Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance – Customer Segmentation

Group AH

Catarina Duarte, number: 20191211

Inês Nascimento, number: 20170746

Susana Dias, number: 20220198

January, 2023

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

INDEX

1. Introduction

This report addresses the final project of the Data Mining course, inserted in the Data Science and Advanced Analytics master's degree of NOVA IMS. For its development, we were provided with a database from an insurance company, A2Z, that contains information about their customers and the type of premiums they bought in 2016. Our main goal is to segment customers, finding relevant clusters, and helping the company understanding the value and demographics of each customer segment. With that information, it is possible to understand which types of insurance each cluster will be more interested in buying, as well as develop targeted marketing strategies.

2. Data and Variables Description

Firstly, we started by importing the provided data to our notebook so we could get an overview of what type of data and variables we had. This initial dataset was composed by 14 variables - Table 1 - and 10.296 observations, clients. Since one of the features was *CustID*, and upon checking that this was composed only by unique values, we decided to set it as the index for the project.

3. Data Exploration & Visualization

The team started the exploration part by checking the data types of the variables we had - Table 1-, as well as understanding which variables might have missing values. Regarding the data types, *EducDeg* was the only variable stored as an object, while and the rest of them were stored as a float. By taking into consideration different forms of missing values besides a null value, we found that almost every variable has absent data. The total values are also presented in Table 1.

Furthermore, the team found 3 duplicated values, which will be removed in the Data Preprocessing chapter, and, after interpreting the descriptive statistics of the numerical features – Table 2, we understood that almost every numerical variable might have extreme values, outliers. Finally, we verified that all existing values in *Children* and in *GeoLivArea*, made sense in the context of the problem.

Moving forward to the data visualisation step, and since the metric and non-metric features are visualized differently, we first started by splitting out features into metric and non-metric. The division of these features can be found in Table 3. For the metric features, we made boxplots -Figure 1-, histograms -Figure 2- and their pairwise relationship among each other -Figure 3. It is relevant to say that with these visualizations, we confirmed that every metric feature was in the presence of outliers. Concerning the other two visualization tools, once the presence of outliers disturbs the scale of the visualizations, we were not able to identify any other important information, therefore we decided to perform visualizations, again, after the pre-processing step. About the non-metric features, we first plotted its absolute frequencies -Figure 4, where we understood the variables distribution, and then we plotted each non-metric feature boxplot along all metric features – Figures 5 to 7. With these, we concluded that the *GeoLivArea* feature boxplots do not change with any of the metric variables, therefore it doesn't present discriminating power useful for our segmentation. In addition, most customers have at least 1 child, and most of the clients have either a bachelor's or a master's degree.

4. Data Preprocessing

Concerning the 3 previously mentioned duplicates in our dataset and considering that it is very unlikely two costumers to have the exact same behaviour in all categories, we believe that the duplicated rows refer to the same costumers due to an error of registration, as being another *CustID*. For this reason, we removed the duplicated rows.

Moreover, the team decided to create 7 new features: *Age, Longevity, Total_Premiums, Annual_Salary*, each premium proportion on the total spent in premiums, Commitment, and a binary variable for negative premiums. More details about these that can be viewed in Table 4.

Incoherencies can be described as values that are not correct given the context of the problem. In this sense, it is important to look for them, and treat them, to obtain a clean dataset to be worked on. Hence, we started by looking for customers whose first year as a customer was before their birth year. In fact, we found that 1.997 customers suffered from this problem. Secondly, we looked for customers whose first year as a customer was after 2016, the year of the database, where only 1 customer was found, and further removed. Then, in our understanding, it might not be the best approach if A2Z has customers who have not reached the age of majority yet, 18 years old, and we found that 116 customers were not 18 years old yet. After this, we looked for customers who might already be dead, using the European Union average age of death, 83 years old, and we found 1 observation where the *Birth_Year* was 1028, which we consider an error, so this observation ended up being removed. Finally, the last two incoherencies we looked for were: customers who spent more money on premiums than their annual salary, and customers who did not spend any money on premiums in the year of the database. For the first one, we found 1 customer, and ended up removing this observation; for the last incoherence, no customers were found. Having this said, in this step, we removed 3 rows.

Once the 2 main inconsistencies previously identified, those related to customers whose first year as a customer was after 2016 and customers were minors, were determined to be related to the variables Birth_Year or Age, our team agreed to continue without addressing them in order to further investigate these variables and address these inconsistencies in later steps.

Through the simple interpretation of the boxplots made before, the group could identify outliers in every metric feature, and decided to take two approaches to treat these: the IQR[1] Method and the Manual Filtering method. However, we must take into consideration that 6 observations were already removed, therefore, we already removed some outliers, as we can see in Figure 8, which is the boxplots right before starting to treat the outliers. On one hand, we executed the IQR method, where around 20% of the observations were removed. This is considered a high loss on information, and we will not keep this approach. On the other hand, we performed the Manual Filtering method, where we manually set a restriction for each variable we want. Our team agreed to perform this only for the extreme values of each variable for two main reasons: because there were many outliers, and removing them all, as seen in the previous method, would remove much information, and because those values might represent real behaviour of A2Z customers. The restrictions applied can be seen in

[1] Inter Quartile Range

Table 5. This method removed around 3% of the observations, which is an acceptable percentage; therefore, we proceeded with this method.

After treating the outliers, more visualizations were performed – see figures 9 to 11 in the Appendixes, and, once again, we checked how the non-metric features behaved along with the metric features, including the new variables created. After this, and because we believe the *Educ* feature must be stored as a numerical ordinal feature, so we converted its values to numbers, being 1 the lowest academic degree – Basic – and the maximum the value 4, which stood for Phd[2].

At this point, our team proceeded to data scaling, so the features would be all in the same range. We started by checking the histograms of the variables without the extreme outliers - Figure 10, and understood that not all features represent a normal distribution of their values. Taking that into considerations, the Standard scaler and the MinMax scaler were performed. We cannot fail to state that these scalers were not only applied for the metric features, but also for the *Educ*. Due to this feature being ordinal it must be treated either as if it was continuous, or if it was a categorical, and consequently encode their values. We decided to take the first approach, and, for the reason mentioned before, we chose to keep going with the MinMax scaler, having our features scaled between 0 and 1, since the presence of negative values was represented in the *Negative_Prem* variable. Take into consideration that we will use the data standard scaled for the execution of PCA[3] in further steps. Moving on to the encoding, the One-Hot encoder was applied to the *Children*, *GeoLivArea* and *Negative_Prems* features. Once the missing values had not been treated yet, the encoder created a variable *Children_nan* to the one and only missing value of the *Children* feature, which we further converted into a NumPy missing value and deleted the column.

Considering the pre-processing already done in prior steps, when checking again for the presence of missing values at this point, only 27 observations showed up: 14 in the *BirthYear* and *Age*, and 13 in the *Children_1.0* features. Two different executions were tried to overcome this: deleting the observations – which removed round 0.27% of data; and imputing values with the KNN[4], using 5 nearest neighbors. Due to not wanting to remove more information from the data, our group decided to keep the imputation with KNN approach and verified that the results were acceptable values.

The feature selection part was initialized with the performance of some filter methods, more specifically, by checking the variance for the non-metric features where no univariate variables were found. After that, and although it doesn't detect non-linear relationships, we opted to use the Spearman correlation – Figure 12 - due to capturing monotonic relationships. The main insights from this matrix were: the pairs of variables *Age* and *BirthYear*, as well as *Longevity* and *FirstPolYear*, and *Annual_Salary* and *MothSal* had a perfect correlation among each other since they represent the same information. In addition, we were able to state that *Longevity* is not correlated with any other features; *BirthYear*, *Commitment* and *Annual_Salary* as well as the pair *CustMonVal* and *ClaimsRate*, and the pairs *PremMotor* and the other premiums, are highly correlated among each other. Having in mind the existence of plenty observations with the two major inconsistencies found before in the *BirthYear* variable, and fact that this variable is highly correlated with both *Commitment* and *Anual_Salary,* our

[2] PhD stands for Doctor of Philosophy
[3] Principal Component Analysis
[4] K-Nearest Neighbors algorithm

team decided to remove this problematic variable from the project. By the large number of inconsistencies related to this feature, it seems like this variable suffered from some input errors and it is advised for A2Z Insurance to further analyse this internal problem. Thus, regarding the metric features, our team decided to remove *BirthYear, Age, FirstPolYear, Longevity, MonthSal, ClaimsRate* and *Total_Premiums.* As for the non-metric variables, and after analysing figure 14, boxplots of the GeoLivArea not changing with any of the metric variables, therefore not presenting discriminating power useful for our segmentation, our group decided to also remove this variable.

For a better analysis and clustering, and after treating and understanding the data we have, our group decided to create two different views: Product and Demographic. In the Product perspective, information about the products bought by customers - more specifically, the ratios variables[5]- as well as the customer monetary value and the *Commitment* ratio, were included. Regarding the Demographic perspective, only three features were involved: *Educ*, *Annual_Salary* and *Children_1.* With this said, we agreed to split the Clustering Process into three parts: the Product Perspective, the Demographic Perspective, and the final cluster solution, performing, separately, clustering algorithms for each view, and at the end, merge each of the best results. Table 6

In an attempt to reduce the dimension of our data, the team conducted a PCA analysis for the Product view since this was the only with a number of features worth reducing. However, our results found on Figures 15 to 17 showed that the number of PC to retain was not only ambiguous between the methods results but also laid in a large range make it unsuccessful at its original goal of reducing dimensions. For that reason, we didn't follow through with adding the PCA results to the dataset.

5. Clustering Algorithms

5.1. Product Perspective

5.1.1. Hierarchical Algorithm

To start with, it was employed the popular hierarchical algorithm using only the metric features of the product view. Based on the R^2 score for each linkage method - Figure 18, our team opted to use the ward linkage, and test this algorithm for both 3 and 4 clusters due to the results of the dendrogram – Figure 19. We then evaluated which of these attempts led to better interpretability by examining the means of the variables in each cluster solution and the R^2 scores – Figures 20 and 21.

5.1.2. Combination of K-Means and Hierarchical Algorithm

In order to obtain the best results possible using the hierarchical algorithm, our team conducted the latter after performing a K-Means clustering. Since hierarchical clustering can perform poorly on large datasets, this combination of algorithms intends to join the closest data points from the start by initializing K-Means with a large k (in our case, 100) and passing its centroids as an input to build the hierarchical clusters. To choose the linkage method we built a R^2 plot for the various methods – Figure 22- and for the number of clusters a dendrogram – Figure 23. The final results can be found in Figure 24 for 3 clusters with Ward as the linkage method.

[5] Ratios variables were created in the Pre-processing part and can be found on Table 4 of the Appendixes

5.1.3. DBSCAN

Moving forward, our group attempted to use the DBSCAN for clustering, however, we achieved poor results with this algorithm. Despite our efforts to tune the possible parameters, the results always presented a low number of clusters and/or an unbalanced distribution of data among the same. To find the right parameter for the *eps* argument, we plotted the K-distance – Figure 25. This may have happened due to our underlying segments being of different densities and this being a case where DBSCAN performs poorly.

5.1.4. Self-Organizing Maps

Lastly, we employed the self-organizing maps technique. We trained the algorithm over different dimensions but ultimately preserved the 50x50 dimension. Using the component planes - Figure 26- it was detected that *Commitment* had a non-cohesive pattern. In order to address this possible issue, we built models both with and without this variable to compare results. As visualization tools for the outputs, we also used U-maps and Hitmaps that facilitated in the solution evaluation, which can be found in figures 27 and 28, respectively.

On top of the SOM model, we tried to run both K-Means and Hierarchical methods using the units from the first and input to the latter. For K-Means, we first performed an inertia plot - Figure 29- correspondent do different number of clusters from where we retrieve the best number, 3. The results - Figure 30 - were unsatisfactory, seeing that the division of clients among the clusters was very imbalanced. This made unworthy to build marketing strategies for small clusters due to being such a niche of our clients and would make the strategies for the large clusters too general and, therefore, would not appeal to everyone since it was not segment enough.

For Hierarchical, based on the knowledge retained from the employment of this specific algorithm, we carried with 3 clusters and the Ward linkage. Unfortunately, the results - Figure 31- were unsatisfactory in the same way described above due to the disparity of cluster sizes.

5.2. Demographic Perspective

5.2.1. K-Prototypes

Due to having only one metric feature in this perspective, *Annual_Salary*, and performing a cluster algorithm with a single variable is equivalent to binning, our group decided to research algorithms able to handle mixed data. Being the best of both algorithms, K-Prototypes is able to handle numeric data, like K-Means, and categorical data, like K-Modes. Similarly to other times using a partition method, to choose the number of clusters we plotted the inertia for a wide range of possible values - Figure 31- and the final decision was 2 clusters. Then, we trained the algorithm with the categorical variables designated. Even though the distribution of clients per cluster was not the most desirable but acceptable, the characterization of the clusters was very poor – Figure 32- since they had very similar means and the same strategies would appeal to both groups.

5.2.2. Hierarchical Algorithm with Gower distance

Although it was only used for the metric features in the product perspective, hierarchical methods are also able to handle categorical data. For that, it is needed to supply the model with the specific matrix

of distances seeing that Euclidean will not work. To build the distance matrix, our group opted to use Gower distance, a similarity measure capable of dealing with mixed data types. We compared results for both 2 and 3 clusters, since we knew that, due to our number of variables and previous results, a higher number of clusters would not be beneficial for this perspective. In addition, for each number os clusters we also tried both "Complete" and "Average" linkage type, although they equally produce the same clustering solution when creating two agglomerations. In contrast, with the K-Prototypes methods, these clusters presented different behavior – Figure 33- and therefore were more helpful for our goal.

5.3. Final cluster solution

After many attempts with different algorithms and their respective fine tuning, the cluster method chosen for the product perspective was a combination of the K-Means and Hierarchical algorithms with 3 clusters. For the demographic view, the Hierarchical Algorithm with Gower distance was chosen.

Focusing first on the product view, our clusters can be characterized in the following way: the first cluster **(cluster 0)** distinguishes itself by spending most of its total premiums in the health category and presents average values in other categories. Although they have the lowest mean on *CustMontVal*, they are highly committed, this is, spend a lot of their income with the company. The second cluster **(cluster 1)** are clients that spend a very low amount on the motor category, however, standout in the life and work category and by a significant amount on household. They present the highest values in *CustMonVal* and the lowest on *Negative_Prems* which means that is unusual for these clients to cancel their services. Lastly, the third cluster **(cluster 2)** presents a major portion of their spending in the motor category and very low values in the remaining categories. On one hand, almost have of the clients of this cluster had negative premiums, this means that for many it is characteristic to cancel subscriptions. On the other hand, these clients exhibit the highest level of commitment.

In regards to demographics, our clusters can be defined as follows: The first cluster (**cluster 0**) consists of people with higher annual incomes and no children, while the second cluster (**cluster 1**) consists of people with lower annual incomes and children. Both clusters have the same mean regarding the education level so no conclusion can be obtained per cluster.

