esiea
INGÉNIEUR.E.S NUMÉRIQUES UTILES

TD Nº2 : Analyse Factorielle des Correspondances (AFC)
Analyse exploratoire

Auteurs : Manelle Nouar, Eliott Bernardou, Mohamed Abdelkader, Sarra Madad, Alfio Chadoin
Enseignants : Myriam Bertrand & Jordan Gonzalez

25 septembre 2022

1 | Elections presidentielles 2012

1. **Importer le jeu de donnees President12.txt. Dans la suite, le jeu de donnees importe sous R est appele president12. Declarer le departement comme identificateur des individus (nom des cas dans R).**

```
president12 = read.csv2("/cloud/project/President12.csv")
president12 = as.data.frame(president12)

View(president12)
str(president12)
dim(president12)
names(president12)
attach(president12)
summary(president12)
dim(president12)
colnames(president12)
row.names(president12)

rownames(president12) = president12[, 2]
```

Dans ce TP, nous étudions un jeu de données qui représente les présidentielles de 2012. Nous avons 108 observations et 13 variables.

```
> rownames(president12)
  [1] "Ain"                   "Aisne"
  [3] "Allier"                "Alpes de Haute.Provence"
  [5] "Hautes.Alpes"          "Alpes.Maritimes"
  [7] "Ardèche"               "Ardennes"
  [9] "Ariège"                "Aube"
 [11] "Aude"                  "Aveyron"
 [13] "Bouches.du.Rhône"      "Calvados"
 [15] "Cantal"                "Charente"
 [17] "Charente.Maritime"     "Cher"
 [19] "Corrèze"               "Corse.du.Sud"
 [21] "Haute.Corse"           "Côte.d.Or"
 [23] "Côtes d.Armor"         "Creuse"
 [25] "Dordogne"              "Doubs"
```

2. **Representer par un diagramme en batons les nombres de voix obtenues par les différents candidats dans les departements du Bas-Rhin, de Paris, de la Seine Saint- Denis, du Vaucluse.**

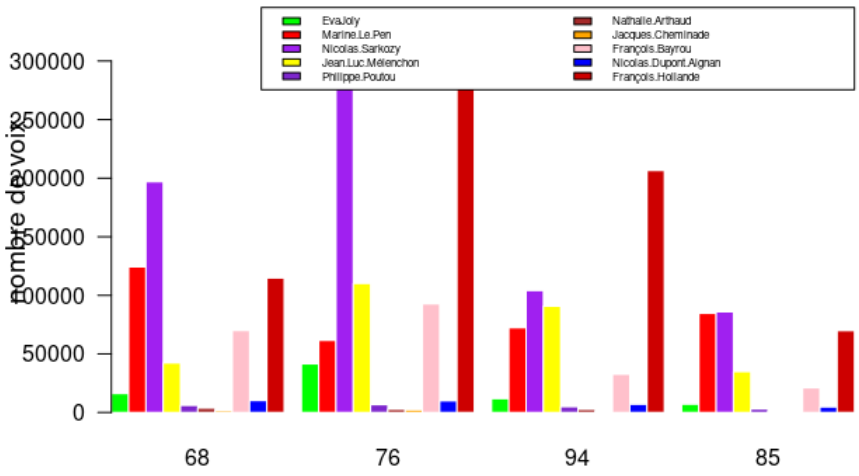
```
color = c("green", "red", "purple", "yellow",
          "purple3", "brown", "orange", "pink", "blue", "red3")

barplot(t(president12[c(68, 76, 94, 85),3:12]),
        col = color, border="white", horiz=F, beside=T,
        ylab="nombre de voix", las=1,  cex.lab=1.2)

legend("topright", cex = 0.5, ncol = 2,
       legend = colnames(president12)[3:12],
       fill = color)

```

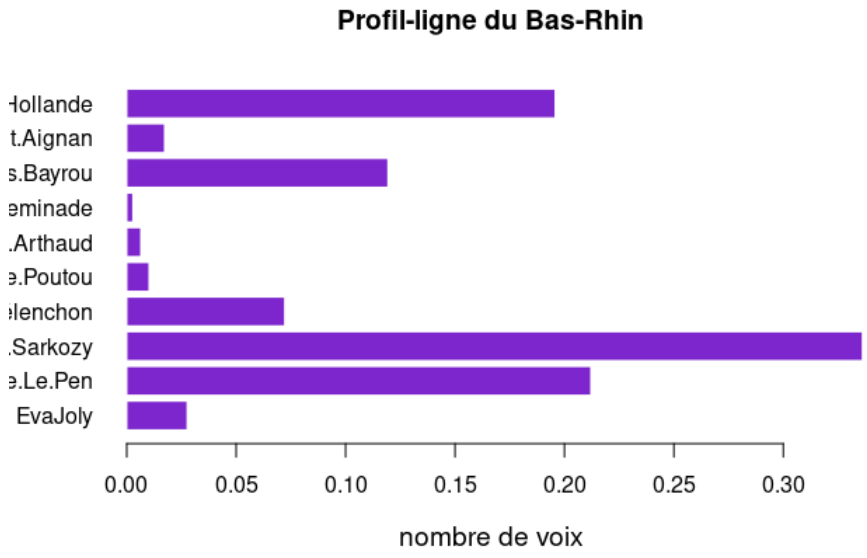
Notre histogramme affiche le nombre de voix dans le Bas-Rhin, à Paris, en Seine Saint-Denis et dans le Vaucluse.



