INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Exploration et Logiciel Statistique

Rapport projet AdultCensus

Louis Delvaux & Sophia Yazzourh

4 GMM - MMS

Année universitaire 2018-2019

Description

Q : Quelle est la population étudiée ? Quel est l'échantillon ?

La popoulation étudiée est la population recensée aux Etats-Unis en 1994. L'échantillon que l'on va utiliser est celui des 30153 individus obtenus après nettoyage.

Q : Commenter, justifier les différentes transformations opérées. Repérer les variables quantitatives, qualitatives. Certaines ('age', 'hoursWeek') sont présentes sous les deux types. Beaucoup de modalités on déjà été regroupées, notamment celles de 'nativCountry' (voir le programme) certaines variables sont rendues qualitatives ('capitalLoss' ou 'Gain').

Pour chaque variable qualitative, on renomme les modalités afin de pouvoir les manipuler plus facilement. Durant cette manipulation, on regroupe certaines modalités afin de ne plus avoir des modalités à trop faible effectif. Pour toutes ces transformations, on utilise la fonction *map* qui de plus nous permet de remplacer automatiquement la valeur des variables des individus possédant un ' ? ' par des 'NaN'.

Pour la variable cat_occup, on décide de ne pas renommer la modalité 'armed_forces' en 'Military'. Cette modalité sera alors affectée en 'Nan' et sera supprimée par la suite.

Pour les variables 'capitalGain' et 'capitalLoss', la distribution est biaisée à droite car il y a quelques valeurs avec une très grande abscisse et donc éloignées de la majorité des valeurs. Nous effectuons donc une transformation logarithmique afin de rapprocher ces valeurs extrêmes.

Ensuite, on transforme les variables quantitatives 'capitalGain' et 'capitalLoss' en variables qualitatives à 3 modalités, obtenues à partir du calcul de la médiane. Si les variables ont des valeurs négatives, elles sont placées dans la catégorie 'None'. Si elles sont comprises entre zéro et la médiane des valeurs positives, elles sont placées dans la catégorie 'cgLow'. Sinon, elles sont placées dans la catégorie 'cgHigh'.

On transforme la variable 'income' en variable binaire par rapport à un seuil de 50K pour avoir des modalités plus éclairantes pour l'analyse.

On crée une nouvelle variable qualitative 'ageQ' qui, à partir de la variable 'Age', situe un individu dans des modalités suivantes : 'Ag1','Ag2','Ag3','Ag4','Ag5'. On souhaite avoir le même effectif pour chaque modalité, on utilise la fonction 'qcut' utilisant les quantiles.

On crée une nouvelle variable qualitative 'hoursWeekQ', qui à partir de la variable 'hoursWeek', situe un individu dans des modalités suivantes : 'HW1','HW2','HW3', en fonction du nombre d'heures travaillées par semaine.

Enfin, on supprime tous les individus possédant des variables de valeur NaN.

Q : Quels sont ces graphiques ? Que dire de la transformation opérée ? La variable CapitalLoss subit le même traitement.

Ce sont un boxplot et un histogramme de la variable LcapitalGain (=log(1+CapitalGain)). La transformation logarithmique effectuée permet de réduire la dispersion de la variable, en rapprochant les valeurs extrêmes.

Q : Quel est ce problème général ? Qu'est-ce qui va se passer si on en calcule l'analyse des correspondances ? Quel est le problème plus spécifique des variables 'relationship' et 'sex' ?

Le problème générale est la redondance de certaines informations.

En effet, pour le tableau entre 'educNum' et 'education', on a beaucoup de cellules d'effectif nul. Pour les relations entre le 'relationship' et 'mariStat' les composantes 'married','husband' et 'wife' regroupent toutes la même information. Pour le tableau recoupant 'origEthn' et 'nativCountry', certaines cellules présentent des effectifs très faibles et sont donc non significatives.

Il faudra faire attention à n'utiliser qu'une des deux variables de chaque couple lors de l'analyse de correspondances afin d'éviter la redondance.

Le problème spécifique des variables 'relationship' et 'sex' est le fait qu'il y ait une femme dans la catégorie époux et un homme dans la catégorie épouse. Par la suite, on ne considérera que la variable 'sex'. Étant donné les effectifs importants des modalités 'Female' et 'Male', on ne supprime pas les deux individus à problème et on considère que ceux-ci sont des erreurs lors du recensement.

