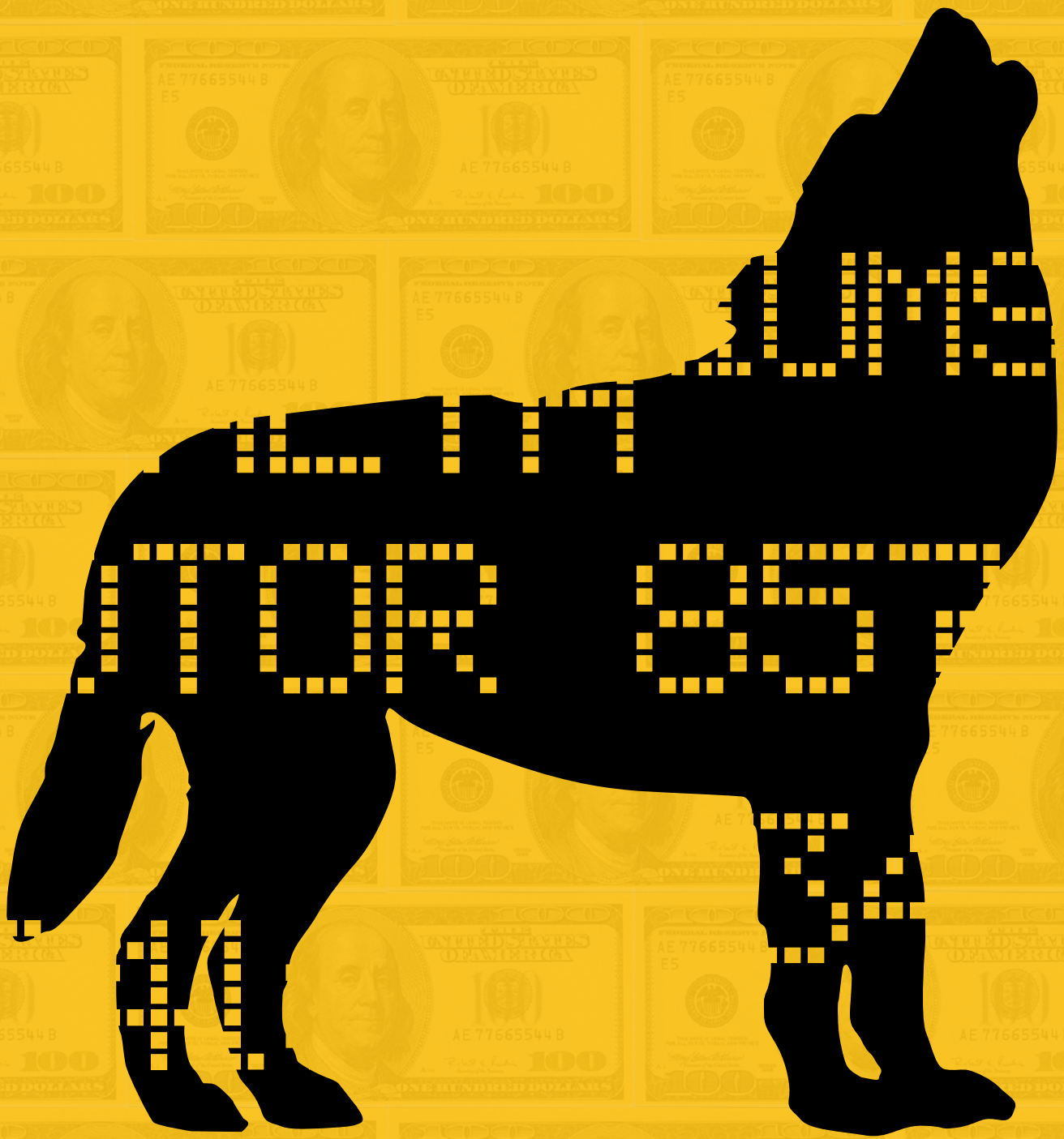
M20180054
VITOR MANITA

M20180061
RODRIGO UMBELINO

M20180081
RITA FRANCO

DATA MINING

CUSTOMER INSURANCE SEGMENTATION ANALYSIS



Index

1. Introduction

The aim of this project was to test the knowledge and skills acquired during the semester, regarding the lectures of the Data Mining course. It was proposed to develop a Customer Segmentation, for a fictional insurance company. After our analysis, the aim of our work would be to provide a clear and proven profiling of Customers within the data set, in order to be delivered to the Marketing Department for further development.

For these goals to be achieved, we followed the classic KDD process framework. Having said that, we started by pre-processing the given data, followed by transforming the said data. After these steps, we played around with several Data Mining techniques and different clustering algorithms, and decided based on the best results, by comparing the several methods. Finally, we elaborated succinct marketing approaches hinged on our final clusters.

The details of such analysis, as well as the decision factors throughout the project, are fully described within the scope of this report.

2. Variable Analysis

LOB				
Variable	**Description**	**Min**	**Max**	**Mean**
Motor	Annual Premiums (in euros) for Motor Insurance	-4.11	11604.42	300.47
Household	Annual Premiums (in euros) for Household Insurance	-75.00	25048.00	210.43
Health	Annual Premiums (in euros) for Health Insurance	-2.11	28272.00	171.58
Life	Annual Premiums (in euros) for Life Insurance	-7.00	398.30	41.86
Work_compensate	Annual Premiums (in euros) for Work Compensation Insurance	-12.00	1988.70	41.28

Table 1. Variable description - LOB

ENGAGE					
Variable	**Description**	**Categories**	**Min**	**Max**	**Mean**
First_Policy_Year	Year of the first policy made by the client	-	1974	53784	-
Birthday	Client's birth date	-	1028	2001	210.43
Living_Area	Area in which the client lives	LA1; LA2; LA3; LA4	-	-	-
Educ	Client's highest level of education attainment	1 – Elementary; 2 – High School; 3 – Bsc/Msc; 4 - PhD	-	-	-
Child	If an individual has children or not	Binary (0;1)	0	1	-
C_value	The relative value of a customer for the company	-	-165680.42	11875.89	177.89
Claims_rate	The amount paid by the company (in euros)	-	0.00	256.20	0.74

Table 2. Variable description - ENGAGE

In figure 1. it is represented the evolution of the number of clients that the insurance company had, per year. In this sense, it is possible to infer the growth that the company had in the first years, maintaining relatively stable for a big period of time (averaging around 450 clients per year). In most recent years, it is possible to observe a decrease in the number of clients.

