le cnam
Paris

PROJET STA101

Analyse de données : méthodes descriptives

Étude de qualité sur un échantillon de cafés

PLUSQUELLEC Valérie

Janvier 2025

TABLE DES MATIERES

I. INTRODUCTION

Cet écrit a pour objet l'étude d'un jeu de données collectées en 2023 par le *Coffee Quality Institute* (CQI), organisation internationale à but non lucratif, qui œuvre pour améliorer la qualité du café de spécialité et la vie des personnes qui le produisent. L'une des missions du CQI est de maintenir une base de données qui sert de ressource pour les professionnels et les passionnés, et qui contient une variété d'informations sur la production, la transformation et l'évaluation sensorielle du café.

L'étude présentée dans ce document porte sur 207 individus décrits par 21 variables (extrait en annexe A). Un individu est défini comme étant un café produit par une ferme agricole identifiée. Parmi les variables :

- 2 portent sur le lieu de plantation et répertorient, pour chaque ferme agricole, le pays et l'altitude ;
- 5 informent sur le café récolté et les méthodes de préparation des grains, en précisant la masse produite, la variété, la couleur du fruit cueilli, la méthode de traitement utilisée pour extraire le grain du fruit (*Processing.Method*), et le taux d'humidité du grain avant torréfaction ;
- 10 correspondent à l'évaluation sensorielle selon les critères : arôme, saveur, arrière-goût, acidité, corps, équilibre, uniformité, propreté de la tasse, douceur, impression générale (annexe B) ;
- 1 variable, appelée score total, est obtenue en sommant les 10 variables sensorielles ;
- 3 variables évaluent la présence de défauts dans une quantité définie de grains de café : nombre de défauts de catégorie 1, nombre de défauts de catégorie 2, nombre de quakers (annexe C).

On dénombre parmi ces 21 variables : 17 variables de nature quantitative, et 4 variables de nature qualitative (avec un nombre de modalités pouvant aller de 6 à 24 modalités selon la variable).

L'objectif principal de cette étude est de répondre au mieux à la problématique suivante : peut-on affirmer que certaines notes sensorielles sont corrélées avec les conditions géographiques des plantations, ou avec d'autres paramètres de production mis en place par les fermes de café ?

Dans un premier temps, l'étude débutera par une analyse univariée et bivariée. Cette partie aura pour intérêt de déterminer si un traitement préalable des données est nécessaire, et si des corrélations peuvent être déjà observées entre deux variables. A l'issue de cette partie, un choix concernant les variables actives et illustratives sera proposé pour la suite de l'étude.

Dans un second temps, une classification des cafés sera réalisée à partir des variables actives, afin de mettre en évidence l'existence de groupes de cafés similaires et les éléments de similitude des cafés d'un même groupe.

Enfin, une analyse factorielle viendra compléter l'étude pour explorer les critères sur lesquels se reposent la classification précédemment obtenue.

II. ANALYSES PRÉLIMINAIRES

A. ANALYSE UNIVARIÉE ET PRÉPARATION DES DONNÉES

Une première prise de contact avec les données consiste à étudier un résumé d'indicateurs statistiques (figure 2.1), étape qui permet d'observer rapidement si des variables quantitatives ne sont pas à supprimer du jeu de données, comme des doublons ou des variables constantes.
Trois suppressions peuvent être effectuées pour les raisons suivantes :

- la première colonne du jeu de données est composée des numéros des individus et ne correspond pas à une variable ;
- les variables Douceur et Propreté de la tasse sont constantes et égales à 10.

Statistic	N	Min	Pctl(25)	Median	Mean	Pctl(75)	Max	St. Dev.
X	207	1	52.5	104	104.000	155.5	207	59.900
Altitude	205	139	1,000	1,300	1,293.380	1,600	5,400	668.167
Weight.kg	207	1	30	300	66,857.960	18,000	6,144,000	601,852.400
Aroma	207	6.500	7.580	7.670	7.721	7.920	8.580	0.288
Flavor	207	6.750	7.580	7.750	7.745	7.920	8.500	0.280
Aftertaste	207	6.670	7.420	7.580	7.600	7.750	8.420	0.276
Acidity	207	6.830	7.500	7.670	7.690	7.875	8.580	0.260
Body	207	6.830	7.500	7.670	7.641	7.750	8.250	0.233
Balance	207	6.670	7.500	7.670	7.644	7.790	8.420	0.256
Uniformity	207	8.670	10.000	10.000	9.990	10.000	10.000	0.103
Clean.Cup	207	10	10	10	10.000	10	10	0.000
Sweetness	207	10	10	10	10.000	10	10	0.000
Overall	207	6.670	7.500	7.670	7.677	7.920	8.580	0.306
Total.Cup.Points	207	78.000	82.580	83.750	83.707	84.830	89.330	1.730
Moisture.Percentage	207	8.100	10.100	10.800	10.784	11.500	13.500	0.999
Category.One.Defects	207	0	0	0	0.135	0	5	0.592
Quakers	207	0	0	0	0.691	1	12	1.687
Category.Two.Defects	207	0	0	1	2.251	3	16	2.950

Figure 2.1 : Indicateurs statistiques des variables quantitatives.

Country.of.Origin	Altitude	Weight.kg	Variety
0	2	0	19
Processing.Method	Aroma	Flavor	Aftertaste
5	0	0	0
Acidity	Body	Balance	Uniformity
0	0	0	0
Clean.Cup	Sweetness	Overall	Total.Cup.Points
0	0	0	0
Moisture.Percentage	Category.One.Defects	Quakers	Color
0	0	0	0
Category.Two.Defects			
0			

Figure 2.2 : Tableau des valeurs manquantes dans le jeu de données.

On s'intéresse désormais à la présence de données manquantes. On constate sur la figure 2.2, qu'en plus des 2 valeurs manquantes pour la variable quantitative Altitude, il en manque 19 et 5 respectivement pour les variables qualitatives Variété et Méthode de Traitement.
Le nombre total de valeurs manquantes n'étant pas négligeable par rapport au nombre total d'individus, on ne peut pas envisager de supprimer les individus concernés (qui pourraient représenter de 9% à 12% de l'ensemble des individus), sans affecter fortement la qualité du jeu de données. Pour combler la base de données, une imputation a été effectuée. Le choix de l'analyse factorielle utilisée pour l'imputation est expliqué dans le paragraphe III.A.

Une observation que l'on peut aussi faire à partir du résumé (figure 2.1) est qu'il y a, parmi les variables quantitatives, 3 variables discrètes avec comme valeur prépondérante zéro. Une transformation logarithmique semble nécessaire pour réduire l'effet d'asymétrie et stabiliser la variance. La préparation des variables quantitatives est finalisée par une standardisation, étape nécessaire pour la suite de l'analyse car les données sont exprimées dans différentes unités.

Statistic	N	Min	Pctl(25)	Median	Mean	Pctl(75)	Max	St. Dev.
Altitude	207	-1.738	-0.417	0.063	-0.000	0.454	6.154	1.000
Weight.kg	207	-0.111	-0.111	-0.111	0.000	-0.081	10.097	1.000
Aroma	207	-4.245	-0.490	-0.178	-0.000	0.692	2.986	1.000
Flavor	207	-3.558	-0.589	0.019	-0.000	0.627	2.701	1.000
Aftertaste	207	-3.370	-0.652	-0.072	-0.000	0.545	2.973	1.000
Acidity	207	-3.315	-0.733	-0.078	-0.000	0.712	3.428	1.000
Body	207	-3.473	-0.604	0.125	0.000	0.467	2.608	1.000
Balance	207	-3.800	-0.562	0.101	0.000	0.569	3.027	1.000
Uniformity	207	-12.781	0.094	0.094	0.000	0.094	0.094	1.000
Overall	207	-3.286	-0.577	-0.022	0.000	0.794	2.948	1.000
Total.Cup.Points	207	-3.298	-0.651	0.025	0.000	0.649	3.250	1.000
Moisture.Percentage	207	-2.688	-0.685	0.016	0.000	0.717	2.720	1.000
Category.One.Defects_log	207	-0.253	-0.253	-0.253	-0.000	-0.253	6.251	1.000
Quakers_log	207	-0.545	-0.545	-0.545	-0.000	0.694	4.040	1.000
Category.Two.Defects_log	207	-1.098	-1.098	-0.210	-0.000	0.678	2.531	1.000

Figure 2.3 : Indicateurs statistiques des variables quantitatives, centrées et réduites.

Pour les variables qualitatives, les fréquences des modalités ont été calculées (annexe D). Certaines modalités rares pourraient être regroupées, mais leur nature ne permet pas que cela soit envisagé. Par exemple, il ne semble pas pertinent de regrouper Madagascar ou Myanmar avec un autre pays, sans dégrader la qualité des données.

Enfin, une analyse univariée graphique, composée de boîtes à moustaches sur les variables quantitatives continues, et de diagrammes en barre pour les variables qualitatives et quantitatives discrètes, est disponible en annexe (annexes E et F). Elle permet de remarquer que la variable Uniformité est constante sauf pour deux individus. On pourra donc considérer cette variable comme illustrative pour la suite des analyses.

B. ANALYSE BIVARIÉE

Dans ce paragraphe, on s'intéresse aux liaisons entre les variables deux à deux. On peut observer, dans la figure 2.4, une représentation graphique des coefficients de corrélation de Pearson. Cette figure démontre un lien linéaire entre sept des variables sensorielles deux à deux (toutes les variables sensorielles restantes sauf la variable Uniformité). On ne soulignera pas le lien avec la variable Score total, avec qui le lien est trivial puisque cette variable est construite à partir des variables sensorielles.

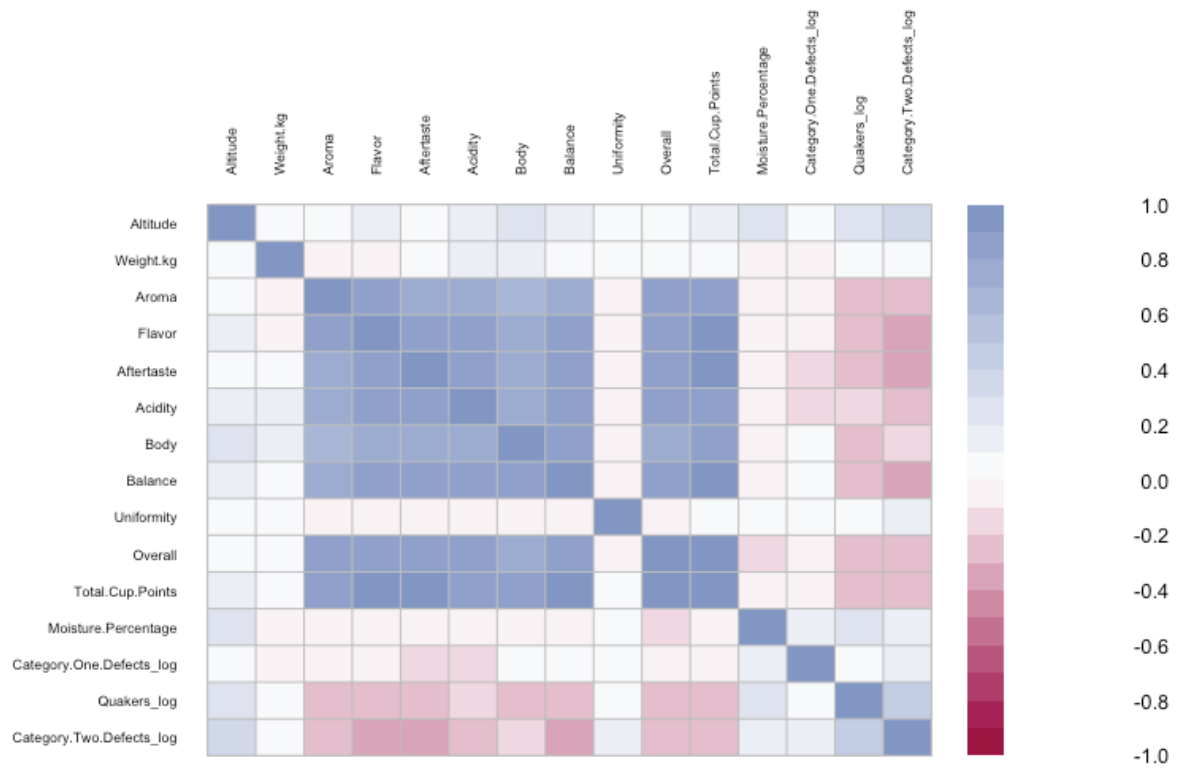


Figure 2.4 : Représentation graphique des coefficients de corrélation de Pearson.

De même, cette figure fait apparaître des coefficients de corrélation négatifs entre chacune de ces 7 variables sensorielles et chacune des 3 variables liées aux défauts.

On notera qu'une recherche des liaisons non linéaires, avec les coefficients de corrélation de Spearman, a été faite en annexe G, mais qu'elle ne fait pas apparaître d'autre liaisons particulières entre deux variables quantitatives.

Il serait intéressant d'approfondir l'étude des liaisons entre l'ensemble des variables regroupant les variables sensorielles, et celles des défauts. Cette étude est brièvement faite en annexe K, sous la forme d'une Analyse en Composantes Principales (ACP) sur les 10 variables quantitatives que l'on vient de citer. Elle met bien en lumière que les 7 variables sensorielles restantes sont fortement liées entre elles, et qu'elles sont indépendantes dans une moindre mesure des variables liées aux défauts.

On peut enfin observer des corrélations entre deux variables qualitatives (annexe H) et entre une variable qualitative et une variable quantitative (annexe I). En particulier, et dans le but de répondre à la problématique, la répartition du score total selon une des variables qualitatives est présenté ci-dessous:

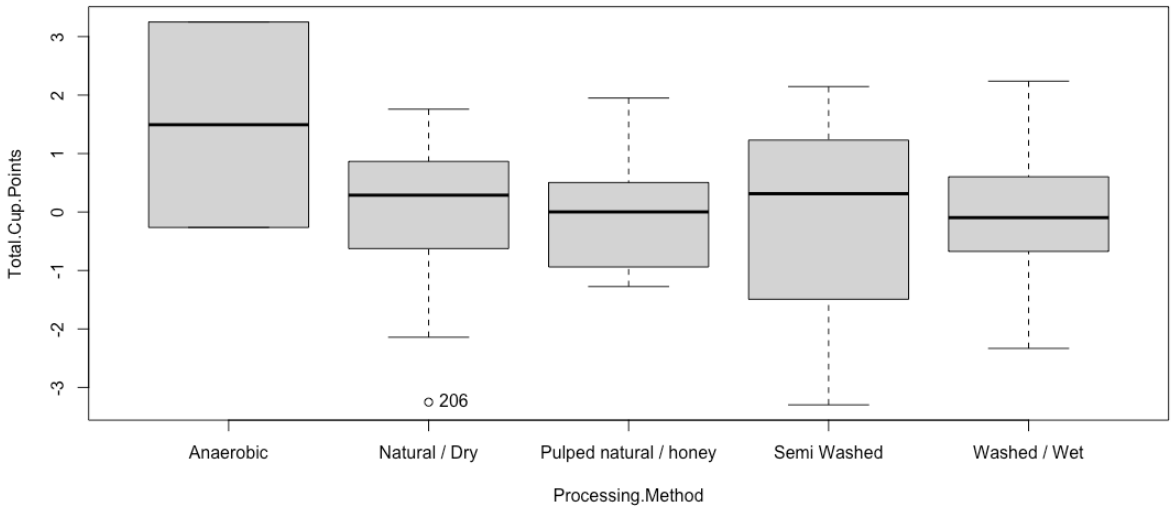


Figure 2.5 : Boites à moustaches de la variable Score total suivant les modalités de la variable Méthode de Traitement.

