Data Mining Report

Maren Leuthner, Catarina Palha and Mafalda Zúquete

January 2020

1 Introduction

Our client, the Antunes&Bação Ltd. company asked us to develop a customer segmentation in such a way that it will make it possible for their Marketing Department to better understand all the different customer's profiles.

They asked us to define, describe and explain the clusters we chose, reasoning how we want to do our clustering, possible approaches, advantages or disadvantages of different decisions and to express the marketing approach we recommend for each cluster.

First, we're going into a exploratory analysis in order to understand possible problems with in the given data, then we're going to solve those problems by cleaning the data.

Then we'll analyse variables in order to understand their importance as clusters features. Finally the cluster analysis itself.

2 Data Description

This database is from the year 2016 and it has 10296 rows and 14 columns, which are:

- Customer Identity:

  Uniquely identifies each customer.

- First Policy's Year:

  The first year as client.

- Brithday Year:

  The year when the customer was born.

- Gross Monthly Salary:

  Salary of the client.

- Geographic Living Area:

  No further information provided about the meaning of the area codes.

- Has Children (Y=1):

  1 means they have children, 0 means otherwise.

- Customer Monetary Value:

$$CustomerAnnualProfit * NumberYearsThatIsCustomer - AcquisitionCost \tag{1}$$

- Claims Rate

$$\frac{AmountPaidByTheInsurance}{Premiums} \tag{2}$$

  In the last 2 years.

- Premiums in LOB: Motor
- Premiums in LOB: Household
- Premiums in LOB: Health
- Premiums in LOB: Life
- Premiums in LOB: Work Compensations

  Annual Premiums, negative Premiums may manifest reversals occurred in the current year, paid in previous ones.

3 Exploratory Phase

In this section we'll look for inconsistent data, outliers and NaN values.

We're going to spend a lot of time in this section because good data is a prerequisite for a good customer segmentation.

3.1 First look

Here's an overview of the data:

	count	mean	std	min	median	max
Customer Identity	10296	5148.50	2972.34	1	5148.5	10296
First Policy's Year	10266	1991.06	511.26	1974	1986	53784
Gross Monthly Salary	10260	2506.66	1157.44	333	2501.5	55215
Geographic Living Area	10295	2.70	1.26	1	3	4
Has Children (Y=1)	10275	0.70	0.45	0	1	1
CMV	10296	177.89	1945.81	-165680.42	186.87	11875.89
Claims Rate	10296	0.74	2.91	0.00	0.72	256.20
Premiums in LOB: Motor	10262	300.47	211.91	-4.11	298.61	11604.42
Premiums in LOB: Household	10296	210.43	352.59	-75-00	132.80	25048.80
Premiums in LOB: Health	10253	171.58	296.40	-2.11	162.81	28272.00
Premiums in LOB: Life	10192	41.85	47.48	-7.00	25.56	398.30
Premiums in LOB: Work Compensations	10210	41.27	51.51	-12.00	25.67	1988.70

We can immediately spot inconsistent values in the Birthday Year columns and in the First Policy's Year column.

The minimum in the Birthday Year column is 1028 and the maximum is 2001 which means we have clients in a range of 15 to 988 years old.

Also, in the column First Policy's Year we can observe that we have the minimum value in the year 1974 and the maximum value in the year 53784 which means that we have at least one client that hasn't even been born yet.

This also suggests that we possibly have clients with Birthday Year greater than the First Policy's Year, which is a problem since we cannot have clients with a First Policy's Year in a year where they are not even born.

We've 1997 rows in these conditions. However, it's known that the Birthday Year column is hand written data, so it can have errors, therefore we'll drop this column.

Another test was done to confirm we don't have any duplicated rows and we also verified that we don't have duplicates on the column Customer Identity, it means each row identify a unique customer.

We know that columns like Geographical Living Area, Has Children (Y = 1), and Educational Degree are categorical variables that can only take a limited number of values. For that reason we confirmed that those variables were filled out correctly.

Finally we noticed that the lowest value for Gross Monthly Salary is 333€, which is lower than the minimum salary for 2016 (530€). Taking into account that that value is for full time workers, it would make sense that a part time employee would be paid less than the minimum monthly salary.

Now, let's get an idea of how many NaN values do we have and also which is the type of our data.

Column Name	**Number of NaN**	**Data Type**
Customer Identity	0	int
First Policy's Year	30	float
Educational Degree	17	object
Gross Monthly Salary	36	float
Geographic Living Area	1	float
Has Children ($Y = 1$)	21	float
Customer Monetary Value	0	float
Claims Rate	0	float
Premiums in LOB: Motor	34	float
Premiums in LOB: Household	0	float
Premiums in LOB: Health	43	float
Premiums in LOB: Life	104	float
Premiums in LOB: Work Compensations	86	float

Regarding the data types the data is consistent: the numeric objects are floats or int and the categorical data are objects.

As can be seen, there are a lot of NaN values, in fact we have a total of 295 rows with NaN values. This represents 2,86% of the data, so we can't just drop the rows because we're going to loose too many information, we'll replace the NaN values.

We should notice that this is an acceptable practice in this industry, but when dealing with, for example, medical data, fabricated values, even if well justified, may lead to conclusions with serious consequences.

3.2 Replacing Nulls

3.2.1 Premiums

We immediately replaced the NaN values in the premiums columns by 0. With this, we're assuming that if there is a NaN value then the customer doesn't have that kind of insurance.

The number of NaN values decreased a lot, but we still have 78 rows with NaN values.

3.2.2 Gross Monthly Salary

Let's start by plotting a histogram of this variable:

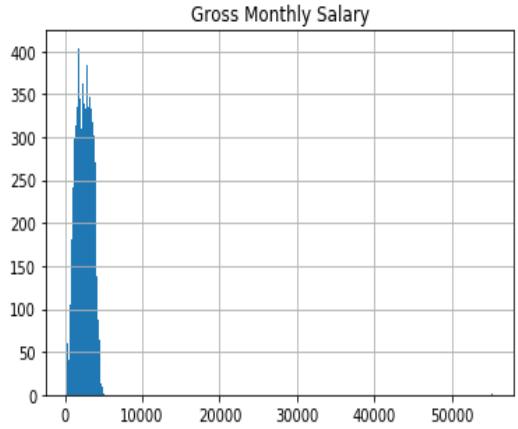


Figure 1: Gross Monthly Salary Histogram

The histogram of this variable appears to be symmetric, however, this figure also suggests there are outliers that are very removed from the rest of the data.

For this reason, we'll replace the NaN values using the median instead of the mean.

3.2.3 Educational Degree

We're going to replace these NaN values looking into the median of gross salary of each level of education, again to avoid the influence of outliers.

Educational Degree	Median
1 - Basic	1714.00
2 - High School	2501.75
3 - BSc/MSc	2609.00
4 - PhD	2644.00

For instances, if a customer with a NaN value on the Educational Degree variable has a Gross Monthly Salary of 2540.89€we will assume that he has a Educational Degree of 3 - BSc/MSc.

3.2.4 Geographic Living Area

Since we only have one NaN value for this variable we're going to replace it with the mode (4).

3.2.5 Has Children (Y=1)

We couldn't find any pattern for this variable with respect to others so we chose to replace the NaN values with the mode.

3.2.6 First Policy's Year

To replace these NaN values in this variables we need to add some variables to aid us:

1. Age As Client
$$2016 - FirstPolicy'sYear \tag{3}$$
2. Annual Salary
$$GrossMonthlySalary * 12 \tag{4}$$
3. Customer Annual Profit or Premiums Sum
$$Motor + Household + Health + Life + WorkCompensations \tag{5}$$

We know that:

$$\text{CMV} = \text{Customer Annual Profit} * \text{Number Years That Is Customer} - \text{Acquisition Cost}$$

$$\iff$$

$$\text{Number Years That Is Customer} = \frac{\text{CMV} + \text{Acquisition Cost}}{\text{Customer Annual Profit}} \tag{6}$$

$$\iff$$

$$\text{Acquisition Cost} = \text{Customer Annual Profit} * \text{Age As Client} - \text{CMV} \tag{7}$$

Now that we have the Acquisition Cost variable we can calculate the First Policy's Year.

