Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance

Customer Segmentation

Group AA

Andrei Macovei, number: 20221358

Tomás Domingos, number: 20221370

December, 2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

About A2Z Insurance 1
Understanding the Data 2
Data Incoherences 2
Feature Extraction 5
Data Normalization 5
Clustering 6
Multidimensional Visualization 8
Clustering by Product 8
Reintroducing Outliers 8
Marketing Proposal 9
References 9
Annexes 10

About A2Z Insurance

About A2Z InsuranceA2Z Insurance is an established insurance company in Portugal. A2Z offers its clients a wide range of products, including car, household, health, life, and work insurance. As a leader in the Portuguese insurance market, A2Z Insurance is committed to using data as a means of greater, more valuable, and more effective competitive differentiation.

The main objective of this report is to first examine the available data set, put it into a processed format, remove incoherences and missing values, and identify the key drivers of the business. Secondly, the key data is used to identify customer segments, grouping each segment into a differentiated cluster and describing the similarities between customers. The report then ends with our suggestions for targeted applications that the company can apply to each cluster to improve its current marketing activities. We'll use our data analysis techniques to determine which demographic group is most profitable for the company to target for future marketing campaigns and identify the key drivers/barriers for each of the clusters.

Understanding the Data

We started with a widely used methodology for data mining and analysis projects, CRISP-DM. It stands for Cross-Industry Standard Process for Data Mining. To start we loaded the dataset and its metadata into our python IDE, finding no available metadata. We checked the data types for each feature, and to have a first glance at the whole dataset, we looked at its descriptive statistics, finding some incoherences that we would later resolve.

After checking for missing values and replacing them with NaN, we realized most features contained empty cells, which would require some preprocessing work afterwards. We didn't find any duplicated rows.

Finally, we changed the EducDeg feature data, from text to numbers, to make all the future steps easier. The numbers won't represent any quantities. Each value will classify a different level of education.

Data Incoherences

After attempting to do some early feature extraction, to calculate the customer's age upon becoming A2Z's customer, we calculated a worrisome 19.39% of Birth Year data that was smaller than the

respective First Policy Year data. Meaning nearly 20% of the A2Z customers weren't alive according to the dataset available. We concluded that the First Policy year data was most likely imputed by the company, opposite to Birth Year which is most likely supplied on the customer's side, and if incoherent, should be dismissed over First Policy Year. With that being said, we removed Birth Year from any future inferences.

Visual Exploration

Before any outlier removal we decided to carry on with our data exploration, by generating the following visual graphics of the dataset:

- Numeric Variables' Histograms (Figure 1)
- Numeric Variables' Box Plots (Figure 2)
- Categorical Cardinality Variables' Absolute Frequencies (Figure 3)

We also tested possible outlier removals that the graphics suggested, counting how many observations would be removed, inferring that the dataset wouldn't need big removal's intrinsically.

Outlier Removal

Explorations conducted before made it known that most variables have outliers. Since algorithms are sensitive to outliers, various outlier detection methods were used to temporarily remove outliers from the dataset. After clustering, the outliers would be reintroduced to find out where the outliers are so that a reasonable and well-rounded marketing campaign can be conducted. To decide on which removal method we used, we decided to apply and infer different methods.

Manual Removal

After testing and counting possible observation removals, we decided to apply a slightly different manual approach (Figure 2) removing close to 0.5% of the available data. Given that the optimal removal percentage is, depending on the case, between 2-3%, we decided to use the algorithm and statistical methods for removal next.

Local Outlier Factor Filter

The Local Outlier Factor (LOF) filter is a machine learning algorithm used to identify anomalous or outlier data points in a dataset. It calculates the local density of data points in a dataset and compares them to the densities of their neighbours. Points that have a significantly lower density than their neighbours are considered outliers because they are not well represented in the local data. After applying the module, we successfully kept 99.08% of the original data.

IQR Removal

The IQR method is a statistical technique used to identify and exclude outliers in a data set. It stands for interquartile range and is based on the concept of a boxplot, which divides a data set into quartiles (four equal parts). The IQR is the difference between the upper and lower quartile of a data set and represents the middle 50% of the data.

To apply the IQR method, the data is first sorted and the median (mean) is calculated. The lower quartile is then defined as the value that represents 25% of the sorted data and the upper quartile is defined as the value that represents 75% of the sorted data. The IQR is then calculated as the difference between the upper and lower quartiles. Outliers in the data set are then identified as values that fall outside the range defined by the IQR. Specifically, values that are smaller than the lower quartile minus 1.5 times the IQR or larger than the upper quartile plus 1.5 times the IQR are considered outliers.

After applying the statistical method, we kept 83.30% of the original data, which is far beyond the 2-3% optimum, making it a non-considerable option.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised machine learning algorithm used to identify clusters of data points in a dataset. It groups points that are close together and have a high density of neighbours, and points that do not fit into these clusters are called "noise". As a con, DSCAN tends to struggle with high overall density datasets, struggling to identify clusters and maybe even outliers.

