Thème : Graine de café - Annexe

Spécialité : Analyse de données exploratoire

Yasemin AKDAG - Melissa Zennaf - Amadou Diakhate DIOP

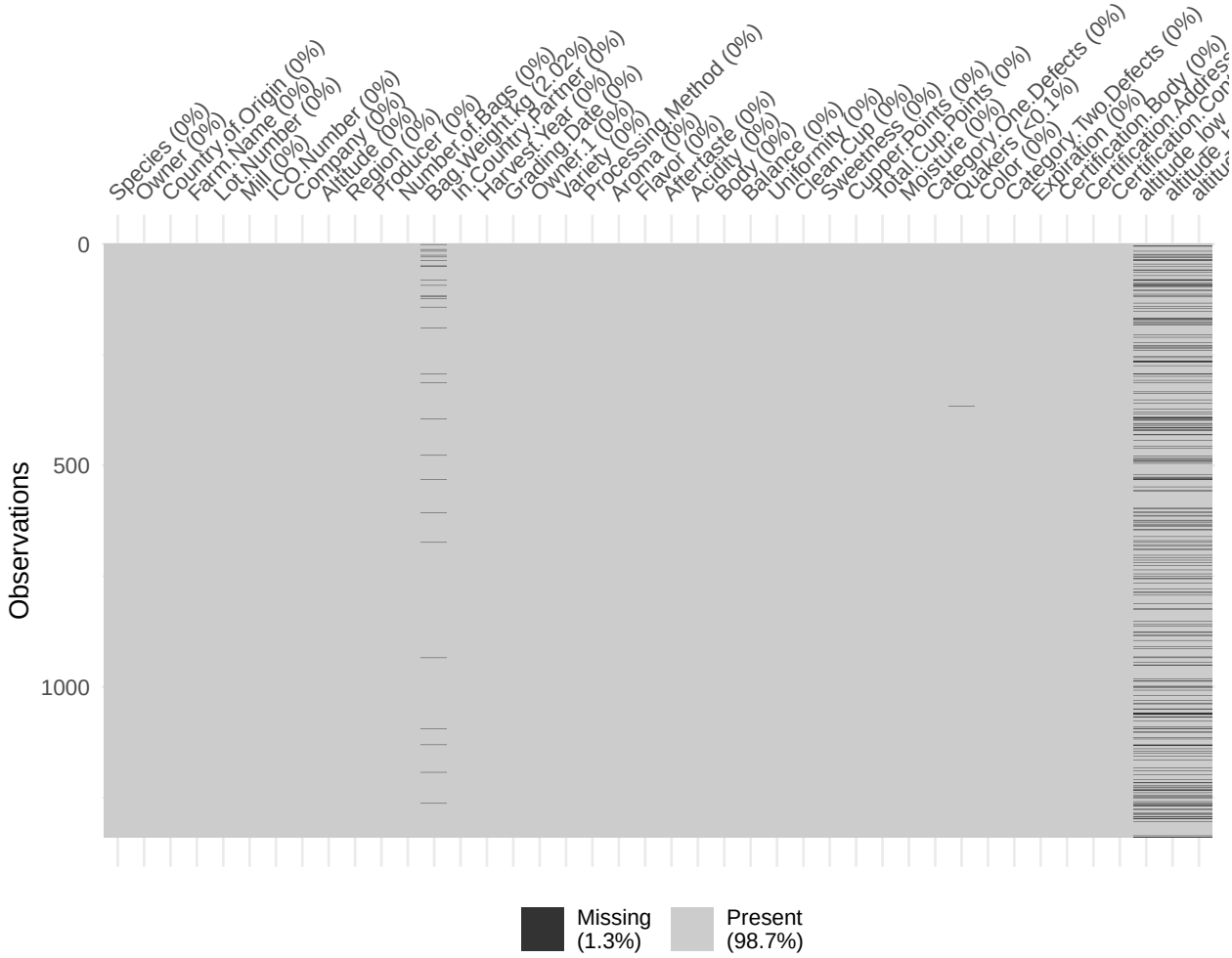
Contents

1 Études des données manquantes **2**

2 Analyse des variables quantitatives sur les sacs de café **5**
- 2.1 Analyse de la variable “Number of bags” . . . 5
- 2.2 Analyse de la variable “Bag.Weight.kg” . . . 5
- 2.3 Analyse des variables de scoring . . . 6
- 2.4 Analyse conjointe des variables “Points bonus” et “Score total” . . . 10
- 2.5 Analyse de la variable “Moisture” . . . 10
- 2.6 Analyse des variables de défauts: qualitatives ou quantitatives? . . . 11
  - 2.6.1 Analyse des variables “Category.One.Defects” et “Category.Two.defects” 11
  - 2.6.2 Analyse de la variables “Quakers”: . . . 12
- 2.7 Analyse des variables d’altitude . . . 13
  - 2.7.1 Analyse des variables “Altitude high meter” et “Altitude low meter” . 13
  - 2.7.2 Analyse de la variable “Altitude mean meters” . . . 14

3 Etude des variables qualitatives **15**
- 3.1 Analyse descriptive . . . 15
  - 3.1.1 Variables Species . . . 15
  - 3.1.2 Variables country.of.origin . . . 15
  - 3.1.3 Variables farm.name . . . 17
  - 3.1.4 Variable Company . . . 17
  - 3.1.5 Variables Region . . . 17
  - 3.1.6 Variable Producer . . . 18
  - 3.1.7 Variable country partner . . . 18
  - 3.1.8 Variable Harvest year . . . 19
  - 3.1.9 Variable grading date . . . 19
  - 3.1.10 Variables variety . . . 21
  - 3.1.11 Variable processing.method . . . 22
  - 3.1.12 Variable color . . . 22
  - 3.1.13 Variables certification.body, certification.address et certification.contact 23
- 3.2 Analyse multivariée . . . 24
  - 3.2.1 Analyse factorielle des correspondances - AFC . . . 24

1 Études des données manquantes



Parmi les 1339 observations, notre base de données possède 245 données manquantes. Pour la suite de l'analyse, nous utiliserons une nouvelle sous base de données , nommée ***data*** qui contiendra 1094 observations. Ce changement n'affectera pas le nombre total de variables étudiées.

Nous pouvons tout de même représenter la répartition de ces données pour la variables "Species" qui décrit le type de café étudié.

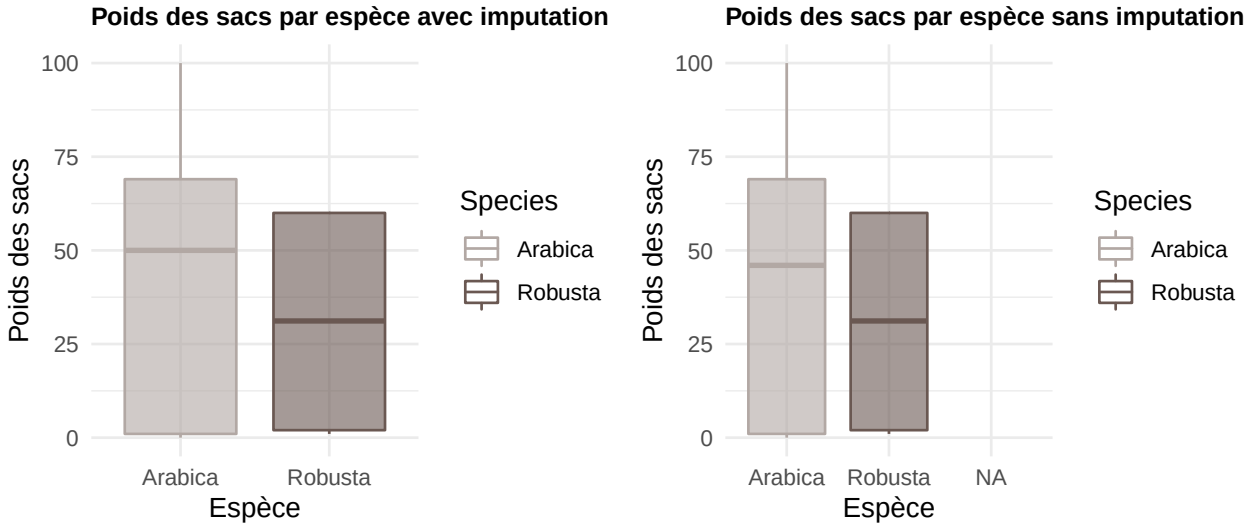
Nous pouvons observer que les données manquantes dans notre base de données se concentrent sur les variables d'altitude (altitude mean meters, altitude low meters, altitude high meters) et sur la variables bag weight km.

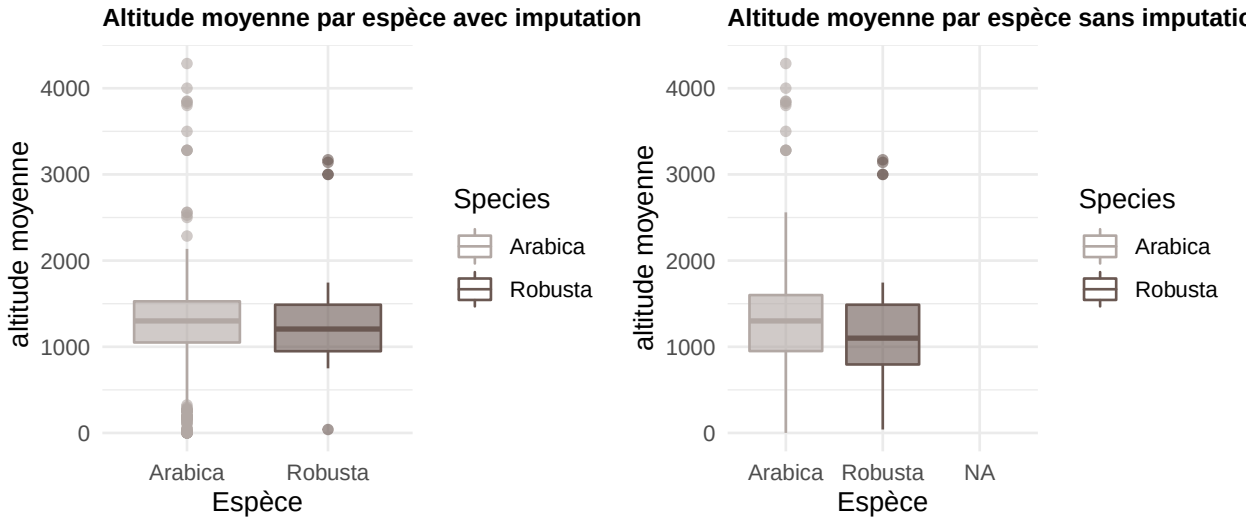
Cependant, il y a bien plus de données manquantes que le graphique nous le laisse entendre. En effet, les variables qualitatives contiennent de très nombreuses données manquantes mais celles-ci sont modélisées par des cases remplies d'un ou de plusieurs espaces, il est donc compliqué de les détecter. Dans l'idéal, il faudrait modifier directement notre base de données pour recoder les espaces en "NA" (avec VBA par exemple). Pour les variables qualitatives,

on conclue qu'il est impossible de les imputer car dans un premier temps, elles ne sont pas détectées. Dans un second temps, admettons qu'on impute la variable "pays d'origine" par la modalité la plus fréquente c'est-à-dire "Amérique", alors on risque de créer une incohérence avec d'autres variables comme par exemple la variable région qui est bien présente dans notre base de données pour cet individu et qui est une région africaine par exemple. L'algorithme d'imputation serait dans ce cas bien trop compliqué car il ferait intervenir trop de critères permettant de conserver la cohérence.

Les données manquantes existent de manière aléatoire, c'est-à-dire qu'on n'a pas trouvé de manière d'expliquer ou de prédire leur apparition. Toutes les données manquantes en Bag.weight manquantes ont aussi des données manquantes dans les variables d'altitude. Il y en a 13. Dans les variables altitudes, il y a 230 données manquantes. S'il y a une donnée manquante dans une des variables d'altitude, alors les autres variables d'altitude seront manquantes aussi. Ces entrées présentent aussi des données manquantes dans d'autres variables mais qui sont qualitatives.

