Data Mining Project

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance Customer Clustering

Group S

Rafael Dinis, number: 20221643

Luca Loureiro, number: 20221750

Maria Arbelaez, number: 20221381

January, 2023

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

INDEX

Contents

1. Introduction

The following project consists of a customer segmentation analysis for the company A2Z Insurance. A2Z became one of the largest insurers in Portugal in 2016, serving a wide range of insurance services: Motor, Household, Health, Life and Work Compensation. So far, the company has been using a mass marketing approach, but now they are interested in differentiating their customers and developing more focused programs. To achieve this, a cluster analysis based on a sample database of 10.296 active customers provided by the company, was performed. This database contained thirteen columns with features describing different perspectives of the customers.

2. Data Exploration

To start the project, an exploration of the database was performed. The goal was to get a better understanding of it and identifying any potential issues that needed to be addressed during the subsequent preprocessing phase.

The first step was to examine the head rows of the data frame to get a sense of the features and their values. It was found that the *CustID* column could be converted into an integer and set as the index of the data frame and that *FirstPolYear* and *BirthYear* could be converted into more interpretable features.

The data types of the columns were checked, and it was noted that *Children* and *GeoLivArea* had the wrong data type, as they were typed as numerical instead of categorical. The number of missing values per feature was also checked and several columns with null values where found. Finally, three duplicated rows where found.

Checking the central tendency metrics of the data it was also possible to identify some strange values, for example in the Min and Max of *FristPolYear* and *BirthYear,* and several features that seemed to have outliers given the difference between the mean and the median and the distribution of the quartiles.

2.1. Visual Exploration

To further understand the data, visualizations were created for both the numeric and categorical variables. The numeric variables were visualized through histograms and box plots, which showed the presence of outliers in most of the features. Pairwise scatter plots were also created to visualize the relationships between the numeric variables but since most features also had multivariate outliers, it was not possible to identify patterns at this point. However, through a correlation matrix it was possible to identify a high correlation between two pairs of features: *BirthYear* with *MonthSal* and *CustMonVal* with *ClaimsRate.*

The categorical variables were visualized through bar plots, showing the count of each category in the variable. There were not low frequency values found in any of the features and the cardinality was considered low, with maximum four possible values for each attribute.

3. Data Pre – Processing

The Data preprocessing was performed to clean, transform, and prepare the data for the analysis, as raw data is often incomplete, inconsistent, and dirty, which can affect the accuracy and reliability of the results. By preprocessing the data, its quality was improved, and it became more suitable for mining.

To start the Data Preparation Treatment, the duplicate rows were removed, as they were only 3, the total amount of rows became 10293.

Feature Engineering was the next step, where the date features *Birthyear* and *FirstPolYear* were transformed into *Age* and *CustDur* (customer duration) respectively, for better interpretability. Both were done by subtracting the original features to the year of the dataset (2016).

The coherence check was the next essential step in the Data Preprocessing, consisting of a group of rules that allowed only realistic data. In the scope of the project three rules were created. The first and second rules were to define *CustDur* and *Age* bigger than zero, as they can't possibly have negative values. Followed by making sure, through the third and last rule, that the *Age* was bigger than the *CustDur*, as a person cannot become a customer before being born.

3.1. Missing Values Treatment

Different methods were applied in accordance with the type of feature, to handle the missing values found in the exploration phase. For the categorical or non-metric features, the statistic of the mode was employed to fill the missing values. The metric features had their missing values filled by the KNN imputer. This is an algorithm that works by completing the missing values using k-Nearest neighbors, in this case, the k parameter was set to 5.

3.2. Outlier Removal

For the outlier removal phase, different methods were explored and checked in terms of the % of data removed. Methods removing more than 3% were deemed unusable.

3.2.1. Standard Deviation Method

This method involved calculating the mean and standard deviation of each feature, and then removing any data points that fell more than three standard deviations away from the mean. This method can usually be sensitive to the presence of a few large outliers, as it will increase the standard deviation and potentially result in the removal of more data points than desired. This method kept only 95.75% of the data, hence it was not used.

3.2.2. Inter-Quartile Range (IQR) Method

This method involved calculating the first (Q1) and third (Q3) quartiles of each feature, and then defining the inter-quartile range (IQR) as the difference between Q3 and Q1. Data points that fell more than 1.5 times the IQR below Q1 or above Q3 were considered outliers. This method kept only 82.14% of the data, hence it was not used.

3.2.3. Manual Outlier Filtering Method

This method involved manually setting upper and lower bounds for each feature based on domain knowledge and analysis of the data. This method allowed for more control over which data points were considered outliers and kept 99.71% of the data.

3.2.4. Isolation Forest Method

This is a multivariate method that involved training a model to identify data points that are anomalous or different from the majority of the data. The method works by randomly selecting a feature and a split value, and then repeatedly partitioning the data based on the selected feature and value. Anomalous data points are expected to be shorter paths in the partitioning process and are therefore identified as outliers. This method kept 86.8% of data, hence it was not used.

3.2.5. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Method

This multivariate method involved grouping data points into clusters based on the density of the points. A data point is considered a core point if it has more than a specified number of points (*minPts*) within a given distance (eps). Data points that are not core points and do not have any core points within a given distance are considered outliers. This method kept 99.54 % of data.

3.2.6. Local Outlier Factor (LOF) Method

This multivariate method involved calculating the local density of each data point and identifying points that had a significantly lower density compared to their neighbors. These points were considered outliers. The local density is calculated using the distance to the k-nearest neighbor, where k is a user-specified parameter. This method kept 98.75 % of data.

3.2.7. Combined Outlier Removal Methods

The three methods yielding an acceptable kept data percentage, the Manual filtering, the LOF and the DBSCAN, were applied in ensemble to achieve a more robust outlier removal treatment. The final percentage of data kept was then 98.84%. A visual exploration was performed after outlier removal to confirm that the distributions of the different variables where now making sense.

3.3. Feature Selection

For the feature selection phase, relevancy and redundancy were checked through a correlation matrix. There were not irrelevant variables found as they were all correlated in some level with other variables. Regarding redundancy, it was found that the following variables were highly correlated and therefore could be redundant for the analysis:

- ***CustMonVal* with *ClaimsRate* (negative correlation):** This seemed logic as the more claims the client makes, the less profitable it is for the company and its *CustMonVal* deteriorates. *CustMonVal* was kept for the analysis as it was a more comprehensive feature.
- ***Age* with *MonthSal* (positive correlation):** This also seemed logic since usually the older a person is, the more experience they have and the higher their salary can be. *MonthSal* was chosen for the analysis as purchasing power of the customer is normally a very relevant attribute for the insurance business.

To confirm that the correct feature selection was made, different clustering exercises with different combinations of feature selection and number of clusters where ran, and their visualizations and basic profiling where analyzed. This will be deepened in section 4.2.

4. Clustering

4.1. Clustering Perspectives

For the clustering, two perspectives were defined. The first one is the **Value Perspective** which describes how valuable a customer is for the company based on their purchase power, years as a client and how profitable they are. The second perspective is the **Needs Perspective** which reveals what are the insurance needs of the customer, based on the range of insurance services offered by A2Z: Motor, Household, Health, Life and Work Compensation.

4.2. Clustering Algorithm

To find the most optimal clustering algorithm for the dataset, the R^2 scores for different clustering solutions (number of clusters and algorithms) were revised. Concluding that the best model would use the K-Means algorithm with three clusters per perspective.

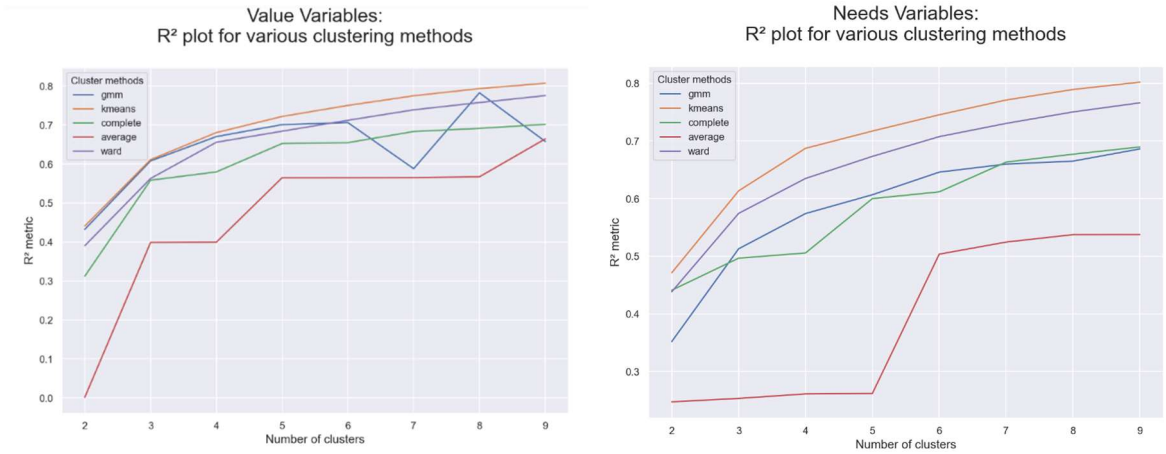


Figure 1 – Selecting the optimal clustering algorithm and number of clusters per perspective

4.3. Number of Clusters

After merging the resulting clustering models of the two perspectives, nine clusters were obtained. To reduce the number of clusters, a Hierarchical approach was performed and by the analysis of the resulting dendrogram, it was defined visually and not through finding the longest line not crossed by horizontals that the appropriate number of clusters would be between four and five.

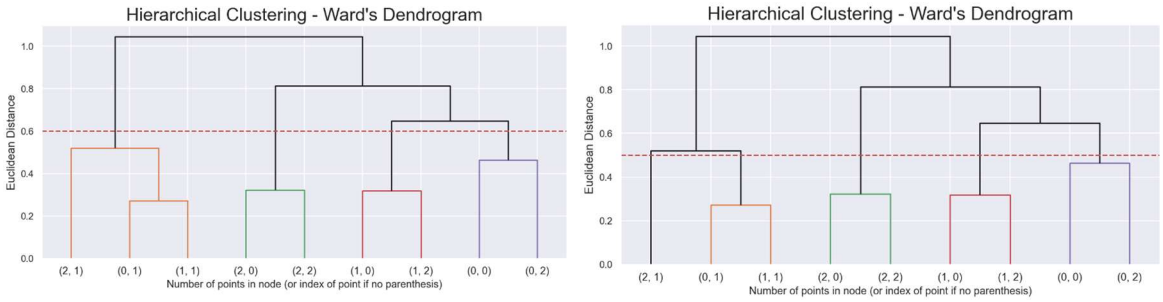


Figure 2 – Dendrograms for 4 and 5 clusters respectively

At this point, eight models were run and visualized with t-SNE to define the best number of clusters and to confirm the best feature selection. Four out of the eight models did not make sense visually. For the remaining four models, a basic profiling was performed with the objective of deciding which clustering solution offered the best interpretability.