Recall that,

$$\text{First Policy's Year} = 2016 - \text{Age As Client} \tag{8}$$

but,

$$\text{Age As Client} = \frac{\text{CMV} + \text{Acquisition Cost}}{\text{Customer Annual Profit}} \tag{9}$$

Therefore we get,

$$\text{First Policy's Year} = 2016 - \frac{\text{CMV} + \text{Acquisition Cost}}{\text{Customer Annual Profit}} \tag{10}$$

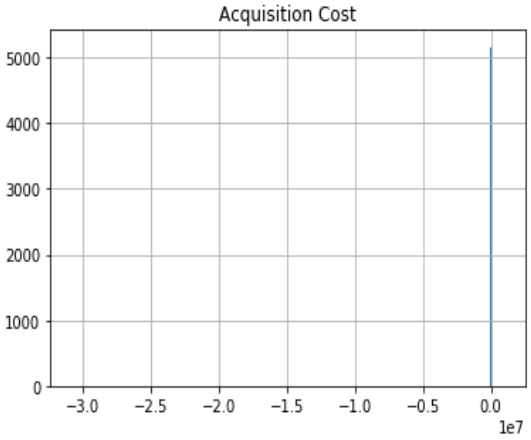


Figure 2: Acquisition Cost Histogram

Before replacing the NaN values of the First Policy's Year variable we need to replace the NaN values in the Acquisition Cost variable. Let's look at its distribution:

We can clearly see that we have outliers on this variable, and for this reason we're going to replace these NaN values with the median.

Finally are ready to replace the NaN values on First Policy's Year variable using equation (10).

3.3 Outliers

In this section we're going to identify the outliers using boxplots and them remove using the Z-score method.

In order to use the Z-score we've to choose a threshold that removes less than 3% of the data and is also good enough to remove only the real outliers instead of outliers and good customers.

We'll try to understand how much data we'll remove using a threshold of 3.0 and 4.0.

But let's first check on which variables we have outliers using boxplots.

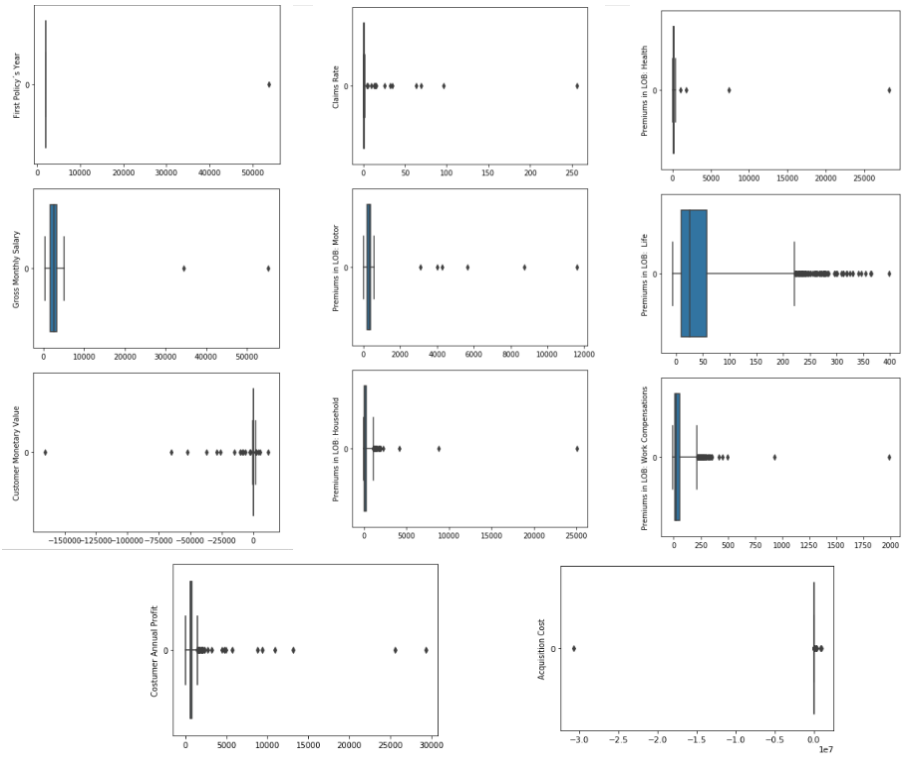


Figure 3: Outliers Boxes Plots

With threshold equal to 3.0 we have:

Variable	**Outliers Number**
First Policy's Year	1
Gross Monthly Salary	2
Customer Monetary Value	13
Claims Rate	12
Premiums in LOB: Motor	6
Premiums in LOB: Household	36
Premiums in LOB: Health	3
Premiums in LOB: Life	210
Premiums in LOB: Work Compensations	162
Customer Annual Profit	13
Acquisition Cost	2

Which means we've 411 outliers in this case, let's now consider a threshold of 4.0:

Variable	Outliers Number
First Policy's Year	1
Gross Monthly Salary	2
Customer Monetary Value	12
Claims Rate	11
Premiums in LOB: Motor	6
Premiums in LOB: Household	12
Premiums in LOB: Health	3
Premiums in LOB: Life	79
Premiums in LOB: Work Compensations	62
Customer Annual Profit	12
Acquisition Cost	1

The number of outliers has decreased to 173.

In order to decide which threshold we're going to use we need to know how much of the data these numbers represent.

Threshold	Outliers Number	Data Percentage
3.0	411	3.9%
4.0	173	1.6%

We don't want to lose too much information. In other words, we don't want to lose more than 3% of the data, therefore we're going to choose a threshold of 4.0.

After removing the outliers we now have 10123 rows and 18 columns.

We still have to drop one row. This row is an example of inconsistent data: here the Premiums Sum variable is greater than the Annual Salary variable, which is not possible given we're talking about insurance.

After cleaning the data we have now 10122 rows and 19 columns. It means we dropped $1,6\%$ of the data, which is an acceptable number.

4 Feature Selection

4.1 Data Set Division

Product Variables:

- Premiums in LOB: Motor
- Premiums in LOB: Household
- Premiums in LOB: Health
- Premiums in LOB: Life
- Premiums in LOB: Work Compensations

Customer Variables:

- First Policy's Year
- Educational Degree
- Gross Monthly Salary
- Geographic Living Area
- Has Children (Y=1)
- Customer Monetary Value
- Claims Rate
- Age As Client
- Annual Salary
- Customer Annual Profit
- Acquisition Cost

Question:
Do we need all these variables to get good clusters?

In order to decide which variables we're going to use to cluster the data we'll use two approaches:

1. Correlation between variables
2. Principal Component Analysis

4.2 Product Variables

As we can see in Figure 4, there isn't any known distribution between the variables and there doesn't seem to be any correlation between them.

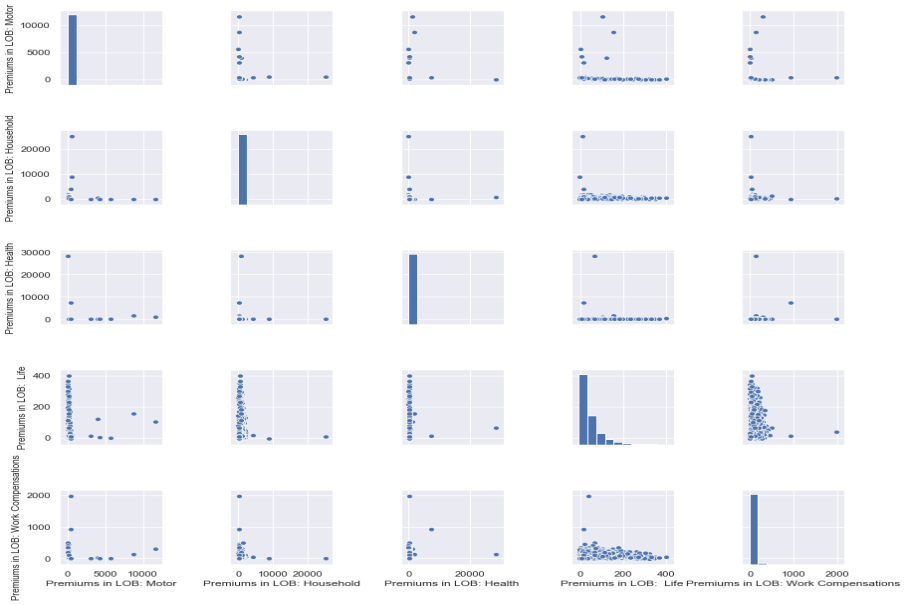


Figure 4: Product Variables Pairplot

4.2.1 Variables Correlation

In fact there isn't a strong correlation between the product variables.

