

**Data Mining Project**

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

**A2Z Insurance – Insurance Company**

Group EU

Diogo Morgado, number: 20221392

Tomás Gama, number: 20221354

December, 2022

INDEX

[1. Introduction iii](#_Toc122788600)

[2. Data Exploration iv](#_Toc122788601)

[2.1. Handling Missing Values iv](#_Toc122788602)

[2.2. Handling Outliers iv](#_Toc122788603)

[3. Data Preprocessing v](#_Toc122788604)

[3.1. Feature Engineering v](#_Toc122788605)

[3.2. Coherence checking v](#_Toc122788606)

[4. Data Partition vi](#_Toc122788607)

[4.1. Feature Selection vi](#_Toc122788608)

[4.2. Data Standardization/Normalization vi](#_Toc122788609)

[5. Dimensionality reduction vii](#_Toc122788610)

[6. Data Clustering viii](#_Toc122788611)

[6.1. K-means viii](#_Toc122788612)

[6.2. SOM viii](#_Toc122788613)

[6.3. K-means on top of SOM units viii](#_Toc122788614)

[6.4. Density based clustering viii](#_Toc122788615)

[6.5. Mean shift clustering viii](#_Toc122788616)

[6.6. DBSCAN viii](#_Toc122788617)

[6.7. GMM viii](#_Toc122788618)

[6.8. K-means and Hierarchical clustering viii](#_Toc122788619)

[7. Cluster Analysis ix](#_Toc122788620)

[7.1. Cluster Visualization using t-SNE ix](#_Toc122788621)

[8. References x](#_Toc122788622)

[9. Appendix (Doesn’t count for the 10page limit) xi](#_Toc122788623)

# Introduction

Acquiring new customers is essential for the success of any business. One way to do this is by gathering information from current customers and using it to understand the needs and preferences of different market segments. This includes analyzing factors such as geographic location, age, personality, and purchasing habits. By dividing the market into smaller groups based on these characteristics, companies can tailor their marketing strategies and make informed decisions about product development, pricing, and targeted advertising. By understanding and effectively targeting specific customer segments, businesses can improve their ability to meet the needs of their current customers and attract new ones.

A2Z Insurance is a reputable Portuguese insurance company that offers a range of services, including motor, household, health, life, and work compensation insurance. A2Z primarily serves customers in Portugal, but a significant number of new customers also come from the company's website. Customers have the option to sign up for A2Z services through branches, by phone, or online.

To effectively understand and target specific customer segments, we used various approaches and perspectives to segment the customers and analyzed the results. A2Z can benefit from gaining insight into the value and demographics of each segment, as well as determining which types of insurance are most appealing to them. This can help the company better serve its customers and make informed decisions about marketing and product development.

# Data Exploration

We start by exploring the dataset and performing some initial exploration tasks. First, we rename the columns to make them easier to understand. Then, we retrieve information from all columns and check for any null values. We also verify that the data types are correct, and we notice that the "BirthYear" and "First\_Policy" columns have float data types, which we need to change to integer because they represent years. During this process, we also observe that the dataset may contain outliers and that several columns have missing values (as shown in Figure 1). To address the missing values, we consider various imputation methods and decide to use Mode imputation and Median Imputation, as the percentage of missing data is not too high to justify the use of other methods.

## Handling Missing Values

We decided to approach the fact of having some missing values, in diverse ways for each case. In the variables “First\_Policy”, “Birth\_Year”, “Salary” and “Area” we decided to fill the missing values with the median, between the data we have for each variable. For this we used the function median(), which splits the higher half of the data or probability distribution from the lower half. For the variable “Children”, we assumed that the Nan means that they don´t have kids, so we filled with the zero value. Finally, for the variable “Education” we filled the missing values using the mode, with the method mode(), that provides with the values that appear most often. Concluding, we do a last check to see if we missed some value, and we transform the variables “First\_Policy”, “Birth\_Year”, “Salary” and “Area” to Integer type.

## Handling Outliers

First, we split the variables in metric and non-metric features, now we take a visualization of the non-numeric and numeric variables before the outlier removal (Figure).

