Methodological Report

Guidance for Consumer Testing Data Analysis & Reporting - Supplement to Step 4

Understanding the Drivers of Trait Preferences and the Development of Multi-user RTB Product Profiles, WP1, Step 4

Montpellier, France, October 2021

Geneviève FLIEDEL, Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpellier, France

Isabelle MARAVAL, CIRAD, Montpellier, France

Aurélie BECHOFF, Natural Resources Institute, University of Greenwich, Chatham, UK

Eglantine FAUVELLE, CIRAD, Montpellier, France (Validator)

This report has been written in the framework of RTBfoods project.

To be cited as:

Geneviève FLIEDEL, Isabelle MARAVAL, Aurélie BECHOFF, Eglantine FAUVELLE, (2022). *Guidance for Consumer Testing Data Analysis & Reporting - Supplement to Step 4. Understanding the Drivers of Trait Preferences and the Development of Multi-user RTB Product Profiles, WP1, Step 4.* Montpellier, France: RTBfoods Methodological Report, 50 p. https://doi.org/10.18167/agritrop/00660

Ethics: The activities, which led to the production of this manual, were assessed and approved by the CIRAD Ethics Committee (H2020 ethics self-assessment procedure). When relevant, samples were prepared according to good hygiene and manufacturing practices. When external participants were involved in an activity, they were priorly informed about the objective of the activity and explained that their participation was entirely voluntary, that they could stop the interview at any point and that their responses would be anonymous and securely stored by the research team for research purposes. Written consent (signature) was systematically sought from sensory panellists and from consumers participating in activities.

Acknowledgments: This work was supported by the RTBfoods project https://rtbfoods.cirad.fr, through a grant OPP1178942: Breeding RTB products for end user preferences (RTBfoods), to the French Agricultural Research Centre for International Development (CIRAD), Montpellier, France, by the Bill & Melinda Gates Foundation (BMGF).

Image cover page © LAJOUS P. for RTBfoods.

This document has been reviewed by:	
Aurélie BECHOFF	09/2020 and 12/01/2022
Final validation by:	
Eglantine FAUVELLE	12/01/2022

CONTENTS

Table of Contents

List of Tables

List of Figures

ABSTRACT

This is a guidance document that gives a step-by-step description of the different analyses that should be included for consumer testing data analysis and reporting, referred to as WP1 Step 4 within RTBfoods project. The objectives of this document are to provide country partners with a guide and a template to follow when reporting on the activity. The main document is divided into Study context; Methodology that includes sampling, consumer testing, data analysis; Results including overall liking (9-point scale hedonic testing), segmentation (Ward cluster analysis), Just-About Right (JAR), Check-all-that-applies (CATA), Sensory mapping (Principal Component Analysis (PCA)); and Conclusions. In addition, there are appendices that give detailed guidance on the use of the XLStat software to conduct the various statistical tests. This appendix is downloadable at https://mel.cgiar.org/reporting/download/report_file_id/25496

Key Words: template, statistical analysis, methodology, consumer testing, roots, tubers and bananas

1 STUDY CONTEXT AND GENERAL OBJECTIVES

The main aim of this Step 4 “Consumer testing” is to understand the consumers’ demand for the quality characteristics of Root, Tuber and Banana products.

Another aim is to provide WP2 with a clear and visual mapping of the most liked products associated with high quality characteristics and high Overall liking scores, and of the least liked products associated with low quality characteristics and low Overall liking scores.

The activity consists in inviting a large number of consumers to test the 4-5 products made in the previous processing step from varieties with very different quality characteristics.

2 METHODOLOGY

2.1 Sampling

The *4-5 products* made by the processors from varieties with very different quality characteristics during the Step 3 “Processing diagnosis”, were tested by a *large number of consumers.*

Mention here the name of the cultivars / number of products / regions / urban and rural areas / number of consumers (women and men) in each location and region …

It should be useful to include here a map of the different locations and a table with regions / areas / villages, small towns, big cities / number of consumers / number of women / men interviewed.

Table 1: Number of consumers interviewed in the rural and urban areas of the two regions

	Total	Region 1					Region 2					Big city
		Village 1	Village 2	Village 3	Village 4	Small town	Village 1	Village 2	Village 3	Village 4	Small town	
Number of Consumers												
Women												
Men												

2.2 Consumer testing

A method including a hedonic test, a just-about-right (JAR) test, and a check-all-that-apply (CATA) test was used. Consumers (*n = 300*) from different locations in rural and urban areas were asked individually to look/touch/smell/taste each Product sample, one after the other, in a random order, and score the overall liking using a nine-point hedonic scale (from 1. “Extremely dislike, to 9. “Extremely like”).

Consumers were also asked to assess how they perceive the intensity of *2-4* characteristics identified as important in the previous Activities 3 & 4, using the 3-point JAR “Just About Right” scale (1 = “Too low”, too weak, not enough, 2= “Just About Right” and 3 = “Too high, too strong, too much”) for each of the Product samples.

Mention here the JAR quality characteristics chosen. Explain why you made this choice?

Consumers were then asked to select the quality characteristics that better describe each Product sample, among a list of 20-25 sensory characteristics -the most liked and the least liked collected during the previous Activities 3 & 4- using a “Check-All-That-Apply” (CATA) approach. Finally, consumers were invited to give their opinion and preferences on the Product samples.

Propose a table with the 20-25 CATA quality characteristics. Mention what quality characteristics were identified during Step 2, those identified during Step 3, and those identified during both Activities by using three different colours. Explain why you made this choice.

Note that the CATA table should contain a balance of the most liked and the least liked quality characteristics related to the appearance, odour, texture between fingers, taste, texture in mouth, aroma, and aftertaste of the final products. The quality characteristics should be mainly sensory characteristics rather than emotional characteristics to be useful for WP2. Complete the table below.

Table 2: Quality characteristics identified during the previous activities 3 & 4 and selected for building the CATA table

	Quality characteristics of the ready to eat product
List of **the most liked** characteristics	Appearance - - odour - - Texture when touching - - Taste - - Texture in mouth - - Aroma - Aftertaste -
List of the least liked characteristics	Appearance - - Odour - - Texture when Touching - - Taste - - Texture in mouth - - Aroma - Aftertaste -

2.3 Data analysis

An analysis of variance (ANOVA) was carried out to identify significant differences in Overall liking scores between the *4-5 Product samples* as tested by *300* consumers. The region or gender effect can be studied. Multiple pairwise comparisons were applied using the Tukey test, with a confidence interval of 95% at $p < 0.05$ (n=300 consumers). For each Product sample, the number of consumers who judged each specific characteristic either Just All Right (JAR), Too weak or Too strong was

counted, and the percentage of consumers (*out of 300*) was determined. A Principal Component Analysis (PCA) was used to describe the relationships between frequencies of citation of CATA sensory characteristics and the mean Overall liking scores for each Product sample. All statistical analyses were performed using XLSTAT 2019 software (Addinsoft).

3 RESULTS

3.1 Overall liking of the product samples

Please refer to the Excel file "Consumer testing WP1 Workshop Benin".

The first worksheet entitled "Raw data" contains all the data collected during our Consumer testing in Benin during the WP1 Workshop.

The Overall liking scores for each Product sample tested by consumers in Benin (here n=40 consumers, but you should have n=300 consumers in the two regions in your country) were extracted from the "Raw data" worksheet and organized in one column in a new worksheet entitled "ANOVA data", one product below the other, with the corresponding Overall liking scores in a second column.

The ANOVA analysis is conducted on another worksheet "ANOVA" using Overall liking scores as dependent variable and Product samples as Qualitative explanatory variable, and using a Turkey test with a confidence interval of 95% (see Tutorial in Appendix A: One-way ANOVA & multiple comparisons).

See below the analysis of our data.

The overall liking of the product significantly differed between the four samples at a significant level of $p<0.05$ (one-way ANOVA) (Table 1).

Table 3: Mean overall liking scores for the four product samples tested

Product Samples	Mean Overall liking scores* (n consumers)	Groups**			
426	8.2	A			
329	6.9		B		
851	4.5			C	
153	2.6				D

*Overall liking was rated on a nine-point scale from 1 = dislike extremely, to 9 = like extremely.
**Different letters correspond to the products, which are significantly different. Tukey test ($p<0.05$).

The most liked product samples were the 426 and the 329 samples with a mean overall liking score close to 8 (like very much) and 7 (like moderately) respectively. The least liked was the 153 sample with a mean overall liking score between 2 (dislike very much) and 3 (dislike moderately). The 851 sample got a medium score, between 4 (dislike slightly) and 5 (neither like nor dislike).

Comment on your results in relation with the varieties or the process used.

The same type of one-way ANOVA analysis can be used with another factor than the Product samples – for example gender factor, or regions' factor or rural/urban areas' factor. But in these cases, you need to use a one-way ANOVA analysis for each product separately to see if there is a significant difference between men and women consumers, or between the two regions, or between the rural areas (consumers interviewed in the 8 villages) and the urban areas (consumers in small towns and in the big city) respectively, regarding the Overall liking of each product.

3.2 Segmentation of consumers into groups of similar overall liking

The overall liking scores for each product tested by all the consumers were extracted from the "Raw data" and organized in a new Excel worksheet entitled "AHC data", with the first column for the consumers and the four other columns for the Overall liking scores of the four Product samples (426, 851, 153, and 329). Please refer to the Excel file "Consumer testing WP1 Workshop Benin").

The AHC analysis is conducted on the Overall liking scores of the four Product samples and will appear in a new worksheet entitled "AHC" (see Tutorial in Appendix B: Agglomerative Hierarchical Clustering).

The aim of an Agglomerative Hierarchical Clustering (AHC) analysis is to create homogeneous clusters of consumers who have similar Overall liking scores. It is useful to classify consumers who have been interviewed randomly, into similar groups.

In our example, by using an Agglomerative Hierarchical Clustering analysis of the mean overall liking scores, we identified three groups of consumers that we have named "426 & 329 likers", "426 likers" and "153 dislikers". These three clusters contained 25%, 15% and 60% of all the consumers interviewed respectively.

There were significant differences ($P < 0.001$) in the overall liking of the three clusters (Figure 1 and Figure 2).

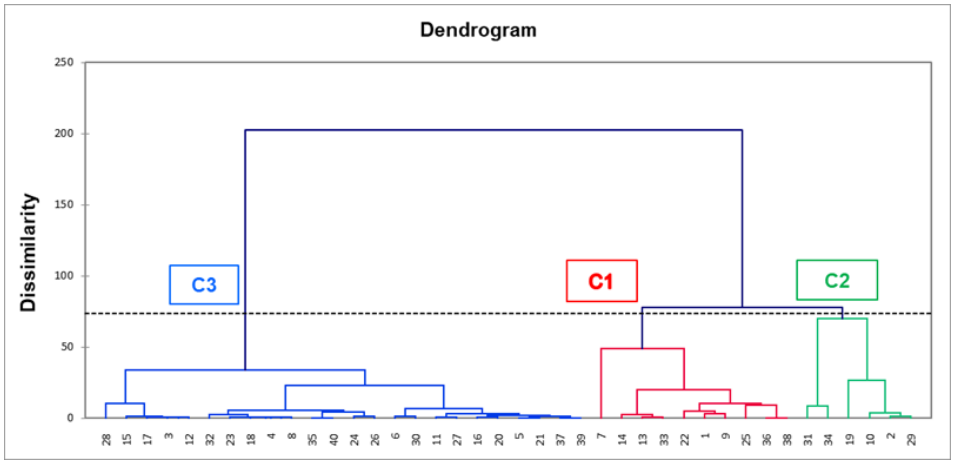
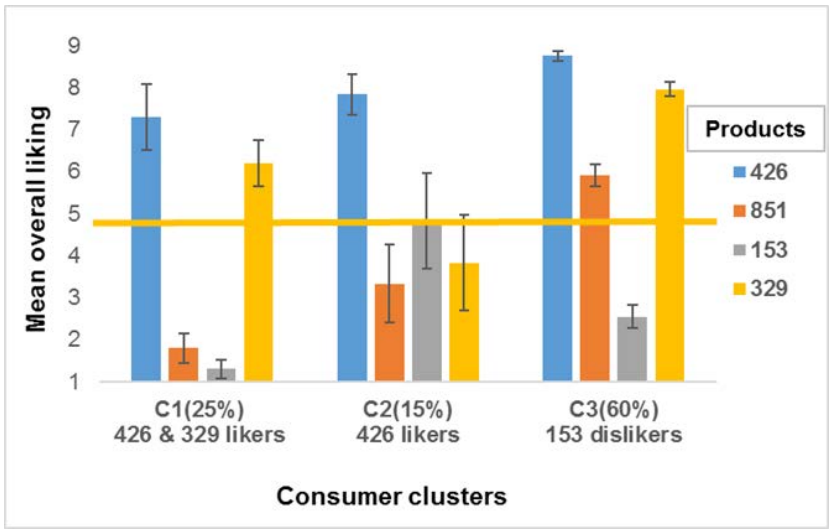


Figure 1: Clustering of the consumers based on their overall liking scores of the product



Where: error bars represent the standard error.

Figure 2: Mean overall liking of the product samples by consumer cluster type (%)

You can calculate the standard error (SE) of the mean Overall liking per cluster and per product (see Tutorial in Appendix B: Agglomerative Hierarchical Clustering) by using a formula (calculation in a new worksheet entitled "AHC standard error").

You can also calculate the standard error (SE) as follows: calculate the standard deviation (STDEV) using a PivotTable (see Tutorial in Appendix B: Agglomerative Hierarchical Clustering) and apply a formula to calculate the SE (calculation in a new worksheet entitled "AHC SE-PivotTable).

3.2.1 Demographic data of the consumers interviewed

Among the *n consumers* interviewed, *x%* were women and *y%* were men. *z%* were 18-25 years old… Most of them were employed as …

Comment if the three clusters differ in terms of sociological characters such as gender, age or level of studies (Table 4).

Table 4: Demographic differences of the consumers with respect to cluster division

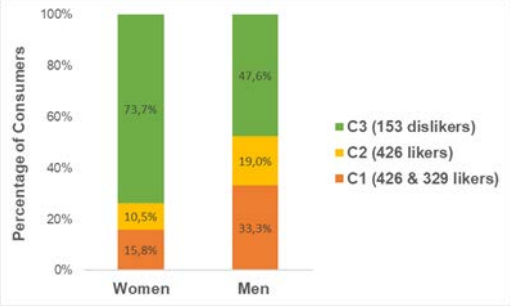
		Total	C1 426 & 329 likers	C2 426 likers	C3 153 dislikers
	Number of consumers (n)				
Gender	Women (%)				
	Men (%)				
Age	18-25 years old (%)				
	26-35 years old (%)				
	36-45 years old (%)				
	56-65 years old (%)				
Ethnicity	Fon (%)				
	Other				
Marital status	Single (%)				
	Married (%)				
	Widower (%)				
Occupation	Student (%)				
	Artisanship (%)				
	Civil servant (%)				
	Trading business (%)				
	Employed (%)				
Education	Unemployed (%)				
	No education (%)				
	Degree and less (%)				
	High level degree (%)				
	Graduated (%)				
Consumption frequency	Daily (%)				
	Several times a week (%)				
	One time a week (%)				
	Several times a month (%)				
Consumption form	One time a month (%)				
	Dry (%)				
	Added with water (%)				
	Added with water and ingredients (%)				
Occasion of consumption	Sprinkled on beans (%)				
	Cooked into piron (%)				
	Breakfast (%)				
	Lunch (%)				
	In between meal (%)				
	Dinner (%)				

3.2.2 Consumption attitudes

Most of consumers interviewed were used to consume the product daily (*x% of answers*). *y%* are used to consume the product several times a week, … or once or several times a month (*z%*)…

The first form of consumption was …. (*x% of answers*), the second was …. (*y%*), and then...

