Nous travaillons sous PYTHON notebook dans cette séance.

Supports de référence :

- [**Cours 1**] « Classification de variables », https://cours-machine-learning.blogspot.com/p/clustering.html
- [**Cours 2**] « Classification de variable qualitatives – Classification de modalités », https://cours-machine-learning.blogspot.com/p/clustering.html
- [**Vidéo 1**] « Clustering de variables quantitatives », https://www.youtube.com/watch?v=Wwq20qxS2E8
- [**Vidéo 2**] « Clustering de variables qualitatives – Classification des modalités », https://www.youtube.com/watch?v=QVK3i44oGx8
- [**Livre**] « Pratique des méthodes factorielles avec Python », https://cours-machine-learning.blogspot.com/p/ouvrages.html

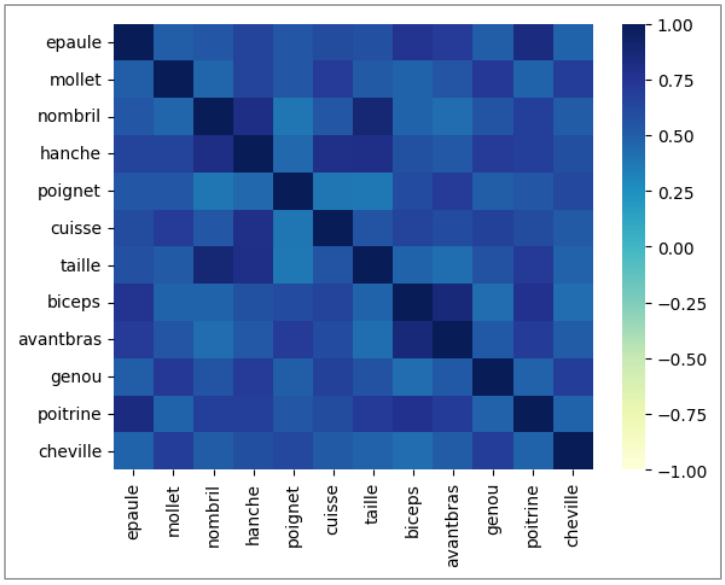
1. Classification de variables quantitatives

Nous travaillons avec le fichier « **body_dataset.xlsx** ». Il recense les circonférences des différentes parties du corps de personnes (poignet, cheville, coude, etc.). Une variable supplémentaire, le poids, permet également de les caractériser. On cherche à savoir si les morphologies des différentes parties de notre anatomie sont liées entres elles c.-à-d. si on a de grosses chevilles, a-t-on forcément des gros genoux, et pour autant est-ce que les poignets ou le tour de taille sont à l'avenant ? Etc.

Importation et inspection des données

1. Importez le fichier « **body_dataset.xlsx** » (pandas.read_excel)(Vidéo 1, 07:45). Affichez les première lignes (head). Affichez les informations relatives à la base de données (info). Combien d'observations disposons-nous ? (247). Combien de variables ? (13)
2. « Poids » fait office de variable illustrative. Isolez les autres descripteurs dans une structure spécifique. Combien de variables disposez-vous maintenant ? (12, forcément) (Vidéo 1, 08:15)
3. Pour ces variables actives, calculez et affichez la matrice des corrélations des variables prises deux à deux (corr, méthode de 'pearson') (Vidéo 1, 08:35). Quel est le type de la structure obtenue ? (type) Que constatez-vous concernant les valeurs des corrélations obtenues ? (elles sont toutes positives)

4. Pour chaque variable, quelle est la valeur minimale de sa corrélation avec les autres (min, voir le rôle de l'option 'axis')
5. Pour chaque variable, laquelle lui est la moins corrélée ? (idxmin) Que constatez-vous dans les résultats ? (pas de symétrie)
6. Affichez le graphique heatmap (https://seaborn.pydata.org/generated/seaborn.heatmap.html ; package « seaborn ») permettant de rendre compte visuellement des corrélations entre les variables (Vidéo 1, 09:19). Assurez-vous que les bornes vont bien de -1 à +1. Essayez de choisir palette (cmap) qui attribue des couleurs foncées aux corrélations élevées (https://www.python-graph-gallery.com/92-control-color-in-seaborn-heatmaps) (j'ai fait le choix suivant, mais les goûts et les couleurs… à vous de voir…)



