Coffee Project

Corentin DUCLOUX & Youssef DIR

2023-01-16

¶ Table des matières :

PARTIE I : ACP

IMPORTATION DES DONNEES

```
df_1 <- read.csv("~/R data/arabica_data_cleaned.csv", sep = ",", row.names = 1)
df_2 <- read.csv("~/R data/robusta_data_cleaned.csv", sep = ",", row.names = 1)
```

Cette étude portera sur des données issues de 2 datasets concernant les cafés **arabica** et **robusta**. Avant de combiner les jeux de données, il faut d'abord vérifier que les noms des variables sont les mêmes avec la commande `colnames(df_1) == colnames(df_2)`. 6 variables n'ont pas le même nom, on peut donc attribuer le nom des colonnes du premier dataframe au second avec `colnames(df_2) <- colnames(df_1)`. On combine les deux *dataframes* avec la commande `df_both <- rbind(df_1,df_2)`

- Nous obtenons finalement un *dataframe* comprenant **1339** observations et **43** variables.

CHOIX DES VARIABLES

Résultats :

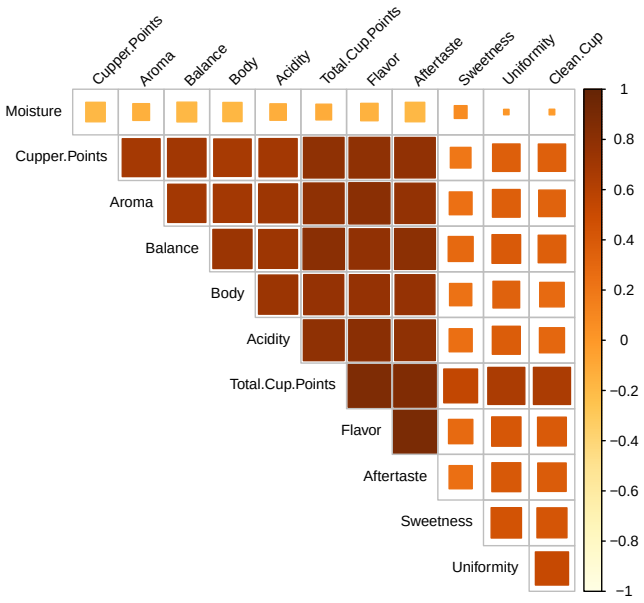
Notre intuition serait de faire une ACP normée (toutes les variables n'ont pas les mêmes unités de mesure) avec les mesures de qualité ($Aroma, Flavor, ..., Moisture$). On pourrait aussi rajouter en caractère qualitatif supplémentaire le type de café ($Species$) ainsi que le score Total ($Total.Cup.Points$) en caractère quantitatif supplémentaire.

Variables retenues pour l'ACP :

Variables	Type
Species	character
Aroma	numeric
Flavor	numeric
Aftertaste	numeric
Acidity	numeric
Body	numeric
Balance	numeric
Uniformity	numeric
Clean.Cup	numeric
Sweetness	numeric
Cupper.Points	numeric
Total.Cup.Points	numeric
Moisture	numeric

On remarque que l'individu **1311** a un score de 0 sur de nombreuses variables numériques ⇒ On peut l'exclure car il risque de fortement influencer l'ACP.

ETUDE DES CORRELATIONS



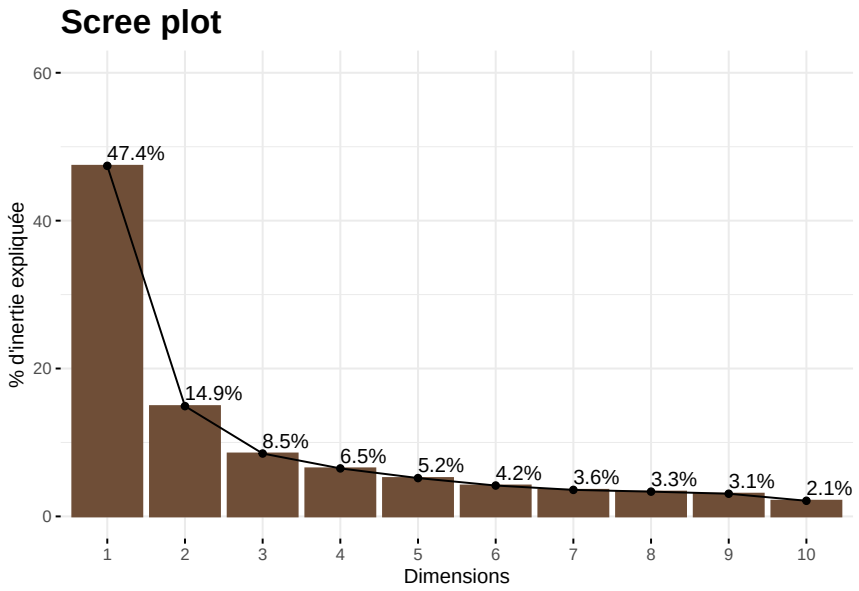
Interprétation des corrélations :

1. On s'aperçoit que les variables $Cupper.Points, Aroma, Balance, ..., Flavor$ sont très positivement corrélés entre elles.
2. La variable $Moisture$ semble quant à elle être légèrement négativement corrélée avec l'ensemble des variables.
3. Les variables $Sweetness, Uniformity, Clean.Cup$ sont globalement corrélés dans le même sens.

ETUDE DES INERTIES

Tableau des inerties :

	Inerties	Inerties relatives (%)	Inerties relatives cumulées (%)
F_1	5.22	47.41	47.41
F_2	1.64	14.91	62.33
F_3	0.94	8.50	70.83
F_4	0.71	6.49	77.33
F_5	0.57	5.18	82.51
F_6	0.46	4.18	86.69
F_7	0.39	3.59	90.27
F_8	0.37	3.35	93.62
F_9	0.34	3.07	96.69
F_{10}	0.23	2.11	98.81
F_{11}	0.13	1.19	100.00



Afin de déterminer le nombre d'axes factoriels F_i à conserver, on peut utiliser la méthode dite du "coude" (On voit que le changement abrupt de pente se produit à partir du troisième axe), mais nous avons préféré tester la méthode introduite dans Detecting Knee Points in System Behavior.

