TP 3

L3 MIASHS - UBO

Dans ce TP, on s'intéresse aux systèmes de santé de 185 pays à travers le monde. Pour mesurer la qualité des soins, on considère l'espérance de vie à la naissance (LIFEEXP). Tous les pays n'ont pas fourni les informations pour chaque variable. Les données manquantes sont notées par 'NA' dans le tableau de données.

Notation des variables :

- REGION : région du monde
- COUNTRY : nom du pays
- LIFEEXP : espérance de vie à la naissance, en années
- ILLITERATE : taux d'analphabétisme des adultes, % 15 ans et plus
- POP : population en 2015, en millions
- FERTILITY : indice synthétique de fécondité, naissances par femme
- PRIVATEHEALTH : dépenses privées de santé en 2004, en % du PIB
- PUBLICEDUCATION : dépenses publiques dans l'éducation, en % du PIB
- HEALTHEXPEND : dépenses de santé par habitant en 2004, PPA (parité de pouvoir d'achat) en USD
- BIRTHATTEND : accouchements assistés par du personnel de santé qualifié (%)
- PHYSICIAN : médecins pour 100000 habitants
- SMOKING : prévalence du tabagisme, (homme) %$ d'adultes
- RESEARCHERS : chercheurs en R&D, par million d'habitants
- GDP : PIB, en milliards d'USD
- FEMALEBOSS : législateurs, hauts fonctionnaires et cadres, % de femmes

1 Préparation des données

1. Importer les données depuis le fichier suivant :

http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/CSVData/UNLifeExpectancy.csv

2. A l'aide de la fonction *head()*, visualiser les données.
3. On ne souhaite pas garder la variable REGION. Proposer une méthode compacte pour supprimer la variable REGION du tableau de données.
4. Commenter les deux lignes suivantes, que ce passe-t-il pour le tableau de données *data* ?

```
row.names(data)=data[,2]
data=data[,-2]
```

5. On souhaite ramener le Le produit intérieur brut (GDP) au PIB par habitant. Proposer une méthode pour modifier la variable GPD en GPD par habitant.
6. Le tableau de données comporte des données manquantes. On souhaite supprimer les pays comportants des données manquantes (NA), proposer une méthode.

2 Analyse descriptive univariée

1. Représenter la distribution des variables sous la forme de boxplot. Commenter.
2. Repréenter la distribution de la variable LIFEEXP sous la forme d'un diagramme en batons (*barplot*). Donner un titre à votre graphique et des noms à l'axe des abscisses et l'axe des ordonnées. Commenter.
3. Choisir deux variables qui vous semblent *a priori* fortement corrélées. Réaliser un nuage de point entre les deux variables afin de visualiser si celles-ci semblent corrélées positivement/négativement ou non.

3 ACP

1. A l'aide des fonctions *PCA* (package "FactoMineR") ou *princomp* réaliser une analyse en composantes principales des données (regarder l'aide de R pour les détails des fonctions).
2. Commenter la qualité de l'ACP :

- Calculer les pourcentages d'inerties expliqués par le premier axe principal puis par le premier plan principal.
- Calculer les contributions des individus aux différentes composantes (CTR). Existe-t-il des individus qui apportent une contribution importante à l'un des deux premiers axes principaux ?
- Calculer les quantités COS2 et discuter de la qualité de la représentation des individus sur le premier axe principal puis sur le premier plan principal.

3. Tracer les graphiques utiles à l'interprétation.
4. Interpréter et analyser les résultats. Notamment en répondant aux questions suivantes :

- Combien d'axes retiendriez-vous pour l'ACP de ce jeu de données ?
- Quelle est la part d'inertie expliquée par le premier plan principal ?
- Visualiser les coordonnées des individus dans le premier plan principal et commenter les contributions des individus aux variables principales.
- Quelles variables sont positivement/négativement corrélées ? Lesquelles ne semblent pas corrélées ?

5. Influence du choix de la métrique : effectuer une ACP sans normaliser les variables (avec "scale.unit=FALSE" dans la fonction PCA). Interpréter les résultats obtenus et comparer avec l'ACP normée.