The product is mainly consumed at *lunch time* for more than *x% of consumers* interviewed. Only *y%* of interviewees consume it …



Percentage of consumers	Gender		Total
Class	Women	Men	
C1	7.5%	17.5%	25.0%
C2	5.0%	10.0%	15.0%
C3	35.0%	25.0%	60.0%
Total	47.5%	52.5%	100.0%

Figure 3: Percentage of consumer cluster type by gender

47.5% of all the consumers interviewed were women and among them, 73.7% were 153 dislikers (Cluster 3) and only 10.5% of them were 426 likers (Cluster 2). At the opposite, 47.6% of the men out of a total of 52.5% of the consumers interviewed, were 153 dislikers (Clusters 3) and 33.3% were 426 & 329 likers (Cluster 2).

Using a PivotTable (see Tutorial in Appendix B: Create a PivotTable), allows you to present your results and disaggregate the data by gender (as in the example above, refer to "AHC Class-Gender" worksheet in the Excel file "Consumer testing WP1 Workshop Benin").

You can also present your results by region, or by rural/urban areas (in villages or in the big city).

3.3 A Just About Right test (JAR)

Just about right (JAR) scale was used to determine the optimum level of intensity as perceived by the consumers for *some important sensory quality characteristics* of the *Product samples*. Such "descriptor diagnostic" may help understand why consumers like or dislike this *Product* sample.

Consumers were asked to give their perception of the Colour, Dryness and Sourness of each Product sample, by using a 3-point JAR scale (1 = "Too low, too weak, not enough", 2= "Just About Right" and 3 = "Too high, too strong, too much").

JAR data were extracted from the "Raw data" in the Excel file "Consumer testing WP1 Workshop Benin", and organized in a new worksheet "JAR data" with the first column for the consumers, the second column for the four Product samples, one product below the other, and the three other columns for the JAR scores of the three sensory characteristics "Colour", "Dryness" and "Sourness".

Using a PivotTable (refer to Appendix B: Create a PivotTable to analyse JAR data), allows you to calculate (in a new worksheet entitled "JAR") the percentage of consumers who scored each Product sample as "Too low, too weak, not enough", or as "JAR Just About Right", or as "Too high, too strong, too much".

Table 5: Percentage of consumers who scored the three specific sensory characteristics

	Colour		
Product samples	**Too clear**	**JAR**	**Too dark**
153	0.0%	7.7%	92.3%
329	18.9%	73.0%	8.1%
426	17.5%	80.0%	2.5%

851	7.5%	32.5%	60.0%

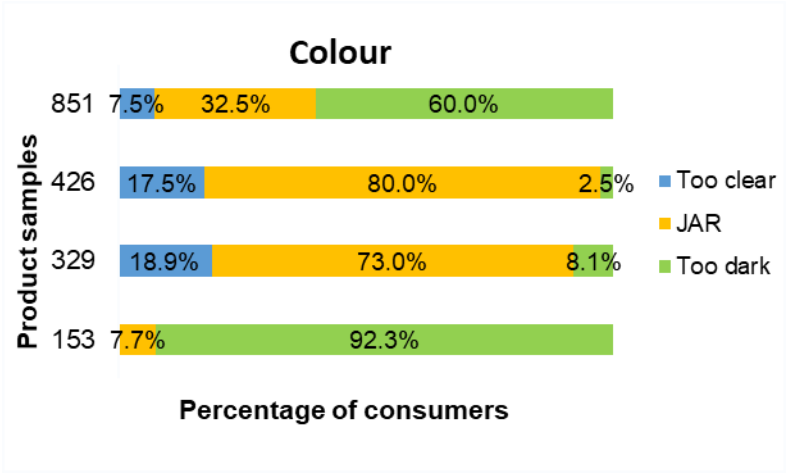
Dryness

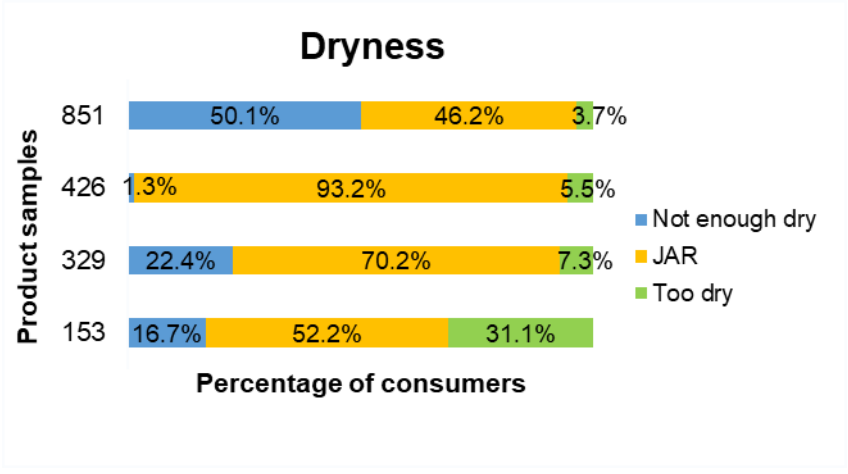
Product samples	Not enough dry	JAR	Too dry
153	16.7%	52.2%	31.1%
329	22.4%	70.2%	7.3%
426	1.3%	93.2%	5.5%
851	50.1%	46.2%	3.7%

Sourness

Product samples	Not enough sour	JAR	Too much sour
153	25.0%	30.0%	45.0%
329	42.1%	47.4%	10.5%
426	17.5%	75.0%	7.5%
851	52.5%	32.5%	15.0%

You can now create a Pivot Chart or a Histogram (see Tutorial in Appendix B: Create a PivotTable to analyse JAR data) to visualize the JAR data (in the worksheet "JAR").





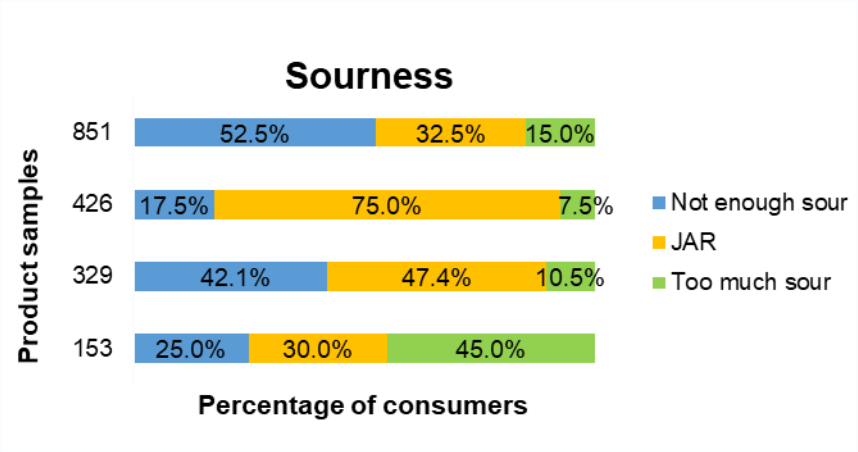


Figure 4: Percentage of consumers who scored the three specific quality characteristics

A majority of consumers was satisfied with the three sensory characteristics of the 426 & 329 product samples: Colour was scored "Just About Right" by 80 and 73% of consumers respectively, the Dryness was scored JAR by 93.2 and 70.2% of consumers respectively, and the Sourness was also scored JAR by 75 and 47.4% of consumers respectively.

The 851 & 153 product samples were perceived "Too dark" by 60 and 92.3% of consumers respectively. Moreover, the 851 Product sample was found "Not enough dry" by 50.1% of consumers, and "Not enough sour" by 52.5% of consumers. The 153 Product sample was scored "Too much sour" by 45% of consumers. Its lower mean overall liking (2.5) could be explained mainly because it was found "Too dark" and "Too much sour" by a majority of consumers.

Sourness was considered "Just About Right" by a high percentage of consumers (75%) only for the 426 Product sample, and by about 50% for the 329 Product sample.

You can also comment your results by gender by calculating the percentage of women/men who scored the three specific quality characteristics.

You can also comment your results by region, or by rural/urban area (in villages or in the big city), by calculating the percentage of consumers who scored the three specific quality characteristics in these regions or areas.

3.4 Check All That Apply (CATA) test

The objective of the CATA test is to show the relationships between hedonic Overall liking scores for each Product sample and the frequencies of citation of each CATA sensory characteristic by all the consumers.

After scoring the Overall liking and the perception of some specific sensory characteristics, consumers were invited to choose the most appropriate terms among 20-25 sensory characteristics that better describe each Product sample.

CATA data were extracted from the "Raw data" in the Excel file "Consumer testing WP1 Workshop Benin", and organized in a new worksheet "CATA data" with the first column for the four Product samples, one product below the other, the second column for the Overall liking scored by the consumers for each product, and the other columns (here 20) for the citations of each quality characteristic by the consumers to describe each Product sample.

The count of citations (in green) and the count of a mean Overall liking (in orange) per product and per characteristic were calculated in a new worksheet entitled "CATA citation frequencies" in the Excel file "Consumer testing WP1 Workshop Benin".

Then a summary of the frequency of citations (in green) with a total of citations, and the mean Overall liking (in orange) per product and per characteristic, were reported in a new worksheet entitled "CATA 2".

The frequency of citations given by consumers to describe each Product sample were calculated (Table 6).

The sensory characteristics most frequently cited by the consumers were considered the best for describing the products. They were the following: “Fine” and “Dry” with a frequency of citation between 100 and 150, followed by “Little Sour” and “White” with a frequency of citation between 75 and 100. The least used terms were “Coarse” and “No taste”.

The 426 Product sample was described as “White” by consumers (35 citations), “Fine” and “Dry” (32 citations respectively), “Little Sour” (26 citations), with a “Good Taste” and “Attractive” (36 and 34 citations respectively). Consumers used the same characteristics to describe the 329 Product sample with almost the same frequencies of citation.

The 153 Product sample was qualified as “Brown” (35 citations), “Burnt” and “Bitter” (24 and 21 citations respectively. However, it was perceived “Dry” (36 citations) as the three other Product samples.

The 851 Product sample ….

Table 6: Frequency of citations of each quality characteristic by all the consumers

Quality characteristics	426	851	153	329	Total
Fibers	2	13	10	11	36
Little sour	26	27	13	26	92
Sweet	14	3	4	9	30
Too acidic	1	2	13	4	20
Good taste	36	15	7	32	90
Beautiful	18	1	2	14	35
Fermented	8	6	16	10	40
Fermented odor	5	5	9	4	23
Bitter	0	2	21	1	24
Coarse	2	3	13	1	19
No taste	0	12	5	0	17
Yellow	2	39	2	0	43
White	35	0	3	36	74
Burnt	0	6	24	0	30
Brown	0	3	35	2	40
Attractive	34	7	5	29	75
Heavy	10	5	14	9	38
Fine	32	24	16	28	100
Dry	32	30	36	34	132
Mean overall liking	8.2	4.5	2.6	6.9	

You can also comment your results by gender by calculating the percentage of women/men who used these sensory terms to describe the Product samples.

You can also comment on your results by region or by rural/urban area (in villages or in the big city), by calculating the percentage of consumers who used these sensory terms to describe the Product samples in these regions or areas.

3.5 Sensory mapping of the sensory characteristics

Principal component analysis (PCA) was used to summarize the relationships between CATA sensory characteristics, Product samples, and mean Overall liking of each product scored by all the consumers.

The PCA analysis is conducted on a new worksheet "PCA" in the Excel file "Raw data Consumer testing Benin".

In the Excel sheet "CATA data 2", select the frequency of citations for all the quality characteristics (in green) as the observations/variables, the Product samples (in blue) as the observation labels, and the Mean Overall liking for each Gari sample (in orange) as a supplementary quantitative variable (refer to Appendix D: Principal Component Analysis PCA).

The PCA plot explained 98.65% of the variance of the sensory characteristics, the first and second axes accounting for 67.93% and 30.73% respectively. Most of the variance was explained by the first axis.

The loading of sensory characteristics on PCA plan (Figure 5) shows that axis 1 was mainly explained positively by the terms such as "White", "Fine", "Sweet" related to the most liked Product samples (426 & 329) and negatively by the terms such as "Burnt", "Coarse", "Bitter", "Fermented odour", "Brown", and "Too acidic" related to the least liked Product sample (153) (*refer to Squared cosines of the variables in Appendix D: Principal Component Analysis*).

Axis 2 was mainly explained positively by the terms such as "Yellow" and "Little Sour" related to the 851 Product sample, and negatively by the terms such as "Heavy", "Dry" and "Fermented" (*refer to Squared cosines of the variables in Appendix D: Principal Component Analysis*).

A high Mean Overall liking scored by consumers was related to the high quality characteristics such as "White", "Fine", "Sweet", "Good Taste (here on the right part of the PCA plan), which were associated to the most liked Gari samples (426 & 329, made from good cassava varieties).

At the opposite, a low Mean Overall liking by the consumers were related to the low quality characteristics such as "Burnt", "Coarse", "Bitter", "Fermented odour", "Brown", and "Too acidic" (here on the left part of the PCA plan), which were associated to the least liked Gari sample (153, bad variety).

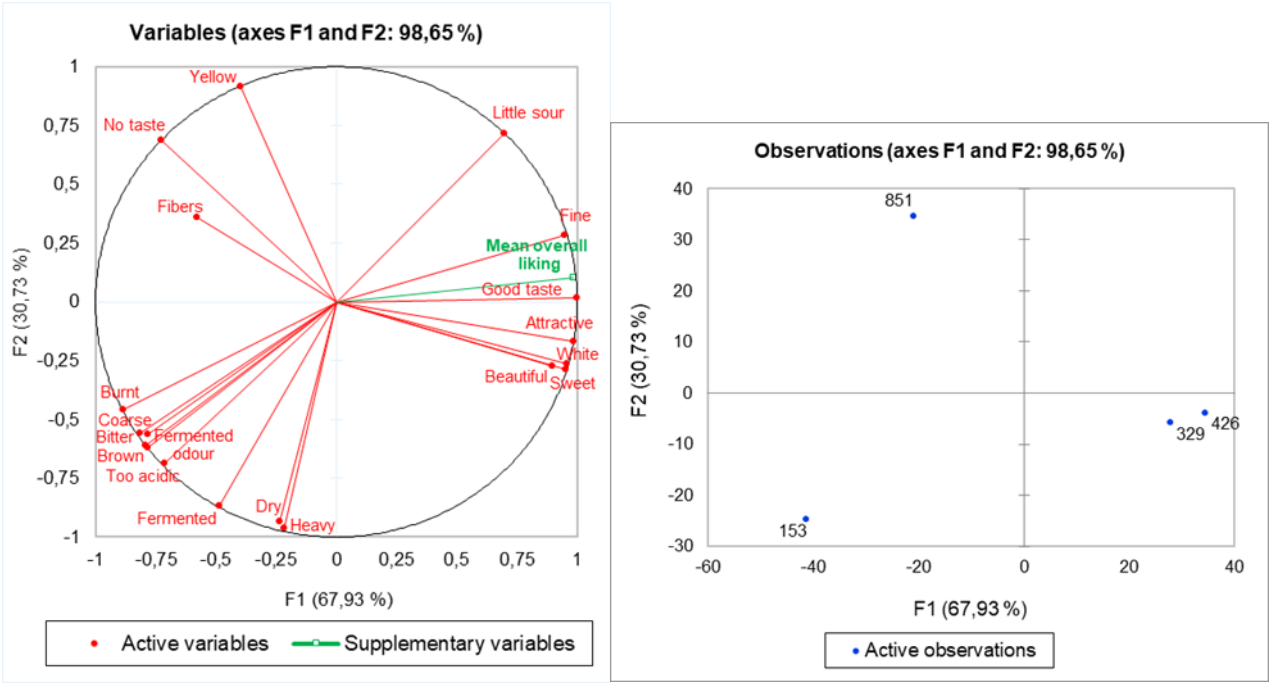


Figure 5: Mapping of the sensory characteristics and the overall liking of the product samples

You can also comment your results by gender, by regions or by rural/urban areas (in villages or in the big city), by mapping the cited sensory characteristics and the overall liking of the product samples scored by women/men, or consumers in each region or area respectively.

