|  |
| --- |
| C:\Users\Cristiano\Downloads\Logo.png |
| Data Mining |
| Segmentation Project |
|  |
|  |
| **Professors:**  André Melo  Roberto Henriques  Vasco Jesus  **Students :**  Illya Bakurov (R2012236)  Pedro Saraiva (R2012262)  Tautvydas Zaliauskas(E2014357) |

Contents

[Introduction 3](#_Toc409455088)

[Executive Summary 3](#_Toc409455089)

[Analysis and Clients Segmentation 3](#_Toc409455090)

[Exploring Data 3](#_Toc409455091)

[Filtering 4](#_Toc409455092)

[Variables transformation 5](#_Toc409455093)

[Clustering 6](#_Toc409455094)

[Product 8](#_Toc409455095)

[Value: Value\_Trans+Gender procedure 13](#_Toc409455096)

[Value: Value\_Pure 18](#_Toc409455097)

[Value proposition for different segments 22](#_Toc409455098)

[Big\_Bronze value segment, one generalized step. 22](#_Toc409455099)

[Particular cases 22](#_Toc409455100)

[Conclusion 23](#_Toc409455101)

[Annexes 24](#_Toc409455102)

[Exploration of Data Quality Issues 24](#_Toc409455103)

[Filtering nodes 24](#_Toc409455104)

[Bibliographic References 30](#_Toc409455105)

# Introduction

In this report you will face a Segmentation Analysis of a sample of 6000 randomly selected clients from Sell Everything supermarket database made by our agency, I.ZÉ.GI, using statistical and analytical tools provided by SAS Enterprise Miner. This document contains two main parts with a description of the technical process used to produce the results: identification of segments and value proposition for each segment.

# Executive Summary

Given a population sample of 6000 clients, we made a segmentation analysis, from which identified 11 segments of, characterized them and also developed a personal approach to the clients in each segment in order to provide the best personalized offer.   
 This process passed through Data Exploration when we get familiar with data, Filtering where we filtered outliers, Variables Transformation where we transformed variables to get reliable indicators in analysis and made Clustering where identified different clusters and joined them.   
 By the end we made a value proposition for retained segments of clients.

# Analysis and Clients Segmentation

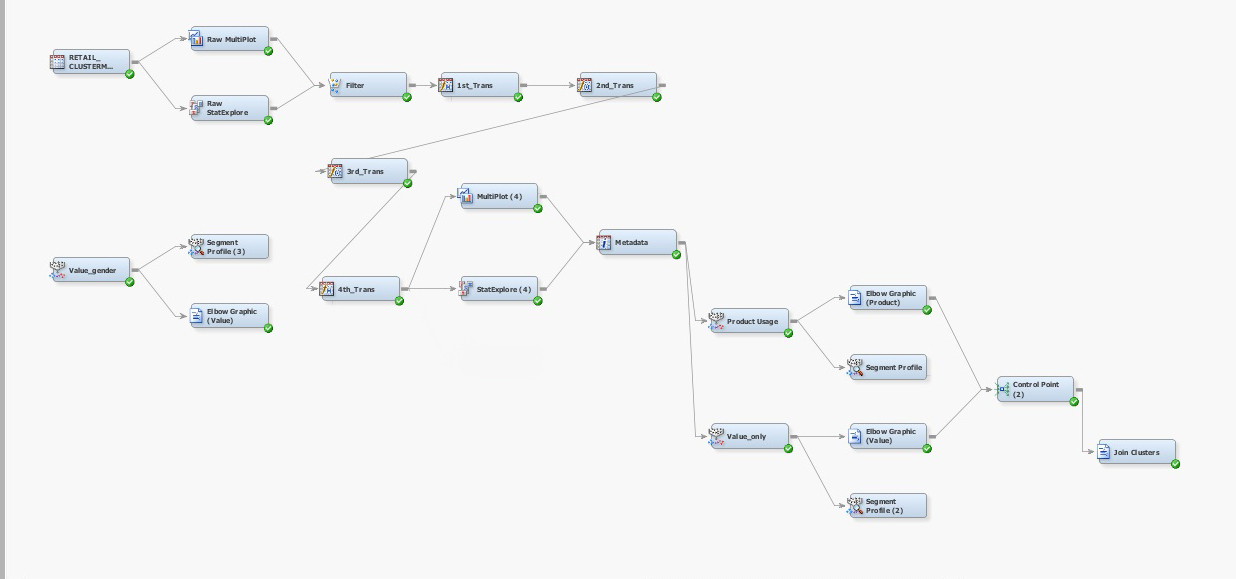


Figure 1 - Model Diagram

To get a tangible and well-marked idea about the clients, group them into similar groups in order of similar characteristics, identify their behaviours and estipulate their needs to provide a better offer, that’s the point of Analysis and Clients Segmentation. To do that, we started from exploring and understanding data that we should manage.

## Exploring Data

First we made a quick review of data, where became familiar with variables in our data-set, their meaning and type. Then, using Stat Explore and Multi Plot nodes, we explored each variable, looked at their distributions, ranges, means and identified some data quality issues.

After exploring data we get two news, one was good and another was bad: the good one was that there were no missing values and the other, the bad one, was that some variables presented strange values that we should manage ([see annexes](#_Exploration_of_Data)). These variables were the following:

* *M\_Clothes*: 4 observations presented negative values;
* *M\_Personal\_Hygiene*: 1 observation presented negative value;
* *M\_Plants\_Animals*: 1 observation presented an approximation of zero;
* *M\_Toys*: 3 observations presented negative values;
* *Monetary*: 1 observation presented an approximation of zero;

*M\_Computers, M\_Sound, M\_Entertainement* and *M\_Sports\_Outdoor* product categories presented, individually, a very low number of costumers, namely *M\_Sports\_Outdoor* - it presented only 1 client who spent only 9,99 euros.

Also we identified that the *Monetary* was not corresponding to its true values. That is, the sum of expenses for each client in every product category was not corresponding to amount spent by each customer during last 52 weeks stored in variable *Monetary*. To manage that we applied a transformation to get true values of expenses for each customer, we will describe this procedure in Transformation Topic.

## Filtering

Having some knowledge about data quality and using *Filter Node*, in this second step, we applied a set of filters to some variables in order to exclude outliers from our analysis and to ensure one important thing: clean cluster analysis.

It´s important to underline that the presence of outliers, using *k-means algorithm*, can easily deflect the *centroids* of clusters which leads to incorrect segmentation of individuals. That´s why we paid special attention to observations strangely far away from the main distribution.

What is regarding to strange values identified during Data Exploration step, we assumed that those values symbolized returned articles in each category of product, that is, articles bought in the year preceding the one when data was collected and delivered in the next year. So, we decided to reset all of them to zero because in fact there is no outcome of money in those product categories, just a compensation for money spent in previous period. The details will be explained further in Variables Transformation step.

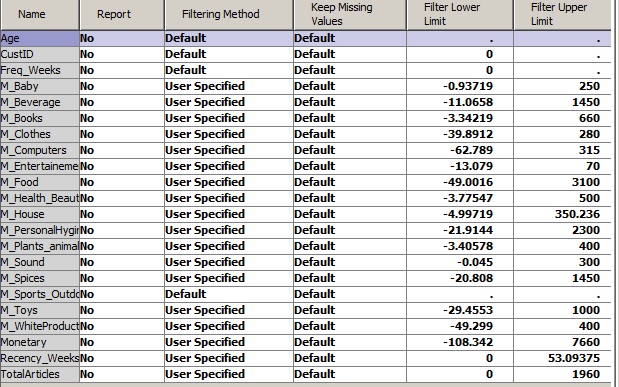
* In the table below, you can find lower and upper limits established for filtering data by looking at distributions in Interactive Interval Filter ([see annexes](#_Filtering_nodes)). Note that we established no lower limit due to further transformation of negative values and all the upper limits were defined taking into account possible outliers. Also we established no *Default Filtering Method*.

Figure 2 - Filtering *Sell Everything* outliers

* Variables such*Age, CustID, Freq\_Weeks, Gender, Recency\_Weeks, M\_Sports\_Outdoor*, suffered no filtering.
* In case of *Age, Freq\_Weeks and Recency\_Weeks*, these variables do not present any outliers in our consideration because we found no observations strangely distant from main distributions;
* It’s not meaningful to filter clients due to their ID - it is just a label to identify each one of them in the database;
* In case of *Gender,* as we have only 2 modalities - male and female and both are well represented.
* In case of *M\_Sports\_Outdoor*, as the variable representing one product category includes only 1 buyer, we decided not to filter it in order to make further transformation, explained in next step.

After whole process, we excluded 189 clients from 6000 sample what corresponds to 3.15%. Thus, now we have total 5814 observations.  
 Despite of ignoring common recommendation of 3%, we fully understand that a consistent cluster analysis is possible if we treat the outliers threat; in our case we removed from the analysis clients that are viewed as very important for the business. The way as we made filtering - to ensure that clusters will be as accurate as possible, required also to loose relevant information for the enterprise, and we kept this in mind while filtering.

## Variables transformation

Given the *Filtering Node* outputs we finally could proceed to the next part where we could invest more with our critical thinking. In topics below you can see which variables suffered transformations and why.