To obtain the final clusters that consider both perspectives, we merged the results and obtained a final label where the first number corresponds to the cluster number from the product view and the second number corresponds to the cluster number from the demographic view. For example, a customer with a final label "20" belongs to the third cluster described in the product clusters and the first cluster described in the demographic clusters. The contingence table can be found in Table 7. Our team wanted to highlight that the cluster 20 has a small proportions of clients and this will be taken into consideration when proposing the marketing campaigns.

6. Marketing Strategies

Cluster 00: This cluster is composed by 1.573 customers who do not have children, have the highest *Annual_Salary* and are the ones who spend more money on Health premiums. Additionally, the customers on this cluster present the highest commitment, that is, spend the highest proportion of their salary on premiums. Our marketing strategy suggestion for these customers, is a type of a Gold Subscription, where we could explore the possibility of cross selling the health insurance these

customers already have, with life premiums, since this is one of the lines of business that they spend the least money on. This pack could raise the interest of these clients, since they are the ones who earn the most annual salary, and, in this sense, make them acquire this pack having of a two in one perspective.

Cluster 01: This cluster includes 2.060 customers who have children, and the premiums they spend the most are Health and Motor premiums. In a marketing perspective, and to make these customers diversify their portfolio for the life premiums, our team suggests betting on a diversification pack: "Pack 4 Life", which includes health, motor and life premiums, with the offer of an insurance associated to a minor child - for example, a life premium. To put this into practice, and assuming that the budget is not a limitation, one suggestion for A2Z to advertise this pack could be through MUPIs[6] scattered around the main European cities with known people, such as actors, football players and musicians.

Cluster 10: This cluster is the one that agglomerates the least clients (984). Here, customers have the highest *CustMonVal* while they are the ones who spend the most in Household and Life, and the second ones in Health premiums. An important definition that A2Z must take into consideration is that customer loyalty and retaining clients is more important than increasing the number of new customers. In this sense, the company must prioritize adopting strategies to try to increase the commitment of these customers. However, and since this cluster is the one with less clients in, our team suggestion to the A2Z company is to first consider doing direct marketing through personalized e-mails promoting other types of services, namely Motor -since this is the premium they spend the least money- and analyze if this strategy has any results. If so, then the company may, in the future, create a cross-selling pack, where there's a promotion when "buying" a Household and Motor premium, having a type of discount for the Life premium.

Cluster 11: In this agglomeration we can find 2.404 customers who are more oriented to pursue family and life related premiums, once they are the ones who spend the most on Work, Life and Household premiums, and have children. Additionally, and since they are the ones who earn the least salary and are also the least committed, our marketing strategy could rely on having a "family" programme, on which the more people of their household they signed for the premiums, on the following year they could have a 5% discount. With this strategy, we would be encouraging the purchase of another premium, as well as almost guaranteeing that these customers stay with A2Z next year.

Cluster 20: Here, there are 2.956 customers who present high values on the Motor premium, and the least on the Household and Health and Life. These customers, in turn, present negative premiums. In this sense, a strategy to be adopted in the first instance for this type of customers could be, for example, to run a kind of promotion in which if the customer insures with A2Z again, gets a 2.5% discount in the first year. Additionally, and again assuming that there is no budget limit, it is proposed to A2Z to hold a raffle for F1 tickets with the purchase of a specific pack for other categories other than motor, since these customers spend almost 70% of their salary on motor premiums and probably are car sport enthusiasts or connoisseurs.

[6] MUPI is a French word that stands for *Mobile Urbain pour l'información*

Like mentioned above, taking into account the small frequency on cluster 20, only with 381 costumers, and that we believe that the marketing strategy developed for cluster 21 will also be effective in these clients, we joined them making the target audience for this campaign a total of 2.956 individuals.

7. Conclusion

In conclusion, our team believes that the project goal was successfully achieved by creating distinctive and useful client segments through the use of clustering techniques. After carefully analyzing and treating the given data and experimenting with several algorithms and paraments, we were capable of uncovering important patterns and grouping the observation according to their similarity. With these groups we were also able to understand the needs and behavior of clients and develop targeted marketing campaigns. Nevertheless, our team would like to once more reinforce the check of integrity of the variable *BirthYear* contained in the original dataset since it showed a large number of problematic information[7]. Overall, the usage of clustering techniques to complete this task was essential and reveled effective.

8. References

Aprilliant, A. (2022, March 31). The K-prototype as clustering algorithm for mixed data type (categorical and numerical). Retrieved January 3, 2023, from https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

Halflingwizard. (2021, September 14). Clustering categorical data using Gower distance. Retrieved January 6, 2023, from https://www.kaggle.com/code/halflingwizard/clustering-categorical-data-using-gower-distance

Kesh, S. (2020, May 20). Concept of Gower's distance and it's application using Python. Retrieved January 6, 2023, from https://medium.com/analytics-vidhya/concept-of-gowers-distance-and-it-s-application-using-python-b08cf6139ac2

3.2 - identifying outliers: IQR method: Stat 200. (n.d.). Retrieved January 6, 2023, from https://online.stat.psu.edu/stat200/lesson/3/3.2

Plot hierarchical clustering dendrogram. (n.d.). Retrieved January 6, 2023, from https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

Ramzai, J. (2021, May 25). Clearly explained: Pearson V/S spearman correlation coefficient. Retrieved January 6, 2023, from https://towardsdatascience.com/clearly-explained-pearson-v-s-spearman-correlation-coefficient-ada2f473b8

[7] To see more about the matter please consult section 4. Preprocessing, page iv

Solyia, S. (2021, Jul 2). Customer Segmentation using k-prototypes algorithm in Python. Retrieved January 5, 2023, from https://medium.com/analytics-vidhya/customer-segmentation-using-k-prototypes-algorithm-in-python-aad4acbaaede

Keany, E. (2021, Nov 1). The Ultimate Guide for Clustering Mixed Data. Retrieved January 2, 2023, from https://medium.com/analytics-vidhya/the-ultimate-guide-for-clustering-mixed-data-1eefa0b4743b

Bação, F. (2022) Data Mining Class slides;

