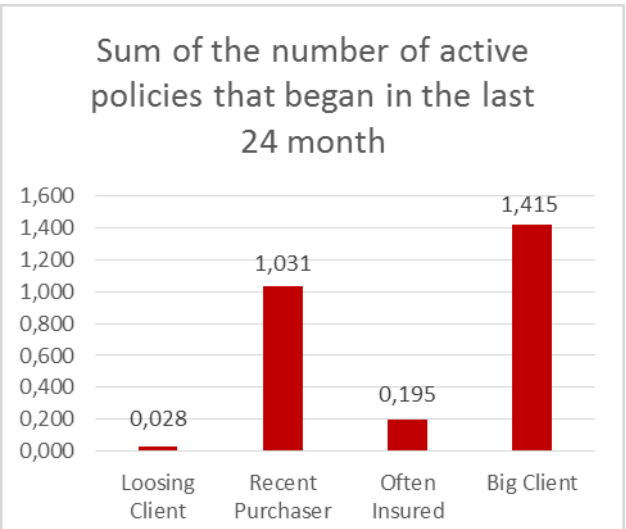

Number of contacts as insured person
1,800
1,600
1,400
1,200
1,000
0,800
0,600
0,400
0,200
0,000
0,038
0,161
1,637
0,764
Loosing Client
Recent Purchaser
Often Insured
Big Client

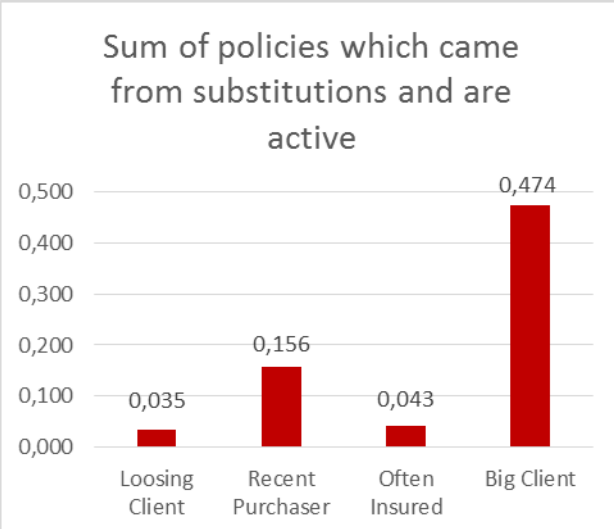

Average of months from the beginning to the end of the policy
140
120
100
80
60
40
20
0
121
27
83
69
Loosing Client
Recent Purchaser
Often Insured
Big Client

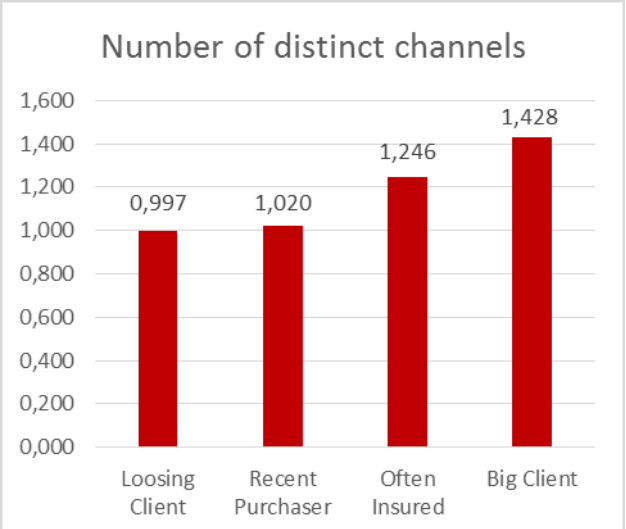

Sum of the number of renewed policies in the last 12 month
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
0,500
0,000
1,432
0,966
1,602
4,686
Loosing Client
Recent Purchaser
Often Insured
Big Client

