2018

LAGOSTE

SEGMENTATION

Ana Rita Marques R2016700
Ana Sofia José R2016718
Catarina Neves R2016723
Joana Neves R2016724
Pedro Alves R2016734

Contents

Tables

1. Executive Summary

The purpose of this project was to identify groups of customers with distinct characteristics, that are as similar as possible to each other, and as different between groups as possible. This way the company can get a sense of the type of customers, have a perspective of what part of its business is attractive to each group, which ones have the most value, what they buy and their socio-economic characteristics. But all this information is important so that the organization can achieve the most important goal - customer satisfaction. This can be achieved by conducting targeted marketing campaigns and meeting the needs and desires of our customers.

For the identification of the clusters, we used two different techniques in order to reach the best solution. One was the K-Means Algorithm, and the other clustering technique used was Self Organizing Maps (SOM), that will be explained later in the report.

Thus, in this project 27 clusters were initially identified, using 3 perspectives - place of purchase, value and product usage. A total of 9 major groups of clients were initially found, taking into account their value - gold (high value - high income, high expenditure), silver (medium value - medium income, medium expenditure) or bronze (low value – less income, less expenditure) - the product usage (product where the clients spent more) - rackets, sneakers, or others (fashion) - and the place of purchase (the preferential place of consumption) – web, catalog, or store.

More specific groups were found by crossing these three views, creating, like it was said, a total of 27 groups. However, many of these 27 groups contained very few clients, thus, not valuable in terms of marketing to target them separately through specific marketing campaigns. So, after identifying 12 major groups, the clients that belonged to the smaller groups were allocated to one of the other 12 major groups using a predictive model. Finally, 12 marketing campaigns were suggested based on the characteristics of each group (generally groups of more than 200 clients).

With this project, the company can more adequately satisfy each client, as well as have information about them, and thus enhance the business.

2. Introduction

This project has the main objective of splitting LaGoste customers into groups that are internally homogeneous and different between them and consistent in time and response to stimulus This partition proves to be very important, especially for the development of new products, creation of different marketing campaigns directed for each segment. This way, LaGoste can adopt strategic moves in the market, taking into consideration the characteristics of each group of its customers.

The customer segmentation will help the organization to tailor the marketing approach in order to have the best “effect” on each group. This is not only good for LaGoste, allowing it to reduce costs and achieve higher profits, by directing the adequate products and campaigns to the right customers, but allowing clients to receive less offers, but targeted at products/services that they are more willing to want/accept/buy through the proper channels.

3. Methodology

With the purpose of segmenting the set of LaGoste customers (5000), we started by a statistical analysis of the data set, treating the outliers and missing values. Then, following the SEMMA process, we checked if there were incoherences in the data and corrected some values. We were then able to create new or transformed variables and compound variables in order to obtain ones with greater explaining power, and reduce dimensionality, while increasing model explainability.

At this moment, we were able to choose perspectives on which to try and segment our clients. We selected Products purchased, to see if there are different groups that buy different product ranges, or if they all buy the same type of products.

The next view was Customer value, to try and understand if we can clearly separate the most and the least important clients in terms of value, and how many groups could we clearly differentiate.

The last view selected was place of purchase, to understand how LaGoste clients prefer to interact, where they do like to purchase, and if the channels are completely separate, or if clients use them interchangeably.

After having these major groups, we set out to understand if the views could be combined, in order to obtain specific groups of individuals knowing their characteristics in the combined views.

When combining the different views we arrive at 27 clusters, some of them with very few clients in it. A cluster reduction was in order. We identified the clusters to remove and the method to reassign them.

We then went on to profiling each of the clusters in the overview by perspective, and each of the final clusters in the final view.

At the end, we were ready to suggest marketing campaigns to each of the final clusters, taking into consideration their total profiles.

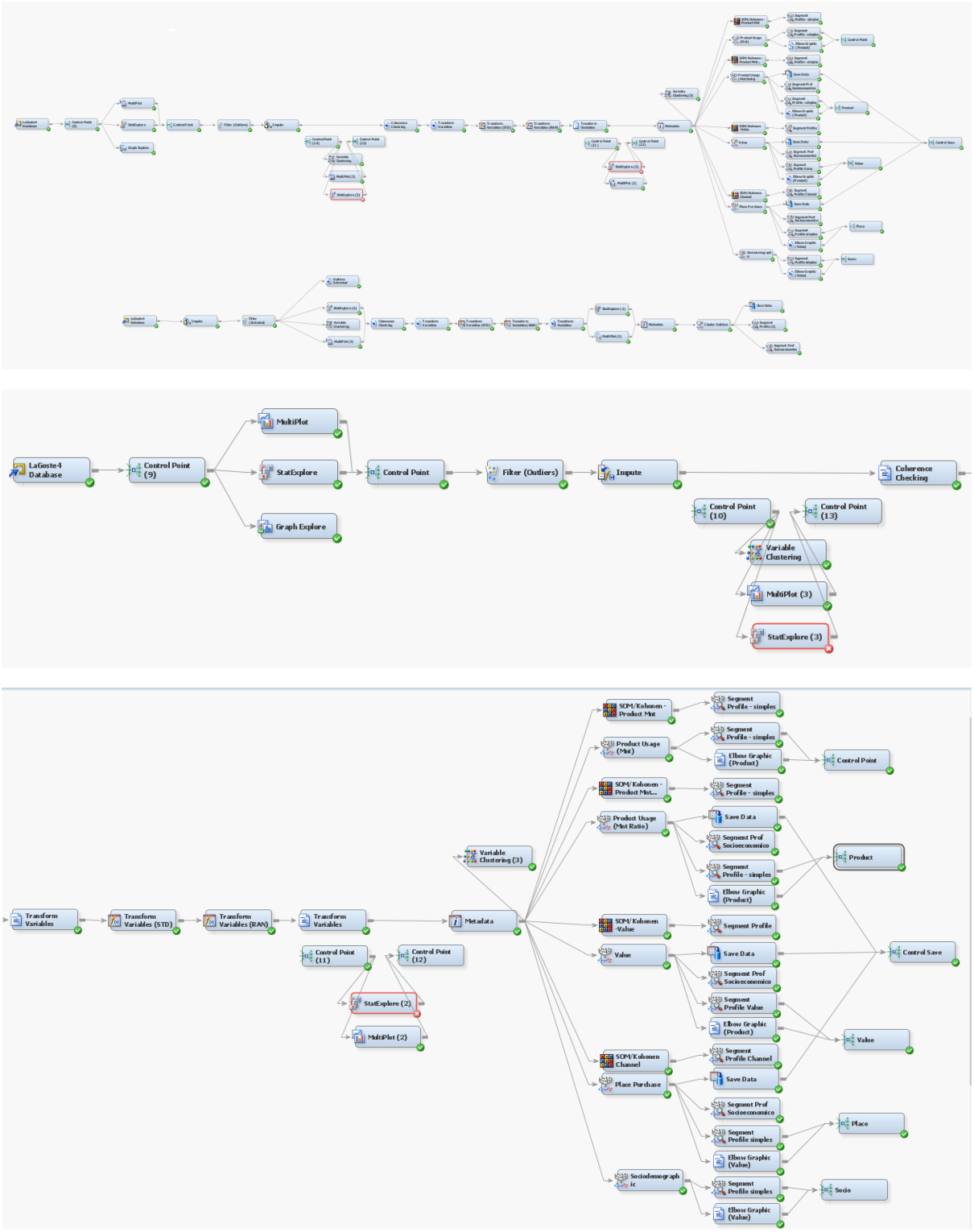

Figure 1 - Process in SAS Miner

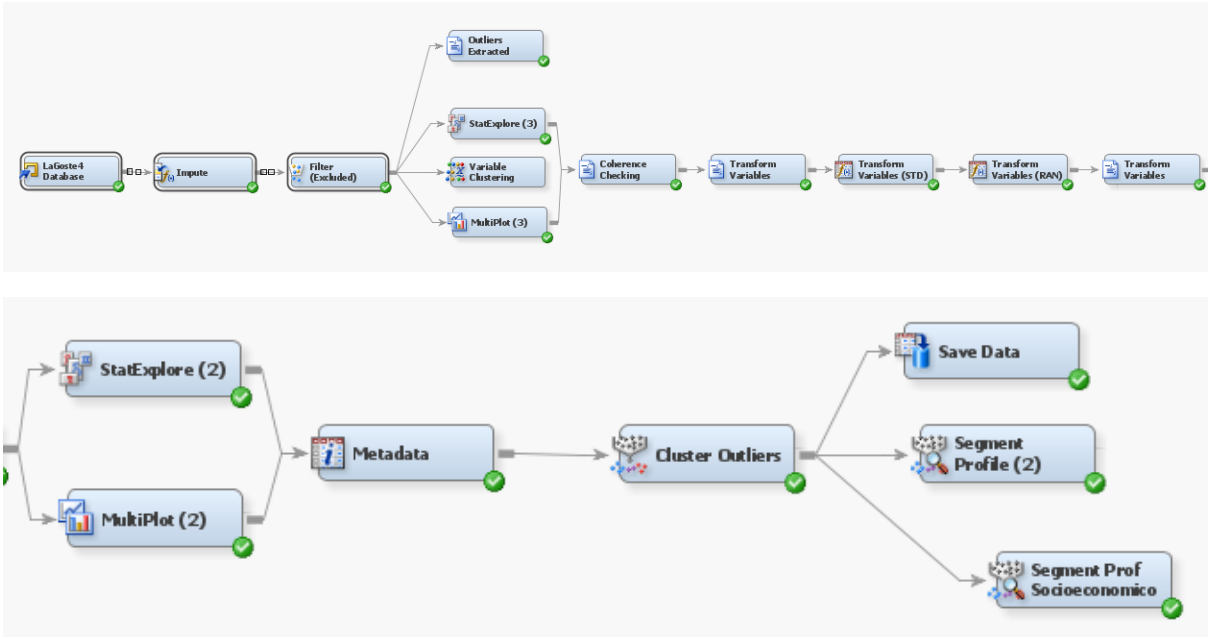

Figure 2 - Process in SAS Miner to Outliers

4. Sample

4.1 Variables Description

At first, we started by defining the role and level for each variable in the dataset. In the role, we set as INPUT the variables used for segmentation modelling and ID for the Custid variable, that has an unique value for each observation, identifying it. At the level setting, we defined it for each variable accordingly to their measuring scales (scales used were nominal, binary and interval). Thus, we started to understand the available data.

The following table identifies the role, level and description of all the variables from the available dataset.

NAME	ROLE	LEVEL	DESCRIPTION
AcceptedCmp1	INPUT	BINARY	Flag indicating customer accepted offer in campaign 1
AcceptedCmp2	INPUT	BINARY	Flag indicating customer accepted offer in campaign 2
AcceptedCmp3	INPUT	BINARY	Flag indicating customer accepted offer in campaign 3
AcceptedCmp4	INPUT	BINARY	Flag indicating customer accepted offer in campaign 4
AcceptedCmp5	INPUT	BINARY	Flag indicating customer accepted offer in campaign 5
Complain	INPUT	BINARY	Flag indicating if customer has complained (last 18 months)
Custid	ID	NOMINAL	Customer ID
Dt_Customer	INPUT	INTERVAL	Date of customer's enrolment with the company
Education	INPUT	NOMINAL	Level of education of Customer
Income	INPUT	INTERVAL	Yearly Income of household of Customer
Kidhome	INPUT	INTERVAL	Number of kids in household
Marital_Status	INPUT	NOMINAL	Marital Status of Customer
MntHats	INPUT	INTERVAL	Amount spent on Hats (last 18 months)
MntPremium_Brand	INPUT	INTERVAL	Amount spent on Premium material (last 18 months)
MntRackets	INPUT	INTERVAL	Amount spent on Rackets (last 18 months)
MntSneakers	INPUT	INTERVAL	Amount spent on Sneakers (last 18 months)
MntTShirts	INPUT	INTERVAL	Amount spent on Tshirts (last 18 months)
MntWatches	INPUT	INTERVAL	Amount spent on Watches (last 18 months)
NumCatalogPurchases	INPUT	INTERVAL	Number of purchases made through catalog (last 18 Months)
NumDealsPurchases	INPUT	INTERVAL	Number of purchases made with discounts (last 18 Months)
NumStorePurchases	INPUT	INTERVAL	Number of purchases made through store (last 18 Months)
NumWebPurchases	INPUT	INTERVAL	Number of purchases made through web (last 18 Months)
NumWebVisitsMonth	INPUT	INTERVAL	Average number of web visits a month to the company site (last 18 Months)
Recency	INPUT	INTERVAL	# days since last purchase
Teenhome	INPUT	INTERVAL	Number of teenagers in household
Year_Birth	INPUT	INTERVAL	Customer's Year of birth

Table 1 - Description of the variables

5. Explore

5.1. Descriptive Statistics

In the StatExplore node, we were able to understand more insights about our data, namely the summary statistics. Regarding the interval variables, we identified missing values in the *Income* (65), *MntWatches* (64) and *MntRackets* (108) variables. We also observed and compared the maximum value, the mean and the median for each variable, in order to understand possible outliers. Therefore, in the *Income* variable, we perceive that with a maximum value of 193685,6, a mean value of 63744,99 and a percentile 50 of 62669, possibly there are some outliers, since the maximum value is very distant from the mean and median value, with huge standard deviations. This occurs in some other variables like *MntRackets, NumCatalogPurchases, NumDealsPurchases* and *NumWebVisitsMonth,* as we can see in the following table.

VARIABLE	MISSING VALUES	N	MIN	MAX	P50	MEAN	STD
Dt_Customer	0	5000	20299	20999	20654	20648,45	201,94
Income	65	4935	1095,6	193685,6	62669	63744,99	29139,83
Kidhome	0	5000	0	2	0	0,43	0,53
MntHats	0	5000	0	398	18	54,16	80,56
MntPremium_Brand	0	5000	0	324	32	56,49	65,12
MntRackets	108	4892	0	2585,3	107	281,31	403,67
MntSneakers	0	5000	0	1497	175	304,66	335,21
MntTShirts	0	5000	0	299	14	40,75	60,53
MntWatches	64	4936	0	180	9	24,44	35,70
NumCatalogPurchases	0	5000	0	25	4	4,89	3,50
NumDealsPurchases	0	5000	0	16	2	2,46	2,33
NumStorePurchases	0	5000	0	14	6	6,71	3,31
NumWebPurchases	0	5000	0	15	8	7,95	2,86
NumWebVisitsMonth	0	5000	0	20	6	5,25	2,69
Recency	0	5000	0	99	50	49,83	29,01
Teenhome	0	5000	0	2	0	0,50	0,54
Year_Birth	0	5000	1944	1999	1973	1971,64	11,99

Table 2 - Summary statistics of the variables

Then, observing the multiplot node, we also analysed the histograms for each variable to have some insights on their empirical distributions. Here, we noticed which variables had missing values and outliers, reinforcing some conclusions already taken.

5.2. Variables Distribution Analysis

In this part, we will analyse the variables before and after the removal of outliers and missing values. Therefore, in the variables where outliers exist, an analysis of the scenario before and after the removal of outliers is made side by side, to verify if outliers have influence in the variables distributions.

We decided not to repeat this analysis after the value imputation of the missing values since when we use a decision tree to impute missing values, it will not cause major changes in the variables distribution.

Observing the histograms of the variables *Accepted Campaigns*, we can conclude that in all of them, there is a high predominance of 0, which means that most of the customers, did not accept the campaigns

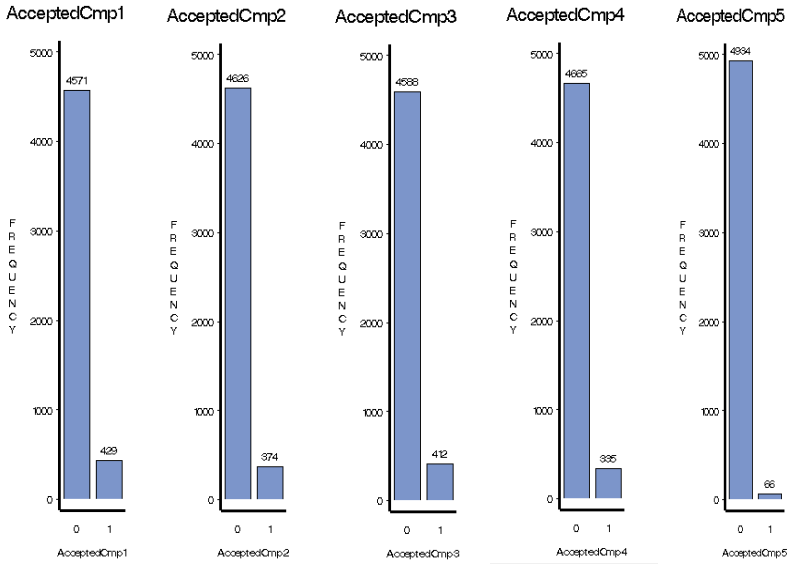

Figure 3 – Histograms of the AcceptedCmp variables

Regarding *Education*, it is possible to conclude that the majority of the clients are graduated, followed by the clients with 2nd cycle.

Almost half of the clients have teens at home, and about 40% of them have kids at home.

Relatively to the *Marital_Status*, we can understand that our clients are predominantly married or are with someone (together) - about 60% - while the others represent less than 40% of the population.

Regarding the *Year_Birth*, we can conclude that all our clients are adults, the distribution is near normal with some small peaks near 1976 and 1981. However, comparing the older and younger customers, there seems to exist a small bias in favour of the younger ones.

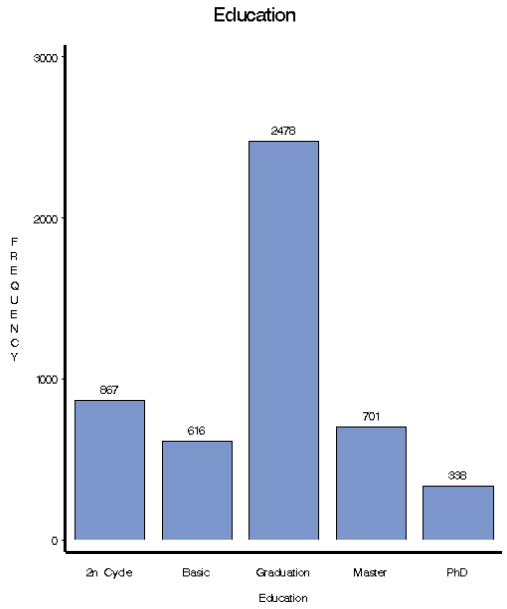

Figure 4 - Histogram of Education variable

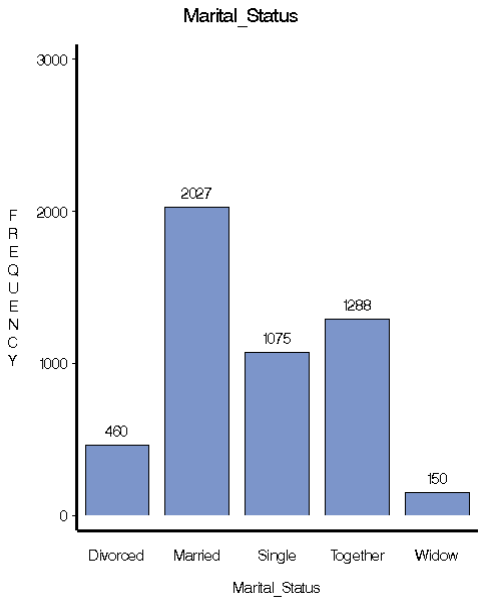

Figure 5 - Histogram of Marital_Status variable

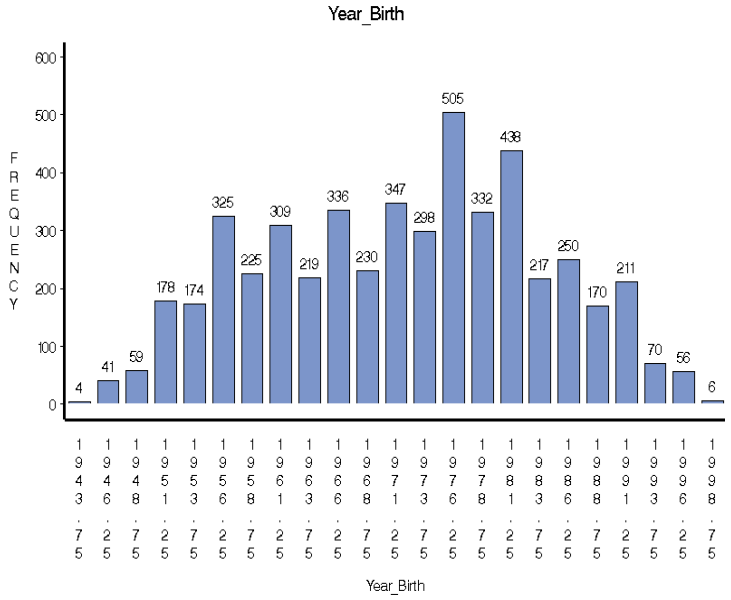

Figure 6 - Histogram of Year_Birth variable

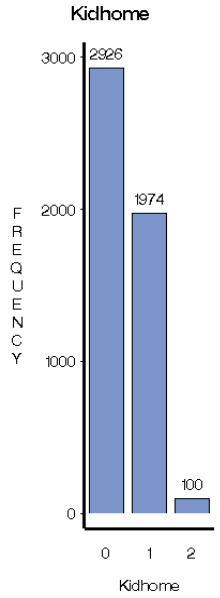

Figure 7 - Histogram of Kidhome variable

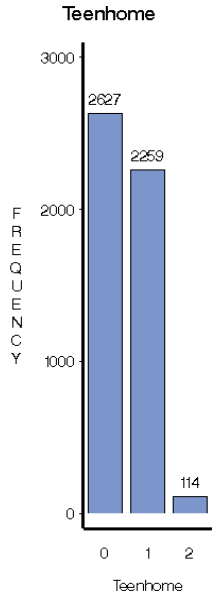

Figure 8 - Histogram of Teenhome variable

Regarding to *Income* (with outliers), first of all, it is to highlight the existence of missing values and outliers, which is not strange, since income is a sensitive issue and some people may choose not to disclose its income. So, observing the distribution, it seems to be close to normal, although some extreme values exist. These values skew the average income and the standard deviation of the distribution, which should be recalculated after the outliers are excluded.

