

**Data Mining Project**

**MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

Segmentation of An Insurance Company

Client Dataset

Group 8

Beatriz Gonçalves, number: 20210695

Diogo Hipólito, number: 20210633

Diogo Pereira, number: 20210657

01, 2022

**Index**

[1. Introduction 3](#_Toc91610232)

[2. Data Understanding 3](#_Toc91610233)

[3. Data Preparation 4](#_Toc91610234)

[3.1. Coherence Checking 4](#_Toc91610235)

[3.2. Identifying and Removing Outliers 5](#_Toc91610236)

[3.3. Filling in the Missing Values 7](#_Toc91610237)

[3.4. Non-Metric Variables 8](#_Toc91610238)

[4. Data Pre-processing 9](#_Toc91610239)

[4.1. Feature Engineering 9](#_Toc91610240)

[4.2. Correlations 9](#_Toc91610241)

[5. Clustering Algorithms 10](#_Toc91610242)

[5.1. K-means 10](#_Toc91610243)

[5.1.1. Product Segmentation 10](#_Toc91610244)

[5.1.2. Value Segmentation 10](#_Toc91610245)

[5.2. Self-Organizing Maps 10](#_Toc91610246)

[6. Reassignment of Individuals to Clusters 11](#_Toc91610247)

[7. Marketing Approaches 11](#_Toc91610248)

[8. Conclusion 11](#_Toc91610249)

[9. References 11](#_Toc91610250)

[10. Appendix 11](#_Toc91610251)

# Introduction

The client, an insurance company, wishes to better understand the scope of its clients, in order to better serve them and increase their return on investment. The given ABT (Analytic Based Table) consists of 10.290 customers and the given task involves analysing the table for evident groups of clusters, extracting the behaviour of those clusters, and providing insights for the Marketing Department to better understand all the different Customers’ Profiles.

The project can be found in a GitHub repository which can be accessed through the following link: <https://github.com/beatrizctgoncalves/project_dm>. The repository contains a Jupyter Notebook with all the relevant analyses. Note that all decisions made in this process are justified in the notebook with theoretical references, appended to the relevant code section that utilizes these references.

# Data Understanding

The company’s ABT from 2016 has 10 296 observations and 14 variablesthat are described in the following table (Table 1). We should note that in the premium variables we can have negative values that manifest reversals occurred in the current year (2016), paid in the previous one(s). This means that the clients with negative values cancelled the respective insurance.

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| CustID | float64 | Customer ID |
| FirstPolYear | float64 | First year as a customer |
| BirthYear | float64 | Customer's Birthday Year |
| EducDeg | object | Customer's Education Level (1-Basic, 2-High school, 3-BSc/Msc, 4-PhD) |
| MonthSal | float64 | Gross Monthly Salary (€) |
| GeoLivArea | float64 | Categorical variable from 1 to 4 that identifies the living area (there is no information about the meaning of the area codes) |
| Children | float64 | Binary variable that tells if the customer has children (1) or not (0) |
| CustMonVal | float64 | Customer Monetary Value/Lifetime value = (annual profit from the customer) x (number of years that they are a customer) - (acquisition cost) |
| ClaimsRate | float64 | Amount paid by the insurance company (€)/ Premiums (€) (Note: in the last 2 years) |
| PremMotor | float64 | Annual premiums in Motor (€) |
| PremHousehold | float64 | Annual premiums in Household (€) |
| PremHealth | float64 | Annual premiums in Health (€) |
| PremLife | float64 | Annual premiums in Life (€) |
| PremWork | float64 | Annual premiums in Work Compensations (€) |

Table 1 - Variable Description

Initially, we found that there are non-metric and metric variables, which requires that these variables be treated separately. The non-metric variables are *EducDeg*, *GeoLivArea* and *Children*, and the remaining variables are metrics.

The next step was to understand the relations between variables, we computed the correlations between them using the Pearson method. The correlation matrix, as can be seen in Figure 1, show us which variables have the most potential for future modelling.