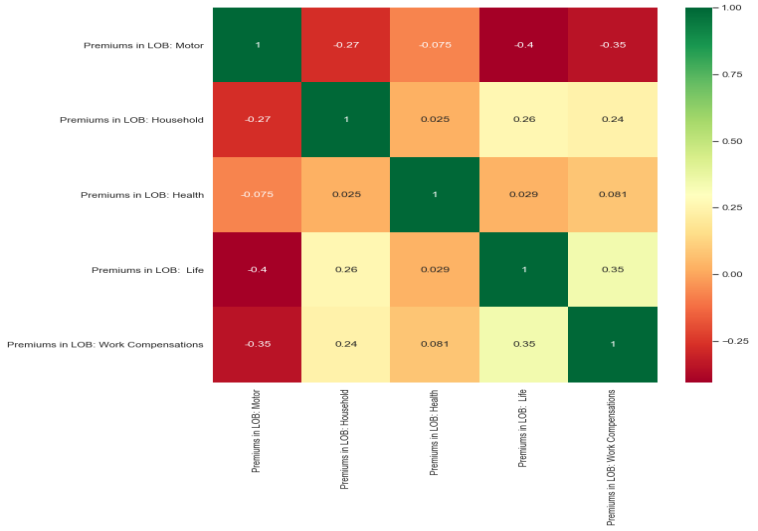


Figure 5: Correlation between Product Variables

This means we probably can't remove any variables, and therefore we can't decrease the number of dimensions. To be sure let's do a principal component analysis.

4.2.2 Principal Component Analysis

For starters we're going to run the PCA with 5 components, in order to plot the elbow graph. Our goal is to know how many components we should use on PCA.

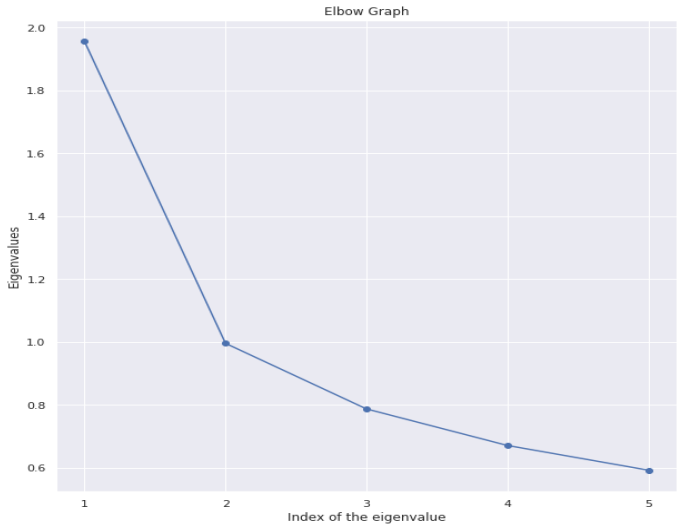


Figure 6: Elbow Graph on the number of eigen values

According this Figure 6 we should use 2 components. But, how many variance do two components explain?

PC1	PC2	PC3	PC4	PC4
0.39108886	0.59015271	0.74750559	0.88160965	1

Using the explained variance ratio we see that it only explains around 60% of the variance, which is usually not good enough.

We're going to run the PCA again, but this time with 3 components, and according with the explained variance ratio we'll preserve around 75% of the variance.

Calculating the loadings we understand how much each variable is correlated with each component.

	Motor	Household	Health	Life	Work
PC1	-0.74417772	0.59227235	0.15499008	0.73497942	0.69773477
PC2	0.01087472	-0.14287702	0.97933784	-0.11889606	0.04057895
PC3	0.17450284	0.79108204	0.09423679	-0.24366487	-0.2496536

We see that,

1. **PC1:** Negatively correlated with Premiums in LOB: Motor and positively correlated with Premiums in LOB: Life and with Premiums in LOB: Work Compensations.
2. **PC2:** Positively correlated with Premiums in LOB: Health.
3. **PC3:** Positively correlated with Premiums in LOB: Household.

Although the principal components are well defined, we can't interpret PC1 in the context of the problem in order to generate a new variable.

Therefore, we won't remove any variable, which means that we're not going to decrease the dimension on this set of variables.

4.3 Customer Variables

As we can see in Figure 7, there isn't any known distribution between the variables. But there it seems to exist a correlation between some variables.

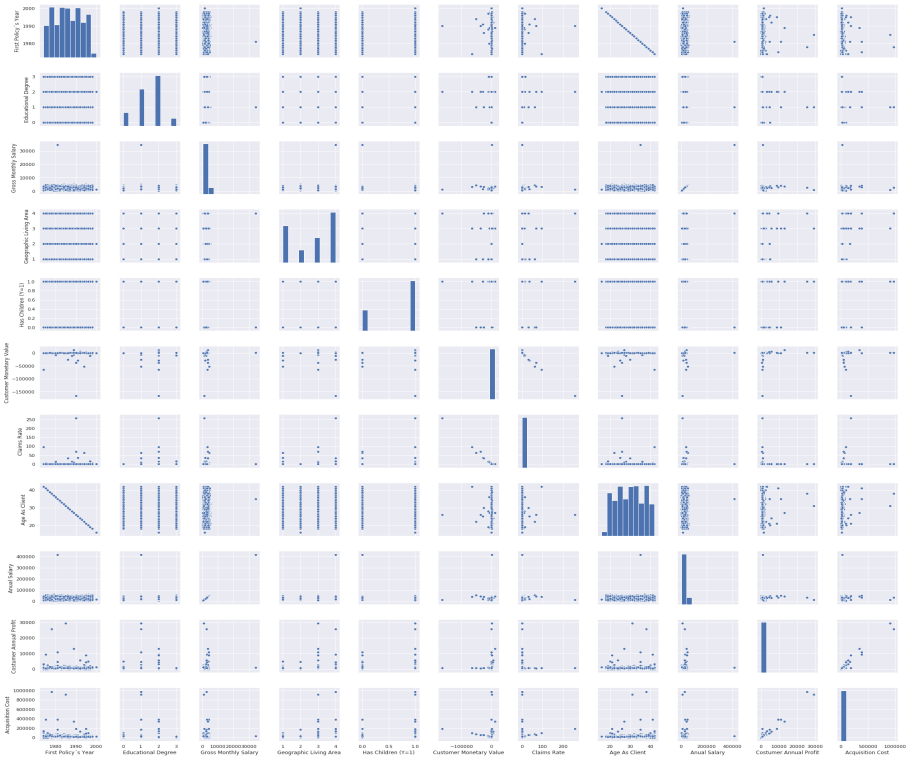


Figure 7: Customer Variables Pairplot

4.3.1 Variables Correlation

In fact there are some variables that are correlated.

Figure 8: Correlation between customer variables

First Policy's Year and Age As Client have a correlation of -1, which is natural because we used the First Policy's Year to calculate the Age As Client.

Gross Monthly Salary and Annual Salary have a correlation of 1, which is natural because we used the Gross Monthly Salary to calculate the Annual Salary.

Geographic Living Area and Has Children ($Y = 1$) aren't strongly correlated with any variable.

Customer Monetary Value and Claims Rate have a correlation of -0.99.

And finally, Customer Annual Profit and Acquisition Cost have a correlation of 0.93.

It seems that we can remove variables, but let's also try a principal component analysis.

4.3.2 Principal Component Analysis

We're going to run the PCA with 11 components, in order to plot the elbow graph.

Our goal is again to know how many components we should use on PCA.

