NOVA IMS
Information
Management
School

2022

DATA MINING - PROJECT REPORT

GROUP CA
Helena Duarte Morais - 20210636
Pauline Richard - 20211019
Mohamed Ali Felfel - 20211322

INDEX

Abstract

This project was made to provide a segmentation model for an insurance company in Portugal. Working as data scientists for the company, our aim was to define the different customer's profiles based on similar behaviours, so that afterwards we could develop a more targeted and efficient marketing strategy.

This study is based on the ABT (Analytical Based Data), containing data regarding 10 296 customers, from the first year they subscribed until the current year - 2016. Our final goal was to perform an accurate cluster analysis. We started by extracting the observations. Then, we checked their quality, to make sure they are efficient for the application in the different clustering techniques. Finally, after cleaning and transforming the data, we worked on several segmentation models, and we came to the conclusion that the best result was obtained with a total of 3 clusters, built using the hierarchical clustering with 50 k-means method.

Keywords: Segmentation, Clusters, Profilers, Data preprocessing

Introduction

As an insurance company, it is crucial to understand the different types of customers and their needs, as well as the value they add to the company. As it will be shown in the following analysis, one of the biggest flaws of this company is the lack of new customers in its database. Hence, in order to help the marketing team to study and to develop new campaigns and more accurate strategies, it is necessary to distinguish the different profiles of customers that the company has, so that it can understand the preferences of the different profiles and identify the ones where the company should invest more. Given that, the ultimate purpose is to know which customers have higher values of premiums for a certain type of insurance and what the insurance company can do and improve to meet such demands.

Methodology

A plan was set to achieve the business objectives. The project started with understanding the metadata by the project description, which explained the features we had available to describe each customer. Some of the titles were self-explanatory, others had a description long enough for us to understand the thirteen variables under study. Then we moved to the next phase - data exploration - where the information is submitted to descriptive analysis, while assisted with the visualization of the appropriate graphics. Afterwards, it was performed the modifying phase, where all the necessary transformations on data were carried out. Finally, it was necessary to adopt the most adequate clustering method to identify the different groups of customers, evaluating the accuracy and usefulness of the different number of clusters and clustering methods.

1. Data Exploring and Analysis

Before we can run a cluster algorithm, we need to understand our data and try to get as many insights as possible. In this first part, we are going to check the quality of our data, the relationships and patterns between the variables, and to spot anomalies and test hypotheses using visualization tools.

1.1 Data and Variables Description

Our Train dataset is composed by 10 296 rows and 14 columns. The types of variables are float and objects.

Observing the summary statistics of the data in Figure 1, it's understandable that there are missing values in almost every variable. It is also visible the high dispersion of the data comparing the minimum, the maximum and the median values of the variables, which indicates the existence of possible outliers. In the birth year it appears that the minimum value is 1028, which is considered a strange value. In the first policy year, the maximum value is 53784, which is a value that does not make sense. After analysing all the values for "FirstPolYear", it was observed that, besides this strange value, the second-highest one is 1998. This shows that the company does not have new clients in a really long time.

Considering the variables that describe the premiums in the different insurances, it is possible to see that in all of them there are negative values, which mean that in 2016 the company had to pay reversals(cancellation) to some customers in the previous year.

Moreover, many variables share the same type, scale, and characteristics, and we can imagine we will have some strongly correlated variables. We can take as an example the Premium variables, such as "PremLife" and "PremMotor".

1.2. Assessing Data Quality

The quality of our Train dataset is going to have an important impact on the accuracy and efficiency of our model, so we will have a deeper look into this.

1.2.1 Checking for missing values and duplicates

First, we checked for missing and strange values such as (!, $, %, ?, *, +, _, @, €), which can be the result of mistakes or failure to record the data. After using *train.isna* and checking for the duplicates, we concluded that we had a few missing values in the dataset, but no duplicates. This is going to be corrected in the preprocessing part.

1.2.2 Checking for outliers

The next step was to check if the dataset contains any outliers. Our first approach was to study the dataset descriptive table. By analysing the values of the minimum, mean, and maximum of each variable, we noticed that, while there is almost no difference between the mean, the min values of the different features suggest that there are no low outliers, except for “CustMonVal” and “BirthYear”. The mean and max show huge differences in values for most of our features, which indicates that we have a problem of extreme values for all the variables.

Our second approach was to build boxplots for all the variables, as we can see on Figure 3. Commonly, we consider as an outlier a data point that is numerically distant from the rest of the data, here defined when it is more than 1.5 IQR above the third quartile or below the first quartile, with IQR= Q3-Q1. Looking at the outcome of our box plots, we could see all of our variables seem to contain outliers with extreme values above the upper quartile, except for “CustMonVal” and “BirthYear”. The latter shows only one outlier on the lower quartile; we can make the assumption that it’s most likely an incorrect value that we will get rid of during the coherence check. Finally, we were able to confirm the existence of extreme values using the IQR method and the z-score.

1.2.3 Checking for variables’ correlations

Secondly, we had a look at the correlation between our variables using a heatmap (Figure 4). We observed that some variables are highly negatively correlated, for instance:

- MonthSal and BirthYear (-0,93)
- ClaimsRate and CustMonVal (-0,97)

We can also conclude that our assumption on the Premium’s variables was correct, since all of them are correlated to the other Premium ‘PremMotor’ (around -0,7).

1.3. Data Visualization

On this step, we printed a few graphics to identify patterns in the different features. We checked the pairwise relationships between the numerical variables and printed histograms to see the values' distribution of the categorical variables.

For the categorical variables, looking at the histograms of Figure 4, we found out that most of the clients on the database have a bachelor/master or high school educational degree, and that the number of customers with children is more than twice as big as the number of people without children.

From the numerical variable’s histograms, We could not gather many conclusions due to the existence of outliers. However, looking at the box plots of Figure 5, besides the outliers’ information, we can observe additional details. We can see by comparing the interquartile ranges, that the length of all the box plots is very small, showing a small dispersion of the data. We can also see that most of the variables are showing a positively skewed distribution, with the median closer to the bottom of the box. It means that the average of all the values (mean) is greater than the middle values of the data (median), and can be linked with the observations we did of the upper outliers.

2. Preparing the database

During this phase, we will be manipulating and transforming data elements into a useful and proper dataset to match the needs of the different algorithms we will run.

2.1 Coherence Check

For the first step of preprocessing our variables, we decided to run a coherence check. We wanted to make sure that the data is logically consistent and can be reliably combined for analysis. During our analysis, we noticed that some values do not make sense. In total, three coherence check have been considered:

Coherence Check 1: we checked if the variables are making sense with the timeframe we are working with. First, the data has been collected until the current year - 2016, so we can drop all rows that contain 'FirstPolYear' bigger than 2016 (1 output). With the same logic, we must drop all BirthYear that are bigger than 2016 (meaning the customer is not born) or smaller than 1900 (that would mean someone is older than 122 years, which is very unlikely) (1 output). In total 2 rows were dropped.

Coherence Check 2: it corrects what we observed - that almost 2000 rows in our dataset indicate that the "FirstPolYear" is lower than the "BirthYear". It does not make sense, since a client cannot be a customer before even being born. These rows represent almost 20 % of our data, so we can't drop this much since we will lose a lot of information. So, to fix this error, Our team agreed to make the following assumption: when the "FirstPolyear" is smaller than the "BirthYear", we are going to assume that BirthYear and FirstPolYear has been switched by mistake when the data was collected, and they need to be switched back.

