Data Mining | Fall Semester, 2019 | Group Project

Abdallah Zaher, M20190684
Cristina Mousinho, M20190303
Gabriel Ravi, M20190925

Table of contents

Introduction **2**
Data Analysis **3**
Multivariate Analysis **10**
Clustering **12**
Cluster analysis **20**
Cluster descriptions for marketing approaches **27**
Outlier Analysis **29**
Conclusions **31**
References **32**

1. Introduction

As part of the 2019 Data Mining course for the Master in Data Science and Advanced Analytics, we were challenged to see ourselves as part of a fictional insurance company located in Portugal. We were given a data set to work with, from each we are expected to find different customers' profiles, to group them and to recommend marketing approaches based on those clusters.

By the end of this report, we should not only have an understanding of how our customers behave and how we can group them, but also a way to explain our findings and what could be done with them to the Marketing Department of our company.

We will begin by getting to know the provided data. To accomplish that, we will start by doing a univariate analysis of our features. As we encounter problems, such as missing values or outliers, we will make sure to apply different methods to solve them.

After, so we know how our variables relate to each other, we will proceed to do a multivariate analysis of them.

Completing those steps will hopefully guide us to a clean dataframe, that we can use to apply the clustering techniques we find to be adequate. We will make sure to visualize our results, plotting the necessary representations. We will also analyse them closely, both from theoretical and practical points of view.

Because the first steps will lead us to removing some of the rows present in our dataset, we will have to make decisions re-introducing (or not) to our dataset. It is important to really give thought to this step, as it might produce big changes in our final results.

Finally, we will reflect on all of our findings, and on how we will deport our knowledge to the Marketing Department.

2. Data Analysis

In this first chapter we will gather all of our efforts to understand, analyze, and improve the provided data. By the end, we should have a new and clean database, ready to be used for further analysis.

A. Classification of the variables

The initial dataset contains 14 different variables to be analysed. Since the application of clustering techniques is dependent on variables' types and natures, we started by classifying 11 of them into three categories - Nominal, Customer, and Premium.

➔ Nominal variables

These categorical variables explain the profile of the customer.

- Education Degree
- Geographic Living Area
- The Customer Has Children? (Yes or No)

➔ Customer variables

These quantitative continuous variables also relate to the profile of the customer.

- Gross Month Salary
- Claims Rate
- Customer Monetary Value

➔ **Premiums in LOB variables**

These variables contain information about premium categories.

- Motor
- Household
- Health
- Life
- Work Compensations

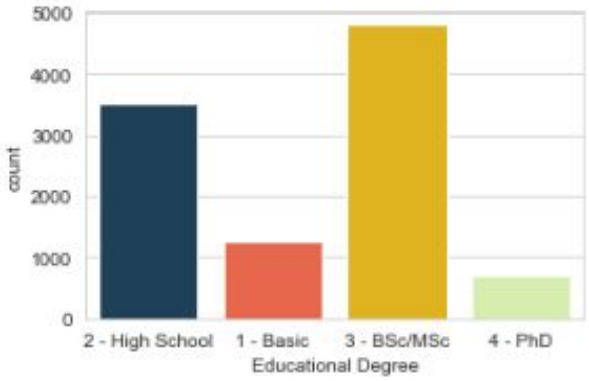
B. Univariate analysis of the data

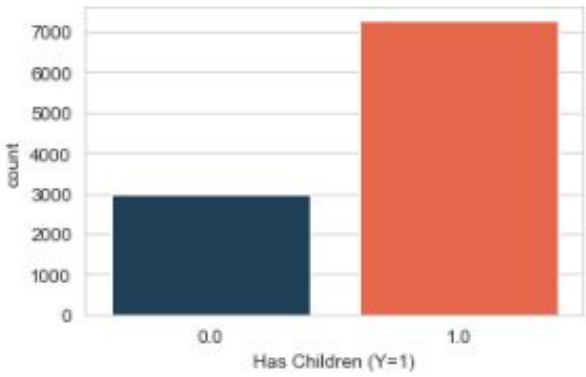
Let's now take a closer look at each one of our variables, starting by the descriptive analysis of the quantitative variables.

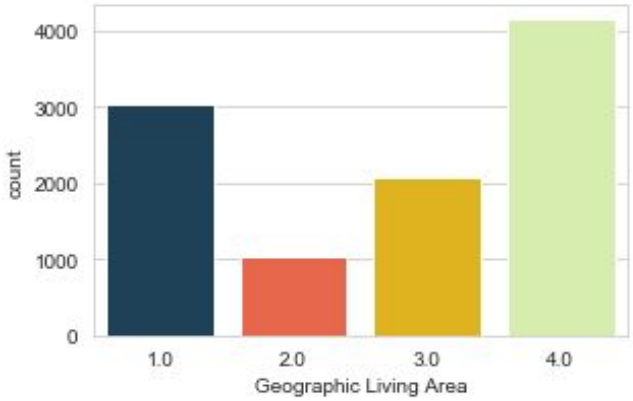
	First Policy´s Year	Brithday Year	Gross Monthly Salary	Customer Monetary Value	Claims Rate	Premiums in LOB: Motor	Premiums in LOB: Household	Premiums in LOB: Health	Premiums in LOB: Life	Premiums in LOB: Work Compensations
count	10266.000000	10279.000000	10260.000000	10296.000000	10296.000000	10262.000000	10296.000000	10253.000000	10192.000000	10210.000000
mean	1991.062634	1968.007783	2506.667057	177.892605	0.742772	300.470252	210.431192	171.580833	41.855782	41.277514
std	511.267913	19.709476	1157.449634	1945.811505	2.916964	211.914997	352.595984	296.405976	47.480632	51.513572
min	1974.000000	1028.000000	333.000000	-165680.420000	0.000000	-4.110000	-75.000000	-2.110000	-7.000000	-12.000000
25%	1980.000000	1953.000000	1706.000000	-9.440000	0.390000	190.590000	49.450000	111.800000	9.890000	10.670000
50%	1986.000000	1968.000000	2501.500000	186.870000	0.720000	298.610000	132.800000	162.810000	25.560000	25.670000
75%	1992.000000	1983.000000	3290.250000	399.777500	0.980000	408.300000	290.050000	219.820000	57.790000	56.790000
max	53784.000000	2001.000000	55215.000000	11875.890000	256.200000	11604.420000	25048.800000	28272.000000	398.300000	1988.700000

Just by a simple overlook at our data, specifically at the minimum and maximum values, we can detect some mistakes (probably outliers) present in it: the first policy's year takes numbers up to 53784, and there's at least one customer born in 1028, which is just not feasible.

