UNIVERSITÉ PARIS 1

PANTHÉON SORBONNE

Master 1 - Économétrie et Statistiques.

Analyses de Données et Machine Learning

Projet présenté à Mr J. Rynkiewicz

NORA AANKOUD, GHALIA JAZIRI , MÉLANIE DADDIO, AYA MOKHTAR.

I) Régression binaire

En choisissant l'accuracy comme critère de sélection des modèles, nous pouvons voir que la valeur d'accuracy la plus élevée est pour le modèle de Gradient Boosting avec le terme quadratique à chaque fois qu'on relance notre code.

```
Régression Logistique Précision : 0.69
Forêt aléatoire Précision : 0.77
Gradient Boosting Précision : 0.79
Arbre de décision Précision : 0.7525
Régression Logistique (Poly) Précision : 0.6775
Forêt aléatoire (Poly) Précision : 0.78
Gradient Boosting (Poly) Précision : 0.8075
Arbre de décision (Poly) Précision : 0.7425
```

```
Régression Logistique Précision : 0.69
Forêt aléatoire Précision : 0.7775
Gradient Boosting Précision : 0.79
Arbre de décision Précision : 0.745
Régression Logistique (Poly) Précision : 0.6775
Forêt aléatoire (Poly) Précision : 0.77
Gradient Boosting (Poly) Précision : 0.8075
Arbre de décision (Poly) Précision : 0.7275
```

```
Régression Logistique Précision : 0.69
Forêt aléatoire Précision : 0.775
Gradient Boosting Précision : 0.79
Arbre de décision Précision : 0.7425
Régression Logistique (Poly) Précision : 0.6775
Forêt aléatoire (Poly) Précision : 0.775
Gradient Boosting (Poly) Précision : 0.8075
Arbre de décision (Poly) Précision : 0.73
```

Nous retrouvons ce résultat dans le fichier texte généré par notre code.

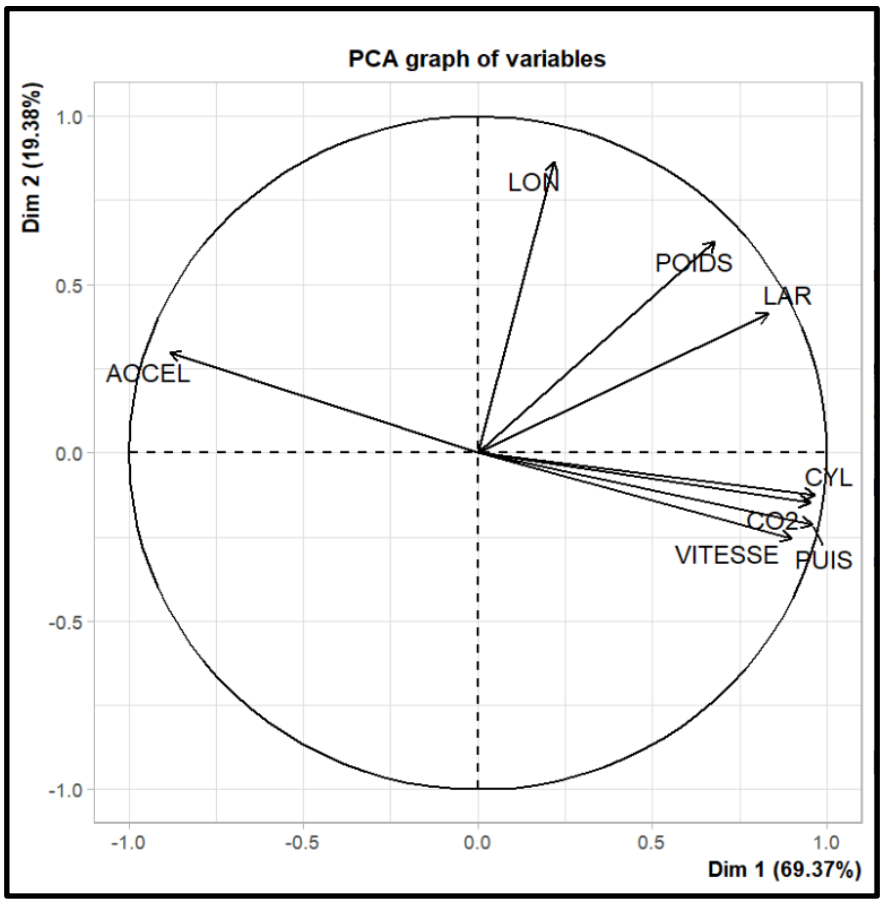
II) Analyse en composantes principales

1) Quel est le pourcentage d'inertie expliquée par les trois premiers facteurs ? Par le premier plan factoriel ?