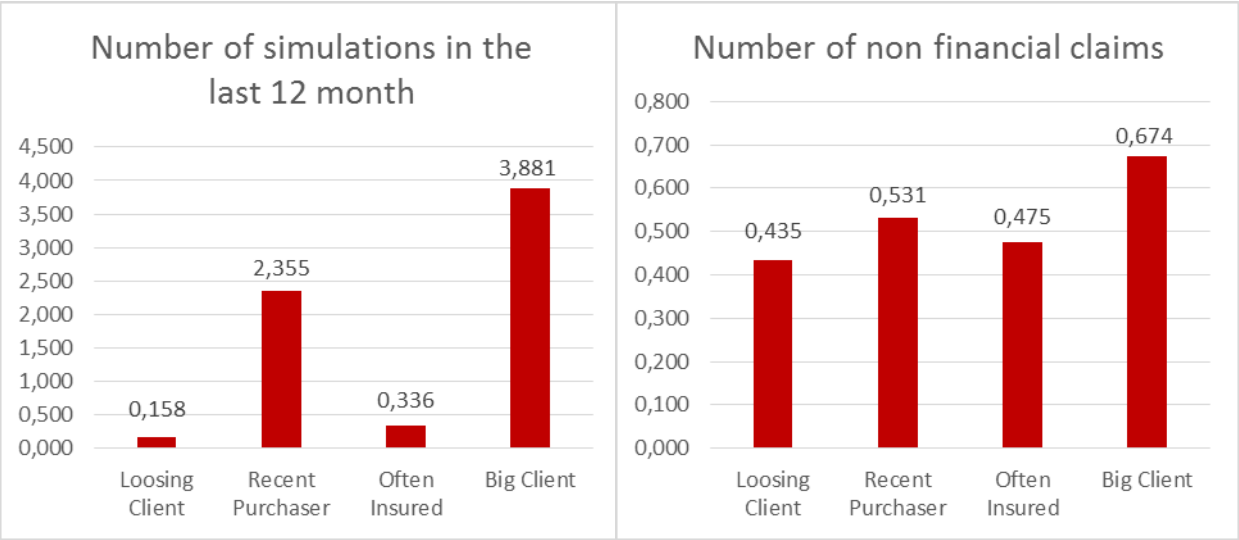

Sum of the number of new policies in the last 12 month
0,700
0,600
0,500
0,400
0,300
0,200
0,100
0,000
0,001
0,649
0,090
0,651
Loosing Client
Recent Purchaser
Often Insured
Big Client

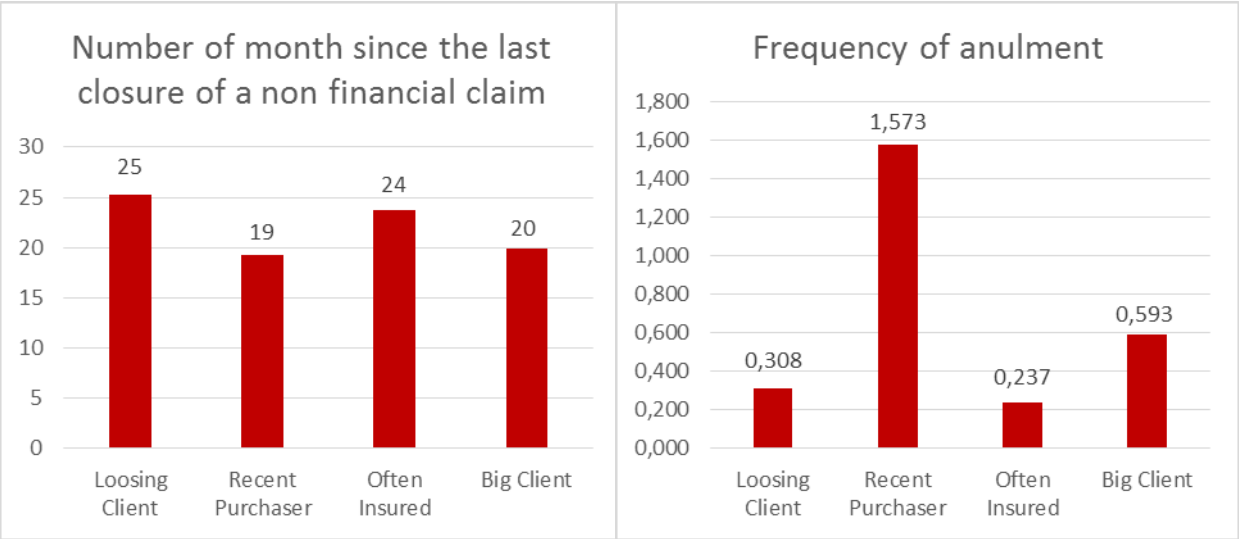

Yearly increase in active product families
1,400
1,200
1,000
0,800
0,600
0,400
0,200
0,000
0,014
0,509
1,206
0,673
Loosing Client
Recent Purchaser
Often Insured
Big Client


