المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN

INSEA

Institut National de Statistique

et d'Economie Appliquée

Rapport de Projet

Filière : M2SI

Relations entre méthodes d'études, personnalité et la performance académique des étudiants de l'INSEA et leur état de santé

Analyse de Correspondances Multiples
ACM

RÉALISÉ PAR:
MANSSOUR WASSIMA
NADI SOKAINA

ENCADRÉ PAR:
Prof. WAFAA HANNOUN

ANNEE UNIVERSITAIRE: 2021-2022

Sommaire

Table de figures

Liste des tableaux

Introduction

L'Analyse des Correspondances Multiples (ACM ou MCA pour multiple correspondence analysis) est une extension de l'analyse factorielle des correspondances pour résumer et visualiser un tableau de données contenant plus de deux variables catégorielles. On peut aussi la considérer comme une généralisation de l'analyse en composantes principales lorsque les variables à analyser sont catégorielles plutôt que quantitatives.

L'ACM est généralement utilisée pour analyser des données d'enquête ou de sondage.

L'objectif est:

- ✓ Détecter les groupes de personnes ayant un profil similaire dans leurs réponses aux questions
- ✓ Identifier Les associations entre les catégories des variables
- ✓ Synthétiser l'information portant sur les variables qualitatives

Caractéristiques de l'ACM :

- L'AFC met en correspondance deux ensembles de caractères : l'ensemble des lignes et l'ensemble des colonnes cependant l'ACM croise un ensemble de lignes avec un seconde ensemble, celui des modalités de réponse à plusieurs questions.
- On pose à n individus (les lignes) m questions.
- Les questions sont sous forme disjonctive complète, c'est-à-dire que, pour chaque question, il y a obligatoirement le choix d'une modalité et d'une seule. C'est de plus un **codage binaire** : 1 ou 0.
- L'ACM permet d'étudier les relations qui existent entre les modalités des différentes questions
- L'ACM est une méthode de description statistique multidimensionnelle d'un tableau de données qualitatives utilisée particulièrement pour l'analyse des fichiers d'enquête.

Intérêt de l'ACM

L'utilisation de l'AFCM présente beaucoup d'avantages:

- le premier avantage tient à ce que les tableaux sont rendus homogènes en (0,1) grâce au codage disjonctif complet.
- Le second avantage est de voir apparaitre explicitement toutes les modalités des variables, ce qui facilite l'interprétation.
- Le troisième avantage est de permettre de décrire les liaisons entre variables quantitatives quand on les suppose non-linaires.

Dans ce qui suit, nous allons décrire comment calculer et visualiser l'analyse des correspondances multiples avec le logiciel R en utilisant les packages *FactoMineR* (pour l'analyse) et *factoextra* (pour la visualisation des données). De plus, nous montrerons comment révéler les variables les plus importantes qui contribuent le plus à expliquer les variations dans le jeu de données. Nous continuons en expliquant comment prédire les résultats pour les individus et les variables supplémentaires. Enfin, nous allons démontrer comment filtrer les résultats de l'ACM afin de ne conserver que les variables les plus contributives.

Exemple de fichier d'enquête

Les variables de l'analyse sont :

StylAprtnsg	Qu'est ce qui décrit le plus votre style d'apprentissage ?	StylAprtnsg 1 : collaboratif StylAprtnsg 2 : compétitif StylAprtnsg 3 : indépendant
IdealMomntTrvail	À quel moment de la journée révisez-vous le mieux ?	IdealMomntTrvail 1 : matin IdlMomtTrvail 2 : après-midi IdealMomntTrvail 3 : soir
StfctEtatSante	Etes-vous satisfait de votre état de santé ?	StfctEtatSante 1 : Oui StfctEtatSante 2 : Non
NbrRatt	Combien de Rattrapage avez-vous ?	NbrRatt 1 : 1 NbrRatt 2 : 2 NbrRatt 3 : plus de 3
MmbrClub	Combien de clubs avez-vous rejoindre ?	MmbrClub 1 : 0 MmbrClub 2 : 1 MmbrClub 3 : 2 MmbrClub 4 : plus de 3

Tableau 1: Exemple de variables tirées d'un questionnaire

Ind	StylAprtnsg	IdealMomnt Trvail	StfctEtatSante	NbrRatt	MmbrClub
Ind1	1	2	1	2	1
Ind2	1	2	2	2	1
Ind3	2	2	2	1	2
Ind4	3	1	1	1	1
Ind5	2	2	2	1	2
Ind6	3	1	1	3	3
Ind7	1	2	2	3	1
Ind8	2	3	1	3	4
Ind9	2	2	2	2	4
Ind10	3	2	2	2	3
Ind11	3	2	1	3	2
Ind12	1	1	2	3	1
Ind13	1	1	1	1	2
Ind14	1	1	2	2	3
Ind15	2	3	1	3	4
Ind16	3	2	1	2	2
Ind17	3	1	2	2	1
Ind18	3	3	1	1	1
Ind19	3	2	2	3	1
Ind20	2	2	1	1	2

Tableau 2: Exemple de données tirées d'un questionnaire

Partie Théorique

Tableau de données, tableau disjonctif complet, tableau de Burt

Tableau de données :

La forme matricielle du tableau de données se présente comme suit :

$$X=\begin{pmatrix} x_{11} & \cdots & x_{12} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Où

- n : le nombre d'individus,
- p : le nombre de variables qualitatives
- x_{hk} : l'individu i qui possède la modalité k

$x_{hk} \in M_k$, avec M_k qui est l'ensemble des modalités de la k'éme variable. On note m_k= card(M_k) le nombre de modalités de k.

- m =m_1+ . . . +m_p : le nombre total de modalités.

NB : Avant toute calcul, il faut être conscient du fait qu'il n'est pas question d'appliquer une méthode d'analyse des données à ce tableau, en effet, les nombres dans le tableau représentent simplement des codages des réponses individuelles, **on n'a pas le droit de les additionner ou de les multiplier**.

On utilise deux méthodes pour transformer ce tableau de données en tableau "analysable".

- ✓ **La première méthode** pour analyser un tel tableau consiste à définir pour chaque variable autant de réponses possibles qu'il y a de modalités : Tableau disjonctif complet.
- ✓ **La seconde méthode** consiste à calculer pour chaque couple de variables, le tableau de leur croisement : Tableau de Brut.

Tableau disjonctif complet T de dimension n x m :

Chaque colonne k est l'indicatrice de la modalité k avec :

$$T=\begin{pmatrix} t_{11} & \cdots & t_{12} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nm} \end{pmatrix}$$

Où : $t_{hk} = \begin{cases} 1, \text{ si l'individu h possède la modalité k} \\ 0, \text{ sinon} \end{cases}$

Prenons la variable StylAprtnsg par exemple. Elle admet 3 modalités de réponses.

On peut la représenter par un vecteur de 3 questions-modalités différentes (StylAprtnsg1, StylAprtnsg2, StylAprtnsg3).

Un individu qui aurait :

La repense 1 à la variable StylAprtnsg aura les réponses :
La réponse 2 à la variable StylAprtnsg aura les réponses :
La réponse 3 à la variable StylAprtnsg aura les réponses :

StylAprtnsg 1	StylAprtnsg 2	StylAprtnsg 3
1	0	0

StylAprtnsg 1	StylAprtnsg 2	StylAprtnsg 3
0	1	0

StylAprtnsg 1	StylAprtnsg 2	StylAprtnsg 3
0	0	1

Tableau disjonctif complet :

	StylAprtnsg			IdealMomntTrvail			StfctEtatSante		NbrRatt			MmbrClub			
ind	StylAprtng 1	StylAprtnsg 2	StylAprtsng 3	IdlMmntTrvail 1	IdlMmntTrvail 2	IdlMmntTrvail 3	StsfctEtatSante 1	StsfctEtatSante 2	NbrRatt 1	NbrRatt 2	NbrRatt 3	MbrClub 1	MbrClub 2	MbrClub 3	MbrClub 4
1	1	0	0	0	1	0	1	0	0	1	0	1	0	0	0
2	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0
3	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0
4	0	0	1	1	0	0	1	0	1	0	0	1	0	0	0
5	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0
6	0	0	1	1	0	0	1	0	0	0	1	0	0	1	0
7	1	0	0	0	1	0	0	1	0	0	1	1	0	0	0
8	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1
9	0	1	0	0	1	0	0	1	0	1	0	0	0	0	1
10	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0
11	0	0	1	0	1	0	1	0	0	0	1	0	1	0	0
12	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0
13	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0
14	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0
15	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1
16	0	0	1	0	1	0	1	0	0	1	0	0	1	0	0
17	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0
18	0	0	1	0	0	1	1	0	1	0	0	1	0	0	0
19	0	0	1	0	1	0	0	1	0	0	1	1	0	0	0
20	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0

Tableau 3: Tableau disjonctif complet

Remarques :

→ La somme des éléments d'une colonne représente le nombre d'individus ayant choisi la modalité représentée par cette colonne.

→ La somme des éléments d'une ligne est constante et est égale au nombre de variables étudiées.

→ La technique de l'analyse des correspondances multiples consiste à appliquer les techniques de l'analyse des correspondances (AFC) à un tel fichier.

Tableau Brut B de dimension m x m :

Il s'agit d'un tableau symétrique qui rassemble les croisements deux à deux de toutes les variables, c'est `à dire tous les tableaux de contingence des variables deux à deux.

On construit un tableau qui regroupe tous les croisements possibles : Tableau de Burt

$$B=TT'=\begin{pmatrix} b_{11} & \cdots & b_{12} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mm} \end{pmatrix}$$

Où $b_{kk'}=\sum_{h=1}^{n} t_{hk}t_{hk'}$: le nombre d'individus qui possèdent la modalité k et la modalité k'.

Sur la diagonale $b_{kk'}=n_k$: le nombre d'individus qui possèdent la modalité k.

Exemple :

	StylAp1	StylAp2	StylAp3	IdlMmtTr1	IdlMtTr2	IdlMmtTr3	StsfctSnt1	StsfctSnt2	NbrRatt1	NbrRatt2	NbrRatt3	MbrClub1	MbrClb2	MbrClub3	MbrClb4	total
StylAprtng1	6	0	0	3	3	0	2	4	1	3	2	4	1	1	0	30
StylAprtng2	0	6	0	4	0	2	3	3	3	1	2	0	3	0	3	24
StylAprtng3	0	0	8	3	4	1	5	3	2	3	3	4	2	2	0	40
IdlMmtTr1	3	4	3	6	0	0	3	3	2	2	2	3	1	2	0	35
IdlMmtTr2	3	0	4	0	11	0	4	7	3	5	3	4	5	1	1	53
IdlMmtTr3	0	2	1	0	0	3	3	0	1	0	2	1	0	0	2	15
StsfctSnt1	2	3	5	3	4	3	10	0	4	2	4	3	4	1	2	50
StsfctSnt2	4	3	3	3	7	0	0	10	2	5	3	5	2	2	1	50
NbrRatt1	1	3	2	2	3	1	4	2	6	0	0	2	4	0	0	30
NbrRatt2	3	1	3	2	5	0	2	5	0	7	0	3	1	2	1	49
NbrRatt3	2	2	3	2	3	2	4	3	0	0	7	3	1	1	2	49
MbrClub1	4	0	4	3	4	1	3	5	2	3	3	8	0	0	0	40
MbrClub2	1	3	2	1	5	0	4	2	4	1	1	0	6	0	0	30
MbrClub3	1	0	2	2	1	0	1	2	0	2	1	0	0	3	0	15
MbrClub4	0	3	0	0	1	2	2	1	0	1	2	0	0	0	3	15
Total	**30**	**24**	**40**	**35**	**53**	**15**	**50**	**50**	**30**	**49**	**49**	**40**	**30**	**15**	**15**	**525**

Tableau 4: Tableau Brut B de dimension m x m

TDP : un tableau de contingence particulier :

En ACM, on traite le tableau disjonctif complet T comme un tableau de contingence :

T=

	1 ... k ...m	
1 . h . n	. . ... t_{hk}	$t_{h.}$=p
	$t_{.k}$= nk	t.. =np

On déduit de ce tableau une matrice de fréquence F particulière (associée à un TDC).

F=

	1 … k …m	
1 . h . n	. . … $\frac{t_{hk}}{np}$ … . .	$t_{h.}=\frac{1}{n}$
	$\frac{n_k}{np}$	

On définit $f_{hk}=\frac{1}{np}$ si l'individu i possède la modalité k et f_{hk} = 0 sinon.

On notera donc :

- r = (. . ., 1 n, . . .)' ∈ R^n : le vecteur des poids des individus.
- c = (. . ., $\frac{n_k}{np}$, . . .)' ∈ R^m : le vecteur des poids des modalités.
- Dr = diag(r)
- Dc = diag(c)

On déduit de ce tableau de fréquences une matrice de profil-lignes L et une matrice des profil-colonnes C.

Matrice des profil-lignes L :

L=

	1 … k …m
1 . h . n	. . … $\frac{f_{hk}}{f_{.k}}=\frac{t_{hk}}{p}$ … . .
c'	$\frac{n_k}{np}$

On a :

- L= Dc^{-1} F
- Profil ligne moyen : c

Matrice des profil-colonnes C :

C =

	1 … k …m	**r**
1 . h . n	. . … $\frac{f_{hk}}{f_{.k}}=\frac{t_{hk}}{n_k}$ … . .	$\frac{1}{n}$

On a :

- C= F Dr^{-1}
- Profil colonne moyen : r

Deux nuages de points pondérés :

On considère dans la suite les deux nuages centrés de point suivants :

- Le nuage des n point-individus centrés de R^m c'est à dire les n lignes de la matrice des profil-lignes centrés $L = Dr^{-1}(F - rc')$. Chaque point-individu est pondère par $\frac{1}{n}$.
- Le nuage des m point-modalités centrés de R^n c'est `a dire les m lignes de la matrice des profil-colonnes centres $C = (F - rc')\,Dr^{-1}$. Chaque point-modalité est pondère par $\frac{n_k}{np}$.

Distance de χ^2:

En ACM, on utilise la distance du χ2 pour comparer deux individus décrits par deux points de R^m (deux profil-lignes) ou deux modalités décrites par deux points de R^n ((deux profil colonnes). En ACP, les individus et les variables étaient les lignes et les colonnes d'une même matrice (la matrice des données centrées-réduites). En ACM, les individus et les modalités sont les lignes et les colonnes de deux matrices différentes (resp. la matrice des profil-lignes et la matrice des profil-colonnes). Pour comparer deux individus ou deux modalités, on utilise en ACM la distance du χ2.

Distance du χ^2 entre deux individus : métrique Dc^{-1}

$$d^2(h, h') = \sum_{k=1}^{m} \frac{1}{f_{.k}} \left(\frac{t_{hk} - t_{h'k}}{p}\right)^2 = \frac{n}{p} \sum_{k=1}^{m} \frac{1}{n_k} (t_{hk} - t_{h'k})^2$$

Donc deux individus sont proches s'ils possèdent les mêmes modalités.

Distance du χ^2 entre deux modalit´es : m´etrique Dr^{-1}

$$d^2(k, k') = \sum_{h=1}^{n} \frac{1}{f_{h.}} \left(\frac{t_{hk}}{n_k} - \frac{t_{hk'}}{n_{k'}}\right)^2 = n \sum_{h=1}^{n} \left(\frac{t_{hk}}{n_k} - \frac{t_{hk'}}{n_{k'}}\right)^2$$

Donc deux modalités sont proches si elles sont possédées par les mêmes individus.

Inertie totale :

En AFC, on pouvait interpréter statistiquement l'inertie des nuages de points (profil-lignes et colonnes) en termes de χ^2 /n mesurant l'indépendance entre les deux variables qualitatives. En ACM ce n'est plus le cas puisque l'on a :

$$I(L) = I(C) = \frac{m}{p} - 1$$

On a donc l'inertie qui dépend de $\frac{m}{p}$, le nombre moyen de catégories par variable.

NB : On en tire l'idée que pour avoir une analyse des données qui ne donne pas systématiquement avantage à telle ou telle variable, il faudra s'arranger autant faire se peut pour que toutes les variables aient à peu près le même nombre de modalités et que les effectifs des différentes modalités ne soient "pas trop dissemblables".

Cette recommandation sert surtout quand il faudra rendre qualitatives des variables quantitatives comme par exemple le revenu des ménages, la taille des entreprises en

nombre de salariés, le chiffre d'affaire des entreprises, ... Il faut résister à la tentation de faire des classes d'amplitude égales pour privilégier des classes d'effectifs égaux.

AFC du tableau disjonctif complet :

Effectuer une ACM consiste à appliquer l'AFC au TDC c'est à dire à effectuer une ACP pondérée des nuages des point-individus et des point-modalités (centrés). On reprend donc les résultats du cours d'AFC.

Coordonnées factorielles des individus et des modalités

L'ACM est l'analyse du triplet suivant :

$$(nTD^{-1} - 1_{n*m}, \frac{1}{n}1_n, \frac{1}{np}D)$$

où D = diag (. . . , n_k. . .) est la matrice diagonale des effectifs des modalités et 1_{n*m}est la matrice de dimension n × m de terme général 1.

Remarques :

- On définit parfois l'ACM comme l'analyse du triplet suivant :

$$(\mathrm{nTD}^{-1}, \frac{1}{\mathrm{n}}1_{\mathrm{n}}, \frac{1}{\mathrm{np}}\mathrm{D})$$

Cela correspond au cas où l'on considère les nuages de points non centrés. Dans ce cas, on observe en plus, un vecteur propre trivial 1_n associé à la valeur propre 1. Le fait de centrer les nuages de points permet donc de l'éliminer.

- En ACM il y a au plus r = min(n − 1, m − p) valeurs propre non nulle dans le cas centré.

Relations barycentriques

$$x_{h\alpha} = \frac{1}{\sqrt{\lambda_\alpha}}\frac{1}{p}\sum_{k\in S_h} y_{k\alpha}$$

$$y_{h\alpha} = \frac{1}{\sqrt{\lambda_\alpha}}\frac{1}{n_s}\sum_{h\in I_k} x_{h\alpha}$$

Où S_h est l'ensemble des modalités prises par l'individu h et I_k est l'ensemble des individus qui possèdent la modalité k. Ces relations quasi-barycentriques donnent trois modes de représentation simultanée des individus et des modalités :

- Comme card(S_h) = p, la première relation dans se lit : "La coordonnées factorielle de l'individu h sur l'axe α est égale, à $\frac{1}{\sqrt{\lambda_\alpha}}$ près, à la moyenne arithmétique simple des coordonnées des modalités qu'il possède".

⇒ Première possibilité de représentation simultanée : les individus au barycentre des modalités.

- Comme card(I_k) = n_k, la seconde relation dans (5) se lit : "La coordonnées factorielle de la modalité k sur l'axe α est égale, à $\frac{1}{\sqrt{\lambda_\alpha}}$ près, à la moyenne arithmétique simple des coordonnées des individus qui la possèdent".

⇒ Seconde possibilité de représentation simultanée : les modalités au barycentre des individus.

- Une troisième représentation simultanée des individus et des modalités consiste à représenter sur un même graphique les moyennes arithmétiques dilatées par $\frac{1}{\sqrt{\lambda_\alpha}}$.

Sur ces graphiques, à $\frac{1}{\sqrt{\lambda_\alpha}}$ près :

- ✓ les individus sont au centre des modalités qu'ils ont choisis,
- ✓ les modalités sont au centre des individus qui les ont choisis.

Règles d'interprétation :

Les individus et les modalités sont représentés sur ses plans de projection dont la lecture nécessite des règles d'interprétation.

Inertie expliquée

On a vu qu'en ACM, l'inertie totale du nuage des individus et du nuage des modalités vaut $\frac{m}{p}$ – 1 et ne dépend donc que du nombre moyen de modalités par variable. De plus l'inertie totale est égale à λ_1 + ... + λr où r = min (n − 1, m − p) est le nombre de valeurs propre non nulles. Le pourcentage d'inertie expliquée par un axe α est donc : $\frac{\lambda_\alpha}{\lambda_1+\cdots+\lambda_r} * 100$

NB: en ACM, les pourcentages d'inertie expliquée par les axes sont par construction "petits" et ne peuvent donc pas être interprétés comme en AFC ou en ACP. Le nombre d'axes retenus pour l'interprétation ou le recodage ne peut pas être choisi à partir de ces pourcentages d'inertie expliquée.

Contributions

On reprend les formules de l'AFC et on trouve qu'en ACM la contribution d'un individu h et la contribution d'une modalité k à l'inertie de l'axe α s'écrivent :

$$\mathrm{Ctr}_\alpha(h) = \frac{1}{n}\frac{x_{h\alpha}^2}{\lambda_\alpha}$$

$$\mathrm{Ctr}_\alpha(k) = \frac{n_k}{np}\frac{y_{k\alpha}^2}{\lambda_\alpha}$$

On en déduit qu'en pratique :

- Les individus les plus excentrés sur les plans factoriels sont ceux qui contribuent le plus.
- En revanche, les modalités les plus excentrées ne sont pas nécessairement celles qui contribuent le plus. En effet, leur contribution dépend de leur fréquence. En ACM, la contribution (absolue) d'une variable j à de l'axe α est définie comme la somme des contributions de ses modalités :

$$\mathrm{Ctr}_\alpha(j) = \sum_k \mathrm{Ctr}_\alpha(k) = \frac{1}{p} * \eta^2_{x^\alpha,j}$$

Où $\eta^2_{x^\alpha,j}$ est le rapport de corrélation entre la composante x^α (quantitative) et la variable j (qualitative). Le rapport de corrélation mesure la part de la variance de x^α expliquée par la variable qualitative j :

$$\eta^2_{x^\alpha,j} = \frac{\sum_k n_k(\mathbb{X}^\alpha_k - \mathbb{X}^\alpha)^2}{\sum_{i=1}^n (x_{i\alpha} - \mathbb{X}^\alpha)^2}$$

Où $\mathbb{X}^\alpha_k$ est la moyenne de x^α calculé avec les individus qui possèdent la modalité k.
En pratique :
La contribution d'une variable qualitative j à un axe α donne une idée de la liaison entre cette variable et la composante principale x^α
On utilise ces contributions pour représenter graphiquement les variables qualitatives de l'analyse sur un plan factoriel (α;α'). On représente en abscisse les contributions des variables à l'axe α et en ordonnée les contributions des variables à l'axe α'. On dessine alors des « flèches » comme dans le cercle des corrélations en ACP.

Cosinus carrés

En pratique, si deux individus sont bien projetés alors s'ils sont proches en projections, ils sont effectivement proches dans leur espace d'origine et on peut alors interpréter leur proximité :
La proximité entre deux individus s'interprète en termes de distance (du $\chi 2$) : deux individus se ressemblent s'ils ont choisis les mêmes modalités. C'est cohérent avec la relation barycentrique qui dit que les individus sont au barycentre des modalités qu'ils possèdent.
La proximité entre deux modalités de deux variables différentes s'interprète en termes de distance (du $\chi 2$) : deux modalités se ressemblent si elles sont possédées par les mêmes individus. C'est cohérent avec la relation barycentrique qui dit que les modalités sont au barycentre des individus qui les possèdent.

Partie Pratique

Contexte du projet

Problématique

Le but de notre étude est d'analyser les relations entre les habitudes et les méthodes d'étude et de travail chez les étudiants de l'INSEA, en plus de leur personnalité sur leur performance académique.

Pour cela, nous étudions un ensemble (fini) de variables catégorisées à valeurs dans un ensemble (fini) d'individus.

Collecte de données

On a opté pour la collecte de données en ligne, par le biais d'un questionnaire rédigé à l'aide de Google Forms. Il était composé de 21 questions en total, répartit en 3 sections: la première est liée aux informations personnelles, la 2ième aux méthodes d'étude et de révision, et la 3ième aux traits de personnalité.

Dans chaque section, il y avait des questions à choix multiples, et l'étudiant est invité à répondre à la question en choisissant une seule réponse.

