

Alex Anthony Panchot (M20190546)

Hugo Saisse Mentzingen da Silva (M20190215)

Rennan Valadares Ornelas Araújo (M20190146)

Final Project

**Data Mining**

# Table of Contents

[Table of Contents 2](#_Toc27369821)

[1. Introduction 3](#_Toc27369822)

[2. Preprocessing 4](#_Toc27369823)

[2.1. *NaNs* and outliers 4](#_Toc27369824)

[2.2. ‘Age’ column assessment 4](#_Toc27369825)

[2.3. Categorical features classification 7](#_Toc27369826)

[2.4. Numerical features regression 8](#_Toc27369827)

[3. Analysis 9](#_Toc27369828)

[3.1. Features correlation 9](#_Toc27369829)

[Quantile Matrix 10](#_Toc27369830)

[Boxplots 10](#_Toc27369831)

[Predictive Modeling 11](#_Toc27369832)

[Summary 12](#_Toc27369833)

# Introduction

This report addresses the final project in the Data Mining course of Master Degree Program in Data Science and Advanced Analytics of Nova IMS. The groups of students received a fictional insurance company database, containing personal customers’ data as well as insurance consumption information. The features contained in the dataset are described in the [Project Description](http://elearning.novaims.unl.pt/moodle/pluginfile.php?file=%2F53291%2Fmod_folder%2Fcontent%2F0%2FProject%20Description.pdf&forcedownload=1) document.

While the team impersonated a data mining consultancy, the final objective of this work was to “develop a Customer Segmentation in such a way that it will be possible for the Marketing Department to better understand all the different Customers’ Profiles”.

Therefore, we present in the following pages the process applied for this dataset treatment, the questions, findings, discussion, and recommendations that arose within it.

This report is accompanied by the file [Data\_Mining\_Project.ipynb](https://1drv.ms/u/s!AvRbdQNKauSThrNiHocr8C_nsKQyfg?e=fDvik3) which contains all the Python code used.

# Data Preparation

The dataset was initially converted into a Pandas DataFrame and the existing year-based columns (Brithday Year and First Policy´s Year) were rescaled to facilitate the interpretation, being substituted by Age and First Policy´s Age. For further analysis, the Premium: Sum feature was also added to the insurance\_df DataFrame, representing the sum of all premiums paid by the customer.

# *NaNs* and outliers

The initial assessment revealed 309 rows (3% of the total) with at least one blank cell (*NaN*) as well as the existence of some outliers or even noisy data (Figure 1). Instead of merely discarding the rows with NaNs, which could result in a significative loss of information, the team decided to apply techniques to estimate numerical and categorical features.

These techniques, regression and classification, respectively applied to numerical and categorical columns, first needed the removal of outliers, since it would provide better estimations.

**Outlier treatment:** since the dataset is not normally distributed, the team used the *interquartile range* (IQR) as the method to remove outliers. The usage of 1.5 IQR as the initial criteria for all columns revealed that some features had a large amount of observations to be considered as outliers (Household, Life and Work Compensations with more than 600 outliers each). In these especific cases, only the most extreme outliers were removed (above 3 IQR).

The criterion above resulted in at most 162 outliers for the same feature and 439 outliers in total (4.3% of the dataset) that were temporarily removed from insurance\_df. The resulting data is presented in Figure 2.

Rows with more than 3 NaNs were sough considering they could also undermine the estimators. However, there were no rows suiting this criterion.

# ‘Age’ column assessment

The visual analysis of the Age feature motivated a deeper assessment regarding its reliability, considering many customers have Age lower than First Policy´s Age. In fact, 1702 customers are in this situation and, additionally, 77.3% of customers below 19 years old already have children, which is quite uncommon.

For this reason, Age was not considered in subsequent analyses.

