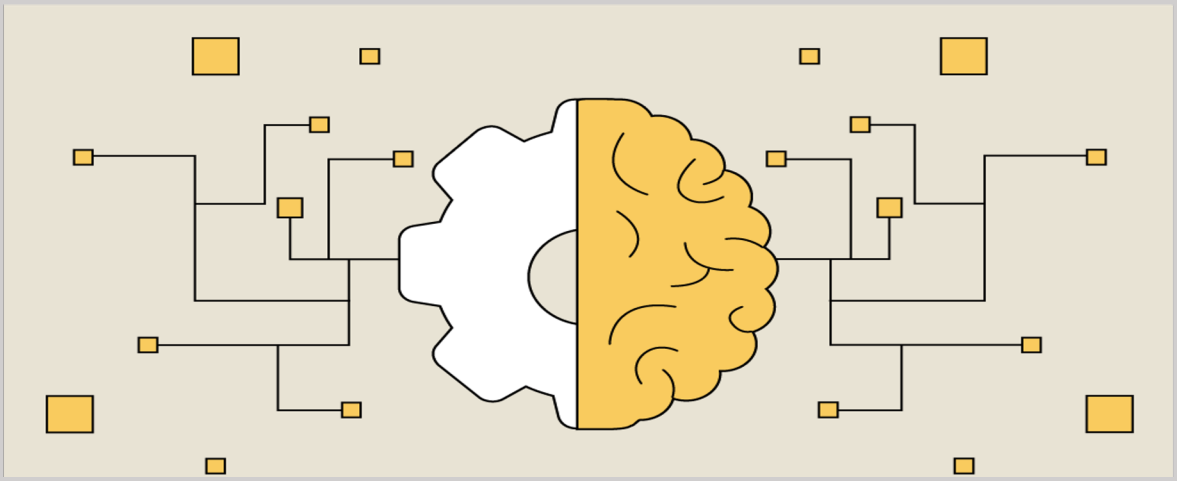
UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Rapport de projet

14/01/2024



Machine Learning et Analyse de données

Analyse de données : ACP et ACM

ABDELJALIL Issame | COLIN Léo| MARZOUK Moustafa | SOPGUOMBUE Brice

Université Paris 1 Panthéon-Sorbonne

Master 1 Econométrie-Statistiques

Année 2023-2024

Sommaire

I / Analyse en composantes principales de voitures

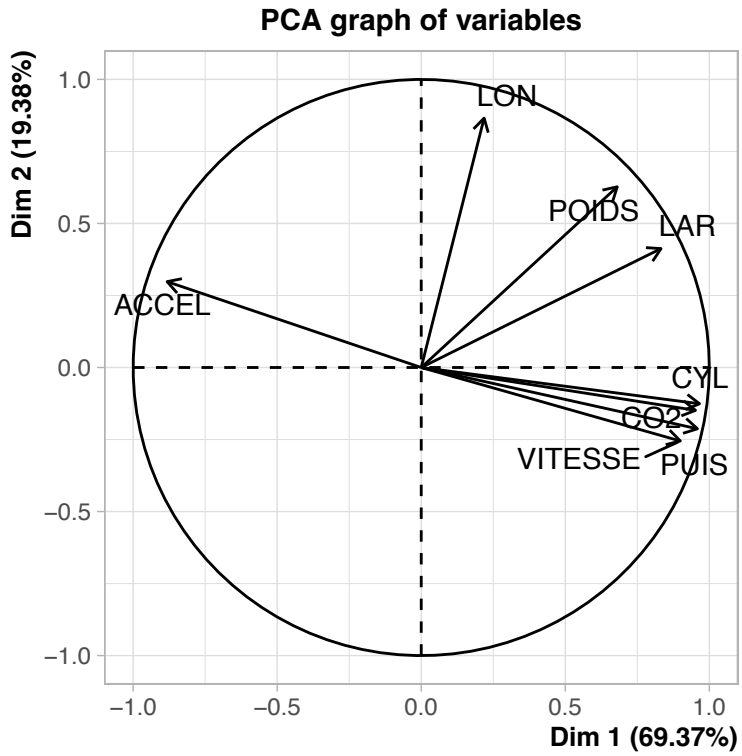
Question 1) Quel est le pourcentage d'inertie expliquée par les trois premiers facteurs? Par le premier plan factoriel?

```
Call:
PCA(X = donnees_centrees_reduites)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance               5.549   1.550   0.481   0.281   0.085   0.035   0.013
% of var.             69.366  19.378   6.008   3.509   1.059   0.435   0.168
Cumulative % of var.  69.366  88.743  94.752  98.260  99.320  99.754  99.922
                       Dim.8
Variance               0.006
% of var.              0.078
Cumulative % of var. 100.000
```