L'ensemble des questions représente les variables et les choix de réponse représente leurs modalités de notre dataset. Les questions sont comme suit:

- Genre
  - Femme
  - Homme
- Filière
  - M2SI
  - DSE
  - DS
  - RO
  - AF
  - SE
  - SD
- Année
  - 1A
  - 2A
  - 3A
- Ecole avant l'INSEA
  - CPGE
  - FS
  - FST
  - EST

  - Autre
- Vous préférez de travailler en : *
  - Groupe
  - Individuel
- Qu'est ce qui décrit le plus votre style d'apprentissage ? *
  - Indépendant
  - Compétitif
  - Collaboratif
- À quel moment de la journée révisez-vous le mieux ? *
  - Matin
  - Après-midi
  - Soir
- Combien de Rattrapage avez-vous ?*
  - 1
  - 2
  - Plus de 3
- Quel est le format de cours que vous préférez *
  - 100% Présentiel
  - 100% à distance
  - Hybride
- Comment révisez-vous pendant la période des examens ? *
  - En groupe
  - Individuellement
- Avez-vous utilisé les plateformes certifiantes en ligne (coursera, Udemy,Openclassroom…) pendant votre formation
  - Oui
  - Non
- Vous aimez utiliser des outils d'organisation comme les horaires et les TO-Do listes ? *
  - Oui
  - Non
- Comment décrivez-vous votre personnalité ?
  - Introverti (Renfermé)
  - Extraverti (Ouvert)
- Vous passez une grande partie de votre temps libre à explorer divers sujets aléatoires qui piquent votre intérêt. *
  - Oui
  - Non
- Suiviez-vous des séries/Films *
  - Tous les jours
  - Assez souvent

    - Rarement
- Etes-vous membre actif d'un club
    - 1
    - 2
    - Plus de 3
- Avez-vous souffert récemment de maux de tête
    - Oui
    - Non
- Avez-vous souffert récemment d'un état dépressif
    - Oui
    - Non
- Avez-vous souffert récemment de mal au dos
    - Oui
    - Non
- Avez-vous souffert récemment de nervosité
    - Oui
    - Non
- Etes-vous satisfait de votre état de santé
    - Oui
    - Non

Importation des libraires nécessaires

Avant de commencer, on spécifie le répertoire de travail:

```
#setwd("C:/Users/HP/OneDrive - Institut National de Statistique et d'Economie Appliquee/2022/classes/S1_M2SI/R/acm")
```

Maintenant on installe les package nécessaires:

```
#install.packages(c("FactoMineR", "factoextra","Rcpp","tidyverse"))
#install.packages(c("Rcpp", "readr"))
library("FactoMineR")
library(factoextra)
library("corrplot")
library(readr)
library(kableExtra)
```

Importation des données

Après la collecte de données, on les a enregistrés comme un fichier .csv, maintenant on importe notre dataset:

```
formResponse <- read_csv("formResponse.csv")
```

```
#head(formResponse[, 1:21], 3)
kable(head(formResponse))%>%
  kable_styling(bootstrap_options = "striped", font_size = 10, full_width = F)
```

Horodateur	Genre	Filière	Année	Ecole avant INSEA	Vous préférez de travailler en :	Qu'est ce qui décrit le plus votre style d'apprentissage?	À quelle moment de la journée révisez-vous le mieux ?	combien de Rattrapage avez vous	Quel est le format de cours que vous préférer	Comment révisez-vous pendant la période des examens?	Avez-vous utilisé les plateformes certifiants en ligne (coursera, Udemy,Openclassroom...) pendant votre formation	Vous aimez utiliser des outils d'organisation comme les horaires et les TO-Do listes?	Comment décrivez-vous votre personnalité?	Vous passez une grande partie de votre temps libre à explorer divers sujets aléatoires qui piquent votre intérêt.	Suiviez vous des séries/Films	Etes-vous membre actif d'un club	Avez vous souffert récemment de maux de tête	Avez-vous souffert récemment d'un état dépressif	Avez vous souffert récemment de mal au dos	Avez vous souffert récemment de nervosité	Etes vous satisfait de votre état de santé
28/02/2022 19:16:52	Femme	M2SI	2A	FS	Groupe	Collaboratif	Soir	1	Hybride	En groupe	Oui	Oui	Extraverti (Ouvert)	Oui	Assez souvent	1	Non	Oui	Non	Oui	Oui
28/02/2022 19:24:19	Femme	DS	1A	CPGE	Groupe	Collaboratif	Matin	1	Hybride	Individuellement	Oui	Non	Extraverti (Ouvert)	Oui	Rarement	Plus de 3	Oui	Non	Oui	Non	Oui
28/02/2022 19:24:38	Homme	DSE	1A	CPGE	Groupe	Compétitif	Matin	1	100% Présentiel	En groupe	Oui	Non	Extraverti (Ouvert)	Non	Tous les jours	1	Non	Non	Non	Non	Oui
28/02/2022 19:24:53	Homme	AF	1A	CPGE	Individuel	Indépendant	Soir	1	Hybride	Individuellement	Non	Oui	Extraverti (Ouvert)	Oui	Tous les jours	1	Non	Oui	Non	Oui	Non
28/02/2022 19:25:10	Homme	SE	1A	CPGE	Groupe	Collaboratif	Matin	1	100% Présentiel	En groupe	Oui	Non	Extraverti (Ouvert)	Oui	Assez souvent	2	Non	Non	Non	Non	Oui
28/02/2022 19:25:26	Femme	SD	1A	CPGE	Groupe	Indépendant	Soir	1	Hybride	Individuellement	Non	Oui	Extraverti (Ouvert)	Oui	Rarement	1	Oui	Oui	Non	Non	Non

Figure 1: Importation de données brutes

Les données contiennent 94 lignes (individus) et 21 colonnes (variables) où chaque variable a des modalités spécifiques. Nous utiliserons tous les individus (donc 94) mais uniquement certaines variables pour effectuer l'ACM.

Préparation des données

Après la collecte des données suit la préparation des données. La préparation des données, parfois appelée « pré-traitement », est l'étape pendant laquelle les données brutes sont nettoyées et structurées en vue de l'étape suivante du traitement des données.

Pendant cette phase de préparation, les données brutes sont vérifiées avec soin afin de déceler d'éventuelles erreurs. L'objectif est d'éliminer les données de mauvaise qualité (redondantes, incomplètes ou incorrectes) et de commencer à créer les données de haute qualité qui peuvent garantir la qualité de l'étude.

On commence par supprimer la 1ière colonne contenant la date de reçu de la réponse:

```
# SUPPRIMER LA 1ERE COLONE 'Horodateur'
df1 <- formResponse[ -c(1) ]
```

Ensuite on va renommer les colonnes:

```
# SUPPRIMER LA 1ERE COLONE 'Horodateur'
colnames(df1) <- c('Genre','Filiere','Annee','EclAvINSEA','MethdTravail','StylAprtnsg','IdealMomntTrvail','NbrRatt','FormatCrs','MethodPrepaExam','OnlinePlatfrms','OutilOrgnst','Personnalite','ExplorerNvSujet','SuivideFilm','MmbrClub','MauxTete','Depression','MalAuDos','Nervosite','StfctEtatSante')

kable(head(df1))%>%
  kable_styling(bootstrap_options = "striped", font_size = 10, full_width = F)
```

On vérifie si on a des valeurs manquantes, puis on les affiche par variable:

```
# VALEURS MANQUANTES
sum(is.na(df1))

## [1] 41

apply(df1, 2, function(col)sum(is.na(col)))

##         Genre       Filiere         Annee     EclAvINSEA
##             0             0             0             0
##   MethdTravail   StylAprtnsg IdealMomntTrvail       NbrRatt
##             0             0             0             0
##     FormatCrs MethodPrepaExam  OnlinePlatfrms    OutilOrgnst
##             0             0             4             0
##   Personnalite ExplorerNvSujet    SuivideFilm       MmbrClub
##             0             0             0            22
##      MauxTete     Depression       MalAuDos       Nervosite
##             4             3             2             3
##  StfctEtatSante
##             3
```

On remplace les NA de la variable 'MmbrClub' représentant le nombre des clubs dont l'étudiant est inscrit, par 0 avant de traiter les autres valeurs manquantes :

```
# REMPLACER LES NA DE LA VARIABLE MmbrClub PAR 0
df1$MmbrClub[is.na(df1$MmbrClub)] <- 0
```

Après on supprime les lignes avec des NA

```
#df[rowSums(is.na(df)) == 0, ]
df <- na.omit(df1)
apply(df, 2, function(col)sum(is.na(col)))

##         Genre       Filiere         Annee     EclAvINSEA
##             0             0             0             0
##   MethdTravail   StylAprtnsg IdealMomntTrvail       NbrRatt
##             0             0             0             0
##     FormatCrs MethodPrepaExam  OnlinePlatfrms    OutilOrgnst
##             0             0             0             0
##   Personnalite ExplorerNvSujet    SuivideFilm       MmbrClub
##             0             0             0             0
##      MauxTete     Depression       MalAuDos       Nervosite
##             0             0             0             0
##  StfctEtatSante
##             0
```

On fixe le problème des accents et des caractères spéciaux:

```
# FIXER LE PROBLEME DES ACCENTS ET CARACTERES SPECIALS
df1$StylAprtnsg <- iconv(df1$StylAprtnsg, from = 'UTF-8', to = 'ASCII//TRANSLIT')
df1$FormatCrs <- iconv(df1$FormatCrs, from = 'UTF-8', to = 'ASCII//TRANSLIT')
```

Notre dataset maintenant est prête pour l'analyse. La figure suivante présente un aperçu des données :

Genre	Filiere	Annee	EclAvINSEA	MethdTravail	StylAprtnsg	IdealMomntTrvail	NbrRatt	FormatCrs	MethodPrepaExam	OnlinePlatfrms	OutilOrgnst	Personnalite	ExplorerNvSujet	SuivideFilm	MmbrClub	MauxTete	Depression	MalAuDos	Nervosite	StfctEtatSante
Femme	M2SI	2A	FS	Groupe	Collaboratif	Soir	1	Hybride	En groupe	Oui	Oui	Extraverti (Ouvert)	Oui	Assez souvent	1	Non	Oui	Non	Oui	Oui
Femme	DS	1A	CPGE	Groupe	Collaboratif	Matin	1	Hybride	Individuellement	Oui	Non	Extraverti (Ouvert)	Oui	Rarement	Plus de 3	Oui	Non	Oui	Non	Oui
Homme	DSE	1A	CPGE	Groupe	Compétitif	Matin	1	100% Présentiel	En groupe	Oui	Non	Extraverti (Ouvert)	Non	Tous les jours	1	Non	Non	Non	Non	Oui
Homme	AF	1A	CPGE	Individuel	Indépendant	Soir	1	Hybride	Individuellement	Non	Oui	Extraverti (Ouvert)	Oui	Tous les jours	1	Non	Oui	Non	Oui	Non
Homme	SE	1A	CPGE	Groupe	Collaboratif	Matin	1	100% Présentiel	En groupe	Oui	Non	Extraverti (Ouvert)	Oui	Assez souvent	2	Non	Non	Non	Non	Oui
Femme	SD	1A	CPGE	Groupe	Indépendant	Soir	1	Hybride	Individuellement	Non	Oui	Extraverti (Ouvert)	Oui	Rarement	1	Oui	Oui	Non	Non	Non

Genre	Filiere	Annee	EclAvINSEA	MethdTravail	StylAprtnsg	IdealMomntTrvail	NbrRatt	FormatCrs	MethodPrepaExam	OnlinePlatfrms
Femme	M2SI	2A	FS	Groupe	Collaboratif	Soir	1	Hybride	En groupe	Oui
Femme	DS	1A	CPGE	Groupe	Collaboratif	Matin	1	Hybride	Individuellement	Oui
Homme	DSE	1A	CPGE	Groupe	Compétitif	Matin	1	100% Présentiel	En groupe	Oui
Homme	AF	1A	CPGE	Individuel	Indépendant	Soir	1	Hybride	Individuellement	Non
Homme	SE	1A	CPGE	Groupe	Collaboratif	Matin	1	100% Présentiel	En groupe	Oui
Femme	SD	1A	CPGE	Groupe	Indépendant	Soir	1	Hybride	Individuellement	Non

OutilOrgnst	Personnalite	ExplorerNvSujet	SuivideFilm	MmbrClub	MauxTete	Depression	MalAuDos	Nervosite	StfctEtatSante
Oui	Extraverti (Ouvert)	Oui	Assez souvent	1	Non	Oui	Non	Oui	Oui
Non	Extraverti (Ouvert)	Oui	Rarement	Plus de 3	Oui	Non	Oui	Non	Oui
Non	Extraverti (Ouvert)	Non	Tous les jours	1	Non	Non	Non	Non	Oui
Oui	Extraverti (Ouvert)	Oui	Tous les jours	1	Non	Oui	Non	Oui	Non
Non	Extraverti (Ouvert)	Oui	Assez souvent	2	Non	Non	Non	Non	Oui
Oui	Extraverti (Ouvert)	Oui	Rarement	1	Oui	Oui	Non	Non	Non

Figure 2: Aperçu du Dataset nettoyée

Analyse de données

Pendant cette étape, les données importées et nettoyées lors de l'étape précédente seront traitées pour interprétation.

Exploration des données

```
# Nombre de ligne
nrow(df)

## [1] 87

# Nombre de colones
ncol(df)

## [1] 21

# Afficher les 4 premières lignes du dataset
head(df[, 1:21], 4)

## # A tibble: 4 x 21
##   Genre Filiere Annee EclAvINSEA MethdTravail StylAprtnsg  IdealMomntTrvail
##   <chr> <chr>   <chr> <chr>     <chr>       <chr>       <chr>
## 1 Femme M2SI    2A    FS        Groupe      Collaboratif Soir
## 2 Femme DS      1A    CPGE      Groupe      Collaboratif Matin
## 3 Homme DSE     1A    CPGE      Groupe      Compétitif   Matin
## 4 Homme AF      1A    CPGE      Individuel  Indépendant  Soir
## # ... with 14 more variables: NbrRatt <chr>, FormatCrs <chr>,
## #   MethodPrepaExam <chr>, OnlinePlatfrms <chr>, OutilOrgnst <chr>,
## #   Personnalite <chr>, ExplorerNvSujet <chr>, SuivideFilm <chr>,
## #   MmbrClub <chr>, MauxTete <chr>, Depression <chr>, MalAuDos <chr>,
## #   Nervosite <chr>, StfctEtatSante <chr>
```

Pour notre étude, nous utiliserons tous les individus mais uniquement certaines variables pour effectuer l'ACM. Les coordonnées des variables restantes seront prédites.

Variables et Individus actifs

Nos données contiennent donc des:

- **Individus actifs** (lignes 1:87): individus qui sont utilisés dans l'ACM.
- **Variables actives** (colonnes (5:8, 10:21)): 16 variables en total vont être utilisées dans l'ACM.
- **Variables supplémentaires**: elles ne participent pas à l'ACM. Les coordonnées de ces variables seront prédites.
  - Variables qualitatives supplémentaires (quali.sup: Colonnes [1:4, 9:9] correspondant aux colonnes 'Genre','Filiere','Annee','EclAvINSEA','FormatCrs' respectivement. Ces variables seront utilisées pour colorer les individus par groupes.

Nous commençons par extraire les individus actifs et les variables actives pour l'ACM:

```
df.active <- df[, colnames(df)[c(5:8, 10:21)]]
head(df.active[, 1:16], 4)
```

```
## # A tibble: 4 x 16
##   MethdTravail StylAprtnsg  IdealMomntTrvail NbrRatt MethodPrepaExam
##   <chr>       <chr>        <chr>            <chr>   <chr>
## 1 Groupe      Collaboratif Soir             1       En groupe
## 2 Groupe      Collaboratif Matin            1       Individuellement
## 3 Groupe      Compétitif   Matin            1       En groupe
## 4 Individuel  Indépendant  Soir             1       Individuellement
## # ... with 11 more variables: OnlinePlatfrms <chr>, OutilOrgnst <chr>,
## #   Personnalite <chr>, ExplorerNvSujet <chr>, SuivideFilm <chr>,
## #   MmbrClub <chr>, MauxTete <chr>, Depression <chr>, MalAuDos <chr>,
## #   Nervosite <chr>, StfctEtatSante <chr>
```

On applique par suite la fonction *summary()*. A ce niveau, la fonction retourne la taille des variables (nombre d'individus par variables) et intuitivement sera 87 nombre total des lignes après suppression

```
summary(df.active)[, 1:16]
```

```
##  MethdTravail      StylAprtnsg       IdealMomntTrvail     NbrRatt
##  Length:87         Length:87         Length:87         Length:87
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##  MethodPrepaExam   OnlinePlatfrms    OutilOrgnst       Personnalite
##  Length:87         Length:87         Length:87         Length:87
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##  ExplorerNvSujet   SuivideFilm       MmbrClub          MauxTete
##  Length:87         Length:87         Length:87         Length:87
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##   Depression        MalAuDos          Nervosite         StfctEtatSante
##  Length:87         Length:87         Length:87         Length:87
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
```

Graphes de fréquence des modalités de variables

le but de cette partie est de visualiser la fréquence des catégories de chaque variable.

La 1ière étape sera de transformer les variables (initialement de *type char*) en *facteur*

```
str(df.active)
```

```
## tibble[,16] [87 x 16] (S3: tbl_df/tbl/data.frame)
##  $ MethdTravail   : chr [1:87] "Groupe" "Groupe" "Groupe" "Individuel" ...
##  $ StylAprtnsg    : chr [1:87] "Collaboratif" "Collaboratif" "Compétitif" "Indépendant" ...
##  $ IdealMomntTrvail: chr [1:87] "Soir" "Matin" "Matin" "Soir" ...
##  $ NbrRatt        : chr [1:87] "1" "1" "1" "1" ...
##  $ MethodPrepaExam : chr [1:87] "En groupe" "Individuellement" "En groupe" "Individuellement
```

```
" ...
##  $ OnlinePlatfrms  : chr [1:87] "Oui" "Oui" "Oui" "Non" ...
##  $ OutilOrgnst     : chr [1:87] "Oui" "Non" "Non" "Oui" ...
##  $ Personnalite    : chr [1:87] "Extraverti (Ouvert)" "Extraverti (Ouvert)" "Extraverti (Ouvert)" "E
xtraverti (Ouvert)" ...
##  $ ExplorerNvSujet : chr [1:87] "Oui" "Oui" "Non" "Oui" ...
##  $ SuivideFilm     : chr [1:87] "Assez souvent" "Rarement" "Tous les jours" "Tous les jours" ...
##  $ MmbrClub        : chr [1:87] "1" "Plus de 3" "1" "1" ...
##  $ MauxTete        : chr [1:87] "Non" "Oui" "Non" "Non" ...
##  $ Depression      : chr [1:87] "Oui" "Non" "Non" "Oui" ...
##  $ MalAuDos        : chr [1:87] "Non" "Oui" "Non" "Non" ...
##  $ Nervosite       : chr [1:87] "Oui" "Non" "Non" "Oui" ...
##  $ StfctEtatSante  : chr [1:87] "Oui" "Oui" "Oui" "Non" ...
##  - attr(*, "na.action")= 'omit' Named int [1:7] 13 26 35 70 82 84 93
##   ..- attr(*, "names")= chr [1:7] "13" "26" "35" "70" ...
```

En R, un facteur (factor, en anglais) est un vecteur dont les éléments ne peuvent prendre que des modalités prédéfinies. Ce qui caractérise un facteur en R est le fait qu'elle dispose de l'attribut ***Levels*** *(niveaux). En pratique, un facteur est typiquement utilisé pour stocker les valeurs observées d'une variable catégorielle (couleur, sexe, jours de la semaine, religion, …)*

```
df.activeFctr<-as.data.frame(lapply(df.active, as.factor))
str(df.activeFctr)

## 'data.frame':    87 obs. of  16 variables:
##  $ MethdTravail    : Factor w/ 2 levels "Groupe","Individuel": 1 1 1 2 1 1 2 2 1 1 ...
##  $ StylAprtnsg     : Factor w/ 3 levels "Collaboratif",..: 1 1 2 3 1 3 2 3 2 1 ...
##  $ IdealMomntTrvail: Factor w/ 3 levels "Après-midi","Matin",..: 3 2 2 3 2 3 3 2 2 3 ...
##  $ NbrRatt         : Factor w/ 3 levels "1","2","plus de 3": 1 1 1 1 1 1 2 1 3 1 ...
##  $ MethodPrepaExam : Factor w/ 2 levels "En groupe","Individuellement": 1 2 1 2 1 2 2 2 2 1 ...
##  $ OnlinePlatfrms  : Factor w/ 2 levels "Non","Oui": 2 2 2 1 2 1 1 2 2 2 ...
##  $ OutilOrgnst     : Factor w/ 2 levels "Non","Oui": 2 1 1 2 1 2 1 2 2 2 ...
##  $ Personnalite    : Factor w/ 2 levels "Extraverti (Ouvert)",..: 1 1 1 1 1 1 1 1 2 1 ...
##  $ ExplorerNvSujet : Factor w/ 2 levels "Non","Oui": 2 2 1 2 2 2 2 2 2 2 ...
##  $ SuivideFilm     : Factor w/ 3 levels "Assez souvent",..: 1 2 3 3 1 2 1 1 1 1 ...
##  $ MmbrClub        : Factor w/ 4 levels "0","1","2","Plus de 3": 2 4 2 2 3 2 3 2 3 3 ...
##  $ MauxTete        : Factor w/ 2 levels "Non","Oui": 1 2 1 1 1 2 2 2 2 1 ...
##  $ Depression      : Factor w/ 2 levels "Non","Oui": 2 1 1 2 1 2 2 2 2 1 ...
##  $ MalAuDos        : Factor w/ 2 levels "Non","Oui": 1 2 1 1 1 1 2 2 2 1 ...
##  $ Nervosite       : Factor w/ 2 levels "Non","Oui": 2 1 1 2 1 1 2 2 2 1 ...
##  $ StfctEtatSante  : Factor w/ 2 levels "Non","Oui": 2 2 2 1 2 1 1 2 1 2 ...
```

On applique de nouveau la fonction *summary()* qui va renvoyer maintenant la taille des catégories de chaque variable.

```
summary(df.activeFctr)

##     MethdTravail      StylAprtnsg   IdealMomntTrvail     NbrRatt
##  Groupe   :50   Collaboratif:38   Après-midi: 6     1       :45
##  Individuel:37   Compétitif  :17   Matin     :28     2       :20
##                  Indépendant :32   Soir      :53     plus de 3:22
```

```
##
##        MethodPrepaExam OnlinePlatfrms OutilOrgnst             Personnalite
##  En groupe      :30    Non:31        Non:52     Extraverti (Ouvert)  :51
##  Individuellement:57    Oui:56        Oui:35     Introverti (Renfermé):36
##
##
##  ExplorerNvSujet       SuivideFilm     MmbrClub  MauxTete Depression
##  Non:33         Assez souvent :52   0       :18   Non:29   Non:37
##  Oui:54         Rarement      :22   1       :38   Oui:58   Oui:50
##              Tous les jours:13   2       :21
##                             Plus de 3:10
##  MalAuDos Nervosite StfctEtatSante
##  Non:28   Non:39    Non:39
##  Oui:59   Oui:48    Oui:48
##
##
```

Graphes

La fonction ci-dessous, permet d'afficher les graphes des 16 variables à la fois:

```
for (i in 1:16) {
  plot(df.activeFctr[,i], main = colnames(df.activeFctr)[i],
      ylab = "Count", col="springgreen4", las = 1, ylim=c(0,60))
}
```

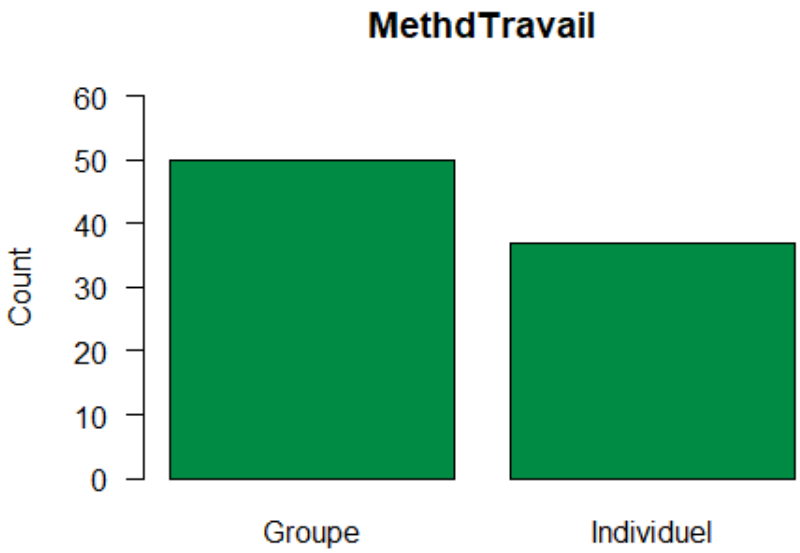


Figure 3: Graphe de la variable 'Methode de Travail'

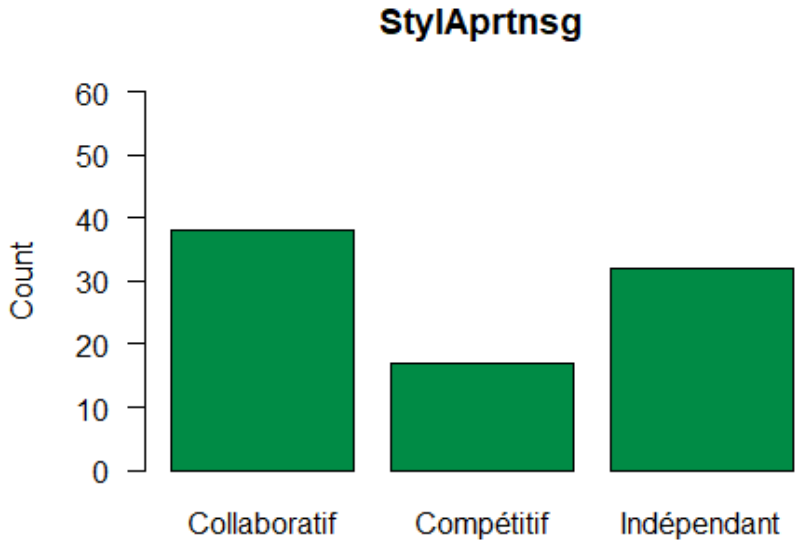


Figure 4: Graphe de la variable 'Style d'apprentissage'