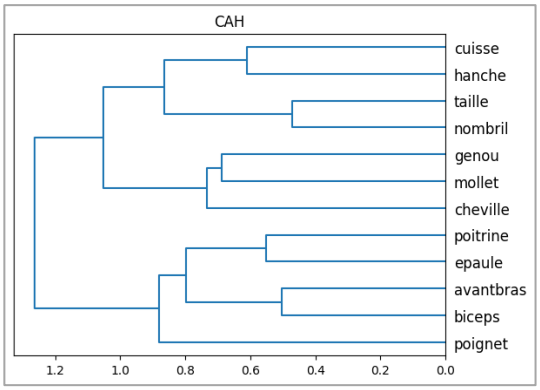
CAH sur variables

7. Dérivez une matrice de distances à partir des corrélations croisées (r) (Vidéo 1, 12:38). Elle est égale à :

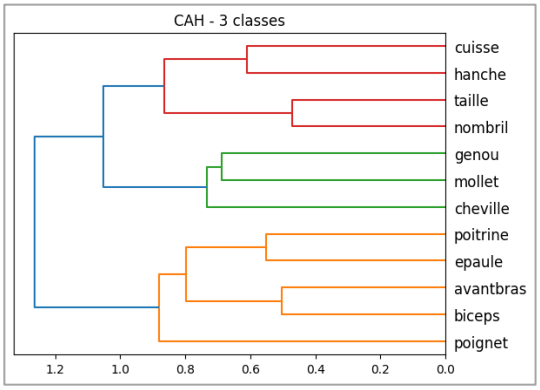
$$D = \sqrt{1 - r^2}$$

8. Vérifiez que D est bien symétrique et que la diagonale est formée par des valeurs nulles (quelle stratégie utiliseriez-vous pour cela ? pour ma part, je suis passé par les fonctions de la librairie « numpy », diag, transpose, …)
9. Pour que la librairie « scipy » puisse lancer la CAH à partir de la matrice de distances, il nous faut vectorisez cette dernière à l'aide de la fonction squareform (https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.squareform.html). Comparez le résultat avec la matrice D (Vidéo 1, 13:36). Retrouvez-vous vos valeurs ?

10. Lancez la CAH via la méthode de Ward (ward, librairie « Scipy », https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.ward.html) (Vidéo 1, 15:00). Affichez le résultat. Comment comprendre la structure proposée ?
11. On peut la comprendre mais elle n'est pas très avenante, nous en convenons tous. Affichez le dendrogramme (https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html ; dendrogram) (Vidéo 1, 15:40), sans préjuger du seuil de partitionnement pour l'instant (color_threshold = 0).



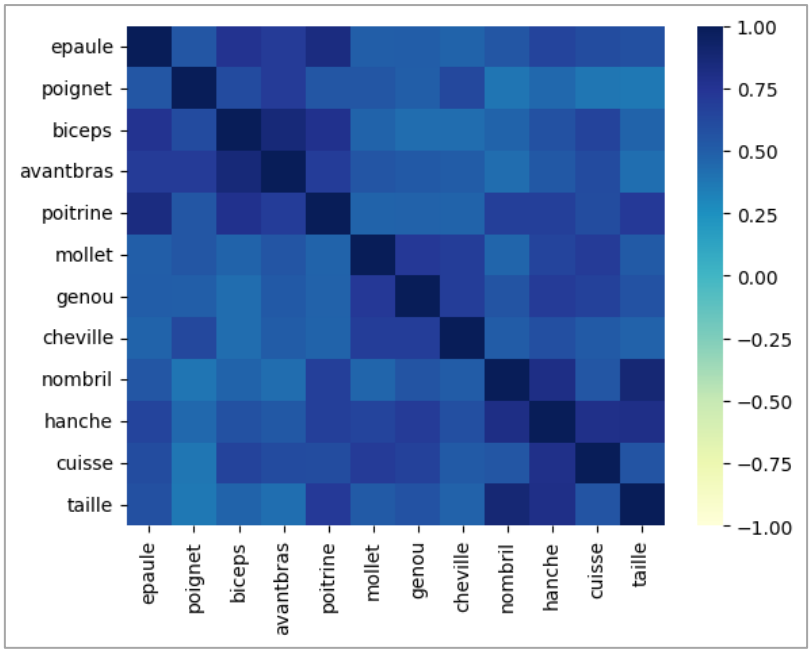
12. Quelles sont les solutions possibles en termes de nombre de groupes ? (2 ou 3 ; mais « 2 » paraît toujours « évidente » dans une CAH).
13. Nous optons pour une **partition en 3 groupes**. En termes de distance d'agrégation, quelle serait le seuil de coupure pour les matérialiser dans le dendrogramme ? Affichez le dendrogramme avec l'option adéquate (color_threshold = ???) (Vidéo 1, 16:43 ; prenez une valeur en adéquation avec votre dendrogramme, ne reprenez pas telle quelle la valeur utilisée dans la vidéo). Comment sont constitués les différents groupes ? Quelles parties du corps présentent des morphologies similaires ? Est-ce que ces résultats sont en adéquation avec les (le peu de) connaissances que nous avons du corps humain ?



