Mamadou NIANG
Dioncounda DIARRA

Projet d'Analyse de Données

1. Enquête budget des familles

1- Déterminer le meilleur modèle pour l'emploi d'un(e) employé(e) de maison, en considérant les variables explicatives possibles

```
mod0<-glm(qualDOMTRAV~qualTypmen2+qualCC+REVTOT+qualDIPLOPR+qualDIPLOCJ,family=quasibinomial(logit),datafr,weights = COEF)

mod1<-glm(qualDOMTRAV~qualTypmen2+qualCC*REVTOT+qualDIPLOPR+qualDIPLOCJ,family=quasibinomial(logit),datafr,weights = COEF)

mod2<-glm(qualDOMTRAV~qualTypmen2*qualDIPLOPR+qualCC+REVTOT+qualDIPLOCJ,family=quasibinomial(logit),datafr,weights = COEF)

mod3<-glm(qualDOMTRAV~qualCC+qualTypmen2*REVTOT+qualDIPLOPR+qualDIPLOCJ,family=quasibinomial(logit),datafr,weights = COEF)

mod4<-glm(qualDOMTRAV~qualTypmen2+qualCC+REVTOT*(qualDIPLOPR+qualDIPLOCJ),family=quasibinomial(logit),datafr,weights = COEF)

mod5<-glm(qualDOMTRAV~qualTypmen2*qualCC*REVTOT+qualDIPLOPR+qualDIPLOCJ,family=quasibinomial(logit),datafr,weights = COEF)
```

```
> X2test1<-deviance(mod0)-deviance(mod1)
> proba<-1-pchisq(X2test1,df.residual(mod0)-df.residual(mod1))
> cat("\n Test du rapport de vraisemblance :\n",X2test1," p-value : ",proba,"\n")

 Test du rapport de vraisemblance :
 33.70077  p-value :  0.02822154
> X2test2<-deviance(mod2)-deviance(mod1)
> proba<-1-pchisq(X2test2,df.residual(mod2)-df.residual(mod1))
> cat("\n Test du rapport de vraisemblance :\n",X2test2," p-value : ",proba,"\n")

 Test du rapport de vraisemblance :
 30.07208  p-value :  0.01166422
```

```
> X2test3<-deviance(mod3)-deviance(mod1)
> proba<-1-pchisq(X2test3,df.residual(mod3)-df.residual(mod1))
> cat("\n Test du rapport de vraisemblance :\n",X2test3," p-value : ",proba,"\n")

 Test du rapport de vraisemblance :
 19.2639  p-value :  0.255216
> X2test4<-deviance(mod3)-deviance(mod4)
> proba<-1-pchisq(X2test4,df.residual(mod3)-df.residual(mod4))
> cat("\n Test du rapport de vraisemblance :\n",X2test4," p-value : ",proba,"\n")

 Test du rapport de vraisemblance :
 12.40919  p-value :  0.2586035
> X2test5<-deviance(mod3)-deviance(mod5)
> proba<-1-pchisq(X2test5,df.residual(mod3)-df.residual(mod5))
> cat("\n Test du rapport de vraisemblance :\n",X2test5," p-value : ",proba,"\n")

 Test du rapport de vraisemblance :
 53.47308  p-value :  0.154995
```

On réalise plusieurs régressions logistiques en tenant compte de la dépendance ou non de certaines variables :

- Pour notre premier modèle, on considère la dépendance entre le degré d'urbanisation et le revenu.
- Pour notre second modèle, une dépendance entre la taille de la famille et les diplômes.
- Pour notre troisième modèle, une dépendance entre le revenu et la taille de la famille.
- Pour notre quatrième modèle, une dépendance entre le revenu et les diplômes.
- Pour notre cinquième modèle, une dépense entre le revenu, la taille de la famille et le degré d'urbanisation.

Au seuil de 5%, le modèle 3 est systématiquement non rejeté après des tests de rapport de vraisemblance.
On en conclu que c'est le modèle 3 qui explique au mieux l'emploi d'un(e) employé(e) de maison.