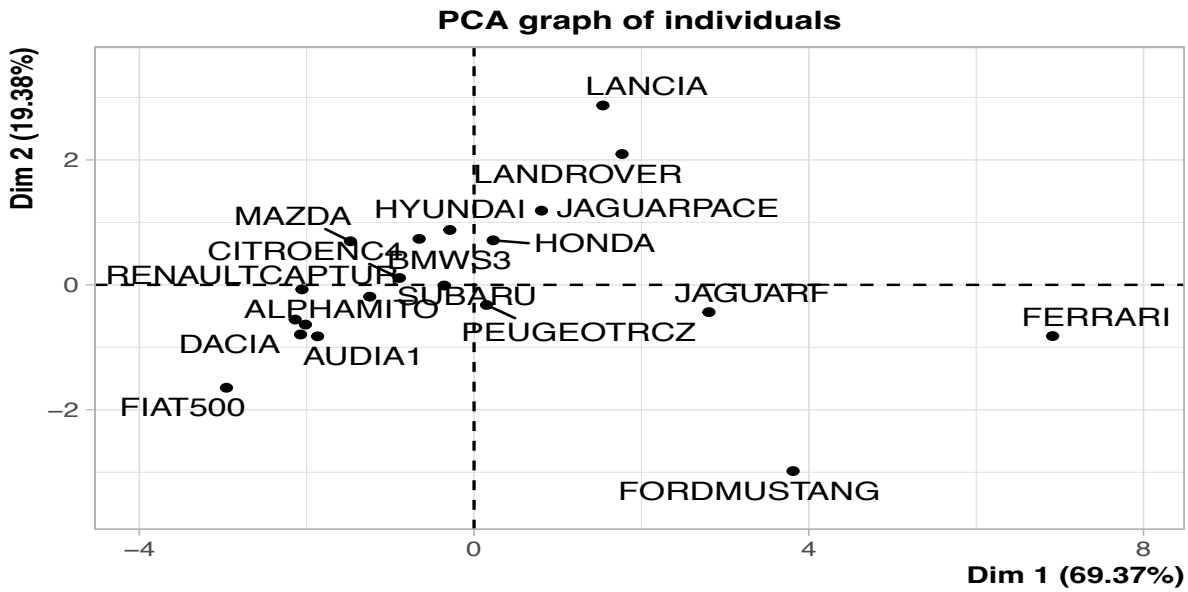
Le pourcentage d'inertie expliquée par les 3 premiers facteurs est d'environ 95%. Le pourcentage d'inertie expliquée par le premier plan factoriel est d'environ 89%.

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



Sur l'axe 1 on retrouve la performance du véhicule. La variable "accélération" est opposée aux autres variables de performance car plus une voiture accélère rapidement moins la valeur associée dans la base de données est grande (temps pour faire 0-100km/h).
Sur l'axe 2 on retrouve les caractéristiques physiques, d'apparence, de design des voitures, …

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




```
> res.pca$ind$cos2[,1:2]
                   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
```

Question 3a) Les *individus sont-ils bien représentés sur le premier plan factoriel ?*

Oui plutôt car les individus sont bien répartis sur le premier plan factoriel. Ils sont aussi bien le long du premier axe factoriel qui discrimine au mieux les observations que le long du second axe. Cependant il y aussi quelques individus qui sont mal représentés car la somme de leurs cosinus carré sur les deux premiers axes est faible. Par exemple l'individu Subaru a une somme de cosinus de 0,2637 + 0,0002= 0,2639 ce qui est inférieur à 0,5 donc plutôt faible.

Question 3b) Quelles sont les caractéristiques des individus en haut du graphe ?

Ce sont des voitures longues, larges et lourdes type SUV haut de gamme.

Question 3c) Quelles sont les caractéristiques des individus à droite du graphe ?