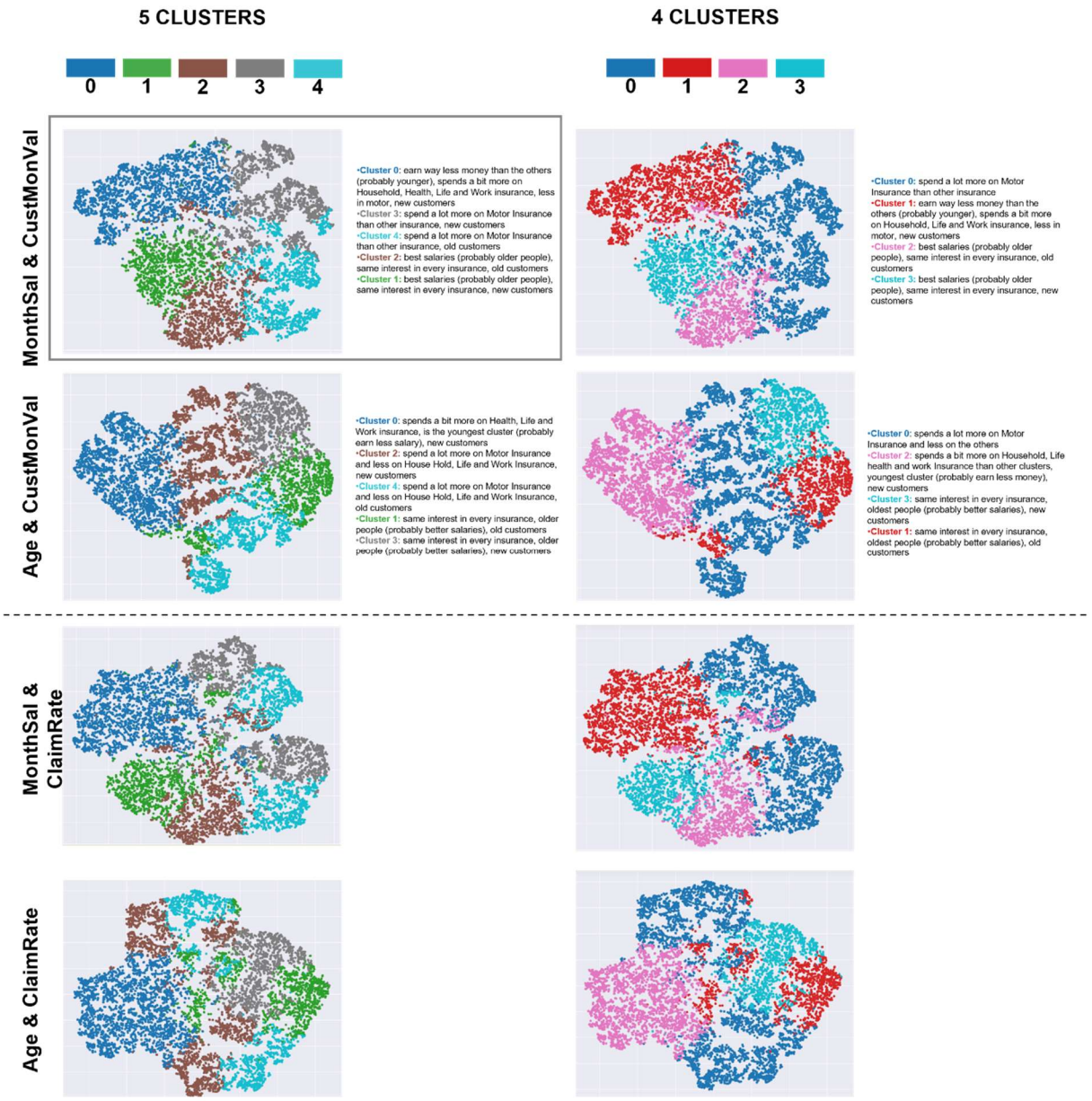


Figure 3 – Feature and number of clusters selection

Figure 3 show the eight t-SNE visualizations of the explored models. The selected model had five clusters instead of four as this allowed to segment customers for whom their principal need is Motor Insurance into recent and old customers. It also kept the feature *MonthSal* and dropped *Age*, as the visualization made more sense since the three clusters whit recent customers where closer together in the top side of the graphic and the two clusters with older customers where closer together in the bottom part of the graphic.

5. Customer Segmentation

After the cluster analysis, the final customer segmentation for A2Z Insurance can be defined as follows:

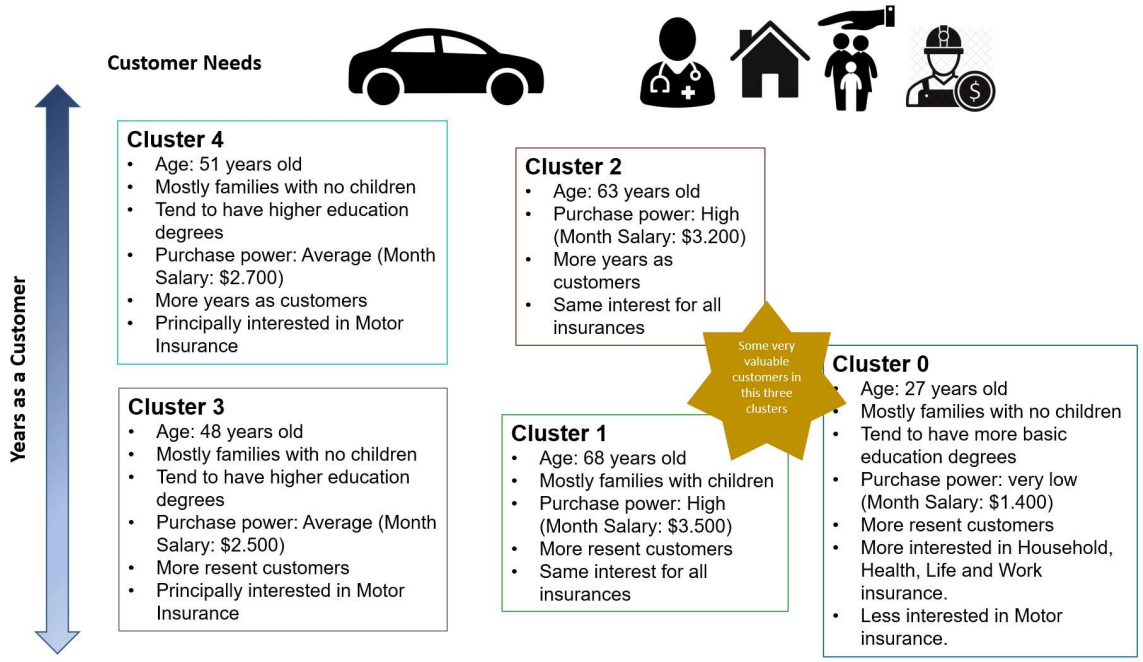


Figure 3 – Customer Segmentation

The clusters size can be appreciated in the following pie charts:

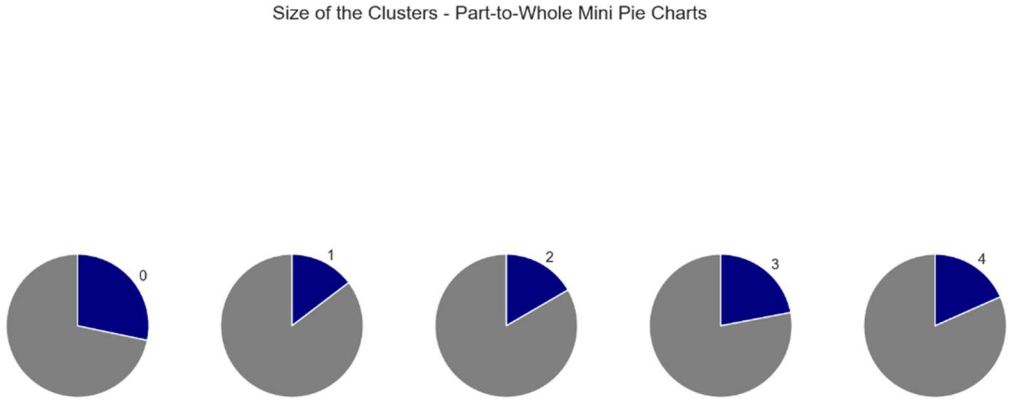


Figure 4 – Size of the clusters

This customer segmentation is mostly defined by the years the customer has been with the company, their insurance needs (specially if they are interested in Motor Insurance or not) and their purchase power defined by their monthly salary.

The most important cluster is Cluster 0, and it is made of young people who are probably starting their working life, have low salaries and most likely still don't own cars because they are interested in all the other insurances except for motor.

The next most important clusters are 3 and 4, this clusters are characterized by the fact that its members are principally interested in Motor Insurance. They are mostly middle-aged people who do not have children and earn average salaries. The difference between these two clusters is that customers in cluster 3 had been less years with the company.

Finally, clusters 1 and 2 are the least important ones. They are composed of the oldest customers in age and hence, the ones with the highest purchasing power. These customers do not have a special interest in a specific type of insurance but more on a comprehensive insurance package that can cover everything. The difference between these two clusters is that customers in cluster 1 had been less years with the company.