L'analyse des correspondances permet de préciser une liaison entre variables. Si certaines informations sont redondantes entre deux variables, l'analyse sera biaisée.

Q : Quel graphique ci-dessus ? Quelle interprétation ? Quel est le test ci-dessous ? Que doit vérifier la table pour que ce test soit valide ? Quelle est l'hypothèse H0 testée ? Conclusion.

C'est un mosaic plot de la table de contingence entre les variables 'sex' et 'occup'. La superficie de chaque case est proportionnelle à l'effectif de la modalité associée. On observe alors que certaines catégories de métiers sont plus exercées par des hommes ('blue-Collar'=ouvrier) et d'autres par des femmes ('admin'). De plus, cela met en évidence les métiers les plus exercés au sein d'un même sexe.

On effectue alors un test du Chi2. Il teste la liaison ou non entre les deux variables qualitatives 'sex' et 'occup' (H0 : les deux variables sont indépendantes en probabilité). Pour que ce test soit valide, il faut que tous les individus présentent une modalité et une seule et que chaque modalité soit observée au moins une fois.

La p-valeur de ce test étant nulle, on rejette H0, on considère donc que les deux variables sont liées. Il faudra donc faire attention lors des prochaines analyses à considérer cette liaison.

Q : Quel est le graphique ci-dessous ? Comment interpréter ? Quel test permettrait de confirmer ? Que dire de l'intérêt de cette variable [fnlwgt](Final sampling weight) ?

Ce graphique représente les boxplots de la variable 'fnlwgt' expliquée en fonction des modalités de la variable 'income' (revenus discrétisés en 2 modalités). Les boxplots semblent relativement identiques, on peut donc supposer qu'il n'y a pas de liaisons entre ces deux variables. Pour confirmer cette hypothèse, on peut effectuer un test du chi2.

Si il n'y a effectivement pas de liaison entre ces variables, on peut considérer que la variable 'fnlwgt' n'a pas d'intérêt pour expliquer la répartition de la variable 'income'.

Q : Que dire dans la figure ci-dessous de la liaison entre les variables 'educNum' et 'age' et de la localisation des points noirs (>50k) par rapport aux rouges (<50k).

Il ne semble pas avoir de forte liaison entre les variables 'educNum' et 'age'. Cependant, on observe un regroupement des points noirs (incHigh) dans la partie supérieure droite.

On observe que le nombre de points noirs augmente avec le nombre d'années d'études. Les revenus augmentent avec le nombre d'années passées sur les bancs de l'école. De plus, on observe que le revenu augmente également avec l'âge, ce qui correspondrait aux augmentations acquises avec l'expérience professionnelle.

Les revenues augmentent donc avec le nombre d'années d'étude et l'expérience.

Analyse en Composantes Principales

Q : Quelle est le graphe ci-dessous ? Quel est le cercle ? A quoi sert-il ?

Le graphe représente l'analyse des deux premières composantes principales. Avant de réaliser notre ACP, nous avons réduits nos données à la même échelle afin de pouvoir les placer sur le cercle des corrélations. Plus la variable se rapproche du cercle plus elle est bien représentée dans le plan factoriel et permet donc d'expliquer les composantes principales.

Le premier axe représente le nombre d'année d'études, l'âge et le nombre d'heures travaillées par semaine. Le second axe représente la différence entre le capital perdu et le capital gagné sur les investissements en bourse.

Le premier axe représente donc l'expérience professionnelle et le second les revenus en bourse.

Q : La représentation ci-dessous montre un artefact avec 3 paquets d'individus. A quoi est dû cet artefact ? Est-il utile à la compréhension des données ?

On observe trois groupes d'individus, ceux qui gagnent de l'argent en investissant en bourse en haut, ceux qui perdent de l'argent en investissant en bourse en bas et ceux qui n'investissent pas au centre. Cette représentation en 3 groupes provient de l'explication du second axe de l'ACP.

Ce graphe n'est pas particulièrement utile à la compréhension des données, il met en évidence un comportement logique.

Analyse factorielle des correspondances de la table 'occup' vs. 'education'

Q : Que conclure du test ci-dessous sur l'intérêt d'une analyse des correspondances ? Quelle est la table étudiée, ses dimensions ?