2- En déduire une description des ménages qui emploie un ou une employée de maison

summary(mod3)

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -3.723e+00  5.628e-01  -6.616 5.44e-11 ***
qualCC1            -6.231e-01  3.892e-01  -1.601  0.10962
qualCC2            -9.814e-01  4.709e-01  -2.084  0.03735 *
qualCC3             4.807e-02  2.909e-01   0.165  0.86881
qualCC4            -1.000e+00  4.208e-01  -2.377  0.01762 *
qualCC5            -2.069e+00  8.252e-01  -2.508  0.01228 *
qualTypmen23       -6.051e-01  6.604e-01  -0.916  0.35969
qualTypmen24       -2.829e+00  8.780e-01  -3.222  0.00130 **
qualTypmen25       -8.217e-01  7.233e-01  -1.136  0.25619
qualTypmen27       -6.747e-01  9.822e-01  -0.687  0.49224
REVTOT              3.168e-06  1.116e-06   2.840  0.00459 **
qualDIPLOPR1        1.202e+00  5.635e-01   2.134  0.03305 *
qualDIPLOPR2        8.666e-01  5.607e-01   1.546  0.12246
qualDIPLOPR3        3.011e-01  7.715e-01   0.390  0.69635
qualDIPLOPR4        1.022e+00  6.933e-01   1.474  0.14077
qualDIPLOPR5        1.595e+00  6.334e-01   2.519  0.01191 *
qualDIPLOPR6        1.649e+00  6.497e-01   2.538  0.01127 *
qualDIPLOPR7        1.725e+00  6.081e-01   2.836  0.00464 **
qualDIPLOCJ1       -3.007e-01  4.530e-01  -0.664  0.50687
qualDIPLOCJ2       -5.565e-01  5.318e-01  -1.046  0.29560
qualDIPLOCJ3       -1.069e+00  5.943e-01  -1.799  0.07227 .
qualDIPLOCJ4        3.860e-01  6.664e-01   0.579  0.56253
qualDIPLOCJ5        5.566e-01  4.799e-01   1.160  0.24637
qualDIPLOCJ6       -4.131e-02  5.373e-01  -0.077  0.93873
qualDIPLOCJ7        6.475e-01  5.237e-01   1.236  0.21658
qualTypmen23:REVTOT -2.197e-06  1.926e-06  -1.141  0.25402
qualTypmen24:REVTOT  7.878e-06  2.635e-06   2.990  0.00284 **
qualTypmen25:REVTOT  3.272e-06  2.270e-06   1.442  0.14960
qualTypmen27:REVTOT -6.222e-07  2.829e-06  -0.220  0.82598
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les ménages employant un(e) employé(e) de maison sont les ménages diplômés avec un revenu élevé, habitant dans des zones très urbanisés et ayant deux enfants.

2. Températures et précipitations

1- Qu'elle est le pourcentage d'inertie expliquée par le premier plan factoriel ?

