Introduction to exploratory data analysis

Illustrated with XLSTAT

Jean Paul Maalouf
webinar@xlstat.com

www.xlstat.com
October 26, 2017

PLAN

- **XLSTAT**: who are we?
- Statistics: **categories**
- Reminder: **Variables**, **individuals**, **Descriptive Statistics**
- Toward exploratory data analysis: scatter plot colored by group
- **Exploratory statistics** & Data Mining
- **Principal Component Analysis (PCA)**: concept and practice
- **Agglomerative Hierarchical Clustering (AHC)**: concept and practice

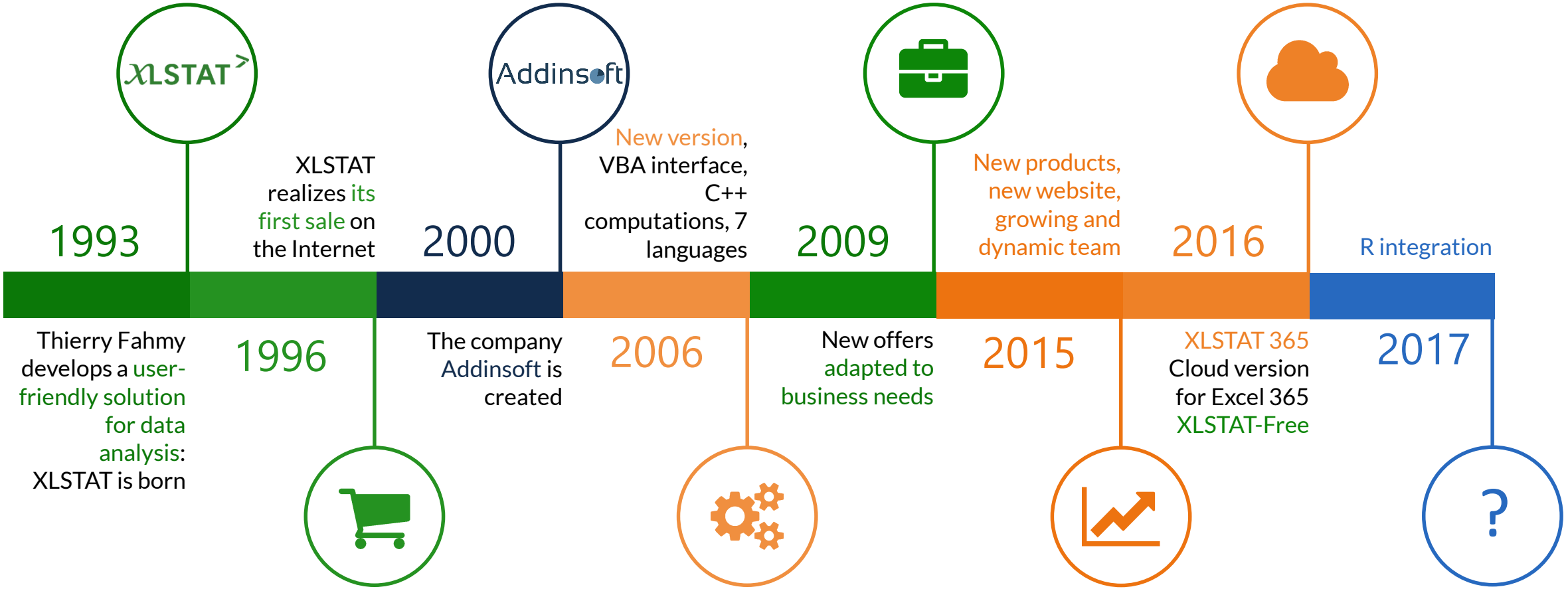
All the data in this webinar were made up unless otherwise specified

XLSTAT: Who are we?

XLSTAT is a user-friendly statistical add-on software for Microsoft Excel®

XLSTAT

A growing software and team



XLSTAT in a few numbers

200+ statistical features

General or field-oriented solutions

100k users

Across the world. Companies, education, research

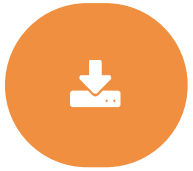
22 employees

Always receptive to the needs of users

220k visits/month on the website

Easy tutorials available in 5 languages

7 languages



10k downloads/month

Statistics: 4 categories

Statistics: 4 categories

Description

I want to **summarize** data using **simple statistics** or **charts** (mean, standard deviation, boxplots...)

Exploration

I want to easily extract **information** from a **large data set without** necessarily **having a precise question** to answer. (PCA, AHC...)

Tests

I want to **accept / reject** a very precise **hypothesis** assuming error risks. (t-tests, ANOVA, correlation tests, chi-square...)

Modeling

I want to understand the way a phenomenon evolves according to a set of parameters. (regression, ANOVA, ANCOVA...)

Reminder: variables, individuals, descriptive statistics

Variables, individuals...

Variable

An element that can take different values

Qualitative variable

A variable that cannot be quantified. *Examples: socioprofessional category, geographical origin, type of licence, blood type...* The possible values it can take are called categories or modalities

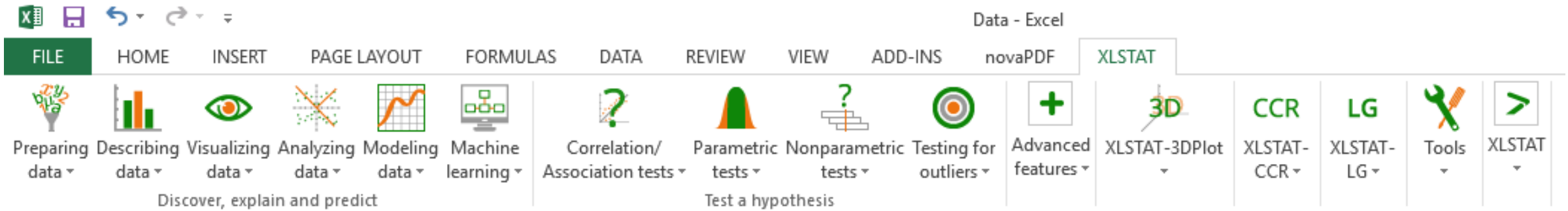
Quantitative variable

A variable that can be quantified. *Examples: invoice amount, number of likes on Facebook, sugar concentration, height...*

Individual

Elementary statistical unit. Can be described with variables. *Examples: customers, surveyed people, patients, laboratory mice...*

Data set: online shoe selling platform



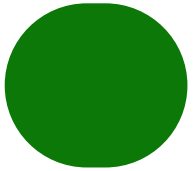
Variables

Individuals

	A	B	C	D	E	F	G	H	I
1	**Client**	**Origin**	**Gender**	**Height**	**Shoe size**	**Weight**	**Time spent on site**	**Preferred brand**	**Invoice amount**
2	Willard	Mars	M	9	26	50	28	Brand A	153
3	Derrick	Pluto	M	102	40	335	39	Brand D	229
4	Olivia	Mars	F	13	31	68	29	Brand A	118
5	Marc	Mars	M	20	29	81	40	Brand C	160
6	Alice	Pluto	F	116	33	378	11	Brand B	248
7	Elijah	Mars	M	22	30	88	44	Brand C	137
8	Clyde	Mars	M	27	29	105	42	Brand C	142
9	Angel	Mars	F	31	28	116	51	Brand A	151
10	Randy	Earth	M	4	29	36	18	Brand A	174

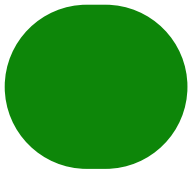
Descriptive statistics

Commonly used tools according to the situation



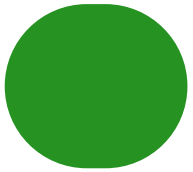
1 qual. variable

Flat sorting, mode, pie charts



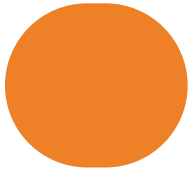
1 qual. variable x 1 qual. variable

Cross tabulation (contingency table)



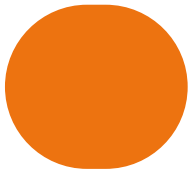
1 quant. variable x 1 qual. variable

Quantitative descriptive statistics per category of the qualitative variable; multiple box plot chart



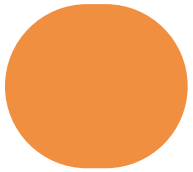
1 quant. variable

Center (mean / median) ; dispersion (variance / std. deviation / quartiles) ; box plot