After applying the statistical method, we kept 96.64% of the original data, which is beyond the 2-3% optimum, but not in a high magnitude that could make it a non-considerable option.

Final Removal

We decided to stick with Local Outlier Factor as it was the method which was closer to the 2-3% removal optimum.

Replacing Missing Values

As noted before, we found observations with empty cells, we removed outliers before replacing the values to remove any possible biases in our removal methods. We also separated Enterprise generated data from Client generated. Enterprise-generated data is what we call MNAR, which stands for Missing Not at Random. This means that features relating to car, household, health, life and labour insurance premiums aren't missing at random and most likely signal the lack of consumption by the customer. Therefore we replaced the respective missing values with 0.

On the other hand, we classified Client generated data as MCAR/MAR (Missing Completely at Random/Missing at Random), meaning that the data is missing in a random and unbiased way. We replaced the values of the Geographic Living Area feature first with the mode value since there was only one missing value in the whole feature. For the Education Degree, First Policy Year and Monthly Salary features, we applied KNN Imputation. KNN (K-nearest neighbours) imputation is a method to fill in missing values in a data set. It involves identifying the nearest neighbours of a data point with a missing value and using the values of these neighbours to replace the missing value. Our function algorithm calculated the distance between the data point with the missing value and all other points in the data set, predicting accurately the value of the missing cells.

Finally, for the Children binary feature, containing values of 0 and 1, we decided to apply a Logistic Regression since the logistic regression model is a statistical method for predicting the probability of a binary outcome, such as the probability of an event occurring or a certain behaviour occurring. It is a widely used method in the field of data analysis and is particularly useful when the goal of the analysis is to identify the factors that influence a binary outcome, in this case, whether the customer has children or not.

Feature Extraction

At this stage, after removing outliers and missing values, we were ready to expand the dataset with new features derived from the original ones. In an attempt to quantify the customer's loyalty, we calculated the Customer Retention feature, using the year of data delivery (2016), and subtracting it from the First Policy Year feature. Total Premium spending was calculated by adding all the car, household, health, life and labour insurance premiums together. The yearly Salary was derived by multiplying the Monthly Salary feature by 12. The total Commitment Rate represents the percentage of the Yearly Salary spent on all insurance products. The same logic was applied to car, household, health, life and labour insurance premiums, generating Motor Commitment, Household Commitment, Health Commitment, Life Commitment, and Labour Insurance Commitment. Concluding the feature extraction process, we calculated an Insurance company's Yearly Cost metric using the Claims Rate feature and equation. Yearly Cost was calculated by multiplying the Claims Rate with the Total Premiums spent by the customer,

$$Claims\ Rate = \frac{Yearly\ Cost}{Total\ Premiums} \Leftrightarrow Yearly\ Cost = Total\ Premiums \times Claims\ Rate$$

Data Normalization

Scaling data can help to avoid distortions that may arise from the different units or scales of the variables. Scaling allows you to compare the variables on a common magnitude and improve the performance of algorithms. Many algorithms use some kind of distance measure to compare observations, and variables on different scales can distort this measure. Scaling variables can help to more accurately reflect the relationships between variables.

In this case, we applied the MinMax Scaler, a preprocessing technique that scales a feature's values between a given minimum and maximum value, typically between 0 and 1. It can be useful for transforming features with a skewed distribution or for features that have a large range of values, as it scales the values to a specific range, usually between 0 and 1. Giving the Dataset which includes a lot of long-tail distributions.

Redo Visual Exploration

Before any clustering was done we decided to visualize the data one last time before applying any clustering methods. We plotted the following visual graphics of the dataset:

- Numeric Variables' Histograms (Figure 4)
- Numeric Variables' Box Plots (Figure 5)

Correlation Matrix

A correlation matrix is a table that shows the correlation between several features. It can be used to summarise the relationship between all pairs of variables in a data set and to identify which variables are most strongly correlated with each other. The values in the matrix are usually Pearson correlation coefficients, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

After inferring on the Correlation Matrix, we realized that the First Policy Year and Customer Retention features weren't particularly correlated with any other features across the dataset. While all other features shared a high enough (!= 0) correlation. (Figure 6)

To understand better CHildren and it's relation with other features we applied a pair plot with Children as the Hue (Figure 7)

Clustering

Clustering by demographic perspective and product perspective can be important in several ways, Demographic clustering can help identify patterns and trends in consumer behavior based on characteristics such as location, education, and children. Clustering products can help to understand how different products relate to each other and which products are most popular with certain consumer groups. Clustering by demographic and product can provide a more complete picture of consumer behavior, allowing companies to better understand the interactions between different product categories and different populations.