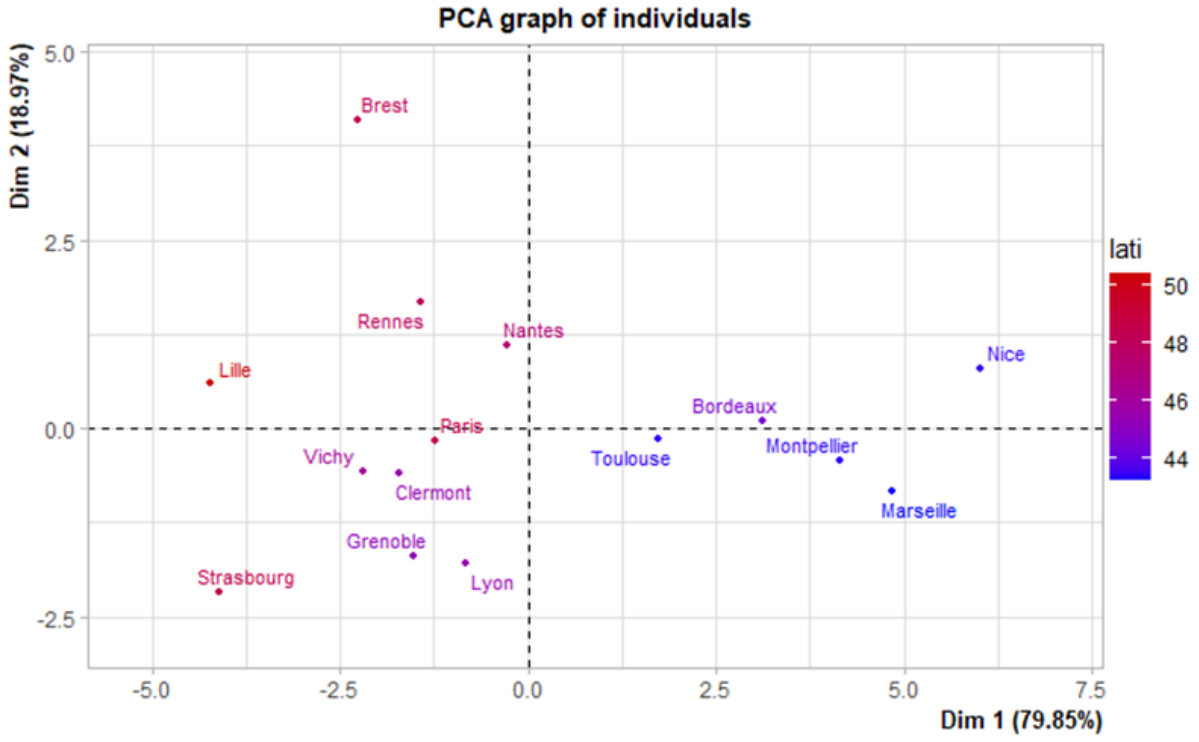
```
library(FactoMineR)
library(factoextra)
villes_donnees.pca<-PCA(villes_donnees,quanti.sup=13:16, graph = FALSE)
villes_donnees.pca$eig
```

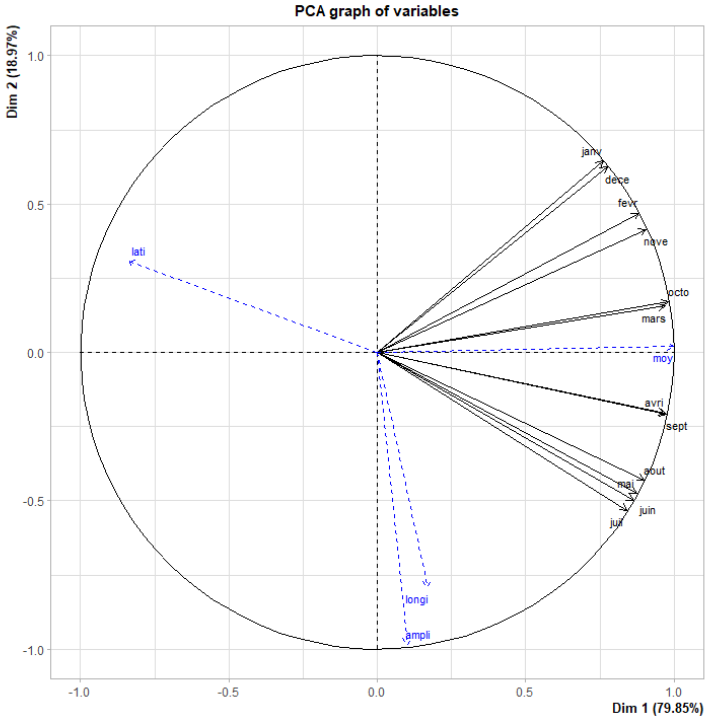
Le pourcentage d'inertie expliqué par le premier plan factoriel est la somme de la variance expliquée par les deux premiers facteurs soit 99% environ.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	9.5817795809	7.984816e+01	79.84816
comp 2	2.2764183987	1.897015e+01	98.81832
comp 3	0.0700144042	5.834534e-01	99.40177
comp 4	0.0396747315	3.306228e-01	99.73239
comp 5	0.0140452901	1.170441e-01	99.84944
comp 6	0.0079815368	6.651281e-02	99.91595
comp 7	0.0060492555	5.041046e-02	99.96636
comp 8	0.0017468930	1.455744e-02	99.98092
comp 9	0.0014921781	1.243482e-02	99.99335
comp 10	0.0004921334	4.101112e-03	99.99745
comp 11	0.0002858357	2.381964e-03	99.99984
comp 12	0.0000197621	1.646842e-04	100.00000

