

**Data Mining Project**

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

**A2Z INSURANCE – A CUSTOMER SEGMENTATION**

<Group Alias>

David Santos: <student number>

Fazoal Islam: <student number>

Tomás Jordão: 20210664

January - 2021

INDEX

[1. Introduction iii](#_Toc92182898)

[2. Exploration of the Dataset iv](#_Toc92182899)

[3. Transforming the Dataset v](#_Toc92182900)

[3.1. Missing Values v](#_Toc92182901)

[3.2. Creating New Variables v](#_Toc92182902)

[3.3. Outliers v](#_Toc92182903)

[3.3.1. Thresholding v](#_Toc92182904)

[3.3.2. IQR vi](#_Toc92182905)

[3.3.3. Z-Score vi](#_Toc92182906)

[3.3.4. Local Outlier Factor – LOF vi](#_Toc92182907)

[3.3.5. One Class SVM vi](#_Toc92182908)

[3.3.6. Technique used for removing outliers - Z-Score and LOF vi](#_Toc92182909)

[3.4. Feature Selection vii](#_Toc92182910)

[4. Title 2 viii](#_Toc92182911)

[4.1. Title 2.1 viii](#_Toc92182912)

[4.1.1. Title 2.1.1 ix](#_Toc92182913)

[5. References x](#_Toc92182914)

[6. Appendix (Doesn’t count for the 10page limit) xi](#_Toc92182915)

# Introduction

In this report we will show a customer segmentation model that has the objective to find clusters and understand the customer´s behavior of an insurance company A2Z Insurance, a fictional Portuguese company. In order to perform this study, we analyzed a total of 13 variables and 10296 observations.

During this project various data preparation techniques were applied such as treatment of data incoherence, missing values, filtering of outliers, data normalization together with clustering techniques like k-means, hierarchical clustering, t-SNE.

Using these techniques and with the insights that they provided us about the data we were able to construct various groups of customers that represent the patterns present in the data and the different customers’ profiles.

In the end considering all the previous work on the development of the clusters and with the knowledge we gained about the data we were able to develop a marketing approach that we think best suits each of the different customers’ profiles.

# Exploration of the Dataset

First, before performing any clustering technique we analyzed and explored the dataset to better understand the variables and any data nuances that may exist. We checked the number of variables, their data types and any incoherence they may have.

We began by loading the a2z\_insurance dataset and set the variable “CustID” as index. Following this action, we looked at the first five rows of the dataset and the variables data types where we observed that the dataset is composed of 10296 records and 13 variables.

After this first observation we noticed that there were various problems with two of the variables, “BirthYear” and “FirstPolYear”. The main problem was that there were 19.4% customers with birthyear that was more recent than the first policy year or the year they became customers of the insurance company. To deal with this problem we decided to drop these two variables because we were unable to determine from which variable the problem originated and the number of records with the problem was great.

In the next step we began by defining the metric and non-metric features, checked for missing values, checked the distribution of the variables and created new variables.

We decided that the variables, “EducDeg”, “Children” and “GeoLivArea” should be considered non-metric features and the rest of the variables are metric.

Regarding the missing values we observed that there were several variables with missing values as it can be seen in the annex XX. These missing values were then filled using measures of central tendency as is explained in the next section of this report.

In relation to the distribution of the variables we used the method “describe(include = ‘all’)” from pandas that allowed us to observe a number of important characteristics of each variable, such as the frequency, unique values, mean, standard deviation and quantiles as it can be seen in annex XX. These metrics showed several important aspects that permitted us to better understand the data.

The first being that the most common academic degree is BSc/MSc, the second is that there are more customers with children than without and finally we were able to observe that most of the variables present in this data set presented outliers. This assumption is easily seen by observing the annex XX where most variables show that the variable max is very distanced from the variable mean implying that it can be outliers.

# Transforming the Dataset

## Missing Values

In this step we began by identifying and selecting techniques to deal with the values that were missing in our dataset. We identified the following number of outliers as it is shown in the table below:



As is possible to observe, there are missing values in both numerical variables and categorical variables.