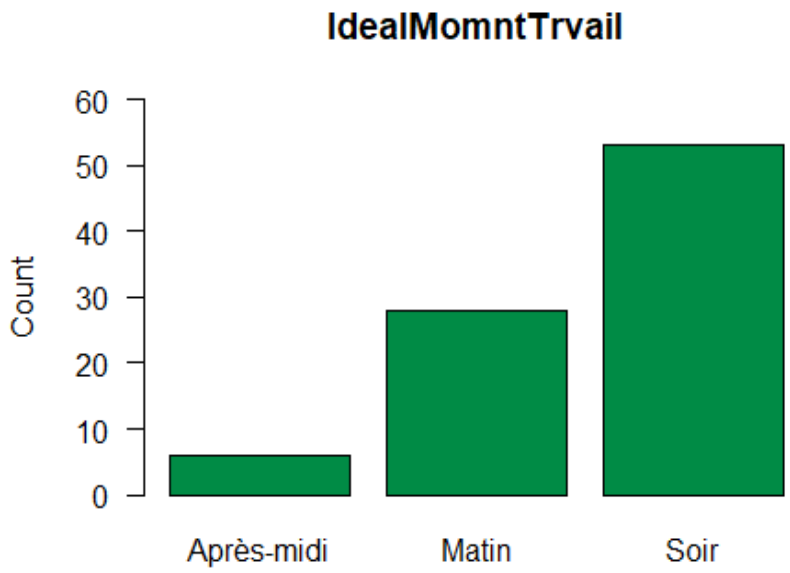


Figure 5: Graphe de la variable 'Ideal moment de travail'

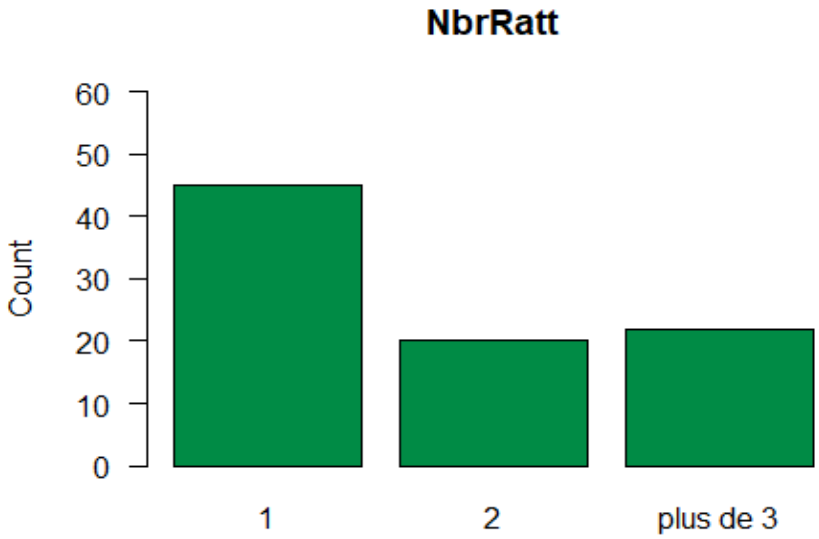


Figure 6: Graphe de la variable 'Nombre de rattrapages'

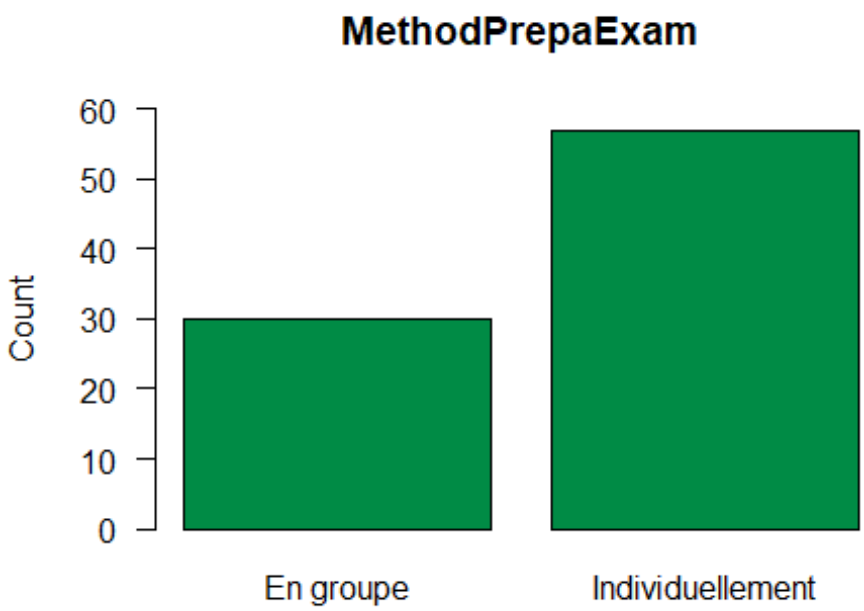


Figure 7: Graphe de la variable 'Methode de préparation d'examen'

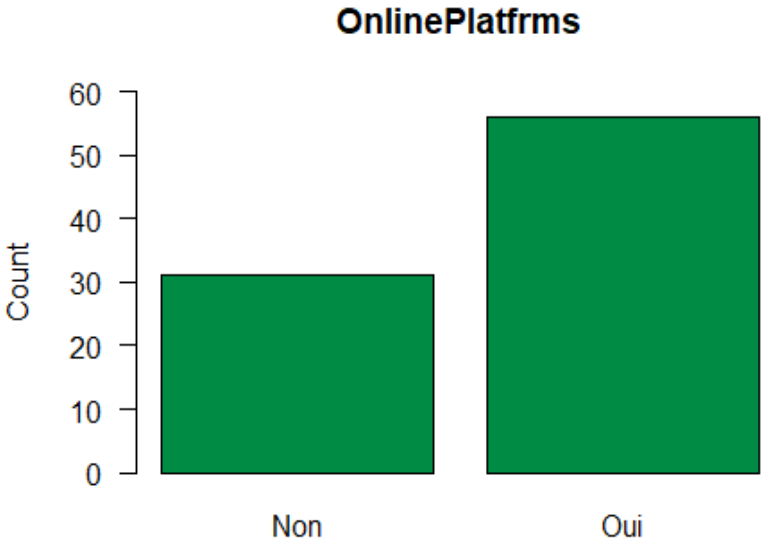


Figure 8: Graphe de la variable 'Utilisation des Plateforme Certifiantes Online'

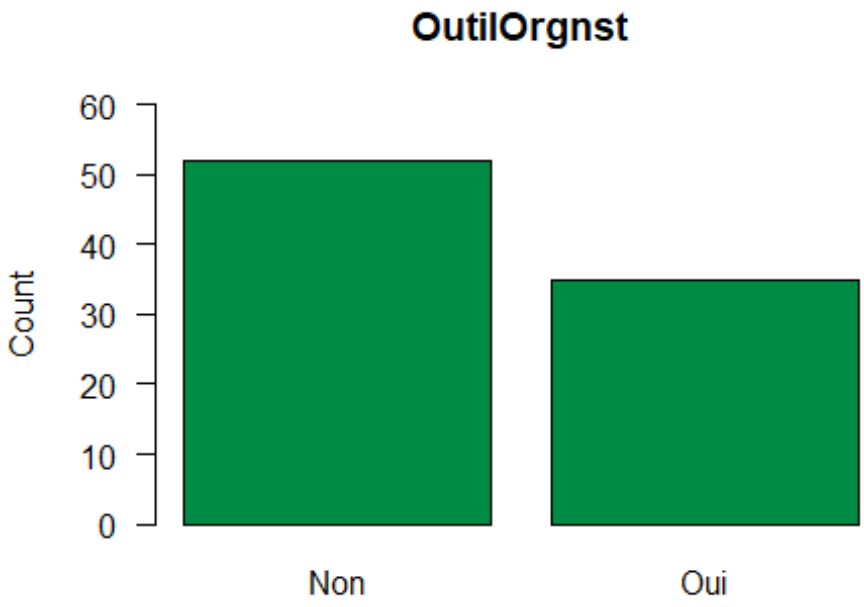


Figure 9: Graphe de la variable 'Outils d'organisation'

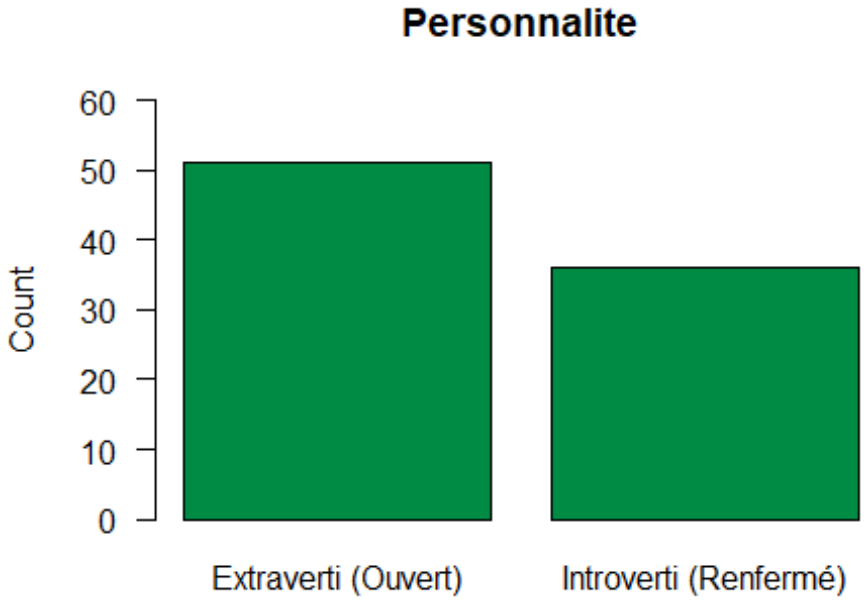


Figure 10: Graphe de la variable 'Personnalité'

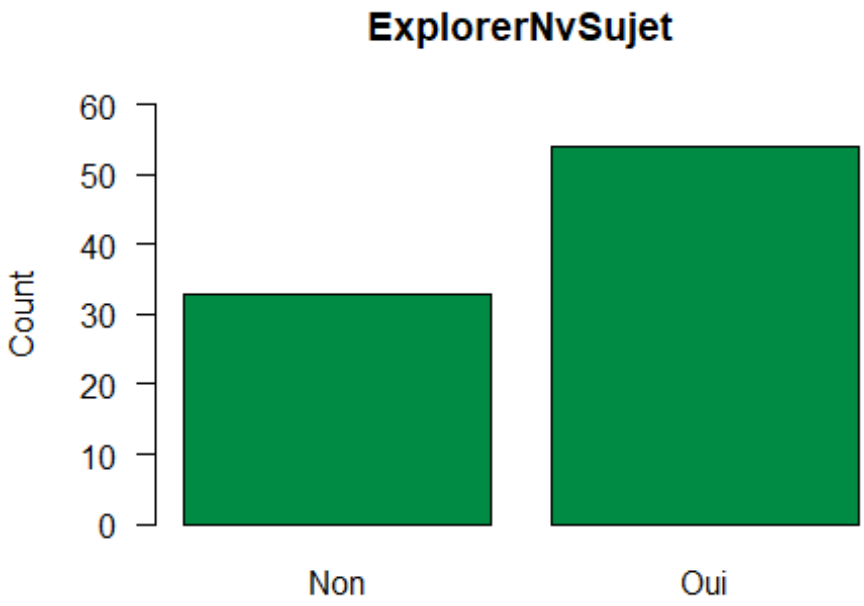


Figure 11: Graphe de la variable 'Exploration de nouveaux sujets'

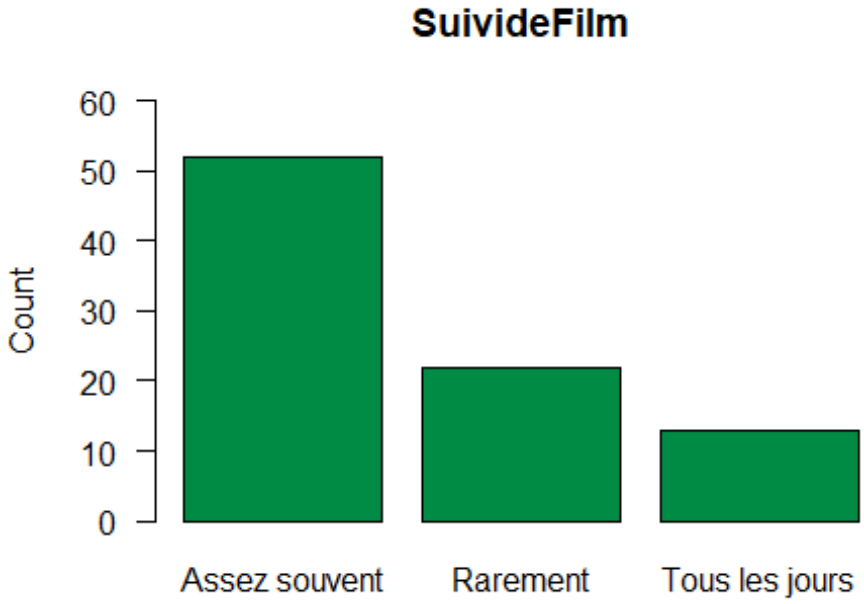


Figure 12: Graphe de la variable 'Suivi de films'

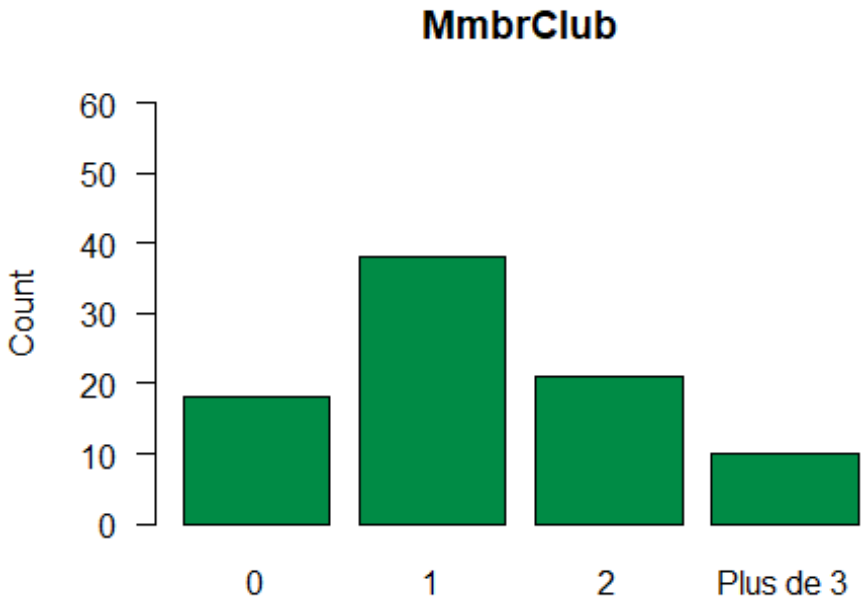


Figure 13: Graphe de la variable 'Membre de clubs'

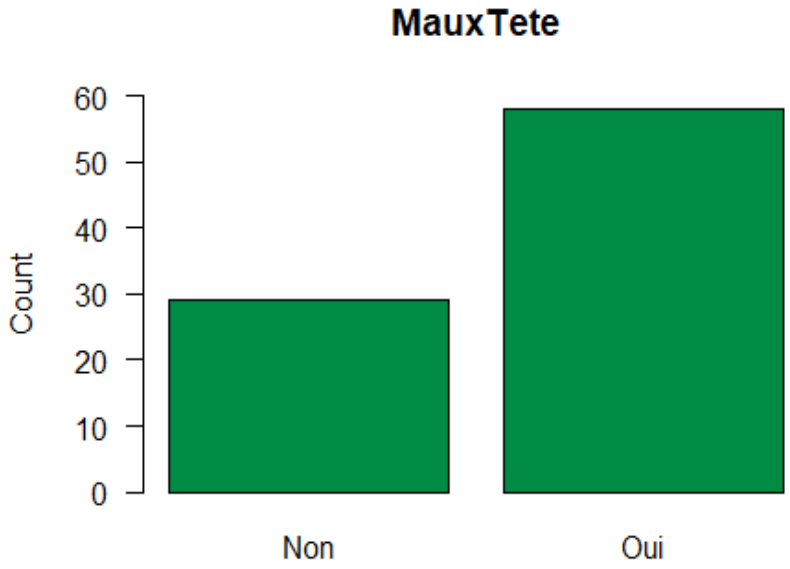


Figure 14: Graphe de la variable 'Maux de tête'

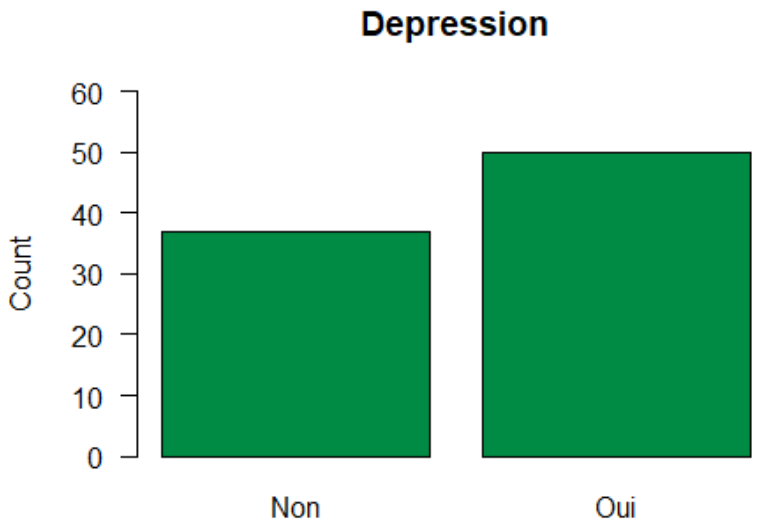


Figure 15: Graphe de la variable 'Depression'

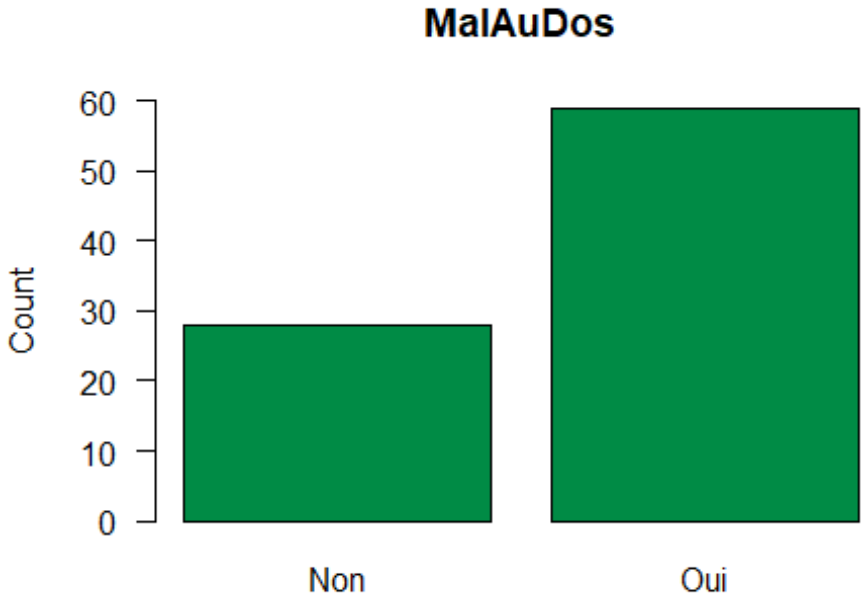


Figure 16: Graphe de la variable 'Mal au dos'

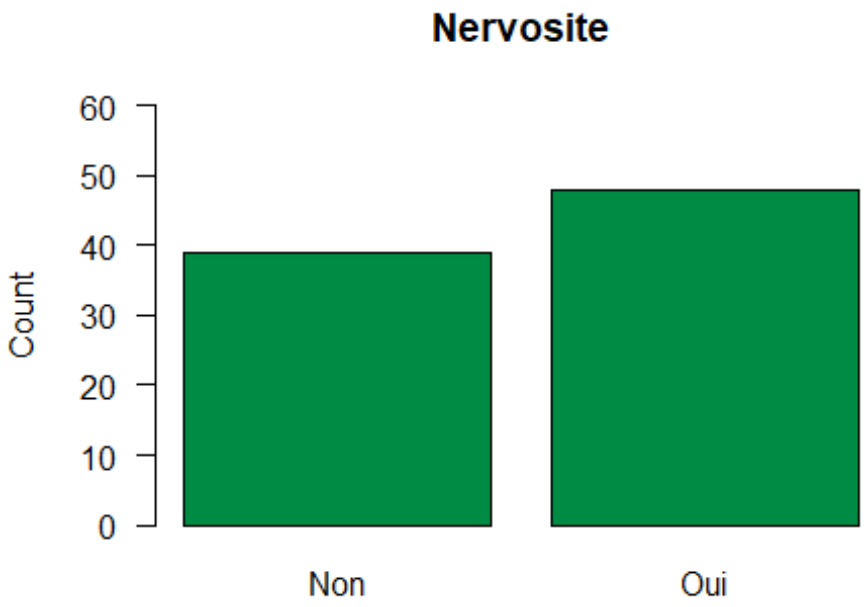


Figure 17: Graphe de la variable 'Nervosite'

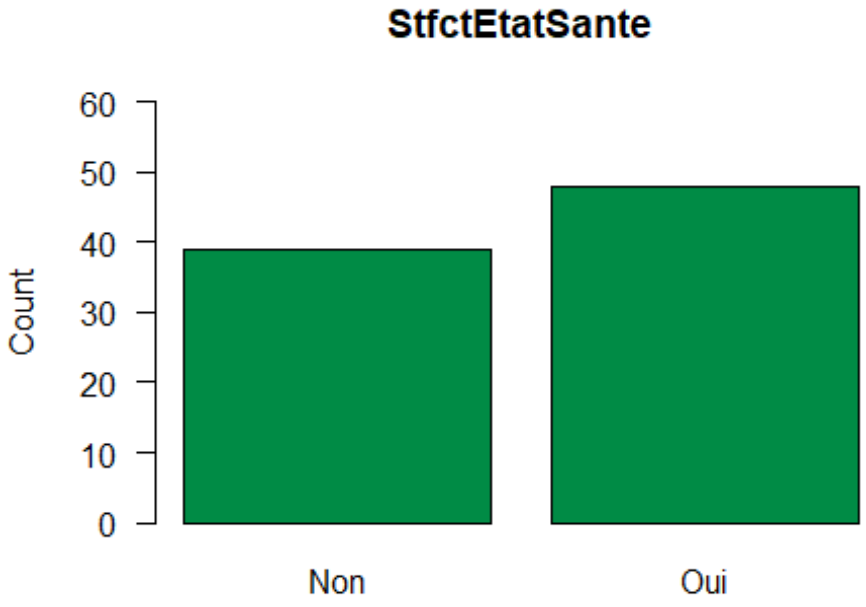


Figure 18: Graphe de la variable 'Satisfaction d'état de santé'

Les graphiques ci-dessus peuvent être utilisés pour identifier les catégories à très faible fréquence. Ce type de variables peut fausser l'analyse et doit être supprimé.