On peut, à partir de la figure 2.5, affirmer que le score total n'est pas réparti de la même façon suivant la méthode de traitement employée sur les fruits. D'autres représentation mettant en lien le score total et les autres variables qualitatives sont disponibles en annexe J.

C. CHOIX DES VARIABLES ACTIVES ET ILLUSTRATIVES

Au regard de cette première partie et de la problématique, il apparaît pertinent de sélectionner comme variables actives celles relatives aux conditions géographiques des plantations, ainsi qu'aux paramètres de production mis en place. Ce choix permettra d'établir une typologie des cafés fondée exclusivement sur ces éléments, sans tenir compte de leur évaluation sensorielle ni de leurs défauts. Les variables sensorielles et celles liées aux défauts seront quant à elles considérées comme illustratives, ce qui facilitera l'estimation de leurs corrélations avec la typologie obtenue.

Un résumé des variables actives et illustratives est disponible ci-dessous.

Variables actives	Nature
Country.of.Origin	Qualitative
Variety	Qualitative
Processing.Method	Qualitative
Color	Qualitative
Altitude	Quantitative
Weight.kg	Quantitative
Moisture.Percentage	Quantitative

Variables illustratives	Nature
Aroma	Quantitative
Flavor	Quantitative
Aftertaste	Quantitative
Acidity	Quantitative
Body	Quantitative
Balance	Quantitative
Uniformity	Quantitative
Overall	Quantitative
Total.Cup.Points	Quantitative
Category.One.Defects_log	Quantitative
Quakers_log	Quantitative
Category.Two.Defects_log	Quantitative

Figure 2.7 : Tableaux des variables actives et illustratives, avec leur nature.

III. CLASSIFICATION

A. CHOIX DE LA MÉTHODE

On veut, dans cette partie, établir des classes d'individus selon les variables actives. Cette classification devra regrouper, dans une même classe, des individus relativement proches, ce qui implique une notion de distance que l'on doit définir.

On procèdera à une classification par approche tandem. Cette méthode consiste à regrouper, en utilisant la distance euclidienne, les projetés des individus dans les plans principaux obtenus à partir d'une analyse factorielle. L'intérêt de l'approche tandem est de nettoyer les données en laissant de côté les derniers axes correspondant à ce que l'on nomme « du bruit ».

Les variables actives étant à la fois qualitatives et quantitatives, l'Analyse Factorielle des Données Mixtes sera privilégiée. La métrique utilisée pour l'AFDM combine la distance euclidienne normée pour les variables quantitatives (comme en ACP normée), et la distance du khi-deux, pour les variables qualitatives (comme en ACM), pondérée pour assurer l'équilibre entre les variables de natures différentes.

On remarquera que c'est justement à partir de l'AFDM que les données manquantes ont été imputées dans la partie II.A. (*)

La classification se fera en deux étapes : une Classification Ascendante Hiérarchique (CAH), selon le critère d'agrégation de Ward, suivie d'une consolidation par les centres mobiles.

(*) Le logiciel R a utilisé un algorithme itératif régularisé appliqué à la matrice de corrélations (pour les variables quantitatives) concaténée avec le tableau disjonctif complet (pour les variables qualitatives). L'algorithme repose sur une initialisation des valeurs manquantes par la moyenne de la variable ou par la proportion de chaque modalité, selon la nature de la variable, puis sur deux étapes (estimation des paramètres via l'AFDM et imputation des valeurs manquantes à l'aide de la matrice ajustée) qui sont répétées jusqu'à convergence.

B. CHOIX DU NOMBRE DE CLASSES

La CAH permet de déterminer le nombre de classes qu'il faut choisir. Pour cela, on observe les valeurs propres obtenues par l'AFDM, et leur inerties (annexes L et M).

On constate que l'on peut conserver plus de 90% de l'inertie du nuage de données avec les 39 premiers axes. On procède donc à une CAH sur les 39 premiers axes, selon la méthode de Ward. Cette approche ascendante consiste, à chaque pas, à regrouper deux éléments (individus isolés ou classes déjà formées) en minimisant la variabilité au sein de chaque groupe (l'inertie intra-classe), et en maximisant les différences entre les groupes (inertie inter-classes)

Le diagramme des gains d'inertie (figure 3.1) indique que le premier saut d'inertie, en diminuant le nombre d'axes, est celui qui permet de passer d'une partition de 5 classes à 4 classes. On retient donc, selon ce critère, 5 classes pour la partition.

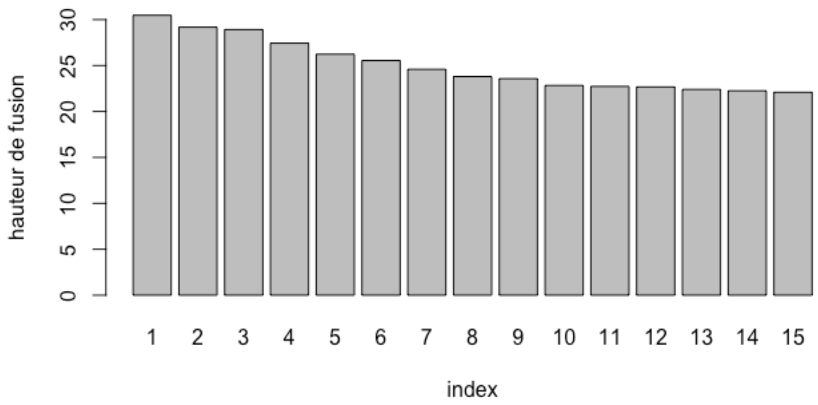


Figure 3.1 : Diagramme des gains d'inertie inter-classes.

On remarque que le nombre de classes aurait également pu être fixé à 7, pour des raisons similaires à celles mentionnées précédemment, bien que l'écart d'inertie à franchir soit légèrement moindre. Cette option a été envisagée, mais écartée, les résultats s'étant révélés moins utiles pour répondre à la problématique de l'étude.

C. PROFIL DES CLASSES

On procède alors à une consolidation des 5 classes obtenues par la CAH en les introduisant comme partition initiale de l'algorithme des centres mobiles. Cet algorithme calcule le centre de chaque groupe et réassigne ensuite les individus aux groupes en fonction de leur proximité avec les centres obtenus. Ce processus est répété jusqu'à ce que la répartition des individus soit stable.

Les effectifs des classes consolidées sont les suivants :

- 106 pour la classe 1, soit environ 51,2 % des individus ;
- 67 pour la classe 2, soit environ 32,4 % des individus ;
- 21 pour la classe 3, soit environ 10,1 % des individus ;
- 10 pour la classe 4, soit environ 4,8 % des individus ;
- 3 pour la classe 5, soit environ 1,4 % des individus.

```
Link between the cluster variable and the categorical variables (chi-square test)
================================================================================
                          p.value df
Country.of.Origin 6.106874e-114 84
Variety            1.176741e-90 92
Color              1.973887e-03 24
Processing.Method  2.029356e-03 16

Link between the cluster variable and the quantitative variables
================================================================
                                Eta2      P-value
Altitude                  0.34477092 1.022957e-17
Category.Two.Defects_log  0.23155146 6.827262e-11
Moisture.Percentage       0.22251736 2.139700e-10
Weight.kg                 0.19355900 7.527188e-09
Quakers_log               0.16516120 2.134907e-07
Acidity                   0.13049090 1.043159e-05
Overall                   0.10544336 1.509072e-04
Aroma                     0.09970624 2.735073e-04
Aftertaste                0.09887839 2.978459e-04
Flavor                    0.09527243 4.310502e-04
Total.Cup.Points          0.09475112 4.546079e-04
Balance                   0.07193619 4.391710e-03
```

Figure 3.2 : Tableaux des corrélations avec la partition.

D'après la figure 3.2, la partition obtenue est corrélée fortement avec les variables qualitatives Pays d'origine, Variété et la variable quantitative Altitude. Dans une moindre mesure, elle est aussi associée aux variables qualitatives Couleur et Méthode de Traitement, et aux variables quantitatives Taux d'humidité, Masse, Défauts de catégorie 2, Quakers, et les 6 notes sensorielles : Acidité, Impression générale, Après-goût, Arôme, Saveur et Équilibre.

Classe 1 (106 cafés, 51,2 % des individus) :

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Variety=Caturra	96.666667	27.3584906	14.492754	7.596615e-09	5.777173
Country.of.Origin=Guatemala	100.000000	19.8113208	10.144928	2.691592e-07	5.143854
Country.of.Origin=Colombia	100.000000	17.9245283	9.178744	1.264710e-06	4.845221
Variety=Bourbon	90.909091	18.8679245	10.628019	4.593517e-05	4.075401
Processing.Method=Washed / Wet	60.937500	73.5849057	61.835749	3.900060e-04	3.546756
Variety=Catuai	92.857143	12.2641509	6.763285	9.371109e-04	3.308757
Country.of.Origin=Brazil	100.000000	9.4339623	4.830918	9.979458e-04	3.291105
Country.of.Origin=Honduras	92.307692	11.3207547	6.280193	1.814369e-03	3.119047
Country.of.Origin=Costa Rica	100.000000	7.5471698	3.864734	4.136447e-03	2.867564
Country.of.Origin=Nicaragua	100.000000	6.6037736	3.381643	8.356458e-03	2.637318
Country.of.Origin=El Salvador	100.000000	6.6037736	3.381643	8.356458e-03	2.637318
Country.of.Origin=Tanzania, United Republic Of	100.000000	5.6603774	2.898551	1.679648e-02	2.391133
Color=greenish	66.666667	22.6415094	17.391304	4.330754e-02	2.020732
Country.of.Origin=United States (Hawaii)	0.000000	0.0000000	2.415459	2.625649e-02	-2.222397
Color=yellow-green	23.529412	3.7735849	8.212560	1.901018e-02	-2.345331
Color=blue-green	16.666667	1.8867925	5.797101	1.531542e-02	-2.424833
Country.of.Origin=Thailand	16.666667	1.8867925	5.797101	1.531542e-02	-2.424833
Variety=Gesha	27.586207	7.5471698	14.009662	6.510617e-03	-2.720891
Country.of.Origin=Ethiopia	9.090909	0.9433962	5.314010	4.118681e-03	-2.868926
Variety=SL34	0.000000	0.0000000	3.864734	2.771928e-03	-2.991960
Processing.Method=Pulped natural / honey	23.076923	5.6603774	12.560386	2.238378e-03	-3.056633
Variety=Ethiopian Heirlooms	0.000000	0.0000000	4.347826	1.295424e-03	-3.216992
Variety=Catimor	0.000000	0.0000000	5.314010	2.780415e-04	-3.634944
Variety=Typica	8.333333	2.8301887	17.391304	3.765357e-09	-5.894185
Country.of.Origin=Taiwan	0.000000	0.0000000	29.468599	1.171017e-25	-10.471238

$`1`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Altitude	5.309318	0.3602146	-3.159040e-16	0.7829717	0.9975816	1.100361e-07
Moisture.Percentage	4.406287	0.2989478	3.126860e-15	0.8597398	0.9975816	1.051577e-05
Category.Two.Defects_log	4.192209	0.2844235	-1.327441e-16	0.9168382	0.9975816	2.762511e-05
Quakers_log	3.672851	0.2491873	7.562389e-17	1.1155830	0.9975816	2.398592e-04
Acidity	-2.491941	-0.1690676	-6.423203e-15	1.0517263	0.9975816	1.270472e-02
Total.Cup.Points	-3.033169	-0.2057876	2.763222e-15	1.0602999	0.9975816	2.420001e-03
Balance	-3.065377	-0.2079728	3.205166e-15	1.0657319	0.9975816	2.173958e-03
Aroma	-3.153315	-0.2139390	-1.699338e-14	0.9938578	0.9975816	1.614277e-03
Flavor	-3.198523	-0.2170062	-6.800786e-16	1.0284302	0.9975816	1.381335e-03
Overall	-3.250701	-0.2205462	9.109192e-15	1.0566646	0.9975816	1.151208e-03
Aftertaste	-3.565271	-0.2418885	-8.568562e-15	1.0728540	0.9975816	3.634799e-04

Figure 3.3 : Tableaux de description pour les variables quantitatives et qualitatives de la classe 1.

La classe 1 est composée de cafés principalement de provenance du Guatemala, de Colombie, du Brésil et du Honduras.

Les variétés dominantes de cette classe sont les variétés Caturra, Bourbon et Catuai, et la méthode de traitement pour extraire les grains est majoritairement *Washed/Wet* (73 % de la classe contre une fréquence de 61% pour tous les individus).

Ces cafés sont issus de régions situées à une altitude haute (la moyenne de l'altitude de cette classe est plus élevée que la moyenne globale), et le taux d'humidité contenu dans les grains est plus important que ce qui est généralement observé.

Enfin, les cafés de cette classe reçoivent des scores inférieurs sur les 6 critères sensoriels (Acidité, Impression générale, Après-goût, Arôme, Saveur et Équilibre) et présentent plus de défauts de catégorie 2 et de quakers que l'ensemble des cafés.

