## Rym BARKAOUI

## Simo SEFFAR

## Manon FOULQUIER

# PROJECT DATA MINING

# Data Engineering Project



103bis Champs-Elysées,

France, Paris, 75008

26/12/2012

Summary

[Introduction 3](#_Toc341866144)

[I. Goals 3](#_Toc341866145)

[II. Identification of input needed and Data Understanding 4](#_Toc341866146)

[1. Data that we have 4](#_Toc341866147)

[2. Datas that we have to add, or create with other datas 9](#_Toc341866148)

[3. Chosen data 10](#_Toc341866149)

[III. Preprocessing 11](#_Toc341866150)

[1. Cleaning 11](#_Toc341866151)

[a. Removed attributes 11](#_Toc341866152)

[b. Removed entries 12](#_Toc341866153)

[2. Data Viewing 14](#_Toc341866154)

[a. Color manager 14](#_Toc341866155)

[b. Parallel coordinate 15](#_Toc341866156)

[c. Scatter plot 19](#_Toc341866157)

[d. Cross tabulation 23](#_Toc341866158)

[e. Statistics 24](#_Toc341866159)

[f. Conclusion data viewing  24](#_Toc341866160)

# Introduction

This project is a data mining project on a computer shop. We have 2000 rows of data: the 1000 first data are buyers from the shop, the following 1000 are about the website buyers. The data contain demographical data as country age… and questions about the buying process (Shop for the first 1000 rows/Website for the others), starting from these data we have established a series of data mining goals. For each goal we select only the relevant attributes to fulfill the data mining goal. There are attributes that we did not use as for example name as they can’t have any impact on any of our goals. Also we will have to work on our data in order to fit the data mining algorithm for example to calculate the age from the birth date and to organize the age in ranges. A goal can be for example the profile of major buyers in the website or in the shop, to reach this goal we would have to use to demographical data, the budget and the frequency using the 1000 first rows we will have the profile of website major buyers. That’s the way we worked for every goal we had. First we are going to present our goals, then present every attribute and explain it, and then show the attributes we had to add and the attributes we have treated in order to fit, then the data we chose to reach each goal and finally the KNIME representation of data and a decision tree.

# Goals

1. Profiles of major buyers

1. website
2. shop

2. Behavior of shoppers

1. general
2. website
3. shop
4. country depending
5. website
6. shop

3. Identification of important parameters

1. website
2. shop

# Identification of input needed and Data Understanding

Data File Table Description:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Number attributes | Number Entries | Date of the record |
| AllDatasWeb.csv |  | 1000 | 15/11/12 |
| AllDatasShop.csv |  | 1000 | 15/11/12 |