K-Modes and Demographic Features

The features used in this Clustering method were Yearly Cost, the Children, and Education Degree features. Since the last two are categorical, we decided to apply the K-Modes Clustering method. When the data consists of categorical variables, standard clustering algorithms like KMeans may not be effective because they are designed to work with continuous variables. KModes is better suited for clustering categorical data because it uses a distance measure that is specifically designed for this type of data.

To understand the number of clusters we needed to find we plotted the Elbow Line, which indicated that the optimal number of clusters would be 4. To be safe we clustered for 4 and 5 clusters. After studying the centroids of each cluster we inferred that clustering with 5 potential clusters would produce a 5th cluster with low observations (<10) that we could dismiss, and focus on centroids that would give us a better picture of the 4 suggested clusters. We identified that people with no Children, living in region #4 incurred the company the least amount of Costs. And that people with Children tend to have two times higher Costs.(Tabel1)

K-Means and Product Features

For the Product Features which are metrics, we picked Yearly Salary, Customer Monetary Value, and Total Premium Spending. We picked the K-Means method after calculating all R Square Scores for K-Means, Complete, Average, Single, and Ward methods. Since the Elbow method suggested 2 clusters, we tested for 2, 3, and 4 clusters. Even though 2 and 3 Clusters offered good differentiation, 4 showed us the best customer segmentation (Tabel 2). We learned that customers with high salaries didn't correlate to

the highest Customer Value, and tended to spend less in Premiums, compared to the segment with the lowest salaries. We can infer that the lower the salary the more the customer spends, equating to much higher Commitents ratios compared to their Yearly Salaries (Insurance is seen as more essential a product).

Merging Product & Demographic Clusters

From Product and Demographic Cluster Exploration we selected, K=3 Clusters for Product and K=5 Clusters for Demographic, as highlighted before. To merge the clusters we used Hierarchical clustering, which is a method of clustering where a hierarchy of clusters is created. It is flexible, easy to visualize, and can be used to identify clusters at different levels of granularity. We obtained 6 clusters, where clusters (4,0) and (4,2) could be merged further due to their low frequencies (Tabel 3) After merging them we obtained the 4 merged clusters with better frequency distribution (Tabel 4).

Merged Clusters Analysis

After applying Cluster Profiling, we inferred from the Merged Clusters, that Cluster 1 stood out as having the lowest salaries, the highest Premium spending, Customer Value, and Yearly Cost. Cluster 2 stood out for the highest salaries, and highest spending on Health Insurance. Cluster 3 represented a segment with the highest spending on Car Insurance, and the lowest on House, Health, Life, and Work Insurance, producing the lowest Customer Value of them all.

Multidimensional Visualization

After deploying the T-SNE and UMAP we realized that the clusters have a good distinction from each other, at first we noticed there was a high degree of overlapping between Clusters 0 and 3 on the T-SNE, as a byproduct of the lack of Categorical features consideration. Cluster 3 is represented at one of the edges in the 3D UMAP visualization, therefore the overlapping we saw was a lack of dimensional representation.

Clustering by Product

By applying the same logic as we did for the Product and Demographic clustering we used the K-Modes clustering algorithm to analyze Work, Life, Health, Motor, and House. For Life Insurance, according to people's Yearly Salary, we inferred that they tend to spend the most, finding a cohort of 112 observations which spend the least, these customers had Children and an above-average salary (Tabel Life). For Work insurance, although the difference wasn't of a high degree, people with children tend to spend less, while childrenless highly paid people tend to spend more. (Tabel Work) .

When it comes to Health Insurance, people with higher salaries and no children tend to spend more compared to parents with an average salary. (Tabel Health).

For Car insurance, our clustering solution showed that parents with above-average salaries were the ones spending the most. (Tabel Car) . Lastly, House Insurance showed that high salaries and no children spend the most. (Tabel House) .

Reintroducing Outliers

We decided not to reintroduce clusters since we only removed <1% manually.

Marketing Proposal

From the clustering insights we got, the following we're the ones that we're most determining and should be taken into account for future marketing campaigns:

- The main focus should be on lower-income groups since they correlate to the highest customer value in the overall range of products. Therefore campaigns such as basket offer Health, Life, and Work to such customers could be beneficial to A2Z.
- Average salary-earning customers, to which the Household Committent Ratio is high, shouldn't be targeted to any insurance products except Car Insurance due to their low Customer Value and Total Spending.
- Highly paid customers should be targeted with Health Insurance products, with a possible Car Insurance bundle, since these are the highest spenders in Health, which also tend to have no Children.

