Multiple Correspondence Analysis (MCA)

Hervé Abdi

In a nutshell:

- **MCA is Principal Component Analysis for qualitative data**

(Hey, Buddy) Then, "In your nutshell," what is Principal Component Analysis (PCA)?

- PCA analyzes rectangular *quantitative* data tables
- Rows are observations described by the columns (variables)
- PCA creates new best variables called *components* or *factor scores*
- A component is a mixture of the original variables
- The amount in the mixture of a variable is called its *loading* for a component
- PCA makes two maps
- One map for the observations: Components give the coordinates
- One map for the variables: Loadings give the coordinates
- In PCA These maps have different scales (but you can cheat…)

Then. In a nutshell, what is Multiple Correspondence Analysis (MCA)?

- MCA analyzes rectangular *qualitative* data tables (0/1)
- Rows are observations described by the columns (variables)
- MCA creates new best variables called *components* or *factor scores*
- A component is a mixture of the original variables (only the 1's count)
- The amount in the mixture of a variable is called its *loading/score*
- MCA makes two maps
- One map for the observations: Scores give the coordinates
- One map for the variables: Scores give the coordinates
- In MCA These maps have the same scale (but you can cheat …)

Example of qualitative variables

- Gender: M vs F.
- Type of tasting: Blind vs Vision
- Type of Fermentation for Beer: Low vs High
- Color of a Wine: Red, White, or Rosé
- Occupation: Primary, Secondary, or Ternary
- Olive oils, Place of production: Italy, Spain, France, Greece, USA

Disjunctive coding: Code the levels as 0/1 vector

- Gender: M vs F. 2 {0/1} Columns. M = [1 0], F = [0 1]
- Occupation: Primary, Secondary, Ternary. 3 {0/1} Columns.

Disjunctive coding: A story of two tables

Column code for qualitative variables

- Gender: M vs F.   2 columns: {M F}
- Type of tasting: Blind vs Vision. 2 Columns: {Blind|Vision}
- Type of Fermentation for Beers: Low vs High. 2 Columns: {Low|High}
- Color of a Wine: Red, White, or Rosé. 3 Columns {Red|Rosé|White}
- Occupation: Primary, Secondary, or Ternary. 3 Colums: {P|S|T}
- Olive oils, Place of production: Italy, Spain, France, Greece, USA
  5 Columns: {Italy|Spain|France |Greece|USA}

What MCA does with qualitative variables: 0/1

Participant	Gender	Occupation
1	M	Primary
2	M	Secondary
3	F	Primary
4	F	Ternary
5	M	Ternary
…		
N		

⇨

Participant	Male	Female	Primary	Secondary	Ternary
1	1	0	1	0	0
2	1	0	0	1	0
3	0	1	1	0	0
4	0	1	0	0	1
5	1	0	0	0	1
…					
N					

What to do with quantitative variables?

Example of quantitative variables

- Participants: Age. From 20 to 70.
- Wines: Alcohol degree. From 11 to 16 degree.
- Participants: weight. From 30k to 120 k.
- Soda drinks: amount of sugar per liter. From 0 g to 200g.

What to do with quantitative variables?

- Transform them into qualitative variables
- How: Bin them. For example. Age: {Young|Middle-Age|Mature}
- How to cut. Try to get balanced levels (same number of observations per bin)
- Look at the distribution of the variable (histogram) to cut

Quantitative: Bin Them. Example Age

20-30	Young
31-50	Middle
51-100	Mature

Participant	**Age**	**Age factor**	**Young**	**Middle**	**Mature**
1	21	Y	1	0	0
2	40	Mi	0	1	0
3	31	Mi	0	1	0
4	100	Ma	0	0	1
5	53	Ma	0	0	1
…					
N					

“Kind of” quantitative, Example Likert

A Likert scale: A question on wine

Rosé wine should be cheap:

- ☐ Totally agree
- ☐ Agree
- ☐ Disagree
- ☐ Totally disagree

Likert Scale

Rosé wine should be cheap:

- ☐ Totally agree
- ☑ Agree
- ☐ Disagree
- ☐ Totally disagree

Likert Scale

Rosé wine should be cheap:

1	☐ **Totally agree**
2	☑ **Agree**
3	☐ **Disagree**
4	☐ **Totally disagree**

We need to recode from the distribution

1	1
2&3	2
4	3

Participant	Original	Recoded	Rosé-1	Rosé-2	Rosé-3
1	4	3	0	0	1
2	1	1	1	0	0
3	3	2	0	1	0
4	2	2	0	1	0
5	1	1	1	0	0
…					
N					

So the whole data table as a factor is

Participant	Gender	Occupation	Age	Rosé
1	M	Primary	Y	3
2	M	Secondary	Mi	1
3	F	Primary	Mi	2
4	F	Ternary	Ma	2
5	M	Ternary	Ma	1
…				
N				

The factor table is recoded as a 0/1 table

Participant	Gender	Occupation	Age	Rosé
1	M	Primary	Y	3
2	M	Secondary	Mi	1
3	F	Primary	Mi	2
4	F	Ternary	Ma	2
5	M	Ternary	Ma	1
…				
N				

R

Participant	Gender-M	Gender-F	Occ-P	Occ-S	Occ-T	Age-Y	Age-Mi	Age-Ma	Rosé -1	Rosé -2	Rosé -3
1	1	0	1	0	0	1	0	0	1	0	0
2	1	0	0	1	0	0	1	0	0	1	0
3	0	1	1	0	0	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1	0	0	1
5	1	0	0	0	1	0	0	1	0	0	1
…											
N											