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.549271024	69.36588781	69.36589
comp 2	1.550202725	19.37753406	88.74342
comp 3	0.480655817	6.00819771	94.75162
comp 4	0.280682369	3.50852962	98.26015
comp 5	0.084751837	1.05939796	99.31955
comp 6	0.034770185	0.43462731	99.75417
comp 7	0.013450458	0.16813073	99.92231
comp 8	0.006215585	0.07769482	100.00000

Le pourcentage d'inertie expliquée par les trois premiers facteurs est de 94.75%.
Le pourcentage d'inertie expliqué par le premier plan factoriel est de 88.74%.

2) Interpréter les deux axes principaux à partir des corrélations des variables avec ces axes.



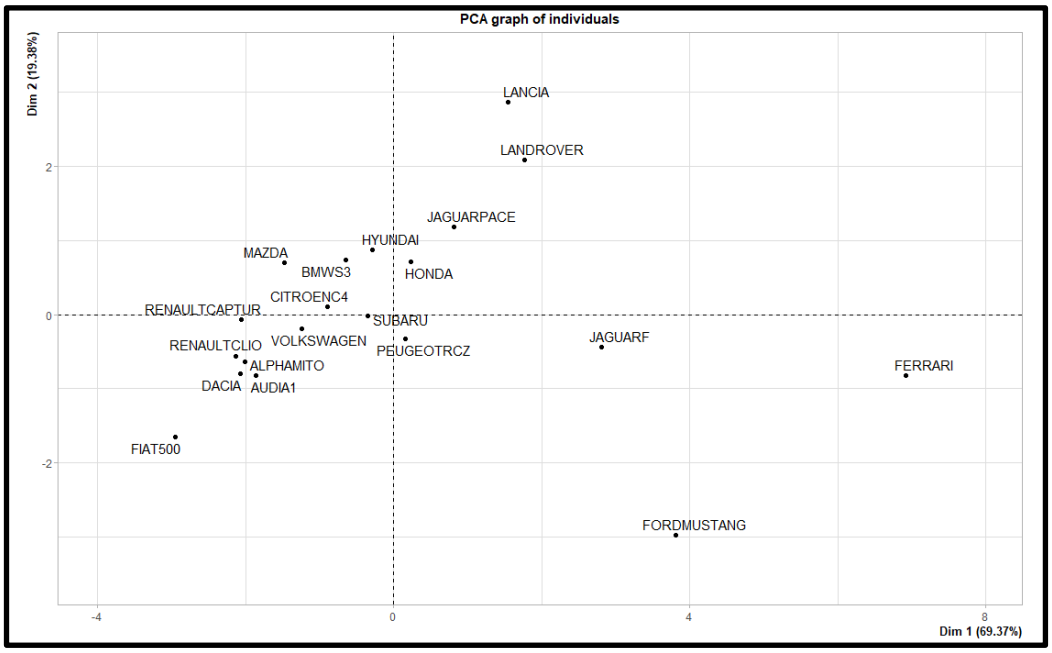
L’analyse de l’axe 1 du cercle de corrélation révèle que les variables CYL, CO2, PUISS et VITESSE sont positivement corrélées avec ce premier axe. On pourra ainsi parler de l’existence d’un effet taille associé à l'axe 1.

La variable “ACCEL” est corrélée à l’axe 1. Cependant, elle est opposée à la variable vitesse, elle est donc corrélée négativement à celle-ci. En effet, en supposant que la variable "accel" soit mesurée en termes de temps, sa corrélation négative avec le premier axe suggère que des valeurs plus petites de "accel" (temps plus court) sont préférables, ce qui signifie une accélération plus rapide. À l'inverse, des valeurs élevées de "accel" indiquent une accélération plus lente.

Le premier axe est donc expliqué par le cylindre, l'émission de CO2, la puissance, la vitesse et l’accélération.

Le second axe représente 19.38% de l’inertie expliquée par le jeu de données, on peut donc dire que la plus grande partie de l’information est représentée sur l’axe 1. Cependant, la variable “LON” semble être très corrélée à l’axe 2. Celui-ci semble donc bien représenter la variable de longueur. Cela suggère qu'il pourrait séparer les voitures en fonction de leur taille en longueur. Ainsi, le deuxième axe pourrait différencier les grandes voitures “en longueur” (en haut) des plus petites voitures (en bas).

3) Représentez les individus sur le premier plan factoriel et répondez aux questions suivantes :



(a) Les individus sont-ils bien représentés sur le premier plan factoriel ?