Data Mining Class Notebooks

9. Appendix

9.1. Figures

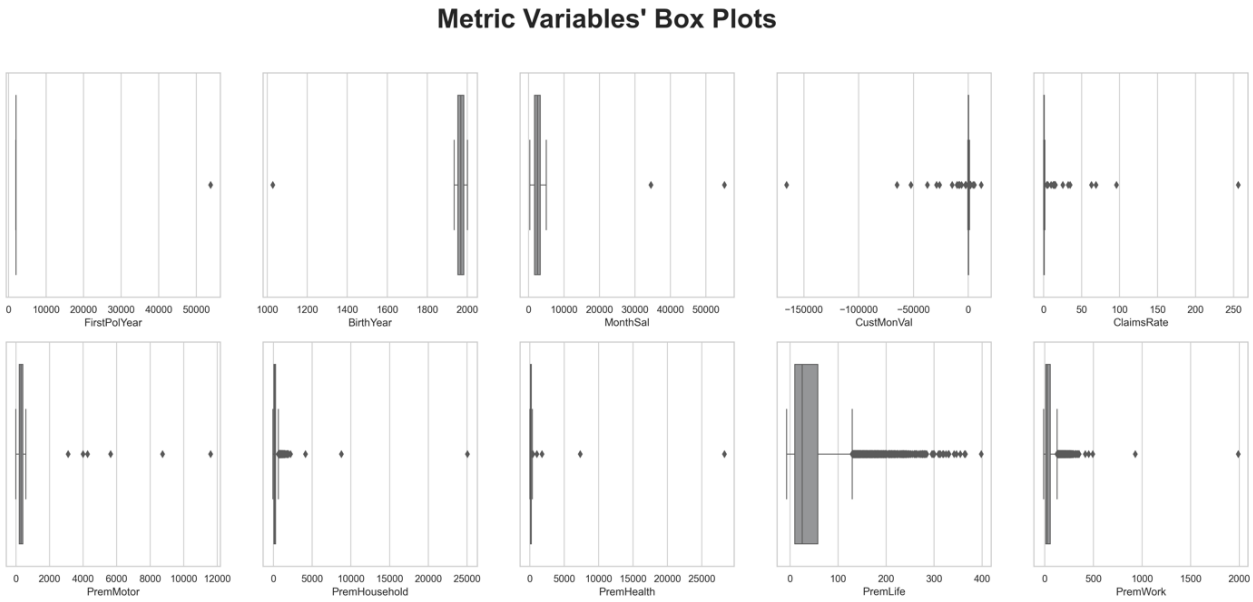

Figure 1 - Initial Boxplots of the Metric Features

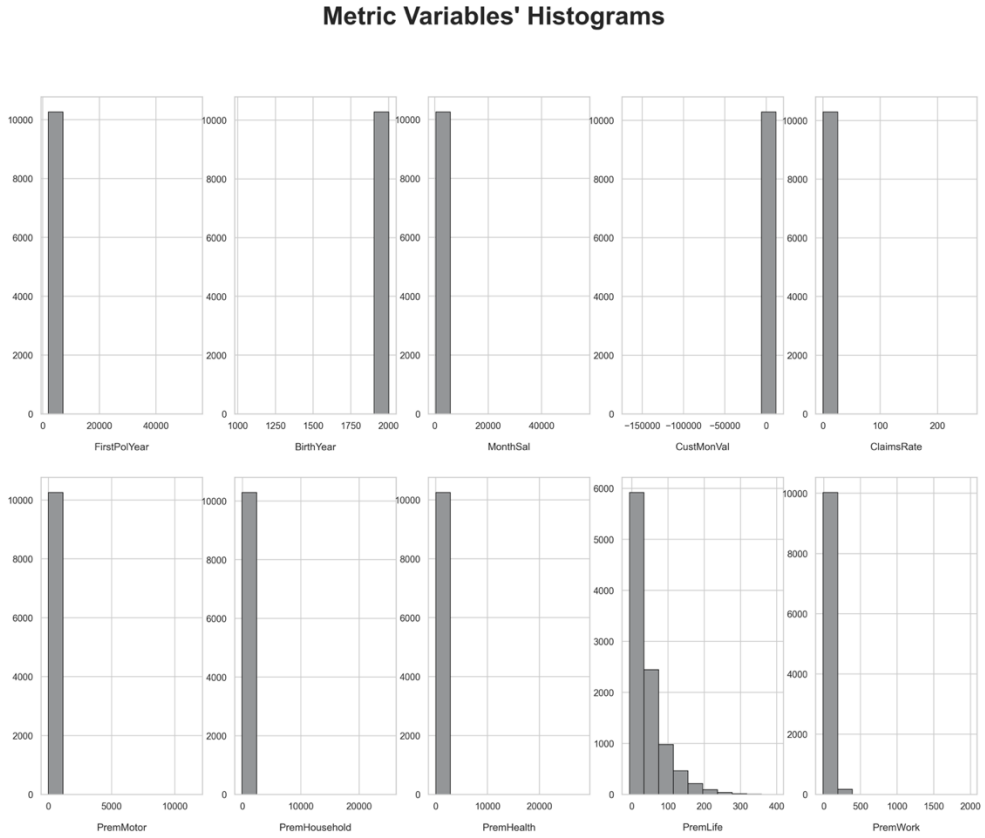

Figure 2 - Initial Histograms of the Metric Features

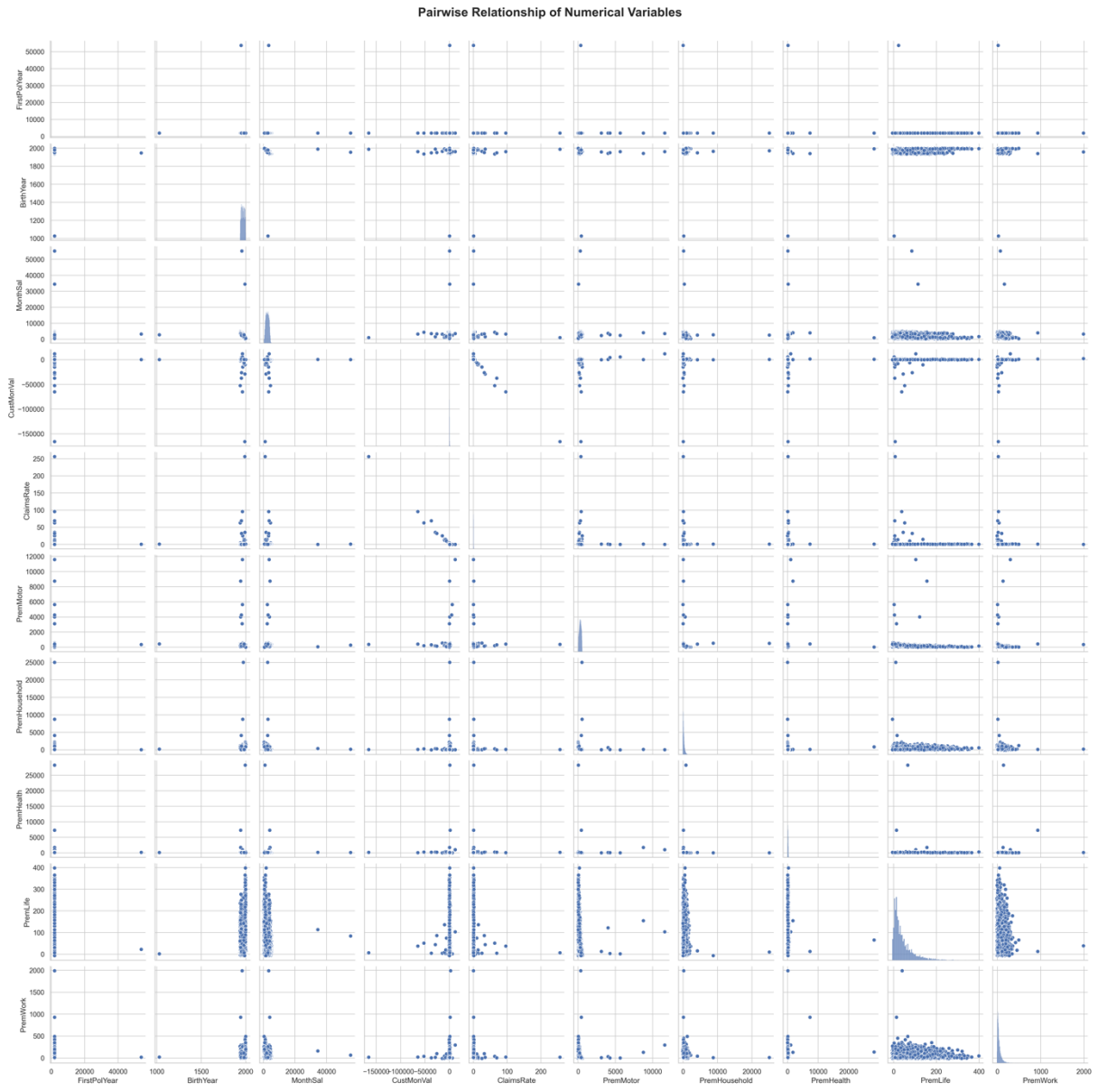

Figure 3 - Pairwise Relationship of the Metric Features

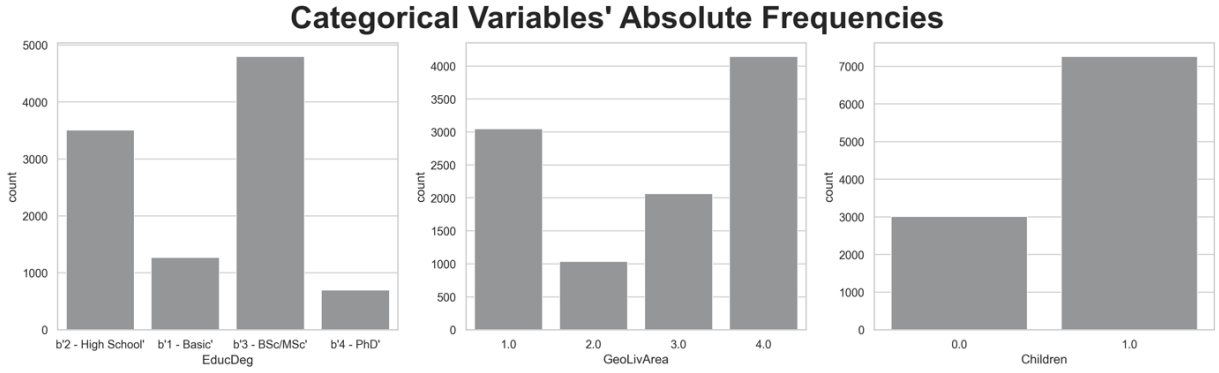

Figure 4 - Categorical Variables' Absolute Frequencies

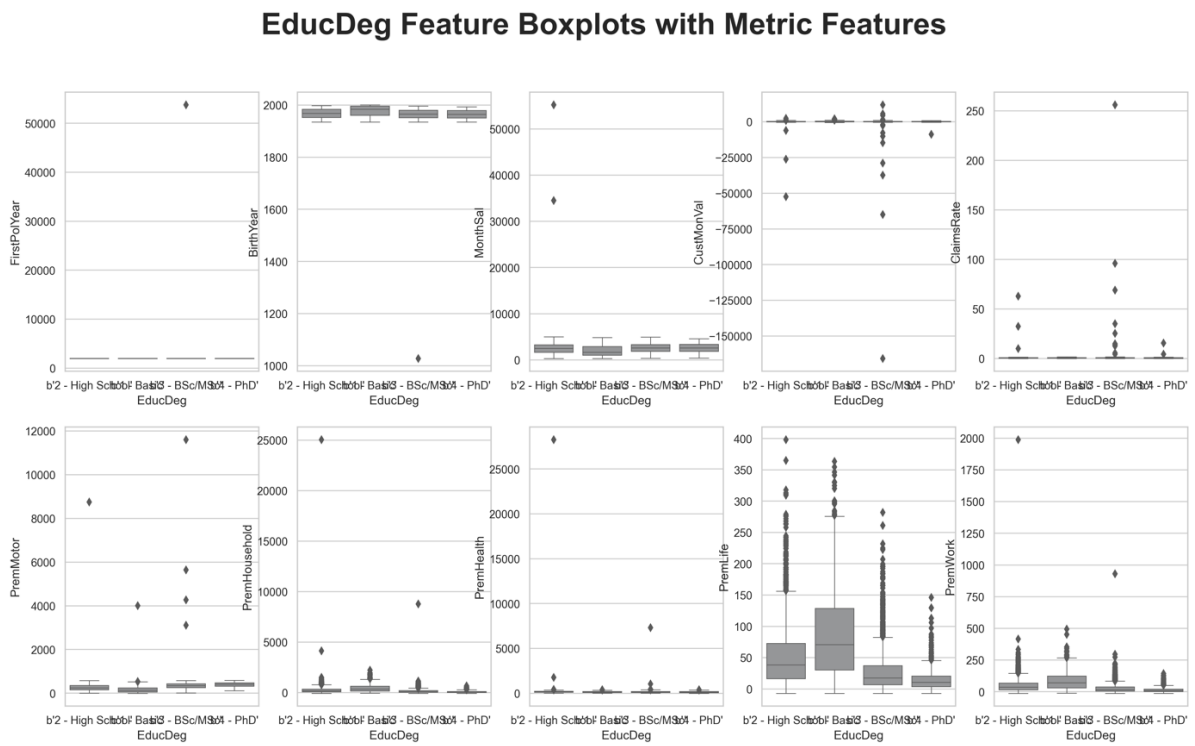

Figure 5 - EducDeg Feature Boxplots with Metric Features

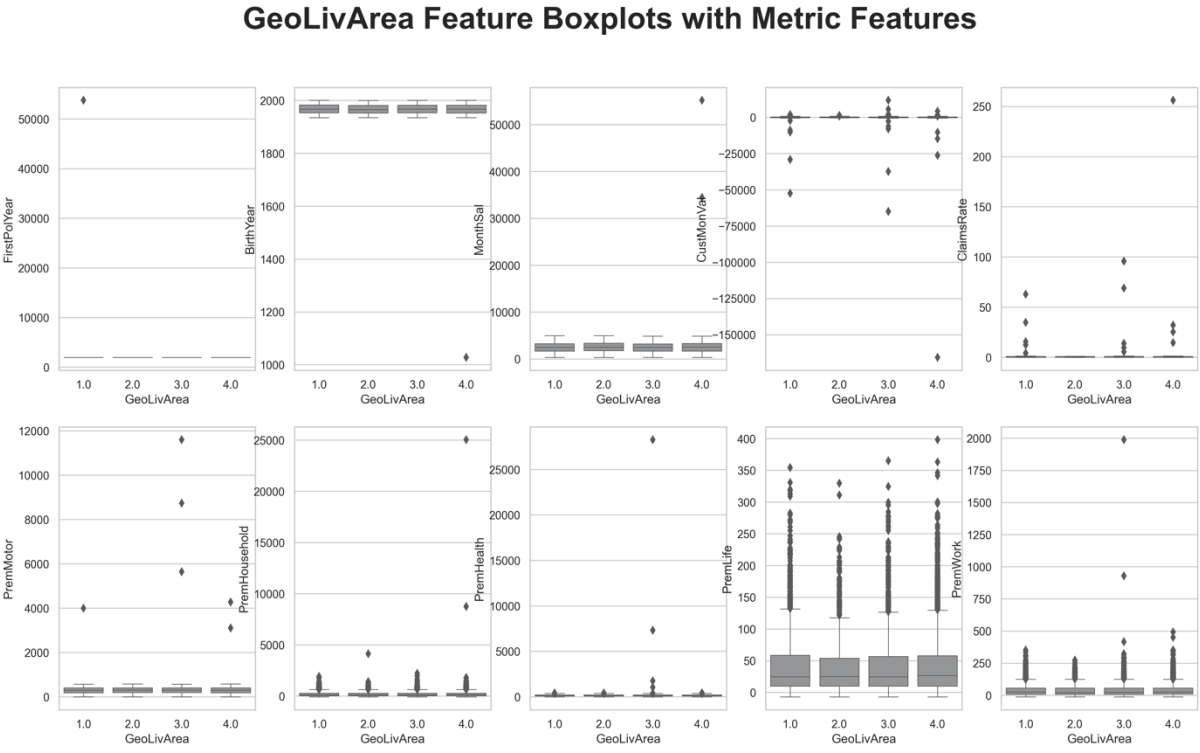

Figure 6 - GeoLivArea Feature Boxplots with Metric Features

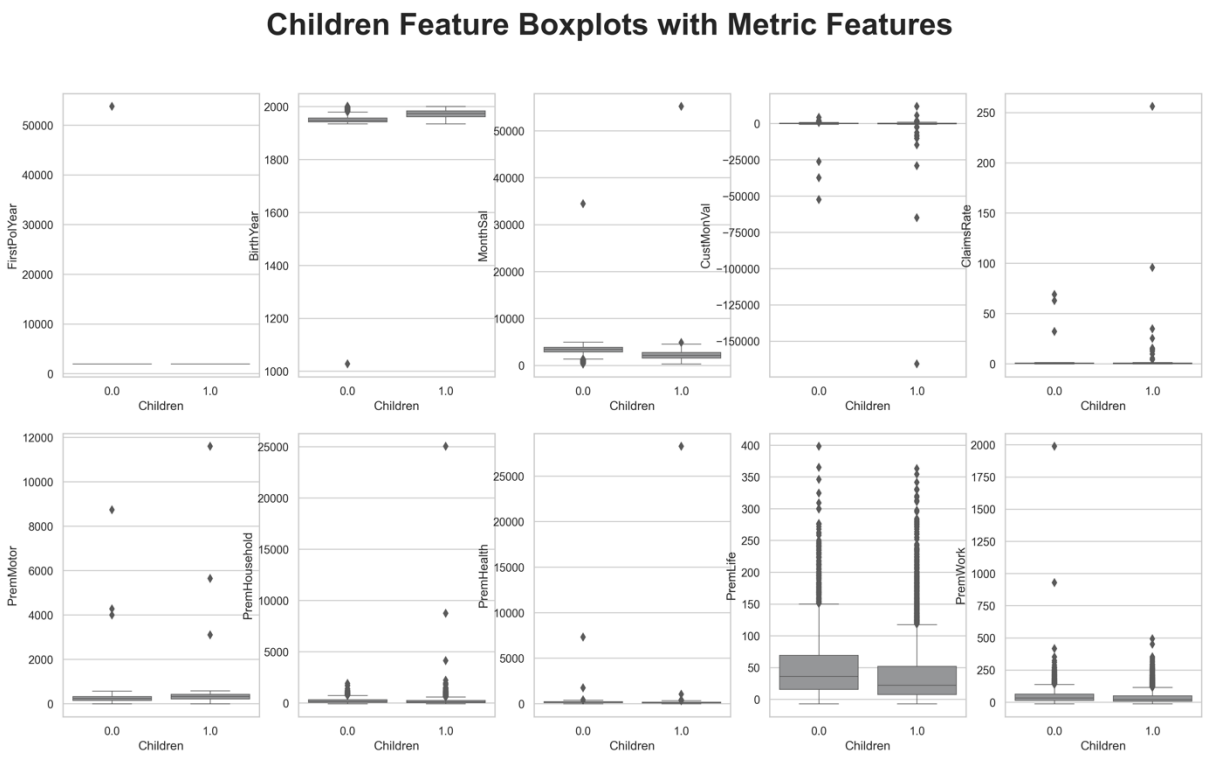

Figure 7 – Children Feature Boxplots with Metric Features

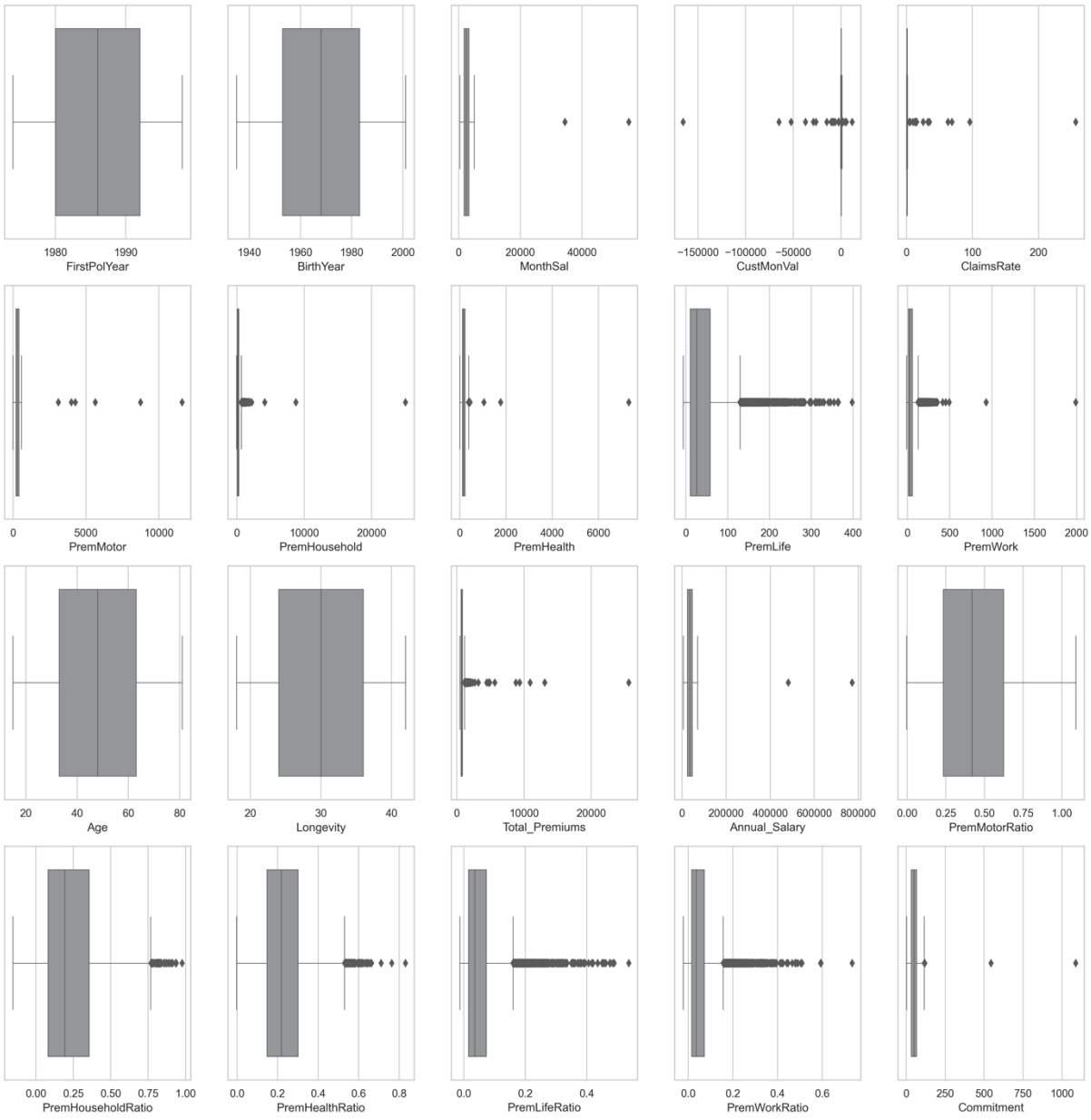

Figure 8 - Metric Features' Boxplots Before Treating the Outliers

