Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z INSURANCES

Group P

Inês Castro, 2022156

Pedro Pereira, 20220684

Rúben Serpa, 20221284

January 2023

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

INDEX

1. Introduction

This project is focused on helping the marketing department of in a Portuguese long standing insurance company called “A2Z Insurance” that serves many insurance services: Motor, Household, Health, Life and Work Compensation. In order to define customer differentiation strategies and to develop more focused programmes, a company data set was made available containing some 10295 customers

During this project, the data was cleaned and treated in several ways, considering the missing values, duplicates, outliers, and some incoherencies to be beneficial for cluster analysis and customer segmentation. After the feature selection, the data set as suggested was split into a demographic approach and one focused on understanding the value of consumers to the company. In these approaches different segmentation techniques such as K-means in conjunction with hierarchical clustering, K-Prototypes, Self-Organizing Maps (SOM), Mean Shift, DBSCAN, and Gaussian Mixture Model were applied.

Knowing all this, the project will be driven by the objective of understanding which are the relevant customer segments that will be allocated the resources of the marketing department. Given the types of insurance that each consumer cluster will be most interested in buying, marketing strategies will not only be focused on increasing the company's profit through sales promotions but also focused on cross selling opportunities, up selling and retaining current

2. Data Exploration

First of all, it was imported all the needed libraries and the data "a2z_insurance.sas7bdat", provided by the insurance company. Next, it was verified the size of the database and was created a backup of that to maintain the original version unchanged.

When looking for **duplicates**, were revealed 3 duplicated records by the duplicated method. When exploring the available data, the second step was to define *CustID* as index of the data frame and check data types, missing values and duplicates for each variable. Moving on to the **checking of missing values**, after replacing by *NaN* the empty cells, it was possible to conclude that most of the variables had missing values which were filled in later on feature engineering, reducing bias.

Through the analysis of **descriptive statistics** of each variable, Table 1, it was proved that all the variables that needed to be greater than 0 (*FirstPolYear, BirthYear, MonthSal, GeoLivArea, Children, ClaimsRate*) have no negative values. Except that it was found that the variables *PremHouseHold, PremLife, PremWork, PremMotor* and *PremHealth* present as minimum values less than 0 which means that a customer has cancelled an insurance policy. Regarding the variable called *BirthYear* which has a minimum value of 1022, this was considered an input error because of the difference between the minimum and the 25th percentile is considerably high. Furthermore, it was revealed the variable *MonthSal* had a high standard deviation. Moreover, the difference between the maximum and the 75th percentile on the *FirstPolYear, MonthSal and ClaimsRate* variables are noticeable. Concerning the CustMonVal variable, it was proved that it has a high standard deviation and a huge difference between the minimum and the 25th percentile as well as between the maximum and the 75th percentile. For the remaining numerical variables, the minimum, maximum, mean and standard deviation seem to be in accordance. For the categorical variables, it is possible to verify the correspondent number of classes. Therefore, a deeper analysis of outliers is needed.

Moving on to the **data visualization**, the dataset was divided into metric and non-metric features since they require different visualization methods. Further the metric features are: *FirstPolYear, BirthYear, MonthSal, CustMonVal, ClaimsRate, PremMotor, PremHousehold, PremHealth, PremLife and PremWork.*

Conductive to explore **metric features** visually, histograms and Boxplots were plotted on Figure 1, to check distributions and to identify possible **outliers**. By analysing them, it was verified that almost all features present outliers.

Afterwards, the relationships between features were checked through heat map of a **correlation matrix** (Figure 2). According to the matrix, it is possible to claim that there is just two strong correlations between *CustMonVal* and *ClaimsRate*, and also *BirthYear* and *MonthSal*, whereby these results may be influenced by missing values and outliers.

In addition, bar charts were plotted, Figure 3 in annex, for each **non-metric feature** *(EducDeg, GeoLivArea, Children*) and none presented any inconsistency. More than that it could be verified that the majority of the consumers have children, have completed secondary school or even a bachelor's/master's degree.

3. Pre-processing

When pre-processing the available data, we started by copying the data into a new data frame called "df_processed" to keep the original version unchanged. Before treating missing values, all duplicates were removed (3 rows in total).

3.1. Treating Missing values and outliers

Regarding the **missing values**, as it was 389 missing values present in the dataset in consequence it was necessary to adopt different approaches that would be suited for each case. Firstly, it was filled all *NaN* values from *Prem* columns with 0, assuming the customers had cancelled insurances. To handle with the 29 lines with more than one missing value, it was decided to eliminate them given their low representativeness in the total data frame.

For the *NaN* values present on the metric features, it was decided to use the *KNN* algorithm to fill them. It is important to note that this method can basically predict values to fill the NaN of one line based on the values of the similar neighbor. Taking into consideration that this algorithm has the limitation on non-metric features, they were filled with the aid of the mode method, which merely fills the missing values with the most frequent value in a specific variable.

Next, just like missing values, were considered approaches to handle **outliers**, but before applying any approach, the graphs of the variables were analysed in detail, in order to verify their distributions and understand which records represented outliers, through Box-Plots and Histograms. Following this, the outliers were divided into global outliers being considered as absurd and contextual outliers being possible to occur but eccentric.

Using automatic limitation, one of the cases considered as an outlier is having a *FirstPolYear* greater than 2016, because the year of the customer's first policy cannot be higher than the current year of the database (2016). Other assumptions considered as outliers were an individual be born before 1930 and having a salary lower than 0, as this is considered untypical. With this approach, 0,02% of the observations would be removed, which is sustainable for the continuation of the study, but it was decided to check how the DBSCAN could contribute.

The **DBSCAN** method, or Density-based spatial clustering of applications with noise, was applied in order to find the radius, and 103 clusters were found. It is important to mention that this density-based clustering technique is better explained in the chapter named *Modelling*.
The percentage of data saved after removing outliers with the two approaches is 0.9775. After treating the outliers, the data was distributed as illustrated in Figure 4.

3.2. Coherence Checks

In this subchapter, it was tested the existence of unreal observations in our dataset. For that reason, restrictions have been placed in order to recognize incoherent observations.

Concerning the variable Age, it was assumed that underage clients could not have an insurance, so it was identified 0 minor clients. Taking into account that there are many data where the first year as a customer (*FirstPolYear*) is less than the *BirthYear*, it was assumed that these variables had been improperly swapped. The values of the rows where this happens have been swapped consequently.

3.3. Feature Engineering

According to Patel 2021, feature engineering is a machine learning technique that leverages data to create new variables that are not in the training set with the goal of simplifying and speeding up data transformations.

So, in order to promote useful and relevant information, the variable *Birth_Year* was transformed into *Age, FirtstPolYear into time_as_customer and MonthSal* into *YearSal*. Moreover, conforming to Patrick 2019, feature transformations can also include aggregating or combining attributes to create new features, depending on the problem at hand but averages, sums and ratios over different groupings can better expose trends to a model. In this sense, multiple features were extracted, such as *TotalPremiums* (sum of each premium), *ProfitsRate* (rate of profits under total premiums) and *CustClassification* (number of premiums active for the client). Additionally, it was computed mapping for the feature *EducDeg* and one hot enconding for the feature *GeoLivArea*, since it's not an ordinal feature. On a final note, the ratio of clients with no premiums (*ChurnRate*) was analysed, and since the percentage was low (0.12%), all the inactive clients were removed from the dataset.

3.4. Feature Selection

Feature selection is a technique with the goal to reduce the input space dimensionality. Meaning, the relevant features are selected, and redundant features are removed. Accordingly, the model complexity decreases, and the model's generalization ability is enhanced, which results in improved performance and better understanding of the problem. This selection will be useful for deciding which are the relevant variables and which can be excluded from the segmentation.

In the first place, the **Kendall Correlation Coefficient** (Figure 5) was employed for the continuous, discrete and ordinal features, to evaluate the relationship between the variables. After an analysis the variable Age is kept with *MonthSal* because both are important for the context; I*nsuredRate* and *CustMonVal* are eliminated instead of *ProfitsRate* because it is better for the final analysis and has better correlation with other characteristics; *TotalPremiums* is eliminated because the variables Prems are also ratios and the *CustClassification* is eliminated because it has no correlation with other characteristics. Also, the Education Degree and Children features appeared to be relevant to the dataset so were kept.

For the **non-metric features**, in order to verify their distributions with all the other variables and understand which records represented, Boxplots were plotted. From the analysis of Figure 6, Figure 7 and Figure 8 it was possible to verify that apart from the *Children* and *EducDeg* variables, the *GeoLivArea* variable does not appear to be relevant since the customers' characteristics do not seem to be influenced by the area in where they live. With this more complete segmentation it will be possible to better define marketing strategies.

Before Modelling, the feature scaling was applied considering that the variables *Age* and *MonthSal* was at different scales from the remaining selected features. It was considered important to standardize the data on equal scales in order to avoid major problems and to make it easier to understand how far the values are from the mean. The **MinMaxScaler** was used as a standardization method. This method was applied to data versions where outliers were removed. So, in observations for these datasets, all values remained to *Age* and *MonthSal* which are non-binary numeric, or rates were now between 0 and 1. It is important to mention that the Standard Scaler was not used because it was not sure that the distributions are normal.

4. Modelling

Regarding modelling, multiple unsupervised and semi-supervised techniques were applied to the dataset to extract unpredictable segments of customers. In order to perform a more selecting approach, two different perspectives were created and modelled apart.

The demographic perspective, consisting in the features *Age, EducDeg, YearSal*, *Children* that characterise consumers demographically. On the other side, the consumption perspective, containing the features *PremMotorRate, PremHouseHoldRate, PremHealthRate, PremLifeRate, PremWorkRate, ProfitsRate* represents the customers' behaviour. With this approach, different perspectives can be segmented separately, allowing the models to detect patterns that otherwise would be harder.