3. **Construire :**
   – le profil-ligne du Bas-Rhin (tableau des frequences).
   – le graphique representant le profil-ligne du Bas-Rhin.

```
1  tab_freq = president12[68, 3:12] / president12[68, 13] # Tableau des
   fre quence
2
3  barplot(as.matrix(tab_freq),names.arg=colnames(president12)[3:12],
4          col="purple3", border="white", main="Profil-ligne du Bas-Rhin",
5          horiz=T, las=1, xlab="nombre de voix", cex.lab=1.2)
```

Nous avons affiché le profil-ligne du Bas-Rhin :



4. **Realiser le test du $\chi 2$ permettant d'etudier le lien de dependance entre les 96 departements de France metropolitaine et les dix candidats.**
   – Extraire des resultats de ce test : la p-value puis les donnees attendues en cas d'independance .
   – Quel est le profil ligne attendu d'un departement quelconque sous l'hypothese d'independance ?

```
sum(president12[1:96,3:12])
min(chisq.test(president12[1:96,3:12])$expected)

resu.chi2 = chisq.test(president12[1:96,3:12])
resu.chi2
```

Ce test permet de vérifier l'absence de lien statistique entre deux variables X et Y. Les deux sont dites indépendantes lorsqu'il n'existe aucun lien statistique entre elles.
Pour effectuer ce test il faut remplir certaines conditions comme, l'effectif total > 50, l'effectif théorique minimum supérieur à 5 et avoir des variables qualitatives.
L'hypothèse nulle (H_0) de ce test est la suivante : les deux variables X et Y sont indépendantes.
En termes de valeur p, l'hypothèse nulle est généralement rejetée lorsque p $\leq 0,05$.

$$H_0 : X \perp\!\!\!\perp Y, \text{ contre } H_1 : X \not\perp\!\!\!\perp Y$$

$$et,$$

$$T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Règle de décision :

- Si $\rho \leq \alpha$: les variables présentent une association statistiquement significative donc on rejette H_0.
- Si $\rho > \alpha$: on ne possédez pas suffisamment de preuves pour conclure que les variables sont associées, donc on ne pas rejeter H_0.

```
> resu.chi2 = chisq.test(president12[1:96,3:12])
> resu.chi2

        Pearson's Chi-squared test

data:  president12[1:96, 3:12]
X-squared = 1226242, df = 855, p-value < 2.2e-16
```

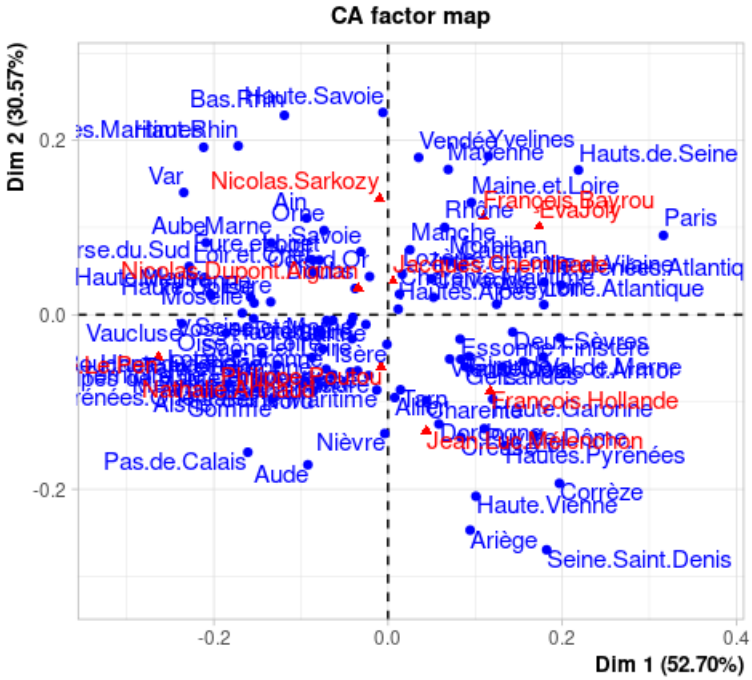
D'après le test de dépendance, la p-valeur est très faible. On rejette donc l'hypothèse nulle d'indépendance des variables. Il existe bien un lien entre ces variables.

5. **Effectuer l'analyse factorielle des correspondances (AFC) du tableau avec les choix suivants :**
   – lignes actives = les 96 departements de la France metropolitaine
   – colonnes actives = les 10 candidats

```r
library(FactoMineR)
AFC = CA(president12[1:96, 3:12], row.sup=NULL, col.sup=NULL)
AFC
```

```
**Results of the Correspondence Analysis (CA)**
The row variable has  96  categories; the column variable has 10 categories
The chi square of independence between the two variables is equal to 1226242 (p-value =  0 ).
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$col"            "results for the columns"
3  "$col$coord"      "coord. for the columns"
4  "$col$cos2"       "cos2 for the columns"
5  "$col$contrib"    "contributions of the columns"
6  "$row"            "results for the rows"
7  "$row$coord"      "coord. for the rows"
8  "$row$cos2"       "cos2 for the rows"
9  "$row$contrib"    "contributions of the rows"
10 "$call"           "summary called parameters"
11 "$call$marge.col" "weights of the columns"
12 "$call$marge.row" "weights of the rows"
```

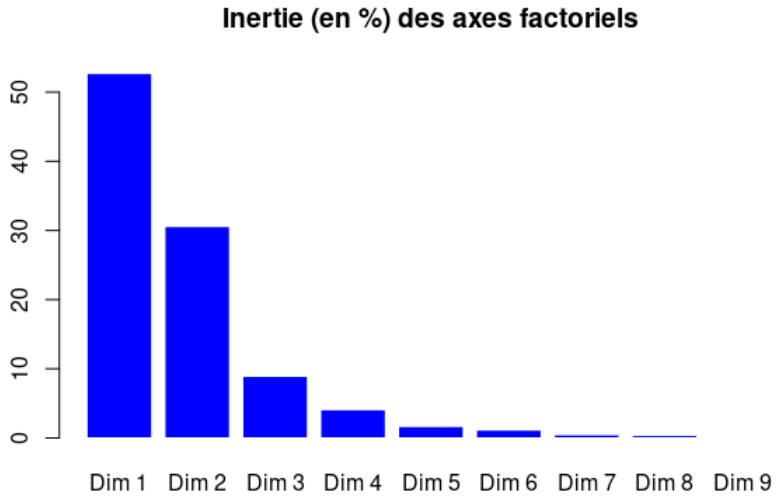


Ce graphe permet de visualiser les similitudes de vote entre les différents départements.