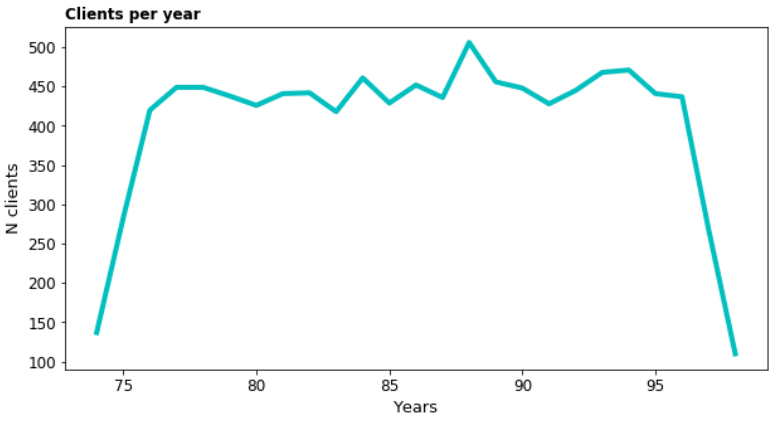
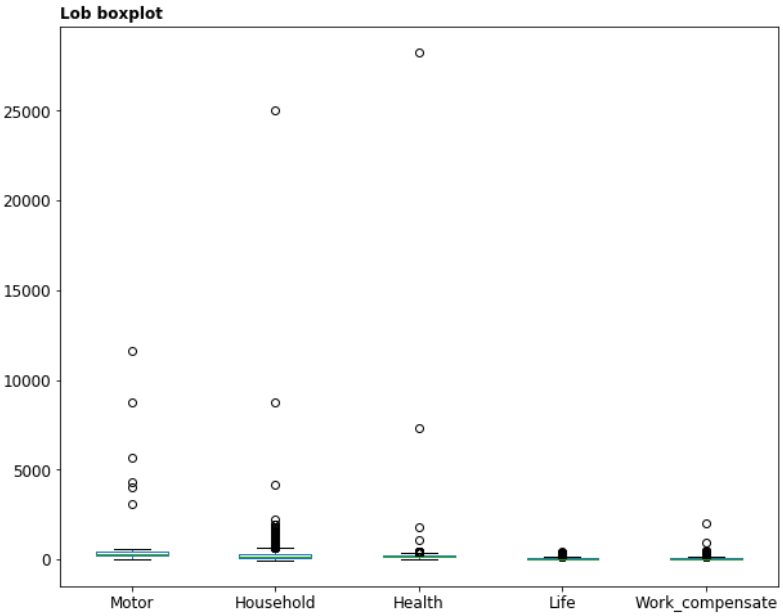


Figure 1. number of clients per year



On figure 2., there are represented the boxplots which provide an explanation in a visual manner for the variability of the variables within the LOB table. Whilst the majority of the variables are not characterized by having a huge variability, it has been proven to be very useful in the detection of outliers, has they been clearly represented in the boxplots.

Figure 2. LOB variables boxplot representation

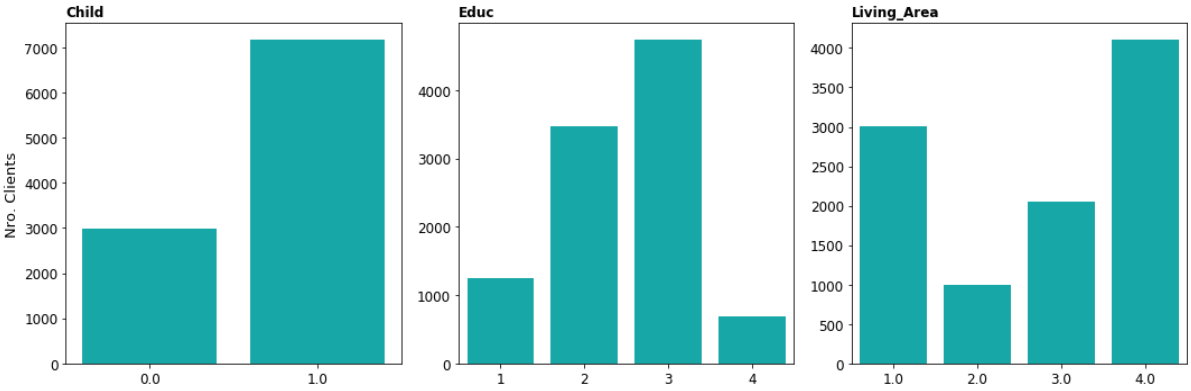


Figure 3. Categorical variables frequency per class

The histograms on figure 3. provide a way of quantifying the number of clients, depending on the study of different variables from the ENGAGE table. In the first graph, we can see that around 7000 clients have kids, opposing to the around 3000 that do not. If we take a glance at the second graph, with regards the highest education attainment level, we can conclude that the majority of the company's clients have a High School diploma or a Bachelor's/Master's degree. Clients with a PHD or that have stopped their academic career in Elementary school are part of the minority.

Finally, on figure 4., we can observe the geographical dispersion of the company's clients, regarding their living area. Amongst the 4 living areas presented in this variable, around 3000 clients live in living area 1, around 1000 live in living area 2, around 2000 on living area 3, and last, but not least, around 4000 in living area 4.

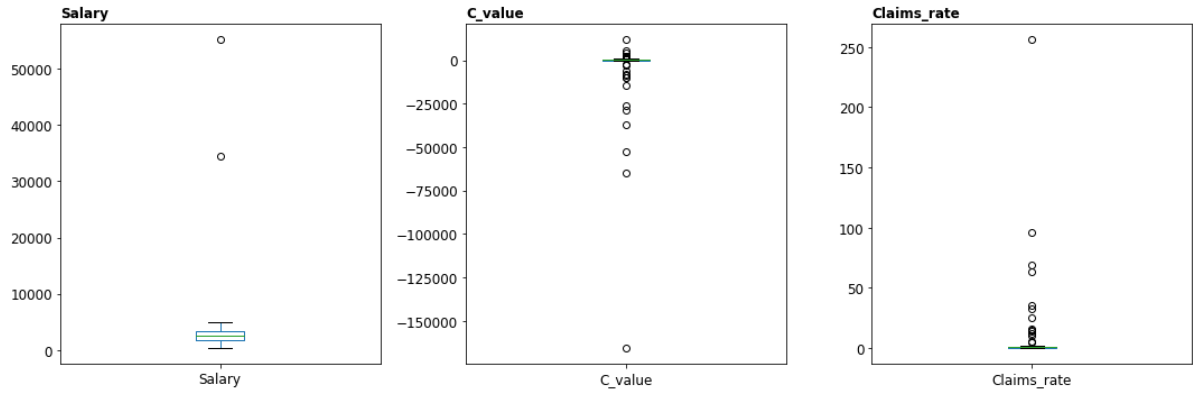


Figure 4. Engage variables boxplot representation

3. Filter Data

In order to filter the data, we decided to merge both tables (LOB and ENGAGE) into one single table. This way, any pre-processing operations only had to be done once, whereas if we kept the split tables, we would've had to duplicate the steps for coherence sake. After the pre-processing phase is completed, we split the two tables into their original structure, to conduct further analysis.

3.1. Coherence Checking

After getting some hands-on with the data set itself, it became quite clear that there were some incongruencies that needed to be taken care of. Firstly, we noticed that there were 1997 rows where the client's birthday date was after their first-year policy. This of course is impossible in any real-world scenario and could be caused by error in user-input, therefore we decided to remove the Birthday column. In addition to this, there was also one record in which the value of the first policy year was 53784, which is obviously an error, so we eliminated it.

