**Introduction:**

In today's increasingly competitive and diverse market, it is no longer effective for companies to rely on generic business approach to attract a large customer base. Instead, businesses must adopt more targeted, personalized marketing strategies to effectively reach and engage their target audience. One effective way to do this is through customer segmentation, which involves dividing the customer base into smaller groups based on common characteristics. A team of analytics consultants was recently tasked with conducting a cluster analysis for an insurance company in order to identify and analyse different customer segments. The goal of this analysis was to help the company develop more targeted marketing strategies and better understand and serve the needs of its different customer groups. The use of cluster analysis and other data mining techniques was crucial in identifying patterns and relationships in customer data, and in helping the company create meaningful segments that could be used to inform its marketing and service efforts. Overall, the implementation of customer segmentation was a powerful tool in helping the insurance company attract and retain a larger, more loyal customer base.

In this project the goal is to segmentate the clients type with the data provided. *[Image 1]*

**1. Data Analysis:**

The data analysis begins by understanding the features and their characteristics. This includes examining the columns in the dataset and understanding the types of data they represent and the range of values they can take on summary statistics that are calculated for the data, including measures such as mean, median, and standard deviation. These statistics provide a general overview of the distribution and characteristics of the data, and can help identify patterns, trends, and potential outliers.

Finally, the data is then split into numerical and non-numerical (also known as categorical) features. This separation is useful for subsequent analysis, as different techniques and methods may be required for each type of feature.

**2. Data Understanding**

**2.1. Data Info & Summary Statistics** [**:**](https://docs.google.com/document/d/19qgsJypEo8z42z_0IVTeK9Bbl_B28_aY/edit#heading=h.1fob9te)

According to the summary statistics of the data set *[image 2]*, it has been observed that the features First Policy Year, Gross monthly salary (€), Customer Monetary Value, Premiums (€) in LOB: Motor and Premiums (€) in LOB: Health all have maximum values that may be considered unrealistic or unexpected based on the context of the data. This may indicate the presence of outliers, or extreme values that differ significantly from most of the data. Outliers can potentially impact the statistical analysis and modelling of the data, and it may be necessary to identify and handle them appropriately. Additionally, the features Birth Year and Customer Monetary Value both have minimum values that may be considered unrealistic or unexpected. This may also be indicative of the presence of outliers in the data. The data set shows that there are more customers with children than those without, as indicated by the mean value being higher than 0.5 for the feature indicating the presence of children. According to the summary statistics, premiums for policies in the Lines of Business (LOBs) Household and Life tend to be higher and lower, respectively, compared to the other policy types. Finally, most of the customers in the data set have a bachelor’s or master’s degree, as indicated by the feature Education Degree.

**2.2 Data Description:**

The insurance company's dataset contains 10,296 observations and 14 variables, the majority of which are floats. One variable is an object. The variables contain information about the customers, including their individual characteristics (such as birth year, education level, and monthly salary) and their relationship with the company (such as customer monetary value and the premiums they pay for various products). In this analysis, the features are the different variables or characteristics being studied in the dataset. It appears that all the features, except for Education Degree, are numerical, which means they are represented by numbers. Numerical features can be either continuous (e.g., age, income) or discrete (e.g., number of children). The analysis also found that there are 9 columns with missing values. This means that there are some observations (rows) in the dataset that have missing or incomplete information for those 9 features. This can potentially affect the accuracy and reliability of any analysis or modelling performed on the dataset, as missing values can introduce biases or distort the results. *[image 2]*

**3. Exploratory Data Analysis**

Exploratory data analysis (EDA) is a method of studying a dataset to understand its characteristics and identify patterns, trends, relationships within the data and feature importance. It involves using a combination of numerical and visualization techniques to summarize the data and provide insights into its characteristics. During EDA, it is important to identify any issues or problems with the dataset, such as variables with large skew or excess kurtosis, gaps in distributions, invalid or unexpected values in categorical variables, and strong correlations with the target variable. It is also important to check for redundancy, missing values, and outliers in the data. After identifying and addressing these issues through data quality verification and cleaning processes, the data can be explored further to gain insights and prepare it for modelling through pre-processing techniques.