6. **Créer un diagramme en bâtons pour étudier la décroissance de l'inertie des axes :**

```r
barplot(AFC$eig[,2], names=paste("Dim",1:length(AFC$eig[,2])),
        main="Inertie (en %) des axes factoriels",
        col="blue", border="white")

```



7. **Retrouver l'inertie totale de deux facons :**
   – a partir du test du 2 ;
   – en calculant la somme des inerties de tous les axes factoriels issus de l'AFC.
   – En deduire que la valeur du V de Cramer est egale a 0.063.

```r
as.numeric(resu.chi2$statistic) / sum(president12[1:96,3:12])

eigenValue = get_eigenvalue(AFC)

inertieTotale = as.numeric(sum(eigenValue[,1]))
inertieTotale

VCramer = sqrt(inertieTotale /(min(nrow(president12[1:96,3:12]), ncol(
president12[1:96,3:12]))-1))
VCramer
```

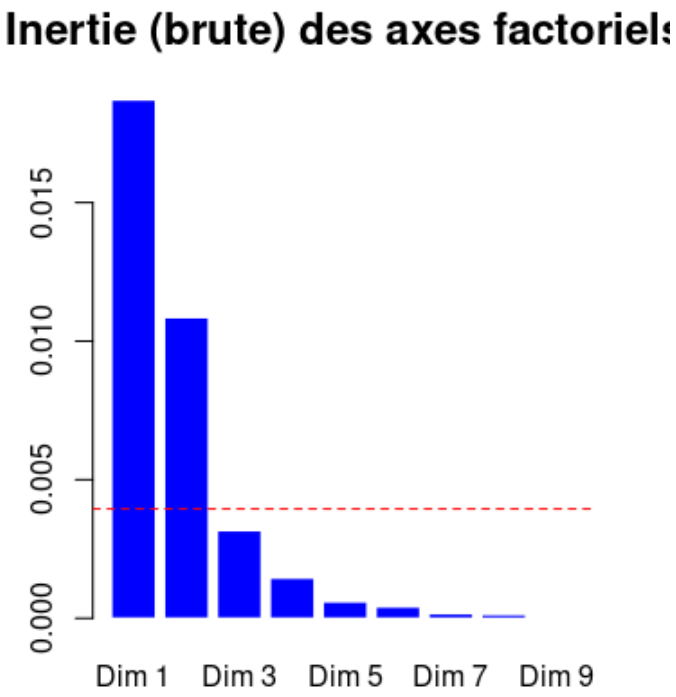
```
[1] 0.03548686
[1] 0.03548686
[1] 0.06279319
```

8. **Calculer la valeur propre moyenne. En deduire le nombre d'axes dont l'inertie est superieure a l'inertie moyenne d'un axe.**

```r
sum(AFC$eig[,1])/9

barplot (AFC$eig[1:9,1],
         names=paste("Dim",1:9),
         main="Inertie (brute) des axes factoriels",
         col="blue", border="white", cex.main=1.5)
abline(h=sum(AFC$eig[,1])/9, lty=2, col="red")
text(6,0.005,"Inertie moyenne",
     col="red", cex=1.5)
```

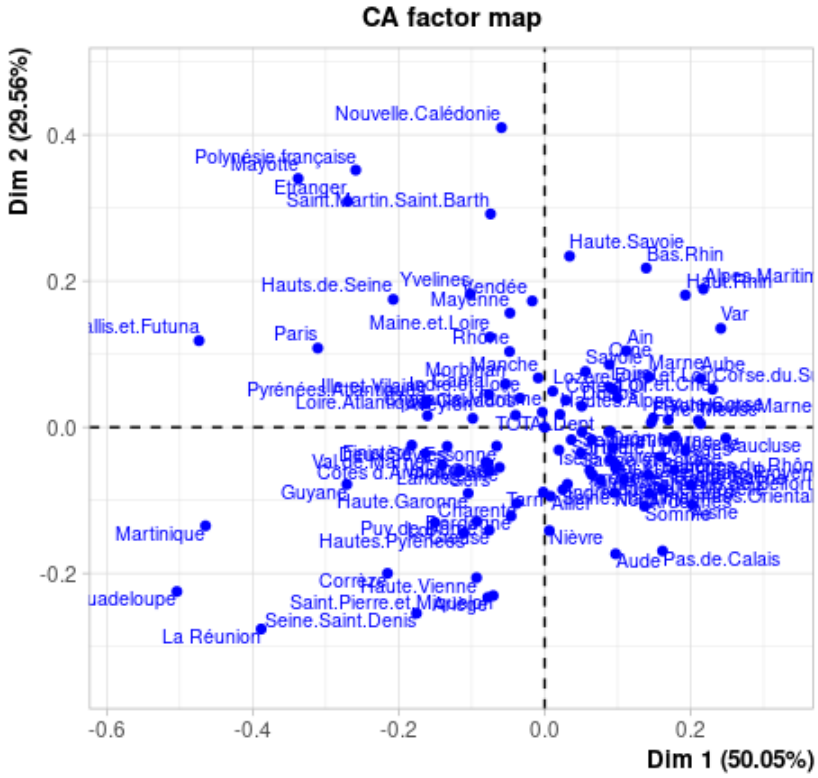
```
> sum(AFC$eig[,1])/9 # Valeur propre moyenne
[1] 0.003942984
```



On peut en conclure qu'il suffit de 2 axes pour que l'inertie soit supérieure à l'inertie moyenne.