## Data that we have

|  |  |  |
| --- | --- | --- |
| Name | Values | Explanation |
| Both shop and website | | |
| 1-Age | ranges of age :  0 [0- 18]  1 [18- 25]  2 [25- 35]  3 [35- 45]  4 [45- 55]  5 [55- 65]  6 +65 | Indicates what kind of people is buying the product (teenagers, adults, third age)  It will help us having an idea about the range of age of buyers.  We will use this data for our first goal which is defining the profile of major shoppers. We will use it also for the third goal which identify some important parameters. |
| 2-Gender | 0 for male, 1 for female | Indicates the gender of the person.  It will help us knowing if women are more likely to buy a product or the contrary.  We will use this data for our first goal which is defining the profile of major shoppers. We will use it also for the third goal which identifies some important parameters. |
| 3-Religion | {Hinduism/Buddhism/atheism/Judaism/Islam} | Indicates one’s religion  This data doesn’t help us in anything with our goals |
| 4-Status | {Single/Married/Divorced/Widowed/civil union/cohabiting} | Indicates one’s status  We will use this data for our first goal which is defining the profile of major shoppers. We will use it also for the third goal which identifies some important parameters. |
| 5-Country | {France/Germany/Italy/Poland/England/Russia/Turkey/Spain} | Indicates one’s country.  It helps us knowing the repartition of all the shoppers  We will use this data for our first goal, which is defining the profile of major shoppers and the second one for the behavior of the shopper. We will use it also for the third goal which identifies some important parameters. |
| 6-When you decide to buy something how much do you spend? | Ranges of price:  [0-300]  [300-600]  [600-900]  [900-1200]  [1200-1500]  [1500-1800]  [1800-2100]  [>2100] | Indicates how much a person spends when going to the website. It will help us knowing the average budget of customers.  We will use this data for our second goal for the behavior of the shopper. We will use it also for the third goal which identifies some important parameters. |
| 7-How frequently do you buy our products? | (less than once a year)  (Once/twice a year)  (Every month)  (Every week) | Indicates how frequently a person purchases on the internet. It will help us knowing the frequency of shopper’s purchases.  We will use this data for our second goal for the behavior of the shopper. We will use it also for the third goal which identifies some important parameters. |
| 11-How easy you change your mind because of attractive offers of other competing brands? | Never  Almost Never  Frequently  Always | Information about the shopper behavior.  It will help us knowing how a customer can easily change his mind.  We will use this data for our second goal for the behavior of the shopper but also the third goal. |
| 19- IDclient | Numerical value | Identification in a unique manner the shopper.  Not relevant in our study case |
| 20- e-mail | String | E-mail of the shopper  Not relevant in our study case |
| 21- IDproduct | String | Identification in a unique manner the product  Not relevant in our study case |
| 22- Name | String | Name of the shopper  Not relevant in our study case |
| 23- Surname | String | Surname of the shopper  Not relevant in our study case |
| 31- Work | 1. Farmer  2. Artisan, shopper, CEO  3. white collar and grey matter  4. Intermediary professions  5. Employees  6. Workers  7. Students | Indicates the shopper category job. It will help us knowing the shopper profiledepending on one’s job.  We will use this data for our first goal, which is defining the profile of major shoppers. We will use it also for the third goal which identifies some important parameters. |
| Only for shop | | |
| 9- What is the main reason for you present product purchase? | Update previous product,  Previous is broken, new requirement | Information about the shopper behavior. It will help us understand what are the main reasons for a shopper for buying a product.  We will use this data for our second goal for the behavior of the shopper but only for shops. We will use it also for the third goal which identifies some important parameters. |
| 10- Do you buy a product only because you think it is a special offer even it is not what originally you were looking for? | Yes/No | Information about the shopper behavior. It will help us knowing if special offers are a main reason of purchases  We will use this data for our second goal for the behavior of the shopper but only for shops. |
| 11-How much time you usually spend in the buying process of a product? | Since conception until you order the product:  1 day,  7 days,  30 days,  180 days. | Information about the shopper behavior.  It will help us knowing after how much time we can consider that a customer has changed his mind.  We will use this data for our second goal for the behavior of the shopper but only for shops. |
| 12- When buying a product, what is the most important factor of the product you look for? | {Cost/Value/Service} | Information about the shopper behavior.  It will help us knowing what is important for a customer  We will use this data for our second goal for the behavior of the shopper but only for shops. We will use it also for the third goal which identifies some important parameters. |
| 27- Do you usually buy products of the same brand as your previous product? | yes  no | Information about the shopper behavior  Not relevant in our study case |
| 28- Is this product brand the one you initially decided to buy? | yes  no | information about the shopper behavior  Not relevant in our study case |
| 29- How much time do you usually spend on the buying process of a product? | 1 day  7 days  30 days  180 days | information about the shopper behavior  Not relevant in our study case |
| 30- What is your most common way of buying products? | online stores  local stores  specialized brand stores  big stores | information about the shopper behavior  Not relevant in our study case |
| Only for website | | |
| 13- Do you go to the website looking for information about the hardware? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website?  It will help us knowing the exact reason why a customer visits the website.  We will use this data for our second goal for the behavior of the shopper but only for websites. We will use it also for the third goal which identifies some important parameters. |
| 14- Do you go to the website looking for information about the software? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website.  It will help us knowing the exact reason why a customer visits the website.  We will use this data for our second goal for the behavior of the shopper but only for websites. |
| 15- Do you go to the website looking for prices? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website.  It will help us knowing the exact reason why a customer visits the website.  We will use this data for our second goal for the behavior of the shopper but only for websites. |
| 16- Do you go to the website looking for special offers? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website.  It will help us knowing the exact reason why a customer visits the website.  We will use this data for our second goal for the behavior of the shopper but only for websites. |
| 17- Do you go to the website looking for shop addresses? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website.  It will help us knowing the exact reason why a customer visits the website.  We will use this data for our second goal for the behavior of the shopper but only for websites. |
| 18- Do you go to the website looking for information about the warranty? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website.  It will help us knowing the exact reason why a customer visits the website.  We will use it for the third goal which identifies some important parameters. |
| 24- Do you go to the website looking for information about accessories? | Binary value:  0 for false  1 for true | What’s the client purpose when going to the website  Not relevant in our study case |
| 25- How many time a month do you search on website? | Numerical value | Frequency of the shopper’s visits  Not relevant in our study case |
| 26- How frequent do you buy this product? | 1 month  6 months  12 months  60 months | Frequency of the shopper’s purchase  Not relevant in our study case |