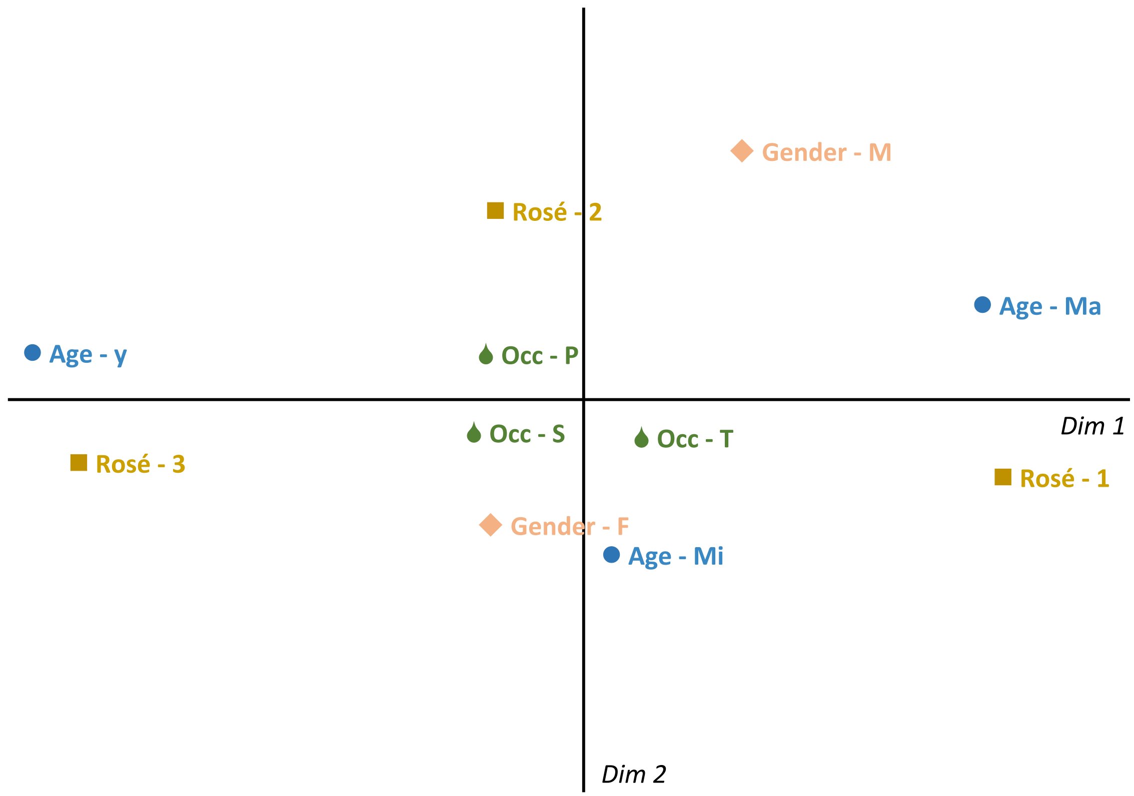
Vocabulary

- Original data table
- Recoded data table: The factor data table
- The 0/1 data table: Often called the complete disjunctive 0/1 table

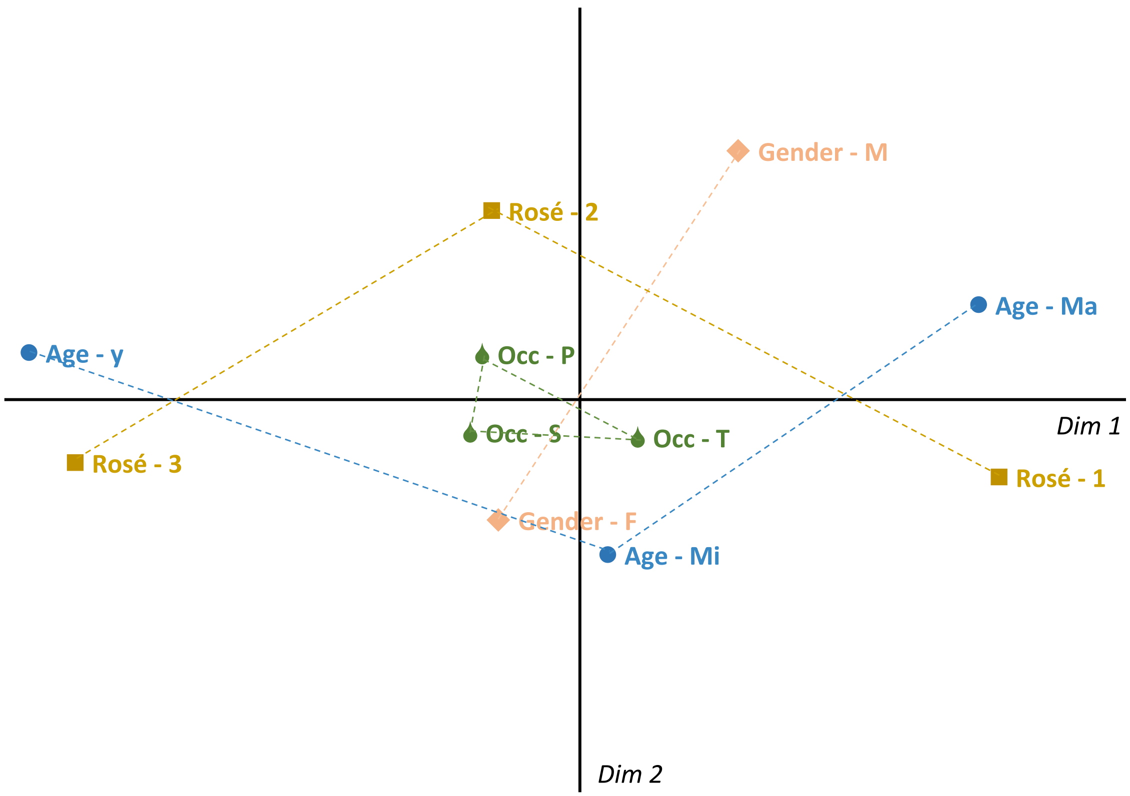
MCA is simply the CA of 0/1 table

Participant	Gender-M	Gender-F	Occ-P	Occ-S	Occ-T	Age-Y	Age-Mi	Age-Ma	Rosé -1	Rosé -2	Rosé -3
1	1	0	1	0	0	1	0	0	1	0	0
2	1	0	0	1	0	0	1	0	0	1	0
3	0	1	1	0	0	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1	0	0	1
5	1	0	0	0	1	0	0	1	0	0	1
…											
N											

