Descriptive Methods of Data Mining
Fall Semester
2019/2020

Data Mining Project

A2Z Company - Client Segmentation

Group 91

Marcos Ramalho (M20190036)

Miguel Pacheco (M20190154)

January 2020

Contents

Index of figures

Index of tables

Introduction

Given Insurance A2Z and their customers base, a database was provided in order to segment the customers based on any criteria that the consultant team believes is valuable and worthy.

The database contains 10.296 lines/customers with relevant data. Above all, and as consultants, our aim is to increase sales, profit and customer engagement. As so, that is what we are going to work for.

In fact, insurance companies have a peculiar business model. People pay a premium in order to be insured. However, and among the values that are paid by the customers, some customers are more profitable than others due to variations in claims and casualties.

In short, the trick here is to make the customer pay a bigger compensation (premiums) for the claims and casualties that are incurred in. To do so, the company needs to design and target properly on what CRM and value proposition is concerned, with the objective of maximizing the benefits that each customer brings[1].

Data that is available to cluster/segment

In fact, and as was previously announced, the data source contains data related to 10.296 customers. The clustering/segmentation in order to maximize the value could be based on any attribute presented in the database or created along the process (new variables). Next, we present the attributes that are depicted in the database.

- ID
- First year as a customer
- Birthday[2]
- Education[3]
- Salary
- Living area[4]
- If the customer has children or not[5]
- Customer Monetary Value[6]
- Claims Rate[7]
- Premiums
    - Motor

[1] The benefits that each customer brings is defined by the formula (premiums – amount paid by the insurance company-acquisition costs)

[2] This column is going to be fundamental to calculate customers age

[3] The education level of the customer could be presented as 1,2,3 or 4, which represents respectively basic, high school, BSc/MSc or PhD educational levels

[4] The living area is divided into 4 main area, area 1,2,3 or 4

[5] Value 0 or 1 accordingly

[6] $CMV = anual\ profit * number\ of\ years\ that\ they\ are\ customers - acquisition\ cost$

[7] $Claims\ Rate = \frac{Amount\ Paid\ By\ The\ Insurance\ Company\ (€)}{Premiums\ (€)}$ (last 2 years)

- Household
- Heath
- Life
- Work Compensations

Going more into detail, each attribute could be classified according to different rules/categories, values that are introduced according to several values types, namely, nominal, binary, ordinal or numeric interval scale or ratio scale (Han, Kamber & Pei, 2012).

Table 1 - Description and type of values that are in each attribute

Attribute	Description	Type
Customer Identity	Customer ID	-
First Policy Year	First year as a company customer and consequently year of the first policy	Numeric Interval-Scale
Birthday Year	Customer birthday Year	Numeric Interval-Scale
Educational Degree	Customer academic degree	Nominal
Gross Monthly Salary	Customer salary	Numeric Ratio-Scale
Geographic Living Area	Customer living area code	Nominal
Has Children	If customer has or not children	Binary Symetric
Customer Monetary Value	Customer monetary value considering the profit and the acquisition cost of each customer.	Numeric Ratio-Scale
Claims Rate	Amount paid by the company in the last 2 years	Numeric Ratio-Scale
Premiums in LOB Motor	Annual Premiums in Household Line of Business (LOB)	Numeric Ratio-Scale
Premiums in LOB Household	Annual Premiums in Motor Line of Business (LOB)	Numeric Ratio-Scale
Premiums in LOB Health	Annual Premiums in Health Line of Business (LOB)	Numeric Ratio-Scale
Premiums in LOB Life	Annual Premiums in Life Line of Business (LOB)	Numeric Ratio-Scale
Premiums in LOB Work Compensations	Annual Premiums in Work Compensations Line of Business (LOB)	Numeric Ratio-Scale

Types of clustering/segmentation

Segmentation is one fundamental concept in marketing (Hair, Anderson, Tattham & Black, 2019). It consists in splitting market costumers in segments based on their common characteristics. This process allows companies to customize products or services to the segment needs. It also enables to increase the efficiency of marketing communication, to reduce the costs of marketing campaigns, to increase profits and customer satisfaction.

There are several segmentation types: demographic, behavioural, geographic and psychographic (LOTAME, n.d.). Demographic segmentation divides customer population according to characteristics like age, gender, family size or income. The second type of segmentation considers customer behaviour related with the brand, the usage level, product knowledge or previous purchases. Geographic segmentation divides customers according to where they are located. Finally, psychographic segmentation uses customer personality to form groups based on their interests, attitudes, values or lifestyle.

Data Exploration

In order to get an overall view of the dataset, we have used basic statistical descriptions (Han, Kamber & Pei, 2012) through the measurement of the central tendency and the distribution of each attribute.

Regarding data central tendency, the following table shows the mean, median and mode of all numeric attributes.

Table 2 - Mean, median and mode of each attribute

Attributes	Mean	Median	Mode
First Policy Year	1991	1986	1988
Birthday Year	1968	1968	1962
Gross Monthly Salary	€ 2.506,67	€ 2.501,50	€ 3.776,00
Customer Monetary Value	€ 177,89	186,87	-€ 25,00
Claims Rate	0,74	0,72	1,00
Premiums in LOB Motor	€ 300,47	€ 298,61	€ 398,74
Premiums in LOB Household	€ 210,43	€ 132,80	€ 39,45
Premiums in LOB Health	€ 171,58	€ 162,81	€ 130,47
Premiums in LOB Life	€ 41,86	€ 25,56	€ 9,89
Premiums in LOB Work Compensations	€ 41,28	€ 25,67	€ 10,89

Through observation, we also concluded that 53% of our customers have a bachelor, master's or PhD degree (5.497 observations on

Figure 1). Furthermore, 71% of the customers have at least one child (7262 observation on **Error! Reference source not found.**). The following charts shows data distribution for these attributes, having in mind respectively educational degree and children/no children statistics.

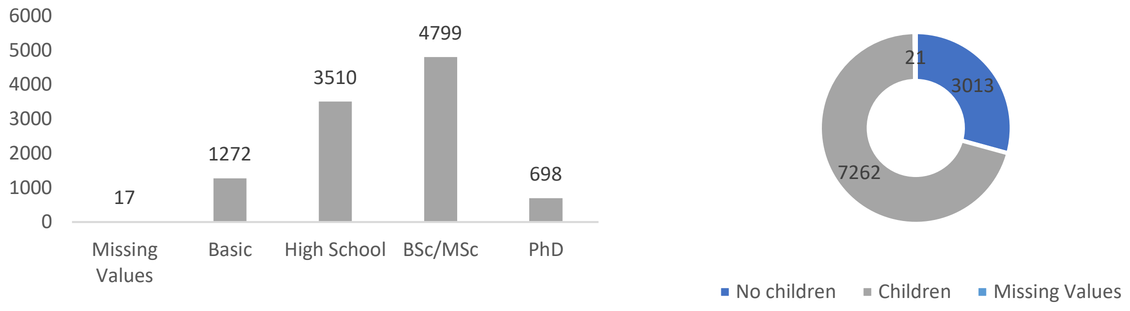


Figure 1- Educational level and children/no children distributions

However, and having in mind the fact that for segmentation purposes other attributes could be also important, we opt to depict/analyse some other value attributes to give a clear idea of the data/customer profiles that were in front of us.

In fact, and adding to the fact that most of this insurance Clients have a degree and one or more children, it was also concluded that most of our population lives in the areas 1 and 4 (70%), have between 25 and 75 years old (83%). Furthermore, more than 9.000 customer have a monthly gross salary between 1.000 and 4.000$ (88%).