## Data that we have to add, or create with other data

(All computed with excel)

|  |  |  |
| --- | --- | --- |
| Name | Values | Obtain with |
| 40-Age Range | ranges of age :  0 [0- 18]  1 [18- 25]  2 [25- 35]  3 [35- 45]  4 [45- 55]  5 [55- 65]  6 +65 | birth date  It will help us knowing exactly in what category we can range customers.  We will use this data for our first goal, which is defining the profile of major shoppers. We will use it also for the third goal which identifies some important parameters. |
| 41-Budget | ranges of price :  0-[0-300]  1-[300-600]  2-[600-900]  3-[900-1200]  4-[1200-1500]  5-[1500-1800]  6-[1800-2100]  7-[>2100] | As 6  It will help us knowing exactly the budget of customers.  We will use this data for our second goal for the behavior of the shopper. |
| 42-Buying frequency | 0-(less than once a year)  1-(Once/twice a year)  2-(Every month)  3-(Every week) | As 7  It will help us knowing the frequency of purchases for customers.  We will use this data for our second goal for the behavior of the shopper. |
| 43- Changing mind frequency because of attractive offers of other competing brands? | 0-Never  1-Almost Never  2-Frequently  3-Always | As 11  It will help us knowing how often a customer can change his mind because of special offers or other brands advertisements.  We will use this data for our second goal for the behavior of the shopper. |

## Chosen data

|  |  |
| --- | --- |
| Goal | Attributes Used to achieve this goal |
| 1a | 1/2/4/5/6/7/8/31/40 |
| 1b | 1/2/4/5/6/7/8/31/40 |
| 2a.i | 6/7/13/14/15/16/17/41/42/43 |
| 2a.ii | 6/7/9/10/11/12/41/42/43 |
| 2b.i | 5/6/7/13/14/15/16/17/5/41/42/43 |
| 2b.ii | 5/6/7/9/10/11/12/5/41/42/43 |
| 3 a | Profile and 11/13-18 |
| 3 b | Profile and 11/ 9/12 |

# Preprocessing

After data understanding, data are inputted in Knime with a CSV Reader.

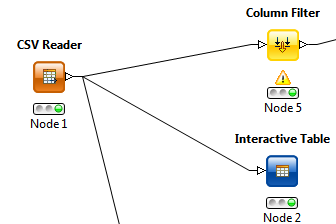
In this part, data are going to be treated, and using different viewing tools, we could start to have an idea of different relations.

## Cleaning

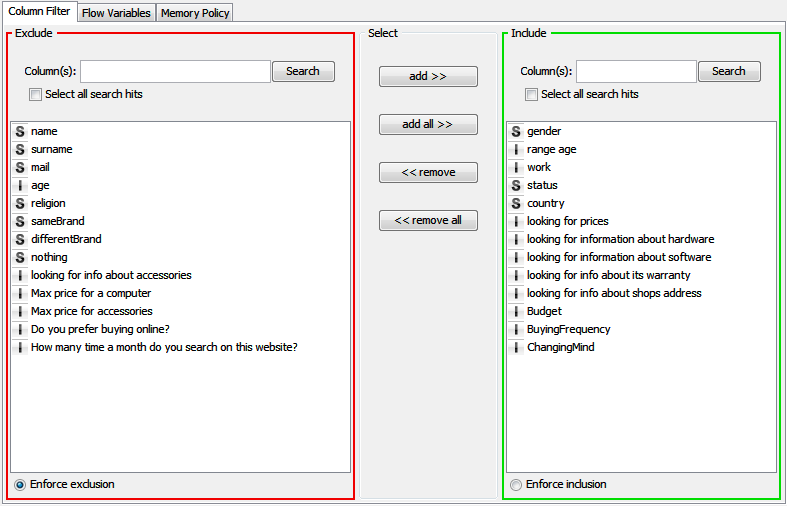
First of all, we need to remove all the data that we don’t need, using the chosen data table.

### Removed attributes

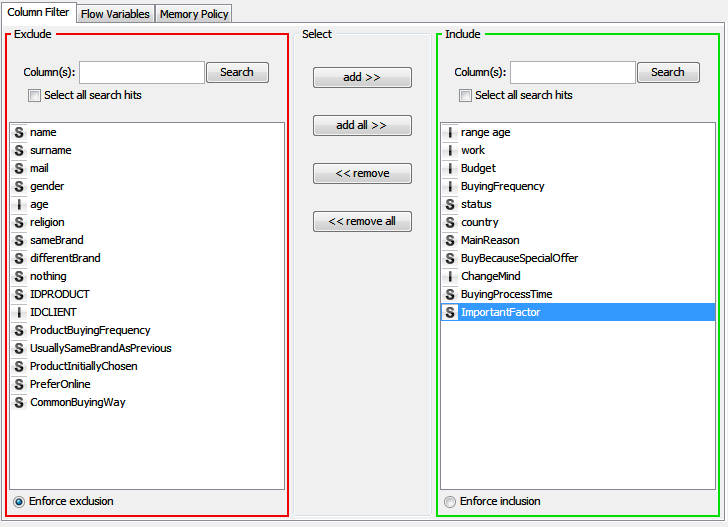
(Interactive table & column filter)



