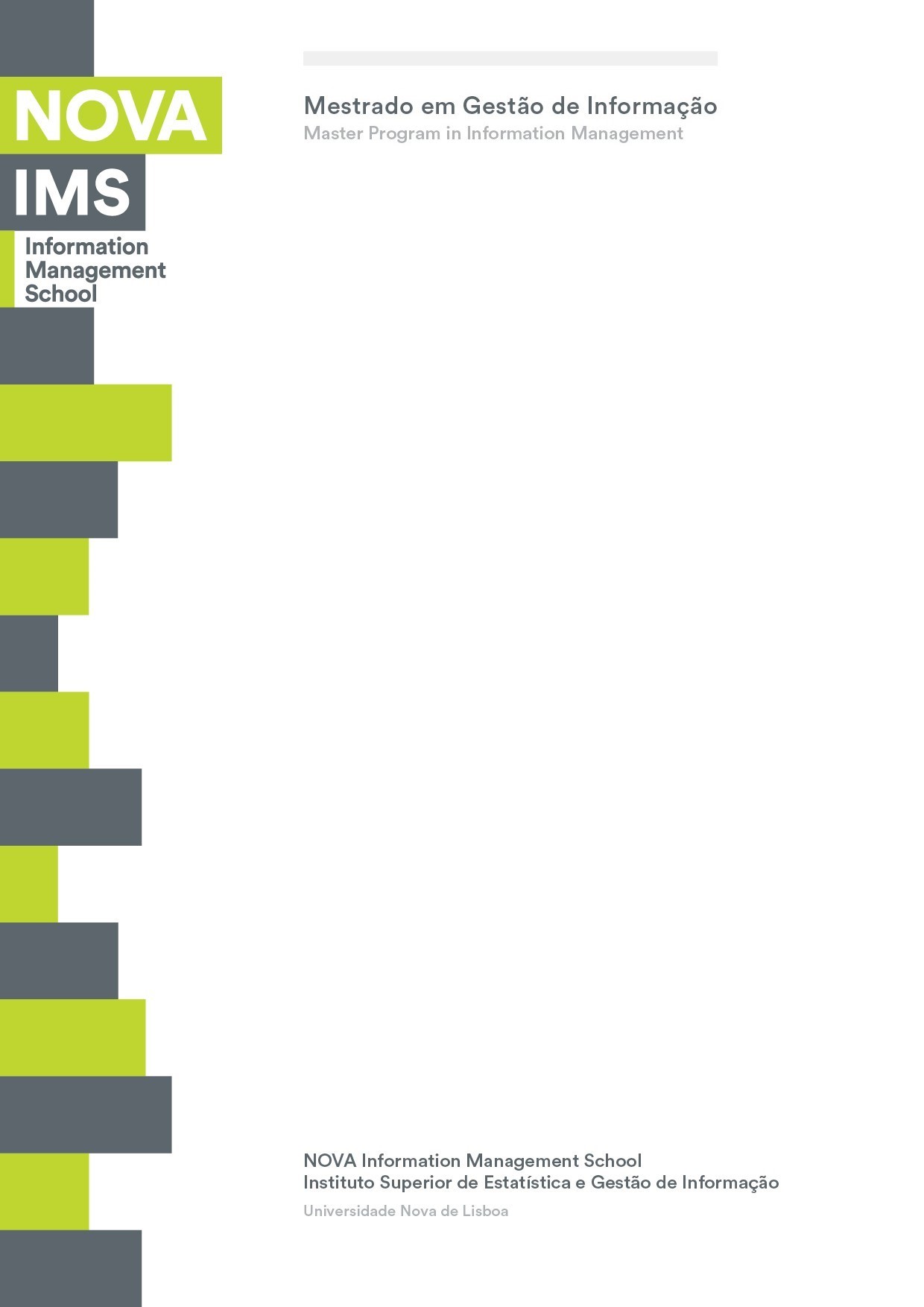
****

Group 28:

Carlos Lopes (M20170532)

Ricardo Gonçalves (M20170300)

Sofia Gouveia (M20180655)

Fall Semester

Academic Year 2018 - 19

**Descriptive Methods of Data Mining**

Index

[Introduction 3](#_Toc533773708)

[Data Exploration 3](#_Toc533773709)

[Input data 3](#_Toc533773710)

[StatExplore (1) 3](#_Toc533773711)

[Multiplot (1) 3](#_Toc533773712)

[Graph Explore and Variable Clustering (1) 4](#_Toc533773713)

[Replacement 4](#_Toc533773714)

[Filter (1) 4](#_Toc533773715)

[Impute (1) 5](#_Toc533773716)

[Transform Variables 5](#_Toc533773717)

[StatExplore (2), Multiplot (3), Graph Explore and Variable Clustering (2) 6](#_Toc533773718)

[Filter (2) 6](#_Toc533773719)

[StatExplore (3), Multiplot (4), Graph Explore and Variable Clustering (3) 6](#_Toc533773720)

[Metadata 7](#_Toc533773721)

[Value Cluster 7](#_Toc533773722)

[Value Cluster Segmentation 8](#_Toc533773723)

[Product Cluster 8](#_Toc533773724)

[Product Cluster Segmentation 9](#_Toc533773725)

[Conclusion 9](#_Toc533773726)

[Annexes i](#_Toc533773727)

[Annex 1: Project Diagram i](#_Toc533773728)

[Annex 2: Initial variables i](#_Toc533773729)

[Annex 3: Variable Summary (StatExplore 1) ii](#_Toc533773730)

[Annex 4: Interval Variables (StatExplore 1) ii](#_Toc533773731)

[Annex 5: Class Variables' Variation (StatExplore 1) ii](#_Toc533773732)

[Annex 6: Histogram from Multiplot (1) – Variable “Life Premium” iii](#_Toc533773733)

[Annex 7: Statistical information (Multiplot 1) iii](#_Toc533773734)

[Annex 8: Variable Clustering (1) iv](#_Toc533773735)

[Annex 9: Variable Correlation (Variable Clustering 1) iv](#_Toc533773736)

[Annex 10: Replacement Values for Class Variables (Replacement) v](#_Toc533773737)

[Annex 11: Limits and Replacements for Interval variables (Replacement) v](#_Toc533773738)

[Annex 12: Filter Limits for Interval Variables (Filter 1) v](#_Toc533773739)

[Annex 13: Number of Excluded Observations (Filter 1) v](#_Toc533773740)

[Annex 14: Impute (1) vi](#_Toc533773741)

[Annex 15: Transformation of variables (Transform Variables) vi](#_Toc533773742)

[Annex 16: Transformation Statistics (Transform Variables) vi](#_Toc533773743)

[Annex 17: Variables after transformation (StatExplore 2) vii](#_Toc533773744)

[Annex 18: Variable Clustering after transformation (Variable Clustering 2) vii](#_Toc533773745)

[Annex 19: Variable Correlation after transformation (Variable Clustering 2) viii](#_Toc533773746)

[Annex 20: Filter limits and excluded observations after transformation (Filter 2) viii](#_Toc533773747)

[Annex 21: Variables after transformation (StatExplore 3) ix](#_Toc533773748)

[Annex 22: Interval variables after transformation (StatExplore 3) ix](#_Toc533773749)

[Annex 23: Variable Clustering after transformation (Variable Clustering 3) x](#_Toc533773750)

[Annex 24: Variables for value segmentation (Value Cluster) x](#_Toc533773751)

[Annex 25: Value Cluster – Elbow Graph xi](#_Toc533773752)

[Annex 26: Imput means plot – Value Cluster (k = 4) xi](#_Toc533773753)

[Annex 27: Segment frequencies (Value cluster) xi](#_Toc533773754)

[Annex 28: Distance plot (Value cluster) xii](#_Toc533773755)

[Annex 29: Tree profile (Value cluster) xii](#_Toc533773756)

[Annex 30: Segment profiles (Value cluster) xiii](#_Toc533773757)

[Annex 31: Product Cluster – Elbow Graph xiii](#_Toc533773758)

[Annex 32: Imput means plot – Product Cluster (k = 5) xiv](#_Toc533773759)

[Annex 33: Segment frequencies (Product Cluster) xiv](#_Toc533773760)

[Annex 34: Distance plot (Product cluster) xv](#_Toc533773761)

[Annex 35: Tree profile (Product cluster) xv](#_Toc533773762)

[Annex 36: Segment profiles (Product cluster) xvi](#_Toc533773763)

# Introduction

The A2Z Insurance Company needed a better understanding of the existing customer profiles in order to establish different marketing approaches to each segment. In this context, we were asked to develop a customer segmentation, based on value and products, taking advantage of the SAS Enterprise Miner Tool (Annex 1).

# Data Exploration

The A2Z Insurance provided us a set of data that included 10,296 customers of five insurance policies (motor, household, health, life and work compensation) in different periods. Listed client attributes were year of birth, academic degree (education), gross monthly salary, living area, with or without children, and specific indicators regarding the interaction with the company, namely the year of first policy, monetary value and claims rate in the past 2 years.

# Input data

For the initial process, the variables were classified as nominal (education and living area), ordinal (birthyear and first policy year) and binary (with children or not). The remaining variables were considered as interval variables (as inputs) [Annex 2].

# StatExplore (1)

The five existing class variables had 86 missing values, and the eight interval variables had 303 missing values. The statistics also indicated motor as the biggest insurance in terms of average value, with life and work-related insurances being less representative. The class variables with the highest variability were having children, followed by education and living area, while the interval variables with the highest coefficient of variation were customer’s monetary value and claims rates (Annex 3, 4 and 5).

# Multiplot (1)

The original histograms show the distribution of the variables, allowing to detect some possible outliers, as well as eventual data errors and missing values. From observing these results, it was possible to confirm that outliers may exist in relation to the different insurance policies, although the life premium presents more of a long tail type distribution, which could mimic an outlier and interfere with the clustering process (Annex 6 and 7).

# Graph Explore and Variable Clustering (1)

By combining these two nodes, we can obtain customized charts that explore the data more effectively and offer an overall view of some important aspects of the initial data set.

In this phase, it was possible to distinguish three different clusters, with a dominant segment mixing the variables of value and product, that explained about 60% of the variance of the data (Annex 8).

Furthermore, a strong negative correlation between two interval variables, monetary value and claims rate, was identified (Annex 9).

# Replacement

With this node, we decided how to treat the missing values.

The chosen options were to use the most common value (mode) for the variables education and location. In the first case, the 17 missing values were replaced by the string 3 and regarding location, 1 missing value was attributed to region 4.

The 21 blanks in the binary variable “children” were interpreted as "no children", since that would correspond to a numeric value of "0".

As for the missing values in birth year and year of first policy, they were assigned as 1973 and 1995 (the values of low/through years), respectively.

The possibility of restricting the data range of the variables, keeping extreme observations within pre-defined limits was also tested. However, there seemed to be no benefit in this strategy, possibly due to the characteristics of the data sets and previously placed filters.

Also, this type of restrictions are not ideal for sensitive data, and might not be applicable for this data set (Annex 10 and 11).

# Filter (1)

In order to perform any analysis, it is essential to find a balance between preserving the variability of the data and eliminating the outliers that could impact the results, without compromising the differentiation that allows clustering.

As such, clustering should minimize the variability within each cluster and, at the same time, maximize the variability between them.

Regarding the treatment of outliers, two class variables, the first policy year and birth year, were filtered to exclude the two wrong values, while the interval variables were adjusted to remove the extreme outliers (Annex 12).

This process led to the exclusion of 4.1% of observations, that is, 422 observations in total (Annex 13).

# Impute (1)

At this stage, the missing values that were previously kept could be filled.

Although this decision can introduce a persistent error throughout the remaining analysis, we felt that this would be a preferable option, rather than eliminating all the missing observations.

As explained in class, we decided to use the surrogate tree for missing class values, and the median to replace missing interval variables.

With this, the impute node was able to substitute 294 missing values, which allowed us to further preserve the data set (Annex 14).

# Transform Variables

This node allows the introduction of new variables to the process. These can be related to the business problem in question and involve the transformation of previous variables, thus permitting further interpretation of the data set and clarification of the segmentation process.

For the purpose of this segmentation process, we decided to explore two different aspects, client value and product profile.

Regarding client value, the newly created variables were customer age and antiquity (number of years as client), total premium, cost to the company, margin, margin percentage, monthly effort percentage and annual salary.

The product segmentation required determining the relative contribution of each of the five insurance policies available in relation to each client’s expenses (product percentage) [Annex 15].

During the integration of the data, we noticed that about 2000 customers were younger than their first policy and about 3000 customers were under 18 years old when signing up on the first insurance policy.

This was affecting about 50% of the data set, and needed to be addressed, in order not to exclude such a significant number of observations.

One plausible explanation would be a recurrent typing error when entering the data, with the year of birth mistakenly switched with the year of first policy.

Using the oldest date to determine age and the most recent date to calculate the number of years as client, we reached a range of clients between 18 and 81 years old and antiquity between 15 and 42 years. These results seemed to be consistent and in agreement with the switching hypothesis (Annex 16).

# StatExplore (2), Multiplot (3), Graph Explore and Variable Clustering (2)

After performing all these transformations, it was necessary to reassess the data. At this point, there were 19 variables and no missing values (Annex 17).

Some variables (customer monetary value, cost to company, health percentage, life percentage, work percentage, margin, margin percentage, monetary effort percentage, and total premium) presented peaks and discontinuous long-tail distributions. As such, the use of additional filtering to remove these extreme values could be considered.

The Graph Explore and Variable Clustering (2) nodes indicated that there was a strong correlation between some pairs of the transformed variables: total premium and household percentage; margin and margin percentage; both margin/margin percentage and customer monetary value; customer monetary value and cost to the company; age and annual salary (Annex 18 and 19)

With this information, it was possible to exclude some variables from the segmentation process, as they would be no additional benefit to the definition of distinct clusters. The excluded variables were customer age, margin, margin percentage and claims rate.

# Filter (2)

After choosing the variables, we tried to further exclude the presence of outliers, keeping in mind that the removal of an excessive number of observations could compromise the variability of the data set.

At this point, the class variables remained unchanged, and only the interval variables were trimmed. This resulted in the additional exclusion of 137 observations (Annex 20).

# StatExplore (3), Multiplot (4), Graph Explore and Variable Clustering (3)

The final exploration of the data set before the segmentation process showed that there were no missing values within the 9737 enduring observations. Relatively to the original data set, a total of 559 observations we excluded (which correspond to 5,4% of all the observations) [Annex 21 and 22].

The results from the Variable Clustering (3) node pointed to the automatic creation of four clusters, with the product and value variables mixed and the retention of about 75% of the variability (Annex 23).

