TP2 : Distances, Similarités et Dissimilarités

1. EXPLORATION DES DONNEES
2. SIMILAIRES ET DIFFERENTS
3. REPRESENTER LES DISTANCES
4. ANALYSE DES PREFERENCES

1. EXPLORATION DES DONNEES

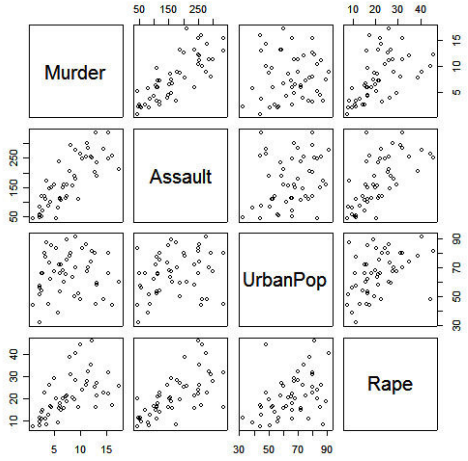
Télécharger le jeu de données « USArrests » dans votre dossier.

a) Calculer les paramètres statistiques élémentaires des variables descriptives.

>data(USArrests)
> USArrests

b) Analysez les résultats des commandes suivantes :

>label<-attributes(USArrests)$row.names
>boxplot(USArrests)
> plot(USArrests)



c) Interpréter le type de relation liant les couples de variables (Murder, Assault), (UrbanPop, Murder) et (Rape, Murder).

2. SIMILAIRES ET DIFFERENTS

L'objectif de cette section est de déterminer les états des US ayant des profiles de violence les plus similaires puis les plus opposés.

a) Étudier le fonctionnement de la fonction « Dist » à travers les commandes suivantes :

> ?dist
>D=dist(USArrests, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
> is.matrix(D)
> is.data.frame(D)

```
>  attributes(D)
> Dmat<-as.matrix(D)
> Dmat[1:6,1:6]
> Dmat<-print(round(as.matrix(D),digits=0))
> Dmat[1:6,1:6]
```

b) Expliquer à quoi est due la différence des résultats des deux commandes suivantes

```
> min(D)
> min(Dmat)
```

c) Interpréter les instructions suivantes d'extraction des états de profils les plus similaires

```
>(1:2500)[Dmat == min(D)]
> col(Dmat)[Dmat == min(D)]
> row(Dmat)[Dmat == min(D)]
>USArrests[c(15,29),]
```

d) Ecrire une fonction R permettant de fournir le(s) couples(s) d'états de profils les plus distants. Testez cette fonction sur « USArrests ».

3. REPRESENTER LES DISTANCES

L'objectif de cette section est de représenter l'ensemble des états sur un axe de telle manière à ce que les distances entre deux états correspondent au mieux aux distances calculées dans la section précédente.

a) Analyser l'effet de la commande « cmdscale » :

```
> Cord<-cmdscale(D)
> Cord
> label<-attributes(Cord)$dimnames[[1]]
> plot(Cord, type="n")
> text(Cord,label)
```

b) Retrouve-t-on les résultats de la section précédente.
c) Analysez les instructions suivantes puis justifiez la différence de résultats citée en b)

```
> Cord3<-cmdscale(D, k=3)
> label3<-attributes(Cord3)$dimnames[[1]]
> plot(Cord3[,1],Cord3[,3], type="n")
> text(Cord3[,1],Cord3[,3],label3,cex=0.7)
```

4. ANALYSE DES PREFERENCES

Une trentaine d'étudiants ont exprimé leurs préférences sur 8 groupes de musique numérotés de 1 à 8. Les données sont enregistrées dans le fichier «PrefMusique.txt » sous forme d'un tableau à 30 lignes (étudiants) et 8 colonnes :

```
  V1 V2 V3 V4 V5 V6 V7 V8
1  7  8  6  5  1  3  4  2
2  1  2  6  7  3  8  4  5
```

3 6 3 2 1 7 8 5 4
…
29 1 6 3 7 2 8 4 5
30 8 7 6 3 1 4 5 2

l'étudiant numéro 1 (ligne 1) préfère dans l'ordre décroissant d'abord le groupe no 7, ensuite le no 8 …jusqu'au moin préféré de no 2.

Codage des groupes : 1= U2, 2= ABBA, 3= Hendrix, 4= Les Chaussettes Noires, 5= Zappa, 6= Doors 7= Bob Marley 8= Léo Ferré.

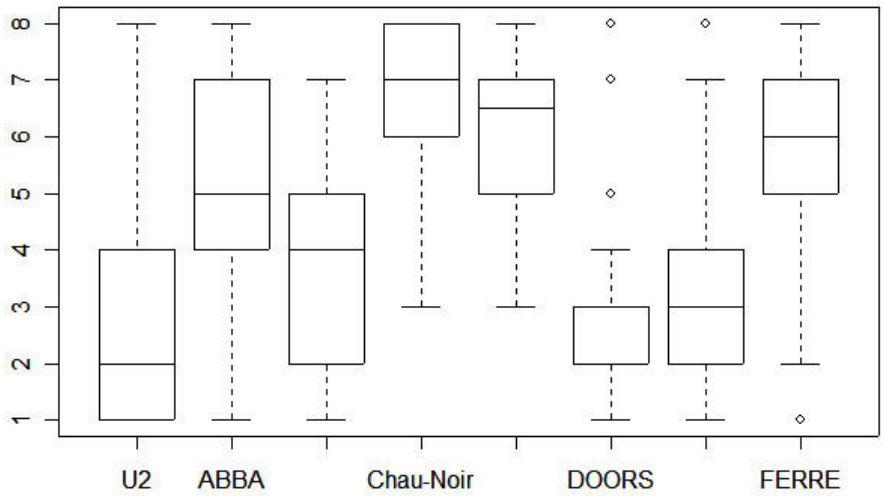
Notre objectif est d'extraire à partir des données de préférence l'ensemble des étudiants ayant fait un choix identique (resp. opposé) et dans le cas ou il n'existe pas de profil identique (resp. opposé) fournir la liste des étudiants ayant des goûts musicaux les plus similaires (resp. les plus opposés).

a) Expliquez pour quelle raison il est erroné d'appliquer les commandes suivantes afin de calculer la matrice de distance entre les choix des étudiants.

>Pref<-read.table(« PrefMusique.txt »)
>Detud=dist(Pref, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)

b) Analyser et interprétez l'effet des instructions qui suit :

> rang<-t(apply(Pref,1,order))
> names(rang)<-c("U2","ABBA","HENDRIX","Chau-Noir","ZAPPA","DOORS","MARLEY","FERRE")
> DR<-dist(rang)
>attributes(DR)
>boxplot(rang)



b) En vous inspirant de ce qui a été fait dans la section 2, déterminez les étudiants ayant effectués les choix les plus similaires (les plus distants) puis fournir une représentation de ces choix. Confrontez les résultats.

d) Écrire une fonction calculant le coefficient de corrélations des rangs de Spearman, l'appliquer ensuite sur le data.frame « *Rang* ».