Sum of the number of active policies that began in the last 24 month
1,600
1,400
1,200
1,000
0,800
0,600
0,400
0,200
0,000
0,028
1,031
0,195
1,415
Loosing Client
Recent Purchaser
Often Insured
Big Client


Sum of policies which came from substitutions and are active
0,500
0,400
0,300
0,200
0,100
0,000
0,035
0,156
0,043
0,474
Loosing Client
Recent Purchaser
Often Insured
Big Client

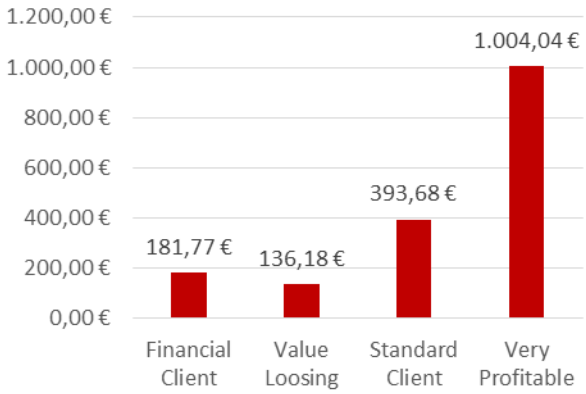

Number of distinct channels
1,600
1,400
1,200
1,000
0,800
0,600
0,400
0,200
0,000
0,997
1,020
1,246
1,428
Loosing Client
Recent Purchaser
Often Insured
Big Client