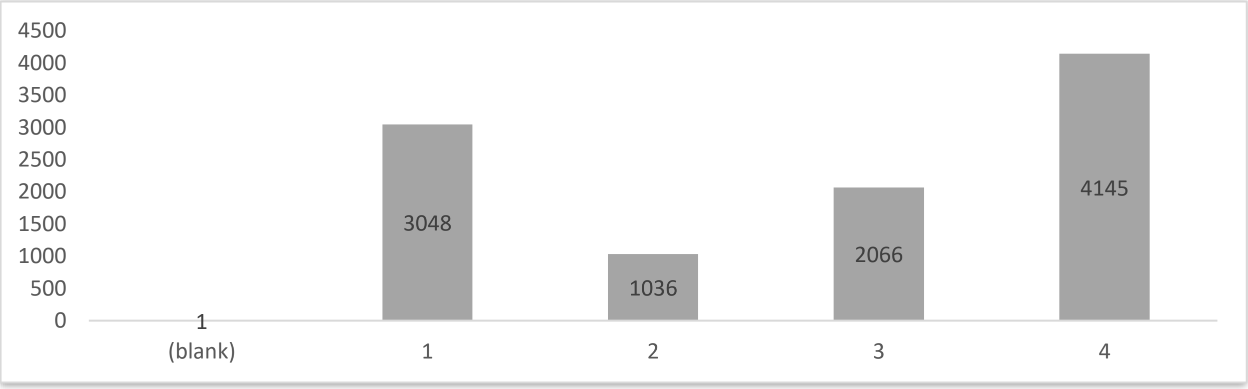


Figure 2 - Geographical distribution

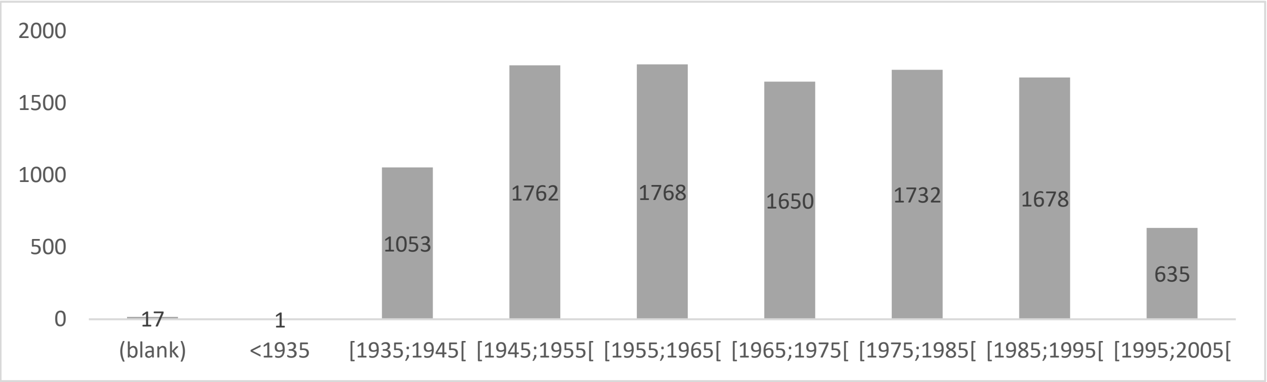


Figure 3 - Birthday year distribution

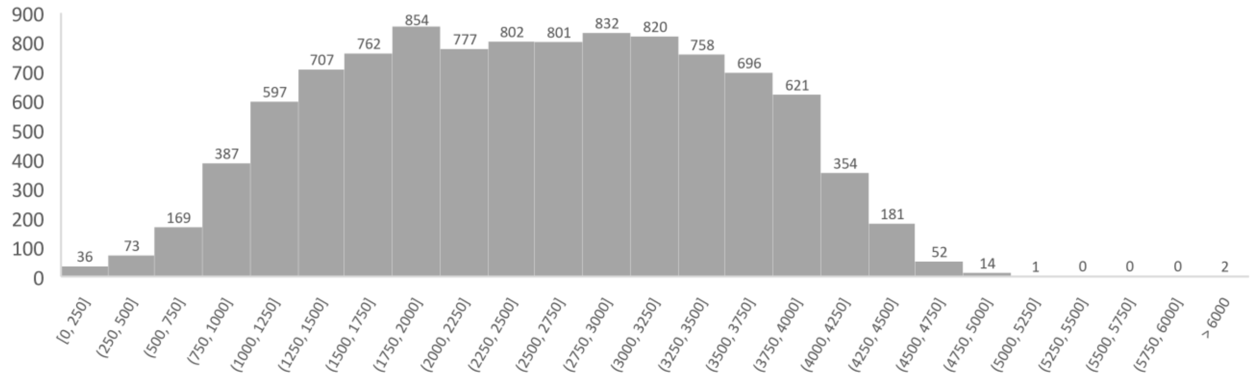


Figure 4- Gross monthly salary distribution

However, besides customers intrinsic characteristic that were depicted on the previous graphs, data is produced when customers interact with the A2Z insurance company. Looking briefly for the sum of the premiums, customers have already paid premiums between 0 and approximately 29.300$.

In the following chapters, and as we move on, additional details are going to be given since, and as was seen by the graphs that were depicted, the graphs suggest the existence of two important concepts that need to be fixed/treated:

- Missing values

- Skewed distributions, which indicates the existence of outliers[8]

In order to find the distribution of numeric attributes and the evidence of outliers, we calculated Interquartile Range (IQR) value for each attribute. We checked also the maximum and minimum values for each attribute, verifying thus that these values are significantly above or below the limits (Q3 + 1.5IQR) and (Q1 – 1.5 IQR). As expected, this evidence suggests the existence of possible outliers.

Table 3 - Interquartile Range (IQR) for each attribute

	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
Min	-165.680,42	0,00	-4,11	-75,00	-2,11	-7,00	-12,00
Q1	-9,44	0,39	190,59	49,45	111,80	9,89	10,67
Median	186,87	0,72	298,61	132,80	162,81	25,56	25,67
Q3	399,78	0,98	408,30	290,05	219,82	57,79	56,79
Max	11.875,89	256,20	11.604,42	25.048,80	28.272,00	398,30	1.988,70
Mean	177,89	0,74	300,47	210,43	171,58	41,86	41,28
Range	177.556,31	256,20	11.608,53	25123,80	28274,11	405,30	2.000,70
IQR	166.080,20	0,98	412,41	365,05	221,93	64,79	68,79
(Q1 - 1.5 IQR)	-249.129,74	-1,08	-428,03	-498,13	-221,10	-87,30	-92,52
(Q3 + 1.5 IQR)	249.520,07	2,45	1.026,92	837,63	552,72	154,98	159,98

The table that is going to be presented next, shows the standard deviation for interval numeric attributes. As can be seen, the standard deviation for these attributes suggests that the data observations tend to be spread over a large range of values (Han, Kamber & Pei, 2012).

Table 4 - Attributes mean and standard desviation

Attributes	Mean	Standard Deviation
Customer Monetary Value	177,89	1.945,72
Gross Monthly Salary	2.506,67	1.157,45
Premiums in LOB Household	210,43	352,60
Premiums in LOB Health	171,58	296,41
Premiums in LOB Motor	300,47	211,91
Premiums in LOB Work Compensations	41,27	51,51
Premiums in LOB Life	41,86	47,48
Claims Rate	0,74	2,92

Fix problems with data

As was seen in the previous chapter, and inspired by Berry, M and Linoff, G (2004), this dataset could contain some typical problems, namely:

- **Category variables with too many values -** Sometimes variables such as zip code, country or telephone provides important insights. However, in this case, there are so many possible outcomes, that this kind of data/information is discarded due to its complexity. However, since

[8] Especially in the birthday year range and gross amount salary distribution just by looking at the distributions

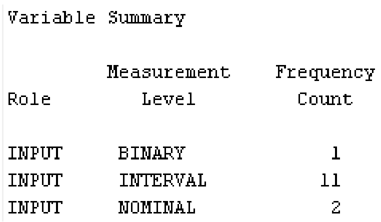
most of the information that is contained in the database is numerical, we believe that we don’t need to fix data problems related to category variables with too many values.

- **Numeric variables with skewed distributions and outliers** – According to data explore exercise, it reveals the existence of possible outliers, namely in the gross monthly salary or birthday year.
  In this case, we could discard values, divide the values into equal sized ranges, such as deciles, or transform the values by, for example, replacing each value with its logarithm.

- **Missing values** – According to Michael J. A Berry and Gordon S. Linoff, “some data mining algorithms are capable of treating “missing” as a value and incorporating it into rules”. Having in mind the same authors, the best way to solve missing values problem not creating bias or spurious information, is to create a model whose target value is the missing one.
  At this work, we are going to consider a model/algorithm in order to fulfill missing values with the objective of utilize all the data that was given to us.

- **Values with meanings that change over time** – An attribute can change its meaning over time. However, and because there is no indication of meaning change related to the variables, we are not going to consider this possibility in our work.

- **Inconsistent data encoding** – When data is collected from different sources for the same attribute, inconsistency can be encountered due to different ways of indicating the same value/king of data or different abbreviations for the same meaning. After carefully looking for the data, that is no evidence of inconsistent data encoding.



```
Variable Summary

          Measurement    Frequency
Role         Level         Count

INPUT     BINARY             1
INPUT     INTERVAL          11
INPUT     NOMINAL            2
```

Figure 5 - Types of data using “StatExplore” node in SaS Miner

Data Preparation

Missing Values

As was previously “denounced” by the caption in Figure 5, in order to study missing values and contrast the results that were plotted in Data Exploration, the “StatExplore” functionality was used in order to check initial missing values.

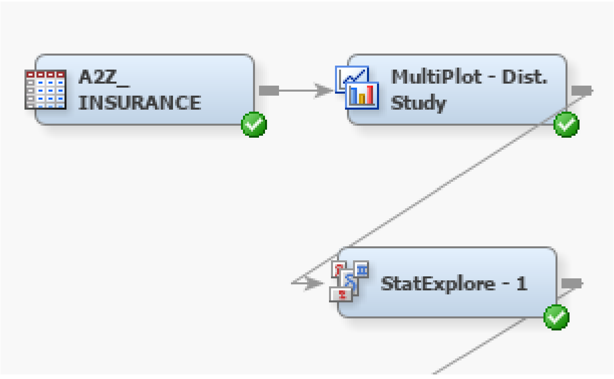
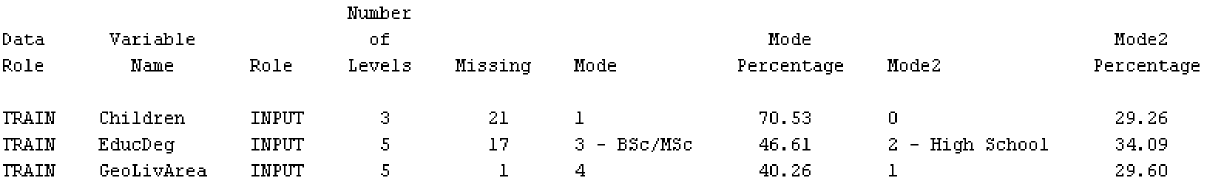