1 quant. variable x 1 quant. variable

Scatter plot

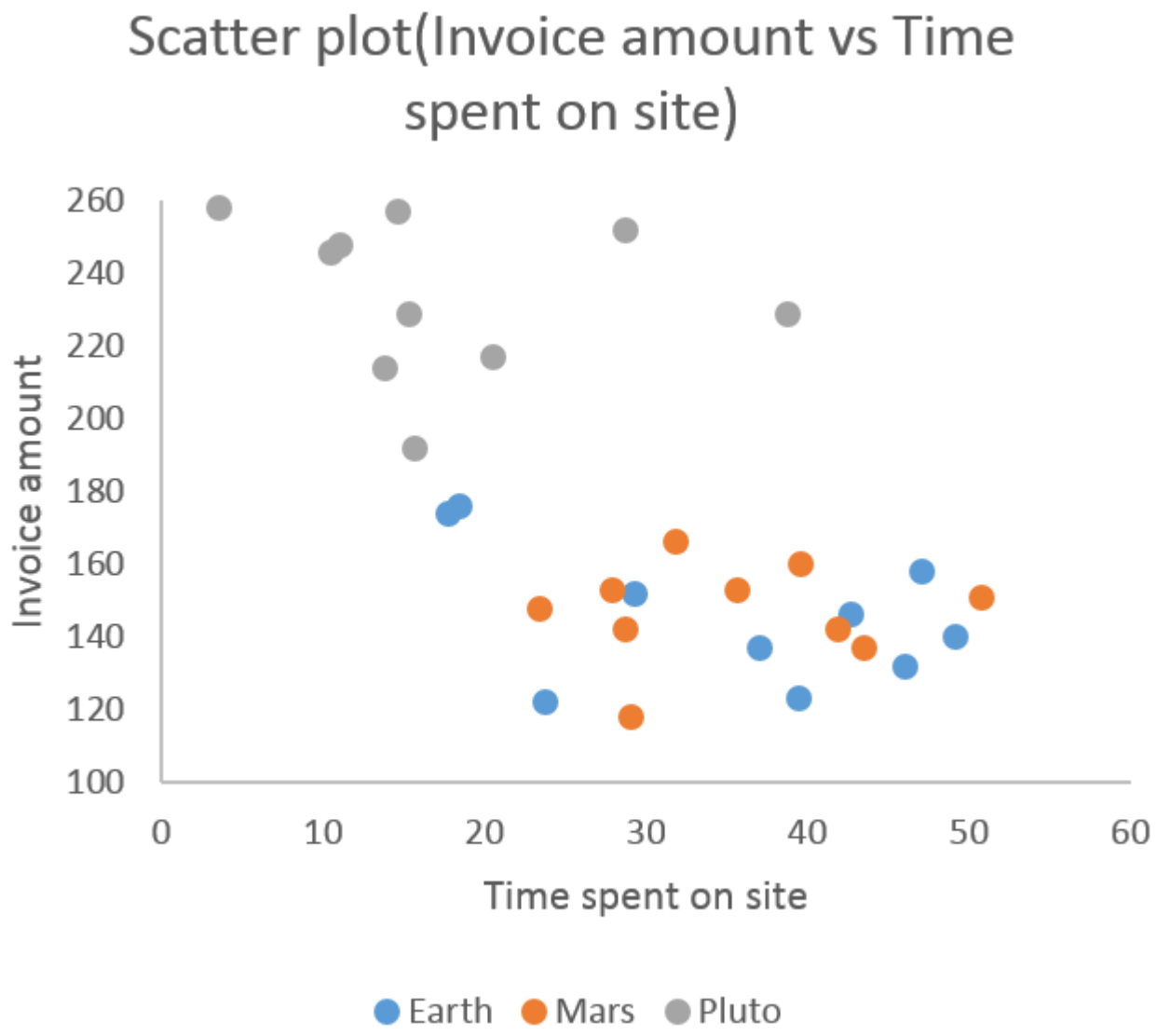


1 quant. variable x 1 quant. variable x 1 qual. variable

Scatter plot with points colored according to the categories of the qualitative variable

Toward exploratory data analysis: scatter plot colored by group

Toward exploratory data analysis: scatter plot colored by group



- Invoice amount decreases with time spent on the website.
- Plutonians spend more money on the website compared to others.
- Martians and humans form a relatively homogeneous group
- ...

Imagine having the same kind of reasoning on a higher number of variables...

Time for Exploratory statistics (or Exploratory Data Analysis)

Example: Principal Component Analysis (PCA)

We want to analyze multiple variables (dimensions) at a time the same way we did with the 2D scatter plot.

	A	B	C	D	E	F	G	H	I
1	**Client**	**Origin**	**Gender**	**Height**	**Shoe size**	**Weight**	**Time spent on site**	**Preferred brand**	**Invoice amount**
2	Willard	Mars	M	9	26	50	28	Brand A	153
3	Derrick	Pluto	M	102	40	335	39	Brand D	229
4	Olivia	Mars	F	13	31	68	29	Brand A	118
5	Marc	Mars	M	20	29	81	40	Brand C	160
6	Alice	Pluto	F	116	33	378	11	Brand B	248
7	Elijah	Mars	M	22	30	88	44	Brand C	137
8	Clyde	Mars	M	27	29	105	42	Brand C	142
9	Angel	Mars	F	31	28	116	51	Brand A	151
10	Randy	Earth	M	4	29	36	18	Brand A	174
11	Robin	Earth	M	5	33	40	40	Brand C	123
12	Elias	Pluto	M	107	32	340	15	Brand C	257
13	Stanley	Mars	M	37	31	133	23	Brand A	148
14	Max	Earth	M	6	26	43	24	Brand C	122
15	Dewey	Mars	M	33	28	123	32	Brand B	166
16	Henry	Earth	M	4	34	39	49	Brand B	140

Exploratory data analysis

I want to easily extract **information** from a **large data set without** necessarily **having a precise question** to answer.

Exploratory data analysis: a few words

Exploratory statistics

Look for information in a multi-variables data set, without having very precise expectations. Exploratory tools are part of Data Mining.

First thing you can do: concentrate the information of big data sets in a few dimensions

Examples: Principal Component Analysis, Correspondence Analysis…

Second thing you can do: classification (= clustering = segmentation)

Examples: Agglomerative Hierarchical Clustering, k-means…

We'll be able to investigate:

- Relationships among variables
- Proximity among individuals
- How individuals relate to variables

Principal Component Analysis (PCA)

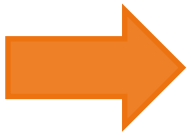
I'd like to summarize a big data set in a few simple charts

PCA: concept

Initial dataset

C	D	E	F
Height	**Shoe size**	**Weight**	**Time spent on sit**
9	26	50	28
102	40	335	39
13	31	68	29
20	29	81	40
116	33	378	11
22	30	88	44
27	29	105	42
31	28	116	51
4	29	36	18
5	33	40	40
107	32	340	15

Amount of information



Artificial data set synthesized by PCA

The information is re-distributed in a way to concentrate most of it on a few dimensions.

F1	F2	F3	F4
-0.799	-1.279	-0.147	0.005
0.980	3.351	0.212	0.004
-0.845	0.046	-0.592	0.023
-1.182	-0.135	0.639	-0.016
2.689	0.981	-0.414	0.020
-1.180	0.217	0.522	-0.008
-0.935	-0.040	0.620	0.012
-1.170	0.065	1.295	0.010
-0.604	-0.829	-1.187	-0.010
-1.556	0.678	-0.297	0.001
2.251	0.850	-0.296	-0.026

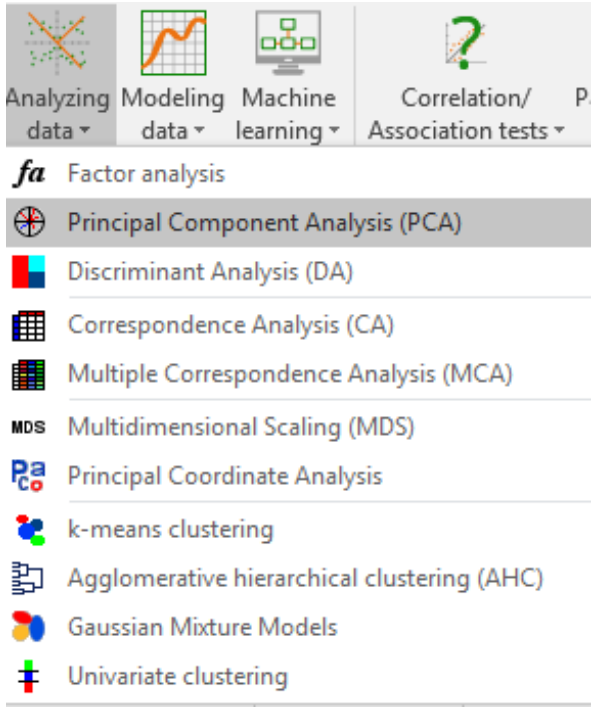
PCA jargon:

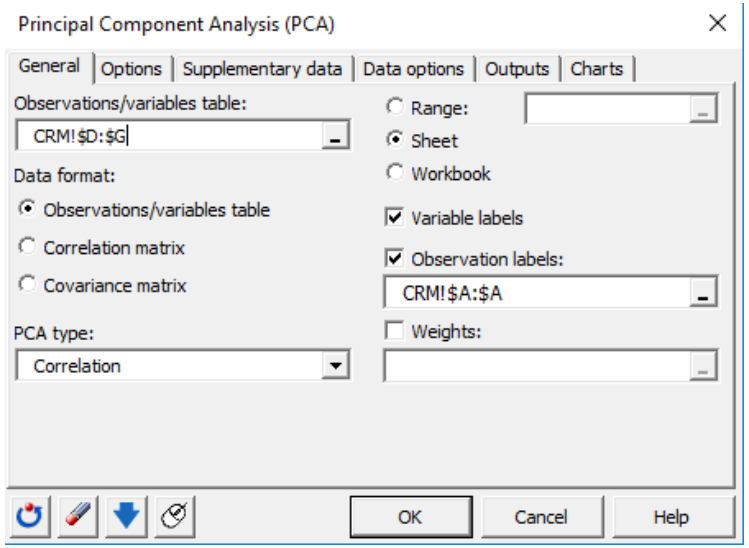
dimension
= axis
= factor

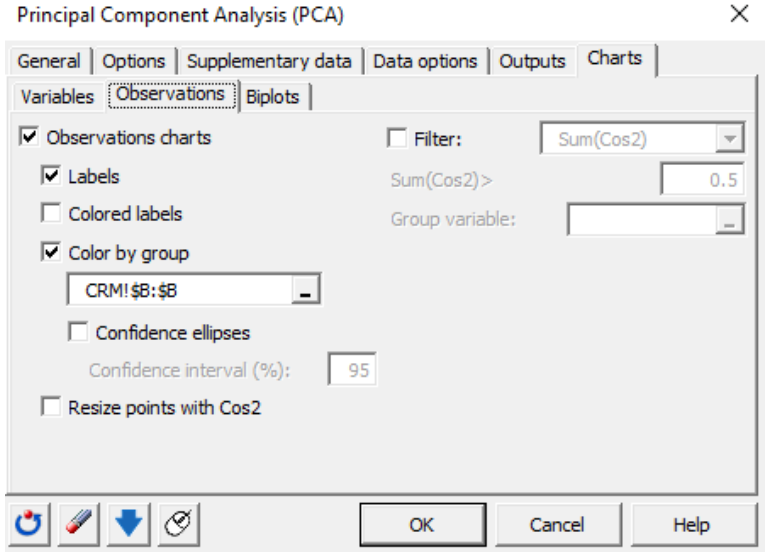
information
= variability
= inertia

Setting up a PCA in XLSTAT

Chart 1: correlation circle



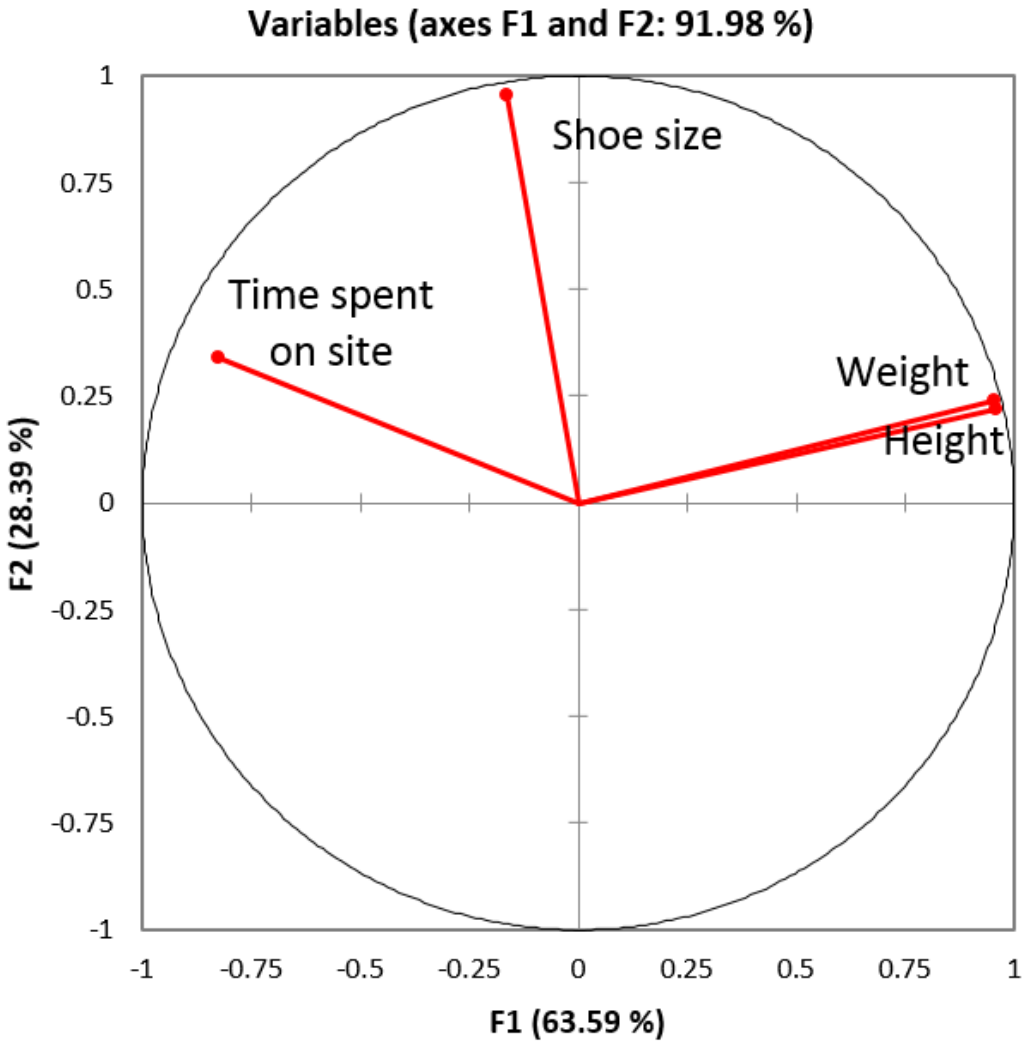




PCA tutorial link

How PCA looks like in reality

Chart 1: correlation circle

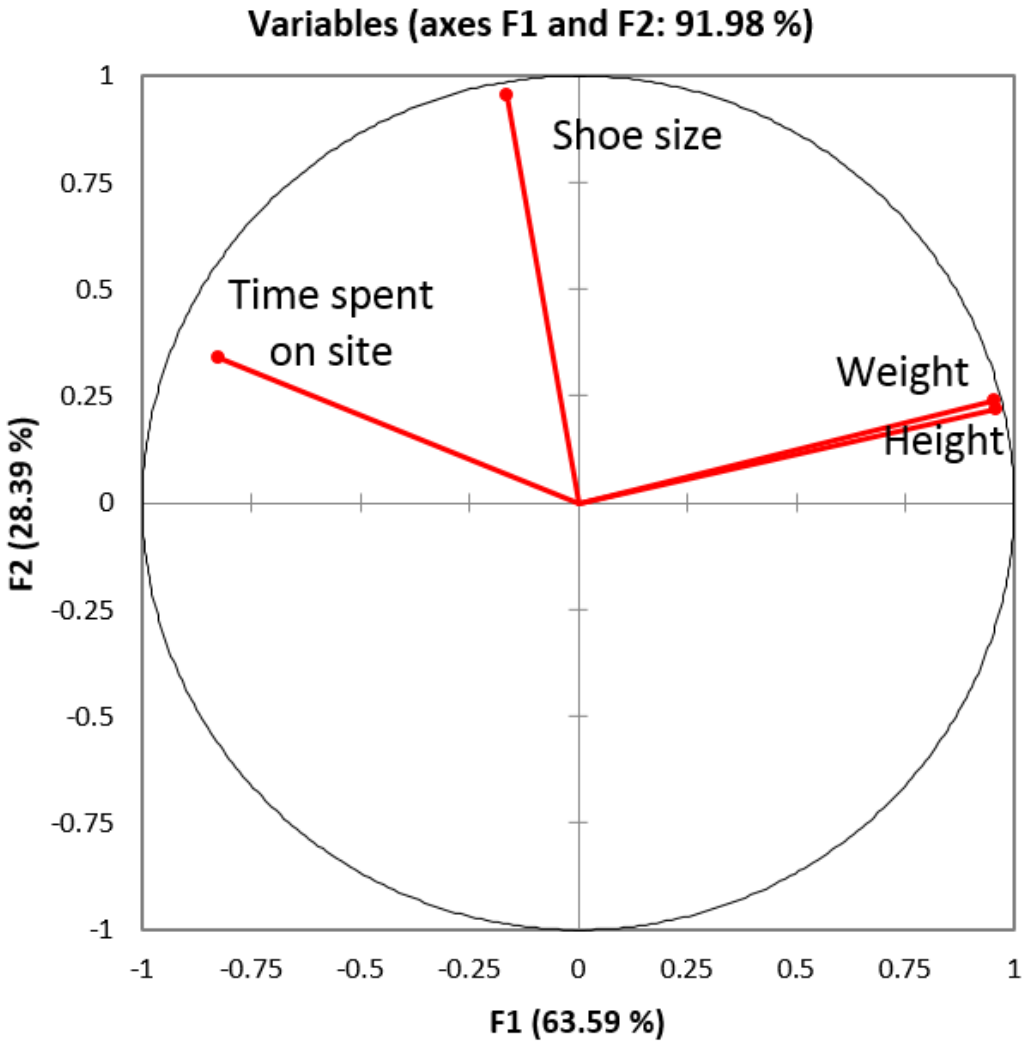


- Acute angle: positively-linked variables (e.g. weight & height)
- Right angle: uncorrelated variables (e.g. height & shoe size)
- Obtuse angle: negatively-linked variables (e.g. weight & time spent on site)

Vector length reflects representativeness in the selected plan (F1/F2 here)