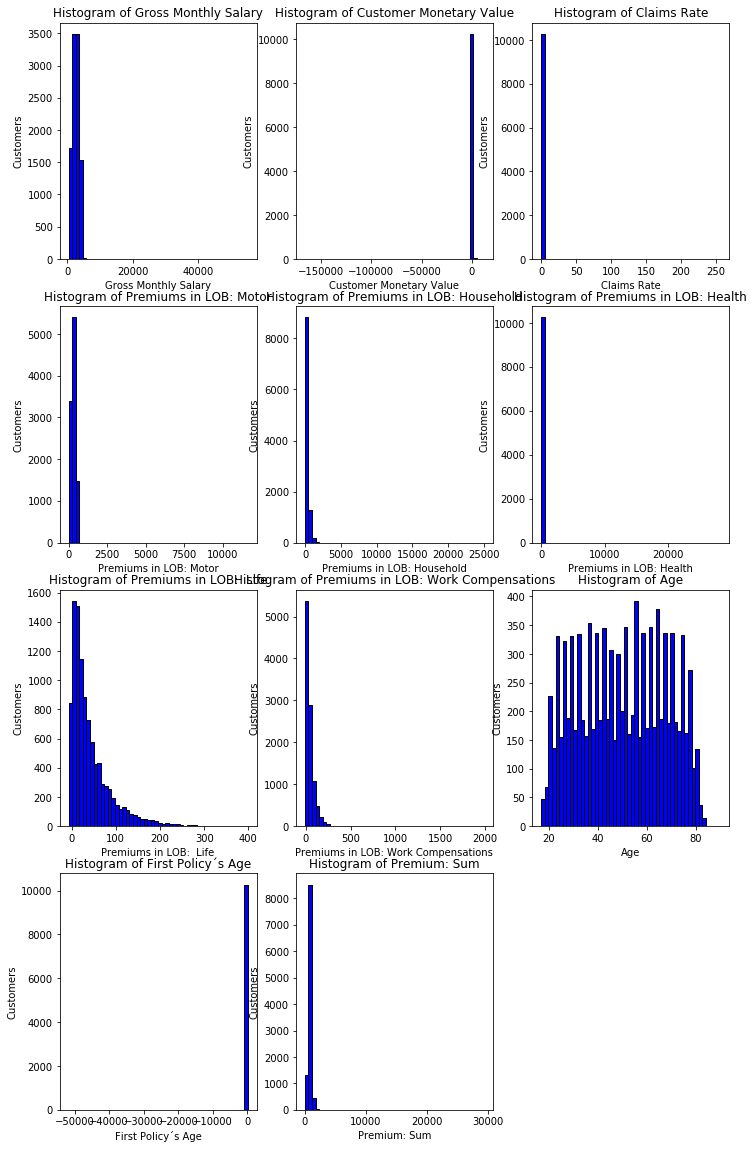


Figure 1 - Features histograms before removing outliers

A picture containing indoor

Description automatically generated

Figure 2 - Features histograms after removing outliers

# Categorical features classification

Initially, the Educational Degree needed to be encoded for numerical values with LabelEnconder from scikit learn. The numerical features were then rescaled to be within the [0,1] range using the MinMaxScaler. Later, the k-nearest neighbor algorithm KneighborsClassifier was applied and its accuracy for each feature was evaluated with a custom function which computed the confusion matrix for each trained model. The results are shown in Figure 3.

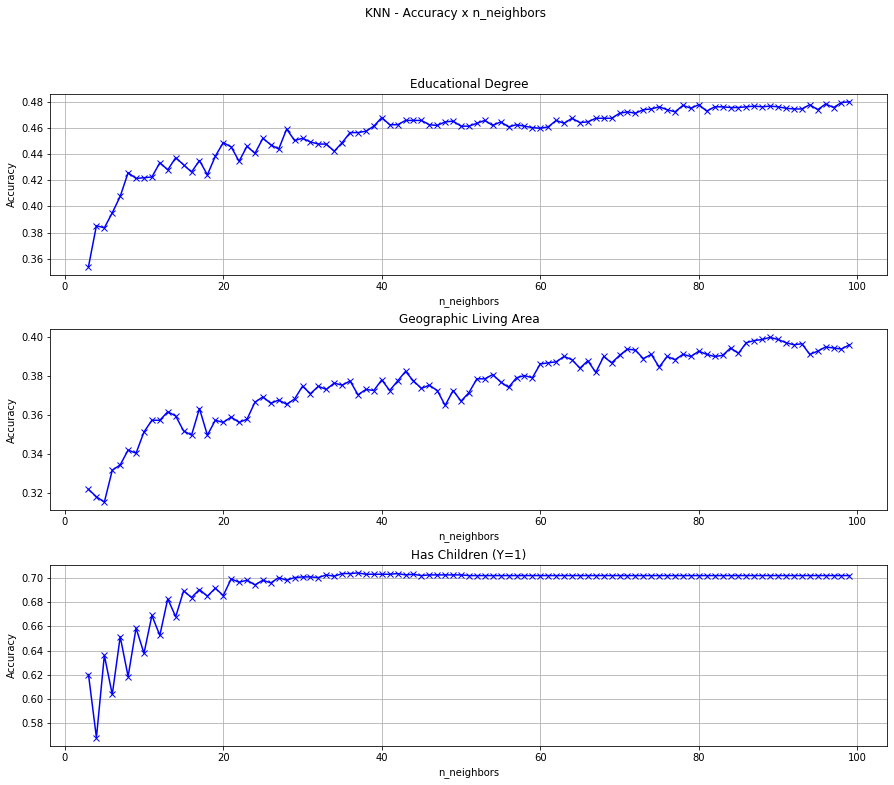


Figure 3 - Classifier accuracy for each categorical feature

Since we didn't achieve good estimations for the first two categorical features, the team decided to state the minimum accuracy at 2/3 and to drop the rows which contained NaNs on these two columns. Next, the optimal value for the number of neighbor samples (k) to Has Children (Y=1) was set to 21.

