A conceptual method to introduce multivariate thinking from a simple scatter plot

Jean Paul Maalouf, Efthalia Anagnostou

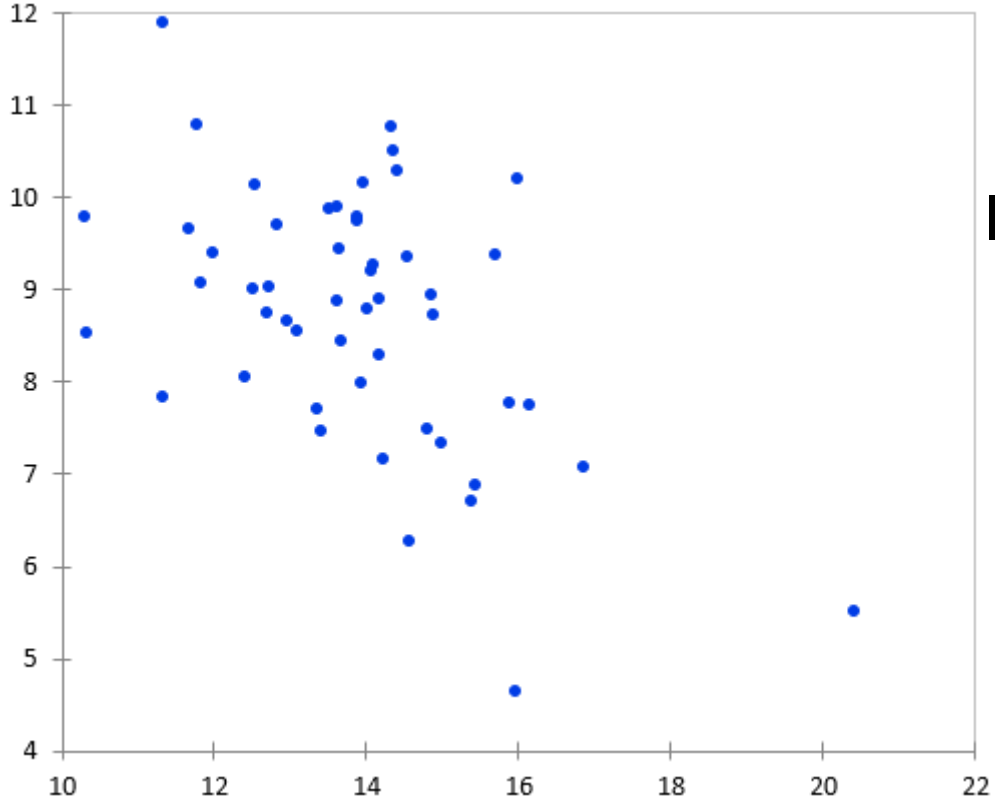
www.xlstat.com

November 18, 2017

Washington DC

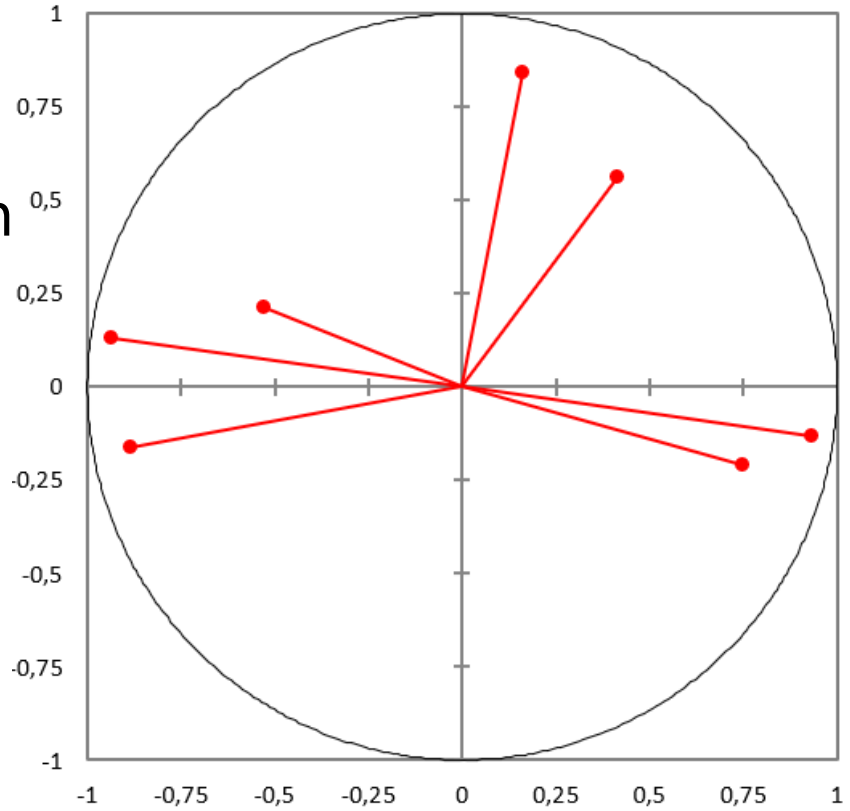
From bivariate to multivariate thinking

Two variables



Information

Many variables



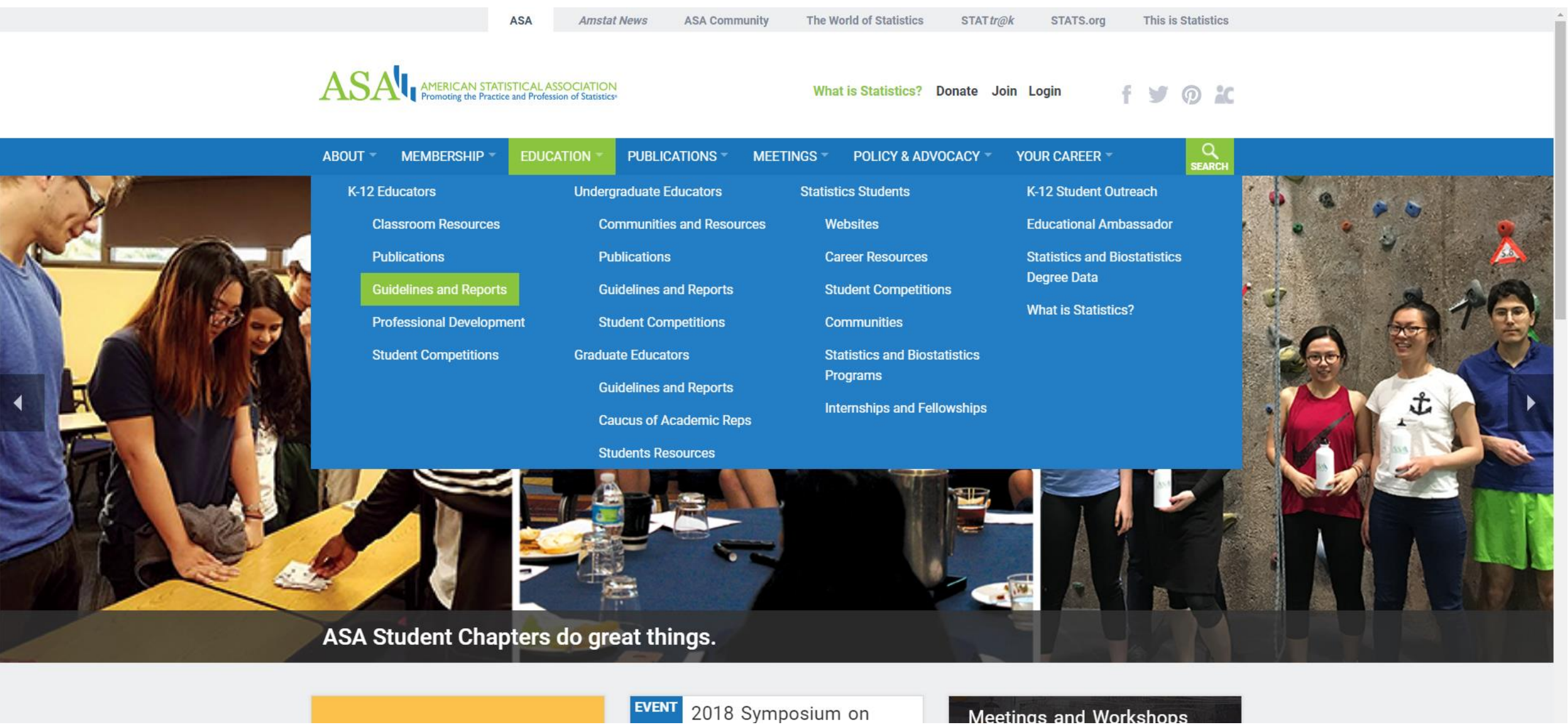
Respects the GAISE recommendations

Guidelines for Assessment and Instruction in Statistics education College Report (2016)

GAISE recommendations

- Teach statistical **thinking.**
- Focus on **conceptual** understanding.
- Integrate **real data** with a context and purpose.
- Foster **active** learning.
- Use **technology** to explore concepts and analyze data.
- Use **assessments** to improve and evaluate student learning.

Finding the report on the ASA website



Steps

1. Find a multivariate data set and highlight a **conceptual representation of information**
2. Generate a scatter plot out of the data