Interpreting the axes

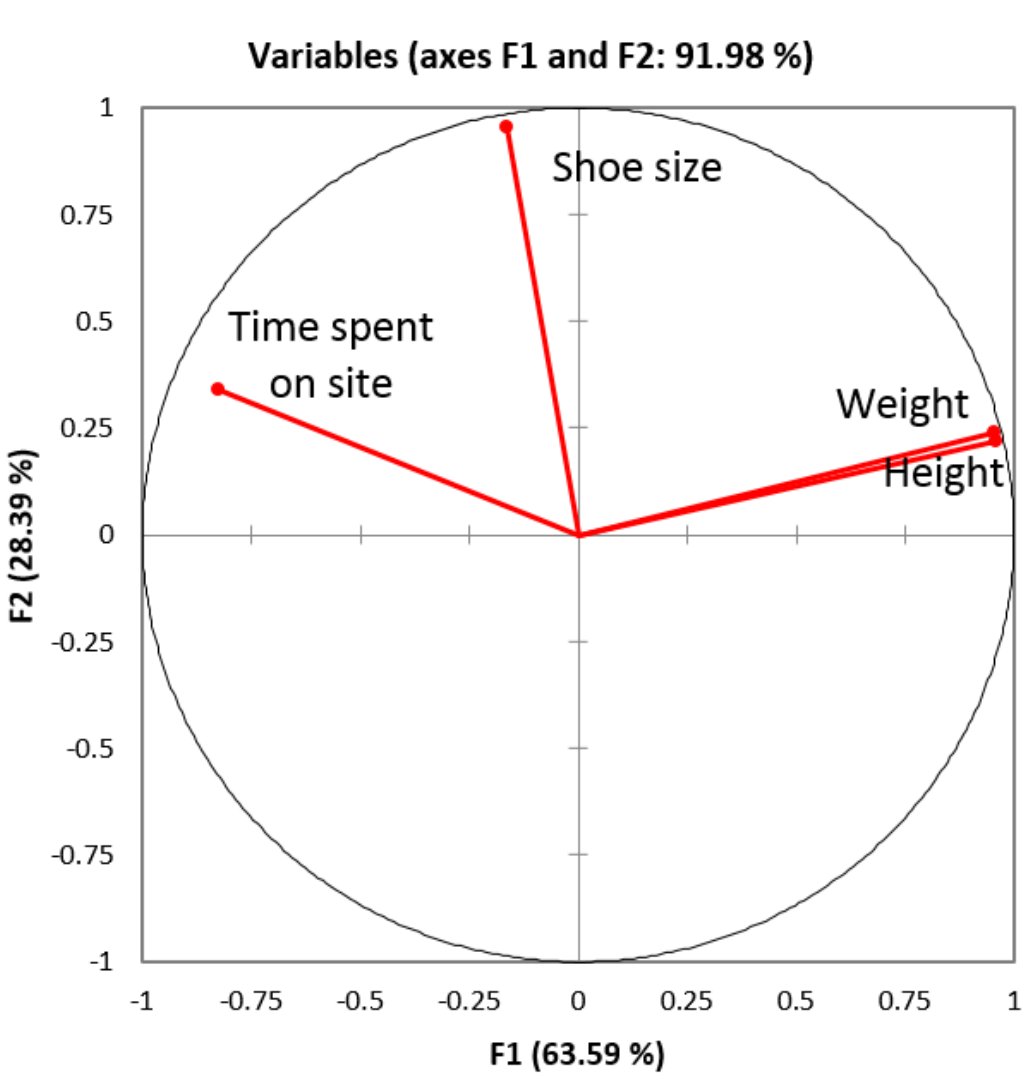
Chart 1: correlation circle

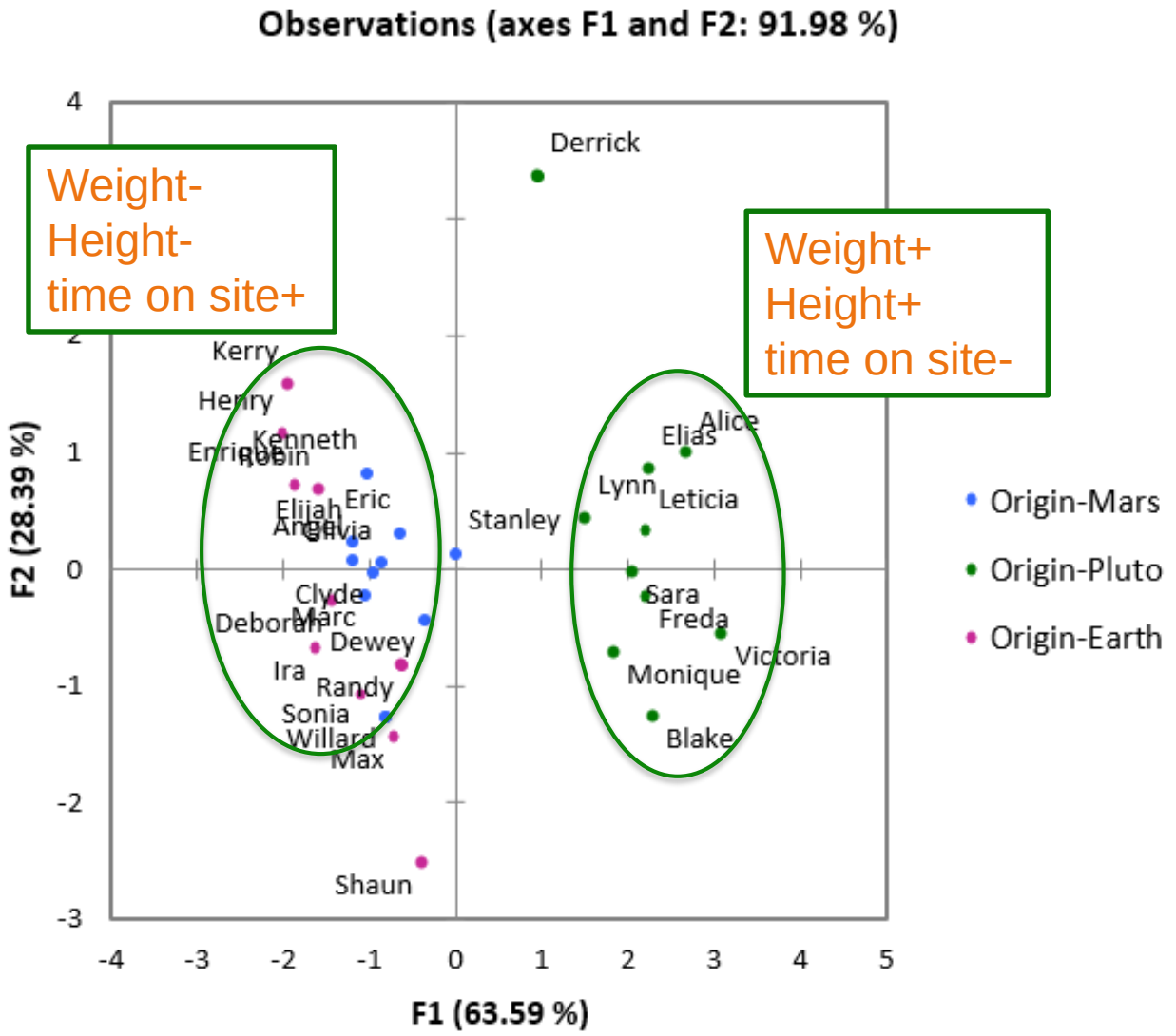


- F1 reflects:
  - High weight & height (right)
  - Long time spent on site (left)
- F2 is strongly related to shoe size:
  - Big shoes (top)
  - Small shoes (bottom)

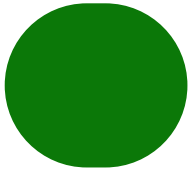
How PCA looks like in reality

Chart 1: correlation circle ; chart 2: observations

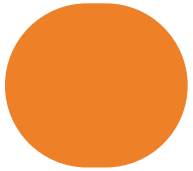




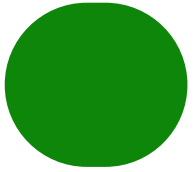
PCA: explorations ...



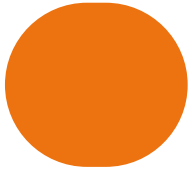
Weight increases with height



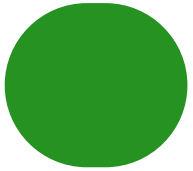
Shoe size is unrelated to weight / height



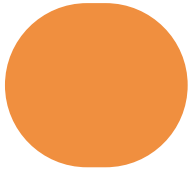
Time spent on site decreases with weight & height



Derrick has big feet. Shaun has small feet.