**k choice for the kNN algorithm:** the ‘rule of thumb’ is to choose k = sqrt(n)where n stands for the number of samples in the training dataset and k is the number of instances that we take into account to determine affinity with classes. Since we chose 80% of the data to be the training dataset, k would be around 89. The Figure 3 instead shows that there would be no improvement in selecting a number higher than 21 for Has Children (Y=1).

# Numerical features regression

(RENNAN)

# Data preprocessing

This chapter presents the techniques applied to analyze the data and the findings that emerged from them.

# Features correlation

Plotting the correlation matrix for all features (Figure 4) we can see that Age is highly correlated to Gross Monthly Salary. Since Age is also an untrustworthy feature (see section 2.2) and the amount of information would remain in the correlated feature, the team decided to drop the Age column for the analysis.

The Claims Rate also has a high inverse correlation with Customer Monetary Value (CMV), which makes sense since the Claims Rate is related to the profit that the company makes with each customer. So, the team decided to again reduce the input space and later clusterize the dataset only on the CMV perspective, since it carries more information than Claims Rate (includes customer retention and acquisition cost).

A screenshot of a cell phone

Description automatically generated

Figure 4 - Correlation matrix and heatmap for all features

What also calls our attention is that Premiums in LOB: Motor Insurance is considerably correlated (inversely) with all other insurance premium totals. Considering that signing a policy in any group could lead to increase proneness to contract other kinds of insurance (complementary products), this behavior is unexpected. Taking into account that the mean premium for Motor is substantially higher than for the other groups (Figure 5), it's contracting may be affecting the capacity or interest in paying for the other products. A possible insight would be to focus more on reverting this tendency.

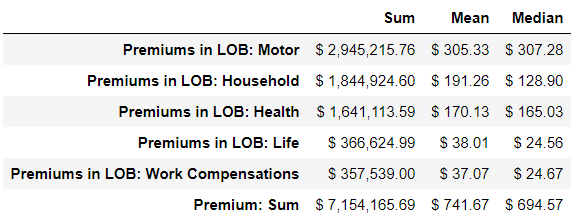


Figure - Premiums measures

Premiums in LOB: Household, Premiums in LOB: Life and Premiums in LOB: Work Compensations correlation show that these groups are decently working as complementary products, and its correlation could also be improved.

Finally, it’s possible to note a perfect correlation between the amount in Household insurance and the sum of the premiums, a fact that can also be commercially explored by the company.

# Analysis

# Quartiles (a priori grouping)

Separating the premiums data into quartiles confirms the inverse correlation between Motor insurance consumption and the sum of premiums (Figure 6). One good strategy would be to focus on inverting the correlation. Regarding the other insurance groups, specifically Health, Life and Work Compensations, they show a scattered but still correlated distribution over the quartiles. Possible approaches for these findings would be to take actions in order to bring customers closer to the main diagonal, i.e., increase the premiums total for customers that already have high premiums in any of these groups, and increase consumption on specific groups when the customer has already a high premiums total.