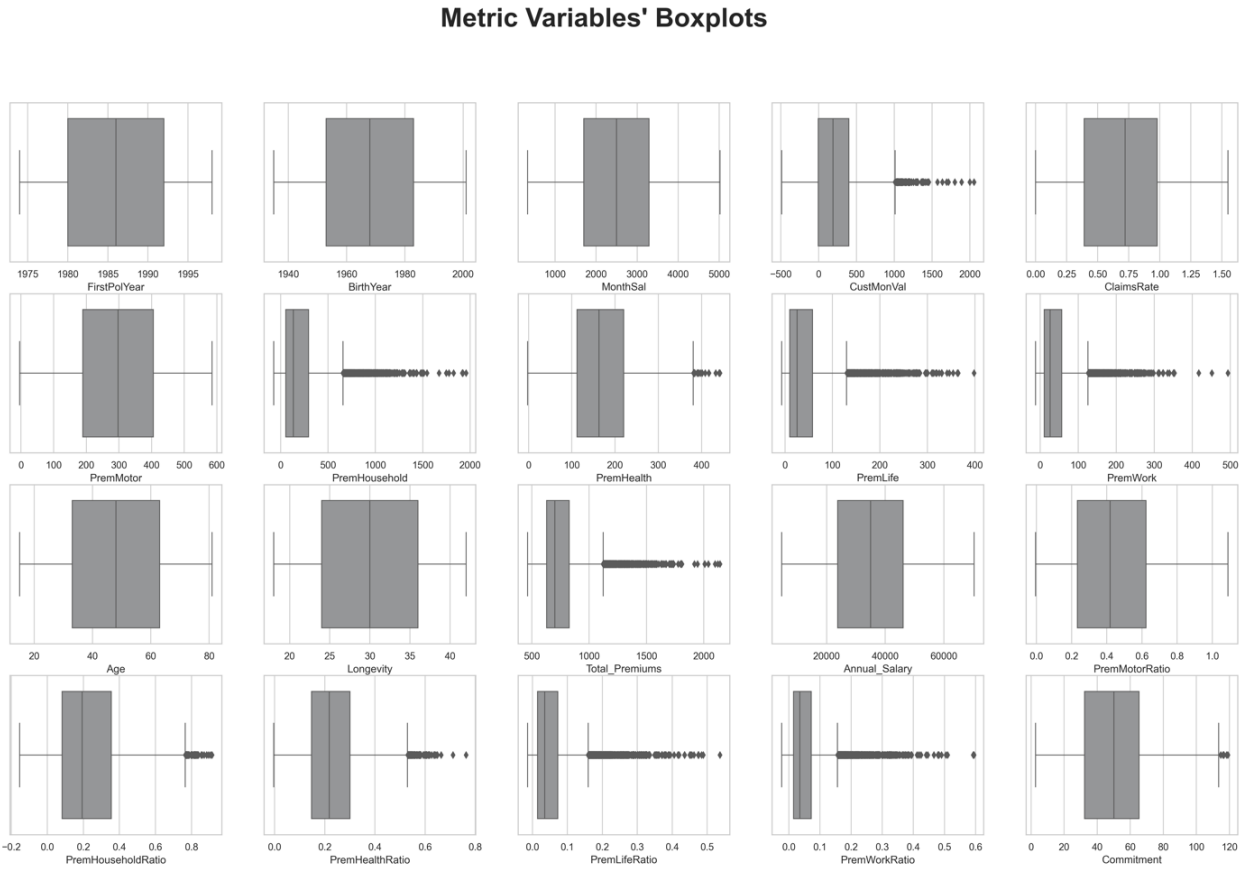

Figure 9 - Metric Features' Boxplots After Treating the Outliers

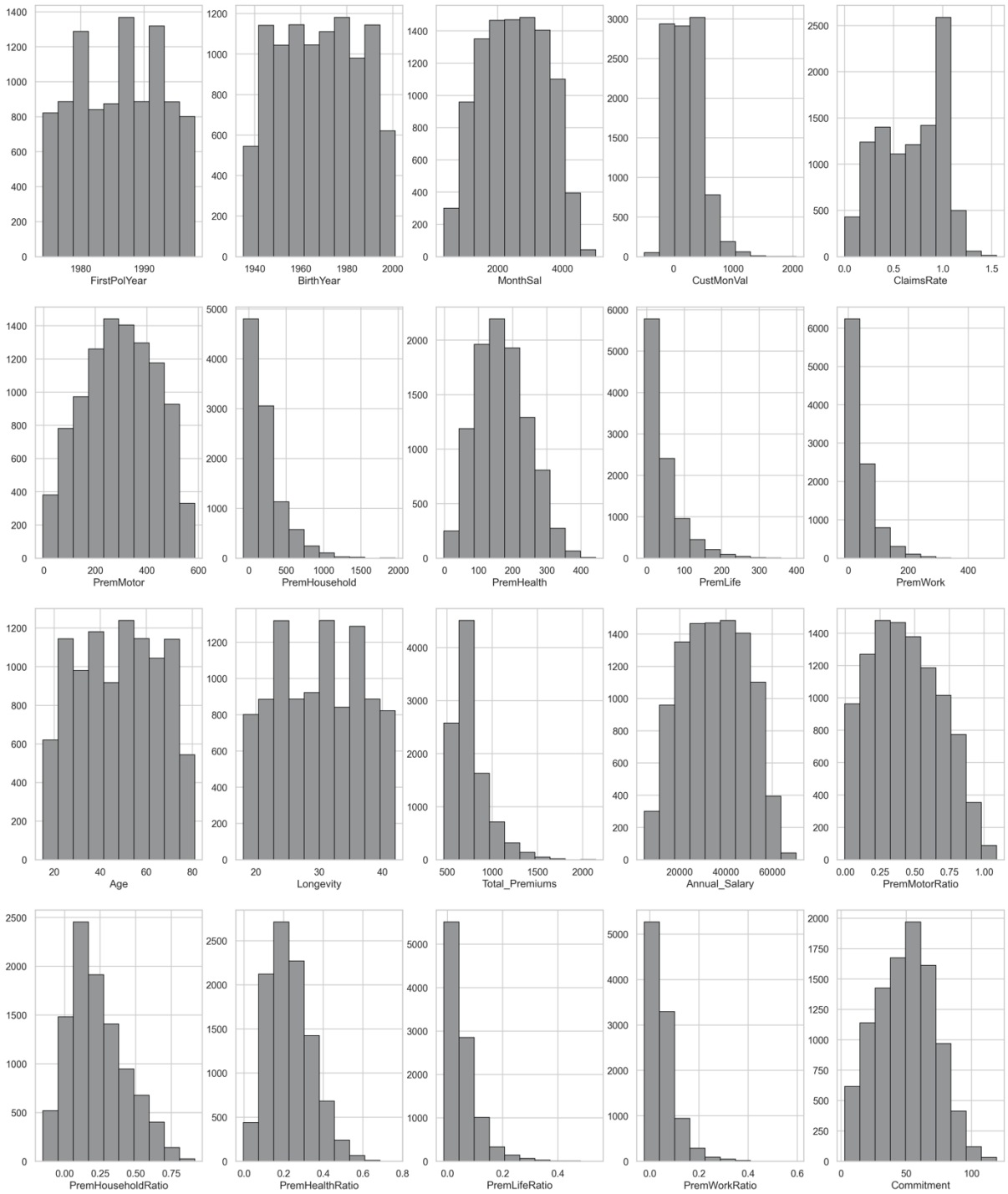

Figure 10 - Metric Features' Histograms after treating the outliers

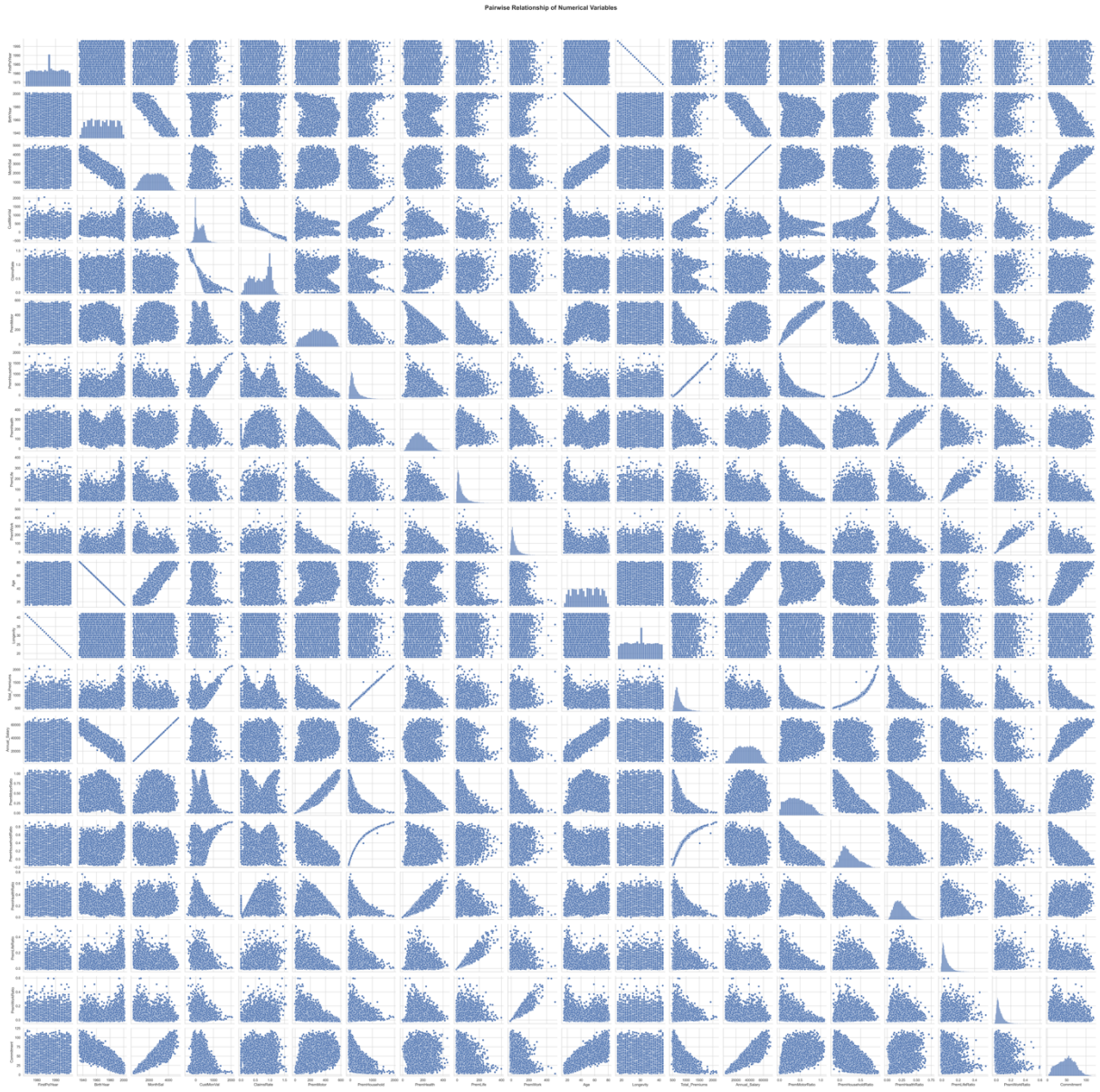

Figure 11 - Pairwise Relationship of the Metric Features

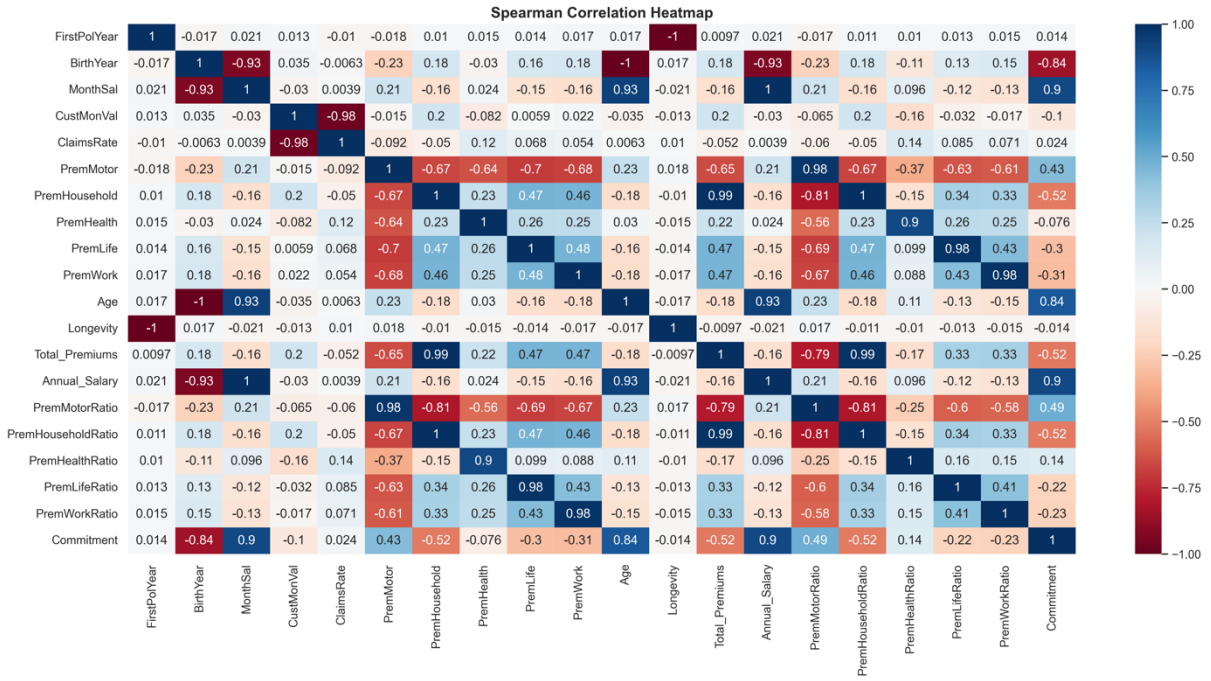

Figure 12 - Spearman Correlation Matrix before Feature Selection

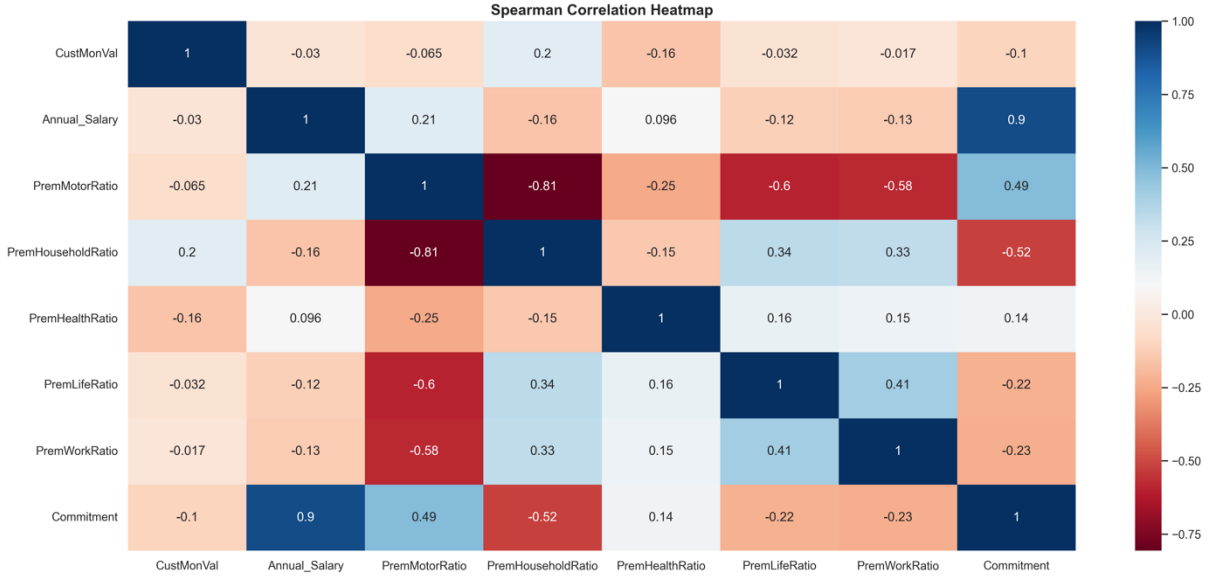

Figure 13 -Spearman Correlation Matrix after Feature Selection

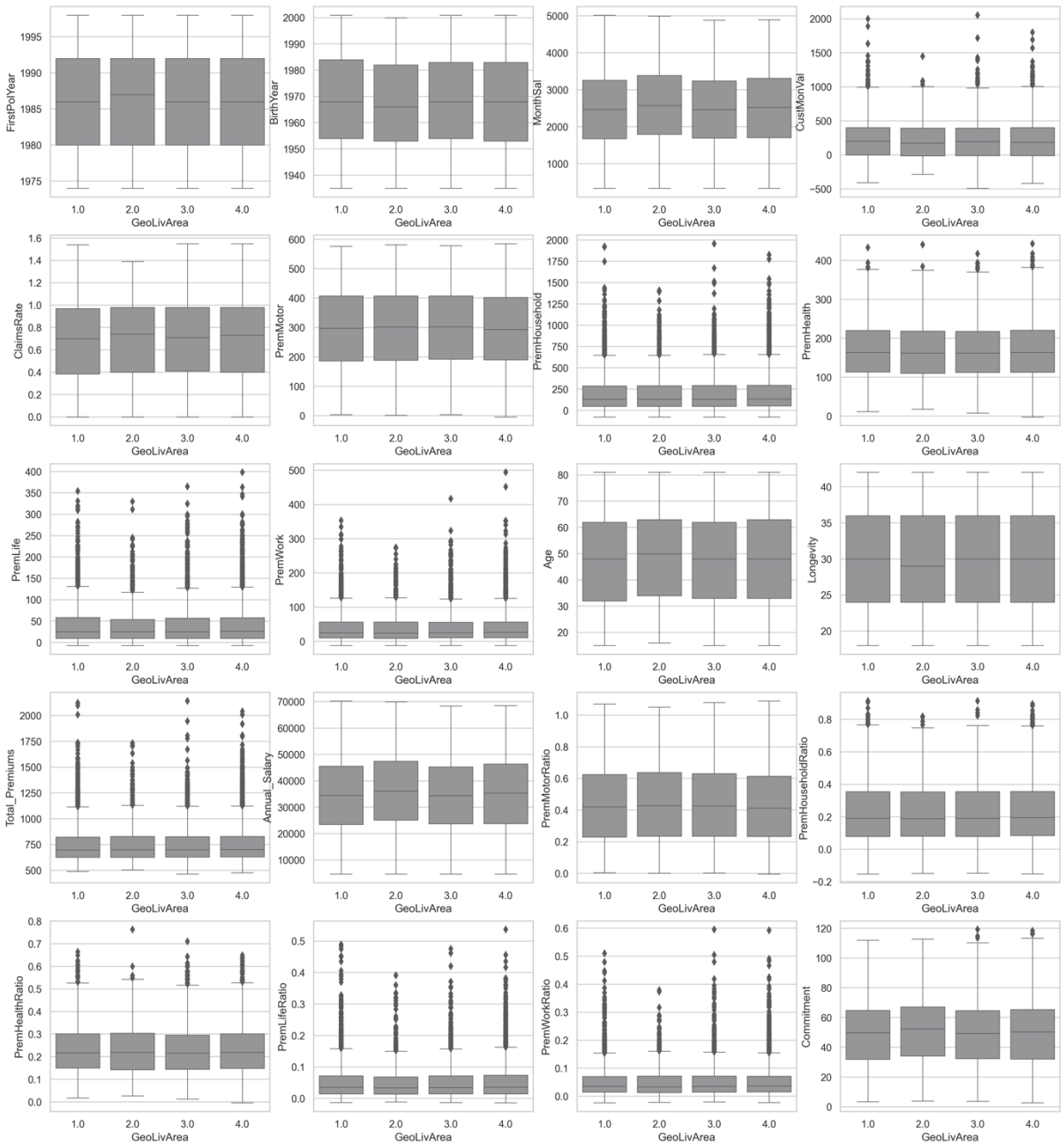

Figure 14 - GeoLivArea Feature Boxplots with the Metric Features

	Eigenvalue	Difference	Proportion	Cumulative
1	9.659355e-02	0.000000	5.610858e-01	0.561086
2	3.206311e-02	-0.064530	1.862460e-01	0.747332