As for the categorical variables, since the Pandas description doesn't provide much information, we decided to plot a histogram for each feature, in order to get a better understanding on how they are distributed.







C. Building new features

To complement the provided variables, we decided to build three more:

- Age - that relates to the customers age in the year of 2016.
- Client Spend - that comes from the summation of all premiums.
- Fidelity - that will allow us to know more about the customers and their time spent as clients. It is calculated by current year (2016) - the year of their first policy.

It is important to keep in mind that because all of these variables were created based on existing ones, we won't be looking at filling their missing values, as they will be fixed automatically.

D. Detecting and treating non available values

It's not possible to deliver a good analysis with a database that contains too many missing values. With knowledge on their amounts and their types, we had to decide what kind of strategy to use to fix these problems. There are uncountable manners to treat the missing values: removing the row that contains them, replacing them using the mean, the mode, a k regressor…

After analyzing the effects of the different methodologies, we kept the ones that don't affect our dataset negatively. These approaches were:

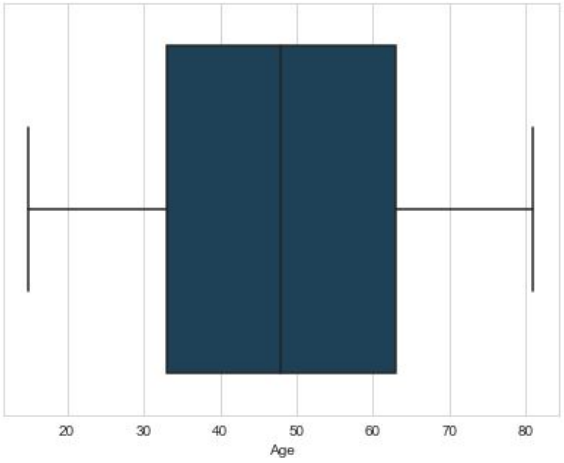
Category	Missing values treatment
Nominal variables	Replaced by the mode
Customer variables	Replaced by the median
Premium variables	Replaced by zeros

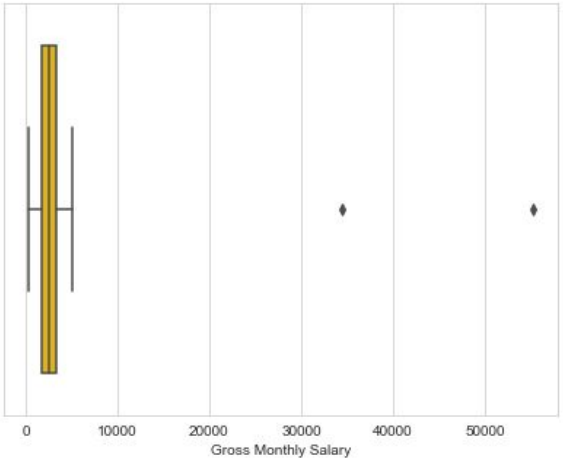
Now that we have a dataset free of missing values, let's move on to taking care of outliers.

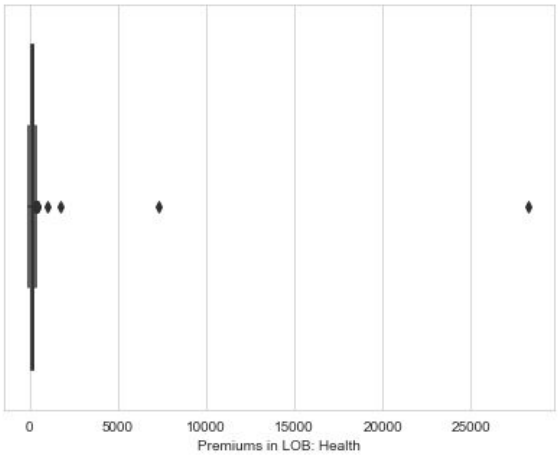
E. Detecting and treating outliers

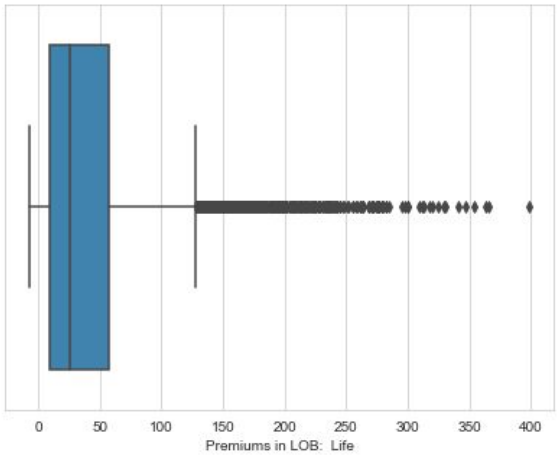
To find the outliers in our database, we are going to analyse the boxplots and the z-scores of each one of the variables.

Though we have plotted box plots associated for all of our features, here we present only 4 of them: Age, Gross Monthly Salary, Premiums in LOB: Health and Premiums in LOB: Life.









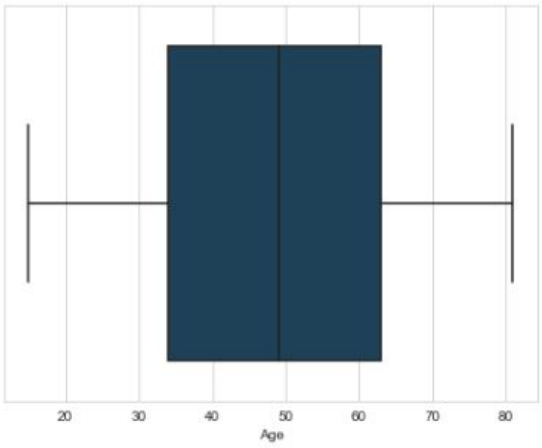
As we can see in the boxplots, there are outliers in almost every variable. Some of them are excepted: premiums depend on a lot of different factors, so the huge amount of outliers in Premiums in LOB: Life is reasonable.