Pour savoir si les individus sont bien représentés, il faut regarder la somme des cosinus carrés pour les deux premiers axes, pour chaque individu. Si cette somme est supérieure à 0.5, l’individu est bien représenté.

	Dim.1	Dim.2
ALPHAMITO	0.88418157	0.0882481546
AUDIA1	0.82447953	0.1603240825
CITROENC4	0.63441116	0.0097161853
JAGUARF	0.87295591	0.0211260379
PEUGEOTRCZ	0.02354612	0.1090728501
LANDROVER	0.31747599	0.4453076078
RENAULTCLIO	0.91588474	0.0619189342
BMWS3	0.34739866	0.4385043460
DACIA	0.83548059	0.1227757130
HYUNDAI	0.07691485	0.7124226564
LANCIA	0.20567125	0.7152890894
RENAULTCAPTUR	0.91218646	0.0011120940
FORDMUSTANG	0.52868209	0.3233183930
FIAT500	0.70301792	0.2181054338
HONDA	0.05681670	0.5506712131
FERRARI	0.92712779	0.0130358167
SUBARU	0.26365614	0.0002406284
MAZDA	0.57219677	0.1274534969
VOLKSWAGEN	0.85640405	0.0198445442
JAGUARPACE	0.21914850	0.4760572094

Les résultats obtenus indiquent que la quasi-totalité des individus sont très bien représentés sur le premier plan factoriel, avec une somme des carrés des cosinus supérieure à 0.5 pour chaque individu. Cela suggère qu'il n'y a pas eu de problème d'écrasement lors de la projection pour ces individus, et que le premier plan factoriel offre une représentation significative. On peut néanmoins noter que SUBARU ($\Sigma\cos^2 \approx 0.26$) et PEUGEOT RCZ ($\Sigma\cos^2 \approx 0.13$) sont par contre mal représentés, ils ont donc été mal projetés.

(b) Quelles sont les caractéristiques des individus en haut du graphe ?

Les individus en haut du graphe sont ceux qui ont des valeurs élevées sur l'axe 2, donc des valeurs élevées pour la variable "LON".

Cependant, pour interpréter la variable, il est nécessaire de voir si elle est bien représentée par l'axe factoriel. Pour cela, on regarde la $\Sigma\cos^2$ pour les deux premiers axes:

	Dim.1	Dim.2
CYL	0.93322153	0.01583087
PUIS	0.91932896	0.04521963
LON	0.04765934	0.74936053
LAR	0.69297159	0.17038452
POIDS	0.46290063	0.39339274
VITESSE	0.80889208	0.06509541
ACCEL	0.77789077	0.08879599
CO2	0.90640613	0.02212303

La variable "LON" est plutôt bien représentée par l'axe factoriel, et d'autant plus sur l'axe 2. Donc, cela oppose des grandes voitures comme Land Rover ou Lancia à des voitures moins longues.

(c) Quelles sont les caractéristiques des individus à droite du graphe ?

À droite du graphique, on trouve les voitures qui ont des caractéristiques associées aux performances élevées. Cela inclut une cylindrée importante, une émission de CO2 élevée, une puissance élevée, une vitesse élevée et une accélération "courte".

Ces caractéristiques sont typiques des voitures de sport, qui privilégient la vitesse et la puissance comme Ferrari et JaguarF.

(d) Quelles sont les caractéristiques des individus en bas à gauche du graphe ?

Les voitures du côté gauche du premier axe ont des caractéristiques associées à des performances plus modérées. Elles ont une cylindrée plus faible, une émission de CO2 réduite, une puissance moins importante et une vitesse plus modérée. De plus, ces voitures ont une accélération plus longue. Ces caractéristiques sont souvent associées aux voitures classiques, conçues pour une utilisation quotidienne et une efficacité énergétique.

Donc les voitures qui sont en bas à gauche représentent les voitures de taille modeste, avec les caractéristiques citées ci-dessus.

(e) Peut-on dire que les individus PEUGEOTRCZ et JAGUARF ont un profil semblable ? Si oui, quel est-il ?

On ne peut pas dire que ces deux voitures sont similaires puisque la somme des cosinus carrés de PEUGEOTRCZ est plus petite que 0.5, on ne peut donc pas interpréter la proximité entre ces deux individus.

(f) Peut-on dire que les individus LANCIA et LANDROVER ont un profil semblable ? Si oui, quel est-il ?