3. Switch to Multivariate thinking based on the **conceptual representation of information**

Amount of information

Gender	Brand loyalty	Price sensitivity	Online buyer
A	6	3	1
B	5	10	0
B	6	3	3
A	9	1	5
B	5	7	5
B	9	5	9
A	5	3	3
A	7	4	4
A	9	5	7
A	8	2	10

Step 1: Find a multivariate data set and highlight a conceptual representation of information density

Grab a question which will need real data to be answered

As a marketing manager for chocolate products, how can you characterize chocolate consumption behavior around you? [and calibrate your marketing campaigns accordingly]

Let students figure out a questionnaire

Rate your loyalty to renowned brands.

Rate your degree of appreciation of crunchiness in chocolates.

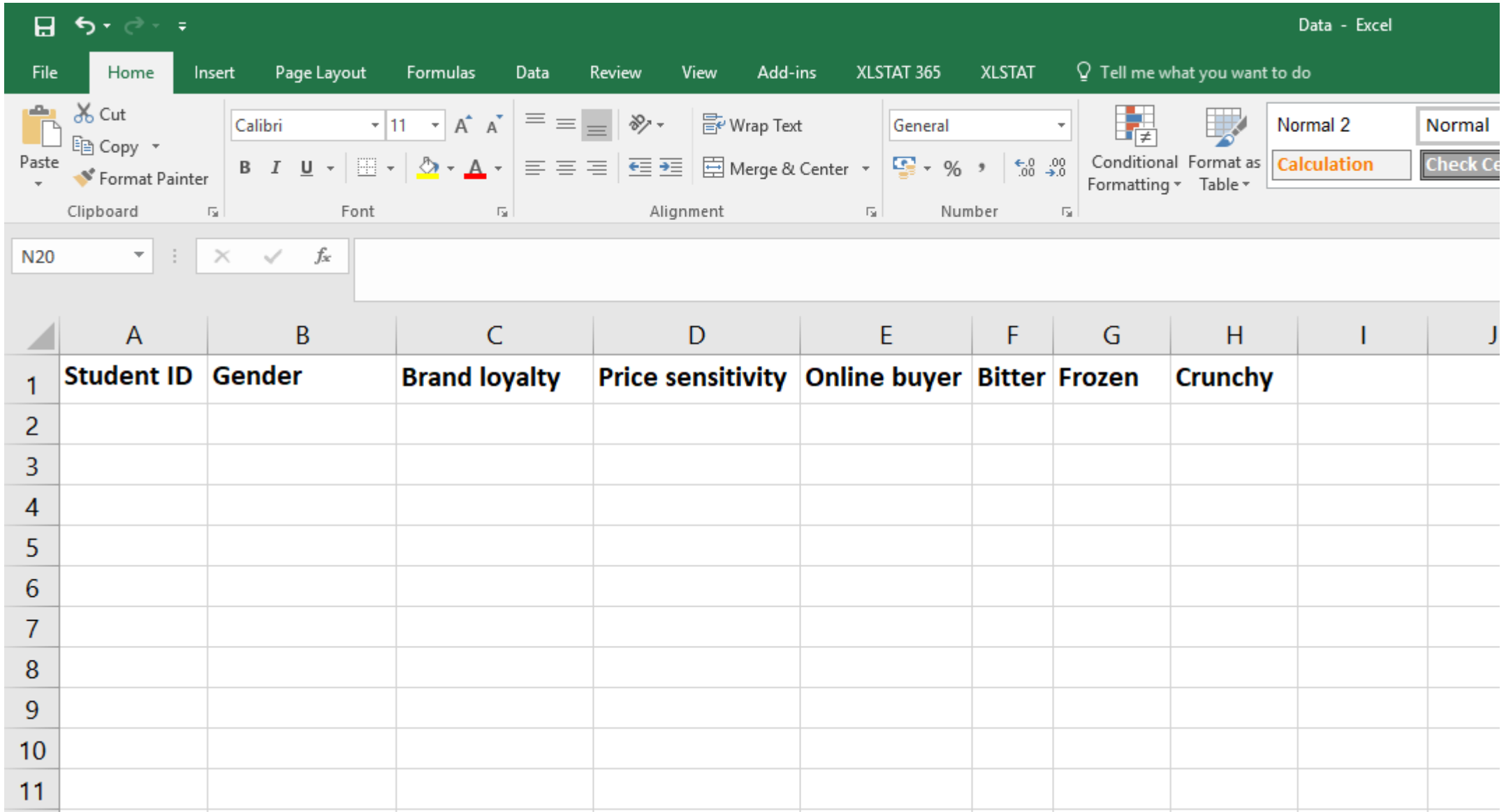
Rate your degree of appreciation of bitterness in chocolates.

How likely would you avoid buying a chocolate if it was too expensive?

Rate your degree of preference for frozen chocolates compared to room-temperature chocolates.