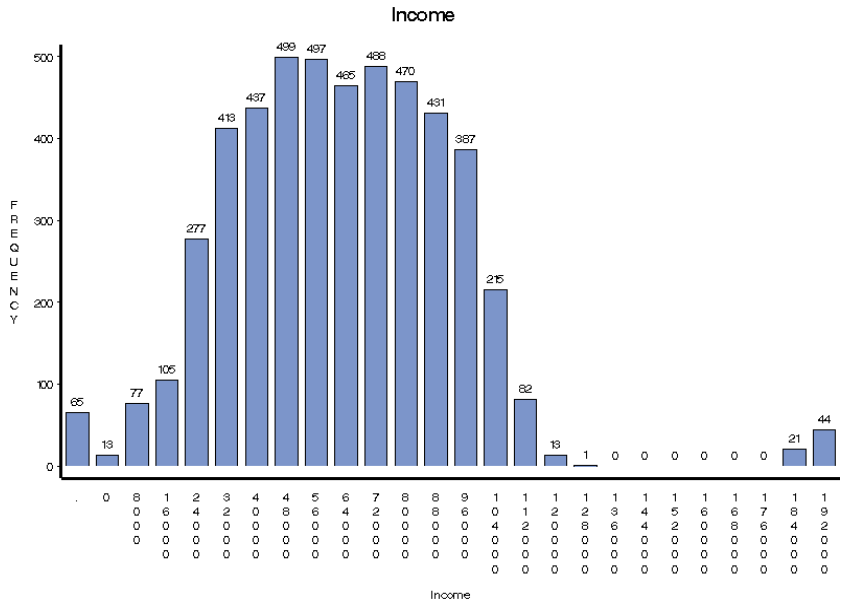

Figure 9 - Histogram of Income variable

After the removal of outliers, we can observe that, foremost, the classes defined are different (more specific), since some observations were removed. Visually we can notice that about 90% of our customers have an income between 30.000€/year and 100.000€/year, which is similar to the last one. It is important to note that the outliers removed had a really high income, so maybe, it should be wise to analyse them, separately to see if we can conclude something "interesting", namely, the creation of a cluster with some of the outliers, or if they should be joined to other clusters.

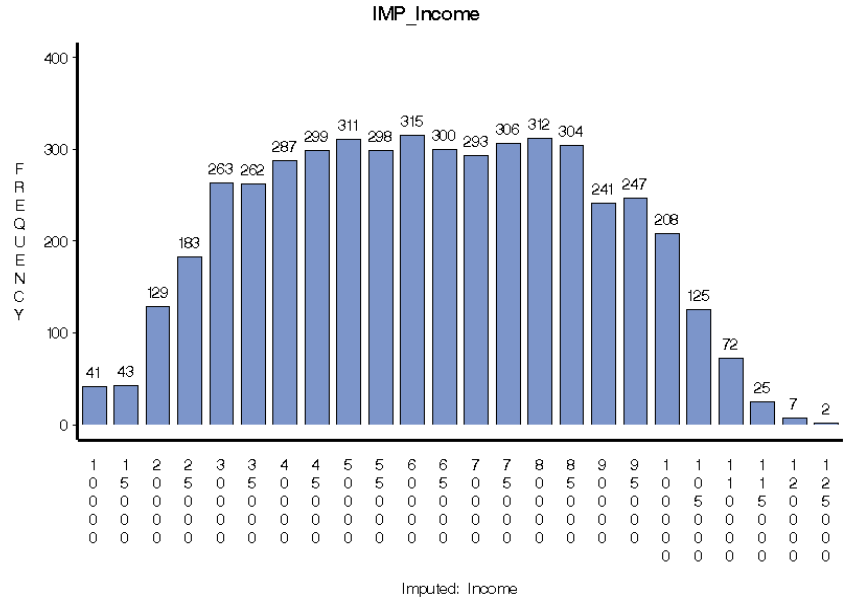

Figure 10 - Histogram of IMP_Income variable

Observing the *Recency* histogram, it seems that the distribution of last date of purchase is uniform.

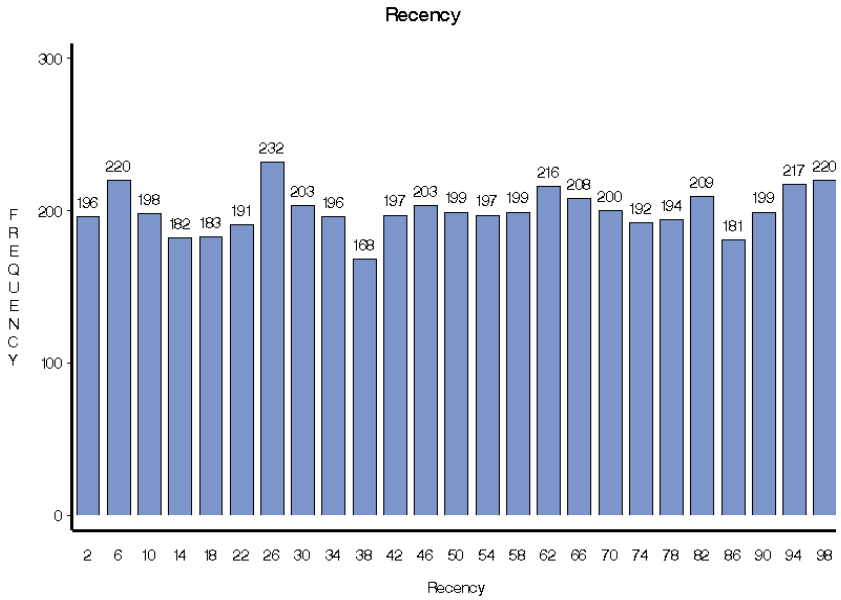

Figure 11 - Histogram of Recency variable

Observing the *MntRackets* distribution, in the first graph, the existence of missing values and outliers is evident. We can conclude that it is a positive asymmetric distribution and the majority of people spent less than 100€ in rackets in the last 18 months. From there, as the money spent increases, the number of clients decreases. It should be noted that this distribution and trend is visible in all the Mnt variables.

It should be noted that regarding this variable, the outliers have a really high monetary expenditure. Therefore, it may be important to analyse this set of individuals. The distribution, after the outlier's removal, remained the same.

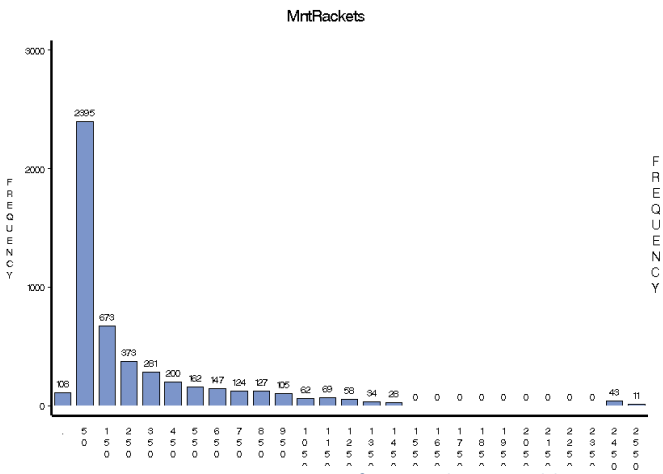

Figure 122 - Histogram of MntRackets variable

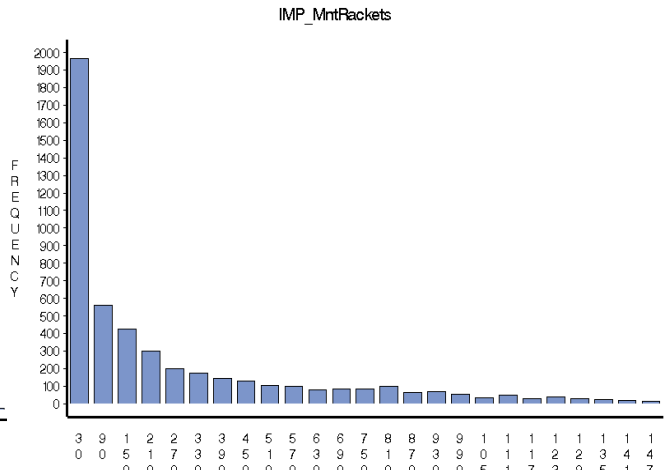

Figure 13 - Histogram of IMP_MntRackets variable

There is a really low number of customers with 4 or less web purchases, and the majority of the clients did 5 to 15 web purchases, with the mode at 5 and decaying from there on. The number of web visits per month is mostly between 6 and 8. Regarding catalog purchases, most of the customers make 2 to 4 purchases and, when we analyse the discount purchases, more than 50% of the clients made just 1 or 2 of them. In the store purchases, we can say that about 1% of the customers did not use this channel, less than 2% made 1 store purchase, and we find that around 40% of clients made between 4 and 5 purchases at the store

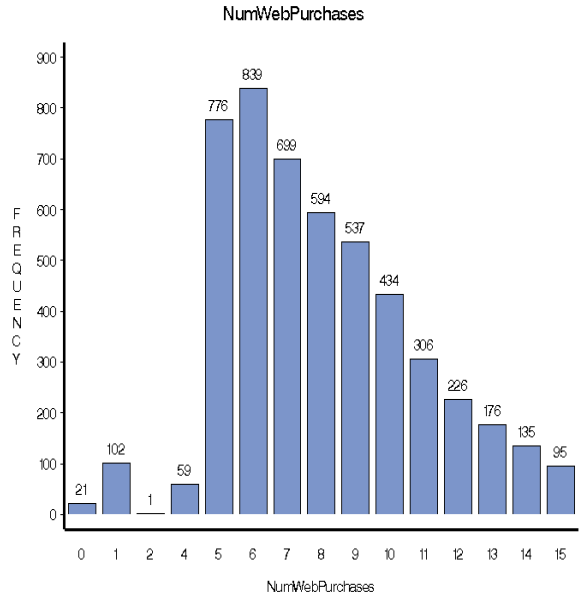

Figure 14 - Histogram of NumStore variable

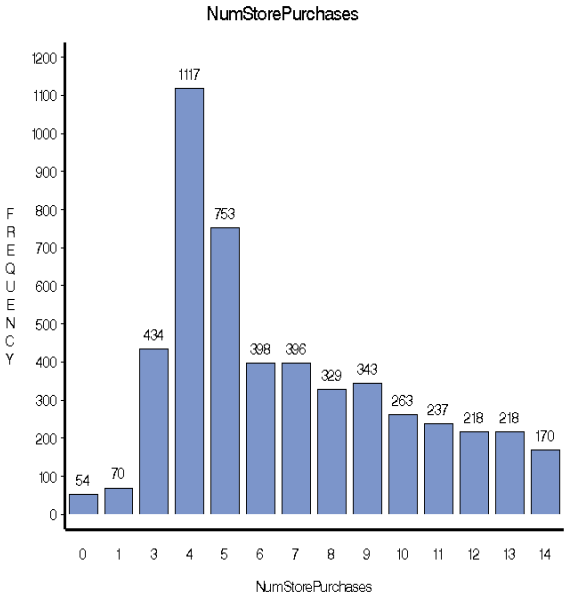

Figure 15 - Histogram of NumWebPurchases variable

Before removal of outliers

After removal of outliers

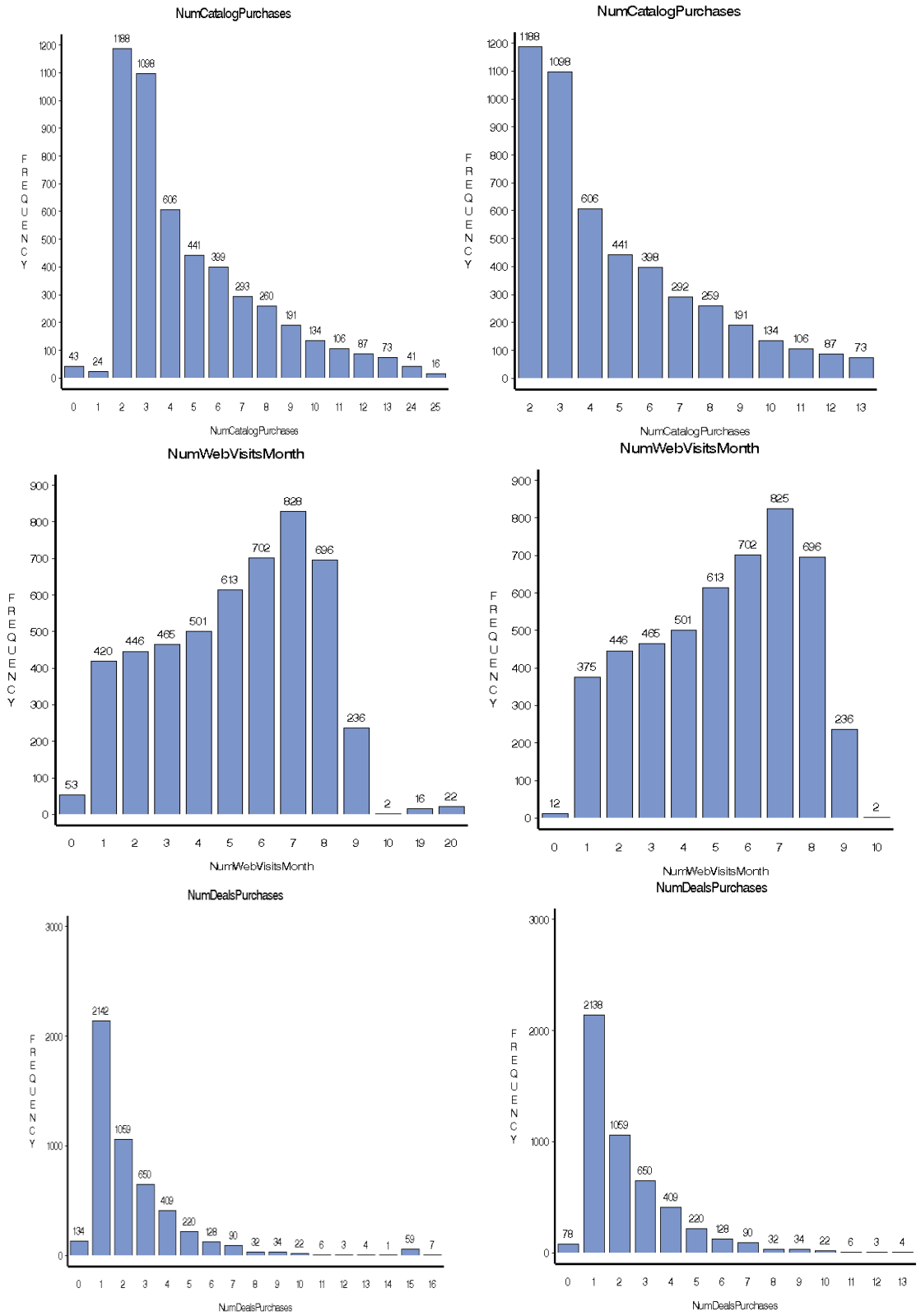

Figure 16 - Histograms before and after removal of outliers

Regarding to the three variables above considering the distributions with and without outliers, it is possible to conclude that the distributions remain the same. Therefore, the analysis is similar to the ones made earlier.

5.3. Outliers

As previously said, we identified outliers in the variables *Income, MntRackets, NumCatalogPurchases, NumDealsPurchases* and *NumWebVisitsMonth*. In the following graphs we can see the cut-off values for each of the variables, for excluding outliers.

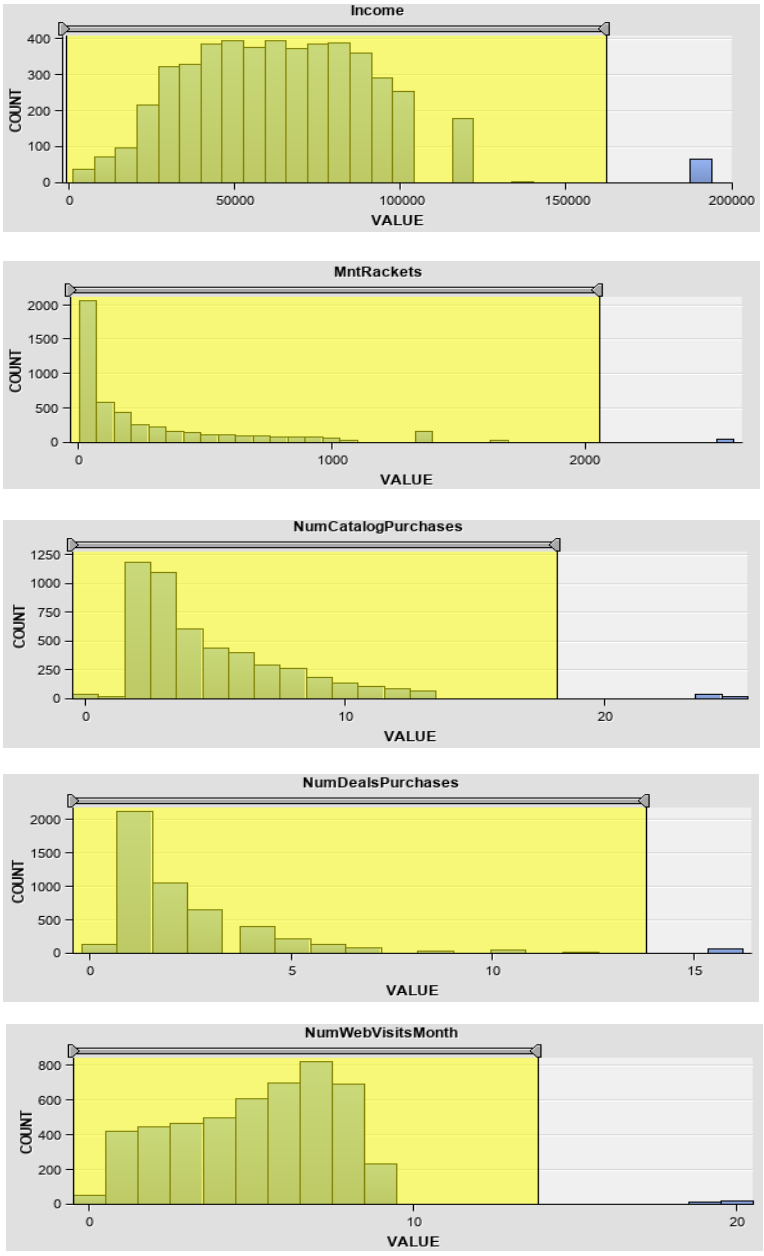

Figure 17 - Removal of outliers

We will discard these observations from the main model, however, we will save them in a sub-dataset. With this new dataset we intend to explore the outliers behaviors, see if the same observations are considered outliers by different criteria, and how do outliers behave regarding the other variables.

With this analysis we can try and better understand if they are the result of obvious errors in our database or if they are individuals with really different values from the rest of the population. We find it important not to exclude them liminarly since we might be excluding clients that are really important for the business as all the exclusions were done on the right side of the distribution (higher values). For instance, we excluded 65 individuals with higher income. So, having more money, the chances of spending more money in LaGoste's products are higher and consequently, there should be a campaign directed to them, if this is the case.

5.4. Missing Values

In the process of analysing our data, it was possible to conclude that some variables - *Income* (65), *MntWatches* (64) and *MntRackets* (108) - had missing values. Therefore, we should manage them in order to ensure data quality of our dataset. To achieve this objective, we used the Impute node, applying a decision tree to assign not completely random values to each missing value. The Impute node is explained further in the annexes.