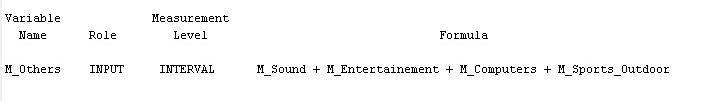
* All the variables that presented negative values were transformed, such that all negative values became equal to zero. To do that, we used SAS Code option in Transformation node and applied the following code for every product type variable:
  + *if variable\_name<0 then variable\_name=0*;
* *M\_Computers, M\_Sound, M\_Entertainement, M\_Sports\_Outdoor* were merged cause we identified a low number of clients buying each one of this type of products, presenting a low individual value for *Sell Everything*. The new product type called was *M\_Others* and is representative now.
  + *M\_Sports\_Outdoor* was not excluded from the analysis despite of having only one client in whole 6000 database sample spending only 9.99 euros, we decided to merge it with other less relevant variables. Initially we thought about excluding it due to assumption of random and representative sampling, if this assumption was ensured while extracting this sample then this product type seems to be not only irrelevant for segmentation but also could be irrelevant for business.

Figure 3 – Transformation 1

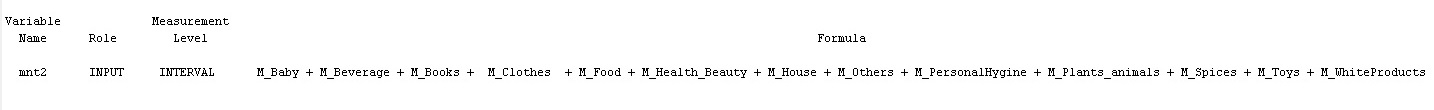
* As we mentioned before, *Monetary* was not corresponding to its true values. To manage this problem we just summed all product categories into *mnt2* variable.

Figure 4 - Transformation 2

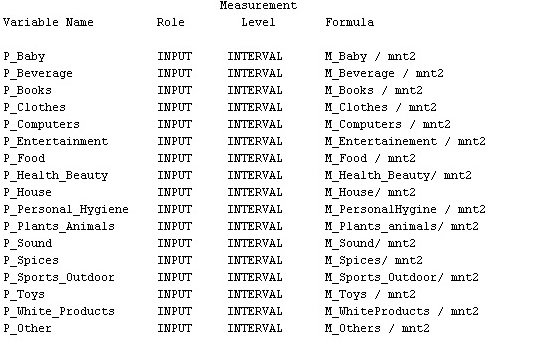
* In order to easily understand where clients spent more money we decided to rewrite all product type expenses in order of proportion, that is, each product type variable was divided by total amount spent in given time period achieving by this the proportion from overall money spent in each type of product.

Figure 5 - Transformations 3

* + In order to test if *mnt2* transformation was made correctly, we created a new variable - t*est*, where we summed all proportions representing expenses in each product category and verified if the sum was equal to 100. Thus, we understood that transformation was made correctly
* Moreover, we created a variable *Value* to easily measure the value of each customer for the enterprise and provide their segmentation. That is, in our opinion having one simple, understandable and worthy variable could benefit our segmentation and further description of segments. To do that we executed the following operations:
  + *Monetary, Freq\_Weeks, Recency\_Weeks* suffered *Range Standardization*.
  + Then, range standardized variables were weighted in order to their relative importance for this type of analysis. Also *Recency* was transformed in order to provide similar interpretation to other two variables used in formula: when client went to supermarket recently, variable will show values near to 1, thus, we made difference between 1 and range standardizer *Recency* value.
    - 0.5 x RS*Monetary* + 0.3 x RS*Freq\_Weeks* + 0.2 x (1 - RS*Recency\_Weeks*)
  + We choose to make weighted formula because we think that one variable from RFM analysis is more important than another for defining clients value. So, we gave the highest weight to *Monetary* because we think, in this study, it has higher influence comparing with others. Moreover, our reasoning is supported by additional study ([see annexes](#_Bibliographic_References)), where authors concluded that these weights for RFM analysis are an optimum choice, in this type of studies.

## Clustering

Figure 6 - Transformations 4

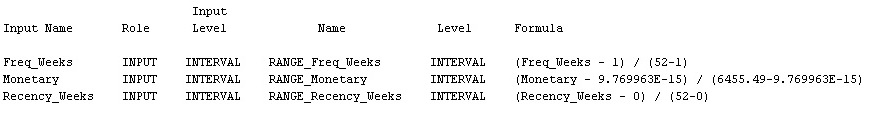


Figure 7 - Transformations 5

Variables used affect the quality of cluster analysis, that´s why we paid special attention to their selection in order to be evolved in theoretical context and provide ones that have higher ability of discrimination.   
 We developed 3 clustering procedures:

* *Product*: using percentage of expenses in each one of product categories;
* *Value\_Trans+Gender*: using created indicator displaying the *Value* of each client to *Sell Everything*, number of *Total Articles* bought, *Age* and *Gender*;
* *Value\_Pure*: using *Age, Frequency, Monetary, Recency*, number of *Total Articles* bought.

|  |  |  |
| --- | --- | --- |
| Product | Value\_Trans+Gender | Value\_Pure |
| P\_Baby | Age | Age |
| P\_Beverage | Gender | Frequency |
| P\_Books | Total Articles | Monetary |
| P\_Clothes | Value | Recency |
| P\_Computers |  | Total Articles |
| P\_Entertainment |  |  |
| P\_Food |  |  |
| P\_Health\_Beauty |  |  |
| P\_House |  |  |
| P\_Personal\_Hygiene |  |  |
| P\_Plants\_Animals |  |  |
| P\_Sound |  |  |
| P\_Spices |  |  |
| P\_Sports\_Oudoor |  |  |
| P\_Toys |  |  |
| P\_White\_Products |  |  |
| P\_Other |  |  |

Figure 8 – Representation of Variables used for each segmentation perspective

It is important to say that we developed two distinct procedures for the same sort of clustering (last two) in order to understand which will provide us more quality. Further, in Value section you will face the comparison and conclusions we achieved.

### Product

#### Elbow Graphic

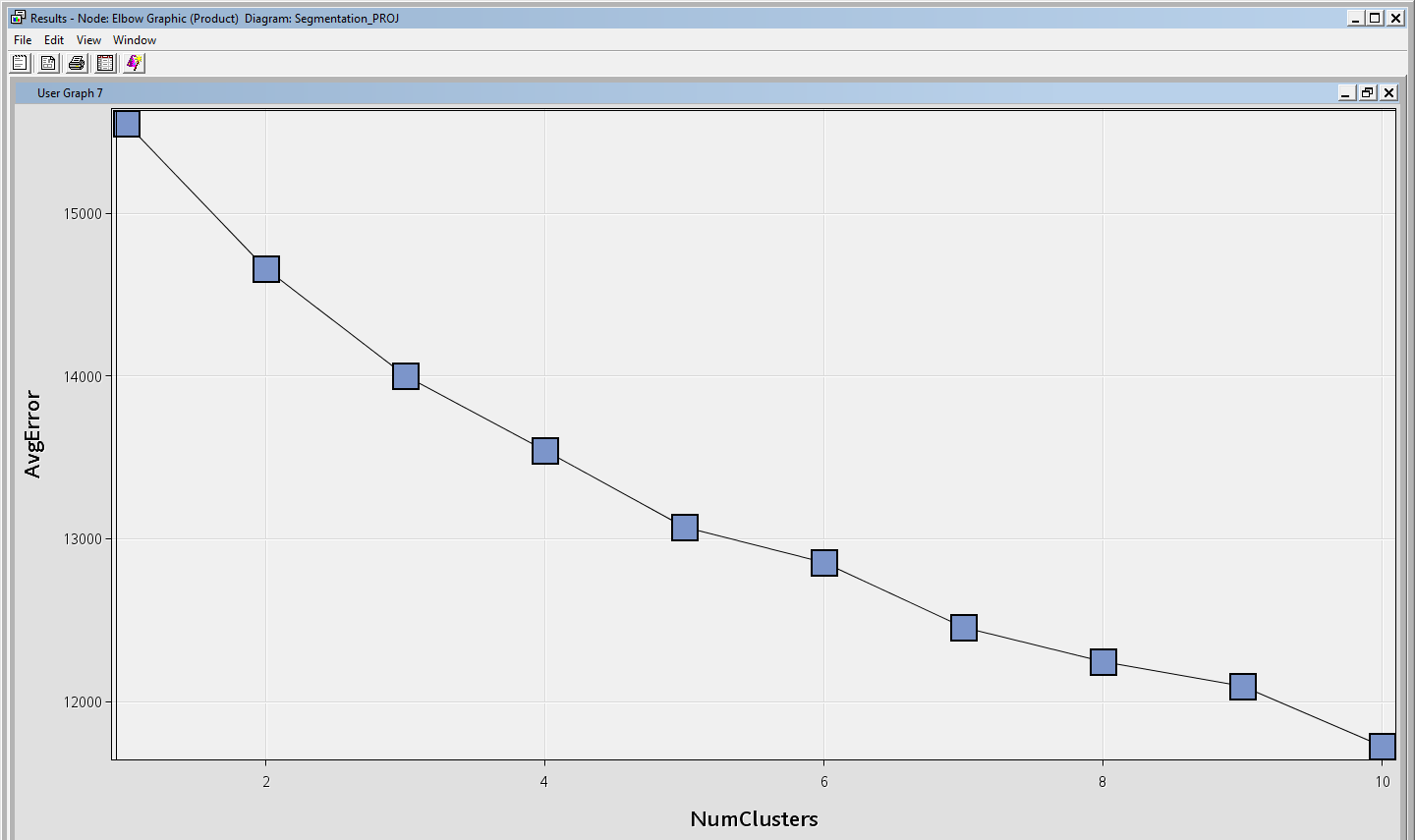


Figure 9 - Elbow graph for product segmentation

Given the output of the *Average Error* representation above, we can say that, finding the "*Elbow*", ideally there can be 5 clusters because the relative decrease from adding one more gets less significant from the 5th cluster. Despite of this interpretation we know that this tool is only to get a general view of how many clusters there could be, that´s why we will use further, for different number of clusters, analysis of *Segment Size Pie Chart, Input Means Plot* and *Mean Statistics*.