The difference between clusters can be better appreciated in the following graphic:

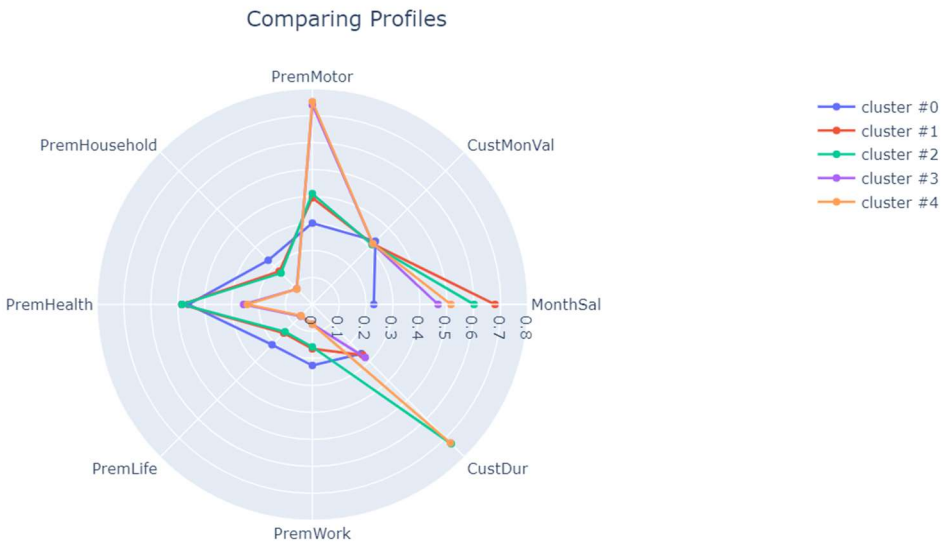


Figure 5 – Comparing cluster profiles

As for the customer monetary value, it is generally the same for every cluster:

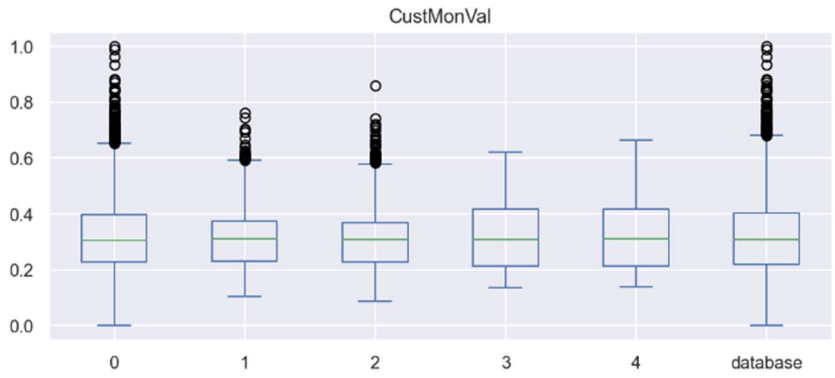


Figure 5 – Comparing Customer Monetary Value between clusters

This means that the focus for the company should be the biggest clusters (0, 3 and 4). Nonetheless, there are a few clients in clusters 0, 1 and 2 that represent a special high monetary value for the company and could be targeted with some niche strategies.

6. Marketing Actions

The proposed marketing actions based on the customer segmentation are the following:

- For the most important cluster (Cluster 0), since it is characterized by their low purchase power, the company can offer credit options with low interest rates for the customers to acquire insurances with better coverage and benefits that tend to be more expensive.
- They could also offer special discounts for young customers that are the children of older customers, this way the incentivize both the older and younger consumers.
- This cluster buys less Motor Insurance probably because they still do not own cars. The company could make an alliance with a car dealership where they give special discounts and credit lines to customers in cluster 0 so they can buy a car and acquire the insurance with A2Z.
- For the clusters that have better purchasing power (clusters 1 and 2) and that are interested in a comprehensive insurance package, the company can offer packages that can be more economic than buying the individual insurance. This way they guarantee that the customer has every insurance with A2Z.
- For clusters 3 and 4, the company should focus on having the best Motor Insurance service like having the quickest response, alliances with tow companies, angel driver service for when people want to drink alcohol, among others. This way they can guarantee the loyalty of these customers.
- But they should also focus on getting these customers to acquire the other insurances with A2Z. A way to do this is guarantying better coverages and prices than the competitors, but they could also focus on making special alliances that can attract the clients. For example, making an alliance with the best hospitals and doctors in town so people who acquire their Health Insurance are guaranteed to have the best health service.
- Regarding newest and oldest clients, the company could have an incentive to keep the customers more years with them. For example, if a customer turns five years with the company, then the insurance renovation is free; if they turn ten year with the company they participate in the raffle of new cars and travel packages, and so on.
- Regarding the few customers that have a high value for the company, A2Z could make an elite membership that gives this customers access to special alliances with luxury service such as VIP lounges in airports, Michelin Star restaurants, luxury hotels, designer brands, and so on, to reward their loyalty and value.

7. References

Scikit-learn developers. (n.d.). Agglomerative clustering. Retrieved from https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

Ismiguzel, I. (2022). Outlier detection with simple and advanced techniques. Retrieved from https://towardsdatascience.com/detecting-outliers-with-simple-and-advanced-techniques-cb3b2db60d03

Pandey, P. (2020). Part3: Visualising Kannada MNIST with UMAP. Retrieved from https://www.kaggle.com/code/parulpandey/part3-visualising-kannada-mnist-with-umap/notebook

UMAP developers. (n.d.). Basic UMAP parameters. Retrieved from https://umap-learn.readthedocs.io/en/latest/parameters.html

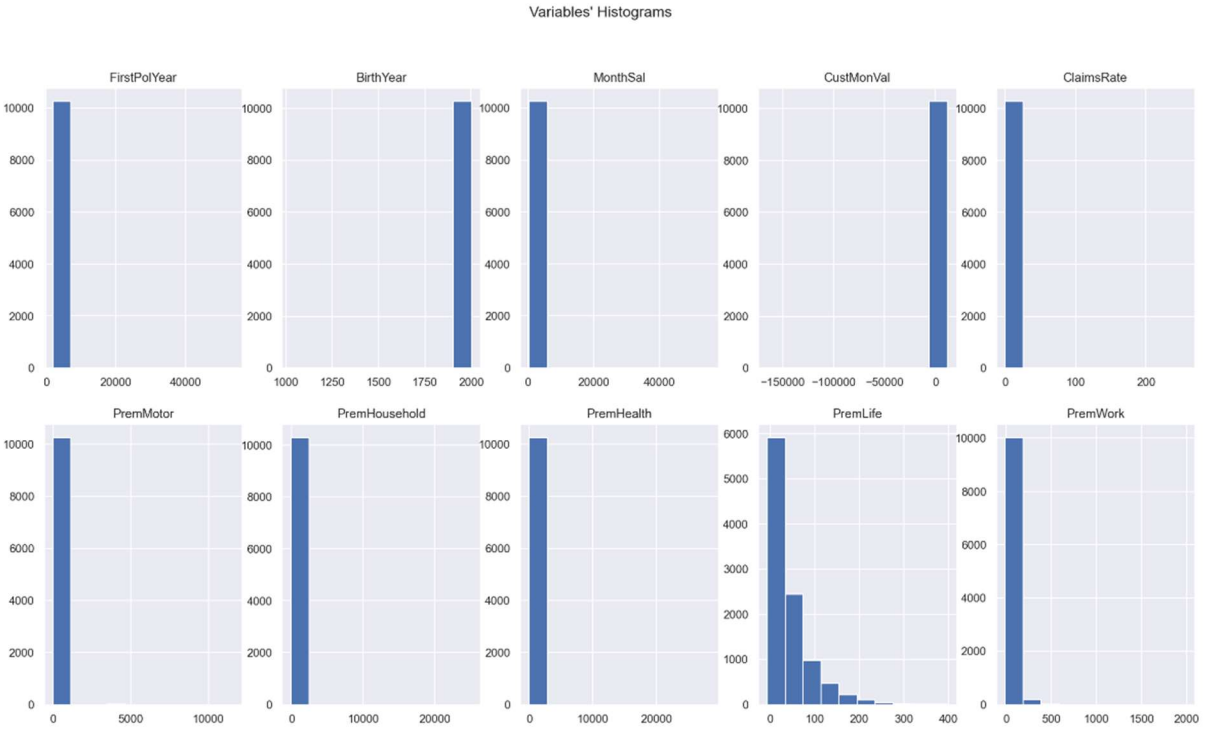
Rink, K. (2022). How to create and visualize complex radar charts. Retrieved from https://towardsdatascience.com/how-to-create-and-visualize-complex-radar-charts-f7764d0f3652

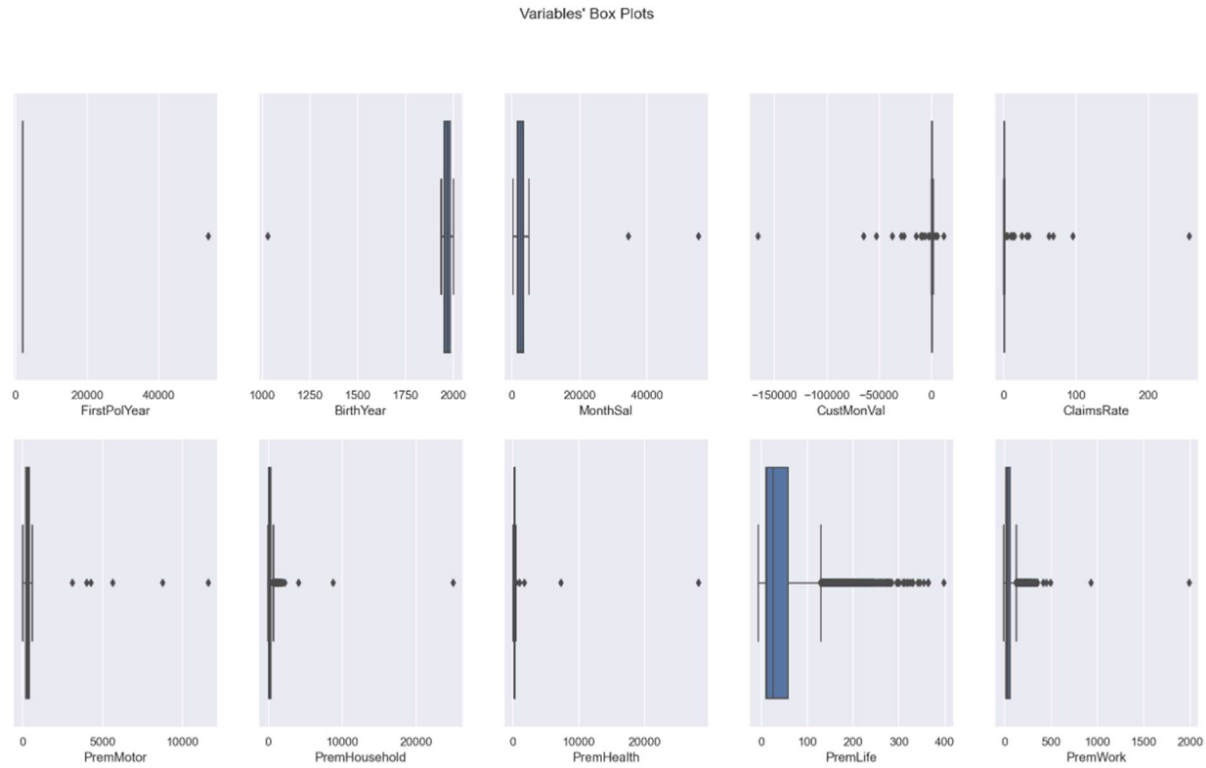
Rishi, R. K. (2022, February 1). Plot multiple boxplots in one graph in Pandas or Matplotlib. Retrieved from https://www.tutorialspoint.com/plot-multiple-boxplots-in-one-graph-in-pandas-or-matplotlib