Coherence Check 3: we checked the coherence of the Age of our clients, and the information we have about them. For example, We checked that no client was earning a salary, while having less than 16 years old, since it's the legal age to work in Portugal. With the same logic, our client cannot be less than 16 and have a Bachelor, Master or even a PhD. We got no output for either check.

To conclude, at the end of checking the coherence, we only dropped 2 rows, which gives us a new shape for our dataset of (10294, 14).

2.2 Dealing with missing values

On this step, we dealt with the missing values that we found in some variables. Those values were: BirthYear (14), MonthSal (34), Children (13), PremMotor (34), PremHealth (43), PremLife (104) and PremWork (85).

Regarding the Premium-related variables, we decided to see if any of them had the value 0 in them, which they did not. Based on this observation, we figured that the missing values could come from premium values equal to zero, and decided to replace the missing values of these four variables with "0". Then, to fill the missing values on the rest of metric features (numeric variables) we used the

K-Nearest Neighbours (KNN imputer) method. The missing values were replaced with a uniform average of the non-null values of their 5 nearest neighbours on the data set. We used this method so that we could get a more accurate imputing of values than the one we would get by using a constant value. Finally, for the non-metric features (categorical variables), we chose to replace the missing values with the most frequent value for that variable – the mode. Hence, the missing values in "EducDeg" were replaced by "3 – BSc/MSc" and the missing values in "GeoLivArea" were replaced by "4.0".

2.3 Dealing with Outliers

As seen in the first part, we have outliers to deal with. However, it is important to mention that determining whether a value is an outlier that should be removed or not is very subjective. And while there are certainly valid reasons for throwing away outliers if they are the result of a computer glitch or a human error, eliminating every extreme value is not always a good idea. In our case, for instance, some outliers are unusual values and impossible to occur, like the one in "FirstPolYear". Nevertheless, some values can present significant information about the data, such as the ones in the column of "PremLife", where they could be considered only as outliers while looking at the box plots. This being said, we tried different options to reduce or remove the outliers:

- The IQR method: not the best for our data set, since it was removing a too large percentage of our dataset (more than 3%);
- The Z-score method: could not be used either, since we are mostly dealing with a positively skewed distribution;
- The filter method: the final method we chose.

We decided to use the filter method. We printed box plots and histograms so that we could better perceive the density distribution of the different numerical variables. Based on the inspection of these graphics, we defined individual conditions to drop values above and/or below specific thresholds (Figure 6 part a). The percentage of data we kept after removing the outliers was 0.9795 of the original data set. We can see an example in Figure 6 regarding the variable 'PremLife'.

2.4 Fixing Wrong Types

As we saw in the first part, all our data types are float, except for Education Degree. We corrected it by factorizing the variable and changing the outputs to: 0.0, 1.0, 2.0 and 3.0, each matching an education level, from the most basic one to the highest.

2.5 Feature Engineering

In total, we created 6 new variables:

- *"Annual_profit*" is based on the sum of all the different premium types of the insurance company, so that we could have a general perspective of the overall total values of each customer's premiums.

- We replaced the variable "BirthYear" with the variable *"Age"*, based on the difference between 2016 (the year of our dataset) and "BirthYear" of the client.

- "*Custumer_lifetime*" is based on the difference between 2016 and the first policy year, to measure how long the person has been a customer. The "FirstPolYear" variable has been dropped after, to avoid having the same information twice.

- "*Yearly_Salary*" is based on 12 times the monthly salary value. Therefore, we dropped the "MonthSal" variable.

- "*gen*" is a categorical variable based on which generation the customer belongs to. To create this variable, we created a set of different conditions based on "age".

- Finally, *"ClientValue"* is a binary variable, based on the "CustMonVal", that indicates if a customer is valuable to the company (=1 and it means "CustMonVal" is positive), or he's not valuable and cost money to the company (=0 and it means "CustMonVal" is negative)

At the end of this process, we redefined our metric and non-metric features. The non-metric features of our dataset are now ["EducDeg", "Children", "GeoLivArea", "gen", "ClientValue"]. The rest of the variables are metric features.

2.6. Feature selection

First, to see if we still had redundant or irrelevant information, we decided to analyse the correlations between the numeric variables we had at this point. To do so, we printed Spearman's correlation matrix (Figure 7). The highest correlations we found were between "Annual_profit" and "PremHousehold" (0,98), "CMV" and "ClaimsRate" (-0.95) and "Age" and "Salary" (0.9). We decided to drop "ClaimsRate" and "Age" to avoid redundancy due to the high correlation with other variables. Moreover, the age of the customer can also be analysed by using the categorical variable "gen". On the other hand, "PremHousehold" being one of the five products of the company, it can be considered as an important feature with key information and we decided to keep it.

In a second part, we analysed the relevancy of our categorical variables with the 'ClientValue' variable, using Pandas' cross tab function and visualizing the summary with a bar chart (Figure 8). Looking at the cross tab between the 4 categorical variables and "ClientValue", we were able to conclude that, except for "gen", the 3 others don't show a major impact on the value of the client for the company. However, "Children" and "GeoLivArea" seem to be the less relevant, with practically the same result for all values, so we will not keep them.

To conclude, after creating new variables and selecting the most relevant ones, the final shape of our dataset is (10085, 12), with a percentage of 97,85% of the data kept from the original dataset.

2.7 Normalizing the Data

Normalizing the data is an important step in data preparation, especially if we are going to deal with an algorithm that relies on the distance between the different observations. For this dataset, we used MinMaxScaler.

2.8 One-hot encoding

Since most models cannot work with categorical variables, this step becomes necessary. One-hot encoding is the process of changing categorical data into numerical ones. Given that all our categorical variables are not ordinal, we decided to turn all of them into dummy values, meaning binary variables with values of either 0 or 1. However, this method creates a column for each option.

3. Performing the cluster algorithm

For this last step, our goal was to cluster our data into groups of customers with similar behavior, so that afterwards we could develop a more appropriate marketing strategy for each cluster.

3.1. Cluster Methods used and Thinking Process

To ensure that our clustering was made in the best way, we performed several clustering algorithms before choosing the one we considered more appropriate. Given our data set, we performed the following clustering methods: hierarchical clustering, k-means clustering (k-means with hierarchical clustering, k-means with raw data, k-means with encoded data, hierarchical clustering with 50 k-means clustering), self-organizing maps, density based clustering (mean-shift clustering, density-based spatial clustering, Gaussian mixture model).

Details of all the different cluster algorithms used can be found in ANNEXE 3. The final method that gave us the best result was the Hierarchical Clustering with 50 kmeans, and we will interpret the results in the following section. Looking at the Ward’s Dendrogram (Figure 9), we can conclude the optimum number of clusters is three. Dividing into four clusters lead us to two relevant clusters and two really similar ones (the red and the purple on the graph), so one of them would be irrelevant, and by dividing into only 2 clusters we would miss information on our customer segmentation.

3.2. Analysis of the Final Cluster

We will now further explore and analyze the previously chosen clusters so that afterwards we can develop our marketing strategy for the insurance company. All the visualization we are using to analyse the three clusters below can be found in ANNEXE 2.