Ce sont les voitures très performantes, les voitures de sport (voitures rapides, puissantes, qui accélèrent vite, consomment beaucoup et ont beaucoup de cylindres)

Question 3d) Quelles sont les caractéristiques des individus en bas à gauche du graphe ?

Ce sont les voitures petites, légères et qui sont peu performantes type citadine

Question 3e) Peut-on dire que les individus PEUGEOTRCZ et JAGUARF ont un profil semblable ? Si oui quel est-il ?

PEUGEOTRCZ est très mal représenté sur le plan factoriel car la somme des cosinus carré sur les deux premières dimensions est largement inférieure à 0.5. En conséquence, il n'est pas possible de conclure sur la proximité entre ces deux voitures à partir du plan factoriel.

Question 3f) Peut-on dire que les individus LANCIA et LANDROVER ont un profil semblable ? Si oui quel est-il ?

D'après les cosinus carré des deux premiers axes pour ces deux individus, ces deux voitures sont bien représentés sur le premier plan factoriel et on peut donc conclure sur leur proximité. Les individus LANCIA et LANDROVER sont proches sur le plan factoriel et on conclue donc qu'ils ont des profils semblables. Ce sont des voitures longues, larges et lourdes type SUV haut de gamme.

Question 3g) Interpréter la représentation graphique des individus.

Le graphique oppose plusieurs types de véhicules. Il y a les voitures de sport à droite, les SUV haut de gamme en haut, les petites citadines légères en bas à gauche, les voitures pour plus "grand public" au centre/gauche qui sont peu performantes. Ces dernières sont des voitures avec des caractéristiques plus générales, de basse/moyenne gamme. Vers le centre/centre-droite ce sont des voitures un peu plus performantes avec des attributs néanmoins courants (dimensions, poids..).

II / Analyse des correspondances multiples de races des chiens

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

```
Call:
MCA(X = chiens, quali.sup = 7, graph = TRUE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance               0.482   0.385   0.211   0.158   0.150   0.123   0.081
% of var.             28.896  23.084  12.657   9.453   9.008   7.398   4.888
Cumulative % of var.  28.896  51.981  64.638  74.091  83.099  90.497  95.385
                       Dim.8   Dim.9  Dim.10
Variance               0.046   0.024   0.008
% of var.              2.740   1.413   0.463
Cumulative % of var.  98.125  99.537 100.000
```

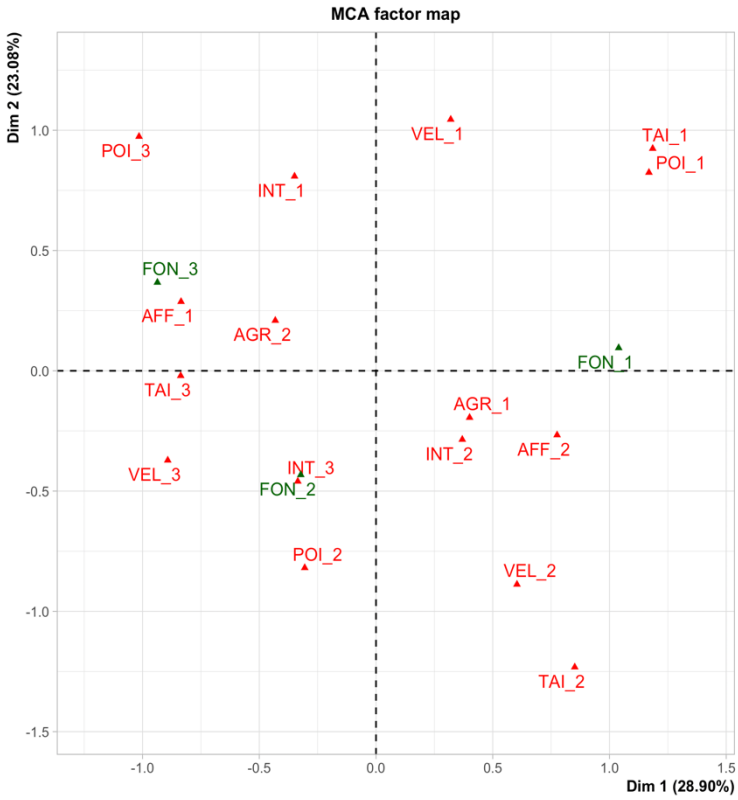
Le pourcentage d'inertie expliquée par le premier plan factoriel est d'environ 52%.
Le premier axe factoriel semble opposer les chiens de par leurs poids, tailles, affectuosité ou encore agressivité. Le second axe factoriel est plus compliqué à interpréter car on ne retrouve pas de tendance explicite qui se dégage.