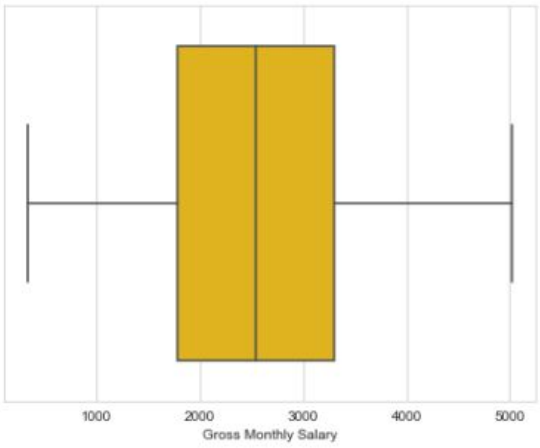
After that, we considered the logarithm of these variables and we checked the new boxplots. Unfortunately, that method didn't lead to any improvements in terms of number of outliers.

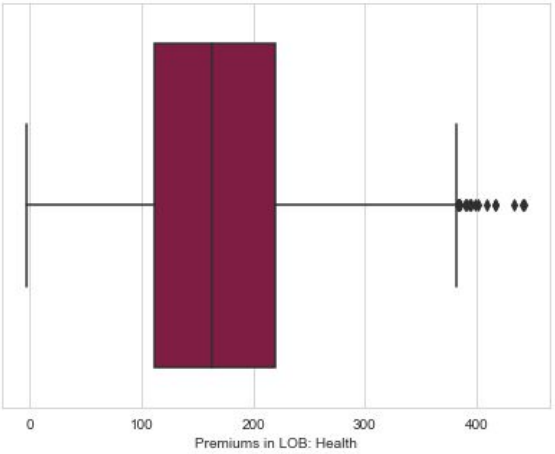
So, we tried a different approach, which is the Z-score. The original database contains 10264 rows. With the z-score segregation, 206 outliers were identified. After removing them, we obtained a new database that contains 10058 rows.

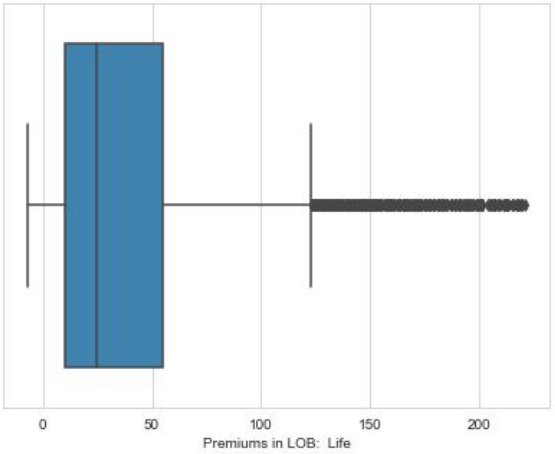
Database	Original	Outliers	Final
Number of rows	10264	206	10058

Let's see what our box plots look like after removing the identified outliers:









It’s clear that we still have values outside the box plots limits, but they will not be considered as outliers. Because the z-score method didn't identify them as ones, we will keep them in the data frame as we continue our analysis.

At the end of the clusters analysis, we will check if it’s feasible to add the outliers in the dataframe again.

3. Multivariate Analysis

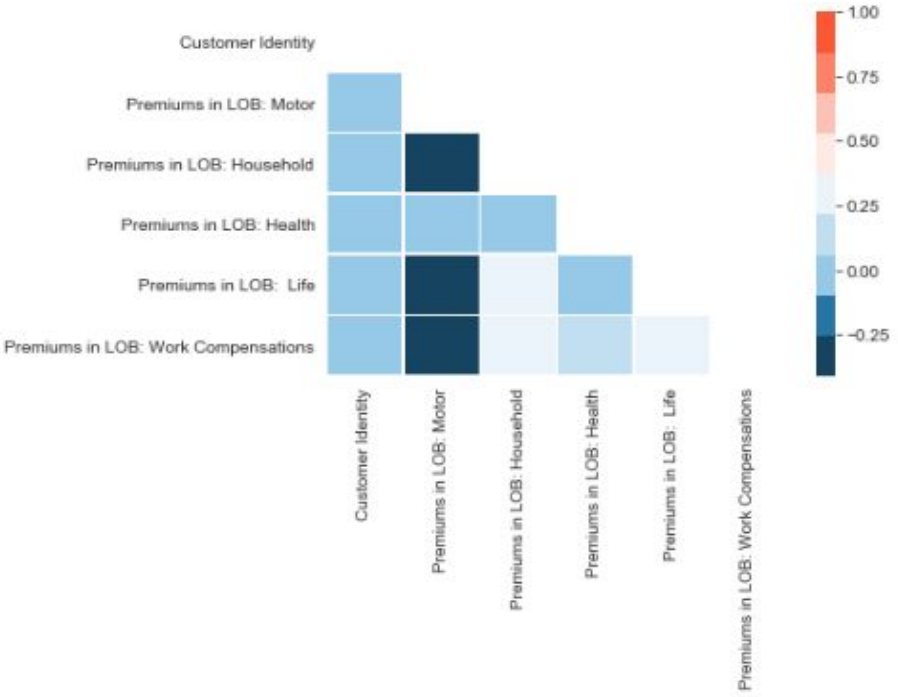
In this section, we will take a look at the correlations between the categorical and quantitative variables.

A. Quantitative variables

Here's the correlation map between the customer variables:

You can see that the correlation between "Claims Rate" and "Customer Monetary Value" is really high, at -0.99. We also obtained a high correlation between "Age" and "Gross Monthly Salary".