4.1. Demographic Features Clustering

Concerning the consumption perspective, it represented a challenge when it comes to the use of categorical variables. Most algorithms are equipped with the capacity of only working with metric features. So, as a workaround, Hierarchical and K-Means clustering were applied to the metric features (Age and YearSal). In order to consider a mixed data type algorithm, K-Prototypes was also tested.

4.1.1. Hierarchical clustering

This technique is defined as an unsupervised learning algorithm which is based on clustering data upon hierarchical ordering. In this sense, it's developed the hierarchy of clusters in the form of a tree known as the dendrogram. In the scope of the project, 4 different linkage methods (single, complete, average and ward linkage) were evaluated, by analysing the R-squared plot. By checking the Figure 9 it's possible to observe that only applying the hierarchical clustering does not give a good solution, however it was chosen the ward linkage to plot the dendrogram (Figure 10). The dendrogram was constructed based on these specifications and the number of clusters chosen with this method was 4.

Hierarchical clustering is easy to understand and construct, which are good characteristics, however it rarely provides the best solution because it involves lots of arbitrary decisions. Also, its main output, the dendrogram, is commonly misinterpreted, therefore there are better alternatives for clustering, such as k-means.

4.1.2. K-means

Classified as one of simplest and best unsupervised models, this tool starts by placing random K points in the dataset, and iteratively, adjusts them until convergence is found, by computing and assigning the mean. One of the main disadvantages is the fact that is necessary to specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. To mitigate this problem, multiple techniques can be use, such as Inertia and Silhouette plots and a hybrid hierarchical k-means model.

From the Inertia and silhouette score plots (Figure 11), 2 to 4 clusters seemed to provide a somewhat good solution but, to complement the decision, a dendrogram over 35 k-means computed centroids (Figure 12) confirmed the choice of 3 clusters.

4.1.3. K-Prototypes

K-Prototypes can be defined as a hybrid model that complements K-Means with K-Modes, an algorithm popular for handling only categorical variables. To find the best number of clusters, k, to use in the method it was computed the cost plot (Figure 13), concluding 4 as the number of clusters.

4.1.4. Assessment

Even though in the comparison R-squared plot (Figure 14) the highest score is represented by k-means, K-Prototypes was the chosen clustering technique for this perspective, being the only model supporting both metric and non-metric features.

4.2. Consumption Perspective

Regarding the consumption perspective, since it only consists in numerical variables, the models' options to use are much wider than the demographic ones. Therefore, Hierarchical, K-means, Self-Organizing Maps, Density Based and Gaussian Mixture algorithms are tested.

4.2.1. Hierarchical Clustering

Just like the hierarchical procedure performed to the demographic perspective, firstly, a multi-line plot with the R-squared scores for each linkage and each number of clusters from 2 to 20 was analysed, with the purpose to find the optimal overall linkage. From the Figure 15, the best linkage shown is ward. In order to find the best number of clusters, both the elbow in the multi-line plot and a dendrogram of hierarchical clustering with the ward linkage (Figure 16) were inspected, concluding that the numbers of clusters to choose was 5.

4.2.2. K-means

For the K-means, the most important aspect is to know the number of clusters to choose. To come up with a value, three types of analyses were made. The Inertia plot aims to help find the best trade-off between number of cluster and the sum of squared errors (SSE). From trying to minimize the SSE with the lest numbers of clusters possible, the elbow is extracted. From both the Inertia and the Silhouette Score plots (Figure 17), it was concluded that 5 clusters would provide a good solution. However, to complement that analysis, also a dendrogram of a hierarchical procedure applied to 35 k-means calculated' centroids was explored (Figure 18). The optimal number of clusters chosen was 5.

4.2.3. Self-Organizing Map

Self-Organizing Map (SOM) is a type of artificial neural network which is also inspired by biological models of neural systems. It follows an unsupervised learning approach, training its network through a competitive learning algorithm. Until the number of epochs is reached, the nodes of a m by n map are continuously updated to closely match the dataset points. Originally SOM was constructed to map multi-dimensional data onto lower-dimensional which allowing data experts to reduce complex problems for easy interpretation.

Notwithstanding, it can also be used to provide clusters by applying unsupervised models to the trained network. From that, to map the clusters back to the original dataset, the cluster from the closest node to each record (best matching unit) is assigned to it.

To proceed with the SOM analysis, it was needed to decide on the map size and nodes' lattice. From trying different solutions, a network of 50 by 50 nodes and hexagonal lattice provided the best solution, with a final quantization error of 0.61.

From the components planes (Figure 19) and U-Matrix (Figure 20) it's possible to conclude that the data can't be well differentiated under the SOM network.

Applying the procedures from previous Hierarchical and K-means models to the network, the solution from Hierarchical (Figure 21) with 5 clusters under SOM demonstrated a better cluster cohesion than K-Means (Figure 22), so its best matching units' clusters were mapped back to a copy of the consumption data frame.

4.2.4. Density-based Clustering

Density-based algorithms are a popular form of clustering that identify distinctive clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other clusters by sparse regions. Although they can provide very well-defined clusters, if the data does not have the conditions to cluster by density, they will not perform as good as other algorithms.

4.2.5. Mean Shift

Being classified as a Density-based algorithm, Mean Shift does not need be provided with a strict number of clusters, but a bandwidth to which is used as the radius of sliding windows. To form a cluster, the mean of the data points in the specified bandwidth of a random point is calculated, shifting the central point to that mean consecutively until the density has reached convergency. That procedure is applied until all data points belong to sliding windows.

For the project, to facilitate the choice of bandwidth, *estimate_bandwidth* function was used with a quantile of 0.12. From that bandwidth, the mean shift algorithm was applied, predicting 3 clusters.

4.2.6. DBSCAN

Just like Mean Shift, DBSCAN does not need to be provided with the number of clusters. Instead, from a radius and a number of minimum samples, the algorithm chooses an arbitrary point and checks if there is the minimum number of points within the space of the radius. When the condition is met, a cluster is defined with all the points inside the radius, having, continuously, the points checked again with the previous condition and added to the cluster in the case of being true. When the condition is not met, the points are labelled as noise (label -1). The process is made until all the points are verified.

By applying these steps, DBSCAN is able to find high density regions and separate them from low density ones.

In order to decide on the radius and minimum samples. a K-distance graph is constructed, plotting the sorted graph of the distance of the 15th neighbour for all the points in the dataset, as shown in the Figure 23. However, many values were tested for the radius and minimum sample and the model did not provide, in any case, a good solution.

4.2.7. Gaussian Mixture Process

Gaussian mixture models consist of probabilistic models that learn independent subpopulations automatically without knowing the data points. To model with this algorithm, there are two key parameters that need to be chosen: number of clusters and covariance type. Since the features opted

to model are not independent, there is a high probability of the model to perform badly. In that sense, there was not a high effort to hyperparameter tune, only testing multiple numbers of clusters on a full covariance type.

From the Figure 24, it's possible to visualize the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) over a range of 20 clusters. Aiming to choose the solution with lower BIC and least clusters, 10 clusters were opted

4.2.8. Assessment

From all the models r2 scores (Figure 25), the one that performed better under an acceptable number of clusters is K-Means, presenting a score of 0.73.

As only the r2 score is not enough to decide upon the model to choose the different labels from each tested model were visualized with Principal Components, UMAP and T-SNE dimensions. From the Figure 26, Figure 27 and Figure 28, the one that present better cohesion amongst all visualizations is also K-Means.

With that, K-Means is the chosen model to cluster the consumption features.

4.3. Merging Perspectives

In order to obtain the final clusters, both the perspectives must be merged together. To implement the merging process, an agglomerative clustering model was applied to each perspective's centroids, obtaining the dendrogram in Figure 29. Choosing 4 as the number of clusters, the final R-squared score was of 0.68.

4.4. Reclassifying Data

In order to assess the feature importance among the clustering process, a classification tree was modelled with the merged labels. This application able the data experts to analyse which are the most relevant and redundant features. Being a very simple and intuitive model, decision trees can also provide a visualization of each partition of the decisions. There are other models to apply semi-supervised classification, however, this was the chosen classifier due to the simplicity of application and interpretation. Obtaining an accuracy of 96.18% on 20% test data, the importance of the features can be visualized in Table 2, concluding that *PremMotorRatio, Age* and *ProfitsRate* are responsible for 98.2% of the splitting decisions.

5. Marketing Strategy

With the final clusters, the mean (Table 3), mode (Table 4) and median (Table 5) were applied to each label. After careful analysis, a marketing strategy was defined to each of the segments, with the goal of gathering more clients and more sales.

5.1. Sales Promotion

This segment is composed by the largest quantity of customers who has the minor purchasing power (1545€ per month on average) and with the lower education level. These customers are characterised by buying the most of household, life and work insurances.

Taking into account that it is becoming increasingly financially profitable to invest in electric/hybrid cars given the tax incentives and benefits as well as the importance of sustainability, creating a sales promotion strategy for an electric motor insurance package could be profitable. In this sense, end-of-year advertisements would be published appealing for safe and environmentally friendly driving during the festive seasons passing the message that two months' motor insurance would be offered by the company for new customers.

Further, to ensure the company is competitive on price, a package with home and health insurance could be designed to make it more attractive to consumers than buying them separately. This would be advertised through email marketing, social media and on the website of the company in order to effectively reach this generation.

5.2. Cross-selling

The second cluster with a monthly salary significantly higher than the average (3493€) is composed of retired consumers without children. The insurances most invested in by this group are mainly motor, house and health. It is important to note that they are the ones that invest the most in health insurance among the other clusters.