References

Sklearn.AgglomerativeClustering.Retrieved January 06, 2023, from
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Sklearn.OneHotEncoder.Retrieved January 06, 2023, from
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

Sklearn.Ensemble.RandomForestClassifier. Scikit-Learn. Retrieved December 18, 2022, from
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
Sklearn.TSNE .Scikit-Learn.Retrieved January 06, 2023, from
https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Sklearn.StandardScaler.Scikit-Learn.Retrieved January 06, 2023, from
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Sklearn.Preprocessing.MinMaxScaler. Scikit-Learn. Retrieved November 21, 2022, from
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Sklearn.Tree.DecisionTreeClassifier. Scikit-Learn. Retrieved December 18, 2022, from
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn-tree-decisiontreeclassifier

Sklearn.DBSCAN.Scikit-Learn.Retrieved January 06, 2023, from
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

UMAP .UMAP .Retrieved January 06, 2023, from
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

Stobierski, T. (2021, January 19). Data Wrangling: What It is & Why It is Important. Harvard Business School Online. Retrieved November 19, 2022, from https://online.hbs.edu/blog/post/data-wrangling

Annexes

Tabel 1

demographic_clusters_4	Children	GeoLivArea	Yearly_Cost
1	0.790975	2.739711	0.259039
2	0.694303	2.739711	0.262830
3	0.000000	1.269896	0.274962
4	1.000000	1.000000	0.253543

Tabel 2

product_clusters_4	Year_Sal	CustMonVal	Prem_Total
1	0.524164	0.487999	0.404126
2	0.322422	0.295041	0.380149
3	0.176871	0.471375	0.575823
4	0.690495	0.308585	0.399230

Tabel 3

product_labels behavior_labels	0	1	2
0	1274	2284	2757
1	513	510	1165
3	118	662	111
4	95	642	107

Tabel 4

product_labels behavior_labels	0	1	2
1	1787	2794	3922
4	NaN	1304	NaN

Tabel Life

prem_life_clusters_4	PremLife	Children	Year_Sal
1	0.181034	0.018027	0.610395
2	0.147537	1.000000	0.399210
3	0.061552	1.000000	0.451230

Tabel Work

prem_work_clusters_	PremWork	Children	Year_Sal

4			
1	0.136773	0.996423	0.401133
2	0.160752	0.000000	0.613409

Tabel Health

prem_health_clusters_4	PremHealth	Year_Sal	Children
1	0.444798	0.608699	0.025057
2	0.353604	0.400102	1.000000

Tabel Motor

prem_motor_clusters_4	PremMotor	Year_Sal	Children
1	0.547790	0.401133	0.996423
2	0.418973	0.613409	0.000000

Table HouseHold

prem_house_clusters_4	PremHousehold	Children	Year_Sal
1	0.197791	0.025057	0.608699
2	0.167201	1.000000	0.400102

Figure 1

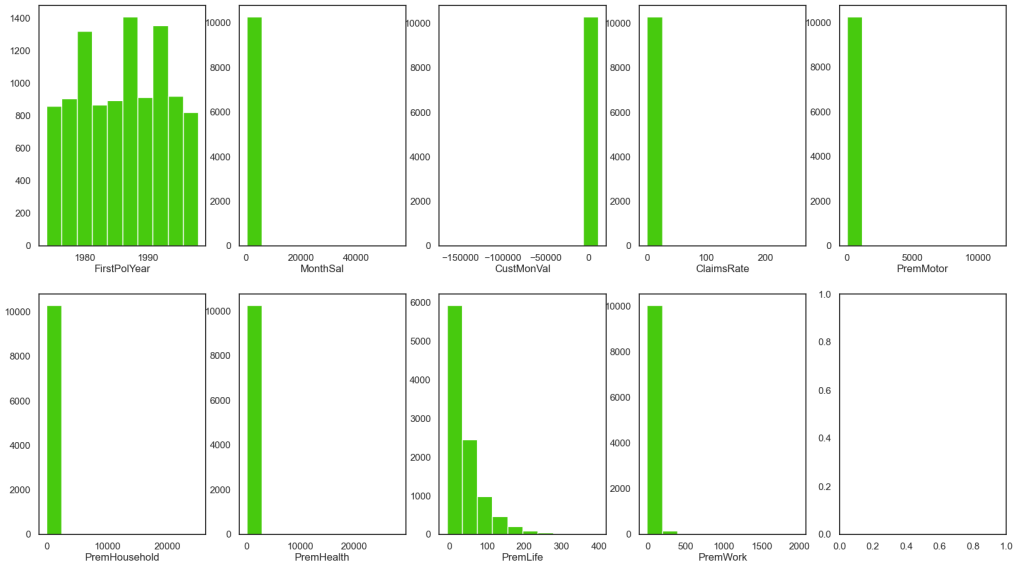

FirstPolYear
MonthSal
CustMonVal
ClaimsRate
PremMotor
PremHousehold
PremHealth
PremLife
PremWork

