M1 Economiste d'entreprise
Projet Analyse de données exploratoire

La qualité du café dépend-elle de sa provenance ?

Aybuké BICAT, Basma GHAFFOUR

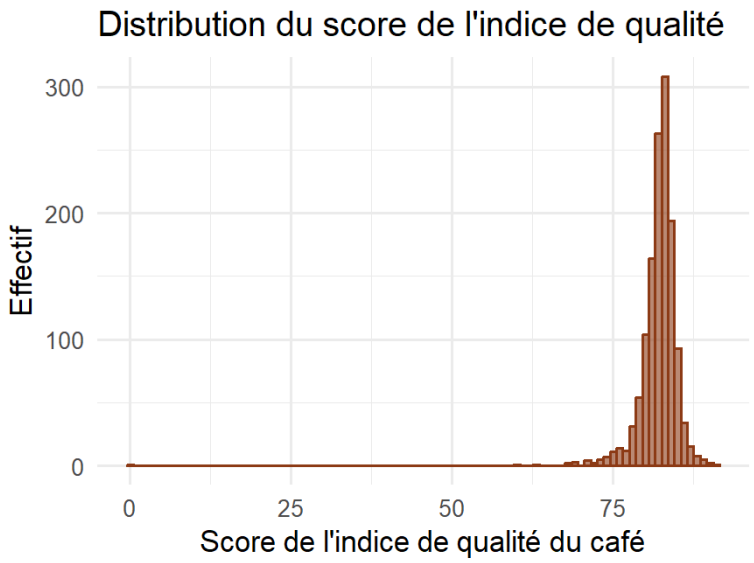
Année universitaire 2022/2023

1 Introduction

On a choisi de travailler sur les bases de données[1] portant sur le café. La problématique est la suivante : la qualité du café dépend-elle de sa provenance ? On a travaillé sur une base de données qui contient 1339 observations et 44 variables. Pour y répondre on a fait le choix de faire une analyse factorielle des correspondances dans un premier temps.

1.1 Analyse descriptive

Les variables sur lesquelles l'étude a été faite dans un premier temps sont les variables portant sur la provenance du café et sur l'indice de qualité du café. La variable de l'indice de qualité du café est une variable quantitative. Cet indice nous indique pour chacune des observations le score qui a été attribué. Plus la note est élévée et plus c'est un café de bonne qualité.



Sur ce graphique on peut voir la distribution de l'indice de qualité du café, elle est concentrée essentiellement autour de 80. La variable à été recodé en variable qualitative qui prendra comme modalité plusieurs intervalles.

La seconde variable nous donne le pays de provenance du café pour chaque observation. Les pays que les observations peuvent prendre sont les suivants: Brazil, Burundi, China, Colombia, Costa Rica, Cote d?Ivoire, Ecuador, El Salvador, Ethiopia, Guatemala, Haiti, Honduras, India, Indonesia, Japan, Kenya, Laos, Malawi, Mauritius, Mexico, Myanmar, Nicaragua, Panama, Papua New Guinea, Peru, Philippines, Rwanda, Taiwan, Tanzania, United Republic Of, Thailand, Uganda, United States, United States (Hawaii), United States (Puerto Rico), Vietnam, Zambia. On les a recodé pour que cette variable prenne 7 modalités qui sont 7 régions du monde.

Table 1: Proportion d'observations par région

	AmeriqueSud	Afrique	AsieEst	AmeriqueCentrale	AsieSud	AmeriqueNord	Oceanie
proportion	24.5%	12.1%	6.9%	25.9%	6.7%	18.4%	5.5%

[1]Sources des bases de données

Sur la table 1 on peut voir la proportion des observations que prennent chacune des modalités de la variable portant sur la région de provenance. On observe notamment que l'Océanie est la modalité la plus rare avec seulement 5.5% des observations qui prennent cette modalité.

Table 2: Repartition des pays par régions (1/2)

AmeriqueSud	AmeriqueCentrale	AmeriqueNord	Afrique
Colombia: 56 %	Guatemala: 52 %	: NA %	Ethiopia: 27 %
Brazil: 40 %	Costa Rica: 15 %	Mexico: 96 %	Tanzania, United Republic Of: 25 %
Peru: 3 %	Honduras: 15 %	United States: 4 %	Uganda: 22 %

Table 3: Repartition des pays par régions (2/2)

AsieEst	AsieSud	Oceanie
Taiwan: 82 %	Thailand: 36 %	: NA %
China: 17 %	Indonesia: 22 %	United States (Hawaii): 99 %
Japan: 1 %	India: 16 %	Papua New Guinea: 1 %

Pour chaque nouvelle modalité de provenance, on peut voir sur les table 2 et 3 que les pays à l'intérieur de chaque région sont inégalement représentés.[2]

2 AFC

Table 4: Tableau croisé sur lequelle porte l'AFC

	AmeriqueSud	Afrique	AsieEst	AmeriqueCentrale	AsieSud	AmeriqueNord	Oceanie
0 - 80.25	20	3	15	68	16	84	18
80.25 - 81.67	36	30	23	70	19	44	11
81.67 - 82.5	65	19	12	55	14	50	6
82.5 - 83.17	69	23	14	50	20	28	13
83.17 - 84.17	89	25	15	56	14	22	12
84.17 - 90.58	49	62	13	47	7	18	14

Sur la table 4 on peut voir le tableau croisé sur lequel a été fait notre AFC.

2.1 Profils colonnes

Table 5: profil colonne

	AmeriqueSud	Afrique	AsieEst	AmeriqueCentrale	AsieSud	AmeriqueNord	Oceanie
0 - 80.25	6.10	1.85	16.30	19.65	17.78	34.15	24.32
80.25 - 81.67	10.98	18.52	25.00	20.23	21.11	17.89	14.86
81.67 - 82.5	19.82	11.73	13.04	15.90	15.56	20.33	8.11

[2]Sur les tableaux 2 et 3 seul les 3 pays les plus représentés dans la modalité de la variable portant sur la région de provenance sont visibles

	AmeriqueSud	Afrique	AsieEst	AmeriqueCentrale	AsieSud	AmeriqueNord	Oceanie
82.5 - 83.17	21.04	14.20	15.22	14.45	22.22	11.38	17.57
83.17 - 84.17	27.13	15.43	16.30	16.18	15.56	8.94	16.22
84.17 - 90.58	14.94	38.27	14.13	13.58	7.78	7.32	18.92
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Sur la table 5, on peut voir que l'on a environ **38%** des données qui prennent la modalité **Afrique** sont comprises dans l'intervalle **84.17 - 90.58**. On peut aussi voir que parmis les données qui prennent la modalité **AmeriqueNord**, **34%** environ sont comprise dans l'intervale **84.17 - 90.58**.

2.2 Profils lignes

Table 6: profil ligne

	AmeriqueSud	Afrique	AsieEst	AmeriqueCentrale	AsieSud	AmeriqueNord	Oceanie	Total
0 - 80.25	8.93	1.34	6.70	30.36	7.14	37.50	8.04	100
80.25 - 81.67	15.45	12.88	9.87	30.04	8.15	18.88	4.72	100
81.67 - 82.5	29.41	8.60	5.43	24.89	6.33	22.62	2.71	100
82.5 - 83.17	31.80	10.60	6.45	23.04	9.22	12.90	5.99	100
83.17 - 84.17	38.20	10.73	6.44	24.03	6.01	9.44	5.15	100
84.17 - 90.58	23.33	29.52	6.19	22.38	3.33	8.57	6.67	100

Dans la table 6 on peut voir qu'environ **38%** des observations qui prennent la modalité **0 - 80.25** prennent la modalité **AmeriqueNord**, mais on a aussi **30%** qui prennent la modalité **AmeriqueCentrale**. Parmis les observations qui prennent la modalité **84.17 - 90.58**, environ **30%** proviennt d'**Afrique**. Ce qui nous donne déjà une première indication sur la provenance du meilleur café.