* Website



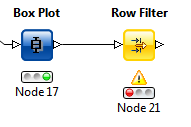
* Shop



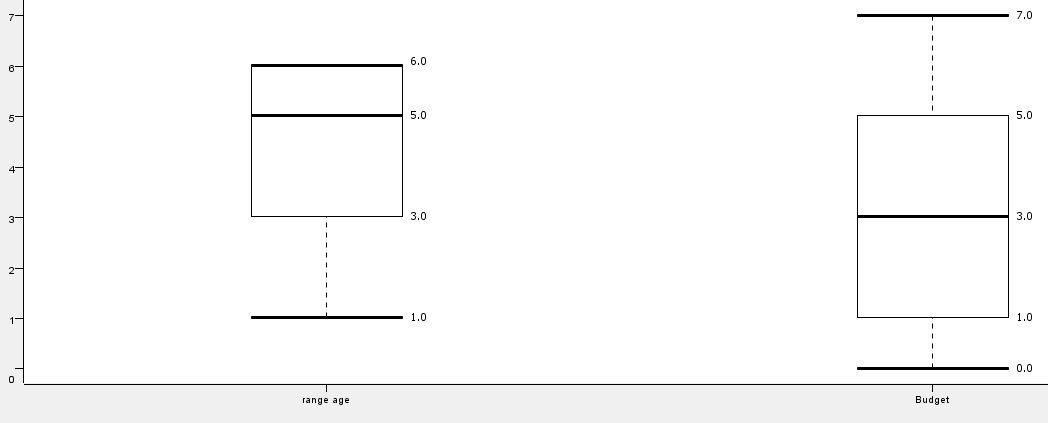
### Removed entries

We can also delete specific entries in the data table, using the row filter. For example, we could see with a box plot that for some data, certain attributes are far beyond the distribution of data: “outbonders”.

(Box plot & row filter)



As we have ranges well represented, we decide not to delete anything. For example we could have deleted some rows with and age>ageMaximum, but we decide to keep them in the last range: +65.



Example of box plot, for range age and budget.

## Data Viewing

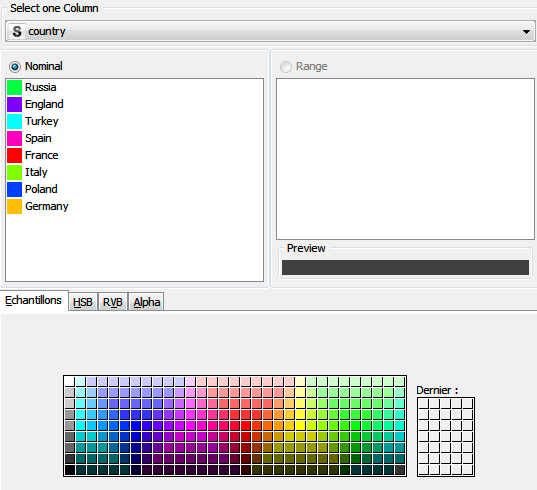
We have different tools useful to have a first idea of our data(more explicit than a table).

NB: for each example, in parenthesis will be given what goal can be helped with this example.

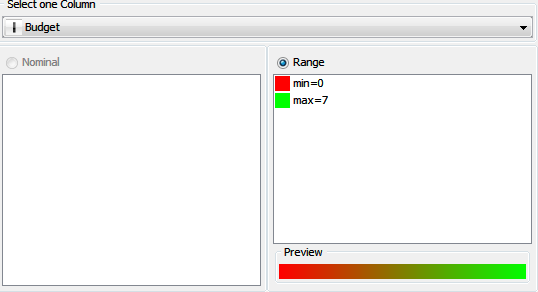
### Color manager

This tool colors data in function of one attribute.

* Example with countries (for goal 2b for example): String values. Each entry will be colored in function of its attribute country.



* Example for budget (goals 1, 2): integer values: each entry will be colored depending of its budget attribute.



### Parallel coordinate

This tool shows data we choose, making a link between them.

### Scatter plot

The scatter plot shows a graphic given two attributes.

### Cross tabulation

The cross tabulation tool shows a matrix using different attributes.

### Statistics

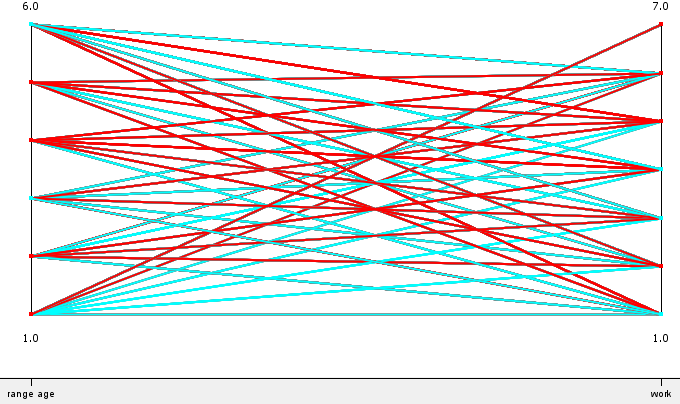
With this tool we can have an idea of the statistics about attributes: min-max, mean, variance …

### Conclusion data viewing

All theese tools give us an idea of how the datas are distributed, linked, maybe we can also have first idea of different groups. Anyway, they are helpful before and after data processing, to show in an explicit way datas, clusters….