#### Additional remarks

* In segmentation we are using one important statistic: *Normalized Mean*. But in this case we have to consider one thing after transforming variables (we transformed from pure variables to percentage proportion between monetary spend on each category divided by total monetary). We have really low means for many of these new variables as you can see in table below, that happens because we have clients who may spend money only few categories, while don't spend anything on another. Naturally, we get low means. While we are doing segmentation, we have to keep it in mind and to take in consideration variables that have relatively higher proportional spending in each category. We cannot forget that when we will be doing the final interpretation of these results.

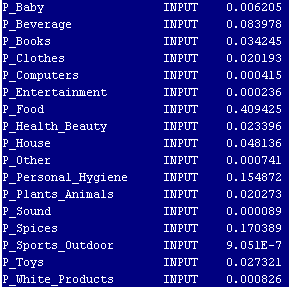


Figure 10 - Normalized mean

* It is important to say that while using *Distance to Nearest Cluster* we are not able to know how is the shape of their distribution, despite of knowing *Maximum Distance from Cluster Seed* (we don’t have idea in which direction is that maximum). The best indicator in this case could be *Minimum Distance Between Two Clusters*, but this indicator does not exists, so that is why we will not pay much attention *to Mean Statistics Table* to infer about dispersion between clusters.

##### 3 Clusters

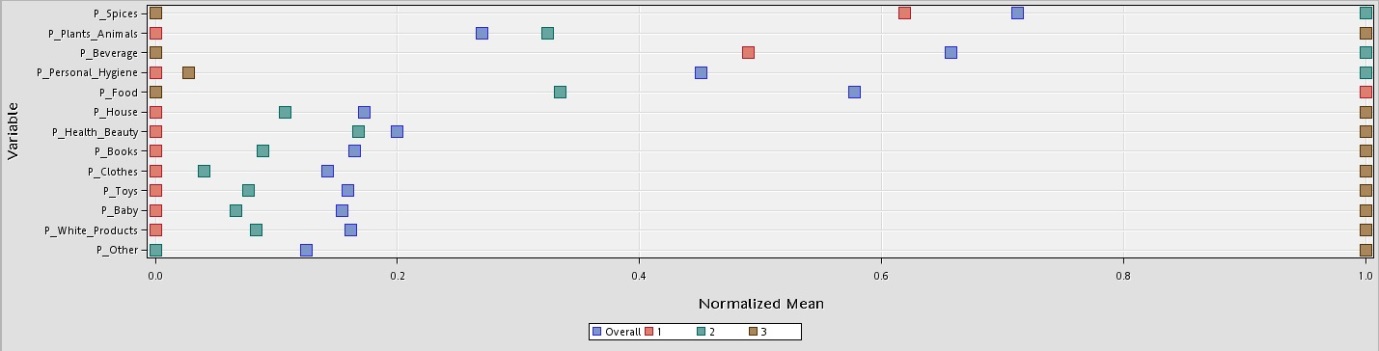


Figure 11 - Input mean plot and segment size of 3 clusters

As you can see in the outputs above, there are 3 possible groups of clients:

1. One big cluster of clients (43%), that almost spend nothing in each product type relatively to overall expenses, except *Spices, Beverages* and *Food*; they spend the biggest part of their money on *Food*, more than the others, while in the other two categories their relative spending is slightly below the average proportion;
2. Another, the biggest cluster of clients (45%), which spend almost the same proportion of their money as the average client in each product type, except *Spices, Beverages* and *Personal Hygiene* where they spend the biggest part of their money, more than every other group; they spend scarcely any money on *Other* products.
3. And the last one, the smallest group (12%), includes all those clients who spend the highest amount of their expenses, relatively to the average costumers and any of the previous groups, almost in each and every type of products, with a denoting exception on *Food, Beverages, Personal Hygiene*, and *Spices*, where they spend significantly lower proportion of their total expenses, almost nothing.

In this clustering option, despite of having relatively heterogeneous distribution of clusters given *Input Means Plot*, we found it poor because it does not bring interesting information about costumers buying behaviours.

##### 4 Clusters

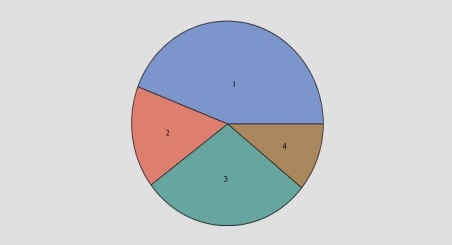


Figure 12 - Input mean plot and segment size of 4 clusters

Comparing with 3 clusters option, this one present more interesting information - we have more distributed segments by product categories:

1. The first cluster of clients is the biggest one, accounting almost a half of our sample - 44%, and includes those clients who proportionally spend nothing or almost nothing in the following types of products: *Personal Hygiene, House, Health and Beauty, Books, White products, Clothes, Toys, Baby* and *Other*; and spend somewhat below the average on *Plants and Animals, Spices* and *Beverages*. Also, there is one category in which they have the highest proportion of spending relatively to average customers and other clusters - Food;
2. Another, relatively smaller (16%), with clients who spend nothing or almost nothing in the following types of products in order of their overall expenses: *Plants and Animals, Personal Hygiene, House, Health and Beauty, Books, Clothes, Toys, Baby* and *Other*; and spend more on *White products* and less on *Food* comparing with average proportion of spending. Moreover, they spend the biggest part of their money on *Spices* and *Beverages*.
3. The next cluster, second in size (29%), has clients whose proportion of expenses is mainly lightly lower than the average client on almost each category of product, except *Health and Beauty* (spend a little bit over average). Also, these clients spend the biggest part of their money on *Plants and Animals* and *Personal Hygiene*.
4. And the last one, the smallest (11%), just like in previous option, includes all those clients who spend the highest part of their expenses in each type of product, much more than the others, with some exceptions. They almost don’t spend money on *Spices, Food, Personal Hygiene* and *Beverage*. Also, these clients spend a little bit lower part of their money comparing with average clients on *Plants and Animals.*

In this clustering option, we can better highlight the "middles", and, despite of having many similar buying behaviours on many product categories in cluster one and cluster two, can be seen a spread out of "lowers" in two clear groups - "Food buyers with some Spices" and "Spices and Drinks buyers". Despite of this improvement, we will be ambitious and try even more clusters!

##### 5 Clusters

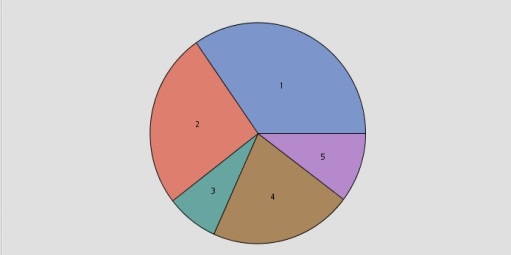


Figure 13 - Input mean plot and segment size of 5 clusters

Comparing with the previous option, having one more cluster, we can see a more confusing and irrelevant differentiation between groups of clients:

1. Just like in previous two options, there remains one big cluster of clients (35%), less than a half, that almost spend nothing in each product type, relatively to their overall expenses except *Food* category;
2. Another, second in size (26%), with clients who spend almost nothing or nothing in all the categories, except *Spices* where they spend the highest part of their expenses; in *Personal Hygiene* and *Food* categories they spend little less than average clients;
3. Another, the smallest (8%), includes clients who spend nothing or almost nothing in all categories, except *Beverages* where they spend the most, comparing to average clients; also, regarding *White Products*, they spend more money, relatively to the average proportion of expenses in this category, more than every other group of clients.
4. This cluster of clients, third in size (21%), represents those who spend mainly as average clients in each product category, except *Personal Hygiene* and *Plants and Animals*, where they spend the highest part of their money.
5. The last, but not the smallest cluster of clients (actually it is almost the smallest – only 10%), includes those who just like in previous options, spend a very big proportion of their expenses in each type of product, more than the others, with an exception on *Beverages, Plants and Animals, Personal Hygiene, Spices* and *Food*, where they spend nothing or almost nothing relatively to other categories.

Given additional group of clients, we can say that this clustering option does not bring much relevant information because it divides existing clusters in even more homogeneous and closer groups. Thus, we difficultly can analyse and infer new and relevant information for us.