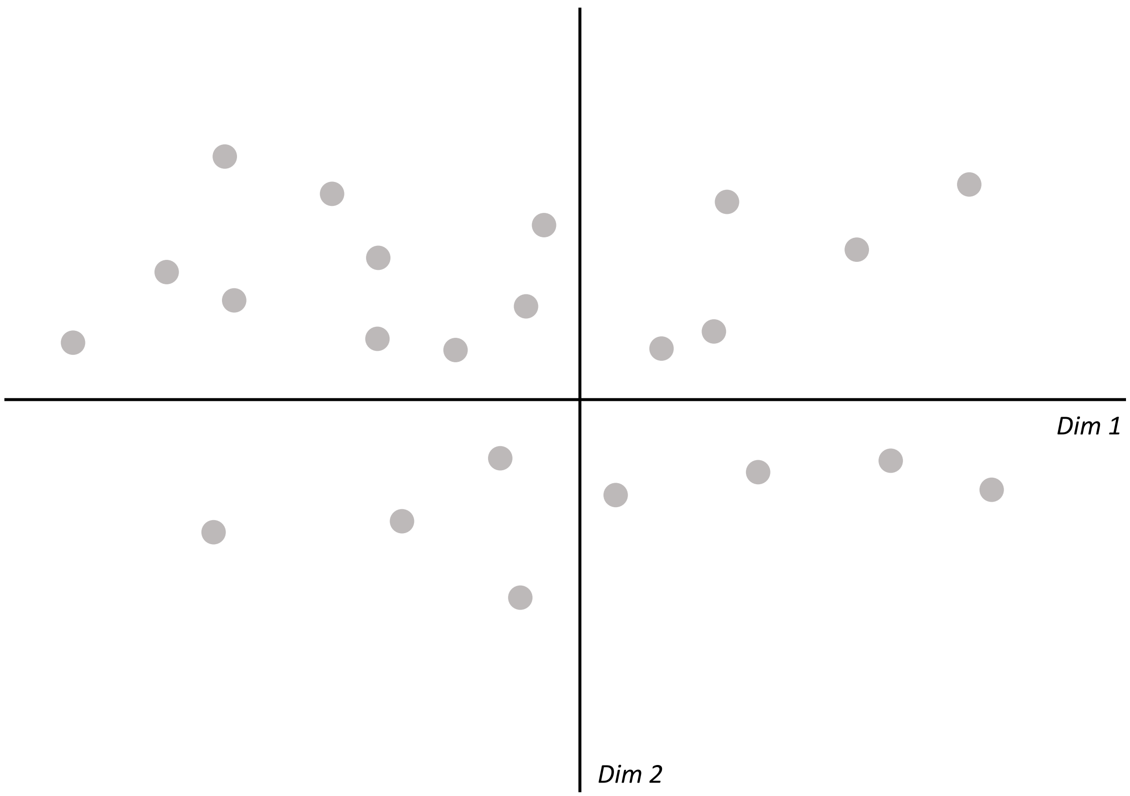
This gives a factor map



Sometimes drawn with lines



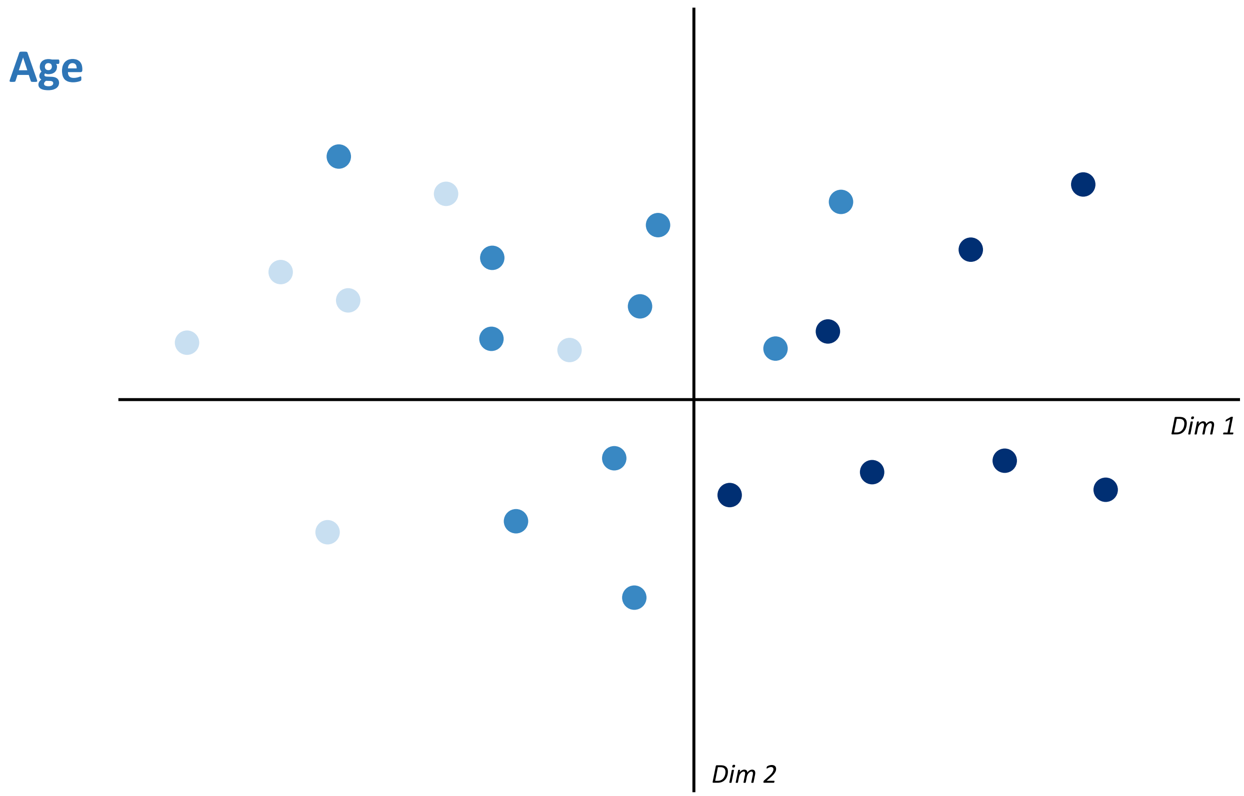
In MCA: there is also a graph for the observations