4 DISCUSSION AND CONCLUSION

The four Product samples were perceived differently by consumers.

The least liked 153 Gari sample got the lowest mean overall liking score (2.5), mainly because it was found “Too dark” and “Too much sour” by the Beninese consumers (JAR test).

The terms that better describe the product samples were …

The conclusion should be focused on the results to deliver to WP2.

You should pinpoint the high quality characteristics related to a high Mean Overall liking by the consumers (here on the right part of the PCA plan), and associated to the most liked Product samples (good varieties).

Conversely, you should pinpoint the low quality characteristics related to a low Mean Overall liking by the consumers (here on the opposite left part of the PCA plan), and associated to the least liked Product samples (low quality varieties).

5 APPENDICES: TUTORIALS FOR CONSUMER TESTING DATA ANALYSIS

5.1 Annex 1 One-way ANOVA & multiple comparisons

This tutorial shows how to set up and interpret a one-way Analysis of Variance (ANOVA) followed by Tukey's HSD multiple comparisons test in **Excel** using the XLSTAT software.

5.1.1 Selecting the dataset for running a one-way ANOVA

The Excel sheet entitled "Raw data" correspond to the Consumer testing fieldwork conducted by the WP1 partners during our WP1 workshop in April 2018, Cotonou Benin (see Excel file "Consumer testing WP1 Workshop Benin"). During this fieldwork, four Gari products coded as follows: 426, 851, 153, 329 were tasted (one after the other) by 40 Beninese consumers who gave an overall liking score for each of the products.

Using the ANOVA function of XLSTAT, we want to find out whether the four Gari samples differ significantly in terms of their overall liking as scored by the 40 Beninese consumers and, if they do differ, which product is the most liked. We use a one-way ANOVA because there is only one **factor - the Gari samples** - and no repetition.

From the Excel sheet "Raw data", create a new Excel sheet that you will name "ANOVA data". The first column will be used for the factor, also called **Qualitative explanatory variable** (here the "Gari samples"), and the second column for the Dependent variable (here the "Overall liking"). The dependent variable (here the "Overall liking") measured during our Consumer testing fieldwork is "dependent" on the independent variable also called factor or explanatory variable (here the "Gari samples").

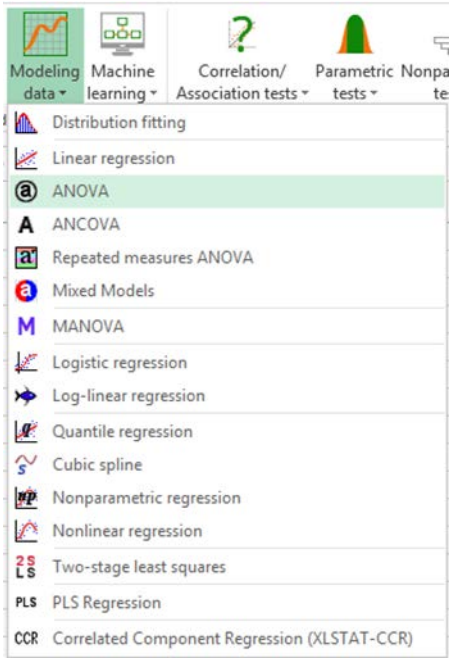
Organize the "ANOVA data" as shown in the Excel file in two columns, with each Gari sample, one below the other, in the first column of the Excel sheet, with the corresponding Overall-liking score in the second column (from Raw data: column AA for 426 Gari sample, column AY for 851, BW for 153 and column CU for 329 Gari sample).

This one-way ANOVA analysis can be used with another **factor** than the **Gari samples** – for example **gender** (men or women in the first column) or **the regions** or **the rural/urban areas**, and Overall liking as a dependent variable (in the second column).

But if we want to know whether gender, regions, or rural/urban areas will affect the Overall liking, then we have to use a one-way ANOVA analysis for **each product separately** to see if there is a significant difference between men and women consumers, between the two regions, or between rural areas (consumers interviewed in the 8 villages) and urban areas (consumers in small towns and in the big city) respectively, regarding the Overall liking of each product.

5.1.2 Conducting a one-way ANOVA

Once XLSTAT is opened, select the **XLSTAT / Modeling data / ANOVA command** (see below).



Once you have clicked on the button, the ANOVA dialog box appears (see below).

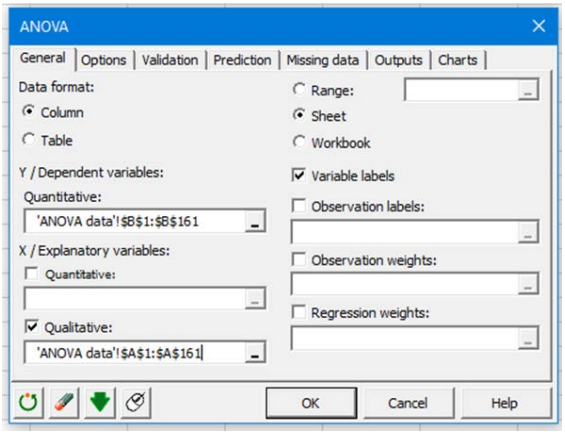
In **XLSTAT**, it is possible to select the data in two different ways for the **ANOVA** (column or table). We will choose the form of column (dialog box below), one column for the **dependent variable**, another for the **explanatory variable**.

Click in the window **Quantitative Dependent Variable**, and select the data in the "ANOVA data" worksheet as follows: Select the second column that contains the "Overall liking" scores $B$1:$B$161.

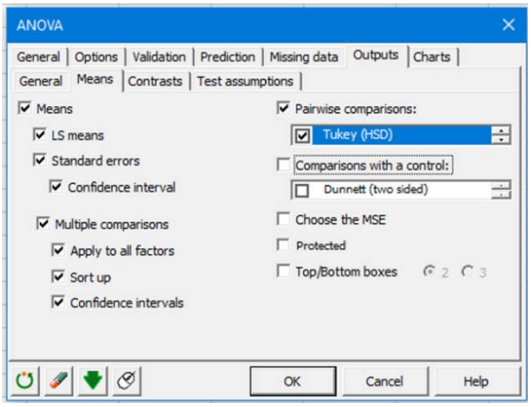
Click in the window **Qualitative Explanatory variable**, and select the data in the "ANOVA data" worksheet as follows: Select the first column that contains the Gari samples codes $A$1:$A$161.

While selecting the column title for both variables, also tick the option **Variable labels** (dialog box below).

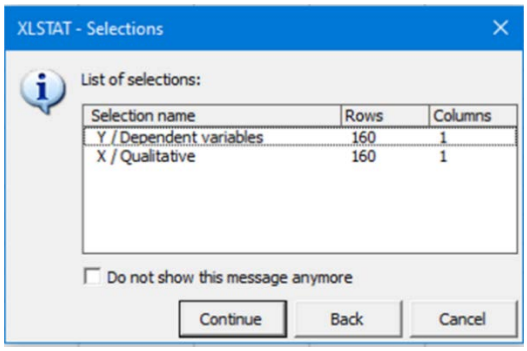
In this example, we want to display the results on another sheet "ANOVA" where the data are stored, so we chose the **Sheet** option (dialog box below).



In the **Outputs** tab (**Means** sub-tab), select the **Pairwise comparisons** option to run a **Tukey's test**.



The computations begin once you have clicked on the **OK** button.



Once you have clicked on the **Continue** button, the computations resume and the results are displayed in a new worksheet called “ANOVA”.

5.1.3 Interpreting the one-way ANOVA results

The first important results displayed by XLSTAT are the **goodness of fit statistics**.

The **goodness of fit** coefficients includes the coefficient of determination R^2, the adjusted R^2 and several other statistical coefficients (see below).

The **coefficient of determination R^2** (here 0.563) tells how much of the variability of the modeled variable (here the “Overall liking”) is being explained by the explanatory variables (here the “Gari samples”); in our case, we have 56.3% of the variability of the Overall liking explained by the Gari samples.

Regression of variable Overall liking	
Goodness of fit statistics (Overall liking):	
Observations	160.000
Sum of weights	160.000
DF	156.000
R^2	**0.563**
Adjusted R^2	0.555
MSE	3.788
RMSE	1.946
MAPE	60.725
DW	2.109
Cp	4.000
AIC	217.029
SBC	229.330
PC	0.459

The next table down is the **Analysis of variance**. This is an important result to look at because it tells whether there are significant differences between the Gari samples in terms of their Overall liking (see below). The null hypothesis H0 is that there is no difference between the Gari samples. If we reject the hypothesis H0 this means that there are significant differences in the mean Overall liking between the Gari samples.

Analysis of variance (Overall liking)

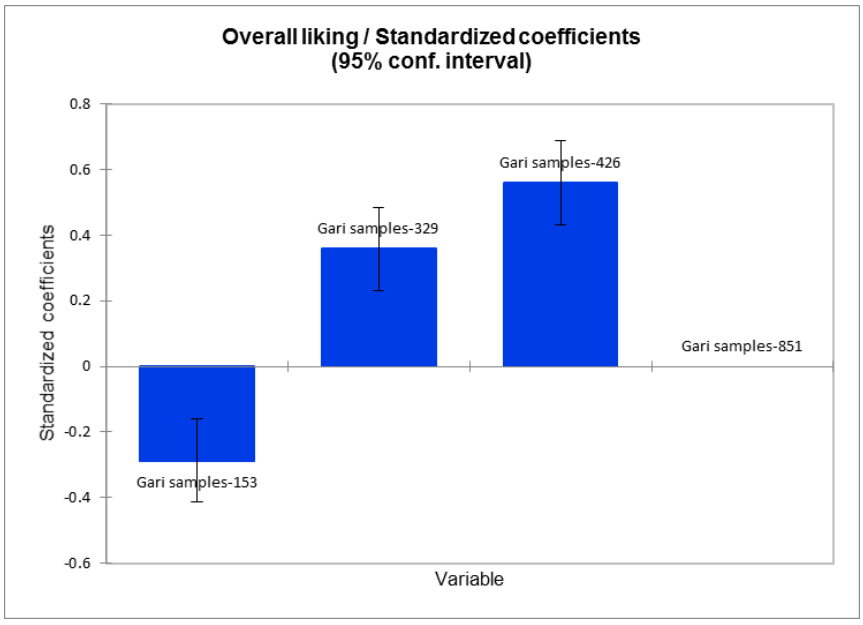
Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	3	762.619	254.206	67.114	**< 0.0001**
Error	156	590.875	3.788		
Corrected Total	159	1353.494			

Computed against model Y=Mean(Y)

The test used here is the **Fisher's F** test. A probability Pr > F less than 0.0001 means that we would take a 0.01% risk to conclude that the null hypothesis (no effect of the type of Gari samples on the Overall liking) is wrong.

We can therefore conclude with confidence that there is a significant influence of the type of Gari samples on the Overall liking as scored by the Beninese consumers.

The bar chart of the **standardized coefficients** shows the relative impact of the categories or explanatory variables (here the Gari samples).



Now that we have obtained an answer to our initial question: is there a significant difference between the Gari samples, the next step is to classify the difference between the samples.

As shown in the next table, the **Tukey's HSD** (Honestly Significantly Different) test is applied to all pairwise differences between means. The analysis of the differences between the categories is done with a **confidence interval of 95%** (this means that Pr > Diff <0.05 will indicate a significant difference between the Gari samples).

Gari samples / Tukey (HSD) / Analysis of the differences between the categories with a confidence interval of 95% (Overall liking)

Contrast	Difference	Standardized difference	Critical value	Pr > Diff	Significant
426 vs 153	5.675	13.041	2.597	**< 0.0001**	Yes
426 vs 851	3.750	8.617	2.597	**< 0.0001**	Yes
426 vs 329	1.350	3.102	2.597	**0.012**	Yes
329 vs 153	4.325	9.938	2.597	**< 0.0001**	Yes
329 vs 851	2.400	5.515	2.597	**< 0.0001**	Yes
851 vs 153	1.925	4.423	2.597	**0.000**	Yes
Tukey's d critical value:			3.673		

Based on the p-values below (Pr > Diff), all the pairs appear to be significantly different. This can also be confirmed by the **95% confidence intervals** (last four columns). If an interval does not contain zero, then we can reject the null hypothesis that there is no significant difference between the two means.

The overall scores for the different Gari samples are then ordered in hierarchical order. Different letters (A, B, C, D indicate significant differences in the mean Overall liking of the Gari samples). The table gives a summary of all pairwise comparisons of Gari samples using the Tukey's HSD test.

Summary of all pairwise comparisons for Gari samples (Tukey (HSD)

Category Gari samples	LS means Overall liking	Groups			
426	8.250	A			
329	6.900		B		
851	4.500			C	
153	2.575				D

Other multiple comparison tests can be applied such as the **Dunnett's test** that compares each category with a control category.

5.1.4 Concluding on the one-way ANOVA analysis

The conclusion is that the four Gari samples significantly differ in terms of their Overall liking as scored by the Beninese consumers.

5.2 Annex 2 Agglomerative Hierarchical Clustering (AHC)

This tutorial will help you set up and interpret an Agglomerative Hierarchical Clustering (AHC) in **Excel** using the XLSTAT software.

5.2.1 What is Agglomerative Hierarchical Clustering?

Agglomerative Hierarchical Clustering (AHC) is a classification method whose principle is simple. The process starts by calculating the dissimilarity between the N objects. When two objects minimize a given agglomeration criterion, they are clustered together, thus creating a class comprising these two objects. Then the dissimilarity between this class and the N-2 other objects is calculated using the agglomeration criterion. When the two objects or classes of objects minimize the agglomeration criterion, they are then clustered together. This process continues until all the objects have been clustered.

These successive clustering operations produce a binary clustering tree (dendrogram), whose root is the class that contains all the observations. This dendrogram represents a hierarchy of partitions.