Next, we opted for using two different methods of removing the outliers, manually and using the IQR method. Using the manual method, we defined by ourselves, with graphic assistance and interpretation, to remove the values which we thought would be right to remove, this method can vary from session to session, because it depends on the interpretation and the view, may be different from person to person, using this method we kept 77% of the data.

Using the IQR method we managed to keep 85% of the data. This method consists of defining an upper and lower limit of the quantile removing the values that are out of the range (Figure).

# Data Preprocessing

After exploring the dataset, we have a better understanding of the variables and how we can use them for data clustering. To enhance the performance of the clustering, we perform feature engineering to create new variables that might give us an advantage. We also conduct a coherence check to ensure that the data in our dataset is consistent and make sense, and we remove any outliers that we find during this process.

## Feature Engineering

During the feature engineering process,process of creating and selecting features that can improve the performance of machine learning models**,** we created several new variables: "Age", "Customer\_Years", "Total\_Premium", and "Salary\_Rate". To obtain the "Age" variable, we subtracted the year of the database from the "Birth\_Year". We calculated the "Customer\_Years" variable by subtracting the current year of the database from the year of the "First\_Policy". We transformed Education into Ordinal Encoding, meaning 4 would be corresponding to PhD and 1 to Basic Education, and all negative values of premiums into 0, because it meant the customer had already left the insurance or did already pay, and created the "Total\_Premium" variable as the sum of all the customer's premiums. Finally, we created the "Salary\_Rate" variable by dividing the "Total\_Premium" by the customer's annual salary (calculated by multiplying their salary by 14) and multiplying the result by 100, which gives us the rate of salary they invest in the company's insurance.

## Coherence checking

When performing coherence checking, we looked for values in the dataset that did not make sense. We began by analyzing the difference between the "Age" and "Customer\_Year" variables to ensure that a customer cannot be older than they have been a customer of the insurance company. We also checked for underage individuals with children, individuals with a PhD but not the minimum age to obtain one, and individuals with ages below the minimum required for a BSc/MSc. As these values could potentially influence the results, we removed them from the dataset.

# Data Partition

In this step we do a Data Partition, consisting of splitting in Demographic and Insurance Data. The Demographic Data contains “Birth\_Year”, “Age”, “Education”, “Salary”, “Area”, “Children”, basically the information more focused on the client’s information and Insurance Data the information related about the insurance company such as “First\_Policy”, “CMV”, “Claims”, “Motor”, “Household”, “Life”, “Health”, “Work”, “Customer\_Years”, “Total\_Premium”, and “Salary\_Rate”.

## Feature Selection

For feature Selection we plotted two correlation matrixes, each one for Demographic and Insurance Data. After interpreting both the graphics and the correlation between each variable we decided to drop “Birth\_Year”, “Age” from Demographic Data and “Total\_Premium”, “Claims” from the Insurance Data, due to high correlation.

## Data Standardization/Normalization

# Dimensionality reduction

Despite previously performing feature selection, we decided to use dimensionality reduction to further simplify the dataset while preserving as much information as possible. The high dimensionality of the data can make it difficult to process and visualize, and it can also increase the risk of overfitting. Therefore, we applied principal component analysis (PCA) to our insurance dataframe to improve the output and performance of machine learning algorithms. From the results of PCA, we selected 4 components and applied PCA again with those components. We then interpreted the values of each principal component and decided to drop "PC2" because it did not add additional value to the dataframe.

In conclusion, using principal component analysis (PCA) for data clustering can be an effective way to reduce the dimensionality of the data and improve the performance of clustering algorithms. By transforming the data into a new set of linearly uncorrelated variables called principal components, PCA allows us to capture the most important information in the data and use it to cluster the observations into groups.

# Data Clustering

Data clustering is a technique used to group similar observations together into clusters, based on their characteristics and patterns. In the context of an insurance company, data clustering can be used to understand and analyze the characteristics of different groups of policyholders, identify trends and patterns in their behaviors and characteristics, and make informed decisions about how to best serve their needs.

By applying data clustering to insurance data, companies can better develop targeted marketing and underwriting strategies. Clustering can also be used to identify subgroups within the policyholder population that may have different needs or preferences, and tailor insurance products and services to meet those needs.