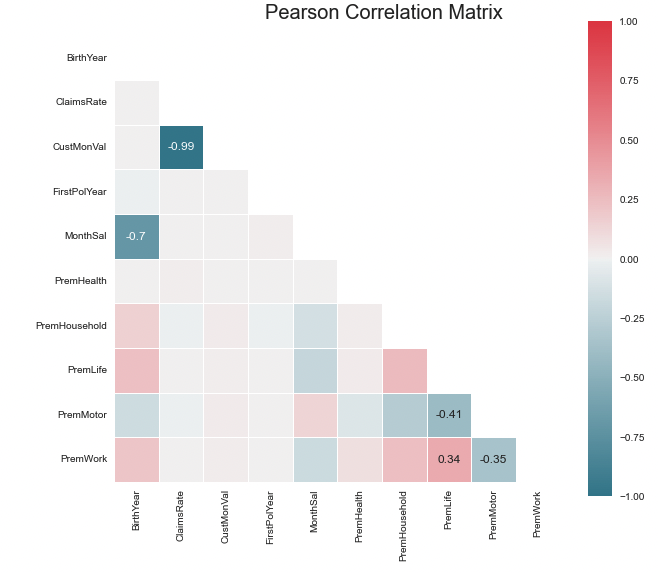


Figure 1 – Metric Correlation Matrix

After analysing it, we ended up noticing that there is a high correlation between *ClaimsRate* and *CustMonVal*, in fact it is almost a perfectly negative correlation, which would mean that with one variable we can predict the value of the other. However, looking at the problem description, it does not initially look like one variable was created using the other. According to the description, *CustMonVal = Total profit from customer \* number of years as customer*, and *ClaimsRate = Amount paid by insurance / Premiums.* In the future we will decide if we keep both columns.

Other than that, there isn't any meaningful correlation between any of the variables, which is not a good start for our model. We will have to go through a deep data preparation process in order to have a workable dataset.

# Data Preparation

Data preparation is a crucial first task in this project since the team has come across various issues with the variables in the dataset, such as missing values to different scales among the variables and outliers. Consequently, these can have a negative impact on the latter analysis of the clusters, therefore they need to be rectified. First, we’ll begin by checking the coherence of our dataset, then identifying and removing outliers, and the last filling in the missing values.

## Coherence Checking

As this dataset supposedly comes from a real-life situation, a first check should be to verify that the data is coherent with reality. For this, a few sanity checks were performed in accordance with standard questions of the area.

The second rule is about the *FirstPolYear*. It does not make sense for the *FirstPolYear* to be smaller than 1143, since it is the year that Portugal was founded and we cannot have a record of something that did not happen yet, so it cannot be bigger than 2016. After applying this rule, we found out that there is also only one observation that does not obey it. This is clearly not right so we decided to delete this row.

The third rule is also about *FirstPolYear*. Since this variable represents the first contact of the client with the insurance firm, it does not make sense that this contact happened before the person was born - *FirstPolYear* cannot be smaller than *BirthYear*. After applying this rule, we found that 1997 observations did not comply with it. Since this is a huge number, we could not delete all the incoherent rows. So, as *FirstPolYear* was calculated by the company and *BirthYear* was probably submitted by the customers, we decided that it is more likely that the customers gave wrong information. Assuming this, it would mean that 1997 customers filled the forms wrongly or that these customers inherited all this data from their parents. However, as we cannot be sure about that, we decided to delete the *BirthYear* column.

## Identifying and Removing Outliers

The outlier removal process was done in a several step protocol because the dataset presents some serious challenges.

Firstly, taking into consideration the Metric Variable’s Box Plot on Figure 2, we concluded that most of the variables have outliers. Therefore, we will go step by step throughout the variables, analysing the histograms and boxplots for each one that raises a red flag.Table

Description automatically generated with medium confidence

Figure 2 – Metric Variables’ Box Plot

We first choose a simple approach, by deleting the values that were extremely out of context just by looking into Figure 2.

For that, we dropped every row with the ClaimRate above 4, the rows with CustMonVal bellow -2000, FirstPolYear higher than 2017 and values of MonthSal, PremHealth, PremHousehold, PremMotor and PremWork above 30000, 5000, 4000, 2000 and 750 respectively. With this we conserved around 99,71% of the original data.

With this, we noticed that we still had dubious data, so we decided to try to filter even more using the IQR method for outliers’ detection. Where we get the 0.25 quartile and the 0.75 quartile so that we could get the Interquartile Range (iqr), then, we defined a threshold and with this we defined a margin (*Threshold \* IQR*) and for each variable every value that is higher than 0.75 quartile plus the margin or lower than 0.25 quartile minus the margin are considered outliers. With this method we kept 83,43% of the data from previous detection. With this method we conserved around 83,18% of the original data.

After applied de IQR method, we also did outlier removal using the Z-Score method which conserved around 93,31% of the original data.