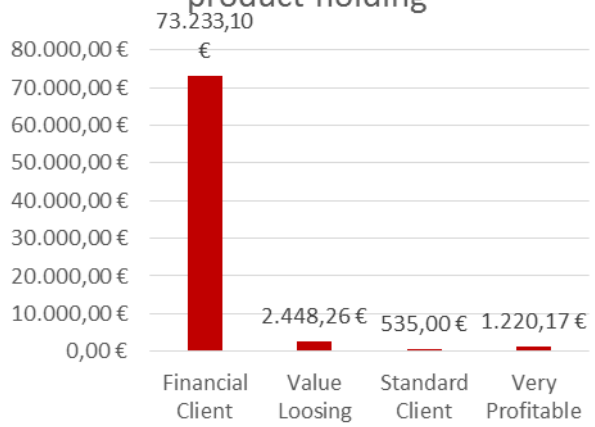


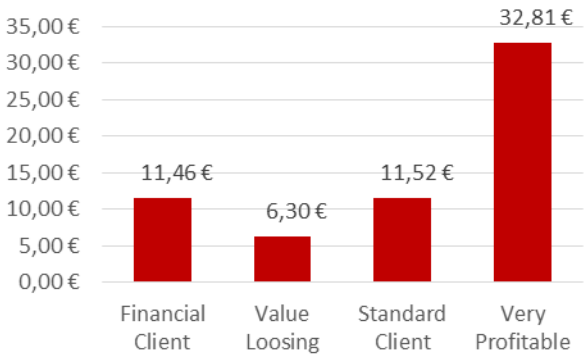


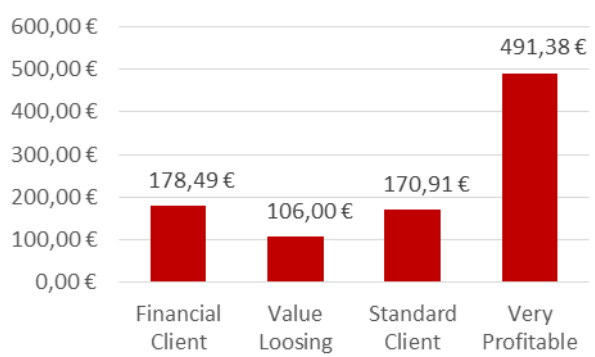
7.5. PROFILING OF VALUE SEGMENTS

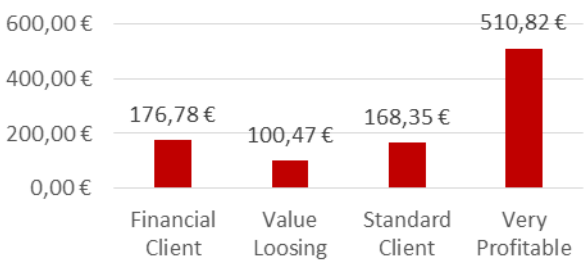
This section gives conclusion about the average values that have been reached by the specific value segments in the variables that have been unsed to create the segments. The main findings are explaned in the principle part of this work. The here presented statistics serve to complete the profiling.

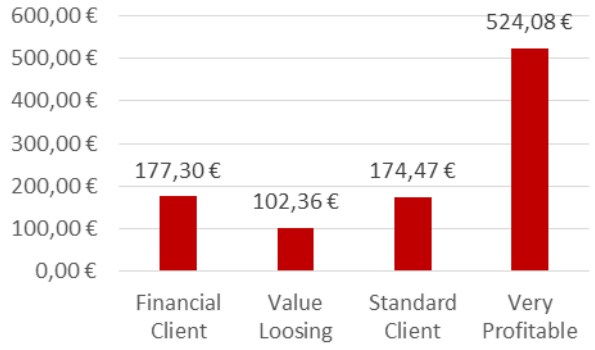

Total PCA of policies in force (non financial products)
1.200,00 €
1.000,00 €
800,00 €
600,00 €
400,00 €
200,00 €
0,00 €
181,77 €
136,18 €
393,68 €
1.004,04 €
Financial Client
Value Loosing
Standard Client
Very Profitable


Current value of financial product holding
80.000,00 €
70.000,00 €
60.000,00 €
50.000,00 €
40.000,00 €
30.000,00 €
20.000,00 €
10.000,00 €
0,00 €
73.233,10 €
2.448,26 €
535,00 €
1.220,17 €
Financial Client
Value Loosing
Standard Client
Very Profitable

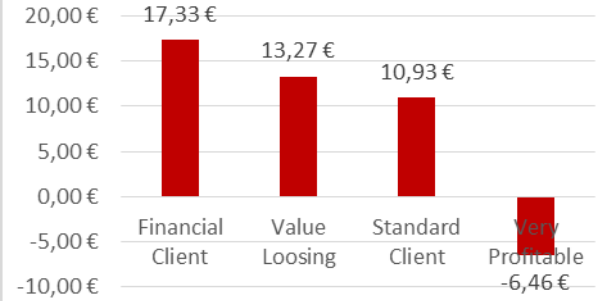

Average of total processed comissions for active policies (non financial products)
35,00 €
30,00 €
25,00 €
20,00 €
15,00 €
10,00 €
5,00 €
0,00 €
11,46 €
6,30 €
11,52 €
32,81 €
Financial Client
Value Loosing
Standard Client
Very Profitable

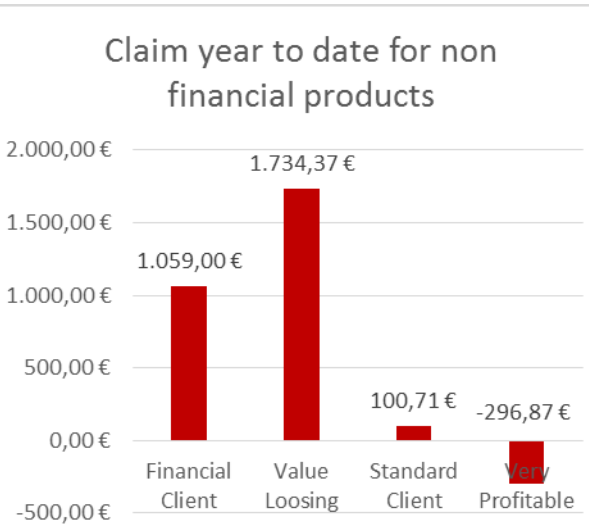

Average of the PCA of the homologous year (non financial products)
600,00 €
500,00 €
400,00 €
300,00 €
200,00 €
100,00 €
0,00 €
178,49 €
106,00 €
170,91 €
491,38 €
Financial Client
Value Loosing
Standard Client
Very Profitable