Classe 2 (67 cafés, 32,4 % des individus) :

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Country.of.Origin=Taiwan	100.000000	91.044776	29.468599	5.134517e-46	14.240518
Variety=Typica	88.888889	47.761194	17.391304	8.707438e-15	7.756836
Variety=Gesha	68.965517	29.850746	14.009662	1.627083e-05	4.310741
Variety=SL34	100.000000	11.940299	3.864734	8.946030e-05	3.917532
Processing.Method=Pulped natural / honey	61.538462	23.880597	12.560386	1.272336e-03	3.222148
Color=blue-green	75.000000	13.432836	5.797101	2.606380e-03	3.010710
Country.of.Origin=United States (Hawaii)	100.000000	7.462687	2.415459	3.201363e-03	2.947711
Country.of.Origin=Honduras	7.692308	1.492537	6.280193	4.511319e-02	-2.003597
Country.of.Origin=Costa Rica	0.000000	0.000000	3.864734	4.095214e-02	-2.044014
Variety=Blend	11.111111	2.985075	8.695652	3.952691e-02	-2.058659
Variety=Catuai	7.142857	1.492537	6.763285	3.148213e-02	-2.150925
Variety=Ethiopian Heirlooms	0.000000	0.000000	4.347826	2.716423e-02	-2.209150
Country.of.Origin=Brazil	0.000000	0.000000	4.830918	1.797229e-02	-2.366188
Variety=Catimor	0.000000	0.000000	5.314010	1.185989e-02	-2.516286
Country.of.Origin=Ethiopia	0.000000	0.000000	5.314010	1.185989e-02	-2.516286
Processing.Method=Washed / Wet	25.781250	49.253731	61.835749	1.122479e-02	-2.535622
Variety=Bourbon	9.090909	2.985075	10.628019	1.004686e-02	-2.574212
Country.of.Origin=Thailand	0.000000	0.000000	5.797101	7.805744e-03	-2.660359
Country.of.Origin=Colombia	0.000000	0.000000	9.178744	3.867689e-04	-3.548951
Country.of.Origin=Guatemala	0.000000	0.000000	10.144928	1.597418e-04	-3.775414
Variety=Caturra	3.333333	1.492537	14.492754	5.150338e-05	-4.048698

$`2`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Flavor	3.927677	0.3946183	-6.800786e-16	0.8020466	0.9975816	8.577013e-05
Aroma	3.883319	0.3901616	-1.699338e-14	0.9095490	0.9975816	1.030403e-04
Aftertaste	3.858365	0.3876545	-8.568562e-15	0.7808972	0.9975816	1.141479e-04
Overall	3.808231	0.3826174	9.109192e-15	0.7881721	0.9975816	1.399647e-04
Total.Cup.Points	3.372115	0.3388003	2.763222e-15	0.8117499	0.9975816	7.459341e-04
Balance	3.250525	0.3265840	3.205166e-15	0.8305733	0.9975816	1.151922e-03
Acidity	2.536237	0.2548187	-6.423203e-15	0.7373126	0.9975816	1.120508e-02
Quakers_log	-5.052508	-0.5076313	7.562389e-17	0.3004709	0.9975816	4.360470e-07
Category.Two.Defects_log	-6.670182	-0.6701610	-1.327441e-16	0.6978431	0.9975816	2.554862e-11
Moisture.Percentage	-6.679986	-0.6711460	3.126860e-15	0.9209648	0.9975816	2.389648e-11
Altitude	-8.119676	-0.8157933	-3.159040e-16	0.6580811	0.9975816	4.674289e-16

Figure 3.4 : Tableaux de description pour les variables quantitatives et qualitatives de la classe 2.

La classe 2 regroupe des cafés principalement originaires de Taiwan et de Hawaï.

Les variétés majoritaires sont les variétés Typica, Gesha et SL34, et la méthode de traitement *Pulped Natural/Honey* est plus présente que dans l'ensemble des individus.

Ces cafés sont cultivés à basse altitude (l'altitude moyenne de la classe est nettement inférieure à celle de tous les individus), et le taux d'humidité contenu dans les grains est faible.

Les cafés de cette classe se distinguent par des notes modérément supérieures aux autres cafés, sur les 6 critères sensoriels (Acidité, Impression générale, Après-goût, Arôme, Saveur et Équilibre), et des défauts (de catégorie 2 et quakers) peu présents.

Classe 3 (21 cafés, 10,1 % des individus) :

```
$`3`
                                Cla/Mod  Mod/Cla    Global      p.value    v.test
Variety=Catimor               100.00000 52.38095  5.314010 6.170230e-13  7.196667
Country.of.Origin=Thailand     83.33333 47.61905  5.797101 6.640678e-10  6.174420
Country.of.Origin=Vietnam     100.00000 19.04762  1.932367 8.054782e-05  3.942764
Variety=Java                  100.00000 14.28571  1.449275 9.128753e-04  3.316087
Country.of.Origin=Laos        100.00000 14.28571  1.449275 9.128753e-04  3.316087
Country.of.Origin=Indonesia   100.00000 14.28571  1.449275 9.128753e-04  3.316087
Color=yellowish                40.00000 19.04762  4.830918 1.235264e-02  2.501907
Variety=Blend                  27.77778 23.80952  8.695652 2.795394e-02  2.197932
Processing.Method=Semi Washed  66.66667  9.52381  1.449275 2.863546e-02  2.188469
Variety=Gesha                   0.00000  0.00000 14.009662 3.516293e-02 -2.106477
Variety=Caturra                 0.00000  0.00000 14.492754 3.101450e-02 -2.156887
Processing.Method=Washed / Wet  6.25000 38.09524 61.835749 2.312875e-02 -2.271301
Color=greenish                  0.00000  0.00000 17.391304 1.436832e-02 -2.447921
Country.of.Origin=Taiwan        0.00000  0.00000 29.468599 4.128405e-04 -3.531737

$`3`
                     v.test Mean in category  Overall mean sd in category Overall sd    p.value
Moisture.Percentage  2.454211        0.5076607  3.126860e-15      0.8855271  0.9975816 0.01411941
Aroma               -1.970764       -0.4076582 -1.699338e-14      0.9887083  0.9975816 0.04875091
Overall             -1.986062       -0.4108227  9.109192e-15      1.0366719  0.9975816 0.04702644
Acidity             -2.276334       -0.4708663 -6.423203e-15      0.9948488  0.9975816 0.02282603
```

Figure 3.5 : Tableaux de description pour les variables quantitatives et qualitatives de la classe 3.

La classe 3 regroupe des cafés principalement originaires de la Thaïlande, du Vietnam, du Laos et d'Indonesie.

La variété majoritaire est Catimor, avec la présence de la variété Java qui se trouve exclusivement dans cette classe. La méthode de traitement *Semi/Washed* est plus présente dans cette classe que dans l'ensemble des individus.

Le taux d'humidité dans ces cafés est plus important que dans l'ensemble des individus.

Les cafés de cette classe présentent enfin des notes inférieures aux autres cafés, sur les 3 critères sensoriels : Acidité, Impression générale et Arôme.

Classe 4 (10 cafés, 4,8 % des individus) :

```
$`4`
                              Cla/Mod Mod/Cla    Global      p.value    v.test
Country.of.Origin=Ethiopia   90.90909     100  5.314010 3.446215e-16  8.156588
Variety=Ethiopian Heirlooms 100.00000      90  4.347826 6.203187e-15  7.799746
Color=yellow-green           17.64706      30  8.212560 4.358511e-02  2.018059
Country.of.Origin=Taiwan      0.00000       0 29.468599 2.772401e-02 -2.201170

$`4`
                           v.test Mean in category  Overall mean sd in category Overall sd      p.value
Weight.kg                6.312998        1.9475275  3.016910e-17      4.0749508  0.9975816 2.736810e-10
Acidity                  3.881084        1.1972945 -6.423203e-15      0.6358599  0.9975816 1.039919e-04
Altitude                 3.246990        1.0016798 -3.159040e-16      0.3034326  0.9975816 1.166323e-03
Quakers_log              3.122421        0.9632508  7.562389e-17      0.9899251  0.9975816 1.793704e-03
Category.Two.Defects_log 2.724986        0.8406441 -1.327441e-16      0.8961830  0.9975816 6.430430e-03
Body                     2.569394        0.7926450  1.485661e-14      0.8060520  0.9975816 1.018764e-02
Total.Cup.Points         2.140084        0.6602050  2.763222e-15      0.5265350  0.9975816 3.234796e-02
```

Figure 3.6 : Tableaux de description pour les variables quantitatives et qualitatives de la classe 4.

La classe 4 regroupe des cafés exclusivement originaires d'Éthiopie, appartenant en majorité à la variété Ethiopian Heirlooms, et cultivés à très haute altitude.

Ces cafés se distinguent par une quantité produite importante.

Ils présentent plus de défauts de catégorie 2 et de quakers, mais obtiennent des notes nettement supérieures en Acidité et en Corps, ce qui en font des cafés avec de très bons scores totaux.

Classe 5 (3 cafés, 1.4 % des individus) :

```
$`5`
                         Cla/Mod Mod/Cla   Global      p.value  v.test
Variety=SL14                 100     100 1.449275 6.863724e-07 4.96518
Country.of.Origin=Uganda     100     100 1.449275 6.863724e-07 4.96518
```

Figure 3.7 : Tableau de description pour les variables qualitatives de la classe 5.

ID	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Overall	Total.Cup. Points	Category. One.Defe cts_log	Quakers_ log	Category. Two.Defec ts_log
44	0,970	0,627	0,545	0,230	0,467	1,077	0,794	0,747	-0,253	-0,545	-0,210
119	0,101	0,019	-0,362	-0,733	-0,261	-0,250	0,239	-0,166	-0,253	-0,545	0,678
121	0,101	-0,267	-0,072	-0,733	0,125	-0,562	-0,022	-0,218	-0,253	-0,545	1,197

Figure 3.8 : Tableau des notes sensorielles et défauts des individus de la classe 5.

La classe 5 est composée des cafés d'Ouganda, exclusivement de la variété SL14. Les 3 cafés qui composent cette classe (individus 44, 119 et 121 dans la figure 3.8) obtiennent un score dans la moyenne haute en arôme, et ne présentent aucun quaker.

Bilan de la classification :

Sur la base des observations que l'on vient de faire, on peut constater que même si la partition n'a pas été faite à partir des notes sensorielles et des variables liées aux défauts, des liens entre ces variables et les classes obtenues peuvent être établis.

L'observation de ces classes, dans les premiers plans de l'analyse factorielle, pourra apporter des détails à propos des critères sur lesquels se reposent ces regroupements.

IV. ANALYSE FACTORIELLE

Comme expliqué dans la partie III.A, le choix de l'analyse factorielle s'est porté sur l'AFDM. Cette analyse permet de visualiser les individus dans les plans principaux à partir de variables à la fois quantitatives et qualitatives. Comme expliqué dans l'annexe N, le nombre d'axes non triviaux est conséquent et s'élève à 57. Il en résulte que les inerties des premiers axes sont faibles (contrairement aux inerties obtenues par l'ACP), et que l'interprétation des axes devra être consolidée par l'observation des données brutes.

A. CHOIX DU NOMBRE D'AXES

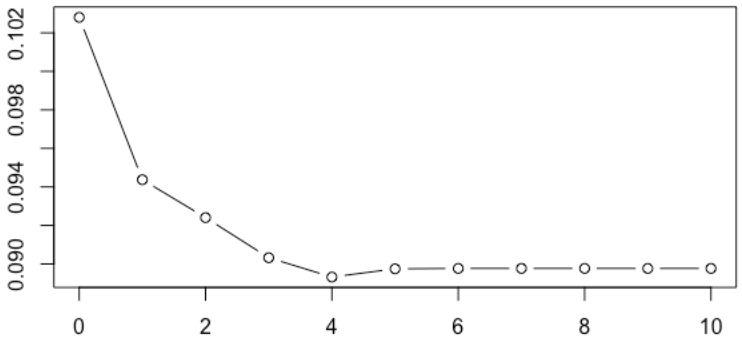


Figure 4.1 : Graphique représentant l'erreur de validation croisée en fonction du nombre d'axes.

La validation croisée est utilisée ici pour obtenir le nombre d'axes factoriels. Cette technique consiste à diviser les données en deux parties : une partie d'entraînement qui sert à construire le modèle, et une partie de test qui permet d'évaluer la qualité du modèle.

A l'issue de 40 simulations, l'erreur mesurée entre le modèle et les valeurs de test est représentée en figure 4.1. Sur ce graphique, l'erreur est minimale pour 4 axes. Ces 4 axes représentent environ 17% d'inertie cumulée (annexe I).

B. PREMIER PLAN FACTORIEL (9,1% D'INERTIE)

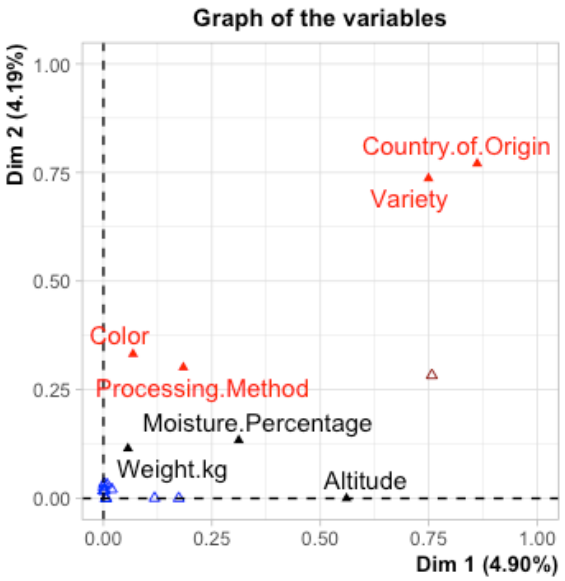


Figure 4.2 : Graphe des variables qui contribuent le plus.

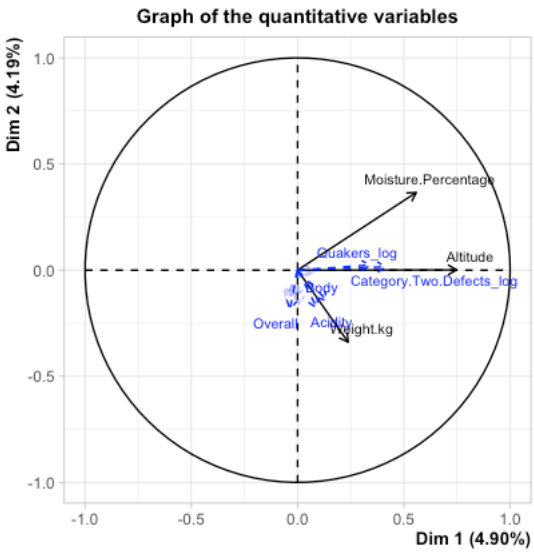


Figure 4.3 : Cercle des corrélations des variables quantitatives.