# Metadata

The Metadata node allows changing the attributes of variables, nevertheless, the analysis previously produced did not require any changes in the variables’ characteristics.

For the purpose of this segmentation process, the included variables were annual salary, costs to the company, monetary value, monthly effort percentage, the total premium and the percentage of each insurance policy relative to the total expense of the client.

After completing the data preparation, we decided to generate a segmentation based on value and another on product. The value-based segmentation included the variables: annual salary, costs to the company, monetary value, monthly effort percentage and total premium. As for the product segmentation, we exploited the relative percentages of the five insurances policies subscrived by each client.

The cluster analysis method was based on the K-means algorithm

# Value Cluster

This node allows for the selection of desired variables. As mentioned, we decided that the value segmentation would include the annual salary, costs to the company, monetary value, monthly effort percentage and total premium (Annex 24).

Before the k-means algorithm could be properly applied, it was also necessary to standardize the data and use a principal components analysis for the initial seeds.

To determine the most suitable number of clusters, that is, the k-value, we created the associated elbow graph, for k-values ranging between 2 and 10.

For each value of k, the sum of squared errors (SSE) is calculated and plotted. This represents the distance between the different points in each cluster and its centroid. As the distance gets smaller, the capability of the model to resemble the behavior of the data set increases.

The point in which there is the most pronounced inflexion of the line (also referred as the knee or elbow) should be the most appropriate value for k.

Although a higher number of clusters will decrease the SSE, it will also result in a more complex model that can be difficult to interpret. With this in mind, the trade-off between gaining variability at the expense of increased complexity should always be sought.

In this case, we targeted a k-value between 4 to 6 (Annex 25).

When evaluating these options, we found that there were only small differences in the relative position of each cluster, throughout the different variables, when increasing the number of clusters. As so, it would probably be of little interest to consider increasing the complexity of the segmentation.

As such, we concluded that k = 4 would provide the least complex interpretation for clustering, while maintaining reasonable differentiation between the clusters (Annex 26 and 27).

The input means plot represents the relative position of each cluster comparatively to the overall population and each other. For the selected segmentation (k = 4) it pointed to:

* Cluster 1 (45% of customers) - high salary and cost to the company with low premium, value and effort.
* Cluster 2 (12% of customers) – high premium, effort and cost to the company with low value and salary.
* Cluster 3 (37% of customers) – high salary and value cost with low premium, effort and cost to the company.
* Cluster 4 (6% of customers) – high premium, effort and value with low salary and cost.

The proximity diagram reveals four distinct clusters that are paired together in opposite extremes: cluster 1 closest to cluster 3 and cluster 2 closer to cluster 4 (Annex 28).

The decision tree regarding this segmentation process is represented in (Annex 29)

# Value Cluster Segmentation

This node allows to compare the variables of each clusters with those of the overall population. Regarding cluster behavior, it reveals similar information to that which was already described above (Annex 30).

With this data, it seems that both cluster 1 and cluster 3 could be targets for further businesses. On the other hand, cluster 2 and cluster 4 may be at risk of being lost due to excessive effort.

# Product Cluster

To achieve a product-based customer segmentation, we followed a similar approach to the one described for value segmentation. In this case, the chosen variables were the percentage of the five existing insurance policies (motor, health, life, work and household) relative to the total expense of the client

Once again, we opted for a number of clusters that allowed a more distinct segmentation, in order to reduce complexity and create a model that would be easier to interpret when developing the business strategy.

Given the results from the elbow graph and exploration of the input means plot from k-values between 5 and 7, we decided to choose a model where k = 5 (Annex 31, 32 and 33).

Through the analysis of the input means plot for k = 5, regarding the average characteristics of the clusters it was possible to observe the following:

* Cluster 1 (6% of customers) – high work and life insurances premiums with low values motor policies and intermediate on the remaining variables.
* Cluster 2 (22% of customers)– high household premiums and low values on all other variables.
* Cluster 3 (34% of customers) – high motor premium with low values on all other insurance policies.
* Cluster 4 (10% of customers) – high work and health premiums with low values on the remaining variables.
* Cluster 5 (28% of customers)– higher health insurance premiums with low values on all other policies.

These five clusters seem to be relatively distinct, particularly cluster 3, which is the furthest apart from all other clusters (Annex 34).

Annex 35 depicts the decision tree associated with this segmentation process.

# Product Cluster Segmentation

The information obtained through this node seemed to be in line with the previously elaborated results, adding more accurate details on the composition of the clusters compared to the overall population (Annex 36).

As there seems to be a very defined segmentation based on product subscription in almost all clusters, there may be potential for additional insurances being sold to the different groups, if specific marketing campaigns are put to place.

# Conclusion

As mentioned, we chose to explore the data set trough two different segmentation processes.

The value-based segmentation produced 4 distinct clusters, where:

* Cluster 3 (V3) is considered the most profitable, due to lower cost, higher salary and value.
* Cluster 1 (V1) should be targeted in order to reduce cost, given the customers’ high salary and low effort.
* Cluster 2 (V2) can also benefit from a strategy of cost-reduction, despite lower salaries, in order to prevent losing these clients to more affordable competitors.
* Cluster 4 (V4) is a delicate group, where value is high, nevertheless clients are also at risk of being lost if not properly handled, due to excessive effort.

The segmentation based on products created 5 clusters, almost all of them specifically directed to a particular type of insurance: Cluster 1 (P1) for work and life; Cluster 2 (P2) for household; Cluster 3 (P3) for motor; Cluster 4 (P4) for work and health; Cluster 5 (P5) for health.

This unidimensional tendency suggests an opening for new business opportunities, trying to expand the portfolio of policies for the different segments, as mentioned above.

The table below represents the combination of both segmentation processes.

Table 1: Combination of Value and Product Clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Product  Value | Cluster 1 (P1) | Cluster 2 (P2) | Cluster 3 (P3) | Cluster 4 (P4) | Cluster 5 (P5) |  |
| Work + Life | Household | Motor | Work + Health | Health | **Total** |
| Cluster 1  (V1) | 190 | 544 | 1720 | 483 | 1417 | **4354** |
| Cluster 2  (V2) | 180 | 610 | 20 | 147 | 175 | **1132** |
| Cluster 3  (V3) | 130 | 585 | 1566 | 288 | 1051 | **3620** |
| Cluster 4  (V4) | 99 | 421 | 4 | 59 | 48 | **631** |
| **Total** | **599** | **2160** | **3310** | **977** | **2691** | **9737** |

Considering that clients in the V1 and V3 segment presented higher salary with lower value and effort, this large group could be the primary focus of further marketing campaigns, offering additional insurance policies. For this purpose, selecting clients that were also included in clusters P2, P3 and P5 would probably be a more cost effective initial strategy, since these are the clients that tend to have only subscribed to one particular type of product.

Customers in the V2 segment were identified as representing higher cost for the company. To reverse these negative outcomes, their contracts could be reviewed, particularly in terms of coverages, service levels and deductible fees. At the same time, a broader range of policies could be offered, in order to increase profit margins (again, in this situation, clients also in clusters P2, P3 and P5 should be the first to be targeted).

Albeit, it should be kept in mind that these clients may not withstand the increased expenditures and be lost to competitors. As such, the marketing strategies should be less aggressive than those targeted at segments V1 and V3.

As mentioned, clients in the V4 segment represent the more delicate group, where effort and premiums are high, compared to a lower salary. Most of them subscribe to household, work and life coverages, which may be associated to real-estate mortgages.

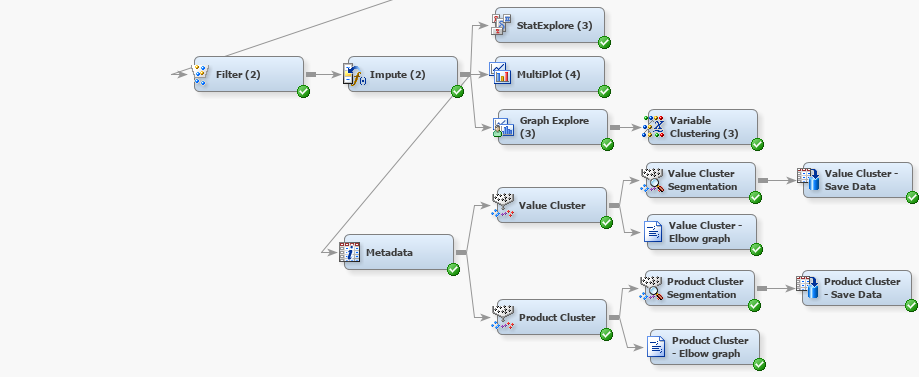
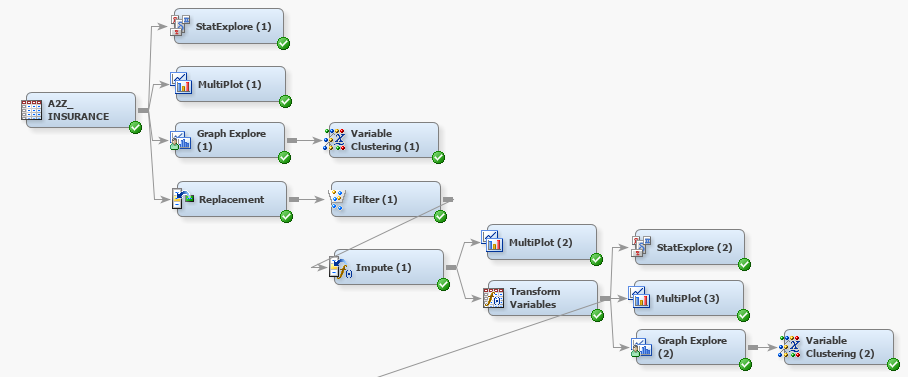
An interesting approach for this segment could be to reduce effort through more personalized services, as well as developing new products or partnerships related with financial institutions.

Considering the products, motor insurance, which is already the most profitable, should be offered to clients that subscribed to one of the other policies (segments P4 and P5) and vice-versa.

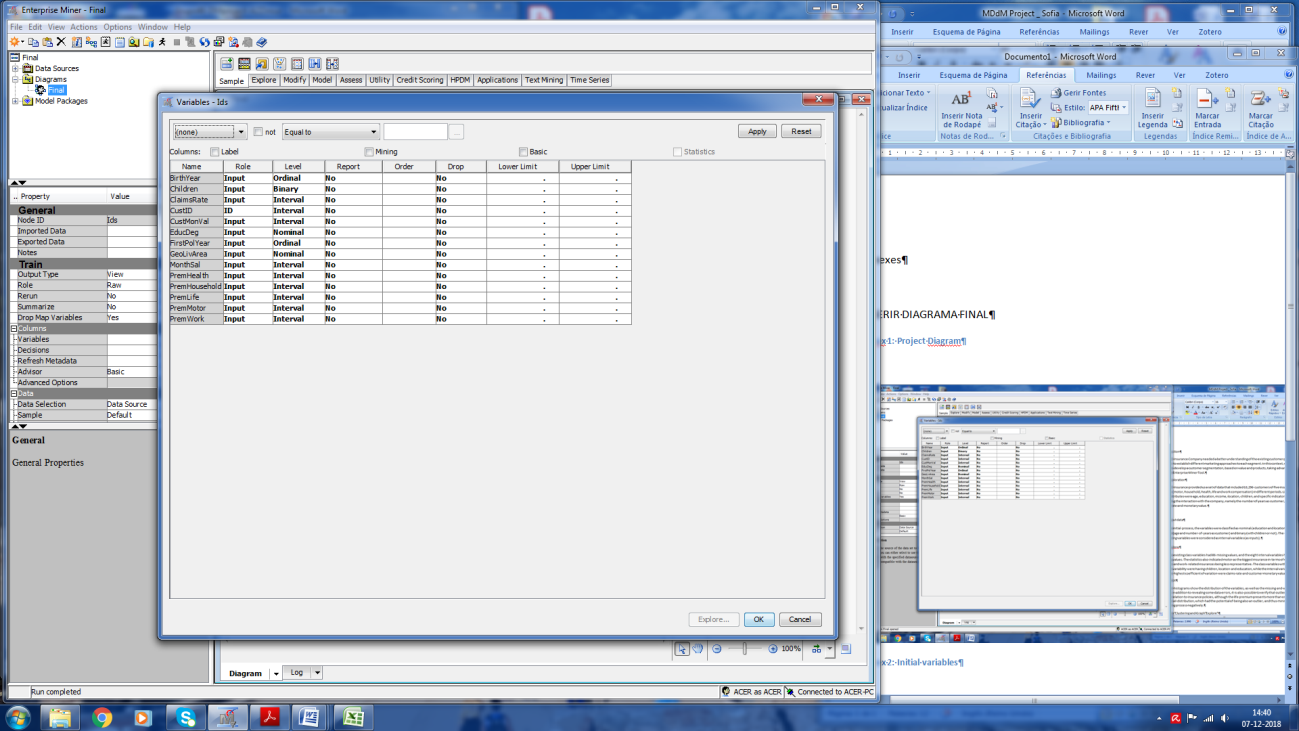
These business opportunities should be more aggressively focused on segments V1 and V3, while clusters V2 and V4 could benefit from a more tailored approach, possibly involving the creation of new products, policy combinations or stratified types of coverages.

Using this approach, the company could not only increase the value and profit margins from clusters V1 and V3, but also adjust the coverages provided to segments V2 and V4 customers, trading off between their costs and effort with loyalty and value.