C'est un test du Chi2. Il teste la liaison ou non entre les deux variables qualitatives 'occup' et 'education' (H0 : les deux variables sont indépendantes en probabilité). La p-valeur de ce test étant nulle, on rejette H0, on considère donc que les deux variables sont liées.

On peut donc faire l'AFC pour préciser la liaison entre les deux variables. On étudie alors la table de contingence des 2 variables de dimension 7x7 (7 modalités pour 'occup' et 7 modalités pour 'education').

Q : Quelles sont les ACPs considérées dans cette analyse des correspondances, avec quelles métriques ?

L'AFC effectue une double ACP du tableau disjonctif. On calcule l'ACP des profils-colonnes avec les métriques D_r^{-1} et D_c et on calcule l'ACP des profils-lignes avec les métriques D_c^{-1} et D_r. La matrice D_r est la matrice diagonale des fréquences par lignes et la matrice D_c est la

matrice diagonale des fréquences par colonnes.

Q : Quelle est la matrice diagonalisée ?
La matrice diagonalisée dans l'AFC est le tableau de Burt mais ici, on effectue l'AFC sur le tableau disjonctif, ce qui est relativement identique.

Q : Quelle est la signification des valeurs (%) présentes dans les légendes ?
Les pourcentage dans les légendes représentent la part de chi2 expliqué.

Q : Donner en une ligne une signification à l'Axe 1.
L'axe 1 représente le niveau d'étude et le niveau hiérarchique du métier exercé. Les niveaux d'éducation 'bachelors', 'Masters', 'Doctorate' et 'Prof-School' sont liés avec les métiers 'Professional', 'White-Collar' et 'Sales' au niveau de l'axe 1. Parallèlement, les personnes exerçants des métiers de salariés sont liées à un nombre d'années d'études plus faible.

Analyse factorielle multiple des correspondances

Q : La section 2 montre le peu d'intérêt de l'ACP. L'AFCM est donc utilisée. Quel prétraitement a été mis en œuvre pour y faire intervenir toutes les variables ? Que faire des couples de variables posant problème (première section) ?
Pour faire intervenir toutes les variables dans l'AFCM, on a transformé les variables quantitatives en variables qualitatives. Pour les couples posant problème, nous décidons de ne considérer qu'une variable de chaque couple.

Q : Combien d'axes est-il raisonnable de retenir selon la figure ci-dessous ?
Selon l'éboulis des valeurs propres, on ne considère que les 2 premiers axes. On voit distinctement apparaître deux coudes (changement de signe dans la suite des différences d'ordre 2). On pourrait être pointilleux et en distinguer un troisième. Mais les variables sont trop proches et la séparation n'est pas net.

Q : Comment interpréter les axes 1 et 2 ci-dessous ? Que signifient les valeurs (%) ?
Il est difficile d'interpréter les axes directement. Cependant, on observe sur la diagonale montante que le revenu et le niveau d'éducation sont fortement liés (cf AFC) et sur la diagonale descendante, le sexe, la profession exercée et le nombre d'heure travaillée sont liés. Ces deux diagonales étant orthogonales, cela met en évidence une discrimination. On observe que les femmes se voient confier des postes à moindre responsabilité et avec moins d'heures de travail. Alors qu'il n'y a pas de différences de répartition entre hommes et femmes pour le niveau d'études, il y a tout de même une répartition inégale des responsabilités entre sexes. Les pourcentages des axes représentent l'inertie du nuage de valeurs propres. Elles ne sont pas des indicateurs de qualité pour l'AFCM.