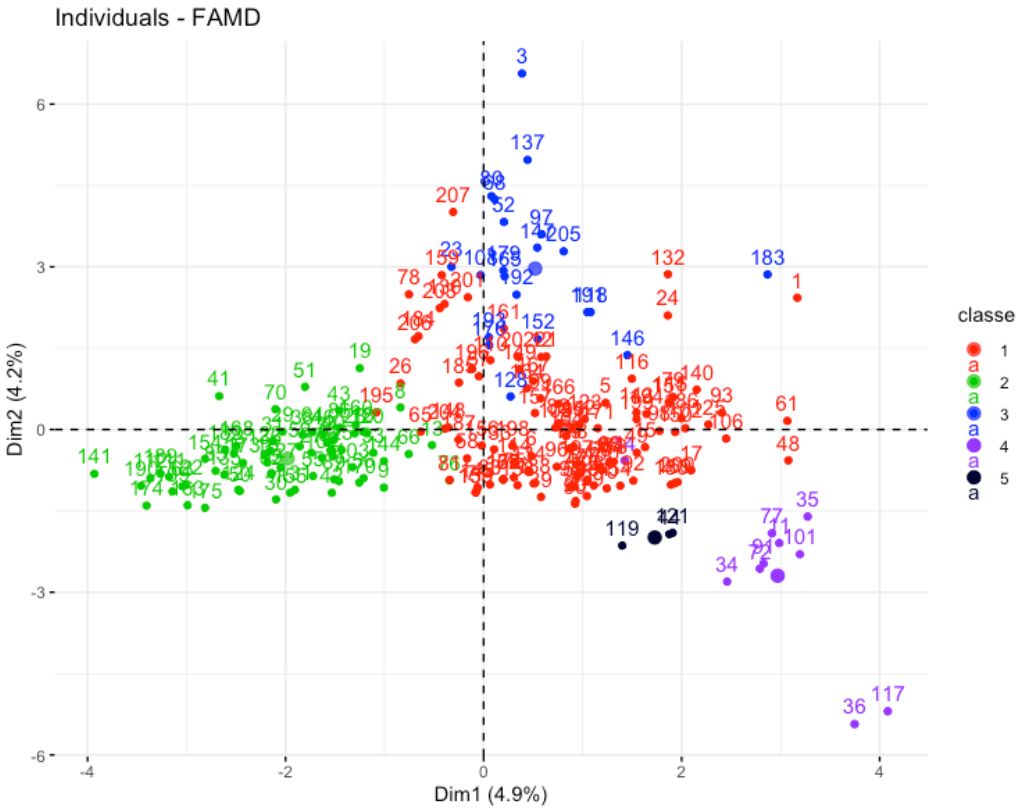


Figure 4.4 : Graphe des individus dans le premier plan factoriel.

1ère dimension :

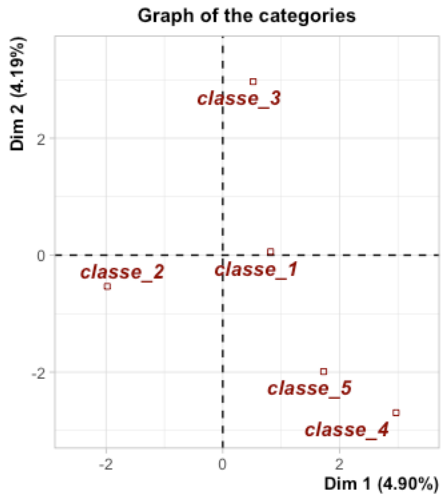


Figure 4.5 : Graphe des centres des classes.

	Coord		Contrib	
Individu	Dim.1	Dim.2	Dim.1	Dim.2
1	3.1686	2.4303	1.7373	1.1939
3	0.3878	6.5646	0.026	8.7106
34	2.4593	-2.8035	1.0466	1.5887
35	3.2729	-1.6047	1.8536	0.5205
36	3.7468	-5.431	2.4291	5.9621
68	0.1107	4.2467	0.0021	3.6453
80	0.0799	4.3033	0.0011	3.7431
101	3.1945	-2.2996	1.7658	1.0689
117	4.0823	-5.1934	2.8837	5.4518
137	0.4441	4.9721	0.0341	4.997
141	-3.9311	-0.8162	2.674	0.1347
174	-3.4039	-1.4032	2.0049	0.398
190	-3.4585	-1.0375	2.0697	0.2176

Figure 4.6 : Tableau des coordonnées et contributions des individus contribuant le plus sur les 2 premiers axes.

```
$quanti
                          correlation p.value
Altitude                        0.748   0.000
Moisture.Percentage             0.558   0.000
Category.Two.Defects_log        0.417   0.000
Quakers_log                     0.343   0.000
Weight.kg                       0.238   0.001
Body                            0.139   0.046

$quali
                    R2 p.value
Country.of.Origin 0.862   0.000
classe            0.755   0.000
Variety           0.749   0.000
Processing.Method 0.184   0.000
Color             0.068   0.027
```

```
$category
                                        Estimate p.value
classe=1                                   0.007   0.000
Country.of.Origin=Ethiopia                 2.343   0.000
Variety=Ethiopian Heirlooms                2.768   0.000
classe=4                                   2.158   0.000
Country.of.Origin=Colombia                 1.355   0.000
Variety=Caturra                            0.985   0.000
Processing.Method=Washed / Wet             0.226   0.000
Country.of.Origin=Guatemala                0.550   0.001
Processing.Method=Anaerobic                2.307   0.033
Variety=Castillo                           2.141   0.033
Variety=Bourbon                            0.334   0.036
Color=greenish                             0.669   0.043
Country.of.Origin=Peru                     1.126   0.043
Processing.Method=Natural / Dry           -0.645   0.039
Variety=Gesha                             -1.162   0.006
Variety=SL34                              -1.969   0.006
Color=blue-green                          -1.204   0.004
Processing.Method=Pulped natural / honey  -1.734   0.000
Variety=Typica                            -2.685   0.000
Country.of.Origin=Taiwan                  -2.593   0.000
classe=2                                  -2.791   0.000
```

Figure 4.7 : Description de l'axe 1 par les variables quantitatives et qualitatives.

Les résultats obtenus et présentés dans les figures 4.2, 4.3 et 4.7 montrent que les variables qui contribuent le plus à l'axe 1 sont les variables Altitude, Pays d'origine, et Variété, et celles qui contribuent d'une façon plus modérée sont les variables Taux d'humidité, Défauts de catégorie 2 et Quakers.

D'après les figures 4.5 et 4.7, et les descriptions de la classe 4, les modalités positivement associées à l'axe 1 sont : les pays Éthiopie, Colombie et Guatemala, les variétés Ethiopian Heirlooms et Caturra, une altitude élevée, et un taux d'humidité et de défauts (catégories 2 et quakers) plus élevés que pour l'ensemble des cafés.

De même, en exploitant les figures 4.5 et 4.7, et les descriptions de la classe 2, on peut affirmer que les modalités négativement associées à l'axe 1 sont : le pays Taiwan, la variété Typica, une altitude basse et très peu de défauts (de catégorie 2 ou quakers).

On peut vérifier cette interprétation avec des individus (annexe A) aux fortes coordonnées (figure 4.4 et 4.6). Par exemple, l'individu 141 est bien issu de Taiwan, cultivé à basse altitude et de variété Typica, sans aucun défaut, alors que les individus 48 et 35, qui sont à l'opposé du 141 sur l'axe 1, sont bien des cafés d'Éthiopie ou du Guatemala, cultivés à haute altitude et de la variété Ethiopian Heirlooms et Catuai, avec des défauts (de catégorie 2 et quakers).

2e dimension :

$quanti

	correlation	p.value
Moisture.Percentage	0.366	0.000
Aftertaste	-0.138	0.047
Body	-0.143	0.039
Total.Cup.Points	-0.163	0.019
Acidity	-0.178	0.010
Overall	-0.184	0.008
Weight.kg	-0.339	0.000

$quali

	R2	p.value
Country.of.Origin	0.771	0
Variety	0.737	0
classe	0.584	0
Color	0.332	0
Processing.Method	0.301	0

$category

	Estimate	p.value
classe=3	3.406	0.000
Variety=Catimor	2.693	0.000
Processing.Method=Semi Washed	3.328	0.000
Country.of.Origin=Laos	4.494	0.000
Color=brownish	1.750	0.000
Color=yellowish	1.712	0.000
Country.of.Origin=Vietnam	2.874	0.000
Country.of.Origin=Thailand	1.366	0.000
Country.of.Origin=Indonesia	3.240	0.000
Variety=Java	3.121	0.000
Variety=Mundo Novo	2.342	0.000
Country.of.Origin=Brazil	1.302	0.000
Variety=Blend	0.655	0.001
Color=bluish-green	0.228	0.009
Processing.Method=Anaerobic	1.016	0.015
Country.of.Origin=El Salvador	0.893	0.020
Variety=Castillo	1.757	0.037
Variety=Pacamara	1.667	0.046
classe=5	-1.552	0.025
Variety=SL14	-2.501	0.025
Country.of.Origin=Uganda	-2.438	0.025
Color=green	-0.899	0.007
Country.of.Origin=Taiwan	-0.956	0.002
Variety=Typica	-1.244	0.002
classe=2	-0.096	0.001
Processing.Method=Natural / Dry	-0.881	0.000
Color=greenish	-1.546	0.000
Country.of.Origin=Ethiopia	-2.965	0.000
classe=4	-2.256	0.000
Variety=Ethiopian Heirlooms	-3.440	0.000
Processing.Method=Washed / Wet	-2.131	0.000

Figure 4.8 : Tableau de description de l'axe 2 par variables quantitatives et qualitatives.

Les résultats obtenus et présentés dans les figures 4.2, 4.3 et 4.8 montrent que les variables qui contribuent le plus à l'axe 2 sont les variables Pays d'origine, et Variété, et d'une façon plus modérée, les variables Méthode de Traitement, Couleur, Masse produite et Taux d'humidité.

D'après la figure 4.8, les cafés observés en haut de l'axe sont des cafés issus d'Asie du Sud-Est (Laos, Indonésie, Vietnam), de couleur brunâtre ou jaunâtre, souvent traités par une méthode de traitement *Semi Washed*, de variétés Catimor, Java ou Mundo Novo, et produits en petite quantité, tandis que les cafés observés le plus en bas sont des cafés provenant d'Ethiopie et de variété Typica ou Ethiopian Heirlooms, de couleur verte/verdâtre, traités par la méthode *Washed/Wet*, et produits en grande quantité.

On retrouve ces observations avec les classes : la classe 3 se situant en haut de l'axe, et la classe 4 en bas.

Cette interprétation peut être validée par les individus 137 et 36, choisis pour leurs oppositions et leurs contributions (figure 4.4 et 4.6). On constate tout de même les limites de cette interprétation en observant des individus à coordonnées plus modérées, ayant pourtant des coordonnées opposées sur l'axe 2, et qui n'ont pas les critères précisés dans ce paragraphe (voir les individus 1 et 101). Cela est dû à la faible inertie de la seconde composante (environ 4%).

C. SECOND PLAN FACTORIEL (7,7% D'INERTIE)

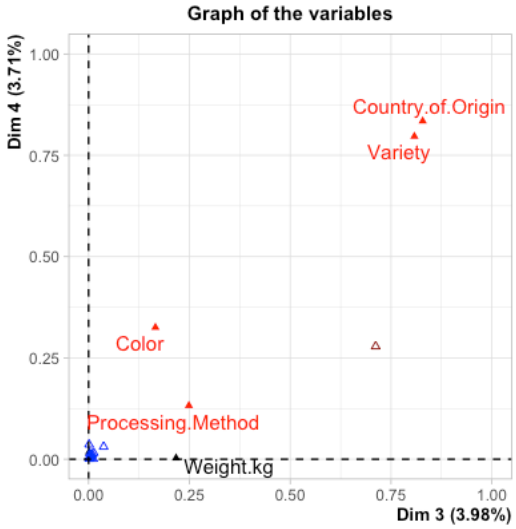


Figure 4.9 : Graphe des variables aux plus fortes contributions

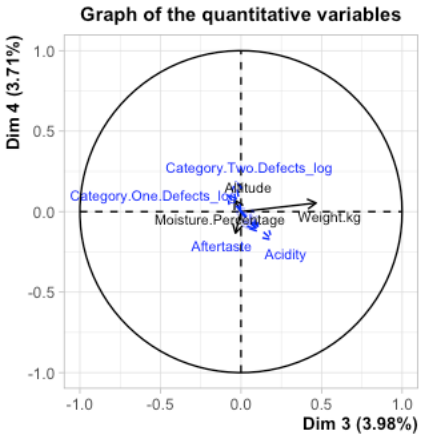


Figure 4.10 : Cercle des corrélations des variables aux plus fortes coordonnées

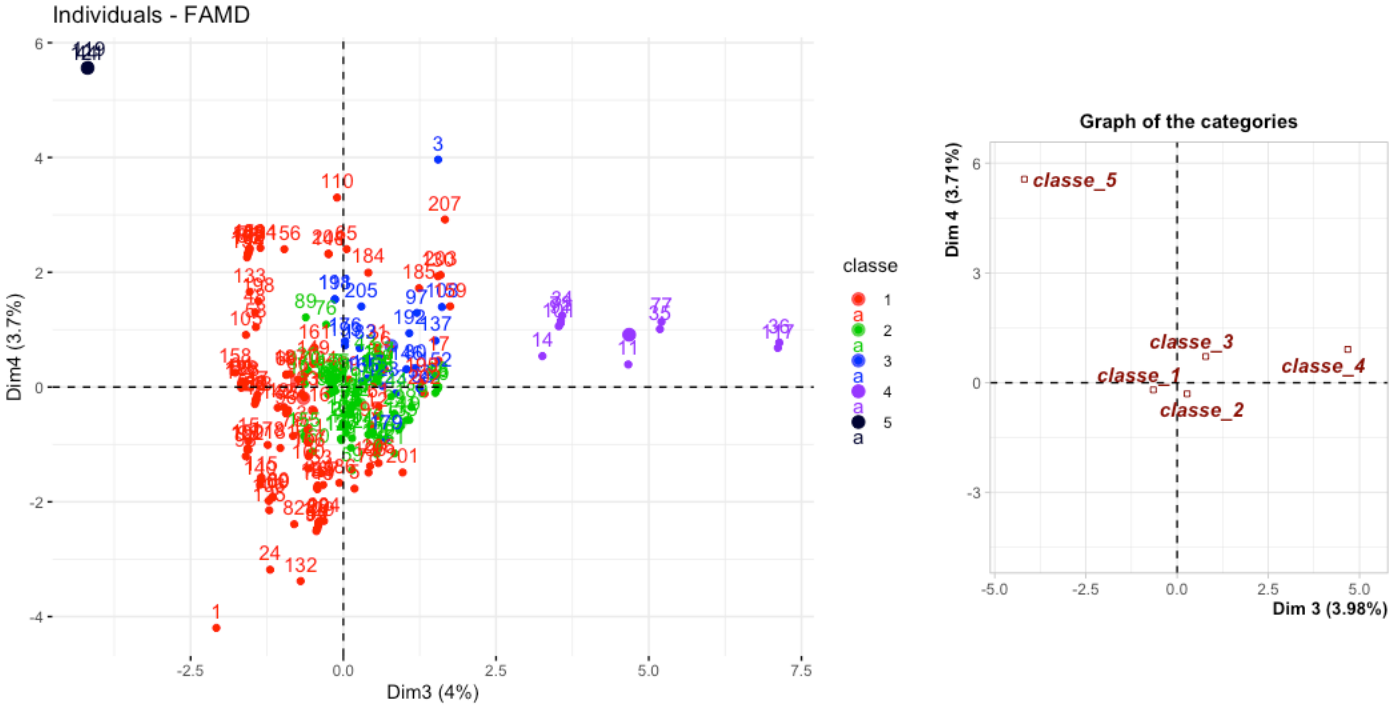


Figure 4.11 : Représentations graphiques des individus et des centres des classes dans le premier plan factoriel

Figure 4.12 : Tableau des coordonnées et contributions des individus contribuant le plus sur les axes 3 et 4.