3	1.784422e-02	-0.014219	1.036523e-01	0.850984
4	1.141507e-02	-0.006429	6.630707e-02	0.917291
5	8.411867e-03	-0.003003	4.886226e-02	0.966153
6	5.826851e-03	-0.002585	3.384660e-02	1.000000
7	4.316863e-33	-0.005827	2.507549e-32	1.000000

Figure 15 - Results of the PCA

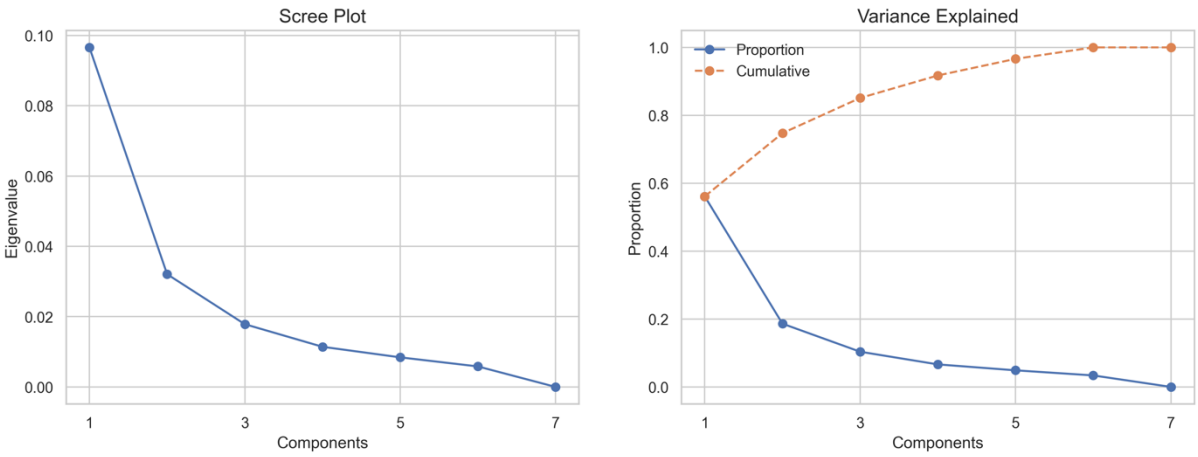

Figure 16 – Scree Plot of the Eigenvalue and Elbow Plot of the Variance Explained

	PC0	PC1	PC2	PC3	PC4
PremMotorRatio	-0.940310	-0.330889	0.076020	-0.018680	-0.004597
PremHouseholdRatio	0.897444	-0.212388	-0.341410	-0.117233	-0.135240
PremHealthRatio	0.029619	0.869409	0.330359	-0.362655	0.041608
PremLifeRatio	0.466430	0.271575	0.122822	0.664312	0.247949
PremWorkRatio	0.450309	0.248847	0.102590	0.513436	0.161796
CustMonVal	0.207432	-0.317207	-0.303937	-0.322364	0.812222
Commitment	-0.717509	0.448385	-0.528869	0.065375	-0.012333