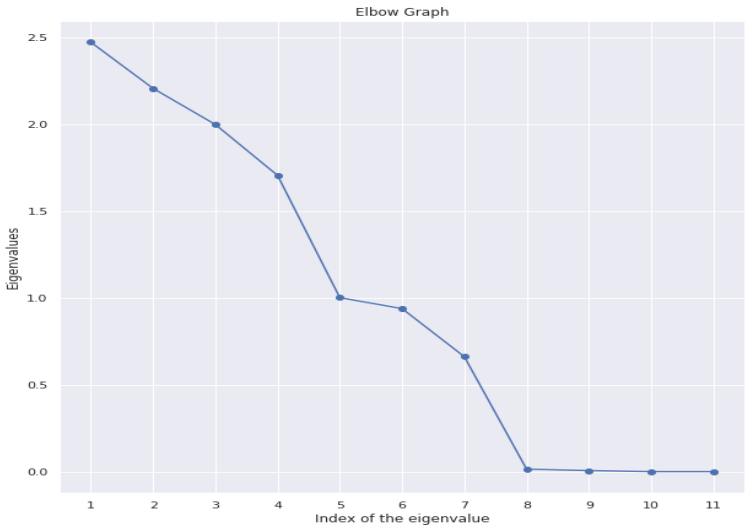


Figure 9: Elbow Graph on the number of eigen values

We only have two elbow points in this graph, but after the second one the eigenvalues of the correlation matrix are already too close to 0 for us to consider using the elbow graph to determine the number of principal components. So we checked the explained variance ratio for 5 principal components:

PC1	**PC2**	**PC3**	**PC4**	**PC5**
0.22470131	0.42522628	0.60686691	0.76184719	0.85284571

Since 5 principal components explain around 85% of the variance, that's the number of principal components we're using.

We'll run the PCA again but this time with 5 components.

	PC1	PC2	PC3	PC4	PC5
First Policy's Year	-0.477816	-0.635335	0.237698	0.555153	0.011438
Educational Degree	-0.285223	0.075554	0.008880	-0.264159	0.159527
Gross Monthly Salary	-0.735000	0.615264	-0.029811	0.114077	0.018619
Geographic Living Area	-0.000308	0.037278	0.007976	0.031846	-0.982825
Has Children (Y=1)	0.422387	-0.461872	0.018761	-0.229326	0.057239
Customer Monetary Value	-0.078590	-0.171181	-0.969897	0.142429	-0.000733
Claims Rate	0.099576	0.179426	0.970140	-0.113093	0.000230
Age As Client	0.477816	0.635335	-0.237698	-0.555153	-0.011438
Anual Salary	-0.734998	0.615264	-0.029811	0.114077	0.018619
Costumer Annual Profit	0.507760	0.310152	-0.021457	0.778597	0.054575
Acquisition Cost	0.63342197	0.51330129	0.025276	0.547595	0.049777

And we define the following principal components:

1. **PC1:** negatively correlated with Gross Monthly Salary and Annual Salary, positively correlated with Acquisition Cost;
2. **PC2:** negatively correlated with First Policy's Year and positively correlated with Age As Client;
3. **PC3:** negatively correlated with Customer Monetary Value and positively correlated with Claims Rate;
4. **PC4:** positively correlated with Customer Annual Profit;
5. **PC5:** negatively correlated with Geographic Living Area;

The principal components are mostly explained by two variables that we had already established as very strongly correlated or by only one variable.

This tells us that there's no difference in reducing dimensions by using the PCA or by dropping correlated variables.

Moreover, we should note that the variable Educational Degree barely contributes to the variance of the first 3 principal components. However, we are choosing not to drop this variable as it can still help with the segmentation.

As we concluded from the correlation matrix, the variables Claims Rate and Customer Monetary Value have such high correlation that it doesn't make sense to use both in the same cluster analysis. However, we don't know yet which one will provide a better clustering, so we'll do the analysis using one and then the other and choose the one that performs best.

We're going to drop the First Policy's Year and keep the Age as Client because it's easier to interpret.

Also we'll keep the Gross Monthly Salary instead of the Annual Salary.

Finally, we're going to keep the Customer Annual Profit instead of the Acquisition Cost because it's an important variable on the customer set and tells us more about the customers than Acquisition Cost.

Notice that the variables we're choosing to drop not only have high correlations but are actually obtained from one another.

4.4 Negative Variables

We have negative values on Customer Monetary Value and on the Premiums.

4.4.1 Negative Customer Monetary Value

We have 2723 customers with negative CMV, this represents 26% of our data.

This variable highlights the lifetime value of this client so we considered that it's still valuable for customer segmentation to keep the negative values as they separate the clients that bring profit to the company from the ones who don't.

4.4.2 Negative Premiums

We have 2194 customers with some kind of negative premium, where:

Premium Type	**Number of Customers**
Premiums in LOB: Motor	1
Premiums in LOB: Household	1075
Premiums in LOB: Health	1
Premiums in LOB: Life	660
Premiums in LOB: Work Compensations	910

We know that a negative premiums signal reversals, that is the client doesn't have that kind of insurance anymore and so is not paying for it. This was considered important for the analysis and so we created new variables referring to each of the premiums but where we replaced the negative values by 0, signaling what the client actually paid for the premium.

4.4.3 Premium Ratios

We'll also calculate how much weight each premium has in the total that the customer pays.

5 Segmentation on Product Variables

Clustering uses the data set and divides it in different groups, so called clusters. The process of grouping objects into the same cluster depends on their similarity or dissimilarity. Through the clustering process, the user does not have knowledge of the turn out of the resulted clusters, since the data does not have any class label. In the development of this report we're going to consider different methods of clustering, such as Partitioning methods, Hierarchical methods, Density-based methods and Model-based methods.

5.1 K-Means

The K-Means algorithm belongs to the Partitioning methods of clustering and is a centroid-based technique. The algorithm aims a partition of n observations into k groups of clusters, e.g, grouping similar objects. Each cluster contains a cluster center, which is called centroid. The number of k clusters is specified by the user with the condition that k is smaller or equal to n observations. The functional objective is to assess the partitioning quality and form clusters with objects that are similar to each other but also dissimilar to other clusters objects. Subsequently, a high intra cluster similarity and a low inter cluster similarity is aimed for in the process. The following requirements have to be met during classifying the data into k groups:

- Each group must contain at least one point
- Each point must belong to exactly one cluster

The algorithm works as follows:

1. Choose random seeds
2. Each individual is associated with the nearest seed
3. Calculate the centroids of the formed clusters
4. Go back to step 2
5. End when the centroids cease to be recentered

K-Means can be slow to converge in most cases. Subsequently, it takes a long time to converge exponentially for accurate conditions. In order to produce results without compromising the accuracy, a reasonable threshold value needs to be specified.

In order to do K-Means we are going to consider three possible sets of variable to cluster:

- The original Premium variables
- The ratio Premium variables

5.1.1 Original Premium variables

Elbow Graph

The elbow method can be used for clustering in order to decide on the number of clusters that are going to be used for the clustering algorithm.

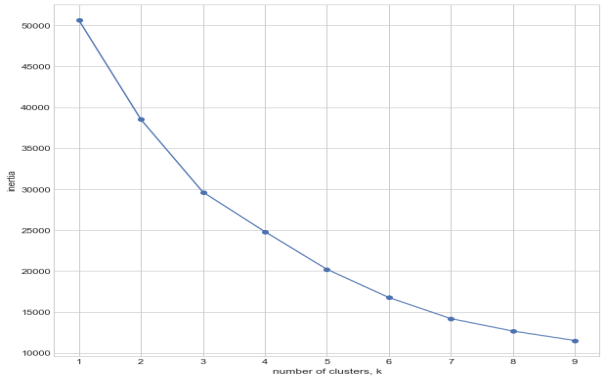


Figure 10: Elbow Graph for KMeans

Looking at the elbow graph, we can see that we should use 3 clusters, but first confirmed with the silhouette score.

Silhouette Score

The Silhouette Score is from -1 to 1 and show how close or far away the clusters are from each other and how dense the clusters are. The closer your silhouette score is to 1 the more distinct your clusters are.

Number of Cluster	Average Silhouette Score
2	0.46747179013752865
3	0.4672289552700388
4	0.4655022191829592
5	0.4691630108984542
6	0.3341479589429703
7	0.34133165411352845
8	0.33937812458678196
9	0.28870571093188885

As we can see from the table, for k-means if the number of clusters is between 2 and 5 the silhouette score doesn't suffer any significant changes. Since the elbow graph indicates that we should use 3 clusters for this algorithm, that's what we'll do.