##### 6 Clusters

Given additional cluster we mainly understood that this option don’t give us any more relevant information because we saw that there is higher partition of clusters in less heterogeneous between themselves groups.

#### Conclusions on number of Clusters in Product Type

Having a look on each clustering option we mainly understood that *Elbow Graphic* suggestion was not an appropriate option, in our case. We decided to retain 4 clusters because:

* Interpreting each one of cluster options we decided that 4 clusters present more reasonable analysis as they will bring enough relevant information about customer behaviours;
* With this solution we tried to retain not only the relevant information about groups but also kept in mind the relative distance between clusters, that is, we were aware to not divide costumers too much, trying to retain heterogeneous groups between themselves.

In the table below you can observe, for chosen clustering option, main highlights. *Strengths* represent the most important categories for each group, they present maximum *Normalized Mean* in *Input Means Plot* (close to 1); *Weaknesses* the less important (close to zero); *Opportunities* the ones that are above the average proportional spending.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster # | Size | Strengths | Weaknesses | Opportunities |
| 1 | 44% | Food | **Plants and Animals**, **Personal Hygiene**, House, Health and Beauty, Books, White Products, Clothes, Toys, Baby, Others | Spices, **Beverages** |
| 2 | 16% | Spices and Beverages | Plants and Animals, **Personal Hygiene**, House, Health and Beauty, Books, Clothes, Toys, Baby, Others | Food, **White products** |
| 3 | 29% | Plants and Animals, Personal Hygiene | White Products, Clothes, Toys, Baby, Others | Books, Spices, **Food**, Health and Beauty, House, **Beverages,** |
| 4 | 11% | House, Health and Beauty, Books, White Products, Clothes, Toys, Baby, Others | Spices, Food, Personal Hygiene, Beverages | **Plants and Animals** |

Figure 14 - Summary table of product segmentation

It is important to underline that due to the situation described in [Additional Remark](#_Additional_remarks), we gave more relevancy to the first four variables (stated in the [table](#_Additional_remarks)) that present the highest P*roportion Mean*. For example, in the third cluster, *Food* is considered as an *Opportunity* and *Books* as *Weakness*, despite of *Food* being further below the mean than *Books*, because mean of proportion for *Food* is 0,41 and for *Books* is 0,03.

### Value: Value\_Trans+Gender procedure

#### ELBOW GRAPHIC

Given the output of average error representation in Elbow Graphic above, we could interpret, by finding an "elbow", that ideally there **can be 4 clusters because the relative decrease from additional one gets less significant from 4th cluster**. Despite of this interpretation we know that this tool is only to get a general view about how many clusters there could be, that´s why we will use further, for different number of clusters, analysis of Segment Size Pie Chart, Input Means Plot and Mean Statistics.

Figure 15 - Elbow graph of first value segmentation

##### 3 Clusters

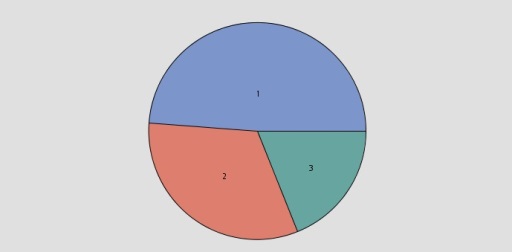


Figure 16 - Input mean plot and segment size of 3 clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Gender | Value | Tot. Articles |
| 1 | 49% | 22 | Mostly Fem (82%) | 0,21 (below av.) | 85 |
| 2 | 32% | 35 | Fem (59%) | 0,41 (above av.) | 345 |
| 3 | 19% | 50 | Mostly Fem (81%) | 0,63 (above av.) | 926 |

Figure 17 - Summary table of value segmentation of 3 clusters

As you can see in the outputs above, there are 3 possible groups of clients:

1. The biggest cluster of clients includes mostly females (82%) of 22 year old. This segment does not bring much value for the enterprise - value indicator is lower than for an average client and it is the lowest between all clusters, as the total number of articles bought.
2. The next cluster is smaller comparing with previous one and is largely composed by women - 59%. This group includes middle aged people who bring a notably higher value, comparing to previous cluster - it is slightly above the average customer value, as total number of articles bought.
3. The last cluster, smallest in size, includes mostly aged females who represents the highest value for company comparing with others clusters and the average client. The same happens with total articles; this group bought the highest number of articles in period of study.

##### 4 Clusters

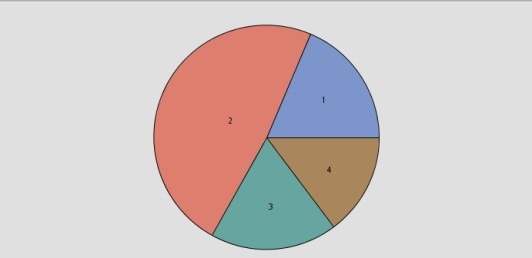


Figure 18 - Input mean plot and segment size of 4 clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Gender | Value | Tot. Articles |
| 1 | 19% | 24 | Man | 0,28 (below av.) | 174 |
| 2 | 48% | 23 | Females | 0,25 (below av.) | 134 |
| 3 | 18% | 47 | Mostly Fem (77%) | 0,46 (above av.) | 426 |
| 4 | 15% | 47 | Mostly Fem (82%) | 0,66 (above av.) | 1036 |

Figure 19 - Summary table of value segmentation of 4 clusters

Comparing with 3 clusters option, this one present more interesting information - we have more clear segments:

1. The first cluster of clients includes only young males, representing roughly 19% of filtered sample, who do not bring much value for the enterprise - value indicator is significantly below average, as the total number of articles bought.
2. Another, the biggest group of clients, retaining 48% of filtered sample, represents clients who bring almost no value for the enterprise. This cluster is composed only by young women.
3. The next cluster, brings much more value than the previous two, seeing that the value indicator is notably higher than sample average and the total number of articles is slightly above the average. The interesting fact is that these people are predominantly aged woman.
4. And the last one, the smallest, just like in previous cluster, includes predominantly aged woman (they even have the same age), however the value they bring to the company is much higher than any other cluster. Also, they buy more products than all the other ones together.

In this clustering option, we achieved more heterogeneous distribution between clusters (given *Input Means Plot*), where we can highlight the "high-middle" clients from "golden" ones, and spread out the "lowers" in two clear groups - only *man* and only *women*. Despite of this improvement, we will be ambitious and try even more clusters!

##### 5 Clusters

Figure 20 - Input mean plot and segment size of 5 clusters



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Gender | Value | Tot. Articles |
| 1 | 35% | 23 | Female | 0,20 (below av.) | 69 |
| 2 | 20% | 26 | Male | 0,29 (below av.) | 191 |
| 3 | 19% | 27 | Female | 0,43 (above av.) | 389 |
| 4 | 14% | 54 | Mostly Fem (85%) | 0,47 (above av.) | 443 |
| 5 | 12% | 48 | Mostly Fem (73%) | 0,68 (above av.) | 1105 |

Figure 21 - Summary table of value segmentation of 5 clusters

Comparing with 4 clusters option, this one presents more interesting information - we have more clear segments:

1. The first cluster, the biggest group of clients, includes only females representing roughly 35% of filtered sample and has very low value for the enterprise, the value indicator is significantly below average as number of total articles – it is very low.
2. The next one is an average sized cluster, composed by young men. The value that they bring to the enterprise is slightly below average as is the total number of articles bought.
3. The following cluster, despite of being almost of the same size as the last one, brings a notable increase in value to the enterprise, its value is slightly above average as well as the number of total articles. It is a cluster composed by young women, however these ones are four years older than the ones described in the first cluster, in average.
4. Fourth cluster includes clients who bring plentiful value for the enterprise and they buy more than the average clients. The people grouped in this cluster are mainly aged women.
5. The last and the smallest one, includes the most valuable clients and those who buy bigger number of products than each one of previous groups. They are not young people, predominantly women (73%).

##### Conclusions on number of Clusters in Value using Value\_Trans+Gender procedure

Having a look on each clustering option we mainly understood that Elbow Graphic suggestion is quite appropriate, in our case. But essentially we decided to retain 4 clusters because:

* Interpreting each one of cluster options we decided that 4 clusters will present more reasonable analysis as they will bring enough relevant information about customers profile and value they bring to the enterprise;
* With this solution we tried to retain not only the relevant information about groups of clients but also kept in mind the relative distance between clusters, that is, we were aware to not divide costumers too much, trying to retain heterogeneous groups between themselves due to profile characteristics.