Considering the high purchasing power and the age of consumers, it would be relevant to develop a cross-selling strategy by offering life insurance, as a complementary good, for one year to clients who already invest in home insurance. This way the customer will be able to prevent, on an economic level, the consequences of death once shared by transferring the responsibilities to the insurance company.

5.3. Customer Retention

This is the cluster with the smallest number of customers, but they represent the majority of the company's profit. Although it is the cluster with the 2nd lowest monthly salary and has dependents, it is characterised for being the one that invests the most in motor insurance.

Given that this group represents the highest profit margin, most marketing efforts should focus on them. Therefore, the marketing plan should focus mainly on loyalty strategies such as: if you have been a customer for 1 year you get a 50% discount on health insurance and with gifts for children so as not only to keep the customer but also to increase a little their investment in health insurance as they are the second most interested in this segment. So, in the period close to the expiry of the insurance, these consumers would be contacted so that the salesperson would explain the advantages of renewing with their insurer and explain the benefit regarding the discount.

Moreover, in order to promote customer confidence in the other insurances offered by the insurer, content could be delivered by the means most appropriate to the age group of customers with a view to highlighting the solutions to their uncertainties

5.4. Up-selling

This segment are middle-aged clients, who despite having financial capacity do not invest much in insurance. They are still the second group that has invested the most in motor insurance.

Taking into consideration the low profit rate and their monthly salary, a good strategy to tap their potential in order become much more profitable an up-selling strategy would be perfect. Therefore, it could offer a premium motor insurance at times when customers are most likely to invest for instance in the summer when they receive their holiday allowance. This insurance would cost around 25% more than the standard motor insurance in order to not only improve the customer's experience with the service but also to generate more value for the company.

In addition, a premium package for health insurance could be created with higher costs than the standard package but with more benefits for households with dependents, as it is the second most invested in by this segment with children.

6. Conclusion

Throughout the development of this project, it was faced challenges that required not only Data Mining knowledge acquired during classes but also some investigation. To improve the company's marketing strategy, Data Exploration and Pre-Processing was an essential phase of the project in order to maintain the quality of the data set as it would influence all further analysis.

Importantly, it was attempted to create as many meaningful variables as possible in order to extract as much information as possible from the data set and to maintain the consistency and coherence of the original data.

To perform a more segmented clustering technique, two perspectives were created: demographic and consumption, each representing the demographic data and the client’s consumption data. With that, multiple models were applied to each perspective, finalizing with the K-Prototypes on the demographic side and K-Means on the consumption side. The merging of both perspective was done with a hierarchical process, obtaining four different clusters of customers.

In summary, we believe we have been able to deliver what "A2Z Insurance" asked us to do. According to the value and the demographic characteristics of the customers its was able to group them in 4 clusters. Consequently, a sales promotion strategy was defined that allies the segment needs to the company's price competitiveness; a cross-selling strategy that allows the consumers to acquire solutions that complement his experience and increase the company's profit; a consumer retention strategy focused mainly on promoting the satisfaction and trust of customers in the company in the long term and finally an up-selling strategy that allows adding value through premium packages with more benefits.

7. References

Patrick, H. (2019, February 19). The Importance of Feature Engineering and Selection. Rittman Mead. https://www.rittmanmead.com/blog/2019/02/the-importance-of-feature-engineering-and-selection/

Patel, H. (2021, September 2). What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. Medium. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

8. Appendix

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
FirstPolYear	10266.0	NaN	NaN	NaN	1991.062634	511.267913	1974.0	1980.0	1986.0	1992.0	53784.0
BirthYear	10279.0	NaN	NaN	NaN	1968.007783	19.709476	1028.0	1953.0	1968.0	1983.0	2001.0
EducDeg	10279	4	b'3 - BSc/MSc'	4799	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthSal	10260.0	NaN	NaN	NaN	2506.667057	1157.449634	333.0	1706.0	2501.5	3290.25	55215.0
GeoLivArea	10295.0	NaN	NaN	NaN	2.709859	1.266291	1.0	1.0	3.0	4.0	4.0
Children	10275.0	NaN	NaN	NaN	0.706764	0.455268	0.0	0.0	1.0	1.0	1.0
CustMonVal	10296.0	NaN	NaN	NaN	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	NaN	NaN	NaN	0.742772	2.916964	0.0	0.39	0.72	0.98	256.2
PremMotor	10262.0	NaN	NaN	NaN	300.470252	211.914997	-4.11	190.59	298.61	408.3	11604.42
PremHousehold	10296.0	NaN	NaN	NaN	210.431192	352.595984	-75.0	49.45	132.8	290.05	25048.8
PremHealth	10253.0	NaN	NaN	NaN	171.580833	296.405976	-2.11	111.8	162.81	219.82	28272.0
PremLife	10192.0	NaN	NaN	NaN	41.855782	47.480632	-7.0	9.89	25.56	57.79	398.3
PremWork	10210.0	NaN	NaN	NaN	41.277514	51.513572	-12.0	10.67	25.67	56.79	1988.7