Finally, we combined different outlier methods and obtained a percentage of data kept of 93,31% which is lower than the data kept after removing outliers manually whereby we choose to use only the manual filtering version.

## Filling in the Missing Values

Most of the missing values are on the *PremLife* and *PremWork* variables as can be observed in Table 2. Therefore, we developed a bar plot of Premium variables as shown in Figure 3, in order to see how many zero values were on each of the premium variables.

|  |  |
| --- | --- |
| Variable | Missing Values |
| FirstPolYear | 30 |
| EducDeg | 17 |
| MonthSal | 36 |
| GeoLivArea | 1 |
| Children | 21 |
| CustMonVal | 0 |
| ClaimsRate | 0 |
| PremMotor | 34 |
| PremHousehold | 0 |
| PremHealth | 43 |
| PremLife | 104 |
| PremWork | 86 |

Chart, histogram

Description automatically generatedTable 2 - Missing values of each Variable

Figure 3 – Premium Variables’ Bar Plot

The only Premium variable with zero values is *PremHousehold*, with 60 total zero values. Therefore, we considered two options. The first one was that zeros represent missing values. However, if so, we do not know why they were only present in *PremiumHousehold*. The second one was that missing values mean zero. Nevertheless, we could not justify why this was not applied in *PremiumHousehold* then.

The fact that we have already encountered errors previously, with the *BirthYear* variable, made us doubt about the integrity of the dataset and given all these factors, we decided not to replace the missing values with 0. Instead, we used an independent approach: imputing the missing values with the mean of its neighbours using *KNNImputer*.

Note that for categorical features, we cannot use the mean as the imputer metric. Therefore, we used Simple Imputer for categorical variables, which replaced missing values with the mode of each variable.

Subsequently, we merged both data frames and concluded that the imputers were successful, since there are no missing values remaining in our data frame as can be seen in Figure 4.

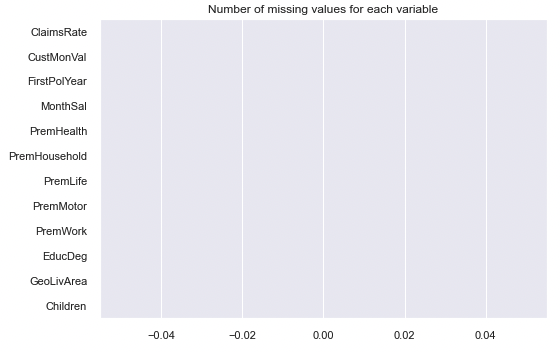


Figure 4 – Number of missing values by variable

## Non-Metric Variables

Graphical user interface, application

Description automatically generatedAs said before, *EducDeg*, Children and *GeoLivArea* are non-metric variables, thus need to be dealt with separately in order to properly add them to the analysis to their full extent. For that, we built the graphics present on Figure 5 composed by the mean estimate (points) along with its confidence interval (width) across different levels of area (x axis), children (colour) and education (columns) for each metric variable (rows). This allowed us to see whether the categorical variables impact or not the distribution of the metric variables and in which way.

Figure 5 – Categorical variable analysis

First of all, we can conclude that *Children* is, in fact, a meaningful variable that could provide useful information for the cluster analysis since customers with children have less salary and minor but still noticeable effect on all the different premiums, that is higher motor and lower everything else.

Regarding *EducDeg*, we can see a clear impact on salary, particularly between basic education and all other kinds of education, this means that people with more education receive a higher salary. Also, there is a visible difference in premium costs since more education relates to a higher *PremMotor* and lower *PremLife* and *PremWork*.

Finally, we did not see any significant impact on any of the metric variables according to *GeoLivArea*, as such we decided to drop this variable.

# Data Pre-processing

## 4.1. Feature Engineering

In order to try to gain explainability and partitioning power for a better and more precise customer segmentation, we decided to create new variables that are basically transformations of variables we already had in the data frame.

Firstly, we converted *MonthlySal* to a new variable, *YearlySalary*, since the premium costs are represented as yearly, we considered it would be better to use yearly instead of monthly salary for consistency’s sake. YearlySal = MonthSal \* 12.

Secondly, *FirstPolYear* was converted into a new variable named *ClientYears* because it measures the number of years since the first policy and this transformation makes the data simpler and easier to analyze. *ClientYears = 2016 - FirstPolYear*.