### Data Viewing / Goal

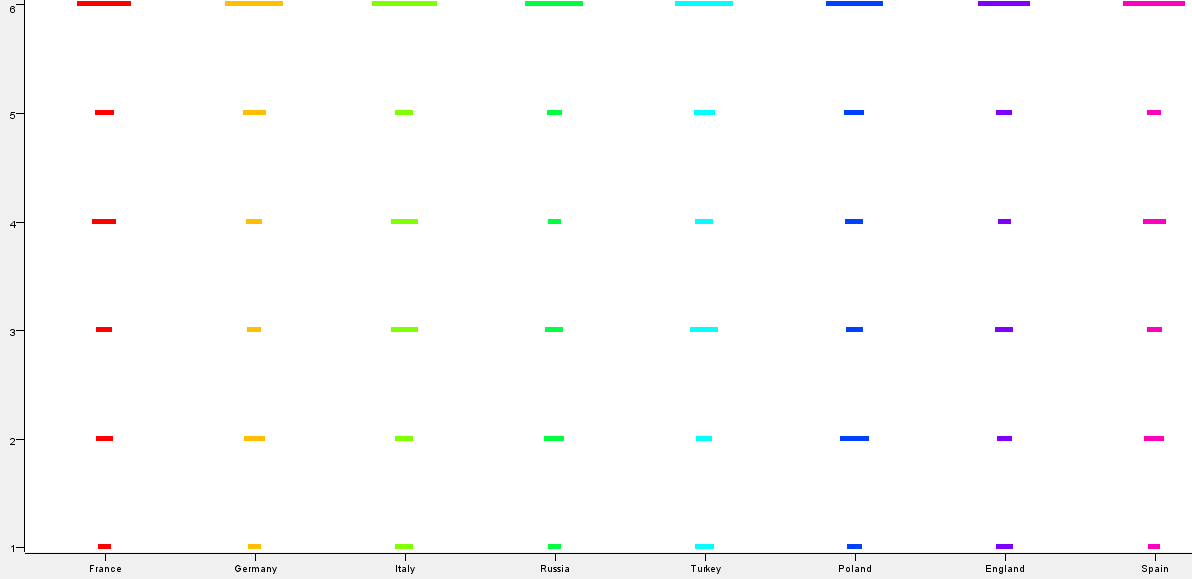
* + Example Profiles(goal 1a) : range age / work (gender colored : red for female, blue for male)



Most of the female in range age 1 (18-25) are in work categories 7(students) and 6(workers).

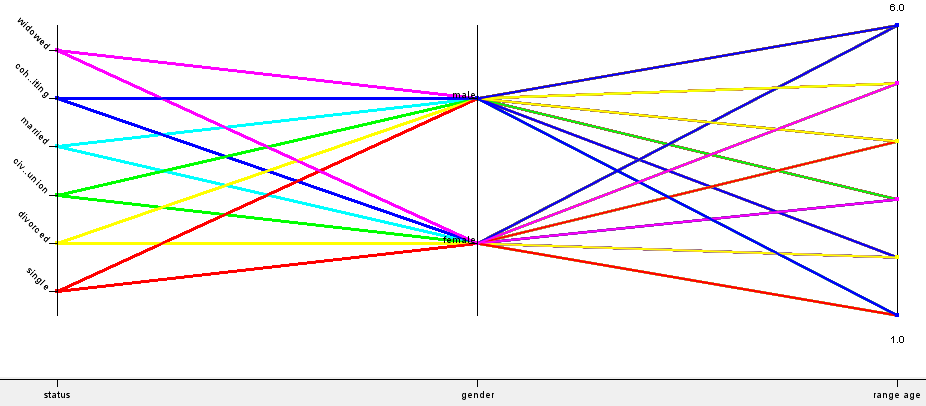
Most of the male in range age 4 (45-55) are in work category 1 (farmers).

* + Example Profiles (goal 1a): range age = f(country) (country colored)



In general we can say that shoppers are mostly seniors (up to 65 years old). Still, in each country, represented age range differ a little, with for example in Turkey, more buyers in age range 3.