Nous allons imputer les données quantitatives grâce à la fonction impute. Nous avons choisi d'imputer les données manquantes de notre base de données par la médiane de chaque variable. En effet, utiliser la moyenne est moins pertinent car il y a pour chaque variable des données très élevées qui tirent trop la moyenne vers le haut et va donc biaiser nos imputations. En utilisant la médiane, nous obtenons les réultats suivants sur nos variables avant et après imputation :

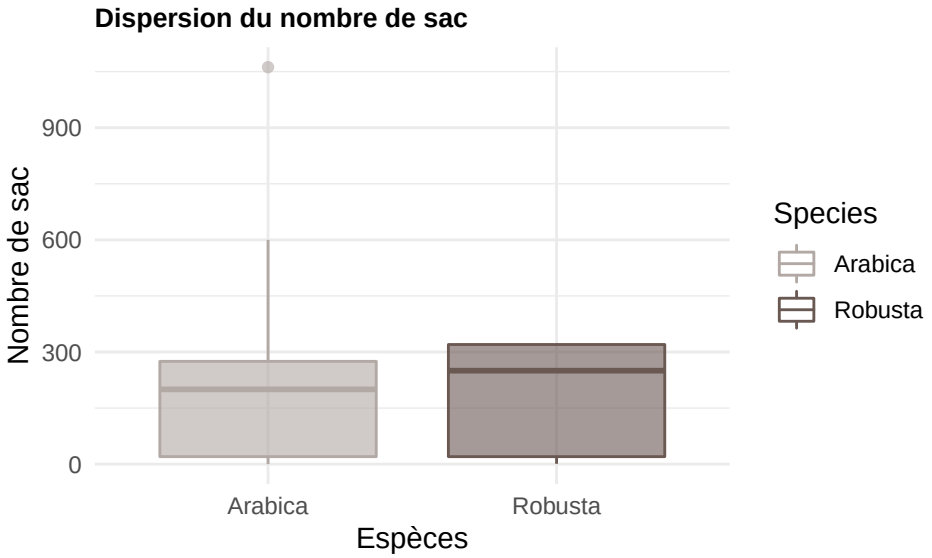




On voit qu'imputer les données manquantes par la médiane ne modifie que très légèrement la structure des données pour les variables concernées. Cela est le signe d'une imputation réussie.

2 Analyse des variables quantitatives sur les sacs de café

2.1 Analyse de la variable "Number of bags"

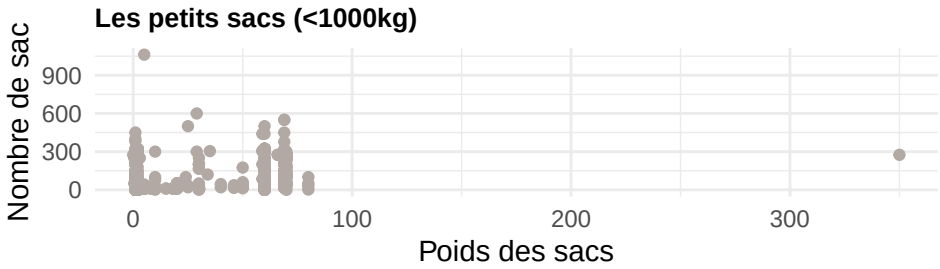


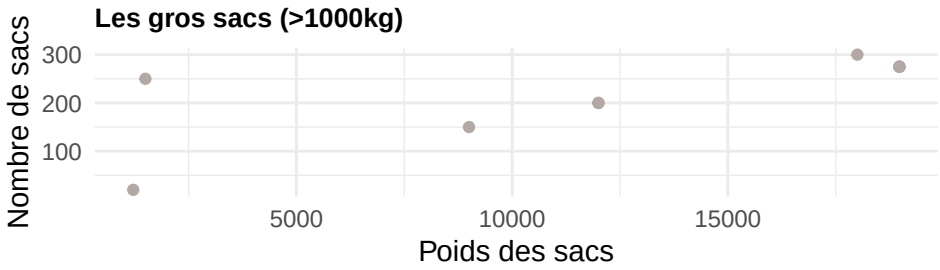
Le nombre de sac pour les cafés robusta est un peu plus élevé que pour les café arabica. Il y a des variables très grandes et des variables proche de zéro.

2.2 Analyse de la variable "Bag.Weight.kg"

Pour une meilleure représentation des données, nous allons séparer nos données en deux groupes. Le premier groupe rassemble les cafés dont les conditionnements ont un poids supérieur à une tonne, le second rassemble les cafés dont le conditionnement est inférieur à une tonne.

	sup	inf
Moyenne	12961.80	36.70
Ecart type	7124.71	32.58
Minimum	1218.00	0.50
Q1	9750.00	1.00
Médiane	15000.00	60.00
Q3	18975.00	69.00
Maximum	18975.00	350.00





Nous avons donc deux groupes qui se distinguent. Le tableau ci-joint présente les statistiques de la quantité produite dans chaque groupe (poids multiplié par le nombre de sacs). On peut observer que les cafés conditionnés en petits sacs ont une quantité produite généralement moins importante que les cafés dans des gros conditionnements. Cela semble logique car on peut s'imaginer qu'il existe des grosses fermes qui produisent beaucoup et donc conditionnent leurs cafés par tonnes alors que les petits producteurs utilisent de plus petits conditionnements.

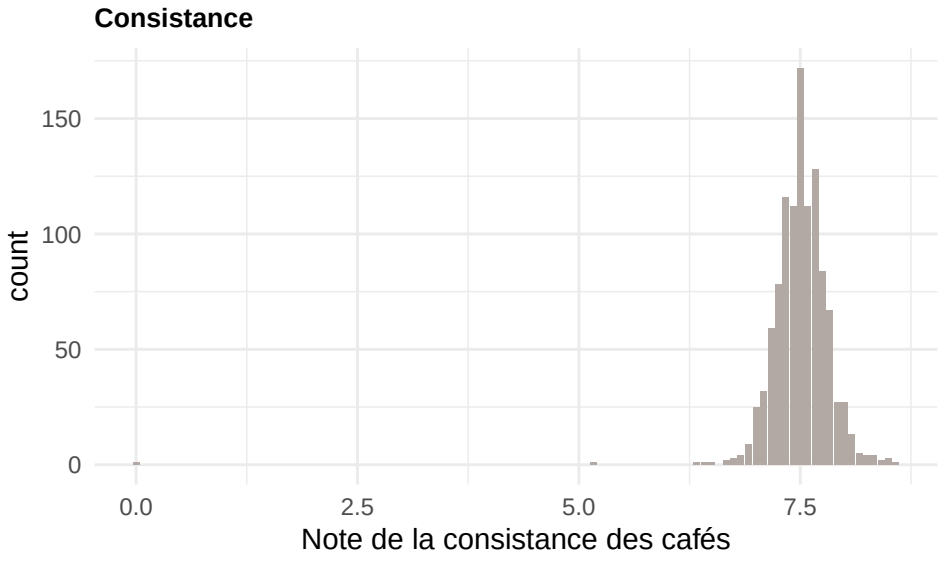
	GrosSacs	PetitsSacs
Moyenne	3282186	6676.620
Ecart type	2206848	8716.357
Minimum	24360	1.000
Q1	1612500	150.000
Médiane	3809062	640.000
Q3	5218125	17250.000
Maximum	5400000	96250.000

2.3 Analyse des variables de scoring

La notation du café, selon la méthode brésilienne, est basée sur plusieurs critères et suit un processus bien particulier. Les critères sont les suivants et sont des notes sur 10 attribuées par un goûteur.

	Arome	Parfum	Arriere gout	Acidite	Consistance	Equilibre	Uniformite	Nettete	Douceur	Points bonus	Score total
Moyenne	7.57	7.52	7.39	7.52	7.50	7.50	9.87	9.85	9.87	7.48	82.08
Ecart type	0.38	0.40	0.40	0.38	0.37	0.42	0.52	0.79	0.60	0.47	3.60
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q1	7.42	7.33	7.25	7.33	7.33	7.33	10.00	10.00	10.00	7.25	81.17
Médiane	7.58	7.50	7.42	7.50	7.50	7.50	10.00	10.00	10.00	7.50	82.50
Q3	7.75	7.75	7.58	7.75	7.67	7.73	10.00	10.00	10.00	7.75	83.58
Maximum	8.75	8.83	8.67	8.75	8.58	8.75	10.00	10.00	10.00	10.00	90.58

Nous allons faire l'étude détaillée de la variable "Consistance". Le corps du café (ou sa consistance) peut être défini comme la sensation tactile du café sur la langue.



Ce graphique rapide nous permet de constater que la plus grande partie des cafés se trouvent avec des notes entre 7 et 8. Plus précisément entre 7,5 et 8.

Nous constatons que nous avons une seule donnée avec un score de 0, cela semble étrange. On s'aperçoit que ce café a une note de 0 pour tous les critères. Cela semble logique de se dire que cette ligne de la dataframe n'est pas à prendre en compte dans notre analyse car les notes ne reflètent pas la réalité de la qualité de ce café. On va l'omettre et réobserver le résumé des variables de scoring:

	Arome	Parfum	Arriere gout	Acidite	Consistance	Equilibre	Uniformite	Nettete	Douceur	Points bonus	Score total
Moyenne	7.58	7.52	7.40	7.53	7.51	7.51	9.88	9.86	9.88	7.49	82.16
Ecart type	0.31	0.33	0.34	0.31	0.29	0.35	0.43	0.73	0.53	0.41	2.61
Minimum	5.08	6.17	6.17	5.25	5.17	5.25	6.00	0.00	1.33	5.17	59.83
Q1	7.42	7.33	7.25	7.33	7.33	7.33	10.00	10.00	10.00	7.25	81.17
Médiane	7.58	7.50	7.42	7.50	7.50	7.50	10.00	10.00	10.00	7.50	82.50
Q3	7.75	7.75	7.58	7.75	7.67	7.75	10.00	10.00	10.00	7.75	83.58
Maximum	8.75	8.83	8.67	8.75	8.58	8.75	10.00	10.00	10.00	10.00	90.58

Les minimums que l'on observe sont logiquement modifiés, mais les moyennes aussi. De même les écarts-types sont assez différents.

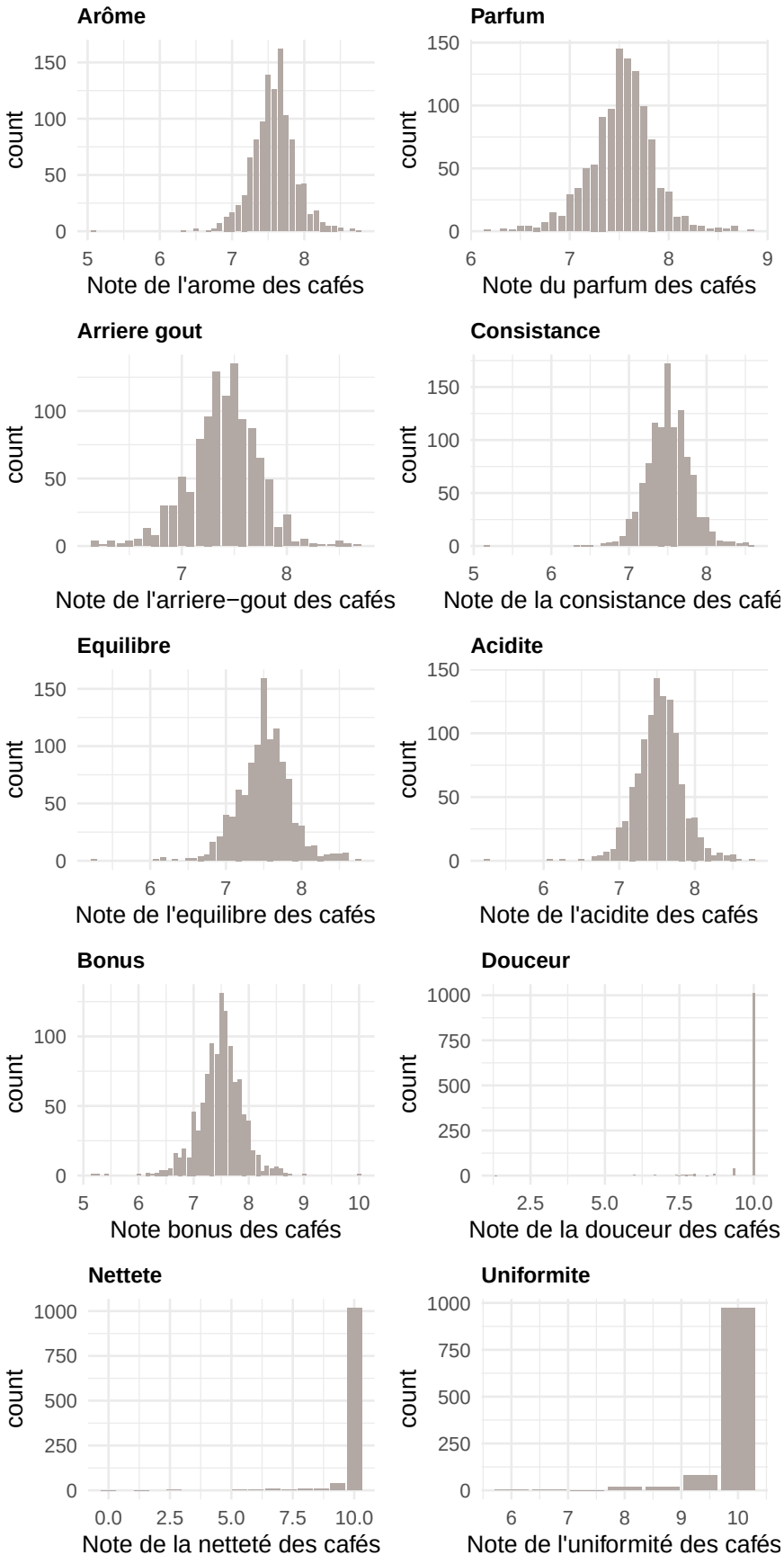
En ce qui concerne les autres critères de notation, ils peuvent être observés comme suit :