#### Join Clusters

In the following table you can see the result of joining two different clustering methods such as we can analyse 16 different clients groups in perspective of their value for the enterprise and their buying behaviours.

|  |  |  |
| --- | --- | --- |
| Value | Product | Nº of customers in cluster |
| Bronze\_Y\_F | Food | 1121 |
| Bronze\_Y\_F | House\_Child\_Beauty\_Others | 429 |
| Bronze\_Y\_F | Plants\_Animals\_Hygiene | 754 |
| Bronze\_Y\_F | Spices\_Beverages | 491 |
| Bronze\_Y\_M | Food | 464 |
| Bronze\_Y\_M | House\_Child\_Beauty\_Others | 128 |
| Bronze\_Y\_M | Plants\_Animals\_Hygiene | 313 |
| Bronze\_Y\_M | Spices\_Beverages | 184 |
| Golden\_Ladies | Food | 486 |
| Golden\_Ladies | House\_Child\_Beauty\_Others | 1 |
| Golden\_Ladies | Plants\_Animals\_Hygiene | 257 |
| Golden\_Ladies | Spices\_Beverages | 111 |
| Silver\_A\_F | Food | 503 |
| Silver\_A\_F | House\_Child\_Beauty\_Others | 65 |
| Silver\_A\_F | Plants\_Animals\_Hygiene | 359 |
| Silver\_A\_F | Spices\_Beverages | 145 |

Figure 22 - Table of joining clusters

As we can see, there are clusters that individually are very small, appealing that they can be joined to other groups of clients that present similar characteristics. Even though, as an example, you can find below the interpretation of each cluster in perspective of mentioned characteristics.

* Cluster "*Bronze\_Y\_M & Food*" represents 8% of the sample and is composed by young male clients that mainly consume *Food* products, some *Spices* and some *Beverages*. Generally they do not consume other categories of products.   
   Despite of being less valuable than an average client, clients of this cluster can be viewed as potential ones.
* Cluster "*Bronze\_Y\_M &House\_Child\_Beaut\_Oth*" represents 2% of the sample and is composed by young males spending proportionally more money on *House* products, *Health and Beauty*, *Books*, *White* *Products*, *Clothes*, *Toys*, *Baby* and *Other* types of products than other young males. Also they spend some proportion of their expenses on *Plants and Animals* and *Spend* almost nothing on *Spices*, *Food* products, *Personal Hygiene* and *Beverages*.   
   What is regarding to their value to the enterprise, it is expressed in the same way as the previous cluster: they can be viewed as potential clients, despite of being less valuable than an average client.

We will not continue the characterization and grouping clusters because, after analysing proportions of individuals in each one of four possible buying behaviour categories belonging to "*Bronze\_Y\_M*" or "*Bronze\_Y\_F*" value categories, we concluded that there is no relative difference in gender. In the table below you can see that, proportionally, there is no significant difference about what young men and young women buy (both presenting a low value for the enterprise).

Figure 23 - Summary table of 2 similar value segments

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| "Bronze" Young Males | nº of customers in segment | % of customers in segment | "Bronze"\_Young\_Female | nº of customers in segment | % of customers in segment |  |
| Food | 464 | 43% | Food | 1121 | 40% |  |
| House\_Child\_Beaut\_Oth | 128 | 12% | House\_Child\_Beaut\_Oth | 429 | 15% |  |
| Plant\_Anim\_Hyg | 313 | 29% | Plant\_Anim\_Hyg | 754 | 27% |  |
| Spice\_Bev | 184 | 17% | Spice\_Bev | 491 | 18% |  |
| Total nº of costumers | 1089 |  |  | 2795 |  | 3884 |

Also, as we remember from value segmentation (see table below), we found no significant difference what is regarding to age (1 year), value indicator (0,03 points) and total articles (40 units) between individuals belonging to cluster 1 and 2. That is, clusters seem to be very similar, the only thing that differentiates them is gender.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Gender | Value | Tot. Articles |
| Bronze\_Y\_M | 19% | 24 | Man | 0,28 (below av.) | 174 |
| Bronze\_Y\_F | 48% | 23 | Females | 0,25 (below av.) | 134 |

Figure 24 - Summary table of 2 value segments

If there would be an evident proof that young men by, for instance, more Spices and Beverages than young women (in relative terms), we could say that this clustering procedure, including gender, is worthy (as we saw, it is not). Also there is one important thing, despite of usefulness of value indicator created in resuming perception of value of each client to Sell Everything, we understood that in this study this is not a point and we can lose some relevant information what is regarding to clients segmentation.   
 Having these conclusions in mind we decided to redo Value segmentation with original variables, excluding gender – Age, Frequency, Monetary, Recency and Number of Total Articles bought.

### Value: Value\_Pure

#### ELBOW GRAPHIC

Figure 25 - Elbow graph of second value segmentation

Given the output of *Average Error* representation in graphic above, we could interpret, by finding an "*Elbow*", that ideally there can be 4 clusters because the relative decrease from additional one gets less significant from 4th cluster. Despite of this interpretation we know that this tool is only to get a general view about how many clusters there could be, that´s why we will use further, for different number of clusters, analysis of *Segment Size Pie Chart, Input Means Plot* and *Mean Statistics*.

##### 3, 4 e 5 Clusters

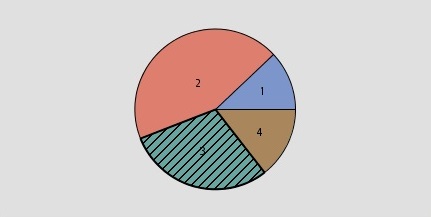
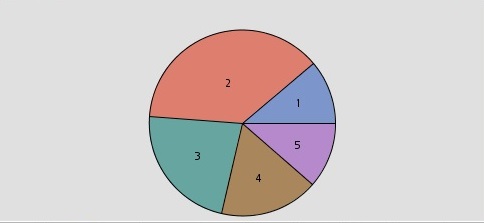
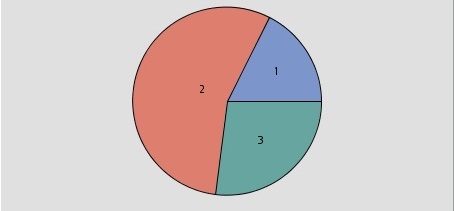


Figure 26 - Comparison of 3, 4, 5 segments size

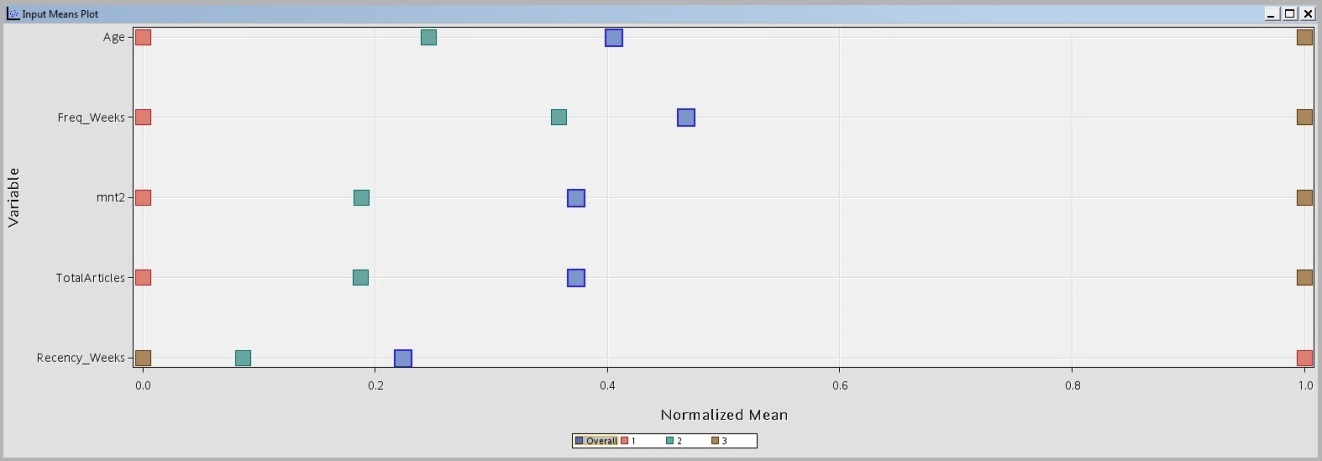


Figure 27 - Input mean plot of 3 clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Recency | Frequency | Monetary | Tot. Articles |
| 1 | 18% | 21 | 28 | 3 | 132 | 42 |
| 2 | 56% | 27 | 3 | 16 | 538 | 186 |
| 3 | 27% | 47 | 1 | 38 | 2288 | 810 |

Figure 28 - Summary table of value segment of 3 clusters

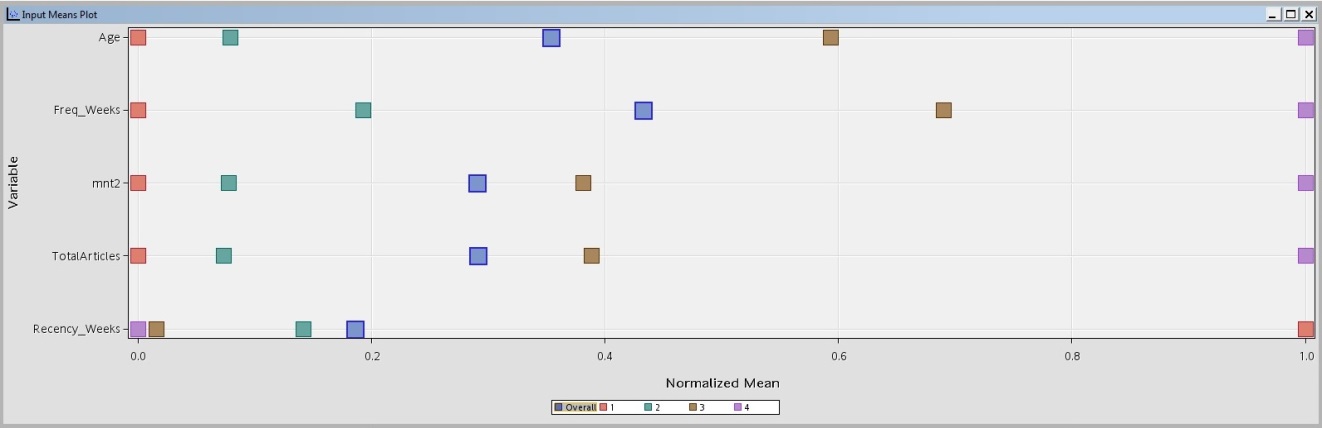


Figure 29 - Input mean plot of 4 clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Recency | Frequency | Monetary | Tot. Articles |
| 1 | 12% | 21 | 35 | 3 | 116 | 38 |
| 2 | 44% | 24 | 5 | 11 | 335 | 112 |
| 3 | 29% | 38 | 1 | 30 | 1192 | 425 |
| 4 | 15% | 50 | <1 | 42 | 2936 | 1033 |