LANCIA et LANDROVER peuvent être interprétées comme semblables puisque la somme carrée des cosinus est plus grande que 0.5. On pourra donc dire que les deux individus ont des caractéristiques semblables : ce sont des voitures de grande taille, mais qui ont une cylindrée, une émission de CO2, une puissance et une vitesse qui ont des valeurs élevées. On peut donc supposer qu'il s'agit de voitures plutôt type familial ou 4x4.

(g) Interpréter la représentation graphique des individus

À droite sur le premier axe, on aura les voitures ayant une cylindrée importante, une émission de CO2 élevée, une puissance élevée, une vitesse élevée et une valeur d'accélération plus petite (bonne accélération dans le temps). Ces caractéristiques suggèrent des véhicules plus puissants et performants tels que Ferrari, JaguarF et FordMustang qui représentent des voitures de sport.

Cependant, à gauche on aura les voitures ayant des valeurs plus faibles de cylindrée, émission de CO2, puissance, vitesse et une valeur d'accélération plus grande (moins bonne accélération dans le temps). Ces caractéristiques représentent les voitures classiques comme Renault Captur, Dacia, etc.

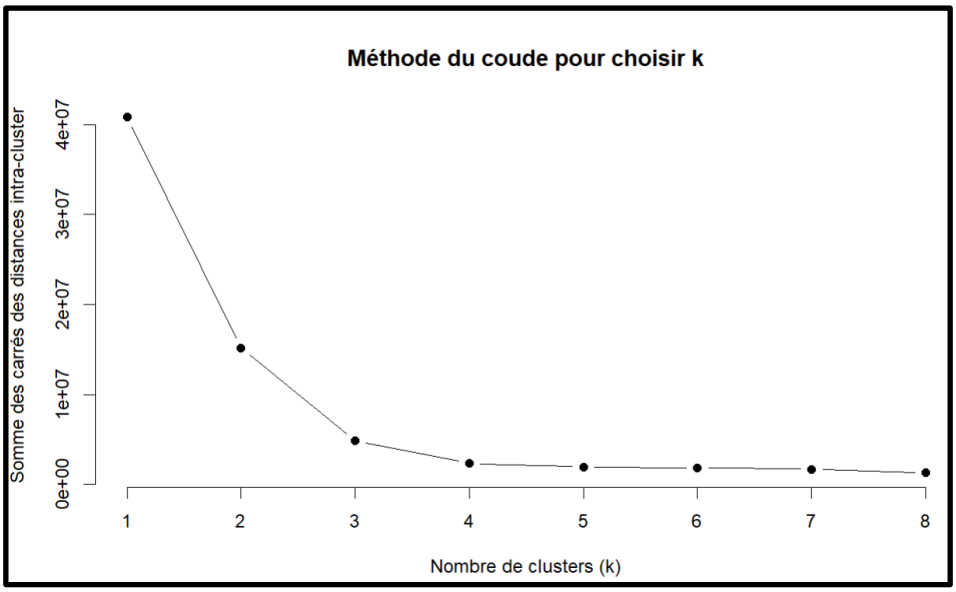
Le deuxième axe pourrait différencier les grandes voitures des plus petites. En haut, on trouve les modèles tel que Landrover et Lancia et en bas les modèles tel que Fiat500.

III) K-Means

1) Avec les données sur les voitures, réalisez une (ou des) classifications avec la méthode des K-mean et interprétez les résultats obtenus.

Pour déterminer le nombre optimal de clusters, nous avons utilisé la méthode du coude en combinaison avec l'interprétation visuelle de la partie 2 de notre analyse.

Suite au graphique résultant de la méthode du coude, nous avons émis l'hypothèse qu'il existe 3 ensembles distincts de clusters. Cependant, ce dernier illustre une stagnation significative de l'inertie après la création de 4 clusters. Ainsi, il est difficile de déterminer s'il existe 3 ou 4 clusters. Nous allons donc tester ces deux cas et voir lequel est le plus représentatif.