	Douceur	Equilibre	Consistance	Uniformité	Arome	Acidité	Arriere-gout	Saveur	Netteté	Bonus
bon	19	84	59	35	99	83	42	75	16	101
excellent	1048	56	22	1051	27	27	101	49	1054	2
mauvais	4	953	1012	6	967	983	950	969	16	83
moyen	21	84	59	1	99	83	42	75	3	907
sans notation	1	56	22	35	27	27	101	49	4	101

Ici, les notes ont été découpées en classes selon les critères suivants :

- les notes entre 0 et 5 portent la modalité "sans notation"
- les notes entre 5 et 7 portent la modalité "mauvais"

- les notes entre 7 et 8 portent la modalité “moyen”
- les notes entre 8 et 9 portent la modalité “bon”
- les notes entre 9 et 10 portent la modalité “excellent”

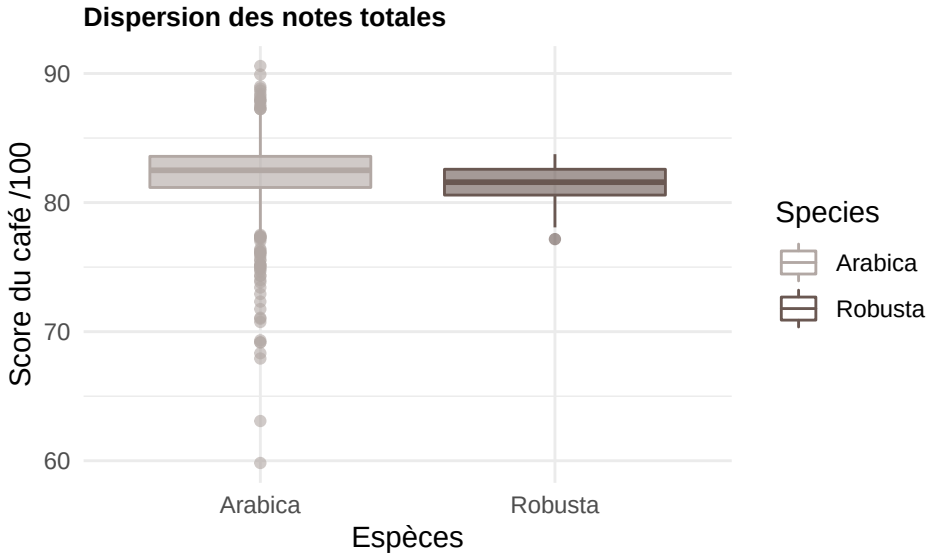

Arôme
count
Note de l'arome des cafés
Parfum
count
Note du parfum des cafés
Arriere gout
count
Note de l'arriere−gout des cafés
Consistance
count
Note de la consistance des café
Equilibre
count
Note de l'equilibre des cafés
Acidite
count
Note de l'acidite des cafés
Bonus
count
Note bonus des cafés
Douceur
count
Note de la douceur des cafés
Nettete
count
Note de la netteté des cafés
Uniformite
count
Note de l'uniformité des cafés

2.4 Analyse conjointe des variables “Points bonus” et “Score total”

Le score total est la note obtenue par le café en question à la fin de son évaluation, c’est en fait la somme des scores obtenus dans les différents critères.

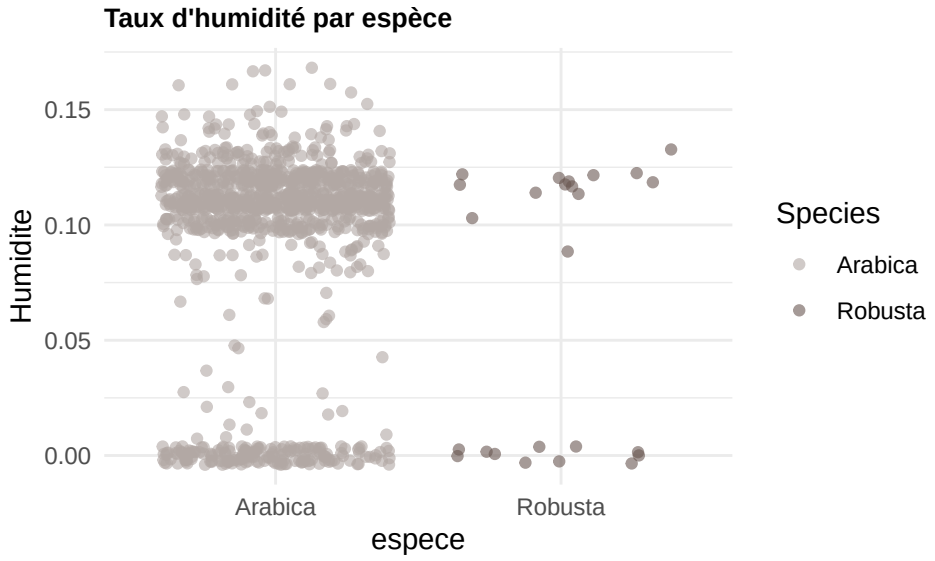
En ce qui concerne la variable points bonus, il s’agit en fait là d’une notation subjective réalisée par le goûteur pour qu’il rende compte de son impression générale sur son expérience de dégustation. La présence de cette variable permet de corriger les écarts qu’il peut exister entre la note du café et l’expérience gustative qu’il a procuré au goûteur.

Les notes influencent positivement le score total et inversement. Cela est d’ailleurs parfaitement logique. La question que l’on peut se poser est donc sur l’intérêt de conserver ces plusieurs variables qui sont très corrélées entre elles et qui nous montrent donc quasiment la même chose. Cela dépendra en effet de la problématique à laquelle nous aurons à répondre, si il est juste question de rendre compte de la note du café, nous n’aurons pas besoin de conserver toutes les variables mais seul le score total sera nécessaire. Au contraire, si le détail des notes est très important dans l’étude, on conservera toute la précision de l’analyse. Nous allons terminer cette étude des variables de scoring par le graphique suivant.



2.5 Analyse de la variable “Moisture”

Un grain de café typique, non transformé, doit contenir environ 45 à 55% d’ humidité après la récolte. La transformation et le séchage ramèneront alors leur teneur en humidité à 10-12%.Nous allons voir si les données observées correspondent à cela.



Effectivement, l'humidité des grains de cafés tourne autour des valeurs attendues (10%). Nous nous demandons quand même comment il se fait qu'il y ait des grains de cafés dont l'humidité est de 0%. En fait, la cible d'humidité est de 11% mais l'intervalle de 10:12% est très acceptable. L'humidité du café n'est pas d'influence directe sur son gout mais un café très sec aura tendance à perdre de sa qualité alors que des taux d'humidité plus élevés favorisent le développement de moisissures.

2.6 Analyse des variables de défauts: qualitatives ou quantitatives?

Certaines de nos variables quantitatives sont des indicateurs de la qualité du café en mettant en valeur les défauts constatés dans les cafés.

2.6.1 Analyse des variables "Category.One.Defects" et "Category.Two.defects"

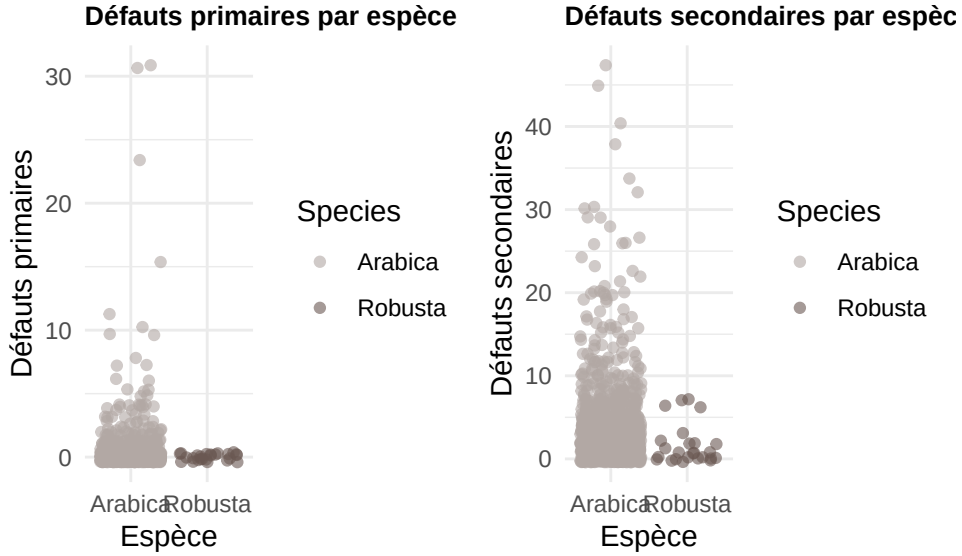
Les défauts de catégorie un ou défauts primaires sont des défauts très importants qui relèvent des défauts intrinsèques du café.Pour les défauts de catégorie deux ou défauts secondaires sont des défauts ayant des incidences sur la qualité du café mais qui ne sont pas totalement rédhibitoires.

Suite à l'observation de ces graphiques nous savons que les variables Category.One.Defect et Category.Two.Defects sont des variables discrètes. Nous sommes tout de même face à un questionnement. Faut -t-il considérer ces variables comme qualitatives ou comme quantitatives?

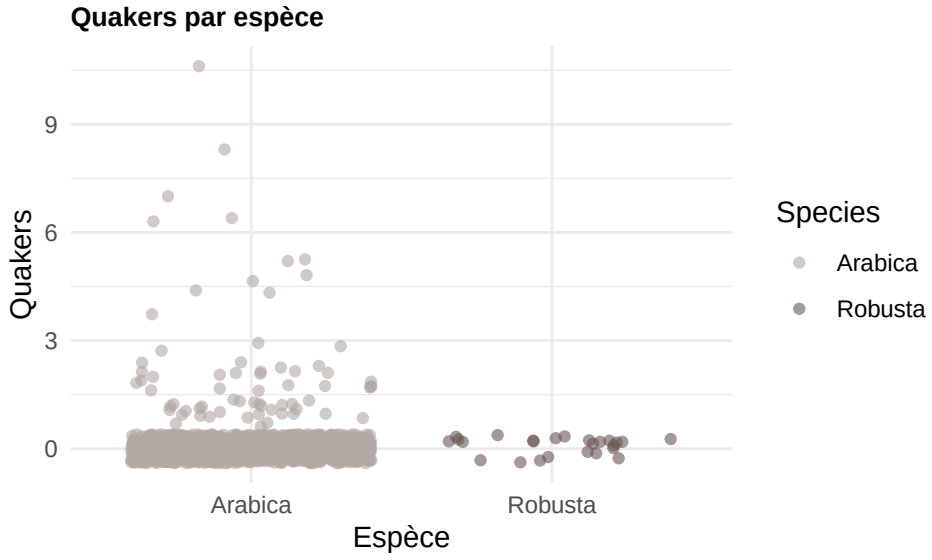
- Si elles sont qualitatives, cela signifie que les entrées du tableau (1,2, etc) correspondent à un codage pour les défauts observés. En effet, il existe des classements des cafés en fonction des défauts primaires: Noir complet, Full Sour, Pod / Cerise: 1 ; Grosses pierres, Pierres moyennes: 2 ; Bâtons moyens, Gros bâtons: 5. Ce classement varie d'un pays ? l'autre ce qui peut expliquer la présence de chiffres différents comme le 3, le 4. La présence d'autres nombres peut être expliquée par des erreurs de recopie

que l'on peut interpréter comme des NAs. Pour ce qui concerne les défauts secondaires voici le classement: Parchemin, Coque / coque, Noir partiel, Sour partiel: 2-3 ; Cassé / ébréché, Flotteur, coquille: 5 ; Dommages causés par les insectes: 2-5 ; Petites pierres ,Petits bâtons: 1 ; Dégâts d'eau: 2-5.