3.2. Dealing with Nulls

Within the given data, there was some presence of null values. In order to tackle this problem, we first had to establish the difference between a null value and a zero value. Zero values are obviously possible, whereas if we decided to delete all null values, we would have eliminated 2.87% of the data, which would have left very little leeway to work with the outliers.

Having said that, we chose two columns with the least number of null values (which ended up being the Life and Work columns) and replaced the said null values with the mean of each column respectively. After this step, we were then able to delete the other records with null values, adding up only to 1.32% of deleted data. This way we had more room to work with the outliers once we separate them from the analysis. After this, we split the table into two, similarly to the initial structure.

3.3. Dealing with Outliers

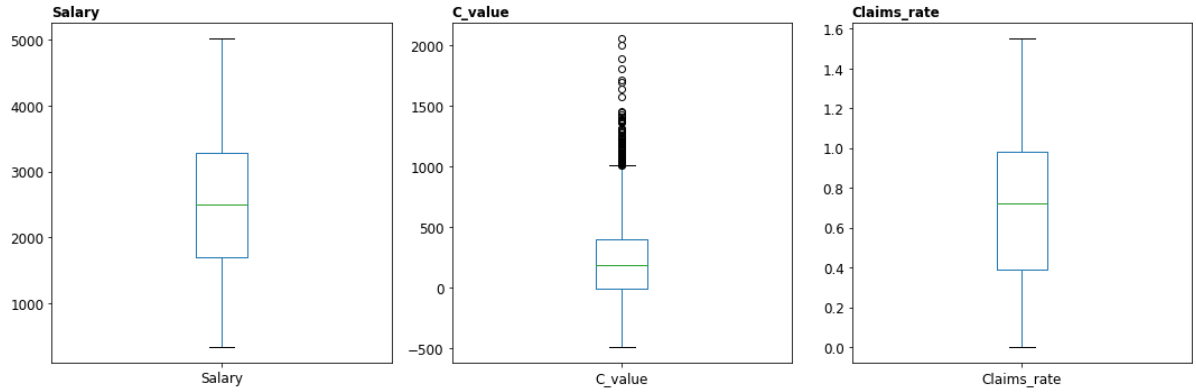


Figure 5. distribution of Engage variables without outliers

After identifying the outliers based on the boxplots presented before, we decided to exclude such data from further analysis, since it would create discrepancies within our model. That is not to say that they were eliminated, they were just put aside and analysed by themselves. Having said that, the 29 identified outliers were removed from the model, which accounts for a grand total of 2.06% of deleted data (212 records to be more precise).

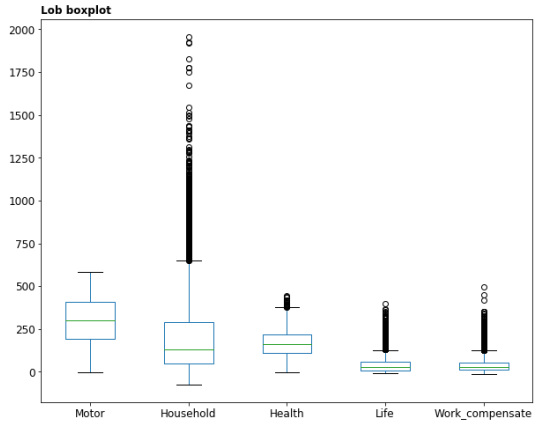


Figure 6. distribution of LOB variables without outliers

Variables	Outlier **cut**
Salary	>20000
C_value	<-2000
Claims_rate	>50
Motor	>2000
Household	>2000
Health	>5000
Work_compensate	> 750

Table 3. Outlier cut in each variable

4. Transform Variables

Another crucial step in the pre-processing phase would be the transformation of variables. Such transformation took place because there were variables that could be enhanced in their intrinsic explanation power, regarding the context of the problem. New variables were also created following the exact same assumptions.

4.1. Engage

In the **Engage** table, a new variable **Total_Premium** was created, which consists on the sum of the premiums of every Insurance field (Motor, Household, Health, Life and Work Compensation) for every client. This allowed a better representation of the monetary investment that each client has applied in the company. Secondly, the **Educ** variable that refers to the client's highest level of education attainment, was

transformed into a binary variable (**High_Educ**) which states whether a given client has pursued college education.

Lastly, we decided to transform the **First_Policy_Year** variable into **Years_as_Cust**, and we decided so because we deemed that the notion of how long a client has been a company customer or not would be much more productive. This transformation was done by subtracting the value of the **First_Policy_Year** to 2016, which is the year when the data was collected.

```
engage['Total_Premium'] = lob[['Motor','Household','Health','Life','Work_compensate']]. sum(axis=1)
engage["High_Educ"] =0
engage.loc [(engage.Educ == 3.) | (engage.Educ == 4.), "High_Educ"] = 1.
engage["years_as_cust"] = 2016. - engage["First_Policy_Year"]
```

4.2. Lob

For the **LOB** table, we decided to convert every value into a ratio, by dividing the value of each insurance field premium, by the **Total_Premium** variable that was created before. This allowed to represent the relative frequency of each Premium, since the absolute values are already present in the Engage table.

```
lob['R_Motor'] = lob['Motor']/engage['Total_Premium'] *100
lob['R_Household'] = lob['Household']/engage['Total_Premium'] *100
lob['R_Health'] = lob['Health']/engage['Total_Premium'] *100
lob['R_Life'] = lob['Life']/engage['Total_Premium'] *100
lob['R_Work_compensate'] = lob['Work_compensate']/engage['Total_Premium'] *100
```

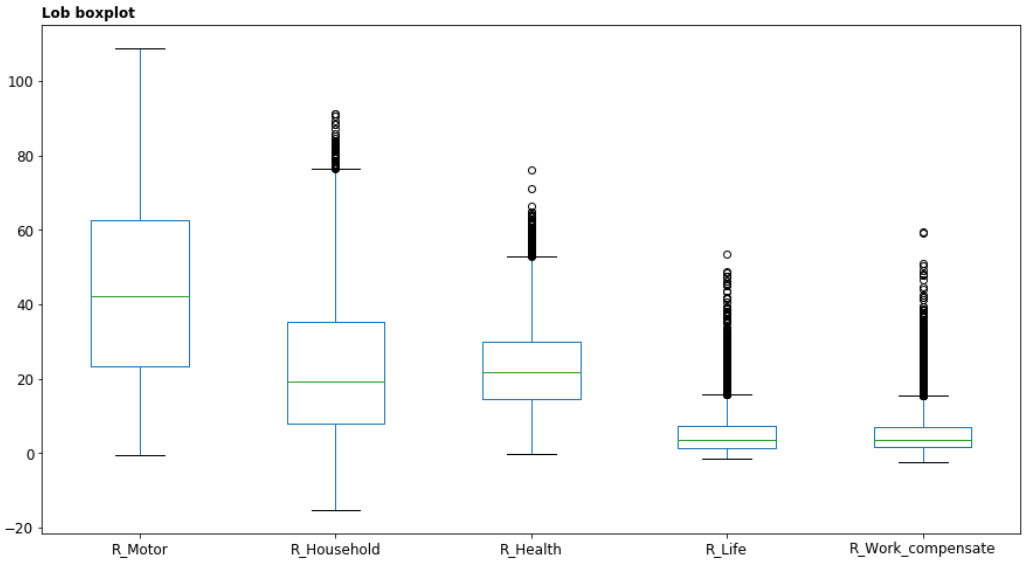


Figure 7. distribution of new LOB variables

5. Principal Components Analysis (PCA)

After having all the variables constructed and ready for clustering, we decided to apply a PCA in order to reduce the variables and be able to cluster in a lower dimensionality, hopefully a 2-D space, on both tables.