**3.1. Distribution analysis:**

Distribution analysis is a process of examining how a variable or variables are distributed within a dataset *[Image 3]*. It involves looking at the patterns, trends, and relationships within the data and identifying any potential issues or problems. This can clarify the characteristics of the data and preparing it for further analysis or modelling. During distribution analysis, it was noted that features have outliers, which will be discussed furtherly in topic *3.2.1 Duplicated Data*, beside outliers’ features “MonthSal”, “PremMotor” and “PremHealth” had a normal distribution, and the rest don’t.

In this data analysis, the categorical and numerical features of the dataset were separated for further processing and analysis. Categorical features are those that represent non-numeric data and can take on a limited number of values and there’s only one categorical feature in the dataset which is: “EducDeg”.

Numerical features are those that represent numeric data and can take on a continuous range of values, such as “CustID”, FirstPolYear”, “BirthYear”, “MonthlSal”, “GeoLivArea”, “Children”, “CustMonVal”, “ClaimRate”, “PremMotor”, “PremHousehold”, “PremHealth”, “PremLife” and “PremWork”, also the new features created in further topic are all numerical.

**3.2 Qualitative Check:**

Qualitative check is a process of examining the quality and characteristics of categorical variables in a dataset *[image 4]*. During this process, it is important to identify any invalid or unexpected values within the categorical variables, as well as any variables with high cardinality (many unique values or categories) or imbalanced distributions (where one level represents a disproportionate number of observations). Identifying and addressing these issues can help to improve the analysis and interpretation of the data as explained in further 3*.6 topic "Checking Irregularities*.

There are issues with the quality of the data in a dataset, specifically with the Education Degree, Living Area, Gross Monthly Salary, Claims Rate, Premium columns, and Birth Year variables. The Living Area values should fall within the range of [1,4], but there is no additional information about these values. The Gross Monthly Salary, Claims Rate, and Premium columns all have extremely high values that are skewing the distributions of these variables in a positive direction. The Birth Year variable has an extreme low value that is skewing the distribution of this variable in a negative direction, was noted during the creation of features “Age” and “FirstPolicyAge” the existence of clients with less than 18-year-old, which is an irregularity for the business case, after analysing these features closely, it was noted that after exchanging values per client, the issue was solved.

**3.3. Duplicated data:**

In this analysis, it was found that there were three pairs of customers with different ID's but identical values *[Image 5]*. Specifically, customers 2076 and 8122 had identical values, customers 2100 and 8014 had identical values, and customers 3507 and 9554 had identical values. To address this issue, it was decided to drop the last duplicated instances of these customers. This means that the most recent version of each pair of duplicate customers (based on the ID) will be retained, while the older versions will be removed from the dataset. This can help ensure that the data is accurate and free of redundant or outdated information. In this case, the decision to drop the last duplicated instances assumed that the most recent version of each record is the most accurate and relevant.

**3.4. Missing Values:**

    It appears that the dataset being analysed has some missing values *[Image 6]*. The percentages of missing values in the data are relatively low, below 3% in all columns with missing values. Given the low percentage of missing values and the lack of clarity about the cause of the missing values, the missing values were dropped, this decision was also assumed because it was noted that all clients that don't have missing values in Premium columns have values in all features a due that was assumed that all clients must regard all Insurance Types. Also imputing the missing values, or estimating the missing values based on the available data, could introduce biases or assumptions about the data that may not be accurate.