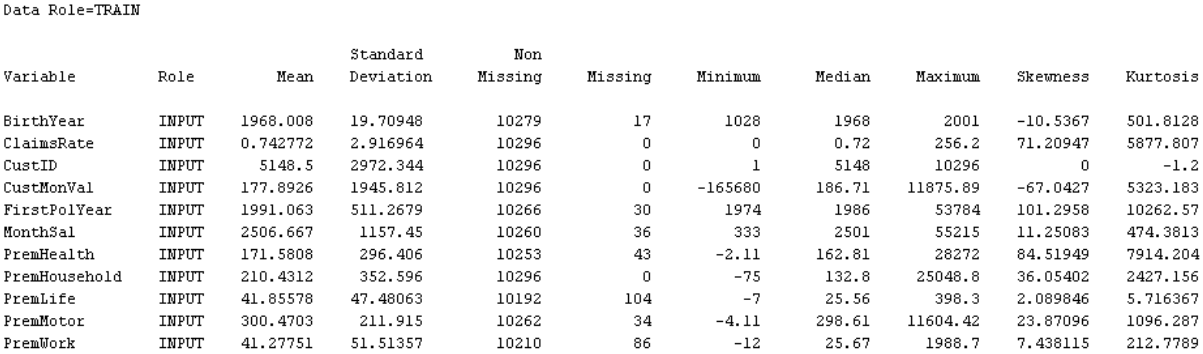
Figure 6 - "StatExplore" SaS Miner functionality representation[9]

Having in mind the results provided, and assuming that all data was taken into account as an input/train data since we want to construct a descriptive model, it can be seen by the "missing" column in Figure 8 that the vast majority of attributes have missing values. However, the missing values are not significative having in mind the 10.296 customers that are provided in the database.



```
                                    Number
Data       Variable                   of                                    Mode                          Mode2
Role         Name        Role       Levels    Missing    Mode             Percentage    Mode2             Percentage

TRAIN     Children      INPUT         3          21      1                  70.53       0                  29.26
TRAIN     EducDeg       INPUT         5          17      3 - BSc/MSc        46.61       2 - High School    34.09
TRAIN     GeoLivArea    INPUT         5           1      4                  40.26       1                  29.60
```

Figure 7 - Some metrics related to both binary and nominal attributes



```
Data Role=TRAIN

                                         Standard       Non
Variable         Role          Mean     Deviation    Missing    Missing    Minimum     Median    Maximum   Skewness   Kurtosis

BirthYear        INPUT     1968.008      19.70948      10279         17       1028       1968       2001   -10.5367   501.8128
ClaimsRate       INPUT     0.742772      2.916964      10296          0          0       0.72      256.2   71.20947   5877.807
CustID           INPUT       5148.5      2972.344      10296          0          1       5148      10296          0       -1.2
CustMonVal       INPUT     177.8926      1945.812      10296          0    -165680     186.71   11875.89   -67.0427   5323.183
FirstPolYear     INPUT     1991.063      511.2679      10266         30       1974       1986      53784   101.2958   10262.57
MonthSal         INPUT     2506.667       1157.45      10260         36        333       2501      55215   11.25083   474.3813
PremHealth       INPUT     171.5808       296.406      10253         43      -2.11     162.81      28272   84.51949   7914.204
PremHousehold    INPUT     210.4312       352.596      10296          0        -75      132.8    25048.8   36.05402   2427.156
PremLife         INPUT     41.85578      47.48063      10192        104         -7      25.56      398.3   2.089846   5.716367
PremMotor        INPUT     300.4703       211.915      10262         34      -4.11     298.61   11604.42   23.87096   1096.287
PremWork         INPUT     41.27751      51.51357      10210         86        -12      25.67     1988.7   7.438115   212.7789
```

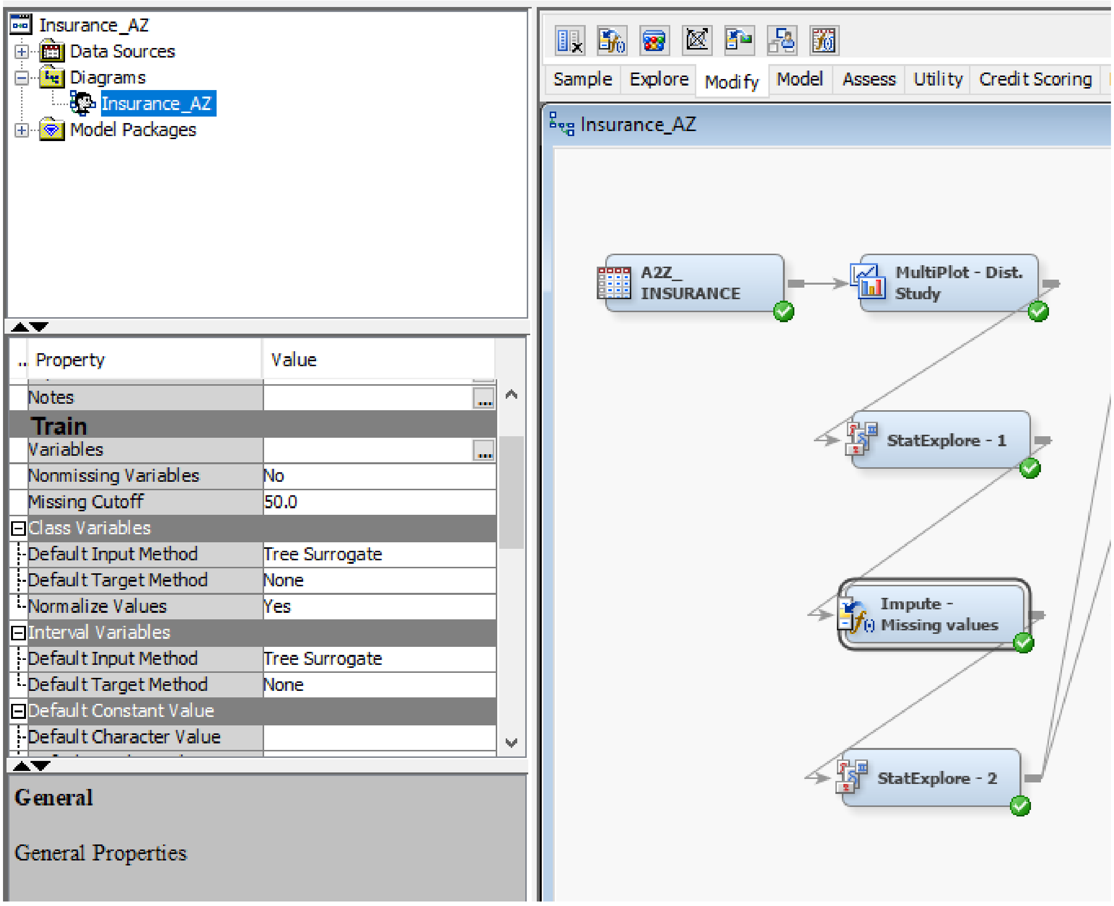
Figure 8 - Metrics related to interval values where the missing values problem is depicted on the "Missing" column

Considering the maximum number of missing values per attribute, it can be easily concluded that:

$$\max \% \; of\ missing\ values = \frac{max(nr.missing\ values)}{Total\ nr.observations} = \frac{104}{10.296} \approx 1\%$$

[9] Functionality that is available in the "Explore" Sas Miner section

Besides the fact that the number of missing values is not significative, we opt to input values on the missing values through “impute” function using “Tree Surrogate” method.



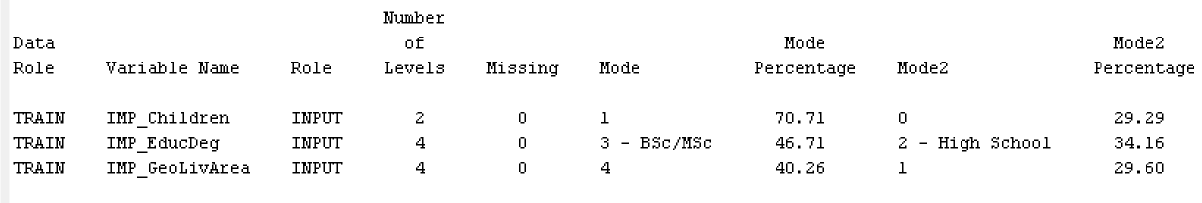
Property	Value
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Default Constant Value	
Default Character Value	

General

General Properties

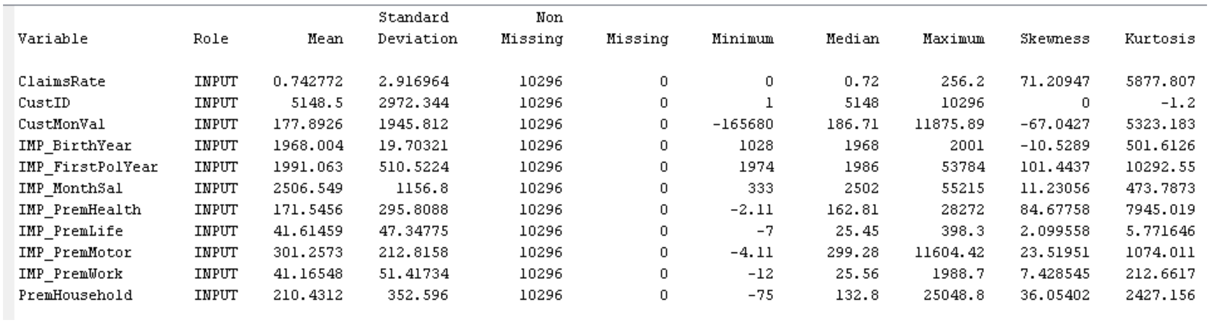
Figure 9 - "Impute" SaS Miner functionality settings representation[10]

In order to verify if the “impute”[11] functionally worked properly for both class and interval variables, an additional “StatExplore” was added after the “Impute” operation to check if there are still missing values. As could be seen by the next two images, there is no missing values now.



```
                                    Number
Data                                  of                                  Mode                              Mode2
Role      Variable Name     Role    Levels    Missing    Mode             Percentage    Mode2               Percentage

TRAIN     IMP_Children      INPUT     2          0       1                  70.71       0                     29.29
TRAIN     IMP_EducDeg       INPUT     4          0       3 - BSc/MSc        46.71       2 - High School       34.16
TRAIN     IMP_GeoLivArea    INPUT     4          0       4                  40.26       1                     29.60
```