Regarding **LOB**, we preformed the PCA on five variables representing the proportion spent on each insurance type by client. In order to decide how many components to use, we considered two rules: the elbow graph and the cumulative explained variance. Although the elbow graph (figure 8.) pointed to use two components, there were needed three principle components to explain at least 80% of the variance, see table 4. Sadly, three variables would not be well explained by neither of the components, as shown in table 5., giving an indicator to drop this analysis.

In addition, we applied PCA on **Engage**, using only the numeric variables: **Salary**, **C_value**, **Claims_rate** and **Total_Premium**. On this segmentation approach, PCA performed better, both the elbow graph and the cumulative variance indicated two components, as shown in table 3. and figure 9., and all variables were well explained by one component or the other, table 6.

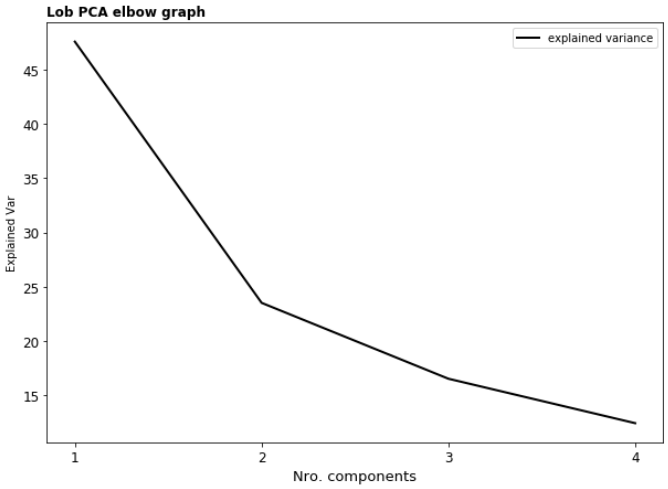


Figure 8. LOB PCA elbow graph

Another aspect to consider was the increase in complexity of analysing the cluster in these new axes given the difficulty in assigning specific meaning to each component. In sum, we decided that the loss of explained variance and the increase in analysis complexity would not be worth it to reduce the dimensionality only by one dimension in each table. Thus, PCA was not applied.

PC	Explained Var.	Cumulative Var.
1	47.6	47.63
2	23.5	71.14
3	16.5	87.60
4	12.4	100.00

Table 5. LOB principle components explained variance

PC	Explained Var.	Cumulative Var.
1	50.2	50.18
2	29.9	80.07
3	19.0	99.07
4	0.9	100.00

Table 4. Engage principle components explained variance

PC	R_Motor	R_Household	R_Health	R_Life	R_Work_compensate
1	-0.62	0.48	0.10	0.43	0.43
2	-0.04	-0.50	0.83	0.16	0.16

Table 6. LOB correlation between variables and principle components

PC	Salary	C_value	Claims_rate	Total_Premium
1	-0.11	0.69	-0.65	0.28
2	-0.72	-0.10	0.29	0.63

Table 7. Engage correlation between variables and principle components

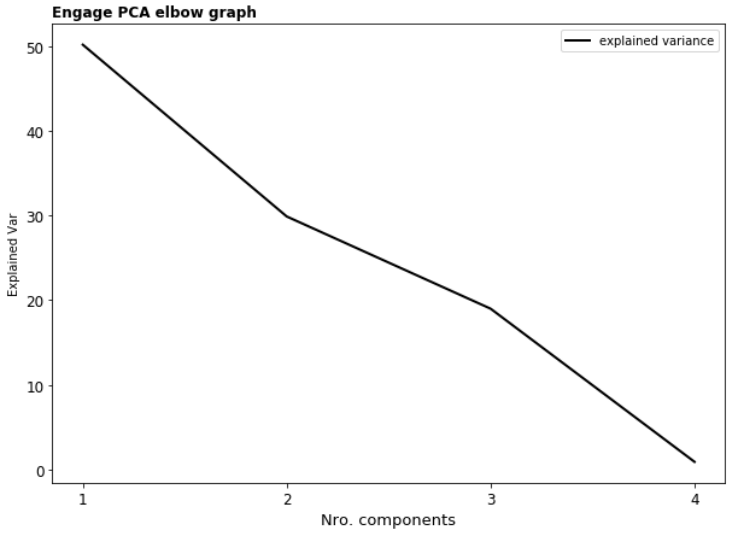


Figure 9. Engage PCA elbow graph

6. LOB Segmentation

Moving on to the clients' segmentation on a product perspective, this analysis was made using four variables: **R_Motor**, **R_Health**, **R_Life**, **R_Work_compensate**, explained in chapter 4.2. The analysis was performed with ratios to enforce that this was a product segmentation, the value perspective will be clustered posteriorly with the social and geographic aspects of the customers. So, in this phase we are prioritizing the products most bought by each client and not their monetary value to the company. The result of the algorithm will be different clusters where clients share similar product choices. Lastly, given the high correlation between **R_Motor** and **R_Household** (-0.8), one can be explained by the other's behaviour, so the proportion of money spent in household insurances was not used to perform the cluster analysis.

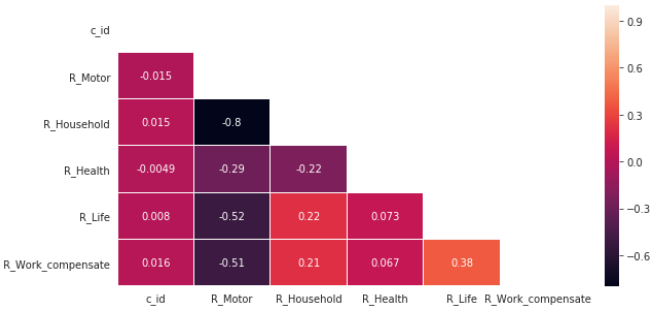


Figure 10. LOB segmentation correlation matrix

The last step before implementing the cluster algorithms was deciding whether to standardize or not the data. Although the data was all in ratio, it did not have the same scale, given that there were negative values and values above 100% due to the indemnity payed to the clients that exceed the premiums for example. Thus, we standardized all data using *standard scaler*.