La commande `kneedle(x = c(1:length(resacp$eig[,2])), y = unname(resacp$eig[,2]))` nous renvoie donc l'axe pour lequel il faut s'arrêter et son inertie associée.

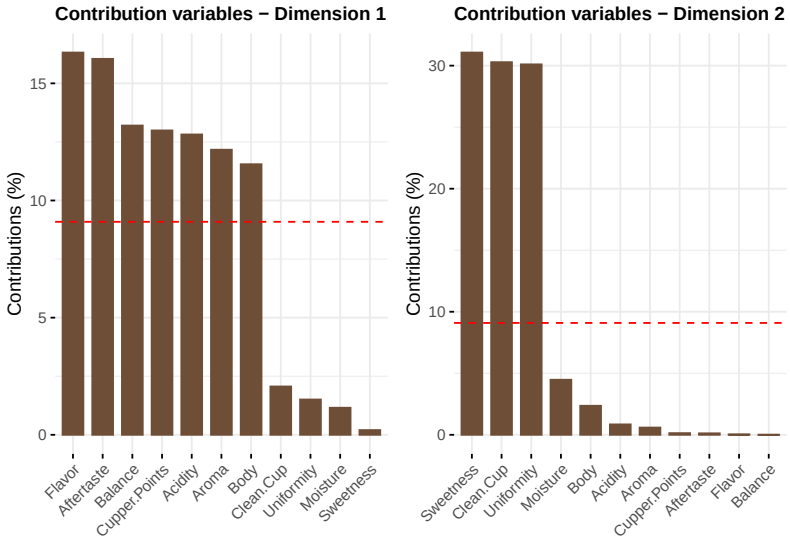
Avec cette méthode, l'algorithme trouve le coude à l'axe **3**. On conserve donc **2** axes pour notre ACP, soit **62.33** % de la variabilité totale expliquée.

ETUDE DES VARIABLES [1]

Contributions et qualité des variables :					
	Cont (F_1)	Cont (F_2)	$cos^2(F_1)$	$cos^2(F_2)$	$cos^2(F_1+F_2)$
Aroma	12.17	0.59	0.63	0.01	0.64
Flavor	16.31	0.03	0.85	0.00	0.85
Aftertaste	16.04	0.12	0.84	0.00	0.84
Acidity	12.82	0.84	0.67	0.01	0.68
Body	11.55	2.36	0.60	0.04	0.64
Balance	13.20	0.00	0.69	0.00	0.69
Uniformity	1.51	30.10	0.08	0.49	0.57
Clean.Cup	2.06	30.28	0.11	0.50	0.60
Sweetness	0.20	31.06	0.01	0.51	0.52
Cupper.Points	12.99	0.13	0.68	0.00	0.68
Moisture	1.15	4.47	0.06	0.07	0.13

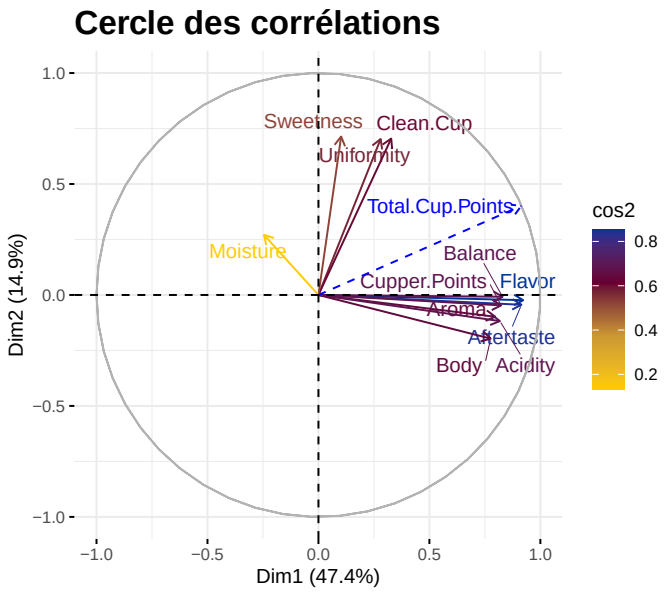
* Les couleurs vertes indiquent les contributions supérieures à la moyenne