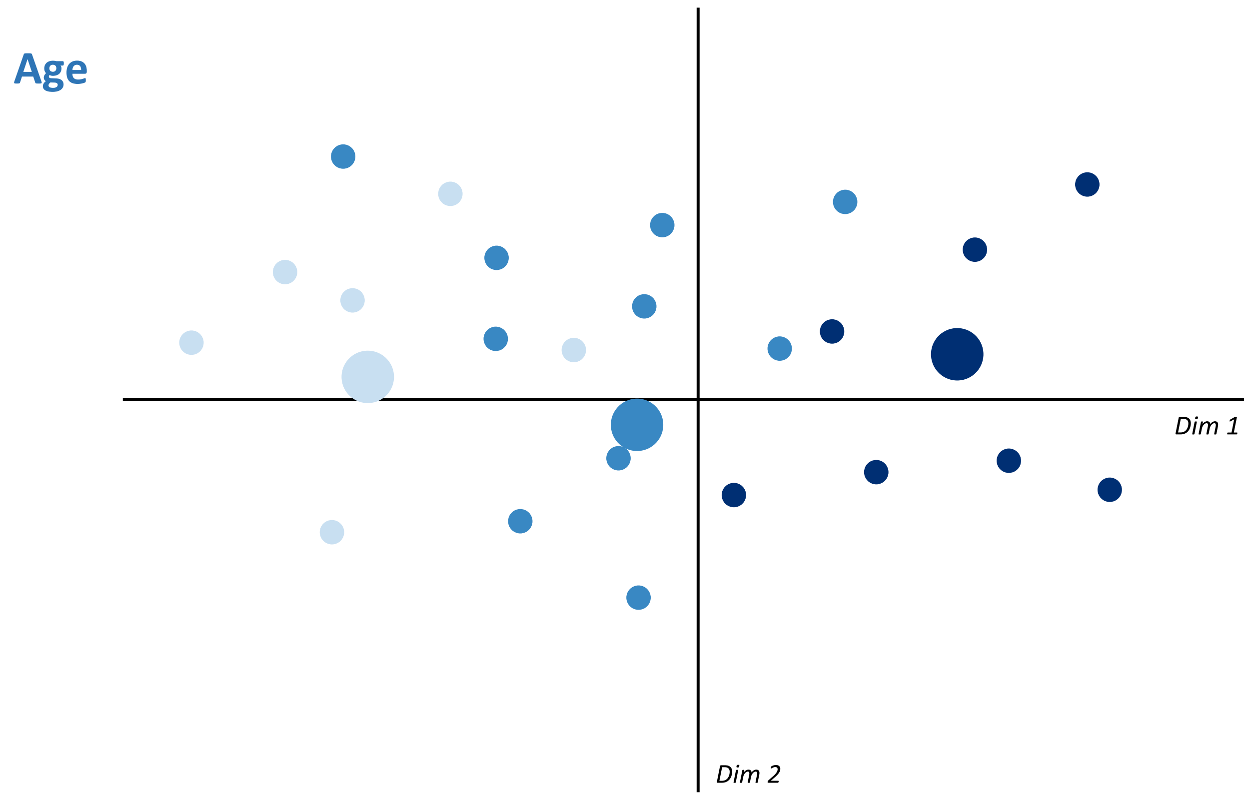
Looks all gray to me …

To me too. Use colors

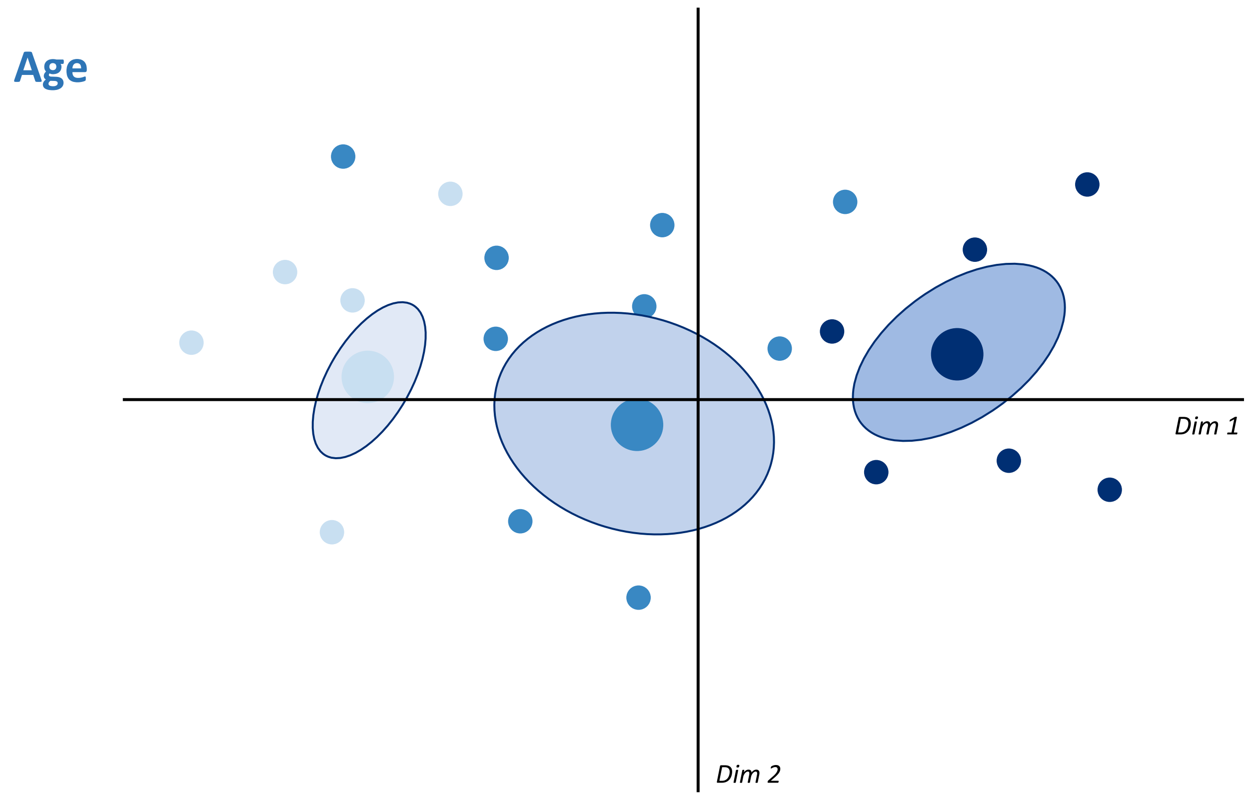
Color the observations by Age groups