Regarding the clustering analysis, we implemented different algorithms: *k-means*, *hierarchical*, *mean-shift*, *DBSCAN*, *expectation-maximization* and lastly *k-means with*

the seeds from hierarchical, the selection of the best method will be explained in the next chapter.

6.1. Lob Final algorithm: Hierarchical

Taking into consideration all the clustering algorithms experimented to group customers in terms of products, the final choice came down to two: *hierarchical* clustering or *k-means with hierarchical seeds*. Regarding the other techniques, a quick glance to the cluster centroids revealed that they were not good methods for the data in hand, the centroids were really like each other. Besides, *DBSCAN* identified a big percentage of the clients as outliers/noise.

Having our choices narrowed down, by looking at the cluster centres, we understood that each technique improved a certain group over the other. However, we needed a real metric to support our decision, thus using the most known measure of distance, the Euclidean distance. This way, we were able to conclude that the *hierarchical* clustering method minimized the distance between each data point and its corresponding cluster centre, the intra-cluster distance.

6.1.1. Clusters

For grouping the clients in terms of product choices, we can denote 3 major groups whose centroids are summarized in table 7. By observing fig 14, we can realize that the clusters are really close to each other and sometimes even overlapping, this may be due the representation only in a two-dimensional space, but it is also due to the proximity of all the clients' product choices. We now present the three clusters that were found and their main characteristics.

Next to each group, there is a small chart representing the distribution of spending's in each category, blues are for **R_Motor**, red is **R_Household**, green is **R_Health**, orange is **R_Life** and grey is **R_Work_compensate**.

Next to each group, there is a small chart representing the distribution of spending's in each category, blues are for Motor, red is Household, green is Health, orange is Life and grey is Work compensate

Motor

A cluster made by 3240 clients who show a strong preference for Motor insurances representing more than 70% of their purchases. Regarding the remaining products, they spend less than 2% in Life or Work insurances, being the group with the lowest sales for these products. Finally, around 16% of their money spent in the company is on Health products.

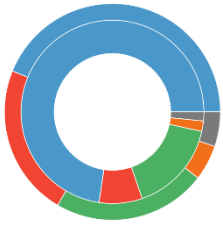


Figure 11. Motor vs Global ratios distribution

Personal

Having only 935 customers is a niche that spends 13,28% in Work and 17.11% in Life insurances, values that are double of the population average. These clients also spend on average a third of their money in Household products. Regarding Motor and Health insurances, these clients spend around 18% in each. This is the group where clients spend their money more evenly throughout the five products.

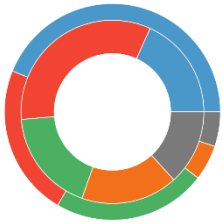


Figure 12. Personal vs Global ratios distribution

Generic + Health

Lastly, we have the biggest group formed by 5955 clients. These clients show a bigger preference to health insurances than the remaining clusters, spending 27% of their purchases in this product. In concern to the other products, this cluster is spending around the population average in all of them, having Household ratio a bit above average and the remaining a bit below.

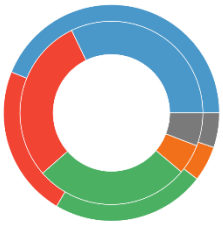


Figure 13. Generic + Health vs Global ratios distribution

LOB	R_Motor	R_Household	R_Health	R_Life	R_Work_compensate
Motor	72.66	7.62	16.31	1.82	1.59
Personal	18.24	33.08	18.28	17.11	13.28
Generic + Health	32.18	29.26	27.42	5.30	5.84

Table 8. LOB clusters centroids

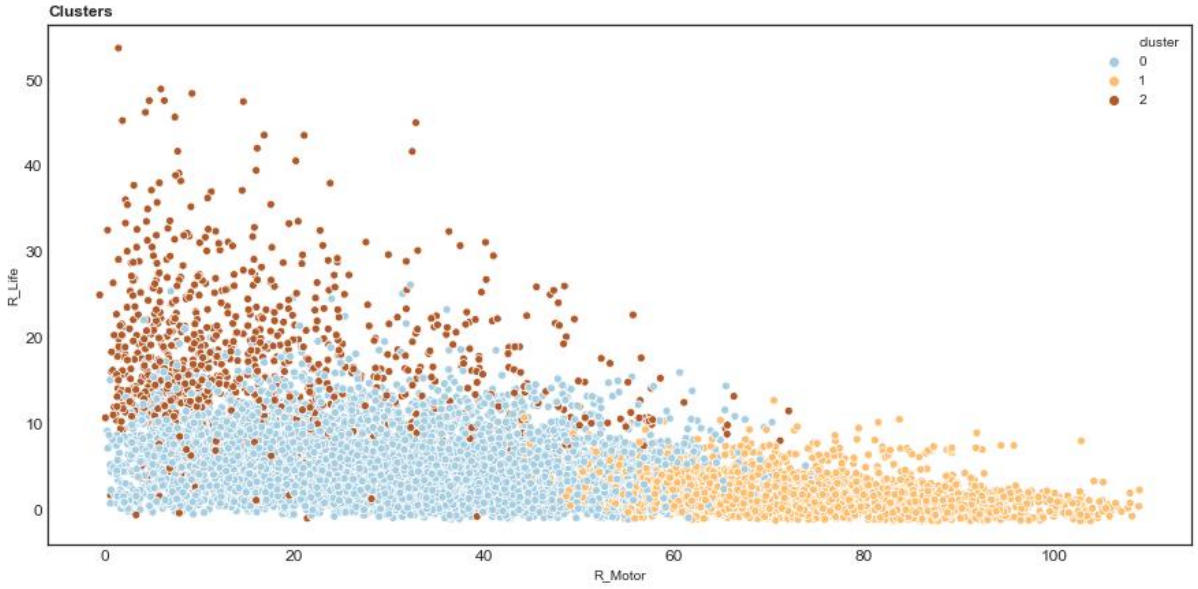


Figure 14. LOB clusters representation *0 – Personal; 1 – Motor; 2 – Generic + Health*

7. Engage Segmentation

Having done a product segmentation, the next step was to perform a clustering analysis by customer value and socio-demographic aspects, using the **Engage** table. For this segmentation, the approach needed to be different given that we had both numerical and categorical variables.

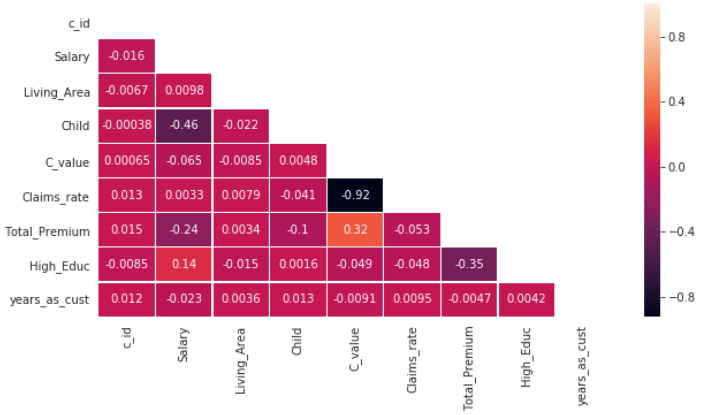


Figure 15. Engage segmentation correlation matrix

Regarding the categorical variables, we one-hot encode them in order to use a standard algorithm such as *k-means* or we could apply *k-modes* separately to this set or variables. Thus, we tested first using one-hot encoding and running the *k-means* algorithm and we realized that the clusters were being influenced by the categorical variables and they did not differentiate the customers in terms of monetary value. Hence, we decided to apply *k-modes* to **Child, High_Educ** and **LA**, then crossing the resulting clusters with the ones created from the variables related with monetary value.