14. Construisez le vecteur qui permet de rattacher chaque variable à son groupe (fcluster ; https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html , voir attentivement l'option criterion) (Vidéo 1, 17:07 ; même remarque que

précédemment). Affichez le vecteur, comment le lisez-vous ? Affichez les noms des variables (columns) pour faire la correspondance.

15. Calculez les effectifs par groupe [de bas en haut dans mon dendrogramme ci-dessus : (5, 3, 4) ; n'oublions pas que la librairie numérote les groupes de manière arbitraire…] (Vidéo 1, 18:04)
16. Créez une copie de la base des variables actives. **Réorganisez ses colonnes de manière à ce que les variables appartenant au même groupe soient contigües (**hé non, ce n'est pas dans la vidéo, à vous de voir comment procéder**)**. Recalculez la matrice des corrélations et affichez le « heatmap ». Le résultat est-il probant ? (oui, on distingue mieux les « blocs » de variables corrélées ; on se rend compte aussi que le clustering est passé à côté de certaines informations, ex. la situation ambivalente de « poitrine »…)



17. « Poitrine » est visiblement corrélée avec les variables de son groupe, mais elle semble également liée aux variables du groupe (nombril, hanche, cuisse, taille). Pour vous en assurer, calculez la moyenne des carrés des corrélations (on aurait pu utiliser les corrélations brutes puisque nous savons qu'elles sont toutes positives) de « poitrine » avec les variables groupées c.-à-d. moyenne des carrés des corrélations avec (épaule, poignet, biceps, avant-bras) (0.52), avec (mollet, genou, cheville) (0.226), avec (nombril, hanche, cuisse, taille) (0.45). Que constatez-vous ? (on peut s'inspirer de Vidéo 1, 21:20 ou 22:58, mais « poitrine » n'est pas une variable supplémentaire dans notre contexte, il faut en tenir compte quand vous calculerez la moyenne des corrélations avec son propre groupe).

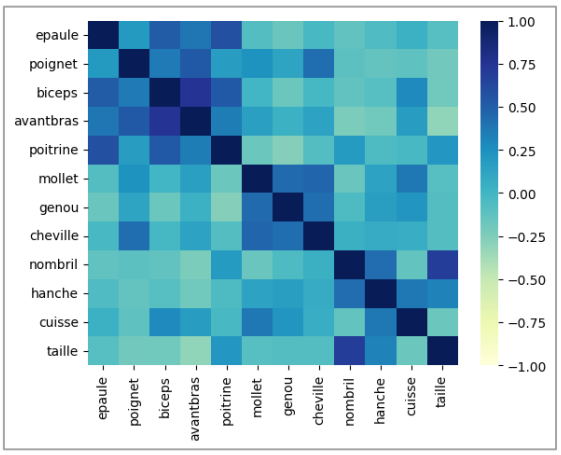
Analyse de la variable illustrative « poids »

18. Il n'est pas nécessaire d'être grand clerc pour savoir que notre morphologie est directement liée à notre poids. Calculez la moyenne des carrés des corrélations de poids avec les variables groupées (0.498, 0.478, 0.654) (Vidéo 1, 22:58). Comment lisez-vous les résultats ? (le poids pèse [😊] sur les différentes parties de notre anatomie, en particulier dans la région de la taille, hé ouais, c'est une lapalissade, mais tellement vraie…)