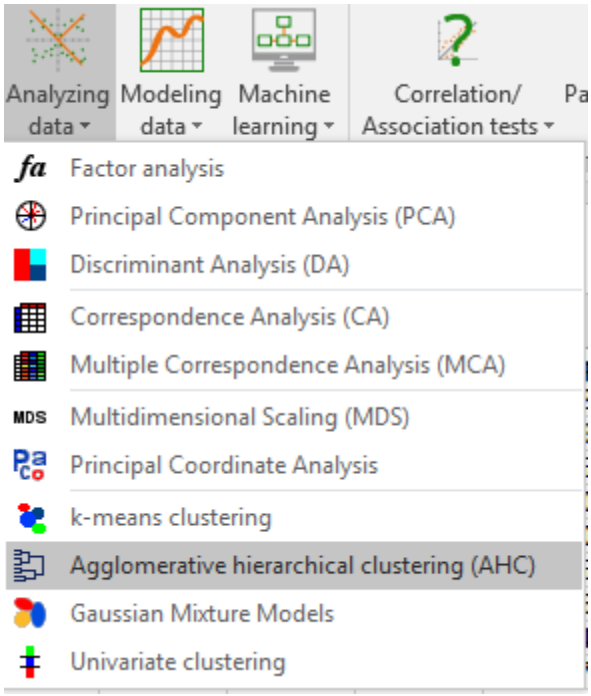
5.2.2 Selecting the dataset to run an Agglomerative Hierarchical Clustering in XLSTAT

From the Excel sheet “Raw data”, create a new Excel sheet that you will name “AHC data”. The first column will be used for the consumers and the four columns for the Overall liking of the four Gari samples (426, 851, 153, and 329).

The aim of AHC is to create homogeneous clusters of consumers who have similar Overall liking scores.

5.2.3 Setting up an Agglomerative Hierarchical Clustering

Once XLSTAT is activated, go to **XLSTAT / Analysing data / Agglomerative Hierarchical Clustering.**

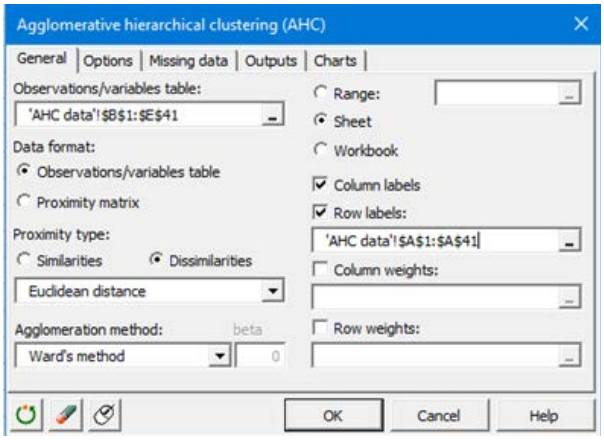


The **Hierarchical Clustering** dialog box will appear. Then select the data on the Excel sheet.

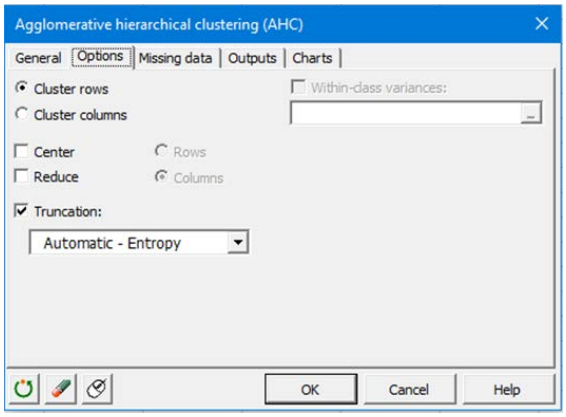
There are several ways of selecting data with XLSTAT. In our example, we will select columns. This explains why the columns are labelled.

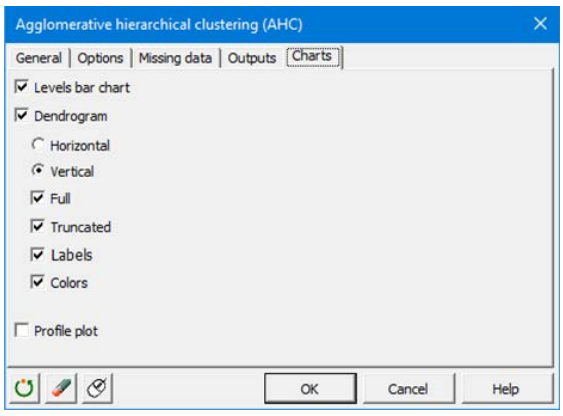
Click in the window **Observations/variables table**, and select the data in the “AHC data” worksheet as follows: Select the columns B, C, D, E containing the “Overall liking” scores for the four Gari samples $B$1:$E$41.

Click in the window **Row labels**, and select the data in the “AHC data” worksheet as follows: Select the first column that contains the Consumers $A$1:$A$41.

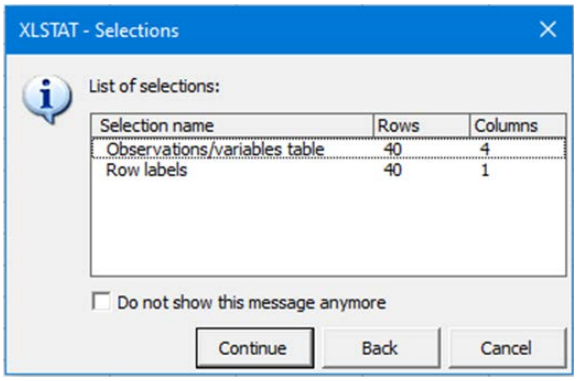


In the **Options** tab, select the **Automatic - Entropy truncation**. The automatic truncation is based on the entropy and tries to create homogeneous groups of consumers having similar Overall liking of the Gari samples.





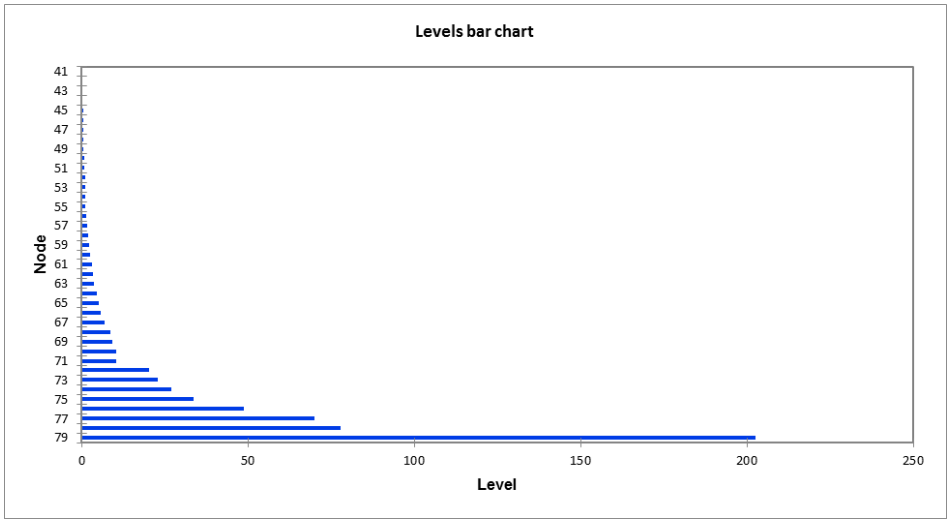
The computations begin once you have clicked on **OK**.



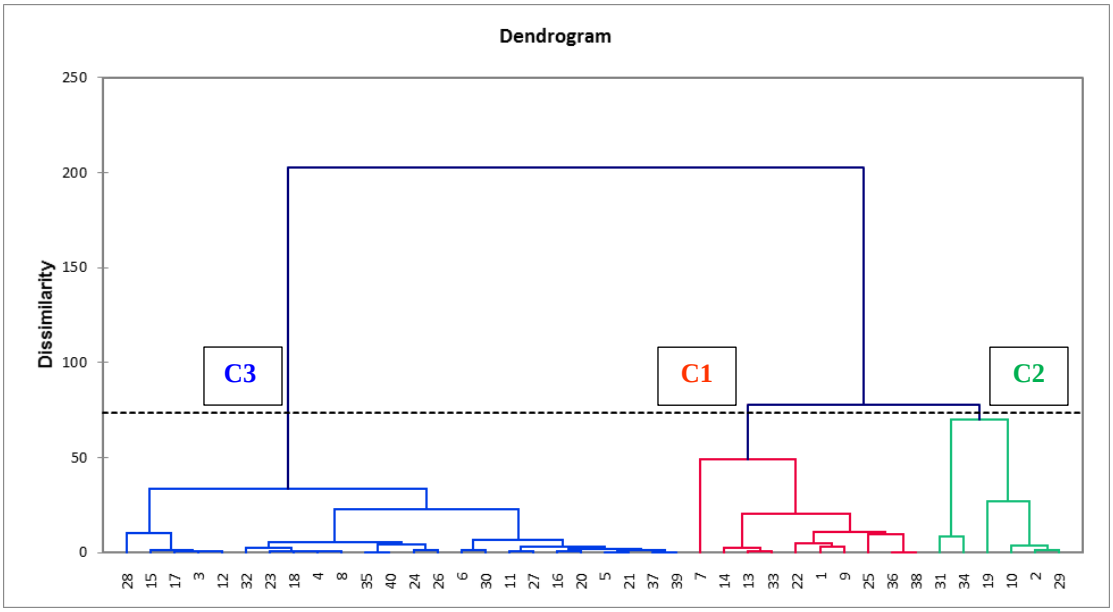
Once you have clicked on the **Continue** button, the computations resume and the results are displayed in a new worksheet called “AHC”.

5.2.4 Interpreting the results of an Agglomerative Hierarchical Clustering

The first result to look at is the levels bar chart. The shape reveals a great deal about the structure of the data. When the increase in dissimilarity level is strong, we have reached a level where we are grouping groups that are already homogenous. Automatic truncation uses this criterion to decide when to stop aggregating observations (or groups of observations, here the consumers).



The chart below is the dendrogram. It represents how the algorithm works to group the observations (the consumers), then the sub groups of observations. As you can see, the algorithm has successfully grouped all the observations. The dotted line represents the automatic truncation, leading to three groups (three Clusters).



The three groups do not necessarily have the same size. In this example, the C3 cluster (displayed in blue colour, 24 consumers) is more homogeneous (it is flatter on the dendrogram) than the C1 cluster (displayed in red colour, 10 consumers), and the C2 cluster (displayed in green colour, 6 consumers).

This is confirmed when looking at the **Within-class variance**. C1 and C2 exhibit higher variance than C3 (see the table below).

Results by class

Class	C1	C2	C3
Objects (40 consumers)	10	6	24
Sum of weights	10	6	24
Within-class variance	11.044	21.967	4.402
Minimum distance to centroid	1.364	2.804	0.604
Average distance to centroid	2.804	4.128	1.868
Maximum distance to centroid	6.623	6.287	5.036

The following table shows how the consumers (here the 40 consumers interviewed in Benin, April 2018) have been classified into each cluster (C1, C2 or C3).

Results by object

Observations (Consumers)	Class
1	1
2	2
3	3
4	3
5	3
6	3
7	1
8	3
9	1
10	2
11	3
12	3
13	1
14	1
15	3

16	3
17	3
18	3
19	2
20	3
21	3
22	1
23	3
24	3
25	1
26	3
27	3
28	3
29	2
30	3
31	2
32	3
33	1
34	2
35	3
36	1
37	3
38	1
39	3
40	3

5.2.5 Create a histogram chart

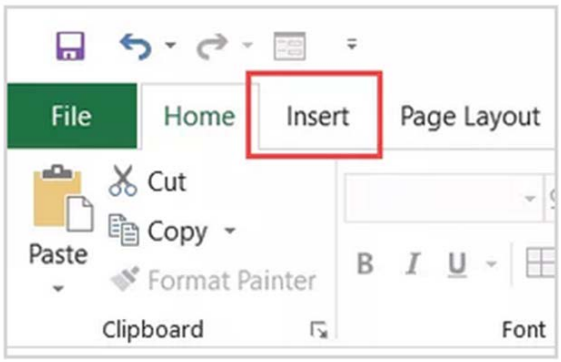
The means of the Overall liking per cluster and per Gari sample are given in the following table (displayed in the Excel sheet “AHC”).

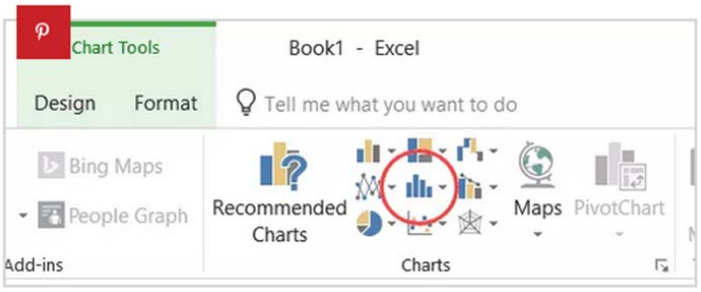
From the class centroid table (here below), you can create a histogram chart to visualize the data.

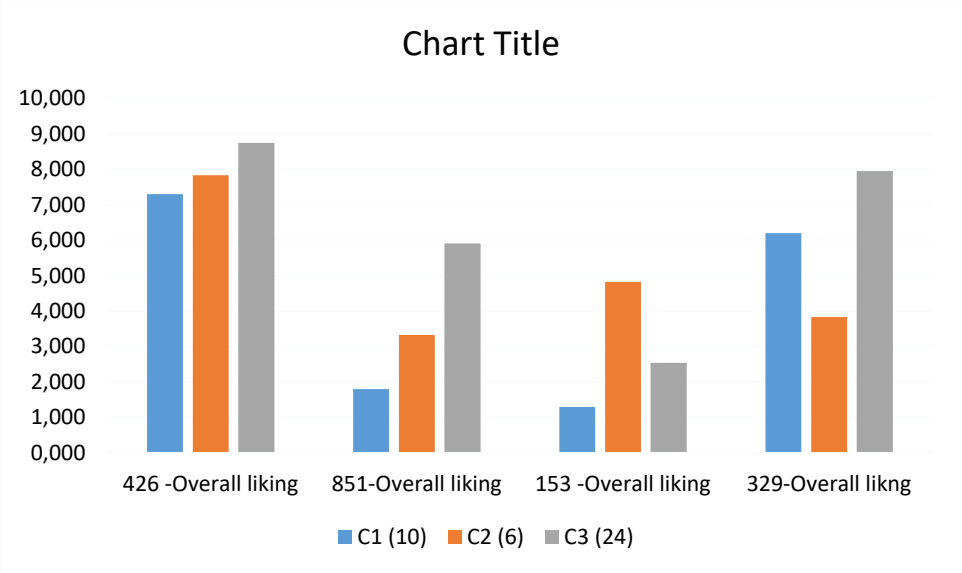
1. Label the clusters C1, C2, C3 in the “Class” column
2. Select your dataset table
3. Click Insert > Chart.
4. In the Insert Chart dialog box, under All Charts, click **Histogram**, and click OK.