Figure 2

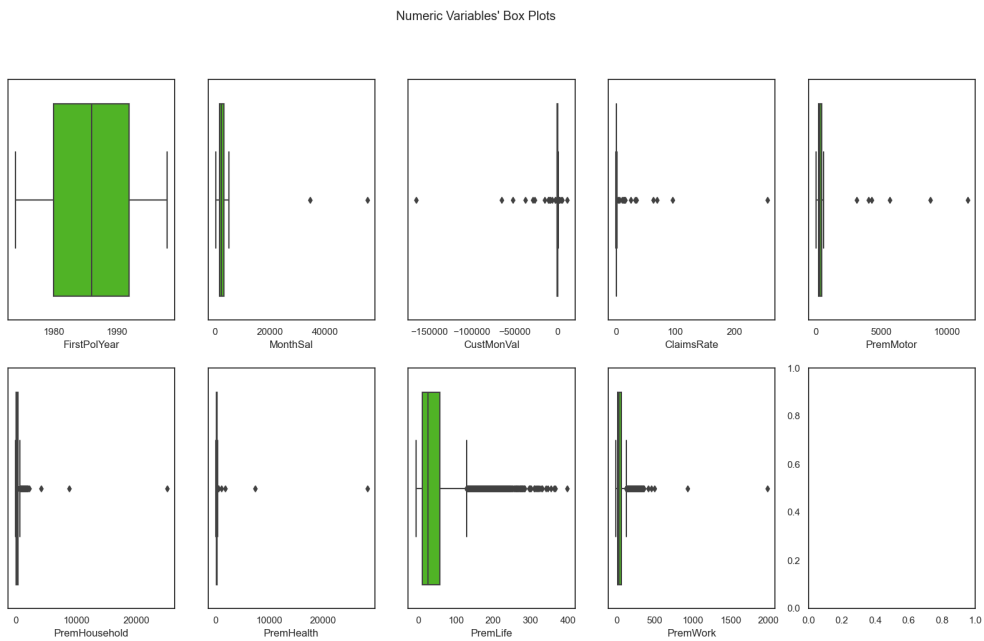

Numeric Variables' Box Plots
1980
1990
FirstPolYear
0
20000
40000
MonthSal
−150000
−100000
−50000
0
CustMonVal
0
100
200
ClaimsRate
0
5000
10000
PremMotor
0
10000
20000
PremHousehold
0
10000
20000
PremHealth
0
100
200
300
400
PremLife
0
500
1000
1500
2000
PremWork
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3

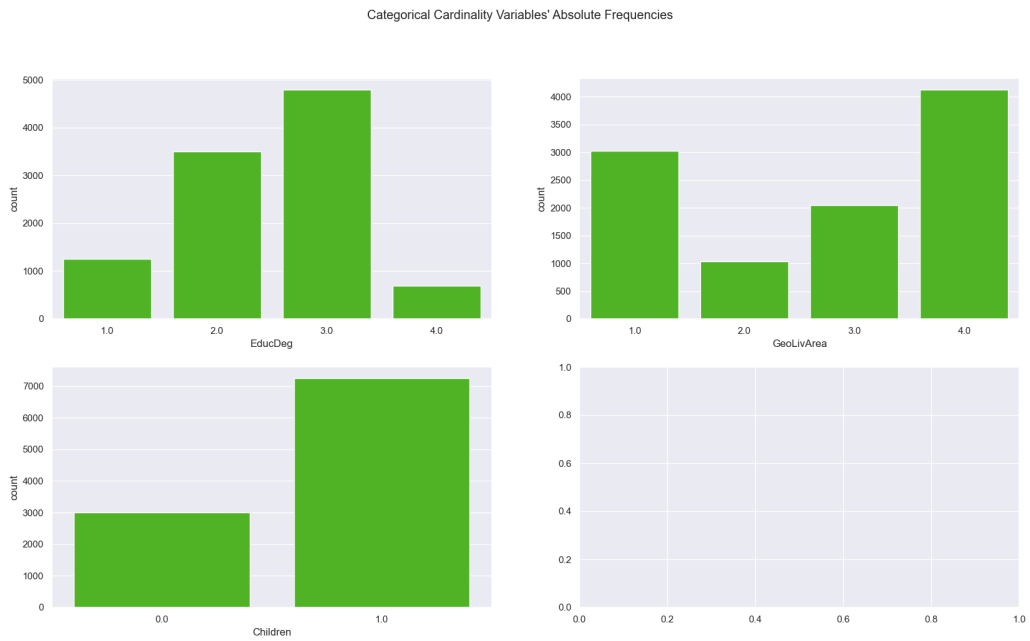

Categorical Cardinality Variables' Absolute Frequencies
count
EducDeg
GeoLivArea
Children