## K-means

## SOM

## K-means on top of SOM units

## Density based clustering

## Mean shift clustering

Based on the results of mean shift clustering using a bandwidth of, approximately, 2.31, the algorithm identified a total of 12 clusters in the data. The R^2 value of 0.2556 for the cluster solution indicates that the clusters can explain approximately 26% of the variance in the data.

One possible interpretation of these results is that the data exhibits a moderate degree of structure, with some clear clusters and some overlap between clusters. The relatively low R^2 value suggests that there may be additional factors influencing the data that are not captured by the clusters.

Overall, the results of mean shift clustering using a bandwidth of, approximately, 2.31 and identifying 12 clusters suggest that the data exhibits a moderate degree of structure, but additional factors may also be influencing the data. Careful evaluation of the clusters is necessary to determine their usefulness for the specific application.

## DBSCAN

Based on the results of DBSCAN clustering, the algorithm identified a total of 2 clusters in the data, with 218 rows classified as noise. The R^2 value of 0.1124 for the cluster solution indicates that the clusters can explain approximately 11% of the variance in the data.

One possible interpretation of these results is that the data exhibits a relatively low degree of structure, with only a small number of distinct clusters and a large proportion of observations classified as noise. The low R^2 value further suggests that the clusters do not capture a significant amount of the variance in the data.

## GMM

Based on the results of Gaussian mixture model (GMM) clustering, we selected a model with 6 components based on the criteria of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The R^2 value of 0.3258 for the cluster solution indicates that the clusters can explain approximately 33% of the variance in the data.

One possible interpretation of these results is that the data exhibits a moderate degree of structure, with distinct clusters and some overlap between clusters. The relatively high R^2 value suggests that the clusters can capture a significant amount of the variance in the data.

Overall, the results of GMM clustering using 6 components suggest that the data exhibits a moderate degree of structure, and the clusters identified by the algorithm can capture a significant amount of the variance in the data. Careful evaluation of the clusters is necessary to determine their usefulness for the specific application.

## K-means and Hierarchical clustering

Based on the results of applying K-means on top of hierarchical clustering to the data, we visualized the R² scores for each cluster solution on demographic variables and insurance variables and selected the right clustering algorithm and number of clusters for each perspective. We decided to go with 4 clusters for each perspective and merged the clusters using hierarchical clustering. We then defined the dataframe centroids for the clusters previously defined, and through our threshold defined 6 clusters for the hierarchical clustering solution.

One possible interpretation of these results is that the data exhibits a moderate degree of structure, with distinct clusters and some overlap between clusters. The use of both K-means and hierarchical clustering allowed us to identify clusters based on both the demographic and insurance variables, and the R² scores suggest that the clusters can capture a significant amount of the variance in the data.

It is important to carefully evaluate the quality and interpretability of the clusters to determine their usefulness for the specific application. Some potential ways to evaluate the clusters might include examining the characteristics of the observations within each cluster, comparing the clusters to external knowledge or expectations, or examining the stability of the clusters over time or across different samples.

# Cluster Analysis

Cluster analysis is a common technique in data mining and machine learning, and it can be used to discover hidden patterns and trends in the data, identify groups of similar observations, and make predictions about future behavior.

There are many different approaches to performing cluster analysis, including K-means clustering, hierarchical clustering, and density-based clustering algorithms, that we previously used. The choice of which algorithm to use will depend on the characteristics of the data and the specific goals of the analysis.

## Cluster Visualization using t-SNE

Cluster visualization is an important step in the cluster analysis process, as it allows us to understand and interpret the structure and patterns in the data. One common technique for visualizing clusters is t-SNE (t-distributed stochastic neighbor embedding), which is a non-linear dimensionality reduction method that can effectively visualize high-dimensional data in two or three dimensions.

Based on the results of using t-SNE to visualize 6 clusters in the data, we can see the relative positions and relationships between the different clusters. This can help us understand how the clusters differ from each other and identify any patterns or trends that may be present.

Overall, cluster visualization using t-SNE can be a valuable tool for understanding and interpreting the results of cluster analysis, and for identifying patterns and trends in the data.

# 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)