5.5. Descriptive Statistics After Removal of Outliers and Treatment of Missing Values

After removing the outliers and missing values, we can understand that the maximum value of some variables decreased, since with the treatment of some outliers, the extreme observations in these variables were all to the right side of the distribution and were removed. By discarding the outliers (that represented 2.54% of our dataset) and imputing the missing values, every variable suffered transformations. However, to minimize the changes in the variables distributions, we used the decision tree method to impute the data and, as we saw above, the distributions remained almost the same, with standard deviation values also decreasing for variables who had extreme values.

VARIABLE	MISSING VALUES	N	MIN	MAX	MEAN	STD
Dt_Customer	0	4873	20299	20999	20648,00	202,30
IMP_Income	0	4873	9000	125757	62749,37	24824,82
IMP_MntRackets	0	4873	0	1499	259,95	331,62
IMP_MntWatches	0	4873	0	180	24,99	35,84
Kidhome	0	4873	0	2	0,44	0,53
MntHats	0	4873	0	398	55,37	81,15
MntPremium_Brand	0	4873	0	324	57,78	65,42
MntSneakers	0	4873	0	1497	311,31	336,61
MntTshirts	0	4873	0	299	41,70	61,01
NumCatalogPurchases	0	4873	2	13	4,72	2,81
NumDealsPurchases	0	4873	0	13	2,31	1,81
NumStorePurchases	0	4873	3	14	6,87	3,20
NumWebPurchases	0	4873	4	15	8,13	2,66
NumWebVisitsMonth	0	4873	0	10	5,22	2,33
Recency	0	4873	0	99	49,93	29,04
Teenhome	0	4873	0	2	0,50	0,54
Year_Birth	0	4873	1944	1999	1971,64	11,97

Table 3 - Summary Statistics after removal of outliers

5.6. Correlation

Observing the correlation matrix, after removing the outliers and treating the missing values, it is possible to conclude that there is no relevant correlation between the variables (there is no correlation higher than 0,8).

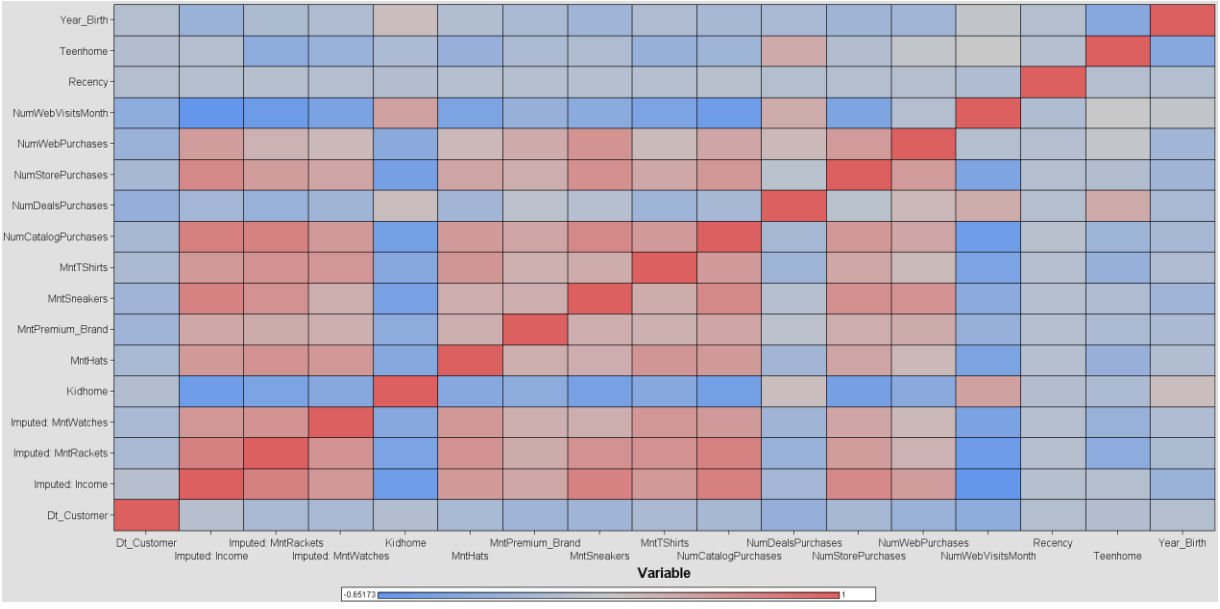

Figure 18 - Correlation Matrix

Looking for the cluster plot of the variables, it is also possible to identify 4 clusters. However, it is relevant to refer that when analysing the same plot before the removal of outliers, we got 5 clusters, instead of 4, and one of them is basically constituted by the variables with outliers - CLUS1. Therefore, these outliers were enough to capture some variables in a unique cluster.

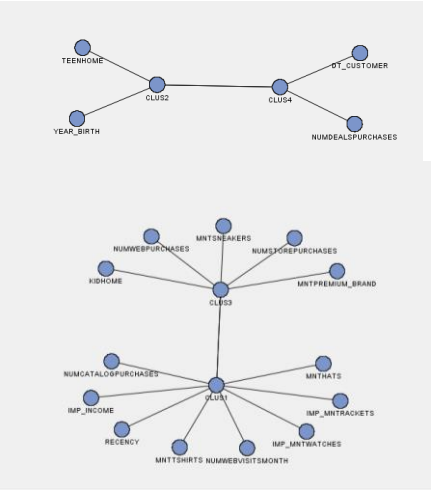

Figure 19 - Clusters of variables without outliers

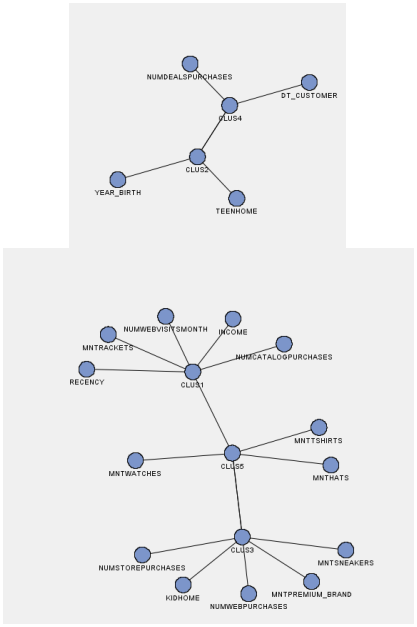

Figure 20 - Clusters of variables with outliers

6. Modify

6.1. Coherence Checking

Before further analysis, it was necessary to check for incoherences in the data, that can lead us to wrong conclusions. This was done using SAS code and in order to detect the possible incoherencies in an easier way, one variable was created for each of the restrictions (IncoherentX, where X=1, 2, ...,13)*.

Here are the coherence rules used:

Incoherence number/variable created*	Incoherence code	Why this restriction
Incoherent1	IMP_Income*1.5 < SUM(MntSneakers, IMP_MntRackets, MntTShirts, IMP_MntWatches, MntHats)	It's not possible for a client to spend more money in LaGoste's products than the total money gained (income)
IncoherentY, where Y=2,3,4,5,6	(AcceptedCmpx NE 0) AND (AcceptedCmpx NE 1), where x=1,2,3,4,5	The campaigns are accepted (1) or not (0), there is no other possibility
Incoherent7	NumDealsPurchases>SUM(NumCatalogPurchases,NumStorePurchases, NumWebPurchases)	A purchase can be made with a deal or not, but the contrary isn't true. It's not possible to have a deal purchase if a purchase wasn't made
Incoherent8	Year_Birth>YEAR(Dt_Customer)	A person that wasn't born at a certain time can't be registered as a client at that moment
Incoherent9	(AcceptedCmp1= 1 OR AcceptedCmp2=1 OR AcceptedCmp3=1 OR AcceptedCmp4=1 OR AcceptedCmp5=1) AND Recency>365	All the campaigns were done in the past year so if a client accepted one or more of the campaigns he has necessarily to have a recency smaller than 365
Incoherent10	SUM(NumCatalogPurchases, NumStorePurchases, NumWebPurchases) < SUM(AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5)	If a client accepts x campaigns, he has necessarily made at least x purchases, because each accepted campaign means a purchase made
Incoherent11	NumWebPurchases > 0 AND (NumWebVisitsMonth*18) = 0	A person can't have any web purchases if he hasn't visited the website
Incoherent12	SUM(NumCatalogPurchases,NumStorePurchases, NumWebPurchases) = 0 AND SUM(MntSneakers,IMP_MntRackets,MntTShirts, IMP_MntWatches,MntHats) > 0	It's not possible for a client to have spent money in LaGoste's products if he didn't do any purchase.
Incoherent13	SUM(NumCatalogPurchases,NumStorePurchases, NumWebPurchases) > 0 AND SUM(MntSneakers,IMP_MntRackets,MntTShirts, IMP_MntWatches,MntHats) = 0	It's not possible for a client to do a purchase and spend no money.

Table 4 - Coherence Checking

After running the Coherence Check code node, we found 6 incoherencies. All of them with the same problem and related to the same coherence check rule (*Incoherent11*) - the person did not make any web visit, however, made web purchases. This can be the result, not of an insertion error, but of the storage format selection for this variable. The variable should, in fact, be a floating-point value obtained by dividing the total visits by the time span, but was instead saved as an integer through rounding. The rounding applied causes loss of information, making someone who already visited us up to 9 times in the last 18 months, to have an average visit number rounded down to 0.

Having this in mind, we decided to change the value of the *NumWebVisitsMonth* variable, on the incoherent observations found. In order to differentiate the clients that had already a value of 1, in the original dataset, we chose to change the incoherent values to 0,5.

6.2. Transform Variables

In order to try to gain explainability and partitioning power for a better and more precise segmentation of the customers, we transformed existing variables and created new ones based on the originals, thinking about different business rules.

The following tables present the transformations made and the respective new variables.

ID	TRANSFORMED VARIABLES	DESCRIPTION
1	Age=2017-Year_Birth	Age of the customer when the database was extracted
2	AmountSpentperPurchase= TotalAmountSpent/TotalPurchases	In average, how much the client spent on each purchase made
3	HigherEducation: 1, if Education in ("Graduation", "Master", "PhD") 0, otherwise	Binary variable that tells if the person has a higher education degree/high level of education (1)
4	Income_18months=IMP_Income*1.5	Average income per 18 months
5	IncomeSpentOnUs=TotalAmountSpent/Income_18months	Proportion of the total income, that was spent on purchases in the company, in the last 18 months
6	InvRecency=99-Recency	Importance of the client in terms of the last purchase
7	Log_IMP_Income=log(IMP_Income)	Logaritm of the IMP_Income variable
8	Log_TotalAmountSpent=log(TotalAmountSpent)	Logaritm of the TotalAmountSpent variable
9	Log_TypeProductsPurchase=log(TypeProductsPurchase)	Logaritm of the TypeProductsPurchase variable
10	MntxRatio=X/TotalAmountSpent, where X = MntPremium_Brand, MntSneakersRatio, IMP_MntRacketsRatio, MntTShirtsRatio, IMP_MntWatchesRatio, MntHatsRatio	Proportion of money spent in each category over the total amount spent in the company
11	RatioWebVisits_Purchases=NumWebPurchases/NumWebVisitsMonth	Proportion of times that the client visited the website and bought any products over the total number of visits
12	STD_InvRecency = InvRecency standardized	Z-score Normalization (obtained values are between -1 and 1)
13	STD_TotalAmountSpent = TotalAmountSpent standardized	
14	STD_TotalPurchases = TotalPurchases standardized	
15	RANGE_InvRecency	Min-Max normalization, to obtain values between 0 and 1 (non negative)
16	RANGE_TotalAmountSpent	
17	RANGE_TotalPurchases	
18	RANGE_IMP_Income	
19	TogetherStatus: 1, if Marital_Status in ("Married" or "Together") 0, otherwise	Binary variable that tells if a person lives with someone or not
20	TotalAcceptedCmp=AcceptedCmp1+AcceptedCmp2+AcceptedCmp3+AcceptedCmp4+ AcceptedCmp5	Total number of accepted campaigns, of the 5 done previously
21	TotalAmountSpent=MntSneakers+IMP_MntRackets+MntTShirt+IMP_MntWatches+ MntHats	Total amount spent in all products
22	TotalChildrenBinary: 1, if Kidhome+Teenhome>0 0, if Kidhome+Teenhome=0	Binary variable that tells if there are any kids or teenagers in the household
23	TotalChildrenDiscrete=Kidhome+Teenhome	Total number of children and teenagers in the household
24	TotalPurchases= NumStorePurchases+ NumCatalogPurchases+NumWebPurchases	Total purchases made in last 18 months
25	Typekids: 0, if Kidhome=0 and Teenhome=0 1, if Kidhome=0 and Teenhome>0 2, if Kidhome>0 and Teenhome=0 3, if Kidhome>0 and Teenhome>0	Multinomial variable, that tells if there are only kids in the household (2), only teens in the household (1), both (3) or none (0)
26	X : 1, if MntX>0 0, if MntX=0 where X=Sneakers, Rackets, TShirts, Watches, Hats	Binary variables for each of the product types, in order to know if the client ever bought that specific kind of product
27	TypeProductsPurchase=Sneakers+Rackets+TShirts+Watches+Hats	Sum of the previous 5 binary variables, that gives how many type of products the client has bought
28	XPurchasesRatio=NumxPurchases/TotalPurchases, where X = Catalog, Store, Deals, Web	Proportion of purchases of each category over the total number of purchases
29	YearsWithUs=2017-Year(Dt_Customer)	Number of years since the person has become LaGoste's client
30	IncomeT=(IMP_Income*(TogetherStatus+1))/(TogetherStatus+1+TotalChildrenDiscrete)	Variable that takes into account the person income and if she is married and how many kids they have in their household

Table 5 - Transformed Variables

RFM Analysis variables - measure the clients value to the company		
ID	**TRANSFORMED VARIABLES**	**DESCRIPTION**
31	ImportanceA=STD_InvRecency + STD_TotalAmountSpent	Recency and Monetary Analysis (sum of Z-score normalized variables)
32	ImportanceB= STD_TotalAmountSpent + STD_TotalPurchases	Frequency and Monetary Analysis (sum of Z-score normalized variables)
33	ImportanceC=STD_InvRecency+STD_TotalAmountSpent+STD_TotalPurchases	Recency, Frequency and Monetary Analysis (sum of Z-score normalized variables)
34	ImportanceRMx=RANGE_InvRecency*RANGE_TotalAmountSpent	Recency and Monetary Analysis (multiplication of min-max normalized variables)
35	ImportanceFMx=RANGE_TotalAmountSpent*RANGE_TotalPurchases	Frequency and Monetary Analysis (multiplication of min-max normalized variables)
36	ImportanceRFMx=RANGE_InvRecency*RANGE_TotalAmountSpent* RANGE_TotalPurchases	Recency, Frequency and Monetary Analysis (multiplication of min-max normalized variables)
37	ImportanceRMp=RANGE_InvRecency+RANGE_TotalAmountSpent	Recency and Monetary Analysis (sum of min-max normalized variables)
38	ImportanceFMp=RANGE_TotalAmountSpent+RANGE_TotalPurchases	Frequency and Monetary Analysis (sum of min-max normalized variables)
39	ImportanceRFMp=RANGE_InvRecency+RANGE_TotalAmountSpent+ RANGE_TotalPurchases	Recency, Frequency and Monetary Analysis (sum of min-max normalized variables)
40	ImportanceIMx=RANGE_IMP_Income*RANGE_TotalAmountSpent	Monetary Analysis with Income (multiplication of min-max normalized variables)
41	ImportanceIFx=RANGE_IMP_Income*RANGE_TotalPurchases	Frequency Analysis with Income (multiplication of min-max normalized variables)
42	ImportanceIFMx=RANGE_IMP_Income*RANGE_TotalAmountSpent* RANGE_TotalPurchases	Frequency and Monetary Analysis with Income (multiplication of min-max normalized variables)
43	ImportanceIMp=RANGE_IMP_Income+RANGE_TotalAmountSpent	Monetary Analysis with Income (sum of min-max normalized variables)
44	ImportanceIFp=RANGE_IMP_Income+RANGE_TotalPurchases	Frequency Analysis with Income (sum of min-max normalized variables)
45	ImportanceIFMp=RANGE_IMP_Income+RANGE_TotalAmountSpent+ RANGE_TotalPurchases	Frequency and Monetary Analysis with Income (sum of min-max normalized variables)

Table 6 - RFM Analysis Variables

Based on the information referred in "Data Mining and Predictive Analytics", 2nd Edition, Wiley (page 722 – 724), we understand that typically individuals who buy from just one or two categories tend not to adhere to campaigns, since they know exactly what they want to buy, and people who buy different products with us, tend to adhere more to campaigns. Therefore, we decided to create binary variables for each of the product types (*Sneakers*, *Rackets*, *T-shirts*, *Watches* and *Hats* - variables 26) in order to understand if a client ever bought that specific type of product and then, sum all of them (variable 27) to know how many types of different product categories the client has bought.

When analysing *Education* and *Marital_Status* we found that we had different categories that could have a similar behaviour, so we decided to join the categories that were identical in the *Education* variable and then did the same with the ones in *Marital_Status*, creating the dummy variables 3 and 19, respectively.

We decided to apply the log transformation to the variables *IMP_Income, TotalAmountSpent*, and *TypeProductsPurchase* (7, 8 and 9) since we wanted to see if it is possible to remove some skewness from the variables, so that we can get a different view of the data. Many researchers want to correct the skewness because of the assumptions many algorithms make regarding

the shape of the data, namely normal distributions. The log transformation is perhaps the most often used transformation to correct for positive skew (all the variables had only positive values in the distribution).

We opted to use the year 2017, as the year when dataset was extracted from the database, to form the variables *Age* and *YearsWithUs*. We made this decision after analysing the dataset and realizing that the last purchases were made in 2017.

Regarding the variable *InvRecency*, the objective of its creation was to consider the value 0 as the worst possible value and have the higher numbers representing better clients. So, we inverted the *Recency* variable by subtracting the value of each observation, from the maximum of the variable.

At this point, we decided to reject some variables, so that they didn't enter in the next phase, where they would or could negatively affect the correlations. This said, we rejected the variables that were only created in order to achieve a final one (the variables *Sneakers*, *Rackets*, *T-shirts*, *Watches* and *Hats*) and the ones created to help us knowing which ones of the coherence check rules were giving us incoherencies.

Regarding the variables *TotalAmountSpent*, *TotalPurchases* and *InvRecency (from Recency)*, we decided to standardize them since we were going to create composite variables that better represent customer importance (*ImportanceA*, *ImportanceB* and *ImportanceC*), and we wanted them to have equal weights, instead of unknown random weights due to different scales, since two have a monetary scale and the third has a day scale. We at first thought about multiplying the variables, but then realized that multiplying Z-score normalized variables wouldn't have the result we were expecting. Explaining further the problem, considering an observation with a value of 0 in the Recency variable, the new variable would always have the value 0, whether the customer was very good or really bad in the other dimensions. Considering a new case with 2 positive values, the multiplication would have the same result than if the 2 were negative. We ended up summing the variables to create the 3 new variables.

However, the previously exposed problem can be solved with a different normalization transformation. Therefore, we applied a min-max normalization which results in a scaled value of a variable equal to (x - min) / (max - min), where x is current variable value, min is the minimum value for that variable, and max is the maximum value for that variable. SAS already allows this range standardization method. We applied it on the variables *TotalAmountSpent*, *TotalPurchases* and *InvRecency* thus allowing for the multiplication or addition between them without any kind of issues to the concept previously explained.

The main differences between the methods are:

1. Z-score normalization subtracts by the mean, so a zero value is equivalent to the mean value of the original distribution, whereas on a Min-max a zero value is the minimum value of the original distribution.
2. Z-score normalization divides by the standard error, so the observation assigned score is the distance from the mean in standard deviations, while in Min-max we divide each value by the amplitude (max-min), so every observation gets a value attributed from 0 to 1. In this transformation, the new mean must be calculated using the same transformation, 0.5 is not the mean.

Consequently, we created the variables *ImportanceRMx, ImportanceFMx, ImportanceRFMx* that are similiar to *ImportanceA*, *ImportanceB* and *ImportanceC,* but instead of having standardized variables, they are composed by min-max normalized variables. Adding to this, we also created *ImportanceRMp, ImportanceFMp and ImportanceRFMp.* These two groups of variables, that only differ in the type of operation, were created to test which one is better to increase our partitioning power.

In addition, we also conceptualize that *IMP_Income* could also be important to define the value of the customer as more wealthy people have more money available, thus being able to purchase with less restrictions. Then, we also applied the min-max normalization on this variable and also multiplied and sum with the variables *RANGE_TotalAmountSpent* and *RANGE_TotalPurchases* forming the variables 40, 41, 42, 43, 44 and 45.

Another transformation to *Income* that we decided to apply was *IncomeT*, where we take into consideration if the individual is together or not, and the total amount of kids. We hope this captures better the actual amount available for shopping. We considered that the income of someone married or together can be 2 times higher than someone who is single. This is certainly incorrect but on average we consider this to be more correct than consider a single income for someone who is together or married. Then, we divided the income by the number of mouths to feed. This was obtained by considering the sum of total of kids and adults in the household, obtained by adding 1 to our binary together value. Alone people had a value of 0, becoming 1, and together had a value of 1, becoming 2.