8. Appendix

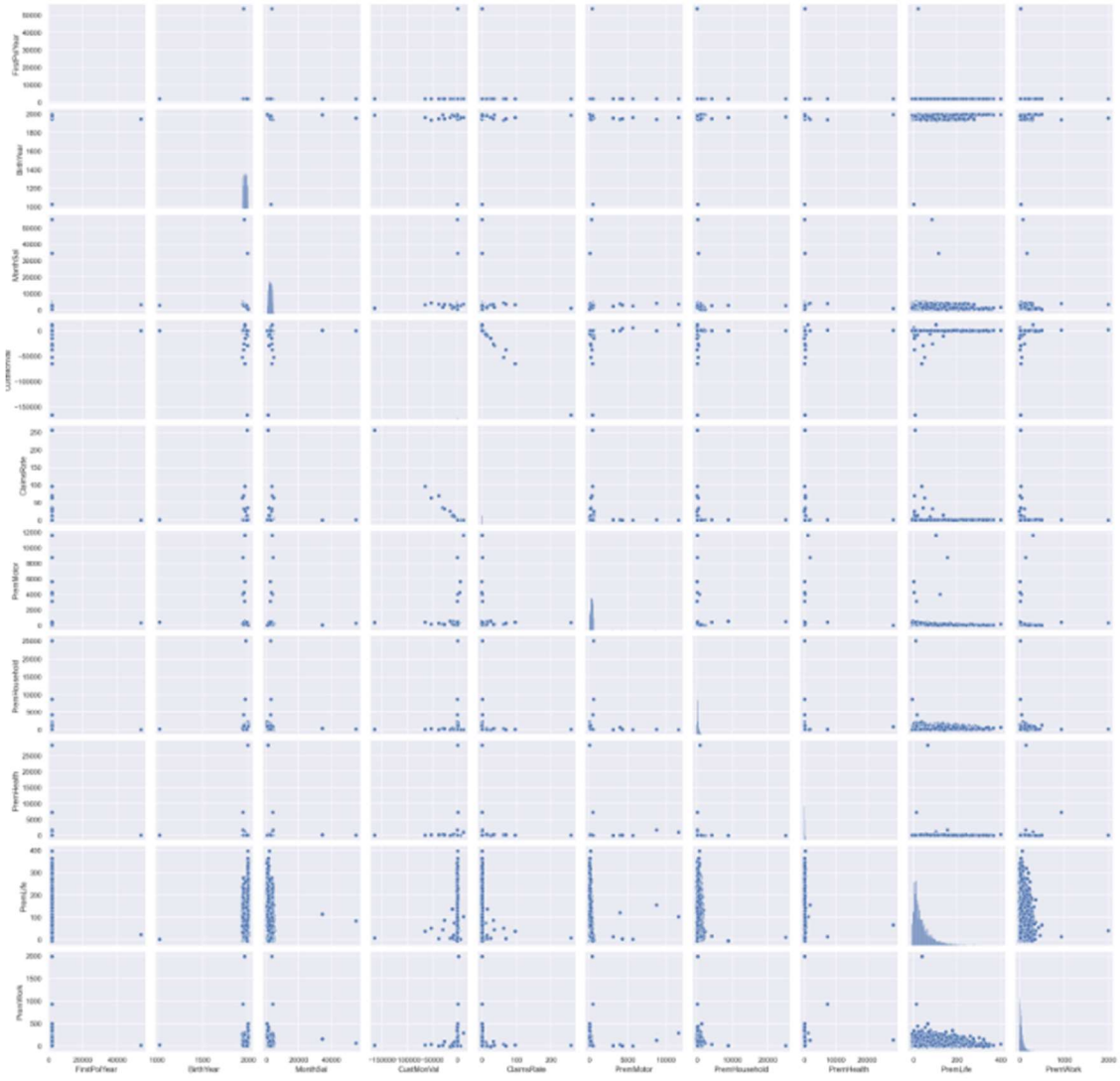
8.1. Visual Exploration Before Preprocessing