Looks like there are two clusters in the data

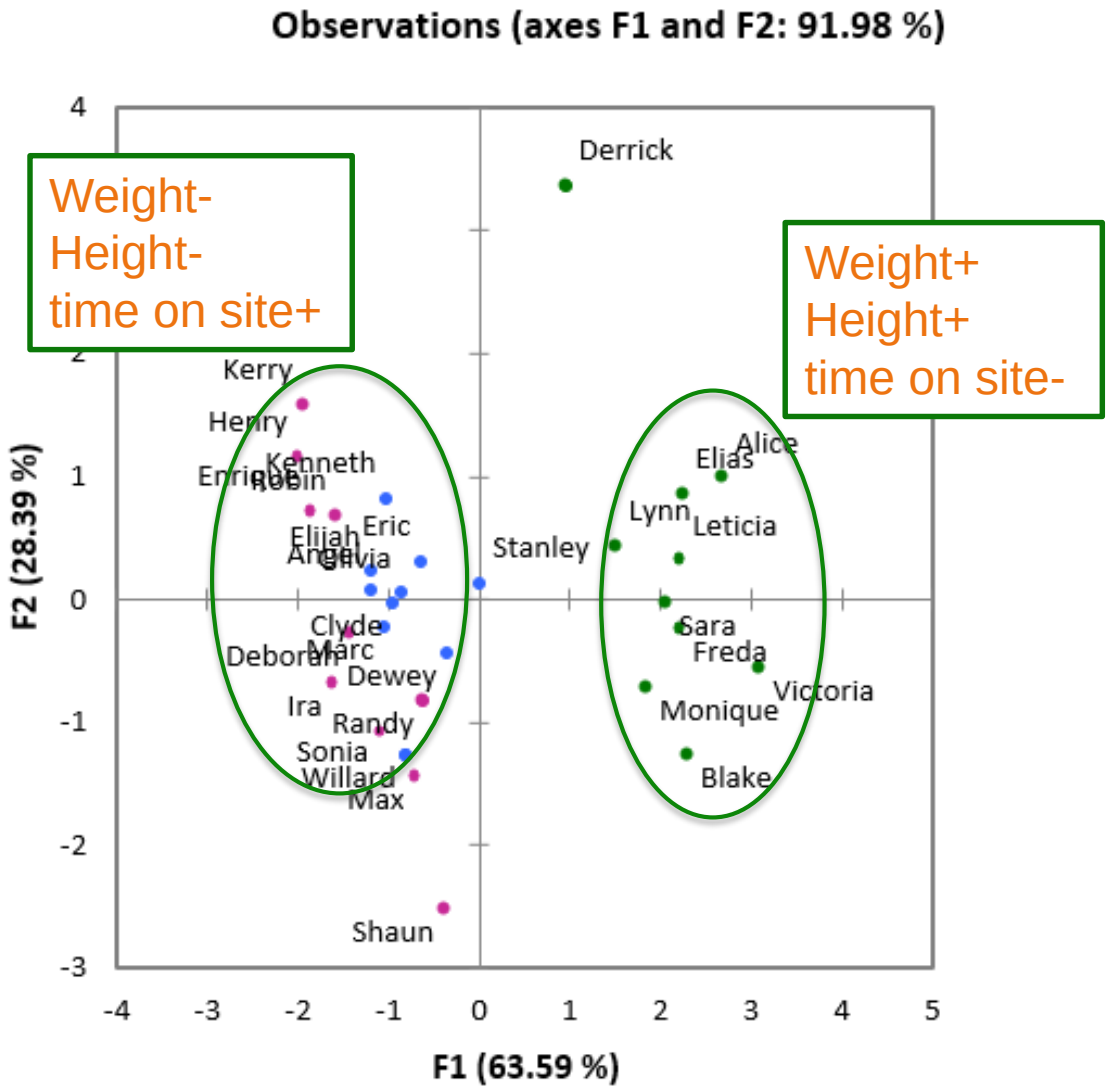


And so on...

PCA tutorial link

PCA works only with quantitative data. Click here to check out other exploratory methods.

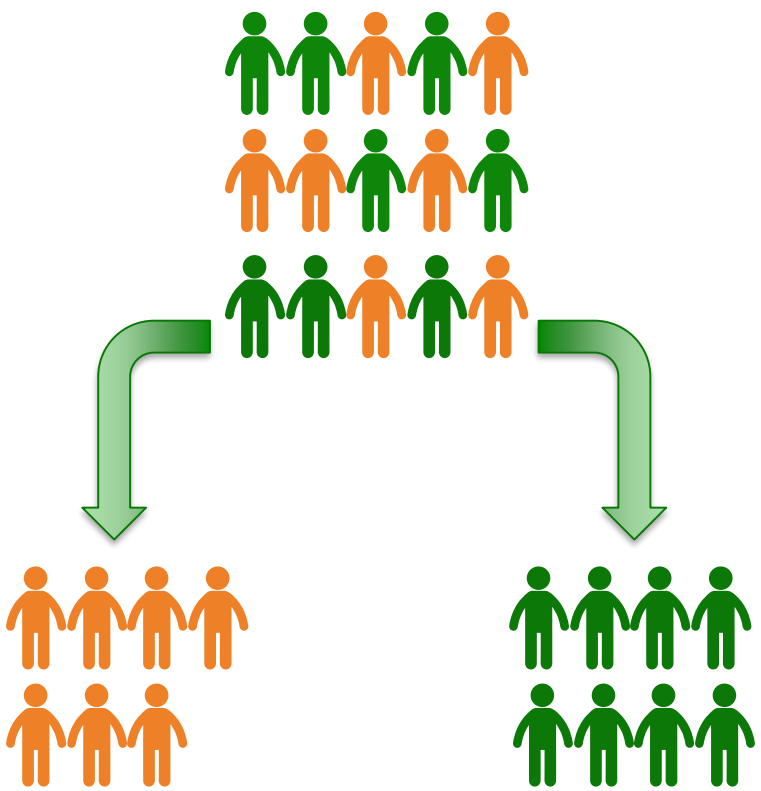
It was easy to detect two clusters of customers. Nice for marketing!



According to our PCA, customers can be split into two clusters characterized by height, weight and time spent on site.

This may help us define tailored marketing campaigns.

But what if groups were not that easy to define visually?



Agglomerative Hierarchical Clustering (AHC)

I want to **cluster** (= **classify** = **segment**) individuals in homogeneous **groups** (= **segments** = **clusters** = **classes**)

Agglomerative Hierarchical Clustering (AHC)

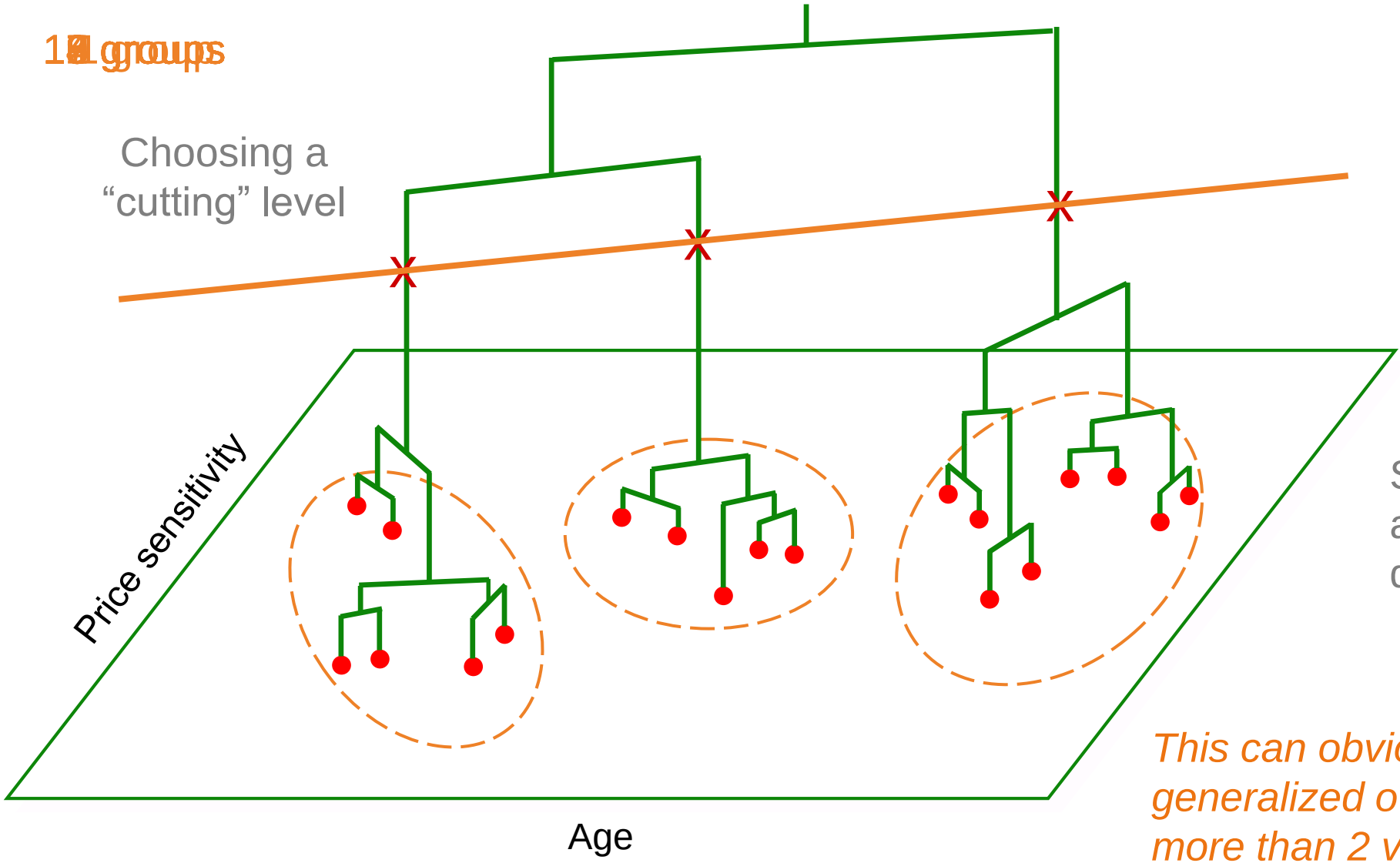
EXAMPLE: sensory analysis, chocolate consumers survey