Figure 17 - 5 Principal Components

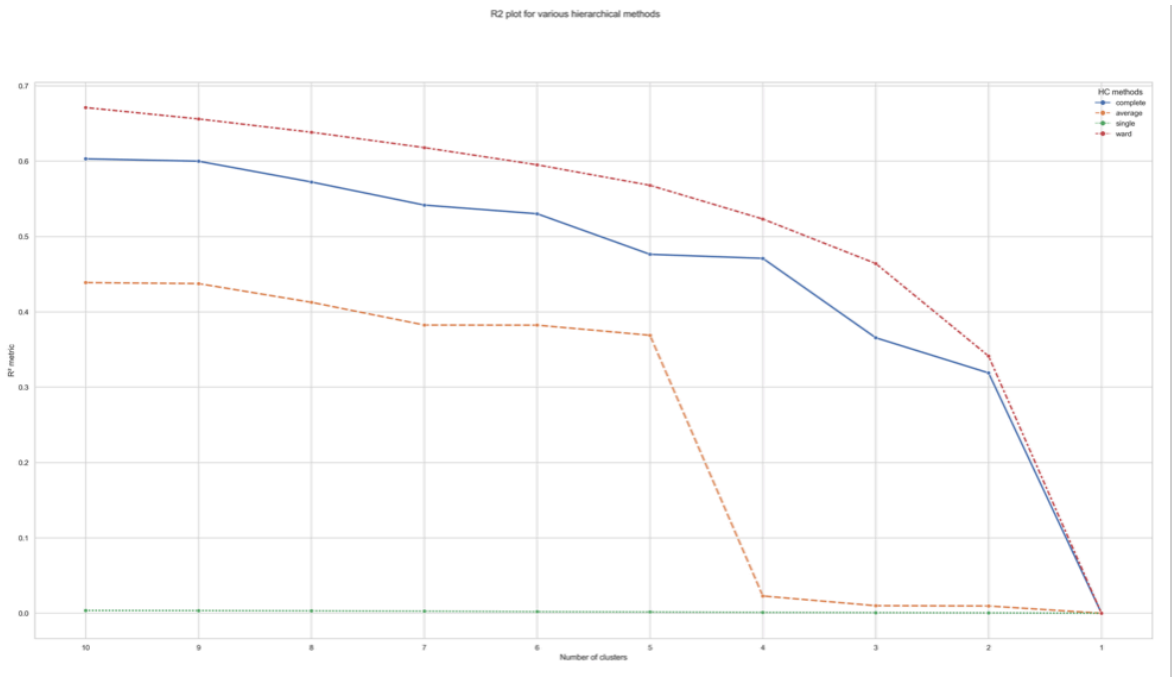

Figure 18 - R^2 Plot for Various Hierarchical Methods

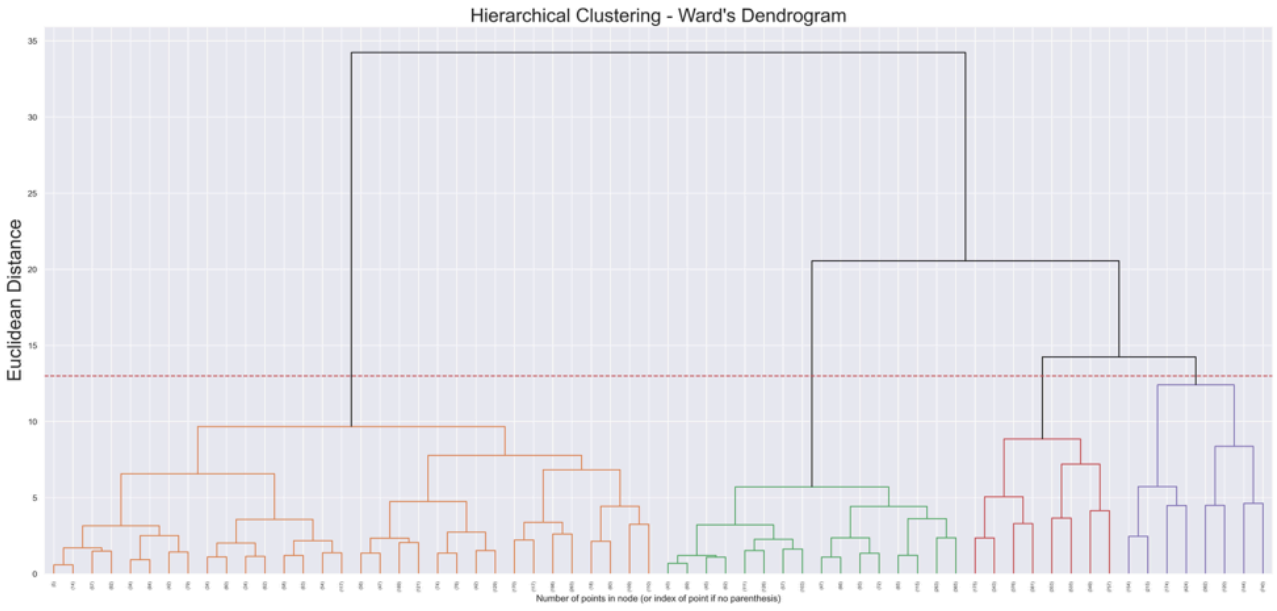

Figure 19- Dendrogram for Ward Linkage

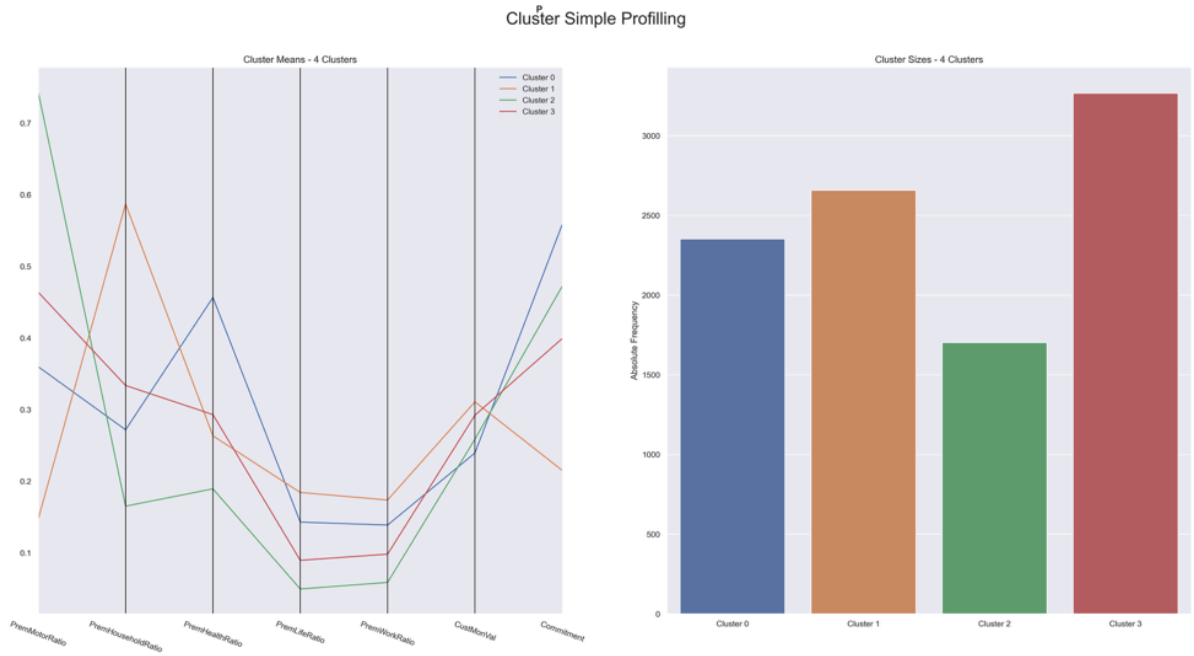

Figure 20- Solution for 4 Clusters with Hierarchical Method

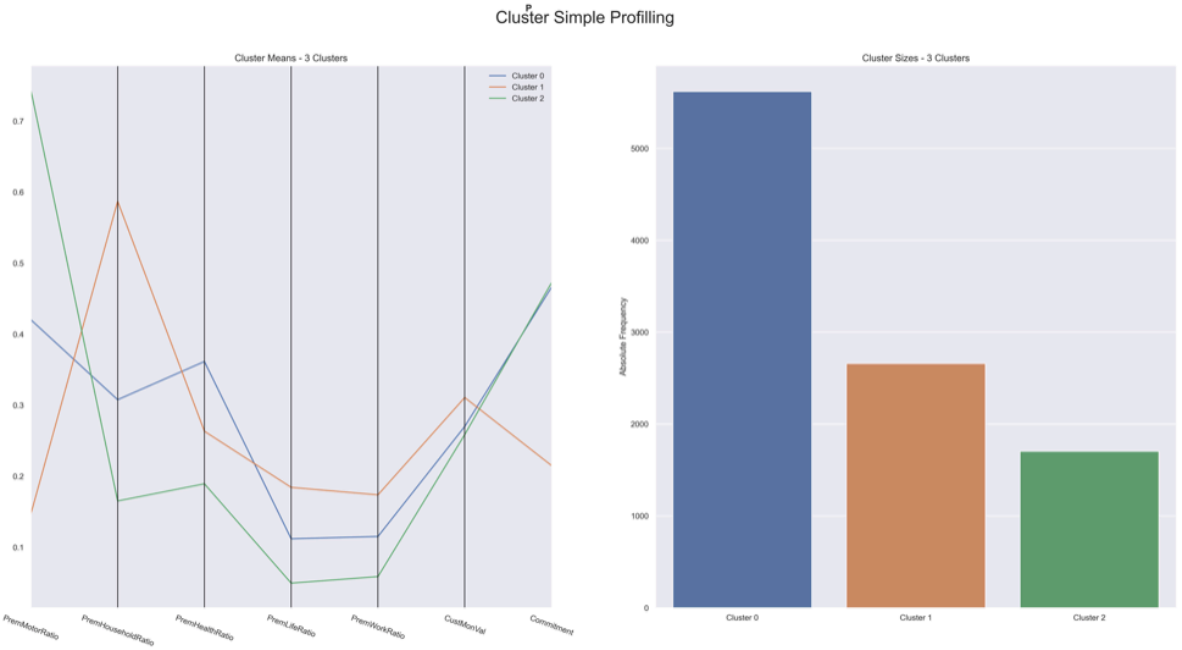

Figure 21- Solution for 3 Clusters with Hierarchical Method

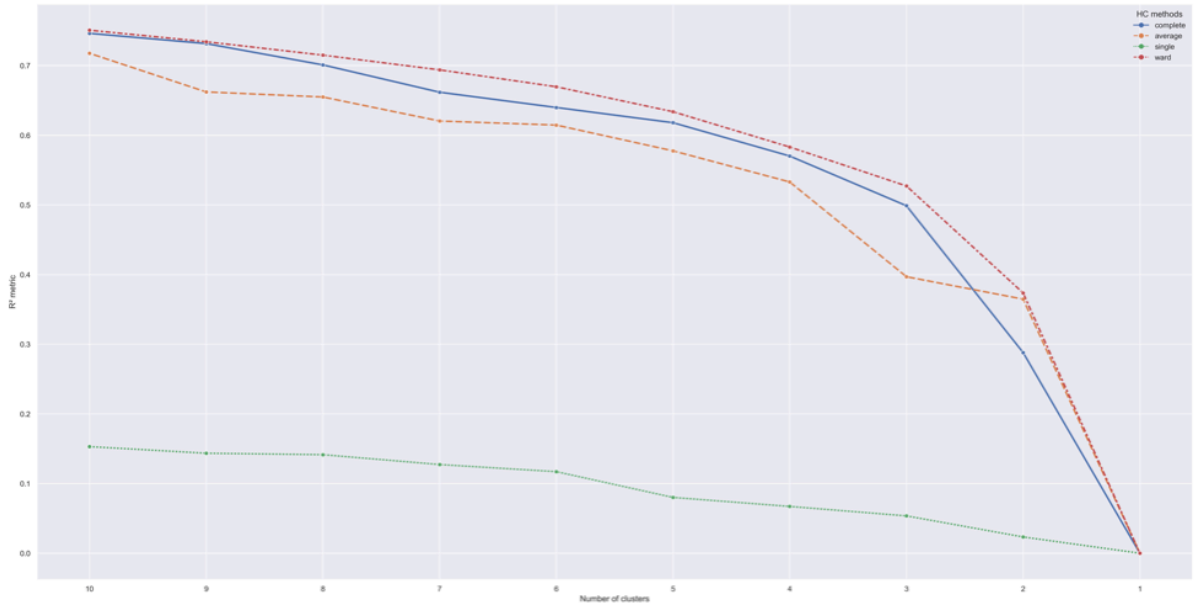

Figure 22- R^2 plot for Various Hierarchical Methods on top of K-Means

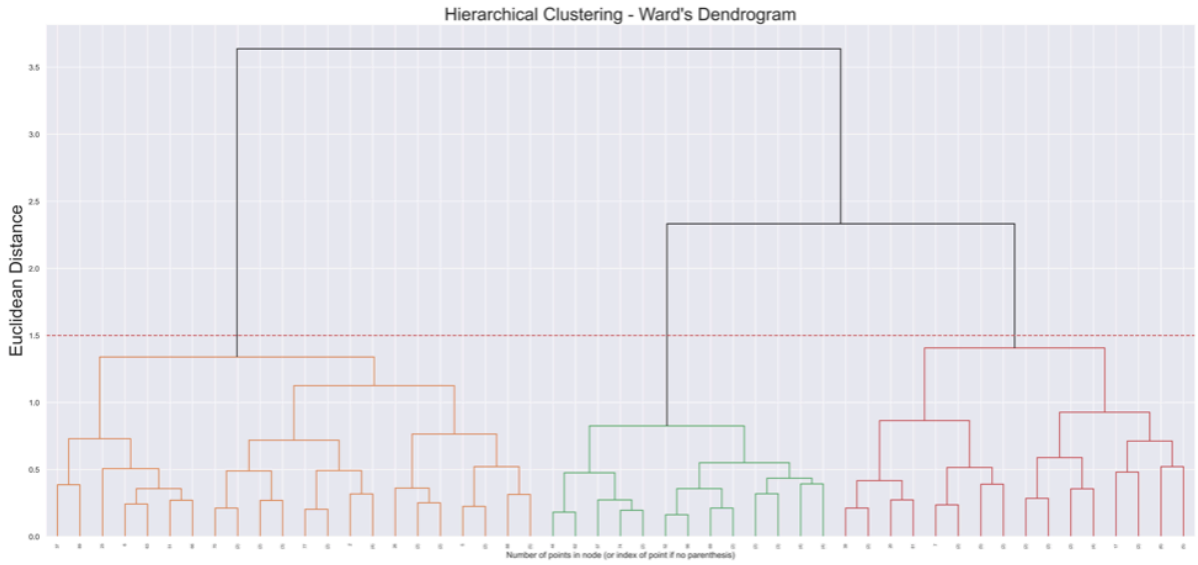

Figure 23- Dendrogram for Ward Linkage on Top K-Means

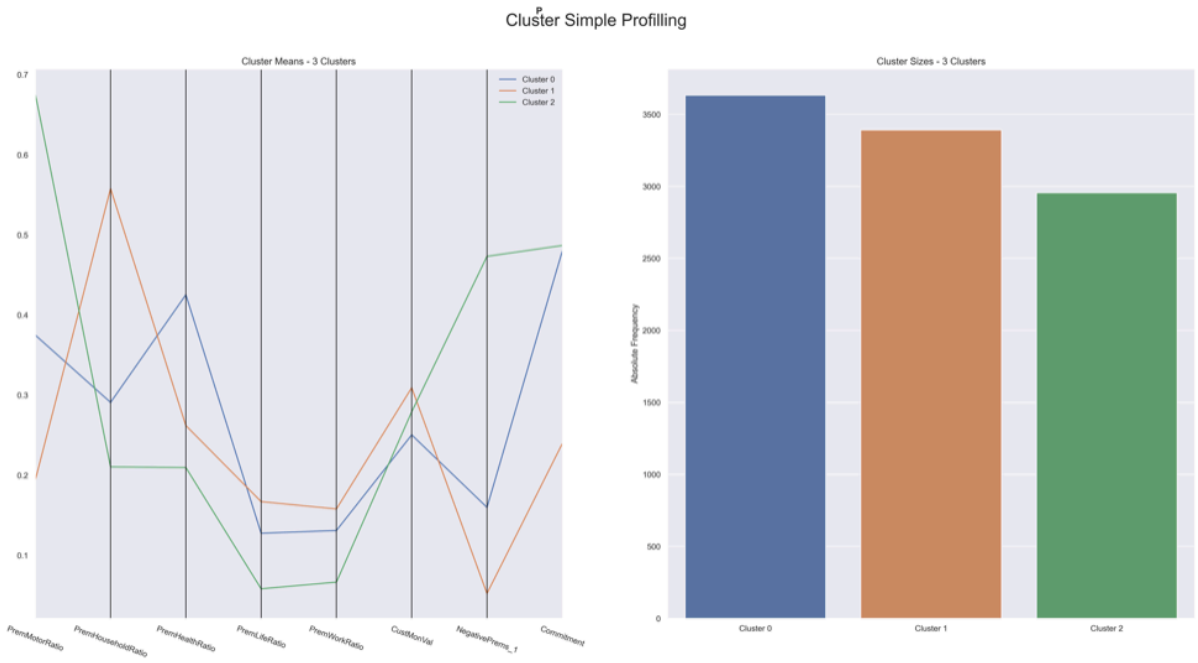

Figure 24- Solution for 3 Clusters with Hierarchical Method on top of K-Means

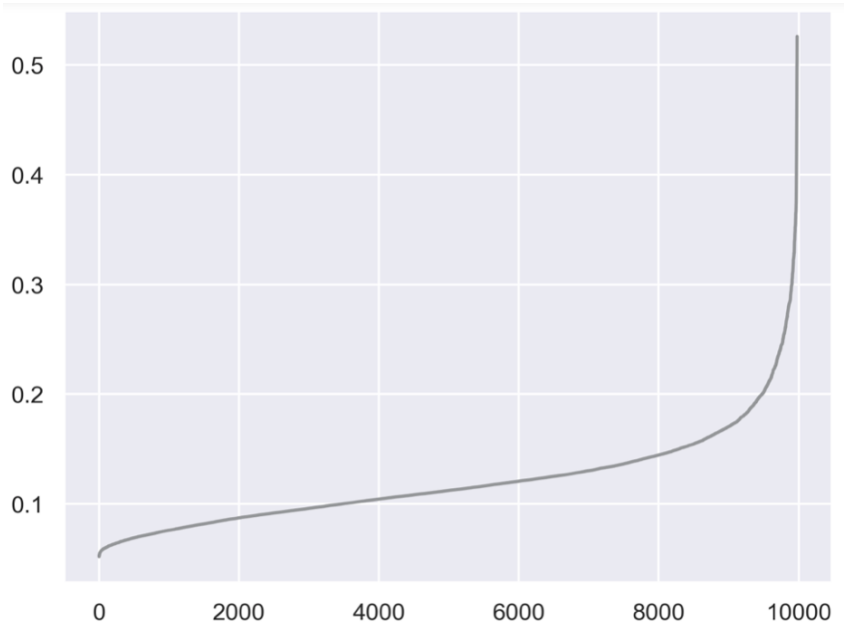

Figure 25- K-distances for DBSCAN

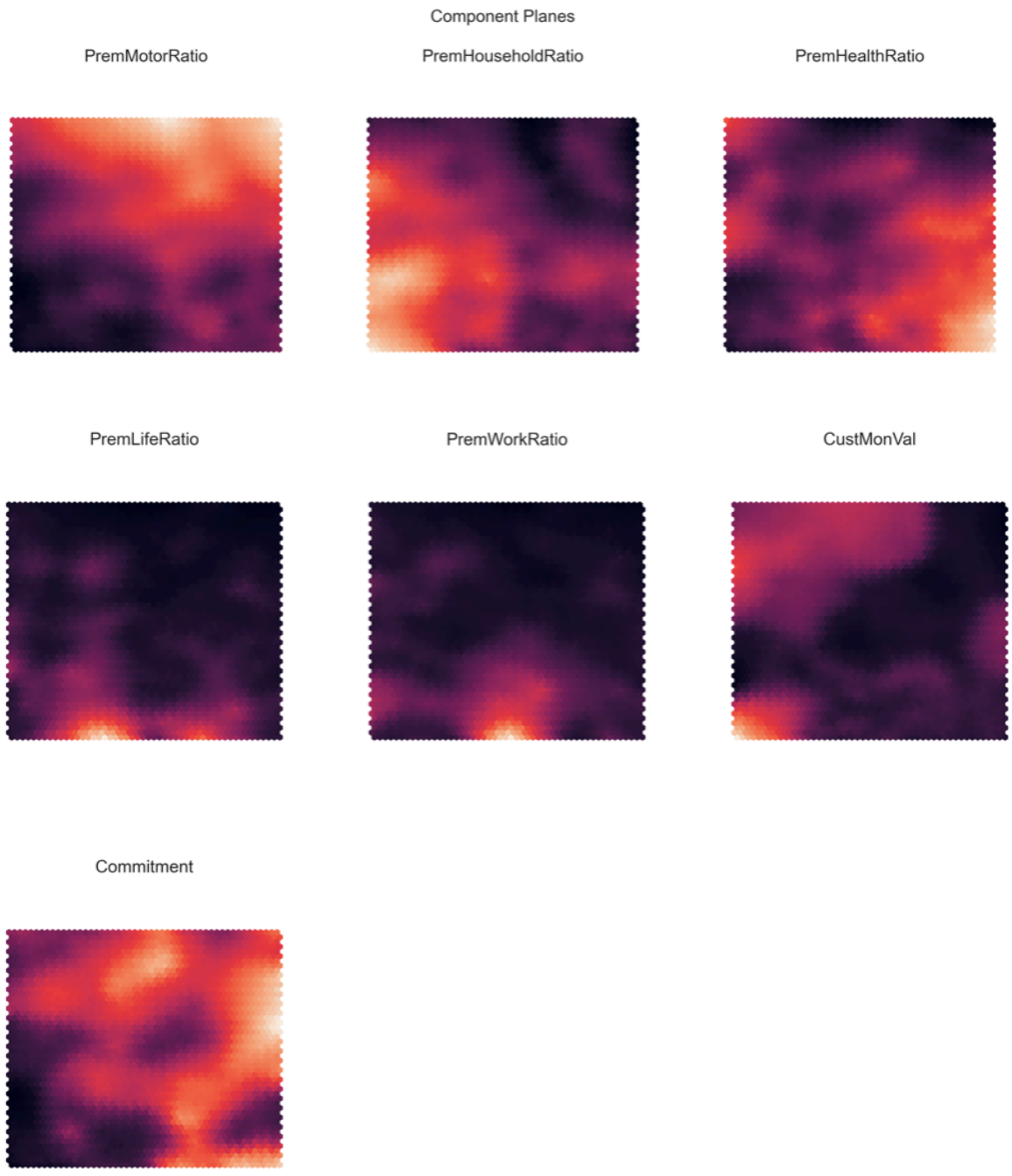

Figure 26- Component Planes

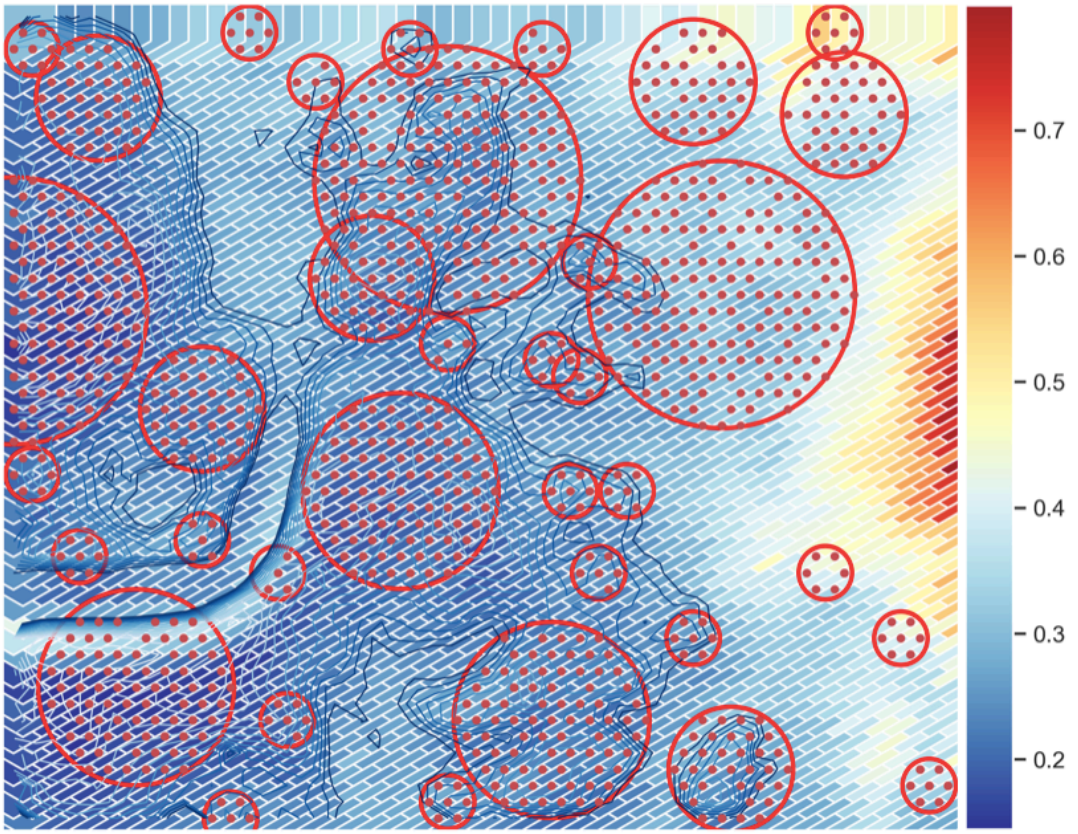

Figure 27- U-Matrix

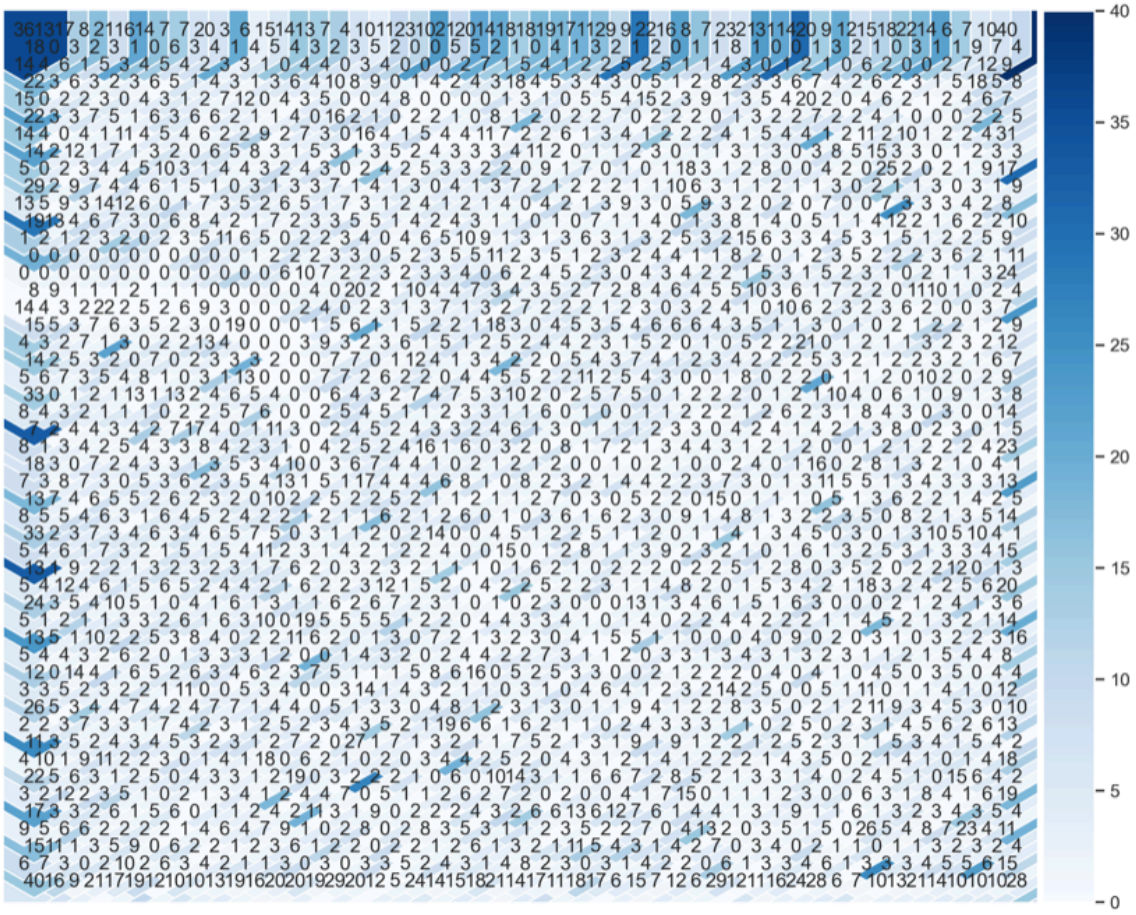

Figure 28- Hitmap

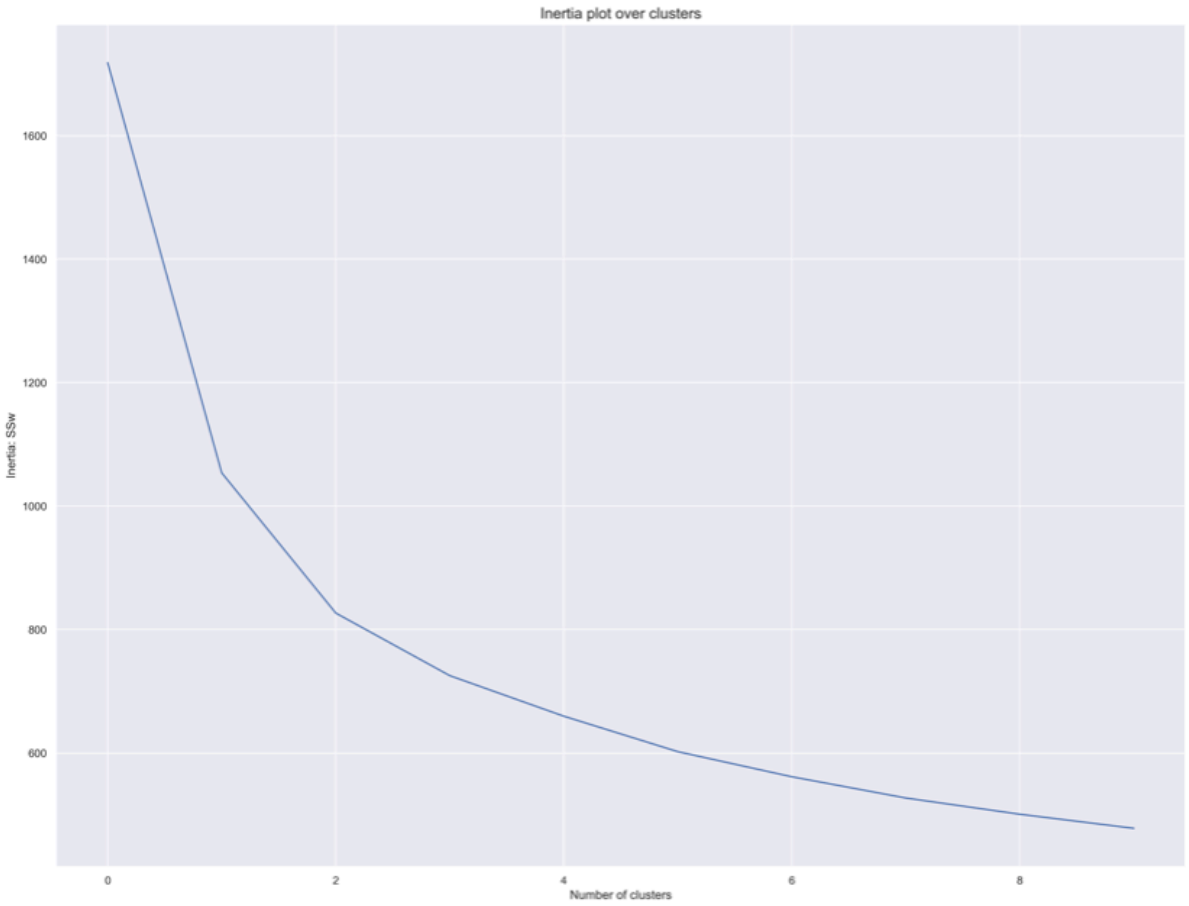

Figure 29- Inertia Plot over Cluster on Top of SOM

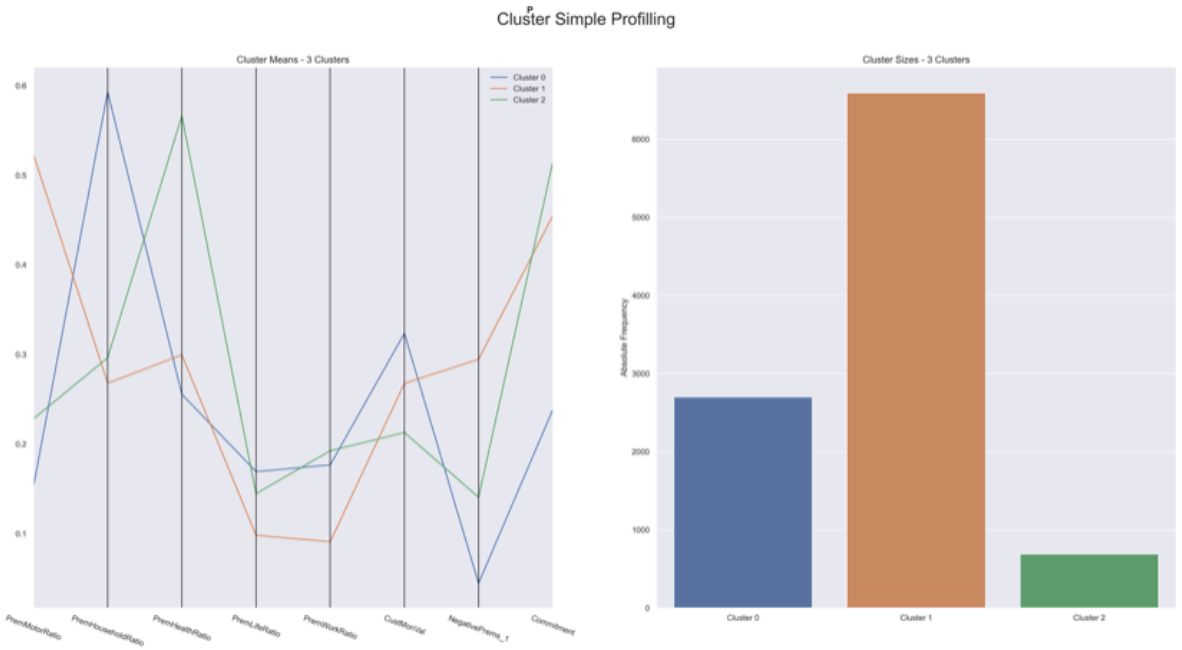

Figure 30- Solution for 3 Clusters with K-Means Method on Top of SOM

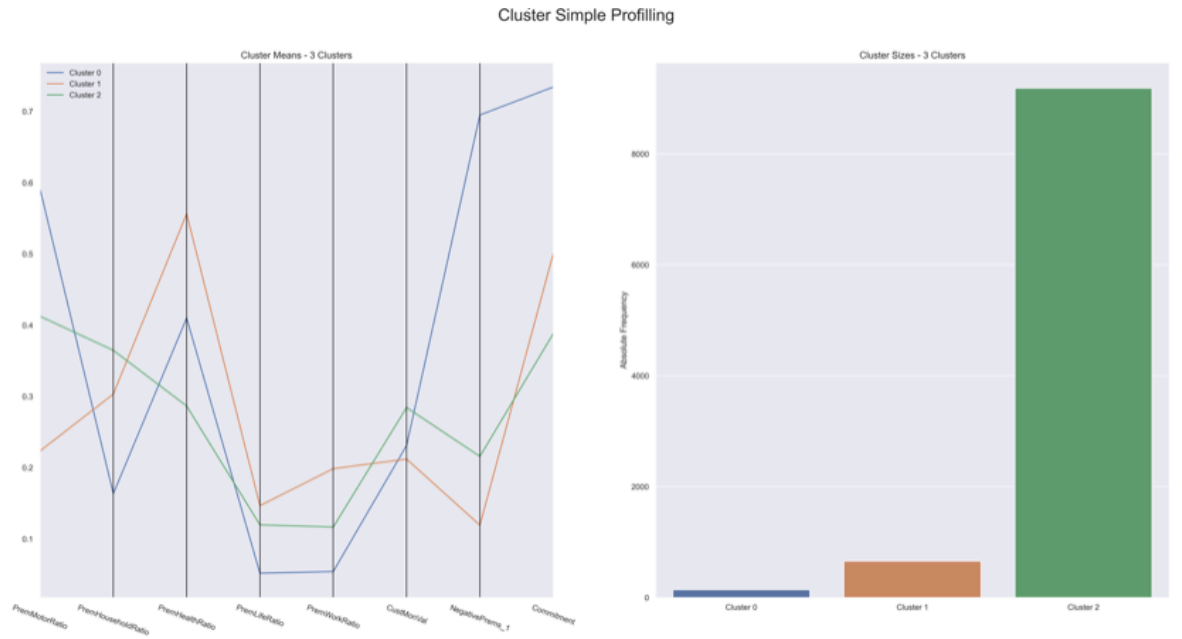

Figure 30- Solution for 3 Clusters with Hierarchical Method on Top of SOM

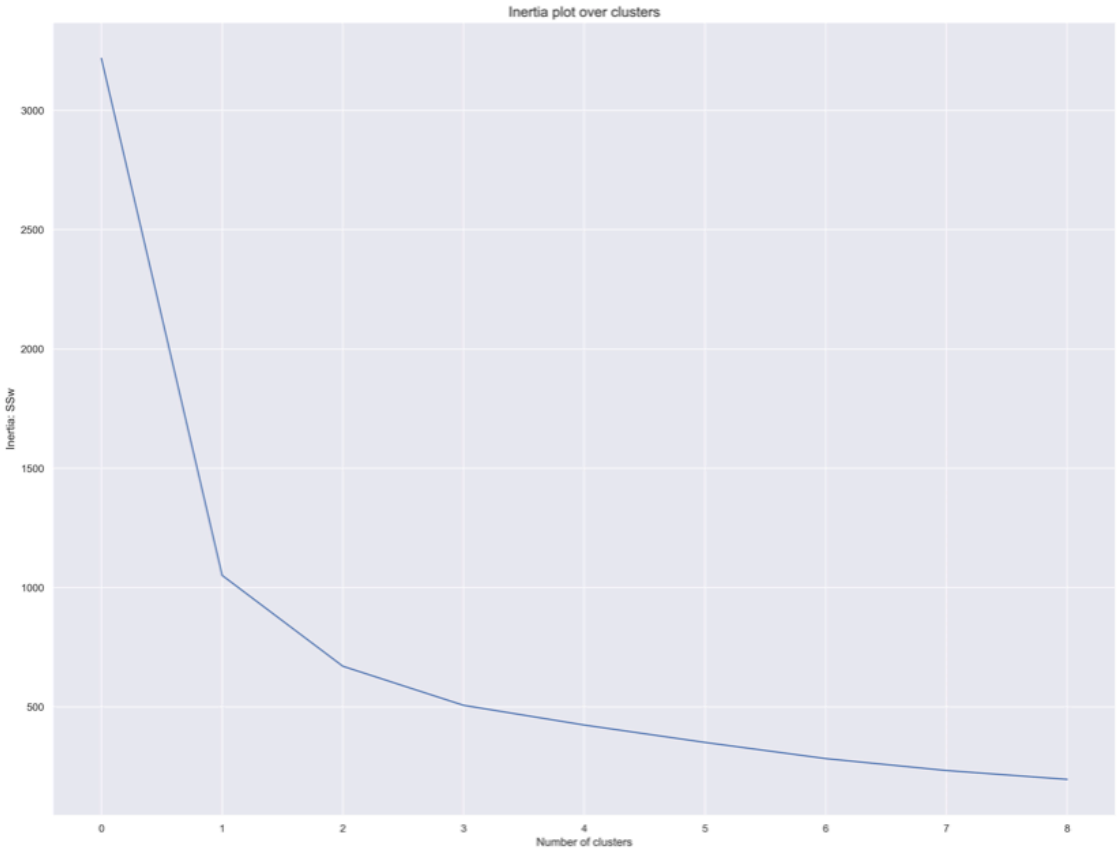

Figure 31- Inertia Plot for K-prototypes

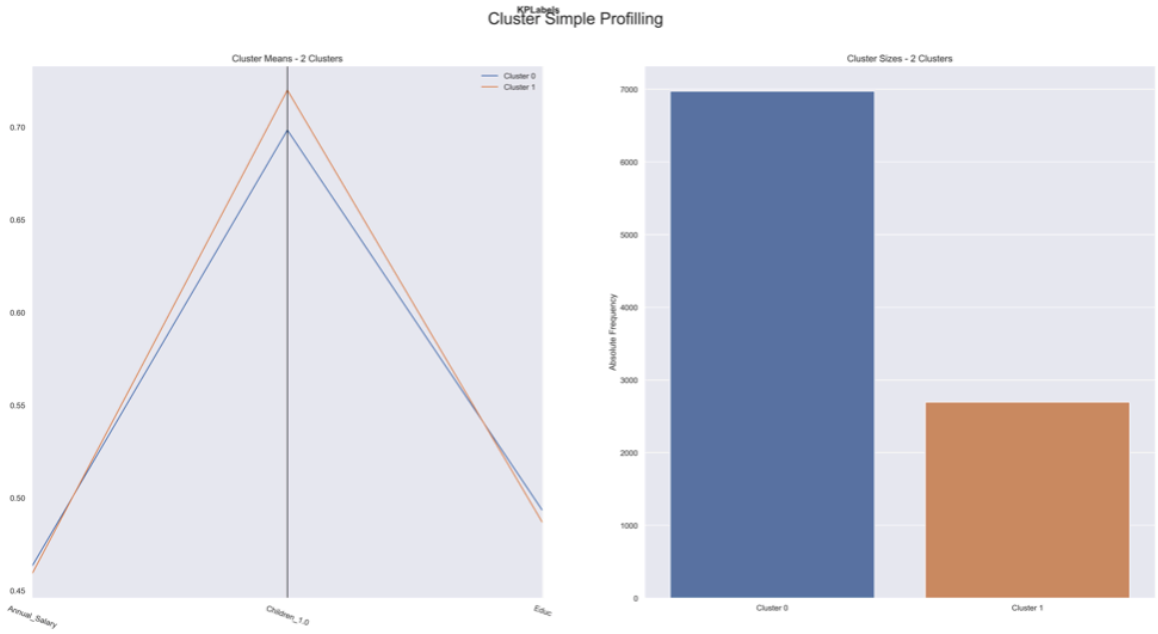

Figure 32- Solution for 2 Clusters with K-Prototypes

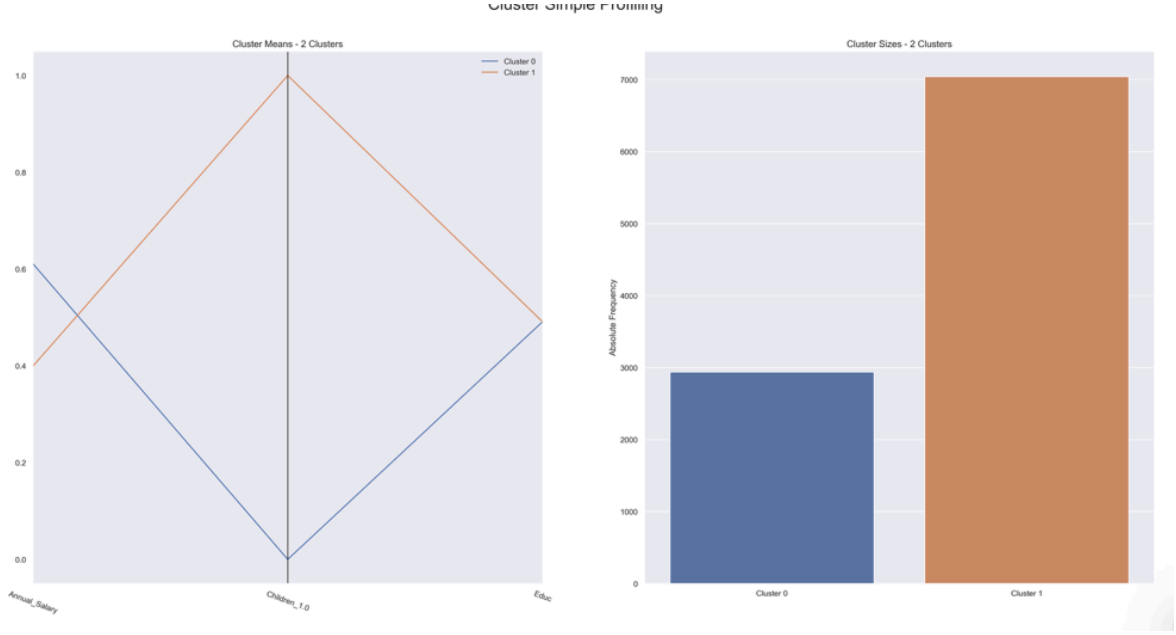

Figure 33- Solution for 2 Clusters with Hierarchical with Gower distance

9.2. Tables

Feature Name	Data Type	Non-null count	Description
CustID	Float64	10296	Unique customer identifier.
FirstPolYear	Float64	10266	Year of the customer's first policy.
BirthYear	Float64	10279	Birth year of the customer.
EducDeg	Object	10279	Academic qualifications of the customer.
MonthSal	Float64	10260	Gross monthly salary in euros.
GeoLivArea	Float64	10295	Living area.