**3.5 Feature creation:**

Features creation is the process of creating new variables in a dataset in order to extract additional insights or facilitate further analysis or modelling. Three specific variables are created: “Age”, “First Policy Age”, and “Prem Total” (Total Premiums (€)). The “Age” variable is created by subtracting the “Birth Year” variable from the analysis year (“2016”), which provides a more interpretable measure of a person's age than the raw “Birth Year” value. The “First Policy Age” variable is created in a similar way, by subtracting the “First Policy Year” variable from “Birth Year”. This may provide a more interpretable measure of how old a person was when taking insurance policies. The “Prem Total” was created by summing the values of the Premium columns for each customer. This may provide a more comprehensive measure of a person's insurance premiums across all their policies.

**3.6. Checking irregularities:**

This process involves identifying and correcting errors or inconsistencies in a dataset by swapping the values of certain variables and updating other variables based on these revised values.

This dataset appears to contain errors or inconsistencies in the values of the Birth Year and First Policy Year variables. Specifically, there are 1948 cases (19.51% of the total) where the Birth Year occurs after the First Policy Year, which is not possible.

The correction for errors or inconsistencies in the values of the Birth Year and First Policy Year variables is applied to the original dataset. This is done by selecting the rows with irregular values (i.e., those where the “First Policy Year” occurs after the “Birth Year”) and then exchanging the values of the “First Policy Year” and “Birth Year” for these rows. The “Age” and “First Policy Age” variables are then updated, respectively. These updates ensure that the “Age” and “First Policy Age” variables are accurate and consistent with the revised values of the “Birth Year” and “First Policy Year”. *[image 7]*

The Claims Rate is a metric that reflects the ratio of claims value to premiums for a customer over a given period, such as the last two years. It is useful to examine the data for customers who have a Claims Rate above 1 and a positive Customer Monetary Value, as these customers may be costing the insurance company more money than they are bringing in.

 It is important to consider both the Claims Rate and the Customer Monetary Value when evaluating the financial impact of a customer on an insurance company. While a Claims Rate below 1 suggests that the insurance company has received more in premiums than it has paid out in claims for a particular customer, this does not necessarily mean that the customer is valuable to the company. There are 497 customers in the dataset (4.98%) who have a Claims Rate below 1 but a negative Customer Monetary Value, indicating that these customers are still costing the company money despite having a favourable Claims Rate. In addition to analysing the Claims Rate and Customer Monetary Value, it is also important to consider the value of premiums paid by customers. It is possible for premiums to be negative due to reinsurance cancellations, reinsurer closures, and other events. There are 2601 customers in the dataset who have at least one negative premium, which should be considered when evaluating the financial impact of these customers on the insurance company. It is also worth noting that the fact that all customers in the dataset have all different policy types with the company, or have acquired the company's services at all, cannot be assumed based on the available data.

**3.7. Redundancy Study:**

Redundancy in data refers to the presence of strong correlations between variables, which can potentially impact the performance of a model. *[Image 8]* In such cases, it may be helpful to remove or combine redundant variables to improve model efficiency.

There are two types of correlations that can be used to assess redundancy in data: Pearson and Spearman correlations. Once our data do not follow a Normal Distribution was used the Spearman correlation due it only assumes that the data variables are monotonically related, meaning that the relationship between the variables falls into one of two categories: (1) as the value of one variable increases, the value of the other variable increases as well, or (2) as the value of one variable increases, the value of the other variable decreases.

In this case, features “Age” and “Birth Year” are perfectly collinear, which means that there is a strong linear relationship between these two variables, it's normal because one was created by the other. This can be problematic in data analysis because it can lead to issues with model stability and interpretability. Similarly, First Policy Age and First Policy Year are also perfectly collinear.

**3.8 Outliers study:**

Outliers are data points that differ significantly from the other observations in a dataset. They can have a significant impact on the results of an analysis, as they can skew the distribution of the data and affect the statistical measures calculated. Therefore, it is important to identify and handle outliers appropriately in order to accurately interpret the results of an analysis.