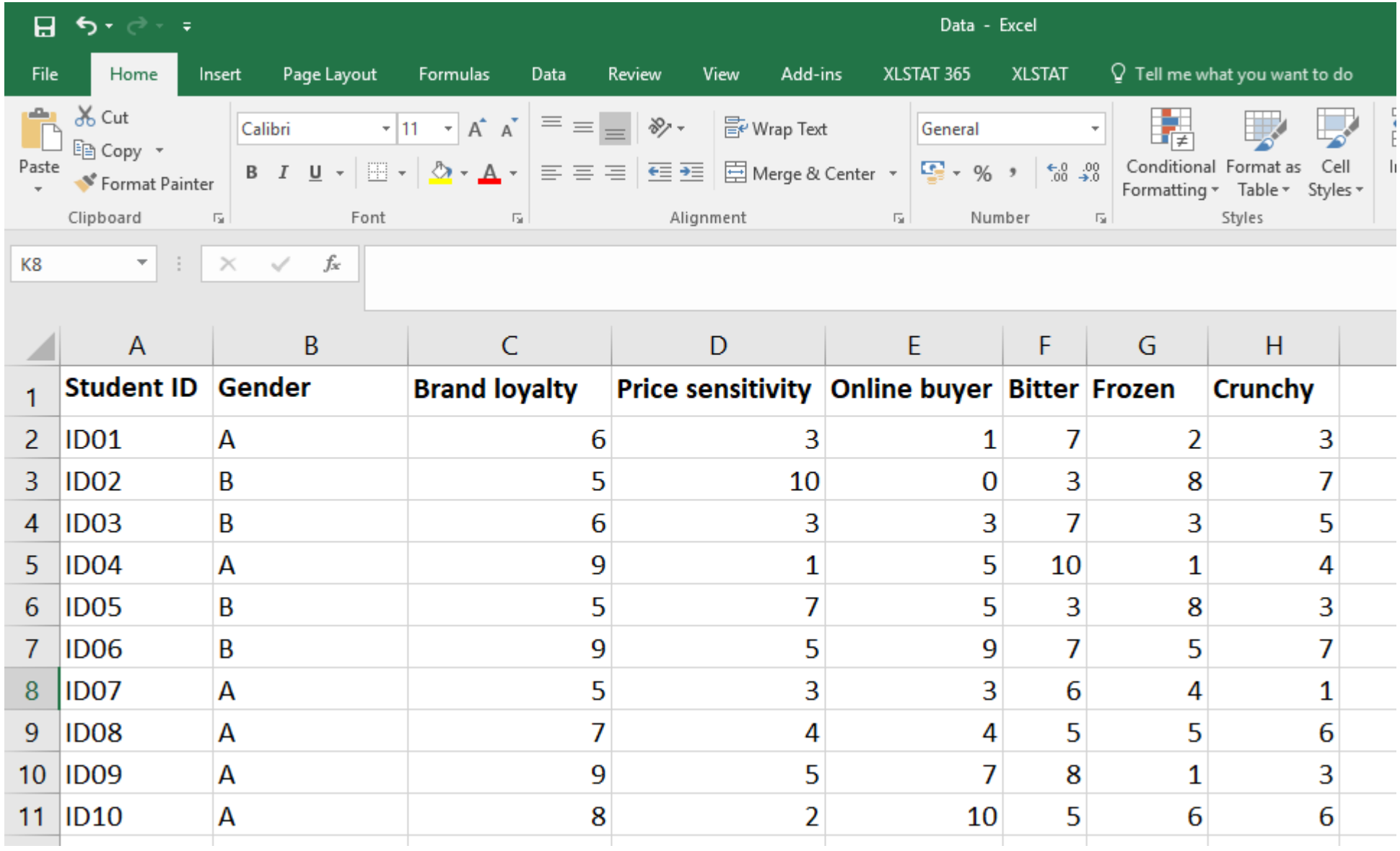
Rate your preference for online shopping compared to real shops.

Let them come up with an appropriate data collection spreadsheet



	A	B	C	D	E	F	G	H	I	J
1	**Student ID**	**Gender**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**		
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										

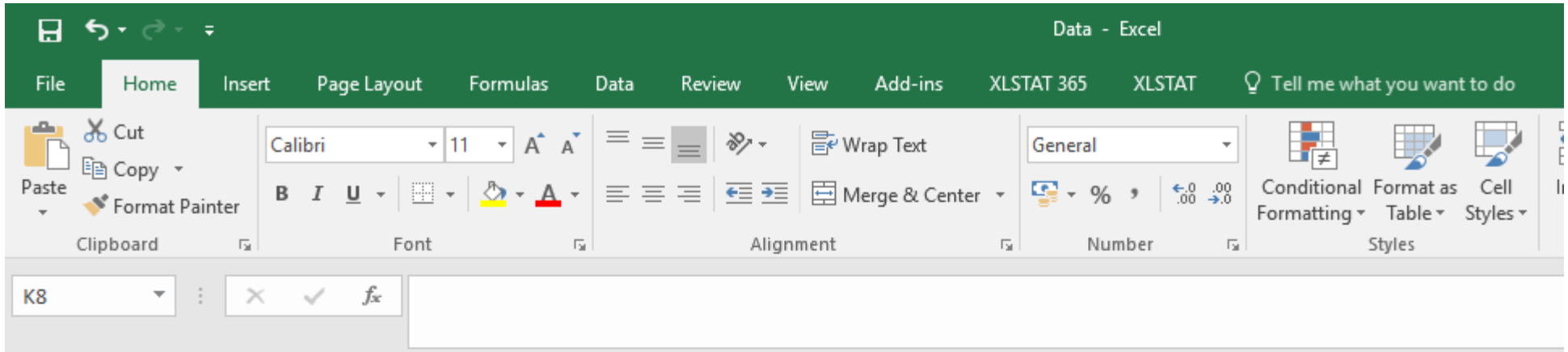
Make them run the survey & fill out the data