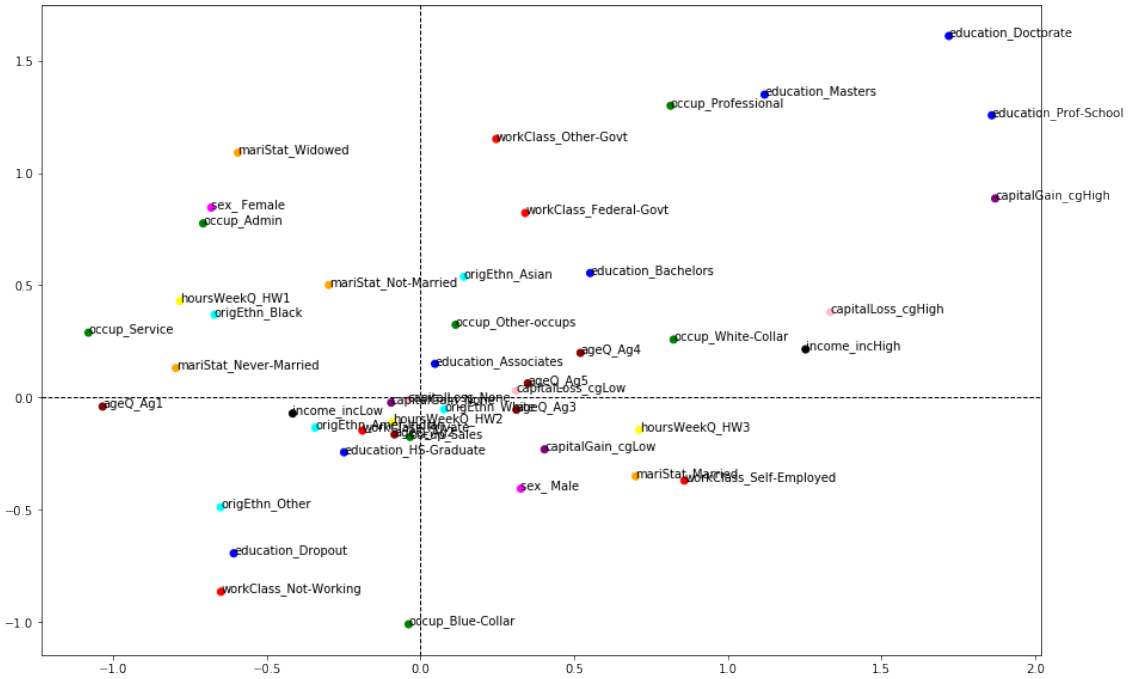

FIGURE 1 – Représentation des variables dans les deux premier axes de l'AFCM

Q : A partir de quelle AFC et donc de quelle ACP, quelle SVD le graphe ci-dessous a-t-il été obtenu ? La discrimination des individus de revenus inférieurs ou supérieurs à 50k sera-t-elle aisée ?

Le graphe ci-dessous a été obtenu par l'AFC du tableau disjonctif complet, ce qui correspond à l'ACP des profils-lignes (car on affiche seulement les individus), et donc à la SVD du tableau disjonctif complet.

La discrimination des individus n'est pas aisée, on observe bien deux groupes mais ils se chevauchent en partie. L'AFCM ne permet donc pas de séparer distinctement les deux groupe.