- Si cette option ne correspond pas à la réalité, alors il faudra interpréter ces deux variables comme des variables quantitatives et dans ce cas là, l'entrée correspond au nombre de défauts primaires ou secondaires observés dans un échantillon. Le problème dans ce cas est que nous ne connaissons pas le nombre d'individus observés et que cela peut tromper nos interprétations. Si cette option est choisie par la suite, nous partirons du principe que l'effectif de l'échantillon est le même pour chaque café, sinon il sera très difficile de se servir de ces deux variables pour nos analyses.



2.6.2 Analyse de la variables "Quakers":



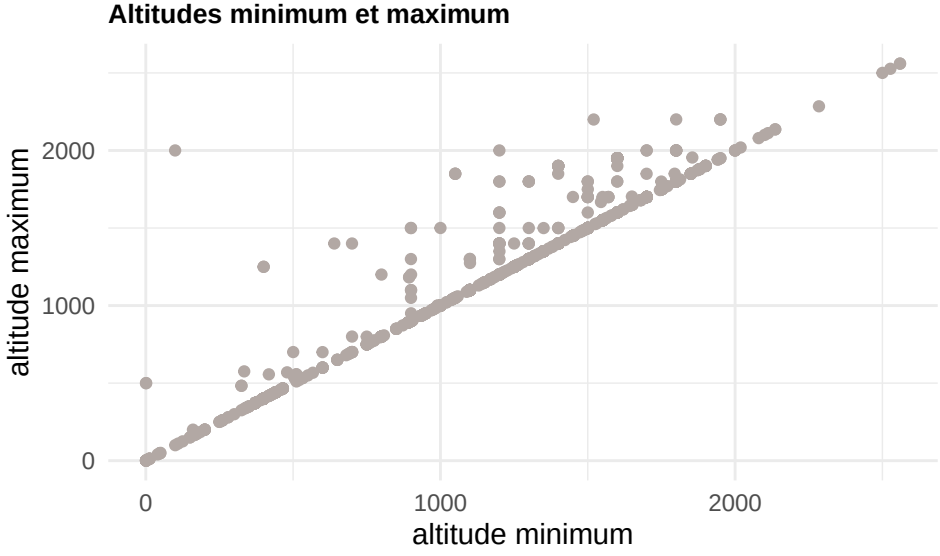
La variable “quakers” est une variable discrète. Là aussi la question de son interprétation se pose. Un quaker est une cerise de café récoltée avant sa maturité. On se demande si il faut l'interpréter comme une variable qualitative: un indice délivré à ce café correspondant au défaut qu'il présente (moisi, trop fermenté etc. . .) ou alors si il s'agit simplement du nombre de quakers constaté dans un échantillon et dans ce cas là d'une variable quantitative. Nous avons le même problème que pour les deux variables précédents: nous ne connaissons pas la taille de l'échantillon.

2.7 Analyse des variables d'altitude

2.7.1 Analyse des variables “Altitude high meter” et “Altitude low meter”

	lowmeters	highmeters
Moyenne	1668.98	1713.82
Ecart type	8736.91	8736.72
Minimum	1.00	1.00
Q1	934.00	950.00
Médiane	1255.00	1300.00
Q3	1554.50	1600.00
Maximum	190164.00	190164.00

Or, le caféier vit une cinquantaine d'années et ne pousse que dans les régions tropicales, à des altitudes qui varient de 200 à 2 000 mètres. Nous avons des données qui nous semblent très extrêmes. Nous savons également que le mont le plus haut sur Terre est le mont Everest qui monte jusqu'à 8800 mètres au dessus du niveau de la mer à peu près. Nous pouvons omettre les données supérieures à 3000 mètres.



Ce graphique nous permet de constater que la plupart des données ont un x et un y (altitude minimale et altitude maximale respectivement) qui sont égaux, c'est-à-dire que l'altitude min et max sont égales.

Nous avons donc revu la dispersion des données d'altitude en excluant les données supérieures à 3000 mètres car nous avons constaté que le café est extrêmement difficile à cultiver au delà

de cette altitude. Mais nous avons également des cafés qui sont cultivés à une hauteur de 1 mètre au dessus du niveau de la mer, ce qui semble étrange. .

2.7.2 Analyse de la variable “Altitude mean meters”

L’ensemble des données de l’altitude moyenne correspondent aux données observées dans les catégories “altitude low meters”" et “altitude high meters”. La question que l’on peut se poser est celle de l’utilité de cette dernière variable qui est en fait la simple moyenne des deux variables précédentes. Elle peut donc aisément être calculée a l’aide de R. Il faut donc faire un choix entre l’utilisation conjointe des colonnes “altitude low meters” et “altitude high meters” et l’utilisation seule de la variable “altitude mean meters” car les informations qu’elles nous fournissent sont les mêmes.De même, nous avons choisi d’exploiter ces trois variables plutôt que la variable “Altitude” qui présente des formats trop divers et qui n’est pas au format numérique mais cette variable nous fourni les mêmes informations.

3 Etude des variables qualitatives

3.1 Analyse descriptive

3.1.1 Variables Species

La variable Species represente les deux espèces de café étudié, Arabica et Robusta. Cette différence est obtenue par une formation de graine sur des arbustes différents ayant des critères de formations apportant des spécificitées à chaque graines qui seront aussi étudiée par la suite.

- Répartition selon les espèces

Species	Effectifs	Proportions
Robusta	25	2.3
Arabica	1069	97.7

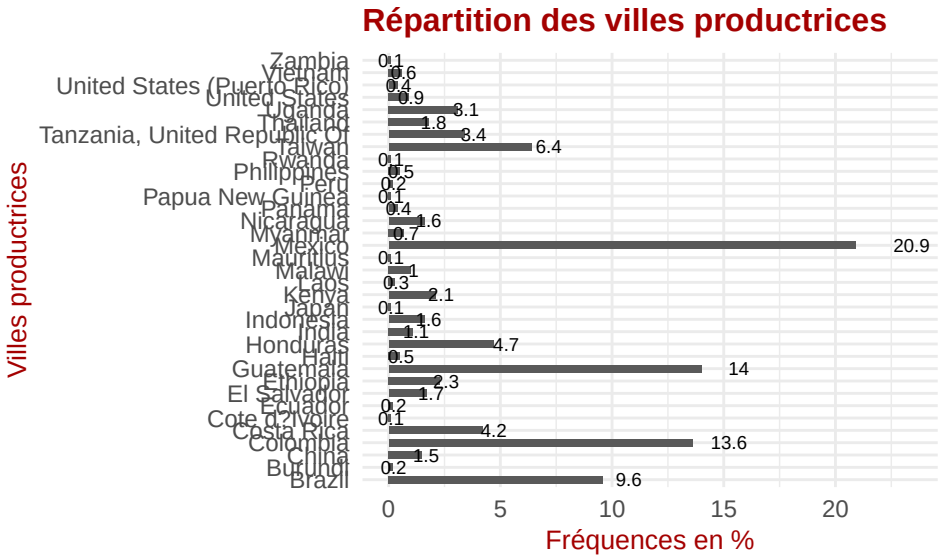
Suite à ce changement, nous remarquons que les proportions n'ont pas été affectés.

3.1.2 Variables country.of.origin

Le tableau suivant illustre les pays producteurs de cafés.

- Effectis et proportions des différentes villes productrices

Country.of.Origin	Effectifs	Proportions
Mexico	229	20.9
Guatemala	153	14.0
Colombia	149	13.6
Brazil	105	9.6
Taiwan	70	6.4
Honduras	51	4.7



Les pays ayant une forte production sont

- le Méxique,
- le Guatemala ,
- la Colombie,
- le Brésil,
- le Taiwan.

Soit 64.5% de la production mondiale.

Pour cette variable, il serait aussi interessant d'étudier la production du café selon les continents. Pour cela, nous avons regroupé les pays selon leur emplacement géographique. Nous obtenons donc quatre sous-groupes qui sont : l'Amérique, l'Afrique, l'Asie et l'Océanie.

- Effectifs et des fréquences de la variable country selon les continents

	Effectifs	Fréquence
Afrique	136	12.4
Amérique	797	72.9
Asie	160	14.6
Océanie	1	0.1

Selon la répartition obtenue au niveau continental, nous apercevons que le continent américain regroupe près de 73 % de la production mondiale.

3.1.3 Variables farm.name

La variable farm.name rassemble le nom des fermes produisant les graines de café.

- Répartition des fermes produisant les graines de café

Farm.Name	Effectifs	Proportions
various	44	4.0
rio verde	23	2.1
several	20	1.8
finca medina	15	1.4
fazenda capoeirnha	13	1.2
los hicaques	11	1.0
capoeirinha	10	0.9
el papaturro	9	0.8
agropecuaria quiagral	8	0.7
cerro bueno	8	0.7
el morito	8	0.7

Parmi les 572 fermes , celle ayant la plus grande part de production est la ferme “Various” avec 4 %. Cet étude de variable nous permet de constater qu'il y a énormément de petites fermes indépendantes. C'est-à-dire que chaque personne ayant un terrain possédant des arbustes de café peut effectivement faire partie du marché des graines de café par le biais des coopératives ou bien leur entreprise.

3.1.4 Variable Company

Nous observons grâce à cette variable la répartition des entreprises productrices. Nous remarquons qu'il y a 282 entreprises ce qui est nettement plus inférieures au nombre de fermes. Cette différence nous indique qu'il y a peut être une liaison entre les fermes et les entreprises.

3.1.5 Variables Region

- Repartition selon les regions productrices

Region	Effectifs	Proportions
huila	93	8.5
oriente	66	6.0
south of minas	66	6.0
veracruz	31	2.8

Les régions les plus productives sont

- la Huila (Colombie) ,
- la South of Minas (Brésil),
- l'Oriente (Équateur),
- la Veracruz (Méxique).

Ces territoires illustrent bien la forte productivité présente en Amérique latine.

3.1.6 Variable Producer

Producer est la variable qui regroupe les coffee shops mettant à disposition les graines de café. Représenter cette variable ne serait pas significatif car nous avons 694 coffee shops différents.

3.1.7 Variable country partner

La variable country.partner représente les coopératives (ou associations) qui permet aux entreprises ou bien les fermes à intégrer le marché dédié aux graines de café.

- Répartition des coopératives ou associations

In.Country.Partner	Effectifs	Proportions
AMECAFE	204	18.6
Specialty Coffee Association	183	16.7
Almacafé	148	13.5
Asociacion Nacional Del Café	135	12.3
Instituto Hondureño del Café	57	5.2
Blossom Valley International	53	4.8
Brazil Specialty Coffee Association	48	4.4
Africa Fine Coffee Association	47	4.3
Specialty Coffee Association of Costa Rica	41	3.7
NUCOFFEE	31	2.8

Pour avoir une meilleur lecture des régions partenaires des fermes productives, nous allons créé des sous-groupes regroupant les associations et coopératives selon leur appartenances continentale.

- Effectifs et fréquences de la répartition des coopératives selon les continents

	Effectifs	Fréquence
Afrique	139	12.7
Amérique	696	63.6
Asie	68	6.2
Coffee Quality Institute	7	0.6
Océanie	1	0.1
Specialty Coffee Association	183	16.7

Nous remarquons une forte similarité de répartition avec la variable country. Ceci nous informe donc sur la gestion locale des productions de graines de café. En revanche, lors des regroupements deux catégories, Speciality Coffee Association et Coffee Quality Institute, n'ont pas pu être intégrées dans des continents précis car ils possèdent des sièges dans plusieurs pays et donc dans différents continents.

3.1.8 Variable Harvest year

Dans cette partie, nous allons étudier l'années de production des graines de café.