9. **Produire le premier plan factoriel (axes 1 et 2) des seuls individus (departements). Remarque : pour rendre le graphique plus lisible, il est possible de reduire la taille des libelles des departements en soumettant a nouveau les commandes de l'AFC apres avoir ajoute l'argument cex=0.7 a la fin de la commande plot.CA.**

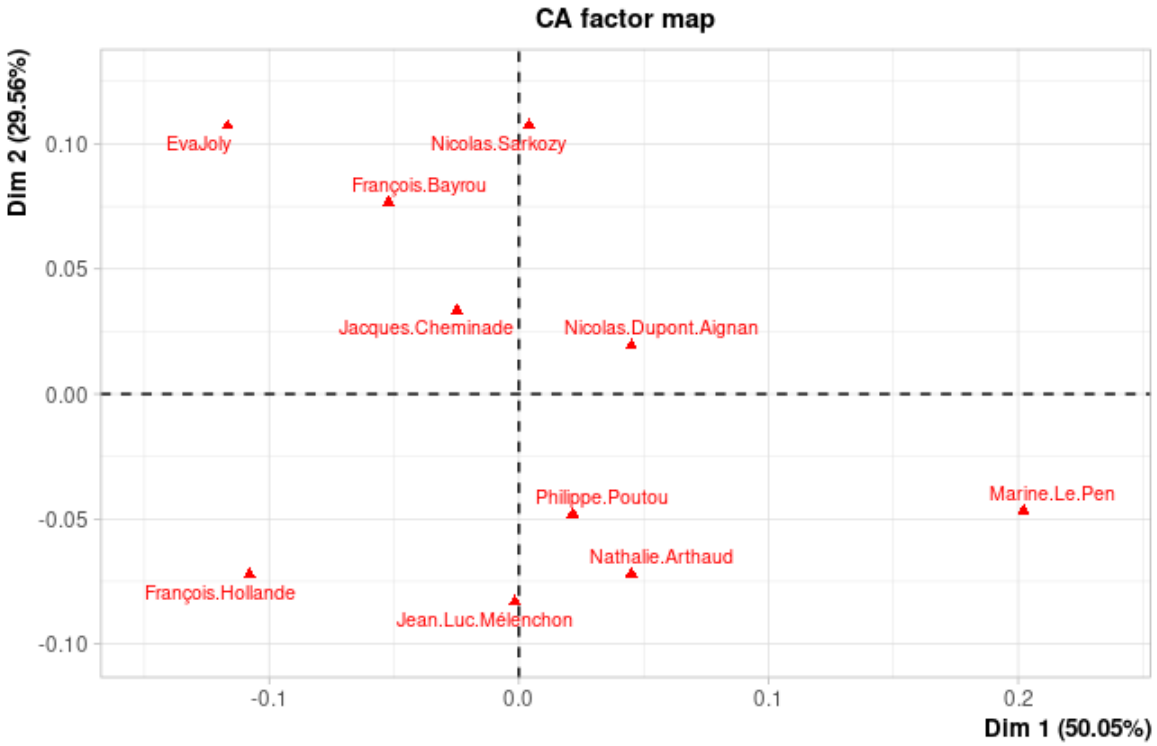
```
AFC = CA(president12[1:108, 3:12], graph = FALSE)
plot.CA(AFC, cex=0.7, invisible = "col")
```



L'ACF permet de visualiser les similitudes de vote entre les différents départements.

10. **Produire le premier plan factoriel des colonnes actives (candidats) et supplementaires.**

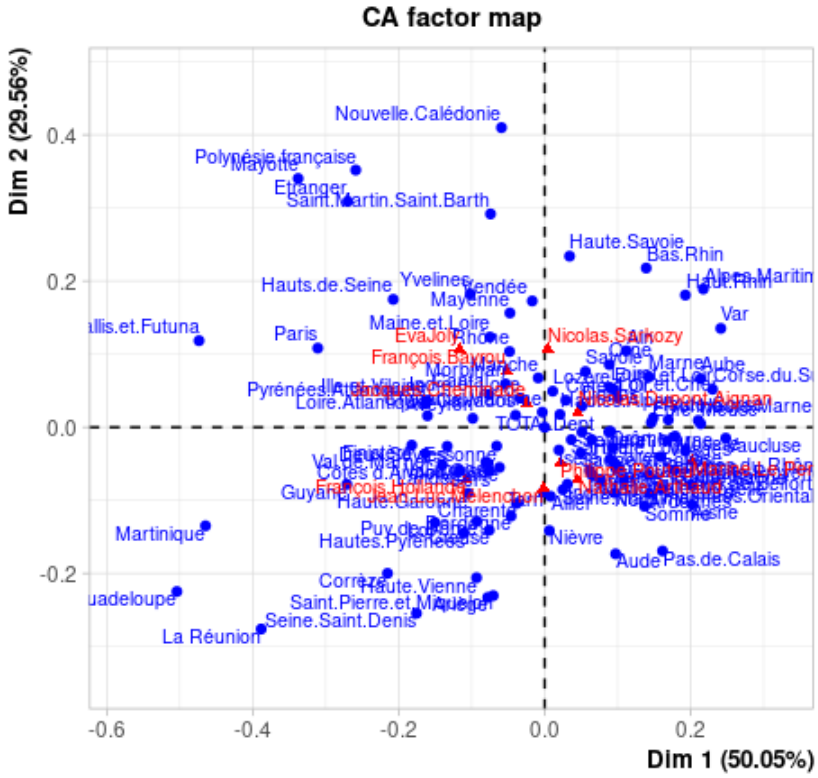
```
plot.CA(AFC, cex=0.7, invisible = "row")
```



Le graphe des colonnes nous permet de voir les similitudes entre les candidats. Deux candidats proches sur le graphique ont un profil de votes identiques dans les départements.