K-Means Clusters

From executing K-Means with 3 clusters we got the following results:

Motor	Household	Health	Life	Work Compensations
360.568747	122.616472	161.879872	21.518776	22.445001
137.309130	444.887902	184.619036	94.205493	90.273434
26.340000	829.050000	28272.000000	65.680000	138.250000

Where each cluster has the following number of elements:

Number of Cluster	Number of Elements
0	7351
1	2770
2	1

5.1.2 Ratio Premium variables

Elbow Graph

The elbow method can be used for clustering in order to decide on the number of clusters that are going to be used for the clustering algorithm.

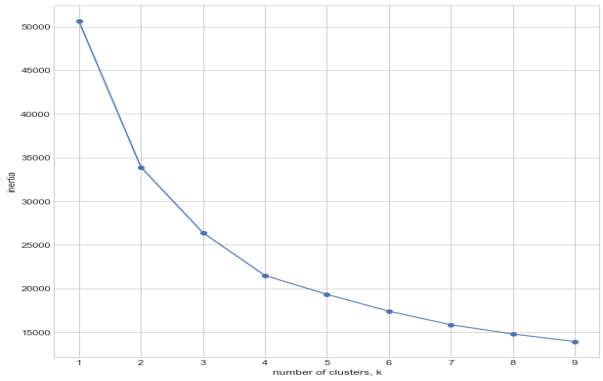


Figure 11: Elbow Graph for K-Means

Looking at the elbow graph, we should use 3 clusters, but we're double checking with the silhouette score.

Silhouette Score

Number of Cluster	Average Silhouette Score
2	0.3202340590500842
3	0.30283113641180825
4	0.322322231006759
5	0.26300346702757016
6	0.26898758032045106
7	0.2492621369123551
8	0.2508187568837002
9	0.24886836552147149

As we can see from the table, for K-Means no matter the number of clusters the silhouette score doesn't suffer any significant changes. Since the elbow graph indicates that we should use 3 clusters for this algorithm, we'll do just that.

K-Means Clusters

From executing K-Means with 3 clusters we got the following results:

Motor	Household	Health	Life	Work Compensations
0.176546	0.461732	0.176498	0.093468	0.091755
0.365947	0.187982	0.339066	0.053893	0.053112
0.691824	0.101984	0.162354	0.020064	0.020559

Where each cluster has the following number of elements:

Number of Cluster	Number of Elements
0	2815
1	3575
2	3732

Looking at the clusters obtained from using the original and the ratio variables it's clear that the clusters we get from the ratio variables are much preferable since they're more homogeneous and less spread out. Notice that the silhouette score between them is very similar, so that's not a differentiating criteria.

We also noticed that in all the clusters the ratios for the premiums for life and work compensation were very low, especially in comparison to the ratios of the other premiums. This suggests that these variables are not very relevant to the clustering so we decided to run the K-Means again.

5.1.3 Reduced Dimension Ratio Premium Variables

Elbow Graph

The elbow method can be used for clustering in order to decide on the number of clusters that are going to be used for the clustering algorithm.

Looking at the elbow graph, we decided that we should use 3 clusters, but we're double checking with the silhouette score.

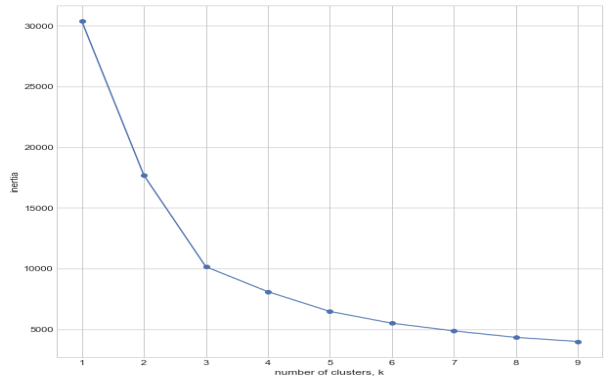


Figure 12: Elbow Graph for KMeans

Silhouette Score

Number of Cluster	Average Silhouette Score
2	0.3818788853319167
3	0.4236460563277549
4	0.3794840527707247
5	0.3514924675495437
6	0.3636880259863281
7	0.3456145520085629
8	0.3211218825579122
9	0.32284283225779153

As you can see from the table, for K-Means no matter the number of clusters the silhouette score doesn't suffer any significant changes. Since the elbow graph indicates that we should use 3 clusters for this algorithm, we'll do just that.

K-Means Clusters

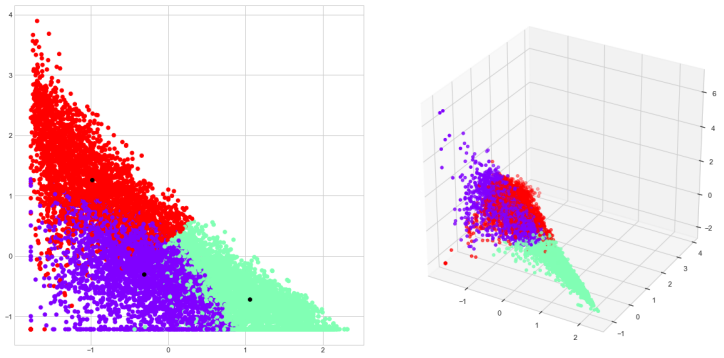
From executing K-Means with 3 clusters we got the following results:

Motor	Household	Health
0.357279	0.174836	0.347183
0.690725	0.095893	0.163224
0.192417	0.474483	0.174488

Where each cluster has the following number of elements:

Number of Cluster	Number of Elements
0	3432
1	3755
2	2935

In comparison to the previous clusters, there are equally as well distributed, but have a better silhouette score and are easier to interpret, so we conclude that these is the best clustering we can get with K-Means.



(a) Clusters 2D (b) Clusters 3D

Figure 13: Visual representation of the clusters obtained from the ratio variables, after reducing the dimensions

5.2 Hierarchical Clustering

5.2.1 Original Premium Variables

By executing a hierarchical clustering for the original premium variables, we got the following three clusters:

Motor	Household	Health	Life	Work Compensations
193.526953	331.053230	199.186178	65.827374	64.388679
422.335172	71.352176	132.061711	13.090484	13.888353
26.340000	829.050000	28272.000000	65.680000	138.250000

The table below shows the number of elements each cluster contains. It is shown that the first two clusters contain the most elements and the third cluster only contains one element.

Number of Cluster	Number of Elements
0	5434
1	4687
2	1

The silhouette score is: 0.3134351388487947.

5.2.2 Ratio Premium Variables

By executing a hierarchical clustering for the ratio premium variables, we got the following 3 clusters.

Motor	Household	Health	Life	Work Compensations
0.157753	0.441645	0.185716	0.104364	0.110522
0.684861	0.093064	0.178345	0.021157	0.022572
0.347122	0.247829	0.302359	0.053019	0.046651

The table below shows the number of elements each cluster contains. Here, the elements are well distributed among the number of clusters in comparison with the elements of the original premium variables in the section before.

Number of Cluster	Number of Elements
0	2283
1	3866
2	3973

The calculated silhouette score is: 0.2462320217433373.

5.2.3 Reduced Dimensions Ratio Premium Variables

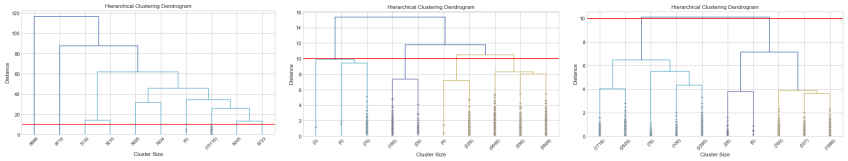
Now, we want to analyze the reduced dimensions ratio premium variables. After executing the hierarchical algorithm, we got the following clusters:

Motor	Household	Health
0.640039	0.101331	0.198848
0.213332	0.439955	0.190023
0.275269	0.184749	0.405089

The three clusters contain the following number of elements:

Number of Cluster	Number of Elements
0	4986
1	3521
2	1615

For the reduced dimensions ratio premium variables, the calculated silhouette score is: 0.3939451280235937.



(a) Dendogram of the original variables (b) Dendogram of the original variables (c) Dendogram of the reduced variables

Figure 14: Various dendograms obtained through hierarchical clustering

The results are analogous to the K-Means results, but the silhouette scores are preferable in K-Means. This is an expected result since we apply k-means on top of the hierarchical clustering.