	Coord		Contrib	
Individu	Dim.3	Dim.4	Dim.3	Dim.4
1	-2.079	-4.1967	0.9197	4.0235
24	-1.199	-3.1828	0.3059	2.3143
35	5.1826	1.0093	5.7155	0.2327
36	7.1365	0.7738	10.8376	0.1368
44	-4.1946	5.5423	3.744	7.0174
77	5.2081	1.1367	5.7718	0.2952
117	7.1136	0.6821	10.768	0.1063
119	-4.1692	5.6054	3.6988	7.1781
121	-4.1969	5.5331	3.7481	6.9942
132	-0.6994	-3.381	0.1041	2.6115

```
$quanti
                correlation p.value
Weight.kg             0.466   0.000
Acidity               0.193   0.005

$quali
                        R2 p.value
Country.of.Origin    0.828       0
Variety              0.808       0
classe               0.712       0
Processing.Method    0.249       0
Color                0.166       0

$category
                                      Estimate p.value
classe=4                                 4.500   0.000
Country.of.Origin=Ethiopia               4.591   0.000
Variety=Ethiopian Heirlooms              4.945   0.000
Processing.Method=Natural / Dry          0.977   0.000
Color=yellow-green                       1.199   0.000
Country.of.Origin=Brazil                 1.404   0.009
classe=3                                 0.604   0.012
Variety=Mundo Novo                       1.746   0.028
Country.of.Origin=Taiwan                 0.550   0.029
Country.of.Origin=Guatemala             -0.566   0.014
Country.of.Origin=Costa Rica            -1.140   0.010
Variety=Catuai                          -0.997   0.004
Country.of.Origin=Colombia              -0.818   0.002
Country.of.Origin=Honduras              -1.157   0.001
Color=green                             -0.862   0.000
classe=5                                -4.367   0.000
Variety=SL14                            -4.079   0.000
Country.of.Origin=Uganda                -3.991   0.000
Variety=Caturra                         -1.259   0.000
classe=1                                -0.831   0.000
Processing.Method=Washed / Wet          -0.741   0.000
```

Figure 4.13 : Description de l'axe 3 par les variables quantitatives et qualitatives.

Sur ce second plan, on n'observe pas de séparation des classes 1, 2 et 3. L'axe 3 porte sur les variables Pays d'origine et Variétés, plus modérément sur la variable Masse produite, et légèrement sur les variables Acidité, Couleur et Méthode de Traitement. Il oppose la classe 4 et la classe 5:

- on retrouve à droite des cafés d'Éthiopie, de variété Ethiopian Heirlooms, et avec une acidité plus marquée et une masse produite plus importante ;
- et à gauche des cafés d'Ouganda de variété SL14.

L'axe 4 positionne en haut des cafés d'Ouganda (individu 44, 119 et 121), tous de variété SL14, traités par la méthode *Washed/Wet,* de couleur verte ; et en bas des cafés de Colombie, de variété Blend ou Castillo, et traités par la méthode *Anaerobic* (individus 1, 24 et 132).

Pour cette interprétation des axes, on aurait aussi pu utiliser d'autres graphiques comme celui des centres des individus prenant la même modalité. Ces graphiques sont disponibles en annexe O, et sont en adéquation avec les descriptions des axes faites précédemment.

Finalement, l'analyse factorielle permet de mettre en évidence les corrélations et les oppositions entre des individus dans les dimensions principales, et pose le cadre permettant d'obtenir les regroupements vus dans la partie III.C. Cette analyse fournit une vision globale et structurelle des données, tandis que la classification permet une interprétation plus ciblée et utile pour cette étude.

V. CONCLUSION

La classification par une approche tandem basée sur l'AFDM permet d'identifier cinq groupes de cafés, définis en fonction de leurs conditions géographiques et des paramètres de production appliqués. Les profils des classes obtenues, en lien avec les notes sensorielles et les défauts, sont les suivants :

- **Cafés d'Amérique du Sud:**
  Cultivés à haute altitude, issus des variétés Caturra, Bourbon, Catuai, et traités par la méthode *Washed/Wet,* ces cafés ont des scores faibles pour la plupart des critères sensoriels et présentent davantage de défauts ;
- **Cafés de Taiwan et de Hawaï :**
  Cultivés à basse altitude, avec un faible taux d'humidité, issus des variété Typica, Gesha ou SL34, et traités par la méthode *Pulped Natural/Honey*, ces cafés se distinguent par de bonnes notes sensorielles sur la plupart des critères et un nombre faible de défauts ;
- **Cafés d'Asie du Sud-Est :**
  Issus en majorité de la variété Catimor, ces cafés présentent plus d'humidité, et ont des scores faibles en Acidité, Impression générale et en Corps;
- **Cafés d'Éthiopie :**
  Cultivés à très haute altitude, issus de la variété Ethiopian Heirlooms, et bien qu'affichant davantage de défauts, ces cafés obtiennent d'excellents scores, surtout en Acidité et en Corps;
- **Cafés d'Ouganda :**
  Issus de la variété SL14, ces cafés sont exempts de défauts de type quakers, et obtiennent une note supérieure à la moyenne en Arôme, bien qu'ils ne se démarquent pas sur les autres critères sensoriels.

A l'aide de ces observations, on peut donc répondre à la problématique : les critères géographiques, la variété et les méthodes de traitement, sont bien liés aux qualités sensorielles et aux défauts des cafés.

Ces résultats peuvent offrir des pistes stratégiques commerciales, comme mettre en avant l'origine des cafés afin de valoriser leurs spécificités gustatives, ou encore dédier certains cafés à des gammes spécifiques en fonction de leur provenance.

REMERCIEMENTS

« Je tiens à remercier les professeurs, Vincent Audigier et Mouhamoudou Ndao, pour la clarté et la rigueur des cours et travaux dirigés, ainsi que pour leurs remarques constructives qui ont grandement contribué à la réalisation de ce projet. »

Valérie Plusquellec

REFERENCES

Cours et TD de l'UE STA101, CNAM
Vincent Audigier et Mouhamadou Ndao

R pour la statistique et la science des données
Sous la direction de François Husson, 2018

Analyse de données avec R
François Husson, Sébastien Lê, Jérôme Pagès, 2009

The World Atlas of coffee from beans to brewing – coffees explored, explained and enjoyed
James Hoffmann, 2014

A System to Assess Coffee Value: Understanding the SCA's Coffee Value Assessment
The Specialty Coffee Association, 2024

Source de la banque de données :
https://github.com/fatih-boyar/coffee-quality-data-CQI/tree/main?tab=readme-ov-file

VI. ANNEXES

A. EXTRAIT DU JEU DE DONNEES

ID	Country of Origin	Altitude	Weight.kg	Variety	Processing Method	Moisture Percentage	Color
1	Colombia	1815	35	Castillo	Anaerobic	11.8	green
2	Taiwan	1200	80	Gesha	Washed / Wet	10.5	blue-green
3	Laos	1300	475	Java	Semi Washed	10.4	yellowish
11	Ethiopia	2000	300	Ethiopian Heirlooms	Natural / Dry	11.8	greenish
17	Ethiopia	2000	40	Blend	Washed / Wet	11.3	green
24	Colombia	1350	30	Castillo		11.3	brownish
35	Ethiopia	2250	19200	Ethiopian Heirlooms	Natural / Dry	12.3	yellow-green
36	Ethiopia	1700	6144000	Ethiopian Heirlooms	Washed / Wet	9.4	greenish
41	Taiwan	400	10	Typica	Natural / Dry	10	brownish
44	Uganda	1905	19200	SL14	Washed / Wet	11	green
48	Guatemala	4700	18960	Catuai	Washed / Wet	11.3	green
52	Indonesia	1200	19200	Catimor		11.9	bluish-green
68	Vietnam	650	15	Catimor	Natural / Dry	11.6	brownish
77	Ethiopia	2361	19200	Ethiopian Heirlooms	Natural / Dry	11	yellow-green
80	Indonesia	1300	60	Blend	Semi Washed	11.5	blue-green
91	Ethiopia	1800	18000	Ethiopian Heirlooms	Washed / Wet	11.3	green
97	Thailand	1350	30	Catimor	Natural / Dry	11.6	yellowish
101	Ethiopia	2000	6000	Ethiopian Heirlooms	Washed / Wet	12	green
105	Mexico	1200	18000	Sarchimor	Washed / Wet	11.7	green
117	Ethiopia	1700	6144000	Ethiopian Heirlooms	Washed / Wet	10.4	greenish
119	Uganda	1650	9000	SL14	Washed / Wet	10.1	green
120	Taiwan	1100	40	Blend	Washed / Wet	10.1	blue-green
121	Uganda	1905	19200	SL14	Washed / Wet	11.1	green
126	Costa Rica	1850	22080	Caturra	Washed / Wet	10	green
132	Colombia	1390	24	Blend	Anaerobic	10.9	brownish
137	Laos	1250	1560	Catimor	Natural / Dry	11.8	brownish
141	Taiwan	275	12	Typica	Pulped natural / honey	8.9	blue-green
174	Taiwan	460	60		Pulped natural / honey	8.1	greenish
190	Taiwan	435	8		Pulped natural / honey	8.5	green
206	El Salvador	1200	2	Maragogype	Natural / Dry	11	bluish-green
207	Brazil	975	36000	Mundo Novo	Semi Washed	11.3	green

ID	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean Cup	Sweetness	Overall	Total Cup Points	Category One Defects	Quakers	Category Two Defects
1	8.58	8.5	8.42	8.58	8.25	8.42	10.0	10.0	10.0	8.58	89.33	0	0	3
2	8.5	8.5	7.92	8.0	7.92	8.25	10.0	10.0	10.0	8.5	87.58	0	0	0
3	8.33	8.42	8.08	8.17	7.92	8.17	10.0	10.0	10.0	8.33	87.42	0	0	2
11	8.08	8.25	8.0	8.08	7.92	7.92	10.0	10.0	10.0	8.0	86.25	0	1	1
17	8.17	8.08	7.92	8.17	7.75	7.92	10.0	10.0	10.0	8.08	86.08	0	2	2
24	8.08	8.0	7.83	8.17	7.75	7.83	10.0	10	10	8.0	85.67	0	0	2
35	8.0	8.08	8.0	8.0	7.67	7.75	10.0	10.0	10.0	7.83	85.33	0	3	4
36	7.92	7.75	7.83	8.17	8.0	7.75	10.0	10.0	10.0	7.83	85.25	0	1	1
41	8.0	7.92	7.83	7.92	7.75	7.75	10.0	10.0	10.0	7.92	85.08	0	0	0
44	8.0	7.92	7.75	7.75	7.75	7.92	10.0	10	10	7.92	85.0	0	0	1
48	7.67	8.0	7.75	7.92	8.0	7.83	10.0	10.0	10.0	7.83	85.0	0	0	4
52	7.83	7.92	7.75	7.83	7.83	7.83	10.0	10.0	10.0	7.83	84.83	0	3	2
68	8.0	7.83	7.75	7.67	7.58	7.83	10.0	10.0	10.0	7.92	84.58	0	0	0
77	7.83	7.75	7.58	7.92	8.0	7.5	10.0	10.0	10.0	7.75	84.33	0	5	4
80	7.67	7.67	7.83	7.83	7.67	7.75	10.0	10.0	10.0	7.83	84.25	0	1	1
91	8.0	7.75	7.5	7.67	7.67	7.67	10.0	10.0	10.0	7.75	84.0	1	1	1
97	7.75	7.67	7.58	7.75	7.75	7.67	10.0	10.0	10.0	7.75	83.92	0	0	3
101	7.83	7.67	7.5	7.92	7.5	7.67	10.0	10.0	10.0	7.75	83.83	0	0	12
105	7.67	7.67	7.58	7.67	7.83	7.58	10.0	10.0	10.0	7.75	83.75	4	0	12
117	7.42	7.42	7.42	8.0	7.92	7.67	10.0	10.0	10.0	7.67	83.5	0	1	11
119	7.75	7.75	7.5	7.5	7.58	7.58	10.0	10.0	10.0	7.75	83.42	0	0	3
120	7.75	7.92	7.67	7.92	7.75	7.83	8.67	10.0	10.0	7.92	83.42	0	0	0
121	7.75	7.67	7.58	7.5	7.67	7.5	10.0	10.0	10.0	7.67	83.33	0	0	5
126	7.58	7.75	7.42	7.67	7.5	7.75	10.0	10.0	10.0	7.67	83.33	0	0	3
132	7.67	7.67	7.58	7.67	7.58	7.58	10.0	10.0	10.0	7.5	83.25	0	0	2
137	7.67	7.75	7.33	7.67	7.58	7.5	10.0	10.0	10.0	7.58	83.08	0	0	7
141	7.5	7.67	7.58	7.58	7.5	7.5	10.0	10.0	10.0	7.58	82.92	0	0	0
174	7.33	7.42	7.42	7.5	7.42	7.5	10.0	10.0	10.0	7.5	82.08	0	0	3
190	7.33	7.5	7.42	7.33	7.33	7.25	10.0	10.0	10.0	7.33	81.5	0	0	0
206	6.5	6.75	6.75	7.17	7.08	7.0	10.0	10.0	10.0	6.83	78.08	0	12	13
207	7.25	7.08	6.67	6.83	6.83	6.67	10.0	10.0	10.0	6.67	78.0	0	0	1

B. LEXIQUE DES VARIABLES SENSORIELLES

Arôme – *Aroma* : perception olfactive du café infusé à deux moments.

Saveur – *Flavor* : perception issue à la fois du goût et de la composante olfactive lorsque le café est en bouche.

Arrière-goût – *Aftertaste* : perception issue du goût et de la composante olfactive après avoir avalé ou recraché le café.

Acidité – *Acidity* : perception du goût acide provoqué par l'infusion, qui peut varier en intensité et en caractère.

Corps – *Body* : perception de l'épaisseur ou de la viscosité du café en bouche.

Équilibre - *Balance* : perception de l'harmonisation des différentes composantes aromatiques.

Uniformité – *Uniformity* : perception de la cohérence du café d'une tasse à l'autre.

Propreté de la tasse – *Clean Cup* : perception de l'absence de tout défaut ou arrière-goût indésirable, comme l'acidité excessive, le goût moisi ou rassis.