11. **Produire le premier plan factoriel avec la totalite des lignes et colonnes (actives et supplementaires).**

```
plot.CA(AFC, cex=0.7)
```



12. **Quels sont les departements ayant le plus fortement contribue a la construction de l'axe 1 ? de l'axe 2 ? Pour obtenir les contributions triées par ordre décroissant des axes 1 et 2, on peut soumettre les commandes :**

```
1  sort(AFC$row$contrib[,1], T)
2  sort(AFC$row$contrib[,2], T)
```

```
> sort(AFC$row$contrib[,1], T)
              Paris          Hauts.de.Seine                     Var          Alpes.Maritimes
       1.535816e+01            5.428419e+00            5.115035e+00            4.045516e+00
  Loire.Atlantique          Pas.de.Calais       Bouches.du.Rhône               Finistère
       3.834126e+00            3.407218e+00            3.371757e+00            3.371119e+00
       Val.de.Marne         Ille.et.Vilaine       Seine.Saint.Denis               Vaucluse
       2.876997e+00            2.869037e+00            2.739564e+00            2.715024e+00
```

```
> sort(AFC$row$contrib[,2], T)
      Seine.Saint.Denis                 Bas.Rhin                 Yvelines           Haute.Savoie
           1.033938e+01             8.168503e+00             6.337401e+00             5.766694e+00
     Alpes.Maritimes          Pas.de.Calais          Hauts.de.Seine                Haut.Rhin
           5.731288e+00             5.628983e+00             5.381849e+00             4.049472e+00
             Vendée                      Var                    Nord             Haute.Vienne
           3.531328e+00             3.147859e+00             2.789018e+00             2.516891e+00
```

Les départements ayant le plus contribué à l'axe 1 sont :
Paris, Hauts-de-Seine, Var, les Alpes Maritimes et la Loire Atlantique.
Les départements ayant le plus contribué à l'axe 2 sont :
Seine-Saint-Denis, Bas-Rhin, Yvelines, la Haute-Savoie et les Alpes Maritimes.

13. **Faire de meme pour les candidats.**

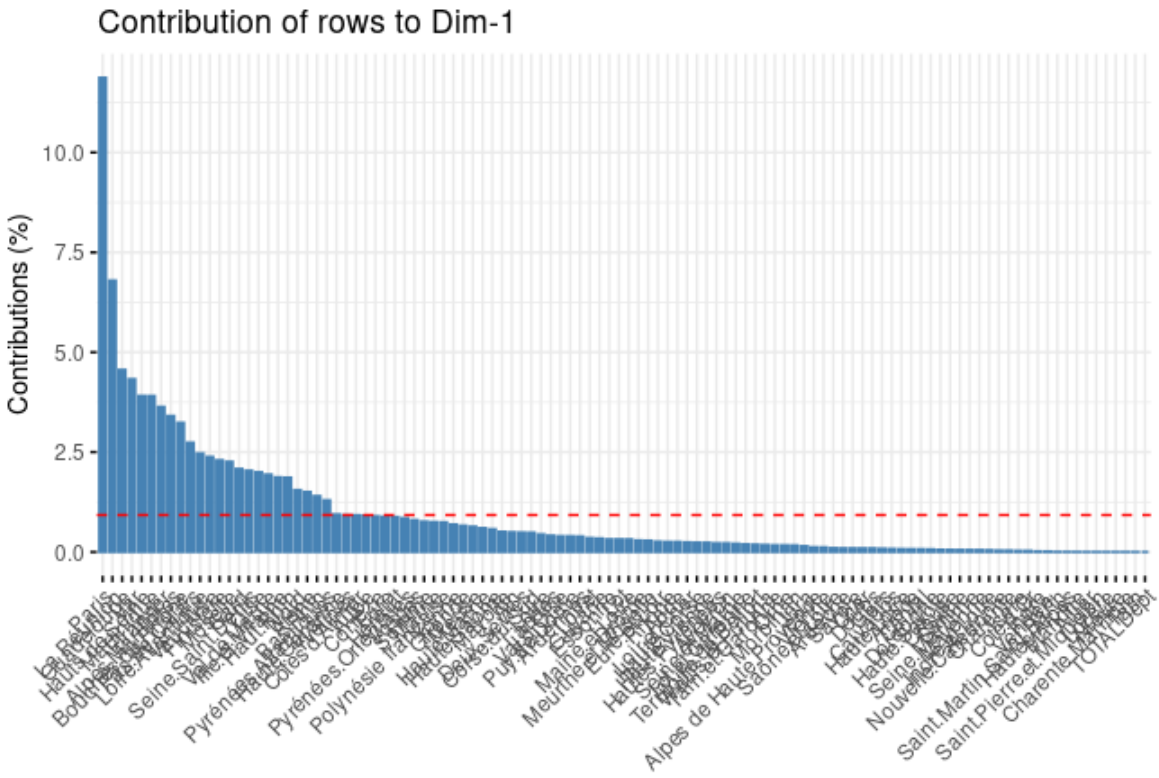
```r
sort(AFC$col$contrib[,1], T)
sort(AFC$col$contrib[,2], T)
```