Table 1: Descriptive Statistics

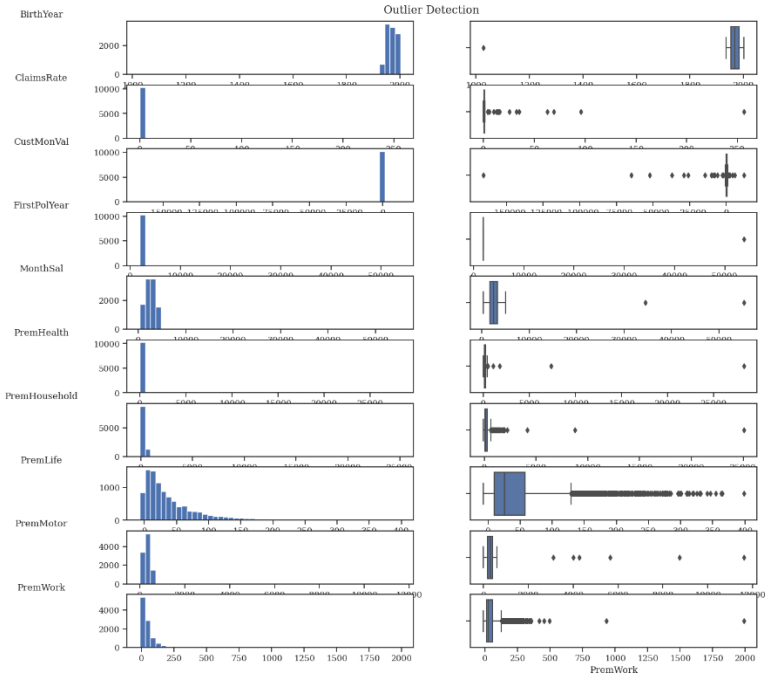

Figure 1: Outlier Detection

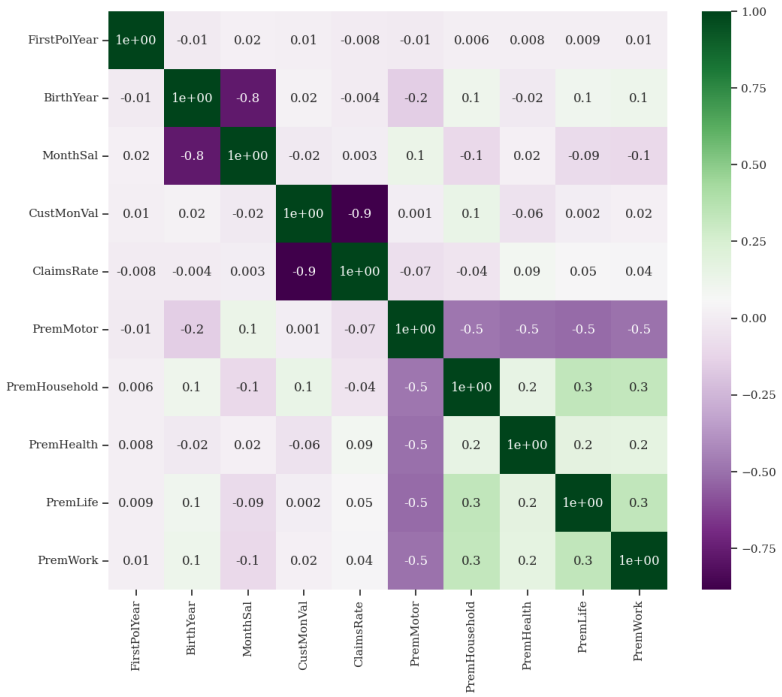

Figure 2: Original Correlation Matrix

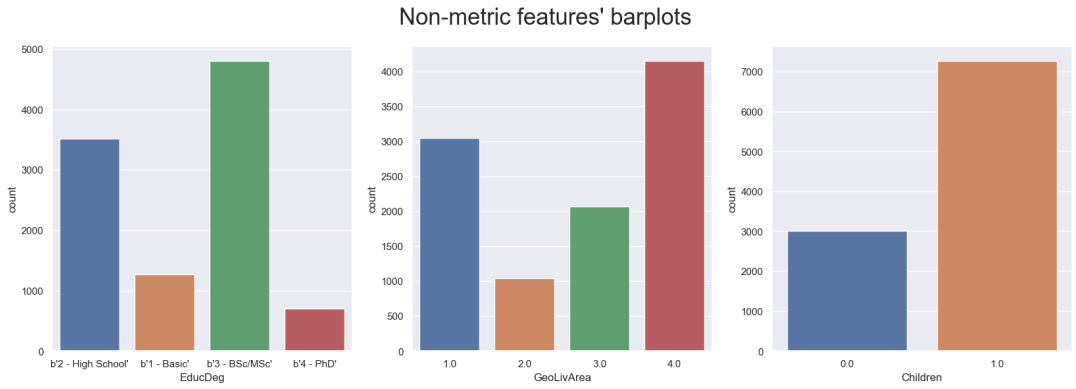

Figure 3: Categorical features' Bar plots

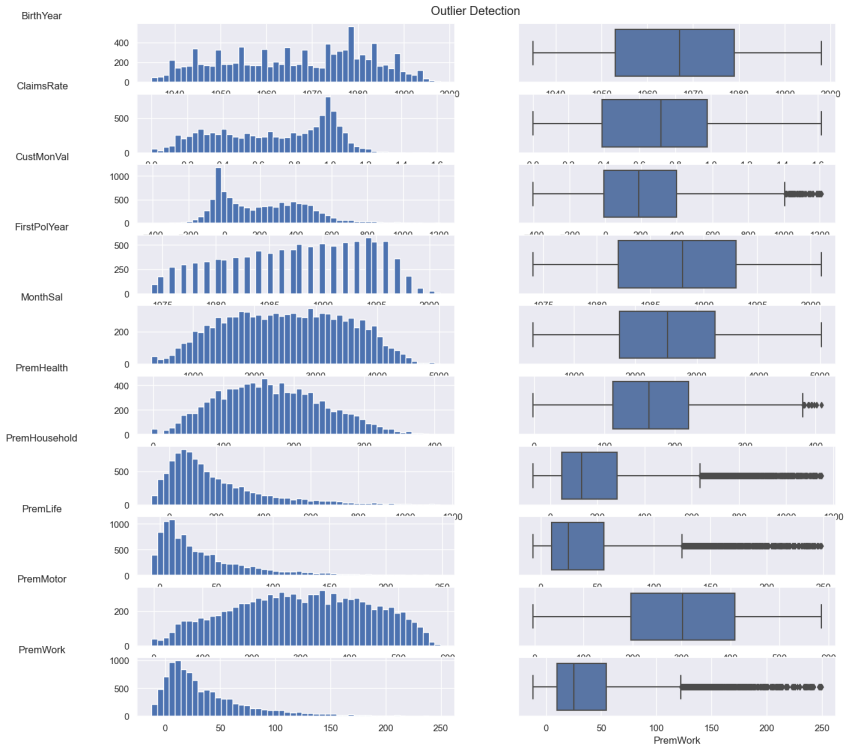

Figure 4: After Outliers' Removal Distribution

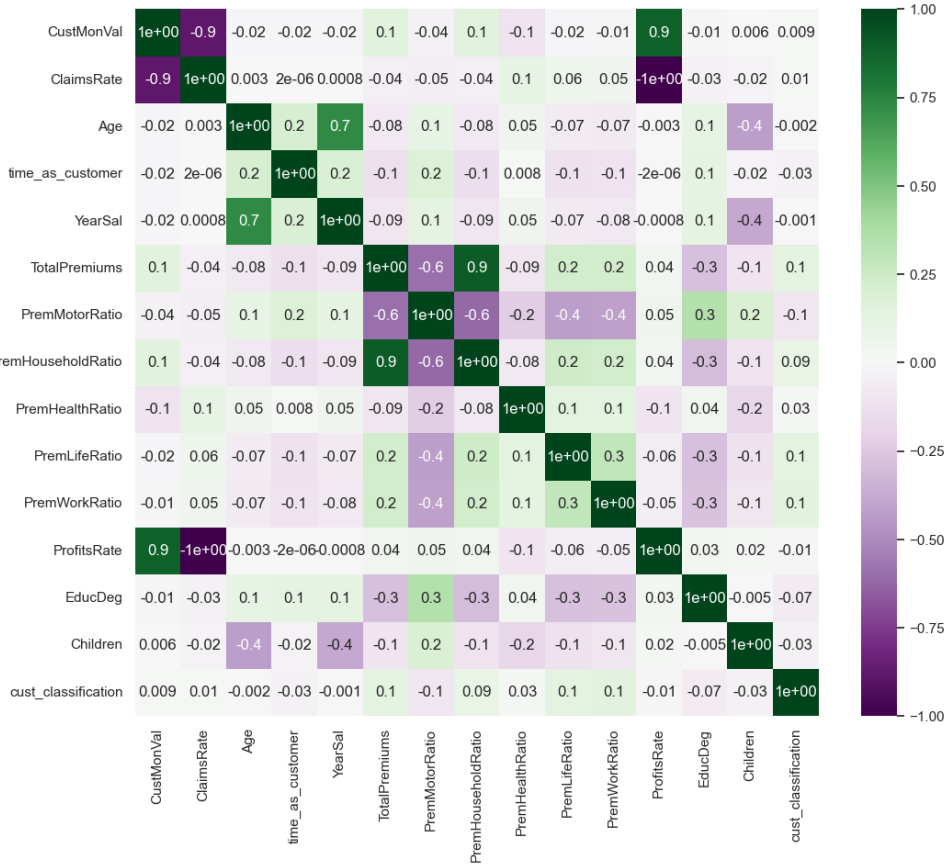

Figure 5: All Features Correlation Matrix

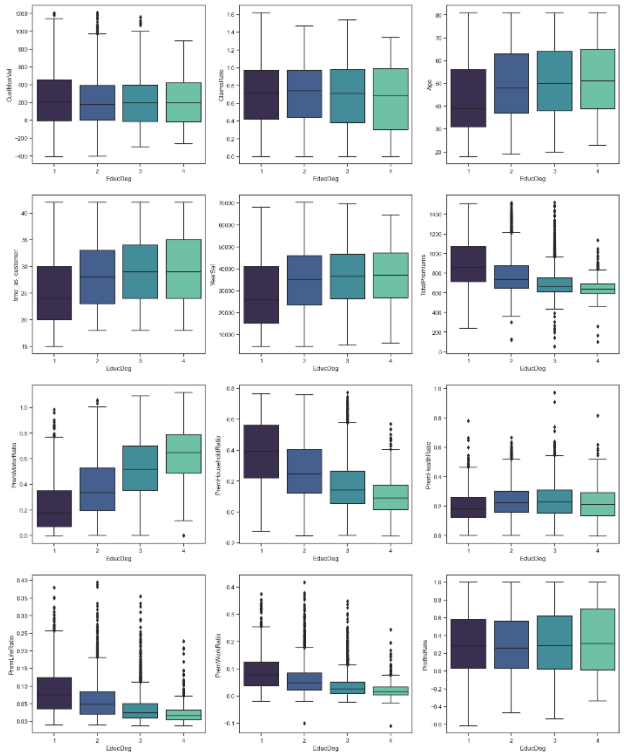

Figure 6: Education Degree Bar Plots

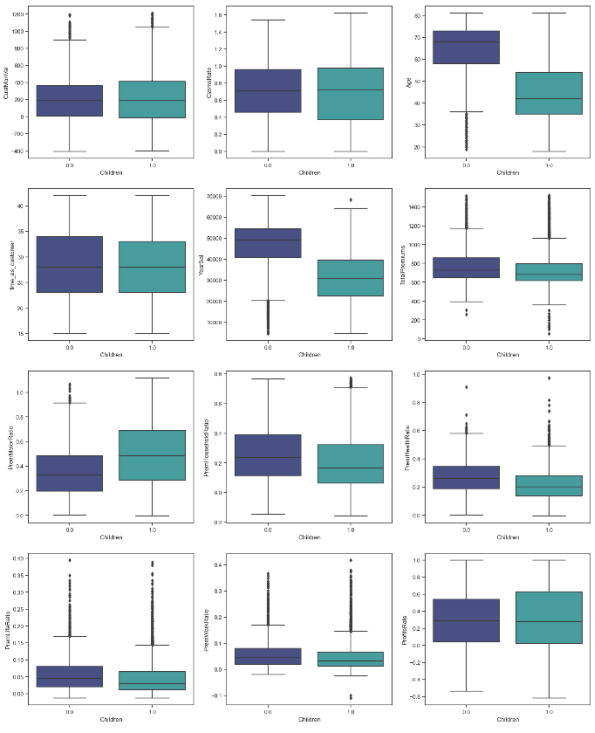

Figure 7: Children Bar Plots

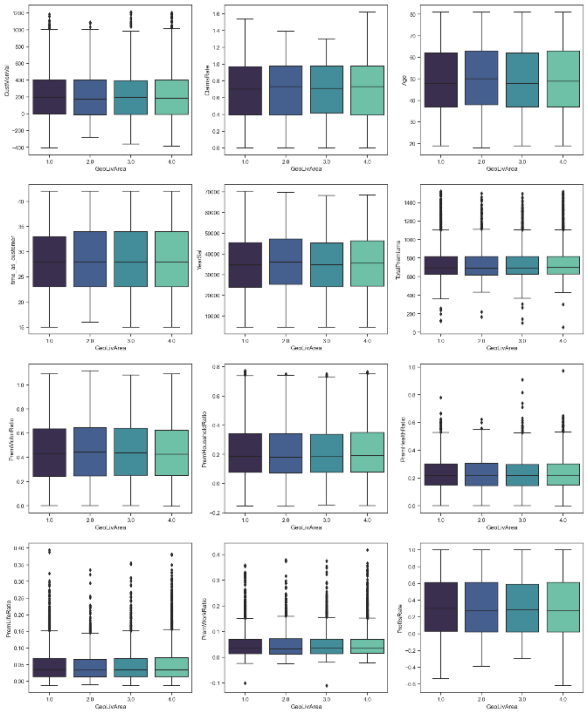

Figure 8: Geographic Area Bar Plots

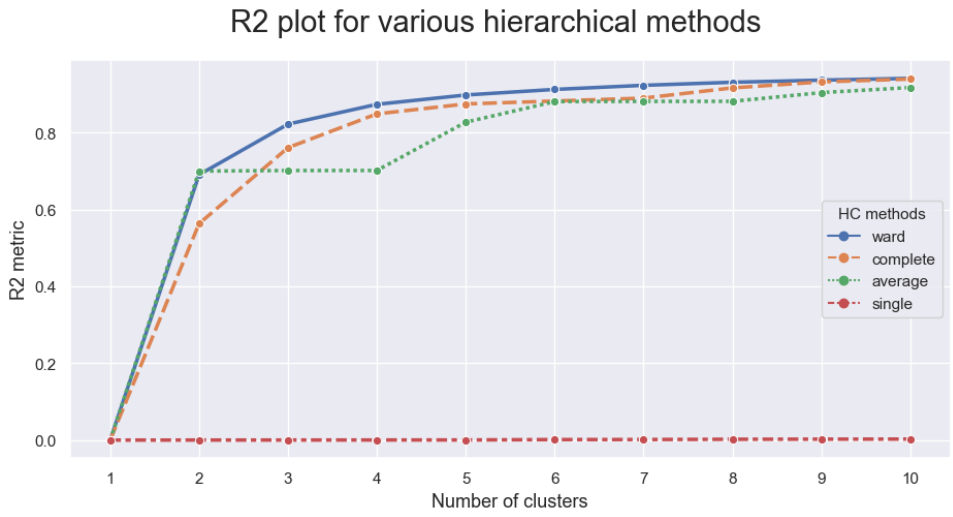

Figure 9: Hierarchical R-squared Plot 1

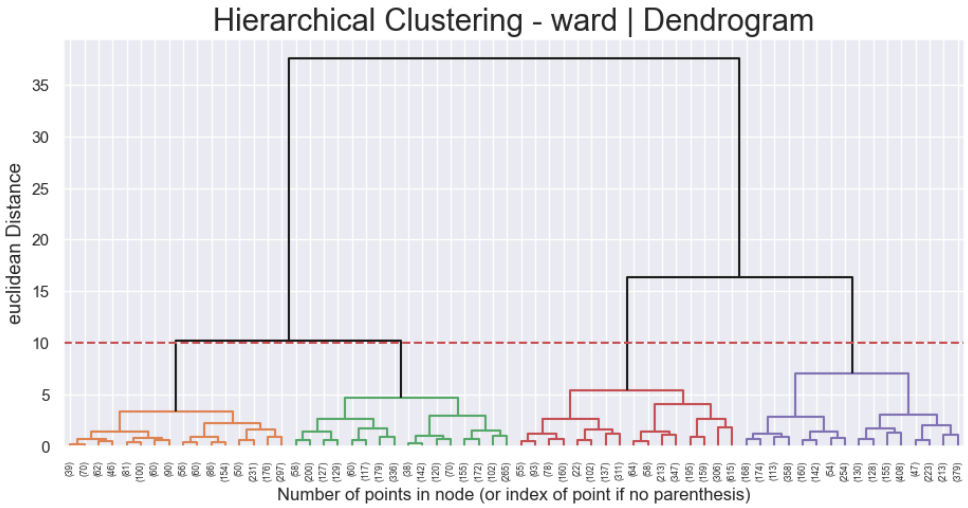