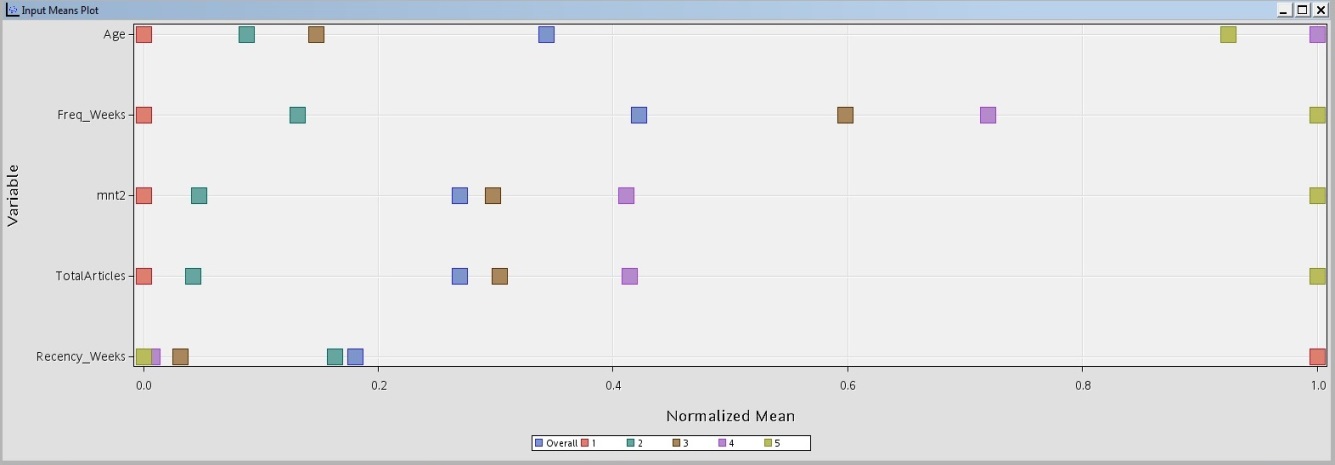
Figure 30 - Summary table of value segment of 4 clusters

Figure 31 - Input mean plot of 5 clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size | Average age | Recency | Frequency | Monetary | Tot. Articles |
| 1 | 11% | 21 | 36 | 3 | 113 | 38 |
| 2 | 38% | 24 | 6 | 8 | 227 | 83 |
| 3 | 23% | 26 | 2 | 27 | 1021 | 364 |
| 4 | 17% | 50 | 1 | 32 | 1368 | 484 |
| 5 | 11% | 48 | 1 | 43 | 3166 | 1114 |

Figure 32 - Summary table of value segment of 5 clusters

##### Conclusions on number of Clusters in Value using Value\_Pure procedure

Having a look on each clustering option we mainly understood that Elbow Graphic suggestion is quite appropriate, in our case, just like it was in previous procedure. But essentially we decided to retain 4 clusters because:

* Interpreting each one of cluster options we decided that 4 clusters will present more reasonable analysis as they will bring enough relevant information about customers profile and value they bring to the enterprise. ;
* With this solution we tried to retain not only the relevant information about groups of clients but also kept in mind the relative distance between clusters, that is, we were aware to not divide costumers too much, trying to retain heterogeneous groups between themselves due to profile characteristics.

###### Characterization of clusters

1. First cluster, the smallest (12%) of clients includes mostly young people. It does not bring much value for the enterprise as these clients spend less money than average clients and every other cluster, are less frequent and last time visited Sell Everything, in average, 35 weeks ago, more than an average client (regarding to date of data collection). Also they by the lowest number of articles.
2. The next cluster is the biggest (44%) and is largely composed by young people who bring more value for Sell Everything, comparing to previous cluster, but still – they are less frequent, spend less money and by less articles in comparison to average clients. Also these clients last time visited Sell Everything, in average, 5 weeks ago – more than an average client .
3. The next cluster (29%), includes mostly middle-aged clients who can be described as “high-middle” because they are slightly more frequent, spend more money and by more articles in comparison to average clients. The same happens with total articles; this group bought the highest number of articles. Also these clients last time visited Sell Everything, in average, 1 week ago – less than an average client.
4. And the last one, the smallest (15%), includes predominantly aged clients, however the value they bring to the company is much higher than any other cluster as they are significantly more frequent, spend much more money and by much more articles in comparison to average clients. Also, they went to Sell Everything less than one week ago – significantly more recent than an average client (regarding to date of data collection).

##### Joining Clusters

In the following table you can see the result of joining two different clustering methods such as we can analyse 16 different clients groups in perspective of their value for the enterprise and their buying behaviours.  
 In comparison to other clustering procedure, we can say that despite of not having an enormous difference in results, this option is more accurate as we are using variables that are worthy to this context.

|  |  |  |
| --- | --- | --- |
| Value | Product | Nº of customers in cluster |
| Big\_Bronze | Food | 989 |
| Big\_Bronze | House\_Child\_Beauty\_Others | 417 |
| Big\_Bronze | Plants\_Animals\_Hygiene | 725 |
| Big\_Bronze | Spices\_Beverages | 446 |
| Big\_Silver | Food | 844 |
| Big\_Silver | House\_Child\_Beauty\_Others | 91 |
| Big\_Silver | Plants\_Animals\_Hygiene | 551 |
| Big\_Silver | Spices\_Beverages | 221 |
| Golden | Food | 475 |
| Golden | House\_Child\_Beauty\_Others | 2 |
| Golden | Plants\_Animals\_Hygiene | 255 |
| Golden | Spices\_Beverages | 106 |
| Low | Food | 266 |
| Low | House\_Child\_Beauty\_Others | 113 |
| Low | Plants\_Animals\_Hygiene | 152 |
| Low | Spices\_Beverages | 158 |

Figure 33 - Table of joining clusters

###### Cluster re-joining

As we can see, there are some segments of clients that are very small and individually do not bring enough value in order to compensate the effort to develop different marketing campaigns and design value propositions of the loyalty programs. Also, as we know from practice, the most important is the “quality” of clients and not their “quantity”, that is, mainly businesses like *Sell Everything* lives from least number of clients that brings the most profit. So that we should take more care of these clients and not forget about the others that can become clients of *Sell Everything* one day.

We decided to re-join the least valuable segments (*Low* and *Big-Bronze)* because they do not present substantial differences from value point of view and called it Big\_Bronze. Also, we included *Golden* and *House\_Child\_Beauty\_Others* segment into Big\_*Silver* and *House\_Child\_Beauty\_Others* because the first segment presented only 2 clients, what is non-reasonable and non-compensating from business point of view and these are the most similar categories from value perspective. Thus, the result is presented in the following table:

|  |  |  |
| --- | --- | --- |
| Value | Product | Nº of customers in cluster |
| Big\_Bronze | Food | 1255 |
| Big\_Bronze | House\_Child\_Beauty\_Others | 530 |
| Big\_Bronze | Plants\_Animals\_Hygiene | 877 |
| Big\_Bronze | Spices\_Beverages | 604 |
| Big\_Silver | Food | 844 |
| Big\_Silver | House\_Child\_Beauty\_Others | 93 |
| Big\_Silver | Plants\_Animals\_Hygiene | 551 |
| Big\_Silver | Spices\_Beverages | 221 |
| Golden | Food | 475 |
| Golden | Plants\_Animals\_Hygiene | 255 |
| Golden | Spices\_Beverages | 106 |

Figure 34 - Table of joining clusters

## Value proposition for different segments

Finally, after identification of segments, we can suggest different value propositions:

### Big\_Bronze value segment, one generalized step.

But if Sell Everything really wants to explore the potential of *Big\_Bronze* segment independently on product type segment, there is one effective policy to be applied: if some client belonging to this value segment will spend more than 30 $ in a purchase, for example, he/she will receive a discount of 5$. Also this discount can be *restrictive,* that is, can be spent only on certain category of products, like the ones that were identified as opportunities or as weaknesses to make clients “decentralize” their buying behaviours.  
 Such policies can contribute to easily involve people in Sell Everything by creating a perception of direct benefit from shopping there.