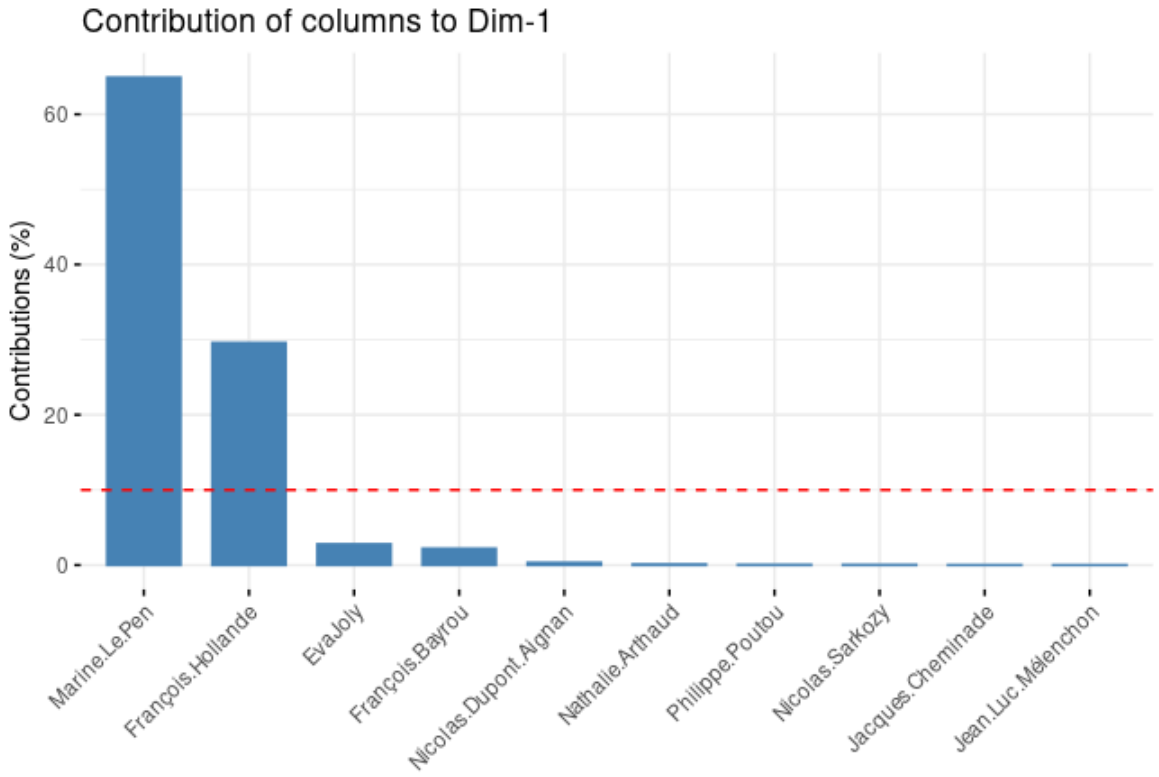
```
> sort(AFC$col$contrib[,1], T)
       Marine.Le.Pen     François.Hollande               EvaJoly      François.Bayrou Nicolas.Dupont.Aignan     Nathalie.Arthaud      Philippe.Poutou
         64.902393775          29.580049222           2.786568303           2.202574998           0.323948133           0.101554809           0.047596475
     Nicolas.Sarkozy     Jacques.Cheminade    Jean.Luc.Mélenchon
         0.039199958           0.013526517           0.002587809
> sort(AFC$col$contrib[,2], T)
     Nicolas.Sarkozy     François.Hollande    Jean.Luc.Mélenchon      François.Bayrou       Marine.Le.Pen             EvaJoly     Nathalie.Arthaud
         47.24443889           22.33617626           11.50434065            8.09484714            5.85284461            3.98969413            0.43787765
     Philippe.Poutou Nicolas.Dupont.Aignan     Jacques.Cheminade
         0.39605324            0.10152503            0.04220239
```

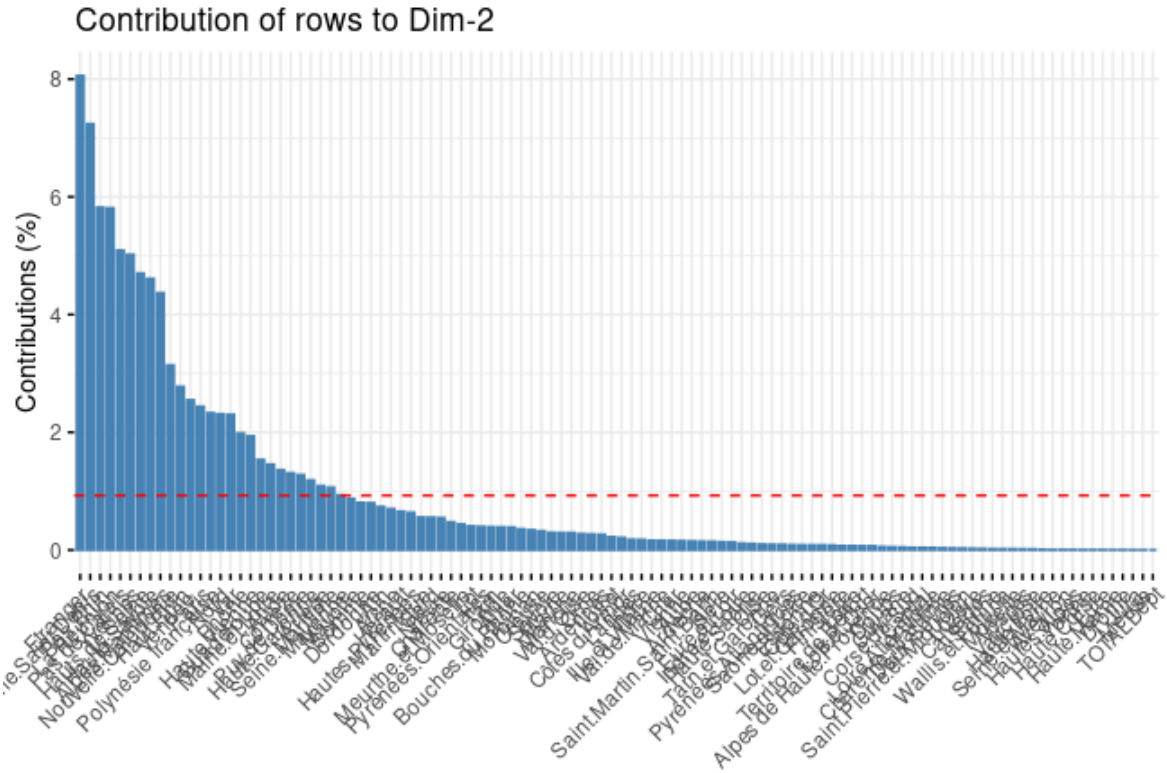
Les candidats qui ont le plus contribué à l'axe 1 sont :
Marine Le Pen et François Hollande.
Les candidats ayant le plus contribué à l'axe 2 sont :
Nicolas Sarkozy et François Hollande.