Douceur – *Sweetness* : perception d'une qualité telle que caramel, fruitée ou florale.

Impression générale – *Overall* : perception générale de la qualité, y compris les aspects non couverts dans les autres sections, comme les préférences personnelles.

C. EXPLICATION DES DEFAUTS MESURES DANS LE CAFE

Quackers : grains non mûrs.

Défaut de catégorie 1 : grains complètement noirs ou fermentés, gousses/cerises, et bâtons ou pierres de taille grande ou moyenne.

Défaut de catégorie 2 : coques/enveloppes, grains cassés ou ébréchés, dommages causés par des insectes, grains partiellement noirs ou fermentés, grains en coque, petits bâtons ou petites pierres, et dommages causés par l'eau.

NB : pour un échantillon de 350 grammes de grains, le café de spécialité ne doit présenter aucun défaut de catégorie 1 et un maximum de cinq défauts de catégorie 2.

D. FREQUENCES DES MODALITES DES 4 VARIABLES QUALITATIVES

$Country.of.Origin

	n	%	val%
Brazil	10	4.8	4.8
Colombia	19	9.2	9.2
Costa Rica	8	3.9	3.9
El Salvador	7	3.4	3.4
Ethiopia	11	5.3	5.3
Guatemala	21	10.1	10.1
Honduras	13	6.3	6.3
Indonesia	3	1.4	1.4
Kenya	2	1.0	1.0
Laos	3	1.4	1.4
Madagascar	1	0.5	0.5
Mexico	4	1.9	1.9
Myanmar	1	0.5	0.5
Nicaragua	7	3.4	3.4
Panama	2	1.0	1.0
Peru	4	1.9	1.9
Taiwan	61	29.5	29.5
Tanzania, United Republic Of	6	2.9	2.9
Thailand	12	5.8	5.8
Uganda	3	1.4	1.4
United States (Hawaii)	5	2.4	2.4
Vietnam	4	1.9	1.9

$Processing.Method

	n	%	val%
Anaerobic	2	1.0	1.0
Natural / Dry	48	23.2	23.2
Pulped natural / honey	26	12.6	12.6
Semi Washed	3	1.4	1.4
Washed / Wet	128	61.8	61.8

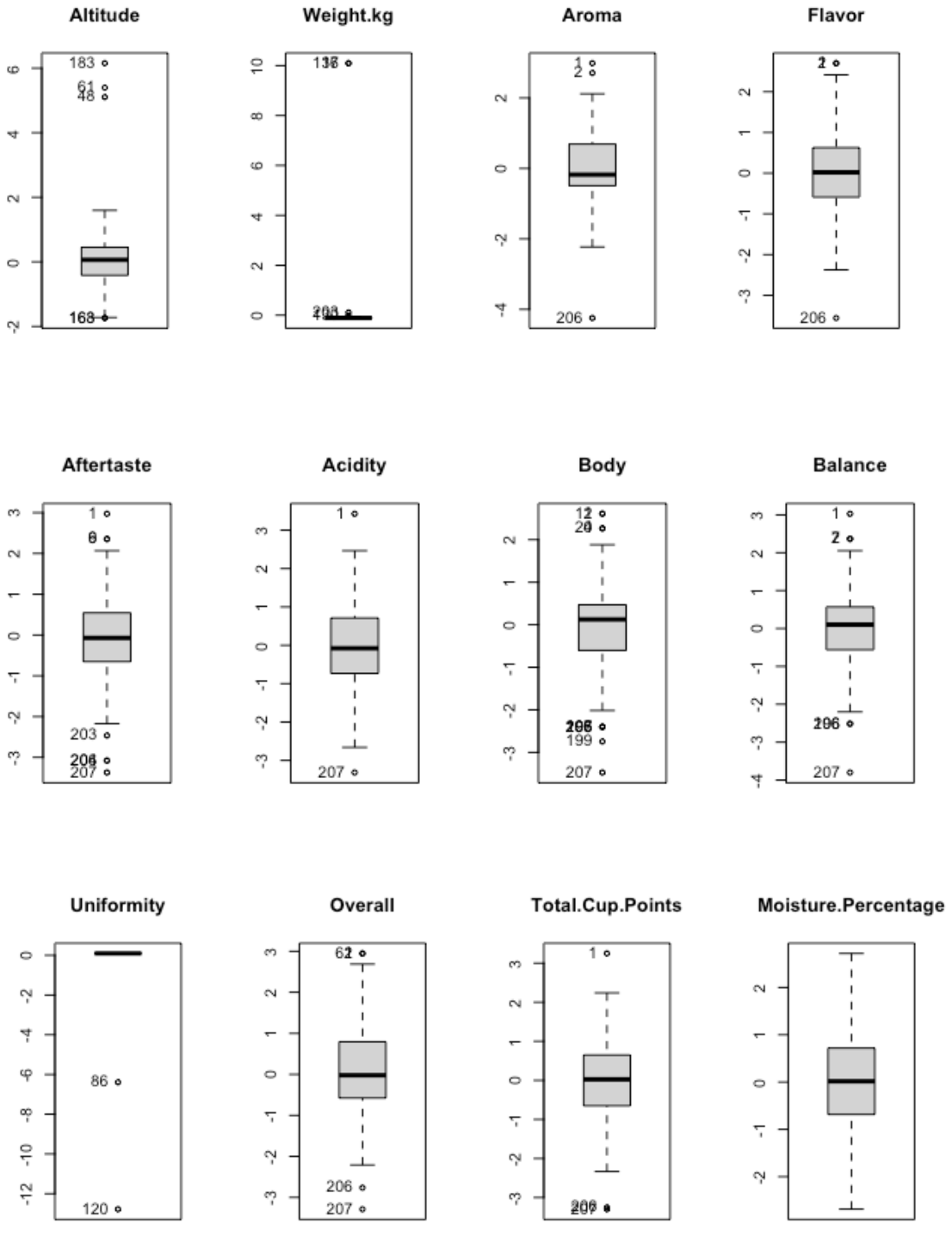
$Color

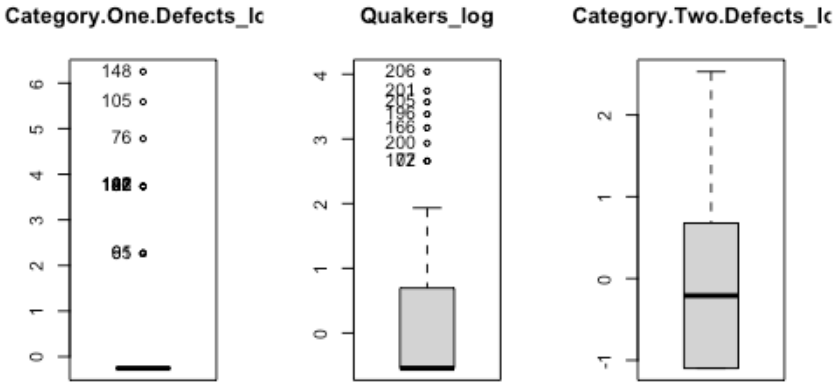
	n	%	val%
blue-green	12	5.8	5.8
bluish-green	21	10.1	10.1
brownish	10	4.8	4.8
green	101	48.8	48.8
greenish	36	17.4	17.4
yellow-green	17	8.2	8.2
yellowish	10	4.8	4.8

$Variety

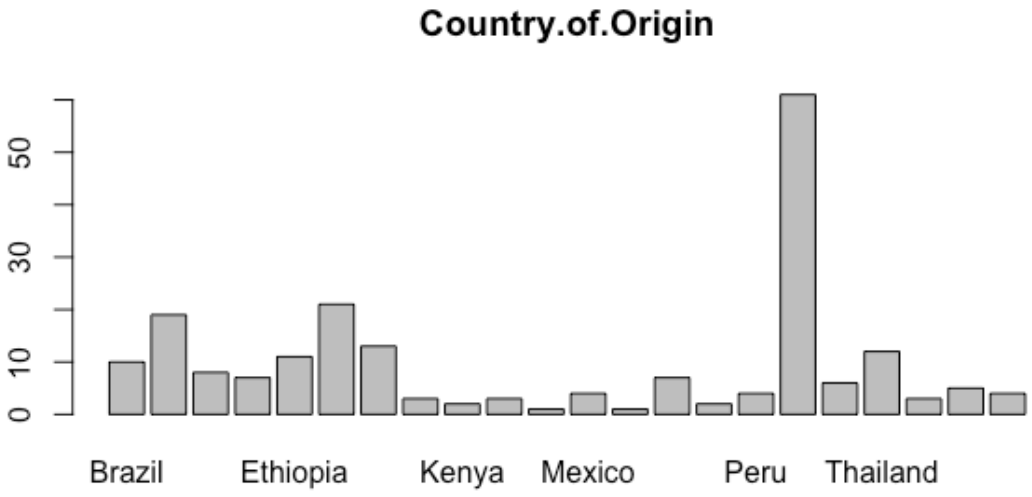
	n	%	val%
Blend	18	8.7	8.7
Bourbon	22	10.6	10.6
Castillo	2	1.0	1.0
Catimor	11	5.3	5.3
Catrenic	1	0.5	0.5
Catuai	14	6.8	6.8
Caturra	30	14.5	14.5
Ethiopian Heirlooms	9	4.3	4.3
Gayo	1	0.5	0.5
Gesha	29	14.0	14.0
Java	3	1.4	1.4
Lempira	1	0.5	0.5
Maragogype	2	1.0	1.0
Mundo Novo	4	1.9	1.9
Pacamara	2	1.0	1.0
Pacas	1	0.5	0.5
Parainema	2	1.0	1.0
Santander	1	0.5	0.5
Sarchimor	2	1.0	1.0
SHG	3	1.4	1.4
SL14	3	1.4	1.4
SL28	2	1.0	1.0
SL34	8	3.9	3.9
Typica	36	17.4	17.4

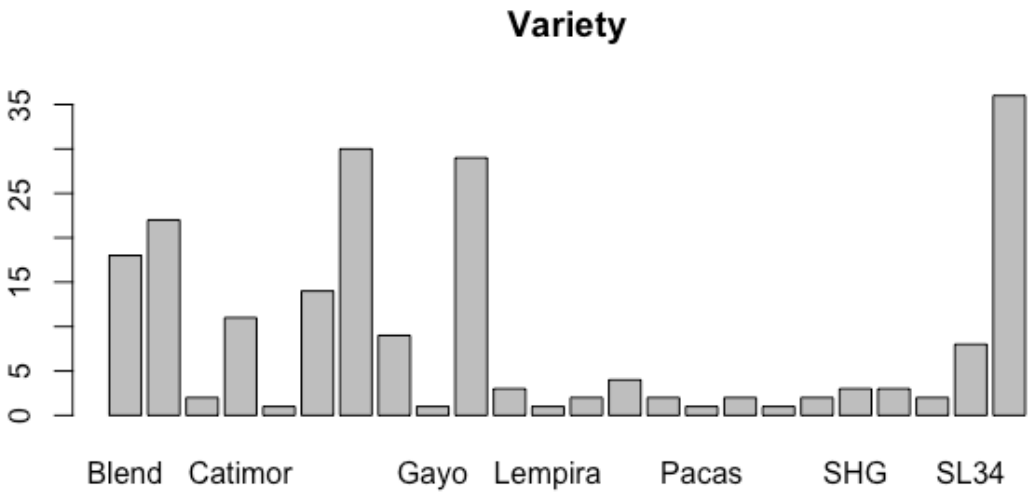
E. BOITES A MOUSTACHE DES VARIABLES QUANTITATIVES