As for the previous segmentation, we used *standard scaler* to standardize the data and then run a set of clustering algorithms: *k-means*, *hierarchical*, *mean-shift*, *DBSCAN*, *expectation-maximization* and lastly *k-means with the seeds from hierarchical*. As well as in **LOB**, there were highly correlated variables: **C_value** and **Claims_rate** and we decided to keep **C_value**, given that it was related to the lifetime value of the client and not only the most recent years.

7.1. Engage Final algorithm: *K-means with Hierarchical Seeds*

In order to have a better grasp of the customer value, we used the three continuous variables we had, namely, **Salary, C_value** and **Total_Premium.** Beforehand, we knew that methods like *DBSCAN* and *Mean-shift* would not work so well, given the shape of the data we have. A quick glance to the results of these methods proved just that. The same happened with *Expectation-Maximization* clustering method. In the end, the choice was between using *Hierarchical*, *K-means and K-means with hierarchical seeds*. For comparing the two *K-means* methods we used the inertia metric, which showed a smaller error in *K-means with hierarchical seeds*. Then, for comparing this method with a simple *Hierarchical* technique, the Euclidean distance was used. The final algorithm used, that minimized the distance from each point to its cluster centre was *K-means with Hierarchical seeds*.

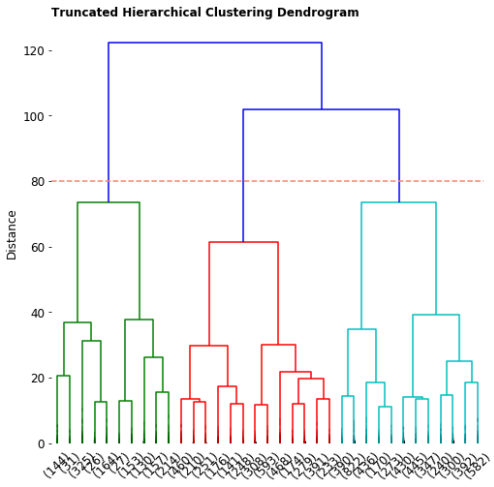


Figure 16. Engage dendogram

7.1.1. Clusters

By accessing the dendrogram and elbow graph, we decided to cluster engage in 3 different groups. In table 8 we can see summarized the centroids of the different clusters, as well as their graphic representation on figure 17. Here the clusters are represented in a two-dimensional space, not their original dimensionality, but we can see that their frontiers are better defined than in the previous segmentation approach. The final three groups for value segmentation as characterized as following:

Less_valuable

A cluster made by 4273 clients (42.18% of data) who reveal themselves as having the lowest customer value, with an average of 179.0, lower than the 244.72 of the population. Then, they are the ones who spent the lowest amount of money on overall premiums, averaging 673.87€. Lastly, notice that these clients, despite having the smallest customer value, represent a good business opportunity, given that they have a Salary average of 2032.98€, very close to the population average of 2218.11€.

Most_valuable

Having 4256 customers (42.01% of data), this group is made by clients who are the most valuable ones, since they have the highest customer value and highest amount of money spent on total premiums, 341.56 and 1020.26€, respectively.

It is important to notice that this cluster, despite being the "most valuable" customers, they are also the ones who earn the less, with an average of 1157.35€.

Wealthier

Lastly, we have the smallest group built by 1601 clients (15.8% of data). These clients have a customer value and total premium spending's like the population mean, 213.59 and 737.66€, respectively, however, they differ themselves from the other two groups by having the highest earnings, around 3463.01€, meaning that they are very good potential customers and have the capital to spend more on premiums.

Engage	Salary	C_value	Total_Premium
Wealthier	3464.01	213.59	737.66
Most_valuable	1157.35	341.56	1020.26
Less_valuable	2032.98	179.00	673.87

Table 9. Engage clusters centroids

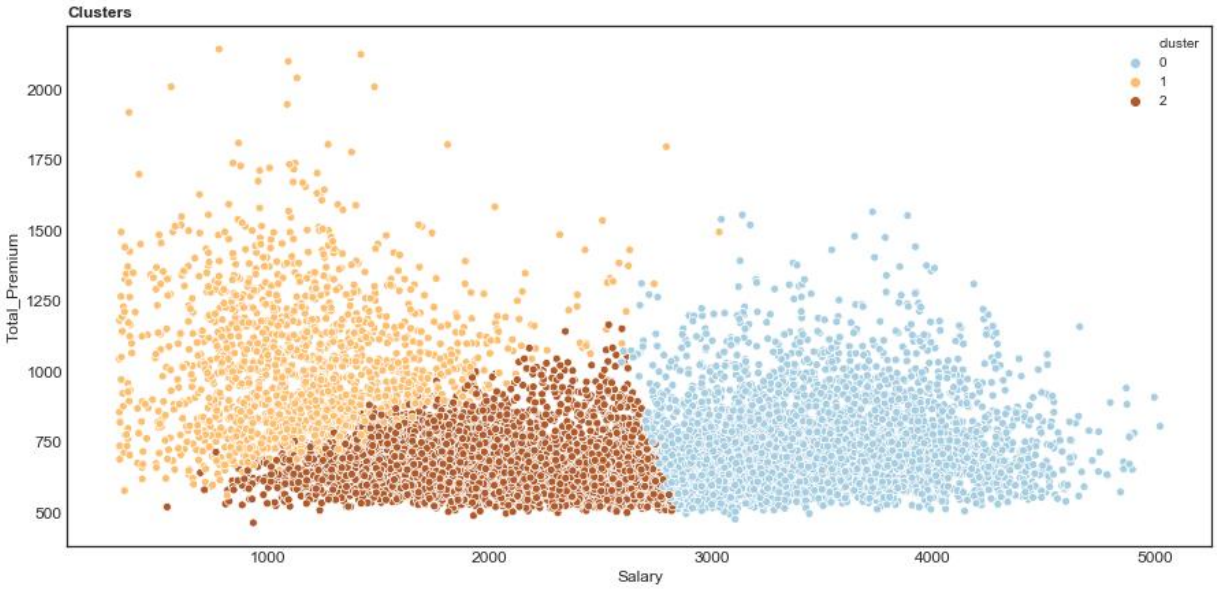


Figure 17. Engage clusters representation *0 – Wealthier; 1 – Most_valuable; 2 – Less_valuable*

7.2. K-Modes

Having a set of categorical and binary variables not used for grouping clients in a value perspective, due to their poor effect on the final clusters, we decided to apply the *K-modes* technique. This method has a similar approach to *K-means*, however, instead of the mean value, it uses the mode of each variable. Thus, we used 3 variables we determined as most relevant: **Child**, **High_Educ** and **LA**.