Nous pouvons voir ici les 3 clusters ainsi que les moyennes de leurs caractéristiques respectives :

```
K-means clustering with 3 clusters of sizes 9, 9, 2

Cluster means:
       CYL      PUIS      LON      LAR    POIDS  VITESSE     ACCEL      CO2
1 2327.444 184.66667 464.3333 187.6667 1778.889 205.8889  9.344444 160.0000
2 1129.667  98.77778 411.8889 174.3333 1147.111 180.6667 11.777778 111.3333
3 5606.500 540.50000 381.5000 193.5000 1800.000 292.5000  4.450000 339.5000

Clustering vector:
    ALPHAMITO        AUDIA1     CITROENC4       JAGUARF    PEUGEOTRCZ     LANDROVER   RENAULTCLIO
            2             2             2             1             1             1             2
       BMWS3         DACIA       HYUNDAI        LANCIA RENAULTCAPTUR   FORDMUSTANG       FIAT500
           1             2             1             1             2             3             2
       HONDA       FERRARI        SUBARU         MAZDA    VOLKSWAGEN    JAGUARPACE
           1             3             1             2             2             1

Within cluster sum of squares by cluster:
[1] 3016577.2  921087.3  931599.2
 (between_SS / total_SS =  88.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Le premier cluster regroupe des véhicules imposants, affichant une taille supérieure en moyenne par rapport aux autres clusters. Ces voitures se distinguent par leur longueur et leur largeur importantes. Malgré leur gabarit imposant, ces modèles présentent en termes d'émissions de CO2, affichant des niveaux plus bas par rapport au troisième cluster. De plus, il représente les voitures ayant le deuxième plus gros moteur en moyenne parmi les clusters étudiés. Ce regroupement englobe les SUV conçus pour une utilisation tout-terrain et classés dans les catégories utilitaires, loisir ou sport. On retrouve notamment les Subaru, les LandRover ou les JaguarF.

Le deuxième cluster rassemble des véhicules caractérisés par la plus petite cylindrée, la puissance la plus basse, le poids le plus léger, la vitesse la plus faible, ainsi que les émissions de CO2 les plus basses en moyenne parmi tous les groupes étudiés. Parmi les modèles inclus dans cette catégorie, on trouve l'Alfa Romeo Alphamito, l'Audi A1, la Citroën C4, la Renault Clio et la Fiat 500. Ces véhicules, tous destinés à une conduite en milieu urbain, sont compacts et économiques, correspondant ainsi à la catégorie des citadines.

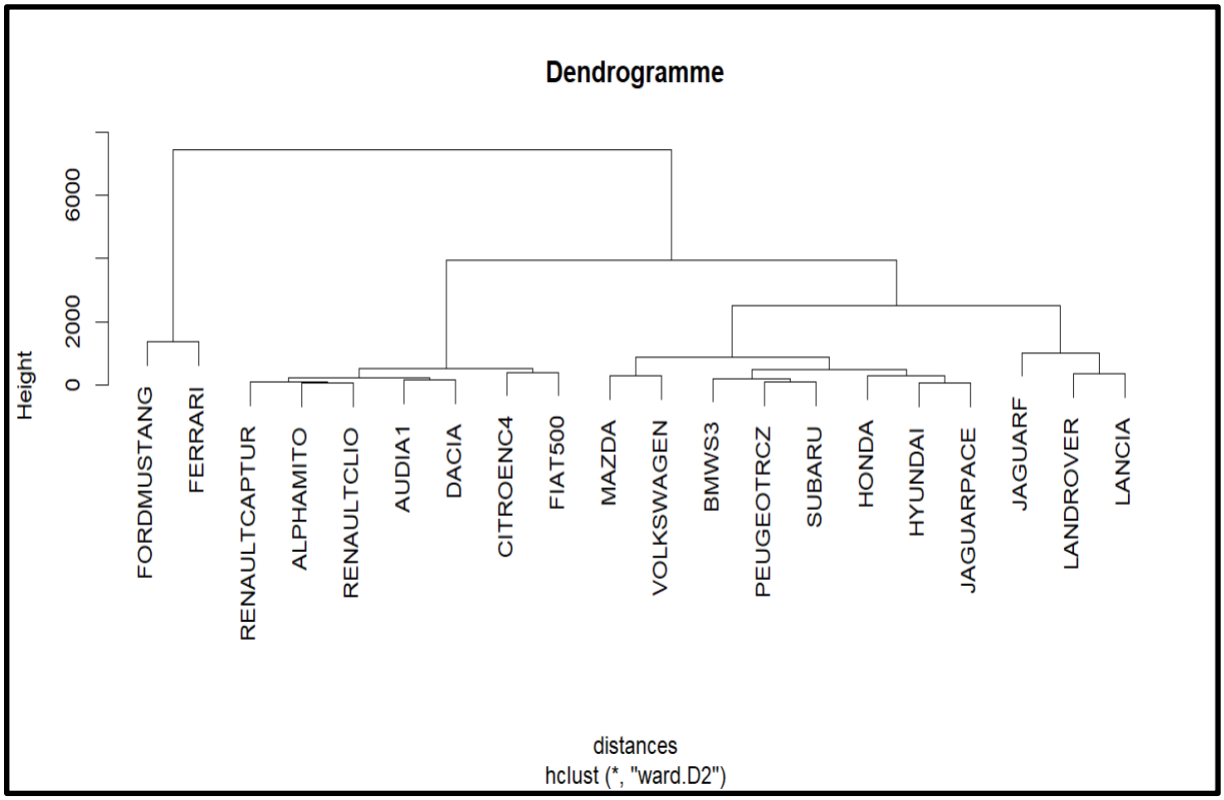
Le troisième cluster se caractérise par des valeurs moyennes plus élevées en termes de puissance, de vitesse, de cylindrée et d'émissions de CO2 par rapport aux autres regroupements. Les modèles de voitures inclus dans ce cluster représentent des marques de luxe telles que Ferrari et Ford Mustang, reconnues pour être à la fois des voitures sportives haut de gamme et donc des véhicules extrêmement puissants.

Maintenant, nous pouvons voir ici les 4 clusters ainsi que les moyennes de leurs caractéristiques respectives :