[1] Détail des variables

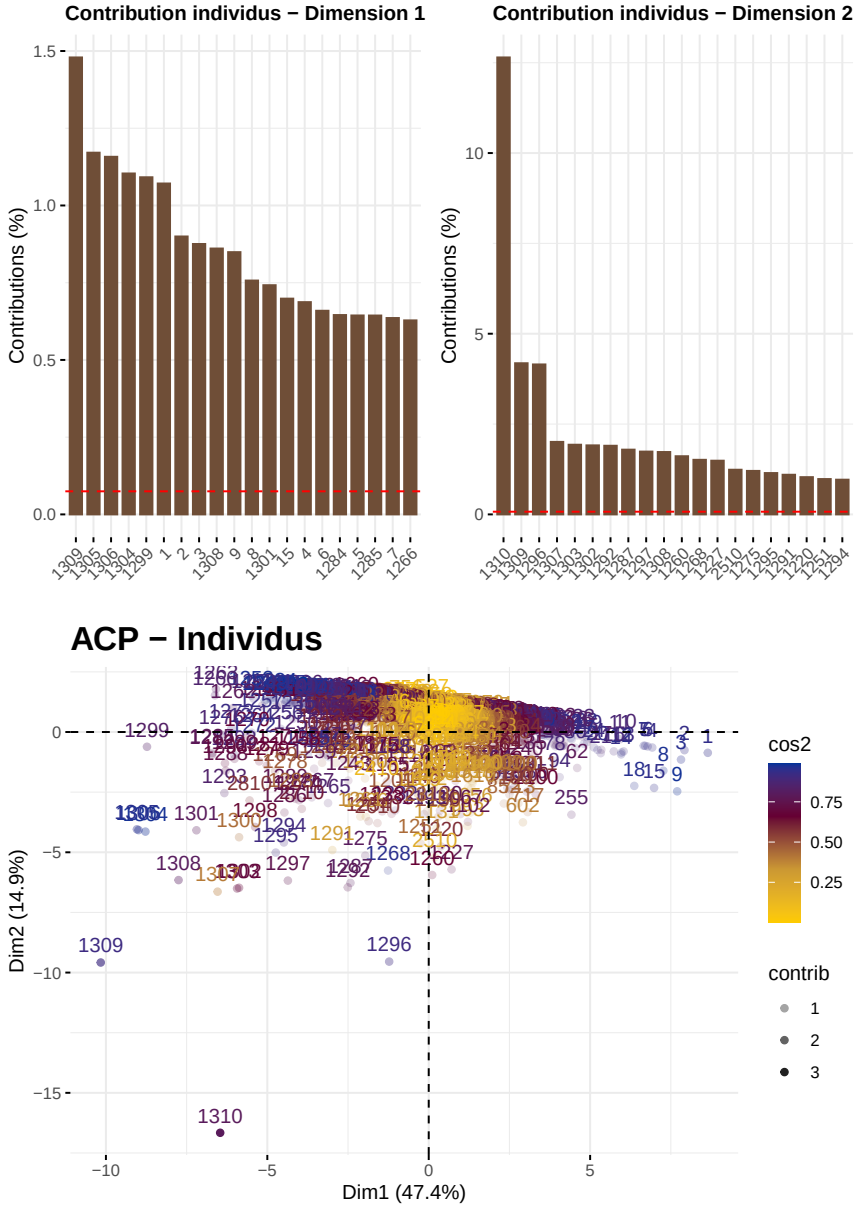


Interprétation des variables :

1. Le tableau des contributions nous indique que les variables ont globalement une qualité de représentation moyenne à bonne ($cos^2 > 0.5$), excepté la variable *Moisture* - il faudra donc être prudent quant à l'interprétation de cette variable dans le plan factoriel (F_1, F_2).
2. Les variables ayant le plus contribué à la construction de l'axe factoriel **1** sont *Flavor*, *Aftertaste*, *Balance*, *Cupper.Points*, *Acidity* et *Aroma* avec des proportions de contribution similaires (10-15 %) $\Rightarrow$ **L'axe factoriel 1 est donc un axe global**.
3. Les variables ayant le plus contribué à la construction de l'axe factoriel **2** sont *Sweetness*, *Clean.Cup* et *Uniformity* avec des niveaux de contribution élévés (au dessus de **20** % chacune). Cependant, l'axe ne porte que 15 % d'inertie.
4. Le Cercle des corrélations nous permet quant à lui d'observer que les 2 cônes de variables contribuent dans le cadran positif du cercle.
5. La variable supplémentaire *Total.Cup.Points* se retrouve projetée entre les 2 cônes de variables - en effet, c'est une combinaison linéaire des autres variables numériques.

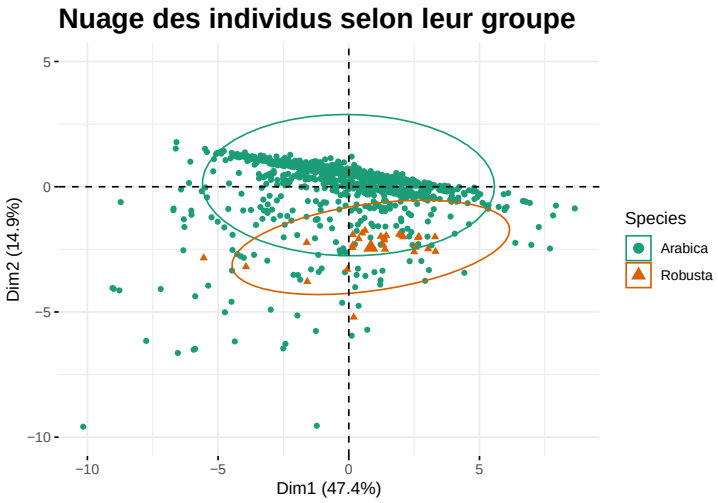


ETUDE DES INDIVIDUS

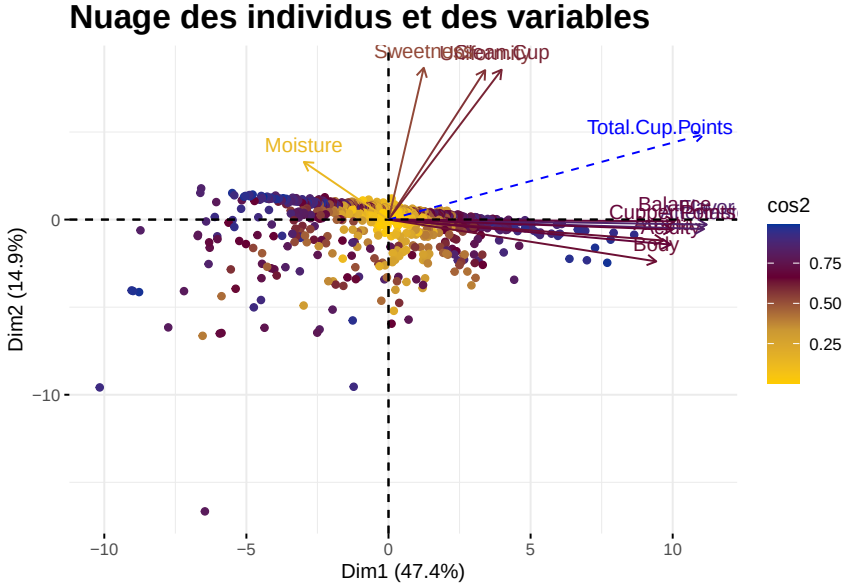


Interprétation des individus :

1. Pour les diagrammes en barres, nous avons sélectionné les 20 individus les plus contributifs, le nombre d'observations rendant difficilement lisible le graphique autrement.
2. Les individus ayant le plus contribué à la construction de l'axe factoriel **1** sont **1309, 1305, 1306, 1304, 1299** avec des proportions supérieures à 1 % chacun.
3. Les individus ayant le plus contribué à la construction de l'axe factoriel **2** sont **1310** (avec une contribution de plus de 10 %), **1296** et **1309**. A eux 3, ils combinent 20 % de l'inertie totale sur l'axe **2** ⇒ **il décrit des individus extrêmes**.
4. La représentation graphique nous montre sans surprise que les individus les mieux représentés et ayant la meilleure qualité sont ceux qui contribuent le plus, tandis que les individus les moins bien représentés se concentrent vers le centre du graphique.