3.2.1. Cluster 0

Cluster 0 can be considered as the “moderate” one, and it is the one that contains a larger amount of data (25 out of the 50 k-means clusters our hierarchical clustering started with, belong to this cluster).

Those customer’s annual profit and loyalty towards the company (“Customer_lifetime”) show moderate values – neither too high nor too low – compared with the other two clusters. Their CMV appears to be very similar to the one of cluster 1 and below cluster 2, and their annual salary is also close to the one in cluster 1 but, on the contrary, significantly higher than the one in cluster 2.

Regarding generational (“gen”) distribution, we can see that all generations (except for generation Z, that only corresponds to a small percentage of our data base) are represented in a balanced way. For what concerns the customers’ educational degree (“EducDeg”), it shows, once again, the most homogenous distribution out of the three clusters, showing a low value only for PhD graduates (the degree that only corresponds to a small percentage of our dataset values).

Finally, the customers belonging to this cluster show the highest interest in the health-related insurance premiums (“PremHealth”) and appear to have a moderate interest in all the remaining premium categories offered by the insurance company.

3.2.2. CLuster 2

Cluster 1 contains the second-largest number of customers from our data set (20 out of the 50 k-means clusters our hierarchical clustering started with).

This is the group that displays the highest customer lifetime and also the highest annual salary. As we mentioned before, their CMV has a value very similar to the one for cluster 0. Despite the higher salaries and loyalty, we can observe that the clients from this cluster are the ones that spent less on the insurance company (lowest “Annual_profit”) and also have the lowest CMV.

The generational distribution on this cluster is mainly concentrated on baby boomers and generation X, which means that most of these customers have ages between 50 and 90 years old. This cluster is the one displaying a higher educational level, with most of its customers having a Bachelor or Master degree, and being also the cluster to which most of the PhD graduates from our data set belong to.

These customers show the highest interest on the vehicle-related premiums (“PremMotor”) and the lowest interest regarding all the other premiums offered by the insurance company. They also show some interest in “PremHealth”, making this the premium in which all of our clusters are interested.

3.2.3. Cluster 3

Finally, cluster 2 is the smallest one, and is significantly smaller than the previous two (it only accounts for 5 of the 50 k-means groups we started with).

Despite being the smallest cluster, this is the cluster that gives the highest annual profit to the insurance company, and it is also the cluster with highest CMV. On the other hand, these customers have the lowest annual salary and the lowest lifetime inside the company.

This cluster is more focused on the middle-age and younger generations. There is a high focus on generation Y, followed by generation X, and it is also the cluster that most of the gen Z clients belong to. Most of its elements have an educational degree of bachelor/master or below (high school).

To conclude, the premium in which this group shows the higher interest is the one related to household ("PremHousehold"), but it is also the cluster with the higher interest in life ("PremLife") and work ("PremWork") premiums. They also show a moderate interest in health premiums.

4. Marketing Strategy

Our marketing strategy aims to develop a more appropriate approach for each of the different types of premiums the insurance company offers to their clients. To do so, we will base on the profiles of the clients we developed through our clustering analysis. We also took into consideration the importance of focusing especially on Health and Motor related products, because these are the products that appeal the most to the customers with the highest engagement value for the company (clusters 0 and 1).

4.1. Health premiums

Health premiums should be able to appeal to all kinds of customers, regardless of their age or salary, since this is the type of premium that all of our groups of clients are interested in. This means that the marketing campaign should cover all of the main communication branches, so that it reaches both younger and older generations. Being health-related, advertisements with a more emotional approach can make the customers "connect" with the company on an emotional level, which can attract new clients and maintain/increase the loyalty of the customers the company already has.

4.2. Vehicle premiums

Motor premiums seem to be more appealing to people in a more advanced life stage and that receive a high salary (the customers from cluster 1). The marketing campaign should take this into account and aim for something classy (to reflect the high salary) and that calls on ethical values (something that the clients within this group usually identifies with) and without being too extravagant. The ethical values can help to reflect confidence and reliance on the company, characteristics that are usually valued by customers when they are choosing an insurance company.

The advertisement should be displayed in more traditional branches, such as the radio, magazine a newspaper ads and television commercials, for example.

4.3. Household, Work and Life premiums

The marketing strategy for these three types of insurance products should be mainly focused on the same age range group – people from 20 to 45 years old. Hence, we are looking for advertising that easily stands out and triggers emotions from our target audience. This can be something funny, for example, especially for household and work, but it can also be something that triggers fear in a way that reminds the customer of why they need this kind of insurance. At the same time, the advertisements should also share a feeling of comfort and trust that makes the customer feel that they can count on the company to "be there" for them.

For this profile we can use a more digital focus advertising, focusing on social media, television, and trying to develop partnerships with other companies the younger generations identify with, so that we increase the number of clients and the monetary value of the younger generations.

Even though we are targeting different customers' profiles, it is important that all the strategies have two things in common: they must be *ethical* – to reflect the values the company stands for and in which the customers can identify themselves – and to *inspire trust*.

Conclusion

To conclude, after working on the customer segmentation of the Insurance company, we can clearly define three distinct profiles. It is essential for every company to know their customers and be able to offer the adequate product for their needs. Indeed, as we could see during the Exploration phase, the newest customer to the database subscribed in 1998, meaning the company is struggling to attract new customers.

As a result of our cluster analysis, we can advise the company to invest more in their marketing strategy and to adapt the advertisements to the target they are aiming for. For example, clients from different generations (older or younger) and with a different education level will not earn the same salary, and they will not look for the same things from the Insurance Company. That's why our work was important to identify in which product each profile was interested in and after advise the right marketing tools to reach the right customers. A more targeted marketing campaign will increase the profit of the company (by adding new customers to their database) but will also reduce its cost by aiming a smaller audience, but in a more efficient way.

REFERENCES

Tallón-Ballesteros, A. J., & Riquelme, J. C. (2014). Deleting or Keeping Outliers for Classifier Training?

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining.

Brownlee, J. (2021). accessed 2 December 2021, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