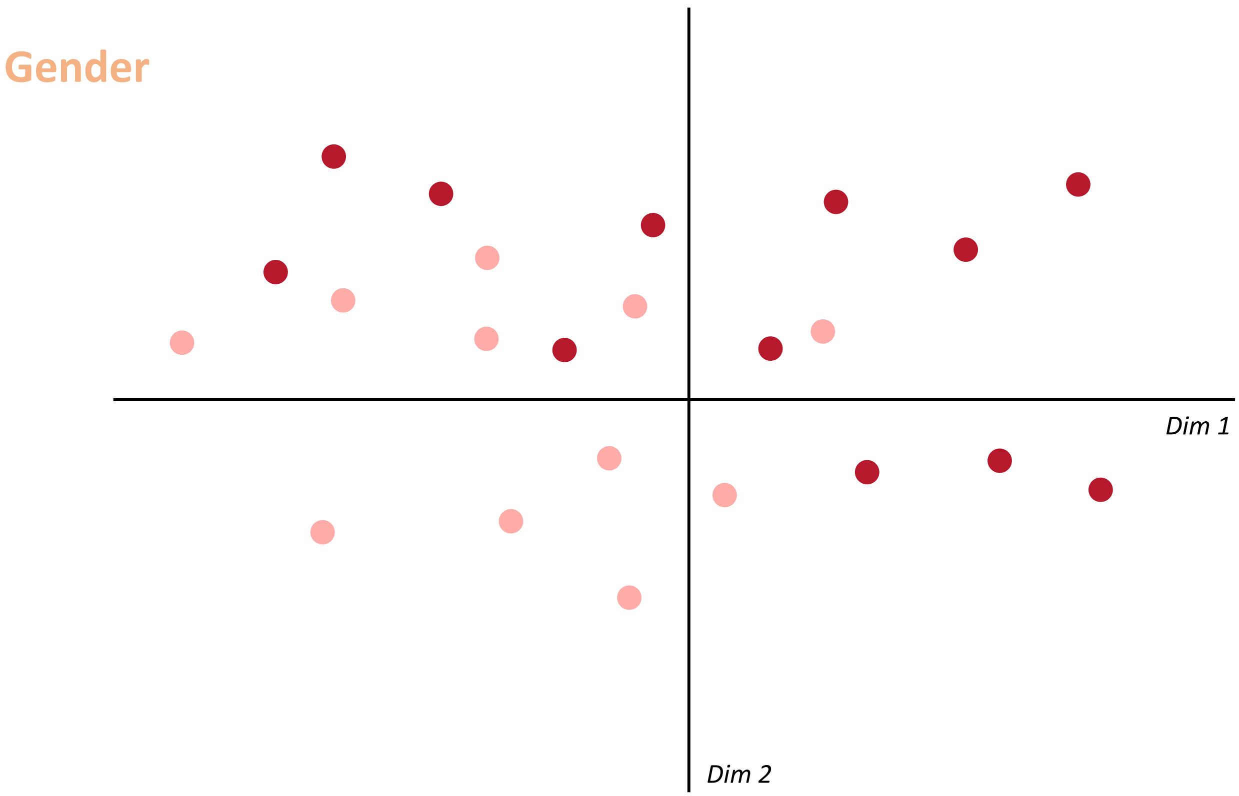
Color the observations by Age groups. With means



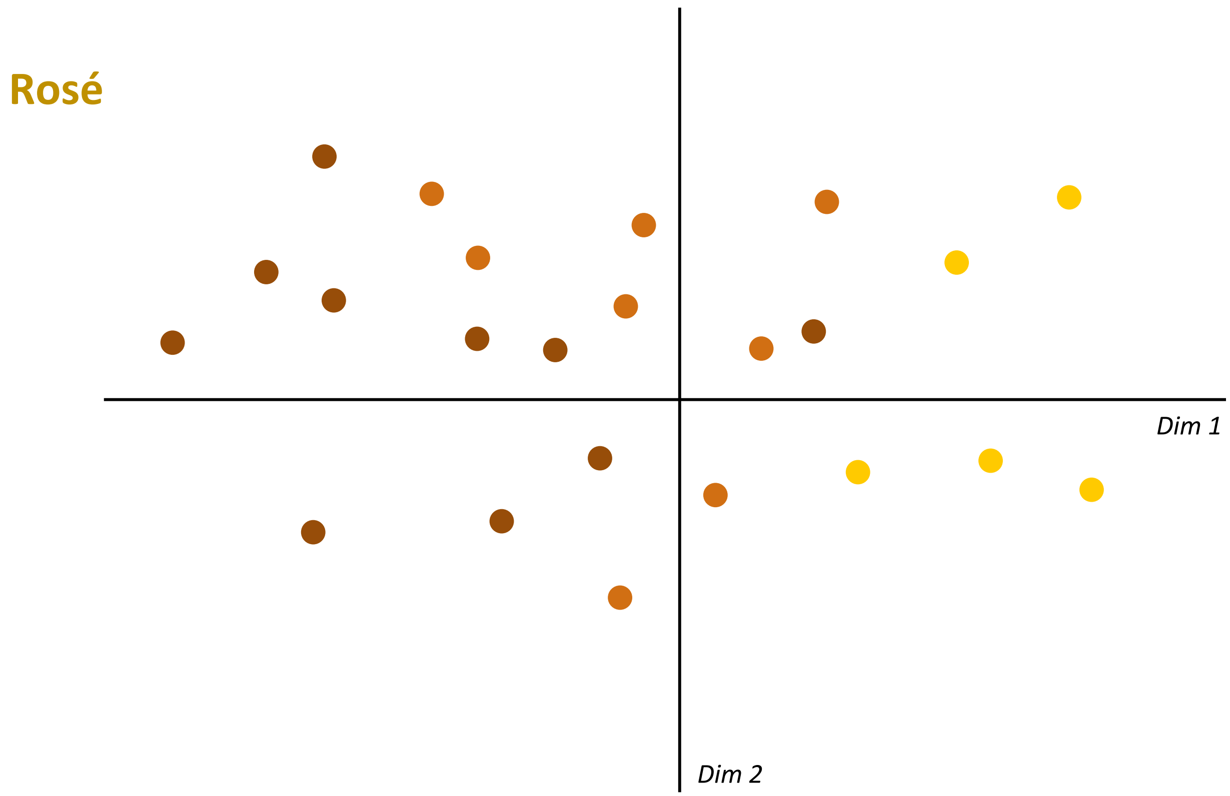
Color the observations by Age groups. Means + CIs