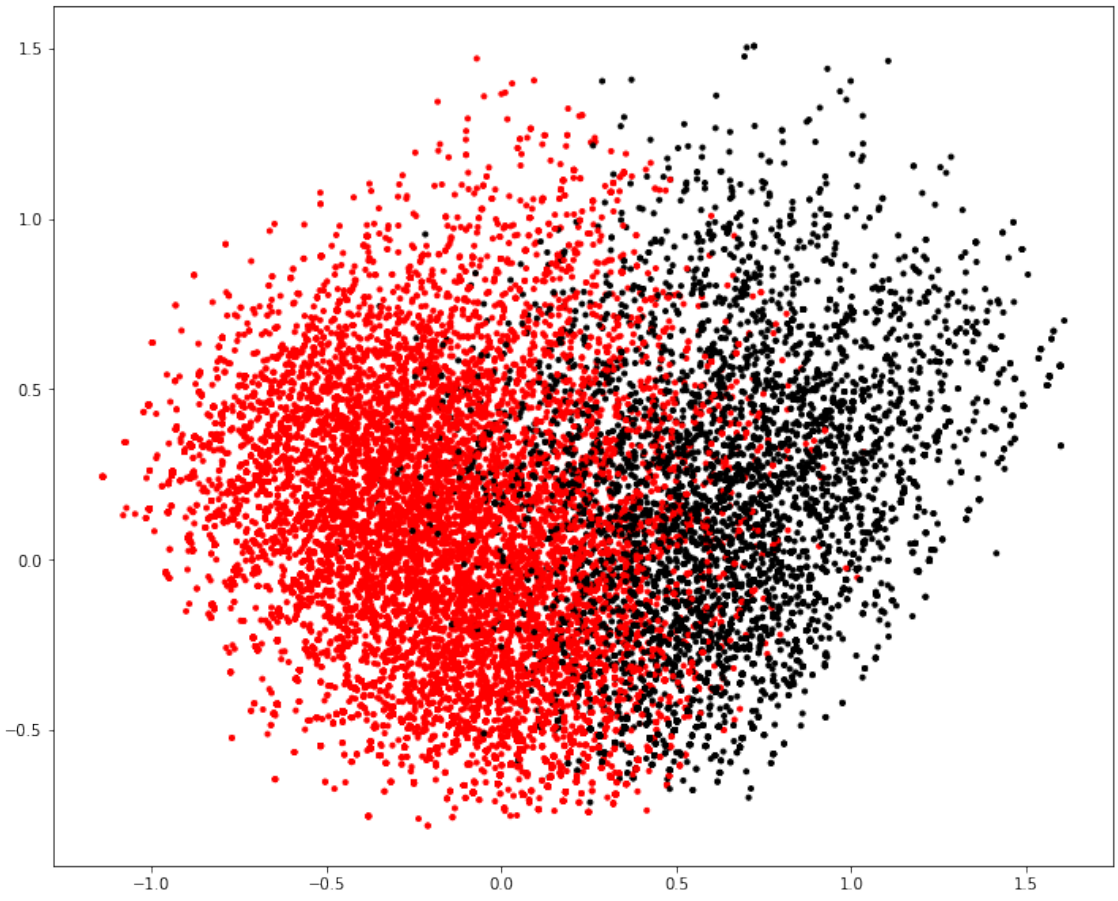

FIGURE 2 – Représentation des individus dans les deux premier axes de l'AFCM (incomeHigh en noir, incomeLow en rouge)

Classification non supervisée

Q : Pourquoi la classification ascendante hiérarchique des données précédentes ne marche-t-elle pas sur un ordinateur portable basique ? Quelle stratégie faudrait-il mettre en œuvre ?

La classification ascendante hiérarchique des données ne marche pas sur un ordinateur portable basique car la matrice des distances calculées est trop importante en mémoire et en temps de calcul. Pour résoudre ce problème, il faut réduire le nombre de données à considérer. Pour cela, on peut tirer aléatoirement un petit nombre de données qui sera représentatif de l'ensemble des données.

Q : Les commandes suivantes sont plus simplement exécutées. Quelle astuce est mise en œuvre ? Comment choisir le nombre de classes ?

Sur R, le dendrogramme et la séparation des classes par les k-means ont été réalisés sur un échantillon tiré aléatoirement. Ici, nous les avons réalisés sur toutes les données. Le temps de calcul est cependant un peu long.

On réalise une classification ascendante hiérarchique en utilisant la distance euclidienne (par défaut) et la méthode de ward pour calculer les distances entre deux classes. On choisit le nombre de classes à l'aide du graphique de la décroissance de la variance interclasses. La présence d'une rupture dans cette décroissance nous aide dans ce choix. Ici, on choisit 5 classes conformément au graphique de R car la décroissance sous Python nous donne généralement moins de classes (2 en général), ce qui nous semble pas suffisant pour notre analyse.