```
K-means clustering with 4 clusters of sizes 2, 7, 3, 8

Cluster means:
       CYL      PUIS      LON      LAR    POIDS  VITESSE     ACCEL      CO2
1 5606.500 540.50000 381.5000 193.5000 1800.000 292.5000  4.000000 339.5000
2 1001.286  95.57143 403.4286 173.5714 1087.714 179.1429 11.285714 109.2857
3 2921.333 257.66667 484.0000 194.3333 2157.333 211.0000  8.333333 214.6667
4 1917.625 138.62500 451.2500 182.5000 1531.000 199.0000  9.750000 129.1250

Clustering vector:
    ALPHAMITO        AUDIA1     CITROENC4       JAGUARF    PEUGEOTRCZ     LANDROVER
            2             2             2             3             4             3
  RENAULTCLIO         BMWS3         DACIA       HYUNDAI        LANCIA RENAULTCAPTUR
            2             4             2             4             3             2
  FORDMUSTANG       FIAT500         HONDA       FERRARI        SUBARU         MAZDA
            1             2             4             1             4             4
   VOLKSWAGEN    JAGUARPACE
            4             4

Within cluster sum of squares by cluster:
[1] 931599.0 249369.7 572542.0 603161.6
 (between_SS / total_SS =  94.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
```

En regardant les moyennes, nous constatons que le cluster 3 et 4 ont des caractéristiques moyennes assez proches. Ainsi, il nous a paru plus judicieux de ne garder que trois clusters.

2) **Faire une classification hiérarchique avec la méthode de “Ward”. Interpréter le dendrogramme. En combien de classes aurait-on envie de couper ce graphique ?**



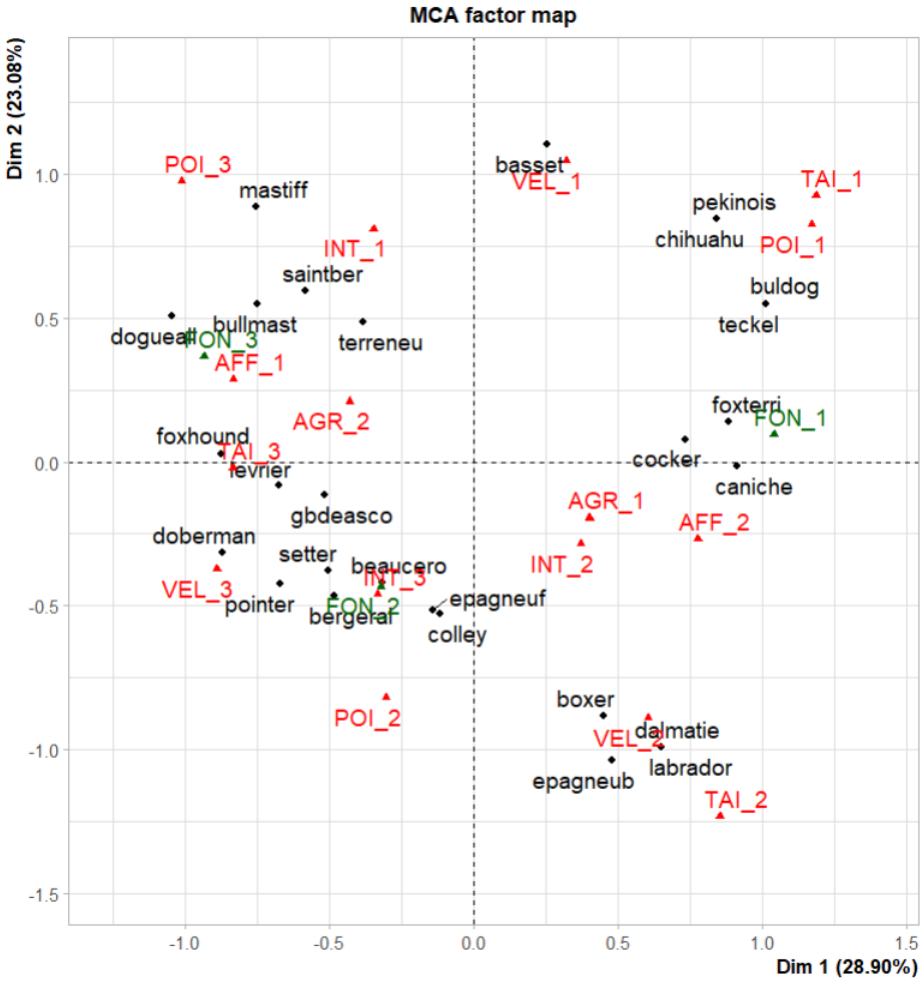
En regardant le dendrogramme, si on le découpe à 4000 height, nous formons 3 clusters comme vu précédemment. Si nous coupons à environ 2000 height, nous obtenons 4 clusters.

Ainsi, au vu de nos résultats pour les K-Means, nous pensons qu’il faut former 3 clusters : les voitures de luxe très puissantes, les citadines et les SUV tout-terrain.

IV) Race de chien

1) **En prenant la variable FON comme variable supplémentaire, faire une analyse des correspondances multiples de ces données.**

Représentation des individus et des variables sur le premier plan factoriel:



Les valeurs propres pour chaque dimension ;

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.481606165	28.896370	28.89637