And for the Premium variables:



It's visible that the correlations between these variables don't go above 0.3, which implies that they are not very correlated among themselves.

B. Qualitative variables

For the qualitative variables, we will calculate the Contingency coefficient of each one of them. The Coefficient C is a hypothesis test based on the Chi-Square Statistics, where the alternative hypothesis is the independence between two categorical variables.

Let's take a look at the table with the statistics:

P-value	Has Children	Educational Degree	Geographic Living Area
Has Children	1	0,003	0,00001
Educational Degree	0,003	1	0,02
Geographic Living Area	0,00001	0,02	1

Considering the confidence value of 99%, we checked that all nominal variables are independent because all the p-values are lower than 0.01.

4. Clustering

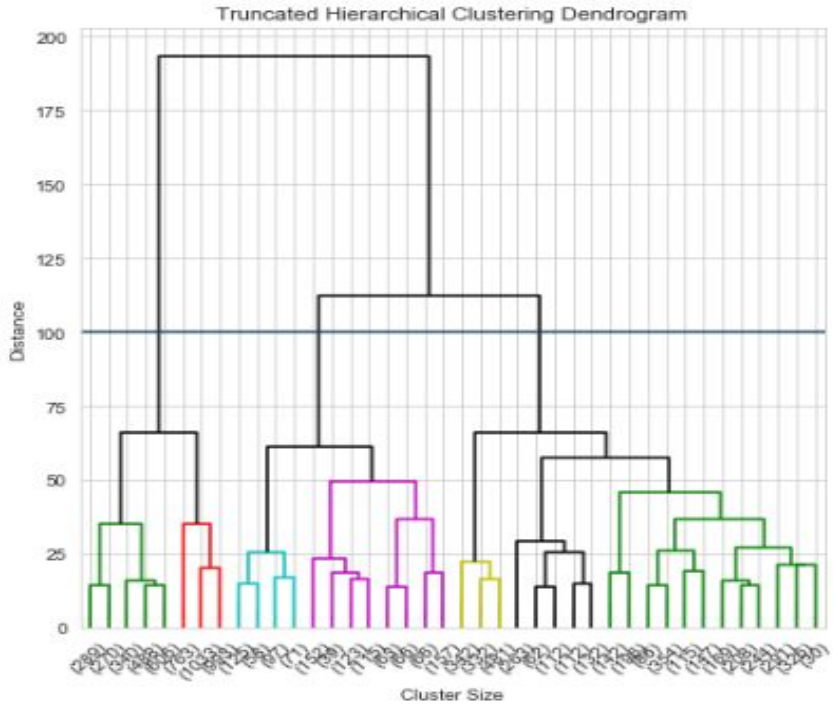
In this chapter will be focusing on clustering. At the end of this part, we are expecting to have different clusters for each one of the variable groups (premium, nominals and customers).

The techniques that will be used are: K-means, K-medoids, Self Organizing Maps, DBScan and Mean Shift. After we apply them, each approach will be complemented with a silhouette graph, dendrogram and/or a decision tree, depending on their compatibility.

A. Premium Variables

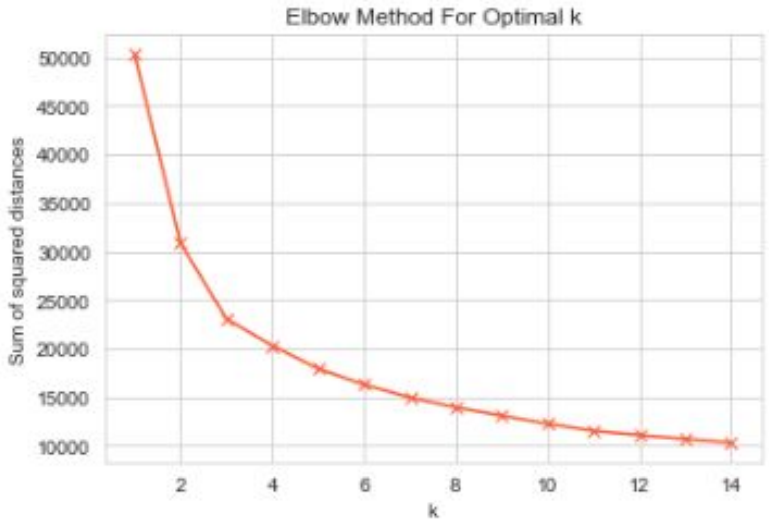
Since this is a group of quantitative continuous variables, we can apply every technique.

I. Dendrogram



The dendrogram clearly suggests the usage of 3 clusters to represent the premiums in our dataset.

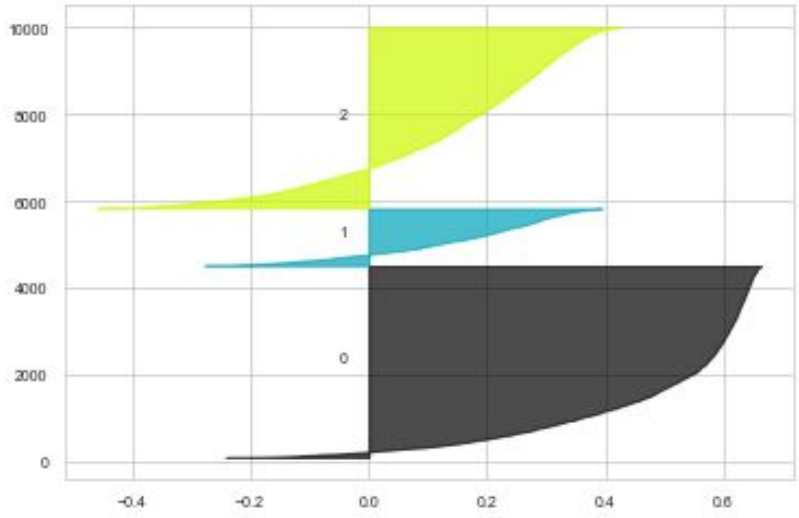
II. K-Means and K-medoids



As you can see by the elbow graph above, the “optimum point” is also at 3 clusters.