- Une visualisation graphique des individus selon qu'ils appartiennent au groupe **Arabica** ou **Robusta** nous semblait intéressante, et on distingue en moyenne des positions différentes selon le groupe. Cela peut s'expliquer par le fait que la médiane de la variable *Sweetness*[2] des cafés Arabica est égale à **10** tandis que celle des cafés Robusta est égale à **7.67**
- Attention cependant, les groupes sont loin d'être homogènes ! En effet, le groupe Arabica ($n_1 = 1311$) contient beaucoup plus d'observations que le Groupe Robusta ($n_2 = 28$)



Conclusion :

La plupart des "Specialty Coffees" ($Total.Cup.Points > 80$) ont les caractéristiques *Sweetness*, $Uniformity$ et $Clean.Cup = 10/10$. Les critères de différenciation des très bons cafés sont donc *Body*, *Balance*, *Acidity*, *Flavor*. L'axe **1** peut être interprété comme une mesure de la qualité du café (Plus l'individu se situe à droite de l'axe, plus le café est de bonne qualité et a un bon score, et inversement). L'axe **2** pourrait être interprété comme un autre indicateur de qualité, mais lequel ? L'axe est plus difficilement interprétable.

[2] Arabica coffee contains almost twice the amount of sugar than any other coffee

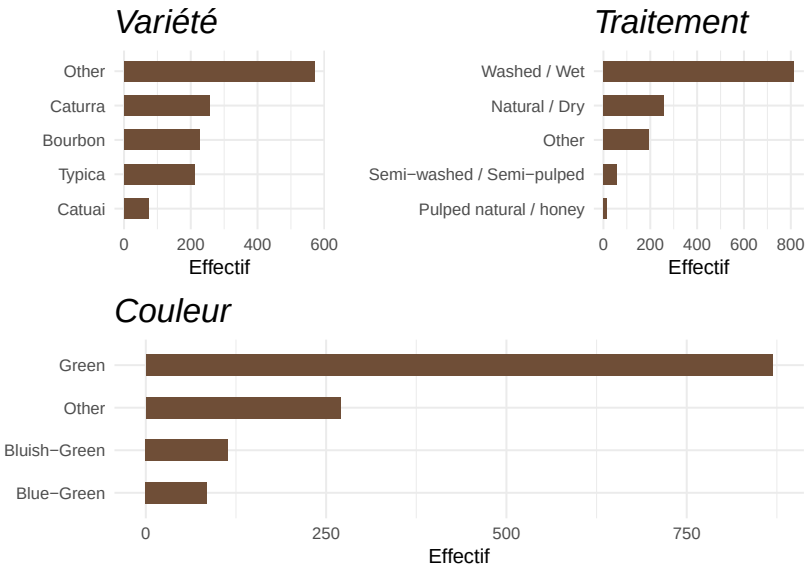
PARTIE II : ACM

- Dans la partie I, nous n'avons pas pu trouver plusieurs groupes d'individus distincts. Nous avons vu grâce à l'ACP qu'avec les **caractéristiques de qualité**, les individus se distribuent de manière globalement uniforme sur l'axe 1.
- Afin de pouvoir effectuer une **Classification hiérarchique**, on opte donc pour l'étude de 3 variables actives : $Variety, Processing.Method, Color$ $\Rightarrow$ La question devient donc : La méthode de traitement et la variété des cafés influe t-elle sur leur couleur ?

MODALITES DES VARIABLES ACTIVES

Recodage préalable :

1. Il convient de noter que la modification des variables est la cause de l'effectif important de la modalité **Other** pour la variable $Variety$.
2. Les modalités rares sont **Catuai** pour la variable $Variety$, **Semi-washed/Semi-pulped** pour la variable $Processing.Method$ et **Pulped natural/Honey**.

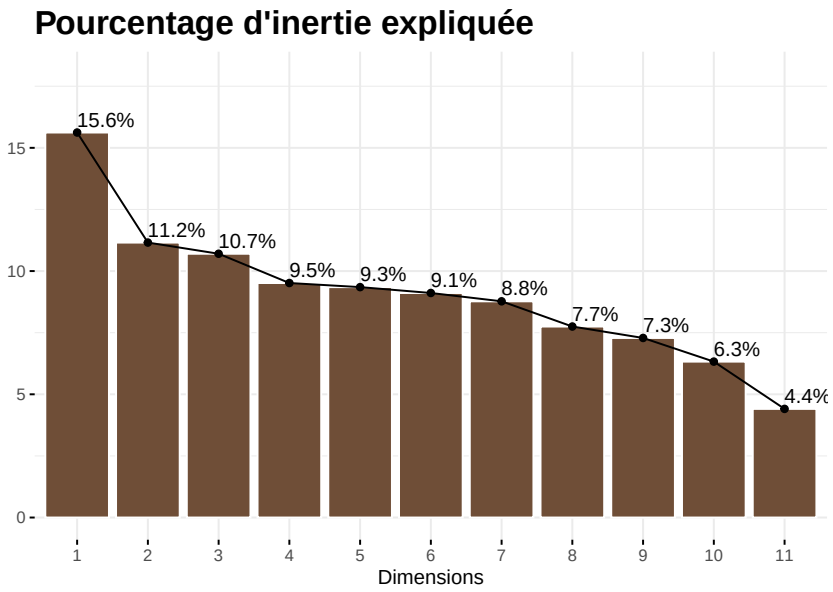


Les V de **Cramer** permettent de mesurer l'intensité des liens entre les variables qualitatives étudiées. On observe alors que :

- Le lien entre la variable $Color$ et $Processing.Method$ est le plus important (**0.28**)
- Le lien entre la variable $Variety$ et $Processing.Method$ (**0.22**).
- Le lien le moins important est entre la variable $Color$ et $Variety$ (**0.17**).