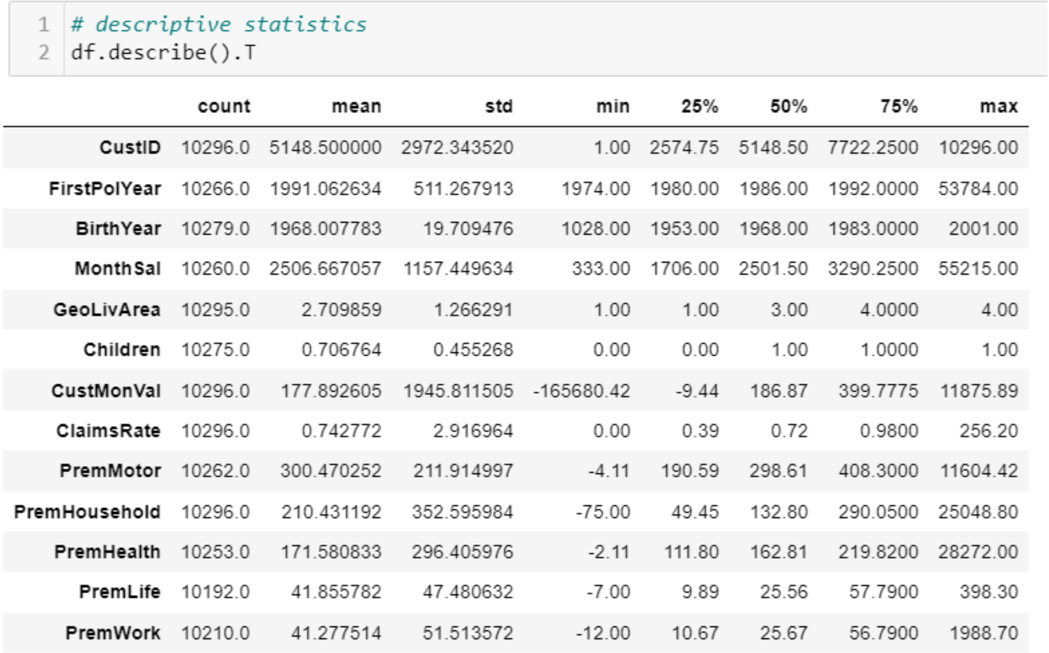
PennStateExtension Magazine (JUNE 16, 2016): Understanding Your Customers: How Demographics and Psychographics Can Help https://extension.psu.edu/understanding-your-customers-how-demographics-and-psychographics-can-help

Eric Hogson (2020/02): HOW TO IMPLEMENT AND UNDERSTAND CLUSTER PROFILING https://www.dotactiv.com/blog/how-to-implement-and-understand-cluster-profiling

DailyFreeExpress revue (2021/03/15): Generation name Explained https://dailyfreepress.com/2021/03/15/generation-names-explained/

ANNEXE 1: Figures

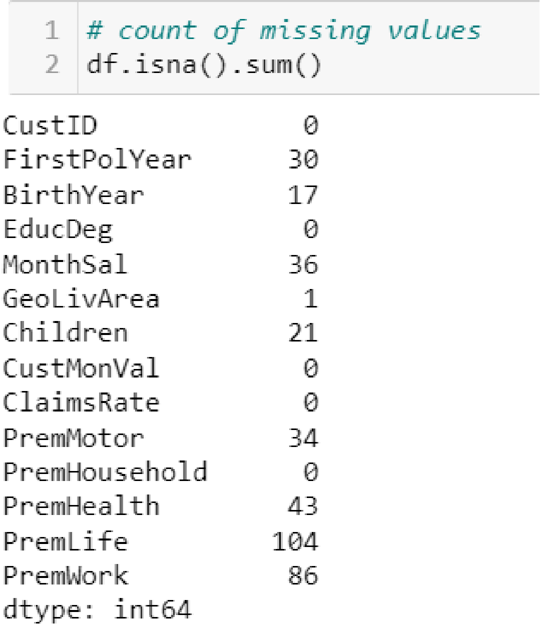
Figure 1: Summary statistics for each variable



```
# descriptive statistics
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
CustID	10296.0	5148.500000	2972.343520	1.00	2574.75	5148.50	7722.2500	10296.00
FirstPolYear	10266.0	1991.062634	511.267913	1974.00	1980.00	1986.00	1992.0000	53784.00
BirthYear	10279.0	1968.007783	19.709476	1028.00	1953.00	1968.00	1983.0000	2001.00
MonthSal	10260.0	2506.667057	1157.449634	333.00	1706.00	2501.50	3290.2500	55215.00
GeoLivArea	10295.0	2.709859	1.266291	1.00	1.00	3.00	4.0000	4.00
Children	10275.0	0.706764	0.455268	0.00	0.00	1.00	1.0000	1.00
CustMonVal	10296.0	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	0.742772	2.916964	0.00	0.39	0.72	0.9800	256.20
PremMotor	10262.0	300.470252	211.914997	-4.11	190.59	298.61	408.3000	11604.42
PremHousehold	10296.0	210.431192	352.595984	-75.00	49.45	132.80	290.0500	25048.80
PremHealth	10253.0	171.580833	296.405976	-2.11	111.80	162.81	219.8200	28272.00
PremLife	10192.0	41.855782	47.480632	-7.00	9.89	25.56	57.7900	398.30
PremWork	10210.0	41.277514	51.513572	-12.00	10.67	25.67	56.7900	1988.70

Figure 2: Checking for missing values