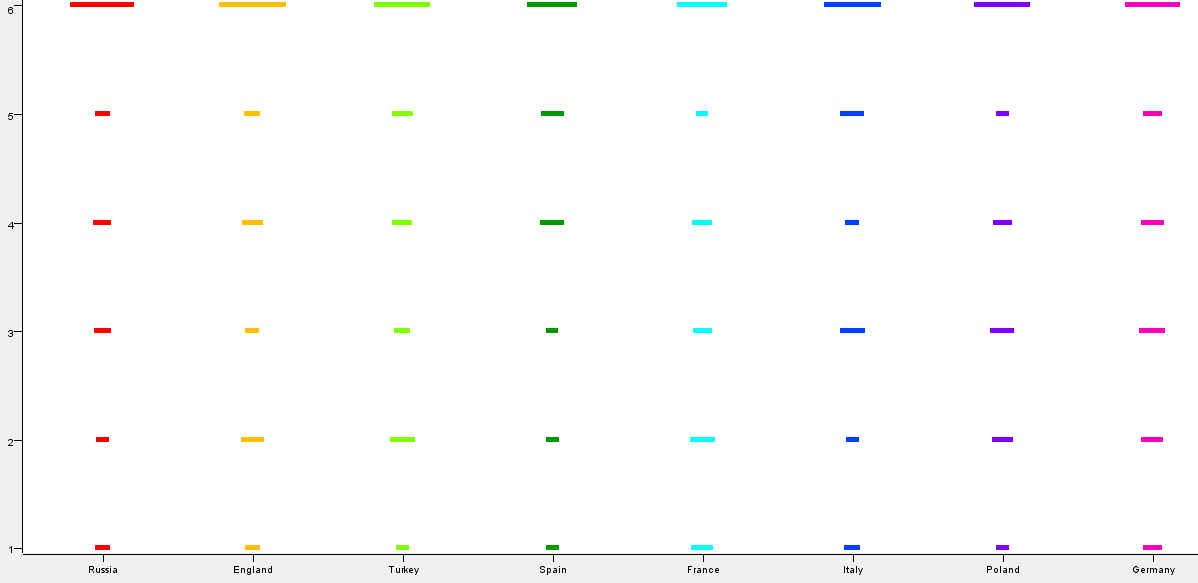
1. Shop
   * Example Profiles (goal 1b) : range age / gender ( status colored)



This diagram gives us many information, we can see for example that most of the buyers in age range 6 (+65) are cohabiting male and cohabiting female. We can also see that most of the buyers in age range 4 are single female and divorced male.

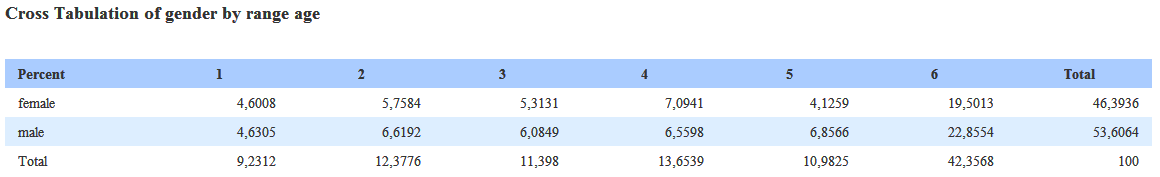
Most of the buyers in age range 1 (18-25) are single female and cohabiting male (which usually correspond to students)

* + Example Profiles (goal 1b): range age = f(country) (country colored)



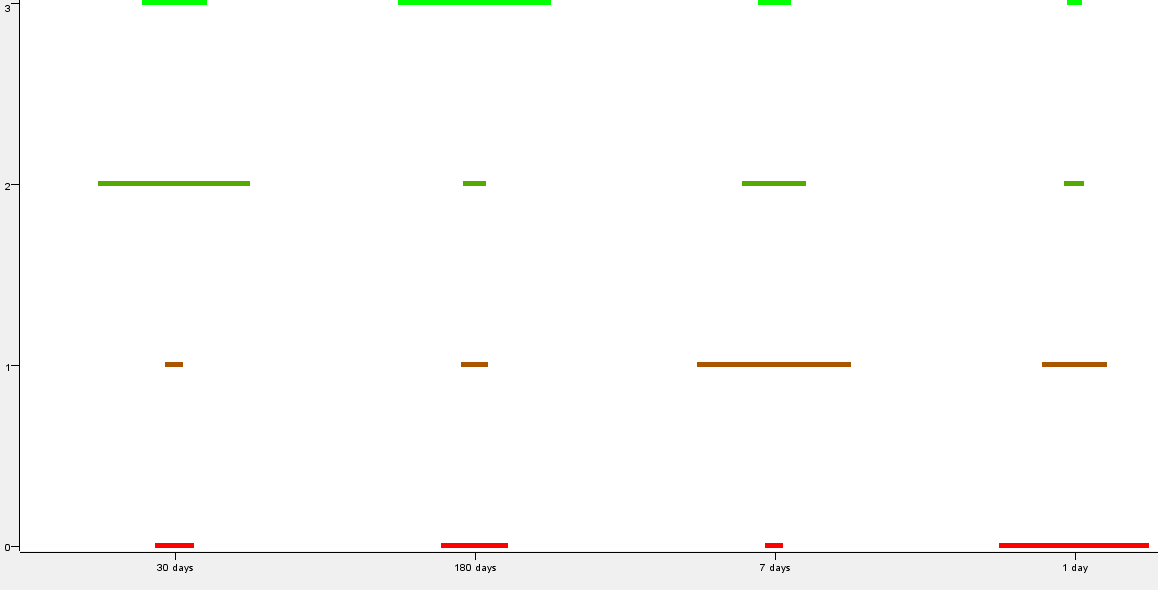
For all the countries, age range 6 is predominant. Moreover, we can see that in Spain for example, there are more people from age ranges 4, 5 and 6 in shops than the others.