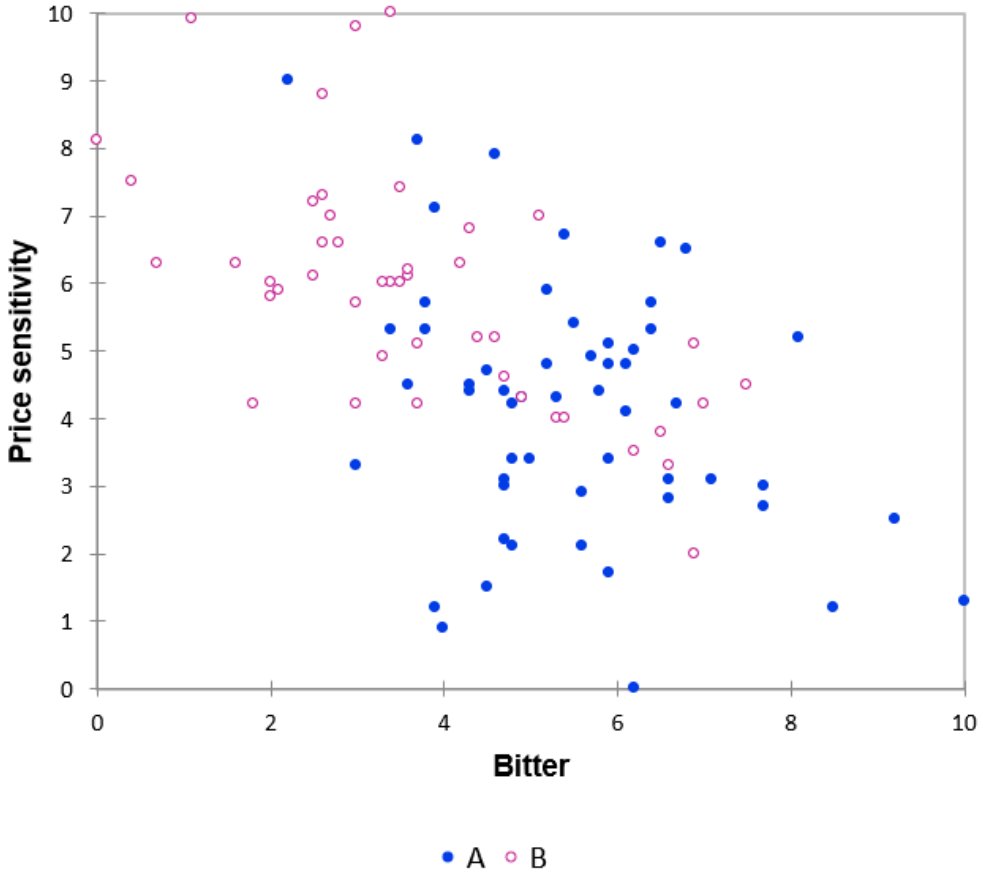
	A	B	C	D	E	F	G	H
1	**Student ID**	**Gender**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**
2	ID01	A	6	3	1	7	2	3
3	ID02	B	5	10	0	3	8	7
4	ID03	B	6	3	3	7	3	5
5	ID04	A	9	1	5	10	1	4
6	ID05	B	5	7	5	3	8	3
7	ID06	B	9	5	9	7	5	7
8	ID07	A	5	3	3	6	4	1
9	ID08	A	7	4	4	5	5	6
10	ID09	A	9	5	7	8	1	3
11	ID10	A	8	2	10	5	6	6

Stress on the idea that there is potentially interesting info everywhere in the data set

(Focus on conceptual understanding)



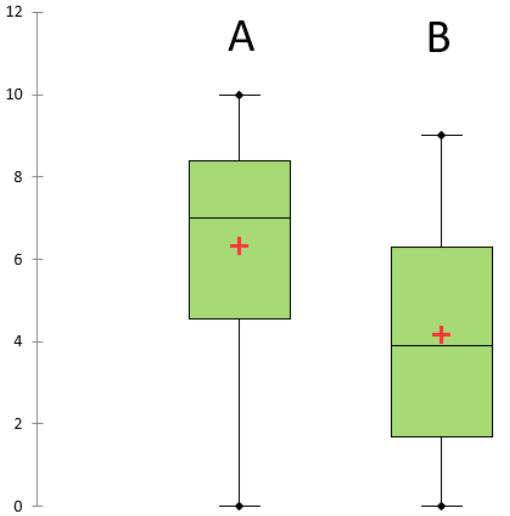
	A	B	C	D	E	F	G	H
1	**Student ID**	**Gender**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**
2	ID01	A	6	3	1	7	2	3
3	ID02	B	5	10	0	3	8	7
4	ID03	B	6	3	3	7	3	5
5	ID04	A	9	1	5	10	1	4
6	ID05	B	5	7	5	3	8	3
7	ID06	B	9	5	9	7	5	7
8	ID07	A	5	3	3	6	4	1
9	ID08	A	7	4	4	5	5	6
10	ID09	A	9	5	7	8	1	3
11	ID10	A	8	2	10	5	6	6

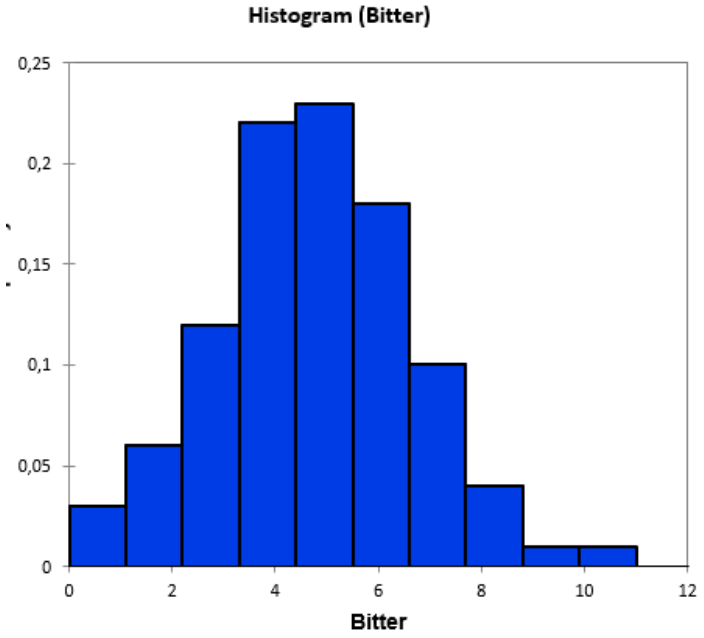


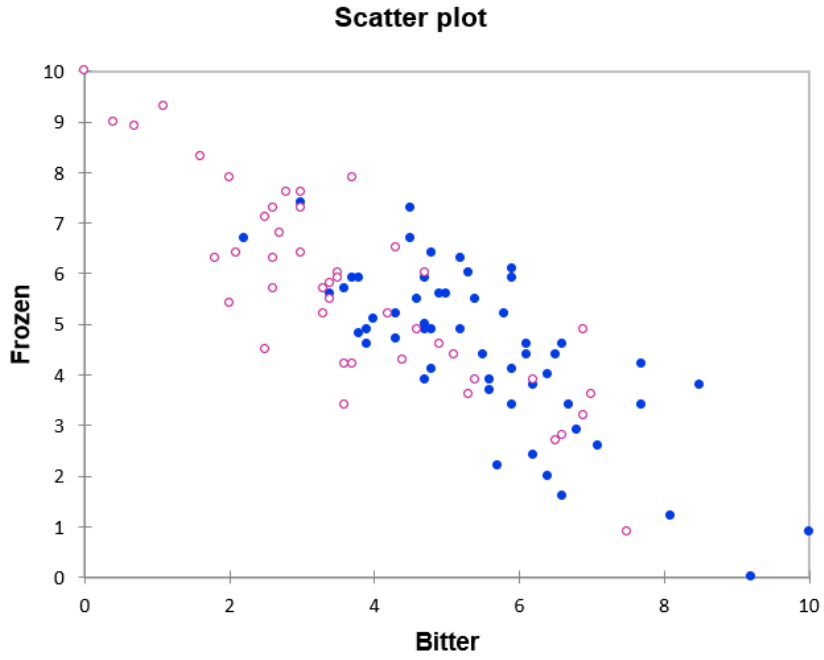
Step 2: Generate a scatter plot out of the data