```
# count of missing values
df.isna().sum()
```

```
CustID            0
FirstPolYear     30
BirthYear        17
EducDeg           0
MonthSal         36
GeoLivArea        1
Children         21
CustMonVal        0
ClaimsRate        0
PremMotor        34
PremHousehold     0
PremHealth       43
PremLife        104
PremWork         86
dtype: int64
```

Figure 3: Box Plots of Numerical Variable

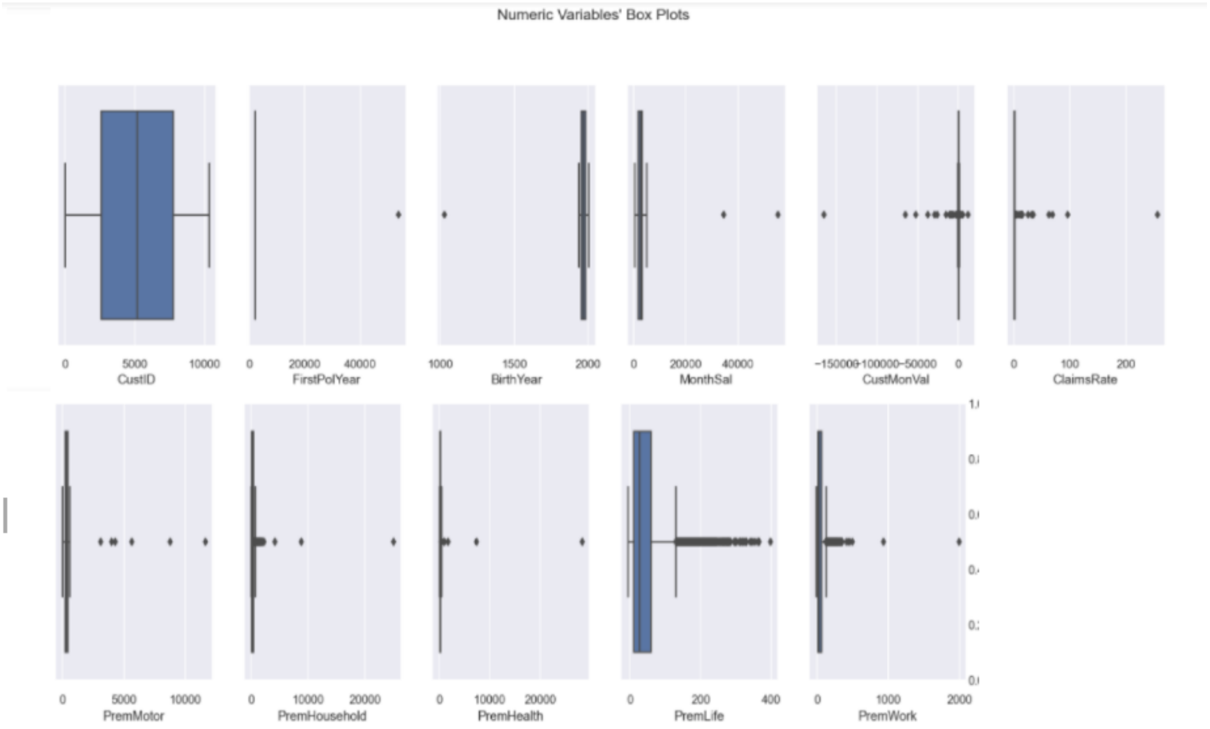


Figure 4: Heatmap of a Correlation Matrix

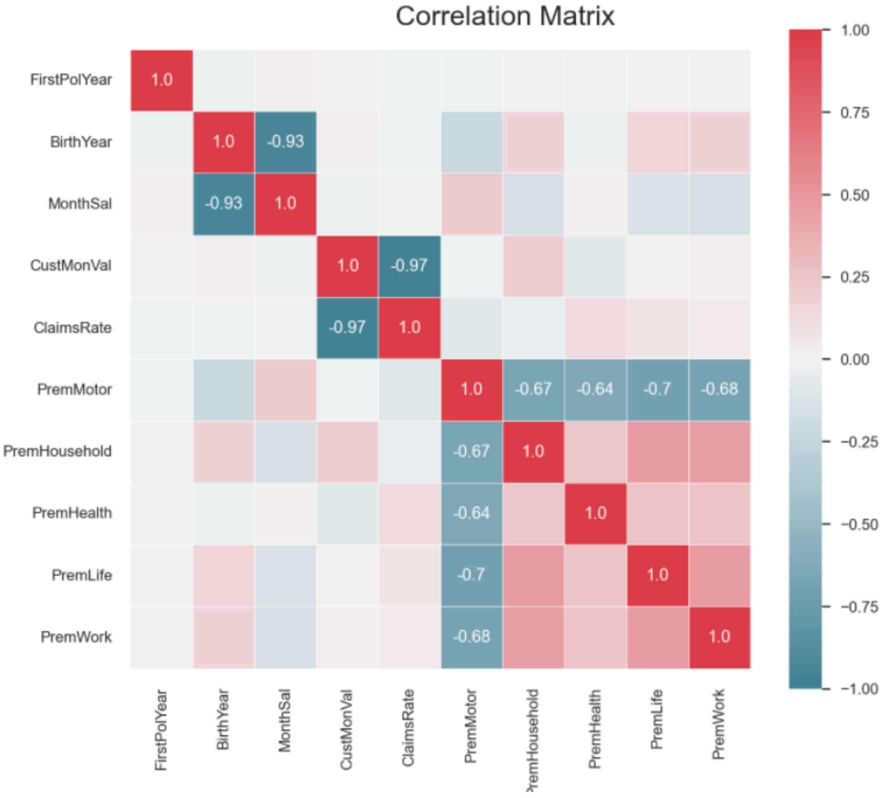


Figure 5: Categorical/ Low Cardinality Variables' Absolute Frequencies

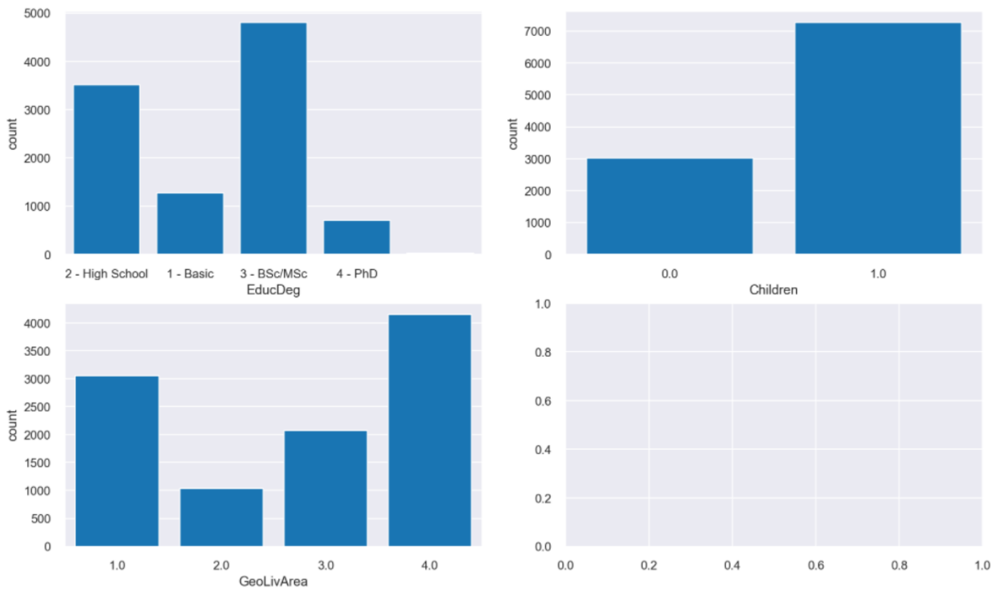
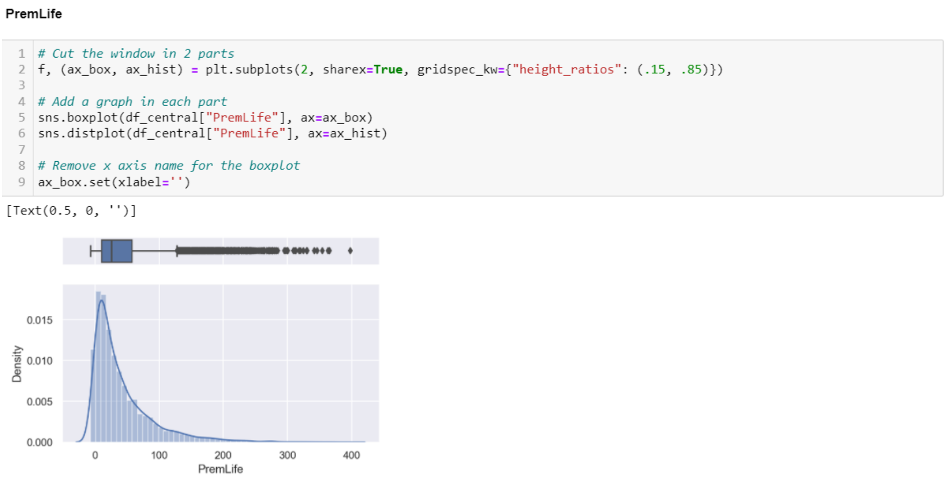


Figure 6 part a: Visualization of the outliers and distribution of 'PremLife'



PremLife

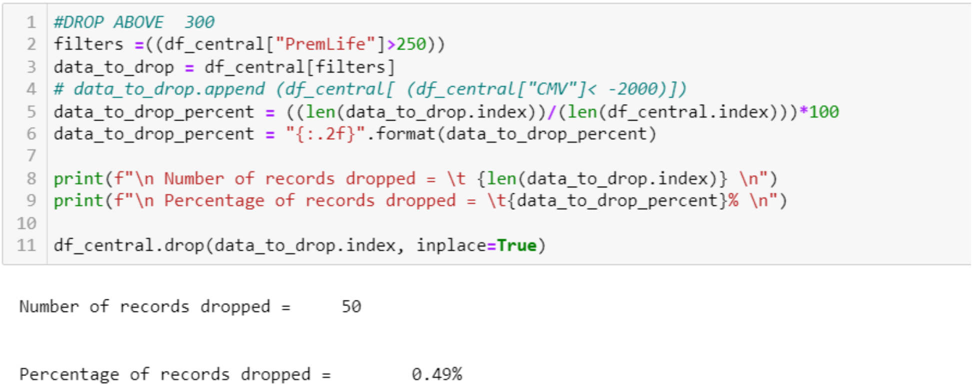
```
# Cut the window in 2 parts
f, (ax_box, ax_hist) = plt.subplots(2, sharex=True, gridspec_kw={"height_ratios": (.15, .85)})