Travailler sur les corrélations partielles

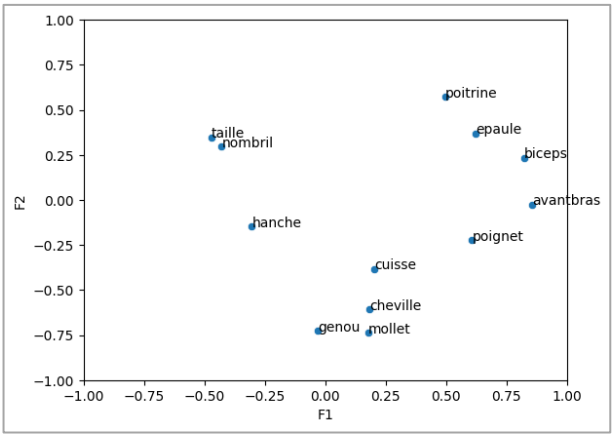
Manifestement, le « poids » pèse (ok, ok …) sur l'ensemble des mesures. Il faudrait retirer son influence en effectuant une étude à « poids égal » c.-à-d. à poids égal (si toutes les personnes avaient la même corpulence), est-ce que celles qui ont de grosses chevilles ont aussi de gros genoux ? Mais aussi de gros poignets ? etc. Pour ce faire, nous travaillons à partir des corrélations partielles (https://fr.wikipedia.org/wiki/Corrélation_partielle).

19. A partir du dataset des variables triées selon leur groupe d'appartenance, calculez la matrice des corrélations des variables actives prises deux à deux conditionnellement au « poids » (reprenez la formule décrite sur la page Wikipédia – il va falloir programmer un peu là, j'ai utilisé une double boucle en ce qui me concerne… ; sinon, des packages la propose, ex. https://pingouin-stats.org/build/html/generated/pingouin.partial_corr.html, ou d'autres, à vous de savoir les utiliser à bon escient). Assurez-vous que la diagonale contient bien la valeur « 1 » (la corrélation d'une variable avec elle-même est égale à 1, partielle ou pas). Que constatez-vous dans les valeurs obtenues ? (certaines sont négatives)
20. Affichez le graphique « heatmap » des corrélations partielles. Commentaires ?

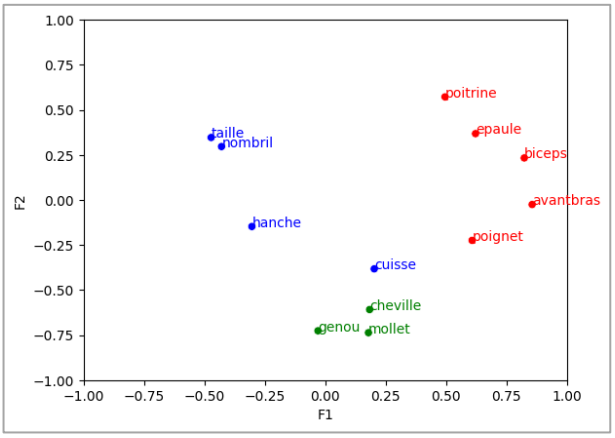


21. On souhaite réaliser une ACP (analyse en composantes principales) à partir des corrélations partielles. Calculez les valeurs et vecteurs propres (numpy.linalg.eig) de la matrice ci-dessus. (Livre, page 39).

22. Affichez les valeurs propres, puis leur somme. Est-ce que le résultat correspond à celui qui est attendu ?
23. Quelle est la part d'information restituée par les 2 premiers facteurs ? (45.06%).
24. Affichez les 2 premiers vecteurs propres.
25. Calculez les corrélations des variables avec les 2 axes (elles sont égales au produit des vecteurs propres avec la racine carrée des valeurs propres – cf. Livre, pages 40 et 41)
26. Créez un graphique qui représente ces corrélations des variables aux axes factoriels. Que constatez-vous en termes de proximité des variables ?



27. Refaites le même graphique, mais en illustrant les points (avec des couleurs par exemple) à l'aide des groupes d'appartenance issus du clustering de variables. Est-ce que les 2 approches sont cohérentes ? (plutôt oui, mis à part peut-être « cuisse » mais, d'une part n'oublions pas que les 2 premiers facteurs ne représentent que 45% de l'information disponible, d'autre part sa position dans le repère factoriel n'est pas si choquant que cela à vrai dire)



2. Classification de variables qualitatives

Le fichier « **zoo_dataset.xlsx** » recense les caractéristiques de (**n = 101**) animaux : présence ou non de poils, de plumes, ponte d'œufs, etc. Chacun d'entre eux est associée à un « genre » (mammal [mammifère], fish [poisson], etc.) qui fait office de variable illustrative. Nous souhaitons mettre au jour les associations de modalités les plus marquantes. Par exemple, les animaux qui pondent des œufs ont-ils des dents (tout le monde connait l'histoire des poules et des dents) ? Produisent-ils du lait ? etc.