2.3 Inerties

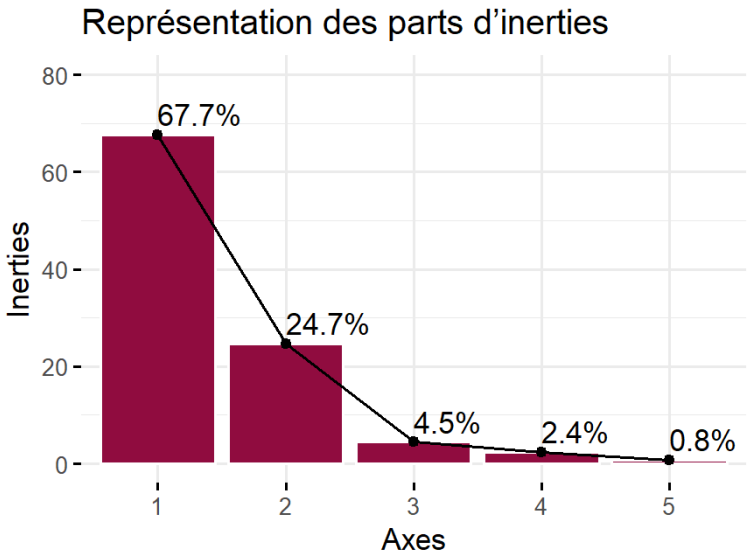
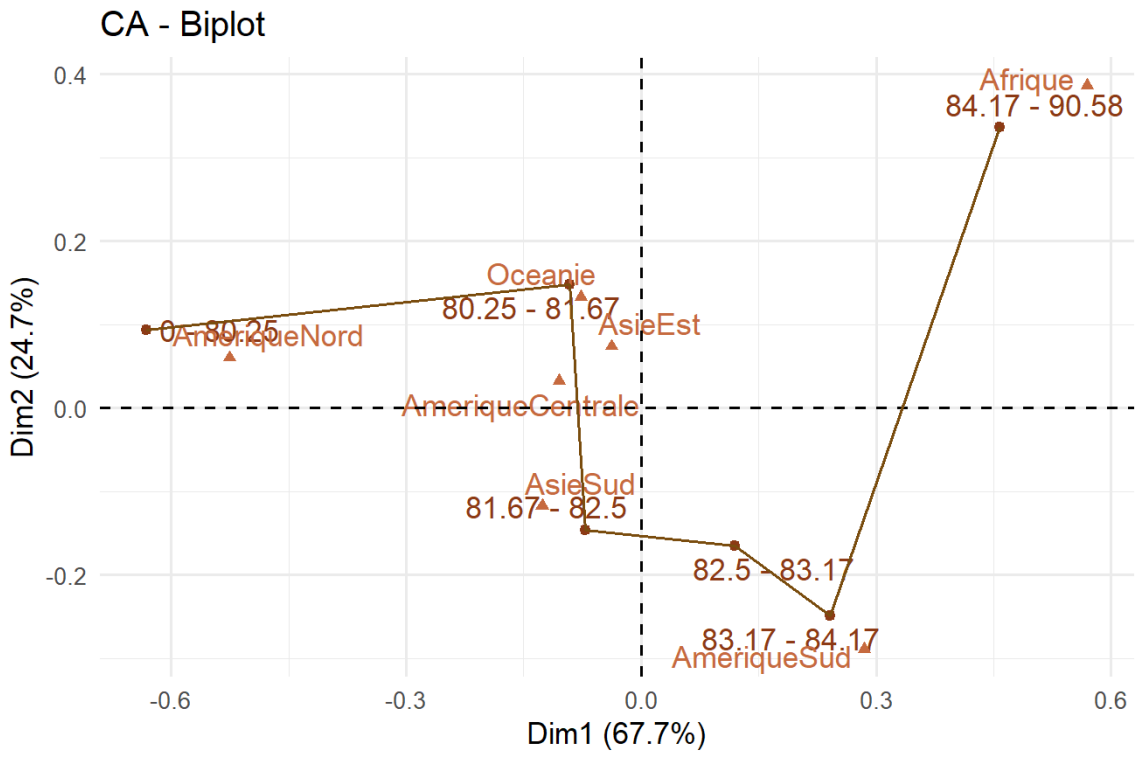


Table 7: Part d'Inertie

	F1	F2	F3	F4	F5
Inerties	0.114	0.042	0.008	0.004	0.001
Inerties relatives (%)	67.650	24.699	4.497	2.376	0.778
Inertie relatives cumulées (%)	67.650	92.350	96.847	99.222	100.000

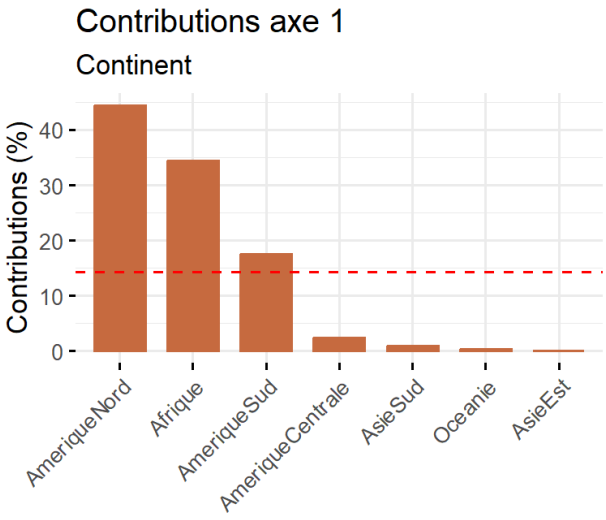
Sur ce graphique, on peut voir les inerties relatives pour chacun des axes. On a fait le choix de garder les deux premiers axes qui cumulent environ **92%** de l'inertie totale.

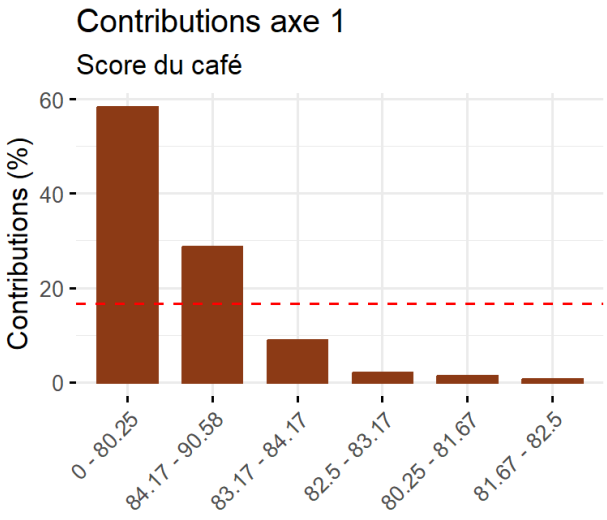
2.4 Graphique



D'après ce graphique, on voit globalement sur l'axe des abscisses une relation croissante pour les intervalles des indices de qualité. Plus on va du coté positif de l'axe 1 plus on aura un intervalle des indices de qualité avec de bornes élevées. Les intervalles **80.25 - 81.67** et **81.67 - 82.5** permettent de différencier l'axe des ordonnées.