Figure 4

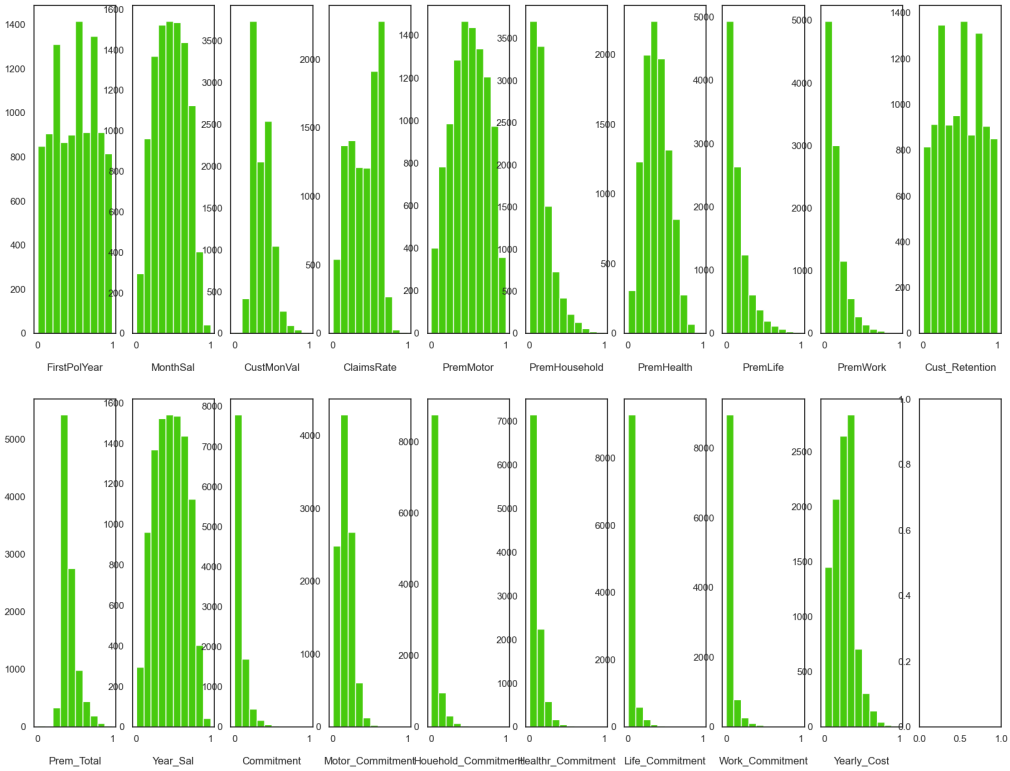

FirstPolYear
MonthSal
CustMonVal
ClaimsRate
PremMotor
PremHousehold
PremHealth
PremLife
PremWork
Cust_Retention
Prem_Total
Year_Sal
Commitment
Motor_Commitment
Household_Commitment
Healthr_Commitment
Life_Commitment
Work_Commitment
Yearly_Cost

Figure 5

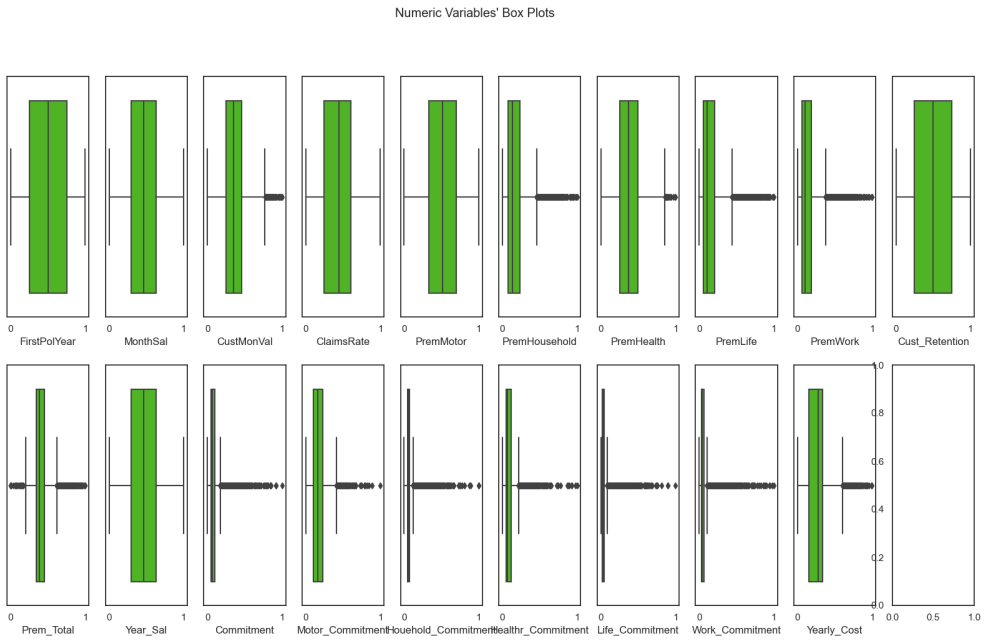

Numeric Variables' Box Plots
FirstPolYear
MonthSal
CustMonVal
ClaimsRate
PremMotor
PremHousehold
PremHealth
PremLife
PremWork
Cust_Retention
Prem_Total
Year_Sal
Commitment
Motor_Commitment
Household_Commitment
Healthr_Commitment
Life_Commitment
Work_Commitment
Yearly_Cost