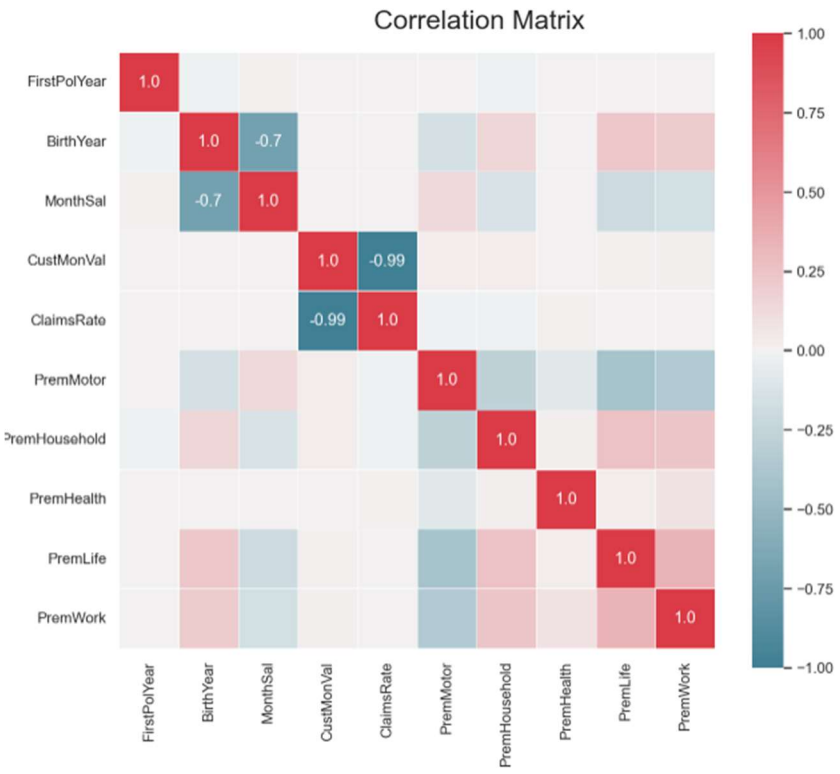
8.1.1. Numeric Variables



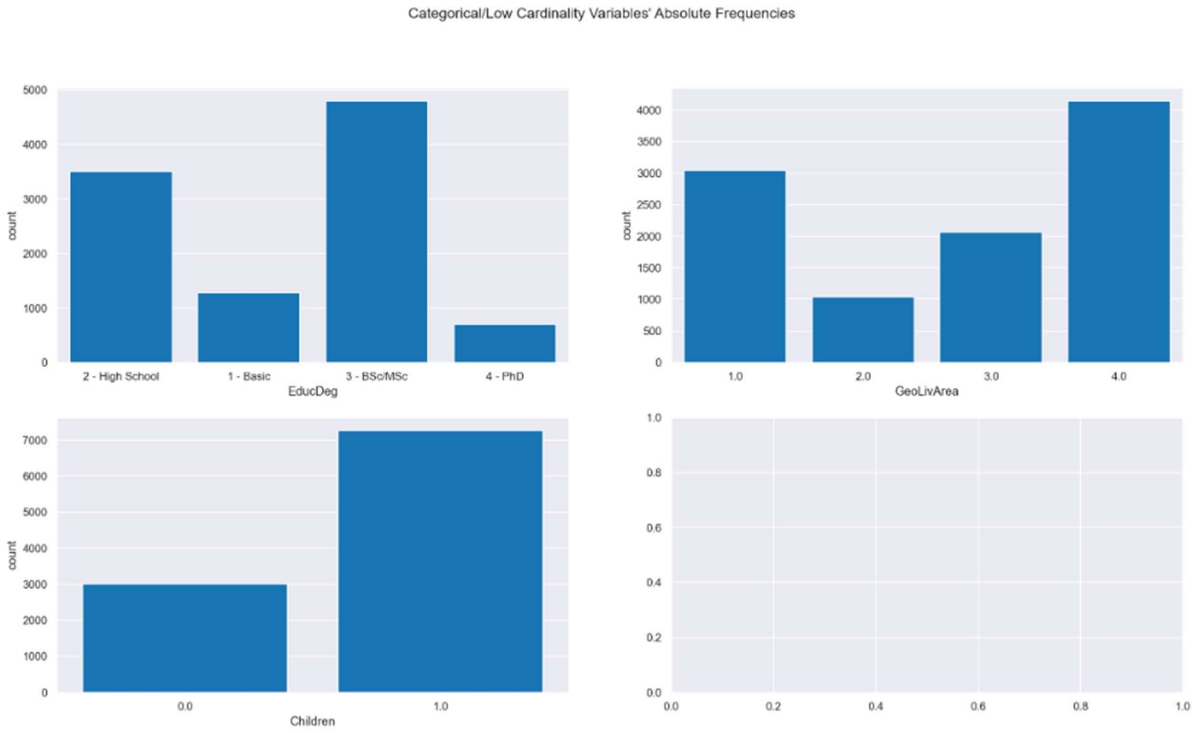


Pairwise Relationship of Numerical Variables

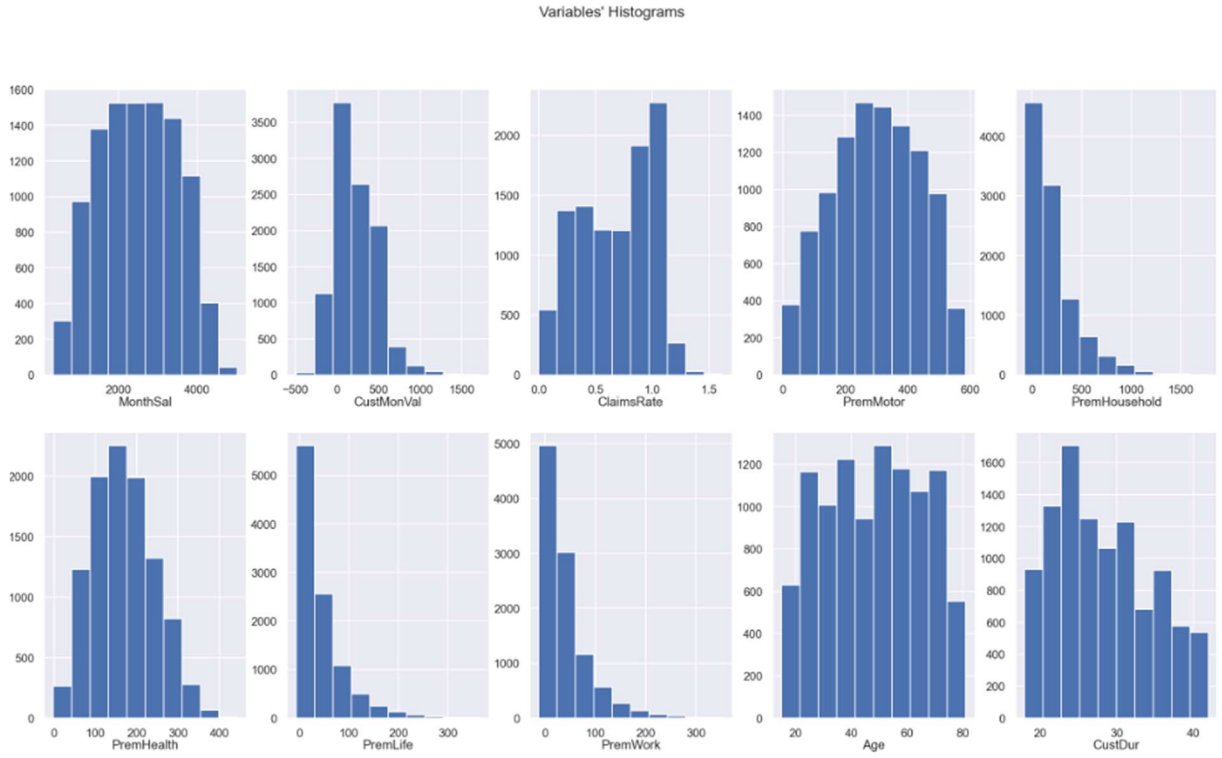


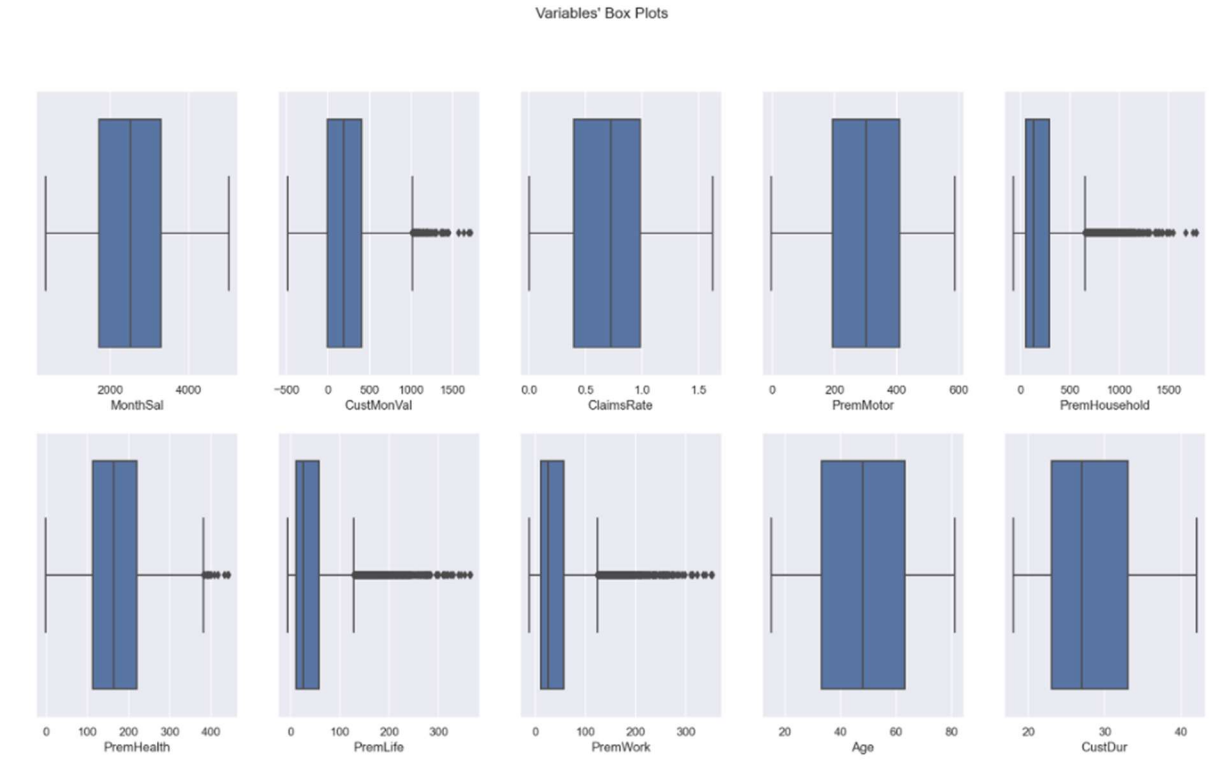


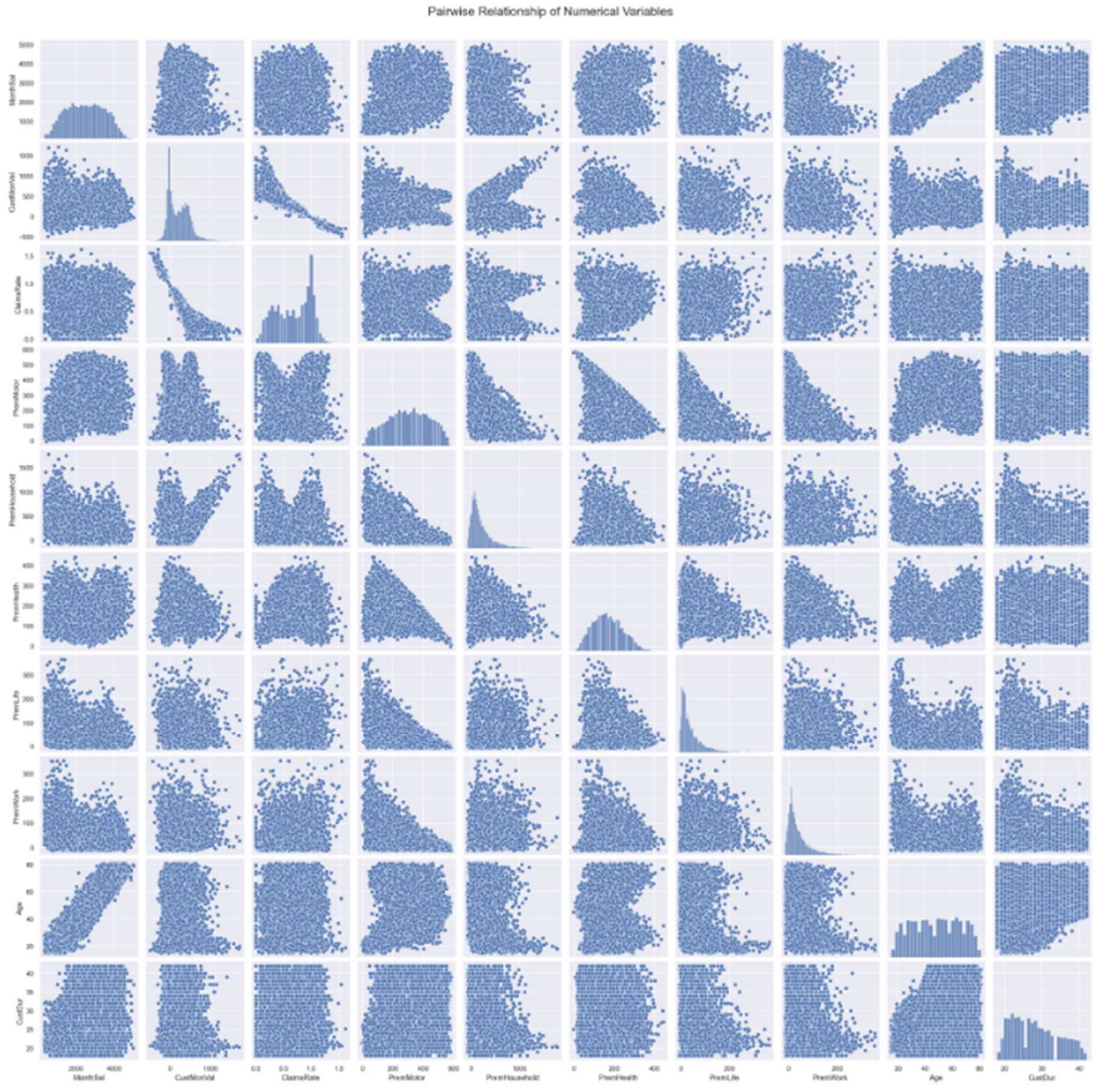
8.1.2. Categorical Variables