dim 2	0.384737288	23.084237	51.98061
dim 3	0.210954049	12.657243	64.63785
dim 4	0.157554025	9.453242	74.09109
dim 5	0.150132670	9.007960	83.09905
dim 6	0.123295308	7.397718	90.49677
dim 7	0.081462460	4.887748	95.38452
dim 8	0.045669757	2.740185	98.12470
dim 9	0.023541911	1.412515	99.53722
dim 10	0.007713034	0.462782	100.00000

L'inertie expliquée par le premier plan factoriel est de 51.98%.

Cos2 pour chaque individu :

	Dim 1	Dim 2
beaucero	0.08863547	0.1536995944
basset	0.03380431	0.6348671357
bergeral	0.15372250	0.1401636585
boxer	0.11133075	0.4325235284
buldog	0.62448464	0.1838806326
bullmast	0.27069077	0.1429582059
caniche	0.38519392	0.0001212751
chihuahu	0.37993129	0.3826952203
cocker	0.27915682	0.0032460020
colley	0.01239617	0.2492609870
dalmatie	0.23628517	0.5530165596
doberman	0.48761169	0.0636477694
dogueall	0.56079391	0.1314738467
epagneub	0.10498339	0.4939526916
epagneuf	0.01753323	0.2221256292
foxhound	0.55831304	0.0004628928
foxterri	0.43627101	0.0108396312
gbdeasco	0.18602321	0.0089387139
labrador	0.23628517	0.5530165596
levrier	0.33881559	0.0051177295
mastiff	0.29999507	0.4136333336
pekinois	0.37993129	0.3826952203
pointer	0.29459212	0.1167506763
saintber	0.20156282	0.2087298540
setter	0.22389437	0.1252980645
teckel	0.62448464	0.1838806326
terreneu	0.08840069	0.1415315741

La somme des cos² sur le premier axe est supérieure à zéro pour les races surlignées sur le tableau ci-dessus. Cela représente la plupart des races de chiens, le reste des n'est donc pas très bien représenté sur le premier axe.

Cos2 pour chaque variable :

	Dim 1	Dim 2
TAI_1	0.49144201	0.2987546600
TAI_2	0.16462520	0.3448030588
TAI_3	0.87503205	0.0005293413
POI_1	0.57531341	0.2861238116
POI_2	0.10044717	0.7221387844
POI_3	0.23420393	0.2155641859
VEL_1	0.06021292	0.6422447857
VEL_2	0.15344741	0.3318791146
VEL_3	0.39792110	0.0691296921
INT_1	0.05129787	0.2752677726
INT_2	0.12673870	0.0756897524
INT_3	0.03207684	0.0603213262
AFF_1	0.64765585	0.0767360421
AFF_2	0.64765585	0.0767360421
AGR_1	0.17292377	0.0406368567
AGR_2	0.17292377	0.0406368567

La somme des cos² sur le premier axe est supérieure à zéro pour la moitié de nos variables catégorielles. Nous considérerons donc celles-ci comme celles qui sont bien représentées et donc interprétables.

Contributions des individus :