5.3 DBSCAN

The DBSCAN algorithm is a density-based spatial clustering method, which is able to find arbitrary shaped clusters by identifying clusters as dense regions and separating them by low dense regions.[5] in order to identify clusters of large spatial data sets by watching the local density of blocks of data using a single input parameter. Determined information classified as noise and outliers can also be identified by the DBSCAN algorithm. Although, DBSCAN is able to identify clusters of arbitrary shape, groups that are located close to each other are likely to belong to the same class.[2] The process of DBScan is as follows:

1. It starts with an arbitrary data point in the data set, which has not been visited. The neighborhood objects of this data point is checked within a given radius, called Eps.

2. If a sufficient number of points in the neighborhood is found, the clustering process begins by setting the current data point as the first point in the new cluster. Otherwise, the point will be marked as noise. In both alternatives, the point is set as "visited".

3. All the points within its Eps distance will become, with the first point, part of the new cluster. This process is repeated for all new points that were added to the group of cluster.

4. Repeat process steps 2 and 3 until all points in the clusters are determined, meaning that all points withing the Eps neighborhood of the clusters ave been marked visited.[2]

DBSCAN convinces based on three advantages:

1. it does not require a pre-set number of clusters

2. it identifies outliers as noise

3. it can also find arbitrarily sized and shaped clusters

The advantages have to be seen in comparison with its disadvantages:

1. it doesn't perform well when clusters contain varying density

2. the setting of the distance threshold Eps and minPoints for identification of the neighborhood points will differ from cluster to cluster due to varying density[2]

5.3.1 Original Premium Variables

By executing DBSCAN with the original premium variables we get:

Number of Cluster	**Number of Elements**
-1	0
2	10121

Although it give us one cluster with a silhouette score of 0.9783202942719289 and zero noise (represented by -1) this isn't good for our propose.

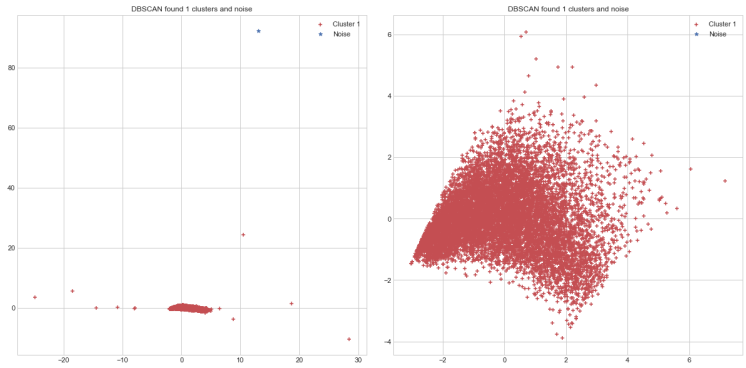
We found this very natural since DBSCAN isn't meant to be used with continuous values.

5.3.2 Ratio Premium Variables

By executing DBSCAN with the original premium variables we get:

Number of Cluster	Number of Elements
0	10122

There is no silhouette score since this clustering only defined one cluster and we need at least two clusters to calculate the silhouette score.



(a) DBSCAN for original variables (b) DBSCAN for ratio variables

Figure 15: Implementation of DBSCAN for product variables

Despite the silhouette score, these are not good clusters as we either only get one cluster or one cluster + noise, neither good segmentation results. However, this was an expected result since DBSCAN is not usually a good algorithm for continuous variables.

6 Self-Organizing Maps

Self-Organizing Maps (SOM) are unsupervised neural networks and, therefore, closely related to clustering.[1] SOM can reduce data dimensions and are able to display similarities among the data. It can, therefore, be used for cluster detection.[6] The algorithm of SOM learning process works as follows:

1. Each node's weight vectors is initialized
2. A sample vector is randomly selected from the training data

3. Each weight vectors' neighbors weight is calculated. The weight, which is most like the input vector is the winning node.

4. All neighbors of the winning node will also become more like the chosen vector.

5. From now on, the number of neighbors and how much each neighbor weight can learn will decrease with time.[7]

This process is repeated usually iterated more than a 1000 times for a successful learning process.

Different colors represent the values of the elements and their distances between the respectively neighboring neurons.

In the figure below, the grey scale is used on the SOM. A dark color between the neurons corresponds to a relatively low distance to the other neurons, which means that the vectors are relatively close to each other in the input space. A light color represents a big distance between the neurons.

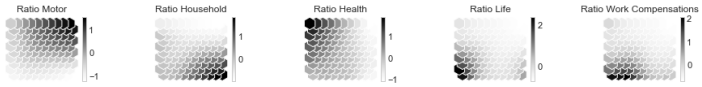


Figure 16: Implementation of SOM

In conclusion, dark areas represent a low distance in the self-organizing maps and can, therefore, be seen as clusters, while light areas can be thought as separators of clusters once they represent a bigger distance.

The SOM algortihm also created these clusters:

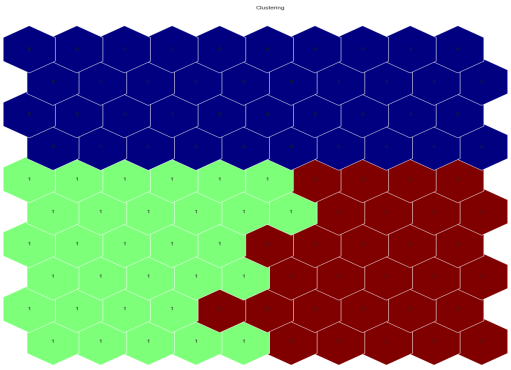


Figure 17: Clusters using SOM

But didn't give us any other information regarding centroids or any way to

calculate the silhouette score so we were not able to interpret this result and discarded this algorithm.

7 Mean-Shift Clustering

The Mean-Shift algorithm is a kernel-density estimation for a clustering based segmentation. The algorithm is based on sliding-window as it tries to identify dense areas of data points.[3] It is therefore seen as a centroid-based algorithm aiming to locate the center points of each class.[4] The process works as follows:

1. Begins a circular sliding window movement centered at a randomly selected point and having the radius selected as kernel.
2. The sliding window is shifted toward high density regions at every iteration by shifting the center point toward the mean of the points within the window.
3. Continue shifting the sliding window based on the mean until their is no further direction to accommodate more points inside the kernel.
4. Repeat process steps 1 to 3 with as many slides as points lie within a given window. The data points are clustered based on the sliding window they are located in.[3]

Original Premium Variables

From Mean-Shift we get 20 clusters:

Number of Cluster	Number of Elements
0	8537
1	812
2	179
3	199
4	210
5	17
6	155
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1

With a silhouette score of 0.43685253475458996.

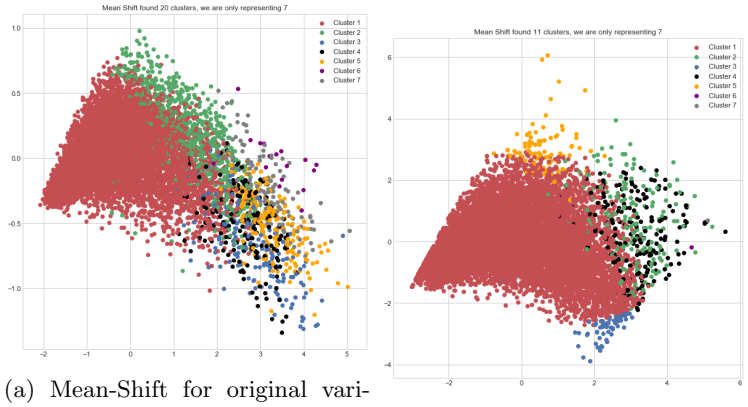
Ratio Premium Variables

From Mean-Shift we get 11 clusters:

Number of Cluster	Number of Elements
0	9440
1	210
2	74
3	206
4	89
5	2
6	1
7	1
8	15
9	83
10	1

With a silhouette score of 0.21969305584594814.

This clustering algorithm also wasn't successful, since we ended up with many clusters with only one element.



(a) Mean-Shift for original variables

(b) Mean-Shift for ratio variables

Figure 18: Implementation of Mean-Shift for product variables

Given all these possible clusters for the premium variables, we selected the K-Means clustering using the premium ratios variables as the being the best clustering.