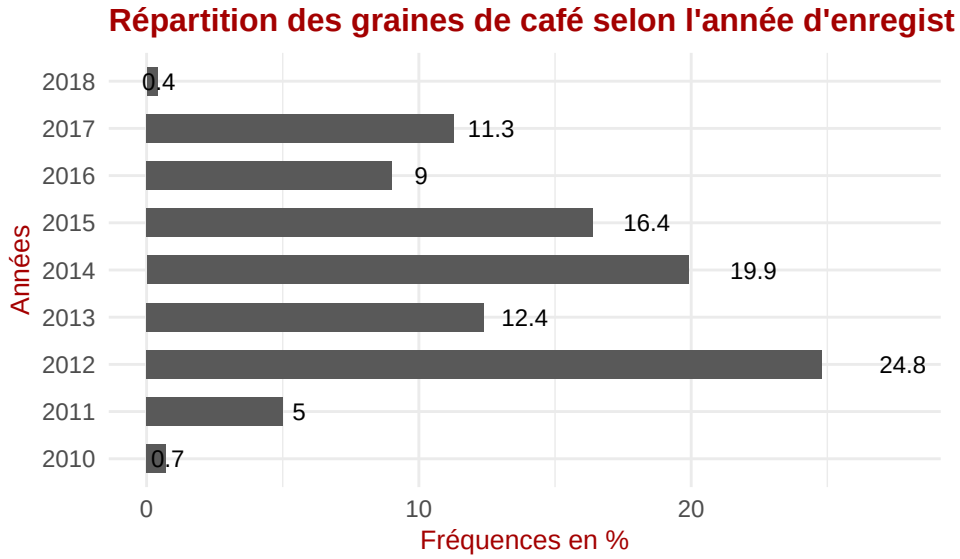
- Effectifs et fréquences de la répartition de l'année de récolte

	Effectifs	Fréquence
2009-2011	55	5.0
2012-2014	700	64.0
2015-2018	323	29.5
other	16	1.5

La récolte des graines de café était le plus élevée entre les années 2012-2014 et environs de moitié entre 2015-2018. Nous constatons donc une diminution de la production de graines de café.

3.1.9 Variable grading date

Grading date correspond au classsement des graines de café. Selon les informations obtenues, cela équivaut à la date d'enregistrement de la récolte car nous retrouvons une répartition semblable aux années de récolte.

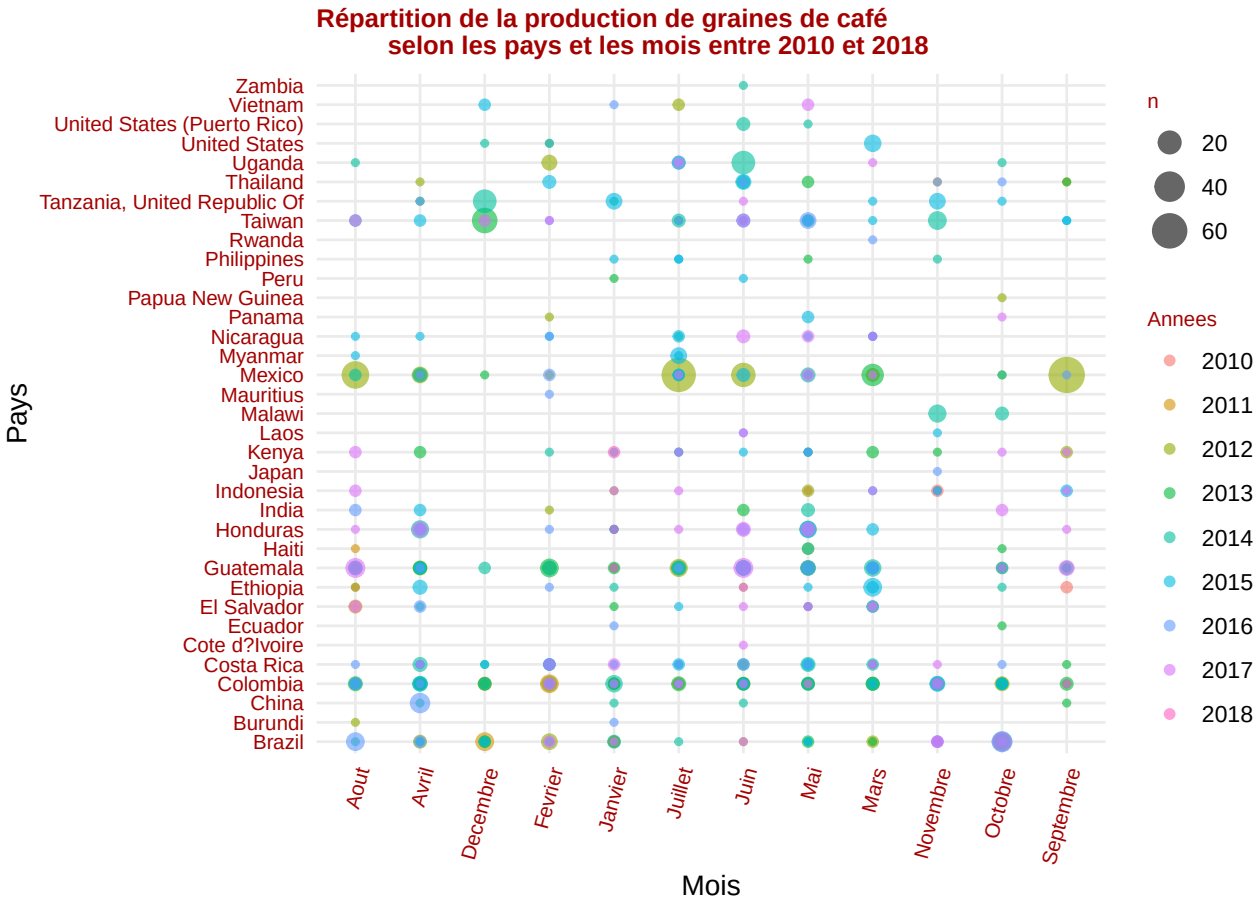


Les années entre 2012 - 2015 représente 73,5% des enregistrements de récolte effectués selon notre base de données. Durant cette période, nous pouvons remarquer un pic de production. En effet, après 2015 nous observons une diminution des enregistrements.

- Répartition des enregistrments selon les mois

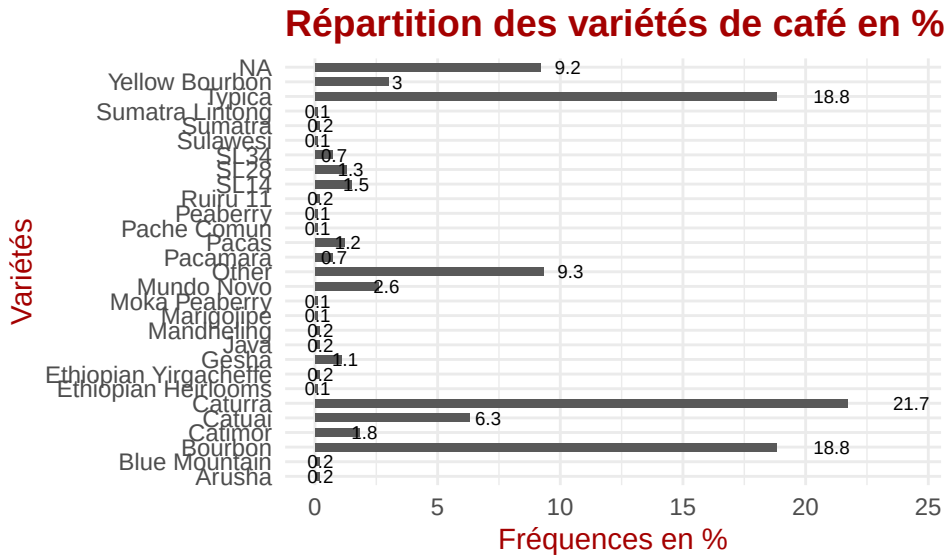
Mois	Effectifs	Proportions
Juin	134	12.2
Juillet	126	11.5
Mai	105	9.6
Avril	102	9.3
Aout	99	9.0
Septembre	97	8.9
Mars	90	8.2
Decembre	77	7.0
Fevrier	76	6.9
Octobre	73	6.7
Novembre	59	5.4
Janvier	56	5.1

Le mois de juin et de juillet sont les deux où les enregistrements sont les plus élevés mais nous pouvons remarquer qu'il y a des récoltes tout au long de l'année. Mais le pic d'enregistrement nous laisse croire que la récolte des graines de café se fait plutôt durant les mois d'été ce qui varie énormément selon la situation géographique des zones de récolte.



Les mois où les enregistrements ont été les plus élevés sont le mois de septembre 2012 et juillet 2012 au Méxique.

3.1.10 Variables variety



Les variétés “Caturra”,“Typica” et “Bourbon” sont des différentes variétés de cafés obtenues

par des mutations ou croisements des gênes des graines de café dans le but de diversifier les saveurs et arômes. Dans notre base de données, nous avons donc 29 variétés de café.

3.1.11 Variable processing.method

Dans cette partie, nous analysons les différentes méthodes qui permettent d'éliminer les fruits de la graine, la peau, la pulpe, le parchemin et la peau d'argent. Ce procédé permet à l'agriculteur de récupérer la graine principale du café.

- Répartition des méthodes de nétoyage des graines fraiches de café

Processing.Method	Effectifs	Proportions
Washed / Wet	740	67.6
Natural / Dry	184	16.8
Other	107	9.8
Honey / Semi-washed	63	5.8

Il y a plusieurs méthodes:

- Processus lavé (ou humide) (washed / wet) : Dans les pays où l'eau est abondante, les cerises sont généralement passées dans une machine à pâte, triées en fonction du poids et déposées dans une cuve de fermentation. Ici, des enzymes naturels dissolvent la pulpe jusqu'à ce qu'elle puisse être lavée de la fève, un processus qui prend 12 à 72 heures en fonction de nombreux facteurs tels que la température et l'humidité. Une fois fermenté, le café lavé, toujours dans son parchemin (filmeux, revêtement semblable à du papier), est étendu jusqu'à ce qu'il atteigne environ 11% d'humidité. À ce stade, la graine est stabilisée et ne germe pas.
- Miel (semi-lavé) (pulped naturel / honey): le café est mis à sécher au soleil et le fruit adhère à la fève. Cela ressemble au style traditionnel de traitement brésilien «pulp-natural».
- Procédé naturel (ou sec) (natural /dry): dans d'autres pays où l'eau n'est pas aussi facilement disponible, les cerises fraîchement cueillies sont réparties sur des bâches, des patios et même parfois sur la route, chaque fois que le fruit peut mieux sécher au soleil. Atteindre une teneur en humidité optimale peut prendre des semaines. Tout au long de ce processus de séchage, les cafés sont régulièrement retournés avec des râteaux pour assurer un séchage uniforme.

Nous remarquons que la méthode la plus utilisée est le processus lavé, c'est-à-dire washed / wet.

3.1.12 Variable color

- Répartition des graines selon leurs couleurs

Color	Effectifs	Proportions
Green	769	70.3
Other	171	15.6
Bluish-Green	83	7.6
Blue-Green	71	6.5

Comme prévu, notre échantillon comporte 70,3 % de graine de café de couleur verte. Blue-Green et Bluish-Green correspondent au couleur naturelle des graines avant la torréfaction.

3.1.13 Variables certification.body, certification.address et certification.contact

D'après les données trouvées dans notre base de donnée concernant les certifications, elles ne peuvent pas pour le moment nous renseigner de certains caractéristiques du café. Par ailleurs, les organismes de certifications pourraient être utiles pour distinguer ceux qui dominent le marché du café en terme de certification et les conséquences qu'elles généreraient sur la qualité du café en fonction des organismes.