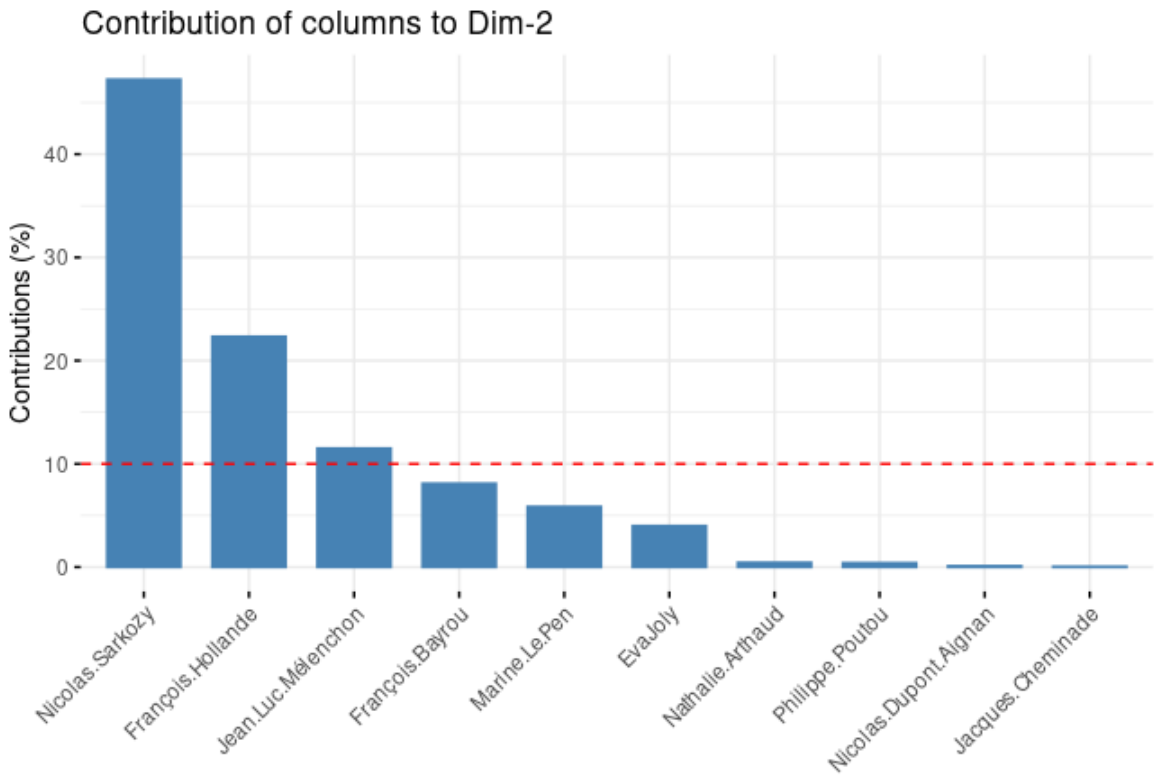
14. **Quels sont les departements et les candidats ayant le plus fortement contribue a l'inertie globale ?**

```r
fviz_contrib(AFC, choice="row", axe=1)
fviz_contrib(AFC, choice="col", axe=1)
fviz_contrib(AFC, choice="row", axe=2)
fviz_contrib(AFC, choice="col", axe=2)
```




Contribution of columns to Dim-1
Contributions (%)
60
40
20
0
Marine.Le.Pen
François.Hollande
Eva.Joly
François.Bayrou
Nicolas.Dupont.Aignan
Nathalie.Arthaud
Philippe.Poutou
Nicolas.Sarkozy
Jacques.Cheminade
Jean.Luc.Mélenchon


Contribution of rows to Dim-2
Contributions (%)
8
6
4
2
0



Les candidats qui ont le plus contribué sont :
Marine Le Pen, Nicolas Sarkozy et François Hollande.
Les départements ayant le plus contribué sont :
Paris, les Hauts de Seine, la Seine-Saint-Denis et le Var.

15. **Calculer la distance de chacun des departements au centre de gravite du nuage. Comment interpreter la distance la plus courte ?**

```
distance = 0
for(i in 1:nrow(AFC$row$coord)){
  distance[i]= sqrt(AFC$row$coord[i,1]^2 + AFC$row$coord[i,2]^2)#Calcul des
distances
}
names = row.names(AFC$row$coord)

df = data.frame(Department = names, Distance_center = distance)
df

df[df["Distance_center"] == min(df["Distance_center"])]
```

Department <chr>	Distance_center <dbl>
Ain	0.15309298
Aisne	0.22805920
Allier	0.09482413
Alpes de Haute.Provence	0.11012843
Hautes.Alpes	0.02730670
Alpes.Maritimes	0.28816967
Ardèche	0.08533303
Ardennes	0.16919743
Ariège	0.24557028
Aube	0.22374941

1-10 of 108 rows Previous 1 2 3

```
[1] "TOTALDept"  "0.00000000"
```

Les Hautes-Alpes sont le département le plus proche au centre de gravité du nuage de points, c'est-à-dire que c'est le département le plus proche du comportement moyen de tous les départements.

16. **Interpreter les axes factoriels 1 et 2.**

    Les départements ayant le plus contribué à l'axe 1 sont : Paris, Hauts-de-Seine, Var, les Alpes-Maritimes et la Loire-Atlantique.

    Les candidats qui ont le plus contribué à l'axe 1 sont : Marine Le Pen et François Hollande.