2. Interpréter les 2 axes à l'aide du cercle des corrélations et de la représentation des villes sur le premier plan factoriel

```
plot(villes_donnees.pca,choix="ind", habillage=13, cex=0.7)
plot(villes_donnees.pca,choix="var", habillage=13, cex=0.7)
```





La moyenne des températures mensuelles calculées sur 30 ans représente bien les villes car elle est couchée sur le premier axe factoriel, l'amplitude aussi car elle est presque couchée sur le 2nd axe factoriel.
La variable latitude est corrélée négativement à la moyenne des températures mensuelles calculées sur 30 ans car plus un pays a une latitude élevée et plus la ville se situe au nord et donc les températures sont basses.

Le 1er axe oppose les villes de France ayant une latitude élevée (situé au nord) à gauche aux villes de Frances ayant une attitude faible (situées au Sud) alors représentées à droite.
Le 2nd axe oppose les villes avec une amplitude des moyennes mensuelles élevé (Lyon, Grenoble, Strasbourg, …) aux villes avec une amplitude faible (Brest, Rennes, Nantes, …).

3. Les individus "villes" sont-ils bien représentés sur le premier plan factoriel ?

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

	Dim.1	Dim.2
Bordeaux	0.94668773	0.001161224
Brest	0.23436246	0.763393814
Clermont	0.87988441	0.103705112
Grenoble	0.42894041	0.522580994
Lille	0.97152116	0.019355705
Lyon	0.17813711	0.817127272
Marseille	0.96419529	0.028358560
Montpellier	0.98575843	0.010862202
Nantes	0.05645333	0.886324192
Nice	0.98005143	0.016920844
Paris	0.88935998	0.014094539
Rennes	0.41985296	0.566502170
Strasbourg	0.77565410	0.217137845
Toulouse	0.95255524	0.005855863
Vichy	0.92150642	0.062910418

Globalement tous les « individus » villes sont bien représentées sur le premier plan factoriel car pour chacun de ces individus la somme des cos2 sur les 2 premières dimensions est supérieur à 0,5.
Par ailleurs les villes les mieux représentées sont Brest (0,9977) et Nice (0,9969).

4. Quelles sont les caractéristiques des villes Lille et Strasbourg ? A quelles villes sont-elles opposées ?

Les villes de Lille et Strasbourg sont situées au Nord (latitude élevée) avec des températures moyennes mensuelles basses. Ces caractéristiques s'opposent aux villes de Marseille et Nice (latitude faible), villes avec des températures moyennes mensuelles élevées au Sud.

5. Quelles sont les caractéristiques des villes Brest, Rennes et Nantes ? A quelles villes sont-elles opposées ?

Les villes de Brest, Rennes, Nantes ont des amplitudes de moyennes mensuelles faibles. Ces villes sont opposées aux villes de Strasbourg, Grenoble et Lyon dont les amplitudes sont élevées.

2.2 Classification

1. Effectuer une classification grâce à la méthode des K-means. Interprétez cette classification. Quel nombre de classes vous semble le plus approprié ?