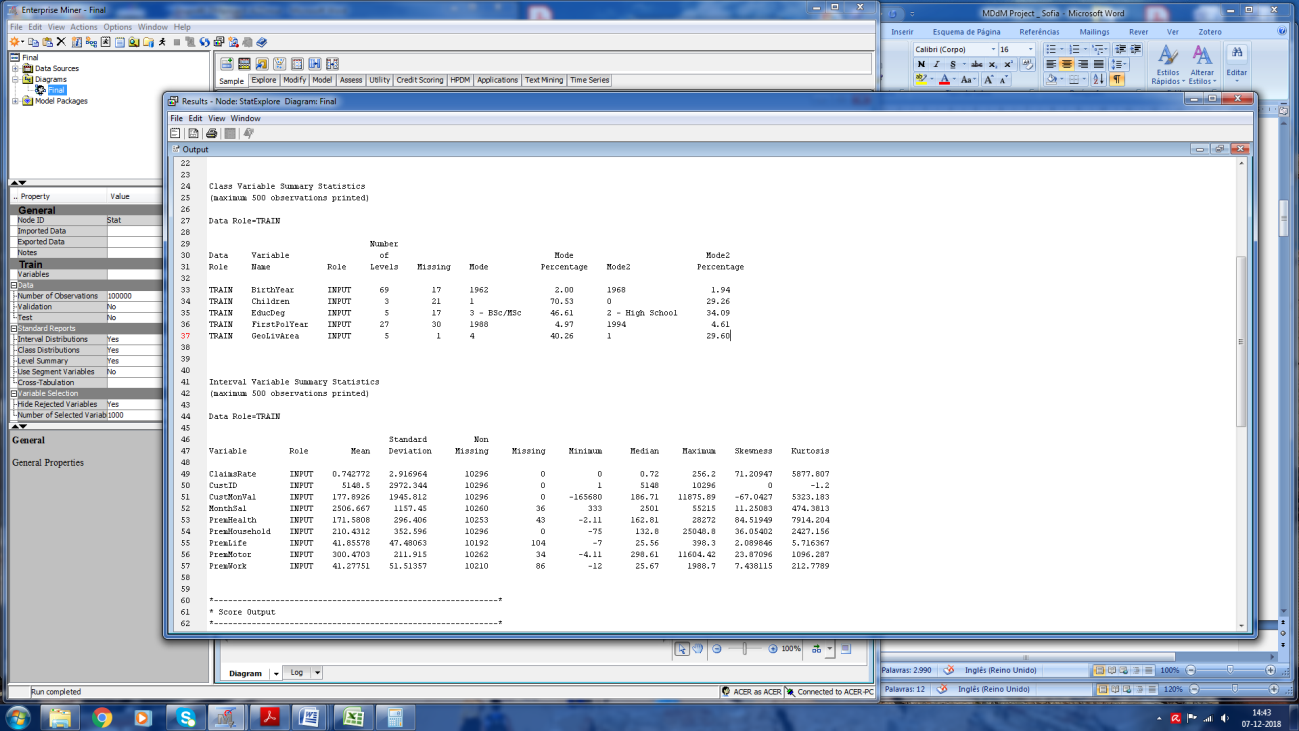
# Annexes



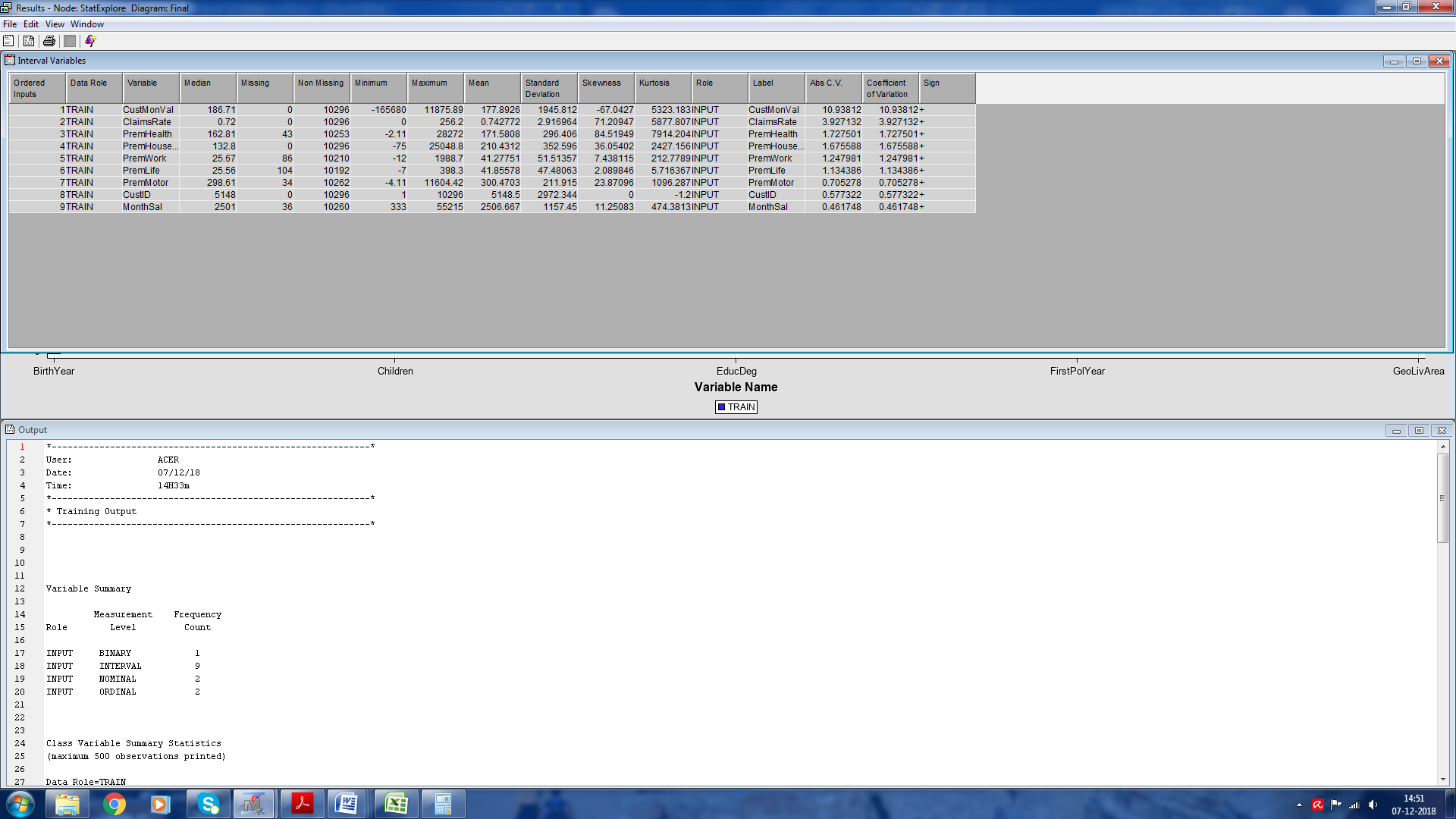
## Annex : Project Diagram



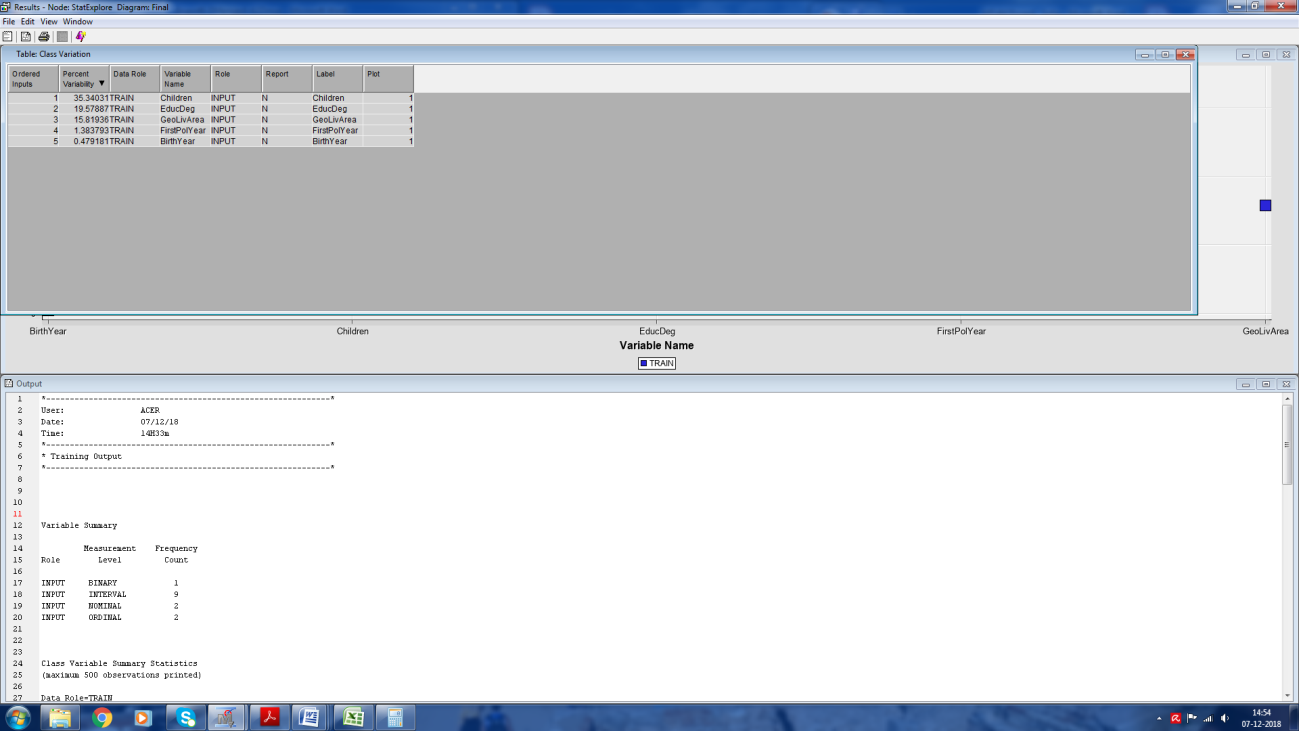
## Annex : Initial variables



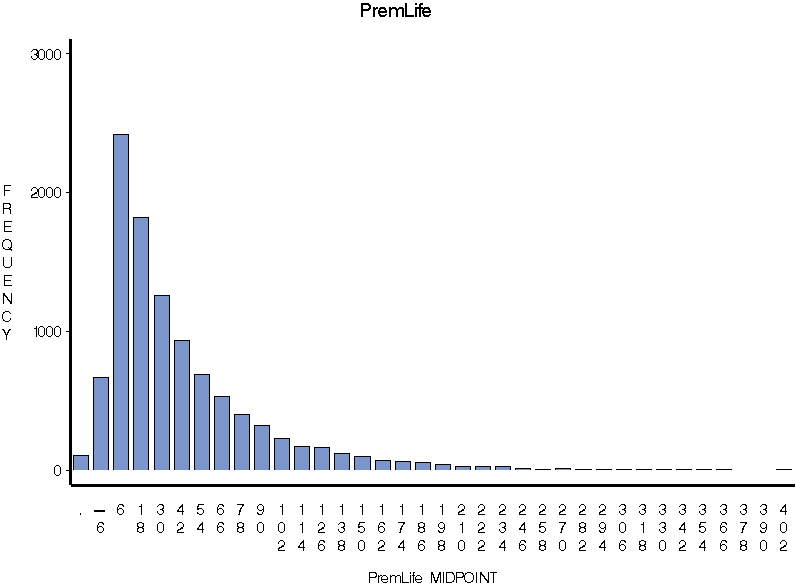
## Annex 3: Variable Summary (StatExplore 1)



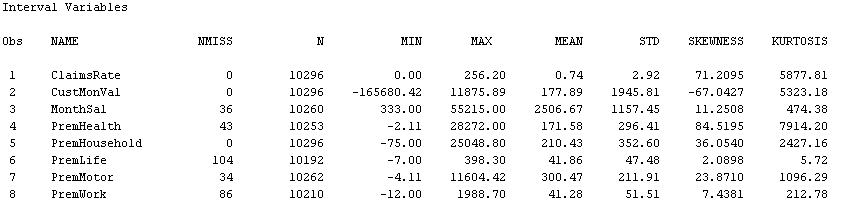
## Annex 4: Interval Variables (StatExplore 1)



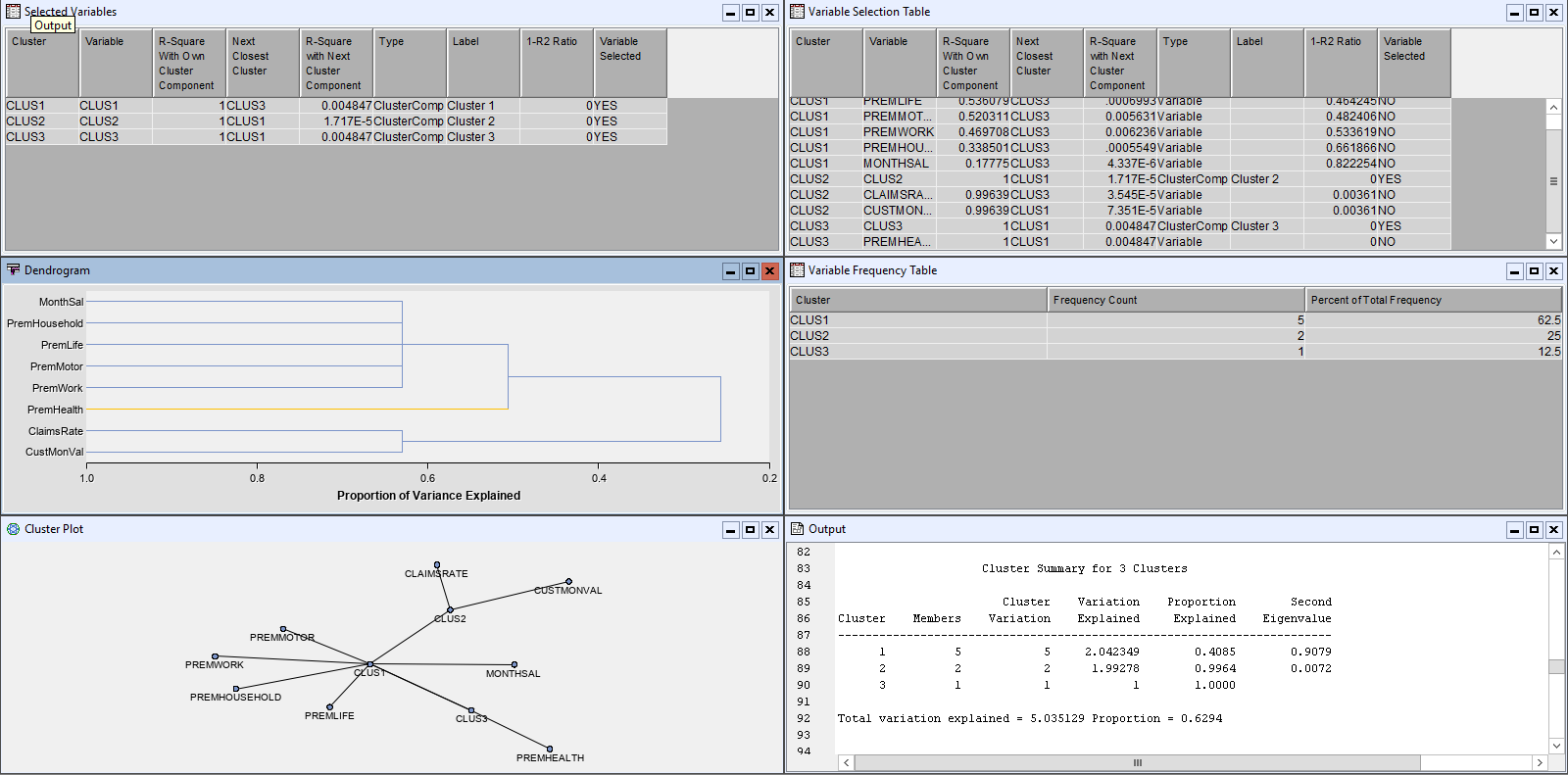
## Annex 5: Class Variables' Variation (StatExplore 1)