This approach has some problems, including the selected multiplying factor of 2, or not considering other people that may also live in the household, contributing or not, or if the other member of the couple actually has an income or not, or if they are at all related. The point is, these are all factors that are also not captured by the original income, and it is our belief that this variable can in fact capture part of this effect better than income alone.

6.3. Correlation

After the metadata node, we decided to reanalyse the correlations because it is important to see the correlation with transformed and new variables. This is relevant for choosing the variables in the forward nodes since highly correlated values can inflate importance of variables, leading to wrong cluster definitions.

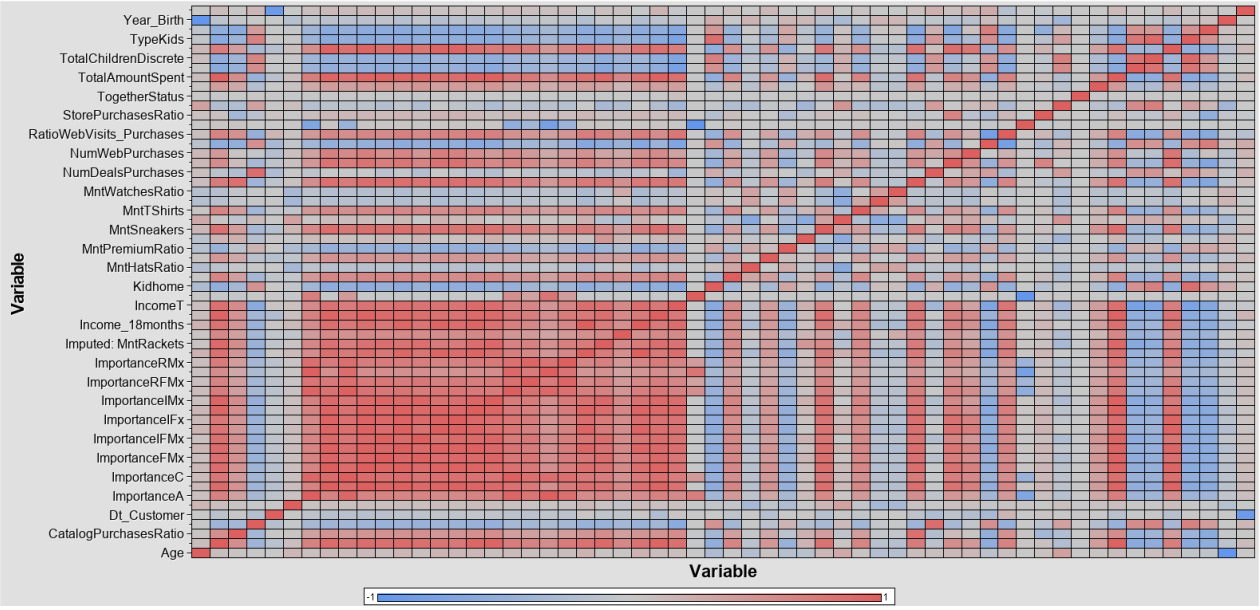

Figure 21 - New Correlation Matrix

Variable	Variable Correlated	Correlation
Age	Year_Birth	-1
AmountSpentperPurchase	TotalAmountSpent	0.971157356
	ImportanceIMx	0.948325052
	IncomeSpentOnUs	0.9429758
	ImportanceIMp	0.933412263
	Imputed: MntRackets	0.892154432
	ImportanceIFMp	0.885297693
	ImportanceB	0.883460565
	ImportanceFMp	0.877000585
	ImportanceFMx	0.849049969
	ImportanceIFMx	0.833515061
	MntSneakers	0.822838838
	Imputed: Income	0.814567954
	Income_18months	0.814567954
	ImportanceIFp	0.800320781
CatalogPurchasesRatio	NumCatalogPurchases	0.841918763
DealsPurchasesRatio	NumDealsPurchases	0.878983732
Dt_Customer	YearsWithUs	-0.913506217
ImportanceA	ImportanceRFMp	0.991751871
	ImportanceRMp	0.988447222
	ImportanceC	0.920343194
	ImportanceRFMx	0.860309779
	ImportanceRMx	0.853079244
ImportanceB	ImportanceFMp	0.999886475
	ImportanceIFMp	0.984523161
	ImportanceIFx	0.958038905
	ImportanceIFp	0.956284667
	TotalAmountSpent	0.949805735
	TotalPurchases	0.949805735
	ImportanceFMx	0.949399657
	IncomeSpentOnUs	0.944821512
	ImportanceIMp	0.94348938
	ImportanceIMx	0.906550321
	ImportanceIFMx	0.90576733
	AmountSpentperPurchase	0.883460565
	ImportanceC	0.882120085
	MntSneakers	0.859188563
	Imputed: Income	0.8558722
	Income_18months	0.8558722
	NumCatalogPurchases	0.850865264
	Imputed: MntRackets	0.82128052
	NumStorePurchases	0.802897069
ImportanceC	ImportanceRFMp	0.962881913
	ImportanceA	0.920343194
	ImportanceB	0.882120085
	ImportanceFMp	0.882024137
	ImportanceRMp	0.868324756
	ImportanceIFMp	0.867714105
	ImportanceRMx	0.853597585
	ImportanceIFx	0.845835476
	ImportanceIFp	0.842500311
	ImportanceFMx	0.839307752
	TotalPurchases	0.83792979
	TotalAmountSpent	0.837755641
	IncomeSpentOnUs	0.83354841
	ImportanceIMp	0.831078024
	ImportanceRFMx	0.800708463
	ImportanceIFMx	0.800618405
ImportanceFMp	ImportanceB	0.999886475
	ImportanceIFMp	0.98423322
	ImportanceIFx	0.960050662
	ImportanceIFp	0.958314551
	TotalPurchases	0.954411705
	ImportanceFMx	0.946998303
	TotalAmountSpent	0.944984112
	IncomeSpentOnUs	0.941446685
	ImportanceIMp	0.940386136
	ImportanceIFMx	0.902234098
	ImportanceIMx	0.900684989
	ImportanceC	0.882024137
	AmountSpentperPurchase	0.877000585
	MntSneakers	0.856963352
	Imputed: Income	0.854790679
	Income_18months	0.854790679
	NumCatalogPurchases	0.850979284
	Imputed: MntRackets	0.814964248
	NumStorePurchases	0.807793854

Variable	Variable Correlated	Correlation
ImportanceFMx	ImportanceIFMx	0.982890925
	ImportanceB	0.949399657
	TotalAmountSpent	0.949364998
	ImportanceFMp	0.946998303
	ImportanceIMx	0.938065213
	ImportanceIFMp	0.922556357
	ImportanceIFx	0.916796758
	IncomeSpentOnUs	0.907608468
	ImportanceIMp	0.904865408
	ImportanceIFp	0.865894899
	TotalPurchases	0.854125479
	Imputed: MntRackets	0.851665809
	AmountSpentperPurchase	0.849049969
	ImportanceC	0.839307752
	MntSneakers	0.832743577
	NumCatalogPurchases	0.828465192
ImportanceIFMp	ImportanceB	0.984523161
	ImportanceIFp	0.984515038
	ImportanceFMp	0.98423322
	ImportanceIMp	0.977596067
	ImportanceIFx	0.964960477
	TotalAmountSpent	0.938805015
	Imputed: Income	0.933107015
	Income_18months	0.933107015
	TotalPurchases	0.931406473
	ImportanceFMx	0.922556357
	ImportanceIMx	0.910163389
	IncomeSpentOnUs	0.906396906
	ImportanceIFMx	0.892318402
	AmountSpentperPurchase	0.885297693
	ImportanceC	0.867714105
	NumCatalogPurchases	0.844227578
	MntSneakers	0.843038442
	Imputed: MntRackets	0.81631318
ImportanceIFMx	ImportanceFMx	0.982890925
	ImportanceIMx	0.95487993
	TotalAmountSpent	0.931525975
	ImportanceB	0.90576733
	ImportanceFMp	0.902234098
	ImportanceIFx	0.896760379
	ImportanceIMp	0.89663351
	ImportanceIFMp	0.892318402
	Imputed: MntRackets	0.860062609
	IncomeSpentOnUs	0.851258119
	AmountSpentperPurchase	0.833515061
	ImportanceIFp	0.830756729
	NumCatalogPurchases	0.802514776
	ImportanceC	0.800618405
ImportanceIFp	ImportanceIFMp	0.984515038
	ImportanceIFx	0.97036796
	ImportanceFMp	0.958314551
	ImportanceB	0.956284667
	TotalPurchases	0.952683632
	ImportanceIMp	0.941812779
	Imputed: Income	0.940366619
	Income_18months	0.940366619
	ImportanceFMx	0.865894899
	TotalAmountSpent	0.86388569
	ImportanceC	0.842500311
	IncomeSpentOnUs	0.834186209
	ImportanceIMx	0.831831742
	ImportanceIFMx	0.830756729
	NumCatalogPurchases	0.825445178
	NumStorePurchases	0.822431521
	AmountSpentperPurchase	0.800320781
ImportanceIFx	ImportanceIFp	0.97036796
	ImportanceIFMp	0.964960477
	ImportanceFMp	0.960050662
	ImportanceB	0.958038905
	TotalPurchases	0.953977601
	ImportanceFMx	0.916796758
	ImportanceIMp	0.911415312
	ImportanceIFMx	0.896760379
	Imputed: Income	0.879461707
	Income_18months	0.879461707
	TotalAmountSpent	0.86592409
	ImportanceIMx	0.852777874
	NumCatalogPurchases	0.849428705
	ImportanceC	0.845835476
	NumStorePurchases	0.82518371
	IncomeSpentOnUs	0.820172177

Variable	Variable Correlated	Correlation
ImportanceIMp	ImportanceIFMp	0.977596067
	TotalAmountSpent	0.958338154
	Imputed: Income	0.956515134
	Income_18months	0.956515134
	ImportanceIMx	0.951260468
	ImportanceB	0.94348938
	ImportanceIFp	0.941812779
	ImportanceFMp	0.940386136
	AmountSpentperPurchase	0.933412263
	ImportanceIFx	0.911415312
	ImportanceFMx	0.904865408
	ImportanceIFMx	0.89663351
	IncomeSpentOnUs	0.894573191
	Imputed: MntRackets	0.861329057
	TotalPurchases	0.833925094
	MntSneakers	0.831856622
	ImportanceC	0.831078024
	NumCatalogPurchases	0.810156046
ImportanceIMx	TotalAmountSpent	0.980687468
	ImportanceIFMx	0.95487993
	ImportanceIMp	0.951260468
	AmountSpentperPurchase	0.948325052
	ImportanceFMx	0.938065213
	Imputed: MntRackets	0.916460208
	ImportanceIFMp	0.910163389
	ImportanceB	0.906550321
	ImportanceFMp	0.900684989
	IncomeSpentOnUs	0.900588951
	ImportanceIFx	0.852777874
	Imputed: Income	0.83937488
	Income_18months	0.83937488
	ImportanceIFp	0.831831742
	MntSneakers	0.819358105
ImportanceRFMp	ImportanceA	0.991751871
	ImportanceRMp	0.966731653
	ImportanceC	0.962881913
	ImportanceRMx	0.868482112
	ImportanceRFMx	0.85613973
	ImportanceRMx	0.906503215
	ImportanceA	0.860309779
	ImportanceRFMp	0.85613973
	ImportanceRMp	0.834938996
	ImportanceC	0.800708463
	ImportanceA	0.988447222
	ImportanceRFMp	0.966731653
	ImportanceC	0.868324756
	ImportanceRFMx	0.834938996
	ImportanceRMx	0.80737126
ImportanceRMx	ImportanceRFMx	0.906503215
	ImportanceRFMp	0.868482112
	ImportanceC	0.853597585
	ImportanceA	0.853079244
	ImportanceRMp	0.80737126
Imputed: Income	Income_18months	1
	ImportanceIMp	0.956515134
	ImportanceIFp	0.940366619
	ImportanceIFMp	0.933107015
	ImportanceIFx	0.879461707
	IncomeT	0.86821448
	ImportanceB	0.8558722
	ImportanceFMp	0.854790679
	ImportanceIMx	0.83937488
	TotalAmountSpent	0.833349838
	AmountSpentperPurchase	0.814567954
Imputed: MntRackets	Imputed: MntRackets	1
	ImportanceIMx	0.916460208
	TotalAmountSpent	0.909261229
	AmountSpentperPurchase	0.892154432
	ImportanceIMp	0.861329057
	ImportanceIFMx	0.860062609
	IncomeSpentOnUs	0.852014194
	ImportanceFMx	0.851665809
	ImportanceB	0.82128052
	IncomeT	0.820129251
	ImportanceIFMp	0.81631318
	ImportanceFMp	0.814964248
Income_18months	Imputed: Income	1
	ImportanceIMp	0.956515134
	ImportanceIFp	0.940366619
	ImportanceIFMp	0.933107015
	ImportanceIFx	0.879461707
	IncomeT	0.86821448
	ImportanceB	0.8558722
	ImportanceFMp	0.854790679
	ImportanceIMx	0.83937488
	TotalAmountSpent	0.833349838
	AmountSpentperPurchase	0.814567954

Variable	Variable Correlated	Correlation
IncomeSpentOnUs	TotalAmountSpent	0.965238959
	ImportanceB	0.944821512
	AmountSpentperPurchase	0.9429758
	ImportanceFMp	0.941446685
	ImportanceFMx	0.907608468
	ImportanceIFMp	0.906396906
	ImportanceIMx	0.900588951
	ImportanceIMp	0.894573191
	MntSneakers	0.854153057
	Imputed: MntRackets	0.852014194
	ImportanceIFMx	0.851258119
	ImportanceIFp	0.834186209
	ImportanceC	0.83354841
	TotalPurchases	0.829554822
	ImportanceIFx	0.820172177
IncomeT	Imputed: Income	0.86821448
	Imputed: MntRackets	0.820129251
	Income_18months	0.86821448
	TotalAmountSpent	0.850329251
InvRecency	Recency	-1
Kidhome	TypeKids	0.857491411
MntSneakers	TotalAmountSpent	0.860237504
	ImportanceB	0.859188563
	ImportanceFMp	0.856963352
	IncomeSpentOnUs	0.854153057
	ImportanceIFMp	0.843038442
	ImportanceFMx	0.832743577
	ImportanceIMp	0.831856622
	AmountSpentperPurchase	0.822838838
	ImportanceIMx	0.819358105
NumCatalogPurchases	ImportanceFMp	0.850979284
	ImportanceB	0.850865264
	ImportanceIFx	0.849428705
	ImportanceIFMp	0.844227578
	CatalogPurchasesRatio	0.841918763
	ImportanceFMx	0.828465192
	ImportanceIFp	0.825445178
	TotalPurchases	0.812529539
	ImportanceIMp	0.810156046
	TotalAmountSpent	0.803783874
	ImportanceIFMx	0.802514776
	DealsPurchasesRatio	0.878983732
	TotalPurchases	0.866157032
	ImportanceIFx	0.82518371
	ImportanceIFp	0.822431521
	ImportanceFMp	0.807793854
	ImportanceB	0.802897069
NumWebVisitsMonth	RatioWebVisits_Purchases	-0.811950859
RatioWebVisits_Purchases	NumWebVisitsMonth	-0.811950859
Recency	InvRecency	-1
TotalAmountSpent	ImportanceIMx	0.980687468
	AmountSpentperPurchase	0.971157356
	IncomeSpentOnUs	0.965238959
	ImportanceIMp	0.958338154
	ImportanceB	0.949805735
	ImportanceFMx	0.949364998
	ImportanceFMp	0.944984112
	ImportanceIFMp	0.938805015
	ImportanceIFMx	0.931525975
	Imputed: MntRackets	0.909261229
	ImportanceIFx	0.86592409
	ImportanceIFp	0.86388569
	MntSneakers	0.860237504
	IncomeT	0.850329251
	ImportanceC	0.837755641
	Imputed: Income	0.833349838
	Income_18months	0.833349838
	TotalPurchases	0.804261867
	NumCatalogPurchases	0.803783874
TotalChildrenBinary	TotalChildrenDiscrete	0.808630926
TotalChildrenDiscrete	TypeKids	0.907411968
	TotalChildrenBinary	0.808630926
TotalPurchases	ImportanceFMp	0.954411705
	ImportanceIFx	0.953977601
	ImportanceIFp	0.952683632
	ImportanceB	0.949805735
	ImportanceIFMp	0.931406473
	NumStorePurchases	0.866157032
	ImportanceFMx	0.854125479
	ImportanceC	0.83792979
	ImportanceIMp	0.833925094
	IncomeSpentOnUs	0.829554822
	NumCatalogPurchases	0.812529539
	TotalAmountSpent	0.804261867
TypeKids	TotalChildrenDiscrete	0.907411968
	Kidhome	0.857491411
Year_Birth	Age	-1
YearsWithUs	Dt_Customer	-0.913506217

Table 7 - Variables Correlation

It should be noted that we only took register of correlations that were equal or higher than 0,8, since those are a strong indication that we should not use these variables at the same time to perform a clustering technique.

At a first look, we can see that the majority of the Importances variables have a really high correlation between them and with a few more, namely, *AmountSpentperPurchase, TotalAmountSpent, TotalPurchases, NumCatalogPurchases, MntSneackers, IMP_MntRackets* and all the variables built from the *Income* variable. Therefore, we have to take special attention when using these variables.

The other variables that have high correlation are the ones that were built directly using the original variables. For example, the variable *Age* is correlated with the *Year_Birth*, as well as the *Recency* with *InvRecency*. The ratio variables (*CatalogPurchaseRatio and DealsPurchaseRatio*) are correlated with the variables that form them, namely *NumCatalogPurchases* and *NumDealsPurchases*.

Concluding, we should not include, at least in the same "view" highly correlated variables. Furthermore, we must have attention to the Importances.

7. Model

After "cleaning" and transforming the dataset, we used the Metadata node to remove some variables. We rejected the incoherences indicators, the log variables and the variables that formed the *TypeProductsPurchases* because we considered that it would not make sense to include them, in any of the views. We also rejected the normalized variables (RANGE and STD) because every variable that enters in the Clustering node can be automatically standardized, in order to form the clusters.

After that, we had to choose the variables to include in each view we want to have of the customers. We chose these variables taking into account the correlations.

At this point we started to cluster the observations. At first, in each view, we used the elbow graphic to understand what would be the best number of clusters to maintain. Then the K-Means Algorithm was used to perform the clustering as well as the SOM-VQ which will be explained ahead, since it constitutes a self-study node.

Regarding to the K-Means this is an algorithm that starts with a predefined number of seeds, initialized in a random way, and then connects the observations with the nearest seed, iteration after iteration - while changing the seeds coordinates, that become the centroids of the cluster - until none observation changes seed.

7.1. Product Usage

The Product Usage dimension tries to segment the customers based on the products usage. That is, we want to segment them based on the type and monetary amount of products bought by them.

For the Product Usage dimension, we decided to test four alternatives:

(1) IMP_MntRackets, IMP_MntWatches, MntHats, MntPremium_Brand, MntSneakers, MntTShirts

(2) MntHatsRatio, MntPremiumRatio, MntRacketsRatio, MntSneakersRatio, MntTShirtsRatio, MntWatchesRatio

(3) SOM-VQ of (1)

(4) SOM-VQ of (2)

The Mnt variables and the Ratio variables are not correlated with each other, thus, everything seems good for these combinations of variables. We chose to test the Mnt's and the Ratios, and also the SOM output of these two combinations, because we were not confident of which one would give us the best results.

For this approach, we selected no standardization, since ratios are already standardized from 0 to 1, and when considering the original values, they all use the same scale, and there is meaning in the differences of values, so when standardizing we would consider arbitrarily an equal value on the different variables instead of considering different values among them.

After doing the elbow graphic for the ratios in K-Means we understood that the best number of clusters to maintain would be 2, 3 or 4, as we can see below. We also did the elbow graphics for the Mnt variables in K-Means and SOM and for the ratios in SOM, and the results were identical.

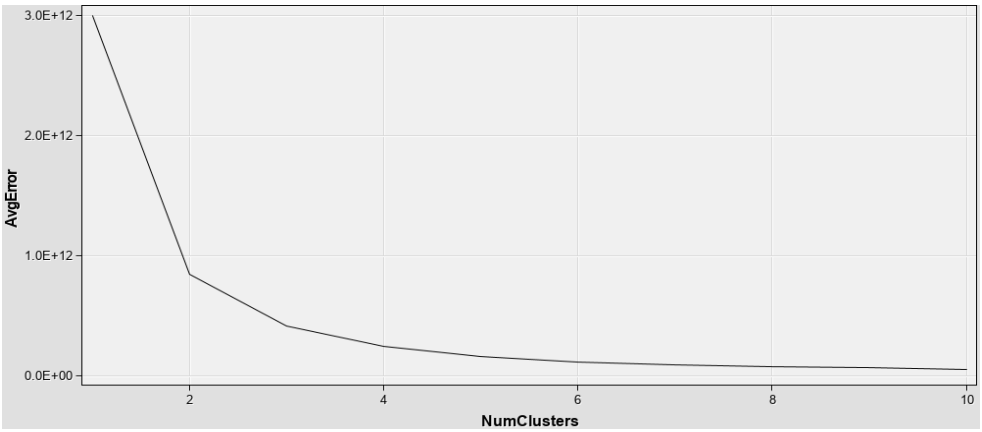

Figure 22 - Elbow graphic

After analysis we concluded that, with 4 clusters, two of them were not distinguishable, with 2 of them, we actually had the ones who buy and the ones who did not, with both including various clients who bought, but not too much.