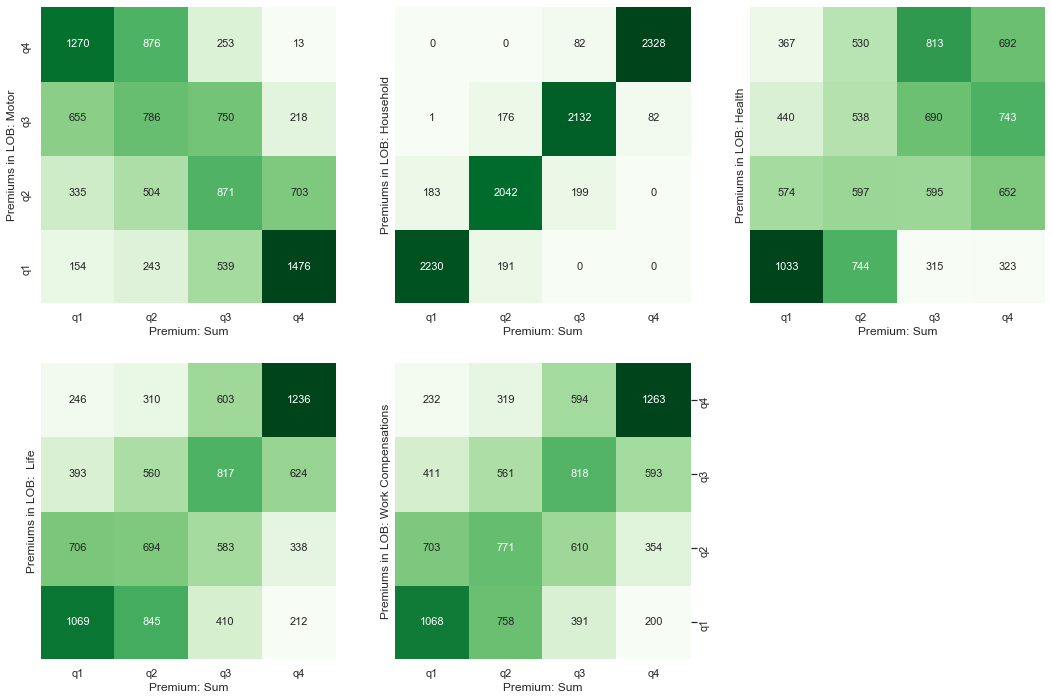


Figure - Quartiles plot for insurance groups vs premium sum

Another perspective that was also analyzed by the team was the premium amount for each group versus the gross monthly salary (Figure 7), which, in principle, could reveal the customer potential for contracting insurance. Ther quartiles show that this potential is not being totally tapped. Stands out, for instance, the customers' group in Q3 for Gross Monthly Salary and Q1 for Premiums in LOB: Health.



Figure - Quartiles plot for insurance groups vs gross monthly salary

# Boxplots

A further investigation continued with boxplots. The team was particularly interested in understanding the changes in the distribution of the insurance consumption features under the different customer categorical features.

Considering that people with children and higher education levels would be more prone to pay higher premiums in Health and Life, since the premium is proportional to the insured sum and the coverages, there is an unexpected behavior on the plots of Figure 8 and Figure 9. The same occurs with Work Compensations and Household, and the opposite is observed for Motor insurance. It seems that clients with high insurance buying potential are not being accessed for events that can cause major changes in their lives.

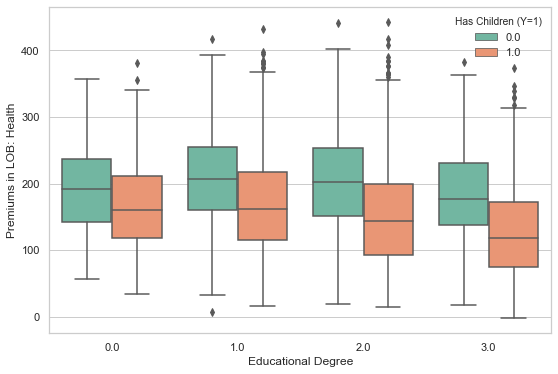


Figure 8 – Health premiums distribution on the educational degree and children in the family.

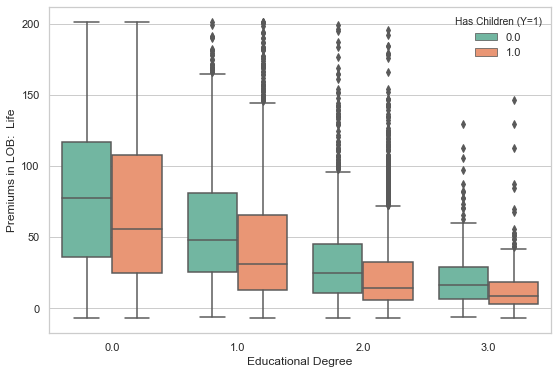


Figure 9 – Life premiums distribution on the educational degree and children in the family.

# Decision tree

Trying to predict the Premium: Sum or the Customer Monetary Value quartiles of a customer from it’s profile (Educational Degree, Geography, Children, Salary and First Policy´s Age), resulted in a maximum accuracy of only 0.357. On the other hand, the graph confirms the behavior observed so far and reveals that the geography is not a relevant variable for customer potential measurement (Figure 10).

A close up of a piece of paper

Description automatically generated

Figure - Decision tree for quartiles of Premium: Sum

# Clustering

(ALEX)

# Predictive Modeling

# Summary