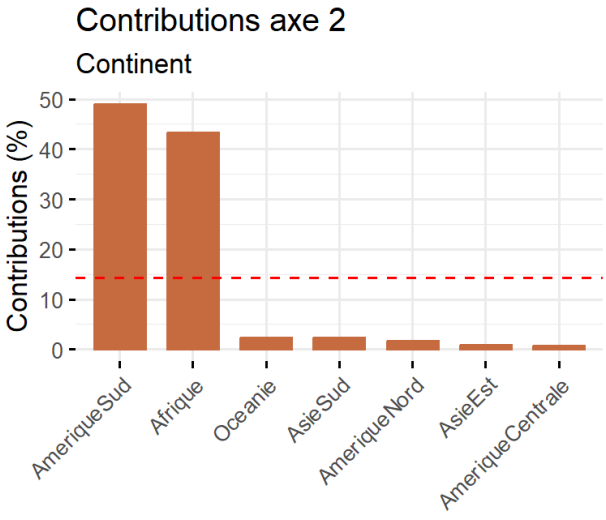
2.5 Etude du 1er axe

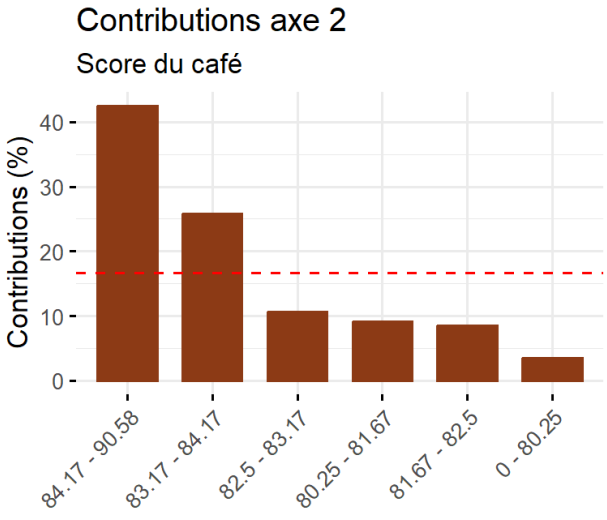




On peut voir que pour l'axe 1 on a l'**Amérique du Nord** et l'**Afrique** ainsi que les intervalles **0 - 80.25** et **84.17 - 90.58** qui contribuent le plus. Sur le graphique précédent, on note qu'à l'extrémité gauche on a l'**Amérique du Nord** et l'intervalle **0 - 80.25** qui sont associés. Et à l'autre extrémité de l'axe, la variable **Afrique** et l'intervalle **84.17 - 90.58** qui sont liés.

2.6 Etude du 2nd axe





Pour l'axe 2, les modalités **Amérique du Sud** et **Afrique** ainsi que les deux intervalles ayant les bornes les plus élevées du score du café contribuent le plus. On peut voir sur le graphique précédent que sur l'extrémité de l'axe 2, les modalités **Afrique** et **84.17 - 90.58** sont associées. Sur l'autre extrémité la variable **Amerique du Sud** et **83.17 - 84.17** se rassemblent.

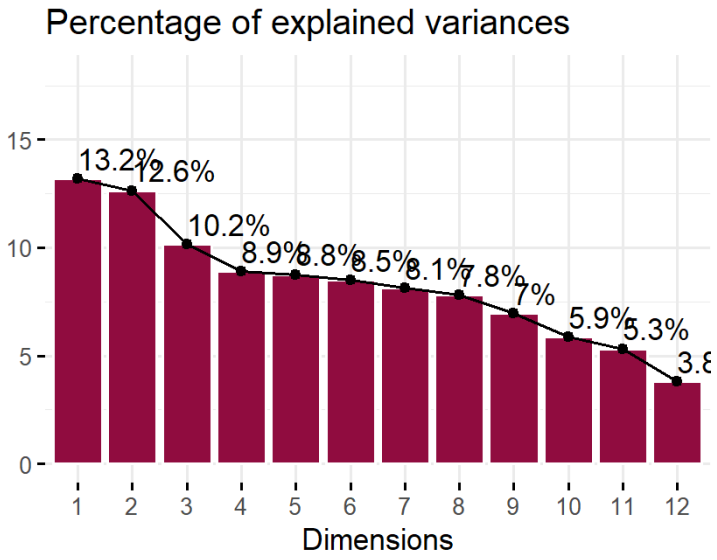
On peut donc en conclure que les meilleurs cafés, ceux qui ont obtenu **le meilleur score proviennent d'Afrique** essentiellement, suivi de l'Amérique du Sud. **Ceux qui ont le score le plus faible proviennt**

d'Amérique du Nord. On trouve qu'en fonction de la région de provenance du café, on a des scores significativement différents, mais pourquoi? Est-ce qu'il y a aurait des caractéristiques propres à la région de provenance qui pourraient expliquer que le café serait meilleur? Des caractéristiques propres aux régions et indépendantes des choix que le producteur pourrait faire. On essaira d'y répondre avec une analyse par composante multiple.

3 ACM

Pour apporter plus d'explications à nos précédents résultats, on a choisi d'étudier, en plus des variables précédentes, **l'altitude moyenne en mètres** ainsi que le **taux d'humidité**. En effet, l'altitude est propre à chaque région et le producteur n'a aucun contrôle dessus. On a decidé de recoder l'altitude en variable qualitative qui prend 3 modalités : 2 intervalles et une modalité NA pour les données absentes ou aberrantes. Pour l'humidité, on l'a aussi recodé en variable qualitative, mais on a fait le choix de la laisser en **variable qualitative supplémentaire** car on ne savait pas si c'était l'humidité du climat ou celle qui a été choisi de laisser dans le produit final par le producteur. Pour la variable de provenance, elle est identique à celle de la partie précédente, mais pour celle de l'indice de qualité on n'a plus que 5 intervalles et non 6.

3.1 Inerties

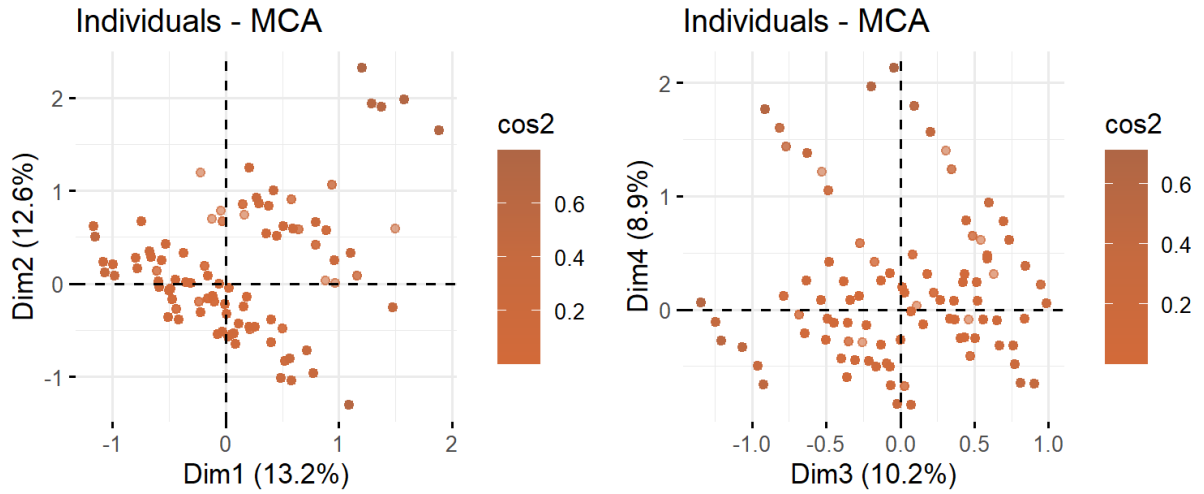


D'après le graphique des inerties, on peut voir que :

- Les 2 premiers axes conservent **26%** de l'inertie totale.
- L'axe F3 conserve **10.2%** de l'inertie.
- Les axes F4, F5 et F6 conservent une part similaire de l'inertie, avec environ **8%** de l'inertie pour chacun.