Class centroids

Class	426-Overall liking	851-Overall liking	153-Overall liking	329-Overall liking
C1 (10)	7.300	1.800	1.300	6.200
C2 (6)	7.833	3.333	4.833	3.833
C3 (24)	8.750	5.917	2.542	7.958

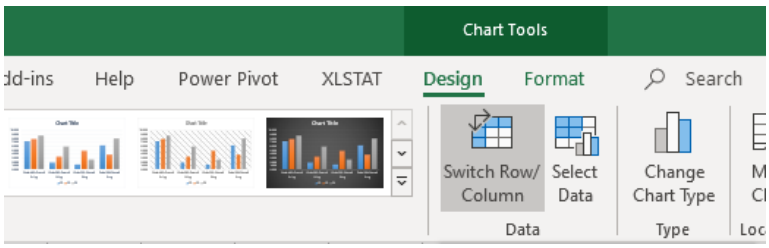


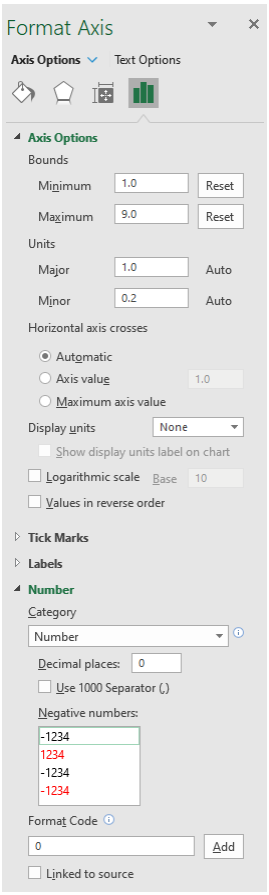




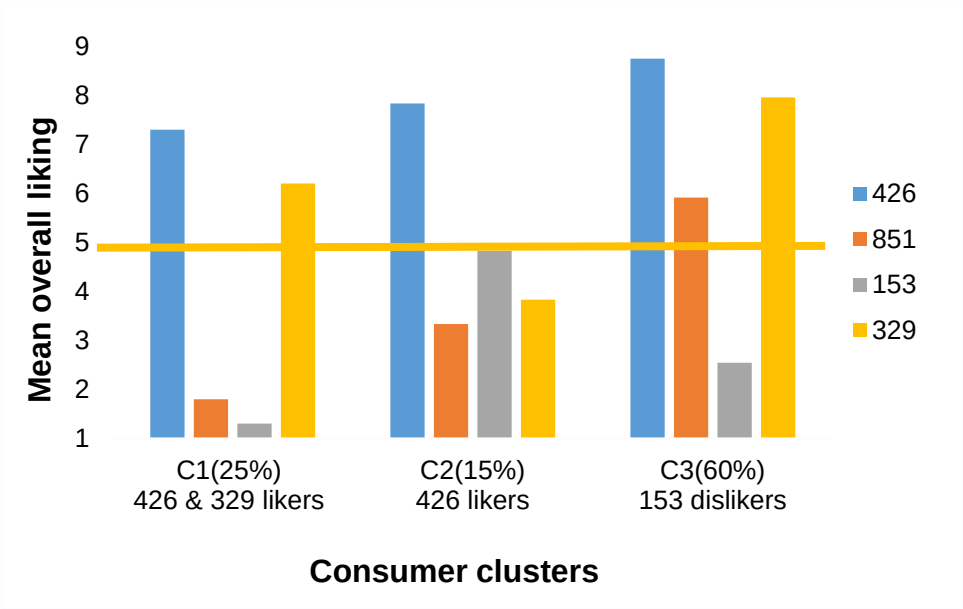
Now you can customize this chart by selecting:

- Design, Switch row/column, Ok





- Delete the grids by clicking on them and press the “delete” button
- Delete the “Chart Title” by clicking on it and press the “delete” button
- Format the legend: “Legend format”, “Right”
- Format the y-axis: “y-axis”, “format axis
    - “Axis options”, bounds, minimum 1, maximum 9
    - “Number”, decimal places 0
- Click on the graph, click on the **+** on the right, title of the axes
- Insert “connectors” “line” to create a line and move it midway on the graph at y-axis = 5 (neither like nor dislike)
- Name the clusters based on the Overall liking of the Gari samples as in the example below and calculate the percentage of the consumers in each cluster.



5.2.6 Calculating standard errors

Create a new Excel sheet entitled “AHC Standard error”. Add the column “Class” found in the “AHC” worksheet (and here above) to the five columns of the “AHC data” worksheet.

Organize the data in the Excel sheet as follows:

- Click on “Sort” the data
- Select “Custom sorting”, Expand selection
- Sort on “Class” data, on the Cell value, from the smallest to the biggest.

Calculate the **standard error (SE)** of the Mean Overall liking for each Gari sample scored by the Consumers in each cluster.

The standard error of the mean (SE) can be expressed as the **standard deviation (STDEV)** of the Mean Overall liking divided by the square root of the number of consumers in each cluster.

SE = STDEV /√(n)

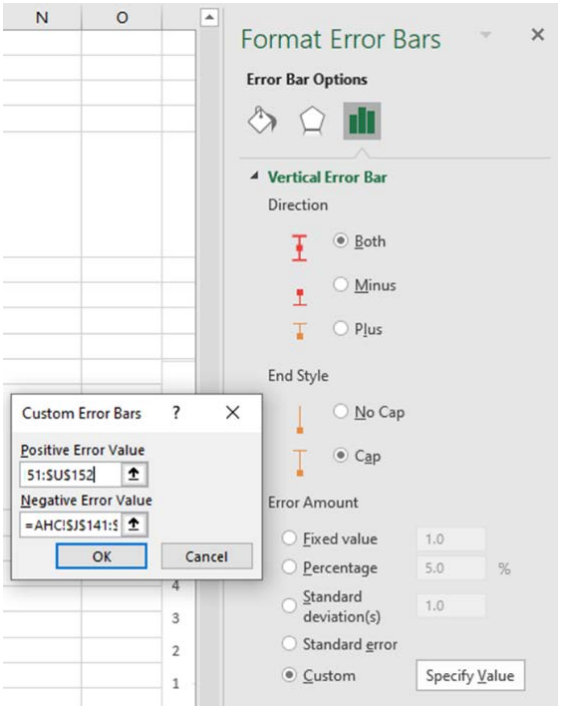
Where n is the number of consumers in each cluster: $n_{(C1)} = 10$; $n_{(C2)} = 6$; $n_{(C3)} = 24$.

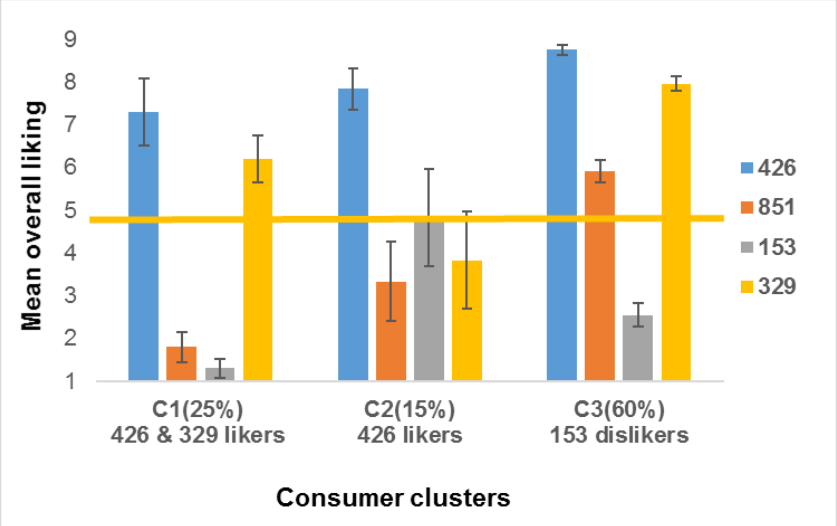
SE

Class	426	851	153	329
C1(10) 426 & 329 likers	0.78951462	0.35901099	0.21343747	0.55377492
C2(6) 426 likers	0.47726070	0.91893658	1.13773654	1.13773654
C3(24) 153 dislikers	0.10851434	0.25478512	0.27570178	0.17527584

5.2.7 Adding standard errors to the histogram

- Click on a histogram bar (ex: blue bar, Gari sample 426)
- Click on “Design”, “Add a chart element”
- Select “Error bar”
- Click on “More error bar options”
- Select “Custom”, “Specify a value”
- For the positive standard error of each Gari sample (e.g. code 426, blue bar), select the data in the column corresponding to the Gari sample 426 in the table SE above
- For the negative standard error, select the same data in column corresponding to the Gari sample 426 in the table SE above
- Click on “Ok”
- Repeat the same procedure for each histogram bar (each Gari sample).





5.3 Annex 3 Create a PivotTable to analyse JAR data

A PivotTable is a powerful tool to calculate, summarize, and analyse data that lets you see comparisons, patterns, and trends in your data.

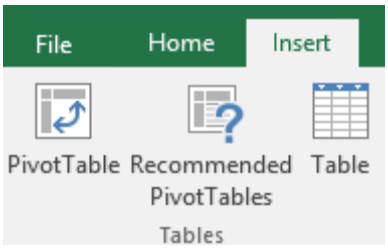
5.3.1 Create a PivotTable

1. From the Excel sheet “Raw data”, create a new Excel sheet entitled “JAR data”. The first column will be used for the Consumers, the second column for the Gari samples, and then three columns for the JAR scores of the three selected descriptors (Colour, Dryness, Sourness).
   Organize the “JAR data” as shown in the Excel sheet “JAR data” with each Gari sample, one below the other in the second column, and the corresponding Consumers and JAR scores for the three descriptors.

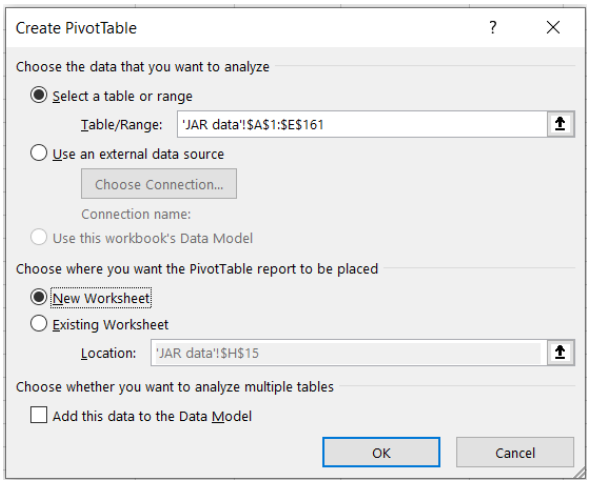
2. Select the dataset “JAR data” you want to create a Pivot Table from.

   Note: Your data should not have any empty rows or columns. It must have only a single-row heading.

3. Select in **Excel, Insert** > **PivotTable**.



4. Under **Choose the data that you want to analyse**, s**elect a table or range**.



5. In **Table/Range** box, select all the data from “JAR data”: $A$1:$E$161.
6. Under **Choose where you want the PivotTable report to be placed**, select **New worksheet** to place the PivotTable in a new worksheet.
7. Select **OK**.

5.3.2 Building out your PivotTable

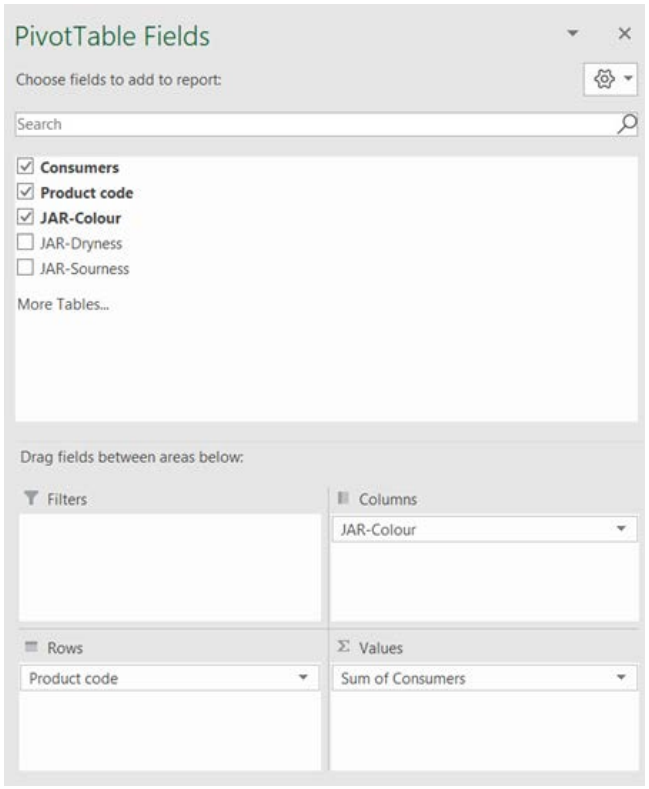
The Field List has a field section in which you pick the fields you want to show in your PivotTable, and the Areas section (at the bottom) in which you can arrange those fields the way you want.

Use the areas section (at the bottom) of the Field List to arrange fields the way you want by dragging them from the Field list to the target area. To delete a field from the PivotTable, drag the field out of its areas section.

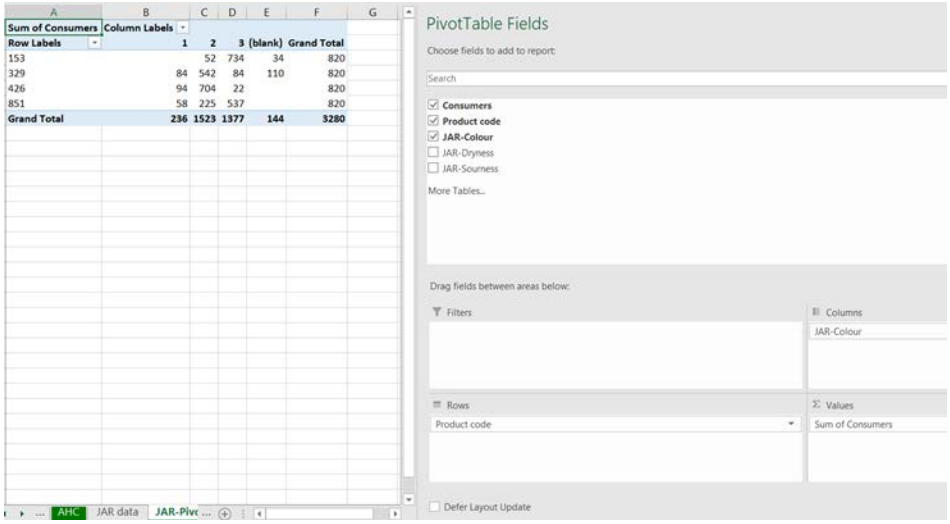
Fields that you place in different areas are shown in the PivotTable as follows:

- **Columns** area fields are shown as **Column Labels** at the top of the PivotTable.
- **Rows** area fields are shown as **Row Labels** on the left side of the PivotTable **Values** area fields are shown as summarized numeric values in the PivotTable

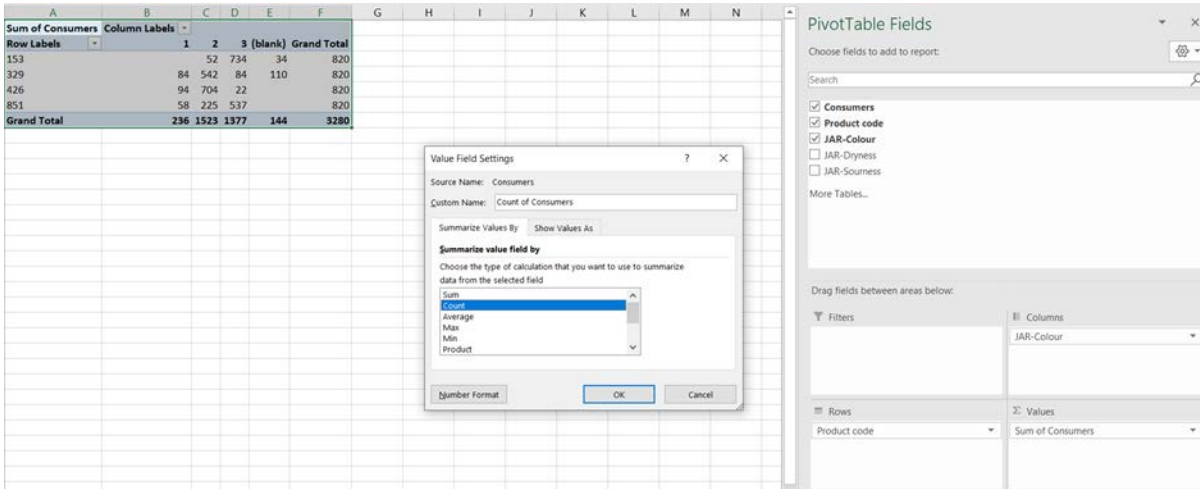
NOTE: Typically, nonnumeric fields are added to the **Rows** area (here **Gari samples**), numeric fields are added to the **Values** area (here **Consumers**), and Online Analytical Processing (OLAP) (here **JAR sensory descriptors**) are added to the **Columns** area, one after the other.