After some trial and error, we decided to keep 3 clusters: **high_educ_la1** is a group of clients that has no children, has high education and where most clients live in living area 1. **High_Child** differentiates from the previous group by having children and living mainly in living area 4. They have, as well, high education. Finally, clients in **Low_Educ_Child** have similar characteristics to the previous group, being the only difference that they have no High education.

Kmode cluster	Child	High_Educ	LA
high_educ_la1	0	1	1
high_child	1	1	4
low_educ_child	1	0	4

Table 10. K-modes clusters

After analysing *k-modes* results, we decided to cross them with the groups obtained in *engage clusters*, creating 9 new groups. Notice that we will not cross these 9 groups with the results of **Lob**, the purpose here is to have a good perspective of the characteristics of clients in value cluster. Thus, by crossing results, we can see that in the group of ***Less valuable*** customers, we have a higher presence of clients having children that live in LA4, where the high education is somehow frequent. For ***Most valuable*** clients, we see that the most important characteristic is most individuals with low education. At last, for ***Wealthier*** clients, there is a more frequent presence of high education levels and more individuals living in LA1, which indicates us that this is probably a more expensive area to live in.

Engage cluster	Kmode cluster	Client frequency
Less_valuable	high_child	38.4%
	high_educ_la1	23.6%
	low_educ_child	38.1%
Most_valuable	high_child	13.2%
	high_educ_la1	21.9%
	low_educ_child	64.9%
Wealthier	high_child	19.1%
	high_educ_la1	51.3%
	low_educ_child	29.7%

Table 11. Crossing Engage clusters with k-modes

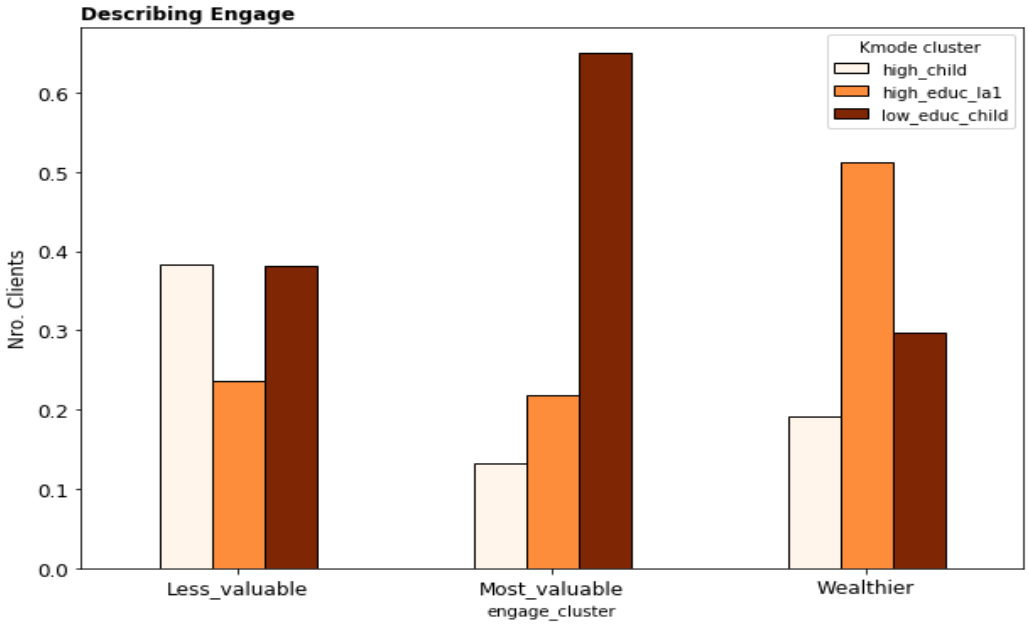


Figure 18. K-modes cluster frequency per engage clusters

8. Crossing clusters

To get a good grasp of our clients and to have the power of redirecting specific and appropriate marketing campaigns to them, we need to cross the results of both value and product (premiums) clusters. Having 3 groups in **Lob** and 3 groups on **Engage**, we can expect 9 new groups of clients. By crossing clients, using their *client id*, we can see in which category they fall. Overall, each premium group is well represented in terms of clients, except the generic group that has the less frequency of clients. Another observation to point out is that in ***Motor*** premiums group, we have only 16 ***Most valuable clients***. A small number like this does not offsets the effort of creating a specific marketing campaign, so, we decided to move these clients to their nearest group, using the *Euclidean Distance*. For all 16 clients, as they have similar characteristics, their closest group was ***Less_valuable*** customers, in the same category.

Lob	Engage	Client frequency (abs)	Client frequency (%)

Generic + Health	Less_valuable	2019	33.90%
	Most_valuable	1170	19.65%
	Wealthier	2766	46.45%
Personal	Less_valuable	269	28.77%
	Most_valuable	415	44.39%
	Wealthier	251	26.84%
Motor	Less_valuable	1968	60.74%
	Most_valuable	16	0.49%
	Wealthier	1256	38.77%

Table 12. Crossing LOB and Engage clusters

Lob	Engage	Client frequency (abs)	Client frequency (%)
Generic + Health	Less_valuable	2019	33.90%
	Most_valuable	1170	19.65%
	Wealthier	2766	46.45%
Personal	Less_valuable	269	28.29%
	Most_valuable	415	43.64%
	Wealthier	251	26.84%
Motor	Less_valuable	1984	61.23%
	Wealthier	1256	38.96%