Par exemple, la modalité *"Après-midi"* de la variable **IdealMommtTravail** a une faible fréquence par rapport au 2 autres modalité (Matin et soir), donc on va l'éliminer:

```
df_momtrv <- df[ df$IdealMomntTrvail != "Après-midi", , drop=FALSE];
df_momtrv$IdealMomntTrvail <- factor(df_momtrv$IdealMomntTrvail)
str(df_momtrv$IdealMomntTrvail)

##  Factor w/ 2 levels "Matin","Soir": 2 1 1 2 1 2 2 1 1 2 ...
```

Implémentation de l'ACM

La fonction *MCA()* [FactoMineR] est utilisée.

```
res.ACM <- MCA (df_momtrv,quali.sup =c(1:4, 9:9), graph = FALSE)
print(res.ACM)

## **Results of the Multiple Correspondence Analysis (MCA)**
## The analysis was performed on 81 individuals, described by 21 variables
## *The results are available in the following objects:
##
##   name              description
## 1  "$eig"           "eigenvalues"
## 2  "$var"           "results for the variables"
## 3  "$var$coord"     "coord. of the categories"
## 4  "$var$cos2"      "cos2 for the categories"
## 5  "$var$contrib"   "contributions of the categories"
## 6  "$var$v.test"    "v-test for the categories"
## 7  "$ind"           "results for the individuals"
## 8  "$ind$coord"     "coord. for the individuals"
## 9  "$ind$cos2"      "cos2 for the individuals"
## 10 "$ind$contrib"   "contributions of the individuals"
## 11 "$quali.sup"     "results for the supplementary categorical variables"
## 12 "$quali.sup$coord"  "coord. for the supplementary categories"
## 13 "$quali.sup$cos2"   "cos2 for the supplementary categories"
## 14 "$quali.sup$v.test" "v-test for the supplementary categories"
## 15 "$call"          "intermediate results"
## 16 "$call$marge.col"   "weights of columns"
## 17 "$call$marge.li"    "weights of rows"
```

On peut aussi générer la table la table disjonctif des variables actives. Dans le tableau disjonctif, les lignes sont les individus et les colonnes sont les modalités des variables qualitatives. Comme le tableau de données contient un grand nombre de variables, nous ne pouvons pas afficher la table complète.

```
#ACM.disjonctif <- tab.disjonctif(df_momtrv)
```

On affiche le résumé du résultat de l'ACM

```
summary.MCA(res.ACM,ncp = 3, nbelements = 5)
```

```
##
## Call:
## MCA(X = df_momtrv, quali.sup = c(1:4, 9:9), graph = FALSE)
##
##
## Eigenvalues
##                  Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance           0.164   0.159   0.112   0.097   0.088   0.081   0.079
## % of var.         12.514  12.086   8.501   7.384   6.716   6.164   6.020
## Cumulative % of var.  12.514  24.600  33.101  40.485  47.202  53.365  59.385
##                  Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance           0.071   0.064   0.053   0.047   0.045   0.043   0.041
## % of var.          5.422   4.875   4.042   3.577   3.440   3.311   3.152
## Cumulative % of var.  64.808  69.683  73.724  77.302  80.742  84.053  87.205
##                 Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20  Dim.21
## Variance           0.037   0.030   0.026   0.025   0.020   0.018   0.012
## % of var.          2.793   2.319   1.978   1.897   1.540   1.341   0.926
## Cumulative % of var.  89.998  ù%  94.296  96.192  97.732  99.074 100.000
##
## Individuals (the 5 first)
##            Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 1          |  0.208  0.324  0.040 |  0.358  0.998  0.120 | -0.444  2.185
## 2          |  0.422  1.342  0.127 |  0.198  0.306  0.028 | -0.339  1.272
## 3          |  0.795  4.750  0.360 |  0.083  0.053  0.004 | -0.255  0.722
## 4          | -0.233  0.407  0.036 | -0.359  1.005  0.086 | -0.168  0.314
## 5          |  0.708  3.764  0.389 |  0.160  0.200  0.020 | -0.213  0.503
##            cos2
## 1          0.185 |
## 2          0.082 |
## 3          0.037 |
## 4          0.019 |
## 5          0.035 |
##
## Categories (the 5 first)
##              Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test
## Groupe       |  0.052  0.061  0.004  0.551 |  0.650  9.669  0.585  6.839 |
## Individuel   | -0.072  0.084  0.004 -0.551 | -0.899 13.365  0.585 -6.839 |
## Collaboratif |  0.062  0.061  0.003  0.470 |  0.951 14.963  0.654  7.236 |
## Compétitif   |  0.231  0.426  0.014  1.064 | -0.435  1.565  0.050 -2.005 |
## Indépendant  | -0.201  0.568  0.024 -1.378 | -0.831 10.088  0.407 -5.704 |
##              Dim.3    ctr   cos2 v.test
## Groupe      -0.004  0.001  0.000 -0.045 |
## Individuel   0.006  0.001  0.000  0.045 |
## Collaboratif -0.149  0.521  0.016 -1.132 |
## Compétitif   0.177  0.369  0.008  0.816 |
## Indépendant  0.068  0.097  0.003  0.469 |
##
## Categorical variables (eta2)
##                 Dim.1 Dim.2 Dim.3
## MethdTravail    | 0.004 0.585 0.000 |
```

```
## StylAprtnsg     | 0.028 0.676 0.018 |
## IdealMomntTrvail | 0.051 0.123 0.030 |
## NbrRatt         | 0.092 0.002 0.275 |
## MethodPrepaExam  | 0.031 0.519 0.077 |
##
## Supplementary categories (the 5 first)
##              Dim.1   cos2 v.test    Dim.2   cos2 v.test    Dim.3   cos2
## Femme       | -0.330  0.117 -3.059 |  0.181  0.035  1.678 | -0.164  0.029
## Homme       |  0.355  0.117  3.059 | -0.195  0.035 -1.678 |  0.176  0.029
## AF          | -0.285  0.027 -1.459 | -0.070  0.002 -0.361 |  0.559  0.102
## DS          |  0.244  0.011  0.955 | -0.283  0.015 -1.107 | -0.226  0.010
## DSE         | -0.591  0.049 -1.985 |  0.010  0.000  0.035 | -0.686  0.066
##            v.test
## Femme       -1.520 |
## Homme        1.520 |
## AF           2.861 |
## DS          -0.883 |
## DSE          -2.301 |
##
## Supplementary categorical variables (eta2)
##              Dim.1 Dim.2 Dim.3
## Genre       | 0.117 0.035 0.029 |
## Filiere     | 0.144 0.087 0.230 |
## Annee       | 0.012 0.059 0.057 |
## EclAvINSEA   | 0.158 0.066 0.007 |
## FormatCrs    | 0.080 0.001 0.034 |
```

```
# Description des dimensions
dimdesc(res.ACM,axes=1:3)
```

L'objet créé avec la fonction MCA() contient de nombreuses informations trouvées dans de nombreuses listes et matrices différentes. Ces valeurs seront décrites dans les sections suivantes.

Visualisation et interprétation

Nous utiliserons le package R **factoextra** pour aider à l'interprétation et à la visualisation de l'analyse des correspondances multiples.

Ces fonctions de factoextra incluent:

- **get_eigenvalue(res.mca)**: Extraction des valeurs propres / variances des composantes principales
- **fviz_eig(res.mca)**: Visualisation des valeurs propres

- **get_mca_ind(res.mca)**, **get_mca_var(res.mca)**: Extraction des résultats pour les individus et les variables, respectivement.
- **fviz_mca_ind(res.mca)**, **fviz_mca_var(res.mca)**: visualisation des résultats des individus et des variables, respectivement.
- **fviz_mca_biplot(res.mca)**: Création d'un biplot des individus et des variables.

Dans les sections suivantes, nous allons illustrer chacune de ces fonctions.

Valeurs propres et variance

La proportion des variances retenues par les différentes dimensions (axes) peut être extraite à l'aide de la fonction **get_eigenvalue()** [factoextra package] comme suit:

```
eig.val <- get_eigenvalue (res.ACM)
eig.val

##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.16425137      12.5143898                  12.51439
## Dim.2  0.15862859      12.0859880                  24.60038
## Dim.3  0.11157571       8.5010068                  33.10138
## Dim.4  0.09691324       7.3838658                  40.48525
## Dim.5  0.08815203       6.7163453                  47.20160
## Dim.6  0.08089599       6.1635037                  53.36510
## Dim.7  0.07901669       6.0203189                  59.38542
## Dim.8  0.07117016       5.4224884                  64.80791
## Dim.9  0.06398196       4.8748159                  69.68272
## Dim.10 0.05304750       4.0417141                  73.72444
## Dim.11 0.04695398       3.5774461                  77.30188
## Dim.12 0.04515109       3.4400834                  80.74197
## Dim.13 0.04345841       3.3111173                  84.05308
## Dim.14 0.04137394       3.1522999                  87.20538
## Dim.15 0.03665719       2.7929291                  89.99831
## Dim.16 0.03044125       2.3193334                  92.31765
## Dim.17 0.02595984       1.9778929                  94.29554
## Dim.18 0.02489586       1.8968277                  96.19237
## Dim.19 0.02021105       1.5398899                  97.73226
## Dim.20 0.01760421       1.3412729                  99.07353
## Dim.21 0.01215993       0.9264708                 100.00000
```

Les valeurs propres (eigenvalues en anglais) mesurent la quantité de variance expliquée par chaque axe principal. Les valeurs propres sont grandes pour les premiers axes et petits pour les axes suivants. Autrement dit, les premiers axes correspondent aux directions portant la quantité maximale de variation contenue dans le jeu de données.

Pour visualiser les pourcentages de variances expliquées par chaque dimension de l'ACM, on uyilise la fonction **fviz_eig()** ou **fviz_screeplot()** [package factoextra]:

```
fviz_screeplot (res.ACM,addlabels = TRUE, ylim = c (0, 13),barfill="springgreen4",linecolor ="red")
```

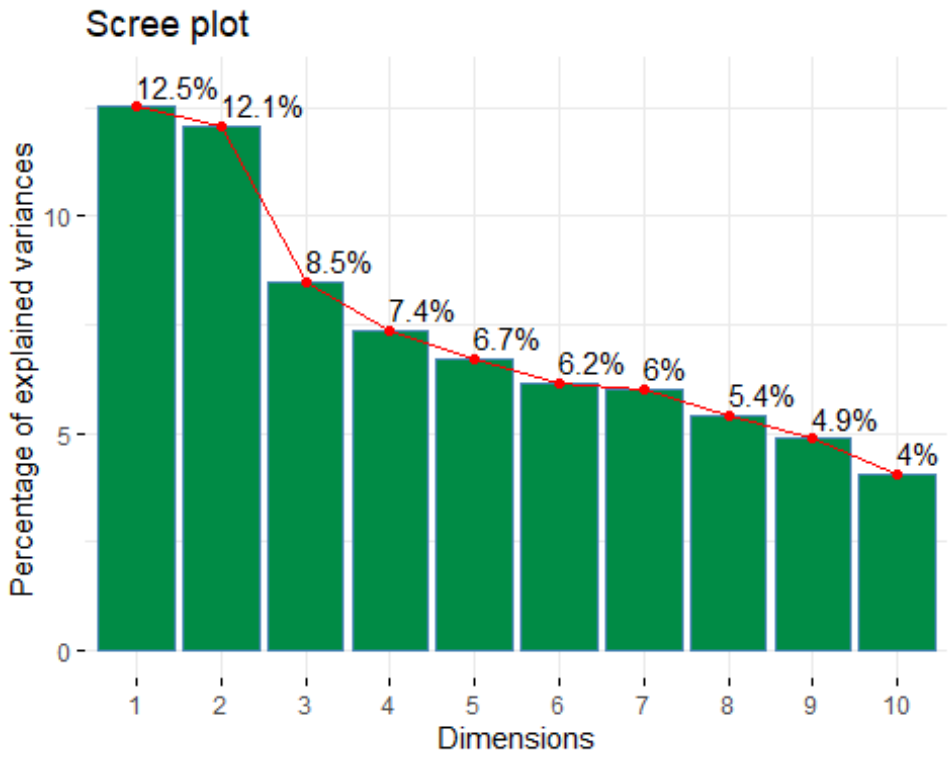


Figure 19: les pourcentages de variances expliquées par chaque dimension de l'ACM

```
#Ou bien avec
#barplot(eig.val[, 2],
#       names.arg = 1:nrow(eig.val),
#       main = "Pourcentage d'inertie",
#       xlab = "Dimensions Principales",
#       ylab = "Pourcentage des variances",
#       col ="steelblue")
#Pour ajouter la ligne
#lines(x = 1:nrow(eig.val), eig.val[, 2],
#     type = "b", pch = 19, col = "red")
```

Biplot des variables et des individus

Dans le graphique ci-dessous, les lignes (individus) sont représentées par des points bleu, les catégories des variables actives sont en rouge et les catégories des variables illustratives sont en vert foncé.

```
fviz_mca_biplot (res.ACM, repel = TRUE,
            ggtheme = theme_minimal())
```

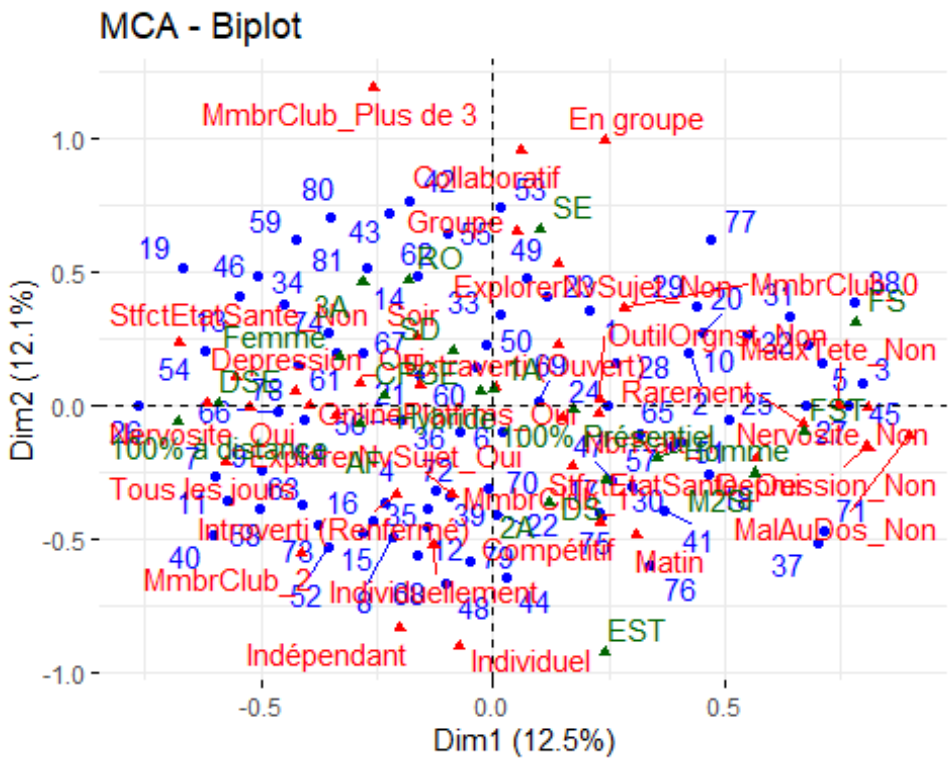


Figure 20: Biplot des variables et des individus

```
# Ou bien avec la fonction
#plot(res.ACM, autoLab = "yes")
```

On peut également visualiser les catégories de variables uniquement:

```
plot(res.ACM,
   invisible = c("ind","quali.sup"),
   cex = 0.8,
   autoLab = "yes")
```

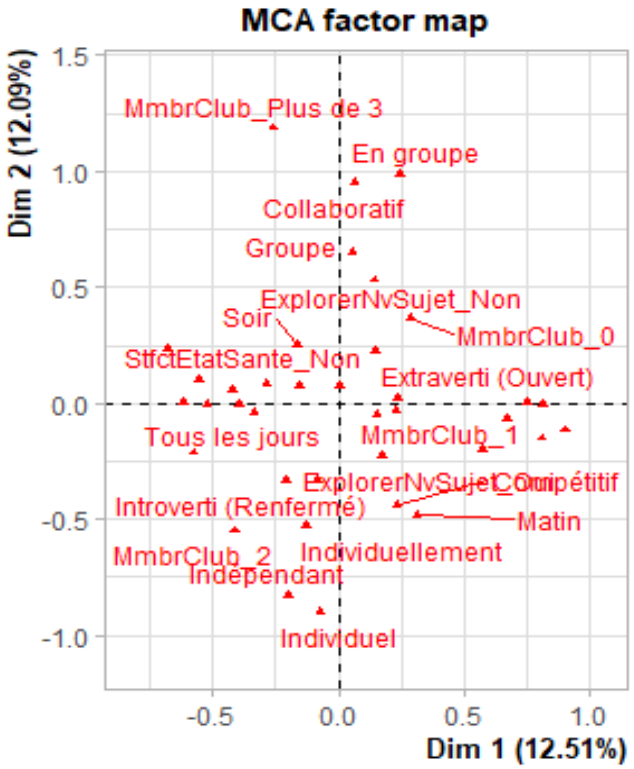


Figure 21: visualiser les catégories de variables uniquement

La distance entre les individus donne une mesure de leur similitude (ou dissemblance). Les individus avec un profil similaire sont proches sur le graphique. Il en va de même pour les variables.

Nous remarquons qu'il est difficile d'analyser le biplot des variables et individus du nombre important d'individus. Nous allons donc par la suite analyser séparément les variables et les individus.

Commençons par l'analyse des variables:

Analyse des variables

Graphique des variables

La fonction get_mca_var() [factoextra] sert à extraire les résultats pour les catégories des variables. Cette fonction renvoie une liste contenant les coordonnées, les cos2 et les contributions des catégories:

```
var <- get_mca_var(res.ACM)
var

## Multiple Correspondence Analysis Results for variables
## ===================================================
##  Name       Description
```

```
## 1 "$coord"   "Coordinates for categories"
## 2 "$cos2"    "Cos2 for categories"
## 3 "$contrib" "contributions of categories"
```

Les composants de get_mca_var() peuvent être utilisés dans le graphique des variables comme suit:

- **var$coord:** coordonnées des variables pour créer un nuage de points
- **var$cos2:** qualité de représentation des variables.
- **var$contrib:** contributions (en pourcentage) des variables à la définition des dimensions.

Extrayons dans un premier temps les coordonnées, les cos2 et les contributions des catégories, nous afficherons uniquement les valeurs pour les cinqs premières dimensions.

- *Les coordonnées des modalités:*

```
head(var$coord)
```

```
##              Dim 1     Dim 2       Dim 3      Dim 4      Dim 5
## Groupe       0.05243239  0.6503161 -0.004325802 -0.02545226  0.03913392
## Individuel   -0.07248007 -0.8989664  0.005979784  0.03518400 -0.05409689
## Collaboratif  0.06177246  0.9511695 -0.148831374 -0.30214189 -0.17621464
## Compétitif    0.23084802 -0.4350445  0.177080366  0.72515972 -0.53050487
## Indépendant  -0.20082267 -0.8314669  0.068330016 -0.06849637  0.50032935
## Matin         0.31010918 -0.4822467 -0.237057261 -0.48702173  0.62020205
```

- *Le cos2 des modalités:*

```
head(var$cos2)
```

```
##              Dim 1     Dim 2       Dim 3       Dim 4      Dim 5
## Groupe       0.003800304 0.58461229 2.586736e-05 0.0008955123 0.002117023
## Individuel   0.003800304 0.58461229 2.586736e-05 0.0008955123 0.002117023
## Collaboratif 0.002760393 0.65448083 1.602397e-02 0.0660393737 0.022462858
## Compétitif   0.014155371 0.05027317 8.329324e-03 0.1396806659 0.074756283
## Indépendant  0.023723379 0.40666899 2.746465e-03 0.0027598543 0.147252621
## Matin        0.050805579 0.12286286 2.968853e-02 0.1253080112 0.203211630
```

- *La contribution des modalités:*

```
head(var$contrib)
```

```
##              Dim 1     Dim 2       Dim 3      Dim 4      Dim 5
## Groupe       0.06069923  9.668529 0.0006082139 0.02424166 0.06300387
## Individuel   0.08390777 13.365319 0.0008407663 0.03351053 0.08709359
## Collaboratif 0.06094734 14.962643 0.5208266759 2.47122732 0.92411398
## Compétitif   0.42558591  1.565056 0.3686507022 7.11751128 4.18784255
## Indépendant  0.56837208 10.088457 0.0968656114 0.11206454 6.57348810
## Matin        1.26494979  3.167452 1.0881527792 5.28769748 9.42729503
```

Corrélation entre les variables et les axes principaux

Pour visualiser la corrélation entre les variables et les axes principaux de l'ACM

```
fviz_mca_var (res.ACM, choice = "mca.cor",col.var ="firebrick", repel = TRUE, ggtheme = theme_mini
mal ())
```

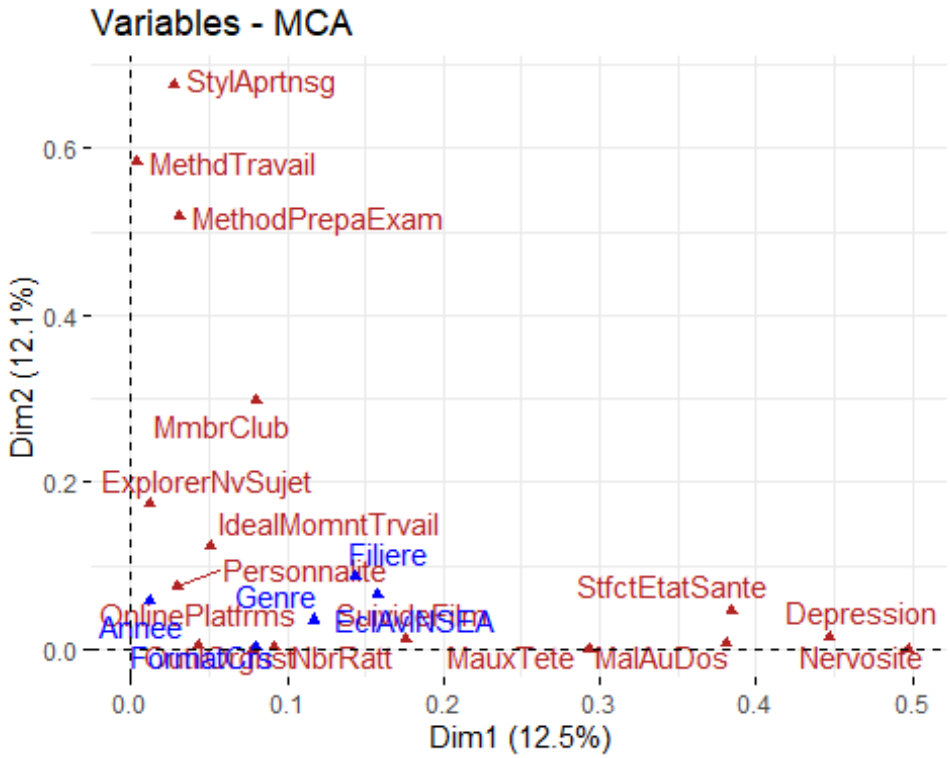


Figure 22: Corrélation entre les variables et les axes principaux

Le graphique ci-dessus permet d'identifier les variables les plus corrélées avec chaque axe. Les corrélations au carré entre les variables et les axes sont utilisés comme coordonnées.

On constate que les variables liées à l'état de santé: ***StfcEtatSante, Depression, MauxAuDos, MauxTets et Nervosite*** sont les plus corrélées avec la dimension 1. De même, les variables liées aux méthodes d'études: ***StylAprntsg, MethdTravailet MethodPreaExam*** sont les plus corrélées avec la dimension 2.

Coordonnées des catégories variables

Dans cette section, nous décrirons comment visualiser uniquement les catégories des variables. Ensuite, nous mettrons en évidence les catégories en fonction soit de leurs qualités de représentation, soit de leurs contributions aux dimensions.