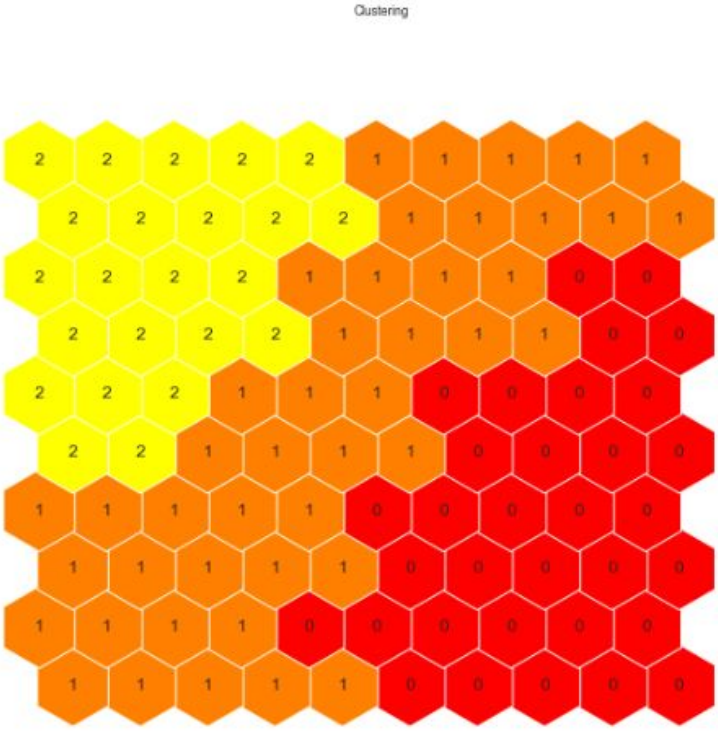
III. Silhouette graph

Let’s build a silhouette graph to check the behavior of these 3 clusters.



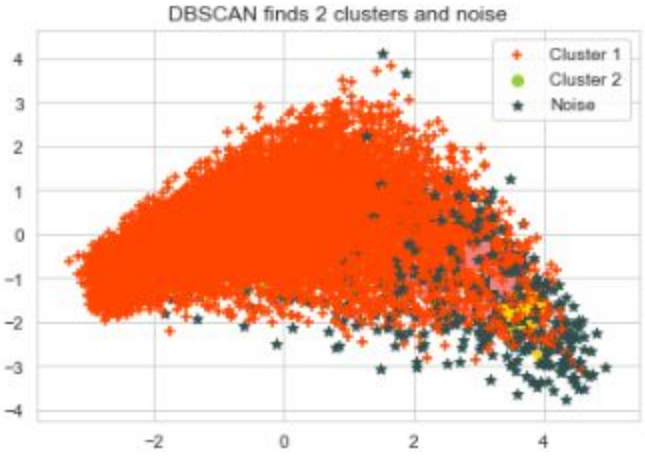
Though not well balanced, they do seem different.

IV. Self Organizing Maps



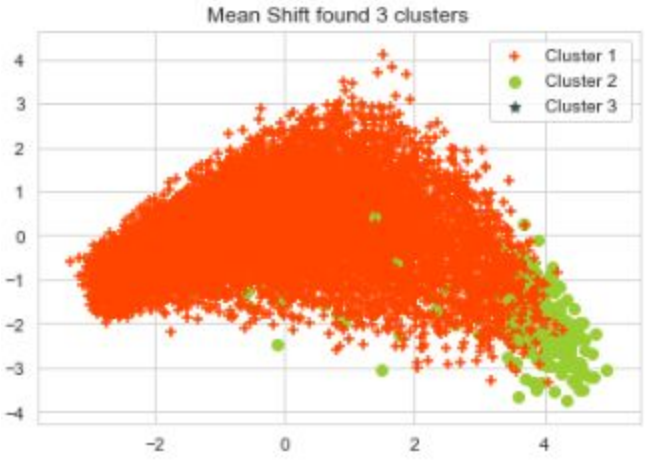
Once again, the used technique suggests 3 clusters. We can notice that the clusters are unbalanced compared to their sizes.

V. DBScan



Density-based spatial clustering for premiums shows some noise, and that cluster 1 is clearly dominant over the whole space. This is expected since our data is continuous.

VI. Mean Shift



Mean-shift is the process of assigning the data points to the clusters iteratively by shifting points towards the mode. Here, you can see that one condensed group dominates over the space and the others are spread all over the graph.

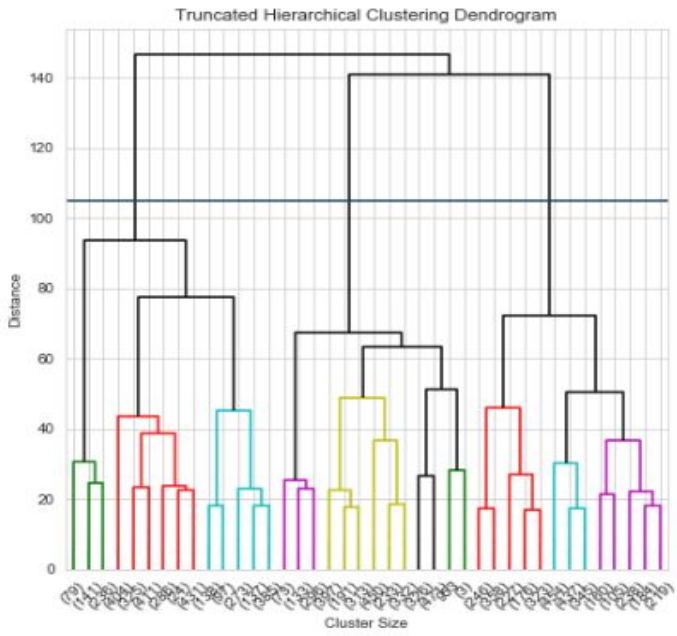
VII. Decision

After applying all the clustering methods, and comparing the efficiencies of the techniques used, we decided to identify the number of clusters based on hierarchical clustering along with k-means, so, three clusters. In the next section, we will confirm the accuracy of our decision