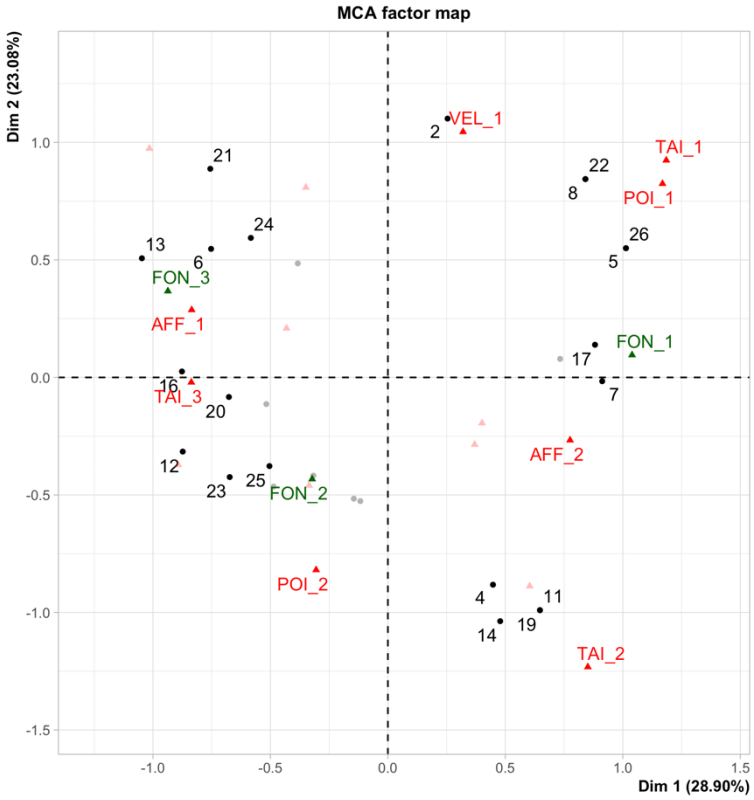
Question 2) En déduire une description des différentes races de chiens

Au milieu à gauche/haut à gauche on retrouve les gros chiens de garde (lourds, gros, peu affectueux, peu intelligents et assez agressifs). On peut citer notamment les mastiffs, les saint-bernards ou les dobermans.
En bas à droite on retrouve un groupe de chiens avec des caractéristiques homogènes (de tailles, vélocités et poids moyens) et qui sont des chiens de chasse ou de compagnie (labrador, boxer...).

Au milieu à droite/haut à droite ce sont plutôt les chiens de compagnie (petits, légers, assez affectueux et peu véloces) comme les caniches, chihuahuas ou teckels.

Au milieu à gauche/bas à gauche les chiens très véloces, intelligents, de poids moyen, plutôt des chiens de chasse (caractéristiques adaptées à la chasse) comme les pointer ou les lévriers.





III / Codes

Exercice 2

ACP.R

```r
library(FactoMineR)
library(MASS)


chemin_fichier <- "../donnees/voitures"
voitures <- read.table(chemin_fichier, header=T)
print(voitures)

donnees_centrees_reduites <- scale(voitures)
print(donnees_centrees_reduites)

res.pca <- PCA(donnees_centrees_reduites)
summary(res.pca)

plot(res.pca,choix="ind", cex=0.7)
plot(res.pca,choix="var", cex=0.7)

res.pca$ind$cos2[,1:2]
```

Exercice 4

ACM.R

```r
library(FactoMineR)
library(MASS)
library(dplyr)

chemin_fichier <- "../donnees/chiens"
chiens <- read.table(chemin_fichier, header=T)
print(chiens)

chiens <- data.frame(lapply(chiens, factor))
chiens.mca <- MCA(chiens, graph = TRUE, quali.sup=7)
plot(chiens.mca)
plot(chiens.mca,select="cos2 20", selectMod="cos2 8")
summary(chiens.mca, nbelements = Inf)

chiens <- read.table(chemin_fichier, header=T)
chiens$Indice <- seq_len(nrow(chiens))
print(chiens)
```