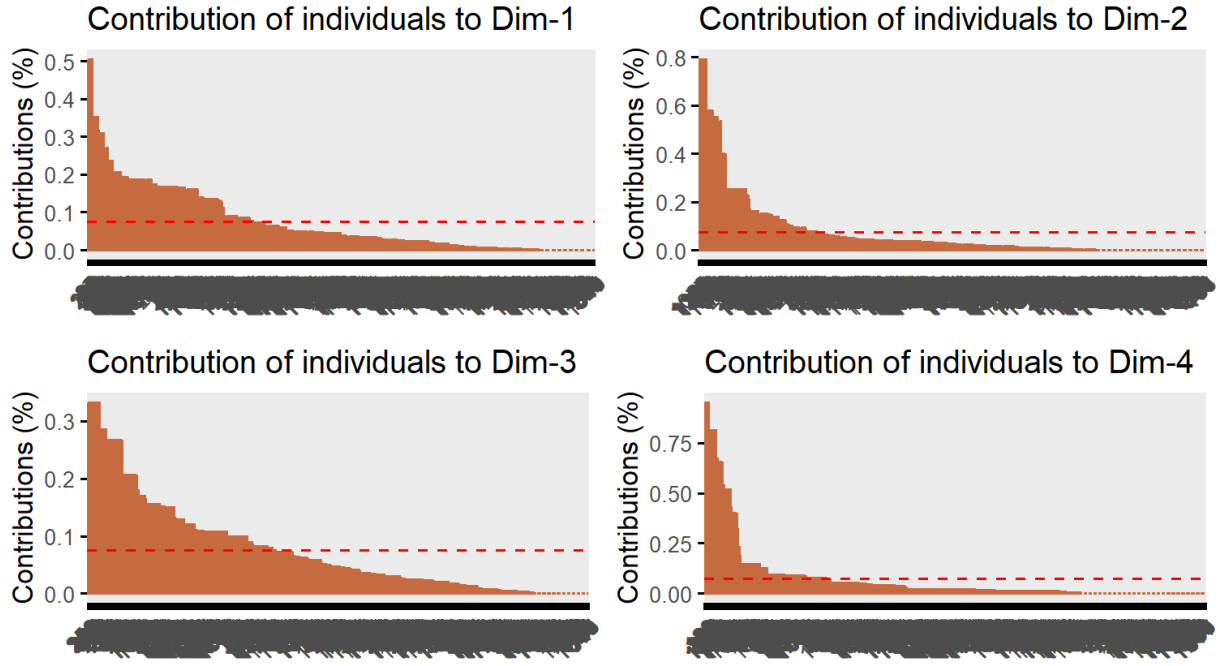
On décide donc de retenir les **4 premiers axes** car ils expliquent environ **45%** de l'inertie totale pour la suite de notre étude.

3.2 Etude des individus



Pour la représentation des individus, on observe la formation de 3 groupes sur **(F1,F2)**. On a un groupe assez détaché des autres, qui regroupe donc les individus extrêmes. Pour les 2 autres groupes on observe une relation linéaire avec un groupe qui comporte beaucoup plus d'individus que l'autre. Sur **(F3,F4)**, on observe + de groupes qui se situent aux extrêmités, ainsi qu'un grand groupe d'individus assez concentrés au milieu.

3.2.1 Contribution des individus



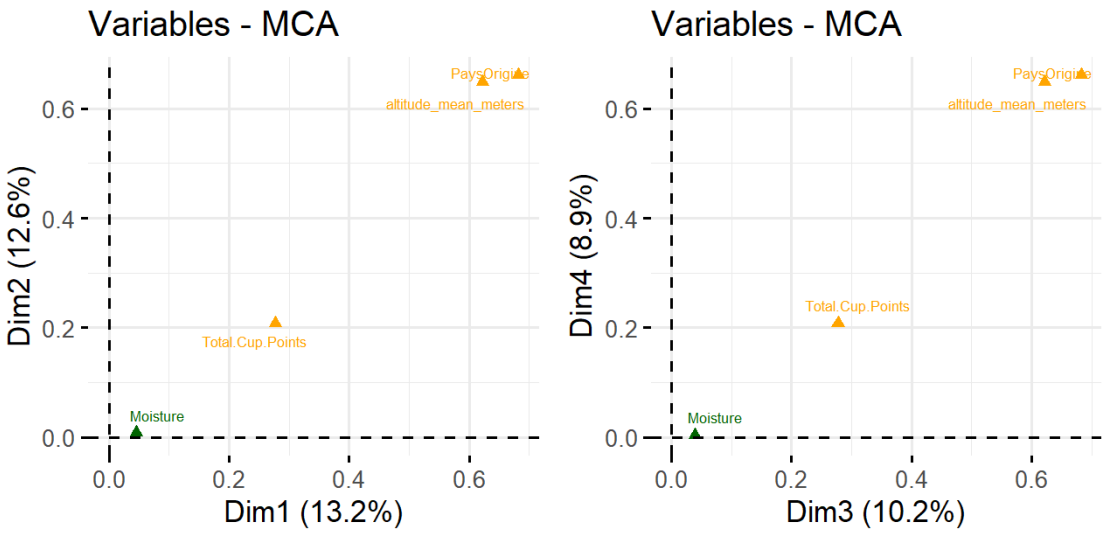
D'après les graphiques de contribution des individus, on s'aperçoit qu'en dimension 1, il y a peu d'individus qui contribuent beaucoup (+ de 0.4%) et on peut voir qu'il y a une relation décroissante assez rapide de ce taux.

En dimension 2, il y a un peu plus d'individus qui contribuent à + de 0.5%, puis il y a une concentration aux alentours des 0.2%.

Pour la dimension 3, les contributions sont comprises entre 0.1% et 0.3% et beaucoup d'individus contribuent à cette dimension de façon moindre que dans les autres dimensions.

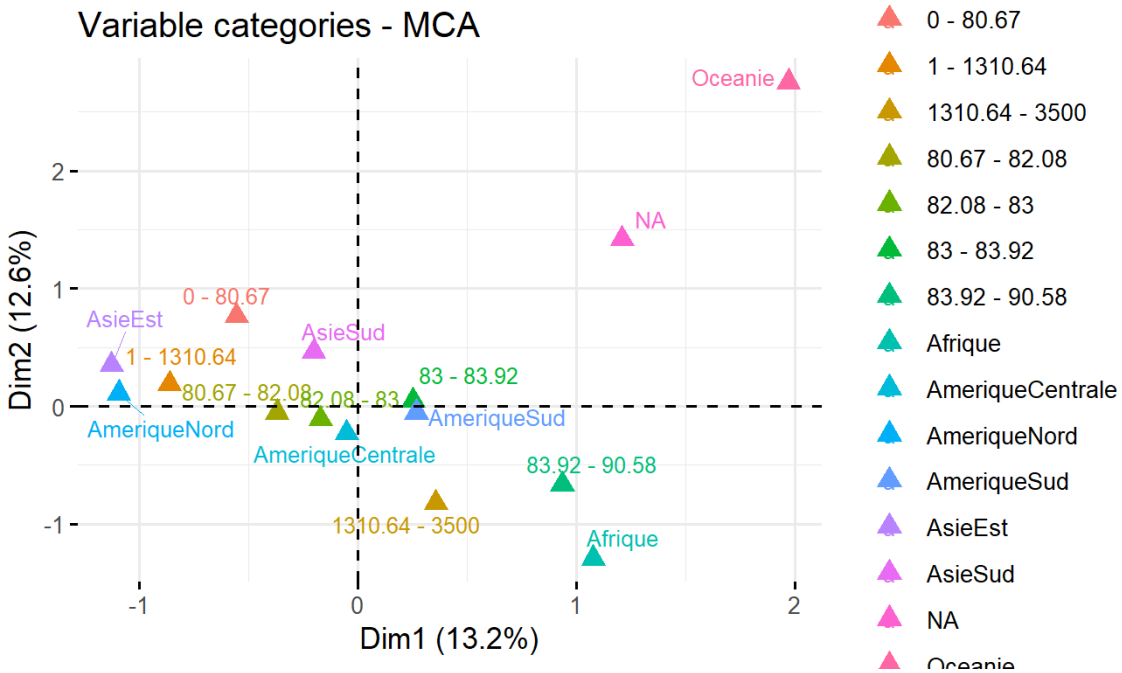
Très peu d'individus contribuent à la dimension 4 et on peut observer une chute des contribution qui passe de 0.75% à environ de 0.2%.