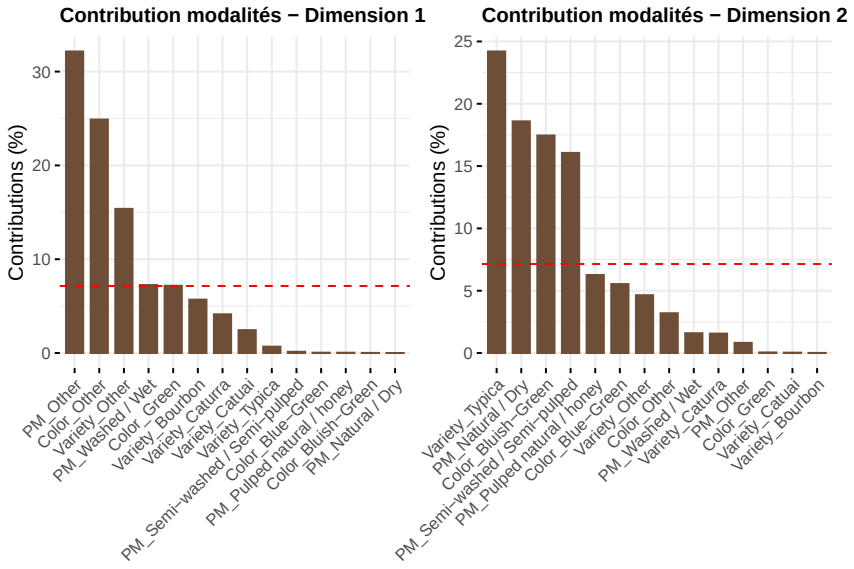
Nous avons aussi ajouté en variables **quantiatives supplémentaires** les variables $Total.Cup.Points$ (pour faire le lien entre l'ACP et l'ACM), $Moisture$ et $Grading.Date$. Celles-ci seront étudiées plus en détail avec un cercle des corrélations (les projections des variables restent quand même proches de 0).

ETUDE DES INERTIES

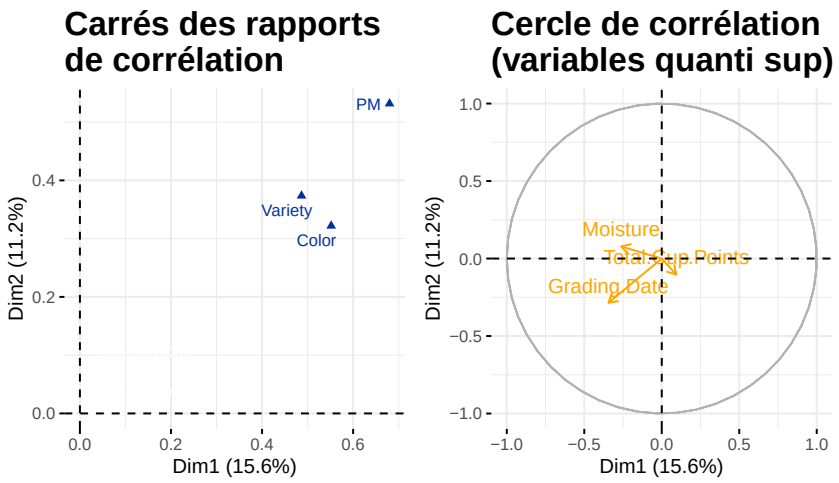


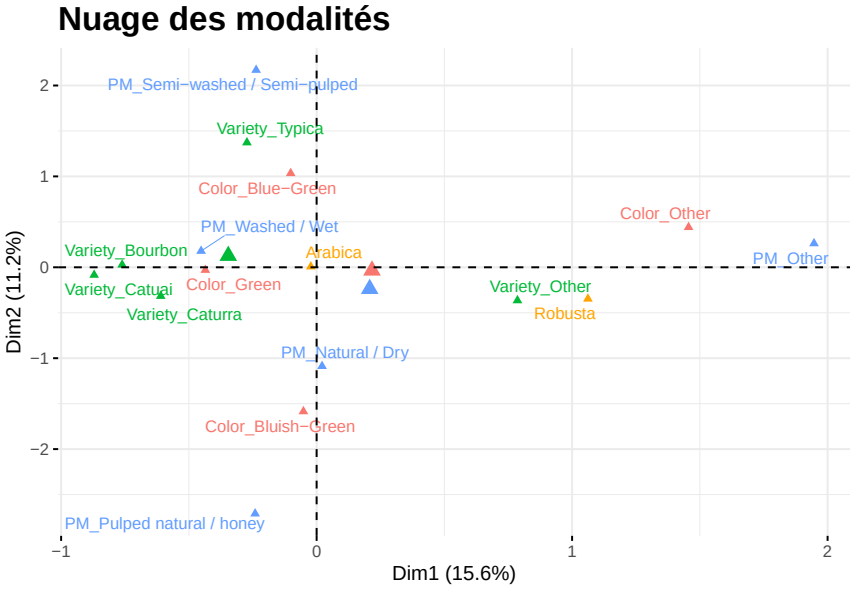
- L'axe F_1 explique **15.62** % de l'inertie totale
- Les axes factoriels F_i avec $(i = 1, \ldots, 4)$ expliquent 47 % de l'inertie totale
- *NB* : Il faudrait théoriquement étudier les 4 dimensions si l'on voulait être précis, mais on peut se restreindre aux deux premières puisque les interprétations restent similaires pour les autres axes.

ETUDE DES VARIABLES



- Sur F_1, les modalités contribuant le plus sont les modalités **Other** des 3 variables actives.
- Sur F_2, la modalité contribuant le plus est la variété **Typica**, s'ensuit la méthode de traitement qui implique un séchage du grain.