Figure 10: Hierarchical Dendrogram 1

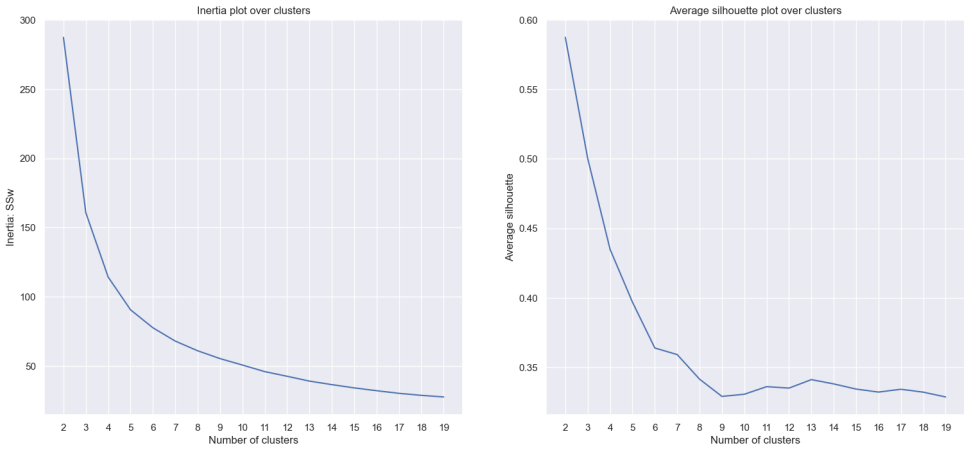

Figure 11: Inertia and Silhouette Plots 1

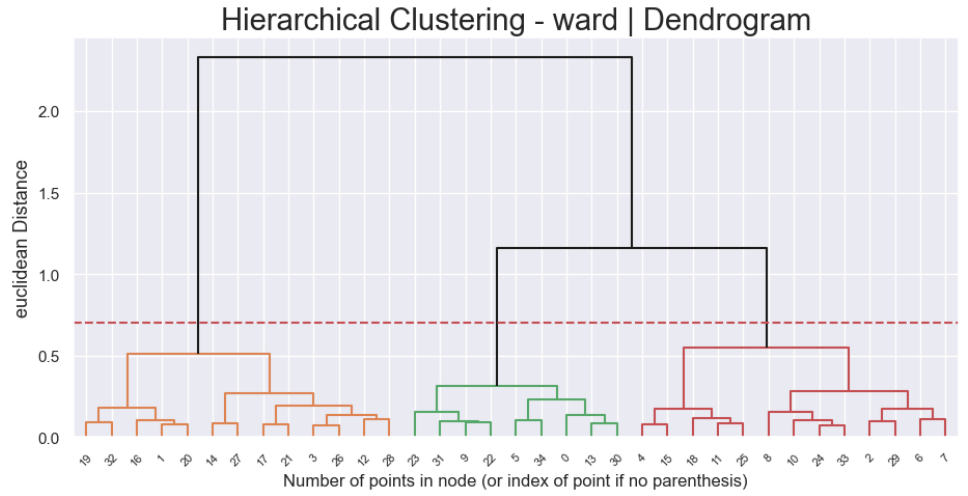

Figure 12: Hierarchical Dendrogram 2

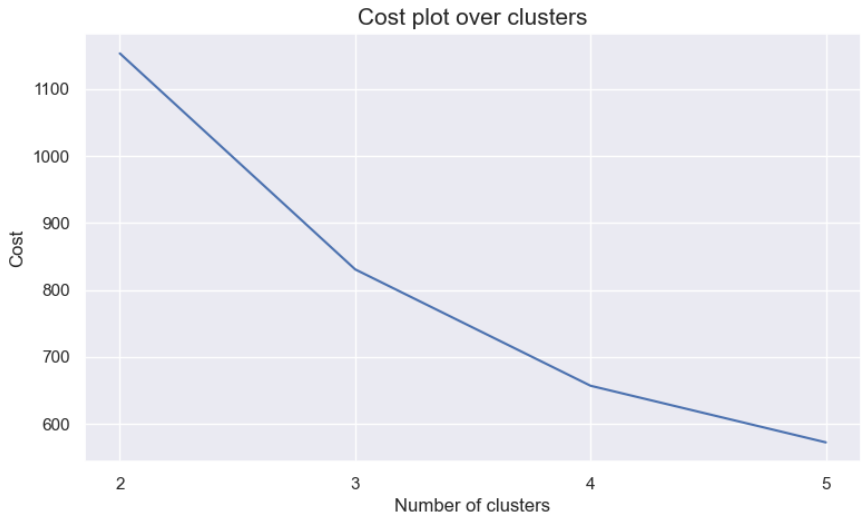

Figure 13: Cost over Clusters

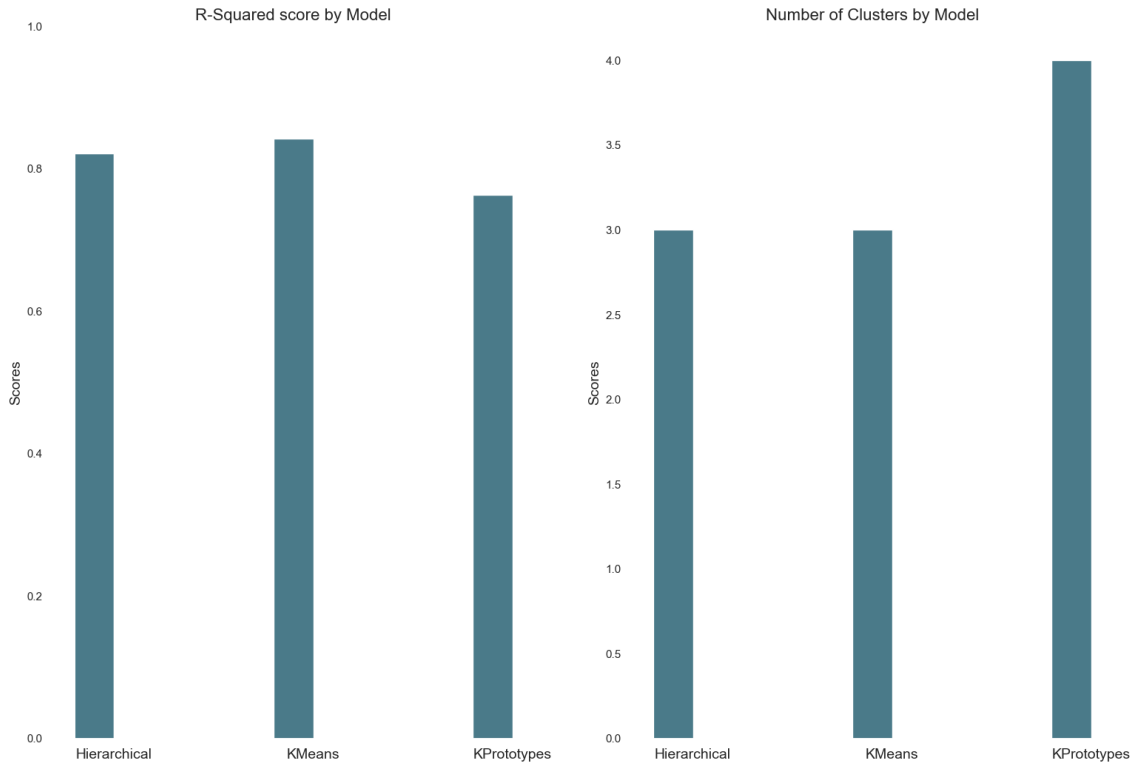

Figure 14: Demographic Models Analysis

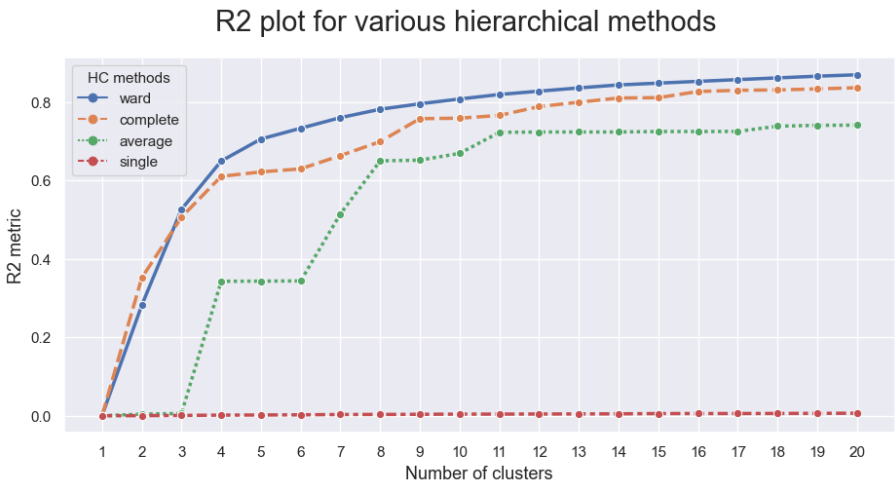

Figure 15: Hierarchical R-squared Plot 2

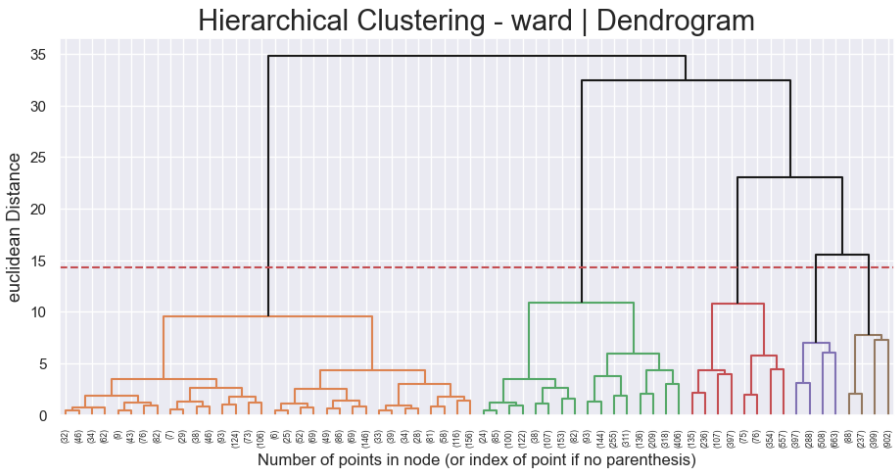

Figure 16: Hierarchical Dendrogram 3

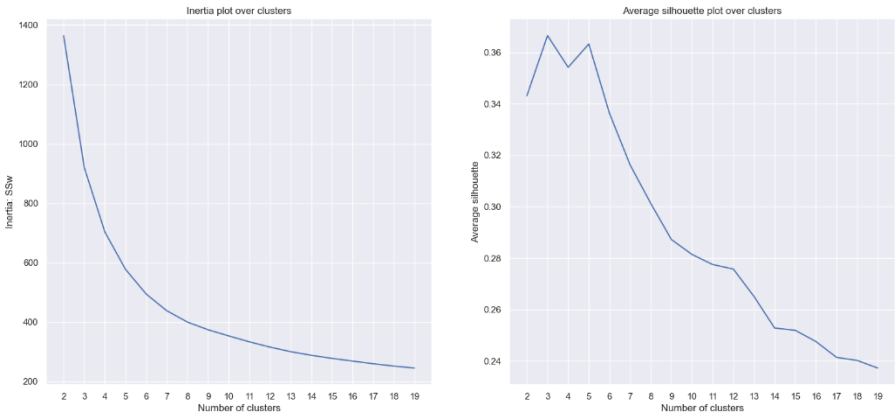

Figure 17: Inertia and Silhouette Plots 2

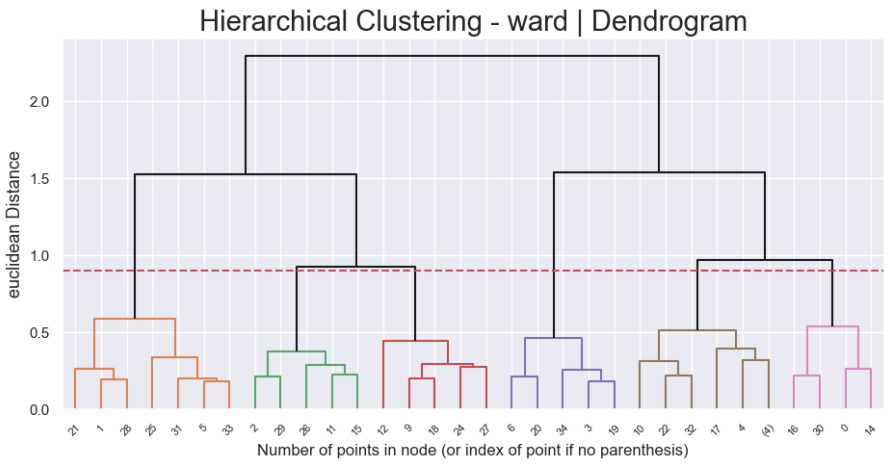

Figure 18: Hierarchical Dendrogram 4

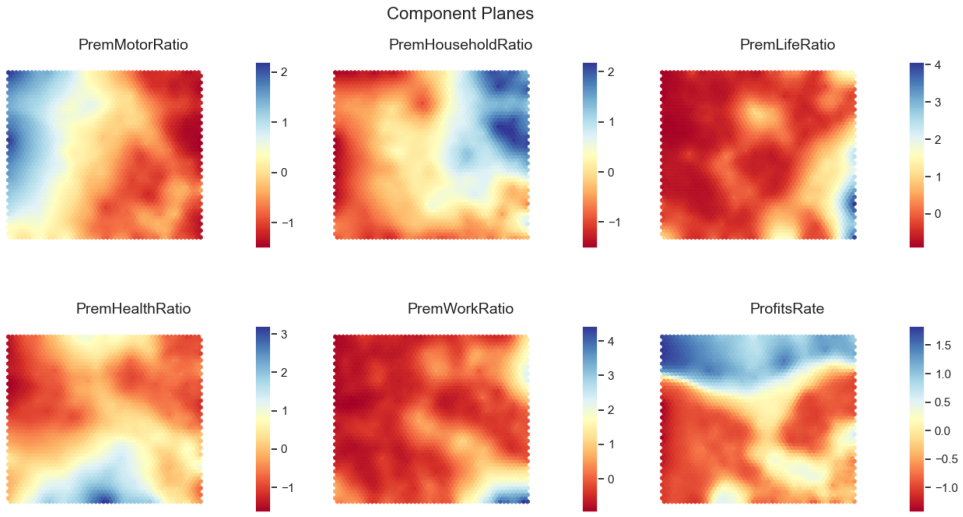

Figure 19: SOM's Component Planes