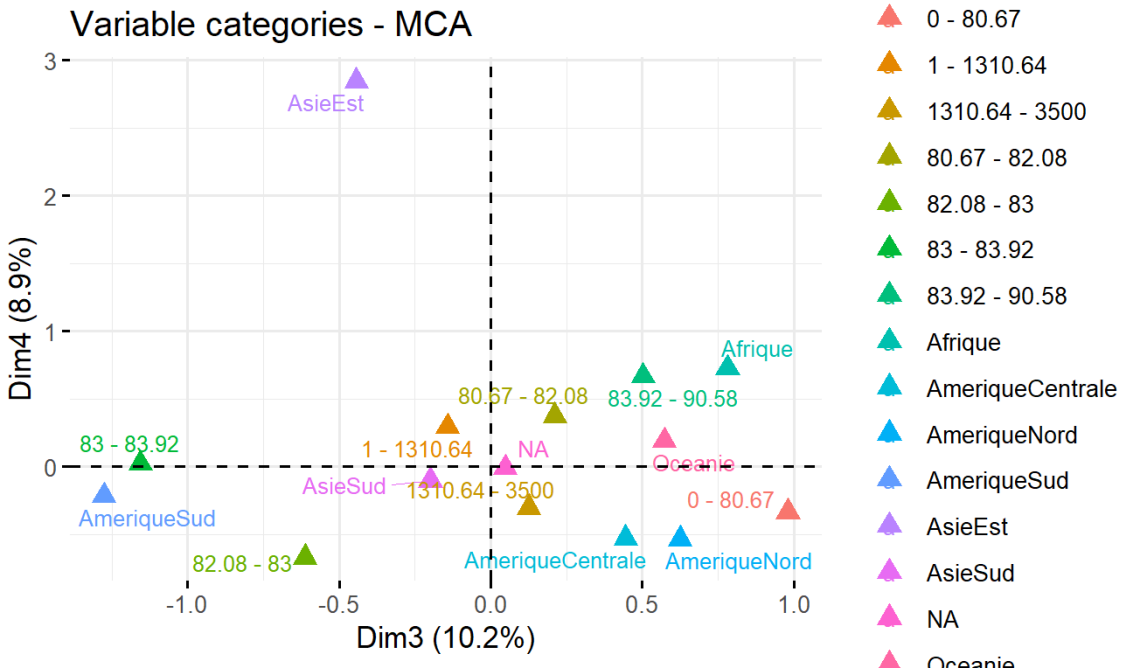
3.3 Etudes des variables et des modalités



On observe ici que les variables sont expliquées de la même manière dans les différentes dimensions. En effet, **Total.Cup.Points** est mieux expliqué par les axes 1 et 3 à la même hauteur. D'autre part, **PaysOrigine** et **altitude_mean_metters** semblent être bien expliqués par les 2 axes dans chacun des graphiques, même s'il y a une petite différence et que les axes 2 et 4 les expliquent légèrement plus. Enfin, la variable qualitative supplémentaire **Moisture** semble être expliquée très faiblement par l'axe 1.

3.3.1 Modalités





Graphiquement sur **(F1,F2)**, on peut voir la formation d'une bissectrice par rapport aux axes 1 et 2, qui fait apparaitre 2 groupes. En effet, on peut voir un groupe dans le côté négatif de l'axe 1, où les modalités des scores les plus bas comme **0 - 80.67** et `nam_bis[2]` s'associent bien avec les pays tels **AsieEst**, **AmeriqueNord** et **AsieSud** qui sont tous liés à la modalité **1-1310.64** mètres d'altitude. On peut déjà dire que les cafés les moins bons viennent plutôt d'endroits d'altitude basse.

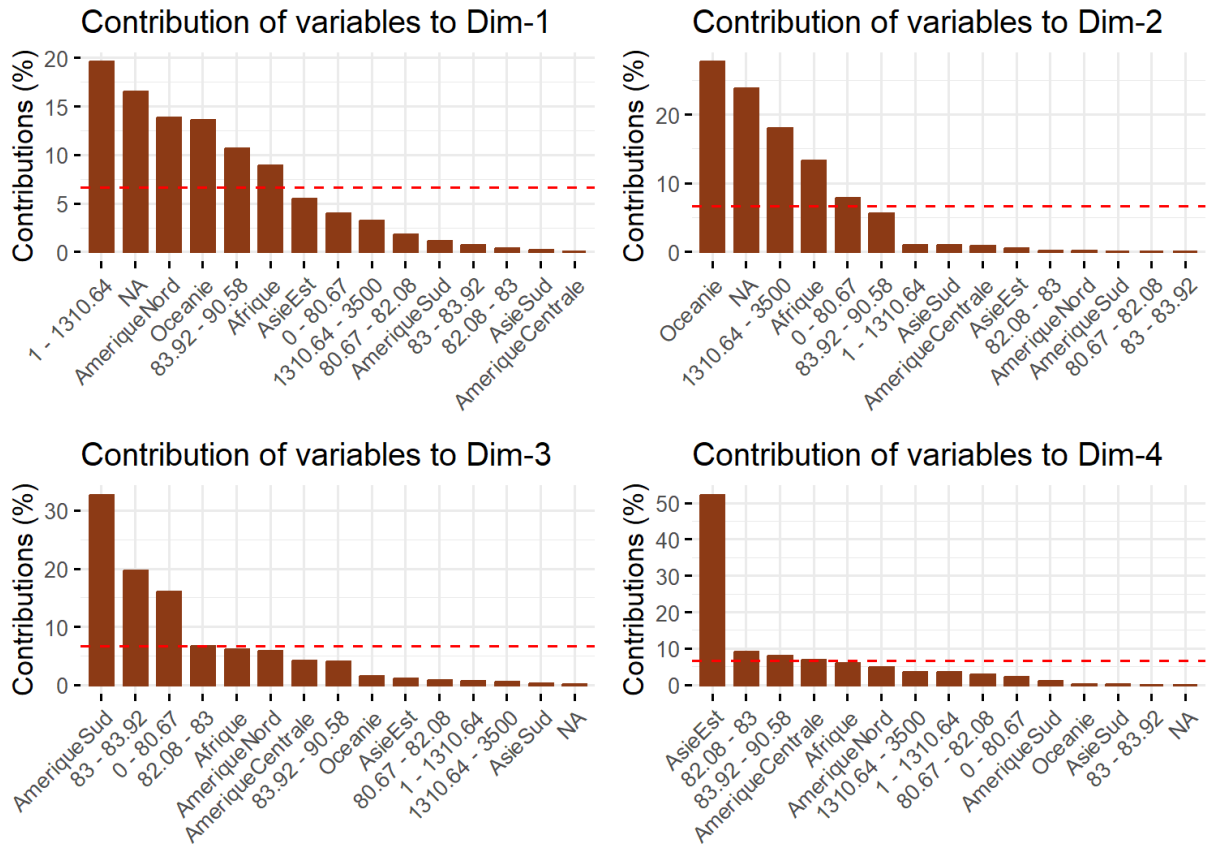
Le 2ème groupe qu'on voit sur cette bissectrice est celui qui regroupe l'**Afrique** et l'**AmériqueSud** qui ont un score de **83-83.92** et qui sont à des altitudes entre **1310.64-3500** mètres.

Par ailleurs, il y a aussi une région, **l'AmériqueCentrale** avec un score de **82.08-83** qui se trouve au centre de ces 2 groupes, on verra par la suite auquel de ces groupes il appartient.

On a aussi un 3ème groupe, **l'Océanie** qui semble être lié à la modalité **NA** d'altitude, on peut penser que ce groupe prend des modalités extrêmes, donc les modalités rares et que les données d'altitude données pour ce groupe étaient fausses, d'où le fait qu'il se retrouve très loin des autres.