On Utilise la **fonction fviz_mca_var()** [factoextra] pour visualiser uniquement les catégories des variables(active et supplémentaires):

```
# Visualiser les modalités des variables et les axes principaux de l'ACM:
fviz_mca_var (res.ACM,
        repel = TRUE,
        col.var="brown4",
        ggtheme = theme_minimal ())
```

```
## Warning: ggrepel: 7 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

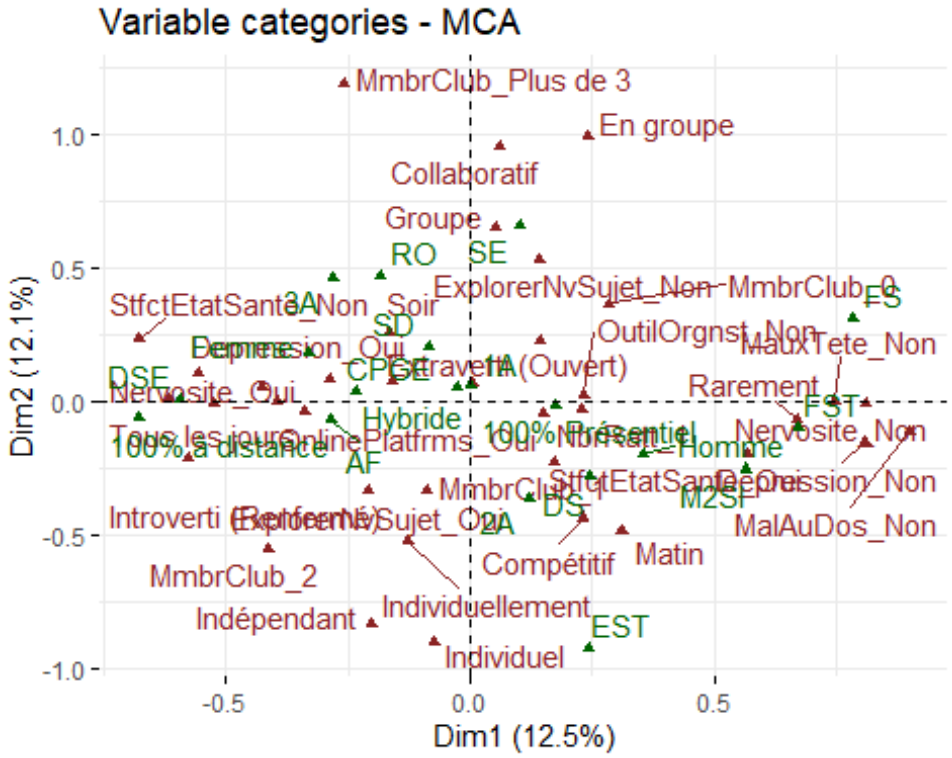
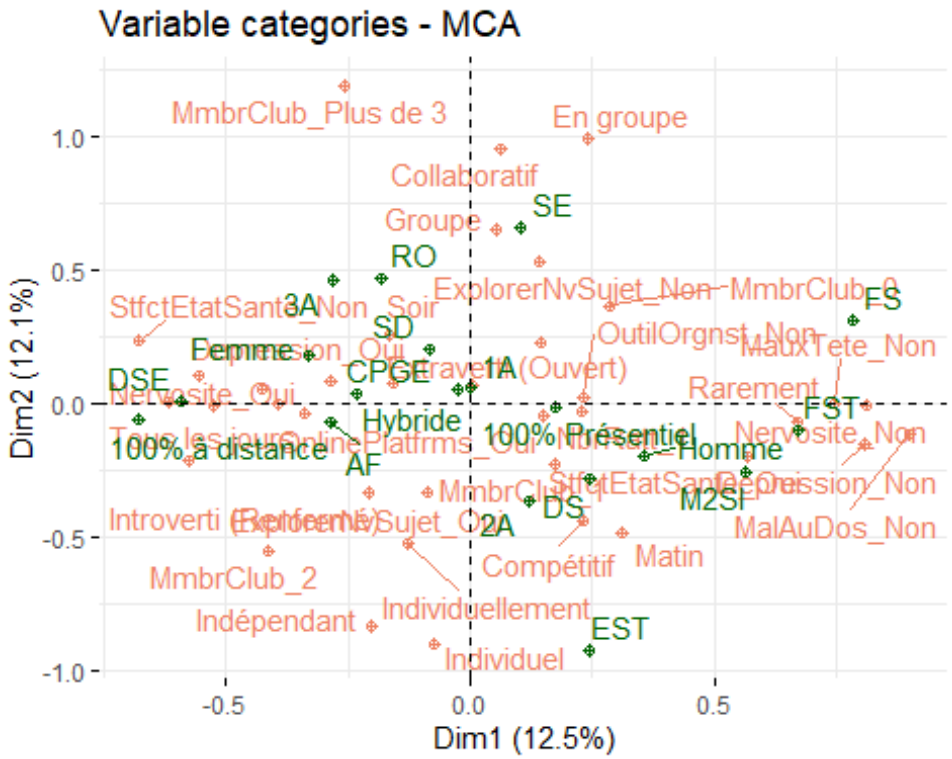


Figure 23: Coordonnées des catégories variables actives et supplémentaires

Il est possible de modifier la couleur et la forme des points à l'aide des arguments **col.var** et **shape.var** comme suit:

```
fviz_mca_var(res.ACM, col.var="salmon2", shape.var = 10,
        repel = TRUE)
```



Le graphique ci-dessus montre les relations entre les catégories des variables. Il peut être interprété comme suit:

- Les catégories avec un profil similaire sont regroupées.
- Les catégories corrélées négativement sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés).
- La distance entre les catégories et l'origine mesure la qualité des catégories. Les points qui sont loin de l'origine sont bien représentés par l'ACM.

Qualité de représentation des catégories des variables

Les deux dimensions 1 et 2 capturent 24.6% de l'inertie totale (variation) contenue dans les données. Tous les points ne sont pas aussi bien représentés par les deux dimensions.

La qualité de représentation, appelée cosinus carré (cos2), mesure le degré d'association entre les catégories des variables et les dimensions. Le cos2 peut être extrait comme suit:

```
head(var$cos2, 5)
```

```
##                Dim 1     Dim 2       Dim 3       Dim 4       Dim 5
## Groupe      0.003800304 0.58461229 2.586736e-05 0.0008955123 0.002117023
## Individuel  0.003800304 0.58461229 2.586736e-05 0.0008955123 0.002117023
## Collaboratif 0.002760393 0.65448083 1.602397e-02 0.0660393737 0.022462858
## Compétitif  0.014155371 0.05027317 8.329324e-03 0.1396806659 0.074756283
## Indépendant 0.023723379 0.40666899 2.746465e-03 0.0027598543 0.147252621
```

Si une catégorie d'une variable donnée est bien représentée par deux dimensions, la somme des cos2 est proche de 1. Pour certains éléments, plus de 2 dimensions sont nécessaires pour représenter parfaitement les données.

Il est possible de colorer les variables en fonction de la valeur de leur cos2 à l'aide de l'argument **col.var** = "cos2". Cela produit un gradient de couleurs. Dans ce cas, l'argument **gradient.cols** peut être utilisé pour spécifier une palette de couleur personnalisée. Par exemple, **gradient.cols** = c("white", "blue", "red") signifie que:

- les variables à faible valeur de cos2 seront colorées en " gold " (jaune)
- les variables avec des valeurs moyennes de cos2 seront colorées en " coral1" (orange)
- les variables avec des valeurs élevées de cos2 seront colorées en " brown4" (marron)

```
fviz_mca_var(res.ACM, col.var = "cos2",
        gradient.cols = c("gold", "coral1", "brown4"),
        repel = TRUE,
        ggtheme = theme_minimal())
```

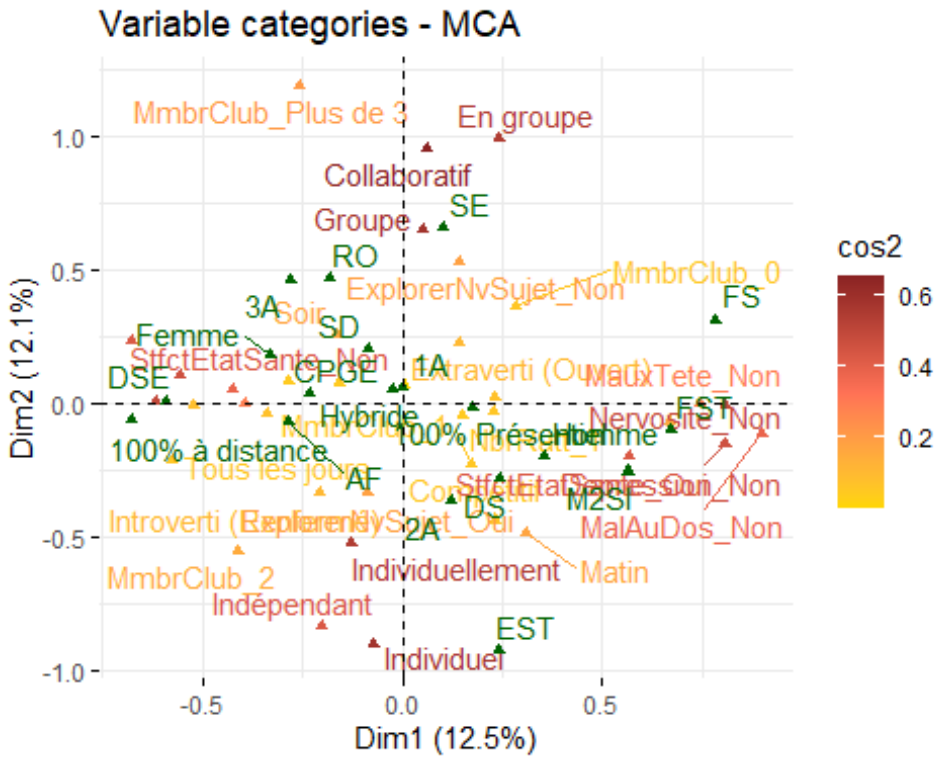


Figure 24: Qualité de représentation des catégories des variables en utilisant le cos2

On peut visualiser le cos2 des catégories sur toutes les dimensions en utilisant le package *corrplot*:

```
library(RColorBrewer)
corrplot(var$cos2,method = 'color', is.corr=FALSE, tl.cex=0.5,tl.col="black", col=brewer.pal(n=8, na
me="BrBG"))
```

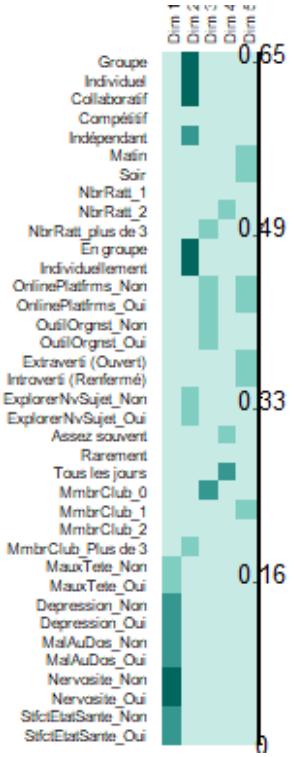


Figure 25: Corrplot de cos2 des variables dans toutes les dimensions

Il est également possible de créer un barplot du cos2 des variables avec la fonction **fviz_cos2()** [factoextra]:

```
fviz_cos2(res.ACM, choice = "var", fill="springgreen4", axes = 1:2)
```

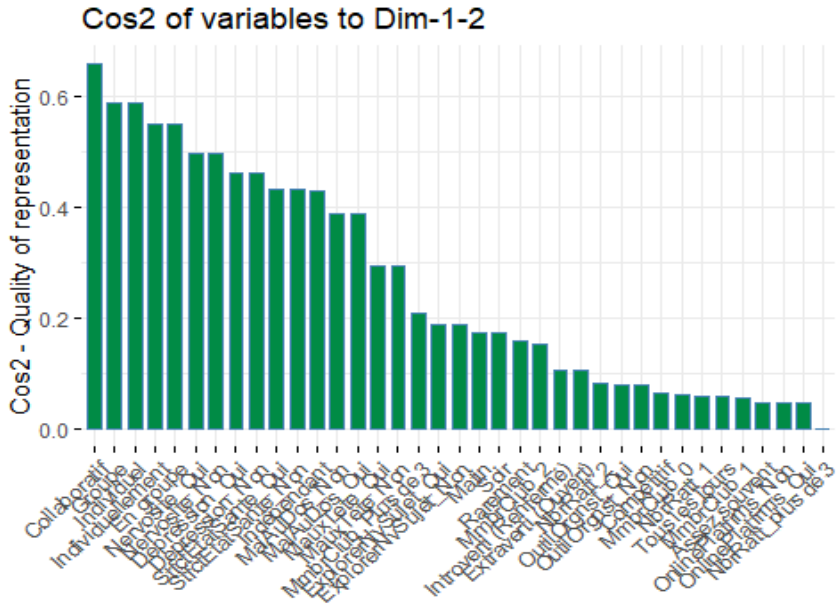


Figure 26: Barplot de cos2 des variables dans les dimensions 1-2

Notez que les catégories NbrRatt_plusde3,OnlinePlatform_Oui et OnlinePlatform_Non ne sont pas très bien représentées par les deux premières dimensions. Cela implique que la position des points correspondants sur le graphique doit être interprétée avec prudence.

Contribution des variables aux dimensions

La contribution des variables (en %) à la définition des dimensions peut être extraite comme suit:

```
head(round(var$contrib,2), 4)
```

```
##             Dim 1 Dim 2 Dim 3 Dim 4 Dim 5
## Groupe       0.06  9.67  0.00  0.02  0.06
## Individuel   0.08 13.37  0.00  0.03  0.09
## Collaboratif 0.06 14.96  0.52  2.47  0.92
## Compétitif   0.43  1.57  0.37  7.12  4.19
```

Les variables avec les plus grandes valeurs, contribuent le mieux à la définition des dimensions. Les catégories qui contribuent le plus à Dim.1 et Dim.2 sont les plus importantes pour expliquer la variabilité dans le jeu de données.

La fonction **fviz_contrib()** [factoextra] peut être utilisée pour faire un barplot de la contribution des catégories des variables. Le code R ci-dessous montre le top 15 des catégories contribuant aux dimensions:

```
# Contributions des variables à la dimension 1
fviz_contrib (res.ACM, choice = "var",fill="springgreen4", axes = 1, top = 15)
```

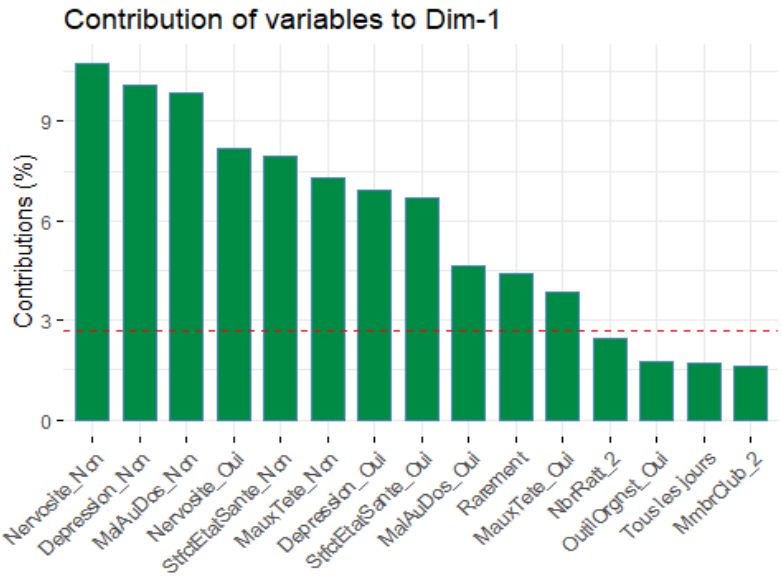


Figure 27: Contribution des variables à la dimension 1

```
# Contributions des variables à la dimension 2
fviz_contrib (res.ACM, choice = "var",fill="springgreen4", axes = 2, top = 15)
```

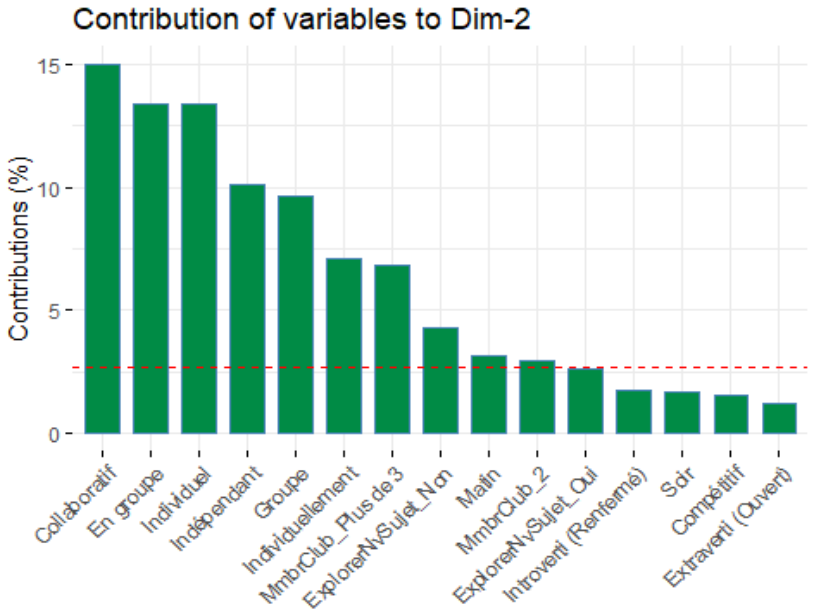


Figure 28: Contribution des variables à la dimension 2

```
# Contributions des variables à la dimension 3
fviz_contrib (res.ACM, choice = "var",fill="springgreen4", axes = 3, top = 15)
```

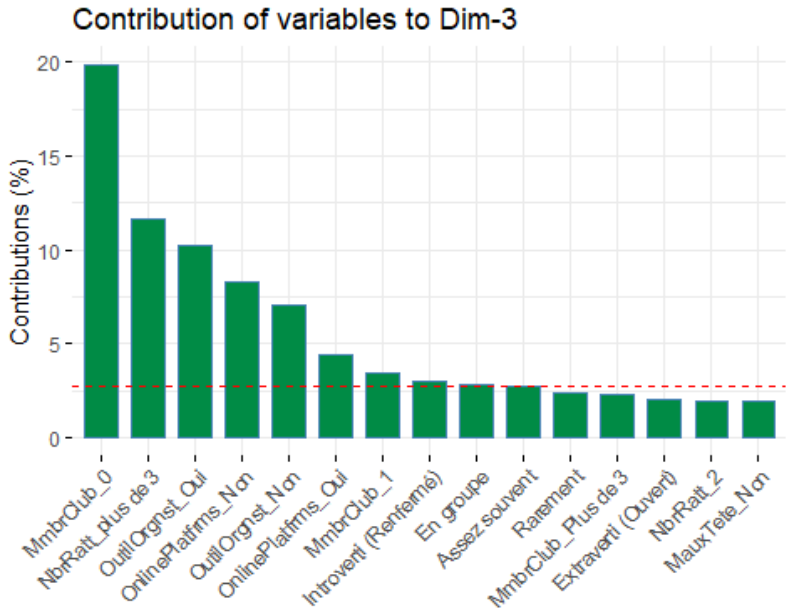


Figure 29: Contribution des variables à la dimension 3

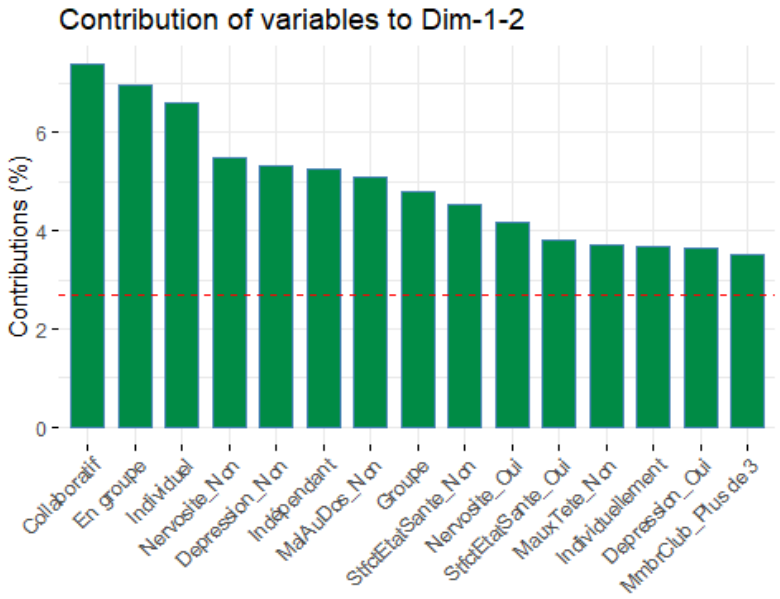
La ligne en pointillé rouge, sur le graphique ci-dessus, indique la valeur moyenne attendue sous l'hypothèse nulle.

On peut voir que:

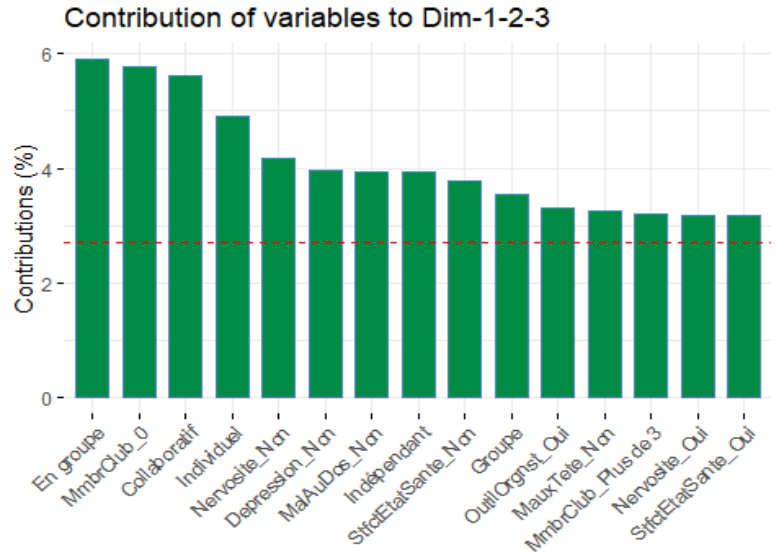
- Les catégories ***Nervosite_non, Depression_Non et MalAuDos_Non*** contribuent le plus à la dimension 1
- les catégories co***llaboratif, En groupe et Individuel*** sont les plus importantes dans la définition de la deuxième dimension.

Les contributions totales aux dimensions 1 et 2 sont obtenues comme suit:

```
fviz_contrib(res.ACM, choice = "var",fill="springgreen4", axes = 1:2, top = 15)
```




```
fviz_contrib(res.ACM, choice = "var",fill="springgreen4", axes = 1:3, top = 15)
```



Les catégories les plus importantes peuvent être mises en évidence sur le graphique comme suit:

```
fviz_mca_var(res.ACM, col.var = "contrib",
             gradient.cols = c("gold", "coral1", "brown4"),
             repel = TRUE,
             ggtheme = theme_minimal()
)
```

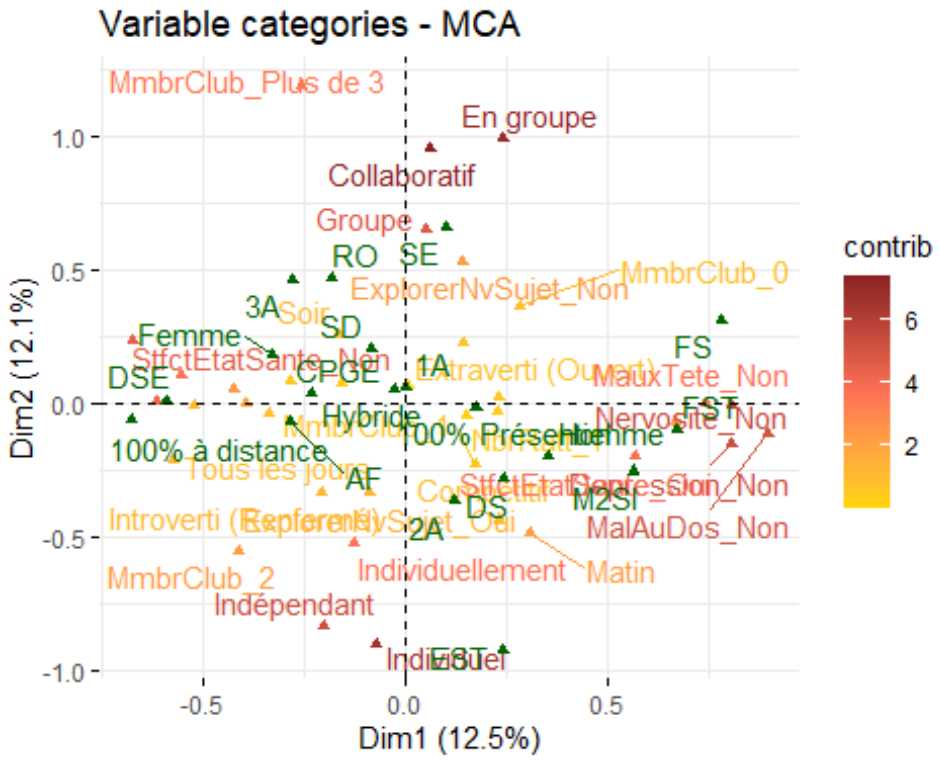


Figure 30: Graphique des modalités selon leur contribusion à la création des 2 dimensions

Le graphique ci-dessus donne une idée du pôle des dimensions auquel les catégories contribuent réellement.

Il est évident que les catégories ***Nervosite_non, Depression_Non et MalAuDos_Non*** ont une contribution importante au pôle positif de la première dimension, tandis que la catégorie ***StfctEtatSante_Non*** a une contribution majeure au pôle négatif de la première dimension, etc.

Graphique des modalités Zoom des quatre quadrants

```
plot(res.ACM,invisible=c("ind","quali.sup"),xlim=c(0,2.5),ylim=c(-0.25,0), hab="quali",repel = TRUE,
palette=palette(c("blue","orange","darkgreen","black","red")))
```

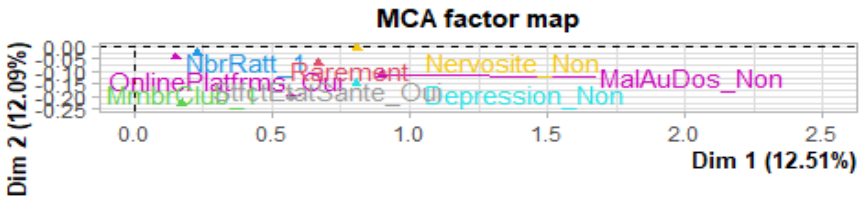


Figure 31: Zoom sur le graphique des modalités - 1ière dimension pôle positif

```
plot(res.ACM,invisible=c("ind","quali.sup"),xlim=c(0,2),ylim=c(0,1), hab="quali",repel = TRUE, palet
te=palette(c("blue","orange","darkgreen","black","red")))
```

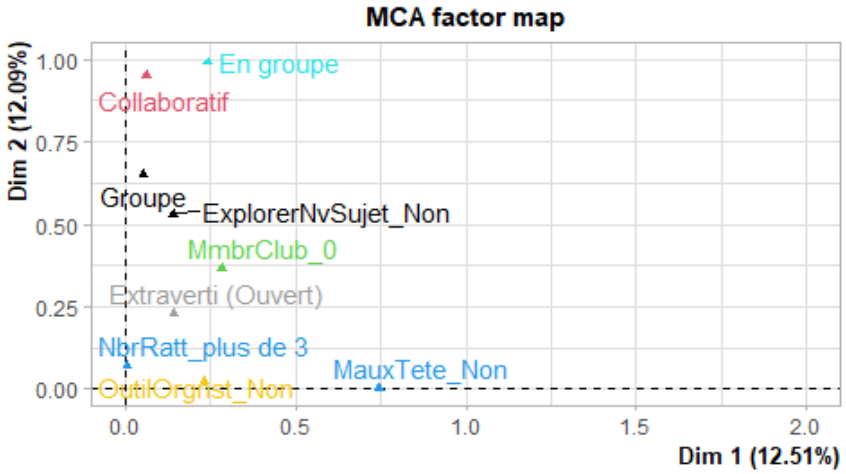


Figure 32: Zoom sur le graphique des modalités - 2ième dimension pôle positif