3.2 Analyse multivariée

3.2.1 Analyse factorielle des correspondances - AFC

3.2.1.1 AFC des varibales Country et Mois

3.2.1.1.1 Test d'indépendance des variables Country et Mois

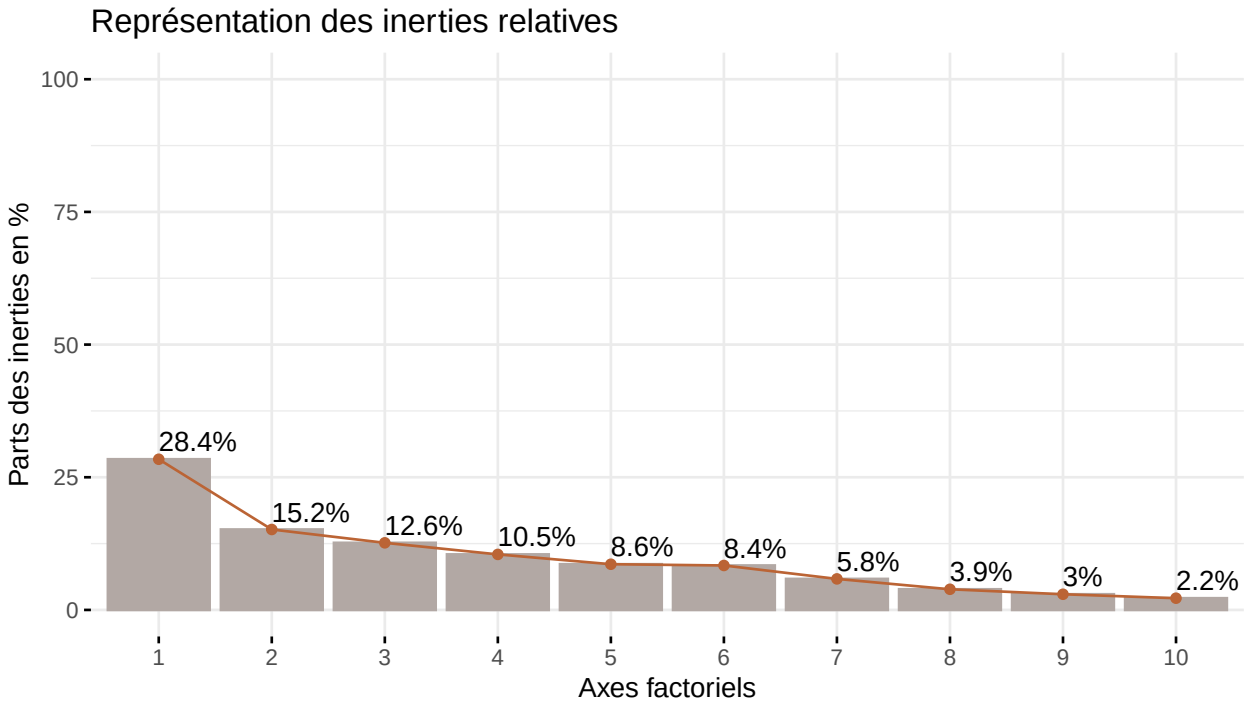
- Résultats du test du Khi2

X2obs	1628.657
df	374.000
pvalue	0.000

Comme la p-value est relativement faible, nous rejettons l'hypothèse H0 d'indépendance des variables Country et Mois. Ces deux variables représentent les pays et le nombre d'enregistrement des récoltes par mois.

Réalisons maintenant l'AFC.

3.2.1.1.2 Valeurs propres - Inerties des axes

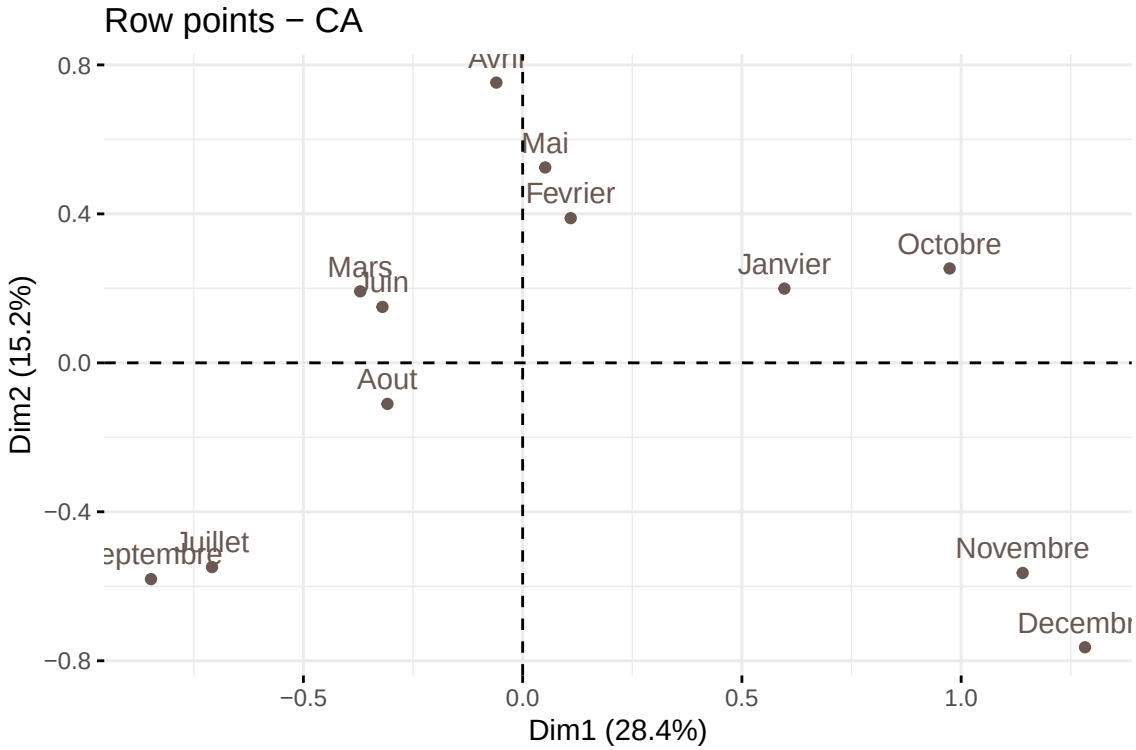


Les deux premiers axes regroupent environs 44% des parts d'inertie totale ce qui reste légèrement faible mais étudier les autres axes n'apporteront pas forcément des informations supplémentaires pertinantes. Notre étude pours ces deux variables se basera donc seulement sur les deux premiers axes.

3.2.1.1.3 Profils lignes - Mois

- Caractéristiques des profils lignes - Mois

	F1	F2	ctr(F1)	ctr(F2)	qlt(F1)	qlt(F2)
Aout	-0.308	-0.110	2.035	0.487	0.224	0.029
Avril	-0.060	0.753	0.079	23.401	0.003	0.412
Decembre	1.282	-0.764	27.385	18.193	0.510	0.181
Fevrier	0.110	0.388	0.197	4.646	0.014	0.173
Janvier	0.597	0.199	4.316	0.901	0.260	0.029
Juillet	-0.708	-0.548	13.673	15.347	0.471	0.282
Juin	-0.320	0.150	2.961	1.224	0.099	0.022
Mai	0.052	0.525	0.060	11.702	0.002	0.254
Mars	-0.370	0.192	2.670	1.343	0.095	0.026
Novembre	1.140	-0.564	16.588	7.607	0.399	0.098
Octobre	0.974	0.253	14.967	1.899	0.331	0.022
Septembre	-0.848	-0.581	15.068	13.251	0.503	0.236

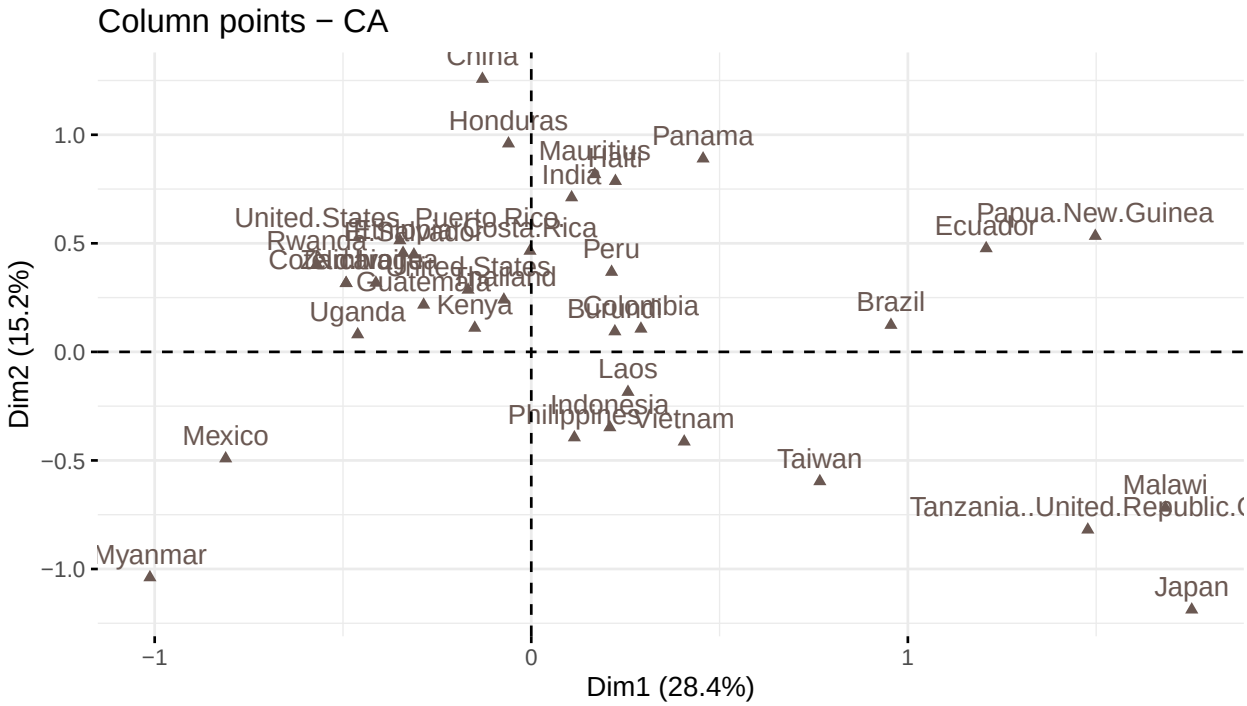


La répartition du nuage de point suit la forme d'un U à l'envers ressemblant à l'effet Guttman. Cette dispersion nous permets d'expliquer l'axe F1 qui représente les mois ayant pour critères les températures du plus chaud au plus froid.

3.2.1.1.4 Profils colonnes - Pays

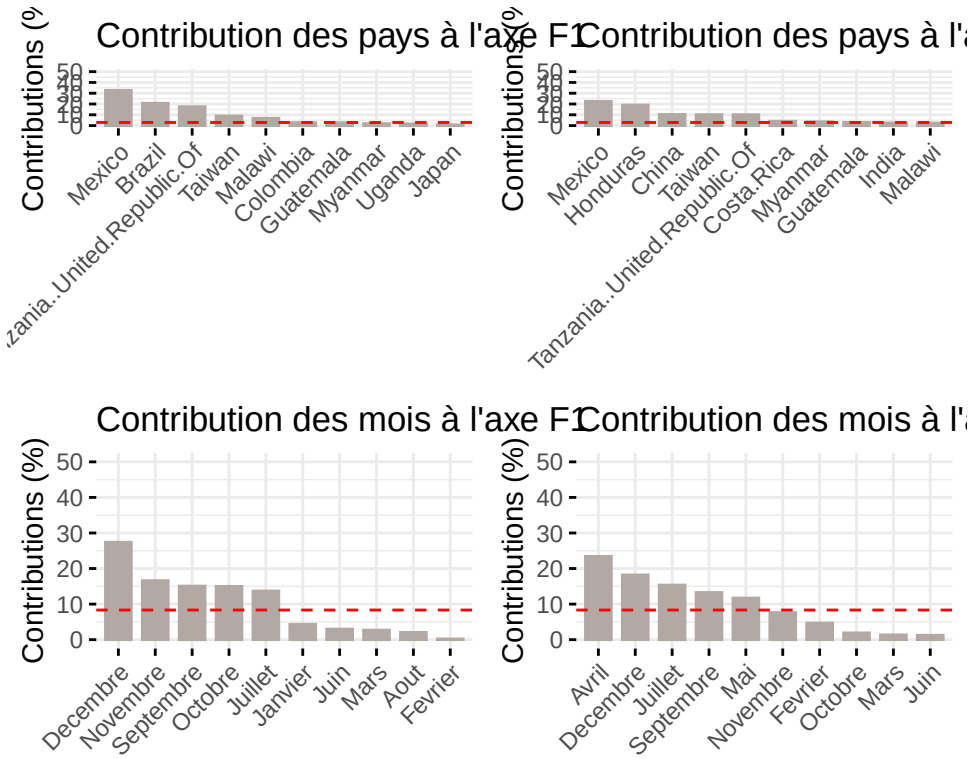
- Caractéristiques des profils colonnes - Pays