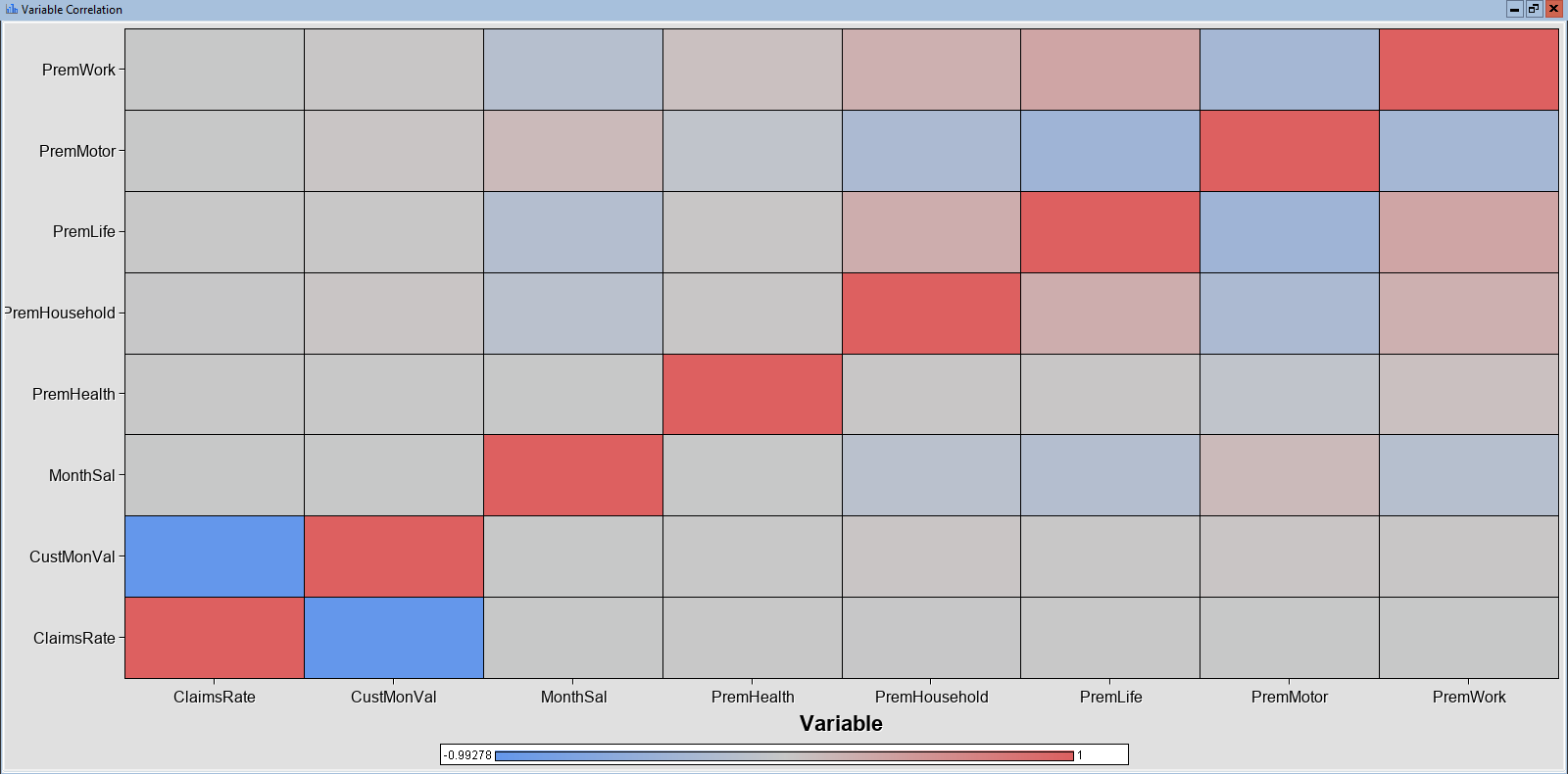
## Annex 6: Histogram from Multiplot (1) – Variable “Life Premium”



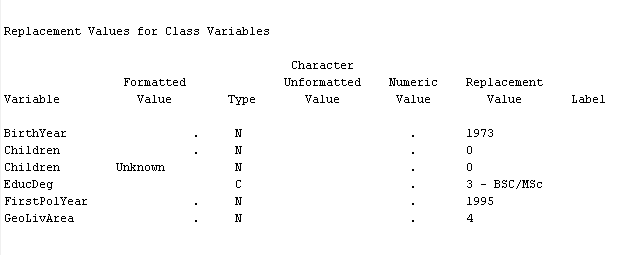
## Annex : Statistical information (Multiplot 1)



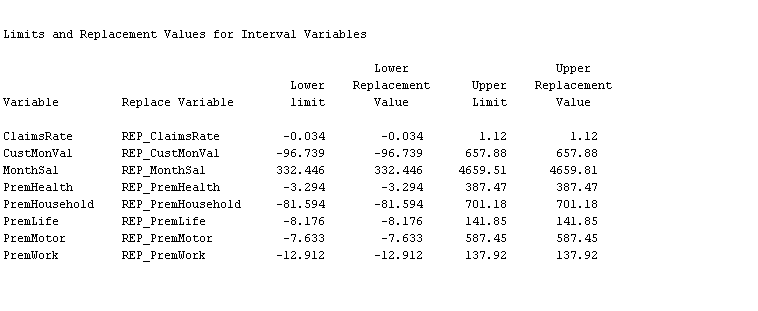
## Annex : Variable Clustering (1)



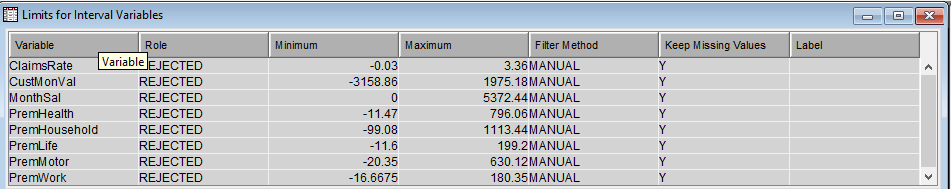
## Annex 9: Variable Correlation (Variable Clustering 1)



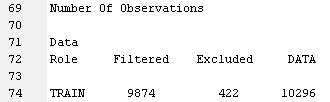
## Annex 10: Replacement Values for Class Variables (Replacement)



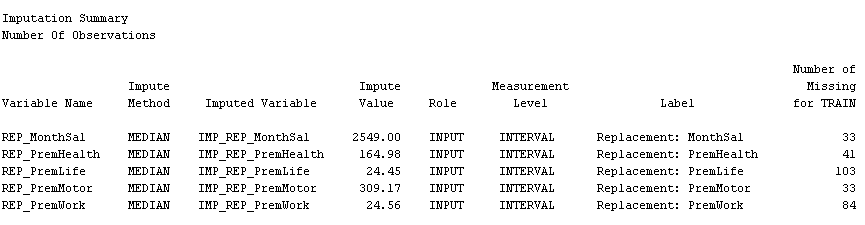
## Annex 11: Limits and Replacements for Interval variables (Replacement)



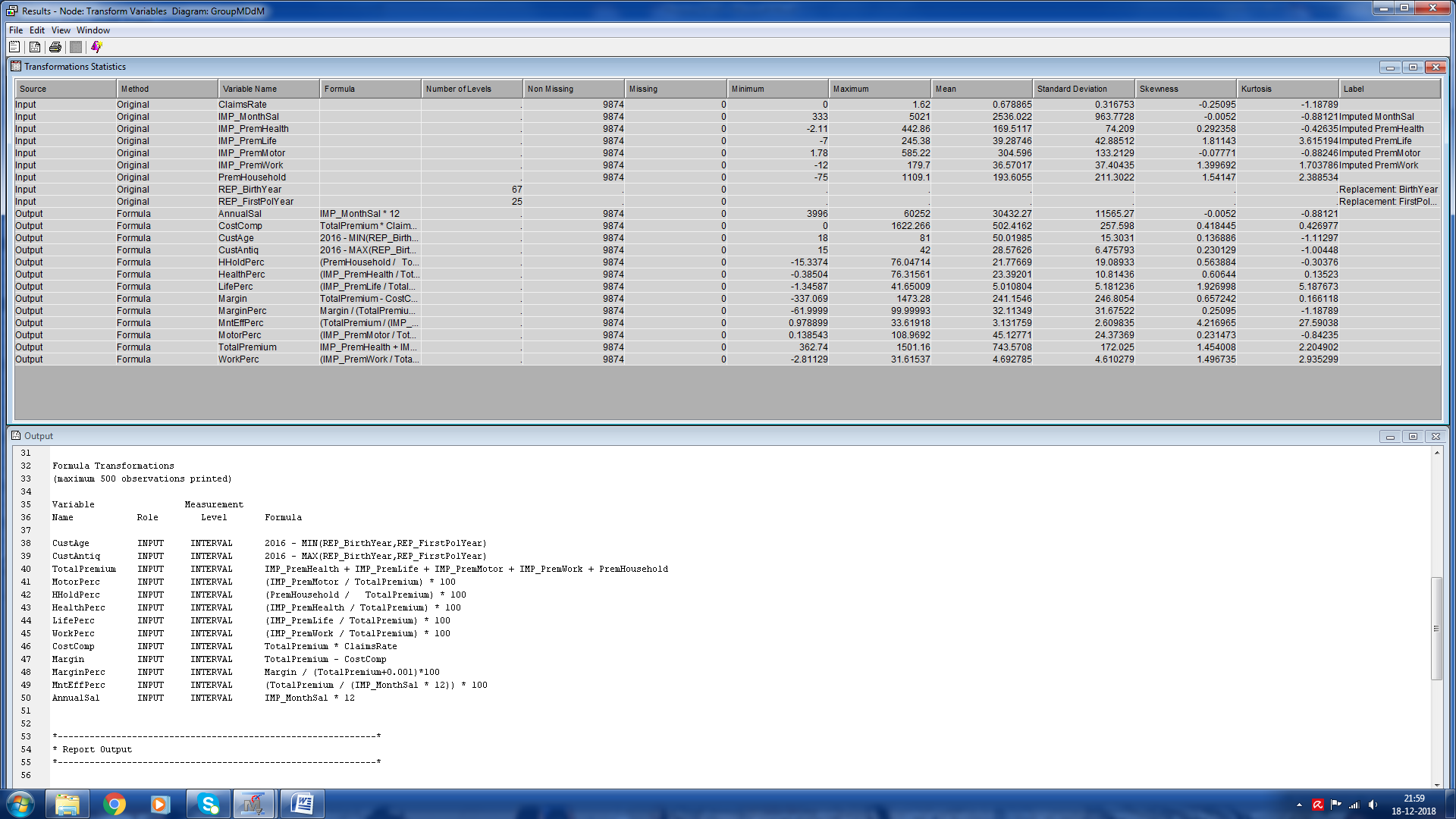
## Annex 12: Filter Limits for Interval Variables (Filter 1)



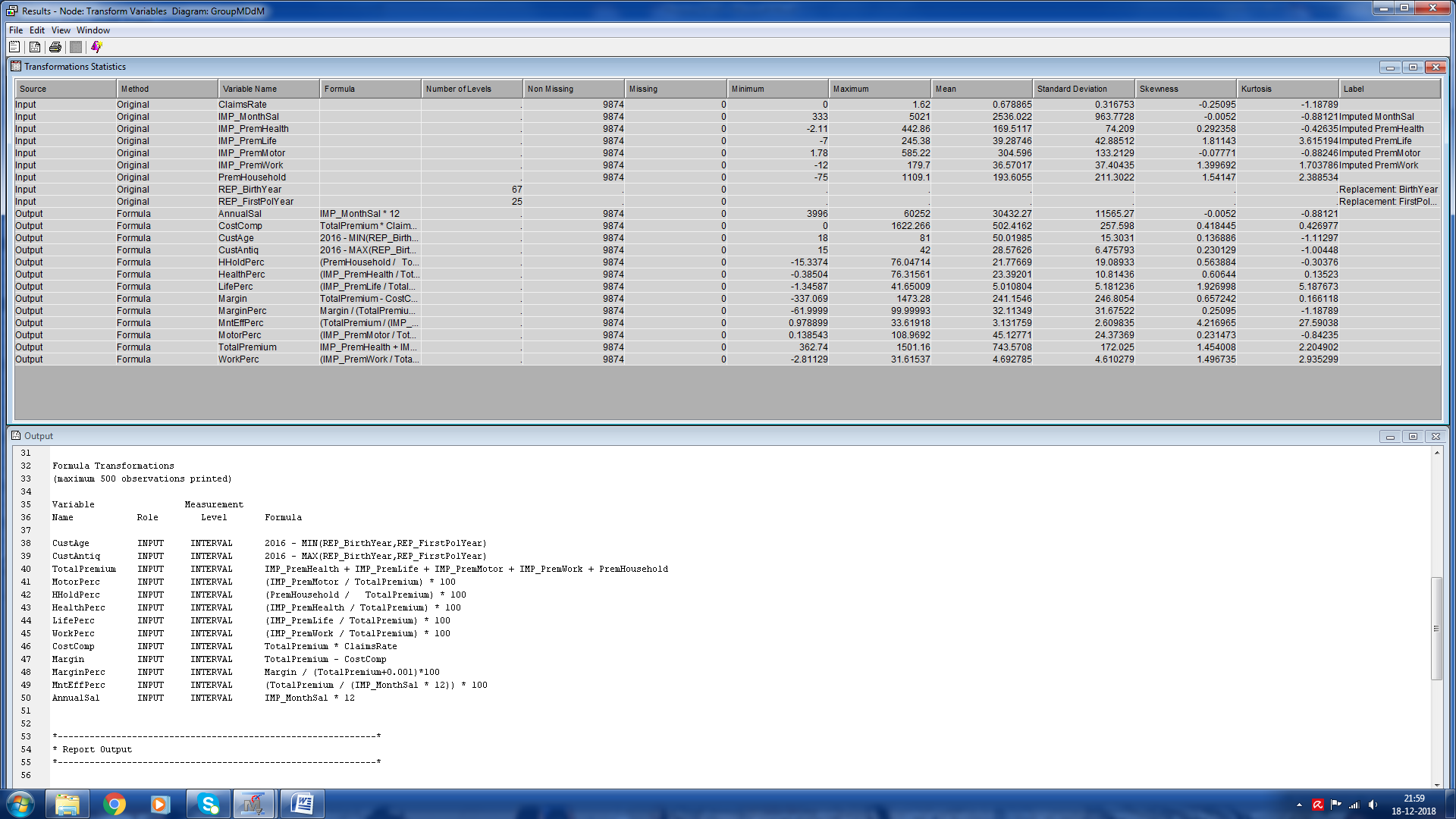
## Annex 13: Number of Excluded Observations (Filter 1)