The class already has some knowledge on univariate and bivariate descriptive methods

Let students fish for information and answer the original question using their own choices of univariate or bivariate methods

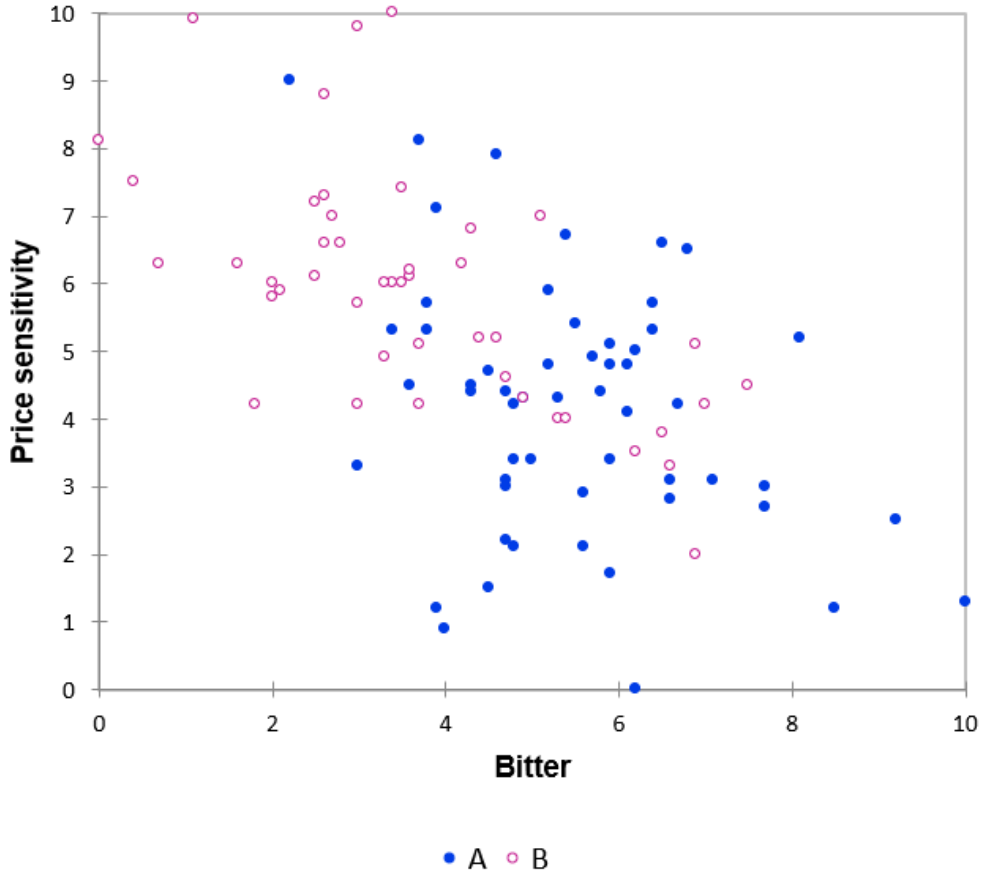






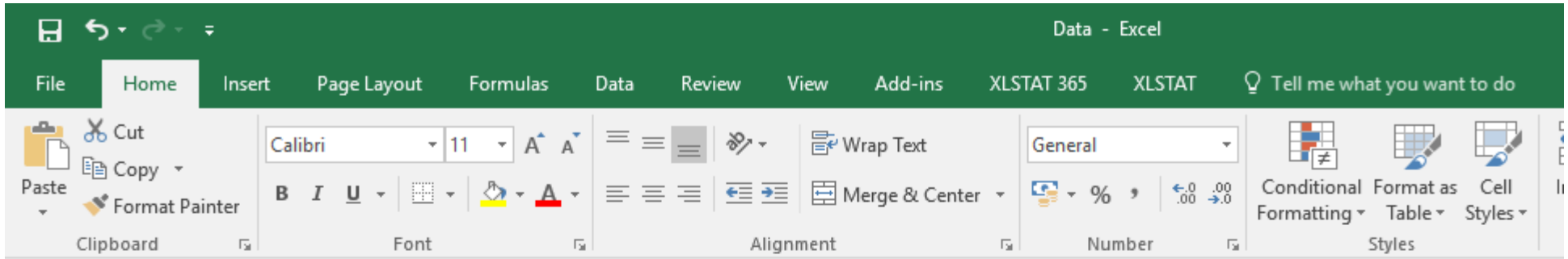
Statistic	Minimum	Maximum	Median	Mean	Standard deviation
Price sensitivity | A	0,000	9,000	4,300	4,127	1,894
Price sensitivity | B	2,000	10,000	6,000	5,889	1,752
Online buyer | A	0,100	9,600	6,000	5,553	2,701
Online buyer | B	0,000	10,000	2,800	3,267	2,472

Display one of the scatter plots the students came up with and let them comment on it



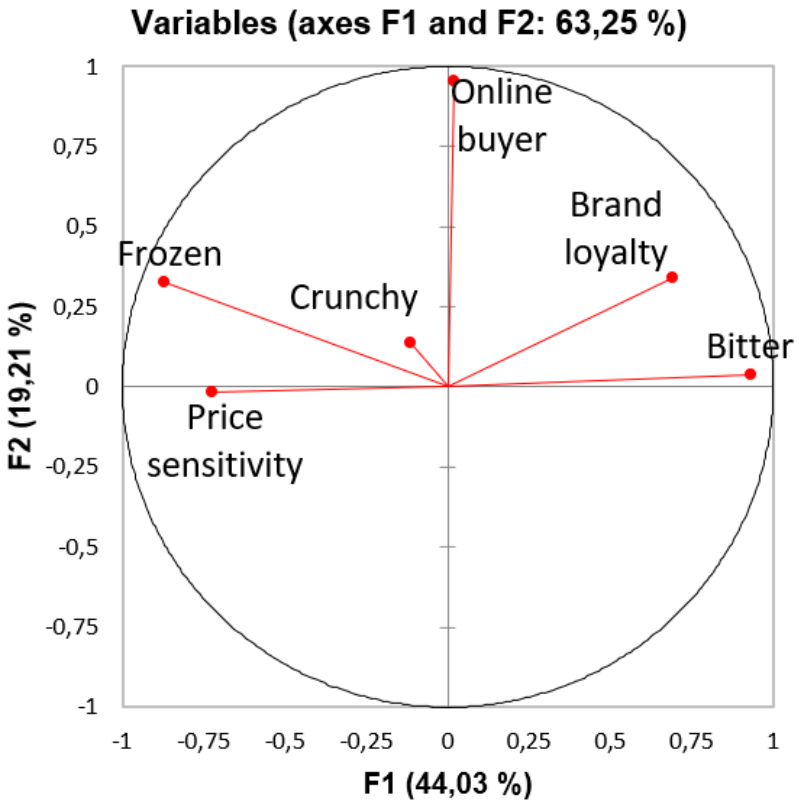
- Price sensitivity decreases with bitterness
- Gender A tends to prefer bitter chocolate
- Gender B tends to be more sensitive to price
- ...More and more information

The scatter plot allowed investigating the information carried by 3 variables...