Children	Float64	10275	Binary variable – yes (1) and no (0).
CustMonValue	Float64	10296	Customer monetary variable.
ClaimsRate	Float64	10296	Claims rate of the customer.
PremMotor	Float64	10262	Amount of money spent on Motor premiums in euros.
PremHousehold	Float64	10296	Amount of money spent on Household premiums in euros.
PremLife	Float64	10192	Amount of money spent on Life premiums in euros.
PremWork	Float64	10210	Amount of money spent on Work premiums in euros.
PremHealth	Float64	10253	Amount of money spent on Health premiums in euros.

Table 1 - Initial Original Features

	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10266.0	1991.062634	511.267913	1974.00	1980.00	1986.00	1992.0000	53784.00
BirthYear	10279.0	1968.007783	19.709476	1028.00	1953.00	1968.00	1983.0000	2001.00
MonthSal	10260.0	2506.667057	1157.449634	333.00	1706.00	2501.50	3290.2500	55215.00
GeoLivArea	10295.0	2.709859	1.266291	1.00	1.00	3.00	4.0000	4.00
Children	10275.0	0.706764	0.455268	0.00	0.00	1.00	1.0000	1.00
CustMonVal	10296.0	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	0.742772	2.916964	0.00	0.39	0.72	0.9800	256.20
PremMotor	10262.0	300.470252	211.914997	-4.11	190.59	298.61	408.3000	11604.42
PremHousehold	10296.0	210.431192	352.595984	-75.00	49.45	132.80	290.0500	25048.80
PremHealth	10253.0	171.580833	296.405976	-2.11	111.80	162.81	219.8200	28272.00
PremLife	10192.0	41.855782	47.480632	-7.00	9.89	25.56	57.7900	398.30
PremWork	10210.0	41.277514	51.513572	-12.00	10.67	25.67	56.7900	1988.70

Table 2 - Descriptive Statistics of the Numerical Data

List	Features
metric_features	*FirstPolYear, BirthYear, MonthSal, CustMonValue, ClaimsRate, PremMotor, PremHousehold, PremLife, PremWork* and *PremHealth*
non_metric_features	*EducDeg*, *GeoLivArea* and *Children*

Table 3 – Metric and Non-metric Features Lists

Feature Name	Description	How was it obtained?
Age	Age of the customer.	2016 – Birth_Year
Longevity	For how many years has this customer been a customer.	2016 - FirstPolYear