[10] Functionality that is available on the “Modify” Sas Miner section
[11] Impute functionality is available in the “Modify” section



```
                                         Standard        Non
Variable          Role          Mean     Deviation     Missing     Missing     Minimum      Median     Maximum    Skewness    Kurtosis

ClaimsRate        INPUT     0.742772     2.916964       10296           0           0        0.72       256.2    71.20947    5877.807
CustID            INPUT       5148.5     2972.344       10296           0           1        5148       10296           0        -1.2
CustMonVal        INPUT     177.8926     1945.812       10296           0     -165680      186.71    11875.89    -67.0427    5323.183
IMP_BirthYear     INPUT     1968.004     19.70321       10296           0        1028        1968        2001    -10.5289    501.6126
IMP_FirstPolYear  INPUT     1991.063     510.5224       10296           0        1974        1986       53784    101.4437    10292.55
IMP_MonthSal      INPUT     2506.549       1156.8       10296           0         333        2502       55215    11.23056    473.7873
IMP_PremHealth    INPUT     171.5456     295.8088       10296           0       -2.11      162.81       28272    84.67758    7945.019
IMP_PremLife      INPUT     41.61459     47.34775       10296           0          -7       25.45       398.3    2.099558    5.771646
IMP_PremMotor     INPUT     301.2573     212.8158       10296           0       -4.11      299.28    11604.42    23.51951    1074.011
IMP_PremWork      INPUT     41.16548     51.41734       10296           0         -12       25.56      1988.7    7.428545    212.6617
PremHousehold     INPUT     210.4312      352.596       10296           0         -75       132.8     25048.8    36.05402    2427.156
```

Figure 10 – “StatExplore” node results to check for missing values after “Impute” node transformation

As could be seen by Figure 10, there is no missing values left. As so, the team believes that this is the right moment to treat outliers since there is no missing values left.

Outliers

As was partially seen in the data exploration phase, there are several interval indicators/attributes that denounces the existence of possible outliers.

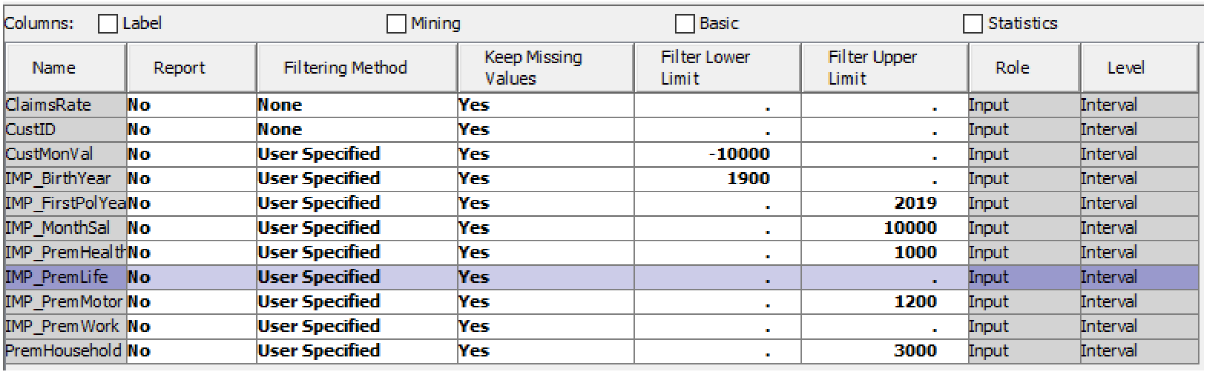
To filter the interval variables in a proper way having in mind the analysis made on the exploration phase, the filter functionality was used with filters/parameters that were specified by the user[12].

Table 5 - Table that demonstrates the outliers that were applied to the "transformed sample" after missing values imputation

Interval Variable Name	Filtering Method	Min	Max	Filter Lower Limit	Filter Upper Limit	Nr. of observations outside the limits (outliers)	Outliers limits rational
ClaimRate	Default	0	256	-	-	-	Since Claims Rates just cannot be negative, no filter limits were applied
CustID	Default	1	10.296	-	-	-	Since Customer ID is an automatic number given by the system, no filters were applied
CustMonVal	User Specifies	-165.680	11.875,89	-10.000	-	9	Aster carefully analysing the data, and knowing that CMV is given by a certain formula, a lower limit was applied to all the data
IMP_BirthYear	User Specifies	1.028	2.001	1.900	-	1	Knowing that no people with insurance could have a lot more than 100 year, as the fact that the insurance business is a recent one, we applied a lower limit of 1900
IMP_FirstPol Year	User Specifies	1.974	53.784	-	2019	1	Knowing that no year should be bigger that the actual one, the upper limit of 2019 was applied
IMP_MonthSal	User Specifies	333	55.215	-	10.000	2	A monthly limit of 10.000 was applies having in mind the distribution analysis
IMP_PremHealth	User Specifies	-2,11	28.272	-	1.000	4	Since, and according to the information provided, negative premium is a possible outcome due to business conditions. An upper limit of 1.000 was set having in mind mean distribution
IMP_PremLife	User Specifies	-7	398,3	-	-	-	Since all values are close to each other, we opt to apply no limits
IMP_PremMotor	User Specifies	-4,11	11.604,42	-	1.200	6	After carefully studying the distribution, an upper limit of 6 was applied

[12] As consultants we opt for this solution since, we already know the industry and the data related to it as could be seen on Data Exploration chapter

IMP_PremWork	User Specifies	-12	1.988,7	-	-	-	After carefully studying the distribution, there was not applied any limit, since the distribution is not that skewed in accordance to our parameters
PremHousehold	User Specifies	-75	25.048,8	-	3.000	3	An upper limit of 3000 was set, with the purpose of taking out values that are far away from the mean
Total						**20**[13]	



Columns: □ Label □ Mining □ Basic □ Statistics

Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
ClaimsRate	No	None	Yes	.	.	Input	Interval
CustID	No	None	Yes	.	.	Input	Interval
CustMonVal	No	User Specified	Yes	-10000	.	Input	Interval
IMP_BirthYear	No	User Specified	Yes	1900	.	Input	Interval
IMP_FirstPolYea	No	User Specified	Yes	.	2019	Input	Interval
IMP_MonthSal	No	User Specified	Yes	.	10000	Input	Interval
IMP_PremHealtl	No	User Specified	Yes	.	1000	Input	Interval
IMP_PremLife	No	User Specified	Yes	.	.	Input	Interval
IMP_PremMotor	No	User Specified	Yes	.	1200	Input	Interval
IMP_PremWork	No	User Specified	Yes	.	.	Input	Interval
PremHousehold	No	User Specified	Yes	.	3000	Input	Interval

Figure 11 - Filter node user-specified limits

Variables creation/transformation

Considering the data that is provided, as the formulas supplied by work criteria and request[14], we have created two new variables given the data provided:

- **Total_Premiums** = IMP_PremHealth + IMP_PremLife + IMP_PremMotor + IMP_PremWork + PremHousehold
- **Aprox_Amount_Paid_By_Insurance** = ClaimsRate[15] * Total_Premiums

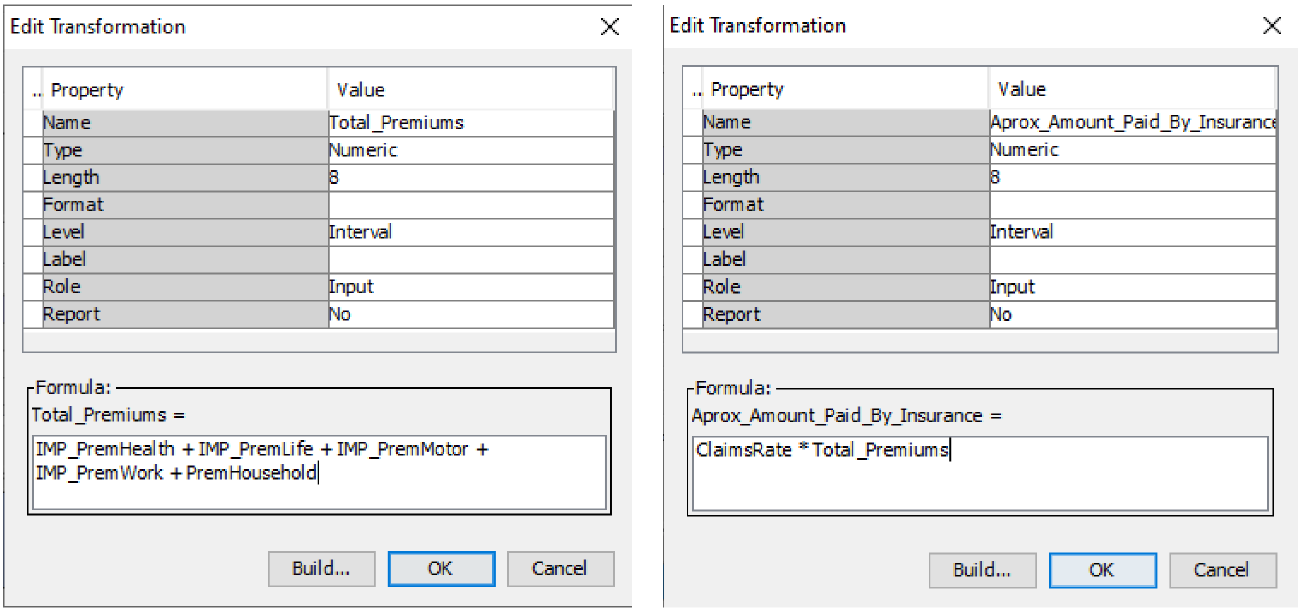
These two variables that were created given the context, are fundamental for our clustering/segmentation purposes.