Figure 6

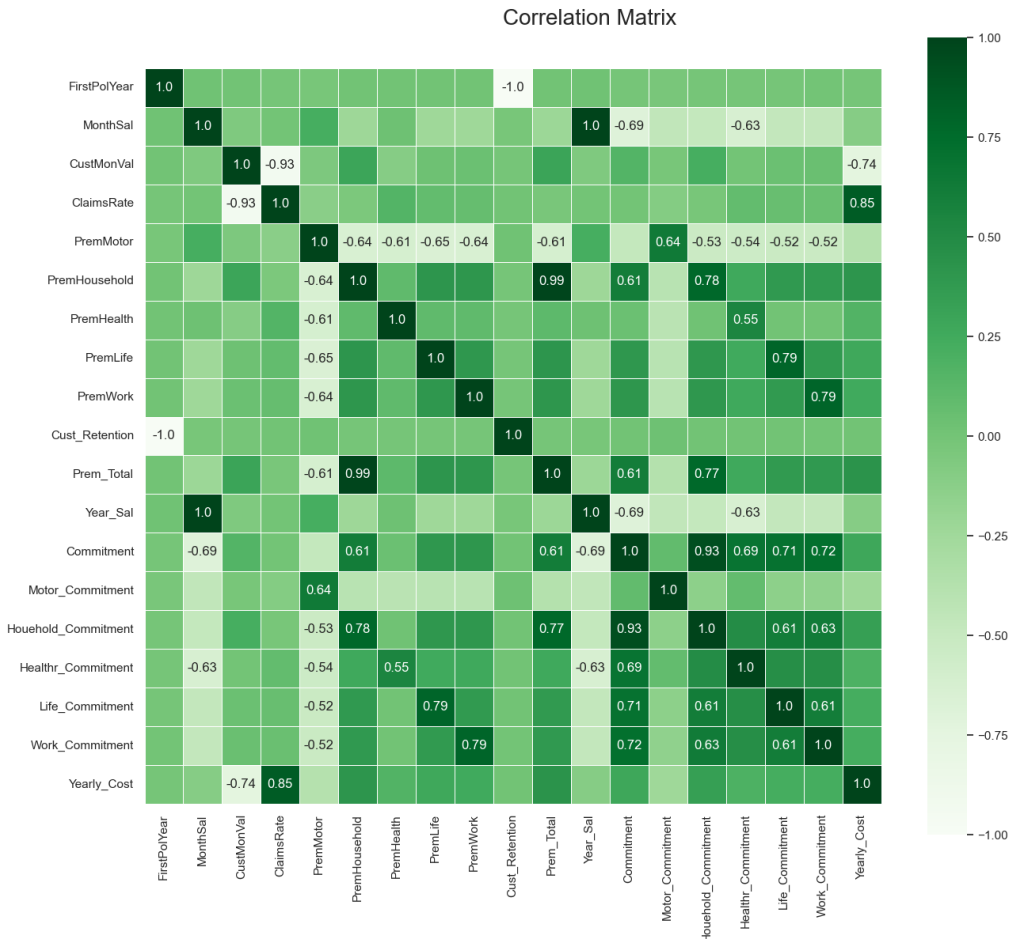

Correlation Matrix
FirstPolYear
MonthSal
CustMonVal
ClaimsRate
PremMotor
PremHousehold
PremHealth
PremLife
PremWork
Cust_Retention
Prem_Total
Year_Sal
Commitment
Motor_Commitment
Houehold_Commitment
Healthr_Commitment
Life_Commitment
Work_Commitment
Yearly_Cost
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 7