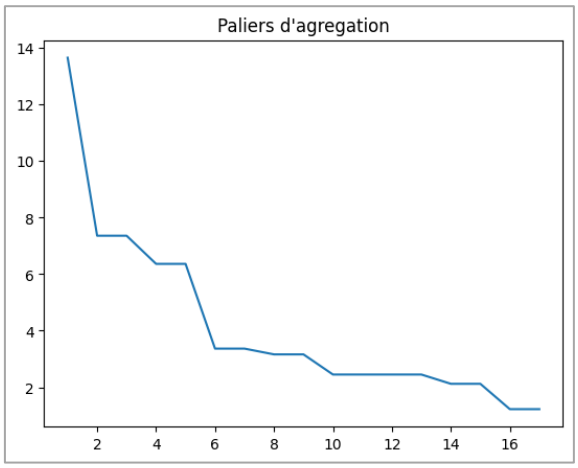
Clustering des modalités

28. Importez « **zoo_dataset.xlsx** » (pandas.read_excel). Affichez les premières lignes (head).
29. Affichez les informations sur le jeu de données (info). Combien d'observations et de variables disposons-nous ? (101, 10). Quel est le type des variables ?
30. Isolez dans une structure spécifique les variables actives (toutes sauf « genre »).
31. Effectuez un codage disjonctif complet (pandas.get_dummies) (Vidéo 2, 08:45), en incluant toutes les modalités (drop_first = False). Quel est le type de l'objet obtenu ? (type) Vérifiez sa structure (info). Quel est le type des variables indicatrices (on parle aussi de « dummy variable ») maintenant ?
32. Calculez les proportions par modalité (describe). Quel est pourcentage des animaux qui ont des poils ? (42.57%) Des plumes ? (19.8%) Etc.
33. Pour pouvoir la manipuler plus facilement, transformez la matrice « pandas » en structure « numpy » (.values) (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.values.html). Affichez les données.
34. Définissez une fonction qui calcule, à partir de 2 vecteurs numériques (contenant des valeurs 0/1), le carré de l'indice de Dice (Vidéo 2, 09:27).
35. Appliquez cette fonction à chaque paire des variables indicatrices (j'ai initialisé une matrice de zéros, puis j'ai effectué une double boucle en ce qui me concerne). Passez le résultat en racine carrée pour obtenir une matrice de distance. Affichez ses 5 premières lignes et colonnes à des fins de vérification (Vidéo 2, 10:03).

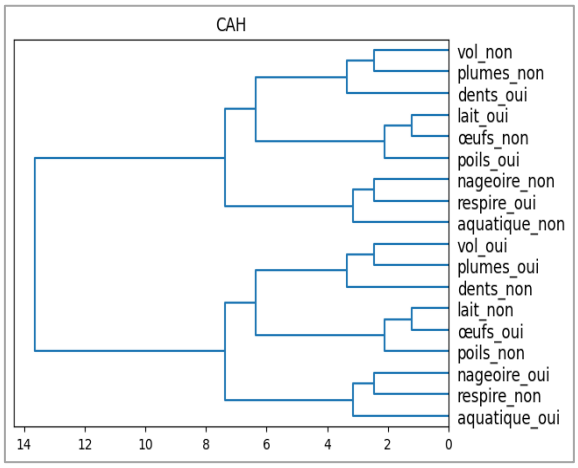
```
[[0.         7.1063352  5.61248608 4.35889894 6.78232998]
 [7.1063352  0.         4.35889894 5.61248608 2.12132034]
 [5.61248608 4.35889894 0.         7.1063352  4.41588043]
 [4.35889894 5.61248608 7.1063352  0.         5.56776436]
 [6.78232998 2.12132034 4.41588043 5.56776436 0.        ]]
```

36. A l'instar de ce que nous avons effectué pour le clustering de variables quantitatives, vectorisez la matrice à l'aide de squareform() puis lancer la CAH avec la méthode ward(). Affichez le contenu du résultat (Vidéo 2, 12:12).

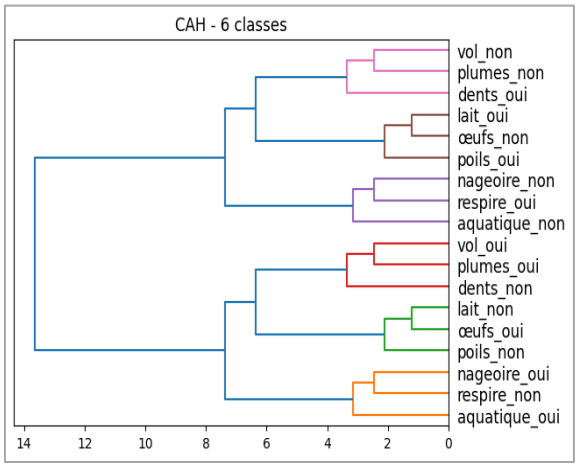
37. A partir des paliers d'agrégation fournis par ward() (la troisième colonne), créez un graphique les représentant pour chaque nombre de clusters. Quel nombre de groupes suggéreriez-vous ?