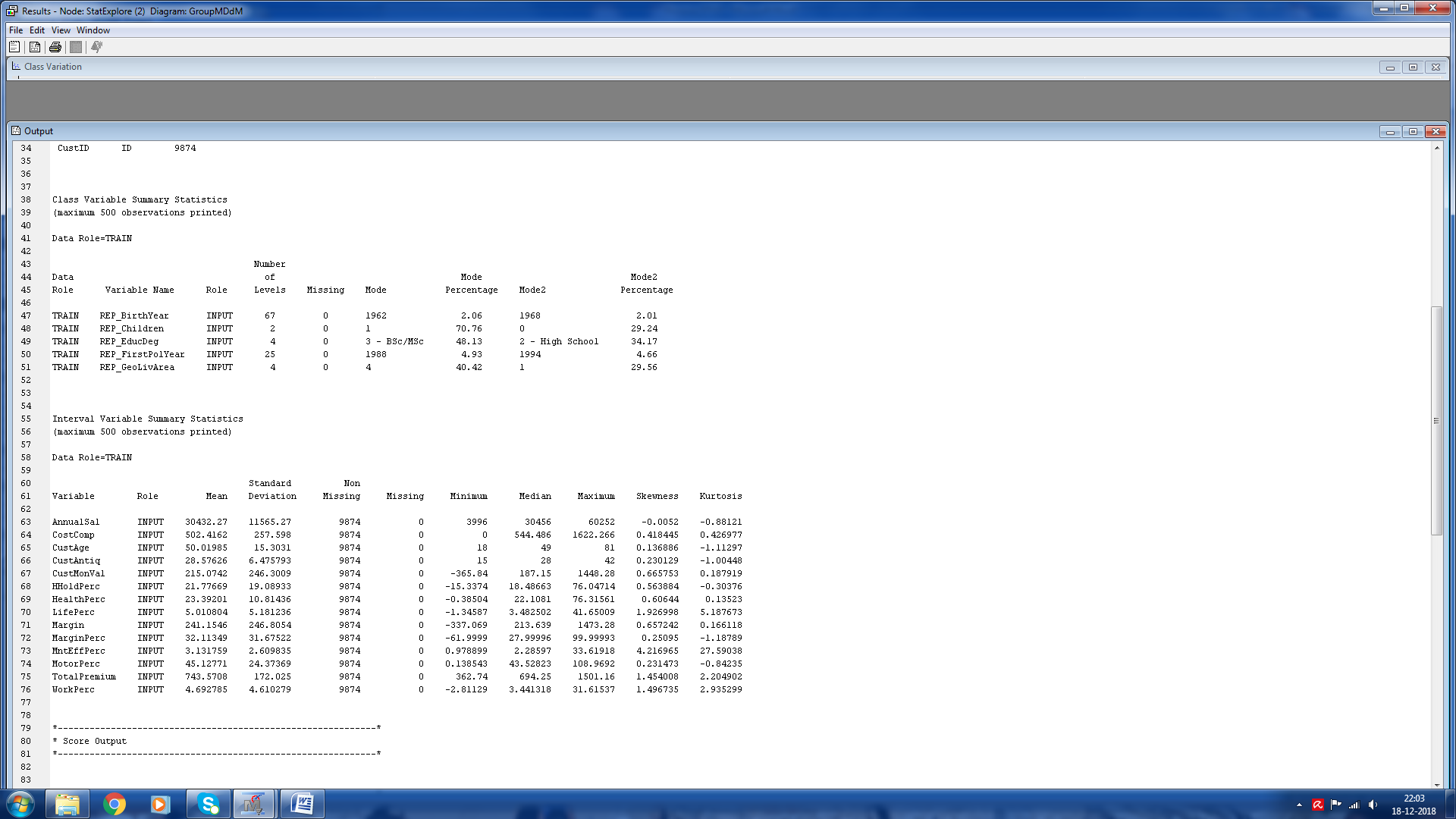
## Annex : Impute (1)

****

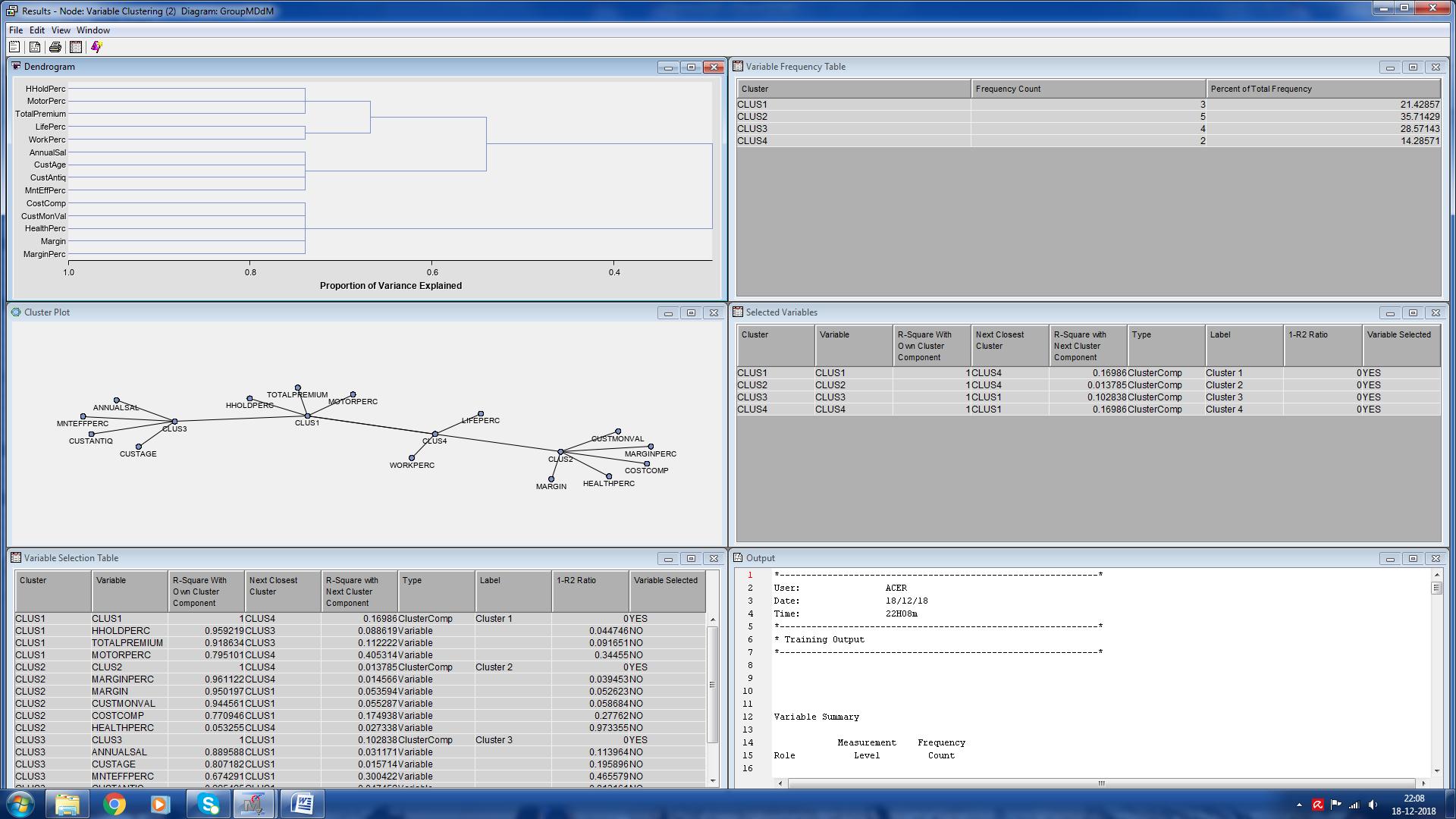
## Annex 15: Transformation of variables (Transform Variables)

****

## Annex 16: Transformation Statistics (Transform Variables)

****

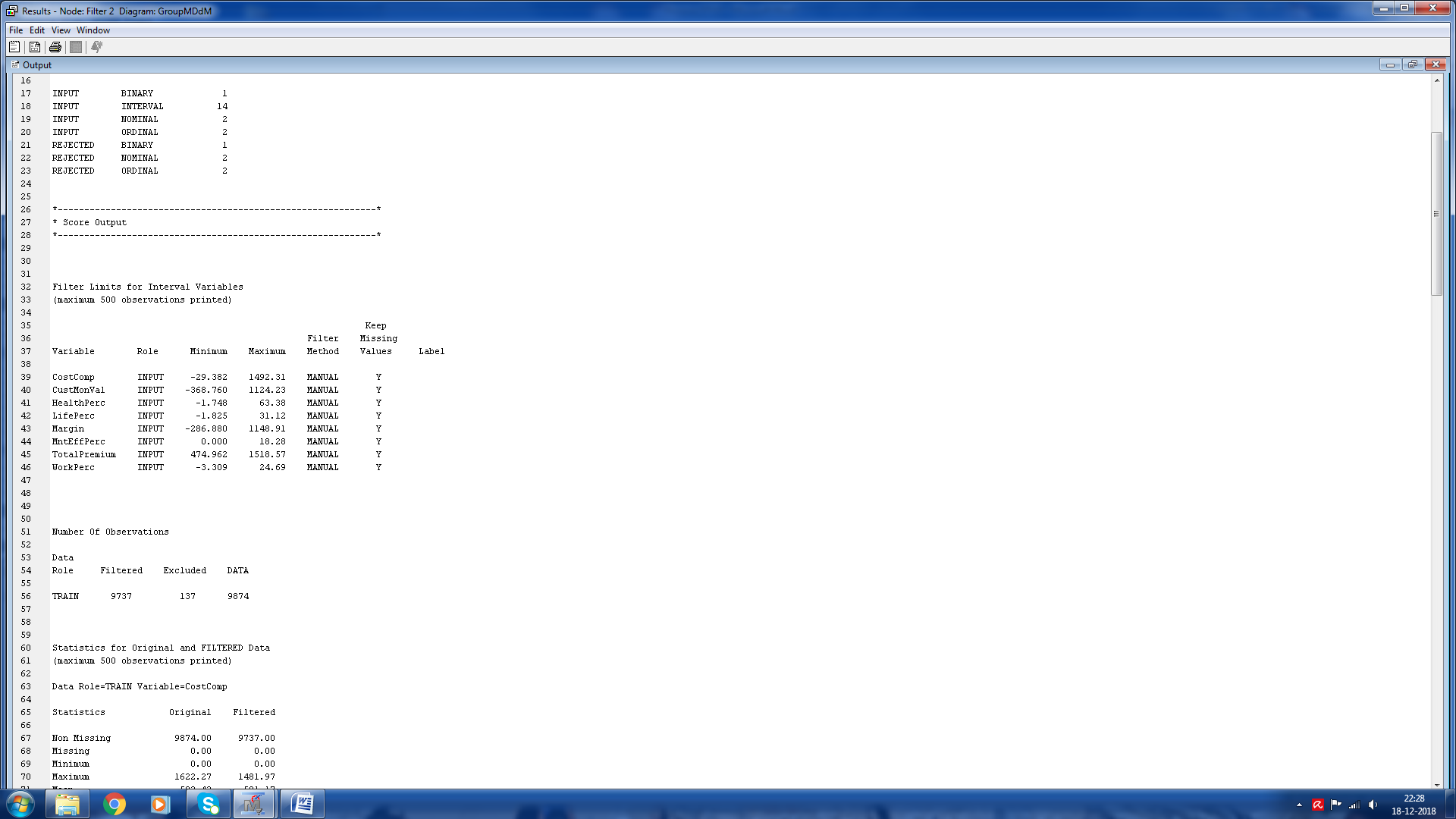
## Annex 17: Variables after transformation (StatExplore 2)



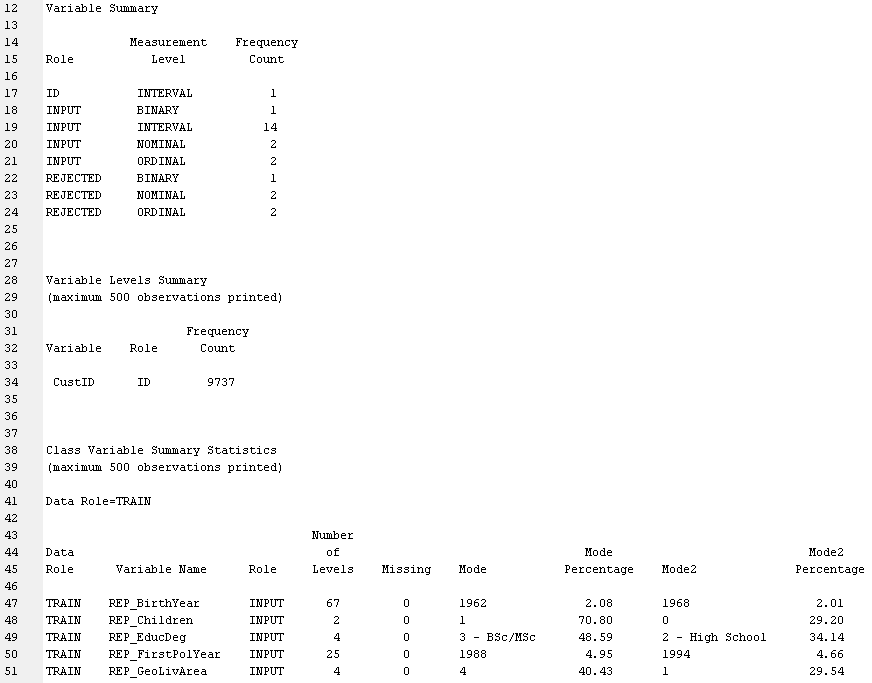
## Annex 18: Variable Clustering after transformation (Variable Clustering 2)