With 3 clusters we had a better overview, and we just had to select among the various methods, since they all provided similar results as can be seen in the graphics below. At this point we decided to include the graphics with the separation variables and also with some indicators, since we decided not to include a socio-economic clustering of our clients.

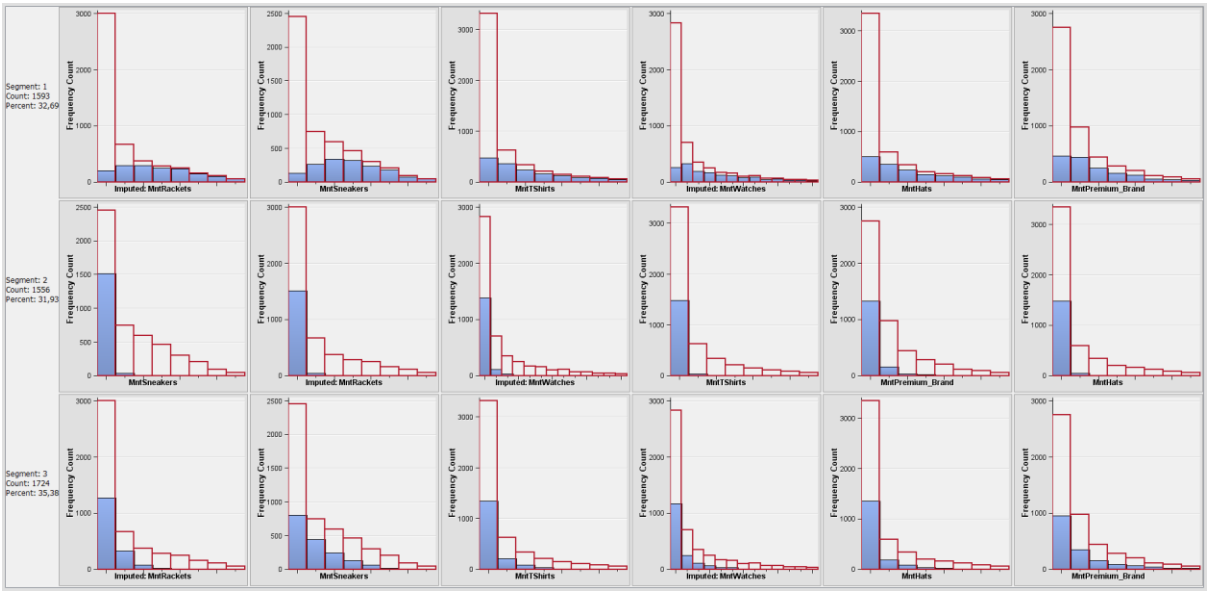

Figure 23 - Product- SOM with Monetary Variables

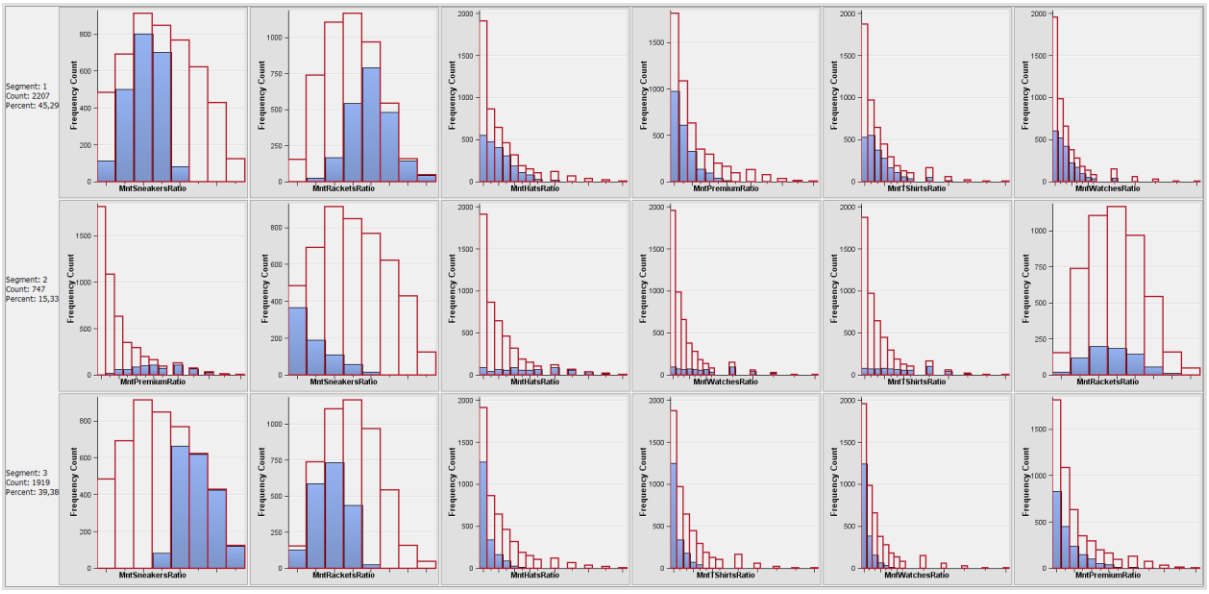

Figure 24 - Product- SOM with Monetary Ratio Variables

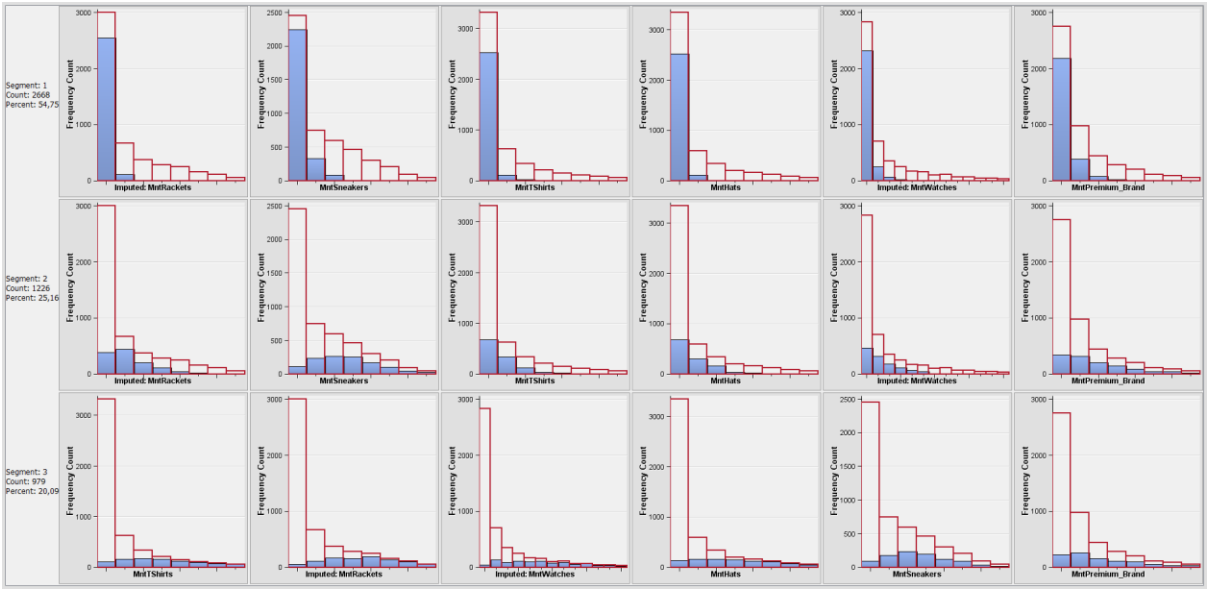

Figure 24 - Product-K-means with Monetary Variables

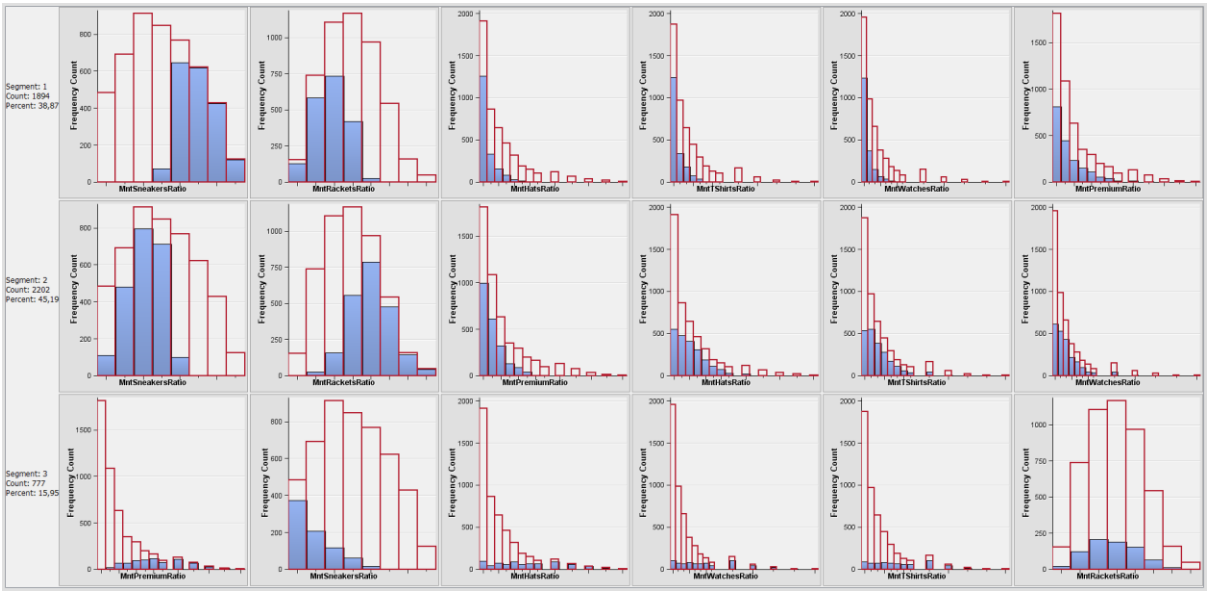

Figure 25 - Product-K-means with Monetary Ratio Variables

In order to see which method was best to move on we analyzed the graphics above to understand which discriminated better the individuals. Between the Ratios and pure Mnt's we considered the Ratios were more discriminant, and then when looking to the Ratios with K-Means and SOM we found that they were very similar, reason why we decided to go on with K-Means.

Although, in a second look in a posterior phase, we were not that sure that the ratios were the best choice, in fact, the pure Mnt's using SOM seemed also to be a good discriminative of the clusters. But considering that the ratios would give us a more interesting interpretation - which product do the clients buy the most, that we will later cross with the Value view, telling us what do the most valuable clients buy the most - than the Mnt's - where do they spend

more money - and also due to time restrictions, we decided not to change our decision. Also, we think the results would not differ that much.

In segment 1 the preferred product is sneakers, although, they also buy a substantial amount of products of the racket type. When we look at cluster 2, it is clear that rackets is the most bought type of product and finally the individuals of cluster 3 seem to buy a lot of Premium Brand products, as well as Watches, Hats and T-Shirts, reason why we decided to label it as Fashion.

Therefore, our labels for each segment were:

- Cluster 1 - Sneakers;
- Cluster - Rackets;
- Cluster 3 - Fashion.

7.2. Value Segmentation

Regarding to the Value Segmentation perspective, we tried some options of variables, to see which combination of them fits the best. That is, has a more discriminant power between the clusters.

The Value Segmentation dimension pretends to segment the customers based on their value as customers to the LaGoste company.

We started with the following variables: *AmountSpentPerPurchase, IMP_Income, TotalAcceptedCmp, TotalAmountSpent, TotalPurchases, IncomeT, NumDealsPurchases, DealsPurchaseRatio* and *Recency.*

It is to note that we also analyzed the possibility of adding the *IncomeSpentOnUS* variable. However, this was highly correlated with the *AmountSpentPerPurchase* and *TotalAmountSpent*. Before continuing our analysis, we notice that the variables *NumDealsPurchases* and *DealsPurchaseRatio* are highly correlated, so we should not have them both. The same happens with the variable *IncomeT* with the variables *IMP_Income* and *TotalAmountSpent*. These correlations have to be taken into account when choosing the variables.

Thus, the first option was: *AmountSpentPerPurchase, IMP_Income, TotalAcceptedCmp, TotalAmountSpent, TotalPurchases, DealsPurchaseRatio* and *Recency.* Our first attempt was with 3 clusters.

After seeing the results, we realized that the variables *TotalAcceptedCmp* and *Recency*, with 2 or even 3 clusters, do not have a good discriminant power as can be seen in the following figures. So, we decided to take them off.

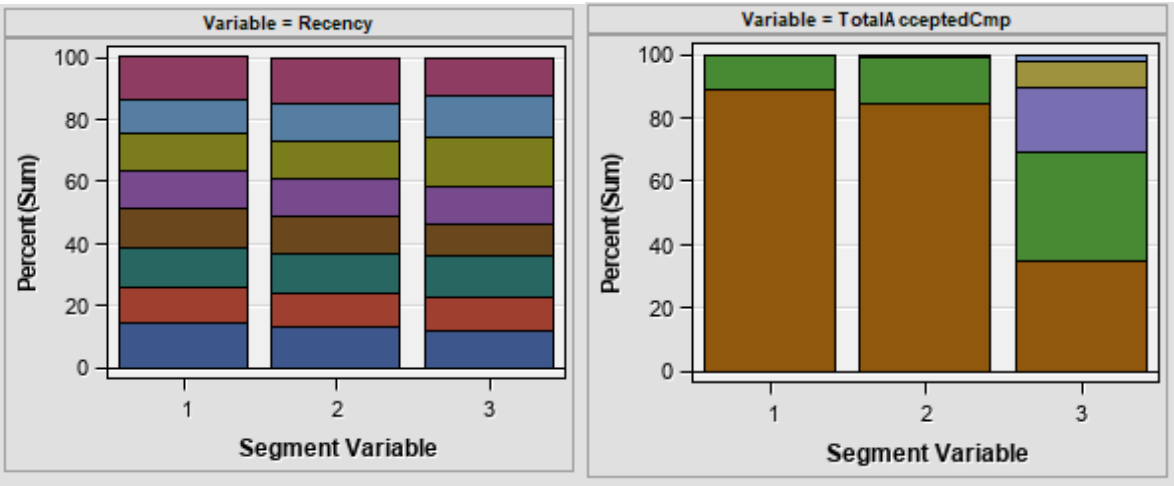

Figure 2613 - Colorful histograms of Recency and TotalAcceptedCmp

Therefore, our second option (with 3 clusters) was: *AmountSpentPerPurchase, IncomeT, TotalPurchases, NumDealsPurchases.* Here, we also wanted to test if the variable *IncomeT* would give us a good result. So, we must decide if we want to have the *IncomeT* variable, once it is correlated with some other two, namely, *IMP_Income* and *TotalAmountSpent*. After analysing the results, we concluded that the three variables seem to have a high discriminant power, however, comparing the graphics, we decided to keep *IMP_Income* and *TotalAmountSpent*. Besides that we could also see that the variables *TotalPurchases* and *DealsPurchasesRatio* were not so discriminant.

Having this, our third attempt (with 3 clusters) was: *AmountSpentPerPurchase, IMP_Income, TotalAmountSpent and NumDealsPurchases.* This one was to check if the variable NumDealsPurchases could give better results.

Looking at the results, we concluded that this combination of variables seems to be good, since it discriminates the 3 groups of individuals in a good manner.

After, we decided that the Importance variables should be relevant for this dimension, once they were formed through the RFM method. From the correlations table, we can see that only the *ImportanceA*, ImportanceRFMp, *ImportanceRFMx, ImportanceRMp* and *ImportanceRMx* variables are not correlated with the rest of the variables in the Value Segmentation perspective. So we considered them, one at a time. We tried to add *ImportanceRFMx* or *ImportanceRFMp* to the previous ones, but they were not good discriminators of the clusters, so we left them. However, when tested with the *ImportanceA*, *ImportanceRMp* and *ImportanceRMx* variables, the results were better discriminating the clusters.

Therefore, given the high correlation between them, we had to choose just one of the Importances and we decided to keep the *ImportanceRMx*, once it seemed more discriminatory for the groups of customers.

Thus, the final solution for Value Segmentation view were 3 clusters with the following variables: *AmountSpentPerPurchase, IMP_Income, ImportanceRMx, TotalAmountSpent and NumDealsPurchases.*

It is to note that we tried this option also with 2 and 4 clusters. However, the best results seemed to be with 3 clusters. This decision was also supported by the results of the elbow graphic.

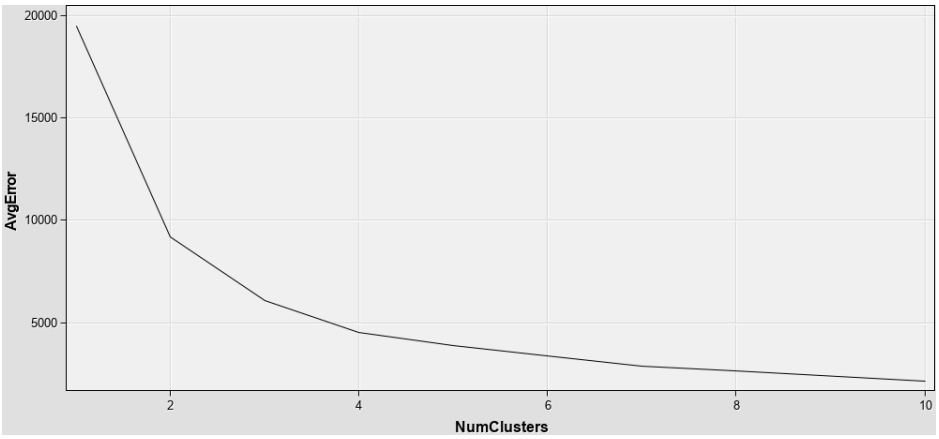

Figure 27 - Elbow graphic

After deciding the number of clusters to go with and the variables to keep, we compared the results between K-Means and SOM and concluded that they give us similar results, as we can see from the comparison of the graphics below.