Total_Premiums	Total value spent in premiums in 2016.	Sum of all premiums
Annual_Salary	Annual salary of the customer, taking into consideration 14 salaries.	MonthSal * 14
PremMotorRatio	Proportion of the Motor premium spent in the total premiums.	PremMotor / TotalPremiums
PremHouseholdRatio	Proportion of the Household premium spent in the total premiums.	PremHousehold / TotalPremiums
PremHealthRatio	Proportion of the Health premium spent in the total premiums.	PremHealth / TotalPremiums
PremLifeRatio	Proportion of the Life premium spent in the total premiums.	PremLife / TotalPremiums
PremWorkRatio	Proportion of the Work premium spent in the total premiums.	PremWork / TotalPremiums
Commitment	Proportion of the annual salary spent in premiums.	Total Premiums/ Annual_Salary
NegativePrems	Binary variable where 1 means that the	np.where((df['PremHouse

	customer has cancelled premiums, and 0 if not.	hold'] < 0) | (df['PremMotor'] < 0) | (df['PremHealth'] < 0) | (df['PremLife'] < 0) | (df['PremWork'] < 0),1,0)

Table 4 - New Features Created

Feature Name	Restriction
FirstPolYear	Must be below or equal to 2016.
Age	Must be below or equal 83 years old
MonthSal	Must be below 7.000
GeoLivArea	Living area.
Children	Binary variable – yes (1) and no (0).
CustMonValue	Must be between -1000 and 7000.
ClaimsRate	Must be below or equal to 100.
PremMotor	Must be below or equal to 2000.
PremHousehold	Must be below or equal to 2000.
PremLife	Must be below or equal to 5000.
PremWork	Must be below or equal to 500.
PremHealth	Must be below or equal to 250.

Table 5 - Restrictions applied on the Manual Filtering Method

Perspective	Metric Features	Non-Metric Features
Product	*PremMotorRatio, PremHouseholdRatio, PremLifeRatio, PremWorkRatio, PremHealthRatio, Commitment and CustMonValue*	*NegativePrems_1*
Demographic	*Annual_Salary*	*EducDeg* and *Children_1.0*

Table 6 – Perspectives Features Lists

Product Demographic	0	1	2
0	1 573	984	381
1	2060	2407	2575

Table 7 – Contingency Table for the Merged Clusters