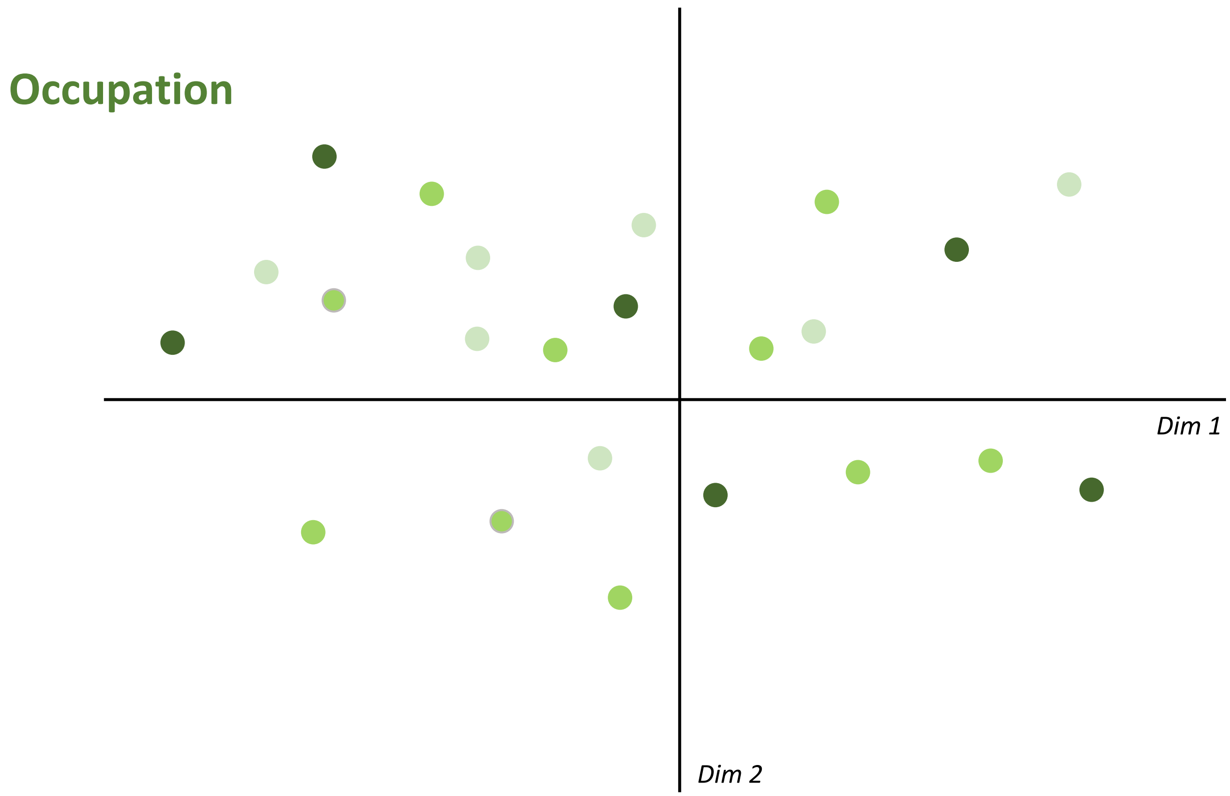
Color by Gender



Color by response to *Rosé*



Color by Occupation



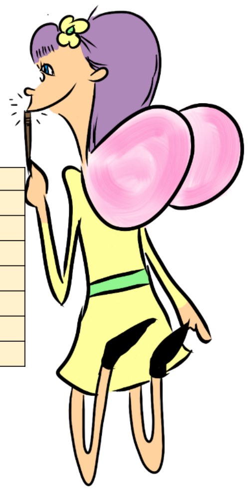
How to interpret an MCA (variables)

- One variable: as many points as levels (compare with PCA).
- Levels of variables close to each other are chosen together.
- Variance of the levels of a variable = importance of the variable.

How is R doing it

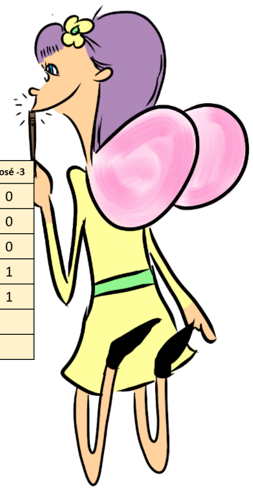
Now give the factor table to R:
R transforms it to the 0/1 table

Participant	Gender	Occupation	Age	Rosé
1	M	Primary	Y	3
2	M	Secondary	Mi	1
3	F	Primary	Mi	2
4	F	Ternary	Ma	2
5	M	Ternary	Ma	1
...				
N				