Let us remember the result:

Motor	**Household**	**Health**
0.357279	0.174836	0.347183
0.690725	0.095893	0.163224
0.192417	0.474483	0.174488

This way we can define,
Cluster 1: Costumers that spend the most in Motor Premiums.
Cluster 2: Costumers that spends the most on Household Premiums.
Cluster 3: Costumers that spends the most on Health Premiums.

8 Segmentation on Costumers Variables

We're going to perform clustering on the costumers variables.

However, we've to divide this set in two because we've categorical variables and continuous variables that must be treated separately and with different cluster methods.

We'll divide the set as follows:

- Educational Degree
- Geographic Living Area
- Has Children ($Y = 1$)

where these are the categorical ones where we'll apply K Modes.

- Gross Monthly Salary
- Customer Monetary Value
- Claims Rate
- Age As Client
- Annual Salary
- Costumer Annual Profit
- Acquisition Cost

And there are the continuous ones where we'll apply K Means.
Notice that we're don't need to use all of them to perform a cluster analysis. We're going to start by the categorical ones.

8.1 K Modes

As we know, we've some variables that are correlated, however we decided not to drop them because we want to understand their performance on the segmentation first.

As consequence, we'll try different combinations of variables in order to decide which ones perform the best clusters using the K Modes algorithm and also to understand how many cluster we should use.

Here are the different combinations:

1. Educational Degree, Geographic Living Area, Has Children ($Y = 1$)
2. Educational Degree, Geographic Living Area
3. Educational Degree, Has Children ($Y = 1$)
4. Geographic Living Area, Has Children ($Y = 1$)

We'll initialize the K Modes with the random mode.

Combination	**Number Of Clusters**	**Average Silhouette Score**
1	2	0.2753701134169115
1	3	0.2753701134169115
1	4	0.2753701134169115
2	2	0.4910480886997046
2	3	0.4910480886997046
2	4	0.4910480886997046
3	2	0.6790202049295938
3	3	0.6790202049295938
3	4	0.6790202049295938
4	2	0.5669514327892092
4	3	0.5669514327892092
4	4	0.5669514327892092

First, the silhouette score barely change when we increase the number of clusters. Also, the combination number 3 has the best score and also represents two variable that we can easily interpret.

Therefore, we're going to chose the combination number 3 to use as variables on the K Modes method.

The best run is the number 5 and the clusters are:

Cluster Number	**Educational Degree**	**Has children ($Y = 1$)**
0	3.0	0.0
1	2.0	1.0
2	3.0	1.0

And in each cluster we've:

Cluster Number	**Number of Costumers**
0	2963
1	3808
2	3351

Before we start the description notice that we're going to encode the Educational Degree variable:

Educational Degree	**Encoding**
1 - Basic	1
2 - High School	2
3 - BSc/MSc	3
4 - PhD	4

We took a further look into the clusters and noticed that there is no cluster with people with and without children. Also, in clusters 1 and 2 the split between educational degrees is also very clear, with cluster 1 only having people with a BSc/MSc and cluster 22 being mostly comprised of people with education only up to High School.

That being said, let's describe the clusters:

Cluster 0: Costumers without children.

Cluster 1: Costumers without superior education and with children.

Cluster 2: Costumers with superior education and with children.

8.2 DBSCAN

Once these variables are categorical we can also try to apply the DBCSAN method. Unfortunately the results aren't good because the DBSCAN only found one cluster which doesn't allow us to segment the costumers.

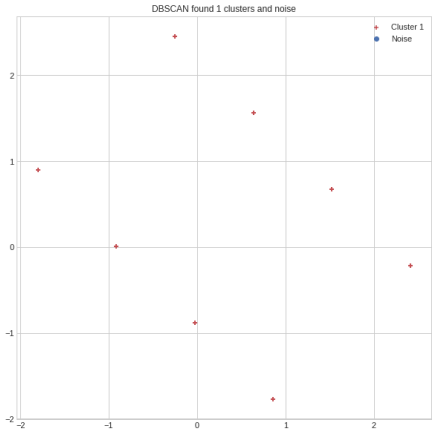


Figure 19: Implementation of DBSCAN for the costumers variable (Categorical)

8.3 K Means

In order to perform a segmentation on the continuous variable we're going to do a similar analysis.

The Gross Monthly Salary variable and the Age As Client variable perform a good cluster together, but we decided to explore multiple options to find out if adding a variable would give better clusters.

This time we used the following combinations:

1. Gross Monthly Salary, Age As Client, Customer Monetary Value
2. Gross Monthly Salary, Age As Client, Claims Rate
3. Gross Monthly Salary, Age As Client, Costumer Annual Profit
4. Gross Monthly Salary, Age As Client, Customer Monetary Value, Costumer Annual Profit
5. Gross Monthly Salary, Age As Client, Claims Rate, Costumer Annual Profit

We also need to decide how many cluster we'll use, so once again we'll calculate the silhouette score:

Combination	**Number Of Clusters**	**Average Silhouette Score**	Elbow Point
1	2	0.3538958844499264	3
1	3	0.3554027463182902	3
1	4	0.36237695950110665	3
2	2	0.35550891199808143	4
2	3	0.35704523636846247	4
2	4	0.3645652113163035	4
3	2	0.3303841167319838	4
3	3	0.3319618616178022	4
3	4	0.32632510503306167	4
4	2	0.34676986643280255	4
4	3	0.34854180686442304	4
4	4	0.3381364170836227	4
5	2	0.34676986643280255	4
5	3	0.34854180686442304	4
5	4	0.3381364170836227	4

The silhouette score barely change when we mix the variables which means that probably all of them have the same importance.

We decided to run all these combinations, usually using 3 clusters as was suggested by the elbow graph, but found out that adding a variable to the analysis always meant that we had a cluster with a singular element if we used 3 clusters for the analysis or, if we used 2 clusters instead, the only differentiating factor would be Age As Client and all the other variables would be irrelevant. So we decided to use the original variables we considered for the clustering.

Here's the elbow graph:

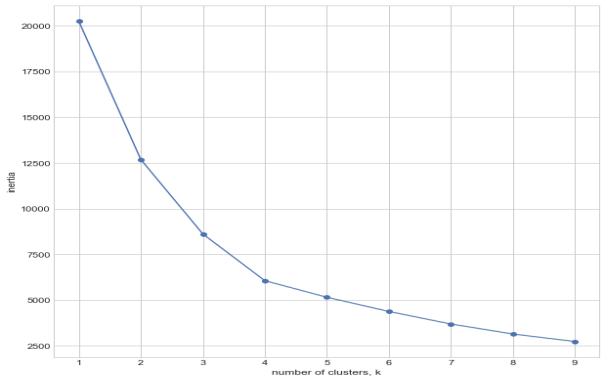
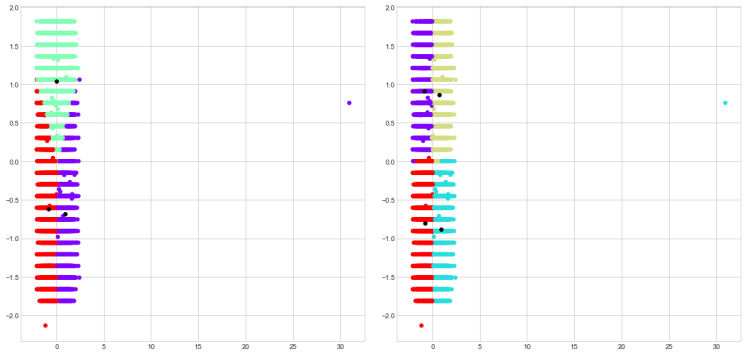


Figure 20: Elbow Graph K-Means Costumer Variables

The elbow graph suggests to use 3 of 4 clusters so we tried both approaches. Here are the results:



(a) K-Means with 3 clusters (b) K-Means with 4 clusters

Figure 21: Implementation of K-Means for continuous customer variables

Remember the silhouette scores:

Number Of Clusters	Average Silhouette Score
3	0.37259158485048743
4	0.38923264519225637

Since the silhouette score is slightly better for 4 clusters and the divisions appear to be clearer in the graphic we decided to use 4 clusters.