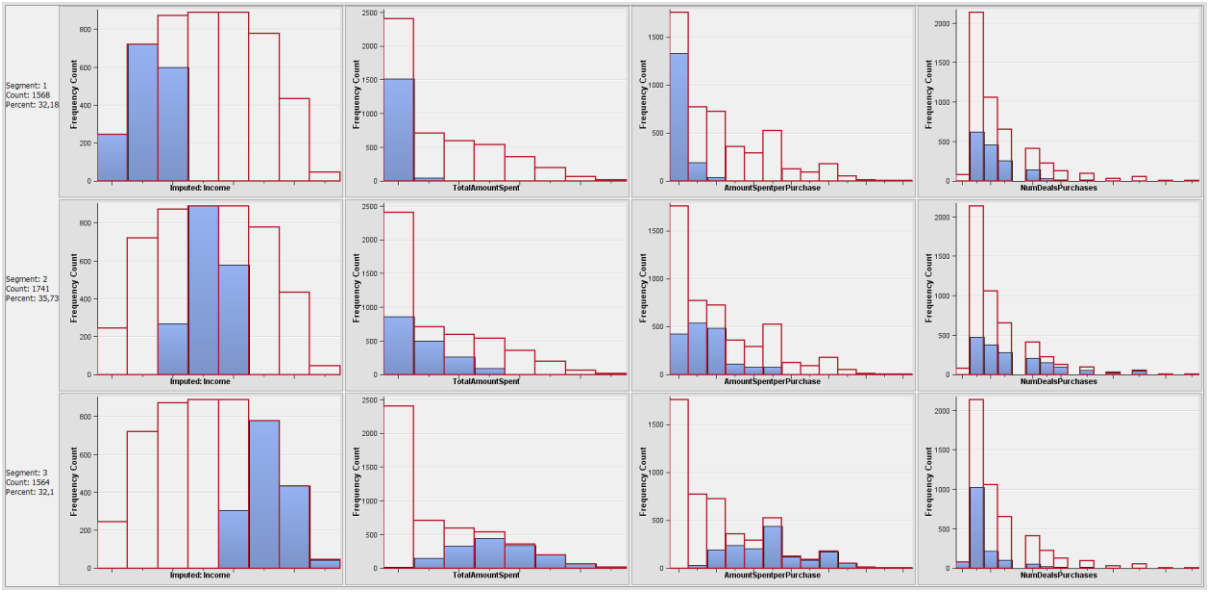

Figure 28 – Value -K-means

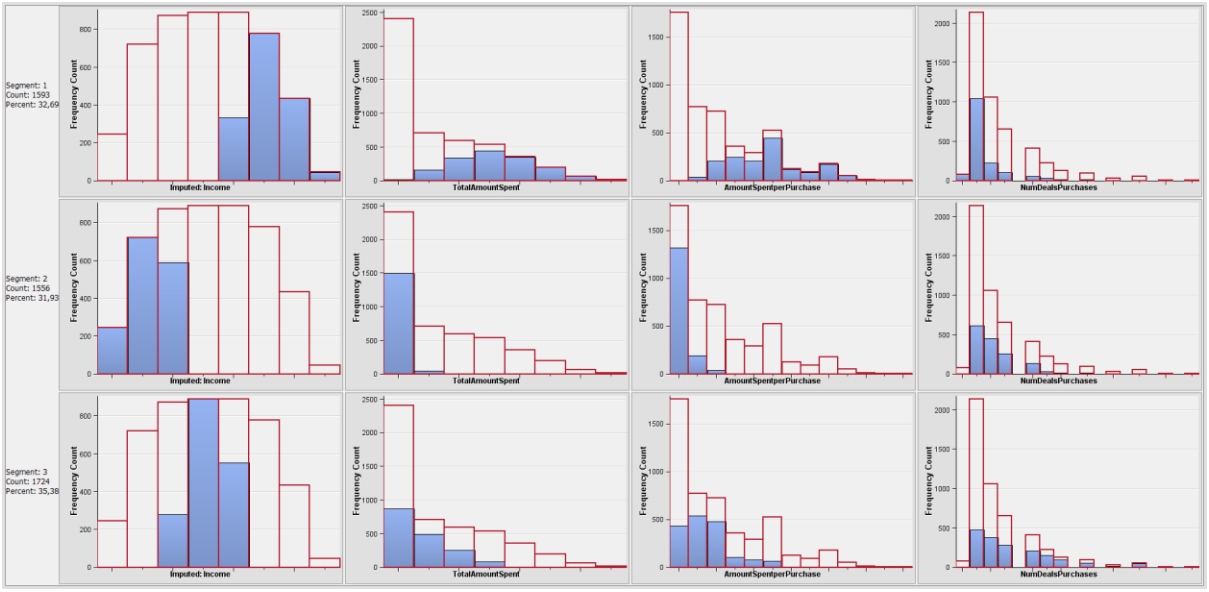

Figure 29 – Value - SOM

Thus, cluster three has individuals with a high income, a big total amount spent, and amount spent per purchase. The first segment is characterized by having a low income and the total amount spent is minimum, as well as the amount spent per purchase. In the second cluster, we have individuals with a medium income, and a total amount spent and amount spent per purchase not very high, but also not very low. These last cluster has something different from the others, which is, the bigger number of deal purchases made.

Looking at the results of the Value Segmentation, we must name the clusters. Having in mind what we just described, we can define the clusters as follow:

- Cluster 1: Bronze
- Cluster 2: Silver
- Cluster 3: Gold

7.3. Place of Purchase Segmentation

Regarding the Place of Purchase Segmentation perspective, we tried to understand if this was useful to discriminate the groups of individuals. Based on the different places of purchase, we tried a first approach using the following variables: *NumCatalogPurchases*, *NumStorePurchases*, *NumWebVisitsPurchases*. We also tried the variable *RatioWebVisitsPurchases*, but the results were not significative.

Then, to know the number of clusters we should keep, we draw the elbow graphic. After doing the elbow graphic, we decided to try this perspective with 2 and 3 clusters, and we ended up concluding that the results were better with 2.

But we could not conclude much more than, the first segment corresponds to the clients that buy less and the other to the clients who buy more, irrespective of the place of purchase. We

could see that in both clusters the place of purchase less used was the catalog and the most used was the web, but we wanted to know more and to discriminate better the clusters, so, we used different variables *CatalogPurchasesRatio*, *StorePurchasesRatio*, *WebPurchasesRatio* to see if they had a better discriminator power.

We did the elbow graphic, to have an idea of the number of clusters to keep.

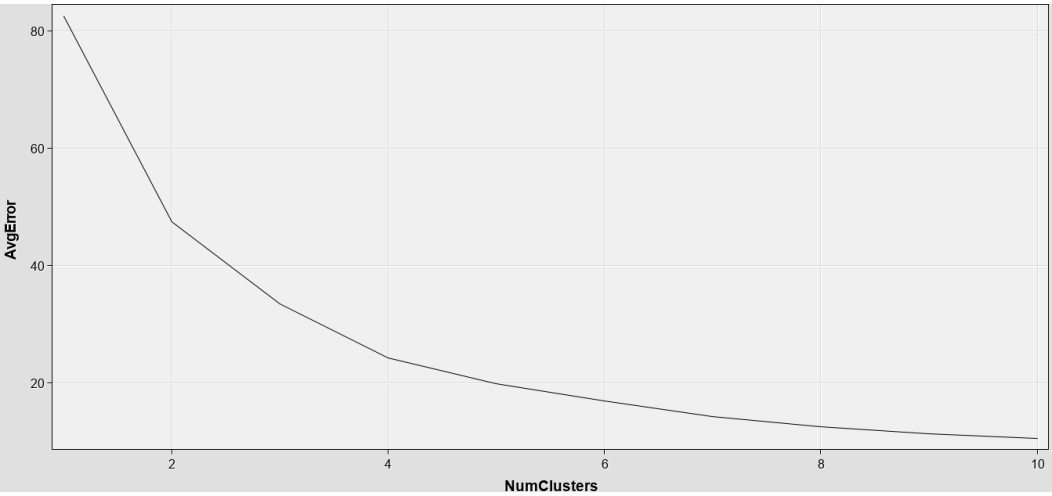

Figure 3014 - Elbow graphic

We tried to see how the results were using 4, 3 and 2 clusters and 3 was the one that gave us better results, so we decided to go with 3 clusters.

Then, we tried performing clustering using SOM and K-Means, in order to see which, one discriminated the 3 clusters better. Although, as you can see in the graphs below, they are very similar, so we decided to use K-Means.

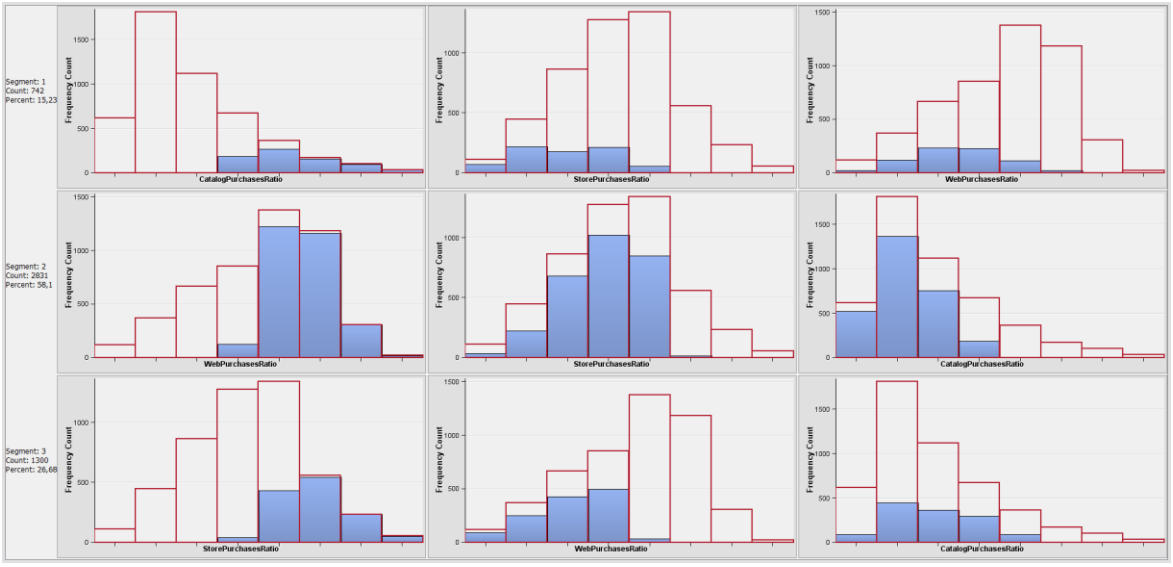

Figure 31 – Place - SOM

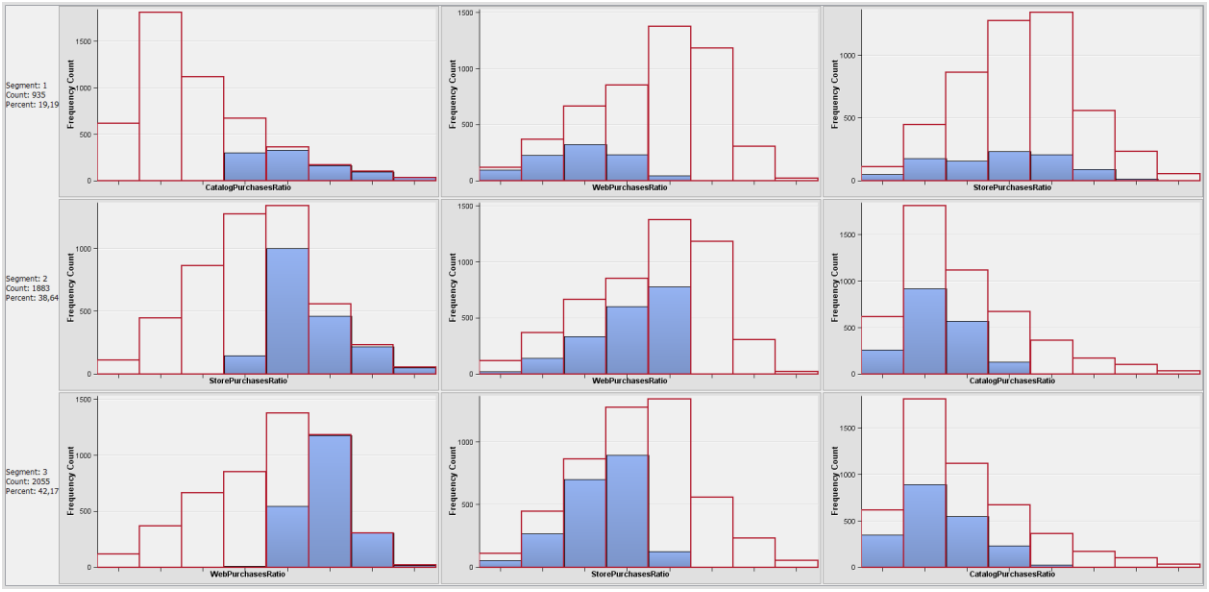

Figure 31 – Value – K-means

Looking at the graphics above, we can conclude that in segment 1 the individuals buy products using mostly the catalog. In segment 2, individuals buy mostly in the store and in segment 3 the web is the preferred channel. Taking this into account, we labelled the 3 segments:

- Cluster 1 - Catalog;
- Cluster 2 - Store;
- Cluster 3 - Web.

7.4. Social Demographic Segmentation

In this case, we wanted to perceive if the social demographic aspect is a good manner to discriminate the individuals. Here, we used different approaches (using different combinations of variables), presented below:

Option 1: *Age, Education, Teenhome, Kidhome, Marital_Status;*

Option 2: *Age, HigherEducation, TogetherStatus, TotalChildrenBinary.*

Since these variables are categorical, we cannot use the mean to evaluate how many clusters should we use and if the variables make sense to discriminate the individuals, so we used the following graphics.

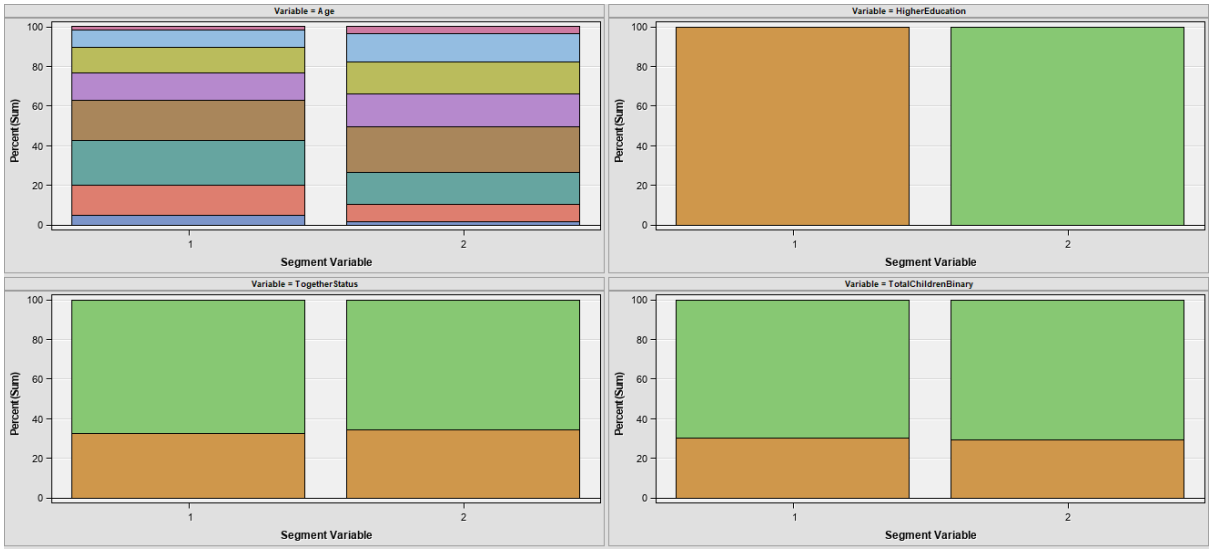

Figure 32 – Colorful Histograms of Age, HigherEducation, TogetherStatus and TotalChildrenBinary

As we can see in the figure of option 2, this was not relevant to discriminate the individuals. The same happened in the first option, and in some other combinations we tried with these variables, so we decided to exclude this perspective.

7.5. Segmentation Profiling

Here we decided to provide an identification of each of our major clusters, before dividing and crossing them into smaller clusters that imply reassignments. This way if we want some disaggregated information to see the social and monetary characteristics of major groups according to their views, it becomes possible.

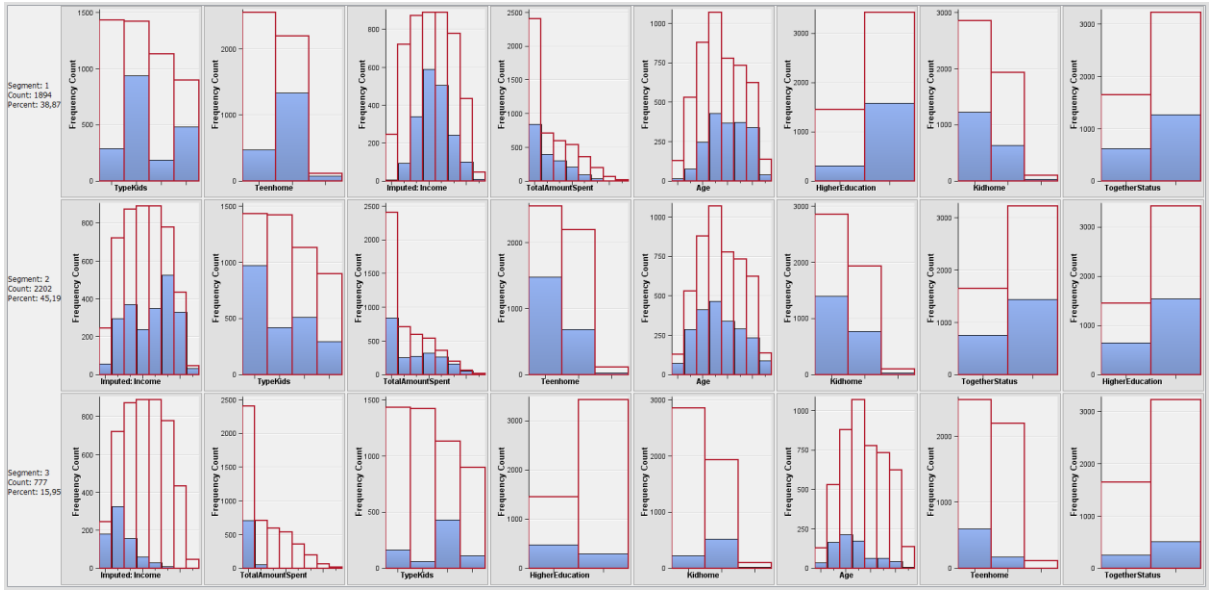

Figure 34- Sociodemographic - By product

- Cluster 1 - Sneakers - Medium income, they are our middle to low spenders and have a higher education. They tend to have less kids and more teens at home or both.
- Cluster 2 - Rackets - Tend to have the higher incomes, and the biggest spenders are here, although the spending is distributed from much to nothing. They are distributed in ages, they have a similar distribution in terms of higher education as the population and tend not to have kids or teens at home.
- Cluster 3 - Fashion - They have the lowest income and spend the least, with much lower than average high education. They are also the youngest and most of them have kids at home, but almost no teens.

They all have in common the same distribution in terms of Together status, so being with someone or not, tends not to make a difference in products they choose.

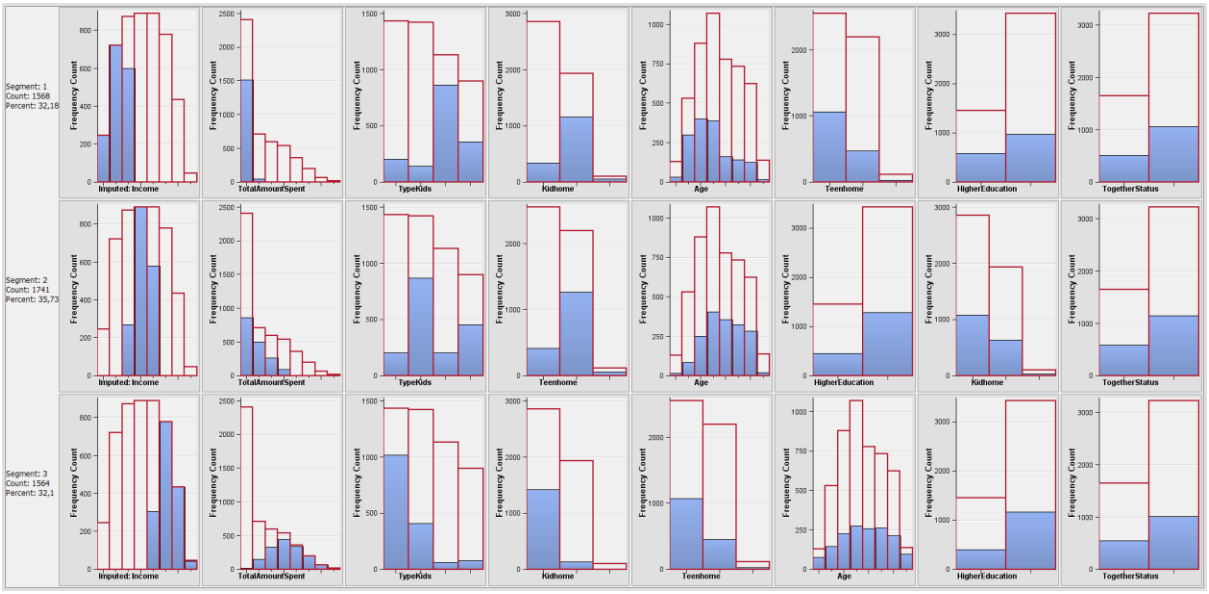

Figure 35 - Sociodemographic- By Value

- Cluster 1: Bronze - Lower income and Amount spent, they tend to be younger and less educated than the others. They have mainly kids and not so many have teens.
- Cluster 2: Silver - Medium Income and lower to medium amount spent. They tend to have the same distribution as the population in Age and Higher Education, they have some kids, but more teens or both at home.
- Cluster 3: Gold - The highest incomes, and highest Amount spent, they seem to have more people with high education than the others, but they almost do not have kids and the individuals who have teens at home are lower than average have teens at home.

  Again, the together status does not seem to change among them.