[13] The total number of outliers that are out (20) is inferior to the sum of observations that are outside the limits (26) since there are outliers that are repeated in each criteria

[14]

Claims Rate = $\frac{Amount\ paid\ by\ the\ insurance\ (€)}{Premiums\ (€)}\ (last\ 2\ years)$

[15] Since at the statement it is said that Claims Rate is based on the data provided in the last 2 years, we despised that fact and we calculated the total amount paid by the insurance assuming that claims rates did not changed that much over the years. That is the reason why the name of the variable is "Aprox_Amount_Paid_By_Insurance"



Edit Transformation

.. Property	Value
Name	Total_Premiums
Type	Numeric
Length	8
Format	
Level	Interval
Label	
Role	Input
Report	No

Formula:
Total_Premiums =
IMP_PremHealth + IMP_PremLife + IMP_PremMotor + IMP_PremWork + PremHousehold

Edit Transformation

.. Property	Value
Name	Aprox_Amount_Paid_By_Insurance
Type	Numeric
Length	8
Format	
Level	Interval
Label	
Role	Input
Report	No

Formula:
Aprox_Amount_Paid_By_Insurance =
ClaimsRate * Total_Premiums

Figure 12 - Variables creation in SaS Miner

However, and after variables creation, consistency must be checked, namely the existence of missing values and outliers (again). As could be easily assumed, the new variables have no missing values since these variables were created having no missing values because the data that was already treated before.

On what outliers of the new variables is concerned, and after carefully analyzing the data, the following options were taken:

Table 6 - Outliers filter given the new variables that were created

Interval Variable Name	Filtering Method	Min	Max	Filter Lower Limit	Filter Upper Limit	Nr. of observations outside the limits (outliers)	Outliers limits rational
Total_Premiums	User Specifies	429,19	0	-	3.000	4	-
CustID	User Specifies	5.672,74	9.287,49	-	5.000	3	-
Total						7	-

As was previously done at the outliners section, the team evaluated carefully both the distributions and values in order to "lose" the minimum number of observations. At the end of this process, there were 10.269 observations/lines at the end.

Correlations and insights

At this moment, all variables are set and defined. So, some correlations among the variables can be checked. Since we are approaching the "segmentation phase", we need to choose the right variables. By this, we want to emphasize the fact that if two or more variables are correlated, its redundant to introduce these same variables for segmentation since it is going to skew the results as the analyst is putting in fact too much "stress" on a single trend.

After carefully analyzing the data through the correlation matrix, it can be easily verified that the variables that were created have a strong correlation with the attributes that contributed to their formation. Being more specific, the Aprox_Amount_Paid_By_Insurance have a strong relationship with Claims Rate (0,86), and Total_Premiums have also a strong correlation with all the premiums that contributed to their origin (correlations around 0,6). To add, monthly gross salary is totally connected to age (0,93), so "as you get old you get more money". To finish, and it is an important insight, Customer Monetary Value is intrinsically related to Claim Rate in an indirect proportionality way (-0,91). As opposed and supposed, Customer Monetary Value and the Aprox_Amount_Paid_By_Insurance are connected in an indirect proportion way (-0,73) too.

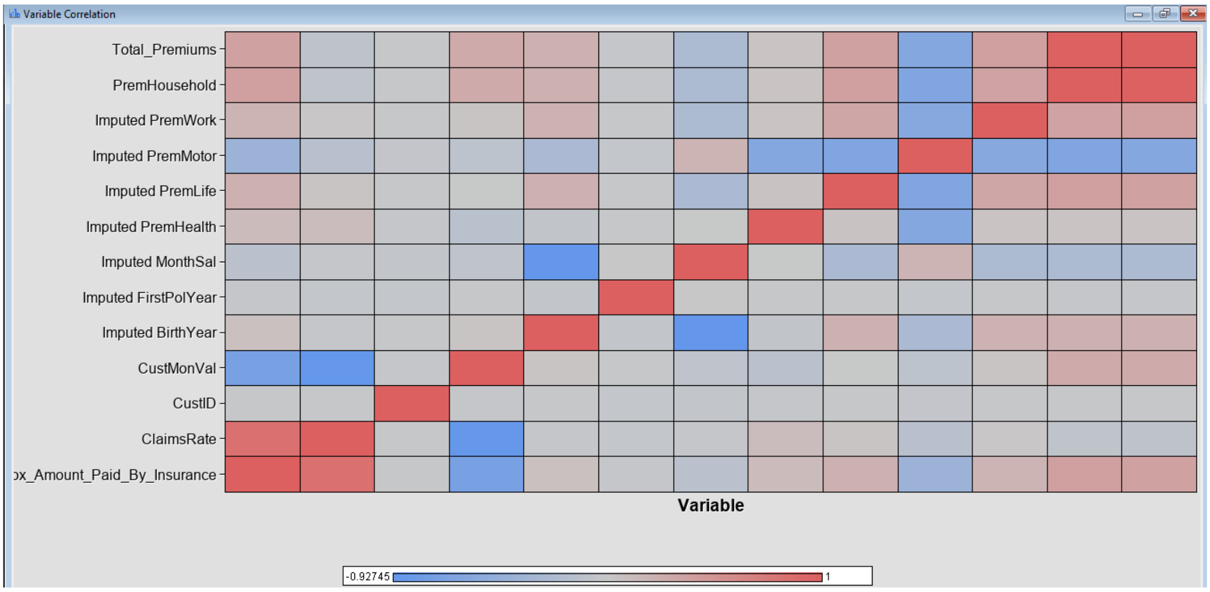


Figure 13 - Correlation matrix[16]

Since the educational level is not an interval variable but a classy one, the team decided to spend some time studying the educational level as an insight. In fact, and according to our analysis, people which educational level is lower tend to be more profitable for the company. In fact, this could be an indicator for future campaigns/moves.

Table 7 - Educational level profitability study

	Number of beneficiaries	Sum of Premiums	Amount paid by insurance	Sum of Premiums/Number of beneficiaries	Amount paid by insurance/Nr of beneficiaries	Profitability
1 - Basic	1.272	119.296.0	808.020$	938$	635$	303$
2 - High School	3.509	276.935.1	192.444.6$	789$	548$	241$
3 - BSc/MSc	4.797	336.156.6	224.401.7$	701$	468$	233$
4 - PhD	698	451.617	300.741$	647$	431$	216$

[16] The correlation matrix is available in the "Variable Clustering" node available in the "Explore" section

Segmentation

Purpose

In fact, the segmentation/division that is going to be made have two main purposes that are explained in the table below.

Table 8 - Segmentation purpose description

Purpose	**CRM**	**Profitability/Potential**
Description	In fact, the insurance company wants to build an omnichannel communication strategy in order to promote up and cross sell via flirting campaigns. To do that, customers must be organized by their tendency to buy more	The company also wants to assign the customers to different KAMs[17] in accordance to their profitability or potential
Chosen attributes[18]	• **Gross Monthly Salary** – This attribute indicates the total purchase power of the customer • **Total Premiums** – This indicates the "loyalty level" of the customer • **Claims Rate**[19] – This indicates the customer profitability when compared to the total premiums paid	

In fact, two different segments could have been done for each description/purpose. However, and after carefully analyzing the problem as the segmentation purposes, we opt to make one segmentation that encompass different styles and personas fulfilling both descriptions and purposes, maintaining business simple as usual.

Segmentation algorithm

Remembering Berry, M and Linoff, G (2004), we are going to use k-means algorithm. K-means is one of the most used algorithm for clustering. In fact, k represents the number of clusters that the algorithm is going to construct. To do this, initial k seed are places, and through an interactive process, seeds coordinates are changed in each iteration until that "initial seed" represents the points that are within the cluster (centroid). In this sense, the initial positions for the seeds are crucial, since they are going to define the result. Intuitively, it is obvious that as the number of clusters increase, the error[20] decreases. However, we need to find the perfect balance between error and the number of clusters not to make everything too complex.

Number of clusters

Given what was said in the last chapter, the elbow graph was designed. The elbow graph is a graph that gives the relationship between the error (distance to centroid) and the number of clusters. As so, it is

[17] Key Account Manager

[18] The attributes were chosen considering the correlation matrix in order not to misrepresent results

[19] Claims Rate = $\frac{Amount\ paid\ by\ the\ insurance\ company\ (€)}{Premiums\ (€)}$(last 2 years)

[20] As the number of clusters increases, each cluster characterizes better the points that are within due to the decrease of the distance between the centroid of each clusters and the points the cluster represents

the team role to choose the number of clusters that gives the best equilibrium between complexity and error.

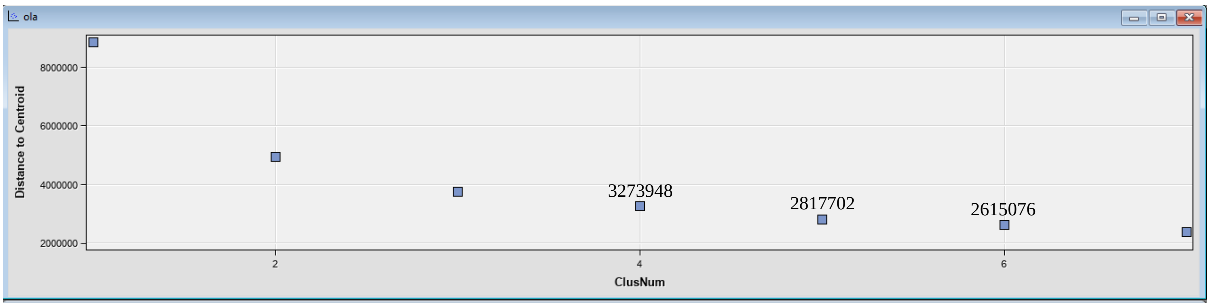


Figure 14 - Elbow graph

After carefully analyzing the graph, all indicators point to 5 clusters. In order to evaluate the elbow graph “suggestion”, we carefully examined the clusters created for each k in order to check the situation using SaS Miner “Cluster” and “Segment Profile” nodes.