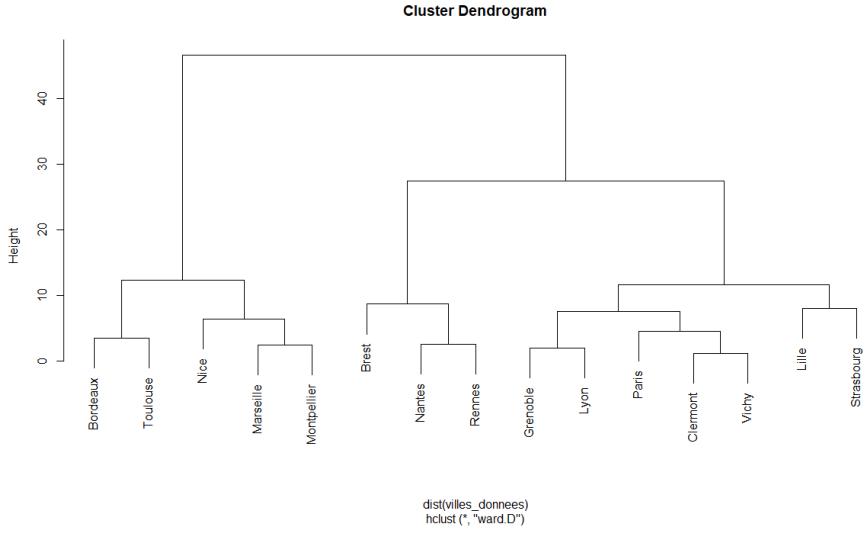
```
km <- kmeans(villes_donnees,5,nstart=10)
km
```

```
Clustering vector:
   Bordeaux       Brest    Clermont     Grenoble       Lille        Lyon   Marseille Montpellier
          1           4           3           3           5           3           2           2
     Nantes        Nice       Paris      Rennes  Strasbourg    Toulouse      Vichy
          4           2           3           4           5           1           3
```

La classification permet de diviser les différentes villes en cinq classes.
Classe 1 : Nord-Est, Lille, Strasbourg.
Classe 2 : Centre, Grenoble, Lyon, Paris, Vichy, Clermont.
Classe 3 : Nord-Ouest, Brest, Rennes, Nantes.
Classe 4 : Sud-Est, Marseille, Montpellier, Nice.
Classe 5 : Sud-Ouest, Bordeaux, Toulouse.

2. Effectuer une classification grâce à une classification hiérarchique. Interprétez cette classification. En combien de classes aurait-on envie de couper le dendrogramme ?

```
library(cluster)
classi_villes=hclust(dist(villes_donnees),method="ward.D")
classi_villes
plot(classi_villes)
```



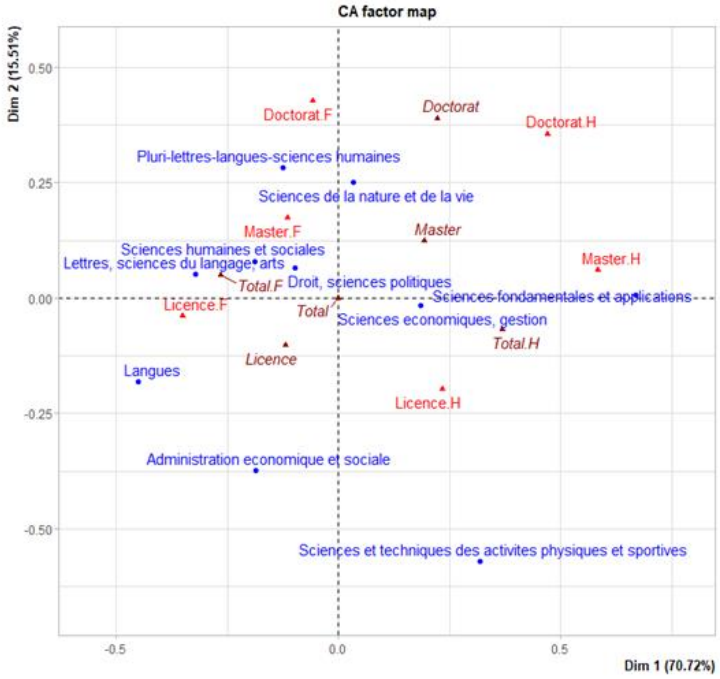
Vraisemblablement, ce dendrogramme classe les villes selon leur position géographique.
Ainsi, les villes proches sur le plan géographique sont représentées proche dans cette classification (Ex : Nantes et Rennes, Marseille et Montpellier)