# Add a graph in each part
sns.boxplot(df_central["PremLife"], ax=ax_box)
sns.distplot(df_central["PremLife"], ax=ax_hist)

# Remove x axis name for the boxplot
ax_box.set(xlabel='')
```

```
[Text(0.5, 0, '')]
```

Figure 6 part b: Defining an outliers' threshold and calculation of how many records were dropped



```
#DROP ABOVE  300
filters =((df_central["PremLife"]>250))
data_to_drop = df_central[filters]
# data_to_drop.append (df_central[ (df_central["CMV"]< -2000)])
data_to_drop_percent = ((len(data_to_drop.index))/(len(df_central.index)))*100
data_to_drop_percent = "{:.2f}".format(data_to_drop_percent)

print(f"\n Number of records dropped = \t {len(data_to_drop.index)} \n")
print(f"\n Percentage of records dropped = \t{data_to_drop_percent}% \n")

df_central.drop(data_to_drop.index, inplace=True)
```

```
 Number of records dropped =      50


 Percentage of records dropped =        0.49%
```

Figure 7: Correlation Matrix after Feature Engineering

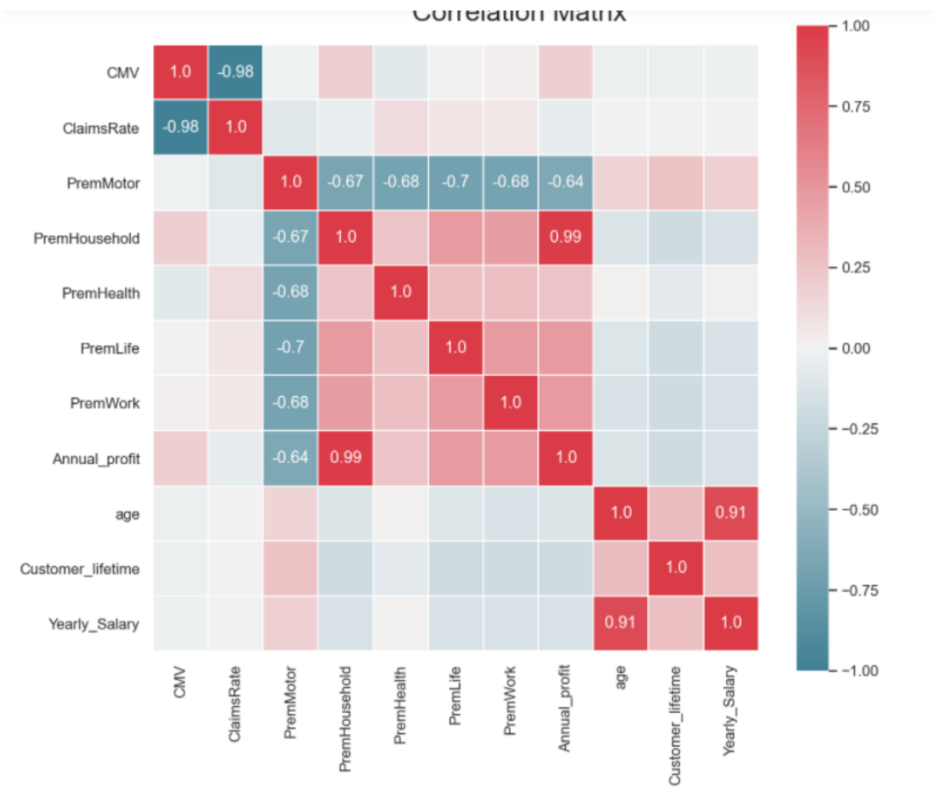


Figure 8: Stacked BarChart of each Categorival Variables vs ClientValue

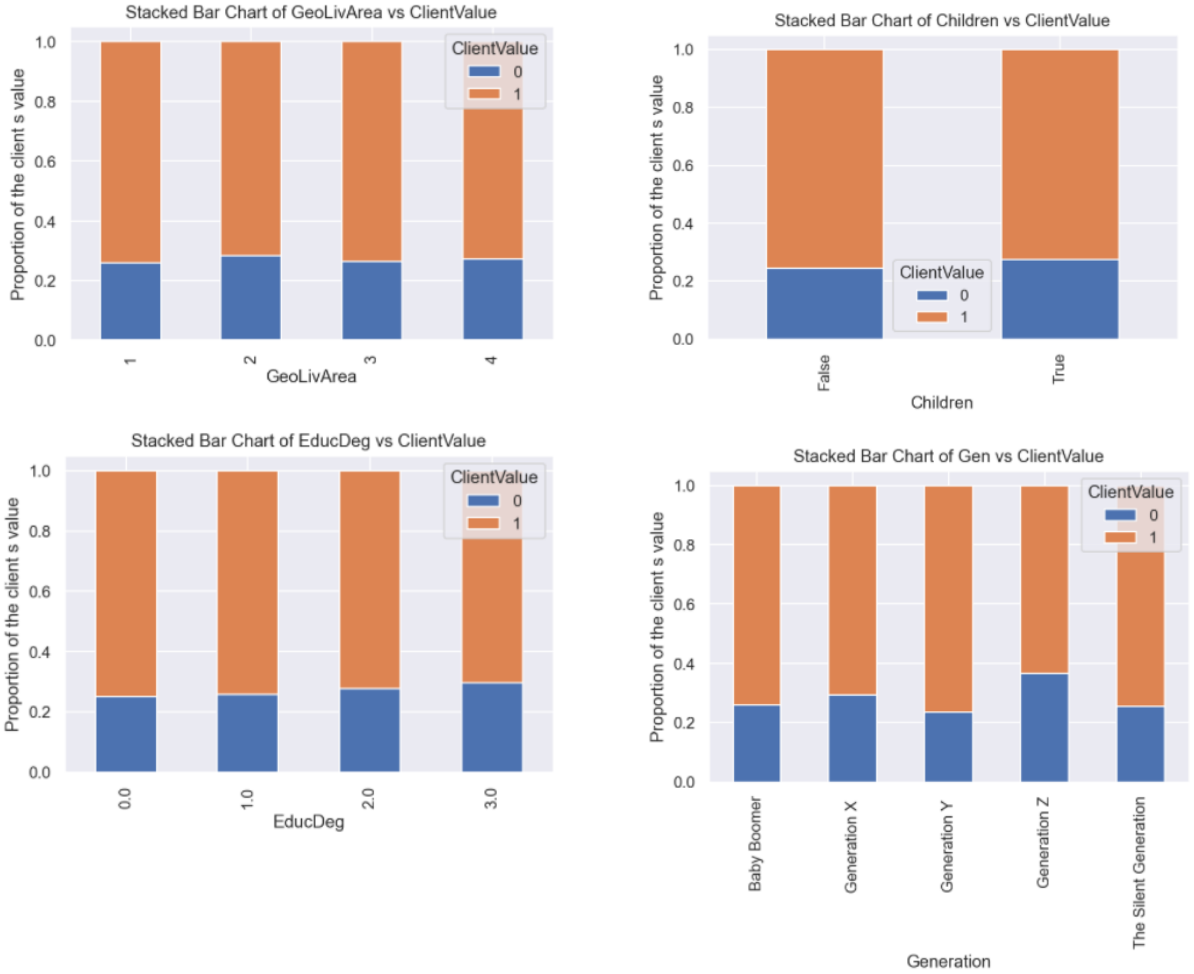
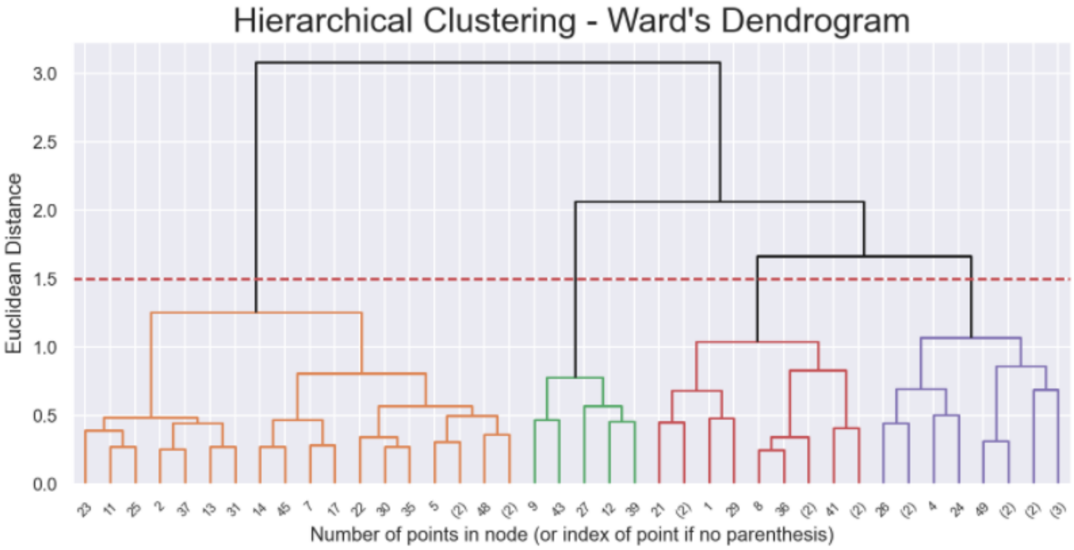


Figure 9: Ward’s Dendrogram (HC)



ANNEXE 2: Visualization from the HC with the 50 K-means cluster

Figure 1 annexe 2: General overview of the numerical variables of the 3 clusters

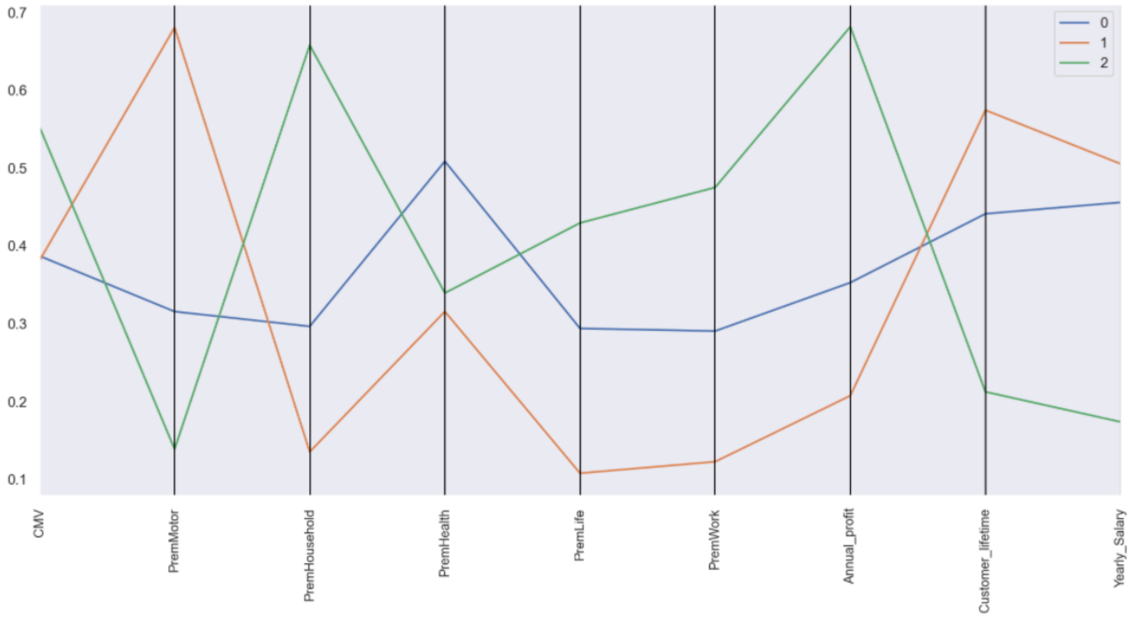


Figure 2 annexe 2: Bar charts of the numerical features of each cluster

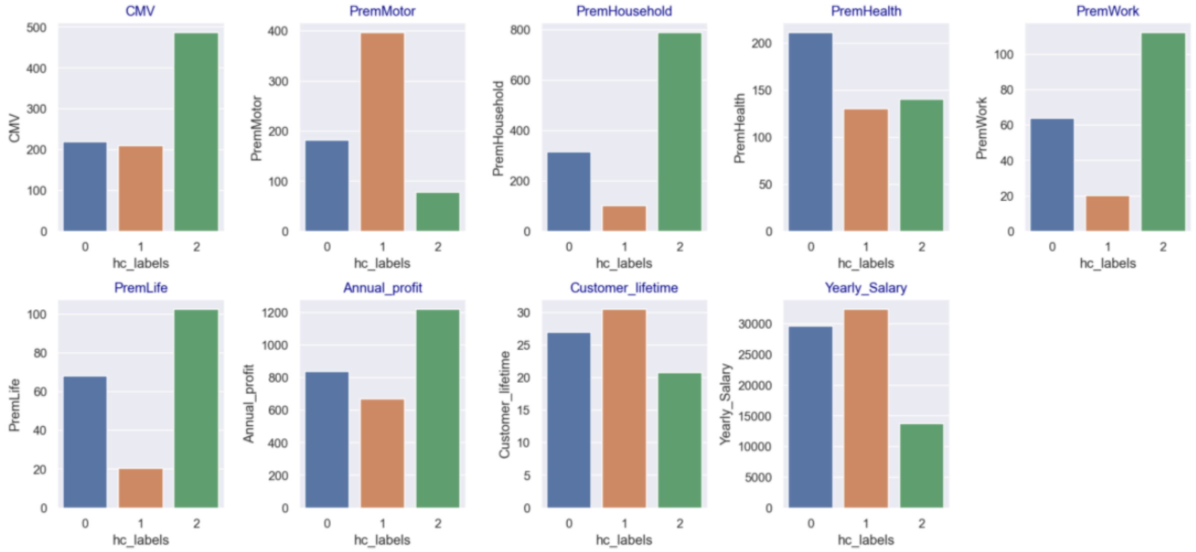