Illustration with 2 variables

	A	B	C	D	E	F	G	H
1	**ID**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**	**Age**
2	Bobby	7	4	7	6	5	9	38
3	Muriel	5	5	5	4	6	5	28
4	Shelia	7	5	9	6	4	1	46
5	Juana	5	6	3	4	4	3	15
6	Tami	2	6	7	4	6	7	34
7	Frank	4	5	4	6	4	9	50
8	Sam	1	7	1	4	6	10	15
9	Marsha	4	7	5	3	7	10	15
10	Dominic	8	5	4	6	4	3	46
11	Kevin	5	3	3	9	0	0	59
12	Melinda	4	6	1	2	6	7	25
13	Candice	5	6	2	4	5	2	31
14	Sherri	2	5	4	3	6	0	37
15	Jordan	1	6	1	2	5	2	21

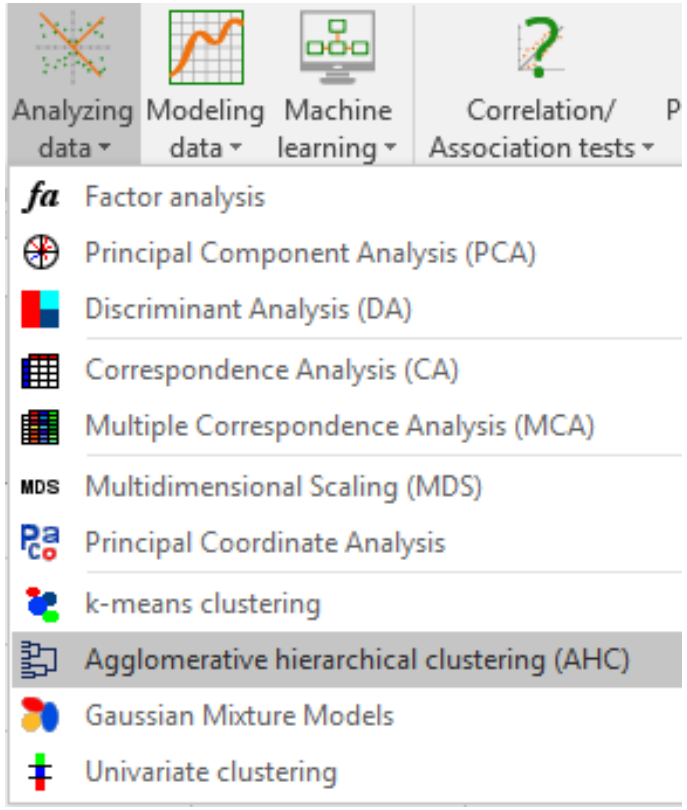
How to cluster consumers into different groups?

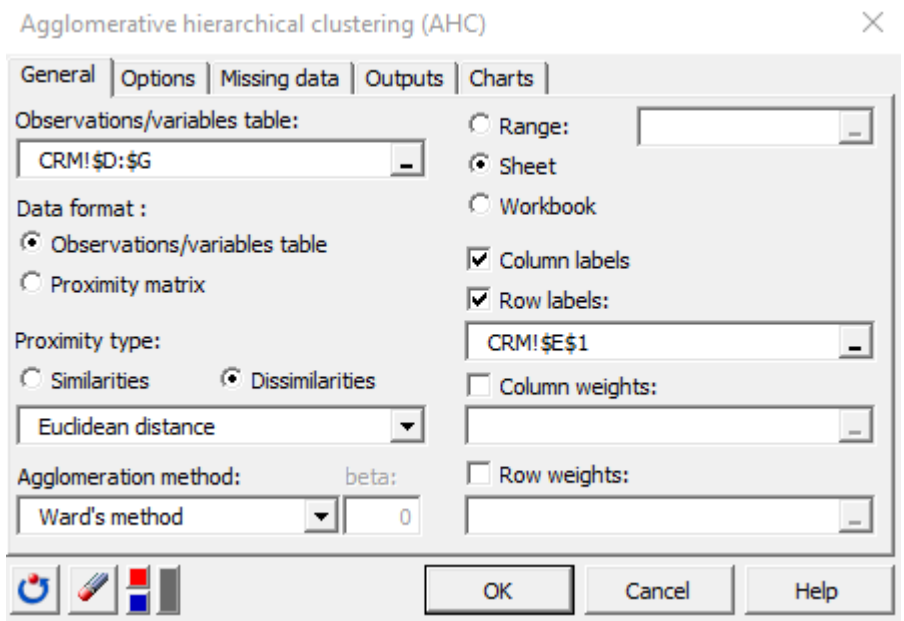
AHC – how it works on 2 variables



Agglomerative Hierarchical Clustering (AHC)

Setting things up in XLSTAT

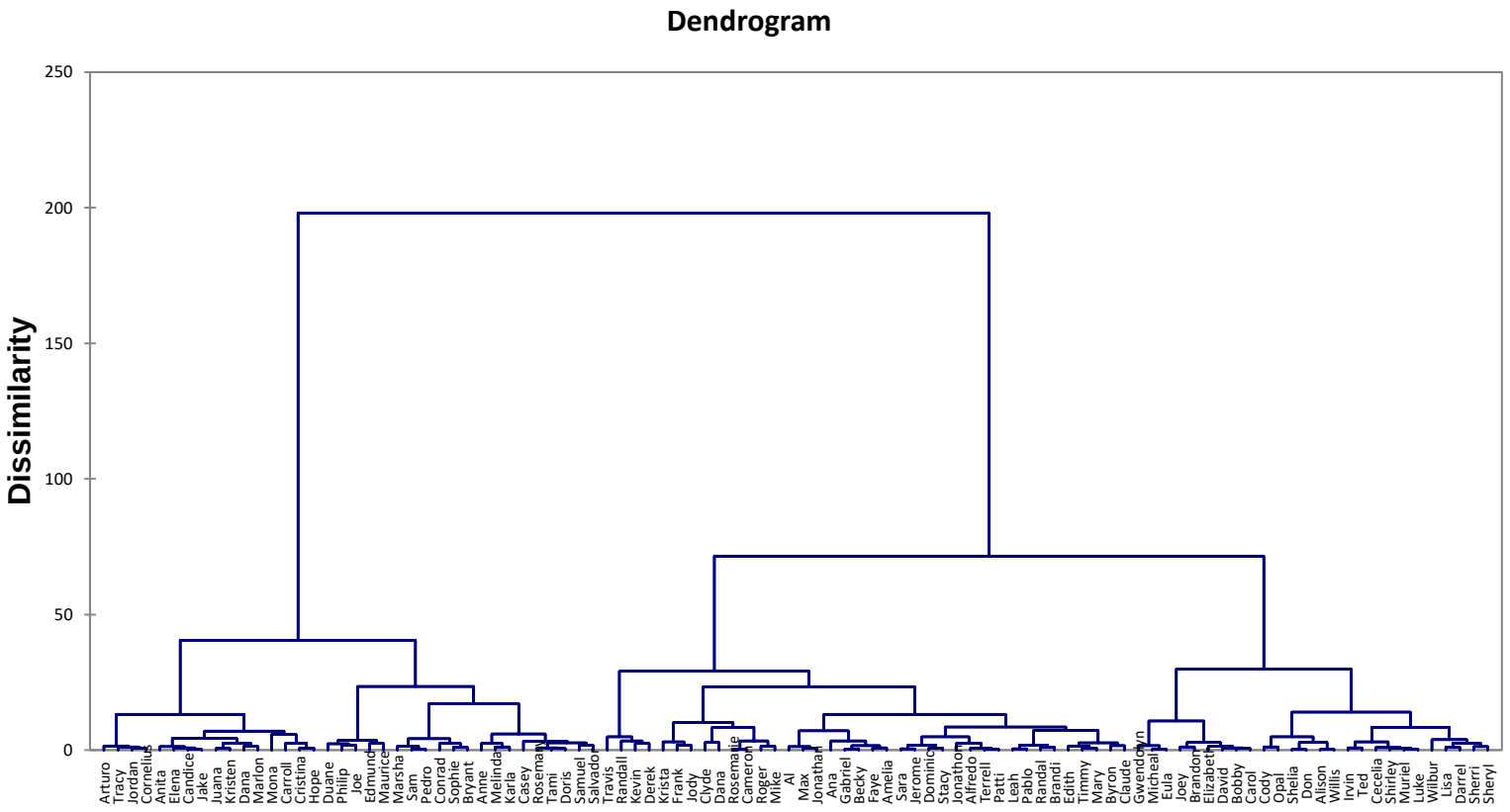




AHC tutorial link

Agglomerative Hierarchical Clustering (AHC)

What it looks like in XLSTAT:

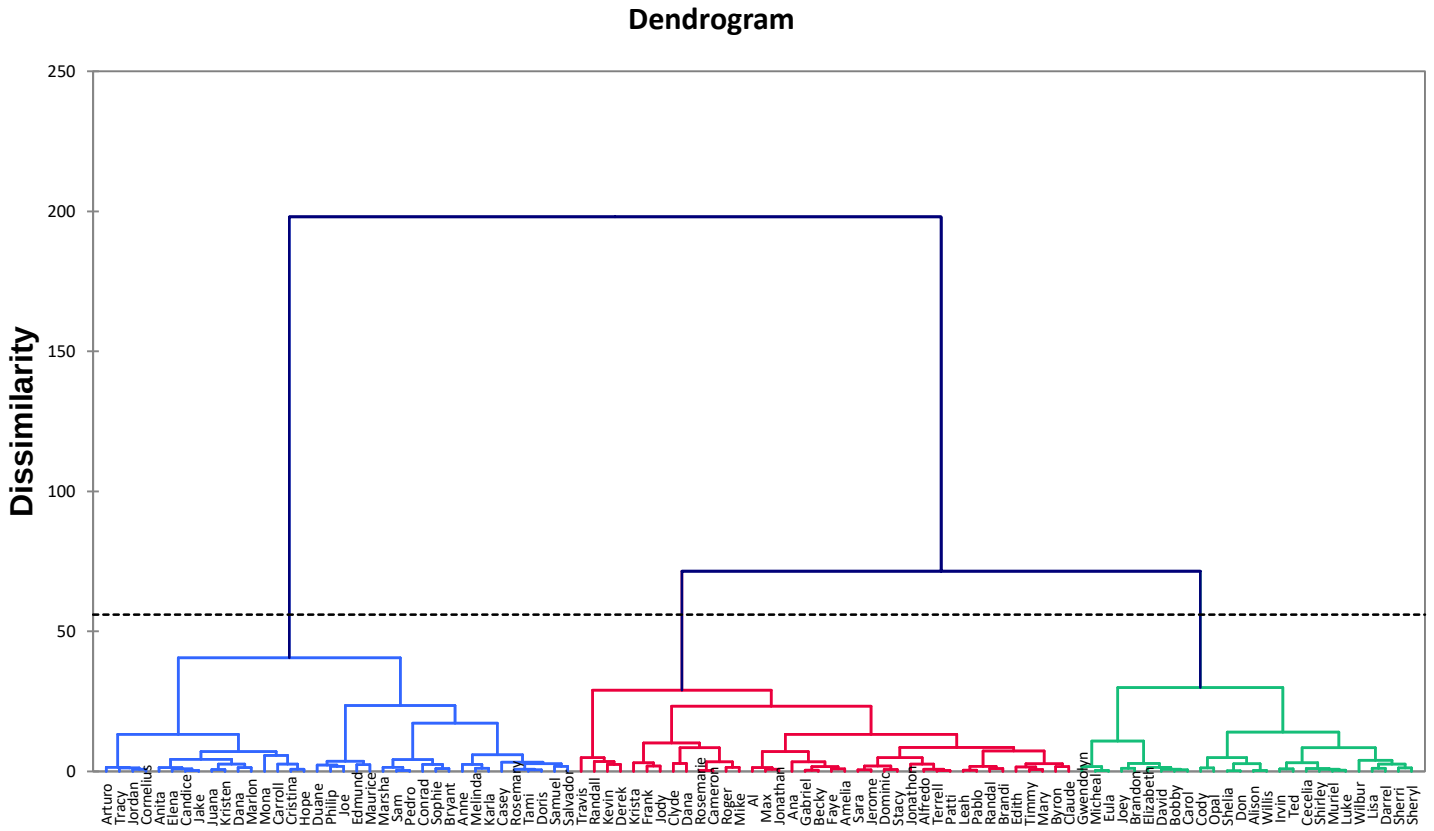


The higher the “vertical distance” between two individuals (or groups), the more different the individuals.

Here we could split the individuals into 3 or 4 homogeneous groups

Agglomerative Hierarchical Clustering (AHC)

3-cluster split:



Okay. And now what?

Let's describe the 3 groups to see how we could take action on a marketing scale

Observation	Class
Bobby	1
Muriel	1
Shelia	1
Juana	2
Tami	2
Frank	3
Sam	2
Marsha	2
Dominic	3
Kevin	3
Melinda	2
Candice	2
Sherri	1
Jordan	2
Anita	2
Alfredo	3

How can I describe segments?

Things become quite straightforward when you extract class membership in the AHC results

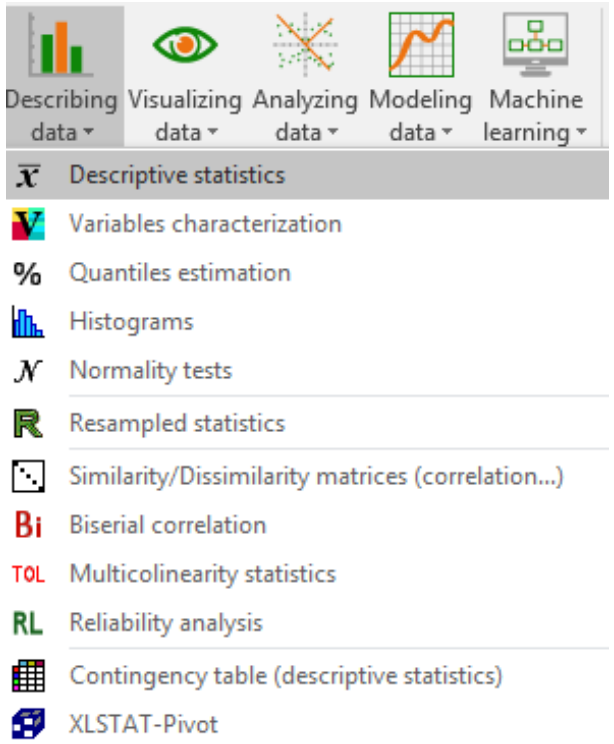
Describing the segments

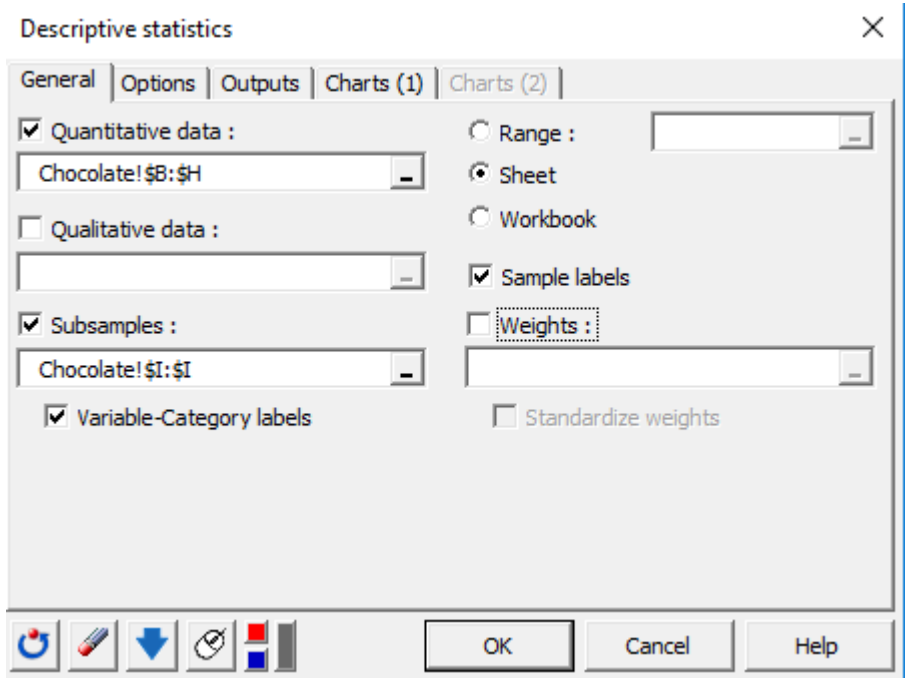
Things you can do

- Split individuals into classes and run descriptive statistics on each segment
- Use Class membership as a supplementary variable in a PCA
- Use Parallel Coordinates Plots

	A	B	C	D	E	F	G	H	I
1	**ID**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**	**Age**	Class
2	Bobby	7	4	7	6	5	9	38	1
3	Muriel	5	5	5	4	6	5	28	1
4	Shelia	7	5	9	6	4	1	46	1
5	Juana	5	6	3	4	4	3	15	2
6	Tami	2	6	7	4	6	7	34	2
7	Frank	4	5	4	6	4	9	50	3
8	Sam	1	7	1	4	6	10	15	2
9	Marsha	4	7	5	3	7	10	15	2
10	Dominic	8	5	4	6	4	3	46	3
11	Kevin	5	3	3	9	0	0	59	3
12	Melinda	4	6	1	2	6	7	25	2
13	Candice	5	6	2	4	5	2	31	2

Describing clusters: descriptive statistics





Tutorial link

Describing clusters: descriptive statistics

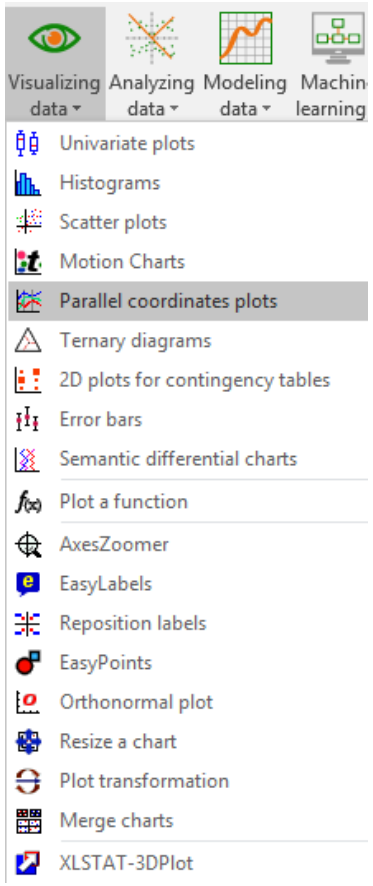
Descriptive statistics (Quantitative data) :