Using input means plot graph as other techniques, we can easily check that for k>5, clusters become “too similar to each other” in terms of characteristics. As so, k=5 seems a good options. Furthermore, with k=5, the dimension of each cluster is balanced on what the samples is concerned.

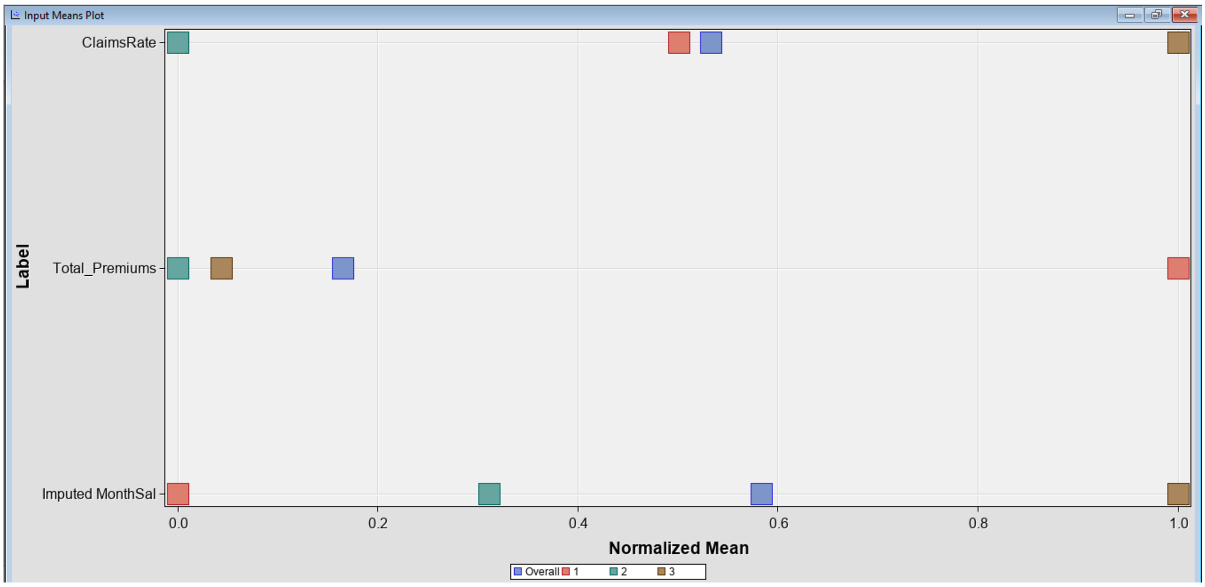


Figure 15 - Input means plot for k=3

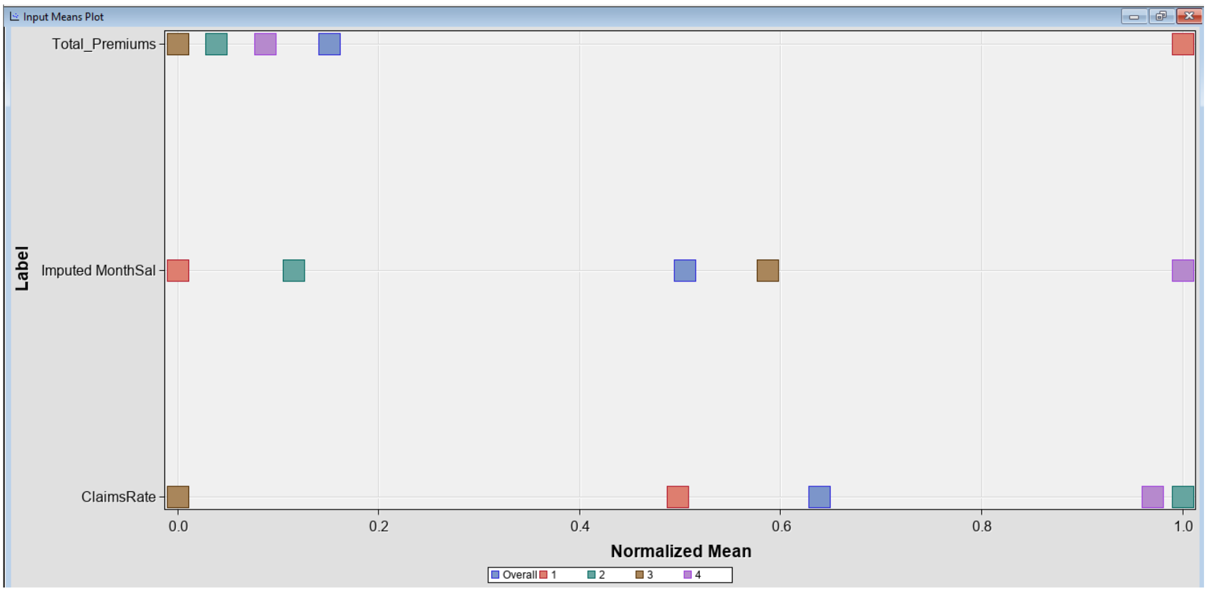


Figure 16 - Input means plot for k=4

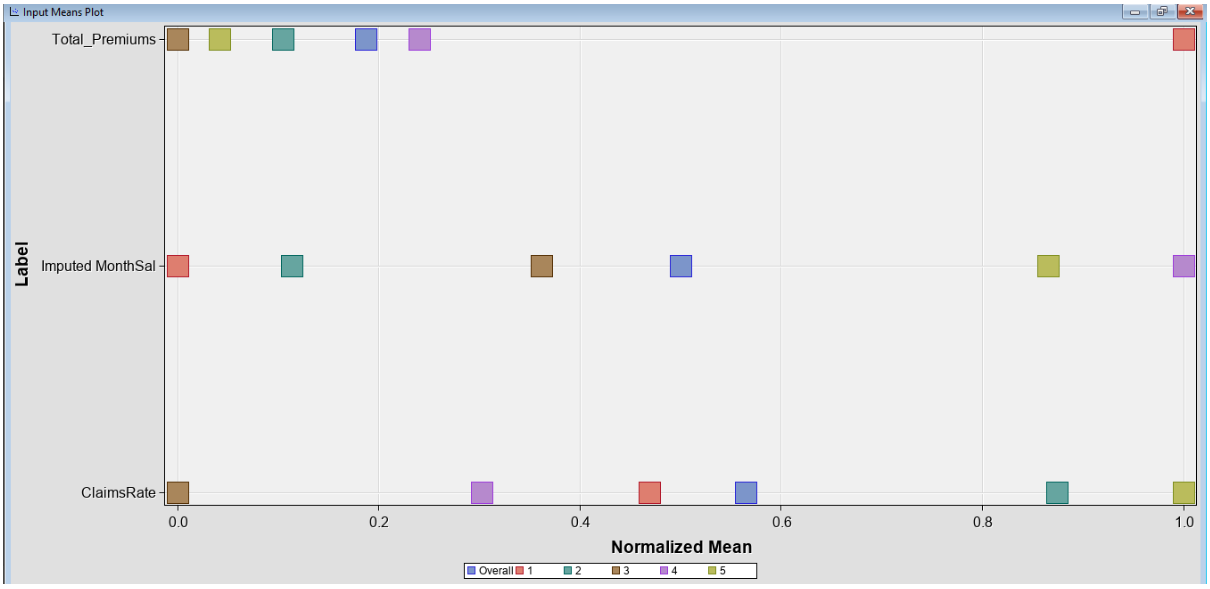


Figure 17 - Input means plot for k=5

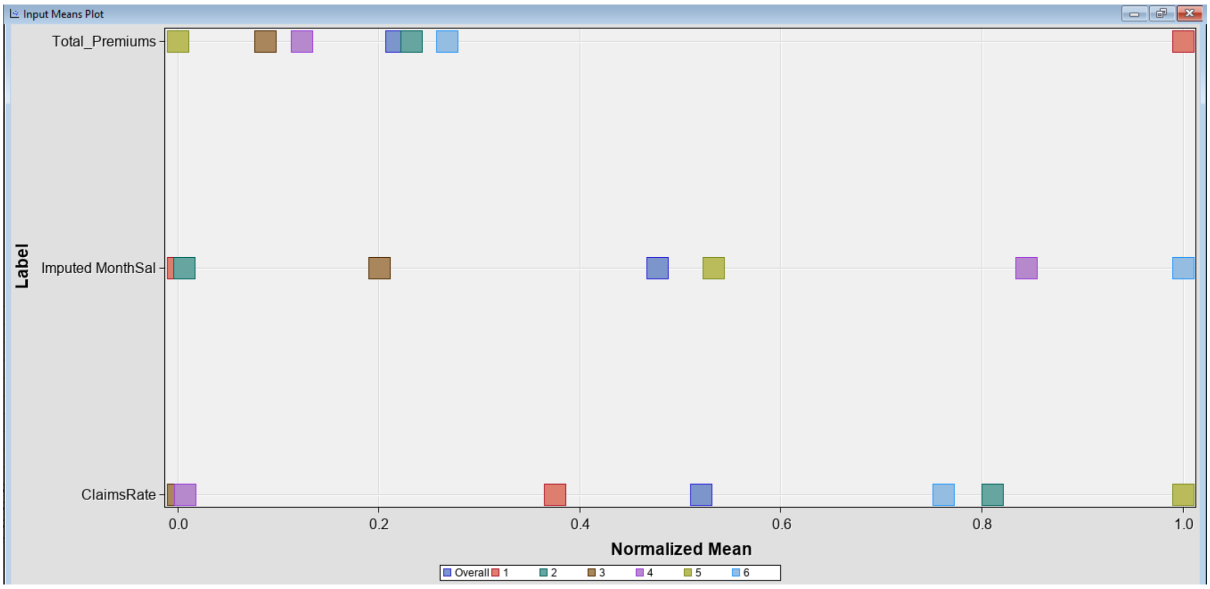


Figure 18 - Input means plot for k=6

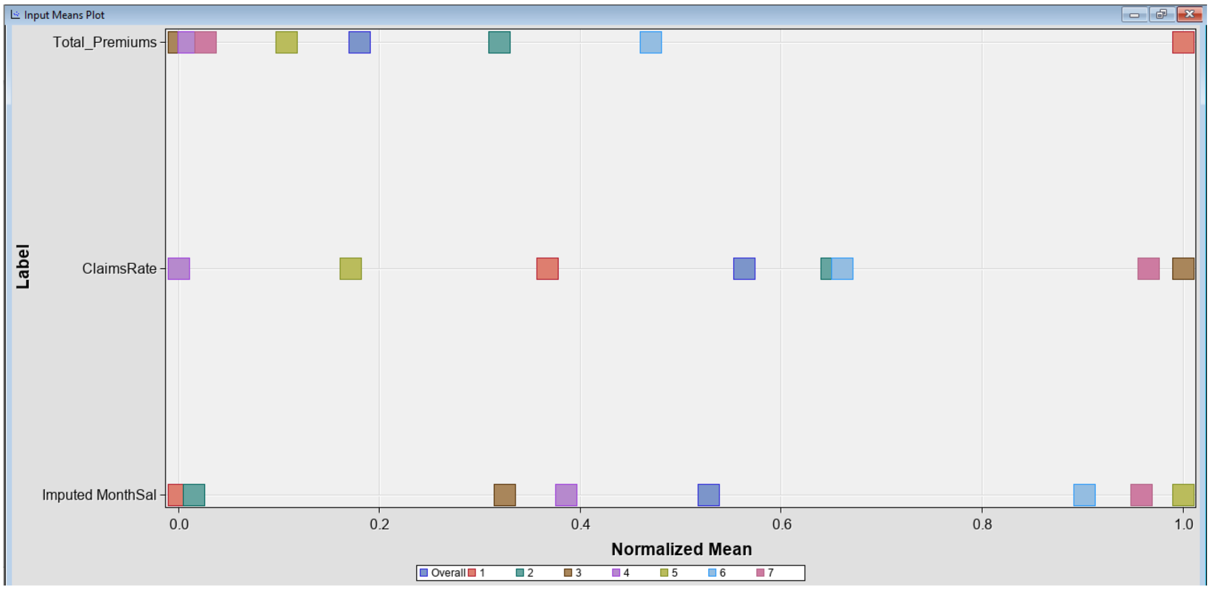
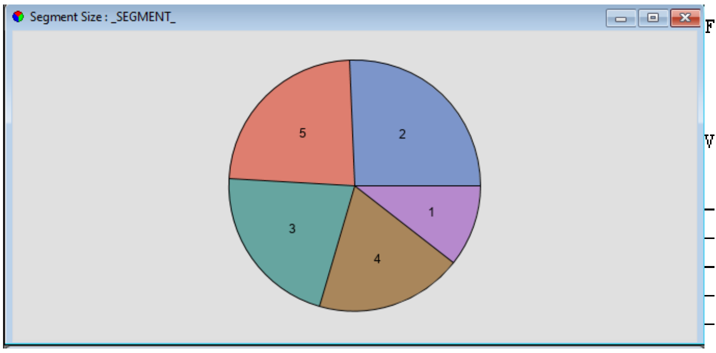


Figure 19 - Input means plot for k=7

Distribution of samples among segments

As was previously mentioned, and taking k=5, the number of samples becomes reasonably well distributed among the 5 segments.



```
Frequencies:  _SEGMENT_

                                                  Percent of
 Segment      Segment    Frequency       Total
Variable       Value        Count       Frequency

_SEGMENT_        2           2627        25.5818
_SEGMENT_        5           2409        23.4590
_SEGMENT_        3           2197        21.3945
_SEGMENT_        4           1959        19.0768
_SEGMENT_        1           1077        10.4879
```

Figure 20 - Samples distribution across segments

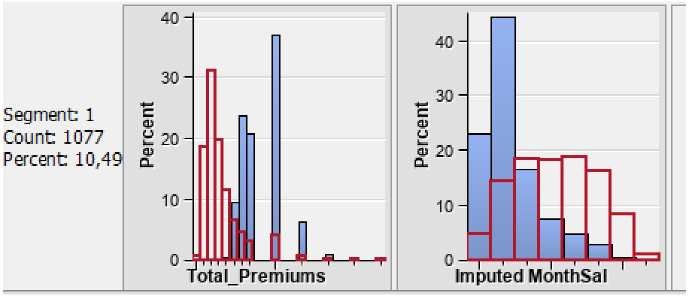
At this moment, since clusters are defined as the points that are within each group, it is time to reference each segment in detail, constituting personas and a defined strategy given the purpose.

Segments characteristics

Segment 1