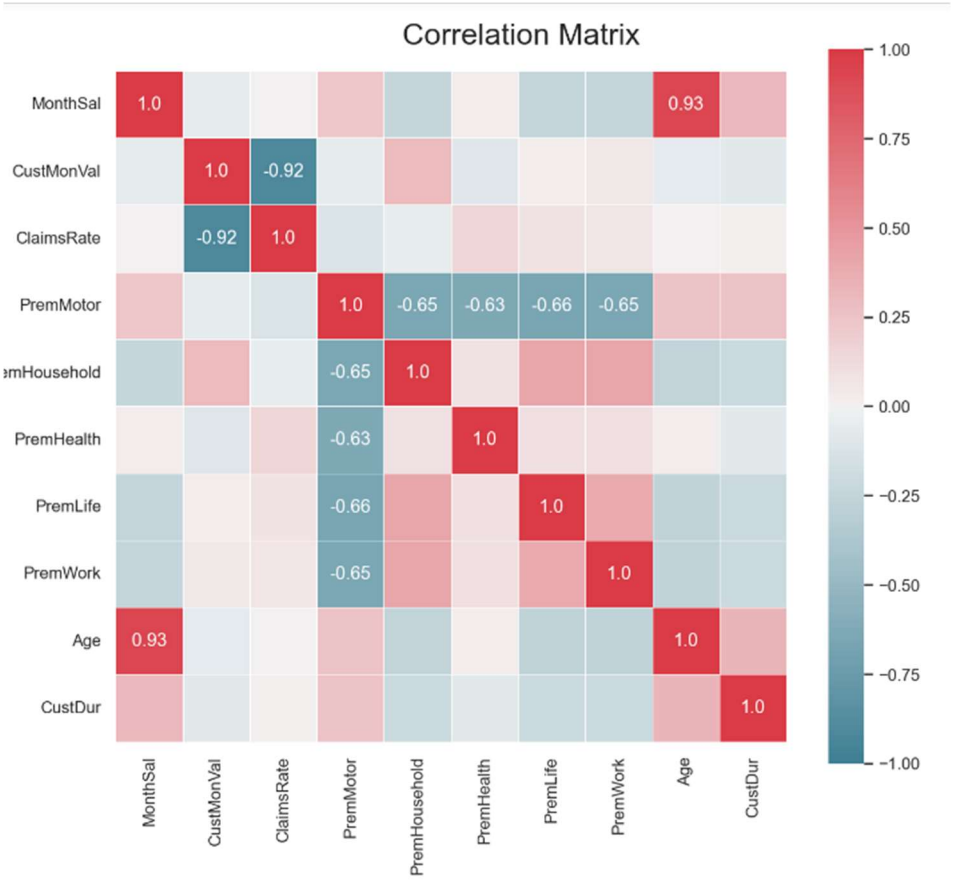


8.2. Visual Exploration after Outlier Removal

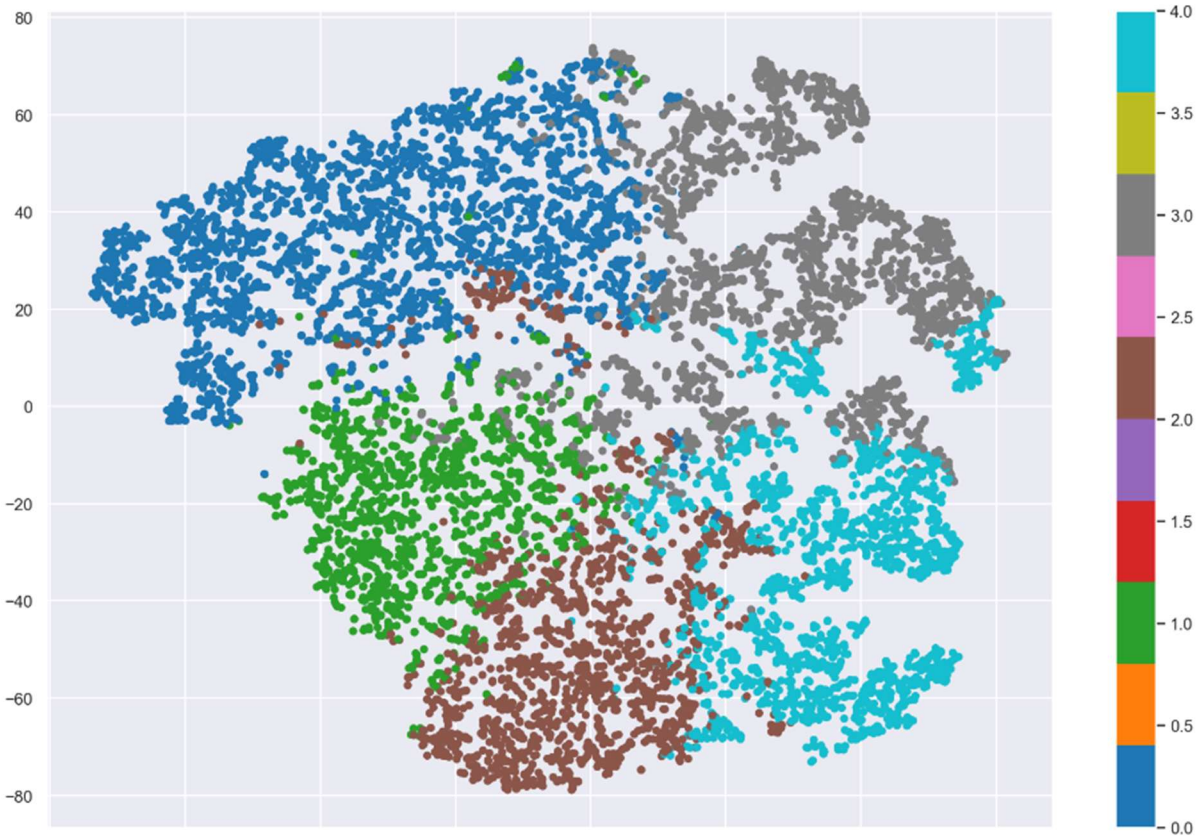




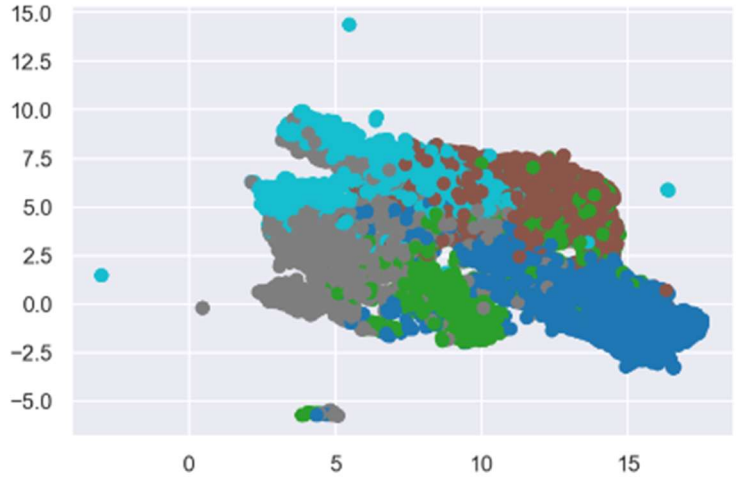

Pairwise Relationship of Numerical Variables
MonthSal
CustMonVal
ClaimsRate
PremMotor
PremHousehold
PremHealth
PremLife
PremWork
Age
CustDur