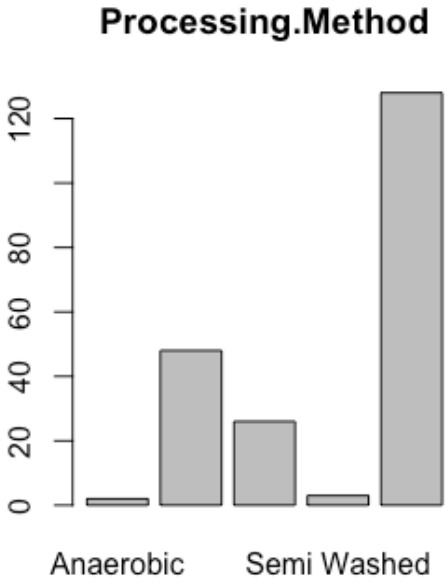


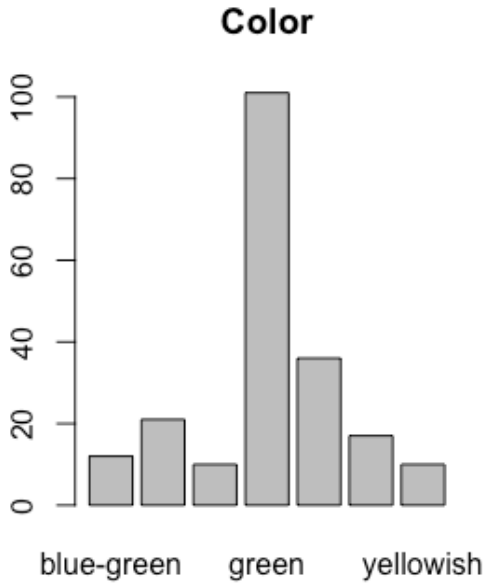


F. DIAGRAMMES EN BARRE DES VARIABLES QUALITATIVES

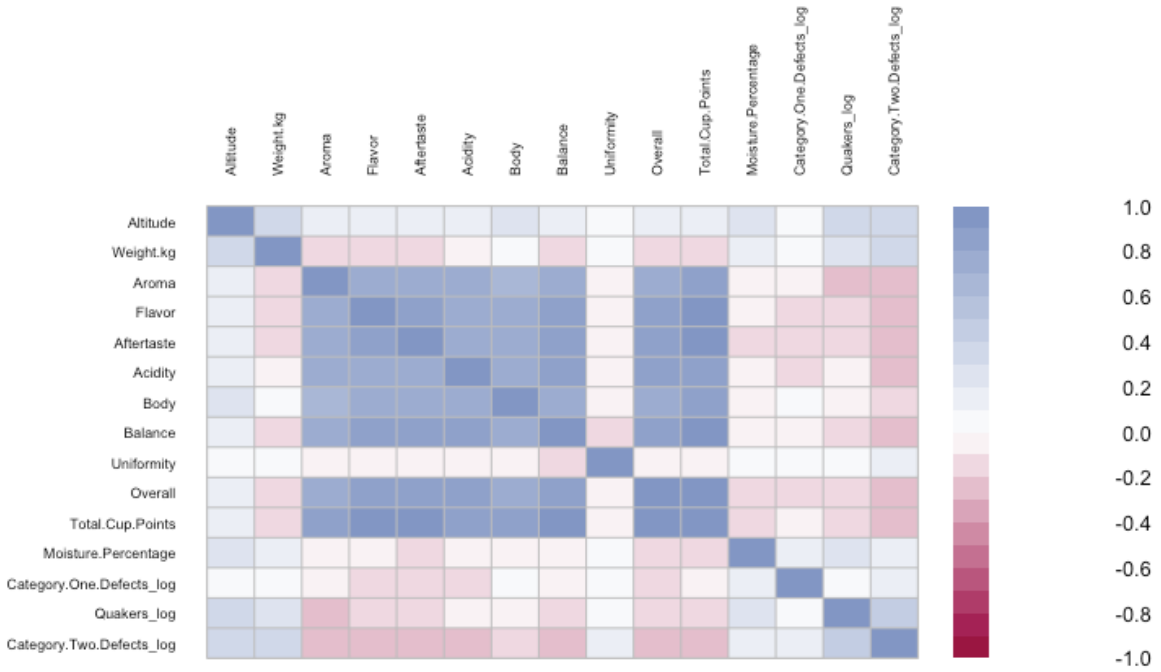



Variety
0
5
15
25
35
Blend
Catimor
Gayo
Lempira
Pacas
SHG
SL34

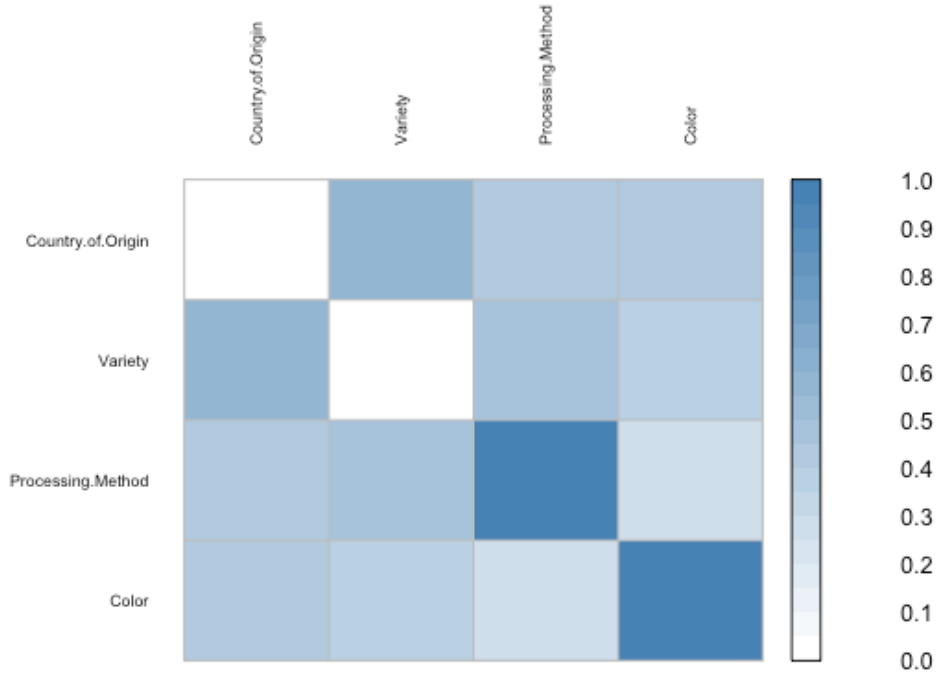

Processing.Method
0
20
40
60
80
120
Anaerobic
Semi Washed


Color
0
20
40
60
80
100
blue-green
green
yellowish

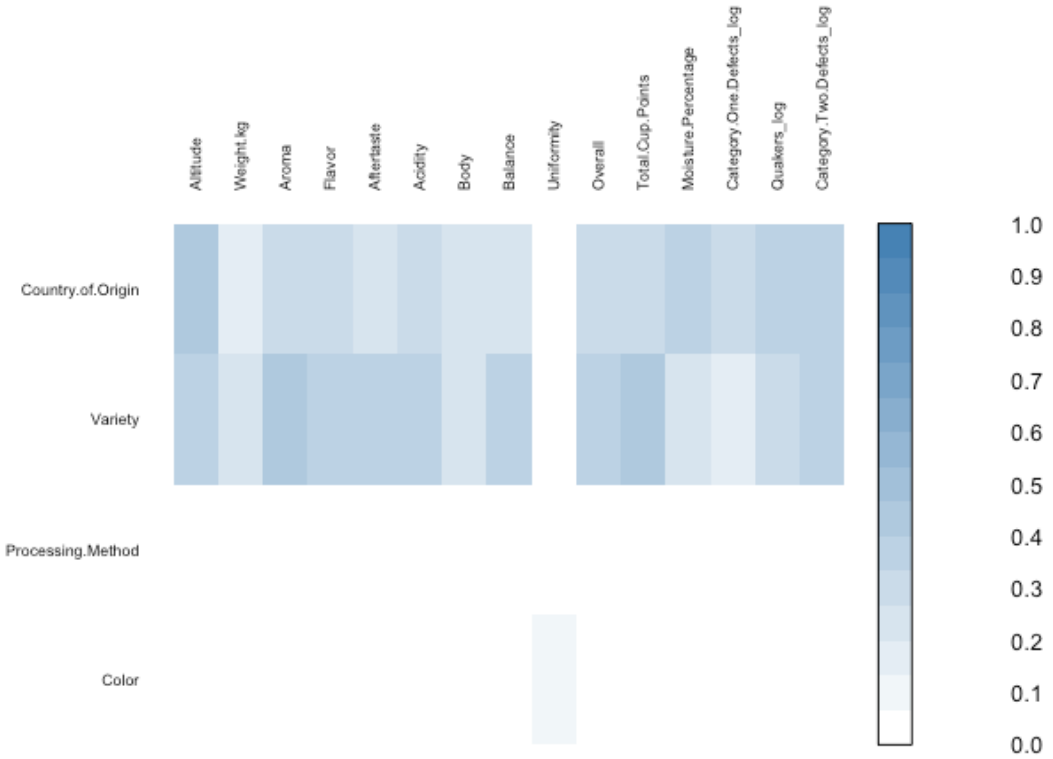
G. REPRÉSENTATION GRAPHIQUE DES COEFFICIENTS DE CORRÉLATION DE SPEARMAN



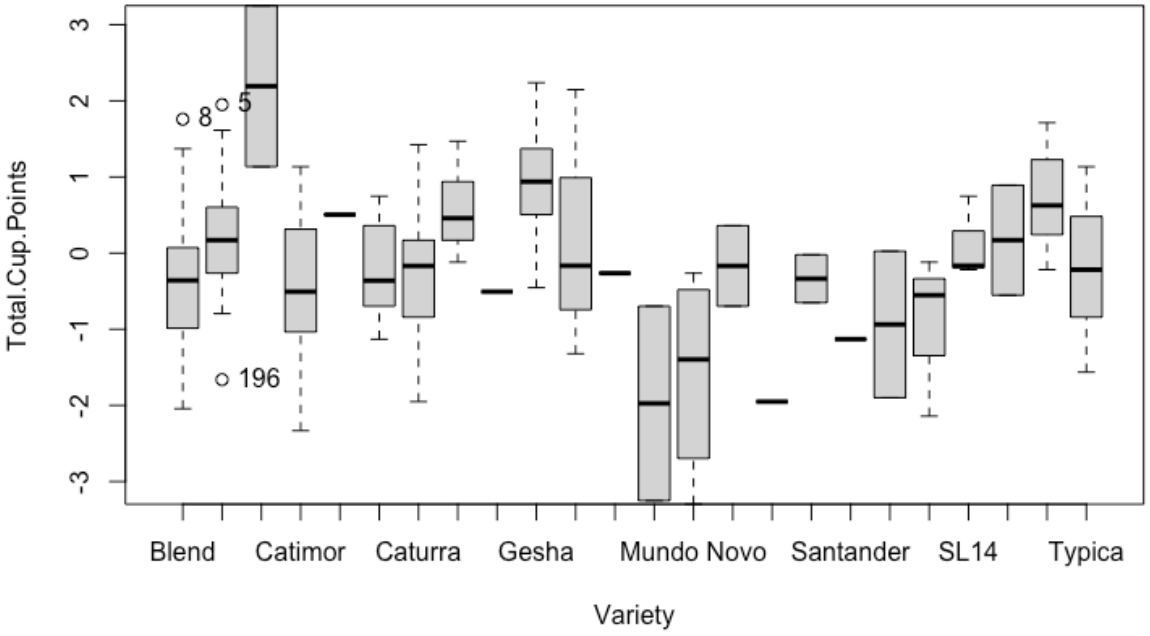
H. REPRÉSENTATION GRAPHIQUE DES COEFFICIENTS V DE CRAMER

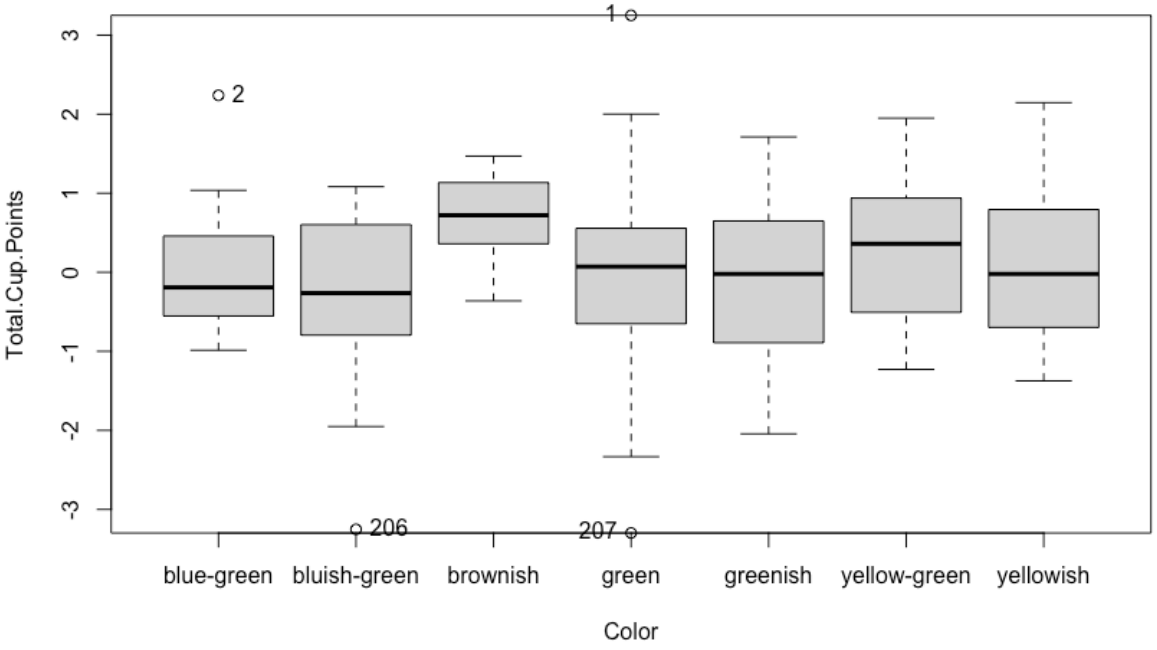


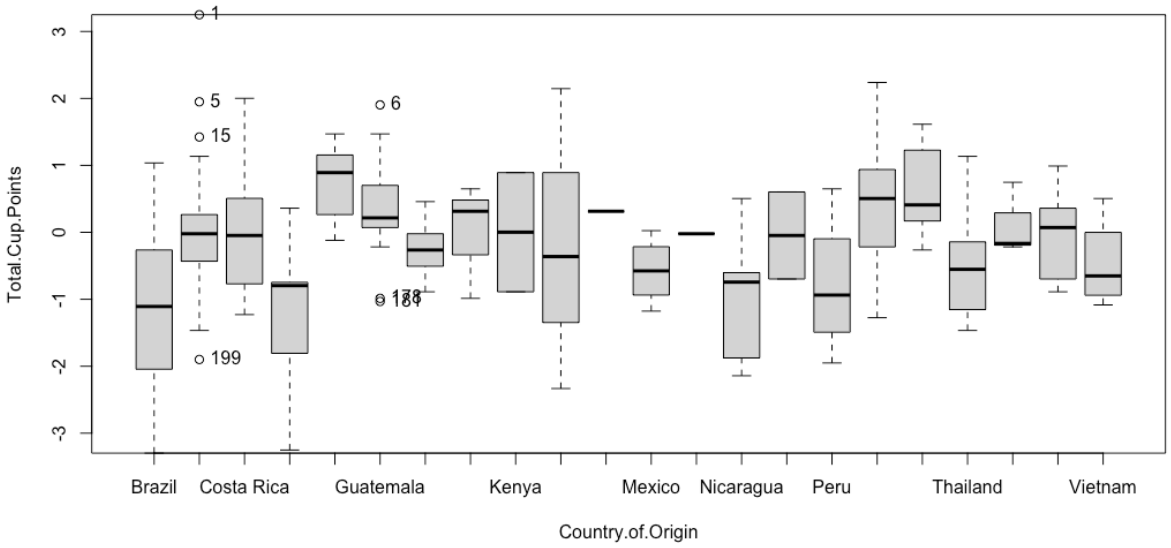
I. REPRÉSENTATION GRAPHIQUE DES COEFFICIENTS ETA^2 ENTRE UNE VARIABLE QUALITATIVE ET UNE QUANTITATIVE.