The next step was to create a variable named *TotalPremiums* that sums of all non-cancelled premiums categories, in order to know how much money each customer spent in our company in 2016. Using these last two new variables, we created the *PremiumsRate*, which measures the proportion of the salary spent in premiums divided by yearly salary in our company. This may be a good measure of a client’s commitment to the company as it measures the effort of each customer to be a client.

Then, we created Premium Proportion for all premiums. These variables express how much a customer spent in one premium relative to the total spent (e.g. *PremWorkProp = PremWork/TotalPremiums*), if *PremWork* is less than 0, its value will be set 0.

Lastly, we created a variable called *Cancelled* that represents which customers have cancelled an insurance contract.

## 4.2. Correlations

At this phase, we find it important to go back to the correlations (Figure 6) and check how the new variables are related to each other and to the original ones, as we should not use high correlated variables when performing clustering analysis because it can inflate the importance of some variables, leading to wrong segmentation definitions.

Chart

Description automatically generated

Figure 6 – Correlation Matrix of cleaned data

By looking to the Correlation matrix, we can see that many variables present perfect or almost perfect correlation, such as *CustMonVal* with *ClaimsRate* (-0.92), the newly created variables of proporcion with their corresponding variable, as it was be expected and we have *PremMotorProp* and *TotalPremiums* with very high correlations with three other variables.

# Clustering Algorithms

## 5.1. Product Segmentation

We decided to use the following methods:

* K-Means for its clarity and trustworthiness. This is a clustering algorithm that identifies k centroids and allocates each observation to the centroid closer to it.
* K-Prototype is a clustering method based on partitioning. Its algorithm is an improvement of the K-Means and K-Mode clustering algorithm to handle clustering with the mixed data types.
* Hierarchical Clustering to check if we could obtain better results and to support the number of clusters decision.

## 5.1.1. Product Segmentation

For the Product Segmentation, we wanted to study how customers were grouped according to what they bought. To find the optimal number of clusters in K-Means, we used two methods.

One of them was Elbow method to find the average distance of each point in a cluster to its centroid and represents it in a plot. The value of k where the average distance suddenly falls was the chosen one. The distortion score served as the main factor for decision making and the silhouette score and *Calinski\_Harabasz* score was used as support. To produce these plots, we decided to use the *KElbowVisualizer* package from *yellowbrick*.

The other one was Silhouette method which computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. The average of the silhouette coefficients for all points returned the silhouette score. To properly analyze the silhouette plots we considered not only the silhouette scores, but also the cluster size and quality for each number of clusters.

Then we went for a product segmentation, analyzing both absolute product values and relative product values (their proportion relative to the total spent by the customer within the company). We finally went for a combined metric and non-metric data clustering analysis focusing on the customer segmentation, by using K-Prototypes, a method that accepts mixed data types and by attempting a K-Means approach with the non-metric variables transformed through one hot encoding (as a comparison method, as this approach can be reliable sometimes). We also employed some approaches to select the best cluster solution among many, however in the following section, we will just present the selected solutions since these approaches can be consulted in detail in the notebook.

To sum up the cluster profiles we used simple correspondence analysis to assess the association between clusters and the non-metric variables. For the metric variables, we employed the parallel coordinates plot to assess the average behaviour for each cluster.

We decided to define a single set of clusters that could be used to propose marketing strategies. We utilized a simple Hierarchical Clustering to group the cluster from both the Product and Customer segmentations. Finally, we proposed a simple set of rules that can be applied to classify customers into the set of defined clusters by using a Decision Tree.

Chart, funnel chart

Description automatically generated

Figure 8 – Silhouette plot for the various clusters

## 5.1.2. Value Segmentation

For the Value Segmentation, SOMETHING

## 5.2. Self-Organizing Maps

# Reassignment of Individuals to Clusters

# Marketing Approaches

# Conclusion

# References

1. YelowBrick: Machine Learning Visualization – <https://www.scikit-yb.org/en/latest/>
2. What Are the Advantages of Decision Trees? – <https://smallbusiness.chron.com/advantages-decision-trees-75226.html>
3. KNN Imputation for Missing Values in Machine Learning – <https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/>
4. The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical) – <https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb>
5. GMM: Gaussian Mixture Models — How to Successfully Use It to Cluster Your Data? –<https://towardsdatascience.com/gmm-gaussian-mixture-models-how-to-successfully-use-it-to-cluster-your-data-891dc8ac058f>

# Appendix