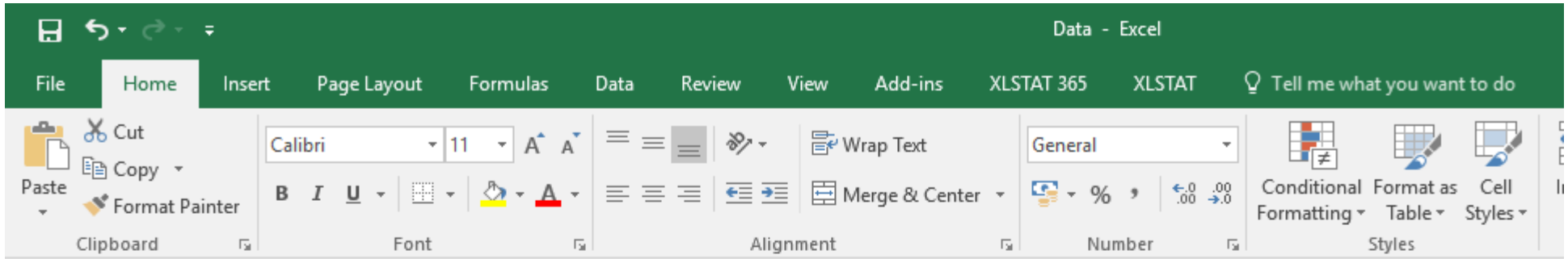
	A	B	C	D	E	F	G	H
1	**Student ID**	**Gender**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**
2	ID01	A	6	3	1	7	2	3
3	ID02	B	5	10	0	3	8	7
4	ID03	B	6	3	3	7	3	5
5	ID04	A	9	1	5	10	1	4
6	ID05	B	5	7	5	3	8	3
7	ID06	B	9	5	9	7	5	7
8	ID07	A	5	3	3	6	4	1
9	ID08	A	7	4	4	5	5	6
10	ID09	A	9	5	7	8	1	3
11	ID10	A	8	2	10	5	6	6

How about having the same kind of reasoning on more variables at the same time?
Problem: the human eye can only see in up to 3 Dimensions



Step 3: Switch to Multivariate thinking based on the conceptual representation of information

Dimensionality reduction techniques allow having the same reasoning on many variables at the same time



	A	B	C	D	E	F	G	H
1	**Student ID**	**Gender**	**Brand loyalty**	**Price sensitivity**	**Online buyer**	**Bitter**	**Frozen**	**Crunchy**
2	ID01	A	6	3	1	7	2	3
3	ID02	B	5	10	0	3	8	7
4	ID03	B	6	3	3	7	3	5
5	ID04	A	9	1	5	10	1	4
6	ID05	B	5	7	5	3	8	3
7	ID06	B	9	5	9	7	5	7
8	ID07	A	5	3	3	6	4	1
9	ID08	A	7	4	4	5	5	6
10	ID09	A	9	5	7	8	1	3
11	ID10	A	8	2	10	5	6	6

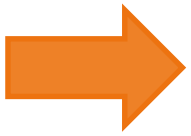
Show the switch as a re-distribution of information

(Focus on conceptual understanding)

Initial dataset

Brand loyalty	Price sensitivity	Online buyer	Bitter	Frozen	Crunchy
6	3	1	7	2	3
5	10	0	3	8	7
6	3	3	7	3	5
9	1	5	10	1	4
5	7	5	3	8	3
9	5	9	7	5	7
5	3	3	6	4	1
7	4	4	5	5	6
9	5	7	8	1	3
8	2	10	5	6	6
6	5	4	5	6	0
8	2	8	5	7	5
8	4	9	6	5	10
9	2	4	6	4	6
8	5	4	6	4	3
2	3	0	6	3	8

Amount of information



Artificial data set synthesized by PCA

The information is re-distributed with high amounts in a few columns

F1	F2	F3	F4	F5	F6
2,189	-1,527	-0,467	-0,017	-0,170	-0,310
-2,520	-1,039	0,991	1,730	0,160	0,839
1,734	-0,709	0,163	-0,162	-0,167	-0,031
4,239	-0,089	-0,089	-0,217	-0,472	0,306
-1,625	0,285	-0,807	0,488	0,571	0,298
1,098	1,989	0,639	0,439	-0,532	0,325
1,096	-0,896	-1,076	-0,615	0,151	0,036
0,351	0,013	0,373	0,144	0,427	-0,045
2,652	0,430	-0,499	0,976	-1,280	-0,302
0,699	2,229	0,155	-0,803	0,927	-0,044
0,034	-0,017	-1,714	0,034	0,418	0,384
0,633	1,597	0,007	-0,965	1,399	0,103
0,625	2,065	1,661	-0,064	-0,220	-0,023
1,865	0,157	0,591	-0,319	1,089	-0,007
1,073	0,025	-0,498	0,597	0,113	0,308
0,554	-1,858	1,593	-1,008	-0,429	0,186

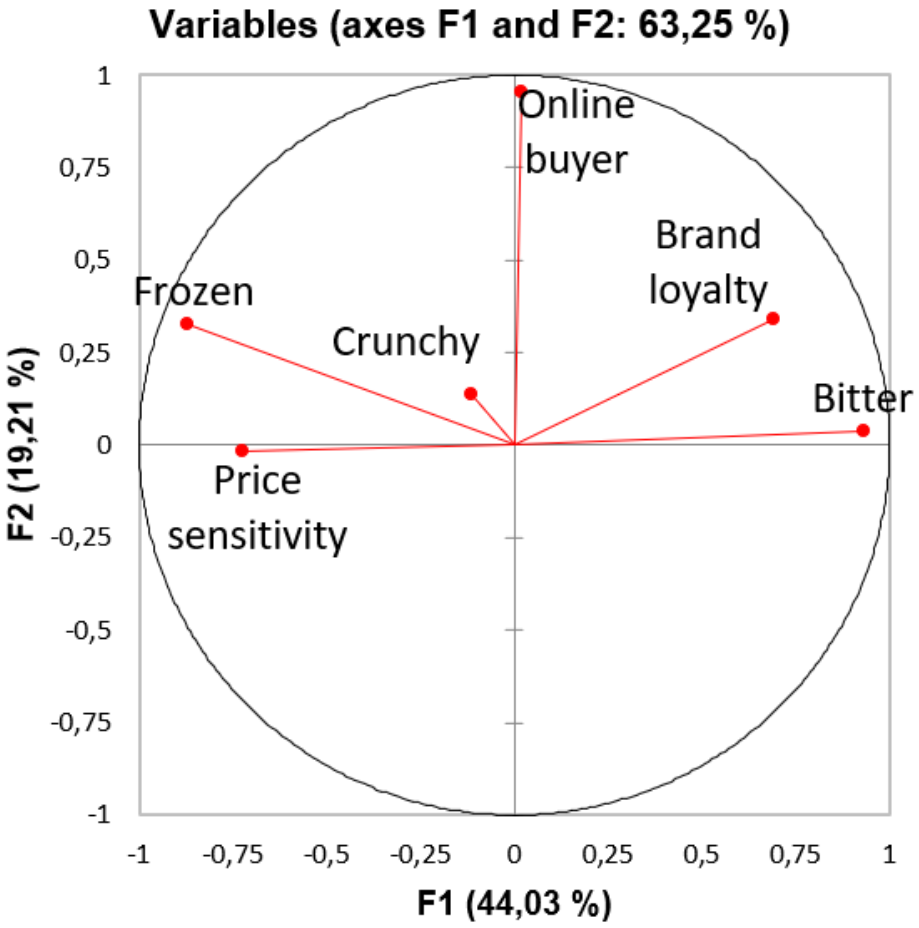
Some jargon:

dimension
= axis
= factor

information
= variability
= inertia

Switch from theoretical to practical

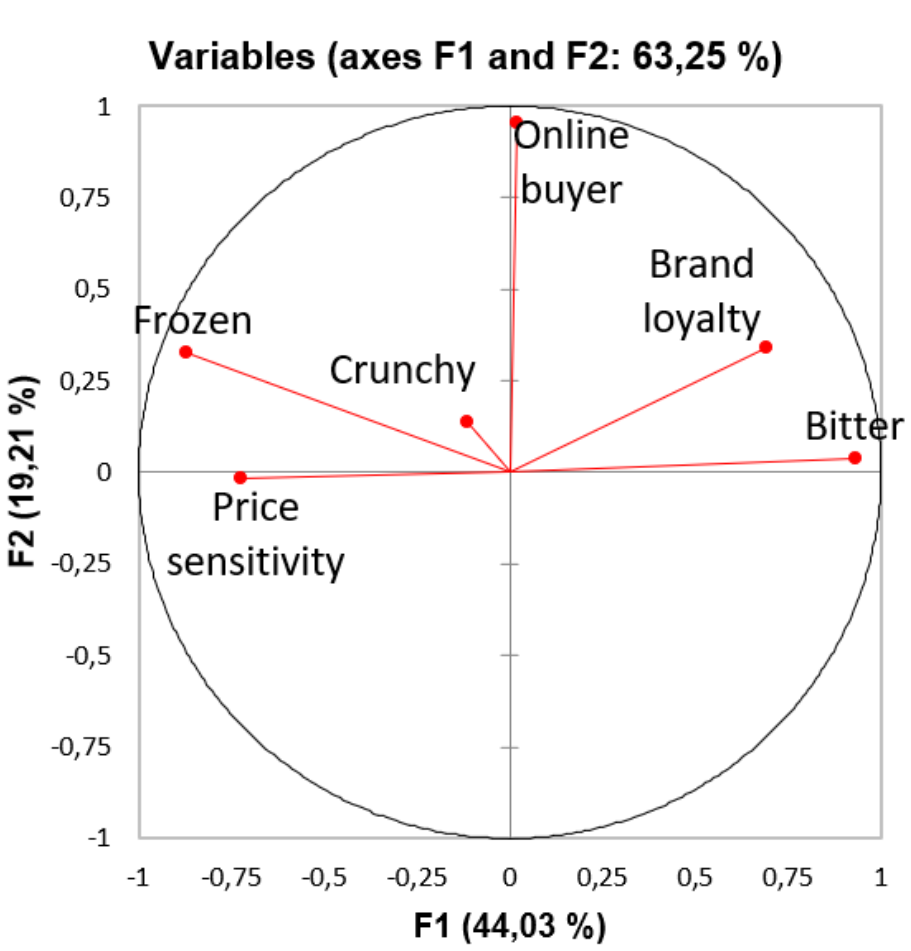
Using user-friendly software and while bypassing complicated maths

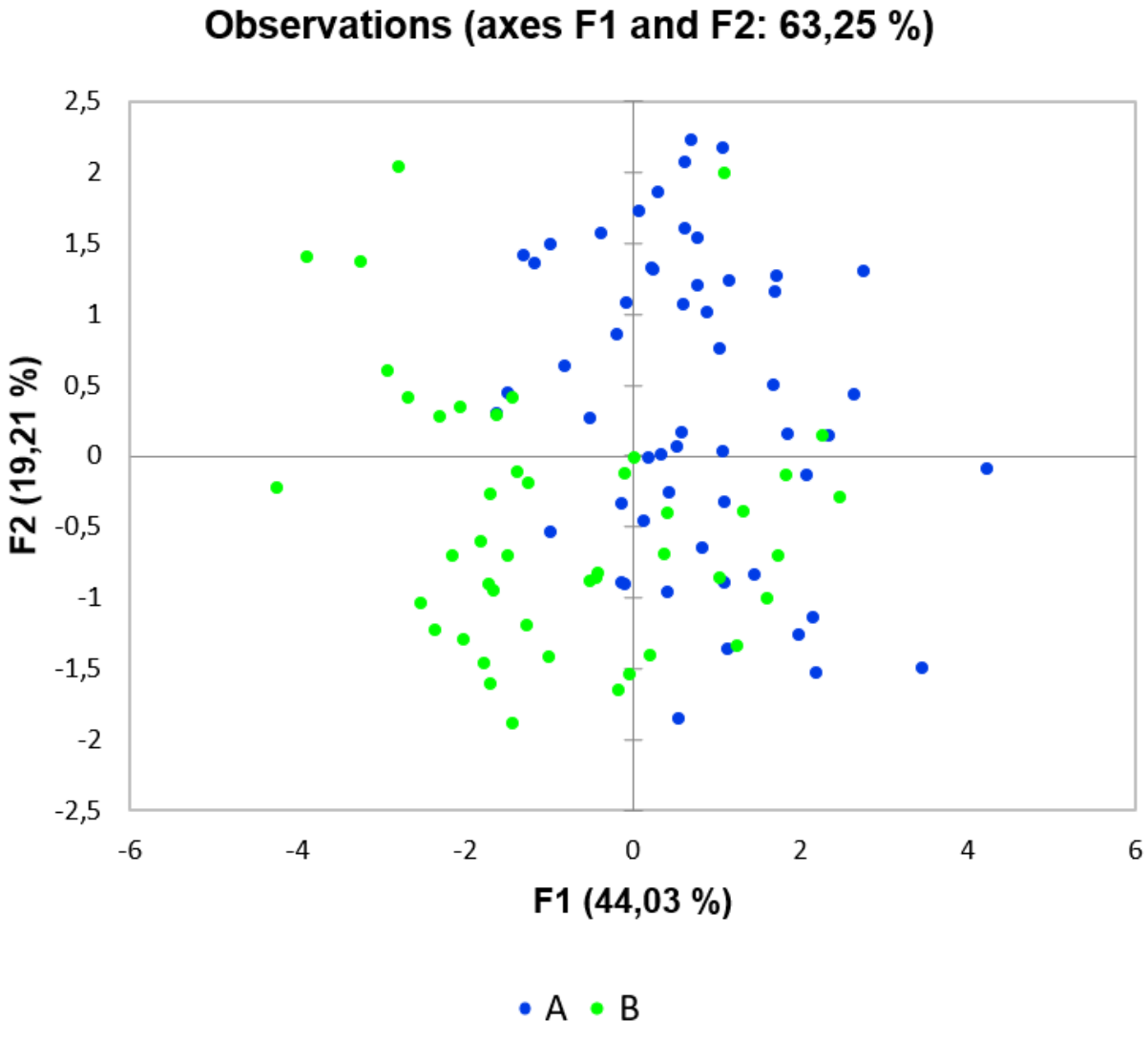


- Relationships investigated in terms of angles
- Quality of representation in terms of vector lengths