Pour **(F3,F4)**, les régions telles que **AsieSud**, **AmeriqueNord** et **AmeriqueCentrale** ainsi que le score de **82.08-83**, semblent expliqués l'axe 4. Puis on a des modalités qui se retrouvent autour de l'axe 3, ils sont moins bien représentés et on n'arrive plus à apercevoir de groupes.

3.3.2 Contributions des variables

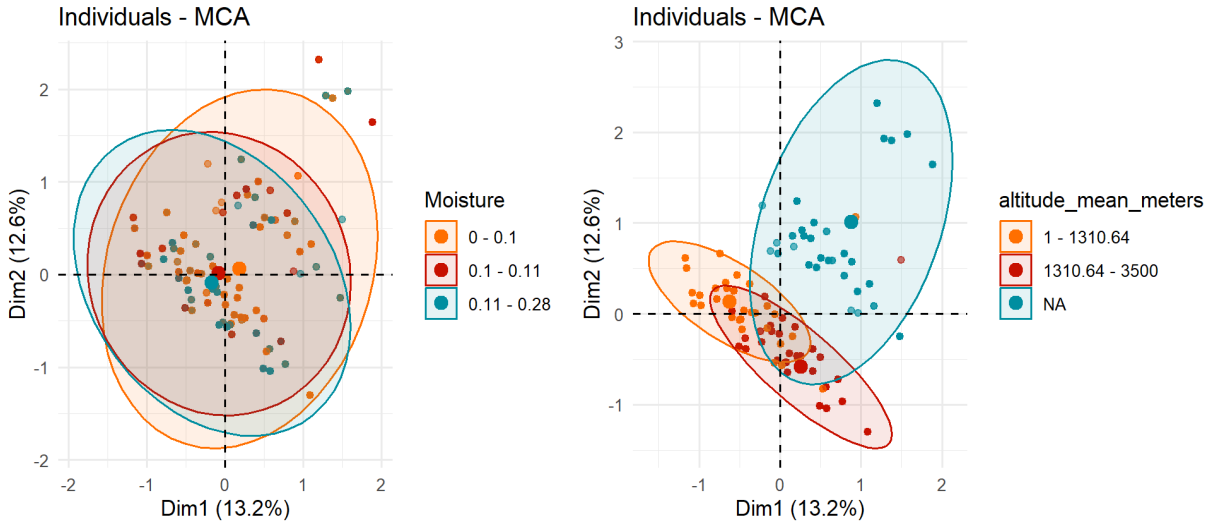


Pour les contributions des modalités, on voit que :

- Les modalités **NA**, **1-1310.64**, **Océanie**, **AmériqueNord**, **Afrique** et le score de **83.92 - 90.58** expliquent mieux la dimension 1.
- La dimension 2 est expliquée par les modalités comme l'altitude de **1310.64-3500** et **NA**, les régions tels **l'Océanie** et **l'Afrique**, et un score de **0-80.67**.
- Pour la dimension 3, ce sont **l'AmériqueSud**, avec des scores de **83 - 83.92**, **0 - 80.67** et **82.08-83** qui contribuent le plus à cette dimension.
- Enfin, **AsieEst** ainsi que les scores **82.08-83** et **83.92 - 90.58** contribuent le plus à la dimension 4.

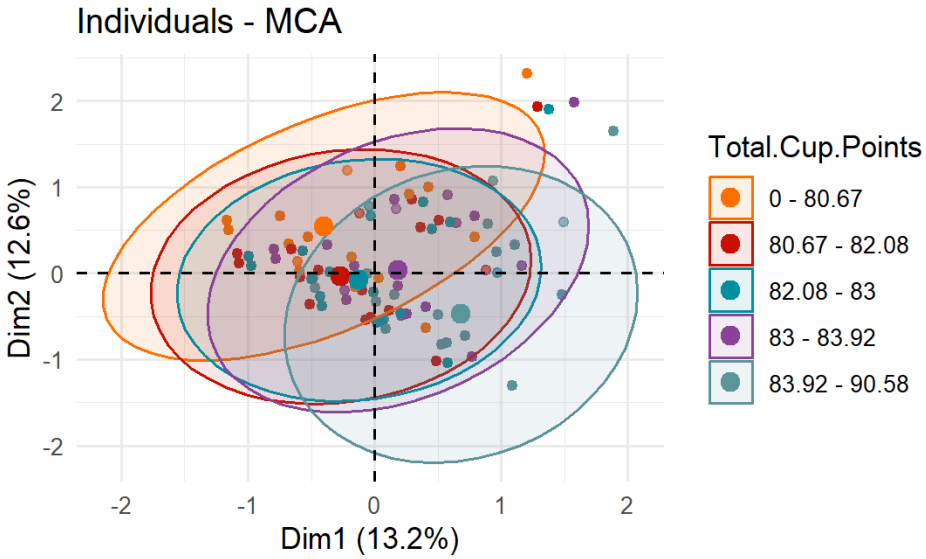
On peut par ailleurs noter qu'aucune modalité d'altitude contribue aux axes 3 et 4.

3.4 Individus en fonction des variables



Pour **l'humidité** qui est la variable qualitative supplémentaire, on s'aperçoit que les barycentres des modalités de cette variable sont très proches, et chaque groupe se supperpose. On en conclue que l'humidité ne sert pas à expliquer le score des cafés.

Pour **l'altitude**, on voit bien l'apparition de 3 groupes en fonction de l'altitude qu'on a observé sur le graphique des modalités. On retrouve bien les 2 groupes qui forment la bissectice avec le 1er groupe qui prend une altitude de **1-1310.64** mètres, le groupe 2 qui comprend les individus ayant une altitude de **1310.64-3500** mètres et enfin le groupe NA qui contient les valeurs extrêmes ainsi qu'une concentration de pays au centre qui se sont retrouvés dans ce groupe du fait qu'ils avaient des valeurs aberrantes d'altitude. On voit qu'il y a quand même une superposition entre ces différents groupes qui doivent sûrement être du à des individus qui prennent des modalités similaires.



Pour **le score du café**, on peut remarquer qu'on a une relation linéaire. En effet, en commençant du haut on peut voir qu'on a les scores de cafés les plus bas et plus on descend et plus on obtient des scores de café

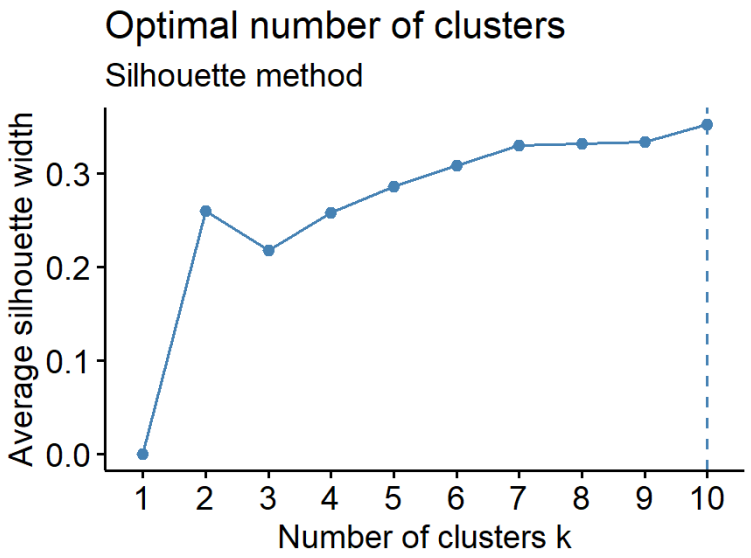
élevés. On voit cette même relation avec les barycentres avec 2 groupes qui sont assez proches, qui sont les individus des groupes **80.67 - 82.08** et **82.08 - 83** .

D'autre part, on constate un groupe d'individus extrêmes en dehors des ellipses qui prend chacune des modalités du score et qui suit aussi cette relation.