As was previously depicted in Figure 20, this segment represents 1.077 “points”, accounting for 10,5% of our sample.

Going more into detail on what the characteristics are concerned, the average born year is 1988, with an average salary of 1.433$. The total premiums average is 1.171$, being the approx. total amount paid by the insurance 720$. 18% of this segment has a bachelor, master or PhD, when compared to more than 53,4% of the initial sample.



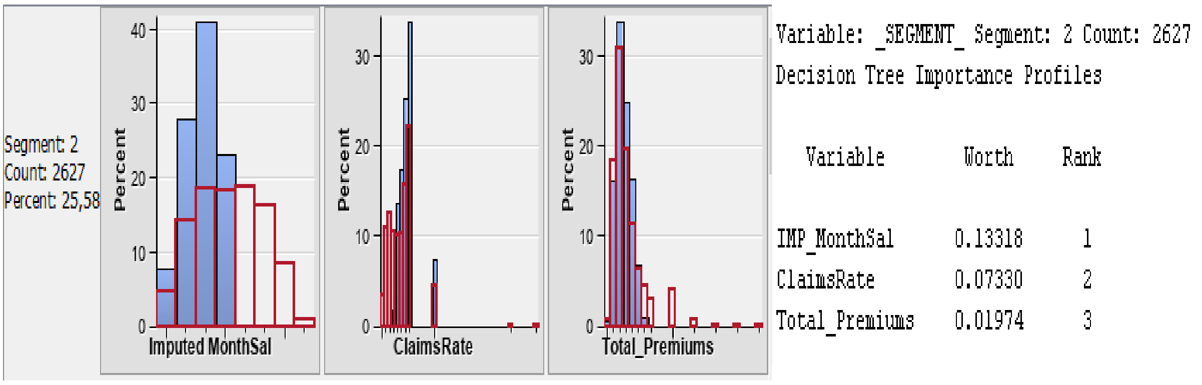
```
Variable:  _SEGMENT_  Segment: 1  Count: 1077
Decision Tree Importance Profiles

   Variable        Worth      Rank

Total_Premiums    0.13359       1
IMP_MonthSal      0.03974       2
ClaimsRate        0.00197       3
```

Figure 21 - Segment 1 SaS Miner metrics

Segment 2

As was previously depicted in Figure 20, this segment represents 2.627 “points”, accounting so for 26,6% of our sample.

Going more into detail on what their characteristics in concerned, the average born year is 1982, with an average salary of 1.635$. The total premiums average is 713$, being the total approx. amount paid by the insurance 632$. 51,4% of this segment has a bachelor, master or PhD, being this number a little bit below the average sample which is 53,4%.



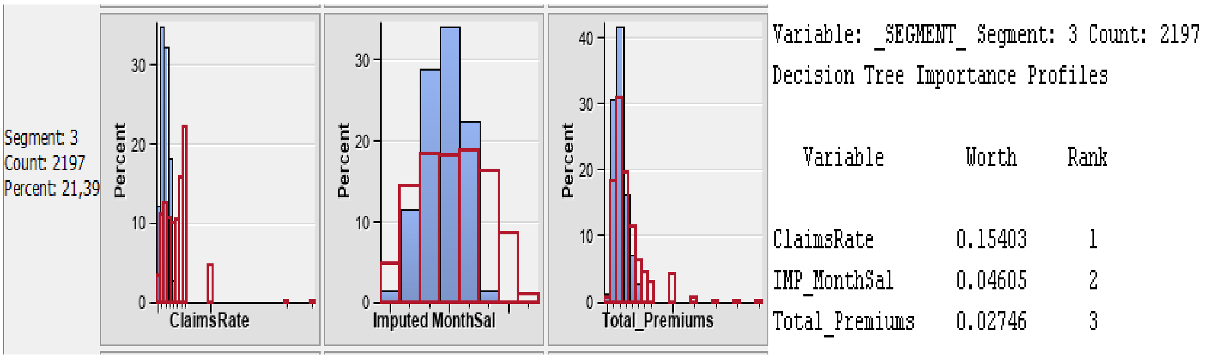
```
Variable: _SEGMENT_ Segment: 2 Count: 2627
Decision Tree Importance Profiles

   Variable        Worth     Rank

IMP_MonthSal      0.13318      1
ClaimsRate        0.07330      2
Total_Premiums    0.01974      3
```

Figure 22 - Segment 2 SaS Miner metrics

Segment 3

As was previously depicted in Figure 20, this segment represents 2.139 "points", accounting so for 21,4% of our sample.