### Particular cases

* For every *Big\_Silver* and *Golden* clients buying more *Food* products Sell Everything can develop marketing campaigns creating percentage discounts on *Food* products, also, applying cross-selling techniques by combining this discounts with *Spices* and *Beverages* products when bought jointly (as we saw in [product](#_Conclusions_on_number), these can be seen as opportunities). It’s important to say that *Golden* clients should have a little bit “closer” approximation in order to underline their exclusivity as valuable clients, to do that up-selling techniques can be applied when buying more expensive products (ex. “10% discount when buying caviar”).
* A similar approach can be taken with *Big\_Bronze* clients buying more *Food* products, but instead of making a percentage discount it could be made in absolute terms like “if you buy 3 units, you will get one more freely”.
* For every *Big\_Silver* and *Golden* clients buying more *Spices* and *Beverages* we can also develop campaigns making percentage discounts on *Spices* and *Beverages* and cross-selling them with *Food* and *White* *Products* when bought jointly (as we saw in [product](#_Conclusions_on_number), these can be seen as opportunities).
  + At the same time, as these clients are more concerned on some specific sort of products, it could be useful to make them buy other types of products they do not use to, promoting *motivational discounts* on *Plants and Animals, Personal Hygiene, House, Health and Beauty, Books, Clothes* and *Other* types of products like “if buying some *Health and Beauty* product, you will get a 10% discount”. The same approachment can be taken with previous two segments, as they don’t use to by almost the same products as this one, with exception on *White Products.*
* The same can happen with *Big\_Bronze* clients buying more *Spices* and *Beverages* products, but instead of making percentage discount, this could be made in absolute terms. Also as we saw before, can be applied a generalized approach to every Big\_Bronze segment independently on product category
* For every *Big\_Silver* and *Golden* clients buying more *Plants* and *Animals and Personal Hygiene* products *Sell Everything* can also develop campaigns making percentage discounts on those products, also cross-selling this discounts with *Books, Spices, Food, Health and Beauty, House, Beverages*, when bought jointly (as we saw these can be seen as opportunities). It’s important to say that *Golden* clients should have a little bit “closer” approximation in order to underline their exclusivity as valuable clients, to do that up-selling techniques can be applied like having exclusive offers in some exclusive type of products regarding to Personal Hygiene, for example.
* The same can happen with *Big\_Bronze* clients buying more Spices and Beverages products but instead of making percentage discount, this could be made in absolute terms to make them involved in Sell Everything. Off course, these clients will not be able to receive discounts on more exclusive sort of products.
* For every *Big\_Silver* clients buying more *House, Health and Beauty, Books, White Products, Clothes, Toys, Baby*, and *Other* products, *Sell Everything* can develop campaigns making percentage discounts on those products, also combining these discounts with *Plants and Animals* when bought jointly (as we saw this can be seen as opportunity). These clients by many different types of products, thus company can develop marketing campaigns making them by different combinations of products – even those that they don’t use to, for example “if buying *Health and Beauty* product, client will receive a discount in *Personal Hygiene* product”.
* Beyond generalized approach described in the beginning, the same can be done with *Big\_Bronze* clients buying more same products. But instead of making percentage discount it could be made in absolute terms as this will make them more involved in *Sell Everything*.

## Conclusion

Segmentation of clients is about understanding customer needs through analysis of their buying behaviours and socio-demographic characteristics; form groups of customers that share similar criteria and determine how to suit them in a better way – that is one of the main goals of an enterprise.   
 Using SAS Enterprise Miner and knowledge acquired from Data Mining course, we performed this task and hope that knowledge acquired during this workful and careful process to help *Sell Everything* to win market share by satisfying better customer needs in a more efficient way and to show an individualized oncoming for everyone.

# Annexes

## Exploration of Data Quality Issues

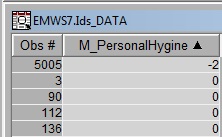
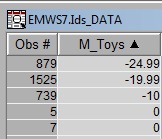
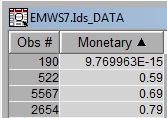
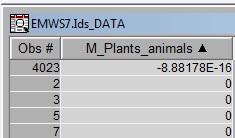
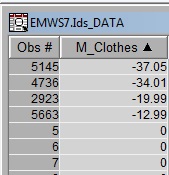