B. Customer Variable

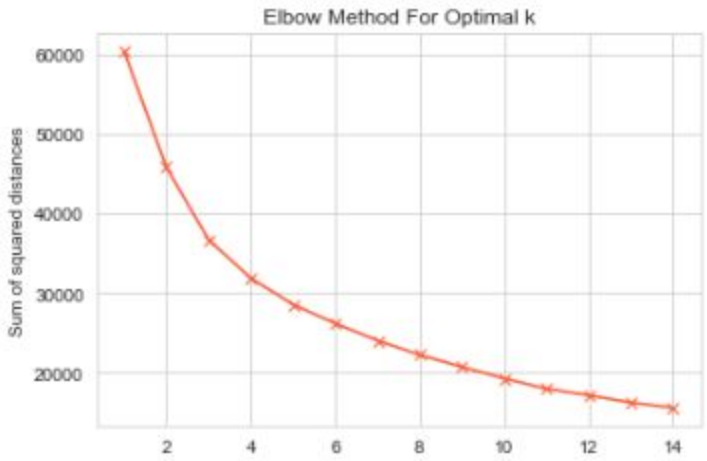
Again, this is a group of quantitative continuous variables, so we will apply every technique, just like we did before.

VIII. Dendrogram



Once again, our dendrogram suggests using 3 clusters.

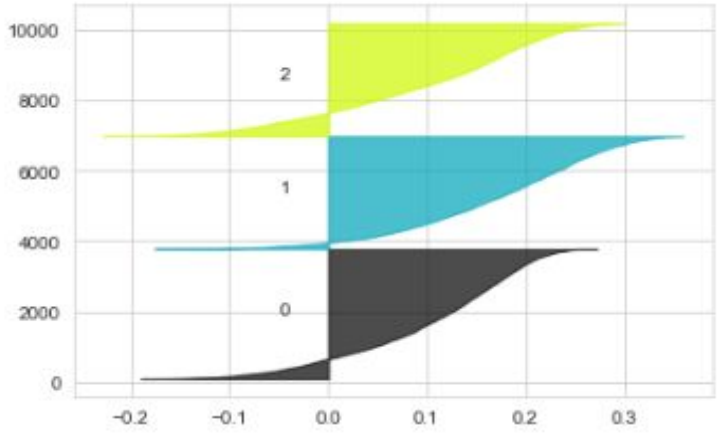
IX. K-Means and K-medoids



As you can see by the elbow graph above, the “optimum point” is also at 3 clusters.

X. Silhouette graph

Let’s build a silhouette graph to check the behavior of these 3 clusters.

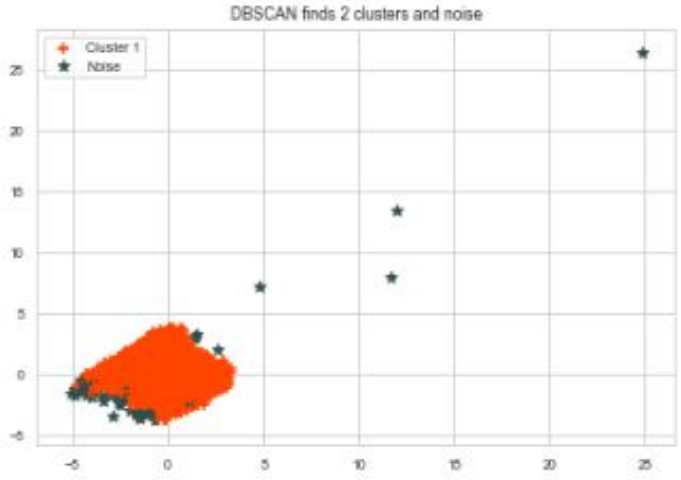


Unlike the silhouette graph from the premium variables, this one shows clusters that are more balanced in size.

XI. Self Organizing Maps

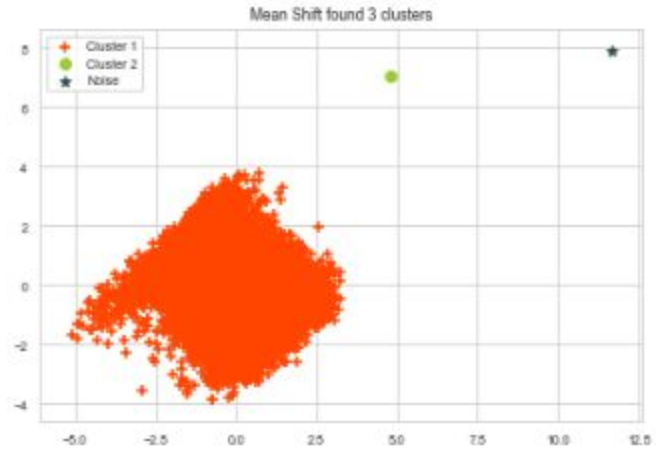
SOM for Customers shows 3 clusters, and this time, a little closer in terms of coherence.

XII. DBScan



Density-based spatial clustering for premiums shows some noise, and that cluster one is clearly dominant over the whole space. This is expected since our data is continuous.

XIII. Mean Shift



Mean-shift density is observed by one condensed group with some outliers.

XIV. Decision

Just like before, and basing ourselves in the same techniques, we will be keeping 3 clusters, whose accuracy we will study in the next chapter.

C. Nominal Variables

XV. K-Modes

```
[['2 - High School' '1.0' '1.0']
 ['3 - BSc/MSc' '4.0' '0.0']
 ['3 - BSc/MSc' '4.0' '1.0']]
```

After going through 50 iterations of the nominal dataset, we obtained 3 clusters as shown in the figure above.

XVI. Decision

In this case, considering that the variables in this categories are categorical, we chose to apply only K-modes, as it is mathematically feasible. And we will analyse and consider 3 final clusters.

5. Cluster analysis

In this section, we aim to develop the clusters and, in the end, to analyse them to build our marketing strategy.

A. Cluster Selection

To select clusters, we will compare the means of the active variables consecutively to the groups. It is possible to quantify the overall amplitude of the differences using the proportion of explained variance (square of the correlation ratio).