Going more into detail on what their characteristics in concerned, the average born year is 1973, with an average salary of 2.204$. The total premiums average is 659$, being the total approx. amount paid by the insurance 197$. 65% of this segment has a bachelor, master or PhD, which is higher than the sample average (53,4%).



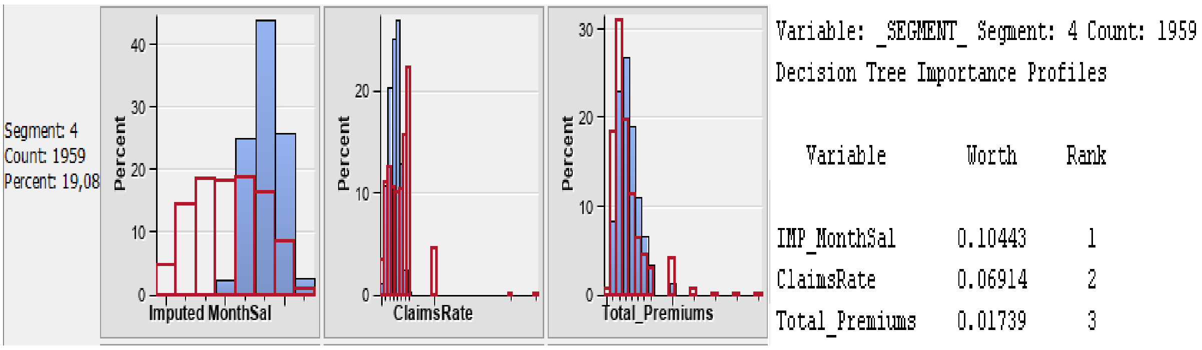
```
Variable: _SEGMENT_ Segment: 3 Count: 2197
Decision Tree Importance Profiles

   Variable        Worth     Rank

ClaimsRate        0.15403      1
IMP_MonthSal      0.04605      2
Total_Premiums    0.02746      3
```

Figure 23 - Segment 3 SaS Miner metrics

Segment 4

As was previously depicted in Figure 20, this segment represents 1.959 "points", accounting so for 19,1% of our sample.

Going more into detail on what their characteristics in concerned, the average born year is 1950, with an average salary of 3.561$, with average total premiums of 782$, being the total amount paid by the insurance 397$. 54% of this segment has a bachelor, master or PhD, which is in line with the total sample numbers (53,4%).



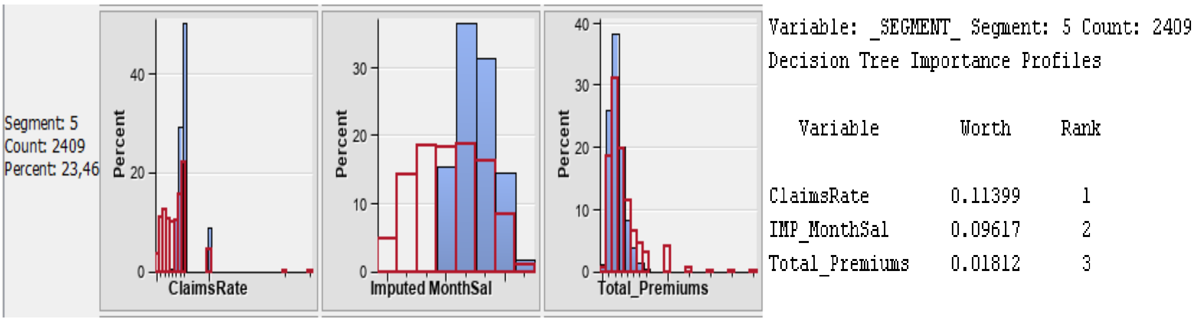
```
Variable: _SEGMENT_ Segment: 4 Count: 1959
Decision Tree Importance Profiles

   Variable        Worth     Rank

IMP_MonthSal      0.10443     1
ClaimsRate        0.06914     2
Total_Premiums    0.01739     3
```

Figure 24 - Segment 4 SaS Miner metrics

Segment 5

As was previously depicted in Figure 20, this segment represents 2.4069 "points", accounting so for 23,5% of our sample.

Going more into detail on what their characteristics in concerned, the average born year is 1955, with an average salary of 3.275$. The total average premiums is 681$, being the approx. total amount paid by the insurance 664$. 60,7% of this segment has a bachelor, master or PhD, which is above the average considering the hole sample (53,4%).



```
Variable: _SEGMENT_ Segment: 5 Count: 2409
Decision Tree Importance Profiles

   Variable        Worth     Rank

ClaimsRate        0.11399     1
IMP_MonthSal      0.09617     2
Total_Premiums    0.01812     3
```

Figure 25 - Segment 5 SaS Miner metrics

Segments resume

As a resume, we depict the characteristics for each segment.

Table 9 - Segments characteristics

Segment	Average birth year	Average monthly salary ($)	Average total premiums ($)	Average Approx. Amount paid by insurance ($)	Average premiums – Approx. Amount paid by insurance ($)
1	1.988	1.433	1.171	720	451
2	1.982	1.675	713	632	81
3	1.973	2.204	659	197	462
4	1.950	3.561	782	397	385
5	1.955	3.275	681	664	20
Total average	**1.968**	**2.498**	**755**	**511**	**280**

Table 10 – Educational level of each segment

Educational level \ Segment	1	2	3	4	5	Total
1 - Basic	438	329	156	183	165	1.271
2 - High School	445	949	613	718	783	3.508
3 - BSc/MSc	193	1.180	1.213	937	1.270	4.793
4 - PhD	1	169	215	121	191	697
Total	**1. 077**	**2.627**	**2.197**	**1.959**	**2.409**	**10.269**

Strategy and personas

Remembering the clustering/segmentation purpose that was explained in the chapter “Purpose”, namely the CRM and the profitability/potential objectives, one persona was “designed” to each cluster with a defined strategy.

Table 11 - Persona and strategy for each segment

Segment/Cluster	Persona	Strategy
1 – “youngsters”	**Mary** Mary is a 33 years old doctor who recently started her career. Mary has a lot of insurances, since she does not want to have problems since she wants to focus on her career at 100% before she starts a family. Mary has a modern lifestyle, bellowing thus to the “millennial” generation.	Due to Mary’s modern lifestyle, the A2Z insurance should develop an omnichannel engagement strategy in order to increase loyalty since there is not “much room” for up and cross sell due to with premiums. As so, and to this segment, A2Z insurance company must target these segment just with content in order to engage them.
2 – “Not so special”	**Radamel** Radamel is a 40 years old Colombian immigrant who came to USA just 10 years ago. Radamel made some insurances in order to establish his life when he arrived at the States. He is married and has two kids. He started his own business just 4 months ago (Colombian restaurant).	Since Radamel has a low profitability, being still young, there an opportunity for up and cross sell. As so, the focus on this segment must be growth through the implementation of a flirting campaigns via e-mail[21].
3 – “wanted ones”	**John** John is a well establish man with an important role at a bank. John took several specializations at universities of prestige all over the world. John has 3 children. He is also interested on politics, economics and investments.	Due to John’s high level of education and sophistication, relevant content related to his interests must be brought to him. Since just a small percentage of his income was allocated to premiums, there is an opportunity for up and cross sell. So, our team of consultants advises the construction of a growth channel to this segments, utilizing all channels.
4 – “oldies”	**Serafim** Serafim has 75 years old. Due to his age, he is starting to have some health problems. Serafim rarely uses computers or mobile phones. Serafim had his own company, that he left to his 3 sons 4 years ago after a lung problem. Besides the fact of having a “stable life”, he likes to maintain his routines, being nicknamed by his family as “the hyperactive one”.	Comparing the total premium to the average monthly salary (3.561 vs 782), we can easily conclude that there is an enormous potential in this segment, namely on what products and services related to health is concerned. However, to this segment average age, traditional channels must be considered such as post and billboards usage.
5 – “unwanted ones”	**Florence** Florence is a 68 years old grandmother who already had some health problems. Florence lives in the countryside, since she got retired from being a teacher (3 years ago). Florence already have a stable life. Florence doesn’t want to have problems anymore.	As could be easily checked by the indicators, this segment has a huge potential for up and cross sell. However, the main difference between this segment and the “oldies” one, is that this segment is not profitable at all. As so, the urgency for action is higher in this segment.

[21] E-mails marketing is still nowadays the strategy that has the highest conversion rates

Conclusion and final remarks

In fact, this document represents the meet of a challenge, the challenge of clustering and define a strategy for a database of 10.296 Customers of A2Z insurance company. Throughout the work, the team corrected and explored the data, correcting missing values and removing/defining outliers, keeping at the end 10.269 observations.Customers.dle of the process, two new variables were constructed to possibly enrich the segmentation that is going to be done. Furthermore, the correlations among the variables were carefully studies in order to select the ones that mean the most[22] giving cluster's objective[23]. At the end, and after studying the perfect of clusters given the defined criteria, personas were defined to each segment, as a strategy to approach each segment.

Last but not the least, and most important, the approach that the consultant team took to this segmentation is not unique, but just a well studies and creative proposal for a challenge that was made.

Bibliography

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques. 3rd Edition. Elsevier.

Berry, J. A Michael, Linoff, S Gordon, Data Mining Techniques, 2th Edition, Wiley Publishing, Inc, 2004

Hair, J., Anderson, R., Tattham, R, and Black, W, Multivariate Data Analysis, 7 th Edition, Prentice Hall, 2019.

LOTAME (n.d). What is market segmentation? Retrieved from https://www.lotame.com/what-is-market-segmentation/

[22] Gross monthly salary, total premiums and claims rate

[23] CRM and take advantage of the profitability/potential that the actual customers database could still give to A2Z insurance company