	F1	F2	ctr(F1)	ctr(F2)	qlt(F1)	qlt(F2)
Brazil	0.955	0.124	20.711	0.649	0.497	0.008
Burundi	0.222	0.094	0.021	0.007	0.007	0.001
China	-0.130	1.257	0.058	10.237	0.003	0.253
Colombia	0.291	0.106	2.733	0.680	0.345	0.046
Costa.Rica	-0.003	0.465	0.000	4.027	0.000	0.490
Cote.d.Ivoire	-0.492	0.316	0.052	0.040	0.034	0.014
Ecuador	1.208	0.477	0.631	0.184	0.191	0.030
El.Salvador	-0.312	0.448	0.399	1.543	0.079	0.162
Ethiopia	-0.340	0.455	0.627	2.098	0.060	0.108
Guatemala	-0.286	0.217	2.702	2.912	0.387	0.223
Haiti	0.223	0.786	0.065	1.503	0.011	0.142
Honduras	-0.060	0.959	0.040	19.008	0.002	0.520
India	0.107	0.711	0.030	2.459	0.011	0.506
Indonesia	0.209	-0.348	0.169	0.883	0.027	0.076
Japan	1.754	-1.188	0.665	0.571	0.175	0.080
Kenya	-0.150	0.110	0.112	0.113	0.064	0.035
Laos	0.257	-0.185	0.043	0.042	0.014	0.007
Malawi	1.684	-0.718	6.745	2.298	0.286	0.052
Mauritius	0.169	0.818	0.006	0.271	0.002	0.050
Mexico	-0.811	-0.491	32.608	22.406	0.664	0.244
Myanmar	-1.013	-1.039	1.774	3.498	0.176	0.185
Nicaragua	-0.413	0.318	0.626	0.695	0.214	0.127
Panama	0.456	0.890	0.180	1.283	0.060	0.230
Papua.New.Guinea	1.498	0.534	0.485	0.115	0.160	0.020
Peru	0.213	0.368	0.020	0.110	0.008	0.023
Philippines	0.114	-0.395	0.014	0.315	0.006	0.067
Rwanda	-0.570	0.404	0.070	0.066	0.029	0.015
Taiwan	0.766	-0.596	8.888	10.087	0.337	0.204
Tanzania..United.Republic.Of	1.478	-0.819	17.474	10.060	0.547	0.168
Thailand	-0.073	0.240	0.023	0.468	0.004	0.040
Uganda	-0.461	0.080	1.562	0.088	0.094	0.003
United.States	-0.168	0.286	0.061	0.330	0.005	0.014
United.States..Puerto.Rico.	-0.349	0.513	0.105	0.427	0.029	0.062
Vietnam	0.406	-0.414	0.250	0.485	0.078	0.081
Zambia	-0.492	0.316	0.052	0.040	0.034	0.014

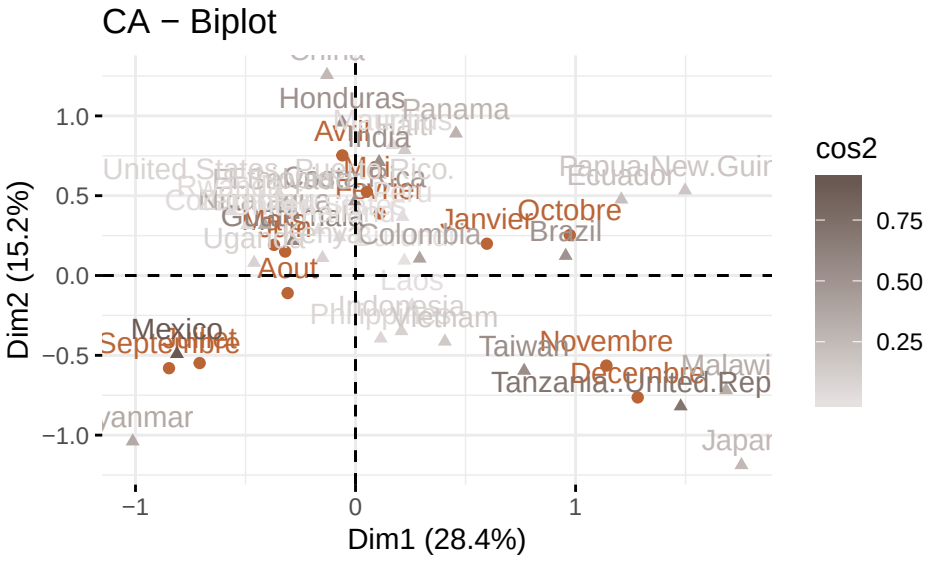


L'axe F2 oppose fortement le Méxique et le Myanmar à la Tanzania, la Malawi et le Japon.

3.2.1.1.5 Analyse des contributions



3.2.1.1.6 Représentation simultanée



La dispersion des pays est similaire à la dispersion des mois, on peut donc en conclure que la saisonalité est présente dans les récoltes des graines de café c'est-à-dire que les récoltes se font essentiellement lorsque le climat est plutôt doux (au printemps ou en automne).

3.2.1.2 AFC des varibales Variete et Mois

3.2.1.2.1 Test d'indépendance des variables variete et Mois

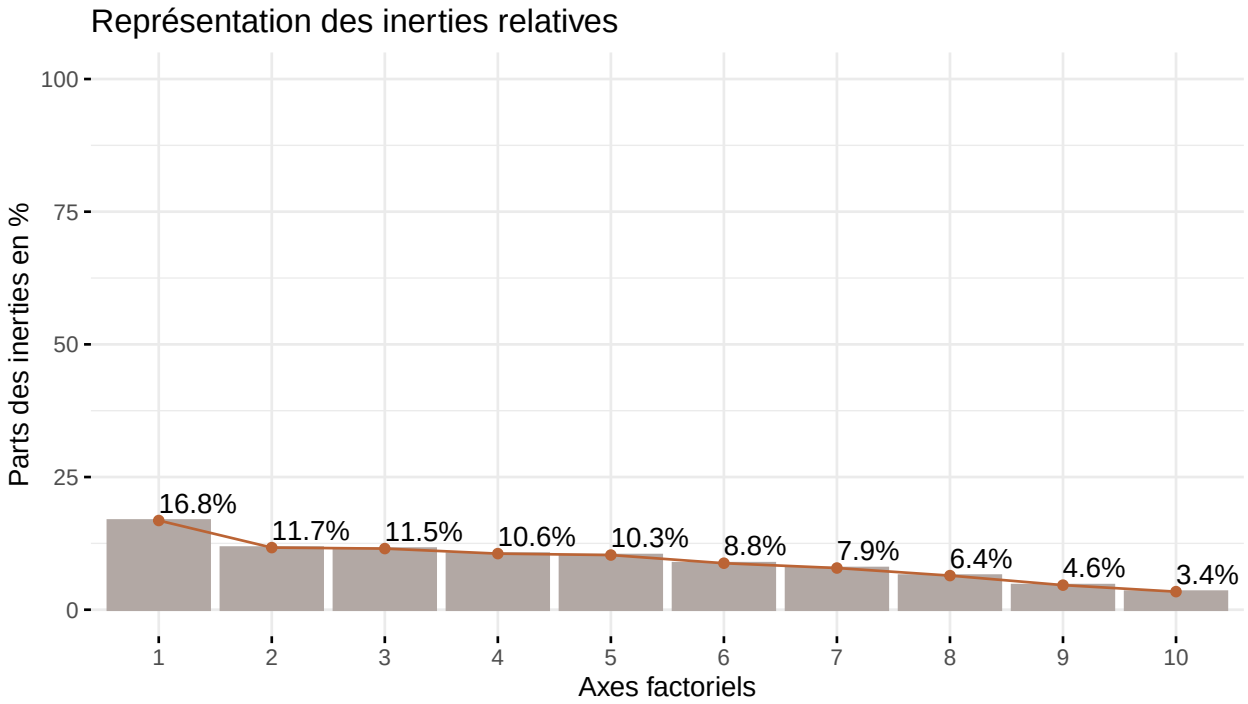
- Résultats du Khi2

X2obs	5949.246
df	810.000
pvalue	0.000

Comme la p-value est relativement faible, nous rejettons l'hypothèse H0 d'indépendance des variables Variete et Mois. Ces deux variables représentent les pays et les différentes variétés de graines de café.

Réalisons maintenant l'AFC.

3.2.1.2.2 Valeurs propres - Inerties des axes

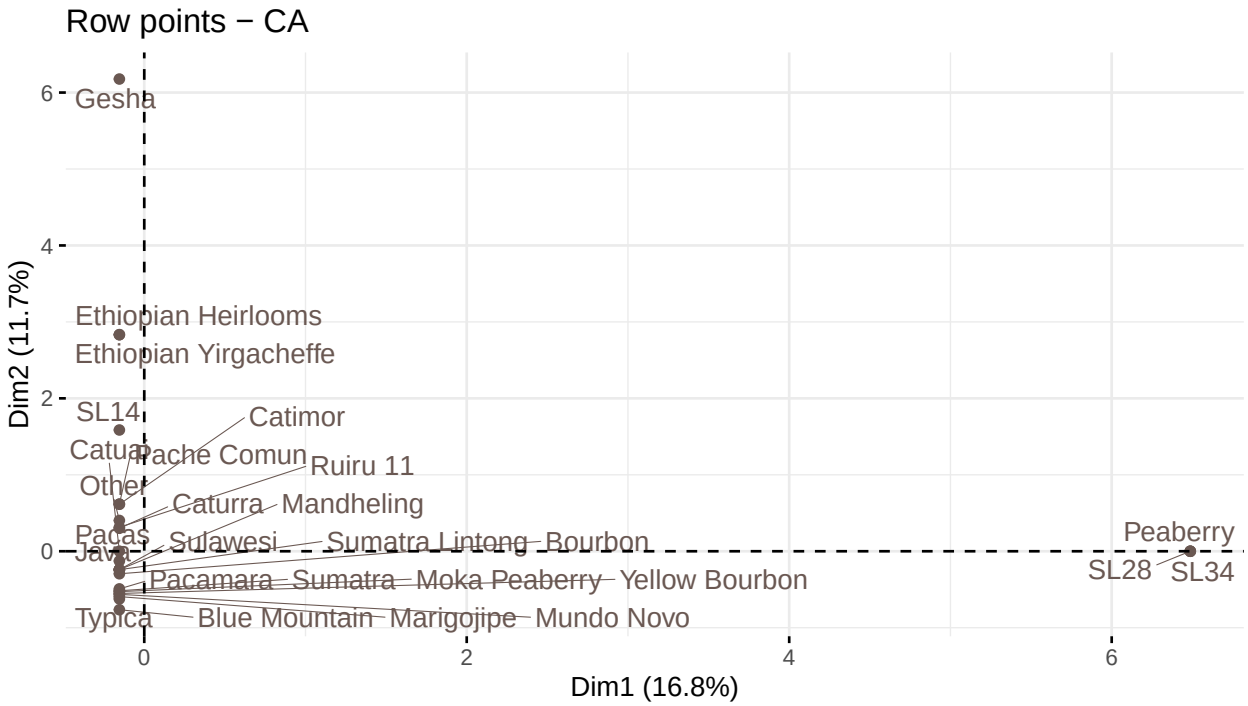


Les deux premères axes regroupent environs 44% des parts d'inertie totale ce qui reste légèrement faible mais étudier les autres axes n'apporteront pas forcément des informations supplémentaires pertinantes. Notre étude pours ces deux variables se basera donc seulement sur les deux premiers axes.

