Data Mining: Final Project

“A2Z Insurance”  
Customer Segmentation

Students: Lilit Tonoyan(M20190930)

Eliane Zanlorense(M20190802)

Norayr Meliksetyan(M20190687)

Instructors: Jorge Antunes

Fernando Lucas BaÇÃo

# Study Objective

# Data Understanding and Processing

* 1. Source Information
  2. Data Exploration
     1. Checking the uniqueness for the primary key
     2. Integrity
     3. Variables information: missing, unique values, outliers
     4. Univariate analysis
        1. Continuous: Histogram
        2. Categoric: Tables
     5. Bivariate Analysis
  3. Data Reduction
     1. Outliers
     2. Correlation

# Modelling

* 1. Premium Segmentation
  2. Decision Tree
     1. Current Customer
     2. New Customers

# Results

**Study Objectives**

For company it is indispensable to know its clients. All clients cannot be treated in the same way, but must be handled in a right manner. One of the ways to better understand all the different Customers’ Profiles is clustering.

* Identify variables used in clustering model and identify parameters such as number of clusters.
* Develop premium clusters based on their premium characteristics by using k‐ means clustering algorithm and identify how this variable influence the cluster.
* Based on premium cluster, geographical and social variables developing a model
* Choose the current best customer based on meaningful interpretation and mathematical decision.
* For the new customers what is the best insurance approach

# **Data Understanding and Processing**

* 1. Source Information

The dataset is composed of 14 variables that have been grouped into two groups:

* Demographics
* Premiums Variables
  1. Data Exploration

Here we `  
Create a new Variable Age computing from the current year assuming the database is of 2016.  
Compare the concordance between Variables “Birthday Year” and “First Policy’s Year”. The inconsistent values are then dropped.  
Eliminate the missing values of categorical variables to “ No Information”.  
Change the missing values of numerical variables to computed mean. As the amount of missing variable are proportionally small, the variance does not experience much change.

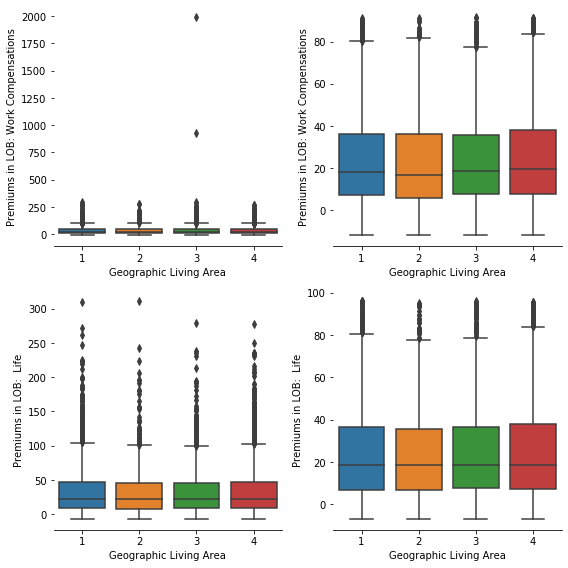
* 1. Data Reduction
     1. Outliers

Outliers give impact to the clustering result as the clustering algorithm uses distance measurement to group the data. It increases the error variance and decreases the power of statistical tests . If outliers are distributed non-randomly, they can reduce normality.

To define outliers we used Extreme Value Analysis technique and for visualization univariate and bivariate graphs were presented in our project. extreme value analysis is determining the statistical tail of the distribution and finds the values at the extreme end of the tails. Our variables are not normally distributed so they are in following upper and lower boundaries

Upper Boundary = 75th quantile +(IQR \* 1.5) Lower Boundary = 25th quantile — (IQR \* 1.5)

We are doing this formula for each continuous variable and deleting the points that are out of our boundaries. On the graph below it is demonstrated the multivariate distribution before and after outliers.



Graphic 1. Premium Life and Work Compensation by Geographic Living area

In graphic 1. we can see how outliers influence the distribution

* + 1. Correlation

Based on dataset exploratory analysis, it can be seen that some variables have high correlation. In this case, the input variables can be reduced. There are 4 sets of variables that have high correlation. The variables that are chosen based on high correlation are: Customer Monetary Value' 'Age', 'Premiums in LOB: Motor' 'Age Policy '.These features having a high correlation are dropped. The threshold is chosen to be 0.7.