**3.8.1 Interquartile Range Method (IQR):**

One way to identify outliers in the data is by using the interquartile range (IQR) *[Image 9]*. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). A decision range is then defined based on the IQR, and any data point outside of this range is considered an outlier. Any data point that is lower than the lower bound (Q1 - criterion \* IQR) or higher than the upper bound (Q3 + criterion \* IQR) is considered an outlier. However, for sensitive data, a stricter criterion of 3 can be used to identify extreme outliers. This criterion states that any observation that falls outside the decision range is considered an extreme outlier. *[Image 10]*

After checking the IQR method for outlier removal using a criterion of 3, 95.823% of the data remained. *[Image 11]*

**3.8.2. Manual approach:**

A manual method was used to identify and remove outliers from the dataset by applying certain conditions based on the extreme values observed in the summary statistics and distribution plots. The resulting dataset after the removal of the identified outliers was then plotted to check the distribution of the numerical features. It was noted that the percentage of data remaining after the manual outlier removal was approximately 97.7%. The manual method resulted in the exclusion of a small percentage of the total data. *[Image 12]*

**3.8.3 Z-score:**

For features “PremHealth”, “PremMotor” and “PremHealth”, that present a Normal Distribution were checked for the existence of outliers by the Z-Score approach. Using this method, the following outlier values can be identified in the dataset: value from row 505 with a value of 440.86, value from row 1069 with a value of 432.97, and value from row 1935 with a value of 442.86. If these outliers are eliminated from the dataset using the Z-score method, it will leave 99.97% of the initial data.

**3.8.4 Outliers Conclusion:**

The Manual approach keeps 99.7% of the data after removals while the IQR approach retains 95.7%. It's clear that the Manual methodology preserves more data while still removing extreme values, so it'll be the method applied, after it was also checked if the detected outliers of Z-Score and was noted that z-score keep tracking some outliers, precisely 4 values, which also where removed. And due that Manual and z-score approaches were implemented.

**3.9 Categorical Features Encoding**

Only categorical feature was “EducDeg” and regards an order already labelled, so was used the same numeric number associated to it as the encoding result.

After this all data was converted to numerical data

**4. Data Normalization**

**4.1 Normalization Methods Analysis**

This part In the process was analysed the accuracy of data preservation from Normalization methods Min Max [0,1], Min Max[-1,1] and Robust Scaler, after normalize data by each one was measured the Mean Absolute Error (MAE) which is a measure the effectiveness of the normalization method, a lower MAE indicates that the normalization method is able to preserve the original structure of the data more accurately, resulting on Min Max [0,1] as the one more accurate. *[Image 13]*

**4.2 Min Max Scaler**

With min max scaler data set was rescaled in all features values to the range [0,1], this was done feature-wise in an independent way. [Image 14]

**5. Clustering**

Clustering is the task of partitioning a set of data points into groups, or clusters, based on their similarity. The goal of clustering is to identify patterns and structure in the data, and to group similar data points together. Clustering is an unsupervised learning technique, as it does not require labelled data or predefined categories.

**5.1 Entire Data Clustering**

Entire data was analysed over several methods to understand the existence of possible clusters.

First approach was using U-Matrix technique to help the interpretation of the distances between weights oof each neuron and its neighbours in the SOM grid, was notable that data have 3 or 4 types of distances [*Image 15],* then applied the Hit-Map technique to understand the units frequency where was also notable 3 types of frequency [*Image 16],* and also measured the correlation between features with component planes, where was possible to see hight correlation, positives and negatives over a few features (described in *image 17).*

* 1. **Clustering Strategy**

After data been analysed in detail was decided to have a strategy of data segmentation, dividing the entire data into two parts, Socialdemographic level regarding “Age”, “GeoLivArea” and “Children” and Customers Value level regarding “MonthSal”, “CustMonVal”, “PremTotal” and “ClaimsRate”.