Figure 3 annexe 2: Bar charts of the Education degree level in each cluster

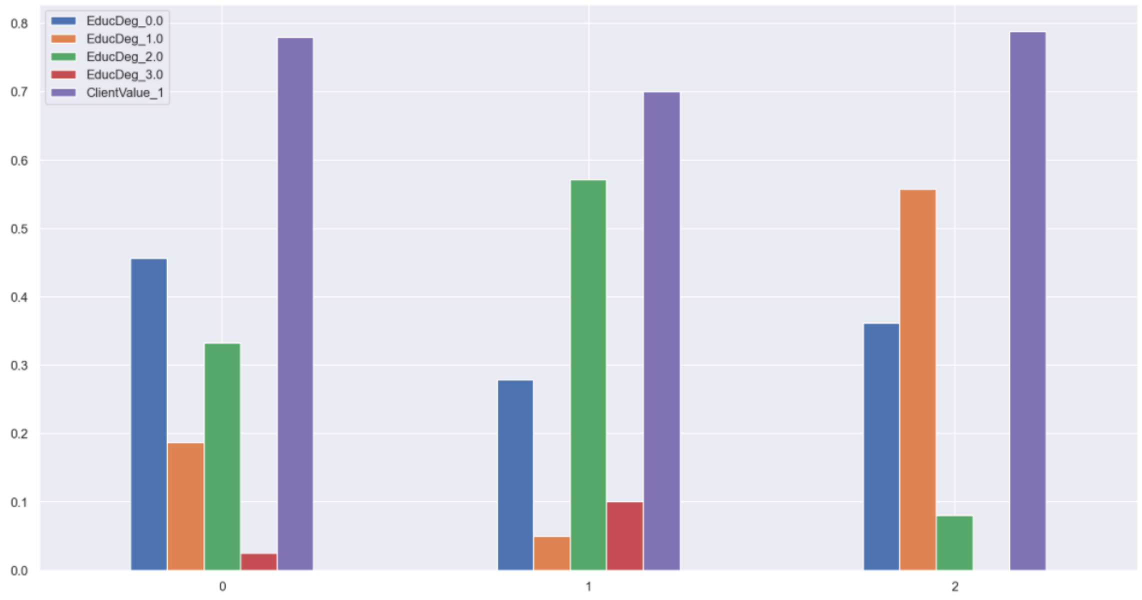
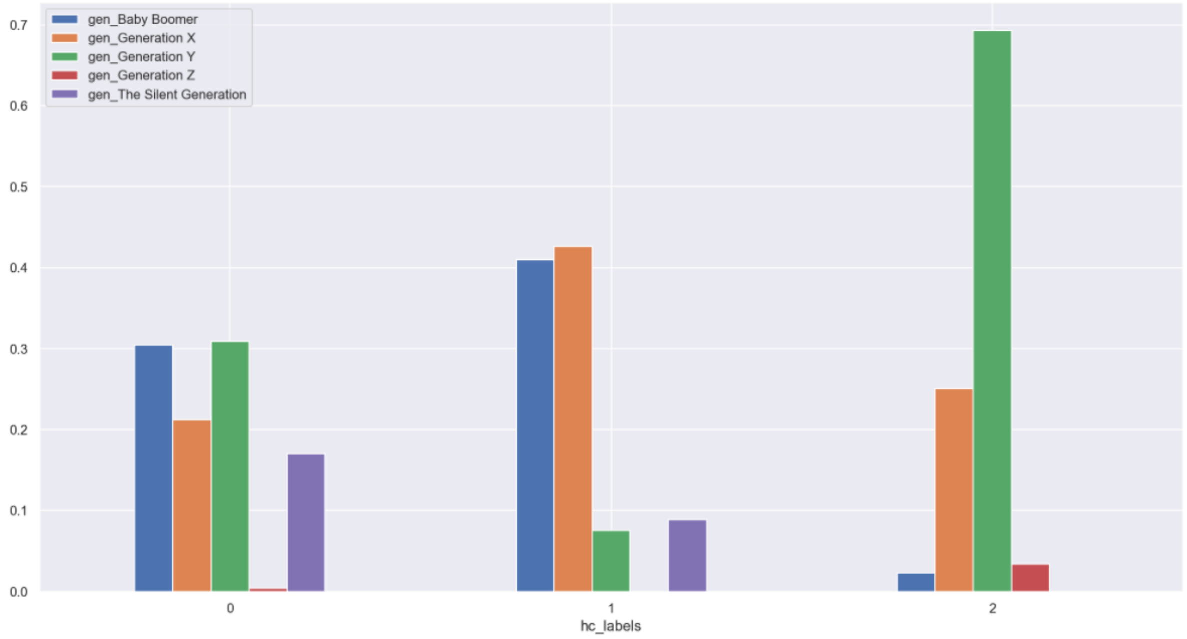


Figure 4 annexe 2: Bar charts of the Education degree level in each clust



ANNEXE 3: Other cluster algorithms we tried using

Hierarchical clustering

Hierarchical clustering bases on grouping data based on the distance (similarity) between different groups. To see the recommended number of clusters we should choose, we built a Ward's Dendrogram. We then used the algorithm to perform hierarchical clustering method with 8 clusters on our normalized data, and we got 9, extremely similar, clusters. Hence, we figured this was not the best method to cluster our data and define our marketing approach for the insurance company.

K-means clustering

K-means clustering is a centroid-based algorithm that, on each iteration, assigns the different points to the closes cluster. Within this method, we performed several different clustering techniques variants:

- "simple" k-means lead us to 2, very look alike clusters, which did not allow us to do a proper profiling of the customers
- K-means with 50 hierarchical clusters resulted in a high number of groups, leading to a higher risk of overfitting and making it difficult to develop a specific strategy for each one of them. However, we can afterwards start from these clusters to group the values into a smaller number of larger clusters.
- K-means with raw data lead us to 50 clusters under similar conditions to the previously mentioned. This also applies to k-means with encoded data.
- Hierarchical clustering with 50 k-means clusters was the method that gave us better results. We built a Ward's Dendrogram to see how we should group the 50 k-means based clusters that we started with, and based on the information we gathered from the graphic, we decided to group our data into 3 clusters. We chose to build 3 clusters instead of 4 because, when clustering the data into 4 groups, two of the clusters came out with very similar results. Using this method, we got three different clusters that we found appropriate to further explore for our marketing strategy.

Self-organizing maps

Self-organizing maps are based on unsupervised neural networks, adjusting the neurons position on each iteration based on their proximity to the data agglomerations.
We visualized the component planes, u-matrix and hit map. Afterwards, we also applied these three methods with k-means. This lead to very high number of clusters that did not fit our purposes.

Density-based clustering

These methods identify distinct clusters in the data based on the different high point density regions.

- Mean shift clustering estimated 5 as the proper number of clusters
- Density-based spatial clustering estimated one single cluster

- Gaussian mixture model was also applied to give us 5 clusters

All of these experiments gave us low r2 values when we computed this test to the clustering solutions.