```
plot(res.ACM,invisible=c("ind","quali.sup"),xlim=c(-1.5,0),ylim=c(0,1.5), hab="quali",repel = TRUE, p
alette=palette(c("blue","orange","darkgreen","black","red")))
```

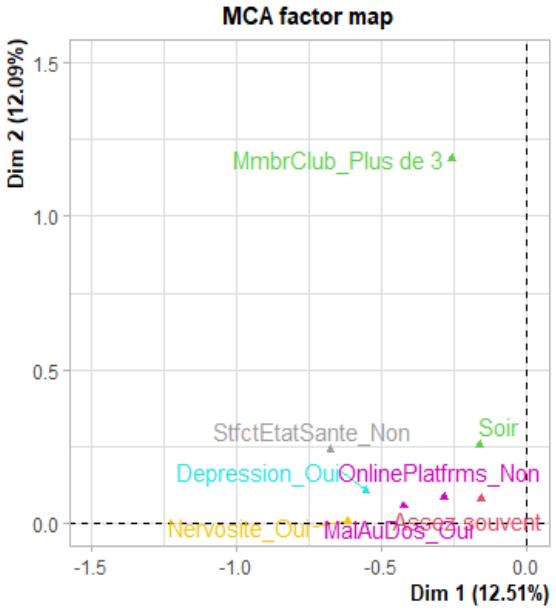


Figure 33: Zoom sur le graphique des modalités - 1ière dimension pôle négatif

```
plot(res.ACM,invisible=c("ind","quali.sup"),xlim=c(0,-2),ylim=c(-0,-1), hab="quali",repel = TRUE, palette=palette(c("blue","orange","darkgreen","black","red")))
```

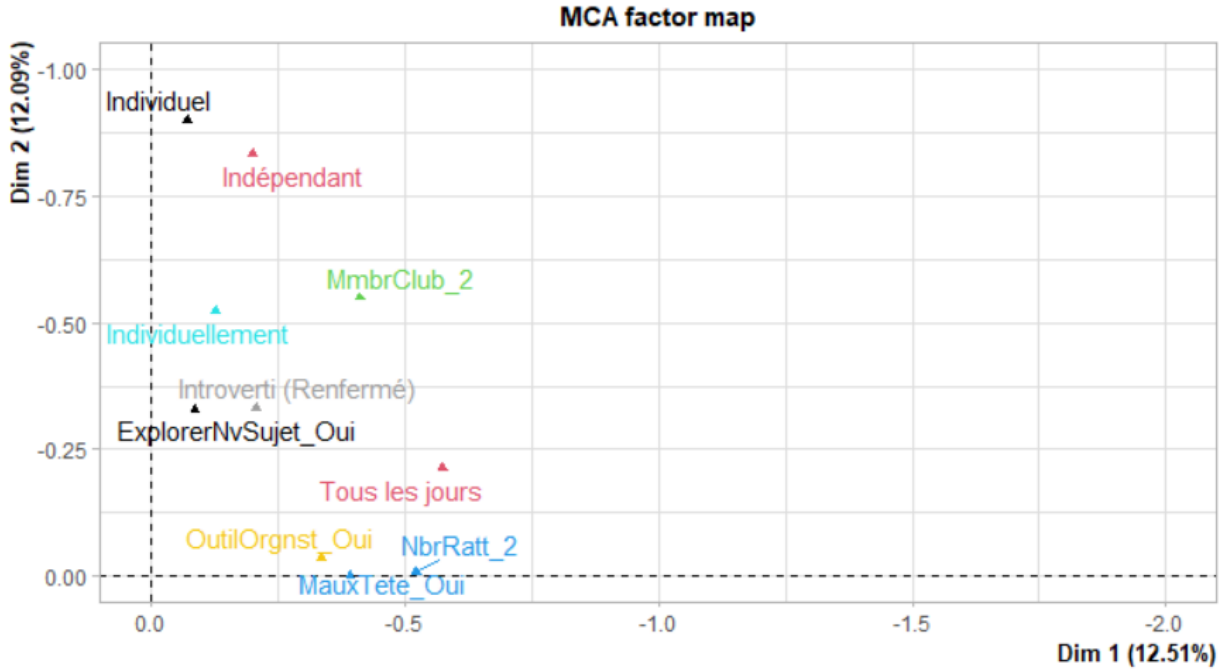


Figure 34: Zoom sur le graphique des modalités - 2ième dimension pôle négatif

Analyse des individus

La fonction **get_mca_ind()** [factoextra] sert à extraire les résultats pour les individus. Cette fonction renvoie une liste contenant les coordonnées, la cos2 et les contributions des individus:

```
ind <- get_mca_ind (res.ACM)
ind

## Multiple Correspondence Analysis Results for individuals
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the individuals"
## 2 "$cos2"    "Cos2 for the individuals"
## 3 "$contrib" "contributions of the individuals"
```

Pour accéder aux différents composants, on utilise:

```
# Coordonnées
head(ind$coord)

##        Dim 1      Dim 2      Dim 3       Dim 4      Dim 5
## 1  0.20760369  0.35811305 -0.4443457 -0.003785363 -0.2052082
## 2  0.42247886  0.19834025 -0.3390832 -0.375369519  0.1912403
## 3  0.79492672  0.08273465 -0.2553719  0.519914574  0.1613811
## 4 -0.23263763 -0.35932727 -0.1683909  0.671569359  0.1828431
## 5  0.70765982  0.16013077 -0.2133062 -0.084820830  0.2387975
## 6  0.02251166 -0.09543031 -0.2000617  0.206956066  0.1280104

# Qualité de representation
head(ind$cos2)

##         Dim 1      Dim 2      Dim 3       Dim 4      Dim 5
## 1 0.0403124156 0.119952409 0.18467609 1.340243e-05 0.03938748
## 2 0.1269146332 0.027971996 0.08175501 1.001889e-01 0.02600523
## 3 0.3596786463 0.003896147 0.03711988 1.538595e-01 0.01482402
## 4 0.0358686076 0.085572645 0.01879281 2.989072e-01 0.02215704
## 5 0.3885886693 0.019897140 0.03530597 5.582723e-03 0.04424871
## 6 0.0004230842 0.007602987 0.03341484 3.575756e-02 0.01368050

# Contributions
head(ind$contrib)

##       Dim 1      Dim 2     Dim 3       Dim 4     Dim 5
## 1 0.32394862 0.99809942 2.1846773 0.0001825353 0.5897557
## 2 1.34157809 0.30616479 1.2722074 1.7949397861 0.5122023
## 3 4.74963444 0.05327315 0.7215915 3.4434664783 0.3647440
## 4 0.40678591 1.00487921 0.3137491 5.7453120387 0.4682089
## 5 3.76404721 0.19956441 0.5034455 0.0916509377 0.7986238
## 6 0.00380909 0.07087714 0.4428668 0.5456173799 0.2294950
```

Graphique des individus

La fonction **fviz_mca_ind()** [factoextra] sert à visualiser uniquement des individus. Comme les variables, il est également possible de colorer les individus en fonction de leurs cos2:

```
fviz_mca_ind(res.ACM, col.ind = "cos2",
             gradient.cols = c("gold", "coral1", "brown4"),
             repel = TRUE,
             ggtheme = theme_minimal())
```

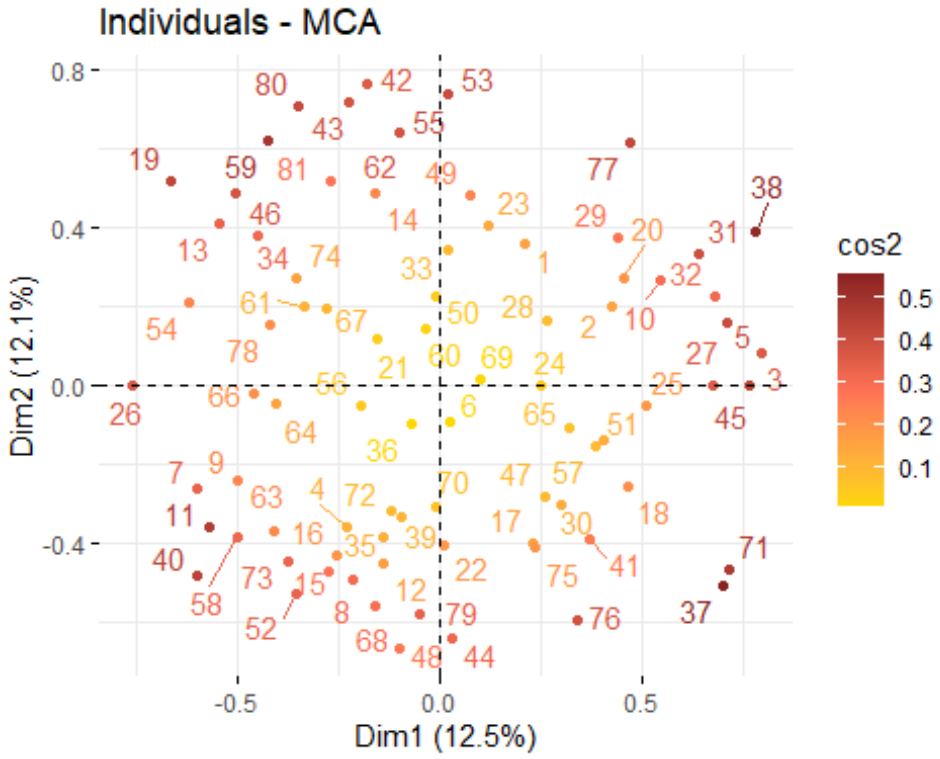


Figure 35: Graphique des individus

Le code R ci-dessous crée un barplot du cos2 et de la contribution des individus:

```
# Cos2 des individus
fviz_cos2(res.ACM, choice = "ind",fill="springgreen4", axes = 1:2, top = 20)
```

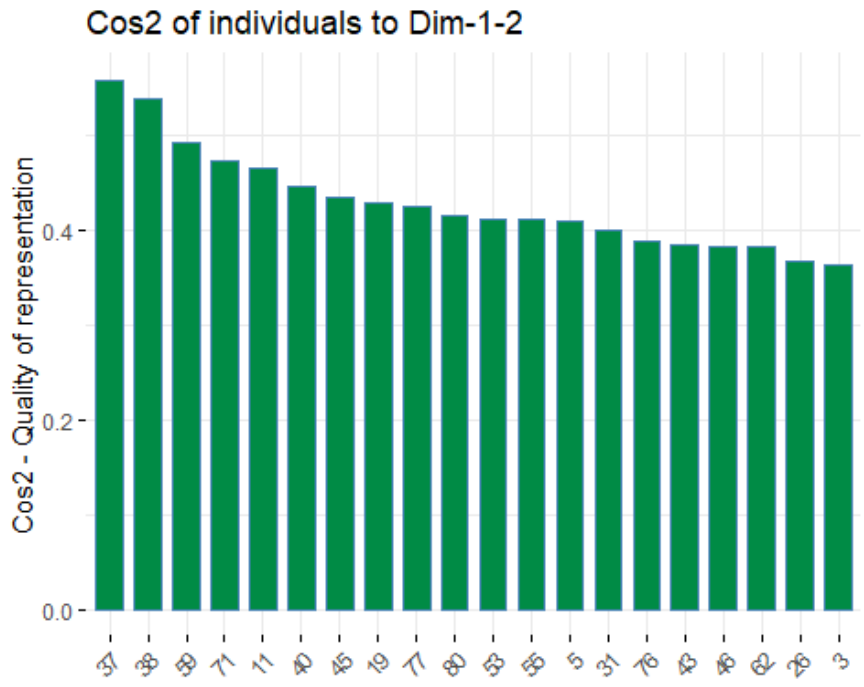


Figure 36: Graphique de Cos2 des individus

```
# Contribution des individus aux dimensions
fviz_contrib(res.ACM, choice = "ind",fill="springgreen4", axes = 1:2, top = 20)
```

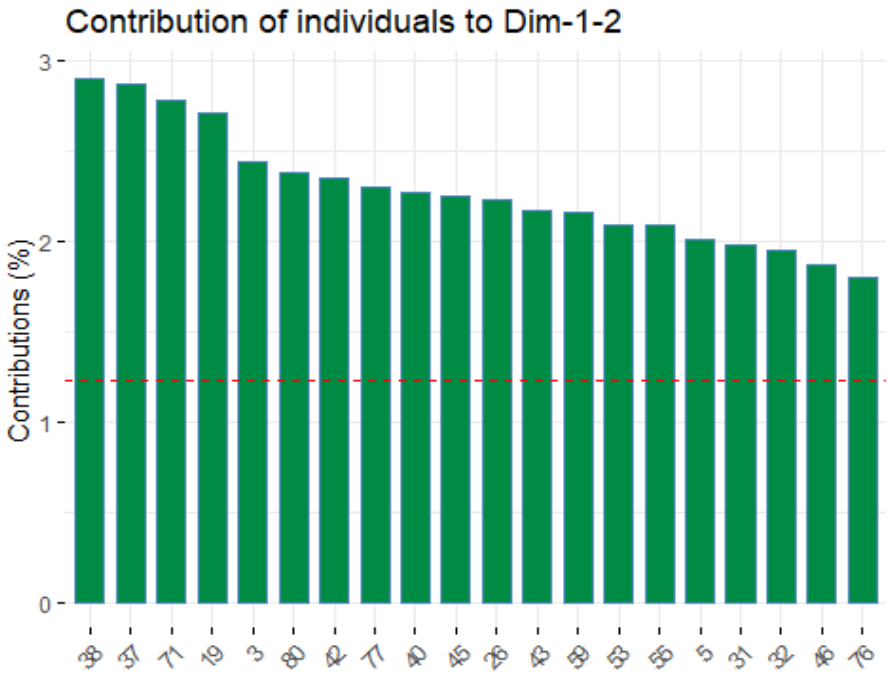


Figure 37: Graphique de contribustion des individus aux dimensions

Colorer les individus par groupes

Le code R ci-dessous colore les individus par groupes en utilisant la variable ***NbrRatt***, puis ***StfctEtatSante***. L'argument habillage sert à spécifier la variable à utiliser pour colorer les individus par groupes. Une ellipse de concentration peut également être ajoutée autour de chaque groupe en utilisant l'argument **addEllipses = TRUE**. Si vous voulez une ellipse de confiance autour du point moyen (centre de gravité) des groupes, utilisez **ellipse.type = "confidence".** L'argument palette permet de modifier les couleurs du groupe.

```
# groupes en utilisant la variable NbrRatt
fviz_mca_ind (res.ACM,
        label = "none", # masquer le texte des individus
        habillage = "NbrRatt", # colorer par groupes
        palette = c ("#00AFBB", "#E7B800","#FC4E07"),
        addEllipses = TRUE, ellipse.type = "confidence",
        ggtheme = theme_minimal ())
```

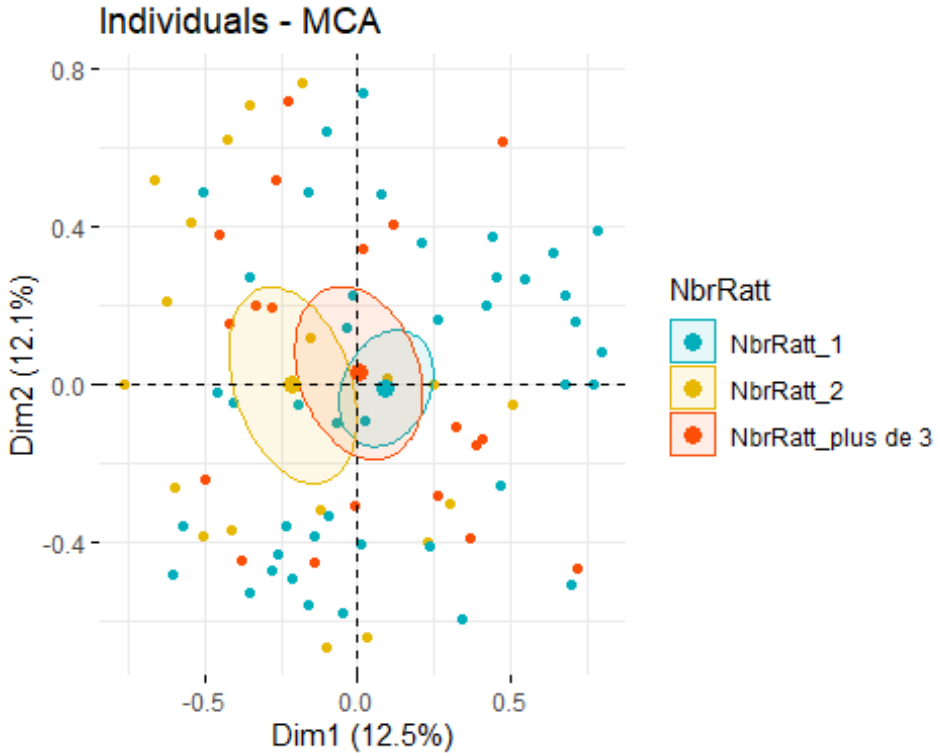


Figure 38: Groupes en utilisant la variable NbrRatt

```
# groupes en utilisant la variable StfctEtatSante
fviz_mca_ind (res.ACM,
        label = "none", # masquer le texte des individus
        habillage = "StfctEtatSante", # colorer par groupes
        palette = c ("#00AFBB", "#E7B800"),
        addEllipses = TRUE, ellipse.type = "confidence",
        ggtheme = theme_minimal ())
```

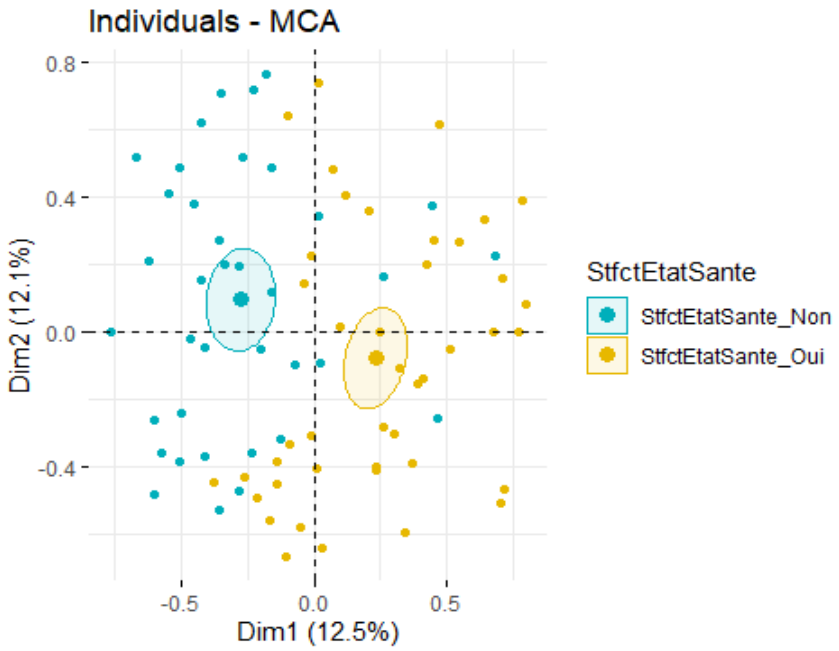


Figure 39: Groupes en utilisant la variable StfctEtatSante

```
## Si vous souhaitez colorer les individus à l'aide de plusieurs variables catégorielles en même temps, utilisez la fonction fviz_ellipses() [factoextra] comme suit:
fviz_ellipses(res.ACM, c("NbrRatt", "StfctEtatSante"),
              geom = "point")
```

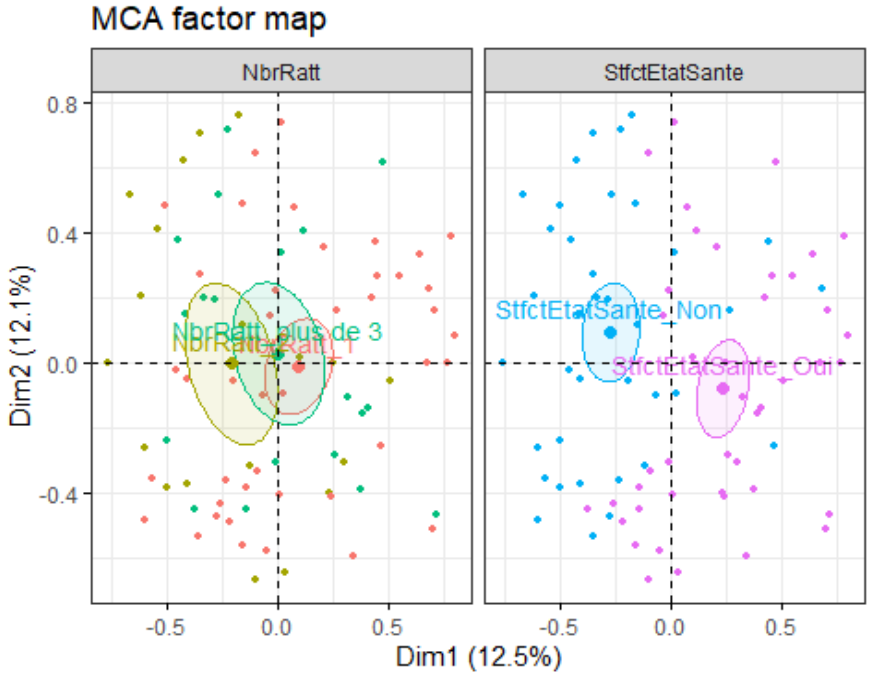


Figure 40: Groupes en utilisant plusieurs variables

Description des dimensions

La fonction dimdesc() [FactoMineR] peut être utilisée pour identifier les variables les plus corrélées avec une dimension donnée:

```
res.desc <- dimdesc (res.ACM, axes = c(1,2))
# Description de la dimension 1
res.desc[[1]]

## $quali
##                   R2     p.value
## Nervosite       0.49772065 1.929051e-13
## Depression      0.44721260 8.936014e-12
## StfctEtatSante  0.38427997 6.789783e-10
## MalAuDos        0.38105367 8.379769e-10
## MauxTete        0.29335653 1.775727e-07
## SuivideFilm     0.17596600 5.270600e-04
## Genre           0.11697063 1.777695e-03
## EclAvINSEA      0.15787990 4.002000e-03
## OutilOrgnst     0.07804844 1.154448e-02
## NbrRatt         0.09198359 2.320888e-02
## FormatCrs       0.07998096 3.873252e-02
## IdealMomntTrvail 0.05080558 4.305314e-02
##
## $category
##                                  Estimate     p.value
## Nervosite=Nervosite_Non          0.28859540 1.929051e-13
## Depression=Depression_Non        0.27579662 8.936014e-12
## StfctEtatSante=StfctEtatSante_Oui  0.25217742 6.789783e-10
## MalAuDos=MalAuDos_Non            0.26793816 8.379769e-10
## MauxTete=MauxTete_Non            0.23077590 1.775727e-07
## SuivideFilm=Rarement             0.27973498 2.554960e-04
## Genre=Homme                      0.13870464 1.777695e-03
## EclAvINSEA=FS                    0.16916355 7.721141e-03
## Filiere=M2SI                     0.24190472 8.510509e-03
## OutilOrgnst=OutilOrgnst_Non      0.11521633 1.154448e-02
## NbrRatt=NbrRatt_1                0.13169626 2.865539e-02
## IdealMomntTrvail=Matin           0.09603914 4.305314e-02
## Filiere=DSE                     -0.22616258 4.650522e-02
## IdealMomntTrvail=Soir           -0.09603914 4.305314e-02
## SuivideFilm=Tous les jours      -0.22453947 4.057164e-02
## MmbrClub=MmbrClub_2             -0.14539816 3.397730e-02
## FormatCrs=100% à distance       -0.20334756 1.531808e-02
## OutilOrgnst=OutilOrgnst_Oui     -0.11521633 1.154448e-02
## NbrRatt=NbrRatt_2               -0.17303247 8.720948e-03
## Genre=Femme                     -0.13870464 1.777695e-03
## EclAvINSEA=CPGE                 -0.24274048 4.413919e-04
## MauxTete=MauxTete_Oui           -0.23077590 1.775727e-07
## MalAuDos=MalAuDos_Oui           -0.26793816 8.379769e-10
## StfctEtatSante=StfctEtatSante_Non -0.25217742 6.789783e-10
```

```
## Depression=Depression_Oui       -0.27579662 8.936014e-12
## Nervosite=Nervosite_Oui         -0.28859540 1.929051e-13
##
## attr(,"class")
## [1] "condes" "list"
```

```
# Description de la dimension 2
res.desc[[2]]
```

```
## $quali
##                    R2     p.value
## StylAprtnsg     0.67553334 8.622497e-20
## MethdTravail    0.58461229 9.851574e-17
## MethodPrepaExam  0.51884839 3.463601e-14
## MmbrClub        0.29897312 4.574539e-06
## ExplorerNvSujet  0.17514068 1.010797e-04
## IdealMomntTrvail 0.12286286 1.336342e-03
## Personnalite    0.07581761 1.284900e-02
##
## $category
##                                Estimate     p.value
## StylAprtnsg=Collaboratif          0.4206988 6.466455e-20
## MethdTravail=Groupe               0.3085257 9.851574e-17
## MethodPrepaExam=En groupe          0.3016123 3.463601e-14
## MmbrClub=MmbrClub_Plus de 3        0.3953682 3.081558e-05
## ExplorerNvSujet=ExplorerNvSujet_Non  0.1714642 1.010797e-04
## IdealMomntTrvail=Soir              0.1467706 1.336342e-03
## Personnalite=Extraverti (Ouvert)    0.1115972 1.284900e-02
## StylAprtnsg=Compétitif           -0.1314054 4.418986e-02
## EclAvINSEA=EST                  -0.3005298 3.341832e-02
## Personnalite=Introverti (Renfermé)  -0.1115972 1.284900e-02
## MmbrClub=MmbrClub_2               -0.2960432 4.263383e-03
## IdealMomntTrvail=Matin           -0.1467706 1.336342e-03
## ExplorerNvSujet=ExplorerNvSujet_Oui -0.1714642 1.010797e-04
## StylAprtnsg=Indépendant          -0.2892934 1.530589e-10
## MethodPrepaExam=Individuellement    -0.3016123 3.463601e-14
## MethdTravail=Individuel          -0.3085257 9.851574e-17
##
## attr(,"class")
## [1] "condes" "list"
```

Eléments supplémentaires

Résultats

Les résultats prédits pour les individus / variables supplémentaires peuvent être extraits comme suit:

```
# Variables qualitatives supplémentaires
head(res.ACM$quali.sup)
```

```
## $coord
##                 Dim 1      Dim 2       Dim 3      Dim 4      Dim 5
## Femme         -0.329568783  0.18076367 -0.163777413 -0.23147800  0.20900975
## Homme          0.354920228 -0.19466857  0.176375675  0.24928399 -0.22508743
## AF            -0.284815279 -0.07045246  0.558587579 -0.04758221  0.23522840
## DS             0.244158807 -0.28315542 -0.225666576  0.29371079 -0.74932112
## DSE           -0.591219720  0.01036762 -0.685522584  0.26360025  0.27871434
## M2SI           0.563705332 -0.25335159 -0.081377042 -0.12720319 -0.02487164
## RO            -0.181909876  0.46805426 -0.014670577  0.47916119 -0.07796826
## SD            -0.084519402  0.20208712  0.708563609  0.08398397  0.30885803
## SE             0.102350371  0.65748297 -0.693814538 -1.04030866  0.15086467
## 1A             0.001716772  0.06079179 -0.017445371  0.03231287  0.23741238
## 2A             0.121567530 -0.36577243 -0.235336008 -0.14938770 -0.40220358
## 3A            -0.280069195  0.46158616  0.623764387  0.14527607 -0.47793024
## CPGE          -0.232958144  0.03696102  0.013134170  0.05516366  0.03750120
## EST            0.242000681 -0.92261530 -0.124664165 -0.41721155 -0.13088038
## FS             0.783387763  0.31273252  0.126403870 -0.33393703 -0.11264377
## FST            0.671521350 -0.09927838 -0.202233415  0.31011032 -0.06167586
## 100% à distance -0.677633210 -0.05710716  0.402496700  0.25728054 -0.28342144
## 100% Présentiel  0.175209142 -0.01389585  0.009674455  0.06782908 -0.08100621
## Hybride        -0.025235634  0.05280783 -0.203020360 -0.24792598  0.28516340
##
## $cos2
##                 Dim 1       Dim 2       Dim 3       Dim 4
## Femme          1.169706e-01 3.518901e-02 0.0288863517 0.0577037594
## Homme          1.169706e-01 3.518901e-02 0.0288863517 0.0577037594
## AF             2.659664e-02 1.627393e-03 0.1023016666 0.0007423168
## DS             1.139670e-02 1.532795e-02 0.0097357389 0.0164920343
## DSE            4.923109e-02 1.513908e-05 0.0661889033 0.0097866327
## M2SI           8.440598e-02 1.704968e-02 0.0017590280 0.0042979852
## RO             3.130249e-03 2.072329e-02 0.0000203592 0.0217184885
## SD             6.757393e-04 3.863168e-03 0.0474923881 0.0006672047
## SE             9.909350e-04 4.089172e-02 0.0455358147 0.1023742530
## 1A             5.284826e-06 6.626668e-03 0.0005457149 0.0018722181
## 2A             4.845464e-03 4.386540e-02 0.0181583727 0.0073169455
## 3A             9.804844e-03 2.663272e-02 0.0486352512 0.0026381421
## CPGE           1.455409e-01 3.663677e-03 0.0004626309 0.0081608511
## EST            3.852916e-03 5.600125e-02 0.0010224443 0.0114516762
## FS             8.643611e-02 1.377488e-02 0.0022504139 0.0157061889
## FST            4.265657e-02 9.323429e-04 0.0038687632 0.0090970116
## 100% à distance 7.215792e-02 5.124787e-04 0.0254577076 0.0104018005
## 100% Présentiel 4.034626e-02 2.537815e-04 0.0001230107 0.0060467448
## Hybride        2.681420e-04 1.174176e-03 0.0173546386 0.0258809656
##                 Dim 5
## Femme          0.0470454681
## Homme          0.0470454681
## AF             0.0181417701
## DS             0.1073421728
## DSE            0.0109410824
## M2SI           0.0001643152
```

```
## RO           0.0005750453
## SD           0.0090236888
## SE           0.0021529871
## 1A           0.1010676274
## 2A           0.0530385964
## 3A           0.0285521648
## CPGE         0.0037715474
## EST          0.0011269523
## FS           0.0017871293
## FST          0.0003598295
## 100% à distance 0.0126229263
## 100% Présentiel 0.0086243516
## Hybride        0.0342392280
##
## $v.test
##                 Dim 1      Dim 2      Dim 3      Dim 4      Dim 5
## Femme         -3.05902766  1.67783207 -1.52016714 -2.1485578  1.9400097
## Homme          3.05902766 -1.67783207  1.52016714  2.1485578 -1.9400097
## AF            -1.45867439 -0.36082051  2.86079243 -0.2436911  1.2047164
## DS             0.95484880 -1.10735556 -0.88252995  1.1486352 -2.9304221
## DSE           -1.98456227  0.03480125 -2.30111109  0.8848337  0.9355675
## M2SI           2.59855318 -1.16789313 -0.37512963 -0.5863777 -0.1146526
## RO            -0.50041974  1.28758041 -0.04035760  1.3181347 -0.2144846
## SD            -0.23250622  0.55592575  1.94920267  0.2310333  0.8496441
## SE             0.28155781  1.80868385 -1.90862914 -2.8618072  0.4150168
## 1A             0.02056176  0.72810264 -0.20894303  0.3870109  2.8434856
## 2A             0.62260509 -1.87329442 -1.20526753 -0.7650854 -2.0598757
## 3A            -0.88565656  1.45966361  1.97251618  0.4594033 -1.5113481
## CPGE          -3.41222418  0.54138169  0.19238105  0.8080025  0.5492939
## EST            0.55518764 -2.11662466 -0.28599921 -0.9571489 -0.3002602
## FS             2.62961763  1.04975720  0.42430309 -1.1209349 -0.3781142
## FST            1.84730233 -0.27310700 -0.55632819  0.8530891 -0.1696654
## 100% à distance -2.40263057 -0.20248036  1.42710077  0.9122193 -1.0049050
## 100% Présentiel  1.79658037 -0.14248692  0.09920108  0.6955139 -0.8306312
## Hybride        -0.14646283  0.30648661 -1.17829160 -1.4389153  1.6550342
##
## $eta2
##                Dim 1      Dim 2      Dim 3      Dim 4      Dim 5
## Genre     0.11697063 0.035189006 0.028886352 0.057703759 0.047045468
## Filiere   0.14382387 0.087397669 0.229629820 0.140356761 0.124233038
## Annee     0.01236636 0.059080480 0.057101538 0.008525608 0.101507055
## EclAvINSEA 0.15787990 0.066465500 0.006591994 0.035039313 0.003976987
## FormatCrs  0.07998096 0.001378813 0.034266164 0.029814533 0.038729545
```

Graphique

Pour visualiser les variables supplémentaires:

```
fviz_mca_var(res.ACM,
             label = "var.sup",
             col.var="brown4",
             ggtheme = theme_minimal())
```

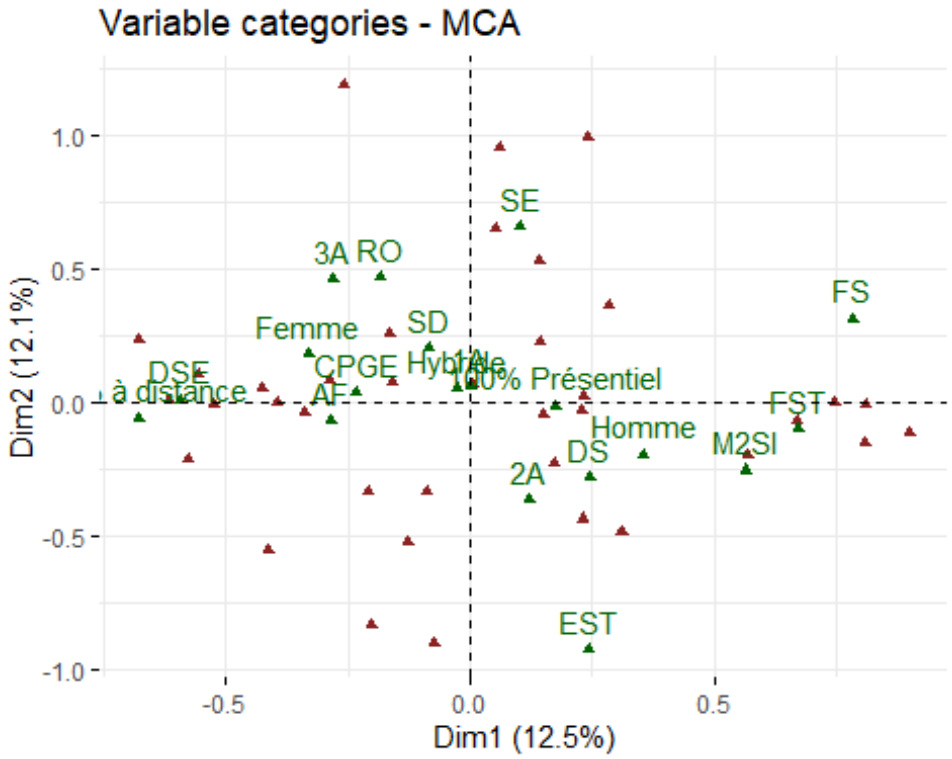


Figure 41: Graphique des variables supplémentaires

Résumé de l'analyse et filtrage des résultats

Dans le cas de plusieurs individus / variables, il est possible de visualiser seulement certains d'entre eux en utilisant les arguments select.ind et select.var.

select.ind, **select.var**: une sélection d'individus / variables à visualiser. Les valeurs autorisées sont NULL ou une liste contenant le nom des arguments, cos2 ou contrib:

- **name**: est un vecteur de caractères contenant le nom des individus / variables à visualiser
- **cos2**: si cos2 est dans [0, 1], ex: 0.6, alors les individus / variables avec un cos2> 0.6 sont montrés. si cos2> 1, ex: 5, le top 5 des individus / variables actifs ainsi que le top 5 des individus/ variables supplémentaires avec le cos2 le plus élevé sont montrés
- **contrib**: si contrib> 1, ex: 5, alors les top 5 individus / variables avec les contributions les plus importantes sont montrés

```
# Visualiser les catégories de variables avec cos2> = 0.4
fviz_mca_var (res.ACM, select.var = list (cos2 = 0.4),title ="Catégories de variables avec cos2> = 0.4")
```

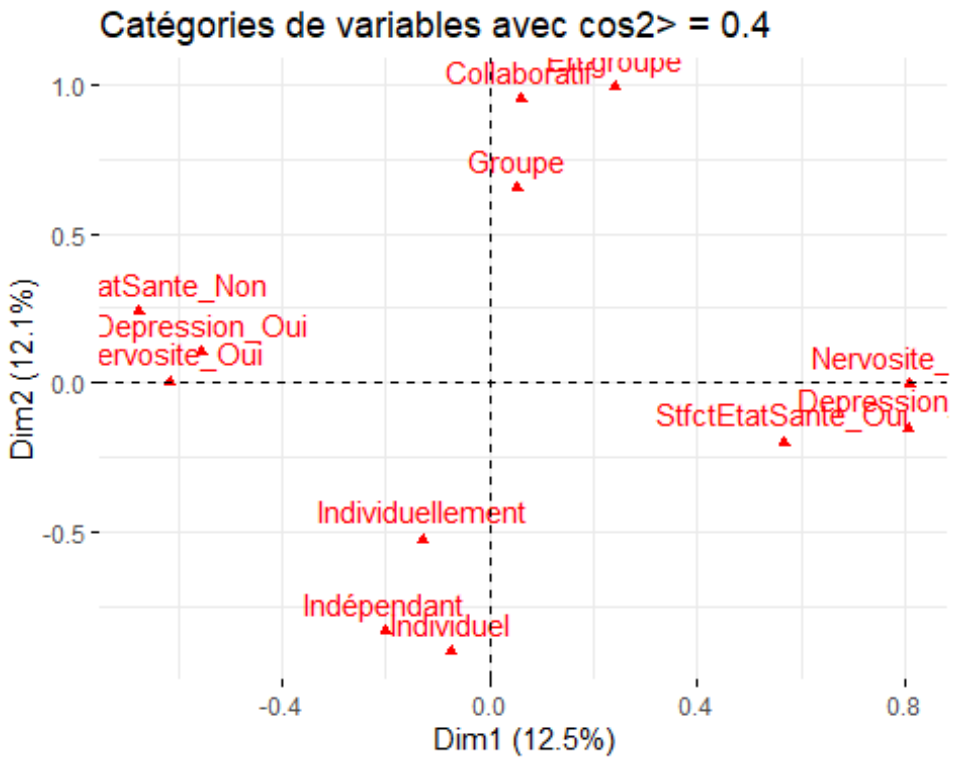


Figure 42: Graphique des Catégories de variables avec cos2> = 0.4

```
# Top 10 des variables actives avec le cos2 le plus elevé
fviz_mca_var (res.ACM, select.var = list (cos2 = 10),title="Top 10 des variables actives avec le cos2 le plus elevé")
```

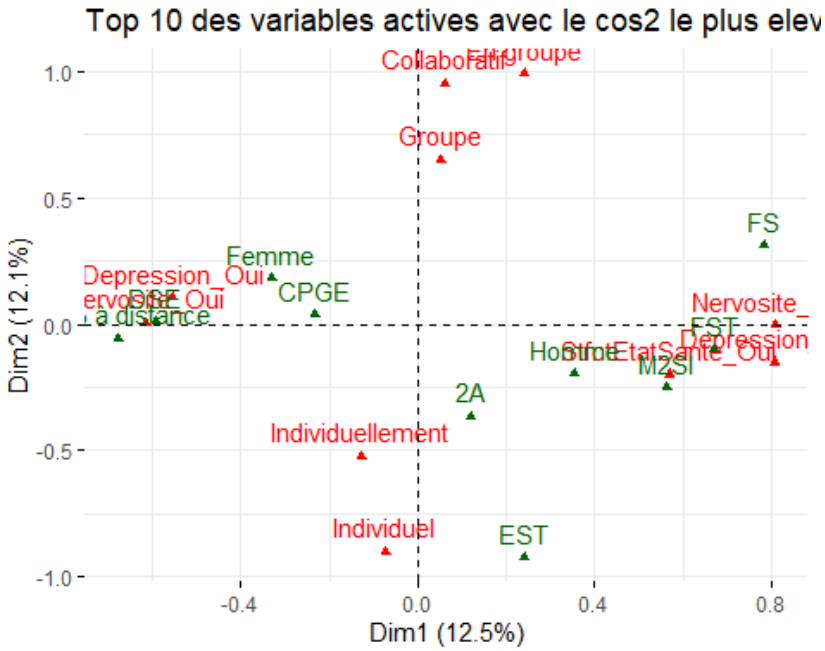


Figure 43: Graphique de Top 10 des variables actives avec le cos2 le plus elevé

```
# Sélectionner par noms
name <- list(name = c("Groupe", "Individuel","Matin","Soir", "StfctEtatSante_Oui","StfctEtatSante_Non", "NbrRatt_1", "NbrRatt_2", " NbrRatt_plus de 3 "))
fviz_mca_var (res.ACM, select.var = name, title="Sélectionner variables par noms")
```

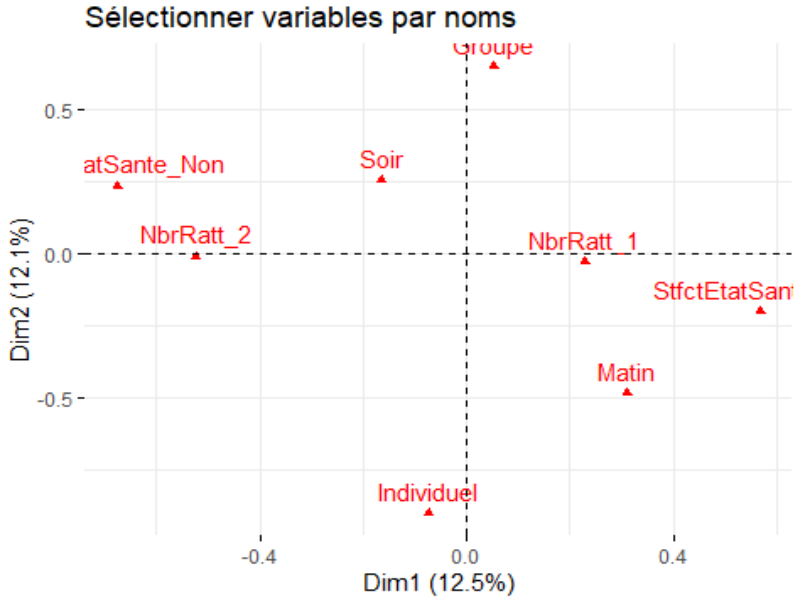


Figure 44: Sélection variables par noms

```
# Top 5 des categories de variables les plus contributifs
fviz_mca_biplot (res.ACM, select.ind = list (contrib = 5),
        select.var = list (contrib = 5),
        title="Top 5 des categories de variables les plus contributifs", shape=18,
        ggtheme = theme_minimal ())
```

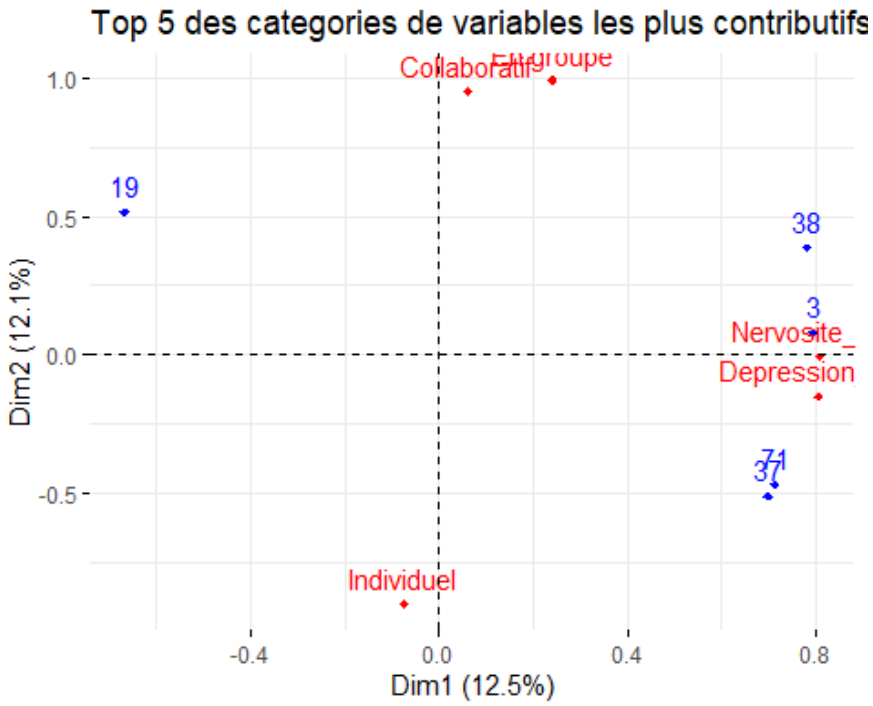


Figure 45: Top 5 des categories de variables les plus contributifs

Réalisation de l'ACM avec une interface graphique

Factoshiny permet d'améliorer facilement et de façon interactive les graphiques pour les rendre beaucoup plus lisibles. En plus de la génération automatique d'un rapport de l'analyse effectuée.

```
#install.packages("shiny")
#install.packages("Factoshiny")
library(Factoshiny)

Factoshiny(df_momtrv)
```

Prédiction et Machine Learning avec Python

Definition

Qu'est-ce qu'apprendre, comment apprend-on, et que cela signifie-t-il pour une machine ? La question de l'apprentissage fascine les spécialistes de l'informatique et des mathématiques tout autant que neurologues, pédagogues, philosophes ou artistes.

Une définition qui s'applique à un programme informatique comme à un robot, un animal de compagnie ou un être humain est celle proposée par Fabien Benureau (2015) : « L'apprentissage est une modification d'un comportement sur la base d'une expérience ». Dans le cas d'un programme informatique, on parle d'apprentissage automatique, ou machine learning, quand ce programme a la capacité d'apprendre sans être programmé. Cette définition est celle donnée par Arthur Samuel (1959). On peut ainsi opposer un programme classique, qui utilise une procédure et les données qu'il reçoit en entrée pour produire des réponses, à un programme d'apprentissage automatique, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

Exemple

Supposons que l'on veuille utiliser ces factures pour déterminer quels produits le client est le plus susceptible d'acheter dans un mois. Bien que cela soit vraisemblablement lié, nous n'avons manifestement pas toutes les informations nécessaires pour ce faire. Cependant, si nous disposons de l'historique d'achat d'un grand nombre d'individus, il devient possible d'utiliser un algorithme de machine learning pour qu'il en tire un modèle prédictif nous permettant d'apporter une réponse à notre question

Utilité de ML sur notre projet:

On a utilisé le machine Learning sur notre projet pour pouvoir prévoir le nombre de rattrapage d'un élevé en se basant sur son état de sante ,son mode de travail ,et les heures de préparation quotidienne, ce projet a pour but d 'améliorer les performances des étudiants pour avoir des résultats satisfaisants

Partie Pratique :

Entrée [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_excel("C:/Users/hp/Documents/MASTER M1 2022/S2/analyse de donnees/ACM.xlsx")
```

Entrée [2]:
```
df.head()
```

Out[2]:

	modeTravail	etatdepr	nbRatt	heureprepa
0	Groupe	Oui	5	1
1	Groupe	Non	0	3
2	Groupe	Non	0	2
3	Individuel	Oui	5	2
4	Groupe	Non	0	2

Entrée [3]:
```
#cleaning data
df = df[df["nbRatt"].notnull()]
df.head()
```

Out[3]:

	modeTravail	etatdepr	nbRatt	heureprepa
0	Groupe	Oui	5	1
1	Groupe	Non	0	3
2	Groupe	Non	0	2
3	Individuel	Oui	5	2
4	Groupe	Non	0	2