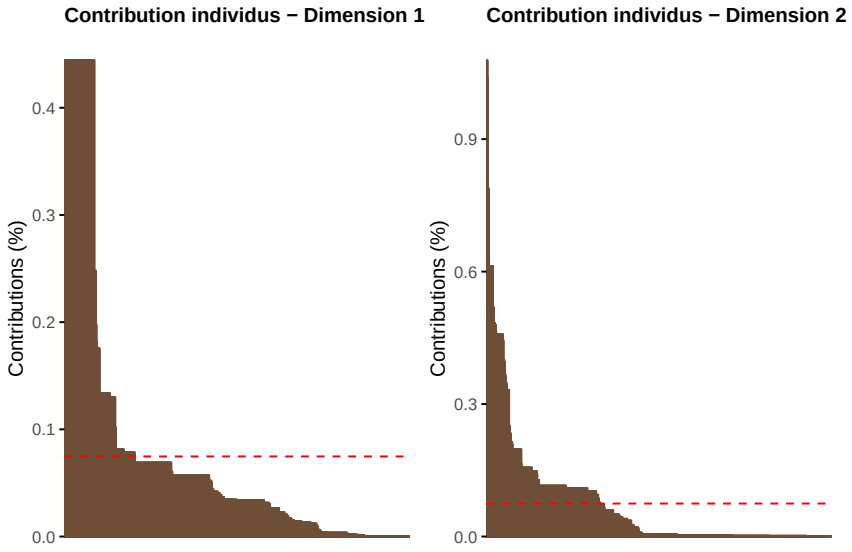




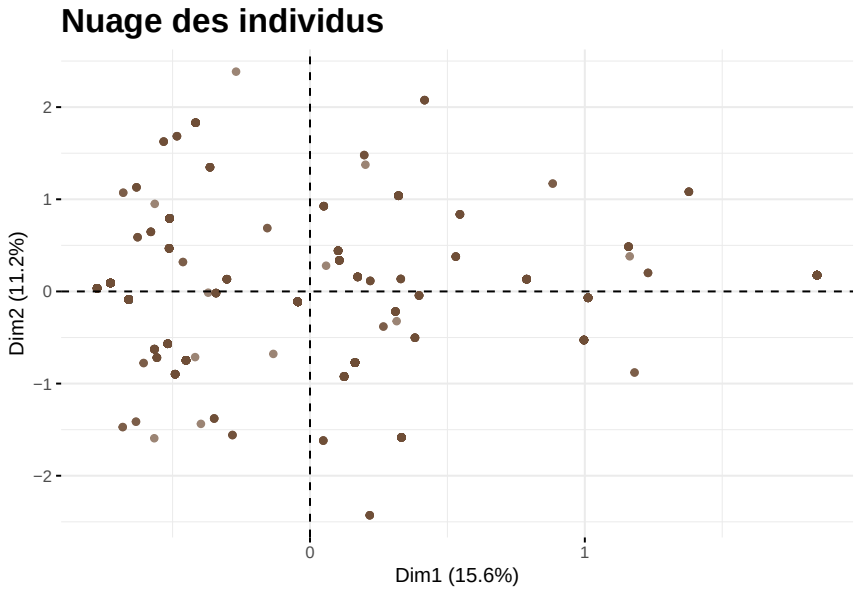
Interprétation des variables :

1. Pour les η^2, on s'aperçoit que la variable la plus structurante sur F_1 et F_2 est $Processing.Method$, suivie de $Variety$ et $Color$ dont les η^2 sont moins importants.
2. Toutes les modalités de $Processing.Method$ semblent être distribuées le long de l'axe F_2, sauf pour la modalité **Other** qui se détache le long de l'axe F_1.
3. Il y a une opposition entre les méthodes de traitement qui impliquent un séchage et celles qui impliquent un nettoyage des grains de café.
4. De plus, on voit que la couleur **Green** s'associe bien avec la méthode de traitement **Washed/Wet**. La variable quanti sup $Moisture$ s'associe bien avec **Washed/Wet**.
5. On remarque aussi que toutes les modalités **Other** se regroupent à droite le long de l'axe F_1.
6. Enfin, les modalités **Blue-Green** et **Bluish-Green** de la variable $Color$ s'opposent sur l'axe F_2.

ETUDE DES INDIVIDUS



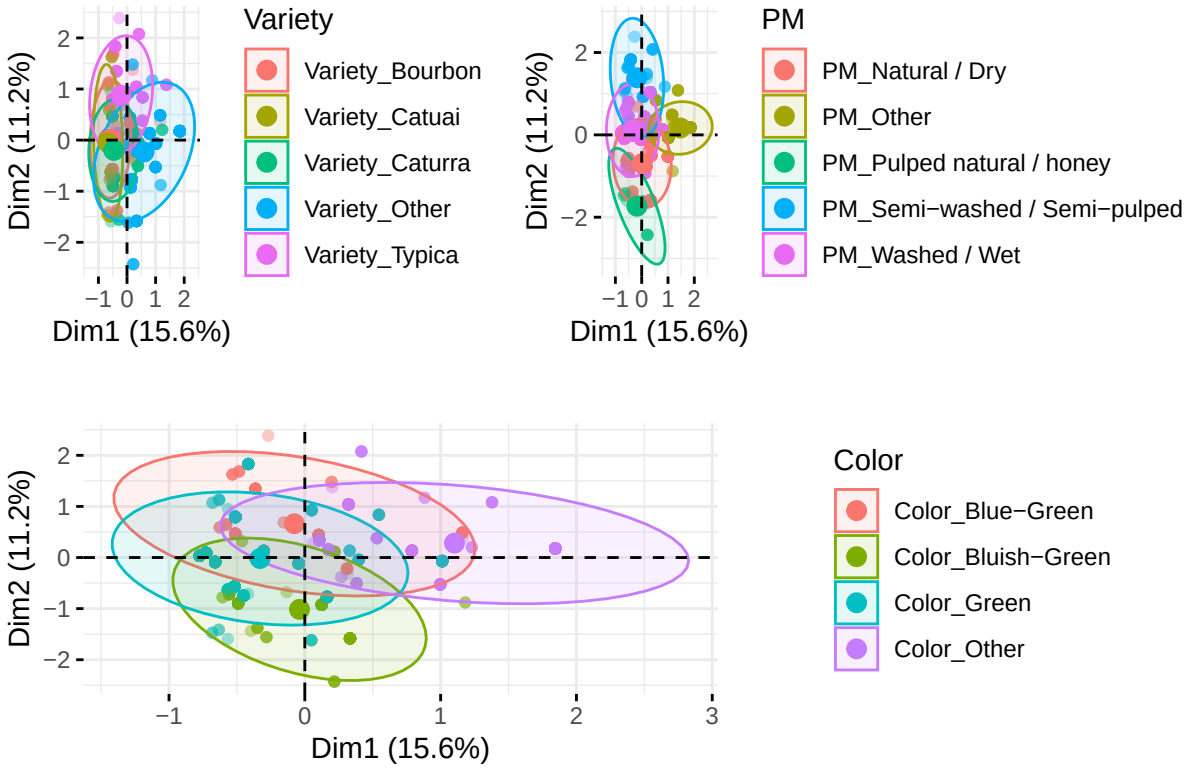
- Sur l'axe F_1, on aperçoit qu'un nombre important d'individus contribue à même proportion (probablement les individus qui prennent la modalité **Other**).
- Sur l'axe F_2, un petit groupe d'individus (ceux qui prennent des modalités rares) contribuent de manière sensiblement plus importante que les autres.
- Pourtant, pour atteindre **50 %** de contribution, il faut à peu près le même nombre d'individus pour F_1 et F_2 $\Rightarrow$ respectivement **113** et **110**.