Figure 35 – Exploration of Data Quality Issues

## Filtering nodes

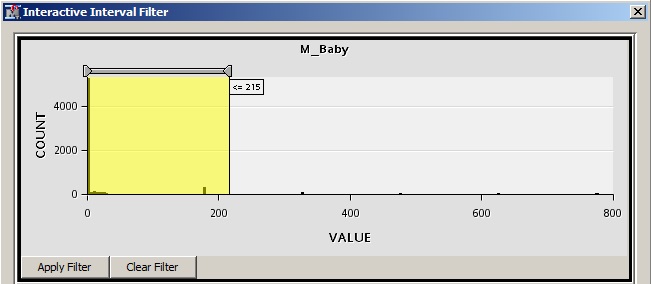


Figure 36 – Filtering M\_Baby

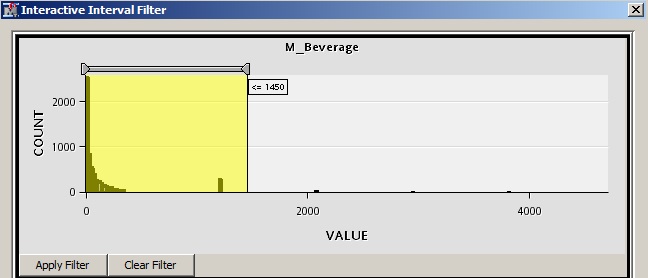


Figure 37 - Filtering M\_Beverge

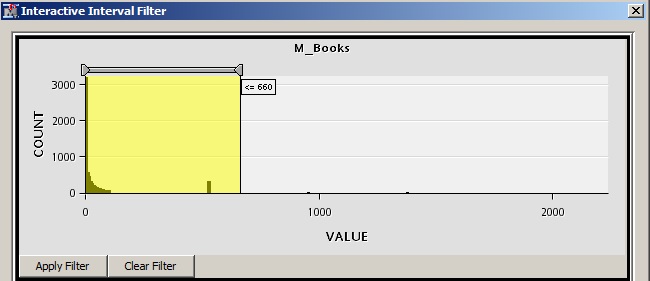


Figure 38 - Filtering M\_Books

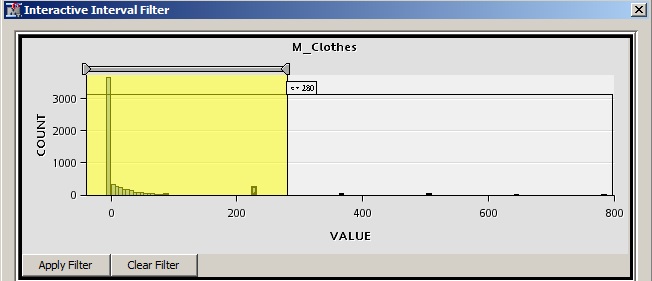


Figure 39 - Filtering M\_Clothes

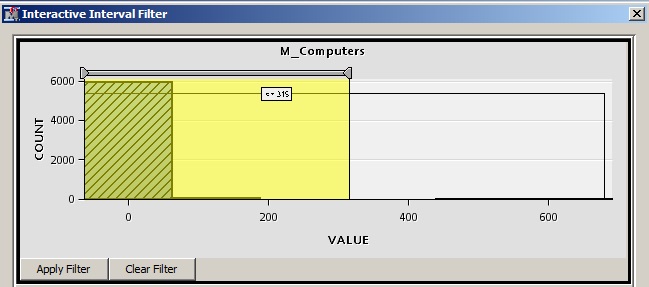


Figure 40 - Filtering M\_Computers

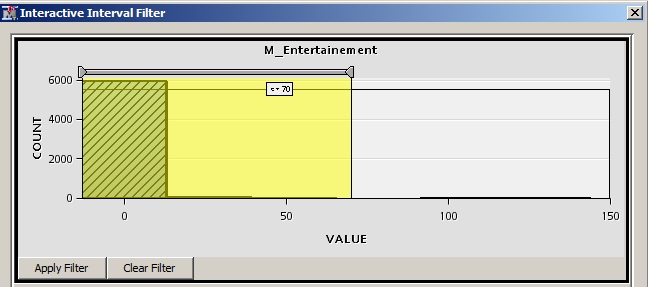


Figure 41 - Filtering M\_Entertainment

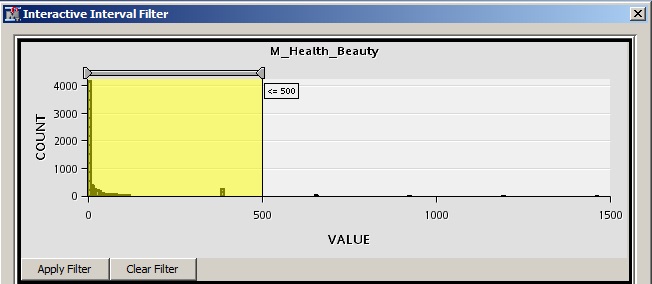


Figure 42 - Filtering M\_Health\_Beauty

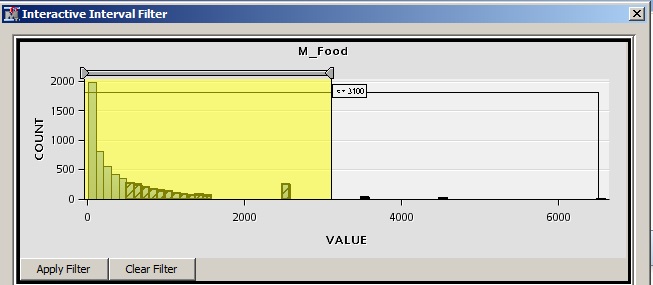


Figure 43 - Filtering M\_Food

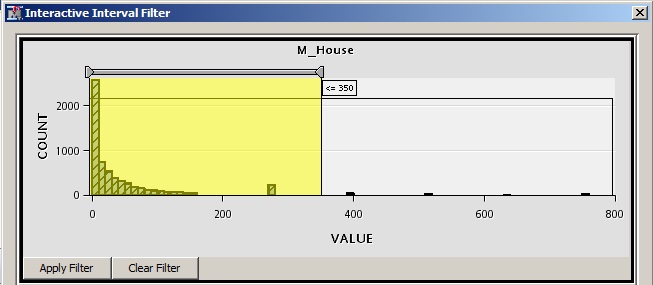


Figure 44 - Filtering M\_House

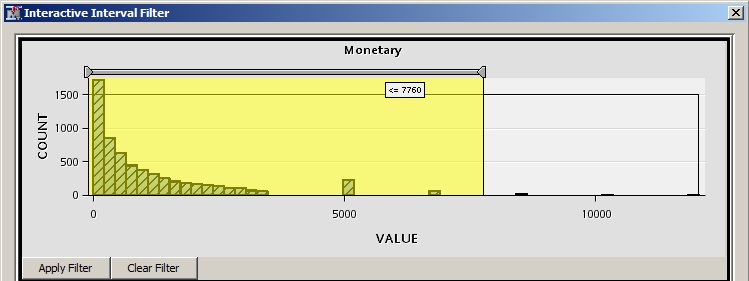


Figure 45 - Filtering M\_Monetary

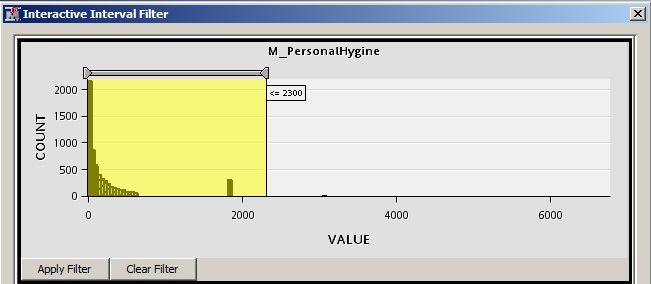


Figure 46 – Filtering M\_Personal\_Hygiene

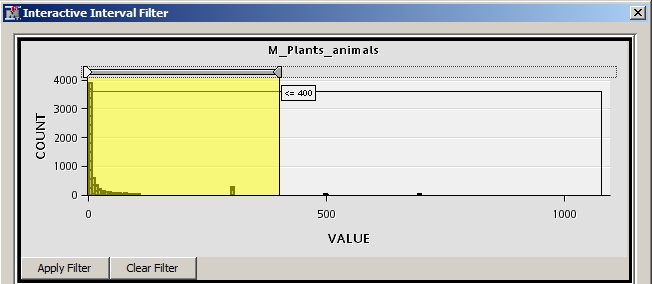


Figure 47 - Filtering M\_Plants\_Animals

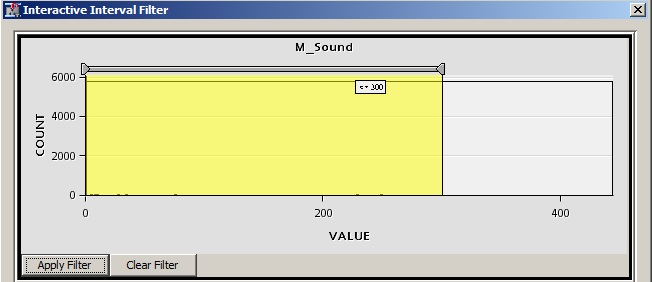


Figure 48 - Filtering M\_Sound

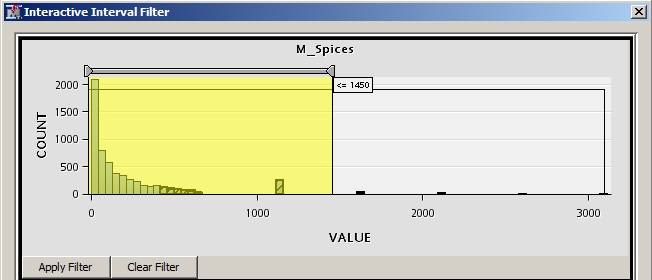


Figure 49 - Filtering M\_Spices

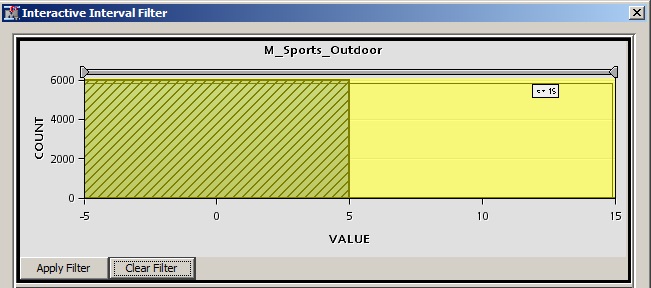


Figure 50 - Filtering M\_Sports\_Outdoor

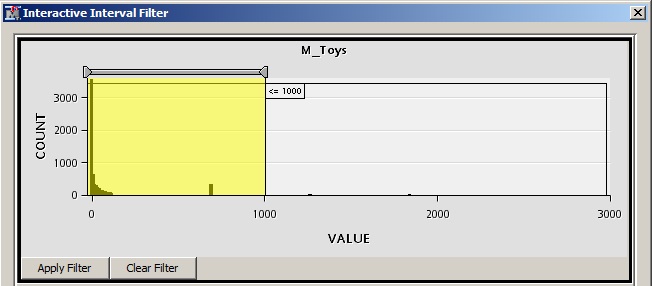


Figure 51 - Filtering M\_Toys

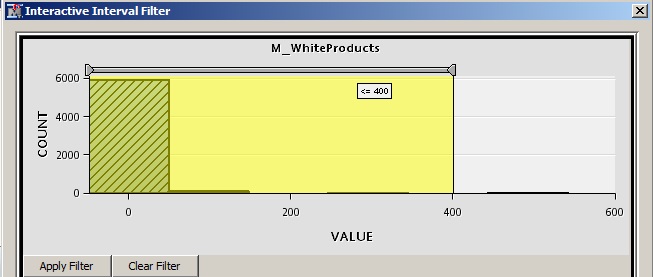


Figure 52 - Filtering M\_WhiteProducts

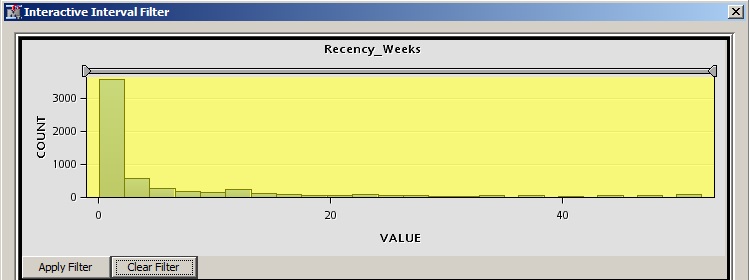


Figure 53 - Filtering Recency\_Weeks

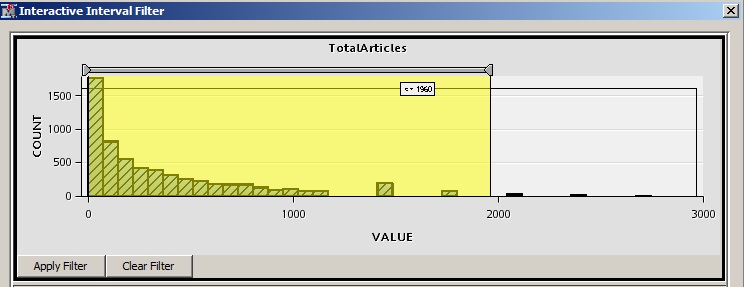


Figure 54 - Filtering Total\_Articles

# Bibliographic References

1. Article "Determination of customer value measurement model

RFM index weights" by Liu Wei-Jiang, Duan Shu-Yong, Wang Xiao-Feng.

Business School, Jilin University, China

Population Resource and Environment Research Institute, Jilin University, China

[https://opus.lib.uts.edu.au/research/bitstream/handle/10453/18891/2011000466OK.pdf?sequence=1](https://opus.lib.uts.edu.au/research/bitstream/handle/10453/18891/2011000466OK.pdf?sequence=1%20)

* Consulted at 18:06 of of 23rd December 2014