

**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](#_Toc91785998)

[2. Exploration of the Dataset iv](#_Toc91785999)

[3. Transforming the Dataset v](#_Toc91786000)

[3.1. Title 1.1 v](#_Toc91786001)

[3.1.1. Title 1.1.1 v](#_Toc91786002)

[4. Title 2 vi](#_Toc91786003)

[4.1. Title 2.1 vi](#_Toc91786004)

[4.1.1. Title 2.1.1 vii](#_Toc91786005)

[5. References viii](#_Toc91786006)

[6. Appendix (Doesn’t count for the 10page limit) ix](#_Toc91786007)

# 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

## Title 1.1

### Title 1.1.1

**The following instructions should be followed for the writing of the report:**

This word file provides a "standard" proposal structure. The format of the report (headers, spacings, fonts, and other formations) should be employed as defined in this template. However, students may change the structure and the titles according to their needs.

The report should be written in English.

In the case references point to the “source code” make sure the code is correctly annotated.

All bibliographic references should be presented in the APA standard; this standard also applies the formatting of references and respective forms of referencing throughout the text[[1]](#footnote-1).

# 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)

1. Example of a footnote. [↑](#footnote-ref-1)