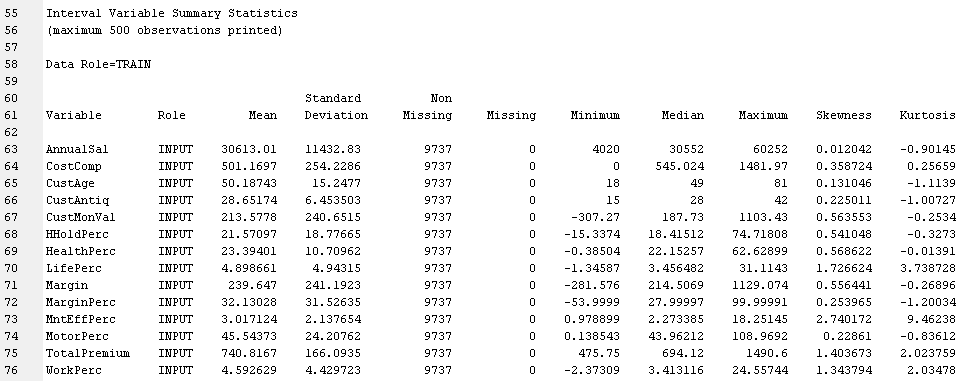


## Annex 19: Variable Correlation after transformation (Variable Clustering 2)

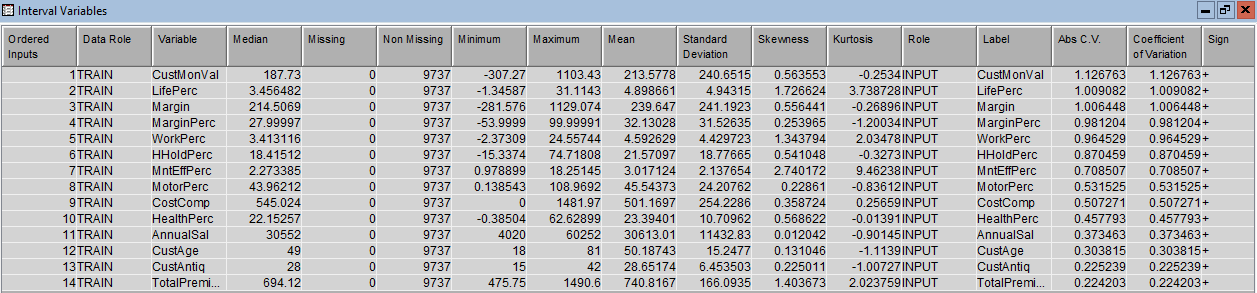


## Annex 20: Filter limits and excluded observations after transformation (Filter 2)

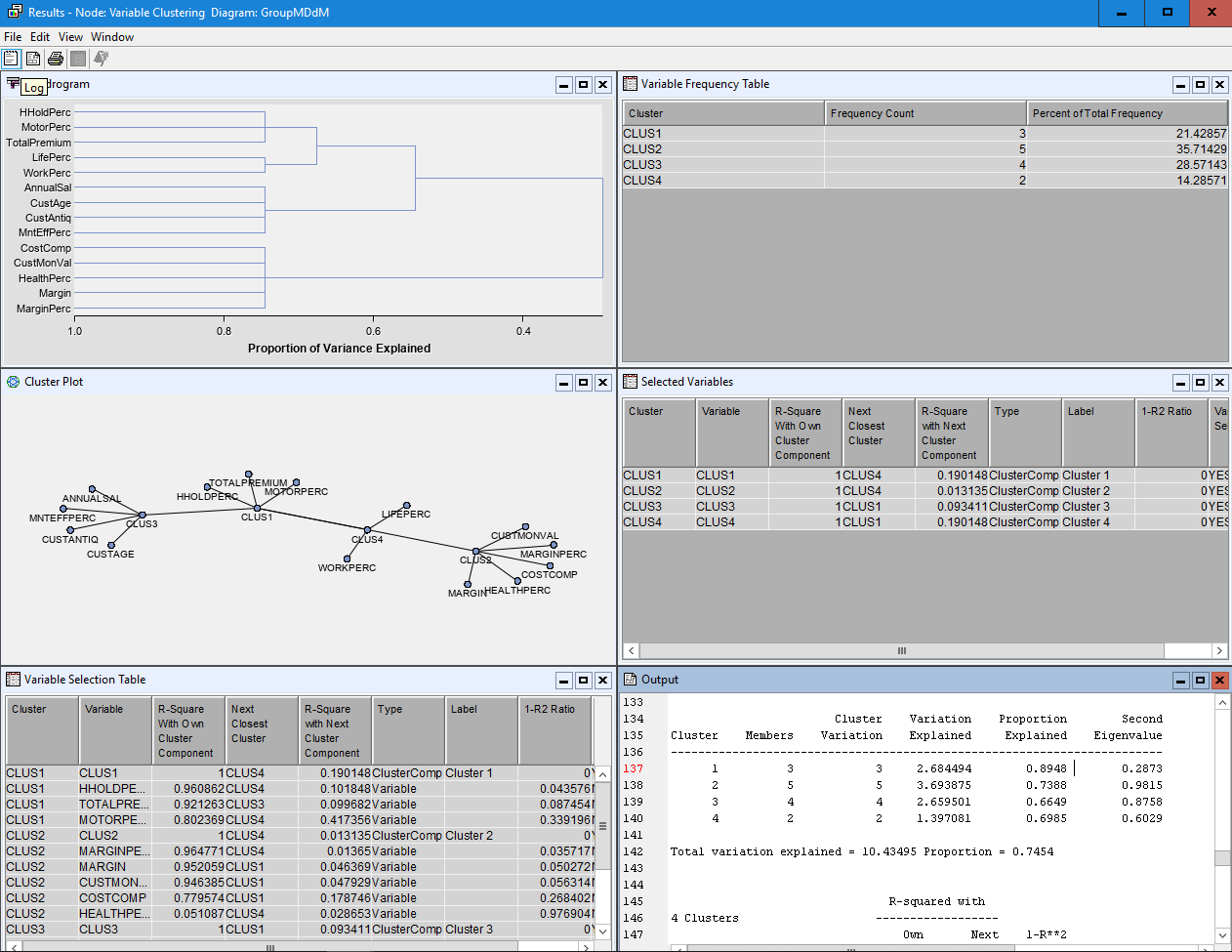




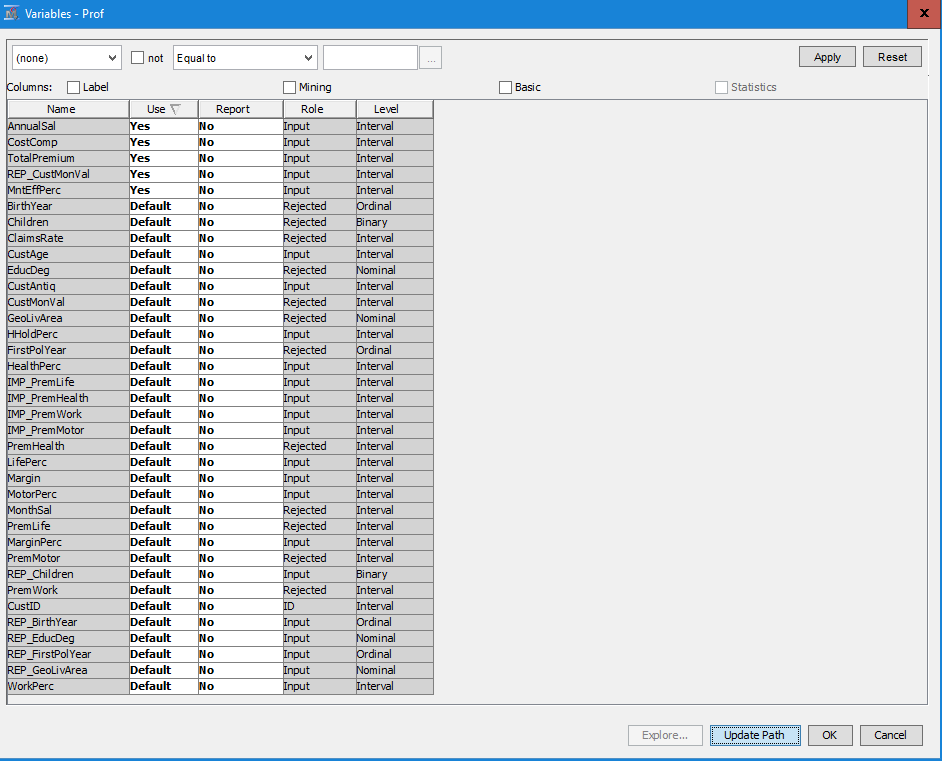
## Annex 21: Variables after transformation (StatExplore 3)



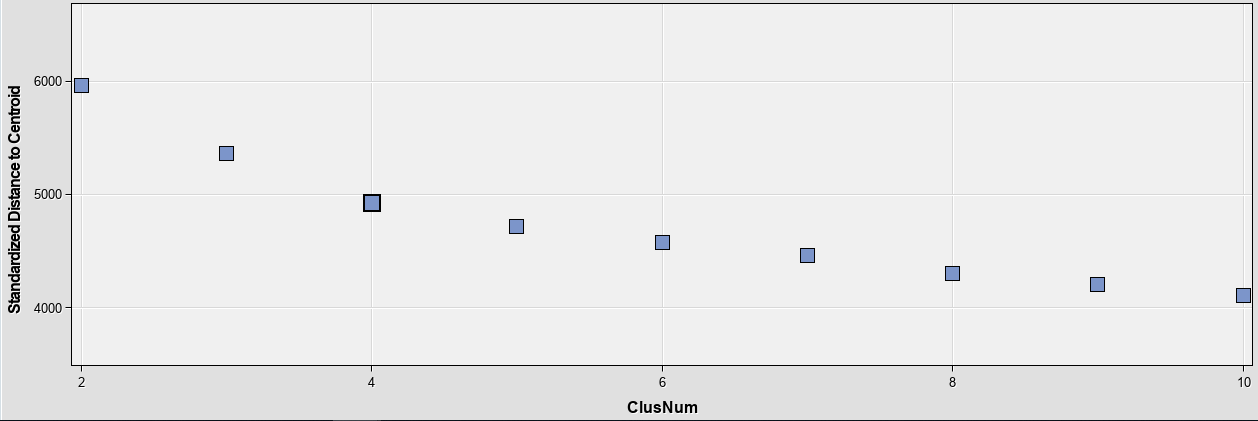
## Annex 22: Interval variables after transformation (StatExplore 3)



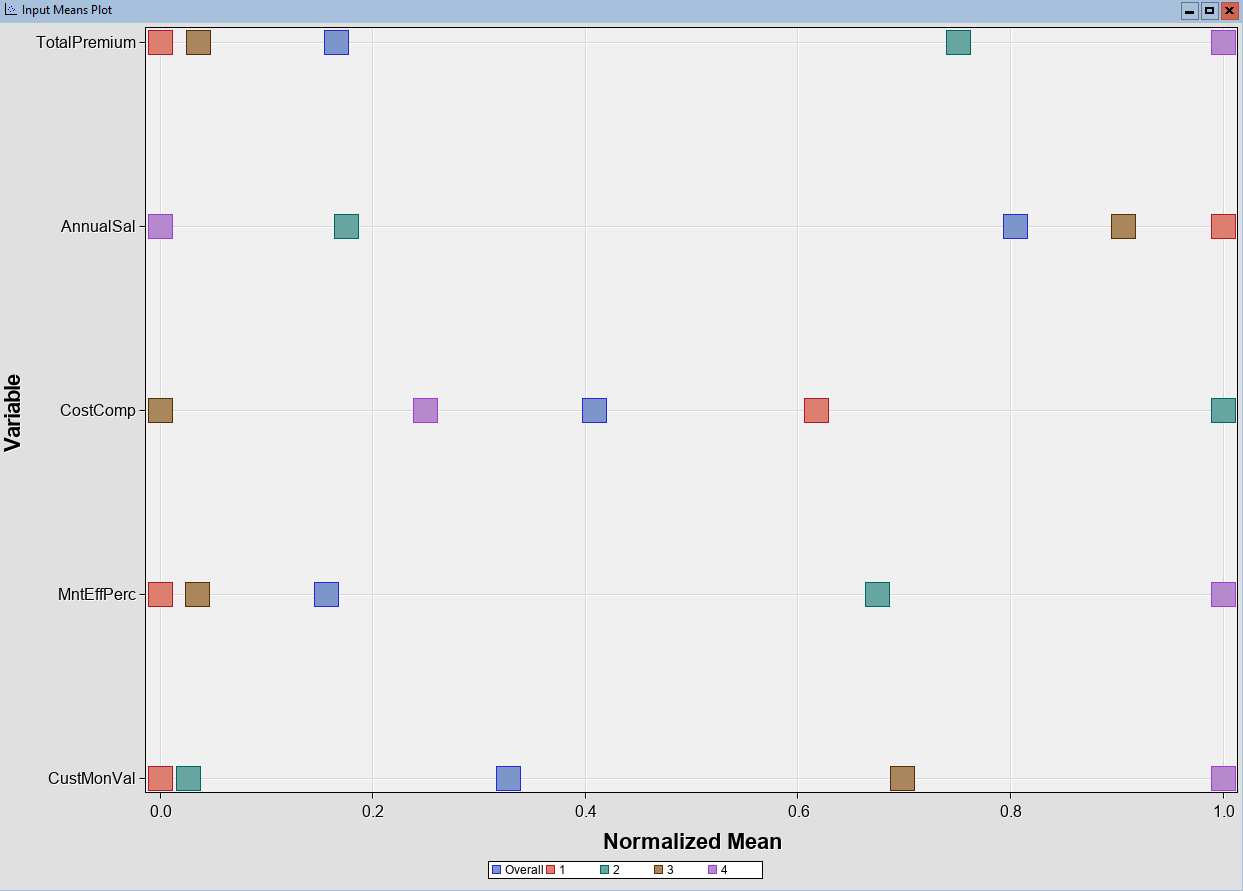
## Annex 23: Variable Clustering after transformation (Variable Clustering 3)



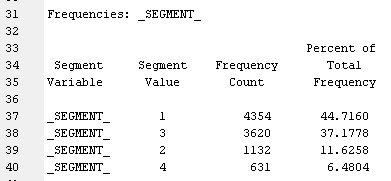
## Annex 24: Variables for value segmentation (Value Cluster)