38. Affichez le dendrogramme (dendrogram). Votre intuition ci-dessus est-elle confirmée pour ce qui est du nombre de groupes ? (K = 2 ou K = 6 sont les 2 solutions possibles)



39. Mettons que nous partons sur la solution (K* = 6). Matérialisez-les dans le dendrogramme.



40. Réalisez le découpage en 6 classes (fcluster). Affichez les indicatrices de groupes. Calculez les effectifs par groupe. Effectuez le rapprochement avec le dendrogramme obtenu à l'étape précédente. Les résultats sont-ils cohérents ? (oui, vaut mieux d'ailleurs…)

Traitement de la variable illustrative « genre »

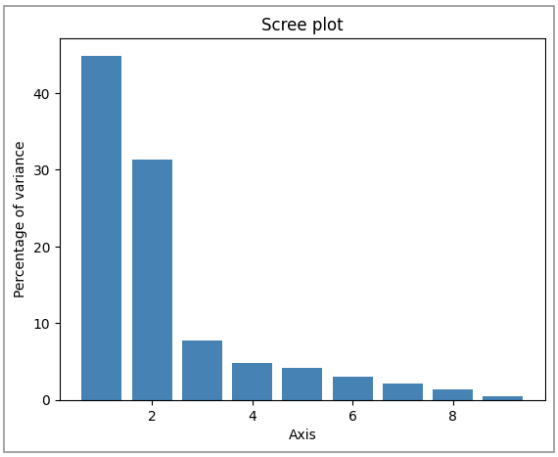
Nous souhaitons identifier les catégories que l'on pourrait associer aux mammifères (genre = « mammal ») (Vidéo 1, à partir de 19:18).

41. Transformez en indicatrices 0/1 la variable « genre » (pandas.get_dummies).
42. Pour chaque catégorie, calculez la moyenne des carrés de l'indice de Dice des variables la composant et l'indicatrice « mammal ». Quel groupe est le fortement lié (qui minimise la distance) avec le caractère mammifère ? ([poils = oui, œufs = non, lait = oui], avec moyenne = 1.5). Que penser de ce résultat ? Il correspond à ce que vous connaissez des animaux ? (un peu non ? les mammifères ne pondent pas des œufs, allaitent leur progéniture, etc.)

Analyse des correspondances multiples (ACM)

Nous souhaitons mettre en parallèle notre catégorisation avec les représentations de l'analyse factorielle des correspondances multiples.

43. Sans sortir du notebook, installez à la volée le package fanalysis (https://github.com/OlivierGarciaDev/fanalysis/) (dans une cellule du notebook, introduisez la commande `!pip install fanalysis` ; mettez-la en commentaire ensuite pour éviter que la librairie soit réinstallée à chaque exécution de votre projet).
44. Ce tutoriel devrait vous aider pour la suite : https://github.com/OlivierGarciaDev/fanalysis/blob/master/doc/mca_tutorial.ipynb ; cette vidéo aussi peut-être https://www.youtube.com/watch?v=oY0mtzcvRYk (à partir de **12:41**)
45. Instanciez l'ACM (MCA), passez en paramètre les labels des lignes et colonnes (row_labels, var_labels). Lancez (fit) l'analyse sur les variables actives non-recodées en 0/1.
46. Affichez les valeurs propres (.eig_). Regardez attentivement la documentation pour lire correctement les sorties. Quel est le pourcentage d'information restitué par les 2 premiers facteurs ? (76.22%).
47. Affichez le graphique des valeurs propres en pourcentage d'information restituée (.plot_eigenvalues)



48. Affichez les coordonnées des variables (.col_topandas).
49. Récupérez les 2 premières colonnes dans une structure data frame. Ajoutez-y l'indicatrice d'appartenance aux groupes.
50. Représentez alors dans le premier plan factoriel les points-modalités en les illustrant à l'aide de leur groupe d'appartenance (voir le graphique ci-dessous). Que constatez-vous ?