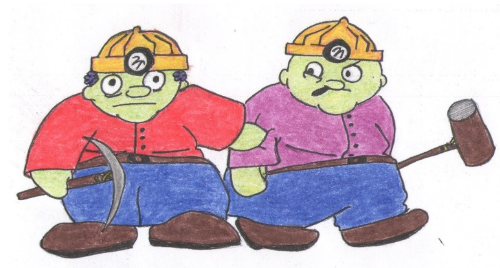
Now give the factor table to R: R transforms it to the 0/1 table

Participant	Gender-M	Gender-F	Occ-P	Occ-S	Occ-T	Age-Y	Age-Mi	Age-Ma	Rosé -1	Rosé -2	Rosé -3
1	1	0	1	0	0	1	0	0	1	0	0
2	1	0	0	1	0	0	1	0	0	1	0
3	0	1	1	0	0	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1	0	0	1
5	1	0	0	0	1	0	0	1	0	0	1
...											
N											

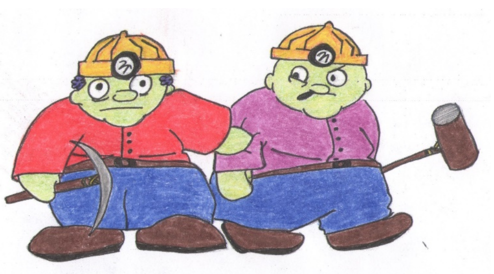


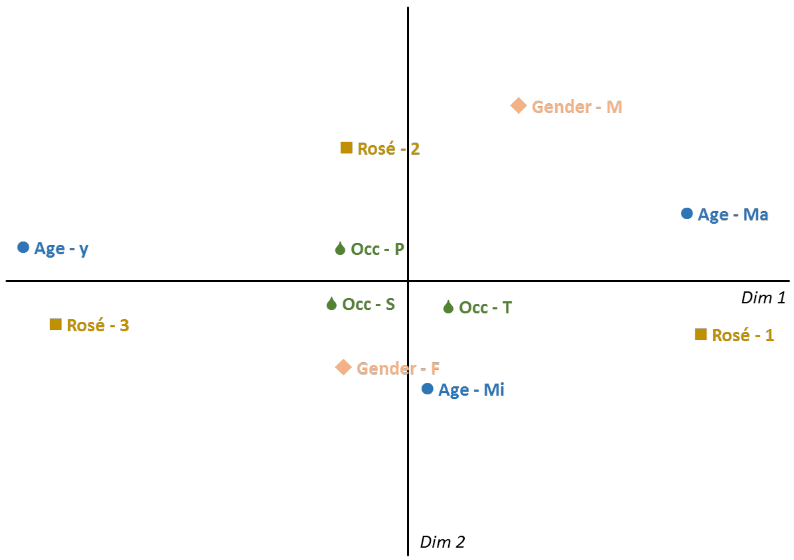
Now the fairy gives the 0/1 to the CA function

Participant	Gender-M	Gender-F	Occ-P	Occ-S	Occ-T	Age-Y	Age-Mi	Age-Ma	Rosé -1	Rosé -2	Rosé -3
1	1	0	1	0	0	1	0	0	1	0	0
2	1	0	0	1	0	0	1	0	0	1	0
3	0	1	1	0	0	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1	0	0	1
5	1	0	0	0	1	0	0	1	0	0	1
…											
N											

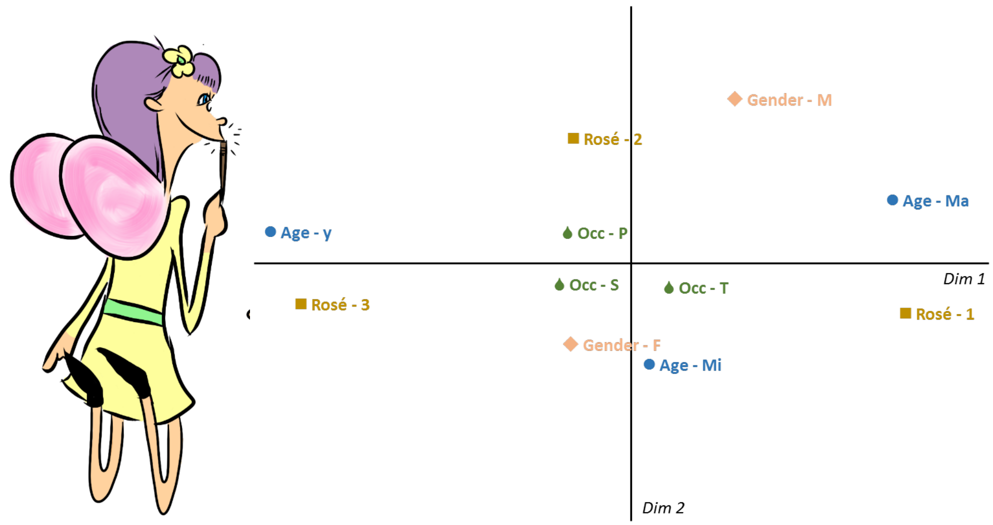


The CA function computes the CA and gives the results to the R fairy





And R gives you back the results



The R fairy calls these steps:
the MCA function for the factor.table

- In R:
- `resMCA <- ExPosition::epMCA(datafac)`

A (way too) small example: 3 Experts, 6 wines

		Expert 1			Expert 2				Expert 3		
wines	Oak-type	fruity	woody	coffee	red fruit	roasted	vanillin	woody	fruity	butter	woody
$wine_1$	1	1	6	7	2	5	7	6	3	6	7
$wine_2$	2	5	3	2	4	4	4	2	4	4	3
$wine_3$	2	6	1	1	5	2	1	1	7	1	1
$wine_4$	2	7	1	2	7	2	1	2	2	2	2
$wine_5$	1	2	5	4	3	5	6	5	2	6	6
$wine_6$	1	3	4	4	3	5	4	5	1	7	5

Data see: Abdi & Valentin (2007): http://www.utdallas.edu/~herve/Abdi-MFA2007-pretty.pdf

Recode (bin + 0/1) as:

		Expert 1							Expert 2									Expert 3					
Wine	Oak Type	fruity		woody			coffee		red fruit		roasted		vanillin			woody		fruity		butter		woody	
W1	1	1	0	0	0	1	0	1	1	0	0	1	0	0	1	0	1	0	1	0	1	0	1
W2	2	0	1	0	1	0	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	1	0
W3	2	0	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0	0	1	1	0	1	0
W4	2	0	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0
W5	1	1	0	0	0	1	0	1	1	0	0	1	0	0	1	0	1	1	0	0	1	0	1
W6	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0	1	0	1
W?	?	0	1	0	1	0	.5	.5	1	0	1	0	0	1	0	.5	.5	1	0	.5	.5	0	1

Data see: Abdi & Valentin (2007): http://www.utdallas.edu/~herve/Abdi-MCA2007-pretty.pdf

Plug the 0/1 matrix into a (plain) CA program

Table 3: Factor scores, squared cosines, and contributions for the observations (I-set). The eigenvalues and proportions of explained inertia are corrected using Benzécri/Greenacre formula. Contributions corresponding to negative scores are in italic. The mystery wine (Wine ?) is a supplementary observation. Only the first two factors are reported.

F	${}_c\lambda$	$\%_c$	Wine 1	Wine 2	Wine 3	Wine 4	Wine 5	Wine 6	Wine ?
			Factor Scores						
1	.7004	95	0.86	−0.71	−0.92	−0.86	0.92	0.71	0.03
2	.0123	2	0.08	−0.16	0.08	0.08	0.08	−0.16	−0.16
F			*Squared Cosines*						
1			.62	.42	.71	.62	.71	.42	.04
2			.01	.02	.01	.01	.01	.02	.96
F			*Contributions* × 1000						
1			177	*121*	*202*	*177*	202	121	–
2			83	*333*	83	83	83	*333*	–

MCA for wines: The variables (whatever they are)

Table 4: Factor scores, squared cosines, and contributions for the for the variables (J-set). The eigenvalues and percentages of inertia have been corrected using Benzécri/Greenacre formula. Contributions corresponding to negative scores are in italic. Oak 1 and 2 are supplementary variables.

			Expert 1							Expert 2									Expert 3							
			fruity		woody			coffee		red fruit		roasted		vanillin			woody		fruity		butter		woody		Oak	
			y	n	1	2	3	y	n	y	n	y	n	1	2	3	y	n	y	n	y	n	y	n	1	2
F	$_c\lambda$	$\%_c$	*Factor Scores*																							
1	.7004	95	.90	−.90	−.97	.00	.97	−.90	.90	.90	−.90	−.90	.90	−.97	.00	.97	−.90	.90	.28	−.28	−.90	.90	−.90	.90	.90	−.90
2	.0103	2	.00	.00	.18	−.35	.18	.00	.00	.00	.00	.00	.00	.18	−.35	.18	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
F			*Squared Cosines*																							
1			.81	.81	.47	.00	.47	.81	.81	.81	.81	.81	.81	.47	.00	.47	.81	.81	.08	.08	.81	.81	.81	.81	1.00	1.00
2			.00	.00	.02	.06	.02	.00	.00	.00	.00	.00	.00	.02	.06	.02	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
F			*Contributions* $\times 1000$																							
1			58	*58*	*44*	0	44	*58*	58	58	*58*	*58*	58	*44*	0	44	*58*	58	6	*6*	*58*	58	*58*	58	–	–
2			0	0	83	*333*	83	0	0	0	0	0	0	83	*333*	83	0	0	0	0	0	0	0	0	–	–

Wines and their variables

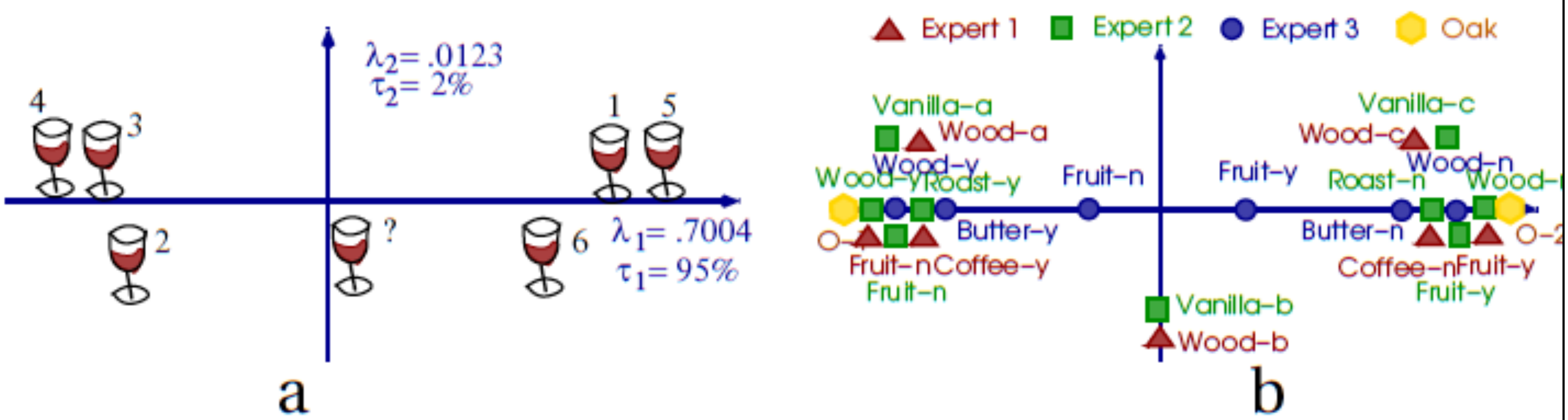


Figure 1: Multiple Correspondence Analysis. Projections on the first 2 dimensions. The eigenvalues (λ) and proportion of explained inertia (τ) have been corrected with Benzécri/Greenacre formula. (a) The I set: rows (*i.e.*, wines), wine ? is a supplementary element. (b) The J set: columns (*i.e.*, adjectives). Oak 1 and Oak 2 are supplementary elements. (the projection points have been slightly moved to increase readability). (Projections from Tables 3 and 4).

Your turn …

Thank you for your attention!