- Le graph ci-dessus permet de visualiser que les individus se distribuent plutôt bien le long de l'axe F_2.

- Pour l'axe F_1, on voit qu'un groupe d'individus se détache à droite - on peut alors vérifier quelles sont les modalités prises par ces individus grâce à un habillage spécifique du nuage de points.

HABILLAGE DU NUAGE DES INDIVIDUS



Interprétation des individus :

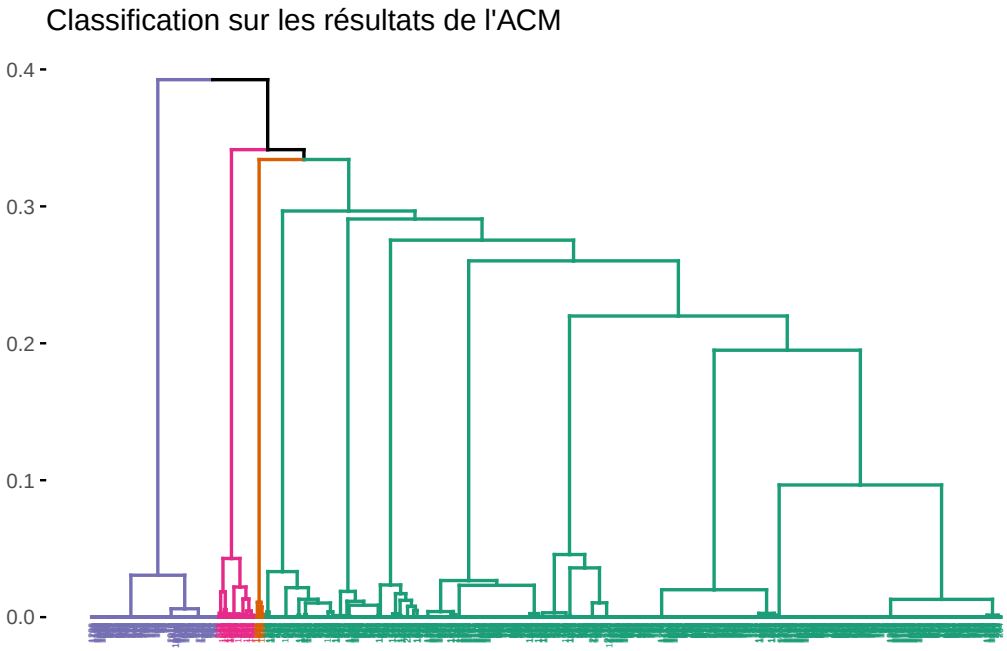
1. L'habillage du nuage des individus par la variable *Variety* n'amène pas d'information particulière, si ce n'est qu'on distingue un détachement des individus prenant la modalité **Other**.
2. L'habillage du nuage des individus par la variable *Processing.Method* montre toujours un détachement des individus prenant la modalité **Other**. On s'aperçoit aussi qu'on distingue beaucoup mieux les individus prenant des modalités différentes.
3. L'habillage du nuage des individus par la variable *Color* confirme une fois de plus le détachement des individus prenant la modalité **Other**.

Dans la prochaine partie de l'étude, nous effectuerons une classification avec les résultats de l'**Analyse des Correspondances Multiples** pour statuer sur la pertinence ou non de notre problématique grâce aux classes observées.

PARTIE III : CAH

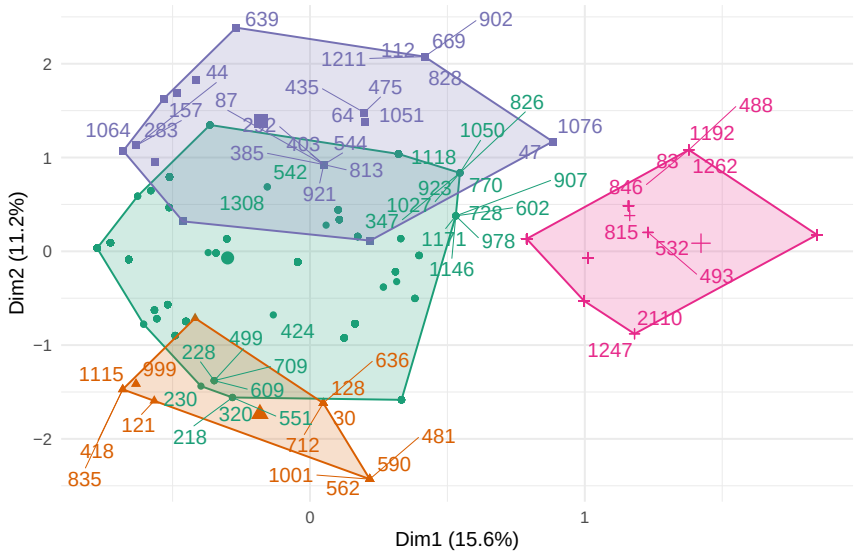
MESURE ET CHOIX DES CLASSES

- On aperçoit des sauts considérables dans les gains d'inertie interclasses à **2**, **4** et **8** classes.
- On souhaite **synthétiser de l'information** ⇒ 2 classes nous semble donc insuffisant, tandis que 8 classes semble trop important : 4 classes est un bon compromis.