For Turkey, age range 2 is more represented than in Russia for example.



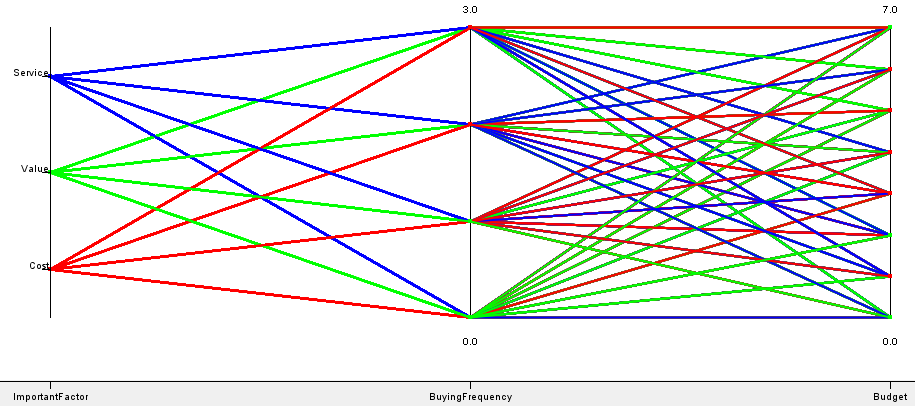
We can see that most of the shoppers are more than 65 years old and in most cases, men are more represented than women

1. Behavior of shoppers
2. general
3. website
   * Example Behavior (goal 1a.i): ChangeMind= f(BuyingProcessTime) (changeMind colored)



After 30 days, people change their mind frequently whereas after 1 day they never change their mind before buying the product. After 180 days they always change their mind, we can consider that they have found another product.

1. shop
   * Example Behavior (goal 2a.ii): buyingFrequency / Budget (important factor colored)

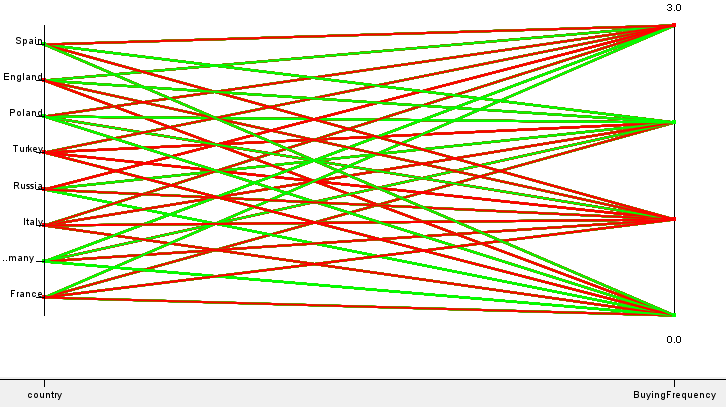


It can be seen that buyers who buy every weeks (buyingfrequency=3) with a high budget (7) think that the cost is the most important factor. If they have medium budget (4), the service is important.

For those who buy less than once a week (buyingfrequency=0) with a small budget (0), the most important factor is the service.

For medium budget(4) who buy every month, the value is important.