3.2.1.2.3 Profils lignes - Mois

- Caractéristiques des profils lignes - Variété

	F1	F2	ctr(F1)	ctr(F2)	qlt(F1)	qlt(F2)
Blue Mountain	-0.154	-0.765	0.005	0.170	0.000	0.012
Bourbon	-0.154	-0.295	0.494	2.590	0.017	0.061
Catimor	-0.154	0.615	0.048	1.095	0.001	0.013
Catuai	-0.154	0.001	0.165	0.000	0.008	0.000
Caturra	-0.154	0.300	0.568	3.101	0.014	0.052
Ethiopian Heirlooms	-0.154	2.833	0.002	1.163	0.000	0.098
Ethiopian Yirgacheffe	-0.154	2.833	0.005	2.326	0.000	0.098
Gesha	-0.154	6.177	0.029	66.348	0.000	0.711
Java	-0.154	-0.241	0.005	0.017	0.000	0.001
Mandheling	-0.154	-0.241	0.005	0.017	0.000	0.001
Marigojipe	-0.154	-0.593	0.002	0.051	0.007	0.105
Moka Peaberry	-0.154	-0.528	0.002	0.040	0.003	0.032
Mundo Novo	-0.154	-0.564	0.067	1.290	0.006	0.078
Other	-0.154	0.403	0.245	2.398	0.008	0.052
Pacamara	-0.154	-0.492	0.019	0.281	0.013	0.130
Pacas	-0.154	-0.127	0.031	0.030	0.006	0.004
Pache Comun	-0.154	0.617	0.002	0.055	0.000	0.003
Peaberry	6.487	0.000	4.247	0.000	0.956	0.000
Ruiru 11	-0.154	0.314	0.005	0.029	0.001	0.004
SL14	-0.154	1.586	0.038	5.834	0.001	0.062
SL28	6.487	0.000	59.457	0.000	0.981	0.000
SL34	6.487	0.000	33.975	0.000	0.956	0.000
Sulawesi	-0.154	-0.241	0.002	0.008	0.000	0.001
Sumatra	-0.154	-0.517	0.005	0.077	0.001	0.014
Sumatra Lintong	-0.154	-0.241	0.002	0.008	0.000	0.001
Typica	-0.154	-0.624	0.494	11.613	0.011	0.182
Yellow Bourbon	-0.154	-0.552	0.079	1.458	0.003	0.043



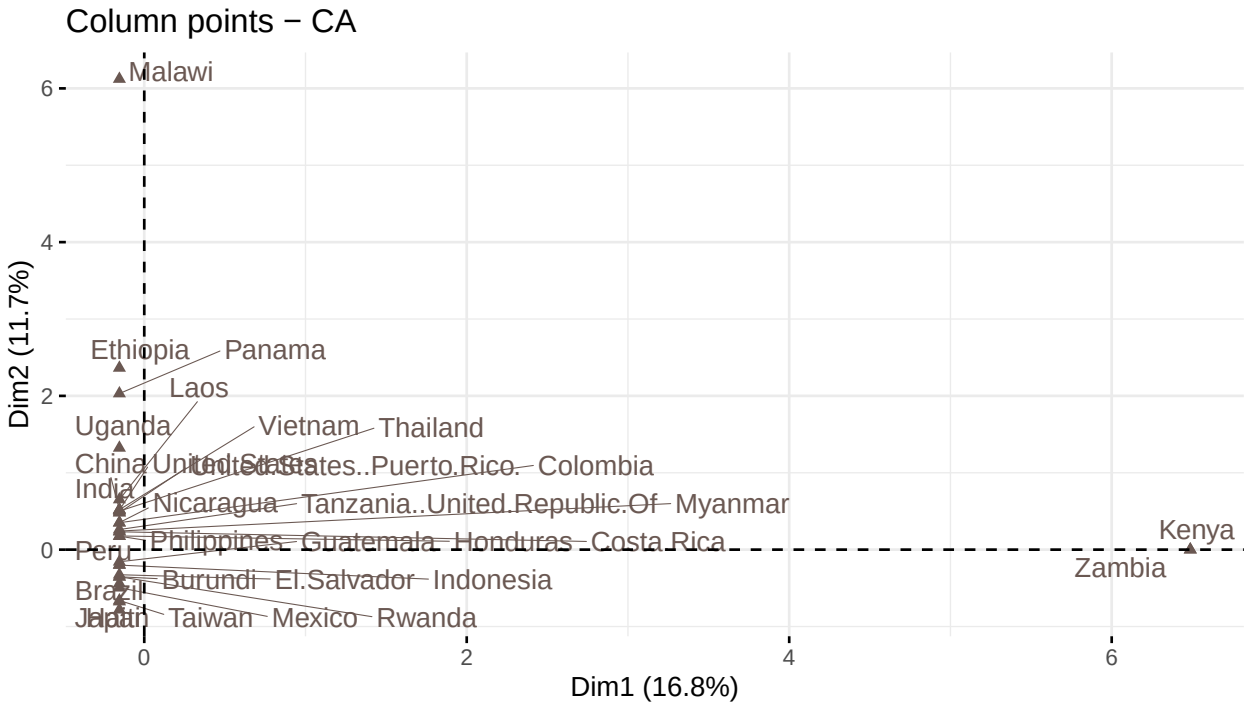
Quelques recherches sur les variétés ayant une forte contribution aux axes :

- F2 pour Geisha : corps crémeux et une acidité maîtrisée, mêlés à de surprenantes notes aromatiques de pêche, de vanille, de caramel et de cerises,
- F1 pour Peaberry : arôme de cette variété est floral et sucré. Ses notes portent sur les agrumes, la noix de coco et même l'ananas,
- F1 pour SL 28-34 (Scott Labs) : corps juteux et une douce saveur, presque tropicale, Notes de Cerise, Cassis, Citron, Orange sanguine, différence entre les 28 et 34(plus faible) est l'altitude de culture .

3.2.1.2.4 Profils colonnes - Pays

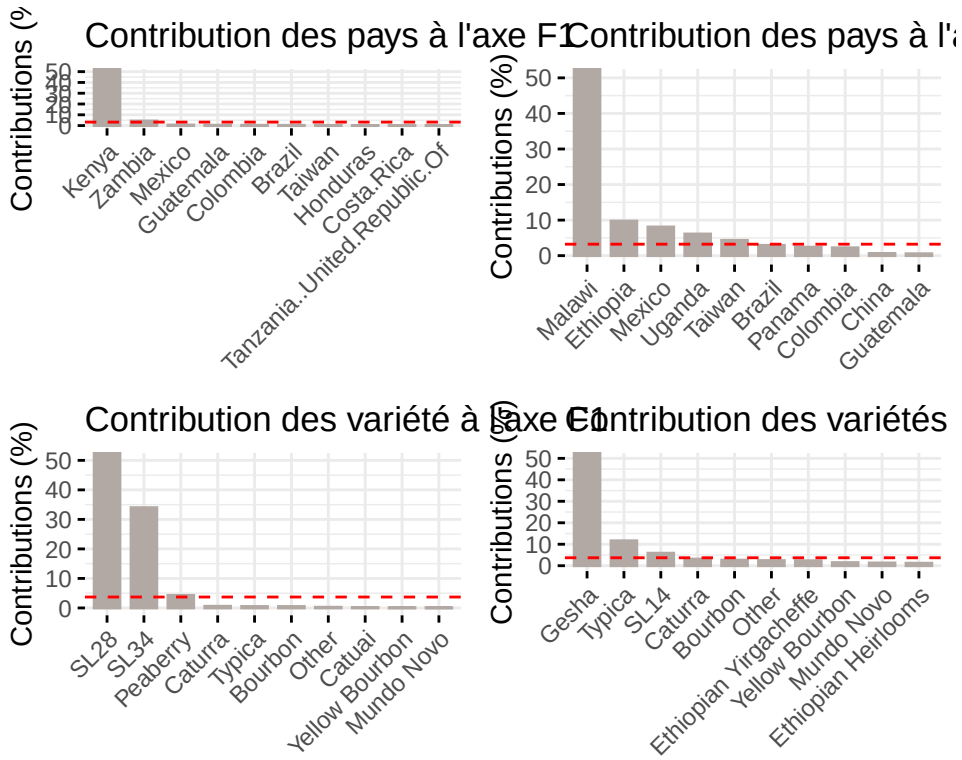
- Caractéristiques des profils colonnes - Pays

	F1	F2	ctr(F1)	ctr(F2)	qlt(F1)	qlt(F2)
Brazil	-0.154	-0.441	0.247	2.900	0.007	0.054
Burundi	-0.154	-0.353	0.002	0.018	0.006	0.033
China	-0.154	0.519	0.038	0.625	0.001	0.010
Colombia	-0.154	0.350	0.300	2.221	0.009	0.048
Costa.Rica	-0.154	0.232	0.105	0.345	0.011	0.025
El.Salvador	-0.154	-0.326	0.036	0.231	0.006	0.027
Ethiopia	-0.154	2.364	0.029	9.718	0.001	0.221
Guatemala	-0.154	-0.158	0.362	0.544	0.016	0.017
Haiti	-0.154	-0.781	0.012	0.442	0.001	0.028
Honduras	-0.154	0.178	0.120	0.230	0.008	0.011
India	-0.154	0.483	0.005	0.068	0.003	0.027
Indonesia	-0.154	-0.201	0.036	0.088	0.001	0.001
Japan	-0.154	-0.676	0.002	0.066	0.001	0.013
Kenya	6.487	0.000	93.432	0.000	0.999	0.000
Laos	-0.154	0.652	0.007	0.185	0.001	0.019
Malawi	-0.154	6.122	0.026	59.737	0.000	0.689
Mexico	-0.154	-0.495	0.544	8.066	0.022	0.226
Myanmar	-0.154	0.242	0.019	0.068	0.005	0.012
Nicaragua	-0.154	0.352	0.034	0.251	0.013	0.067
Panama	-0.154	2.031	0.010	2.391	0.004	0.676
Peru	-0.154	-0.194	0.005	0.011	0.019	0.030
Philippines	-0.154	0.175	0.010	0.018	0.005	0.006
Rwanda	-0.154	-0.353	0.002	0.018	0.006	0.033
Taiwan	-0.154	-0.661	0.163	4.311	0.009	0.163
Tanzania..United.Republic.Of	-0.154	0.262	0.084	0.347	0.004	0.012
Thailand	-0.154	0.501	0.019	0.291	0.003	0.033
Uganda	-0.154	1.324	0.058	6.093	0.001	0.065
United.States	-0.154	0.515	0.019	0.307	0.001	0.012
United.States..Puerto.Rico.	-0.154	0.483	0.010	0.135	0.003	0.027
Vietnam	-0.154	0.520	0.017	0.275	0.005	0.052
Zambia	6.487	0.000	4.247	0.000	0.603	0.000

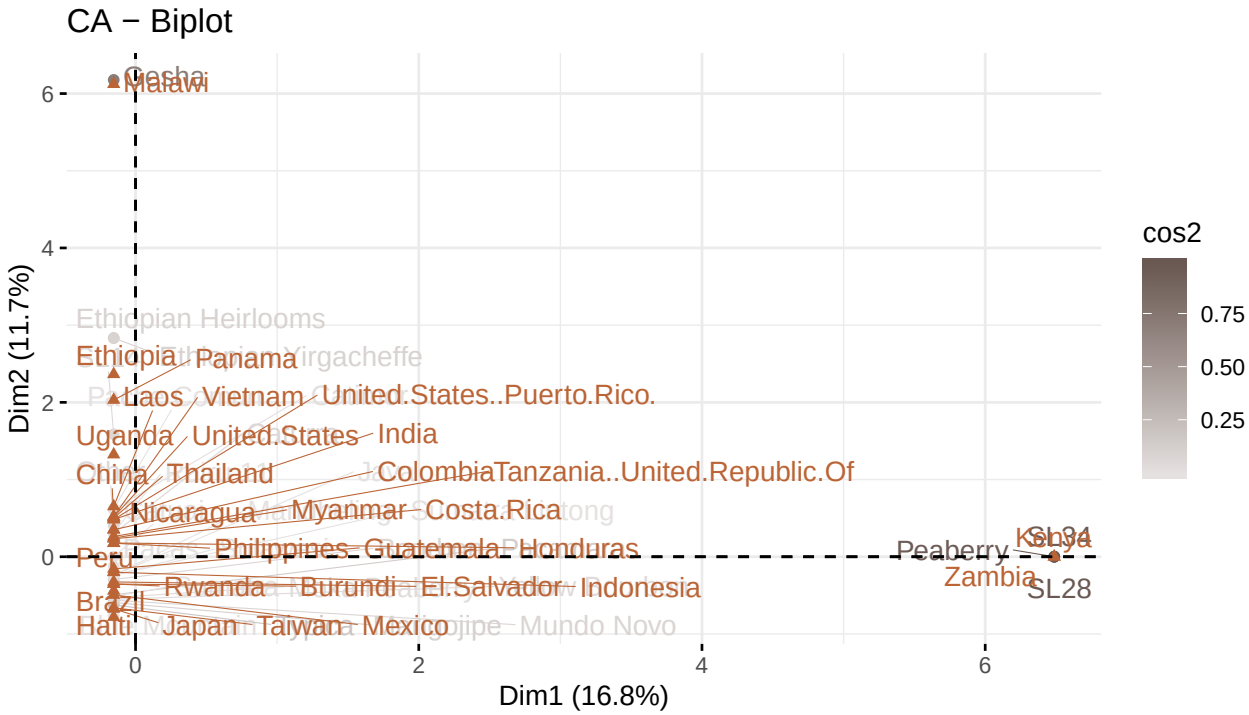


L'axe F2 oppose fortement le Méxique et Myanmar au Tanzania, Malawi et le Japon.

3.2.1.2.5 Analyse des contributions



3.2.1.2.6 Représentation simultanée



La Malawi, le Kenya et le Zambia possèdent une variété de graine de café spéciale qui sont respectivement le Geisha,le SL 34-28 et le Peaberry. Pour les autres variétés, il faudrait faire une analyse sur les autres axes car leurs qualités de représentation sont très faibles sur les axes F1 et F2.