- D'après le *dendogramme* ci-dessus, on obtient 4 **clusters** de taille $n = 1040$, 14, 56 et 228.

VISUALISATION DES CLUSTERS



Interprétation des clusters :

1. Le long de l'axe F_1 on distingue une classe qui se détache des autres (on avait montré dans l'ACM que ces individus prenaient les modalités **Other**) $\Rightarrow$ on retrouve donc bien cette distinction dans notre partition !
2. On distingue de plus 3 classes se chevauchant le long de l'axe F_2.
3. Les individus de la classe située en haut (bleue) doivent probablement prendre la modalité **Semi-Washed/Semi-Pulped**, tandis que les individus de la classe située en bas (orange) doivent prendre la modalité **Pulped Natural Honey** de la variable *Processing.Method*.

TEST DU KHI-DEUX

- On souhaite savoir quelle variable est la plus liée à la partition ci-dessus, on peut utiliser un test du χ^2 pour le découvrir :

Variables caractérisant la partition :		
	$p-value$	ddl
PM	0	12
Color	0	9
Variety	0	12
Species	0	3

- On s'aperçoit donc que la variable caractérisant le mieux la partition est *Processing.Method*.
- La variable la "moins liée" à la partition est *Species* (peut-être à cause du nombre d'individus trop faible pour **Robusta**).

MODALITES ASSOCIES

Modalités caractérisant le mieux le cluster 2 :			
	Cla/Mod	Mod/Cla	Global
PM=PM_Pulped natural / honey	100.00	100.00	1.05
Color=Color_Bluish-Green	3.51	28.57	8.52
PM=PM_Natural / Dry	0.00	0.00	19.28
Color=Color_Other	0.00	0.00	20.18
PM=PM_Washed / Wet	0.00	0.00	60.91

Modalités caractérisant le mieux le cluster 4 :			
	Cla/Mod	Mod/Cla	Global
PM=PM_Other	90.26	77.19	14.57
Color=Color_Other	65.56	77.63	20.18
Variety=Variety_Other	38.29	96.05	42.75
Species=Robusta	75.00	9.21	2.09
PM=PM_Semi-washed / Semi-pulped	0.00	0.00	4.19
PM=PM_Natural / Dry	8.14	9.21	19.28
Variety=Variety_Catuai	0.00	0.00	5.53
Color=Color_Bluish-Green	1.75	0.88	8.52
Species=Arabica	15.80	90.79	97.91
Variety=Variety_Typica	2.37	2.19	15.77
Variety=Variety_Caturra	1.18	1.32	19.06
Variety=Variety_Bourbon	0.44	0.44	16.89
PM=PM_Washed / Wet	3.80	13.60	60.91
Color=Color_Green	4.60	17.54	64.95

Interprétation des modalités (Cluster 2) :

1. 100 % des individus qui prennent la modalité **Pulped natural/honey** de la variable *Processing.Method* se retrouvent dans cette classe.
2. 100 % des individus de cette classe prennent cette modalité quand globalement, seulement 1 % des individus prennent cette modalité.
3. **En conclusion :** ce cluster permet de bien distinguer ce groupe d'individus extrêmes et les regroupe avec une grande précision.

Interprétation des modalités (Cluster 4) :

1. On remarque que les modalités les plus associées sont les modalités **Other** des 3 variables actives.
2. 90.26 % des individus qui prennent la modalité **Other** de la variable *Processing.Method* se retrouvent dans cette classe.
3. 77.19 % des individus de cette classe prennent cette modalité. Globalement, 14.57 % des individus de la population la prennent.
4. **En conclusion :** ce cluster permet de différencier le groupe d'individus **Other** et les regroupe plutôt bien.

PARANGONS

- On peut confirmer les **interprétations des clusters** avec les parangons : ce sont les individus qui caractérisent le mieux une partition (les plus proches du barycentre de la classe).

Tableau des parangons (Cluster 1) :

	Species	Variety	PM	Color
2	Arabica	Other	Washed / Wet	Green
5	Arabica	Other	Washed / Wet	Green
19	Arabica	Other	Washed / Wet	Green
23	Arabica	Other	Washed / Wet	Green
28	Arabica	Other	Washed / Wet	Green

Tableau des parangons (Cluster 2) :

	Species	Variety	PM	Color
30	Arabica	Other	Pulped natural / honey	Green
128	Arabica	Other	Pulped natural / honey	Green
636	Arabica	Other	Pulped natural / honey	Green
712	Arabica	Other	Pulped natural / honey	Green
999	Arabica	Bourbon	Pulped natural / honey	Green

Tableau des parangons (Cluster 3) :

	Species	Variety	PM	Color
87	Arabica	Other	Semi-washed / Semi-pulped	Green
292	Arabica	Other	Semi-washed / Semi-pulped	Green
385	Arabica	Other	Semi-washed / Semi-pulped	Green
403	Arabica	Other	Semi-washed / Semi-pulped	Green
544	Arabica	Other	Semi-washed / Semi-pulped	Green

Tableau des parangons (Cluster 4) :

	Species	Variety	PM	Color
9	Arabica	Other	Other	Other
8	Arabica	Other	Other	Other
14	Arabica	Other	Other	Other
15	Arabica	Other	Other	Other
17	Arabica	Other	Other	Other

Interprétation des parangons :

1. Pour les cluster 1, 2 et 3, les parangons (3 clusters confondus) prennent presque tous les mêmes modalités sauf pour la variable $Processing.Method$, qui change à chaque cluster.
2. On remarque que pour le cluster 4, les parangons prennent tous des modalités **Other** pour les 3 variables actives.

Conclusion :

La méthode de traitement influe effectivement sur la couleur du café (Les cafés **Other** dans $Processing.Method$ sont majoritairement **Other** dans $Color$). Cependant, $Variety$ ne semble pas influer (c'est dû au recodage de la variable et à l'effectif elevé de la modalité **Other**).