For the first level (Socialdemographic) was applied a several techniques to study the best number of clusters, **U-Matrix** view *[Image 18]*, **Elbow Method** resulting in candidates 2, 4 or 5 for number of clusters *[Image 19],* by **Silhouette Score** the best number of clusters is 4 where average silhouette score is 0.577 *[Image 20 and 21],* and due the Silhouette Score suggested that the best value is for 4 was looked for the **R2 value** for **k=4** *[Image 22]***,** by **Hierarchical Clustering Dendrogram** *[Image 23]* the best cut point to 4 clusters by **K-means + Hierarchical Clustering** the best value for k is 2 *[Image 24].*

For the second level (Customers Value) was applied a several techniques to study the best number of clusters, **U-Matrix** view *[Image 25]*, **Elbow Method** resulting in candidates 2, 4 or 5 for number of clusters *[Image 26],* by **Silhouette Score** the best number of clusters is 4 where average silhouette score is 0.356 *[Image 27 and 28],* and due the Silhouette Score suggested that the best value is for 4 was looked for the **R2 value** for **k=4** *[Image 29]***,** by **Hierarchical Clustering Dendrogram** *[Image 30]* the best cut point to 4 clusters by **K-means + Hierarchical Clustering** the best value for k is 2 *[Image 31]*

Annexes

Graphical user interface, text

Description automatically generated*Image 1 – Metadata*

*Graphical user interface, application, table

Description automatically generatedImage 2 – Summary statistics of the data set*

Diagram, shape

Description automatically generated*Image 3 – Distribution analysis*

*Chart

Description automatically generatedImage 4 - Qualitative check*

*Image 5 – Three pairs of customers with different ID's but identical values*

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

A picture containing table

Description automatically generated*Image 6 - missing values.*

*A picture containing table

Description automatically generatedImage 7 – Exchange of feature “First Policy Year” with “Birth Year”*

*Image 8 – Spearman Correlation Heatmap*

*Chart, histogram, waterfall chart

Description automatically generated*

*Chart

Description automatically generatedImage 9 - Interquartile Range Method (IQR)*

*Diagram

Description automatically generated with medium confidenceImage 10 - Interquartile Range boxplots on data*

*Shape, polygon

Description automatically generatedImage 11 – Data distribution after IQR Outliers exclusion*

*Diagram, shape, polygon

Description automatically generatedImage 12 – Data distribution after Manual Outliers exclusion*

*Chart, bar chart

Description automatically generatedImage 13 – Mean absolute error per scaler method*

*Table

Description automatically generatedImage 14 – Data scaled by Min Max*

*A picture containing background pattern

Description automatically generatedImage 15 – U-Matrix of entire data*

*Diagram

Description automatically generatedImage 16 – Hit map from entire data*

*A picture containing application

Description automatically generatedImage 17 – Component planes from entire data*

Chart, line chart

Description automatically generated*Image 19 – Elbow Method for Socialdemographic Level*

Chart

Description automatically generated*Image 20 – Silhouette Score for Socialdemographic Level*

Chart, line chart

Description automatically generated*Image 21 – Silhouette Score for Socialdemographic Level*

*Image 22 – R2 plot for various hierarchical methods*

Chart, line chart

Description automatically generated

*Image 23 – Hierarchical Clustering Dendrogram*

Chart, box and whisker chart

Description automatically generated

*Image 24 – Hierarchical Clustering with Kmeans*Chart, box and whisker chart

Description automatically generated

Chart, line chart

Description automatically generated*Image 25 – Elbow Method for Customer Value Level*

*Image 26 – Silhouette Score for Customer Value Level*

Chart

Description automatically generated

*Image 27 – Silhouette Score for Customer Value Level*

Chart, line chart

Description automatically generated

*Image 28 – R2 plot for various hierarchical methods* Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated*Image 29 – Hierarchical Clustering Dendrogram*

*Image 30 – Hierarchical Clustering with Kmeans*

Chart

Description automatically generated