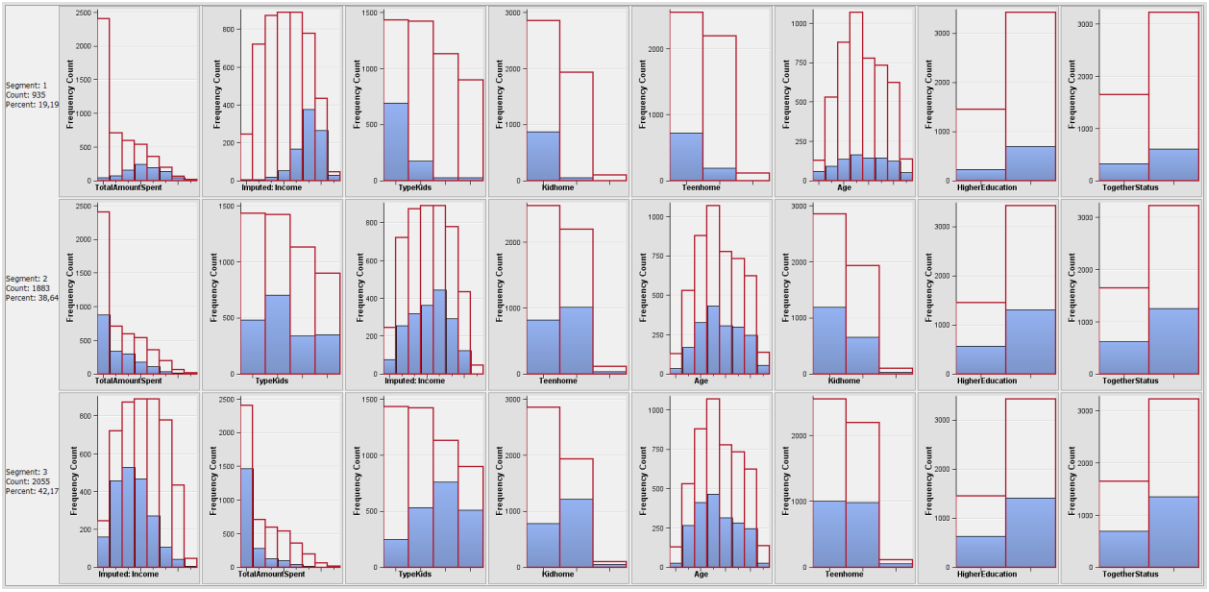

Figure 36 - Socioeconomic- By place

- Cluster 1 - Catalog - They are the highest spenders and have a higher income. They tend to be older, than the others, and mostly they don't have kids at home, while some have teens.
- Cluster 2 - Store - They spend low to medium, have distributed income and age, and some have kids or teens at home or both.
- Cluster 3 - Web - They spend the least and have lower incomes and have kids or teens or both.

Together Status seemed to be have the same distribution and neither did education.

7.6. Reassignment of Individuals to Clusters

As we ended up with 3 clusters per view, the final number of clusters after views crossover was 27, as we can see in the table below.

Clusters ID	Place Segment	Product Segment	Value Segment	Number of Individuals
1	Store	Fashion	Bronze	217
2	Store	Fashion	Gold	5
3	Store	Fashion	Silver	30
4	Store	Rackets	Bronze	334
5	Store	Rackets	Gold	288
6	Store	Rackets	Silver	241
7	Store	Sneakers	Bronze	296
8	Store	Sneakers	Gold	128
9	Store	Sneakers	Silver	344
10	Catalog	Fashion	Bronze	21
11	Catalog	Fashion	Gold	4
12	Catalog	Fashion	Silver	7
13	Catalog	Rackets	Bronze	12
14	Catalog	Rackets	Gold	558
15	Catalog	Rackets	Silver	85
16	Catalog	Sneakers	Bronze	13
17	Catalog	Sneakers	Gold	152
18	Catalog	Sneakers	Silver	83
19	Web	Fashion	Bronze	459
20	Web	Fashion	Gold	1
21	Web	Fashion	Silver	33
22	Web	Rackets	Bronze	440
23	Web	Rackets	Gold	92
24	Web	Rackets	Silver	152
25	Web	Sneakers	Bronze	380
26	Web	Sneakers	Gold	97
27	Web	Sneakers	Silver	401

Table 8 - Reassignments

We find that these are too much clusters for what is intended, so we considered joining the individuals that were part of the smallest clusters (less than 200, which you can see highlighted) with the ones in the bigger clusters (more than 200 individuals). Although we will maintain cluster 17 (152 individuals), because they represent clients of high value to the company and we consider they are worth of having a campaign directed only to them. In the preceding table the clusters to keep are white coded and the ones to be joined are in orange.

In order to do this, we built a predictive model, that received as input the variables used to create the views and as target variable the id of the cluster each individual was assigned to. Before importing this excel file to SAS Miner we deleted from it the individuals that we wanted to reclassify. These ones were presented to the model at the end of its training - this was a table without the target variable, which we wanted to predict.

We decided that it did not make sense to include the variables that we did not use to do the views as predictors, since the clusters (classification/target variable) we used resulted exactly from the combination of these variables.

We stratified the imported file in two, training (70%) and validation (30%), and after this we trained the model using Neural Networks, Decision Trees, Memory-Based Reasoning and Regressions, which gave us the following results:

Method Used	Train: Misclassification Rate	Train: ROC Index	Valid: Misclassification Rate	Valid: ROC Index
Regression	0,021	1	0,022	1
Decision Tree	0,131	0,993	0,131	0,979
Neural Network	0,328	0,985	0,341	0,977
MBR	0,577	0,891	0,684	0,799

Table 9 - Modelling techniques used

The method that gave us the best results was the Stepwise Regression, where we were able to get a Validation ROC Index of 1. In the Regression node we only changed one property, the selection criterion, that was set to Misclassification Rate, as this is what we want to minimize, and it is what is most important to us, in order to know if the model is good or not.

The regression gave us a belonging probability for an individual to be part of each cluster and at the end provided us the final cluster he should be assigned to.

The variables used by the stepwise regression were: *IMP_Income, MntPremiumRatio, MntRacketsRatio, MntSneakersRatio, NumDealsPurchases, StorePurchasesRatio, TotalAmountSpent and WebPurchasesRatio.*

Comparing the results obtained with the ones we had before, we can understand where the individuals of the smallest clusters moved to in the table below.

Count of Custid	Antes															
Depois	2	3	8	10	11	12	13	15	16	18	20	21	23	24	26	Grand Total
1	1	19		6		2										28
14				3	4	4	7	79				1	12	3		113
17			3	1				2	7	73						86
19				11		1	1					29				42
22							1						25	107		133
25									1							1
27										2	1	1	21	28	94	147
4							3	1								4
5	4		11										29			44
6		10						3				2	5	14		34
7									5							5
9		1	114							8					3	126
Grand Total	5	30	128	21	4	7	12	85	13	83	1	33	92	152	97	763

Table 10 - Reassignment (Before and After)

When doing the comparison, we found that there are some individuals, that previously belonged to the same cluster, that got splitted into more than one existent cluster. But there

are also cases where all of them or the big majority got to the same cluster as it happens with the individuals of clusters 2, 8, 11, 15, 18, 21, 24 and 26.

The ones that got separated can be explained if we think that although they were in the same clusters, they are not the same and if we were to see their distances in a map, probably these ones would be closer to a certain cluster and others to another. Even though our model has a really low misclassification rate in the validation data set (0,02), as we are using a predictive model to classify the individuals, it is also possible to have some misclassification.

93% (712/763) of the individuals were assigned to a cluster that only differs from the one where they previously were by one view, what seems to make sense, since the variables used to do the model were the ones we used for the Product, Place and Value Usage nodes.

When seeing more in detail the reassignment of the individuals we can verify that when the biggest mass moves occur, they happen between clusters that only differ in the Value view. For example, 119 from cluster 8 went to cluster 9 (gold to silver), while 107 from cluster 24 went to cluster 22 (silver to bronze). This may occur because the individuals that are moving, were more clearly defined/classified in the Product and Place views than in the Value one, where they were possibly closer to the borders of the cluster extremes (could more easily pass from gold to silver or silver to bronze). This may occur because the individuals that are moving, have a more similar behavior in the Product and Place views than in the Value one, when comparing to all the existent clusters.

We should have these reallocations in mind when thinking about the campaigns to send to each cluster.

7.7. Treatment of Outliers

In this phase we went back to the outliers that we left apart when removing them from the main dataset.

In a parallel flow to the main one we initially imputed the values that were missing to the original dataset. Then we ran the coherence check node.

Here we obtained some incoherent observations. Some of them had to do with having less web visits than purchases in the website, but this was already discussed, and to solve it we imputed the value of 0,5 for the respective observations in the variable *NumWebVisitsMonth*. But this was not the only kind of incoherence we found, we also found incoherent observations in *Incoherent7, Incoherent10* and *Incoherent12* (that can be found in the table 4). Therefore, from 127 observations, we only kept 78.

After the coherence check, the same Transform Variables node applied previously was used. Then, when analyzing the MultiPlot node we understood that there were two different people, one group with a very high income and the other with a much lower income.

Having this in mind we decided to perform K-Means with these observations, to know if these individuals have a behavior that is similar to any of the clusters we already have. These were the variables we considered to make the clusters: *IMP_Income, MntRacketsRatio, MntSneakersRatio, CatalogPurchasesRatio, TotalAmountSpent*, WebPurchasesRatio, StorePurchasesRatio.

Then, we thought it would be important to see the 2 clusters distributions in the variables used as we can see below, and here is what we found:

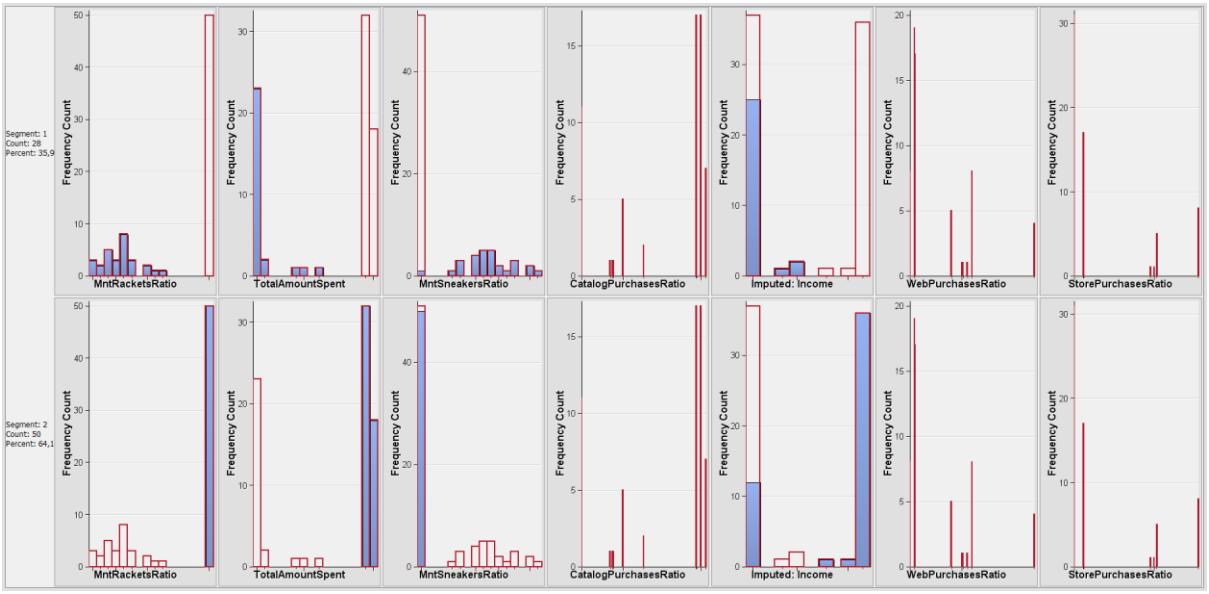

Figure 37 – Outliers K-Means

Cluster 1 - these 28 individuals are characterized by having a high sneakers ratio, so most of the money they spend in LaGoste is usually in sneakers. When it comes to rackets, the ratio is much lower. These clients have a lower income and total amount spent, and when we compare these values with the ones from clients in Bronze, Silver and Gold clusters, we see that they match with Bronze cluster clients. If we analyze the channels of purchase they use the most, we get to the conclusion they use both web and store to buy their products.

If we join these characteristics, and we compare them with the clusters we have, we can conclude that the cluster they are more similar to is 7 - Store, Sneakers, Bronze. However, as we can see on the graphics below, these individuals could also be on cluster 25 (Web, Sneakers, Bronze) since their preference is not only the store, but also the web channel. This decision was really difficult to make but at the end we kept them into the cluster 7.

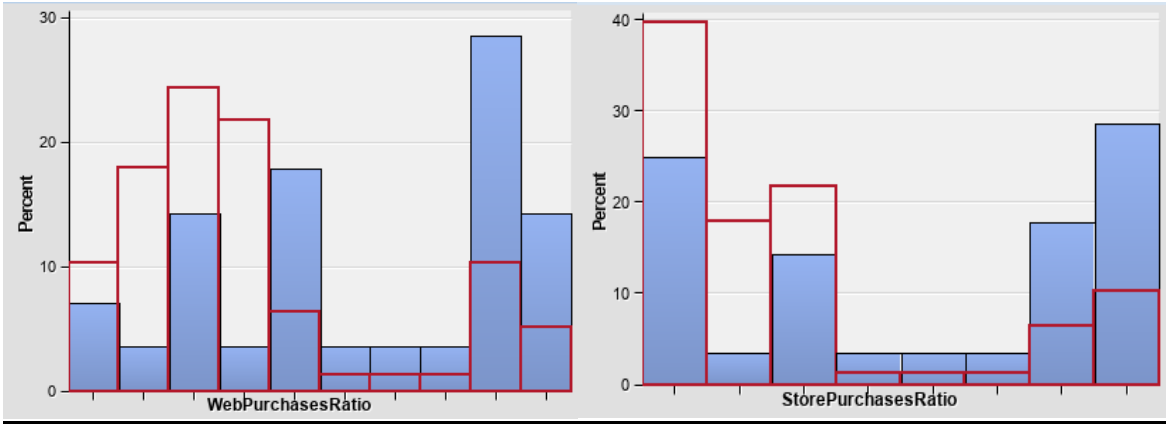

Figure 37 – Histograms of WebPurchasesRatio and StorePurchasesRatio

Cluster 2 - its individuals are characterized by having a high rackets ratio, meaning that for most individuals, this is the type of product they buy the most, a low sneakers ratio, what means exactly the opposite. Their income is very high, as well as the amount spent in the company, what tells us that they should be considered Gold clients if we think in the Value view. Analyzing the variables that are part of the Place view, we can clearly see a preference for the purchases by Catalog.

Knowing they have these characteristics we decided to put them in cluster 14, that we find to be the most similar to them - Catalog, Rackets, Gold.

8. Marketing campaigns

After defining the clusters and interpreting their characteristics, we designed marketing campaigns directed to certain groups of customers. This is essential to improve the relationship with them by understanding their behaviors and make marketing campaigns suitable to their needs and preferences, offering the best product or service to each group.

Therefore, we joined the clusters from all three different perspectives, having a total of 27 clusters (3x3x3). Since 27 is a lot, as we already said, we kept only 12 segments and made one campaign to each one of them.

Segment ID	Place Segment	Product Segment	Value Segment	Number of Individuals
1	Store	Fashion	Bronze	245
2	Store	Rackets	Bronze	338
3	Store	Rackets	Gold	332
4	Store	Rackets	Silver	275
5	Store	Sneakers	Bronze	329
6	Store	Sneakers	Silver	470
7	Catalog	Rackets	Gold	721
8	Catalog	Sneakers	Gold	238
9	Web	Fashion	Bronze	501
10	Web	Rackets	Bronze	573
11	Web	Sneakers	Bronze	381
12	Web	Sneakers	Silver	548

Table 11 - Total number of individuals per final cluster

8.1. Campaign 1 for Segment 1

This cluster is characterized by have less income and spend less (bronze), spend more in Premium Brand products, as well as Watches, Hats and T-Shirts, and buy more in the store.

Taking into account these characteristics, we decided that the best way to reach this type of customers would be giving coupons nearby the local stores. This kind of individuals, in their majority, does not have higher education. Therefore, we decide that the best way to reach them could be by giving the campaigns in hand, so they can use the offer in the store.

At a first look, it seems not reasonable that people with low income, that buy less and spend less money in the store buy premium brand products as well as watches, hats and t-shirts. Therefore, we analyse with more deepness this cluster to verify if there would be people in this cluster that as a not so low income. However, the number of this kind of individuals is low, so we conclude that really exist individuals that, besides their low income, buy premium products.

The campaign gives 30% of discount in the next purchase of any product. Thus, we want to incentive them to spend more money in the store in any kind of product.

Figure 38 - Coupons Cmp1

It is important to note that this cluster is the one with less customers and, since they are bronze, they really do not have so much value to the company. Therefore, since it is not likely that all campaigns are accepted, this could be one of the set of campaigns that did not go forward, since the value/benefit that the company can take by sending this campaign to this kind of customers may not be worth, evaluating costs and benefits.

8.2. Campaign 2 for Segment 2

This cluster is characterized by have less income and spend less (bronze), spend more in rackets and buy more in the store. Considering these characteristics, we decided that the best way to reach this type of customers would be giving coupons nearby the stores.

In this campaign we want to emphasize the 50% discount (title with "promotion" in capital letters, appealing text with high reference to the discount, image with great emphasis for 50%), because they are the ones who spend less and have less income. Was decided that the promotion is for purchases over 50€ in rackets because this is the product where they spent more money. We want to incentive these clients to spend more with us

Figure 39 - Coupons Cmp2

Here it is also to highlight that, since this cluster is composed by bronze people, this campaign could be one of the set of campaigns that did not go forward. However, here we have about 338 customers, so the decision of not making this campaign is not so straightforward, since this cluster still represents a considerable part of LaGoste clients. Although that, since we have 12 campaigns this can be left behind.

8.3. Campaign 3 for Segment 3

This cluster is characterized by having a high income and spend more (gold), spend more in rackets and buy more in the store.

Taking into account these characteristics, we define that the best way to reach this type of customers would also be with coupons in the store and nearby. This campaign is similar to the one before, however here we have Gold people.

Since these are customers who already spend more with us, the goal of this marketing campaign is to ensure that these continue with their high expenses in the company. So, in this case, we do not emphasize the existing promotion (because this type of customers buys a lot, with or without discount), but rather thanking them for being gold customers and increasing their confidence in the brand, offering a gift to them (we want this type of customers to be faithful to the brand and assiduous in their purchases). Thus, we highlight the gift offered, using a more delicate image (using golds). As these customers spend a lot, it was decided that the promotion will be only 10%, once, as stated, we just want to show that these customers are special and make them feel happy with the brand. Although they buy mainly rackets, they also buy another products, so the campaign does not need to be so targeted to rackets, as the discount is in any type of product.

Figure 40 - Coupons Cmp3

8.4. Campaign 4 for Segment 4

This cluster is characterized by have a medium income and have medium money spent in the company (silver), spend more in rackets and buy in the store.

Considering these characteristics, we define that the best way to reach this type of customers would be through coupons giving in the store.

In this campaign, we intend to encourage these customers to buy more and also other things besides rackets (which is the product where they spend the most money). We highlight the promotion, but not too much, because these clients already spent a good amount (so the promotion is only 10%). It is almost a combination of campaign 2 and campaign 3, in the sense that we want to highlight the promotion (to spend more with us), but at the same time also try to build customer loyalty, since that they already spend a significant amount in the company (silver clients). We decided that the promotion is for purchases over 80€ in rackets because this is the product where they spent more money.

Figure 41 - Coupons Cmp4

8.5. Campaign 5 for Segment 5

This cluster is characterized by have less income and spend less (bronze), spend more in sneakers and buy more in the store.

This campaign is very similar to campaign 2, since the only difference between customers in this segment and customers in campaign 2 is that they spend more on sneakers.

Here it is also to note that, since this clusters is composed by bronze people, this campaign may not go forward. This cluster is composed by 329 individuals, like the segment 2, so the decision of staying with this campaign, since the company probably will not proceed with all campaigns giving the costs, may be of not adopt, since the benefits that LaGoste may take from here may not be worth.

Figure 42 - Coupons Cmp5

8.6. Campaign 6 for Segment 6

This cluster is characterized by have a medium income and have medium money spent in the company (silver), spend more in sneakers and buy more in the store. Here, it is to note that we joined 114 individuals that had the same characteristics but were Gold. So, the campaign for these clients was created also taking a little bit into account the fact that there are 114 people Gold. However, as the only thing different is the change of Silver to Gold, the objectives of the campaign are similar for both kind of individuals.

This kind of individuals are very similar to the ones from the segment 4, they have just a little bit more people with a more advanced age. The campaign is similar to the fourth one, just changing the purchases for sneakers.

Was decided that the promotion is for purchases over 80€ in sneakers because this is the product where they spent more money. We highlight the promotion, but not too much, because these clients already spent a good amount (so the promotion is only 10%).

Figure 43 - Coupons Cmp6

8.7. Campaign 7 for Segment 7

This cluster is composed by individuals who mainly buy rackets, buy more from catalog and are Gold clients, meaning, have high income and make a higher number of purchases. Although they buy mainly rackets, they also buy another products, so the campaign does not need to be so targeted to rackets, as the discount is in any type of product. Here, we joined 79 individuals that had the same characteristics, but were Silver. As the only thing different is the change of Gold to Silver, the objectives of the campaign are similar for both kind of individuals.