J. BOITES À MOUSTACHE DE LA VARIABLE SCORE TOTAL SUIVANT LES MODALITÉS DES VARIABLES QUALITATIVES.





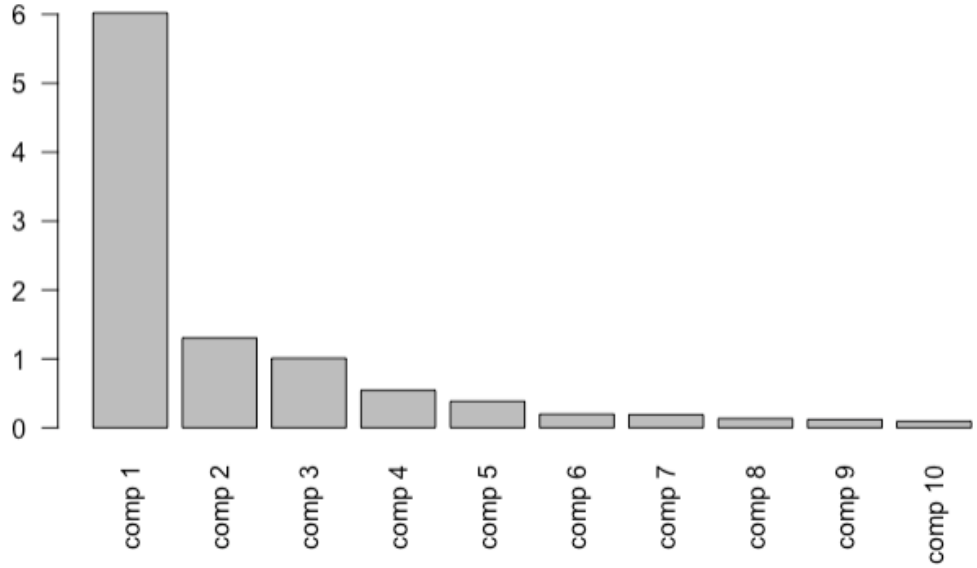

Total.Cup.Points
3
2
1
0
-1
-2
-3
1
5
15
199
6
Brazil
Costa Rica
Guatemala
Kenya
Mexico
Nicaragua
Peru
Thailand
Vietnam
Country.of.Origin

K. ACP SUR LES VARIABLES SENSORIELLES ET DEFAUTS

Figure K.1 : Tableau des valeurs propres et des inerties associées

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	6.02056195	60.2056195	60.20562
comp 2	1.30399458	13.0399458	73.24557
comp 3	1.01029297	10.1029297	83.34849
comp 4	0.54441767	5.4441767	88.79267
comp 5	0.38466113	3.8466113	92.63928
comp 6	0.19737920	1.9737920	94.61308
comp 7	0.19042828	1.9042828	96.51736
comp 8	0.13491544	1.3491544	97.86651
comp 9	0.12039006	1.2039006	99.07041
comp 10	0.09295871	0.9295871	100.00000

Figure K.2 : Valeurs propres associées à chaque dimension



Interprétation :

D'après la règle de Kaiser, on conserve 3 composantes (figure K.1), ce qui est confirmé par la présence d'un coude entre la 3e et la 4e composante (figure K.2).

Les 3 premiers axes cumulent 83% de l'inertie du nuage, ce qui est relativement satisfaisant pour un nombre total de 10 variables.

Figure K.3 : Graphe des individus sur le premier plan

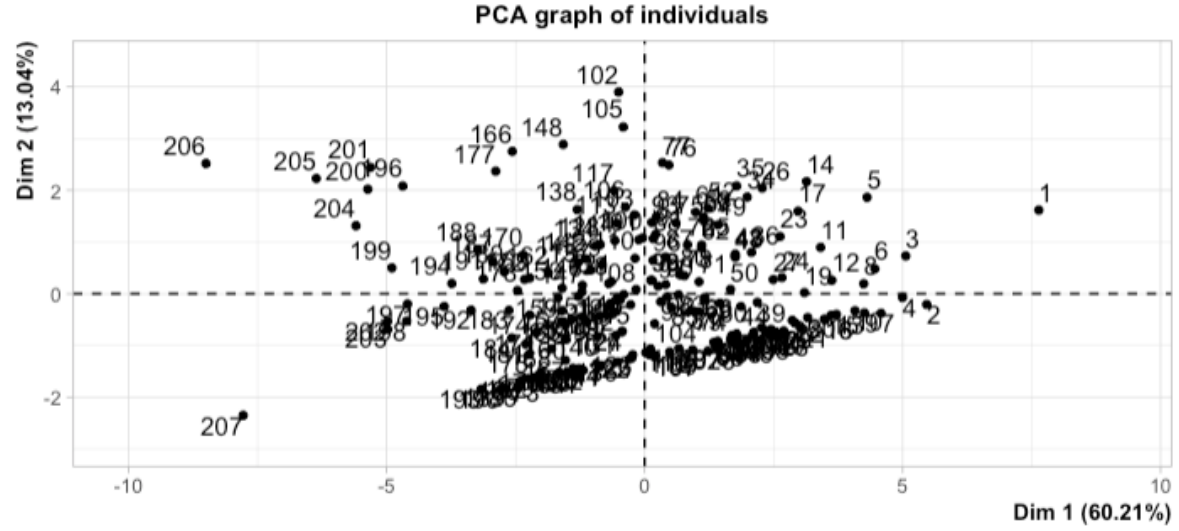
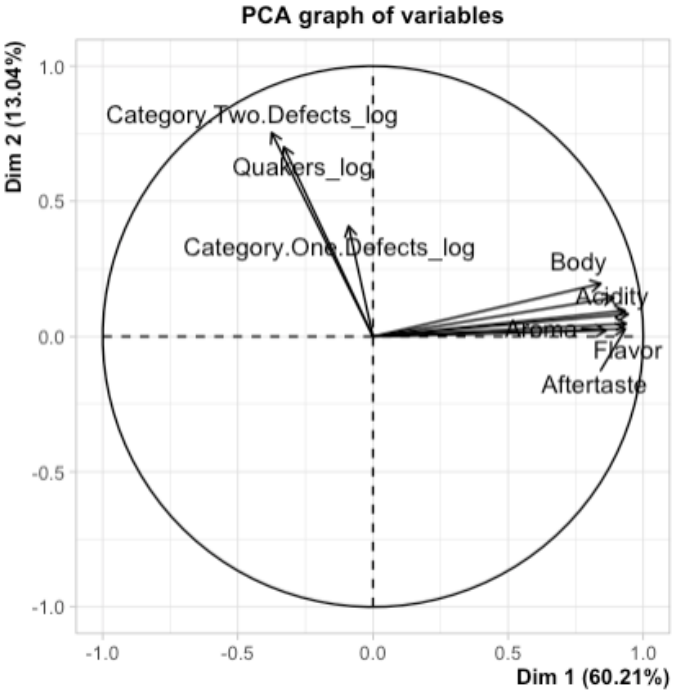


Figure K.4 : Graphe des variables sur le premier plan



Interprétation :

On remarque que sur le premier axe, qui rassemble plus de 60% de l'inertie totale, les 7 variables sensorielles sont bien représentées et peuvent être considérées comme liées linéairement.

On observe aussi que dans le plan qui rassemble plus de 73% de l'inertie totale, les variables *Category.Two.Defects* et *Quakers* sont relativement bien représentées et sont indépendantes des variables sensorielles dans ce plan.

Figure K.5 : Graphe des individus pour la 3e composante

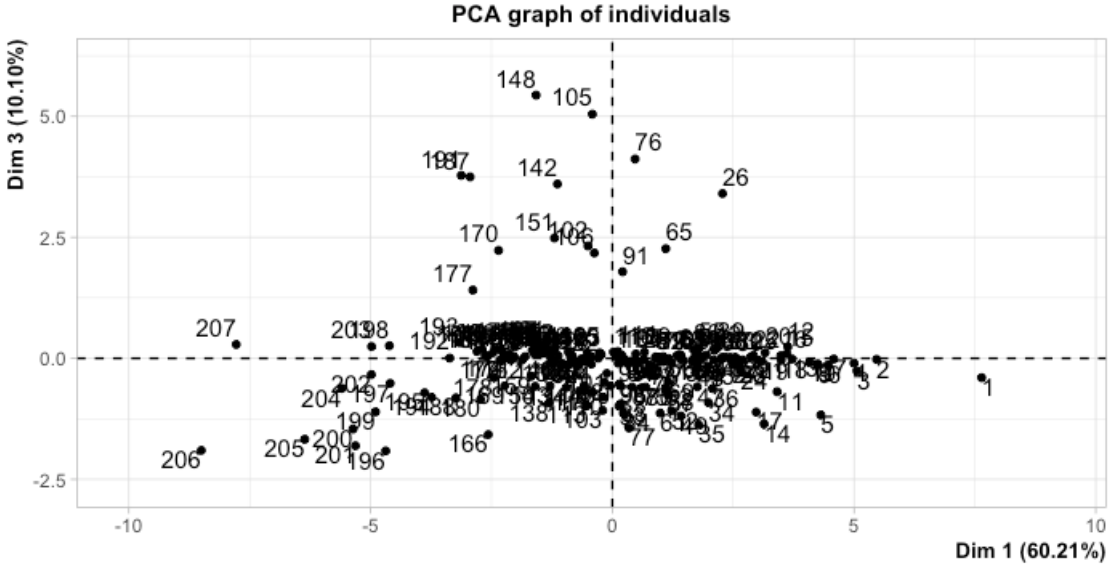
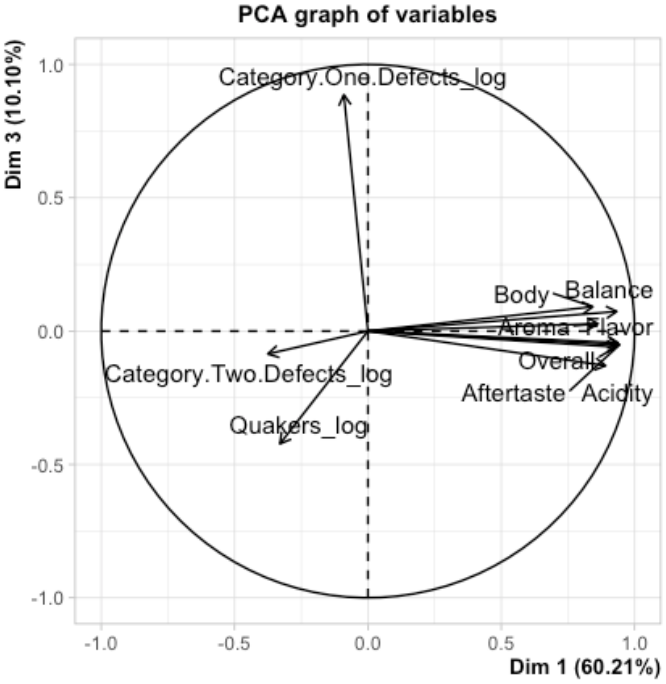


Figure K.6 : Graphe des variables pour la 3e composante



Interprétation :

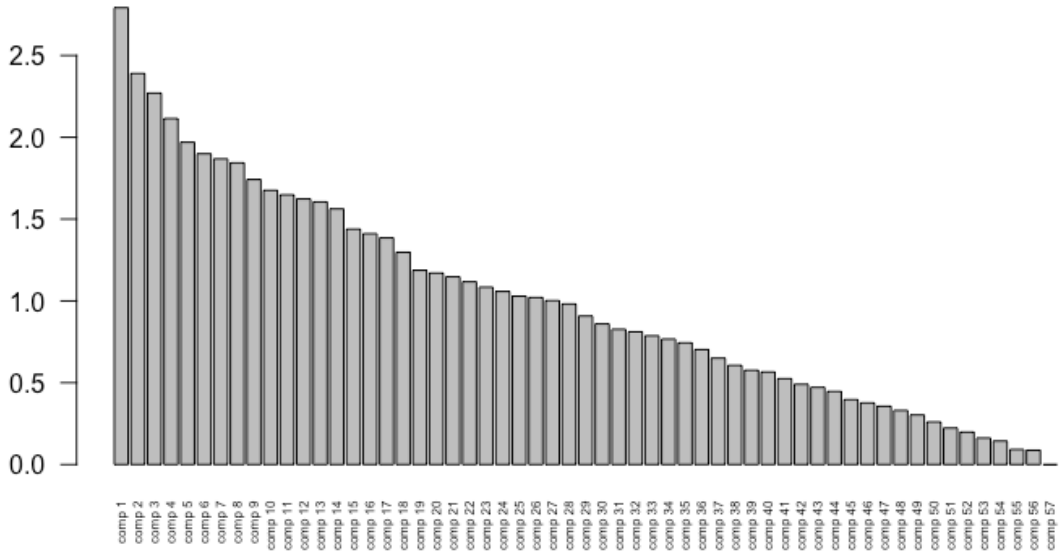
La 3e composante est fortement liée à la variable *Category.One.Defects* et dans ce plan qui rassemble plus de 70% de l'inertie, cette variable est indépendante des variables sensorielles.

L. TABLEAU DES VALEURS PROPRES ET DES INERTIES ASSOCIÉES

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.792	4.898	4.898
comp 2	2.39	4.193	9.091
comp 3	2.27	3.983	13.074
comp 4	2.115	3.71	16.784
comp 5	1.97	3.456	20.24
comp 6	1.899	3.332	23.572
comp 7	1.868	3.277	26.849
comp 8	1.844	3.235	30.084
comp 9	1.742	3.056	33.14
comp 10	1.676	2.94	36.08
comp 11	1.648	2.892	38.972
comp 12	1.623	2.848	41.82
comp 13	1.605	2.816	44.636
comp 14	1.562	2.741	47.377
comp 15	1.439	2.524	49.901
comp 16	1.411	2.475	52.376
comp 17	1.385	2.43	54.806
comp 18	1.297	2.276	57.082
comp 19	1.187	2.083	59.165
comp 20	1.17	2.053	61.218
comp 21	1.147	2.012	63.23
comp 22	1.117	1.959	65.19
comp 23	1.084	1.901	67.091
comp 24	1.058	1.856	68.947
comp 25	1.029	1.806	70.752
comp 26	1.021	1.791	72.543
comp 27	1.002	1.758	74.301
comp 28	0.981	1.722	76.023
comp 29	0.908	1.592	77.615
comp 30	0.859	1.508	79.122
comp 31	0.827	1.45	80.573
comp 32	0.811	1.423	81.996
comp 33	0.787	1.38	83.376
comp 34	0.767	1.345	84.721
comp 35	0.744	1.305	86.026
comp 36	0.703	1.233	87.259
comp 37	0.65	1.141	88.399
comp 38	0.606	1.063	89.463
comp 39	0.576	1.011	90.473
comp 40	0.566	0.992	91.466
comp 41	0.525	0.921	92.387
comp 42	0.491	0.861	93.247
comp 43	0.472	0.827	94.075
comp 44	0.447	0.784	94.858
comp 45	0.397	0.696	95.555
comp 46	0.377	0.661	96.216
comp 47	0.356	0.625	96.841

comp 48	0.331	0.58	97.421
comp 49	0.303	0.531	97.952
comp 50	0.26	0.457	98.409
comp 51	0.223	0.391	98.8
comp 52	0.198	0.348	99.148
comp 53	0.162	0.284	99.433
comp 54	0.144	0.252	99.685
comp 55	0.092	0.162	99.846
comp 56	0.088	0.154	100
comp 57	0	0	100

M. DIAGRAMMES DES VALEURS PROPRES.



N. VERIFICATION MATHEMATIQUE DU NOMBRE D'AXES NON TRIVIAUX

Le nombre de variables quantitatives considérées est de 3.

Le nombre de variables qualitatives considérées est de 4, avec 22, 24, 5 et 7 modalités.

Le nombre d'axes non triviaux est égal au nombre total de colonnes dans la matrice composée des trois variables quantitatives et du tableau de Burt, auquel on enlève une modalité par variable qualitative : $3 + 22 + 24 + 5 + 7 - 4 = 57$.

57 est bien le nombre d'axes obtenus dans les annexes I et J.

O. GRAPHES DES MODALITES DANS LES DEUX PREMIERS PLANS