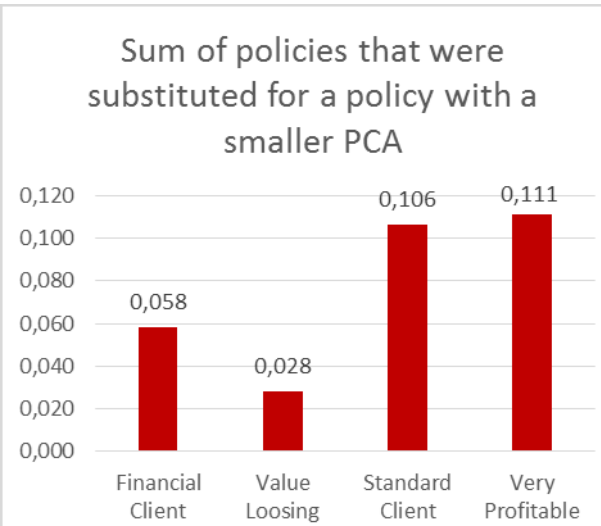

Average PCA of all policies renewed in the last 12 month (by either completing a year or substitution for non financial products)
600,00 €
400,00 €
200,00 €
0,00 €
176,78 €
100,47 €
168,35 €
510,82 €
Financial Client
Value Loosing
Standard Client
Very Profitable

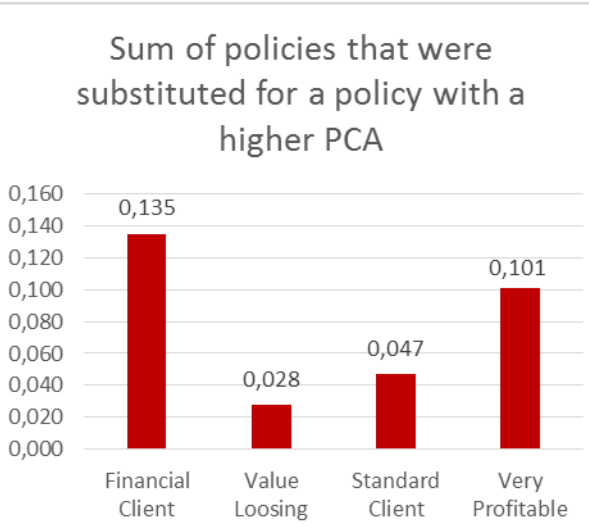

Average of all PCA from active policies (non financial products)
600,00 €
500,00 €
400,00 €
300,00 €
200,00 €
100,00 €
0,00 €
177,30 €
102,36 €
174,47 €
524,08 €
Financial Client
Value Loosing
Standard Client
Very Profitable

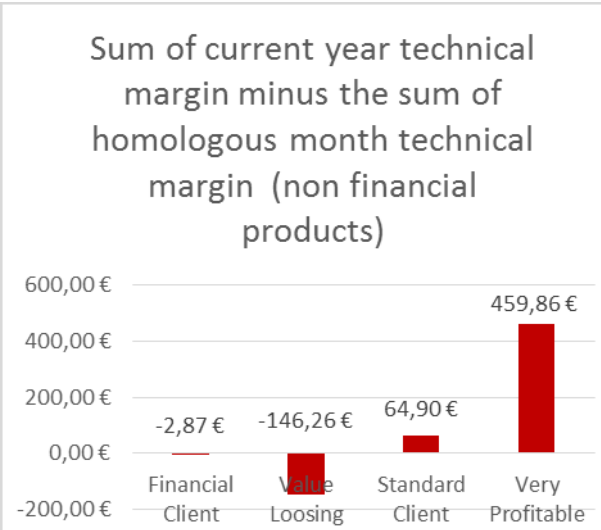
Sum of current year processed comissions divided by the sum of processed premiums (non financial products)
0,120
0,100
0,080
0,060
0,040
0,020
0,000
0,095
0,070
0,099
0,096
Financial Client
Value Loosing
Standard Client
Very Profitable