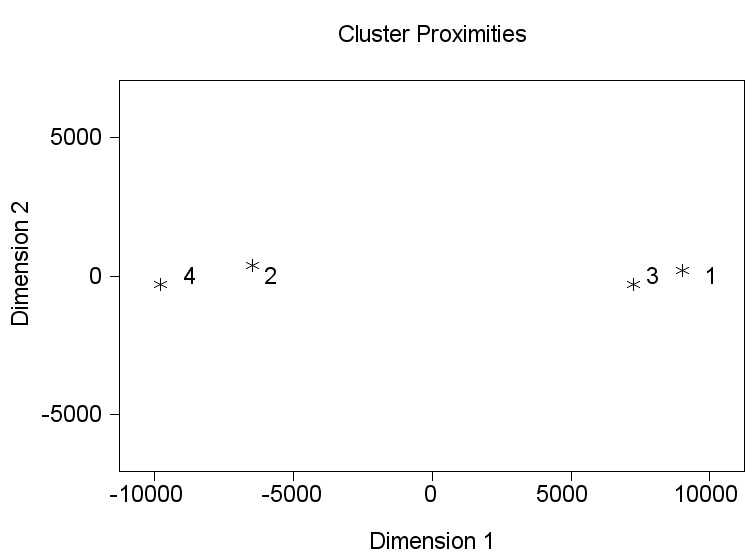
## Annex 25: Value Cluster – Elbow Graph



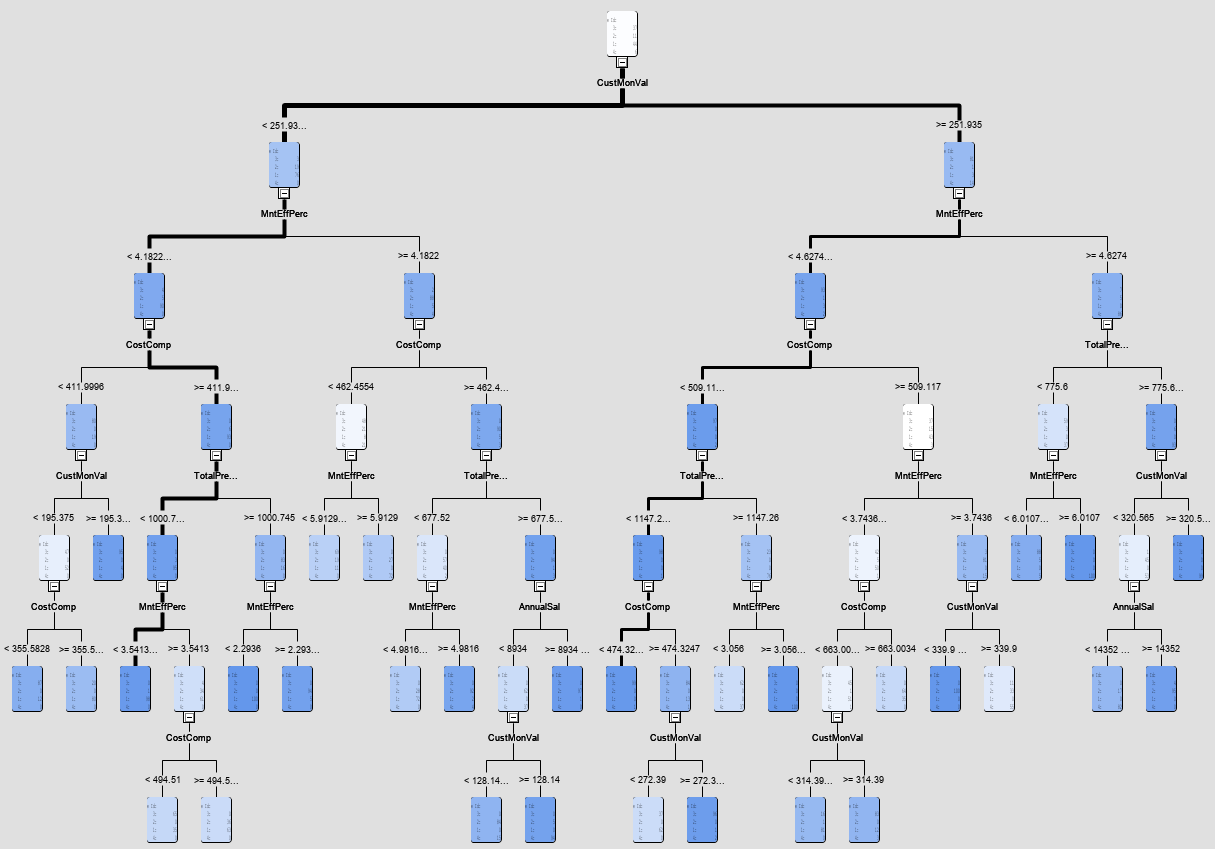
## Annex 26: Imput means plot – Value Cluster (k = 4)



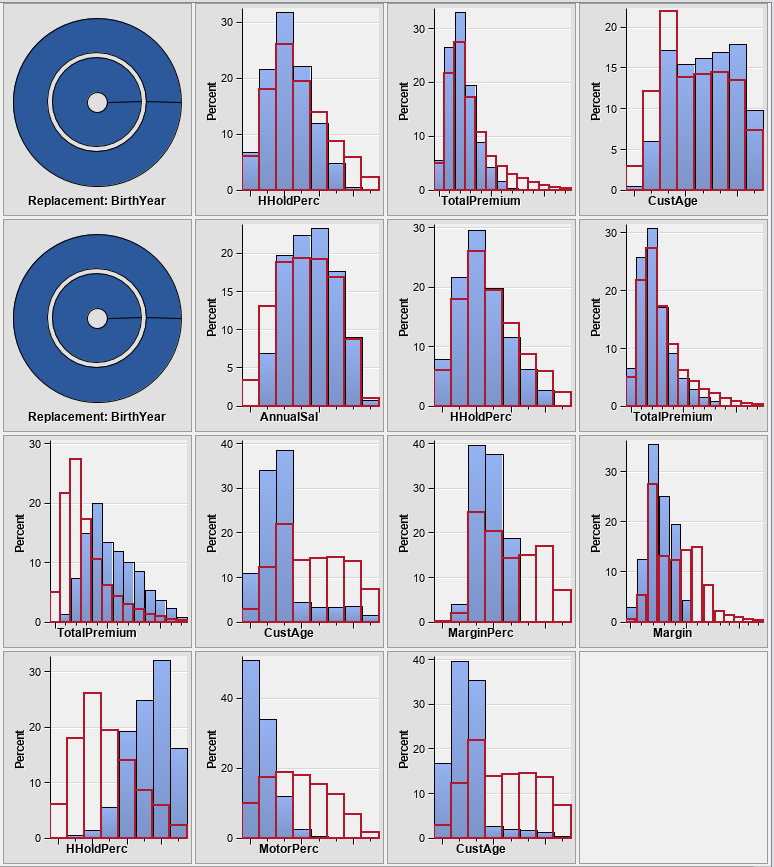
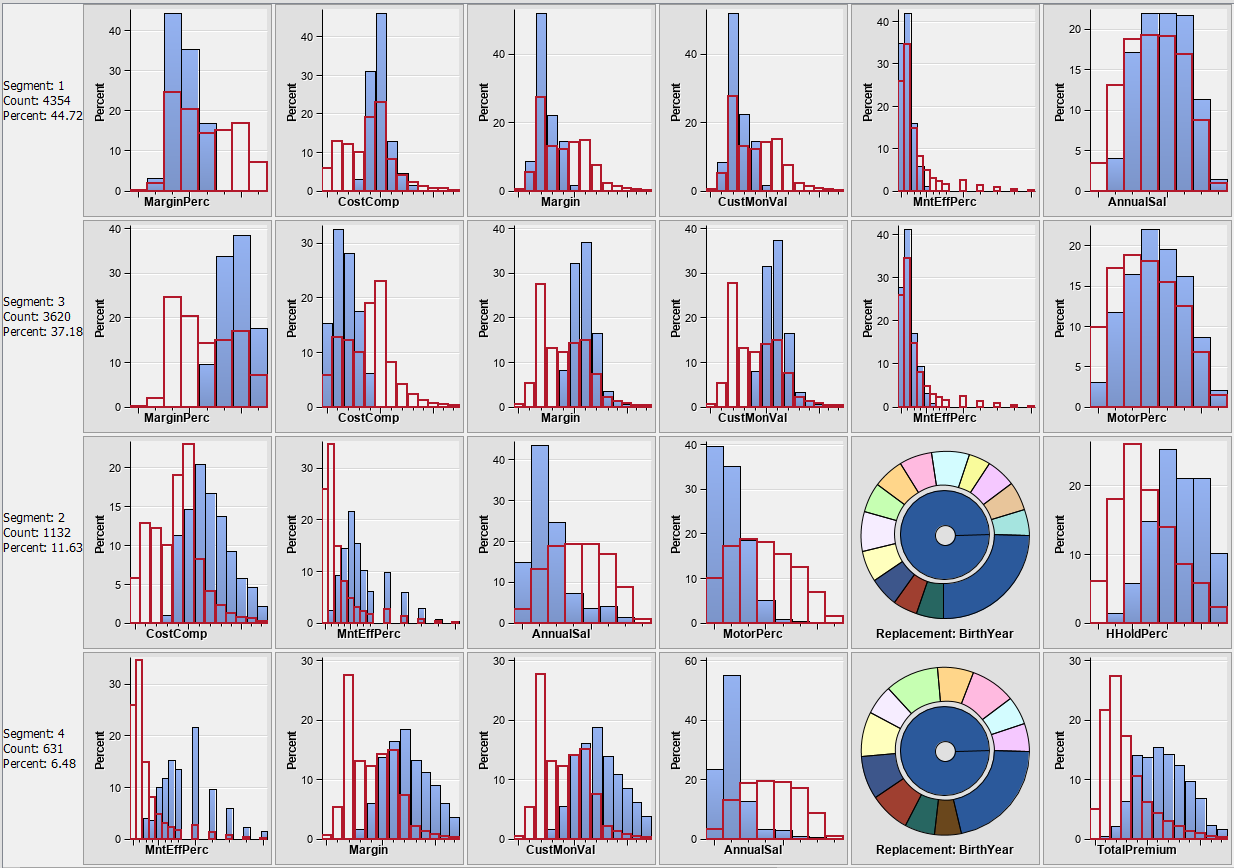
## Annex 27: Segment frequencies (Value cluster)



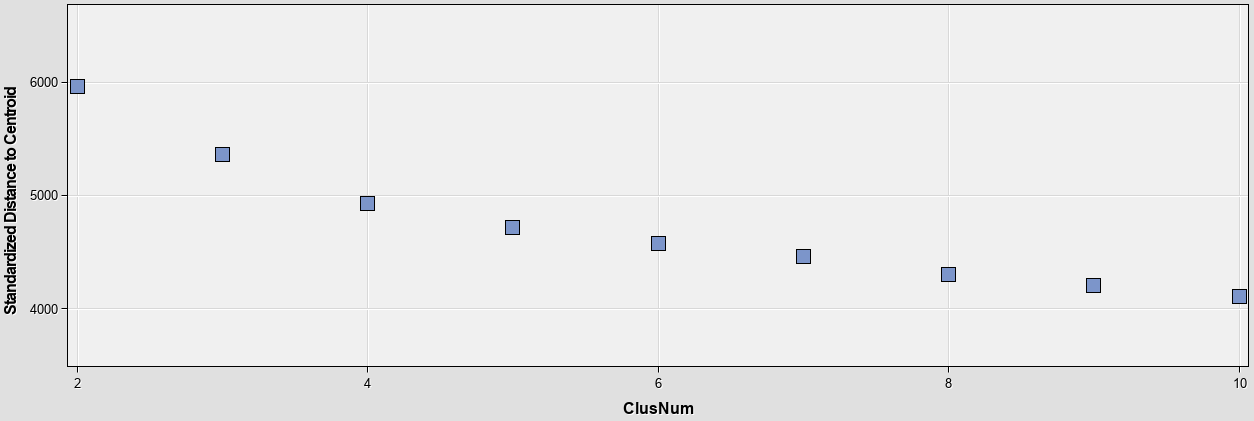
## Annex 28: Distance plot (Value cluster)



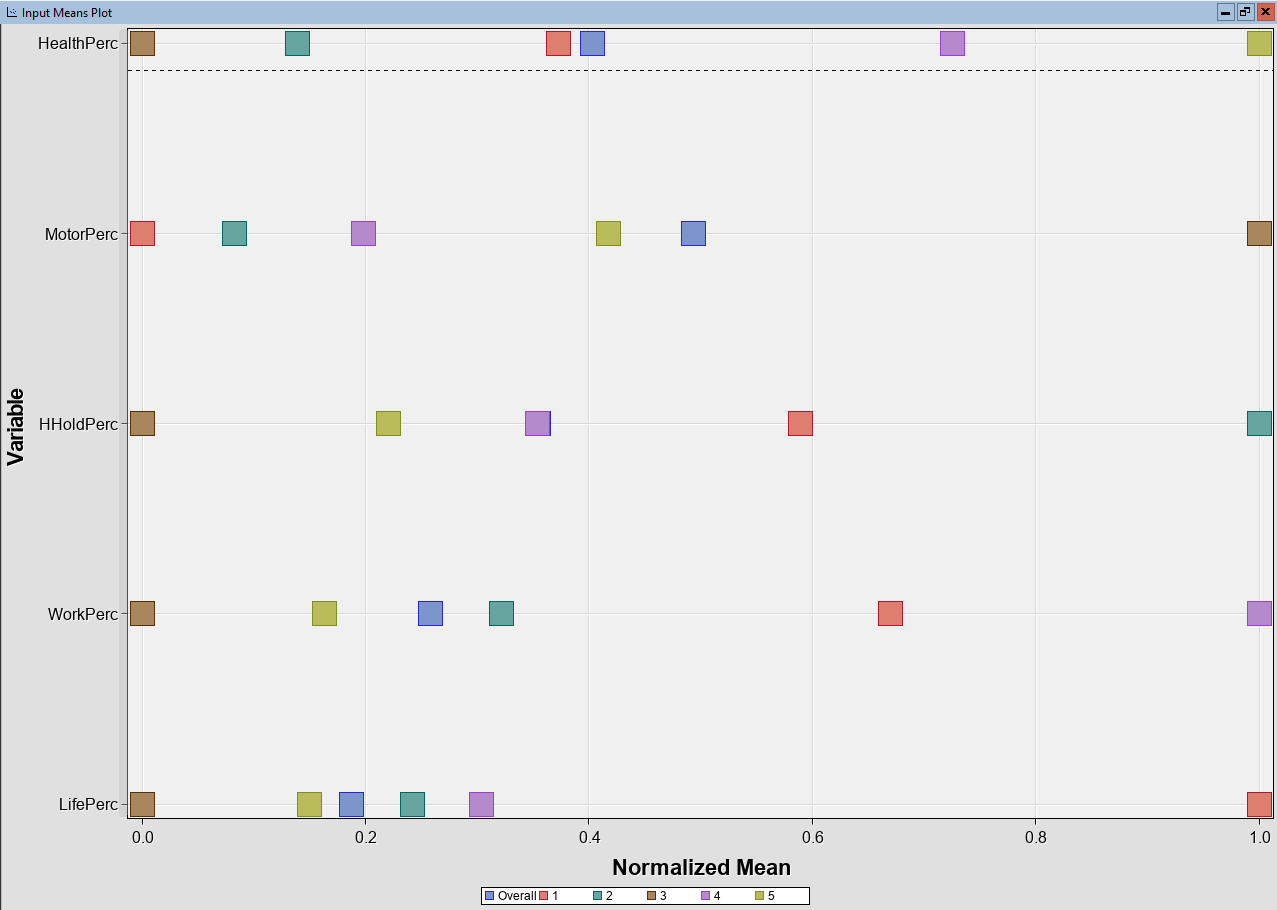
## Annex 29: Tree profile (Value cluster)