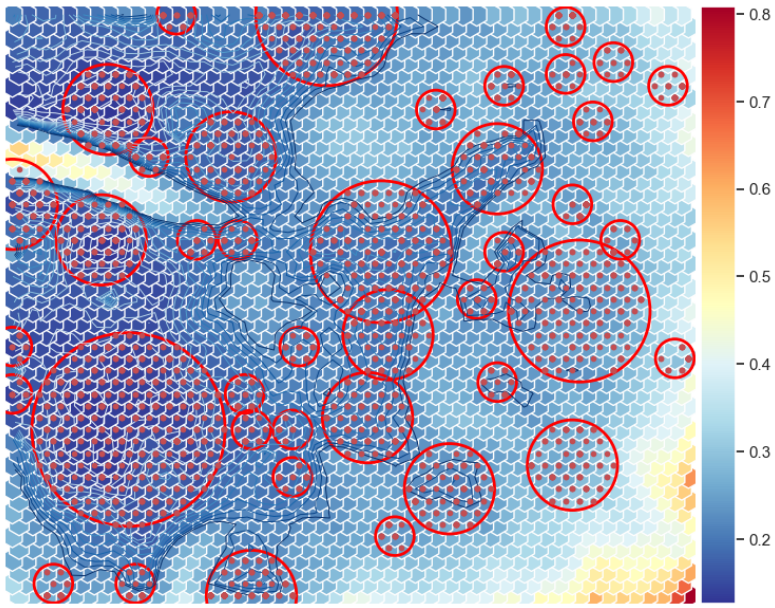

Figure 20: SOM's U-Matrix

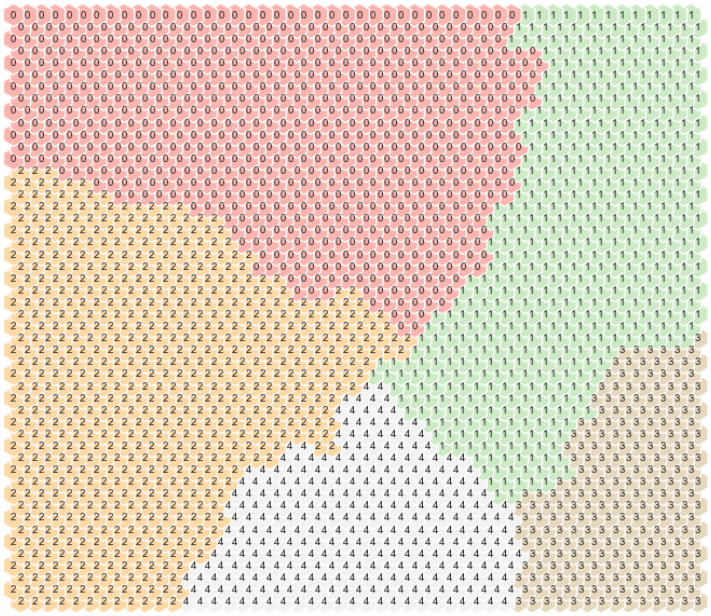

Figure 21: Hierarchical SOM Hit Map

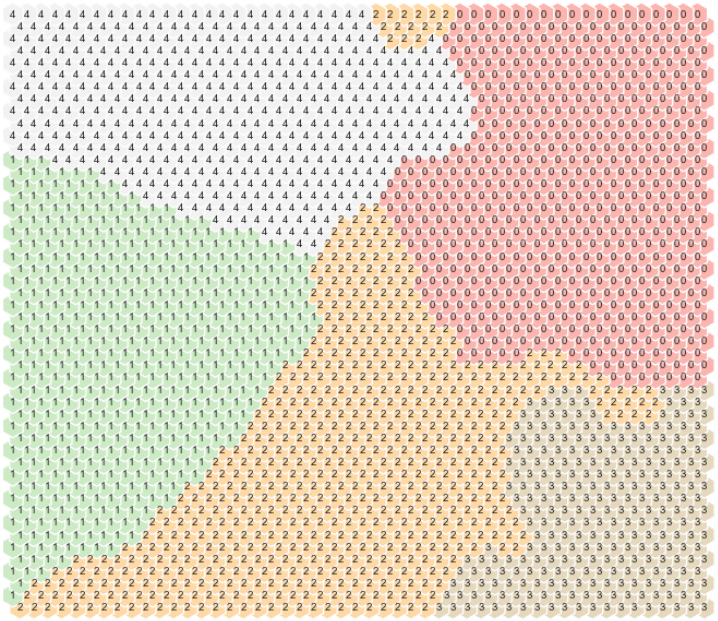

Figure 22: K-Means SOM Hit Map

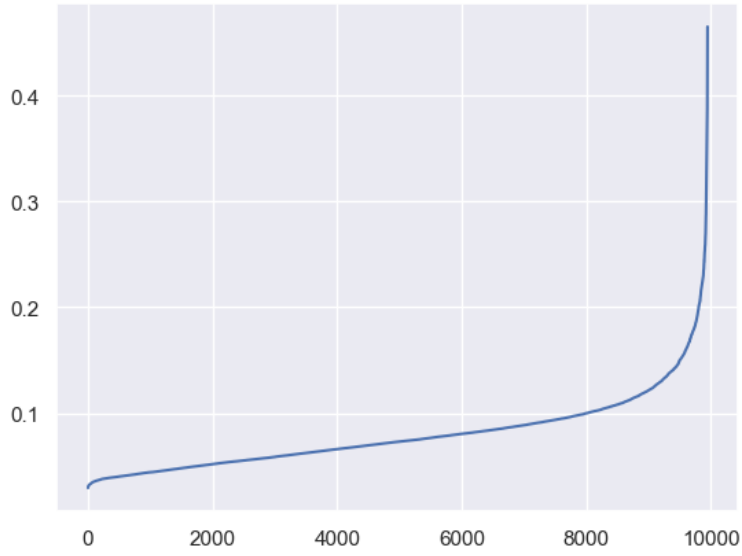

Figure 23: DBScan K-distance Graph

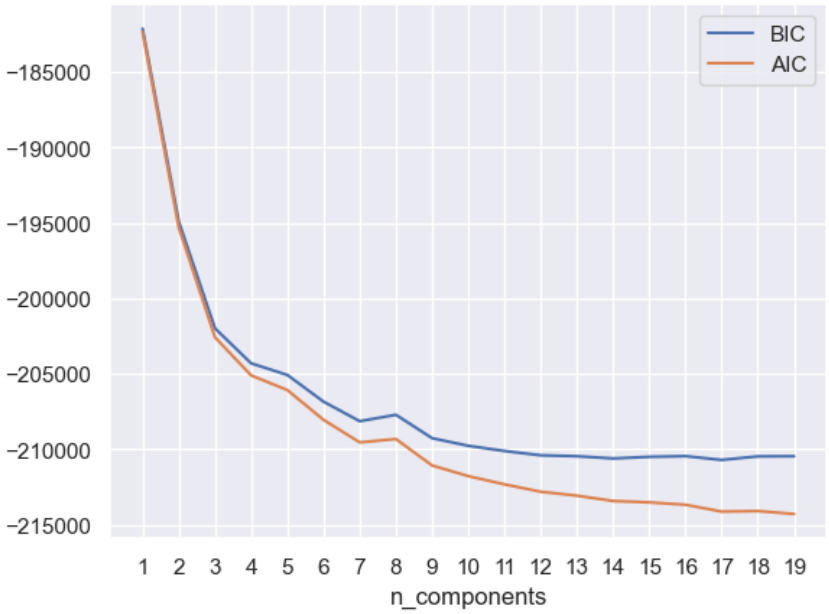

Figure 24: BIC and AIC plots

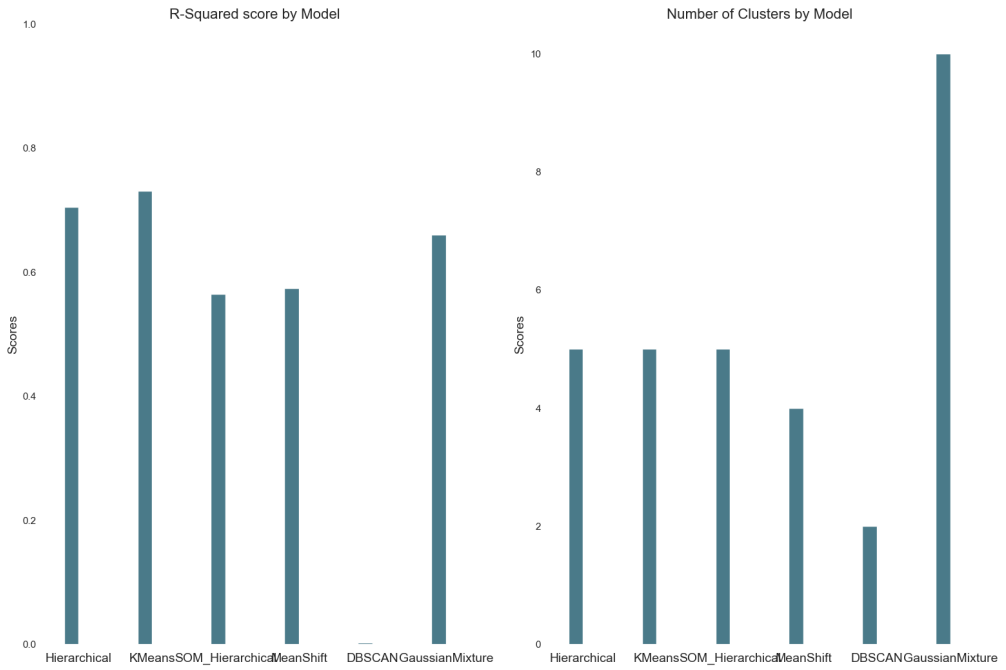

Figure 25: Consumption Models Analysis

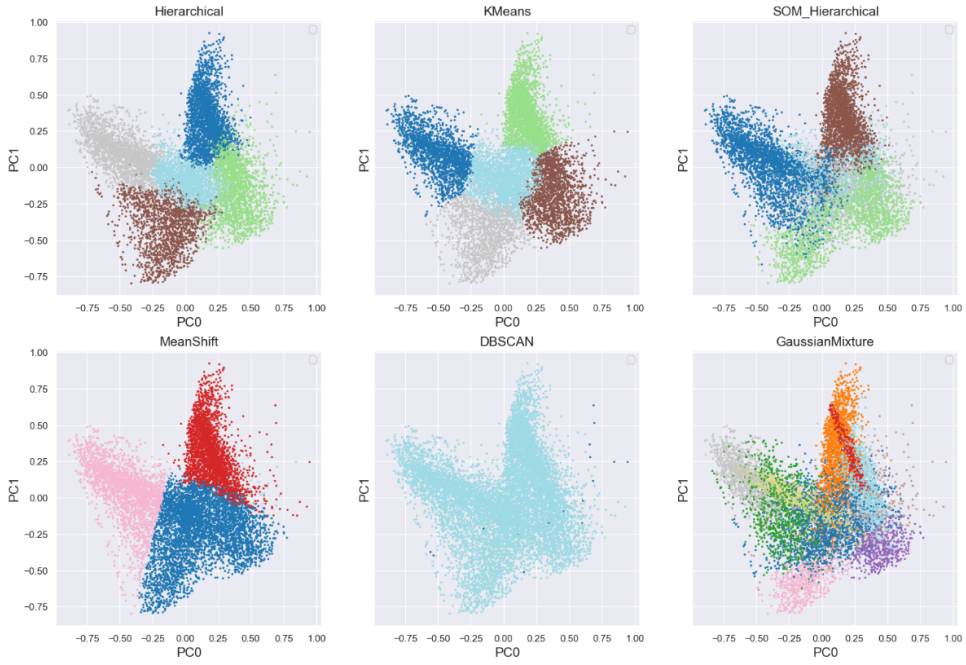

Figure 26: PCA Models Analysis

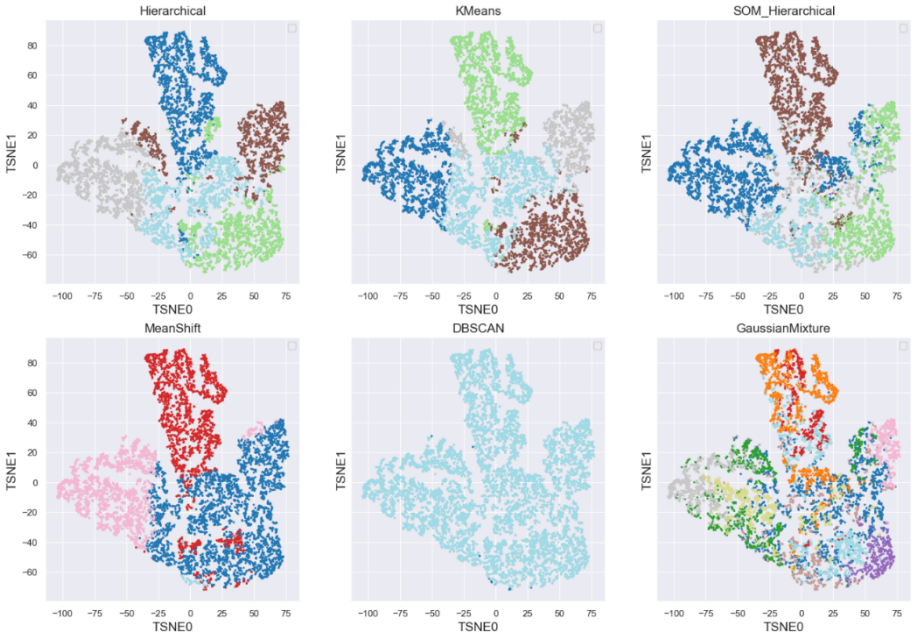

Figure 27: TSNE Models Analysis

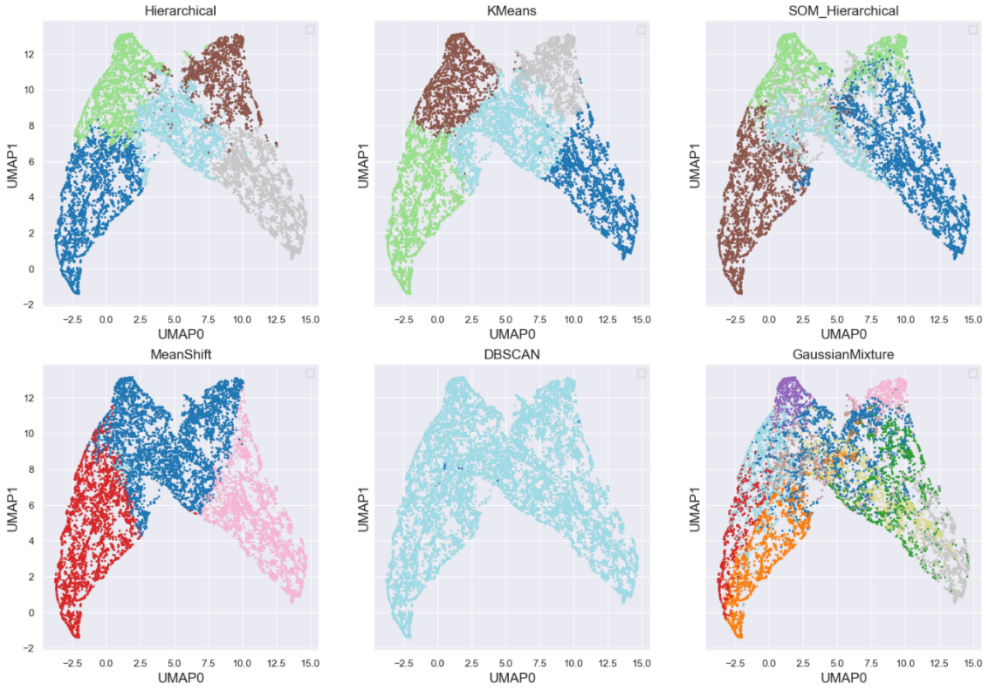

Figure 28: UMAP Models Analysis

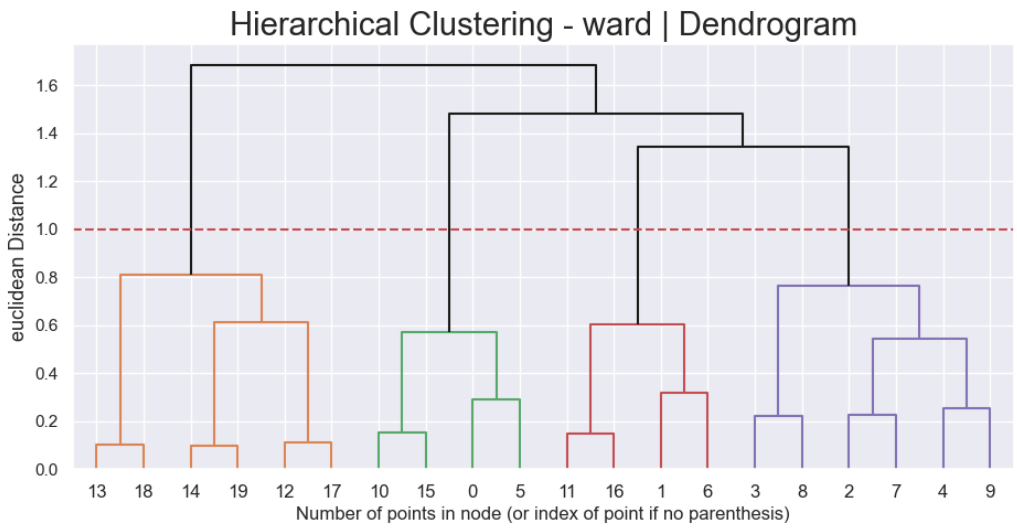

Figure 29: Hierarchical Dendrogram 5

	Importance
PremMotorRatio	0.295640