Considering for that fact we decided to use two techniques:

* For categorical variables or non-metric features we decide to fill the missing values using the mode.
* For numerical variables or metric features we decided to fill the missing values using the median mainly because it is not affected by outliers

## Creating New Variables

In this step we only created a new variable that we thought would give good knowledge about each customer average spending in terms of types of insurances. This new variable “Avg\_Premiuns” give us the mean spent by each customer in insurances.

## Outliers

In this step we used different techniques for identifying and filtering outliers. Those techniques were thresholding, IQR, Z-Score, Local Outlier Factor (LOF) and One Class SVM.

### Thresholding

Firstly, we began by doing a filtering where we removed some outliers. These was necessary since when constructing the histograms and boxplots we were unable to get any relevant information of them.

After this initial filtration we constructed boxplots and histograms and boxplots, seen in annex XX, that enabled us to define thresholds that removed approximately 1.1% of observations we considered outliers.

### IQR

The next technique we tried was intra quartile range. When using the standard values for the classification of observations as outliers we observed that we would remove approximately 14.6% of the observations.

Due to this fact and although we tried different values for defining the upper and lower limit, we could never comply with the rule of thumb of not removing more than approximately 3% of the observations we decided not to use this technique for filtering the outliers from our dataset.

### Z-Score

This technique defines as outlier observations that are at a predefined value of standard deviations of the mean value of what is being observed.

In our case we defined as outlier an observation we a z score greater than 3.5 and considering this value we removed about 2.55% of the observations in our dataset, a value that complies with the rule of thumb stated previously.

### Local Outlier Factor – LOF

LOF is and unsupervised anomaly detection method that measures the local density deviation of a data point in relation with its neighbours. This technique considers outliers observations with substantially lower density than their neighbours.

The main parameter of LOF is “n\_neighbors” and based on the documentation found in scikit-learn a value of 20 usually works well.

In our case we needed to define a value of 50 for the parameter “n\_neighbors” that contradicts the documentation but that could be linked with the characteristics of our dataset.

### One Class SVM

One Class SVM is an unsupervised outlier detection technique that uses support vector machines. It has two main parameters the “kernel” and “nu”. The first parameter cobined with the second define a frontier where observations that lie outside are considered as outliers. The parameter “nu” corresponds to the probability of finding a observations outside of that frontier.

In our problem we used the default value for the parameter “kernel” and decided to use the value 0.025 for the parameter “nu” complying with the rule of thumb stated above for the removal of observations from the dataset.

### Technique used for removing outliers - Z-Score and LOF

For the final filtration of outliers, we decided to combine two of techniques mentioned previously, Z-Score and LOF. First, we removed observations that presented a z-score value lower than 3.5. After this, we used LOF with a value of 50 for the parameter “n\_neighbors” to remove any outliers that were not considered by the z-score method.

In the end, we removed 3.02% of the observations a value that complies with the rule of thumb stated previously.

## Feature Selection

To select the relevant metric features we utilized a correlation matrix that can be seen in annex XX. We began by constructing scatterplots that enabled us to identify relationships between numerical variables as can be seen in annex XX. After observing these graphs, we realized that some of the relationships may not be linear. Considering this fact, we decided to use the method “spearman” to construct the correlation matrix.

In the end we decided to eliminate the variable “PremMotor” because presented a considerable correlation with various variables.

# Title 2

## Title 2.1

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.



Figure 2.1 – Illustrative figure

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

|  |  |
| --- | --- |
| **Title** | **Title** |
| Text | Number |
| Text | Number |
| Text | Number |

Table 2.1 – Illustrative table

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

### Title 2.1.1

Example of unnumbered list:

* Item 1
* Item 2
* Item 3

#### Title 2.1.1.1

Example of numbered list:

1. Item 1
2. Item 2
3. Item 3

# References

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

# Appendix (Doesn’t count for the 10page limit)