The Field List should appear when you click anywhere in the PivotTable.



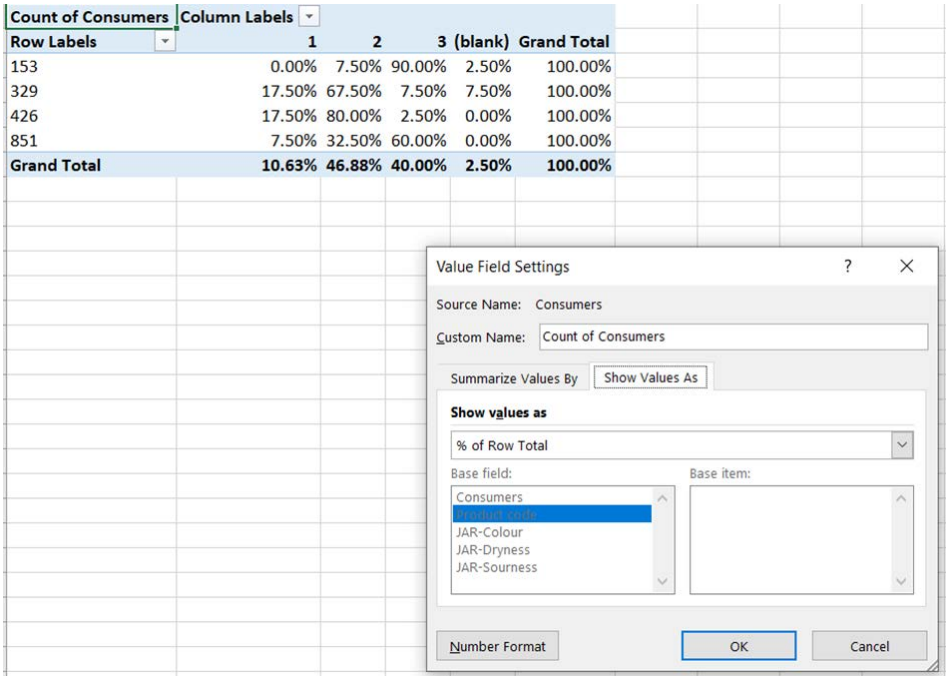
By default, PivotTable fields placed in the **Values** area are displayed in SUM format: Consumers appear as Sum of Consumers and not as Number of Consumers. You can change the default calculation by clicking on the field name “Sum of Consumers” in the Values area, then selecting the Value Field Settings, and in the “Summarize values by” tab, clicking on Count, and Ok.



5.3.3 Display a value as a calculation and a percentage

Instead of using a calculation to summarize the data, you can also display it as a percentage of a field.

In our example, we propose to change the Count of Consumers to show them as a % of the total of Consumers.

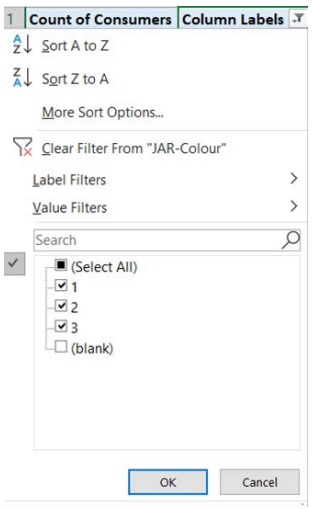


Once you have opened the Value Field Settings dialog box, go to “Show values as” tab.

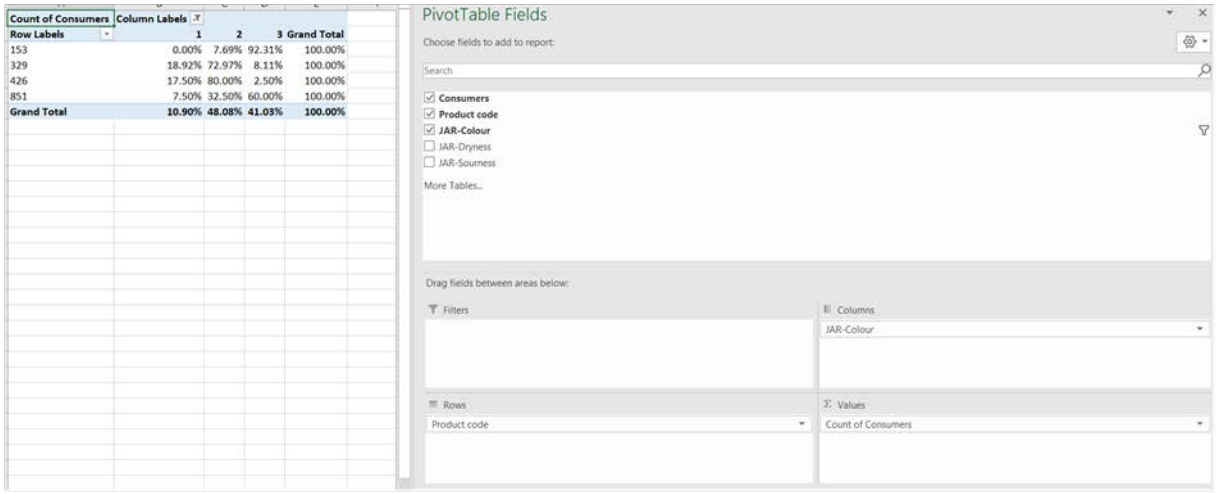
Choose the type of calculation you want to summarize the data in the selected field

“Show values as” option, and here select “% of row total”.

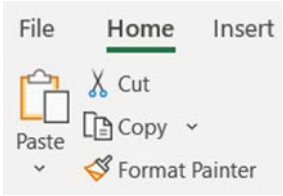
You can remove the “blank column” by clicking on “Column labels” and unticking the “blank column” box.

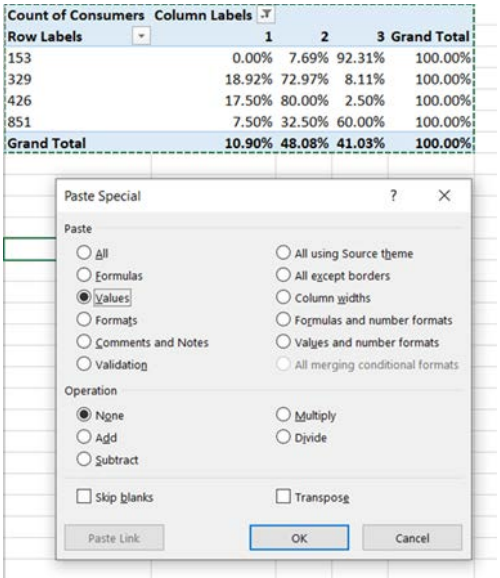


The PivotTable appears with the percentage of consumers who have chosen one of the three options (1= Too clear, 2 = JAR Just About Right as I like, 3 = Too dark) to assess the Colour descriptor for each Gari sample.



Select the pivot table and copy it with “Paste special “Values” into a new worksheet. The table is no longer active when you paste it with “Paste special”.





You can therefore rename the row and column labels in the PivotTable (as shown below). The next step is the creation of a histogram to visualize the data.

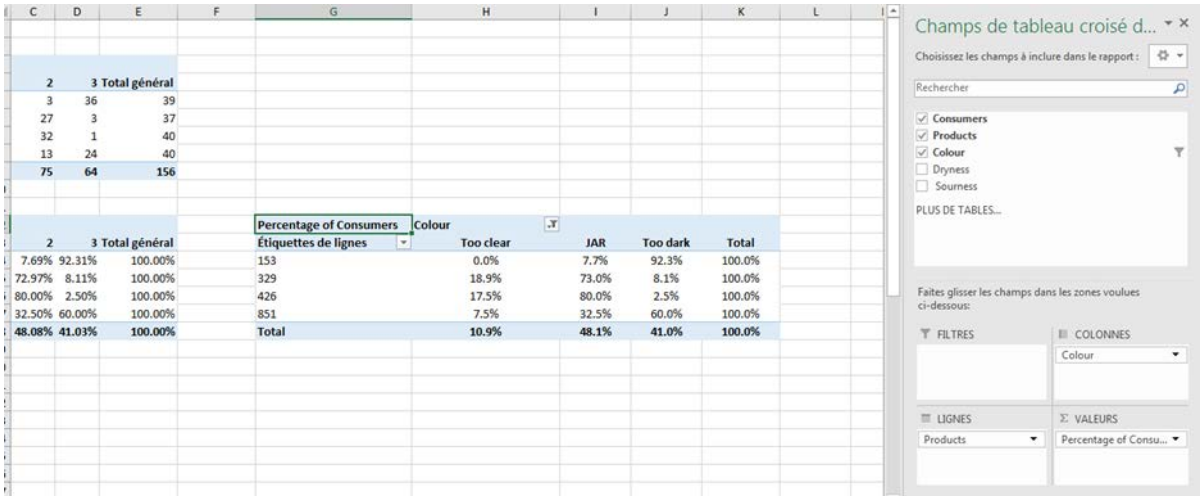
Percentage of Consumers	Colour			
Gari samples	Too clear	JAR	Too dark	Total
153	0,0%	7,7%	92,3%	100,0%
329	18,9%	73,0%	8,1%	100,0%
426	17,5%	80,0%	2,5%	100,0%
851	7,5%	32,5%	60,0%	100,0%
Total	**10,9%**	**48,1%**	**41,0%**	**100,0%**

5.3.4 Create a PivotChart

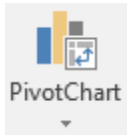
You have two possibilities: create a PivotChart or Insert a Histogram.

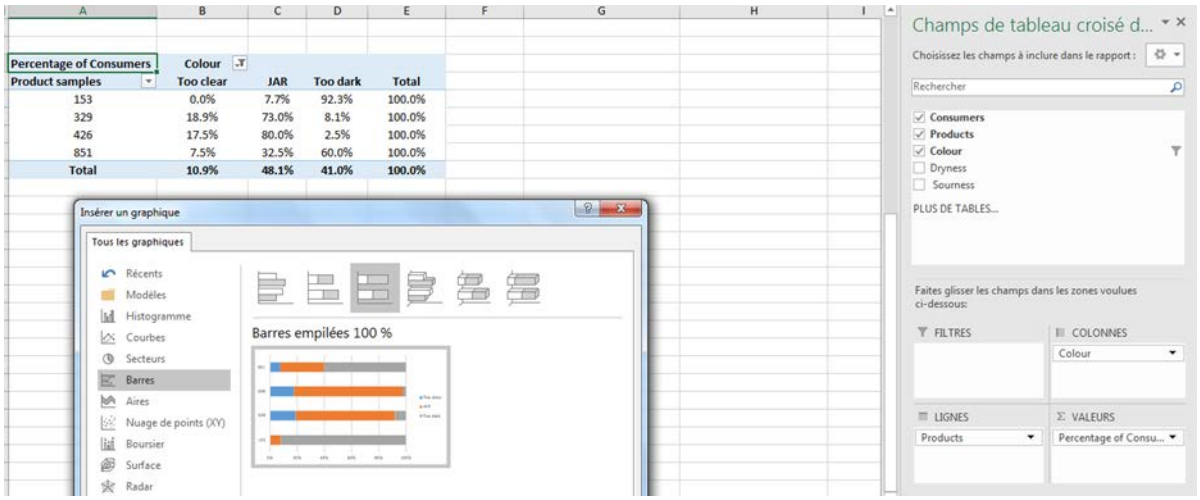
5.3.5 Create a PivotChart

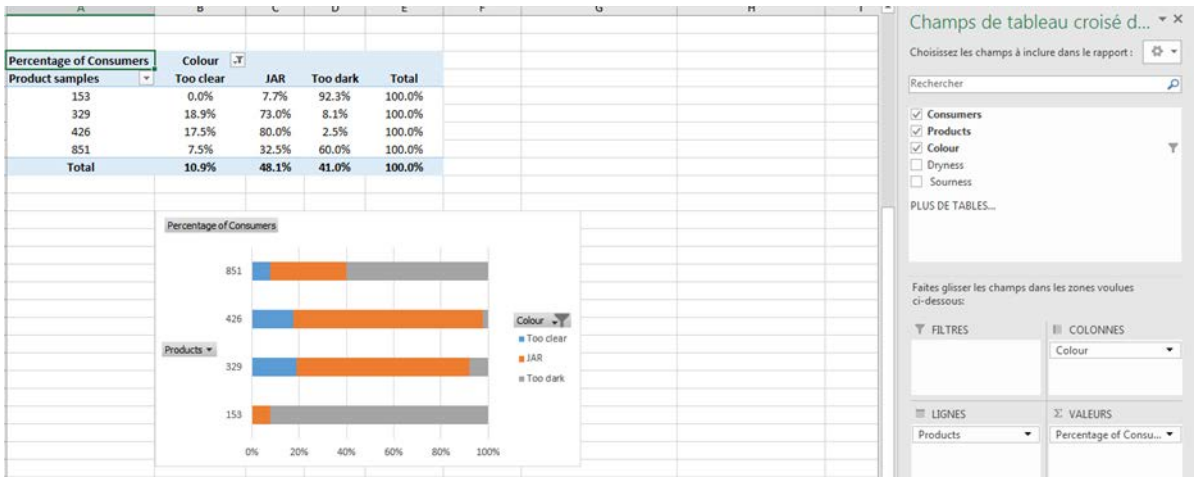
1. Select a cell in your table.



2. Select **Insert > PivotChart**
3. Select **Bars**
4. Select **OK**.







Percentage of Consumers	Colour			
Product samples	Too clear	JAR	Too dark	Total
153	0.0%	7.7%	92.3%	100.0%
329	18.9%	73.0%	8.1%	100.0%
426	17.5%	80.0%	2.5%	100.0%
851	7.5%	32.5%	60.0%	100.0%
Total	10.9%	48.1%	41.0%	100.0%

You can change the colour of the bars, add the values of % of consumers in each bar, remove the X-Axis legend, remove the grid… by right-clicking in the chart.