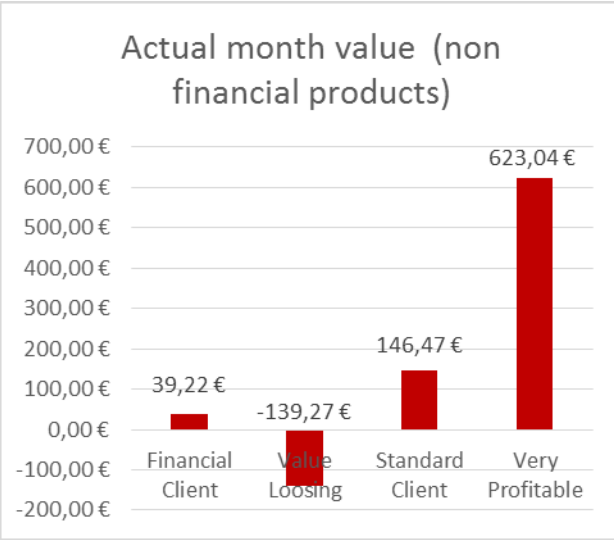

Average of current year PCA minus the average of homologous month PCA (non financial products)
20,00 €
15,00 €
10,00 €
5,00 €
0,00 €
-5,00 €
-10,00 €
17,33 €
13,27 €
10,93 €
-6,46 €
Financial Client
Value Loosing
Standard Client
Very Profitable

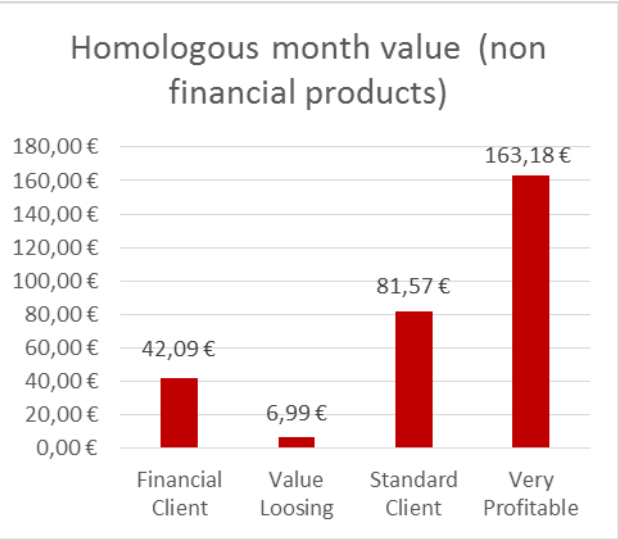

Claim year to date for non financial products
2.000,00 €
1.500,00 €
1.000,00 €
500,00 €
0,00 €
-500,00 €
1.059,00 €
1.734,37 €
100,71 €
-296,87 €
Financial Client
Value Loosing
Standard Client
Very Profitable


Sum of policies that were substituted for a policy with a smaller PCA
0,120
0,100
0,080
0,060
0,040
0,020
0,000
0,058
0,028
0,106
0,111
Financial Client
Value Loosing
Standard Client
Very Profitable


Sum of policies that were substituted for a policy with a higher PCA
0,160
0,140
0,120
0,100
0,080
0,060
0,040
0,020
0,000
0,135
0,028
0,047
0,101
Financial Client
Value Loosing
Standard Client
Very Profitable


Sum of current year technical margin minus the sum of homologous month technical margin (non financial products)
600,00 €
400,00 €
200,00 €
0,00 €
-200,00 €
-2,87 €
-146,26 €
64,90 €
459,86 €
Financial Client
Value Loosing
Standard Client
Very Profitable


Actual month value (non financial products)
700,00 €
600,00 €
500,00 €
400,00 €
300,00 €
200,00 €
100,00 €
0,00 €
-100,00 €
-200,00 €
39,22 €
-139,27 €
146,47 €
623,04 €
Financial Client
Value Loosing
Standard Client
Very Profitable


Homologous month value (non financial products)
180,00 €
160,00 €
140,00 €
120,00 €
100,00 €
80,00 €
60,00 €
40,00 €
20,00 €
0,00 €
42,09 €
6,99 €
81,57 €
163,18 €
Financial Client
Value Loosing
Standard Client
Very Profitable