Master 1 – BDIA
Analyse et visualisation des données

Sujet TP3

Contexte

Le "Forever Pollution Project" est une enquête journalistique qui s'intéresse aux substances per- et polyfluoroalkylées (PFAS), souvent surnommées "polluants éternels" en raison de leur extrême persistance dans l'environnement. Utilisés depuis les années 1940 pour leurs propriétés résistantes à l'eau et à la chaleur, ces composés chimiques sont aujourd'hui une source majeure de préoccupation environnementale et sanitaire. L'enquête se distingue par son approche rigoureuse, basée sur des méthodologies scientifiques reconnues et des prélèvements environnementaux étendus. Un article du journal Le Monde en date du 23 février 2023 résume cette enquête ainsi que la méthode d'analyse retenue : `https://www.lemonde.fr/les-decodeurs/article/2023/02/23/polluants-eternels`-explorez-la-carte-d-europe-de-la-contamination-par-les-pfas_6162942_4355770.html.

Sur le site `https://pdh.cnrs.fr/fr/`, le Centre National de la Recherche Scientifique (CNRS) met à disposition, en accès libre, les données et livre plusieurs analyses.
L'objectif des TP3 et 4 est d'étudier un jeu de donnée, disponible sur ce site, relatif à l'eau du robinet en France.

Exercice 1 : Analyse des données

Sur le site `https://pdh.cnrs.fr/fr/`, télécharger le jeu de données numéro **124**, désigné **france_eaurob** relatif à l'analyse de l'eau du robinet en France. L'objectif de cet exercice est de comprendre la structure ainsi que les données de ce dataset.
Indications :

- summarise() → Agréger des données, souvent utilisée en complément de group_by().
- arrange(desc(xxx)) → Trier xxxx par ordre descendant.
- print(xxx, n = 100) → Limiter le nombre de lignes affichées à 100 ;
  n = Inf → Afficher toutes les lignes.

Travail à faire :

1. Combien de lignes et combien de colonnes dans ce dataset ?
2. Afficher le nom des colonnes.

3. Afficher la première ligne du dataset.
4. Vérifier que les données concernent uniquement la France.
5. Les prélèvements concernent combien de villes différentes ?
6. Est-ce que des prélèvements sont réalisés à Dijon, si oui, combien ?
7. En étudiant *matrix*, préciser les différentes sources de prélèvement (ex. : eau souterraine, eau potable etc) et indiquer le nombre de prélèvements par sources ?
8. Quelle est la période couverte par les prélèvements ? (date la plus ancienne et la plus récente)
9. Indiquer le nombre de prélèvements par an.
10. Citer les 5 villes dans lesquelles on a réalisé le plus de prélèvements ?

Exercice 2 : Statistiques selon les 5 sources de prélèvement

Dans l'exercice 1, question 7, vous avez montré qu'il y avait 5 sources de prélèvement :

- "Groundwater", eau souterraine (nappes phréatiques, puits)
- "Drinking water" : eau potable
- "Surface water" : eau de surface (rivières, lacs, barrages)
- "Sea water" : eau de mer
- "Unknow" : lieu de prélèvement inconnu.

Travail à faire :

- Pour chacune de ces 5 sources de prélèvement, calculer la moyenne, la médiane et l'écart type des PFAS *pfas_ sum*
- Quelles sont les sources de prélèvement pour Dijon ?
- Les prélèvement "Sea water" concernent quelles villes ?

Exercice 3 : Qualité des données

L'objectif de cet exercice est double. D'une part, il s'agit de nettoyer les données. D'autre part, il s'agit de réduire le dataset aux informations pertinentes et de supprimer toutes les autres, notamment celles liées au contexte de l'étude. Par exemple, toutes les lignes du dataset comporte la valeur 124 dans le champs *dataset_ id*. Cette information liée au contexte de l'étude n'apporte pas d'information pertinente dans l'analyse des données. Les fonctions suivantes vous permettront de répondre aux questions de cet exercice.

- unique(x) $\rightarrow$ Trouve toutes les valeurs uniques d'une colonne.
- length(unique(x)) $==$ 1 $\rightarrow$ Vérifie si une colonne n'a qu'une seule valeur unique.
- sapply(data, function(x) ...) $\rightarrow$ Applique la fonction à toutes les colonnes.

- names(data)[...] → Sélectionne les colonnes concernées.
- trimws(data$ville) → Supprime les espaces avant et après ville.
- toupper(...) → Convertir en majuscules
- data$city → Remplacer la colonne city dans data

Travail à faire :

1. Les noms de ville *city* doivent tous être en majuscules.
2. Y a-t-il des valeurs manquantes dans certaines colonnes ?
3. Quels sont les types de données présents dans chaque colonne ?
4. D'où proviennent les relevés *source_ text* ?
5. Quelles colonnes comportent des valeurs uniques comme par ex. *dataset_ id* ?
6. En fonction des réponses obtenues aux questions précédentes, supprimer toutes les colonnes qui ne vous paraissent pas indispensables et sauvegarder votre nouveau dataset dans **data2**. Poursuivre les exercices avec **data2**.
7. Combien de colonnes comporte le nouveau dataset **data2** ?

Exercice 4 : Exploitation des données au format json

Le champ *pfas_ values* contient les données suivantes sous forme de listes au format JSON.

- cas_id : identifiant unique de la substance (CAS).
- unit : unité de mesure.
- substance : nom de la substance mesurée.
- isomer : information sur l'isomère.
- less_than : limite de détection.
- value : valeur mesurée.

Travail à faire :

1. Convertir *pfas_ values* pour l'exploiter en R
2. Lister et compter les différentes substances présentes.
3. Lister les 5 substances les plus fréquentes.
4. Les substances PFOS, PFOA, PFDA et PFNA sont elles présentes, si oui, à quelle fréquence ?

Exercice 5 : Analyse des prélèvements de Dijon

L'objectif de cet exercice est d'étudier les substances prélevées à Dijon et d'établir un classement des villes.

Travail à faire :

1. Quelles sont les substances (*substance*) détectées à Dijon, leurs valeurs ? qu'en déduisez-vous ?
2. Etablir un classement des villes par ordre décroissant sur *pfas_ sum*. Le classement prend en compte que les villes dont la somme est supérieure à 1.

Exercice 6 : Graphique "nuage de mots" (word cloud)

L'objectif de cet exercice est de réaliser un graphique "nuage de mots" afin de mettre en évidence le nom des substances (voir Exercice 4, question 2).

Indications

Différents packages sont nécessaires : **tm** permet d'accéder aux fonctions de nettoyage du texte ; **dplyr** permet de traiter les données, et **wordcloud** permet de créer des graphiques nuages de mots.

- words = `substance_counts$pfas_values_substance` → Utilise le nom des substances.
- freq = `substance_counts$n` → Utilise la fréquence d'apparition.
- colors = brewer.pal(8, "Dark2") → Applique une palette de couleurs.
- random.order = FALSE → Les fréquences les plus présentes apparaissent au centre.

Travail à faire : graphique Nuage de mots

Réaliser un graphique "nuage de mots" à partir des 50 premières substances extraites de la colonne *pfas_ values*.

- La substance la plus fréquente doit être au centre du graphique.
- Une substance est affichée si sa fréquence d'apparition est supérieure à 3.
- Choisir une palette de couleur adaptée.
- Ajouter une légende.