TP5 : Évaluation n° 1

Jeu de données : Mortalité et pollution

1 Consignes

À la fin de cette évaluation, vous devrez nous rendre deux éléments :
- un code propre et commenté
- un compte rendu papier ou électronique comprenant les réponses aux questions théoriques et les justifications et commentaires accompagnant le code et les graphiques.
La note finale prendra en compte ces deux éléments de la même façon.

2 Description des données

Les données traitées ici sont des données réelles sur la mortalité et la pollution de l'air dans 46 villes américaines récoltées au cours de l'année 1960. Chaque ville est identifiée par un numéro (`num`), son nom (`ville`) et son nom abrégé (`code`, reprenant les 4 premières lettres de `ville`) ; c'est cette dernière variable que nous utiliserons comme identifiant. Pour chacune des villes, on dispose des informations suivantes :

- le taux de mortalité, noté `mort` et exprimé en nombre de décès pour 100000 habitants ;
- le taux moyen de sulfates dans l'air, noté `sulf`, en micro-grammes par m^3 multiplié par 10 ;
- le taux de particules en suspension, noté `part` et exprimé dans les mêmes unités que `sulf` ;
- la densité de population, notée `dens`, en habitants par mile-carré multiplié par 0.1 ;
- le pourcentage de familles dont le revenu est supérieur au seuil de pauvreté, noté `ppauv` ;
- le pourcentage de personnes âgées de plus de 65 ans, noté `p3age` et multiplié par 10.

Les données sont disponibles à l'adresse suivante :
`https://synapse.math.univ-toulouse.fr/index.php/s/2uNZpPC28wobKzK`
Pour récupérer les données, utilisez la fonction `read.table` avec les paramètres suivants :
`read.table(`*chemin du fichier*`,colClasses=c("character", rep("numeric",6)), header=F,row.names=1,col.names=C)`.

3 Partie 1 : Analyse en Composantes Principales

1. Pourquoi réaliser ce type d'analyse sur ce jeu de données ?
2. Quelles variables peuvent être utilisées pour réaliser une ACP sur ce jeu de données ?
3. Créer X, la matrice contenant les données. Calculer la matrice des corrélations. Quelles variables semblent corrélées ?
4. Diagonaliser la matrice des corrélations avec la fonction.
   Afin d'avoir une idée du nombre d'axes factoriels à considérer dans la suite, réaliser un diagramme présentant les valeurs propres ainsi qu'un diagramme présentant les pourcentages cumulés de ces valeurs propres. Justifier qu'il est pertinent de considérer deux axes.
5. Calculer le tableau Y des coordonnées des individus dans le nouveau repère.
   Afficher les projections des individus sur le plan correspondant aux deux premiers axes factoriels.
   Ajouter les anciens axes, représentés avec des flèches depuis l'origine, avec le nom des variables associées.
6. Calculer les corrélations entre les variables initiales et les variables correspondants aux axes factoriels (les composantes principales). Créer le graphique correspondant sans oublier le cercle unité.

7. Interpréter les deux premier axes factoriels : quels individus opposent-ils ? Quelles variables y contribuent positivement, négativement ?

4 Partie 2 : Analyse univariée

1. En vous inspirant du TP n° 2, proposez une analyse, la plus détaillées possible, de la variable "taux de mortalité".
2. Donner les indicateurs de tendance centrale de la densité de population, pour les villes où le pourcentage de personnes âgées est respectivement supérieur et inférieur à 8%.