Table 13. Final Clusters

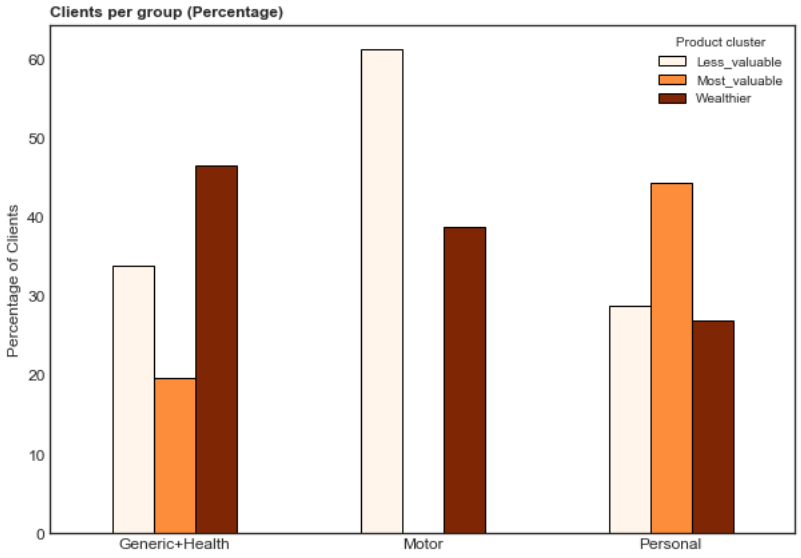


Figure 19. relative frequency of clients per group

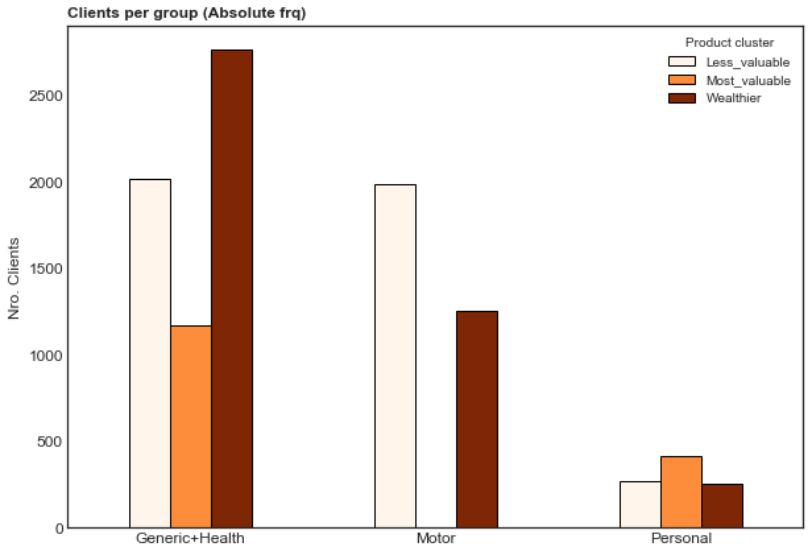


Figure 20. Absolute frequency of clients per group

9. Recovering Outliers

If we recall the early stage of this project, we removed the 29 outliers from our analysis in order not to bias the results. Since these clients have odd characteristics that make them stand out from the rest of the data, is not fair to completely ignore them. So, having them stored, we can now insert them in each corresponding group, using the *Euclidean distance* to check the nearest cluster.

Of course, that, by joining outliers to the formed groups will disrupt their mean values and distribution, however, this step is just not to waste these clients, so we have two final data sets: 1) Clustered clients used in the analysis 2) Clustered clients + outliers. The distribution, with the outliers joined, is shown in figure 21.

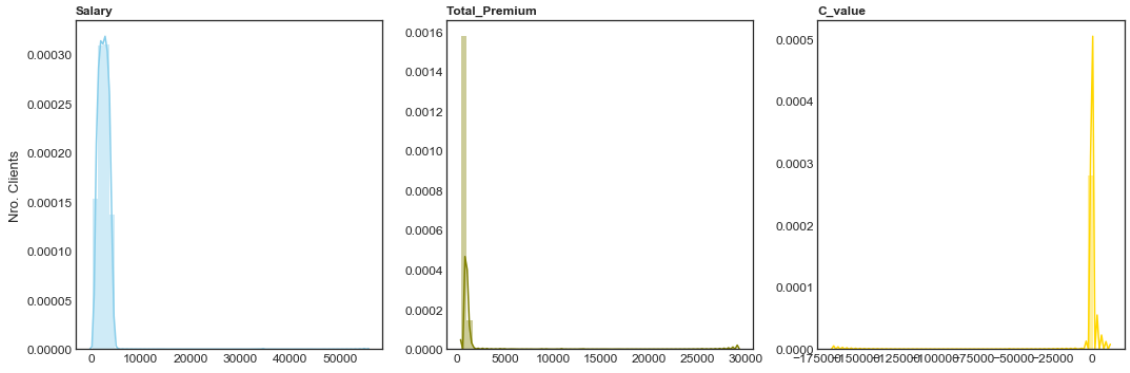


Figure 21. Variables distribution with outliers

10. Marketing Strategies

In the following three pages, a sketch of guidelines is given so with specific possible marketing campaigns for each type of Premiums group and for each type of customer, regarding their value to the company or demographic characteristics of each group.

Motor

Where do these clients spend their money?

73%	7.6%	16%	1.8%	1.6%
MOTOR	HOUSE	HEALTH	LIFE	WORK

Who are they

LESS VALUABLE

WEALTHIER

MARKETING

Loyalty

10% discount in Motor Premiums for a renewal of 3 years regarding Motor Insurance

(Not applied to new clients)

Introductory Price

15% discount on Household Premiums for Less Valuable clients and 10% discount for Wealthier clients

More is Better

For Less Valuable and Most valuable clients, an offer of 5% discount on Health Premiums per each additional Child

Personal

Where do these clients spend their money?

18.2%	33%	18.3%	17%	13.5%
MOTOR	HOUSE	HEALTH	LIFE	WORK

Who are they

LESS VALUABLE

MORE VALUABLE

WEALTHIER

MARKETING

Loyalty

5% discount in bundle Premiums pack for a renewal of 3 years regarding House, Life and Work

(Not applied to new clients)

New Wheels

Tire change offer for new Motor Premium contracts in partner repair shops exhisting in Living Area 1 and 4

More is Better

For Less Valuable and Most valuable clients, an offer of 5% discount on Health Premiums per each additional Child

Generic plus Health

Where do these clients spend their money?

32.3%	29.3%	27.4%	5%	6%
MOTOR	HOUSE	HEALTH	LIFE	WORK

Who are they

- LESS VALUABLE
- MORE VALUABLE
- WEALTHIER

MARKETING

Loyalty

10% discount in Health Premiums for a renewal of 3 years regarding Health Insurance

(Not applied to new clients)

Bundle

15% discount on bundle House, Life and Work Premiums if client purchases these 3 together on contract

More is Better

For Less Valuable and Most valuable clients, an offer of 5% discount on Health Premiums per each additional Child

11. Conclusions

This project was an important bridge between the theoretical and practical components of Data Mining. The segmentation asked was not so linear and the data consistency, completeness and accuracy problems required creative and logical solutions. From the clustering and dimensionality reduction techniques studied in class, we had some a priori knowledge about which ones would perform better for the data in hand. For the project, we segmented the customers under a premiums perspective, represented by the percentage of spending in each premium category, using a *Hierarchical* clustering method, resulting in 3 groups: ***Motor, Generic + Health*** and ***Personal***, each representing a highest amount of spending in one or several premiums. Parallel to this component, we segmented the variables regarding customer value using *K-means with Hierarchical seeds*, resulting in 3 groups: ***Less_valuable, Most_valuable*** and ***Wealthier***. To complement this information, a *K-modes* clustering method was also used in three of the customer demographic characteristics. This way, we were able to cross these results with the ones obtained in the value segmentation and better describe these groups.

Then, the 2 segmentations regarding premiums and value were crossed, where we obtained 8 different groups, after merging two groups given that one had very few customers. For a final step, we also retrieved the initial rejected customers who behaved abnormally and inserted them into each corresponding cluster.

Finally, we denoted some marketing strategies guidelines, with each one of the 8 groups in mind, for a better customer targeting.