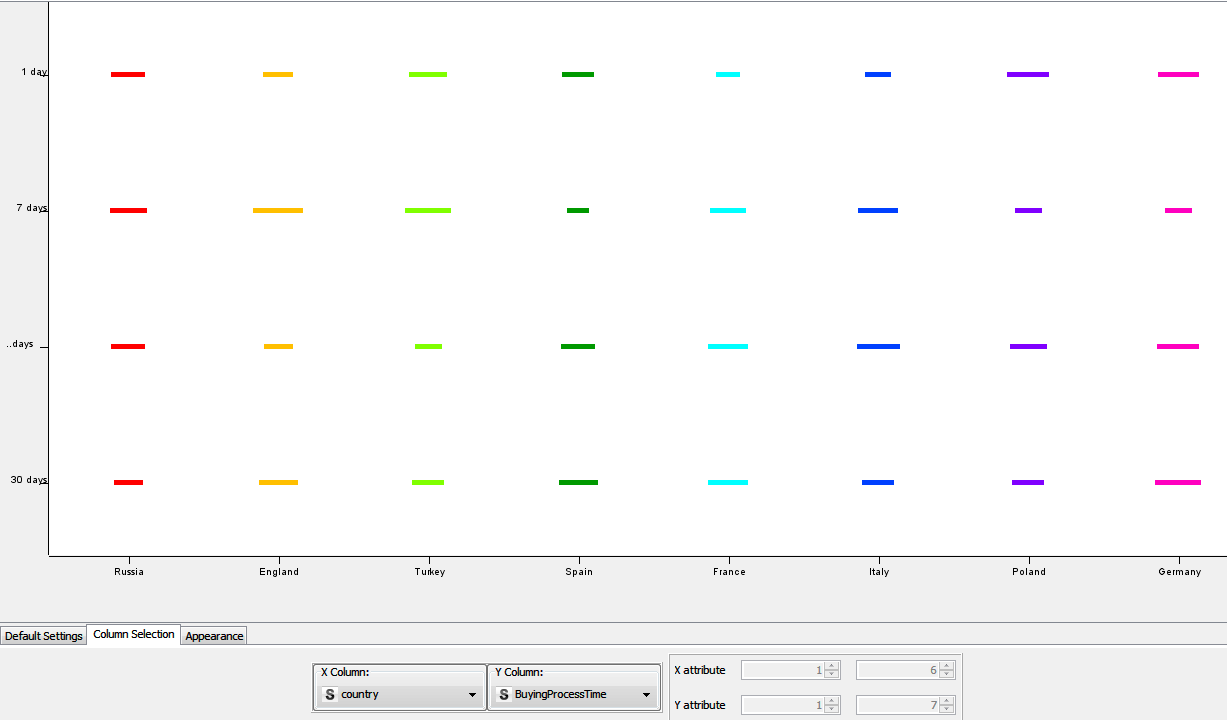
1. country depending
2. website
   * Example Behavior (goal 2b.i): Country / BuyingFrequency (looking for prices colored : Green = looking for prices, red = non looking for prices)



Most of the buyers with a high buyingFrequency in Spain aren’t looking for prices on the website.

Unlike the buyers from Germany with a high budget.

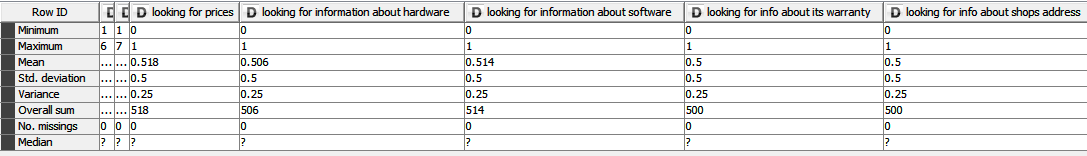
1. Shop
   * Examples Behavior (goal 2b.ii): buyingProcessTime = f (country) (country colored)



We can see that in Russia, buying process times are well represented, whereas in Turkey, people are less likely to buy a product for a buying process time of 14 days.

In Spain, there are fewer buyers who have a buying process time of 7 days.

1. Identification of important parameters
2. Website
   * Example for website (goal 3a):

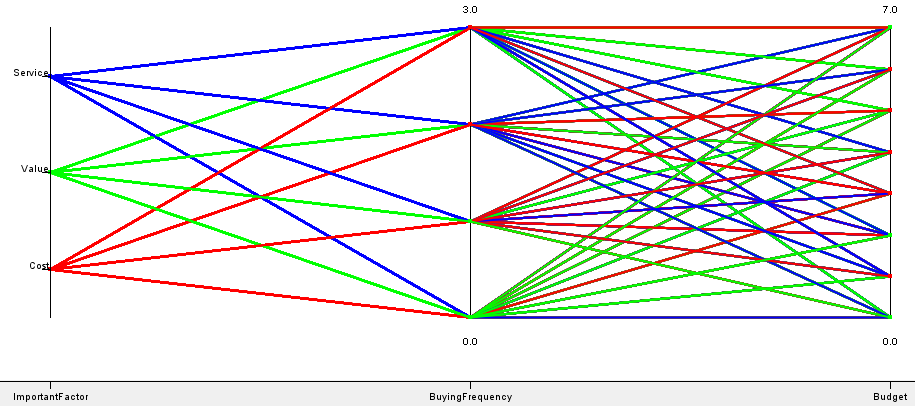


We can see that all the looking for attributes have a mean at 0.5 with a deviation of 0.5 so it’s not relevant, we will have to find a way to identify different groups with more representative values.

1. shop

We can reuse this graphics

* + Example Behavior : buyingFrequency / Budget (important factor colored)



It can be seen that buyers who buy every weeks (buyingfrequency=3) with a high budget (7) think that the cost is the most important **factor**. If they have medium budget (4), the **service** is important.

For those who buy less than once a week (buyingfrequency=0) with a small budget (0), the most important factor is the **service**.

For medium budget(4) who buy every month, the **value** is important.

# Conclusion

With these prepared data sets, and a first view of data, we can now start the modeling steps.