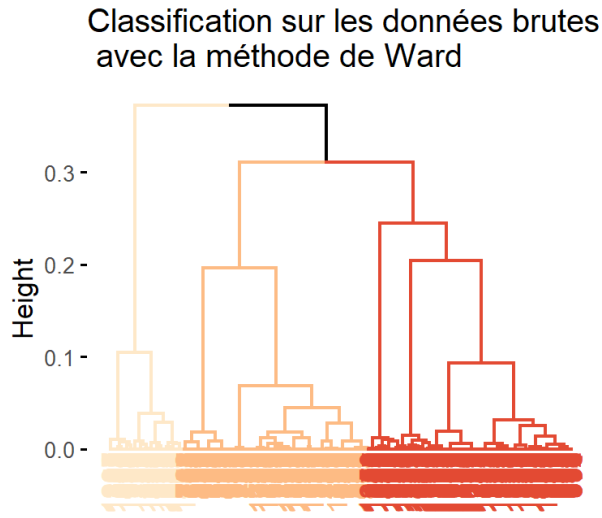
4 Classification

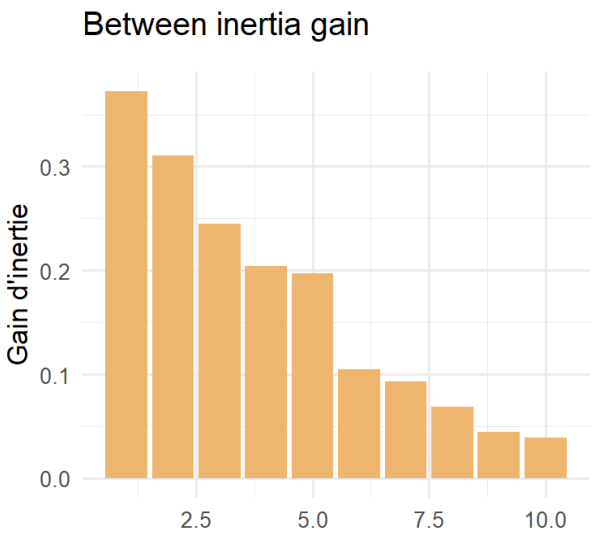
On va faire une classification avec consolidation pour avoir directement des classes homogènes. Par défaut, R fait une classification avec 6 groupes.

4.1 Nombre de clusters



Même si la méthode de Silhouette nous montre qu'il faudrait garder 2 groupes, on décide d'en prendre **3** pour retrouver les groupes trouvés en ACM, car en prenant 2 groupes on va mélanger tous les individus qui sont sur la bissectrice alors qu'on a vu qu'ils se divisaient en 2 groupes.

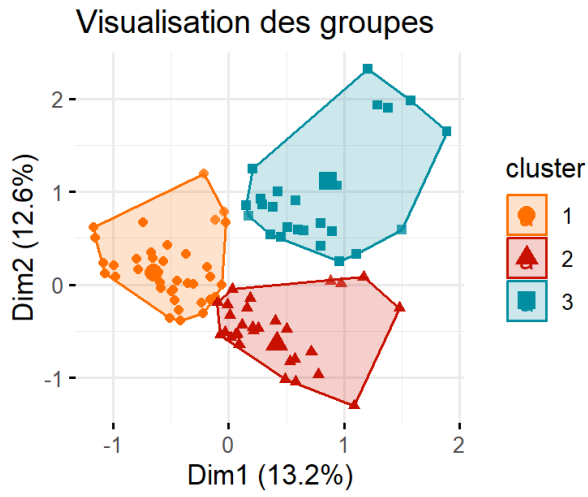


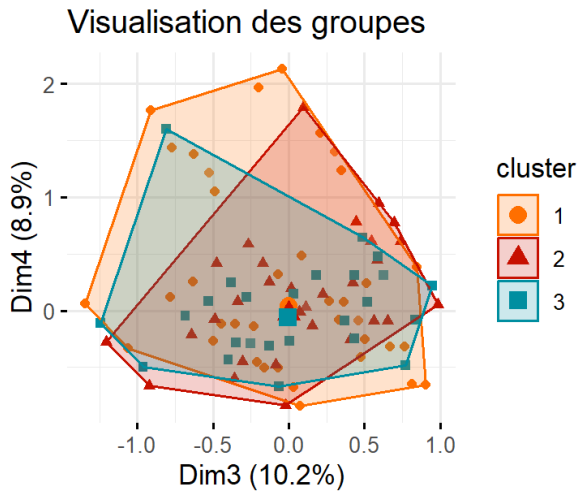


On peut voir grâce à ces graphiques une baisse importante du gain d'inertie intra à partir du 4ème cluster, ce qui se justifie aussi grâce au dendogramme.

En faisant le test du χ^2, on voit que les variables expliquent très bien la partition. On continue donc notre étude avec **3 clusters**.[3]

4.2 Visualisation des groupes





D'après cette représentation graphique, on peut voir que les groupes sont bien distingués sur l'axe **(F1,F2)**, on retrouve 2 groupes qui sont en effet les groupes qu'on a étudié sur la bissectrice, qui comportent les individus prenant pour la plupart des altitudes de **1 - 1310.64**m pour le groupe 1 et **1310.64 - 3500**m pour le cluster 2 comme on avait pu le voir sur le graphique des individus en fonction des altitides.

[3] On peut voir les résultats du test sur la table 17 en annexe

On retrouve aussi le groupe avec les individus prenant les modalités extrêmes, c'est-à-dire **NA** et **l'Océanie**, ainsi qu'un groupe d'individus plus nombreux qui est plus proche du centre qui s'est retrouvé dans le cluster 3 car ils prennent la modalité **NA**.

Sur l'axe **(F3,F4)**, tous les groupes se supperposent, on n'arrive pas à faire de conclusion sur ces dimensions donc on poursuivra notre étude uniquement sur les axes 1 et 2.

4.3 Composition des clusters

Table 8: Composition du cluster 1

	Cla/Mod	Mod/Cla	Global	p.value	v.test
altitude_mean_meters=1 - 1310.64	89.62	81.73	41.78	0.00	28.79
PaysOrigine=AmeriqueNord	93.90	37.68	18.39	0.00	17.79
Total.Cup.Points=0 - 80.67	73.80	32.63	20.25	0.00	10.44
PaysOrigine=AsieEst	92.39	13.87	6.88	0.00	9.80
Total.Cup.Points=80.67 - 82.08	55.36	25.29	20.93	0.00	3.59
PaysOrigine=AmeriqueCentrale	39.88	22.51	25.86	0.01	-2.57
PaysOrigine=AmeriqueSud	33.54	17.94	24.51	0.00	-5.17
altitude_mean_meters=1310.64 - 3500	19.81	17.46	40.36	0.00	-16.09

Dans le cluster 1 :

- Les individus de ce cluster viennent principalement des régions d'Amérique telles que : **38%** des individus sont d'**AmeriqueNord**, **22.5%** d'**AmeriqueCentrale** et environ **18%** d'**AmeriqueSud**.
- **58%** des individus qui prennent les modalité de score de **0 - 80.67** et **80.67 - 82.08** sont dans ce cluster. On peut donc voir que les individus qui prennent les modalités de scores les plus faibles se sont retrouvés dans le cluster 1.
- **82%** des individus de ce cluster ont une altitude de **1 - 1310.64**.

On peut voir l'opposition avec l'altitude **1310.64 - 3500** où seulement **17.46%** des individus de ce cluster prennent cette modalité malgré le fait qu'on avait vu graphiquement que ces 2 groupes d'altitude se supperposaient.