Percentage of Consumers	Colour			
Gari samples	Too clear	JAR	Too dark	Total
153	0,0%	7,7%	92,3%	100,0%
329	18,9%	73,0%	8,1%	100,0%
426	17,5%	80,0%	2,5%	100,0%
851	7,5%	32,5%	60,0%	100,0%
Total	10,9%	48,1%	41,0%	100,0%

5.3.6 Create a Histogram Chart

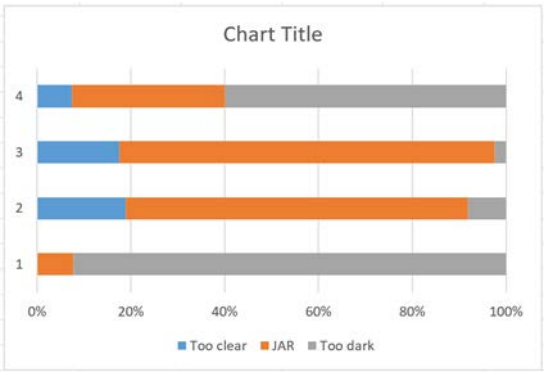
1. Select the pivot table and copy it with “Paste special “Values” beside in the worksheet JAR.
2. Delete the Grand Total and the column labels.
3. Select the data in your table.
4. Select **Insert > Chart**
5. **Select Histogram 2D bar**
6. Select **OK**.

Percentage of Consumers	Colour			
Gari samples	Too clear	JAR	Too dark	Total
153	0,0%	7,7%	92,3%	100,0%
329	18,9%	73,0%	8,1%	100,0%
426		80,0%	2,5%	100,0%
851		32,5%	60,0%	100,0%
Total	10,9%	48,1%	41,0%	100,0%

Percentage of Consumers
Valeur : 18,9%
Ligne : 329
Colonne : Too clear

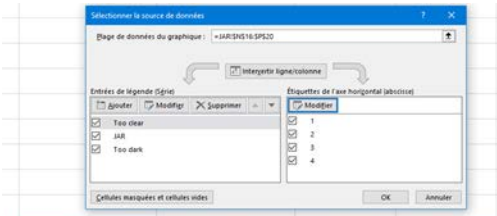
Gari samples	Too clear	JAR	Too dark
153	0,0%	7,7%	92,3%
329	18,9%	73,0%	8,1%
426	17,5%	80,0%	2,5%
851	7,5%	32,5%	60,0%

Gari samples	Too clear	JAR	Too dark
153	0,0%	7,7%	92,3%
329	18,9%	73,0%	8,1%
426	17,5%	80,0%	2,5%
851	7,5%	32,5%	60,0%

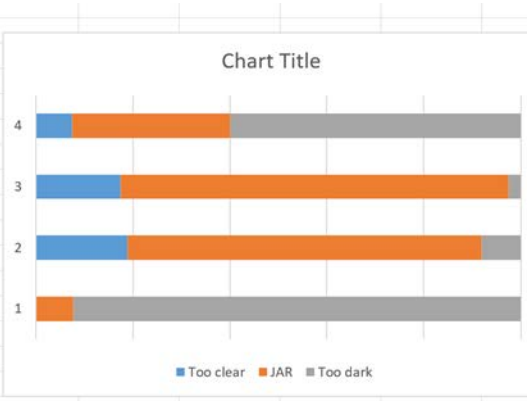


Now you can customize this chart by selecting:

- Click on Y-Axis scale on the Chart
- Design, Select Data, Modify

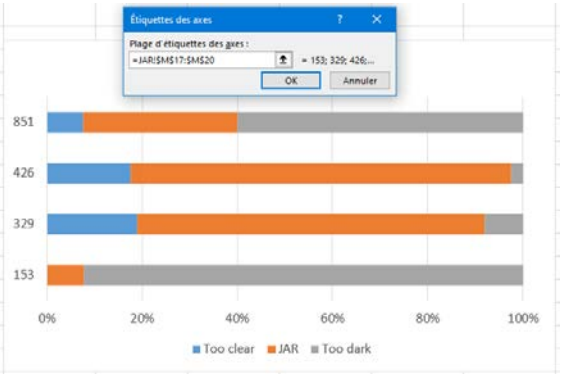


Gari samples	Too clear	JAR	Too dark
153	0,0%	7,7%	92,3%
329	18,9%	73,0%	8,1%
426	17,5%	80,0%	2,5%
851	7,5%	32,5%	60,0%

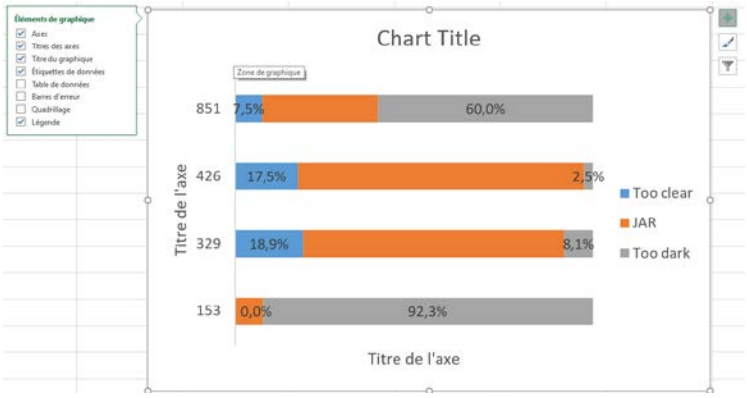


- Click on Modify
- Axis Label, Select Gari samples codes as shown below
- Click on OK.

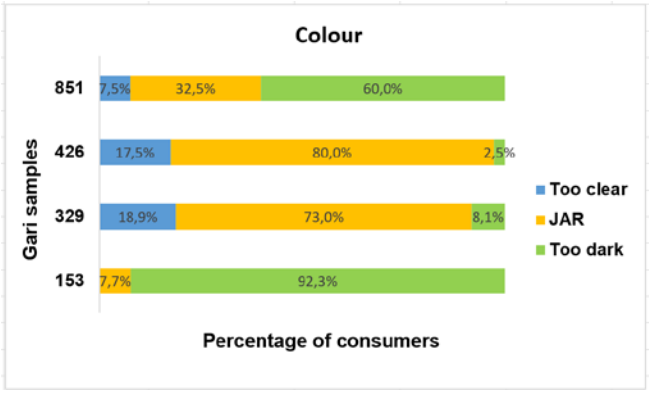
Gari samples	Too clear	JAR	Too dark
153	0,0%	7,7%	92,3%
329	18,9%	73,0%	8,1%
426	17,5%	80,0%	2,5%
851	7,5%	32,5%	60,0%



- Delete the grids by clicking on them and press the “delete” button
- Right Click on the Legend in the Chart, “Legend format”, “Right”
- Right Click on the x-axis scale in the Chart and press the “delete” button
- Click on the graph, click on the **+** on the right, and add Title of the axes
- Click on the graph, click on the **+** on the right, and add data labels on the bar



- You can choose a title for the Chart and for x-axis and for y-axis
- You can modify the colour of each bar by right clicking on the bar and change the colour



5.4 Annex 4 Principal Component Analysis

This tutorial will help you set up and interpret a Principal Component Analysis (PCA) in **Excel** using the **XLSTAT** software.

5.4.1 What is Principal Component Analysis

Principal Component Analysis is one of the most frequently used multivariate data analysis. It investigates multidimensional datasets with quantitative variables.

Principal Component Analysis is a very useful **projection** method to analyse numerical data structured in n observations / p variables. PCA projects observations from a P-dimensional space with p variables to a K-dimensional space (where K<P) so as to conserve the maximum amount of information (information is measured here through the total variance of the scatter plots) from the initial dimensions. It allows to:

- Quickly visualize and analyse correlations between the p variables,
- Visualize and analyse the n observations (initially described by the p variables) in a 2- or 3-dimensional map, in order to identify uniform or atypical groups of observations.

PCA **dimensions** are also called **axes** or **Factors**.

For example, if a table of n observations (here the four Gari samples) are described by p variables (here the CATA quality characteristics or CATA descriptors), and if p is quite high, it is impossible to grasp the structure of the data and the nearness of the observations by merely using univariate statistical analysis methods or even a correlation matrix.

If the information associated with the first 2 or 3 axes represents a sufficient percentage of the total variability of the scatter plot, the observations will be able to be represented on a 2-3-dimensional chart, thus making interpretation much easier.

5.4.2 Dataset for running a Principal Component Analysis

Open two new worksheets in your Excel file named “CATA data” and “CATA citation frequencies” respectively. Copy from your “Raw data” all the CATA data for the four Gari samples and paste them in these two new worksheets.

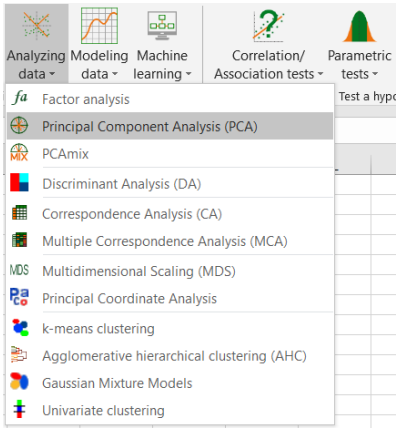
In “CATA citation frequencies” worksheet, calculate the Mean overall liking for each Gari sample (orange line) and the Count of the citations per quality characteristic and per Gari sample (green lines). Alternatively, you can create a Pivot table to calculate those frequencies from “CATA data”.

Copy these summarized data (Mean overall liking and Count of citations per descriptor for each Gari sample) in a new sheet “CATA data 2”.

5.4.3 Setting up a Principal Component Analysis in Excel using XLSTAT

Selecting the data

Once XLSTAT is activated, select the **XLSTAT / Analysing data / Principal component analysis** command (see below).



The Principal Component Analysis dialog box will appear.

Select the data in the Excel sheet entitled “CATA data 2”.

The **Data format** chosen is **Observations/variables**.

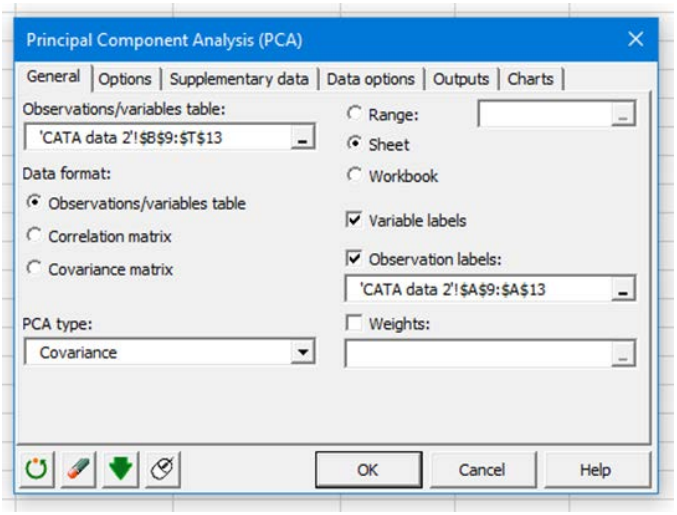
Click in the Observations/variables window and select in the Excel sheet “CATA data 2”, the data of citation frequencies for all the quality characteristics (in green).

Click in the Observation labels window and select the Products (Gari samples) labels (column in blue) in the Excel sheet “CATA data 2”.

Sheet: Activate this option to display the results in a new worksheet.

Principal Component Analysis: what type to choose - Pearson or covariance

The **PCA type** to select here is “Covariance”. Covariance is used when we want to analyse the variance of the variables (here the CATA quality characteristics). Covariance matrices allocate more weight to variables with higher variances.

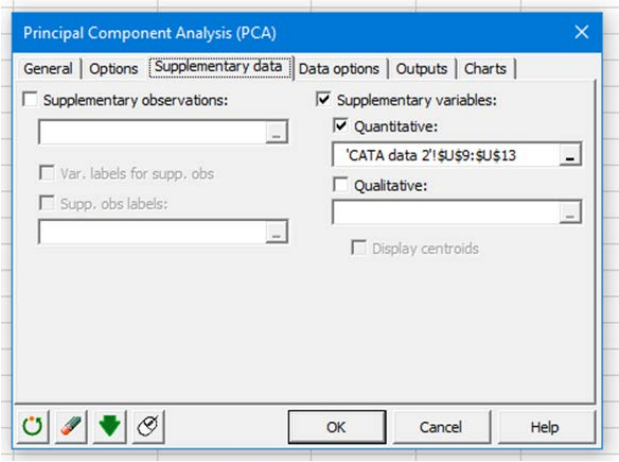


Principal Component Analysis: supplementary variable

XLSTAT allows for additional variables (qualitative or quantitative ones) called **supplementary variables**. This can be used in several contexts. Here are two examples:

- If you want to investigate how a quantitative dependent variable relates to the others (i.e. independent variables) that should be used to build the PCA.
- If you want to see how different categories of observations behave in the PCA space (Men vs Women for example). In this case, a qualitative supplementary variable (gender) may be used to colour observations differently for men and women.

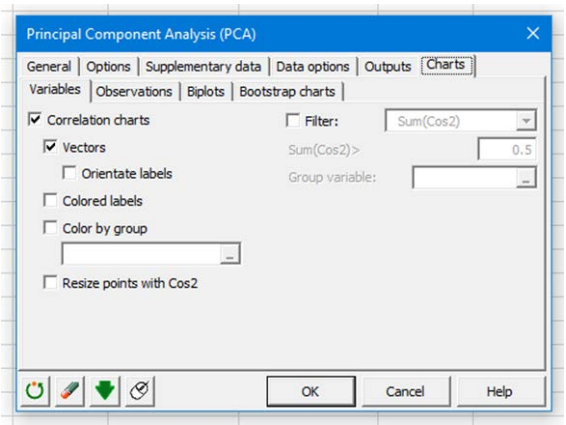
In our example, choose a supplementary quantitative variable in the Supplementary data tab (here below): The Mean Overall liking for each Gari sample (column in orange) in the Excel sheet “CATA data 2”.

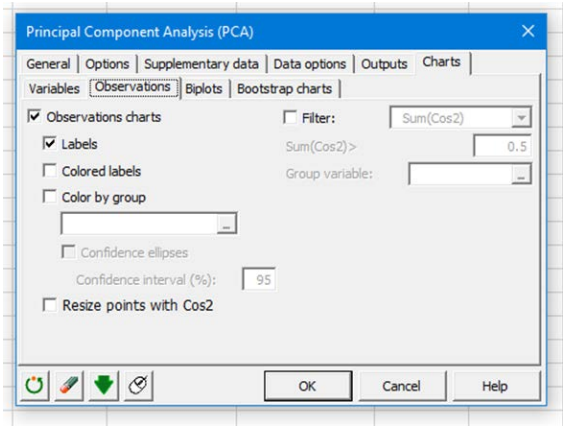


Principal Component Analysis in XLSTAT, configuring charts

In the **Charts** tab, in order to display the labels on all charts, and to display all the observations (observations charts and biplots), the filtering option is unchecked.

