

**Data Mining Project**

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

**<TITLE>**

Group AN

Inês Rocha, number: 20220052

Isabel Dias, number: 20191215

Joana Sousa, number: 20191205

January, 2022

INDEX

[1. Introduction iii](#_Toc123600599)

[2. Data iii](#_Toc123600600)

[1.1. Introduction iii](#_Toc123600601)

[1.1.1. Title 1.1.1 iii](#_Toc123600602)

[2. Title 2 iv](#_Toc123600603)

[2.1. Title 2.1 iv](#_Toc123600604)

[2.1.1. Title 2.1.1 v](#_Toc123600605)

[3. References v](#_Toc123600606)

[4. Appendix (Doesn’t count for the 10page limit) v](#_Toc123600607)

# Introduction

In today's competitive business environment, it is essential for organizations to understand their customers in order to effectively meet their needs and drive growth. One way to achieve this understanding is through market segmentation, the process of identifying groups of customers with similar characteristics and behaviors. By understanding the differences between these segments, organizations can make more informed strategic decisions about opportunities, product definition, positioning, promotions, pricing, and target marketing.

Clustering is a powerful data mining technique that can be used to identify and understand customer segments. It involves grouping similar data points together into clusters, based on their characteristics and behaviors. In this report, we will apply clustering techniques to the customer database of A2Z Insurance, a Portuguese insurance company, in order to identify and understand the different segments within the customer base. Our goal is to not only understand the demographics and value of each segment, but also to identify opportunities for targeted marketing and cross-selling of insurance products.

A2Z Insurance has provided us with a sample of 10,290 customers from its active database, along with data on their characteristics and behaviors. Our analysis will involve cleaning and preprocessing the data, selecting appropriate clustering algorithms, and evaluating the resulting clusters to determine their relevance and value to the organization. By using data-driven approaches to market segmentation, A2Z can better serve its existing customers and improve its targeting of prospective customers.

# Data

We started by importing a variety of libraries for data manipulation, visualization, and Clustering. We loaded the A2Z Insurance dataset, which includes a total of 13 variables, into a pandas DataFrame using the pyreadstat library and set the CustID column as the index. The nunique() method was used to check for duplicate rows based on the CustID column. With these initial steps, we have successfully loaded and prepared the dataset for further analysis.

# Data Exploration

The data exploration process for the A2Z Insurance dataset begins by checking for missing values in the data using the isnull() and sum() methods. The data types of the variables *BirthYear*, *FirstPolYear*, *GeoLivArea*, and *Children* are changed to integers due to being discrete variables, using the astype() method. The describe() method is used to generate summary statistics for the numerical variables where we are able to see that the variable *BirthYear* has an impossible minimum value and the variable *FirstPolYear* has an impossible maximum value.. The describe(include="O") method is used to generate summary statistics for the categorical variables, including count and unique values.

Next, we check for inconsistencies and duplicates in the data. Using the value\_counts() method on the variable *BirthYear*, we conclude that there is only one value that is inconsistent: 1024. We proceeded to check if there are any observations where the first year of policy comes before the birth year. There are 1997 records where this happens, which is a very significant amount of our dataset. We check that our dataset has 117 minors, which cannot have insurance contracts in their name in Portugal. We also verify that for each education category there isn’t any record that seems too young.

We generate value counts and distribution graphs for various variables to gain a better understanding of their characteristics and patterns. With this we see that the *EducDeg* variable does have missing values that are masked as [help como descrever “”]. We change them to Nan values, to have all our missing values marked the same way.

The distribution graphs help us checking for extreme or outlying values that may require further investigation or treatment. We identify several variables with extreme or outlying values which are *MonthSal, CustMonVal*, ClaimsRate, *PremMotor*, *PremHousehold*, *PremHealth*, and *PremWork* have several or extreme outliers. In the following outlier analysis, these values will be explored in more depth and a decision will be made on how to handle them. In this analysis, also several variables are found to have skewed distributions, namely *CustMonVal*, *ClaimsRate*, *PremHousehold*, *PremLife*, and *PremWork*. These variables may require further investigation or transformation in order to better inform the clustering process.

## 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)