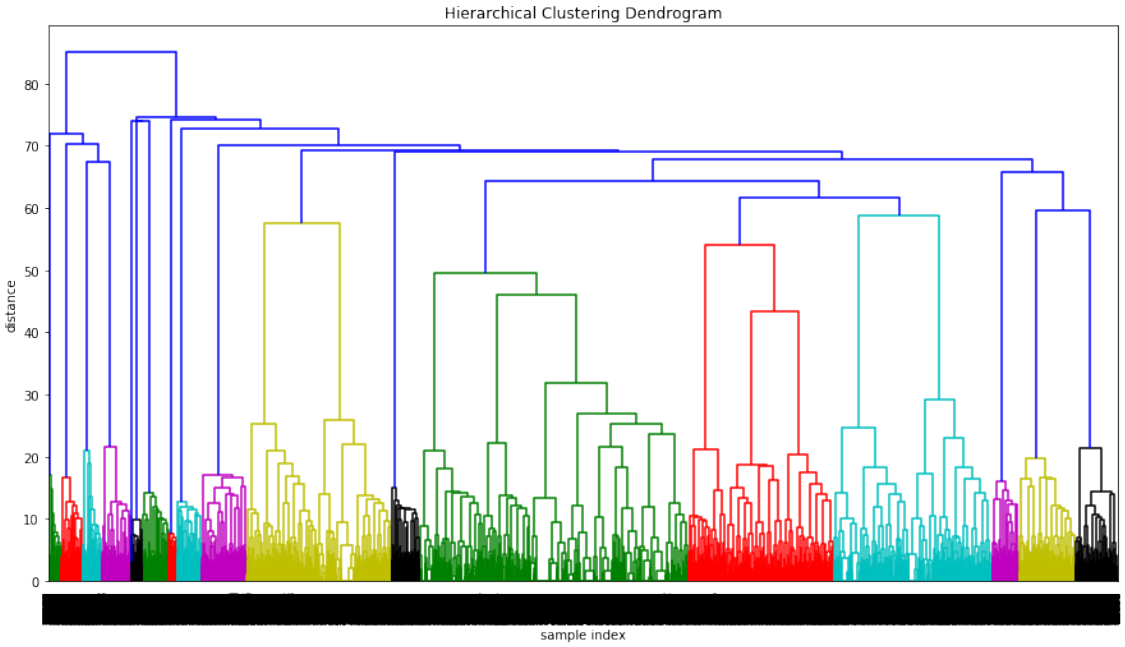

FIGURE 3 – Dendrogramme de la classification ascendante hiérarchique

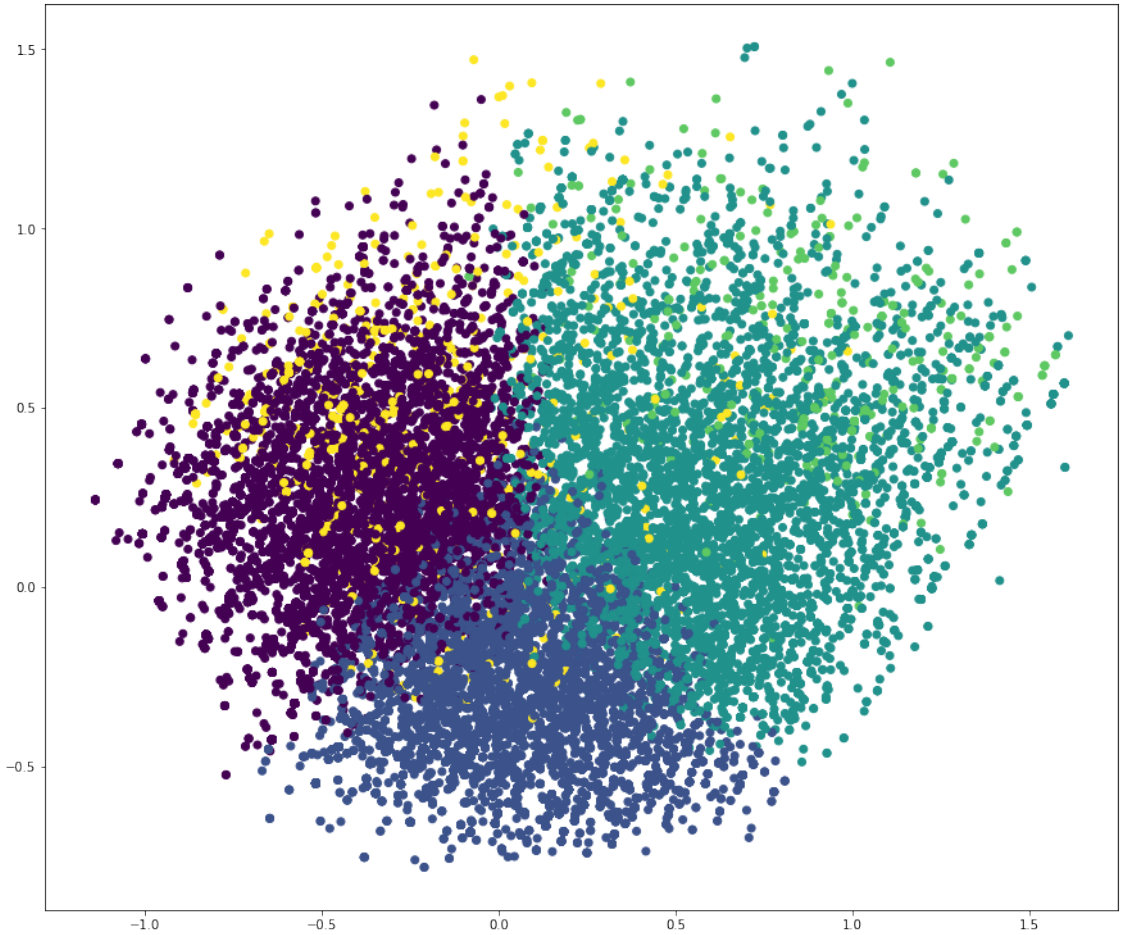

FIGURE 4 – Représentation des classes dans les composantes de l'AFCM

Q : Expliquer comment les commandes ci-dessous permettent de construire une interprétation des classes. Interpréter ces classes.