I. Cluster size

We were previously able to see that the clusters associated with the nominal and the customers variables are balanced, but the ones associated with the premium variables aren't; there's one that is very different in size. So we will remove it, rearrange it and remake the cluster analysis with the new number of clusters.

Here are the sizes of the clusters associated with the customer and the nominal variables:

Cluster / Group	Customer	Nominal
1	3676	5589
2	3194	2484
3	3188	1985

And the before and after sizes of the clusters associated with the premium variables:

Cluster / Size	Initial	Final
1	4160	5559
2	4158	4530
3	1771	-

II. Analysis between clusters

In order to make a decision on the final number of clusters, we will use the cross-table technique between the labels of each one of the variables. We expect that this method will lead to a smaller cluster number.

Here are the tables with the cluster frequency distributions:

Nominal / Customer	1	2	3
1	2183	1172	321
2	1382	388	1418
3	2024	924	246

Premium / Customer	1	2
1	3302	374
2	2499	6
3	1966	1228

Premium / Nominal	1	2
1	4130	1459
2	2137	347
3	1500	485

And, finally, the table of the frequencies of each one of the clusters:

Premium	1			2		
Nominal / Customer	**1**	**2**	**3**	**1**	**2**	**3**
1	1920	1110	272	263	62	49
2	1015	355	1129	376	33	289
3	1195	672	99	829	252	147

You can see that the first premium cluster contains a high concentration of clients in comparison to the second one. You can also observe that there are lots of clusters that contain small frequencies.

We will transfer these clusters to another, and if the meanings of the main variables don't change significantly, we will stick with that change.

III. Each Cluster Description

Let's move on to analyse each cluster, and to check the conditional means for each group of variables.

Here's the table with the conditional mean of each one the variables for the customer cluster:

Cluster / Var	Salary	Claims	Monetary	Client Spend	Fidelity	Age
Cluster 1	2204	0.3	503.5	770.3	29.9	42
Cluster 2	1736	0.92	38.1	755.4	30.2	34.2
Cluster 3	3470	0.78	131.5	723.3	29.7	65.4

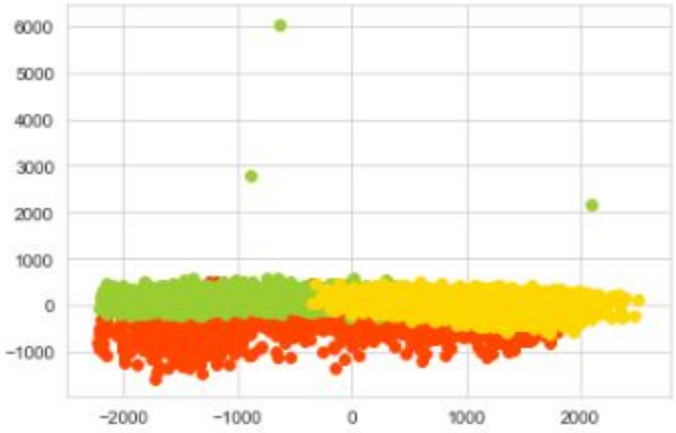
And the table with the square of the correlation ratio, or in other words, the proportion of variance explained for each variable:

Variable	Salary	Claims	Monetary	Client Spend	Fidelity	Age
% Var. Exp.	0.59	0.57	0.55	0.01	0.00001	0.62

The definition of the groups is – above all – dominated by the salary, the claim rate, the monetary value and the age of the customers. On the other hand, fidelity and client spendings are almost useless for the final model.

If you check the conditional mean table, the first cluster is characterized by middle-aged customers with low values in claims, high monetary value, and average salaries. The second cluster is defined by younger customers, with lower salaries than the ones from the first cluster, with high values in claims and a lower (though not much lower) monetary value. The third represents older customers, with higher salaries, and with claims and monetary values in between the other two.

When we combine the cluster analysis with factor analysis, we benefit from the data visualization to enhance the analysis as it allows to obtain a synthetic view of the data. And that's what we decided to do next:



The clusters aren't perfectly separable but you can see that they are not overlapping.

As for the premium clusters, here's a table with the conditional means:

Cluster / Var	Motor	House	Health	Life	Work
Cluster 1	401.4	85.8	132.9	16.9	16.8
Cluster 2	176.1	344.5	211.3	66.1	65.42

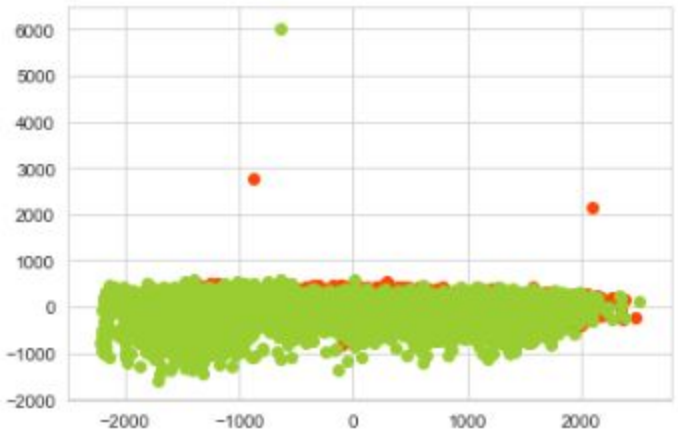
And the square of the correlation ratio or, in other words, the proportion of variance explained for each variable:

Variable	Motor	House	Health	Life	Work
% Var. Exp.	0.67	0.31	0.27	0.33	0.32

The definition of the groups is – above all – dominated by premiums in motor, but all the other premiums are almost equally important.

If you check the conditional mean table, the first cluster has customers with higher values in motor premiums, and the second cluster has customers with higher values in all the other premium categories.

As for the PCA graph for the premium variables:



You can see that one cluster is dominant, overlapping the other.