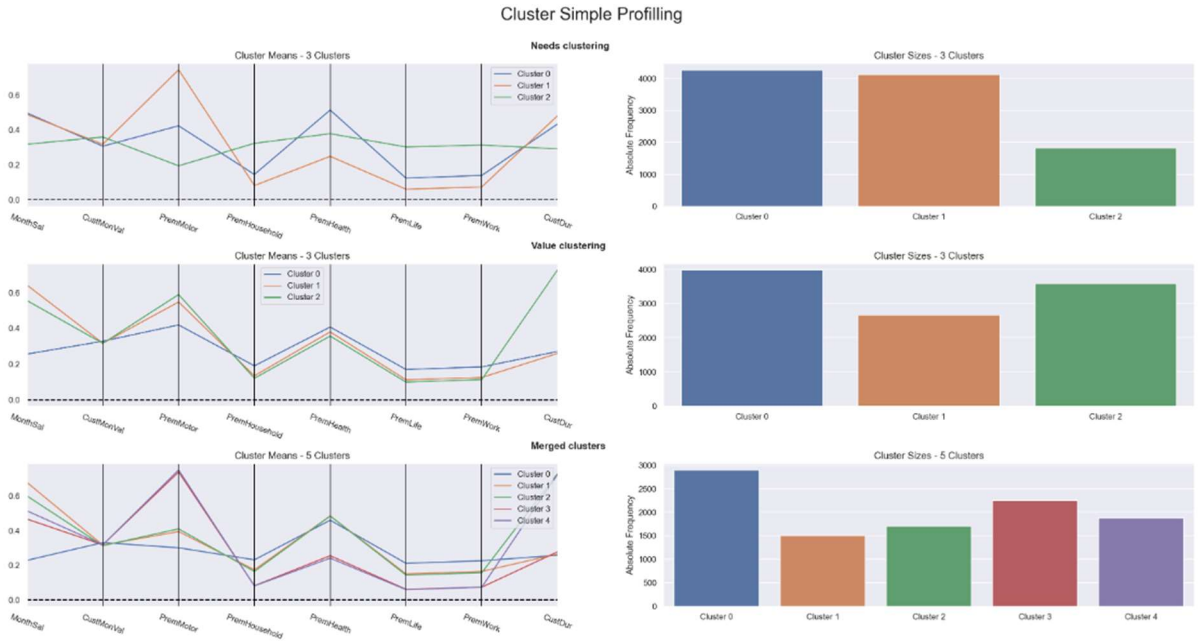
8.3. Cluster t-SNE Visualization



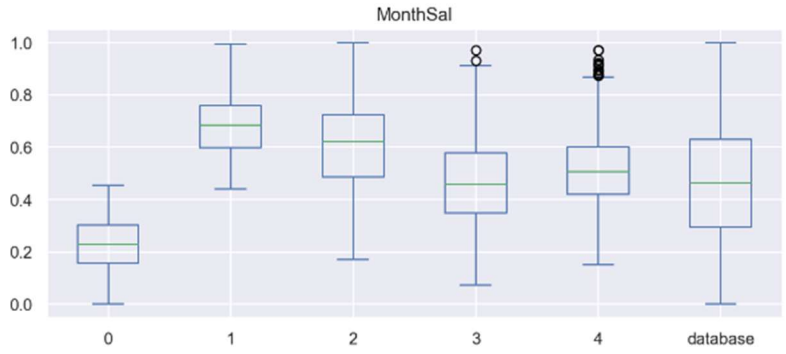
8.4. Cluster UMAP Visualization

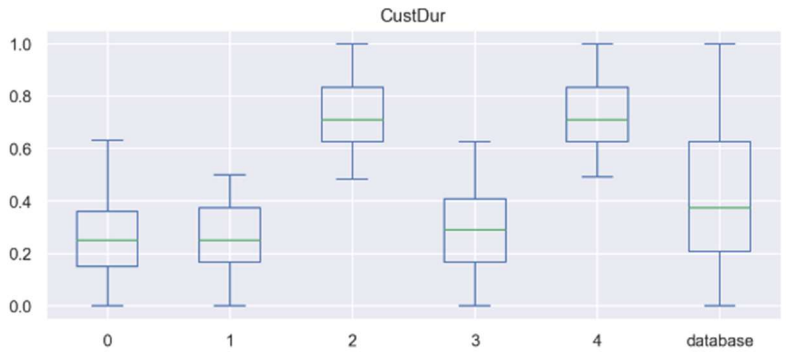


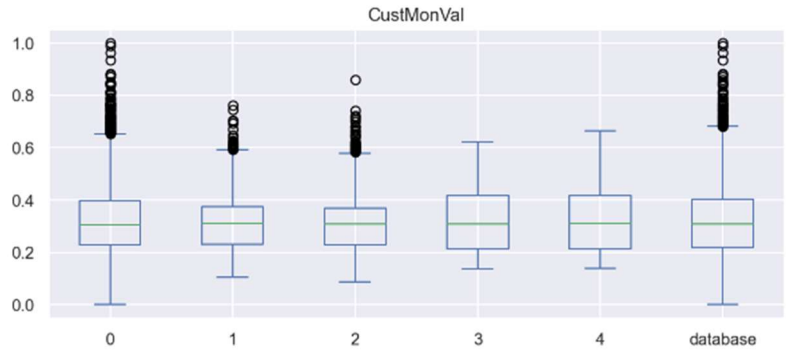
8.5. Cluster Simple Profiling

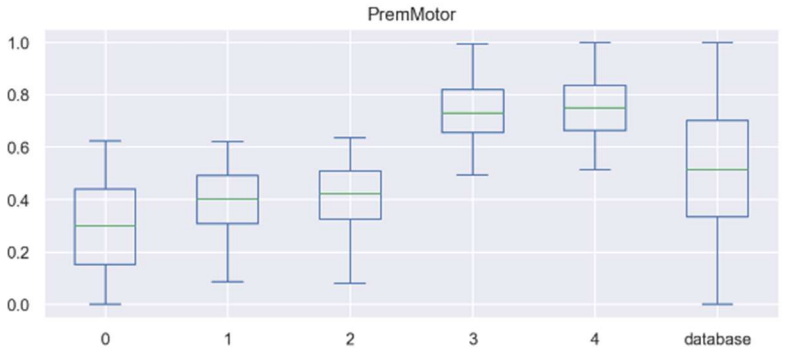


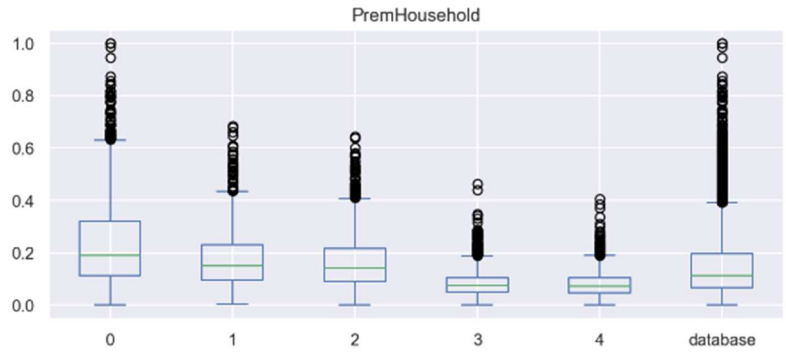
8.6. Profiling (comparing boxplots between clusters and dataset)

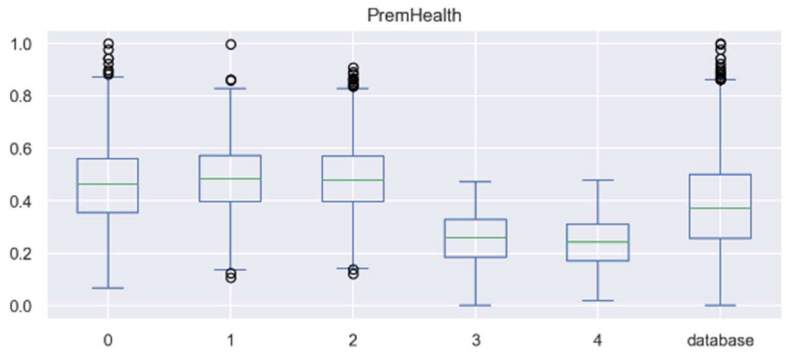


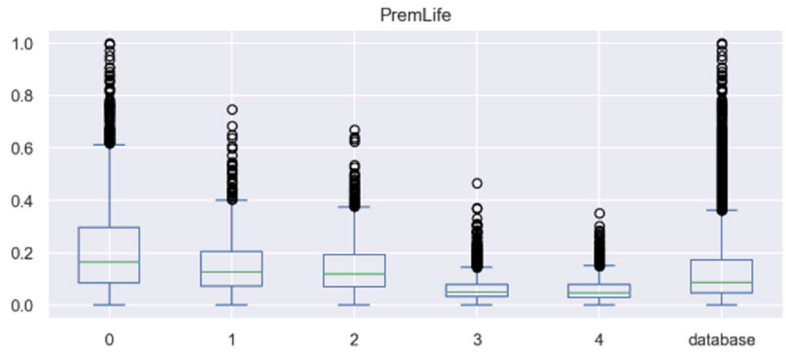


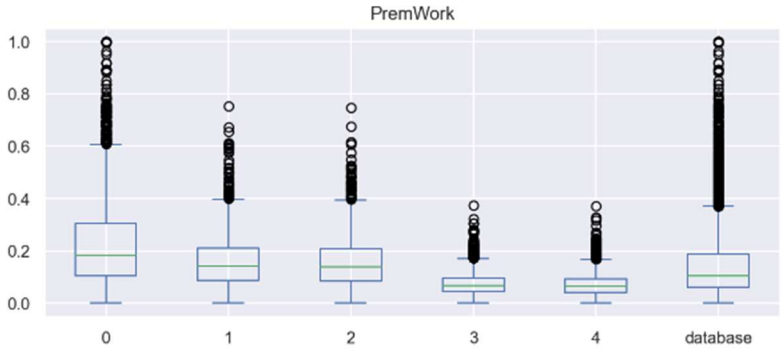




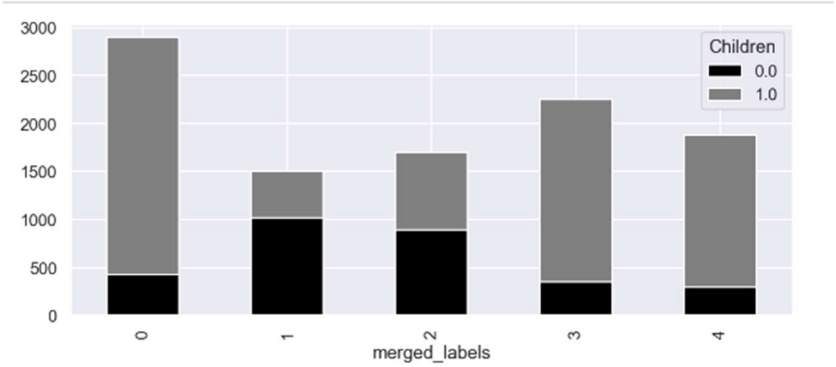

PremHousehold
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
database

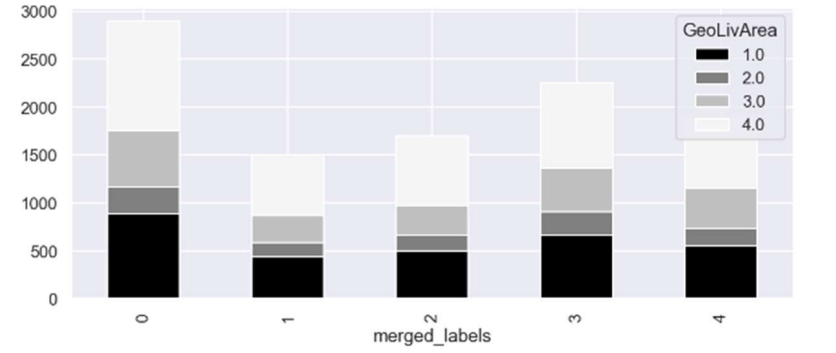

PremHealth
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
database

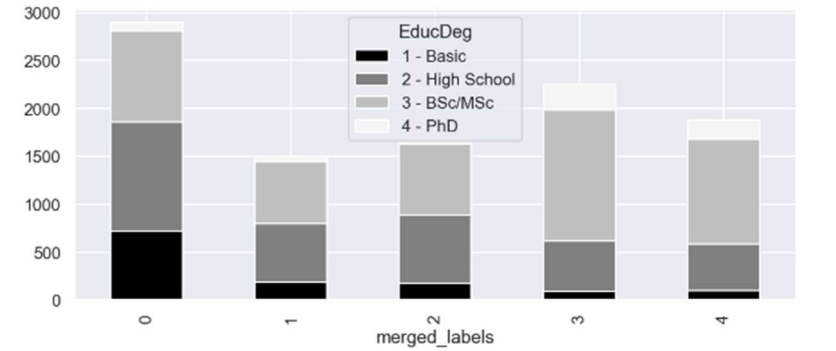

PremLife
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
database


PremWork
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
database

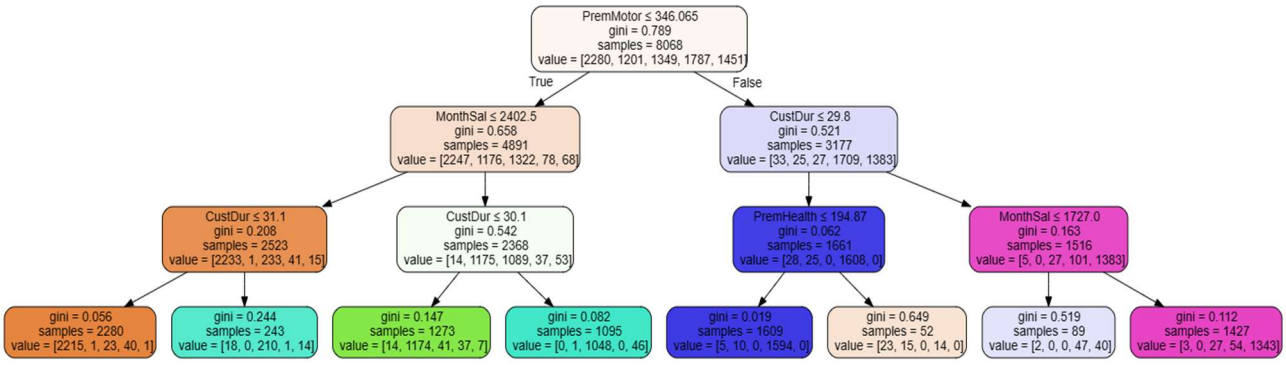
8.7. Density of categorical variables for every cluster