Finally we have a clean data, the table shows the principal statistics for the continuous and categorical variables. On the tables below it is clearly demonstrated that Educational Degree has a high impact on Premium Variables (Household, Health, Work Compensations, Life) and conversely Living area has no influence on those variables.

On the tables below it is clearly demonstrated that Educational Degree has a high impact on Premium Variables (Household, Health, Work Compensations, Life) and conversely Living area has no influence on those variables.

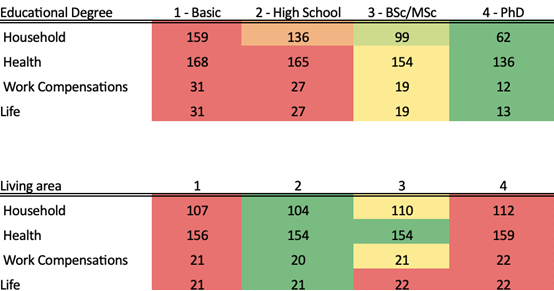


Table 2. Premium Variables distribution by Educational Degree and Living area

Based in the actual clients, in Table 1., can see that the premium variables, in the other words, how much they receive in four categories in the last years, its affect by Education Degree and Living area.

Customers with PHD receive in Household premium 177% less than customer with Basic educational Degree.

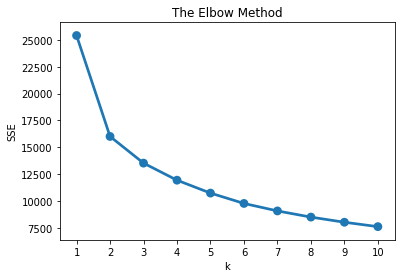
After this result, our main goal, is identify based in Premiums variables, how we can segment actives customers and next step is understand this relation with demographical variables

# **Modelling**

1. Cluster

After chose, the premium variables we all apply K-means, to segment and compare where our customers receive more premium.

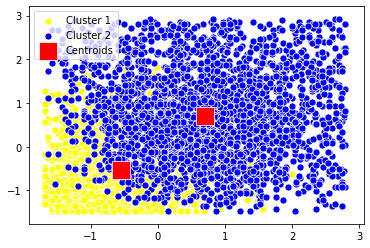
We are using Elbow method to find choose the best K

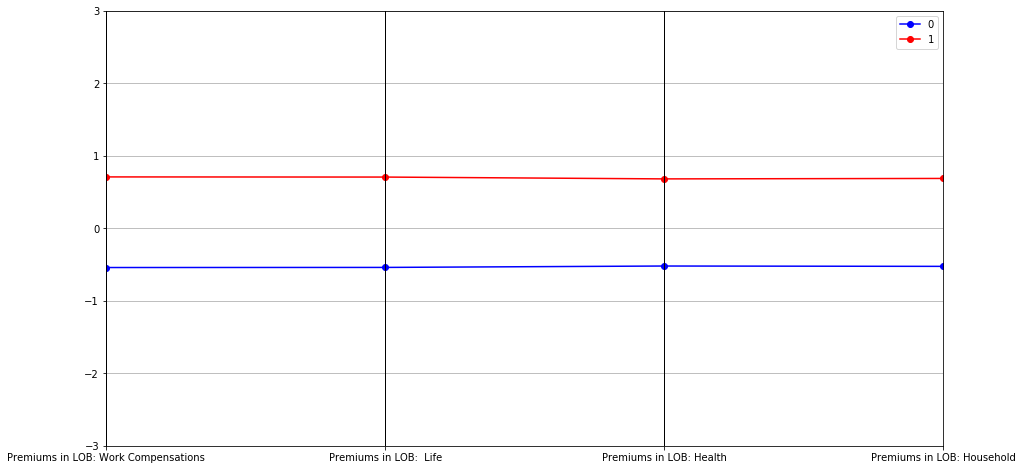


Ideally, we want to choose the point on the elbow chart where the, SSE stops decreasing at an increasing rate — i.e. the point where the change gradient of between the number of clusters becomes constant. For our model we will choose k=2-3.