PremHouseholdRatio	0.005248
PremLifeRatio	0.000407
PremHealthRatio	0.000000
PremWorkRatio	0.000000
Age	0.336510
YearSal	0.004324
ProfitsRate	0.349881
EducDeg	0.007989
Children	0.000000

Table 2: Feature Importance on Decision Tree

merged_labels	Age	YearSal	PremMotorRatio	PremHouseholdRatio	PremHealthRatio	PremLifeRatio	PremWorkRatio	ProfitsRate	EducDeg	Children
0	34.565668	21641.870507	0.236323	0.359537	0.251060	0.076937	0.076133	0.303354	2.146478	0.858460
1	66.078899	48706.225020	0.306190	0.304854	0.271682	0.057578	0.059687	0.313642	2.428345	0.347189
2	50.260740	36435.514894	0.690973	0.078490	0.184602	0.023043	0.022874	0.736256	2.832320	0.835360
3	51.047325	37035.076234	0.664188	0.074955	0.207456	0.026346	0.027051	0.006719	2.771464	0.804645

Table 3: Clusters' Mean

merged_labels	Age	YearSal	PremMotorRatio	PremHouseholdRatio	PremHealthRatio	PremLifeRatio	PremWorkRatio	ProfitsRate	EducDeg	Children
0	35.0	17724.0	0.217	0.279	0.177	0.053	0.031	0.00	2	1.0
1	67.0	52864.0	0.355	0.327	0.204	0.042	0.034	0.00	3	0.0
2	48.0	36610.0	0.632	0.040	0.144	0.000	-0.000	0.76	3	1.0
3	49.0	37618.0	0.658	0.032	0.155	0.000	-0.000	0.00	3	1.0

Table 4: Clusters' Mode

merged_labels	Age	YearSal	PremMotorRatio	PremHouseholdRatio	PremHealthRatio	PremLifeRatio	PremWorkRatio	ProfitsRate	EducDeg	Children
0	35.0	21056.0	0.2280	0.3475	0.240	0.063	0.063	0.29	2.0	1.0
1	67.0	48860.0	0.3095	0.2950	0.258	0.050	0.050	0.32	3.0	0.0
2	49.0	36057.0	0.6810	0.0810	0.177	0.017	0.018	0.74	3.0	1.0
3	50.0	36260.0	0.6550	0.0790	0.196	0.018	0.020	0.00	3.0	1.0

Table 5: Clusters' Median