8.8. Decision Tree to assess feature importance



8.9. Reclassified Outliers

CustID	EducDeg	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife	PremWork	Age	CustDur	outlier	merged_labels
146	1 - Basic	2554.0	1.0	1.0	-96.11	1.12	144.36	-11.10	381.85	47.23	12.78	30.0	34.2	-1	2
282	3 - BSc/MSc	3083.0	1.0	1.0	-175.00	1.30	506.43	-75.00	55.90	5.89	1.78	54.0	39.0	-1	4
288	2 - High School	2193.0	3.0	1.0	618.80	0.00	139.58	82.25	246.49	72.57	102.91	41.0	28.0	-1	0
502	2 - High School	2483.0	1.0	1.0	625.68	0.00	160.03	73.35	129.69	116.91	170.70	48.0	38.0	-1	2
575	1 - Basic	3110.0	4.0	1.0	-233.26	1.26	116.02	255.60	163.81	60.79	208.26	59.0	29.0	-1	1
818	4 - PhD	1541.0	4.0	1.0	640.57	0.00	329.73	83.35	222.93	6.89	22.67	24.0	25.4	-1	0
831	2 - High School	2126.4	4.0	0.0	475.43	0.00	274.83	0.00	180.59	25.45	19.56	74.0	28.0	-1	0
1070	1 - Basic	566.0	4.0	1.0	-71.01	1.02	14.56	1777.55	49.23	46.01	121.69	18.0	20.4	-1	0
1076	2 - High School	3106.0	1.0	1.0	-262.04	1.41	273.05	-25.00	114.91	124.69	87.35	67.0	23.0	-1	1
1238	2 - High School	2777.0	4.0	1.0	-162.58	1.19	308.28	148.35	129.47	119.80	17.78	59.0	27.0	-1	1
1363	2 - High School	3904.0	3.0	0.0	-32.99	1.02	251.49	-60.00	133.47	126.58	74.57	67.0	41.0	-1	2
1468	3 - BSc/MSc	1864.0	4.0	0.0	-118.46	1.13	312.28	143.35	125.69	117.91	24.45	41.0	39.0	-1	2
1900	3 - BSc/MSc	1512.0	1.0	0.0	495.98	0.21	275.72	74.45	137.58	15.78	155.03	29.0	22.0	-1	0
1992	3 - BSc/MSc	3544.0	1.0	1.0	88.59	0.79	305.28	-55.00	154.25	-0.11	136.58	61.0	37.0	-1	2
2009	2 - High School	1902.0	3.0	1.0	382.33	0.28	293.39	-16.10	108.13	16.78	162.92	38.0	24.0	-1	0
2015	2 - High School	3644.0	4.0	0.0	-150.91	1.20	292.72	14.45	107.91	59.01	140.36	64.0	35.0	-1	2
2038	2 - High School	3431.0	1.0	1.0	399.86	0.40	158.03	161.70	58.79	70.68	263.16	59.0	28.0	-1	1
2044	1 - Basic	1089.0	3.0	0.0	1716.00	0.11	67.90	1673.10	65.90	112.02	27.56	20.0	20.2	-1	0
2583	2 - High School	2607.0	4.0	1.0	-184.03	1.20	113.91	264.50	199.15	159.03	41.01	55.0	41.0	-1	2
2594	2 - High School	3745.0	4.0	0.0	391.64	0.26	182.59	0.00	146.14	234.49	-0.22	68.0	22.0	-1	1
2793	1 - Basic	4659.0	1.0	0.0	53.46	0.93	86.35	711.80	142.36	105.02	112.91	74.0	29.0	-1	1
2888	2 - High School	3175.0	4.0	0.0	1191.34	0.20	108.91	1148.55	144.25	24.67	92.46	66.0	27.0	-1	1
3011	1 - Basic	3150.0	1.0	0.0	-131.23	1.17	154.14	77.25	110.91	183.48	102.13	68.0	23.0	-1	1
3388	3 - BSc/MSc	2793.0	1.0	0.0	-298.91	1.54	302.17	-75.00	165.03	-2.00	119.91	55.0	37.0	-1	2
3538	1 - Basic	1812.0	4.0	1.0	-6.33	0.99	30.56	1478.60	129.69	153.14	11.89	26.0	23.4	-1	0
3852	1 - Basic	3324.0	1.0	1.0	302.06	0.54	169.81	157.25	100.02	270.94	15.67	59.0	29.0	-1	1
3982	3 - BSc/MSc	2440.0	4.0	1.0	389.08	0.36	167.92	68.35	211.93	177.81	25.56	56.0	41.0	-1	2
4201	2 - High School	2961.0	1.0	0.0	430.31	0.32	130.47	108.35	107.02	115.91	208.93	60.0	27.0	-1	1
4431	2 - High School	3262.0	1.0	1.0	-166.69	1.23	353.40	24.45	82.46	79.57	86.57	62.0	40.0	-1	4
4976	2 - High School	2627.0	1.0	1.0	-125.68	1.18	240.38	-30.00	129.47	70.68	138.47	57.0	33.0	-1	2
5124	2 - High School	1375.0	1.0	1.0	-202.12	1.34	237.60	-70.00	253.38	62.90	37.12	34.0	26.4	-1	0
5350	2 - High School	2783.0	1.0	1.0	-206.70	1.26	137.25	89.45	241.60	18.78	200.48	60.0	19.0	-1	1
5378	3 - BSc/MSc	1783.0	4.0	0.0	463.98	0.22	246.49	63.35	78.46	139.47	103.02	33.0	30.0	-1	0
5570	1 - Basic	1633.0	1.0	1.0	-241.81	1.37	181.48	-5.55	197.37	65.79	145.47	27.0	25.0	-1	0
5912	2 - High School	3909.0	4.0	0.0	-175.03	1.26	274.94	13.90	145.25	82.46	67.57	61.0	31.0	-1	2
5919	2 - High School	2609.0	1.0	1.0	359.21	0.33	239.38	-11.10	165.03	13.78	167.03	62.0	31.0	-1	2
5961	3 - BSc/MSc	1751.0	4.0	0.0	-278.91	1.49	245.38	-70.00	185.59	48.12	113.91	37.0	32.0	-1	2
6395	2 - High School	3823.0	3.0	1.0	426.65	0.23	204.26	-11.10	134.47	87.46	171.03	58.0	20.0	-1	1
6694	2 - High School	3243.0	3.0	1.0	431.53	0.31	212.93	75.00	129.47	9.78	233.82	64.0	24.0	-1	1
6743	1 - Basic	1482.0	1.0	0.0	1634.97	0.17	32.56	1748.10	51.01	132.47	44.34	19.0	20.6	-1	0
7225	1 - Basic	3035.0	4.0	1.0	1040.32	0.29	116.91	1123.55	195.26	35.45	22.78	55.0	24.0	-1	1
7288	1 - Basic	1889.0	4.0	1.0	307.62	0.45	45.01	62.80	163.81	275.83	56.79	21.0	24.4	-1	0
8444	2 - High School	3056.0	4.0	0.0	369.29	0.38	120.69	64.45	120.69	102.13	227.71	69.0	21.0	-1	1
8598	1 - Basic	3256.0	1.0	0.0	256.17	0.54	94.35	18.90	156.92	167.92	169.03	70.0	21.0	-1	1
8677	1 - Basic	2279.0	4.0	1.0	-291.16	1.62	65.90	0.00	97.13	130.47	135.69	19.0	25.2	-1	0
8982	2 - High School	2285.0	1.0	1.0	-45.00	1.03	44.12	-10.00	168.81	227.82	155.25	34.0	30.6	-1	0
9236	1 - Basic	566.0	4.0	1.0	1691.43	0.15	14.56	1777.55	49.23	46.01	121.69	18.0	20.2	-1	0
9391	1 - Basic	540.0	3.0	0.0	-490.20	1.55	20.78	290.60	47.12	299.50	186.48	17.0	20.6	-1	0
9393	2 - High School	2306.0	4.0	1.0	-85.01	1.09	161.14	106.70	156.03	181.48	60.01	56.0	18.0	-1	0
9952	2 - High School	1356.0	4.0	1.0	567.56	0.00	287.72	-0.55	185.48	75.68	44.23	26.0	24.8	-1	0
9955	1 - Basic	768.0	4.0	1.0	-416.73	1.55	21.67	136.70	27.45	177.81	350.62	18.0	20.6	-1	0
10027	1 - Basic	1855.0	1.0	1.0	-307.27	1.51	92.13	-30.00	237.71	122.80	129.47	30.0	21.0	-1	0
10117	2 - High School	1958.0	2.0	1.0	-223.23	1.39	163.03	-75.00	177.59	48.23	191.26	28.0	26.6	-1	0
10270	2 - High School	1858.0	1.0	1.0	-207.91	1.34	175.59	-40.55	256.16	142.36	11.00	40.0	26.0	-1	0
10293	1 - Basic	2431.0	3.0	0.0	1405.60	0.00	133.58	1035.75	143.25	12.89	105.13	64.0	39.0	-1	2