After develop model with k=2 and k=3 we choose, k=2.

This will help our business, to know customers premiums and behavior



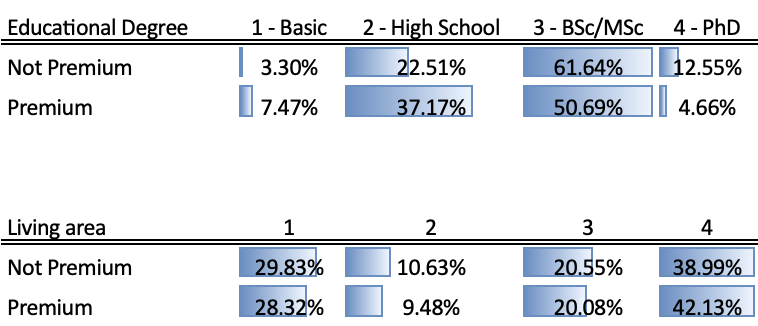


The k-means model, that will be ‘Premium clients’ have a good adjust which is demonstrated n the graph

Our lines are parallels without interruption. For this segmentation, we don’t apply a validation, our advice its to do a k-cross validation and compare results or see the segmentation performance in a few months.

The next step is evaluating the characteristics of cluster model results and the usefulness of the clusters.

Based in our opinion, the living area does not affect the premium behavior, but instead, the education has influence, how we can see. For not premium clients 12% are PhD compare with 4.% on premium clients, this result we can see to in Basic degree



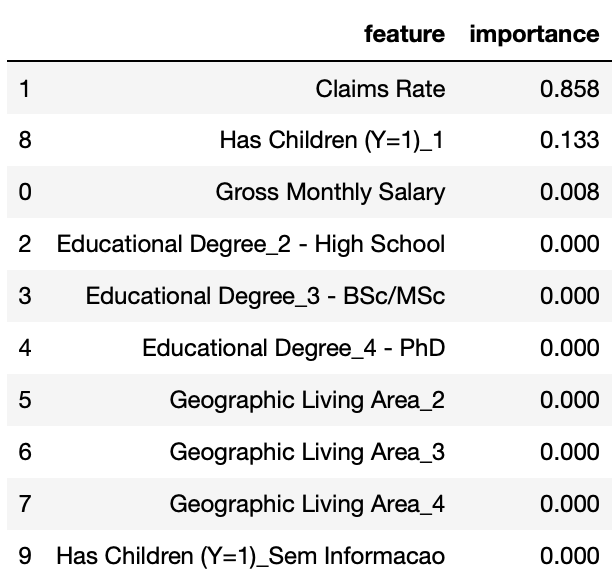
1. Decision Tree

After this results, we apply a decision tree with target: Premium

Our goals are:

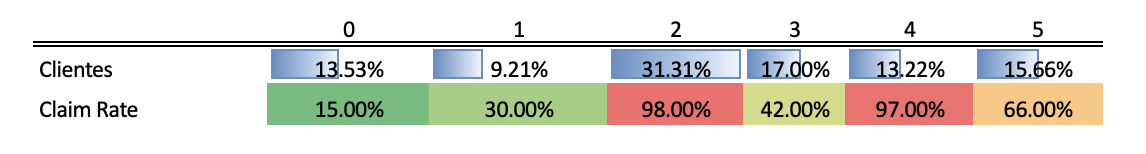
1. We know who are our clients premium and not premium, we want know, how the claim rate affect this, because probably customers with down claims rate can churn his insurance in the future and with high claims rate, understand this behavior

.



# **Results**

Based on these findings the company can choose the most appropriate marketing strategy depending on their end goal. The distributions 0 and 5 are about 30% of our customers, with this we can compare, for example the claim Rate mean for the group 0 it is 15% instead 66% in group 5. If we take a look in to the group 2 it would be clear that there we have about 31% of our customers and claims rate is near 98%.



Our approach in suggesting the company on their marketing strategy would be largely based one the groups 0 and 4.