```
> chiens.mca$ind$contrib[,1:2]
             Dim 1        Dim 2
beaucero 0.7737679  1.679591265
basset   0.4965777 11.674160199
bergeral 1.8193797  2.076582082
boxer    1.5391043  7.484976084
buldog   7.8970523  2.910764017
bullmast 4.3555519  2.879430481
caniche  6.4006040  0.002522556
chihuahu 5.4366197  6.854956241
cocker   4.1352521  0.060190825
colley   0.1058588  2.664533548
dalmatie 3.2216221  9.438522677
doberman 5.8638360  0.958117236
dogueall 8.4304649  2.474089271
epagneub 1.7574399 10.350777943
epagneuf 0.1614884  2.560978462
foxhound 5.9090133  0.006132611
foxterri 5.9773564  0.185906708
gbdeasco 2.0582234  0.123802149
labrador 3.2216221  9.438522677
levrier  3.5214957  0.066583845
mastiff  4.3944998  7.584703763
pekinois 5.4366197  6.854956241
pointer  3.4866395  1.729709149
saintber 2.6172536  3.392717496
setter   1.9545479  1.369226351
teckel   7.8970523  2.910764017
terreneu 1.1310567  2.266782110
```

Bulldog, Caniche et Dogue Allemand sont les chiens qui contribuent le plus à l'axe 1. Pour l'axe 2, ce sont le Basset, Epagneub et Labrador.

Contributions des variables :

```
> chiens.mca$var$contrib[,1:2]
           Dim 1       Dim 2
TAI_1 12.5978150  9.58661729
TAI_2  4.6420727 12.17067028
TAI_3 13.4585463  0.01019149
POI_1 14.0104164  8.72224556
POI_2  1.6736860 15.06207234
POI_3  6.6040417  7.60886705
VEL_1  1.3119931 17.51742290
VEL_2  3.7368537 10.11705778
VEL_3  9.1804174  1.99644722
INT_1  1.2492400  8.39130827
INT_2  2.2742083  1.70014447
INT_3  0.8633836  2.03240794
AFF_1 11.6215827  1.72364624
AFF_2 10.7914697  1.60052866
AGR_1  2.8813167  0.84758676
AGR_2  3.1029565  0.91278575
```

L'axe 1 oppose les chiens de petite taille affectueux et ceux de grande taille moins affectueux. Et l'axe 2 est surtout représenté par la variable vélocité, donc l'axe 2 permet de distinguer les chiens selon leur vivacité.

```
Categorical variables (eta2)
            Dim.1 Dim.2 Dim.3
TAI       | 0.887 0.502 0.291 |
POI       | 0.644 0.725 0.342 |
VEL       | 0.411 0.684 0.291 |
INT       | 0.127 0.280 0.234 |
AFF       | 0.648 0.077 0.004 |
AGR       | 0.173 0.041 0.103 |
```

De plus, en regardant les coefficients de corrélation, on peut voir qu'effectivement les variables taille (TAI), poids (POI) et affectueux (AFF) sont très corrélés à l'axe 1 et les variables poids (POI) et (VEL) très corrélés à l'axe 2.

2) Quelles sont les caractéristiques des différentes races selon les deux axes ?

En haut à gauche, se trouvent les chiens de grande taille et lourds. Ces chiens sont peu affectueux. Si on regarde la variable supplémentaire, il se caractérise par FON 3, ce sont donc des chiens de garde. Le Mastiff et le Dogue Allemand par exemple étant reconnus pour être des chiens de garde.

En bas à gauche, le groupe de chiens est caractérisé surtout par une grande vélocité. Ce sont des chiens très vifs. Cependant, la plupart des races sont mal représentées par le premier plan factoriel.

En haut à droite, on retrouve des chiens petits, très affectueux. Ce sont des chiens de compagnie (FON 1) comme le Chihuahua, le Buldog et le Teckel.

En bas à droite, ce sont des chiens comme le Labrador ou le Dalmatien qui sont des chiens de taille moyenne, de vélocité moyenne et plus affectueux que ceux en haut à gauche.

<u>En conclusion :</u>

Les 3 fonctions qui se distinguent sur le graphique sont les chiens de garde en haut à gauche, les chiens de compagnie en haut à droite et les chiens de chasse en bas. En réalité, en bas, les chiens de chasse sont divisés en deux sous-catégories : les chiens les plus vifs (à gauche) et les moins vifs (à droite).