    Les départements Paris et Hauts-de-Seine ont une contribution plus importante au pôle positif de la première dimension, tandis que les départements Var et Alpes-Maritimes ont une contribution majeure au pôle négatif de la première dimension. De la même manière, c'est l'opposition entre Marine Le Pen et François Hollande qui semble définir la dimension 1 pour les candidats.

    Les départements ayant le plus contribué à l'axe 2 sont : Seine-Saint-Denis, Bas-Rhin, Yvelines, la Haute-Savoie et les Alpes-Maritimes.

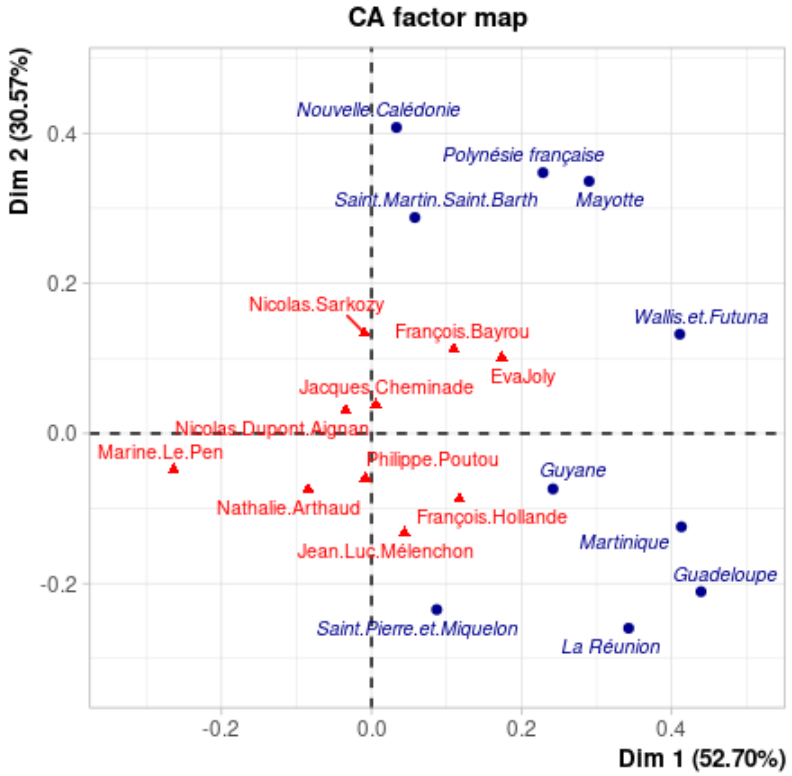
    Les candidats ayant le plus contribué à l'axe 2 sont : Nicolas Sarkozy et François Hollande

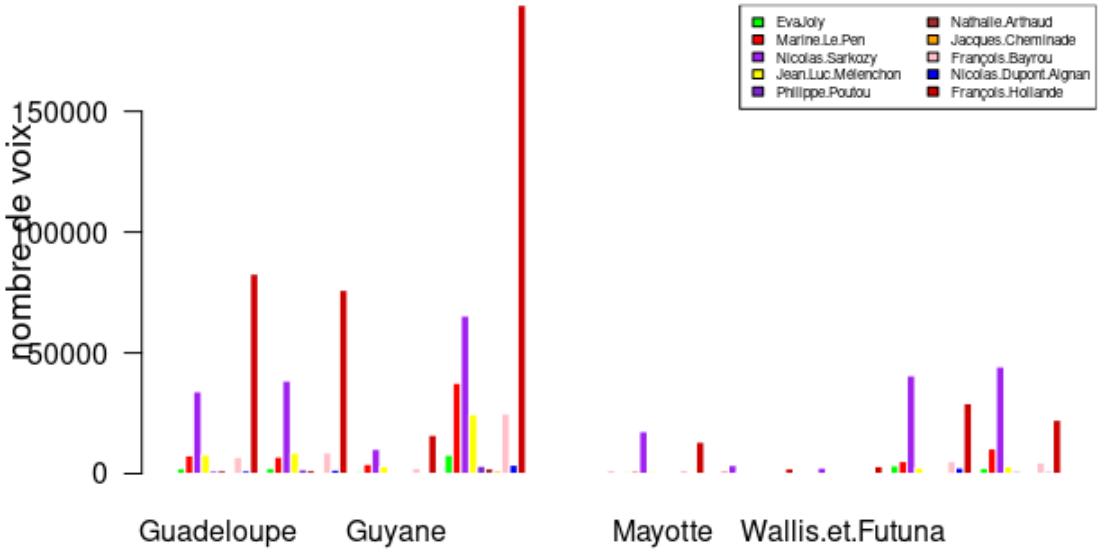
    Pareillement, l'axe factoriel 2 semblent être définies par l'opposition des départements Seine-Saint-Denis et du couple Bas Rhin-Yvelines. Côté candidats, c'est l'opposition du couple Nicolas Sarkozy avec François Hollande.

17. **Refaire l'AFC avec les departements et territoires d'outremer en elements illustratifs (supplementaires). Interpreter la position de ces derniers sur le plan factoriel (valider votre commentaire en construisant les graphiques des profils-lignes de ces nouveaux departements.).**

```
1  plot.CA(CA(president12[1:106,3:12], row.sup = 97:106), cex = 0.7, invisible =
   "row")
2
3  color = c("green", "red", "purple", "yellow",
```

```
4             "purple3", "brown", "orange", "pink", "blue", "red3")
5
6  barplot(t(president12[97:106,3:12]),
7          col = color, border="white", horiz=F, beside=T,
8          ylab="nombre de voix", las=1,  cex.lab=1.2)
9
10 legend("topright", cex = 0.5, ncol = 2,
11        legend = colnames(president12)[3:12],
12        fill = color)
```





Les départements des outre-mer ont un vote très éloigné de la moyenne globale. La Guyane a en grande partie voté pour Marine Le Pen, tandis que les autres départements ont le plus voté pour Nicolas Sarkozy et François Hollande.