Table 9: Composition du cluster 2

	Cla/Mod	Mod/Cla	Global	p.value	v.test
altitude_mean_meters=1310.64 - 3500	80.00	85.38	40.36	0.00	27.18
PaysOrigine=Afrique	95.68	30.63	12.11	0.00	16.75
Total.Cup.Points=83.92 - 90.58	77.13	39.33	19.28	0.00	14.38
PaysOrigine=AmeriqueCentrale	47.11	32.21	25.86	0.00	4.10
PaysOrigine=AmeriqueSud	44.82	29.05	24.51	0.00	2.98
Total.Cup.Points=80.67 - 82.08	31.79	17.59	20.93	0.02	-2.35
PaysOrigine=AsieSud	25.56	4.55	6.73	0.01	-2.52
altitude_mean_meters=NA	7.53	3.56	17.86	0.00	-11.66
PaysOrigine=AmeriqueNord	5.69	2.77	18.39	0.00	-12.75
altitude_mean_meters=1 - 1310.64	10.02	11.07	41.78	0.00	-18.66

Dans le cluster 2 :

- **100%** des individus prenant les modalités **Afrique, AmeriqueCentrale, AmeriqueSud et AsieSud** sont dans le cluster 2. Les individus dans ce cluster se divisent principalement ainsi : **30.63%** des individus de ce cluster viennent d'**Afrique**, **32.2%** d'**AmériqueCentrale**, **29%** d'**AmériqueSud**, et **4.55%** d'**AsieSud**.
- Environ **85.38%** des individus de ce cluster prennent la modalité **1310.64 - 3500**. On retrouve aussi que **11%** des individus du cluster ont une altitude de **1 - 1310.64** mètres.
- Il y a environ **39.3%** des individus qui prennent la modalité de score de **83.92 - 90.58**, **17.6%** qui ont un score de **80.67 - 82.08**, on peut donc voir l'opposition entre les régions qui ont le meilleur score de café et ceux qui ont un des scores les plus faibles.

Table 10: Composition du cluster 3

	Cla/Mod	Mod/Cla	Global	p.value	v.test
altitude_mean_meters=NA	90.38	98.63	17.86	0.00	31.52
PaysOrigine=Oceanie	100.00	33.79	5.53	0.00	16.90
Total.Cup.Points=0 - 80.67	22.88	28.31	20.25	0.00	3.14
PaysOrigine=AmeriqueSud	21.65	32.42	24.51	0.00	2.90
PaysOrigine=AsieSud	24.44	10.05	6.73	0.04	2.04
PaysOrigine=AmeriqueCentrale	13.01	20.55	25.86	0.05	-1.99

- **100%** des individus venant d'**Océanie** sont dans ce cluster et représentent environ **34%** des individus.
- Plus de **60%** des individus viennent d'**AmeriqueSud**, d'**AmeriqueCentrale** et d'**AsieSud**.
- **98%** des individus de cette classe prennent la modalité **NA** pour l'altitude, ce qui explique que ce groupe était détaché comparé aux 2 autres.

Dans ce cluster, la modalité commune de ces régions est **NA**.

4.4 Parangons

Table 11: Parangon du cluster 1

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
1064	0 - 80.67	1 - 1310.64	AmeriqueSud	0 - 0.1
1073	0 - 80.67	1 - 1310.64	AmeriqueSud	0.1 - 0.11
1076	0 - 80.67	1 - 1310.64	AmeriqueSud	0 - 0.1
1095	0 - 80.67	1 - 1310.64	AmeriqueSud	0.1 - 0.11
1115	0 - 80.67	1 - 1310.64	AmeriqueSud	0.1 - 0.11

D'après ce tableau, on voit que l'**AmériqueSud** a un score de **0 - 80.67**, alors qu'en AFC on trouvait un score plus élevé. Cette différence peut venir du fait qu'on a recodé la variable **Total.Cup.Points**.

Table 12: Parangon du cluster 2

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
741	82.08 - 83	1 - 1310.64	Afrique	0.11 - 0.28
1315	82.08 - 83	1 - 1310.64	Afrique	0.11 - 0.28
1316	82.08 - 83	1 - 1310.64	Afrique	0.11 - 0.28
1321	82.08 - 83	1 - 1310.64	Afrique	0.11 - 0.28

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
500	82.08 - 83	1310.64 - 3500	Afrique	0.11 - 0.28

Les parangons du cluster 2 sont des individus d'**Afrique** avec un score de **82.08 - 83** et majoritairement l'altitude la plus basse.

Table 13: Parangon du cluster 3

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
1060	0 - 80.67	NA	AmeriqueSud	0.11 - 0.28
1061	0 - 80.67	NA	AmeriqueSud	0.1 - 0.11
1068	0 - 80.67	NA	AmeriqueSud	0 - 0.1
1102	0 - 80.67	NA	AmeriqueSud	0 - 0.1
1131	0 - 80.67	NA	AmeriqueSud	0 - 0.1

Pour le cluster 3, les parangons viennent d'**AmeriqueSud**, on retrouve bien la modalité **NA**, qui est associée au score de café le plus faible soit de **0 - 80.67**.

4.5 Individus extrêmes

Table 14: Individus extrêmes du cluster 1

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
802	80.67 - 82.08	1 - 1310.64	AsieEst	0 - 0.1
807	80.67 - 82.08	1 - 1310.64	AsieEst	0.11 - 0.28
826	80.67 - 82.08	1 - 1310.64	AsieEst	0.1 - 0.11
845	80.67 - 82.08	1 - 1310.64	AsieEst	0 - 0.1
886	80.67 - 82.08	1 - 1310.64	AsieEst	0.1 - 0.11

Pour les indivivus extrêmes du cluster 1, on se retrouve avec des individus d'**AsieEst** qui prennent la modalité **1 - 1310.64** pour l'altitude, et un des scores de café les plus bas, soit **80.67 - 82.08**.

Table 15: Individus extrêmes du cluster 2

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
2	83.92 - 90.58	1310.64 - 3500	Afrique	0.11 - 0.28
9	83.92 - 90.58	1310.64 - 3500	Afrique	0 - 0.1
5	83.92 - 90.58	1310.64 - 3500	Afrique	0.11 - 0.28
10	83.92 - 90.58	1310.64 - 3500	Afrique	0 - 0.1
15	83.92 - 90.58	1310.64 - 3500	Afrique	0 - 0.1

Les individus extrêmes du cluster 2 viennent d'Afrique avec une altitude de **1310.64 - 3500** mètres avec un score de café assez élevé, soit **83 - 83.92**.

Table 16: Individus extrêmes du cluster 3

	Total.Cup.Points	altitude_mean_meters	PaysOrigine	Moisture
1067	0 - 80.67	NA	Oceanie	0 - 0.1
1077	0 - 80.67	NA	Oceanie	0.1 - 0.11
1093	0 - 80.67	NA	Oceanie	0.11 - 0.28
1101	0 - 80.67	NA	Oceanie	0.11 - 0.28
1114	0 - 80.67	NA	Oceanie	0 - 0.1

Pour les individus du cluster 3, on retrouve encore l'**Océanie** avec cette fois-ci le score de café le plus faible et toujours l'altitude **NA**.

5 Conclusion

On peut conclure qu'en plus de la variable de provenance, l'altitude a un impact sur l'indice de qualité du café. **On a pu voir que plus l'altitude est élevé plus le café est de meilleur qualité**. Sur la classification, les différents clusters nous ont montré qu'en fonction de la région et de l'altitude on pouvait avoir des scores significativement différents. Mais on peut nuancer ces résultats qui sont sûrement du fait que l'on a réduit le nombre de modalités dans le recodage de la variable de l'indice de qualité du café. On peut aussi souligner le fait que géographiquement, les régions sont vastes et les pays composants les régions sont inégalements réparties comme il a été énoncé dans les tables 2 et 3. Au sein d'une même région, les altitudes peuvent aussi être différents ce qui nous pousse à nuancer nos résultats.

6 Annexe

‘

Table 17: résultat du test du khi deux

	p.value	df
altitude_mean_meters	0	4
PaysOrigine	0	12
Total.Cup.Points	0	8
Moisture	0	4

Table 18: proportion d’observations par modalité d’altitude

	1 - 1310.64	1310.64 - 3500	NA
proportion	41.8%	40.4%	17.9%