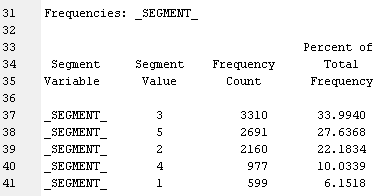
## Annex 30: Segment profiles (Value cluster)



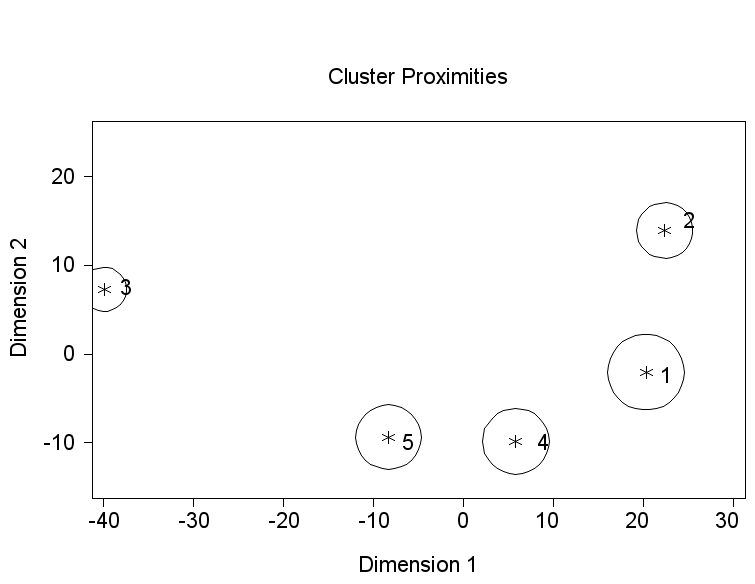
## Annex 31: Product Cluster – Elbow Graph



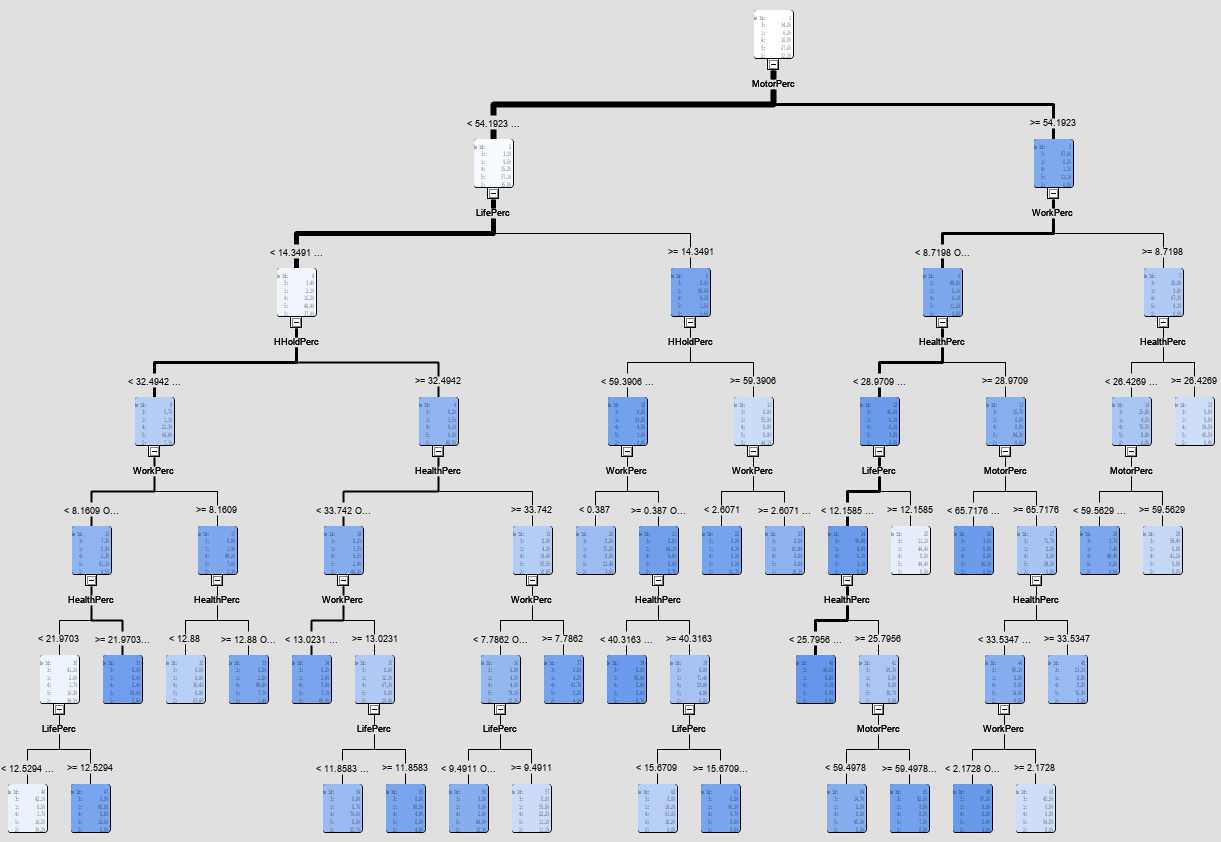
## Annex 32: Imput means plot – Product Cluster (k = 5)



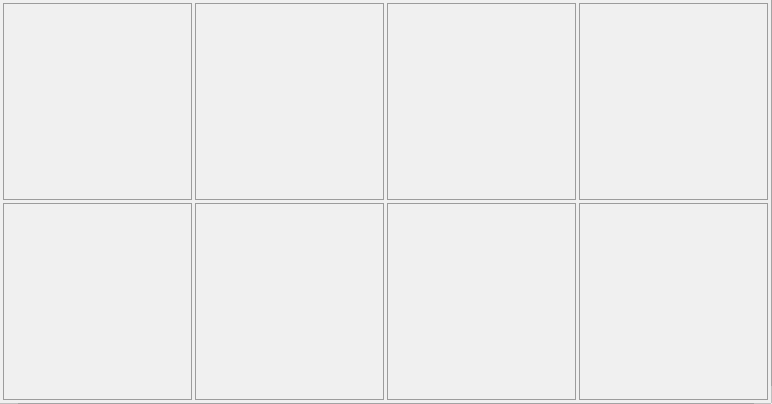
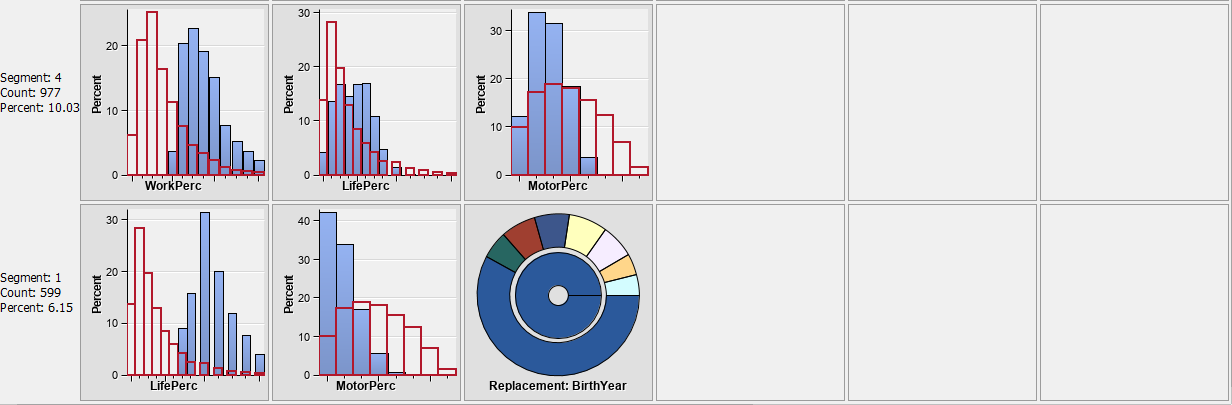
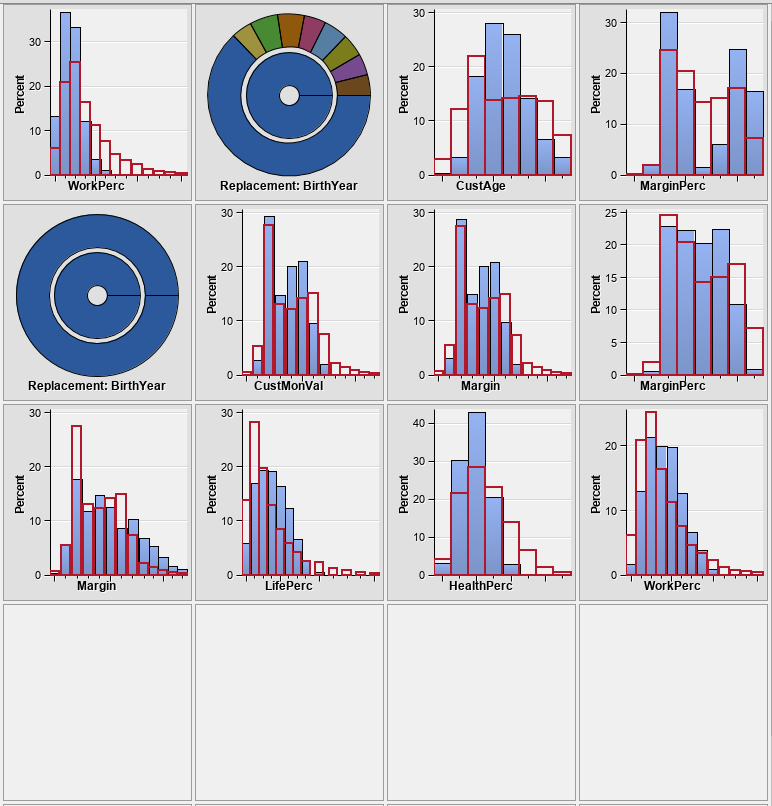
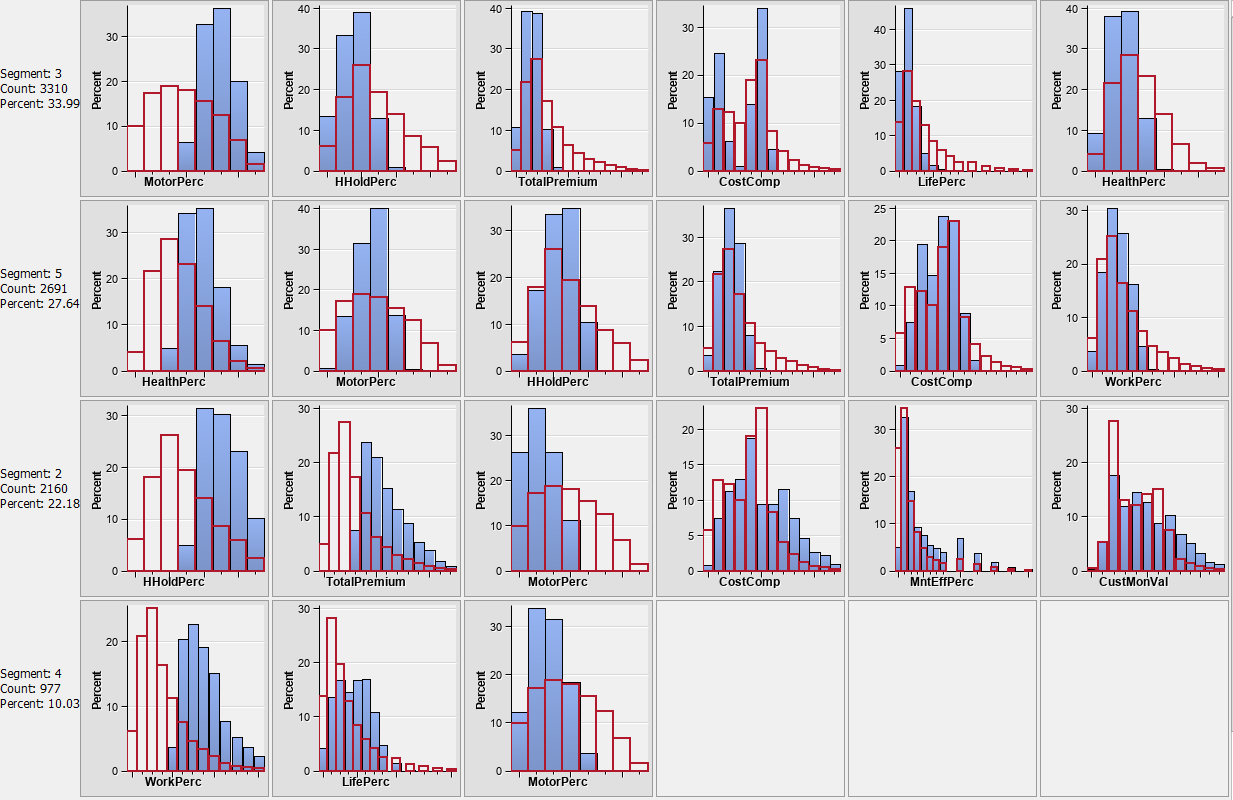
## Annex 33: Segment frequencies (Product Cluster)



## Annex 34: Distance plot (Product cluster)



## Annex 35: Tree profile (Product cluster)



## Annex 36: Segment profiles (Product cluster)