1er cluster (Sud) : Ouest - Bordeaux, Toulouse
Est – Nice et Montpellier, Marseille
2nd Cluster (Nord) : Ouest – Brest et Nantes,Rennes
Est - Lille, Strasbourg et Clermont, Grenoble, Lyon, Paris, Vichy
On aurait envie de couper ce dendrogramme en 3 classes : Nord, Sud et Centre

3. Universités

1- Qu'elle est le pourcentage d'inertie est expliquée par le premier plan factoriel ?

```
library (FactoMineR)
library(factoextra)
res.ca<-CA(univ, col.sup=7:12, graph=TRUE)
eig.val <- get_eigenvalue(res.ca)
eig.val
```



	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.1167660841	70.717688	70.71769
Dim.2	0.0256061155	15.507973	86.22566
Dim.3	0.0180000178	10.901450	97.12711
Dim.4	0.0043369152	2.626590	99.75370
Dim.5	0.0004066786	0.246299	100.00000

Le pourcentage d'inertie expliqué par le premier plan factoriel est la somme de la variance expliquée par les deux premiers facteurs soit environ 86,23 %.

2- Pensez-vous que disciplines attirent surtout les femmes ?

Oui ! toutes les disciplines littéraires sont à gauche du premier axe tout comme les variables niveau-Femmes. Les disciplines littéraires attirent les femmes.

3- Les sciences attirent-t-elles les hommes ?

Oui ! toutes les disciplines scientifiques sont à droite du premier axe tout comme les variables niveau-Hommes. Les disciplines scientifiques attirent les hommes.

4- Les études d'AES sont-elles longues ou courtes ?

Les études d'AES sont des études courtes.

5- Une interprétation des deux premiers axes

```
res.ca$col$cos2
res.ca$row$cos2
```

```
                 Dim 1      Dim 2
Licence.F  0.959440015 0.01203239
Licence.H  0.552700539 0.39492001
Master.F   0.144089823 0.33217275
Master.H   0.952342968 0.01015437
Doctorat.F 0.008951148 0.49191388
Doctorat.H 0.461976568 0.26443143
```

	Dim 1	Dim 2
Droit, sciences politiques	0.29540334	1.280185e-01
Sciences economiques, gestion	0.46018719	3.102281e-03
Administration economique et sociale	0.19943658	7.972010e-01
Lettres, sciences du langage, arts	0.90502330	2.235711e-02
Langues	0.79440050	1.299801e-01
Sciences humaines et sociales	0.84315671	1.515684e-01
Pluri-lettres-langues-sciences humaines	0.03625313	1.802993e-01
Sciences fondamentales et applications	0.97816368	6.961993e-05
Sciences de la nature et de la vie	0.00664339	4.108298e-01
Sciences et techniques des activites physiques et sportives	0.20511247	6.655037e-01

- Le premier axe oppose les disciplines littéraires (Langues, Lettres sciences et arts, Sciences sociales etc.) aux disciplines scientifiques (Sciences fondamentales et applications, Sciences économiques et gestion, AES etc.).
- Le second axe oppose les études longues (Master, Doctorat) aux études courtes (Licence).

6- La proximité de Master F et SVT

La proximité de Master F et SVT n'est pas justifiée car la somme des cos2 pour chaque point est inférieur à 0,5. Les 2 points sont mal représentés dans le 1er plan factoriel, on ne peut donc pas interpréter les proximités.

7- La proximité de lettres, arts et licence F

La proximité de licence.F et lettres, sciences du langage et arts est justifié car la somme des cos2 est supérieur à 0,5 pour chaque points. Les 2 points sont très bien représentés dans le 1er plan factoriel donc on peut interpréter leur proximité. Par ailleurs, cette proximité suggère que les femmes en licence ont tendance à suivre des formations en lettres, sciences du langage et art.