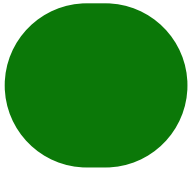
Switch from theoretical to practical

Using user-friendly software and while bypassing complicated maths

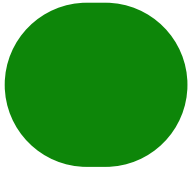




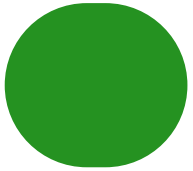
Make them realize they're fishing for information exactly like with Scatter plots



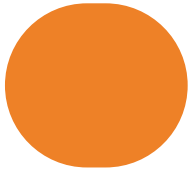
Bitter chocolate lovers are likely not to appreciate frozen chocolate



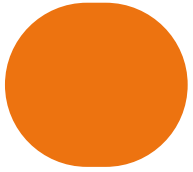
Gender B prefers frozen chocolate



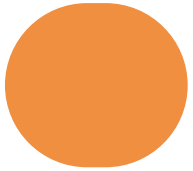
People loyal to brands are less sensitive to price



Affinity to online buying is not related to price sensitivity



Gender A prefers bitter chocolate



And on and on…

And let them take decisions out of these info!

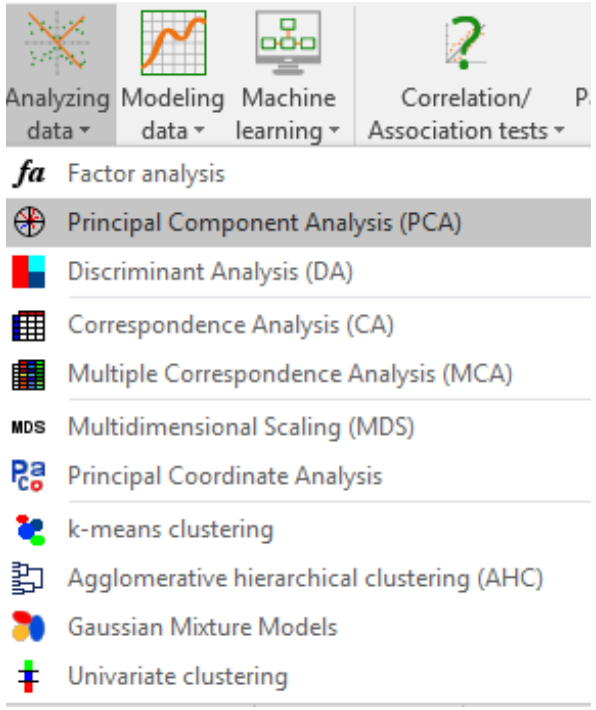
- Promote expensive branded bitter chocolate for gender A
- Promote less expensive frozen chocolate for gender B
- …

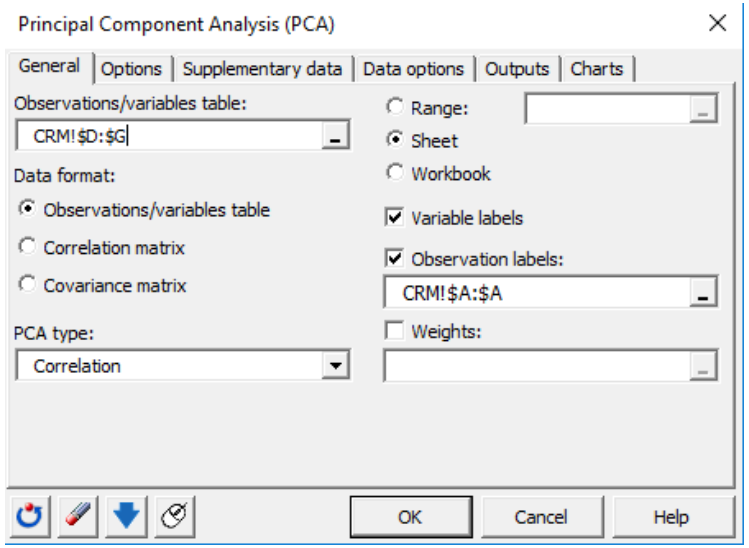
Question	Number of tables	Data description	Tool	Remarks
Exploratory	1	Quantitative variables only	Principal Component Analysis(PCA)	Considers all the variance in the data; components do not necessarily reflect real phenomena
Exploratory	1	Quantitative variables only	Factor analysis (FA)	Considers only the covariance between variables; latent factors reflect real phenomena
Exploratory	1	Proximity matrix	Multidimensional scaling (MDS) /Principal Coordinate Analysis(PCoA)	
Exploratory	1	Contingency table (2 qualitative variables)	Correspondence Analysis (CA)	
Exploratory	1	Qualitative variables only	Multiple Correspondence Analysis(MCA)	

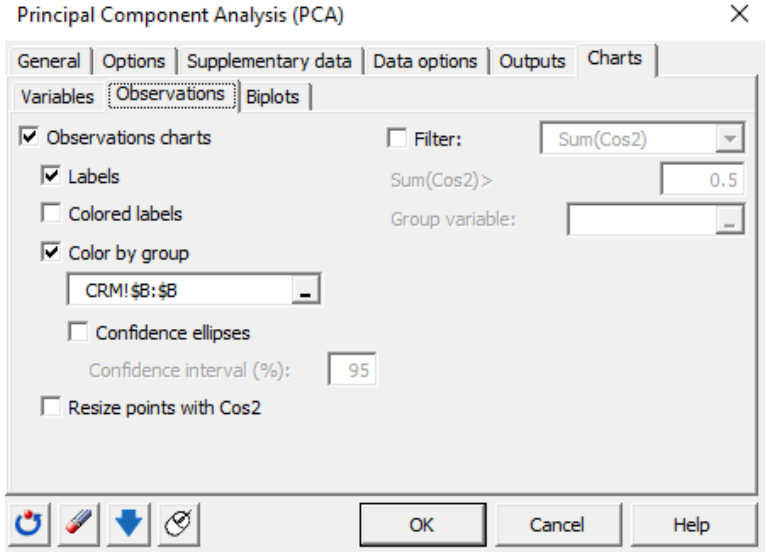
Switching from bivariate to multivariate thinking

1. Find a multivariate data set and highlight a **conceptual representation of information**
2. Generate a scatter plot out of the data
3. Switch to Multivariate thinking based on the **conceptual representation of information**

Setting up a PCA in XLSTAT







PCA tutorial link