On crée un nouveau dataFrame où on rajoute la classe affectée à chaque individu. On réalise une nouvelle AFCM sur le tableau disjonctif complet auquel on a rajouté les classes attribuées précédemment.

La représentation des variables nous permet d'observer quelles variables sont proches des variables "classe" et ainsi de les interpréter.

La classe C0 est représentée par les modalités correspondant aux femmes et à leur position dans le monde du travail. Comme vu précédemment, les femmes occupent des postes administratifs et de services et travaillent peu d'heures par semaine.

La classe C1 est représentée par les modalités correspondant aux ouvriers, aux sans-emplois et aux personnes ayant quittées le système scolaire. D'une manière générale, cette classe représente plutôt les hommes a plus faible niveau d'études.

La classe C2 est représentée par les modalités correspondant aux personnes avec un haut niveau d'études, occupant des postes de cadres et ayant un revenu important. De plus, on peut remarquer qu'il s'agit de personnes investissant en bourse. Il s'agit également plutôt d'hommes au vu de la position de la modalité 'sex_Male'.

La classe C3 est représentée seulement par la modalité 'education_Doctorate' et représente donc les docteurs.

La classe C4 est représentée seulement par la modalité 'mariStat_Widowed' et représente donc les veuves.

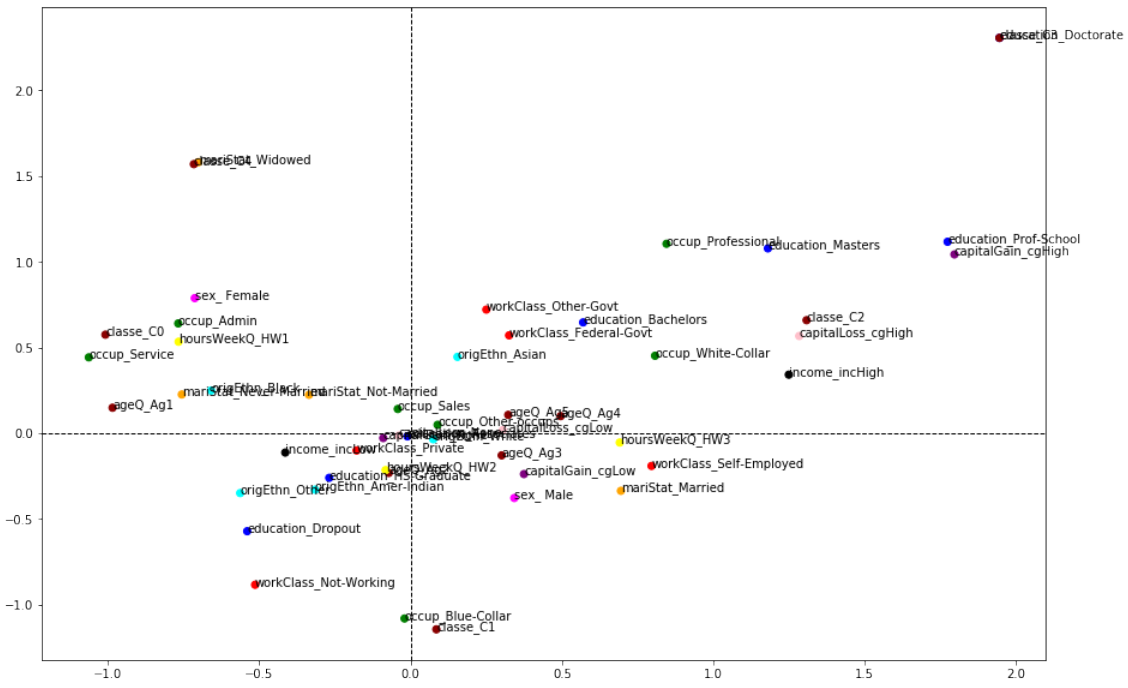

FIGURE 5 – Représentation des variables et des classes dans les deux premier axes de l'AFCM