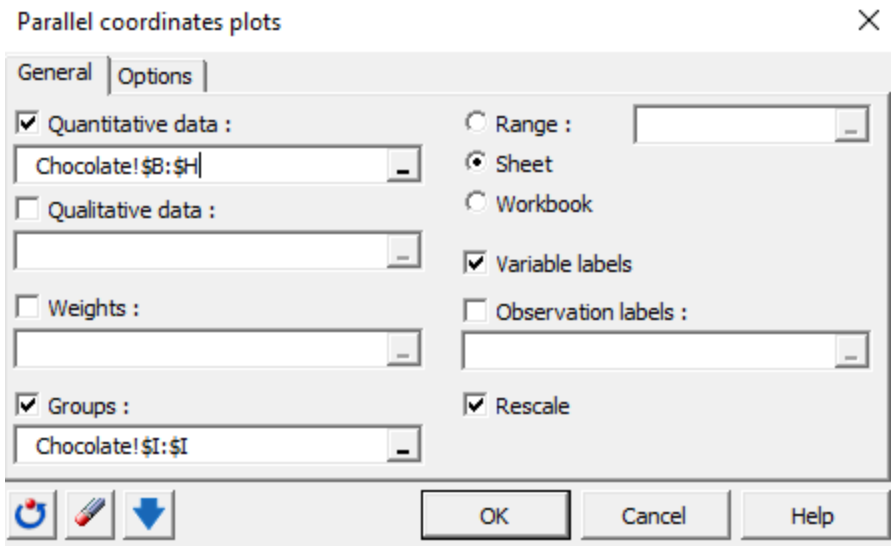
Statistic	Median	Mean	Standard deviation (n-1)
Brand loyalty | 1	6,500	6,392	2,492
Brand loyalty | 2	3,300	3,397	2,712
Brand loyalty | 3	7,300	6,492	2,602
Price sensitivity | 1	4,700	4,577	1,409
Price sensitivity | 2	6,300	6,614	1,574
Price sensitivity | 3	3,450	3,550	1,574
Online buyer | 1	7,400	7,308	1,973
Online buyer | 2	3,200	3,814	2,683
Online buyer | 3	2,950	3,292	2,116
Bitter | 1	4,850	4,881	1,080
Bitter | 2	3,000	2,919	1,223
Bitter | 3	6,200	6,182	1,434
Frozen | 1	5,650	5,585	0,922
Frozen | 2	6,300	6,417	1,534
Frozen | 3	3,850	3,400	1,247
Crunchy | 1	3,600	4,000	3,173
Crunchy | 2	5,000	5,103	2,614
Crunchy | 3	4,650	4,892	2,654
Age | 1	38,000	37,308	7,822
Age | 2	21,500	22,278	8,413
Age | 3	46,000	45,711	9,603

Consumers from clusters 1 & 3 are more loyal to brands than those from cluster 2

Consumers from cluster 2 are younger

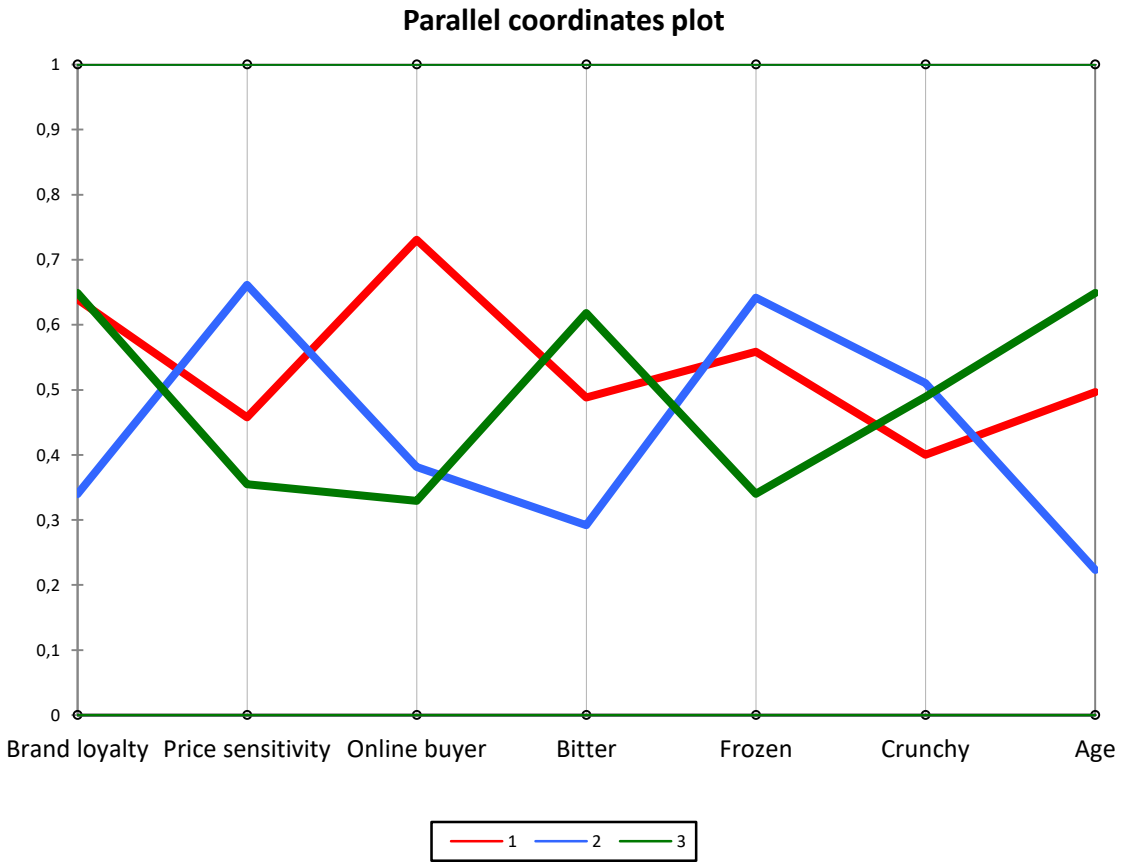
Describing clusters: parallel coordinates plot





Tutorial link

Describing clusters: parallel coordinates plot



Cluster 3: older consumers, loyal to brands, who prefer bitter chocolate and are not online buyers...

Cluster 2: younger consumers, prefer frozen chocolate, are sensitive to prices and care less about brands

…

Consequences :

- Promote branded bitter chocolate to older consumers
- Promote cheaper chocolates to younger consumers
- …

In summary...

Description

I want to **summarize small data sets** (1-3 variables) using **simple statistics** or **charts**. **Leads to hypotheses**.

Exploration

I want to easily extract **information** from a **large data set without** necessarily **having a precise question** to answer. **Leads to hypotheses**.

Tests

I want to **validate** / **reject** a very precise **hypothesis** assuming error risks. (t tests, ANOVA, correlation tests, chi-square...)

Modeling

I want to understand the way a phenomenon evolves according to a set of parameters. (regression, ANOVA, ANCOVA...)

Exploratory statistics: Take Home Message

Exploratory statistics

Allow to gain insight into large data sets

They give a synthetic view of large data sets

Examples: Principal Component Analysis, Correspondence Analysis, MDS…

They allow clustering data sets

Examples: Agglomerative Hierarchical Clustering, k-means

Click here to choose an appropriate exploratory data analysis tool according to your situation

Data exploration inspired us many hypotheses. Are they valid?

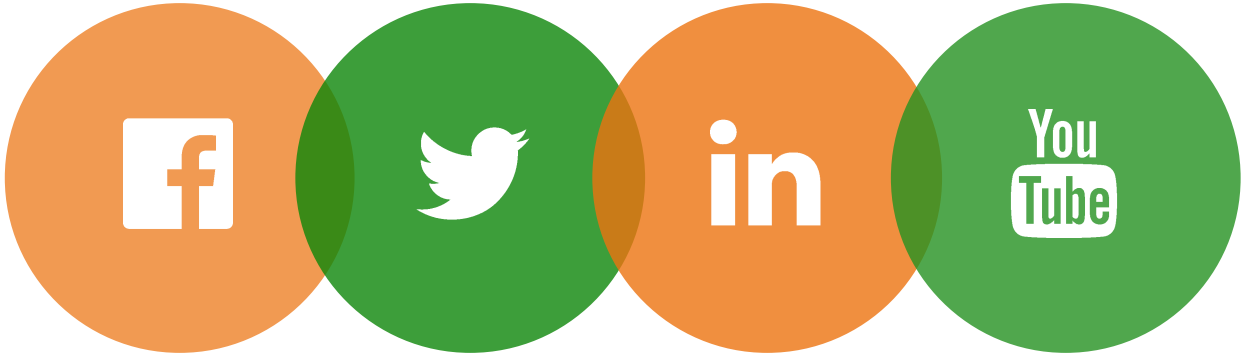
→ Statistical tests

See you on November 16!

Subscribe

Thanks for attending!

All the tools we saw are available in all XLSTAT solutions



Download 30-day Free Trial

Discover our products