It is also to note that from these individuals, only 2% have kids, almost all of them have high education and approximately 100 people have an age between 62 and 74. Therefore, our strategy passed by sending the campaign to email but also give them catalogues nearby the stores and/or sending for their homes. This people almost do not have kids, so they buy products mainly for them (or their companion). So, in this case a good campaign could be incentive them to buy more types of products, since they do not have to buy more from one specific product, as they would if they had children, for example.

It was decided that the promotion is for purchases over 100€ in rackets because this is the product where they spent more money, and we want that this type of client (spent a lot) buy other things too, besides rackets.

Figure 44 - Coupons Cmp7

8.8. Campaign 8 for Segment 8

In this segment, we have individuals who are Gold, that is, customers that have a high income and spend a high amount of money in LaGoste. This kind of clients buy mainly sneakers and they prefer to buy more through catalogues. Although they buy mainly sneakers, they also buy other products, so the campaign does not need to be so targeted to sneakers. Here, it is to note that we joined 73 individuals that had the same characteristics but were Silver. However, the only thing different is the change of Gold to Silver, so the objectives of the campaign are similar for both kind of individuals.

It is also to note that 84% of these individuals have high education and 69% of them have kids.

It was defined that the most suitable channel to send the campaign would be via traditional mail, since all these clients prefer this channel to buy.

As these customers already spend a lot of money on us, we decided that the focus of the campaign should be to try to keep this type of customers, recognizing their importance to the company, so that they feel special. As such, we focus on offering a gift and appreciation, not on the promotion itself, as referred in campaign 3 and 7.

Figure 45 - Coupons Cmp8

8.9. Campaign 9 for Segment 9

The segment 9 is characterized by people that made mainly web purchases and bought more of premium products, watches, hats and t-shirts. This kind of individuals are Bronze, that is, have low income and spend less money in LaGoste. It is to emphasize that only 41% have higher education, the majority of them have an age between 22 and 42 years and approximately 81% have kids. Thus, taking into account these characteristics, we decided that a good way to reach these individuals is certainly through web not only because they made web purchases, but also because of their age. Only 41% have high education, so the language itself used is more accessible and highlight the "Promotion" scope (with a vibrant colour).

The discount is basic giving one product for free if at least 2 purchases in Premium Brand are made. This kind of offer seems to be more beneficial since almost all these individuals have kids, so they are used to buy a more quantity of the same product.

It is important to highlight that, since this cluster is composed by bronze people, this campaign could be one of the set of campaigns that did not go forward. However, here we have about 501 customers, so the decision of not making this campaign is not so straightforward, since this cluster represents a substantial part of LaGoste clients. Therefore, maybe a good option go forward with this campaign, even if they are Bronze.

Figure 46 - Email Cmp9

8.10. Campaign 10 for Segment 10

This segment is similar to the one described earlier. Thus, this is composed by Bronze people, that is people that spend less money and have low income, so have a low value to the company. These individuals made mainly web purchases and, also mainly, of rackets. The only main difference from the earlier segment is that here there are more individuals that have high education (75%). It is to note that we joined 107 individuals that had the same characteristics but were Silver. However, as the only thing different is the change of Bronze to Silver, the objectives of the campaign are similar for both kind of individuals. Thus, the campaign is similar to the campaign 9, but is focused on the purchase of rackets.

Here it is also important to note that although this cluster is composed by bronze people, it has 573 customers. Thus, even if the investment in these campaigns lead to not proceed with all, this is one of the campaigns that maybe should remain, since this segment contains a huge part of LaGoste customers.

Figure 47 - Email Cmp10

8.11. Campaign 11 for Segment 11

This cluster is really similar to the one targeted on the campaign 5, the only main difference is the fact that they buy mainly through the web. This reinforces the fact that the best way of reaching them is by an email, since it seems that they use with regularity the internet. Thus, these kinds of individuals is characterized by having low income and spend less (bronze - low value to the company). From all products, these individuals spend more in sneakers.

It is also to highlight that, since this cluster is composed by bronze people, this campaign could be one of the set of campaigns that did not go forward. However, here we are talking about 381 customers, so the decision of not making this campaign is not so straight and direct, since this cluster represents a considerable part of LaGoste clients. But maybe LaGoste should reach them with a campaign, especially because it would be by email, so it is not so costly.

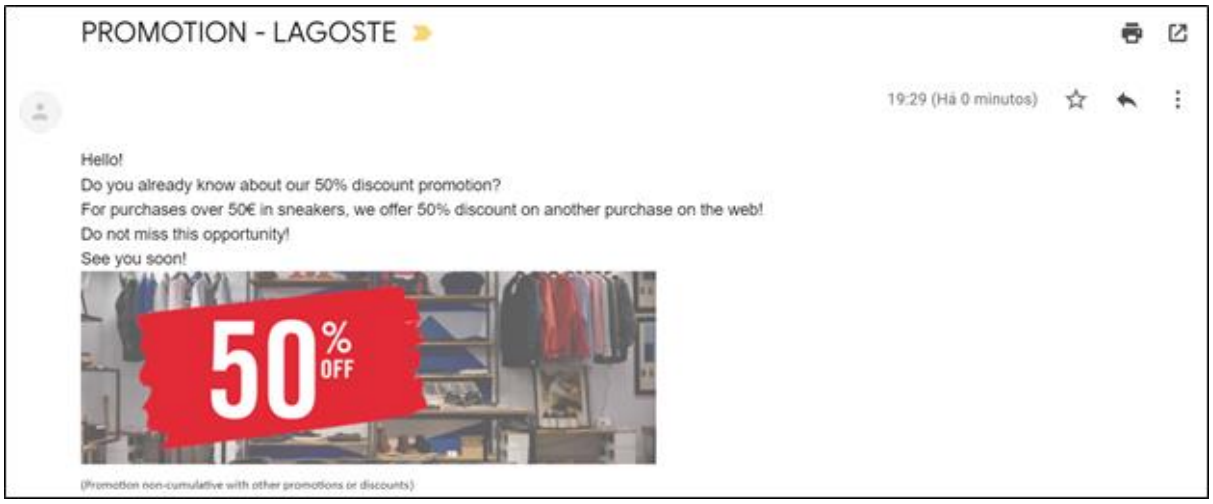

Figure 48 - Email Cmp11

8.12. Campaign 12 for Segment 12

This segment is similar to the sixth, since they also buy more sneakers and they are Silver customers (high income and high amount of money spent). However, this kind of clients make more web purchases and have more kids (43% have kids), comparing to the individuals of the sixth campaign. Here, it is to note that we joined 94 individuals that had the same characteristics but were Gold. So, the campaign for these clients was created also taking a little bit into account these people. However, the only thing different is the change of Silver to Gold, so the objectives of the campaign are similar for both kind of individuals, that is, the fact of the Gold ones receive the campaign made mainly for silver is not a problem.

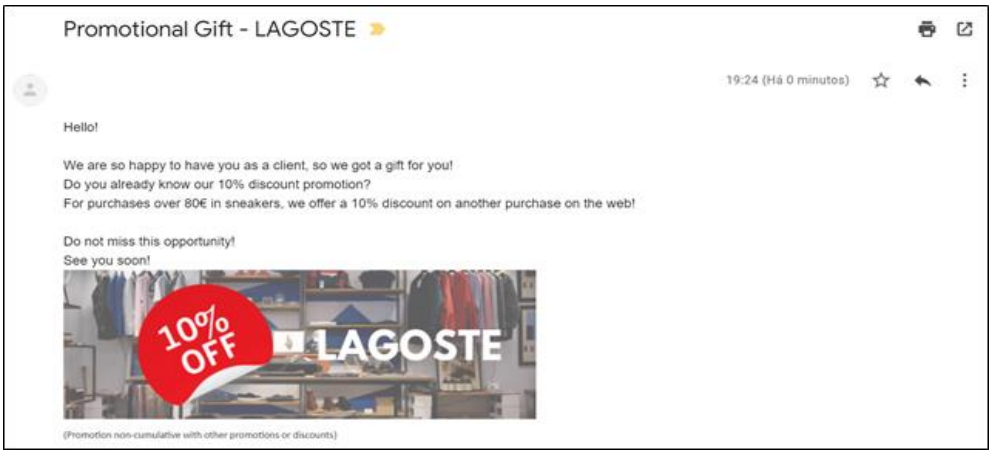

Figure 49 - Email Cmp12

9. Conclusion

With this project it was possible to identify groups of clients of LaGoste (27), in order to create marketing campaigns specific for them. However, some of those groups had a reduced number of customers, therefore, only 12 campaigns were created. We did not have a predefined budget, so some campaigns can be too costly. Having a budget, we must think in what specific groups we really want to invest (probably gold) and how much money we want to spend. The truth is that web campaigns tend to be less costly than the others, so we focus our campaigns in that channel (not only for the money but also because the characteristics of the clients - level of education, preferential channel of purchase, number of kids, age, etc), but inside the web channel, there can be other options to address a campaign, besides the email, although the personalization of the campaigns can be not as easy, as it is by email - we may also have in consideration the company website, facebook page, etc.

Regarding to the other groups, we decided to include them in the major ones using a predictive model so that they also have the chance to receive a campaign, the most suitable possible. Finally, it is to note that, if the budget is tight, we give preference to the 3 campaigns for Gold groups, given the fact that these are the most valuable clients, we want to reach and incentive them to spend more in LaGoste. Nevertheless, it is also to highlight that we have 2 Bronze segments (9 and 10) and 2 Silver segments (6 and 12) that have between 470 and 573 individuals, so it is important to discuss about these and, if the amount available to invest in these marketing campaigns can stand, maybe it could be worth going forward with these campaigns.

We also provide the rest of the marketing department with information on the specific groups but also on the larger groups not only about their patterns but also with socio-economic information, and how to clearly define them, so that they with their marketing experience can devise further campaigns from our efforts.

With this information, the company can better satisfy the clients, having more customer loyalty and, consequently, more profit.

10. References

- ✓ Toolbox, S., & Vesanto, J. (2000). *Neural Network Tool for Data Mining: SOM Toolbox*.
- ✓ Kohonen, T. K. (2001). Self-Organizing Maps. Springer Series in Information Sciences (Vol. 30). Springer-Verlag Berlin. Page 142. http://doi.org/10.1007/978-3-642-56927-2
- ✓ Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research, 174(3), 1742-1759.
- ✓ Vesanto, J. (1999). SOM-based data visualization methods. Intelligent Data Analysis, 3(2), 111–126. https://content.iospress.com/articles/intelligent-data-analysis/ida3-2-03

11. Annexes

11.1 Impute Node

Although the differentiation between types of missing values is rarely used in predictive modelling, we find it important to know if we can determine what type of missing values we have in our dataset since different strategies are suitable to different types of missing values. Missing values are divided in MCAR (Missing Completely At Random), MAR (Missing At Random) or MNAR (Missing Not At Random). Although the names are similar there are nuances among them.

Missing values are considered MCAR, when the fact that there is a missing value, is not dependent on other variable values, being actually independent, and knowing the value of another variable does not gives us information on the actual missing value, although we can sometimes construct a model to try and estimate the missing values. These missing values typically derive from a problem in data collection or storage.

The main difference to MAR is that here, the value is missing not in a random way but actually with a relation to another variable or variables, or with a specific reason. As an example, on a questionnaire a question can be skipped depending on the values of previous questions, but we cannot directly determine the missing value. For instance, if at the time when the data was collected, we were not interested in knowing the total money spent, on a certain good, by people with an income above 100 000€/ year, as they were not part of target population, we know it is related with the income, but we cannot know the missing value.

When we have missing values from MNAR type, we know the values are related with a variable or a set of variables and we can directly know the value. We do not ask if a family with yearly income above 100.000€ /year would have state granted family support since we

know that they are not eligible. In this case the missing values could be replaced by 0. In this case if we used the average, median or other replacement we would be causing a bias in our data.

If we consider the missing values to be MCAR, we can remove the observations, use some substitution like the median, mean, a matching distribution or another imputation method, since the missing values are random. If they are considered MAR or MNAR these substitutions will likely introduce a bias in our model. In all of these cases, advanced imputation should be used, the method applied will be dependent on the particular dataset.

We considered that in the 3 variables where we have missing values (*Income*, *Mnt_Hats* and *Mnt_PremiumBrand*), these are all considered MCAR. In the case of *Mnt_Hats* and *Mnt_PremiumBrand* we first thought that maybe the missing values corresponded to people that started being our clients when these kinds of products weren't sold yet. In this case we would have missing values because columns for the variable would not be available in the database at the time of entry, making these a case of MNAR. When consulting the dataset, we noticed that these missing values had different dates for client creation, so the hypothesis was rejected, and we started treating the variables as MCAR.

We were then faced with the decision of which method to use to impute the variables. The mean is the first option but is not as robust as the median. The median although more robust than the mean, is subject to the same problem, imputing to many values with the mean can cause an increase in a single point of the distribution, causing errors in our model, not in regression algorithms, but mainly in neural networks.

The following best practice is to assign random values which would not change the distribution, but in some cases, we know we can get better results by using other variables (except the target) to predict the missing values, using a tree classification. We decided on this last option. Although SAS does not explicitly how he does the final imputation when the leafs are settled, it may use the mean, median, distribution or predictors weights, using a small regression to obtain the final estimate. This method also has the advantage of tending not to change the distribution of data.

Care must be taken that data that is presented to the model for prediction, should be subject to the same method of imputation, and we should be sure that the reason for missing values remain the same.

Other advanced methods are available, but we decided not to study them at this point.

11.2. SOM/Kohonen - SOM/VQ

The Self-organizing map (SOM) is an efficient data visualization technique, with the objective of reducing the dimensions of data using unsupervised self-organizing neural networks.

"The problem that data visualization attempts to solve is that humans simply cannot visualize high dimensional data as is, so techniques are created to help us understand this high

dimensional data. So SOM networks accomplish two things: they reduce dimensions and they display similarities of observations in those dimensions" (Customer Segmentation and Clustering Using SAS Enterprise Miner, Third Edition).

The SOM is a combined vector quantization and vector projection algorithm. Basically, this method consists of having an organised grid of neurons and each one is represented by a vector with specific weight that is equal to the dimension of the input vector. Therefore, for each dimension of the input vector, the neuron has a certain weight (that is the distance to the origin). The neurons are connected to adjacent neurons by a neighbourhood relation, which dictates the topology or structure of the map.

"The SOM network consists of an input layer and the Kohonen layer. The Kohonen layer is usually designed as a two-dimensional arrangement of neurons that maps N-dimensional input to two dimensions preserving topological order. For the purpose of identifying cluster membership, it is used a one-dimensional Kohonen layer. The SOM input layer of neurons is fully connected to Kohonen layer." (Mangiameli, Paul, Shaw K. Chen, e David West. «A Comparison of SOM Neural Network and Hierarchical Clustering Methods». European Journal of Operational Research 93, n. 2 (Setembro de 1996): 402–17.)

Defining the topology: "If one wants to obtain an approximately uniform lattice spacing in the SOM, the relative numbers of cells in the horizontal and vertical directions of the lattice, respectively, should be proportional to the two largest eigenvalues considered above." (Kohonen, T. K. (2001). Self-Organizing Maps. Springer Series in Information Sciences (Vol. 30). Springer-Verlag Berlin. Page 142. http://doi.org/10.1007/978-3-642-56927-2)

So in order to define the grid size, the first step is to define the proportion of sizes. According to Kohonen, the PCA is already an approximate solution to the SOM/Kohonen grid, and to obtain a better convergence and a better visualization, the proportion of the grid should be proportional to the sizes of the two principal components, the ones that keep most of the variance. As an alternative, if PCA is not done, a good approximation is to use Root of N, as grid size and defining a rectangular or hexagonal grid.

Another parameter to consider is the Neighborhood, or how many neurons are influenced at each presentation. This exhibits a pattern of cooperation, when one neuron pulls the others, allowing for a more efficient grid adjustment.

An important definition is the initial learning rate and the number of epochs to consider, since these two are linked together. The learning rate defines the ratio of information kept and information learned in each presentation, and the number of epochs will define the decaying rate of the learning rate. Why a learning rate and why decay it? Since at the beginning we start with random positions of the neurons, we want to rapidly adjust to our data, so we use a high learning rate(low information kept/ high information learned), but as we adjust, we want to keep more and more of positioning learned and want to start fine tuning the positions.

How fast and how low and what is beginning and end is given by the number of epochs to present the data. We want to do various epochs, with points being presented in random order. If we used only one epoch, the points presented at the beginning would have much higher importance to network definition than the last ones and there's no reasonable

explanation to this, by presenting points multiple times in different orders, we obtain fairly equal importance to each data point in terms of grid definition, and consequently better data representativeness, thus obtaining a better network. Thus, the decaying rate can be set to learning rate, divided by the multiplication of data points and epochs.

The neighborhood should also be decayed in a similar way, as they learn more and more they want to have less influence from their neighbors.

In terms of method, the Kohonen layer computes the Euclidean distance (or other if necessary) between the weight vector for each of the Kohonen neurons and the input pattern. The Kohonen neuron that is closest is considered the Best Matching Unit (BMU) with an activation value of one while all other neurons have activations of zero.

Then the neuron is pulled to the data point according to the learning rate, and the neurons in the Neighborhood are pulled towards the BMU, and then the learning rate decays.

We repeat this process, until all data points are presented to the network, what marks the end of the first epoch. We repeat the process with the different epochs with random allocation of data points presenting order, and in the end we see if the solution converged and if somewhere the neurons stopped moving.

It exhibits various mechanisms of self-organization, neurons compete to be BMU, they cooperate by pulling their neighbours closer to the data, and they self-amplify by coming closer to the data, thus being selected more times as the BMU and representing more data points.

Method application:
Variable selection
We selected variables that could be of importance considering the different channels and views and ended up selecting the same variables we had selected in each prior view for comparability of solutions.
Method: SOM/Kohonen initially selected, since we had no a priori knowledge of the number of clusters Vector quantization (VQ) was not usable, and batch was not necessary since enough computing power was available and the number of points was not massive.
But after analysing the maps produced, we could not discern visually the number of clusters to apply, so we decided to change the method to VQ, selecting the number of clusters as 3 for each view, the same we had obtained in k-means.
We did not use standardization as explained in the k-means, since in each channel they were already normalized or contained relevant information in the scaling and the scale was consistent.
Since we had no a priori knowledge, a wide net was necessary, to allow some units to be attached to possible outliers and allow elasticity to be enough to better cover the data and to have enough stretch on different zones of clients.
Since we had 5000 clients, we started with the square root of N (5000->70) to define a starting ballpark for the number of nodes, arriving to a value of 70, which would result in a 9 by 9 or 8 by 8, if using an equal proportions network.

According to kohonen the proportion on nodes on each direction could be directed by the ratio between the two principal components in a PCA assessment, since we did not use the PCA, there is no reason for a different size side should be chosen when defining the network especially if the seeds are allocated random starting values, we could use an 8x9, that would guarantee the necessary 70 nodes. Since SAS only accepts predetermined values (2,4,6,8,10,20,40,60), we could use an 8x10 network in the initial assessment, but there is no reason to use an 8x10 instead of 10x8, so to allow for the same spatial resolution in both dimensions, we decided to use a 10x10 network to start with. Sas seems to only consider quadrangular networks, although an hexagonal network can exhibit more interesting properties as the resistance to folding.
For initial seeds on the first approach, we selected the outliers method. This way if outliers still exist, that will make us detect them faster, allowing for a faster dataset purity and faster cluster detection, by allowing more nodes and higher elasticity in the cluster differentiation. They can also guarantee greater initial coverage of the network. In a second approach for robustness we can see if there are differences between these initial seeds and and the separate method, that provides for not so extreme initial seeds.
Since they are defined by the outliers, there is no sense in assuming a minimum distance between seeds.
When we ran the algorithm we could neither discern the number of clusters, neither their radius, which is an identified difficulty in SOM.
So, having already established the number of clusters using the elbow and CCC (cubic clustering criterion) over the hierarchical clustering we decided to implement VQ, with 3 clusters to see if we could get better solutions, since k-means is greedy and does not use cooperation between seeds to establish a better solution.
Vector Quantization with the SOM algorithm allows for an unsupervised use of the Learning Vector Quantization(LVQ). LVQ can be described as partitioning the space in distinct regions and have a vector represent each of the distinct reasons. The SOM applied to LVQ allows for the points to self adjust to the space, thus creating the best possible vectors. They reap the benefits of unsupervised learning through self organization from SOM, and the benefits of simple assessment of final results from the VQ, but they need to have prior knowledge on the number of segments to create.
The results were discussed previously in the various views.

Advantages
- They are conceptually easy to understand.
- SOMs tend to work very well.
- In SAS Enterprise Miner, the profiling portion is very similar to the clustering technique

Disadvantages
- SOM networks can be prone to issues with missing data as in all other neural network algorithms and regressions.
- SOMs can produce differing results as they produce maps from sampled data so it may take a number of trials to obtain a map that is consistent with the same training data.
- They are rather computationally intensive.

We also provide the rest of the marketing department with information on the specific groups but also on the larger groups not only about their patterns but also with socio-economic information, and how to clearly define them, so that they with their marketing experience can devise further campaigns from our efforts.