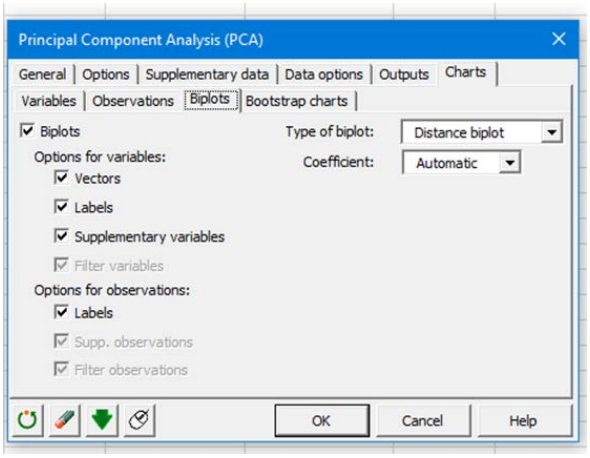
It is possible to simultaneously represent both observations and variables in the factor space. The term biplot is reserved for simultaneous representations which respect the fact that the projection of observations on variable vectors must respect the order and the relative distances of the observations for that same variable, in the input data.



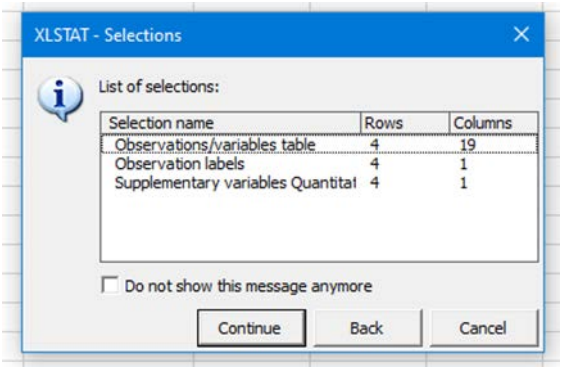


For the graphic representation, a transformation is required in order to make the interpretation precise.

Distance biplot: A distance biplot is used to interpret the distances between the observations. The position of two observations projected onto a variable vector can be used to determine their relative level for this variable. Lastly, the length of a variable vector in the representation space is representative of the variable's level of contribution to building this space.

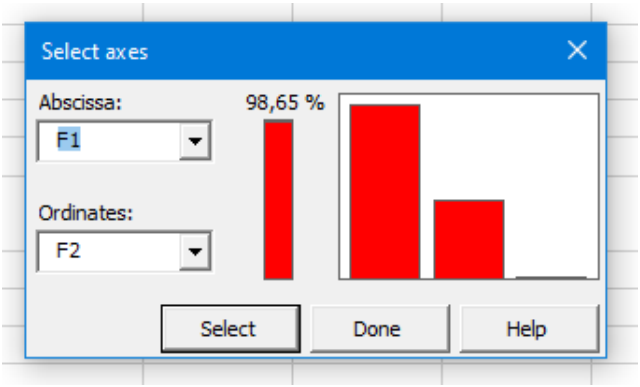


The computations begin once you have clicked on **OK**. You are asked to confirm the number of rows and columns. Then Continue



5.4.4 Principal Component Analysis in XLSTAT - launching the computations

Then you should confirm the axes for which you want to display plots. In this example, the percentage of variability represented by the first two factors is 98.65%.



5.4.5 Interpreting the results of a Principal Component Analysis in Excel using XLSTAT

How to interpret Eigenvalues in Principal Component Analysis

This table and the corresponding chart are related to a mathematical object, the **eigenvalues**, which reflect the quality of the projection of the data. In our example, we can see that the first eigenvalue equals 1025.6 and represents 67.9 % of the total variability. This means that if we represent the data on only one axis, we will still be able to see 67.9 % of the total variability of the data. With two axes, we will see 98.7 % of the total variability, which is very good.

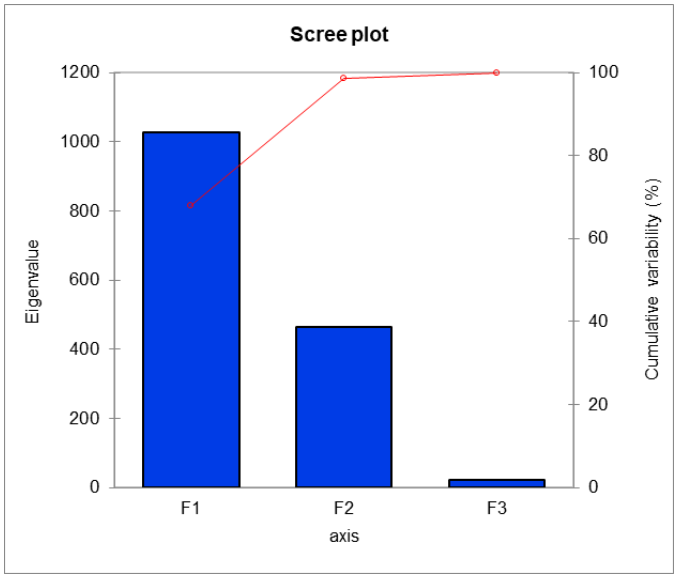
Each eigenvalue corresponds to a factor, and each factor to a one dimension. The eigenvalues and the corresponding factors are sorted by descending order of how much of the initial variability they represent (converted to %).

Ideally, the first two or three eigenvalues will correspond to a high % of the variance, ensuring us that the maps based on the first two or three factors are a good quality projection of the data. In this example, the first two factors allow us to represent 98.7 % of the initial variability of the data. This is a very good result, but we'll have to be careful when we interpret the maps as some information might be hidden in the next factors.

Broadly speaking, factor = PCA dimension = PCA axis

Eigenvalues

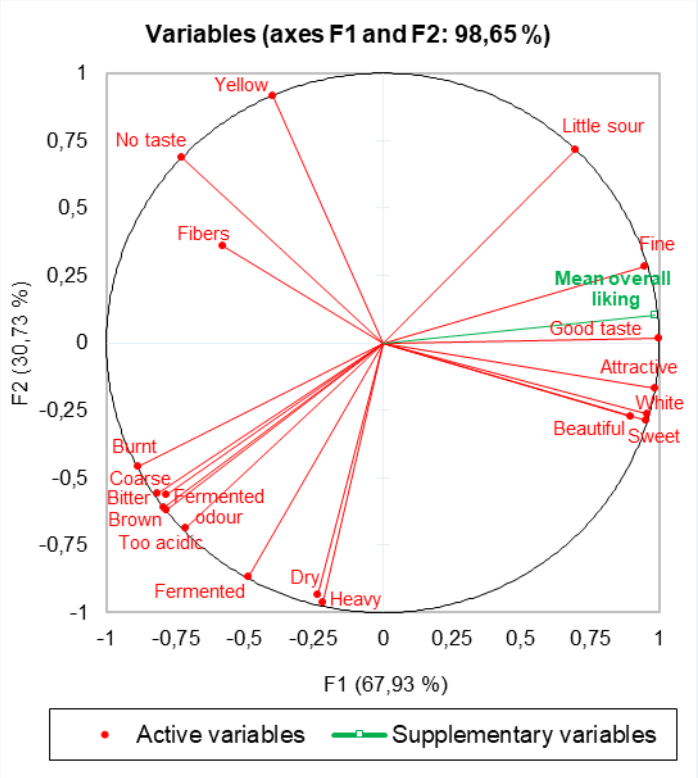
	F1	F2	F3
Eigenvalue	1025.594	463.972	20.309
Variability (%)	67.926	30.729	1.345
Cumulative %	67.926	98.655	100.000



How to interpret results related to variables in PCA

The first map is called the **correlation circle** or **variable chart** (below on axes F1 and F2). It shows a projection of the initial variables in the factor space. When two variables are far from the centre, then, if they are: close to each other, they are significantly positively correlated (r close to 1); if they are orthogonal, they are not correlated (r close to 0); if they are on the opposite side of the centre, then they are significantly negatively correlated (r close to -1). Supplementary variables can also be displayed in the shape of vectors.

When the variables are close to the centre, some information is carried on other axes by looking at the correlation circle on axes F1 and F3.



The correlation circle is useful in interpreting the meaning of the axes. These trends will be helpful in interpreting the next map.

Contributions: Contributions are an interpretation aid. The variables which had the highest influence in building the corresponding PCA axes are those whose contributions are highest.

To confirm that a variable is well linked with an axis, take a look at the squared cosines table: the greater the squared cosine, the greater the link with the corresponding axis. The closer the squared cosine of a given variable is to zero, the more careful you have to be when interpreting the results in terms of trends on the corresponding axis. Looking at this table we can see that the variable “Fibres” would be best viewed on a F1/F3 map.

Contribution of the variables (%)

	F1	F2	F3
Fibres	0.573	0.484	46.145
Little sour	1.578	3.692	0.891
Sweet	1.512	0.308	11.406
Too acidic	1.112	2.281	2.509
Good taste	13.838	0.008	1.427
Beautiful	4.874	0.813	4.546
Fermented	0.321	2.259	1.108
Fermented odour	0.221	0.253	1.227
Bitter	4.608	6.065	0.494
Coarse	1.501	1.565	2.605
No taste	1.247	2.451	0.120
Yellow	4.115	48.273	2.520
White	25.775	5.164	9.568
Burnt	7.403	4.357	3.002
Brown	12.518	17.447	0.808
Attractive	15.675	0.991	4.081
Heavy	0.046	2.053	1.241
Fine	3.053	0.598	4.497
Dry	0.027	0.939	1.803

Squared cosines reflect the **representation quality** of a variable on a PCA axis. Squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines of a variable associated to an axis is low, the position of the variable on this axis should not be interpreted.

Squared cosines of the variables

	F1	F2	F3
Fibres	0.336	0.128	**0.536**
Little sour	0.483	**0.511**	0.005
Sweet	**0.805**	0.074	0.120
Too acidic	**0.507**	0.470	0.023
Good taste	**0.998**	0.000	0.002
Beautiful	**0.914**	0.069	0.017
Fermented	0.235	**0.749**	0.016
Fermented odour	**0.615**	0.318	0.068
Bitter	**0.626**	0.373	0.001
Coarse	**0.664**	0.313	0.023
No taste	**0.529**	0.470	0.001
Yellow	0.158	**0.840**	0.002
White	**0.911**	0.083	0.007
Burnt	**0.785**	0.209	0.006
Brown	**0.613**	0.386	0.001
Attractive	**0.967**	0.028	0.005
Heavy	0.046	**0.929**	0.025
Fine	**0.895**	0.079	0.026
Dry	0.055	**0.871**	0.073
Mean overall liking	**0.969**	0.010	0.021

Values in bold correspond for each variable to the factor for which the squared cosine is the largest

Factor scores: Activate to display the coordinates of the observations (factor scores) in the new space created by PCA.

Factor scores

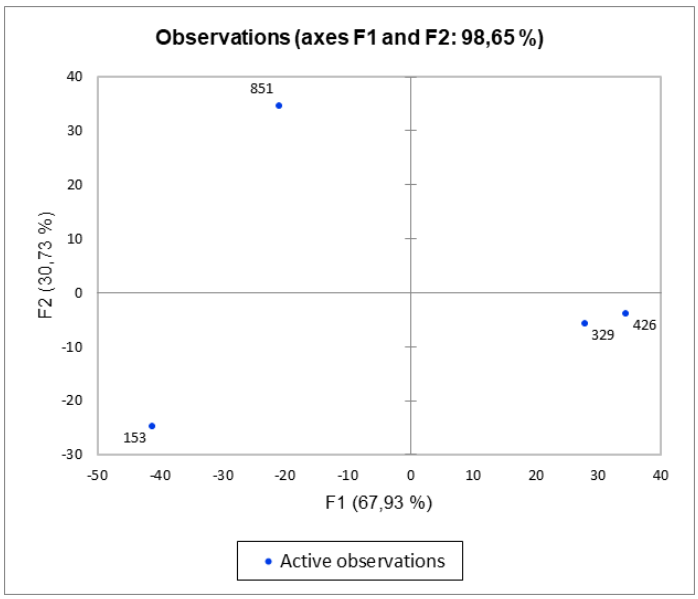
Observations	F1	F2	F3
426	34.438	-3.925	-6.064
851	-20.988	34.535	0.010
153	-41.261	-24.775	-0.589
329	27.811	-5.835	6.642

Factor scores are the observations coordinates on the PCA dimensions. They are displayed in a XLSTAT table.

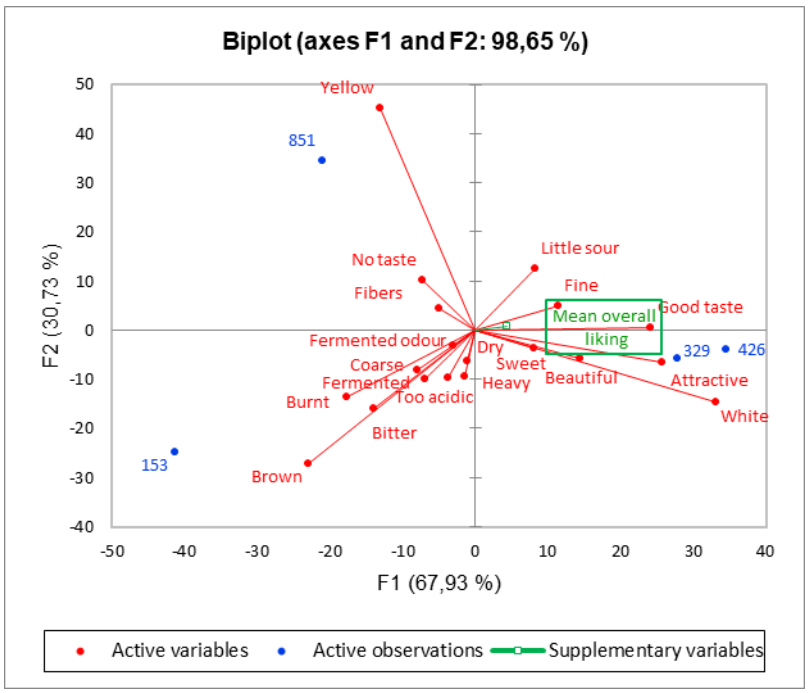
As for the results related to variables, XLSTAT displays contributions of observations (i.e. their contribution in building the PCA axes) as well as squared cosines (i.e. their representation quality on the different axes).

How to interpret results related to observations in PCA

The **observations charts** represent the observations in the PCA space. It enables you to look at the observations on a two-dimensional map, and to identify trends.



It is also possible to display **biplots**, which are simultaneous representations of variables and observations in the PCA space. Here as well the supplementary variable (here Mean overall liking) can be plotted in the form of vectors.



Contributions: The table below shows the contributions of the observations in building the principal components.

Squared cosines: The table below displays the squared cosines between the observation vectors and the factor axes.

Contribution of the observations (%)

	F1	F2	F3
426	28.909	0.830	45.261
851	10.737	64.263	0.000
153	41.500	33.073	0.427
329	18.854	1.834	54.312

Squared cosines of the observations

	F1	F2	F3
426	**0.958**	0.012	0.030
851	0.270	**0.730**	0.000
153	**0.735**	0.265	0.000
329	**0.908**	0.040	0.052

Values in bold correspond for each observation to the factor for which the squared cosine is the largest.

Institute: Cirad – UMR QualiSud

Address: C/O Cathy Méjean, TA-B95/15 - 73 rue Jean-François Breton - 34398 Montpellier Cedex 5 - France

Contact Tel: +33 4 67 61 44 31

Email: rtbfoodspmu@cirad.fr

Website: https://rtbfoods.cirad.fr/