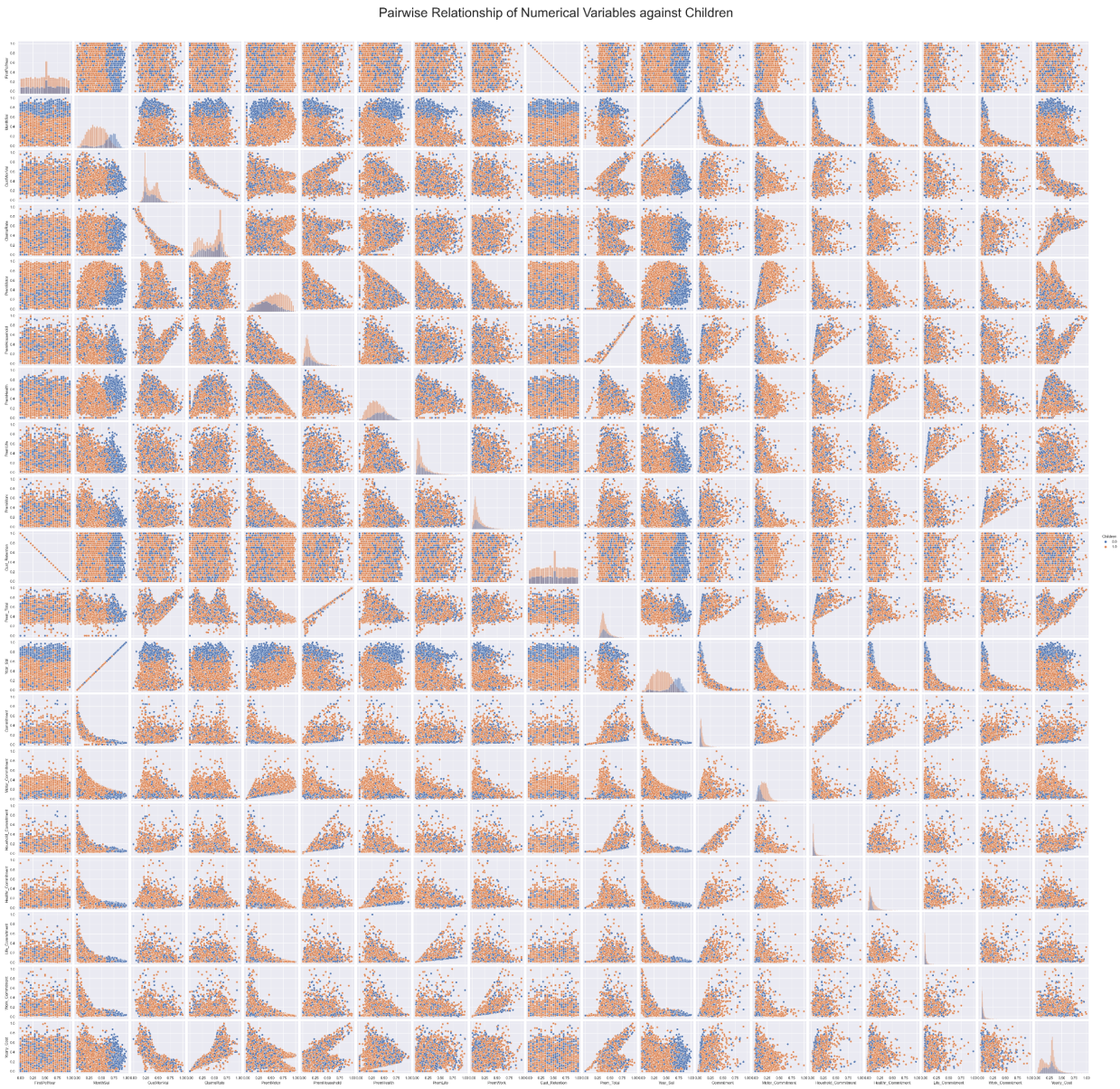
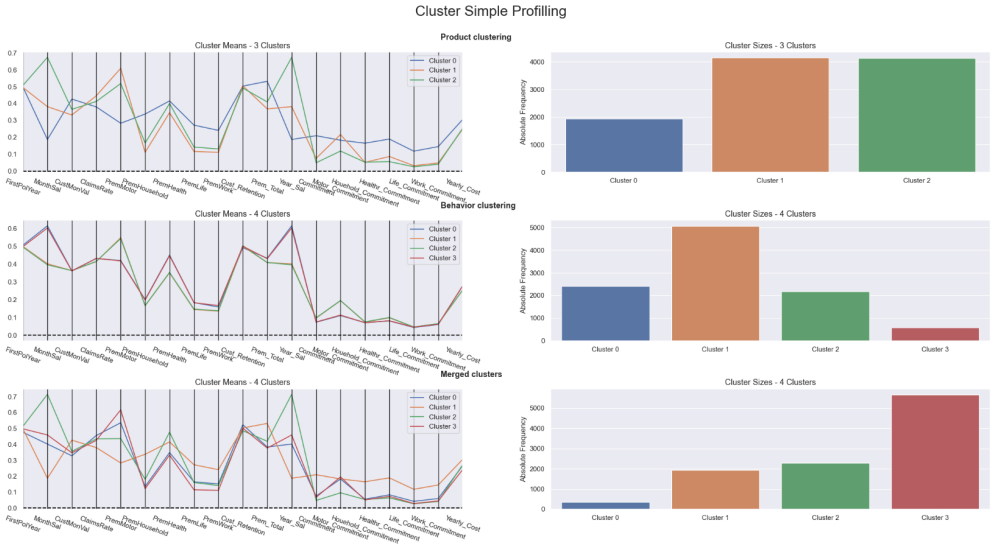

Pairwise Relationship of Numerical Variables against Children

Figure 8


Cluster Simple Profilling
Product clustering
Cluster Means - 3 Clusters
Cluster 0
Cluster 1
Cluster 2
Cluster Sizes - 3 Clusters
Absolute Frequency
Behavior clustering
Cluster Means - 4 Clusters
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster Sizes - 4 Clusters
Absolute Frequency
Merged clusters
Cluster Means - 4 Clusters
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster Sizes - 4 Clusters
Absolute Frequency