Finally, we can check the characteristics of each nominal cluster:

Nominal / Premium	Education Degree	Geographic Area	Has Children?	Centroids
Cluster 1	High School	Area 4	Yes	5589
Cluster 2	Bsc/Msc	Area 1	Yes	2484
Cluster 3	Bsc/Msc	Area 1	No	1985

It's noticeable that the first cluster has a higher distance to the other clusters (based on the centroid positions) and that the clusters 2 and 3 are closer, which is expected since the only difference between them is the variable "Has Children".

And the last cluster contains a higher diversity of labels on each variable.

To complement the analysis, let's check the frequency tables of each variable:

Cluster / Has Children?	Yes	No
Cluster 1	1431	3434
Cluster 2	636	1527
Cluster 3	805	1990

Cluster / Geographic	Area 1	Area 2	Area 3	Area 4
Cluster 1	1488	450	955	1972
Cluster 2	635	251	445	832
Cluster 3	792	283	560	1160

Cluster / Educational	Basic	High School	Bsc/Msc	PhD
Cluster 1	593	1638	2289	345
Cluster 2	231	765	1021	146
Cluster 3	302	3355	4662	680

As you can see, each one of the clusters has a different concentration of variables. For example, for geographic living area, the first cluster contains more customers from area 1, but the second cluster contains more clients living in area 4.

Now that we have analysed all of our clusters extensively, we will move on to reducing the eighteen of them to just five. To avoid extending the report, we decided to keep that process exclusively in our ipynb file. In it, we analysed the different statistics for each one of our clusters (mean, median, standard deviation, proportion, as well as variable frequencies).

The first criteria to merge clusters was their size, but we also took into consideration the behavior of some key variables that we find to be important for the marketing analysis. In the end, we have 5 clusters. The table below contains their size. In the next section, we will look at each one of them more closely.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Size	2702	2370	1195	1876	1924

6. Cluster descriptions for marketing approaches

This chapter analysing our clusters from a practical standpoint, that we hope can be useful when delivering our conclusions to the marketing team.

Final Cluster 1

The first cluster contains customers with ages between 36 to 57, with high salaries (between $1900 and $2500), with high school education, and who predominantly have children. This cluster is interesting since the client spends are higher (especially because of the premiums in motor).

Final Cluster 2

The second cluster contains younger clients (with less than 35 years old), with low gross month salaries (an average of $1736) and who predominantly don't have children. The client spend is low and the majority of this cluster only has high school education too.

Final Cluster 3

This cluster, just like the first two, got high values for premiums in motors. The average age of this cluster is 65 years old, and most of the customers have children (most likely adults). That could be taken as an advantage, since they might give us access to potential new customers. Also, this cluster is the one that contains the higher average gross monthly salaries, which might be related to the fact that their level of education is higher than the ones in the previous clusters: most of them got Bsc/Msc and some of them even PhDs. Unfortunately for us, they got the lowest average in client spending.

Final Cluster 4

The fourth cluster also contains younger clients (on average, 34 years old) but most of them have a Bsc/Msc, receive medium gross month salaries (on average, $2445) and have children. Just like before, the client spendings are lower.

Final Cluster 5

This Cluster is the most different, because unlike the others, these clients have low values on premiums in motor. Another characteristic of this cluster is that it only contains people older than 55 years old, who completed their education in high school, and who have a lower gross monthly salary (on average, $2204).

7. Outlier Analysis

Finally, we will analyse the outliers of each group of variables to see it's feasible to add them again to the final database.

Let's check the conditional mean of each one of the outliers from the customer clusters:

Cluster / Var	Salary	Claims	Monetary	Age	Fidelity
Cluster 1	1088	1.23	-92	20	29
Cluster 2	5702	8.48	-4549	62	30
Cluster 3	1086	256	-165680	27	26

And the square of the correlation ratio or, in other words, the proportion of variance explained for each variable:

Variable	Salary	Claims	Monetary	Age	Fidelity
% Var.	0.15	0.77	0.73	0.87	0.003

The conditional mean of each one the outliers from the premium clusters:

Cluster/Var	Motor	House	Health	Life	Work
Cluster 1	60	420	122	257	70
Cluster 2	471	837	494	58	238

And the variance explained by each one the variables:

Variable	Motor	House	Health	Life	Work
% Var. Exp.	0.03	0.01	0.008	0.83	0.21

As we expected, the results don't show the reality of the dataset. Because of that, for the customer variables, it's recommended to treat them as a different class and we are not going to reintroduce them in the dataset.

8. Conclusions

After analysing, cleaning, and applying clustering techniques to the provided dataset, we are now ready to take our conclusions to the Marketing Department of our insurance company.

Though we have obtained five clusters, and will obviously explain all their specificities to the marketing team, we have to keep in mind that the end goal for the company is to generate the highest amount of profit possible. We'll suggest that customers with higher salaries, client spedings, and monetary values, and with lower claims rates should be prioritized.

One of the conclusions we were able to make is that we have a lot of clients with insurance plans for motor, but not that much money is being spent on the other available premiums. We might want to use that to our advantage and target existing clients, in the hopes of enrolling them in new premium plans.

As we share our results with the team, we will make sure to suggest them to target potential customers with characteristics similar to the ones from cluster **1** and **3**. Though one would expect to see older clients bring more loss to an insurance company, as they are more prone to accidents and health problems, our studies actually show that with their higher education, their salaries are higher, and so are their spendings. Customers represented by cluster 4 should also be paid attention to, since they seem to have space to grow, as they are younger and have higher salaries.

Clusters 2 and 5 seem a little dangerous to bet in, as they already show high claims rate, and a low gross monthly salary.

Just like we mentioned before, we will deliver all the findings to the marketing team, as we are sure their knowledge will be much deeper than ours when it comes to what to do next.

References

- Simmons, B. (2014, April 29). *Smart Vision Europe. What is the CRISP-DM methodology?*. https://www.sv-europe.com/crisp-dm-methodology/
- Rakotomalala, R. *Interpreting cluster analysis results* - Université Lumière Lyon 2.
- R. Madhuri, M. Ramakrishna Murty, J. V. R. Murthy, P. V. G. D. Prasad Reddy, Suresh C. Satapathy (2014). *Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms.*

> **Code available in:** https://github.com/GRaviSantos79/DataMining