Entrée [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 99 entries, 0 to 98
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   modeTravail  99 non-null     object
 1   etatdepr     99 non-null     object
 2   nbRatt       99 non-null     int64
 3   heureprepa   99 non-null     int64
dtypes: int64(2), object(2)
memory usage: 3.9+ KB
```

Entrée [102]:
```
df['modeTravail'].value_counts()
```

Out[102]:
```
Groupe        63
Individuel    36
Name: modeTravail, dtype: int64
```

Entrée [103]:
```
#cette fonction nous permettons convertir la colonne donnee un dictionnaire
def shorten_categories(categories):
    categorical_map = {}
    for i in range(len(categories)):
            categorical_map[categories.index[i]] = categories.index[i]

    return categorical_map
```

Entrée [104]:
```
#appliquer sur la colonne etatdepr
country_map = shorten_categories(df.etatdepr.value_counts())
df['etatdepr'] = df['etatdepr'].map(country_map)
df.etatdepr.value_counts()
```

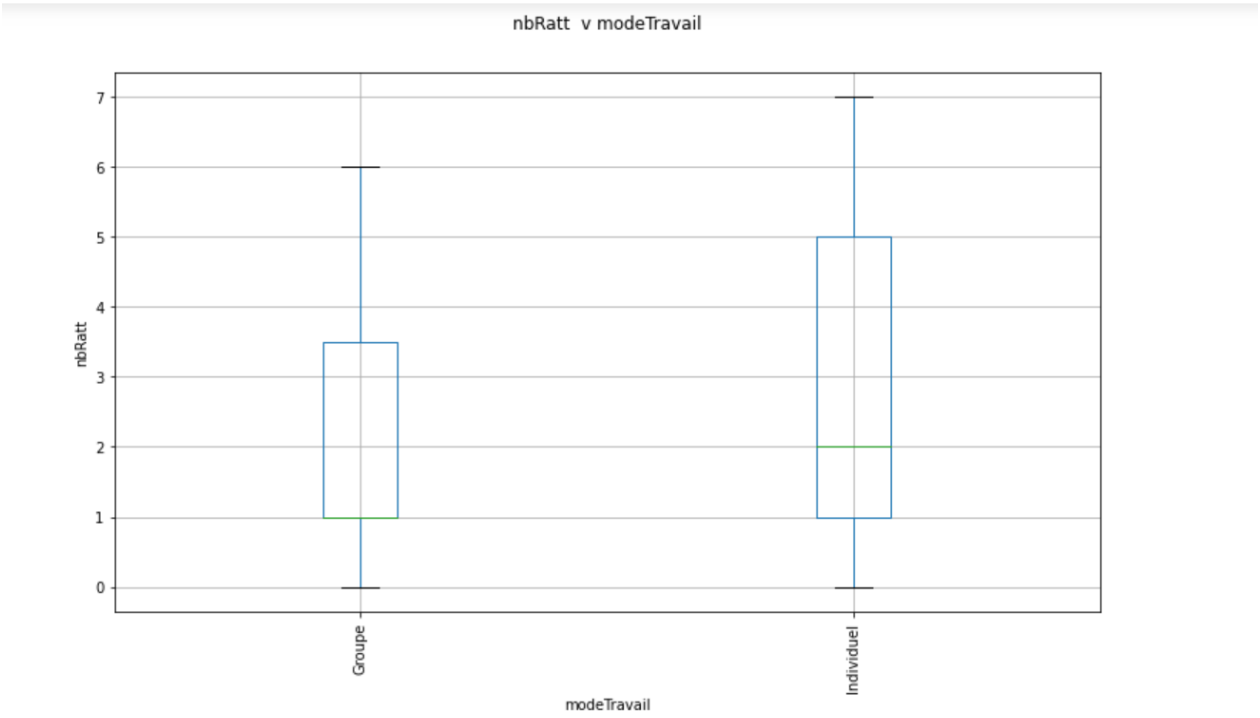
Out[104]:
```
Oui    62
Non    37
Name: etatdepr, dtype: int64
```

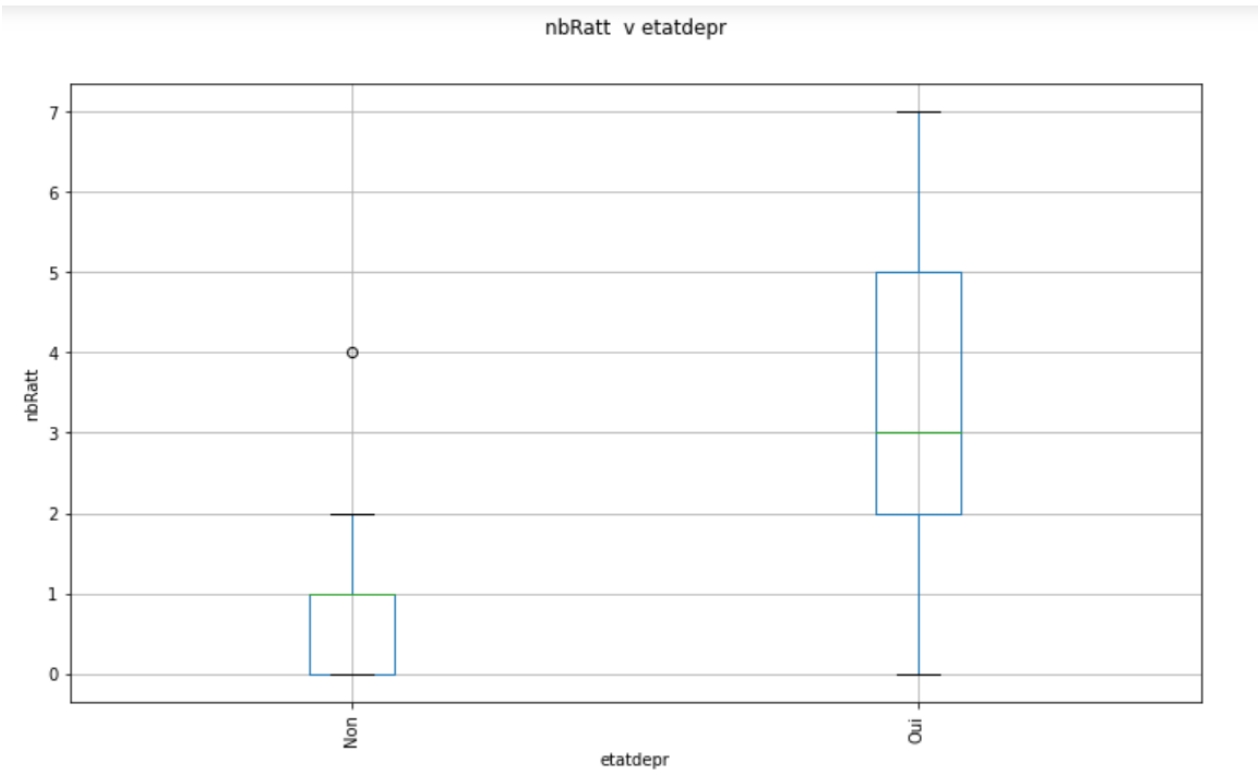
Entrée [105]:
```
country_map = shorten_categories(df.etatdepr.value_counts())
df['etatdepr'] = df['etatdepr'].map(country_map)
df.etatdepr.value_counts()
```

Out[105]:
```
Oui    62
Non    37
Name: etatdepr, dtype: int64
```

Entrée [106]:
```
#dessiner Histogramme
fig, ax = plt.subplots(1,1, figsize=(12, 7))
df.boxplot('nbRatt', 'etatdepr', ax=ax)
plt.suptitle('nbRatt  v etatdepr')
plt.title('')
plt.ylabel('nbRatt')
plt.xticks(rotation=90)
plt.show()
```

Entrée [107]:
```
fig, ax = plt.subplots(1,1, figsize=(12, 7))
df.boxplot('nbRatt', 'modeTravail', ax=ax)
plt.suptitle('nbRatt  v modeTravail')
plt.title('')
plt.ylabel('nbRatt')
plt.xticks(rotation=90)
plt.show()
```


nbRatt v modeTravail
nbRatt
0
1
2
3
4
5
6
7
Groupe
Individuel
modeTravail


nbRatt v etatdepr
nbRatt
0
1
2
3
4
5
6
7
Non
Oui
etatdepr

Entrée [108]:
```
df['heureprepa'].value_counts()
```

Out[108]:
```
2    38
1    31
3    20
5    10
Name: heureprepa, dtype: int64
```

Entrée [109]:
```
df = df[df["etatdepr"].notnull()]
df = df[df["modeTravail"].notnull()]
```

Entrée [110]:
```
df["etatdepr"].unique()
```

Out[110]:
```
array(['Oui', 'Non'], dtype=object)
```

Entrée [111]:
```
df["modeTravail"].unique()
```

Out[111]:
```
array(['Groupe', 'Individuel'], dtype=object)
```

Entrée [112]:
```
from sklearn.preprocessing import LabelEncoder
le_modeTravail = LabelEncoder()
df['modeTravail'] = le_modeTravail.fit_transform(df['modeTravail'])
le_etatdepr = LabelEncoder()
df['etatdepr'] = le_etatdepr.fit_transform(df['etatdepr'])
```

Entrée [113]:
```
df["modeTravail"].unique()
```

Out[113]:
```
array([0, 1])
```

Entrée [114]:
```
df["etatdepr"].unique()
```

Out[114]:
```
array([1, 0])
```

Entrée [115]:
```
X = df.drop("nbRatt", axis=1)
y = df["nbRatt"]
X
```

Out[115]:

	modeTravail	etatdepr	heureprepa
0	0	1	1
1	0	0	3
2	0	0	2
3	1	1	2
4	0	0	2
...	...	...	...
94	0	1	2
95	1	0	1
96	0	1	2
97	0	0	3
98	0	1	1

99 rows × 3 columns

Entrée [116]:
```
#algorithme de prediction
from sklearn.linear_model import LinearRegression
linear_reg = LinearRegression()
linear_reg.fit(X, y.values)
```

Out[116]: LinearRegression()

Entrée [117]:
```
y_pred = linear_reg.predict(X)
```

Entrée [118]:
```
from sklearn.metrics import mean_squared_error, mean_absolute_error
import numpy as np
error = np.sqrt(mean_squared_error(y, y_pred))
```

Entrée [119]:
```
error
```

Out[119]: 1.4022314886063865

Entrée [120]:
```
from sklearn.tree import DecisionTreeRegressor
dec_tree_reg = DecisionTreeRegressor(random_state=0)
dec_tree_reg.fit(X, y.values)
```

Out[120]: DecisionTreeRegressor(random_state=0)

Entrée [121]:
```
y_pred = dec_tree_reg.predict(X)
```

Entrée [122]:
```
#cette fonction dec_tree_reg a un erreur plus petit que les autres methodes
error = np.sqrt(mean_squared_error(y, y_pred))
print("{:,.20f}".format(error))
```

1.31242016485322099406

Entrée [123]:
```
from sklearn.ensemble import RandomForestRegressor
random_forest_reg = RandomForestRegressor(random_state=0)
random_forest_reg.fit(X, y.values)
```

Out[123]: RandomForestRegressor(random_state=0)

Entrée [124]:
```
y_pred = random_forest_reg.predict(X)
```

Entrée [125]:
```
error = np.sqrt(mean_squared_error(y, y_pred))
print("{:,.20f}".format(error))
```

1.31336019497702949366

from sklearn.model_selection import GridSearchCV ¶

max_depth = [None, 2,4,6,8,10,12] parameters = {"max_depth": max_depth}

regressor = DecisionTreeRegressor(random_state=0) gs = GridSearchCV(regressor, parameters, scoring='neg_mean_squared_error') gs.fit(X, y.values)

Entrée [126]:
```
regressor = gs.best_estimator_
regressor.fit(X, y.values)
y_pred = regressor.predict(X)
error = np.sqrt(mean_squared_error(y, y_pred))
print("{:,.20f}".format(error))
```

```
1.31430252506457856398
```

Entrée [127]:
```
X
```

Out[127]:

	modeTravail	etatdepr	heureprepa
0	0	1	1
1	0	0	3
2	0	0	2
3	1	1	2
4	0	0	2
...	...	...	...
94	0	1	2
95	1	0	1
96	0	1	2
97	0	0	3
98	0	1	1

Entrée [128]:
```
#faire un exemple de prediction
X = np.array([["Oui",'Groupe', 1]])
X
```

Out[128]:
```
array([['Oui', 'Groupe', '1']], dtype='<U11')
```

Entrée [129]:
```
X[:, 0] = le_etatdepr.transform(X[:,0])
X[:, 1] = le_modeTravail.transform(X[:,1])
X = X.astype(float)
X
```

Out[129]:
```
array([[1., 0., 1.]])
```

Entrée [136]:
```
y_pred = regressor.predict(X)
y_pred
```

Out[136]:
```
array([0.84615385])
```

Entrée [137]:
```
a=y_pred%1
a=int(a*100)
if a>50:
    y_pred=int(y_pred)+1
else:
    y_pred=int(y_pred)
y_pred
```

Out[137]: 1

Entrée [139]:
```
import pickle
```

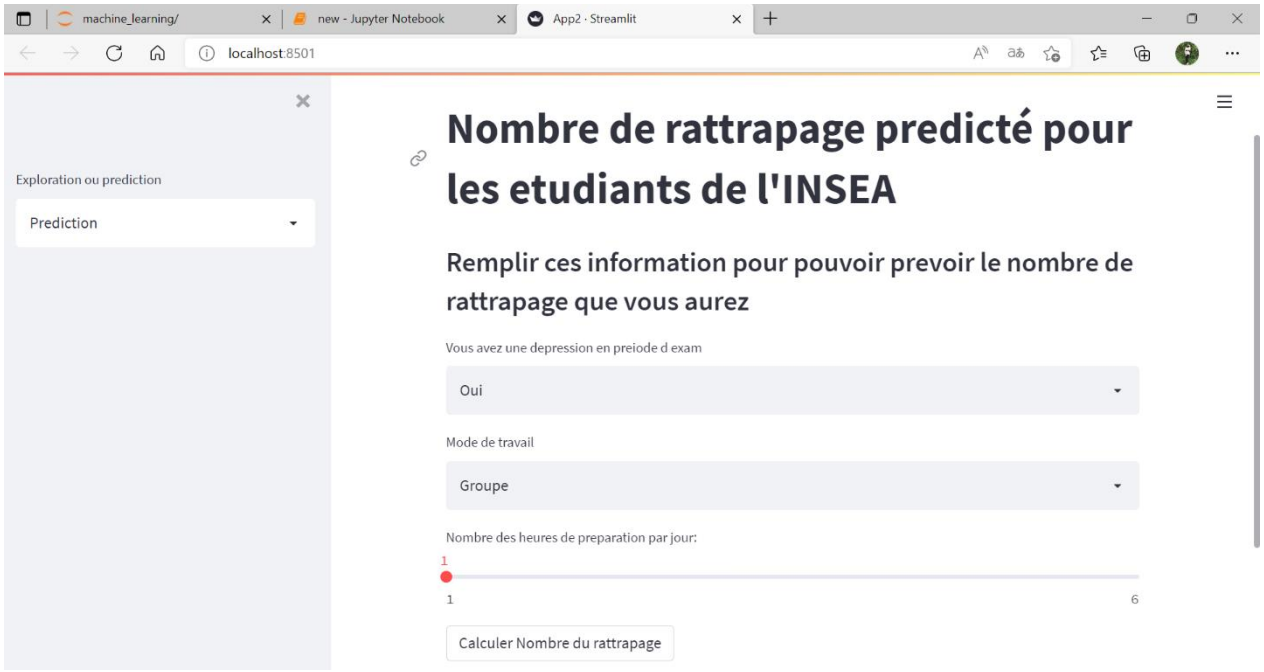
Entrée [140]:
```
#enregistrer ce qu on a fait dans un fichier
data = {"model": regressor, "le_modeTravail": le_modeTravail, "le_etatdepr": le_etatdepr}
with open('saved_steps2.pkl', 'wb') as file:
    pickle.dump(data, file)
```

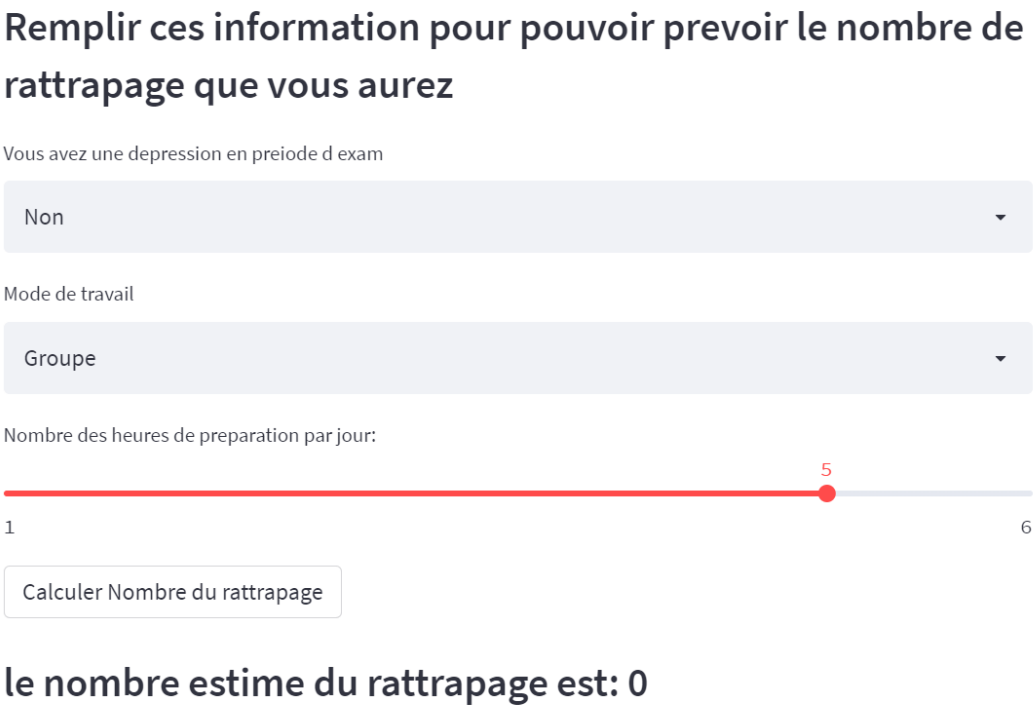
Entrée [141]:
```
with open('saved_steps2.pkl', 'rb') as file:
    data = pickle.load(file)

regressor_loaded = data["model"]
le_modeTravail = data["le_modeTravail"]
le_etatdepr = data["le_etatdepr"]
```

Entrée [142]:
```
y_pred = regressor_loaded.predict(X)
y_pred
```

Out[142]: array([0.84615385])

Interface Graphique :





Remplir ces information pour pouvoir prevoir le nombre de rattrapage que vous aurez

Vous avez une depression en preiode d exam

Oui

Mode de travail

Individuel

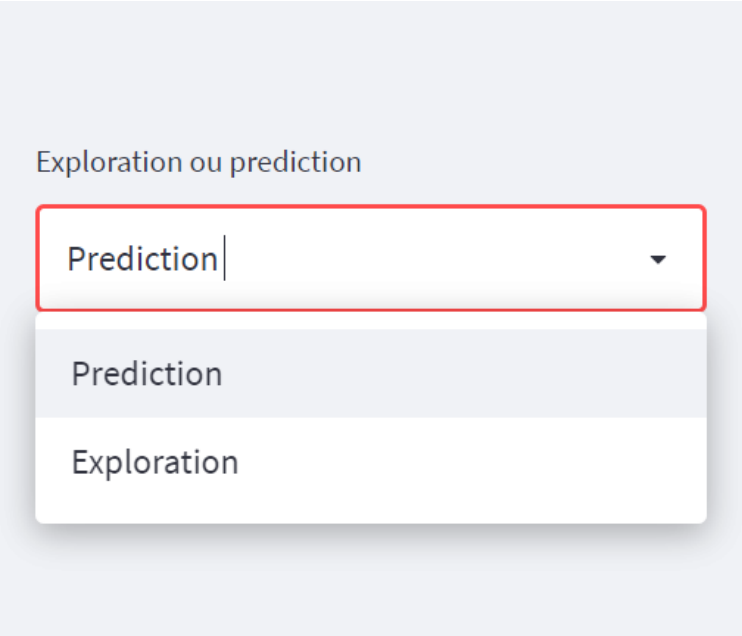
Nombre des heures de preparation par jour:

2

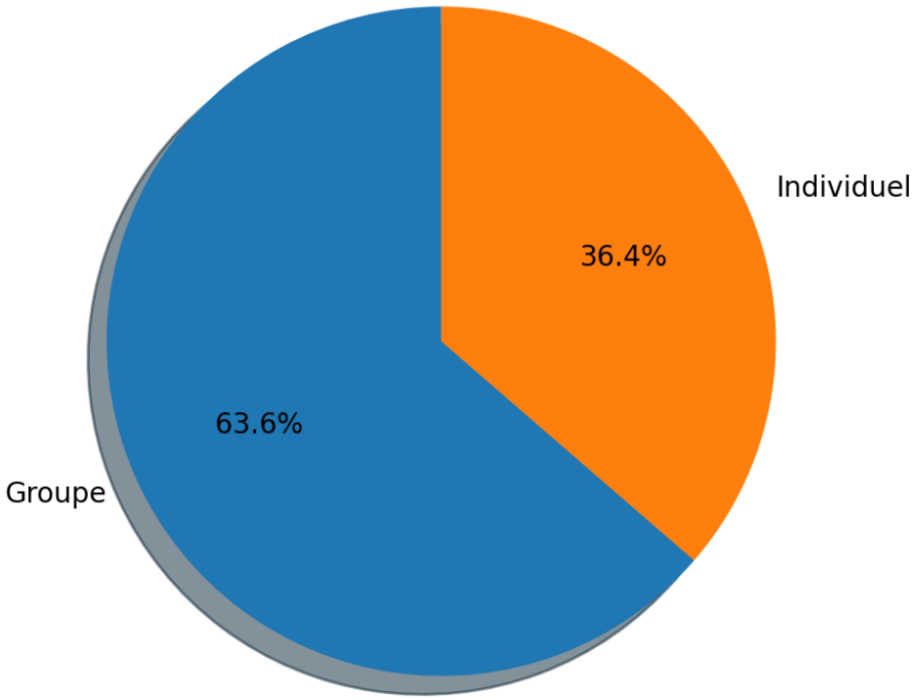
1 6

Calculer Nombre du rattrapage

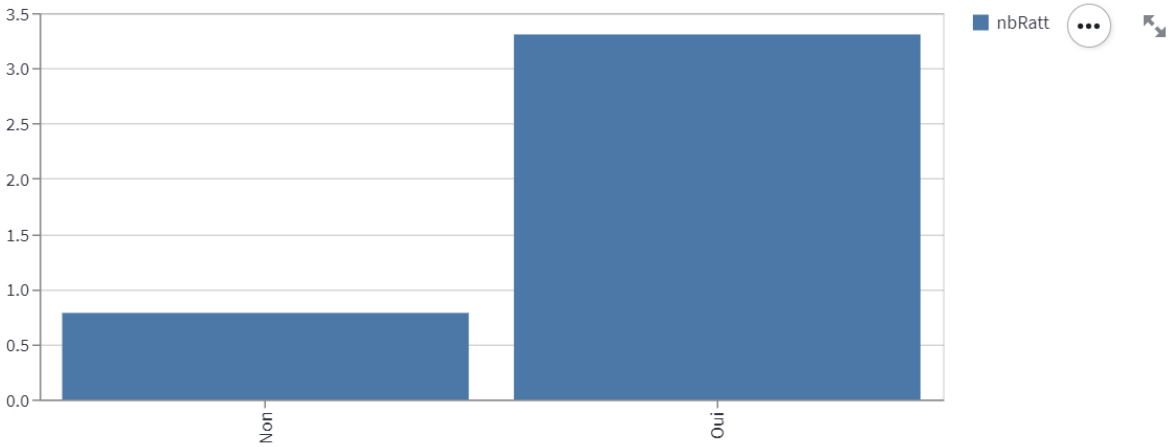
le nombre estime du rattrapage est: 4



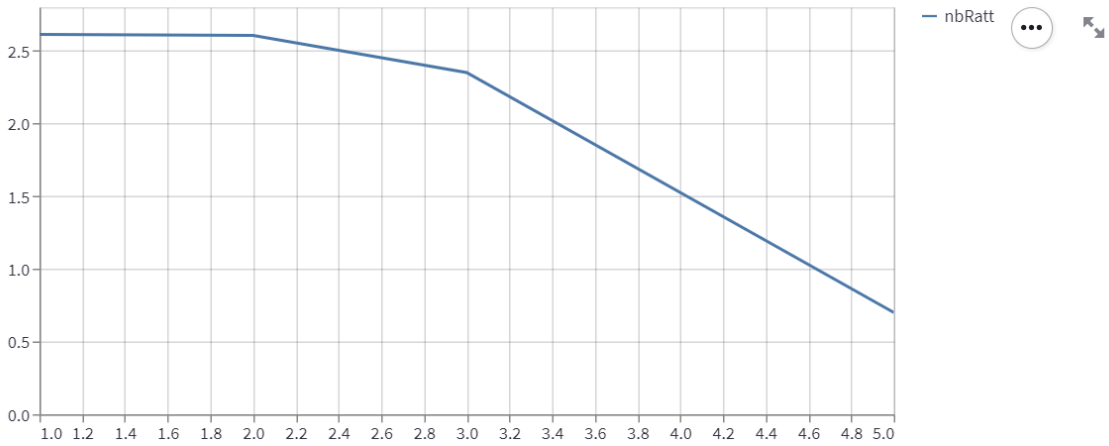
nombre de rattrapage a partir du mode de travail



nombre de rattrapage en se basant sur etat de depression



nombre de rattrapage en se basant sur les heures de preparation par jour



Conclusion

A travers cette étude nous avons constaté les résultats suivants:

- Les étudiants (surtout les hommes) de M2SI, sont les plus satisfaits de leur d'état de santé,(depression_non,nervosite_non, malAuDos_non..), et ils tendent à avoir un 1 seul rattrapage.
    - Leur personnalité est caractérisée d'extroverti, avec un style d'apprentissage plutôt compétitif et préfére étudier matin plutôt que soir.
- Les étudiants (surtout les femmes de DES) venant des CPGE ne sont pas satifaits de leur état de santé (depression_oui, nervosite_oui, malAuDos_oui). Ils tendent à avoir 2 rattrapages.
    - Leur personnalité est introverti, préfèrent étudier soir plutôt que matin et tendent de voir.

En plus, l'étude a montré aussi que:

- Les personnes a caractère introverti(renfermé) ont un style d'apprentissage indépendant, étudient individuellement et pendant la période d'examens aussi préparent individuellement.
- Les personnes a caractère extraverti(ouvert) ont un style d'apprentissage collaboratif, étudient en groupe et préparent les axamens aussi en groupe.

Ce travail peut encore être améliorer et élargir (par exemple utiliser des tests de personnalité validés, suivi des nombre d'heures étudiés par jour et par semaine de chaque étudiant…) pour découvrir les facteurs liés à la personnalité et aux habitudes de travail et leurs influences sur la performance académique et l'état de santé des étudiants, afin de booster notre système éducatif vue qu'il y a une forte corrélation entre l'état morale, les stratégies d 'études et le rendement scolaire des étudiants.

Une première étape d'amélioration c'était la réalisation un modèle de prédiction de nombre de rattrapage en se basant sur le Machine Learning.

Bibliographie

http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/75-acm-analyse-des-correspondances-multiples-avec-r-l-essentiel/

https://larmarange.github.io/analyse-R/analyse-des-correspondances-multiples.html

https://eric.univ-lyon2.fr/~ricco/cours/slides/ACM.pdf

http://factominer.free.fr/factomethods/analyse-des-correspondances-multiples.html

https://www.book.utilitr.org/acp.html