Now, looking into numbers:

Gross Monthly Salary	Age as Client
1619.887636	36.014700
3421.499588	24.159207
3258.863810	35.675662
1709.745297	24.664762

And in each cluster we've:

Cluster Number	Number of Costumers
0	2306
1	2426
2	2625
3	2765

We're ready to describe the clusters:
Cluster 0: Lower class people who have been clients for long.
Cluster 1: Higher class people who haven't been clients for very long.
Cluster 2: Higher class people who have been clients for long.
Cluster 3: Lower class people who haven' been clients for very long.

8.4 Merging Clusters

We're now merging the clusters so we can obtain a single segmentation for the customer variables.

	Children Low Education	**No Children**	**Children High Education**	**Total**
Low Salary Longtime Client	1205 (1)	186 (2)	915 (3)	2306
High Salary Recent Client	984 (4)	830 (5)	612 (6)	2426
High Salary Longtime Client	1140 (7)	737 (8)	748 (9)	2625
Low Salary Recent Client	1495 (10)	194 (11)	1076 (12)	2765
Total	4824	1947	3351	

Now we try to reduce the number of clusters by merging the smaller clusters with the closest, and therefore more similar, clusters:

1. cluster 2 with 12;
2. cluster 11 with 6;
3. cluster 8 with 12;
4. cluster 9 with 5;
5. cluster 6 with 7;
6. cluster 3 with 7;
7. cluster 4 with 1;

We then stopped since all the clusters had a similar size:
We now have these clusters:

	Children Low Education	**No Children**	**Children High Education**	**Total**
Low Salary Longtime Client	2189 (1)			2189
High Salary Recent Client		1614 (5)		1614
High Salary Longtime Client	2861 (7)			2861
Low Salary Recent Client	1495 (10)		1999 (12)	3494
Total	6515	1615	1999	

9 Final Segmentation

	Health	**Motor**	**Household**	**Total**
Low Salary Longtime Client Children Low Education	702 (1)	578 (2)	866 (3)	2146
High Salary Recent Client No Children	572 (4)	696 (5)	290 (6)	1558
High Salary Longtime Client Children Low Education	847 (7)	1341 (8)	634 (9)	2822
Low Salary Recent Client Children Low Education	446 (10)	377 (11)	637 (12)	1460
Low Salary Recent Client Children High Education	802 (13)	705 (14)	458 (15)	1965
Total	3369	3697	2885	

As we just did, now we attempt again to reduce the number of clusters by merging the smaller clusters with the cluster closer to them:

1. cluster 6 with 8;
2. cluster 11 with 15;
3. cluster 7 with 14;
4. cluster 4 with 8;
5. cluster 2 with 4;
6. cluster 9 with 8;
7. cluster 12 with 1;
8. cluster 5 with 14;
9. cluster 10 with 8;
10. cluster 15 with 14;
11. cluster 3 with 7.

We then stopped when all the clusters had a somewhat similar size.

We should note that because cluster 8 was quite bigger than the others, it was natural for other clusters to be closer to it than to other clusters, creating an even greater cluster.

We now have these clusters:

	Health	**Motor**	**Household**	**Total**
Low Salary Longtime Client Children Low Education	1339 (1)		866 (3)	2205
High Salary Recent Client No Children				0
High Salary Longtime Client Children Low Education	1425 (7)	3639 (8)		5930
Low Salary Recent Client Children Low Education				0
Low Salary Recent Client Children High Education		1847 (14)	835 (15)	2682
Total	3630	6321	1701	

And we can describe them:

- Clients who mostly pay for health insurance and have low salaries;
- Clients who mostly pay for health insurance and have high salaries;
- Clients who mostly pay for motor insurance, have high salaries, are long-time clients, and have lower levels of education;
- Clients who mostly pay for motor insurance, have low salaries, are more recent clients, and have higher levels of education;
- Clients who mostly pay for household insurance, are long time clients, and have lower levels of education;
- Clients who mostly pay for household insurance, are recent clients, and have higher levels of education.

10 Marketing Approach

Our suggestions for an innovated marketing approach for Antunes&Bação Ltd. company focuses on improving the relationship between existing customers and attract new customers. Along the already existing marketing activities of Antunes&Bação Ltd., additional measurements in order to raise awareness of the company should be considered. First and foremost, besides a re-make of their website and their mobile application, we suggest Antunes&Bação Ltd. company to set up accounts on Facebook, Instagram and Twitter to approach a wide range of customers by promoting new offers or special discounts in a much faster way. An easier and approachable customer service can also be offered via these social media accounts. Furthermore, the usage of well-known marketing strategies like the following should be implemented:

1. Advertising via Google Ads and social media advertisement on Instagram and Facebook for younger generations on a frequent basis. Additional advertisement streams on TV or radio as well as analog billboards and posters should also be taken into considerations in order to reach a broad range of customers.

2. Introducing campaigns like "Antunes&Bação Ltd. will donate 25% to the bush fires in Australia for every new customer or newly added insurance subscription.". In general, Antunes&Bação Ltd. should make regular donations to trusted environmental charities and nonprofit organizations to gain an appreciated brand image by customers who are concerned by the happenings in the world. A company can benefit from giving back by building respect and a good reputation toward its customers.

3. Raising awareness of the company's brand by visiting universities or company to get in touch with potential new customers. This approach can go in parallel with several sponsoring activities at events at visited universities or companies to deepen the relationship with existing or potential customers. Hosting "Get to know us" events or being present in the city with information stands by handing out flyers and gift bags will also raise awareness of the company's brand and portfolio.

For more specific guidance, several marketing approaches were designed based on the obtained clusters which serve as the base to explore further characteristics or improving areas from the customer segmentation. The marketing approaches are listed below.

1. Offer a loyalty program for long time clients and clients with children. A loyalty club ranks (e.g. Gold, Silver and Bronze) with different insurance and service packages should be introduced, including discounts, promotions and other package deals to long-time clients and clients with children, as they are more likely to register all family members to one insurance company. Package deals could include several insurance policies

offered by the company to a reduced price. Additionally, a family package could include the offer of a lower insurance costs for every additional registered family member.

2. For low income and long time customers, offer flexibility for payments in rates. Offering flexible payment plans to long-time customers with low income as they have proven their loyalty to the company to deepen the customer relationship. Flexible payments allow customers to make their payments on their preferred instead of a fixed date of the month. This loyalty status program could also be given to new customers with low income after a specified amount of time as clients of the company in order to allow more flexibility in payment to low income groups in general.

3. Motor insurance and additional insurance packages: Offer a discount or package deal on Life or Health insurance if the customer has a Motor insurance.

4. Promote discounts for existing customers on their next insurance purchase when they refer the insurance to a friend or family member and bring in a new customer successfully. This promotion is targeted at increasing the number of customers while raising awareness of the company through verbal reference/advertisement of existing customers to potential new customers. Additionally, it will increase sales revenue as the customers are more likely to make use of their promoted discount in adding another insurance plans they have not enrolled in yet.

5. Hosting educational and fitness insight days to deepen the company's relationship to its customers. Improving the customers relationship by promoting educational and fitness insight days offered by the insurance. These insight days are especially important to raise awareness of the importance of Health and Household policies and present the companies offer in the categories Health and Household since many customers may be uneducated or unaware in regard to these policies. Customers can directly register to the mentioned insurance coverage options at the information stands at the insight days.

6. Offer guidance to customers regarding Household policies. Approaching new or long time customers about household policies and offerings. As many young university graduates or workers may be unaware of the responsibilities that come with owning or renting a house.

11 Classifying new data

Although it isn't asked if our client wants to classify new data, e.g, if new costumers want to buy their insurance from Antunes&Bação Ltd., we can use a Decision Tree or a KNN algorithm to classify them, or even the outliers can be classified using these approaches.

If you're interested on these approaches, you'll need to hire us again. ☺

But we recommend a Decision Tree because it's easier to understand the reasons behind a specific classification decision.

References

[1] Fernando Bação, Victor Lobo, and Marco Painho. Geo-self-organizing map (geo-som) for building and exploring homogeneous regions. In *International Conference on Geographic Information Science*, pages 22–37. Springer, 2004.

[2] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[3] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[4] Konstantinos G Derpanis. Mean shift clustering. *Lecture Notes*, page 32, 2005.

[5] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[6] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[7] Y Savant and L Admuthe. Compression of grayscale image using ksofm neural network. *International Journal of Scientific & Engineering Research*, 4(1):2229–5518, 2013.