RDSM 24-25
PROJET D'ANALYSE MULTIVARIEE DES POLLUANTS ATMOSPHERIQUES A PERIGUEUX en 2024

Consignes :

- Produire seul ou en binôme au format html ou pdf un document construit avec RMarkdown qui contient toutes les analyses demandées, avec les codes visibles (en html, possible de proposer une case à cocher "hide/show").
- Déposer sur le moodle du cours les 2 fichiers (le rmd + le résultat), avant le 15/01/2025, à 19h.

Listes des paquets qui peuvent servir, à charger selon les besoins :

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(lubridate) # pour la gestion des format date
library(MASS)
library(FactoMineR)
library(corrplot)
library(factoextra)
library(plotly)
 # manipulation pour les données manquantes
library(naniar)
# si on veut faire de l'ACP en dynamique, ie explorer et voir directement le rendu
library(Factoshiny)
```

1. Introduction - Sources des données et préparation

- A/ Infoclimat

Infoclimat est une association créée en 2003 pour préserver un site internet d'aggrégation de données météo et promouvoir la science participative. Le fichier pdf joint explique le projet INFOCLIMAT et ses actions.

Infoclimat a installé une station météo au beau milieu de Périgueux il y a presque 10 ans dèjà désormais. Vous pouvez trouver la climatologie de 2023 ici : https://www.infoclimat.fr/climatologie/annee/2023/perigueux/valeurs/000BP.html

page Open Data du site ici : https://www.infoclimat.fr/opendata/

- B/ ATMO France

Des données sont accessibles sur la pollution de l'air par l'organisme ATMO France, Fédération des Associations agréées de surveillance de la qualité de l'air ici : https://opendata.atmo-na.org/

En particulier, on peut trouver des données pour Périgueux ici : https://opendata.atmo-na.org/dataset/mesures/horaire/08

De *Infoclimat*, on obtient des données horaires entre le 8/11/2023 et le 07/11/2024 (un an maximum). De *opendata-atmo-na.org* pour la même période du 8/11/23 au 07/11/24, on obtient des données de mesures de

pollution : NO2, NO, O3, PM2.5 et PM10. A noter que PM2.5 contient trop de valeurs manquantes, on décide de ne pas conserver cette variable.

Toutes les données, mises dans le fichier "atmo_polluants_Périgueux.xlsx" et fusionnées, sont dans l'onglet "atmo".

La table globale a été importée dans R, travaillée, puis exportée dans **"ATMO-RDSM.rda"** et peut être rechargée par la commande suivante :

```
load("ATMO-RDSM.rda")
```

Il y a des données manquantes :

```
glimpse(df)
```

```
Rows: 8,696
Columns: 28
$ date_debut              <dttm> 2023-11-08 00:00:00, 2023-11-08 01:00:00, 2023~
$ date_debut_standard     <dbl> 45238.00, 45238.04, 45238.08, 45238.12, 45238.1~
$ NO2                     <dbl> 3, 3, 3, 6, 4, 6, 8, 9, 13, 7, 5, 3, 4, 3, 3, 3~
$ PM2.5                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ O3                      <dbl> 34, 32, 34, 31, 26, 22, 25, 47, 48, 54, 58, 64,~
$ NO                      <dbl> 3.6, 3.9, 3.4, 5.9, 3.7, 7.1, 9.7, 11.7, 20.3, ~
$ PM10                    <dbl> 11, 11, 10, 10, 12, 12, 12, 9, 10, 10, 9, 8, 9,~
$ temperature             <dbl> 7.1, 7.5, 8.1, 8.4, 8.6, 8.9, 8.7, 8.9, 9.6, 10~
$ pression                <dbl> 1018.6, 1018.1, 1018.5, 1018.3, 1018.6, 1018.6,~
$ pression_variation_3h   <dbl> 0.0, 0.0, 0.0, -0.3, 0.5, 0.1, 0.3, 0.3, 0.5, 0~
$ humidite                <dbl> 94, 95, 95, 95, 95, 95, 95, 94, 94, 91, 89, 87,~
$ point_de_rosee          <dbl> 6.1, 6.7, 7.2, 7.8, 7.8, 8.3, 7.8, 8.3, 8.9, 9.~
$ vent_moyen              <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.6, 3.~
$ vent_rafales            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ vent_rafales_10min      <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 8.0, 9.~
$ vent_direction          <dbl> 110, 110, 110, 110, 110, 110, 110, 110, 150, 92~
$ temperature_min         <dbl> 7.1, 7.1, 7.1, 7.1, 7.1, 7.1, 7.1, 7.1, 7.1, 7.~
$ temperature_max         <dbl> 7.1, 7.5, 8.1, 8.4, 8.6, 8.9, 8.9, 8.9, 9.6, 10~
$ pluie_1h                <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
$ pluie_3h                <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
$ pluie_6h                <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
$ pluie_12h               <dbl> 0.6, 0.6, 0.6, 0.6, 0.6, 0.0, 0.0, 0.0, 0.0, 0.~
$ pluie_24h               <dbl> 1.6, 1.6, 1.2, 0.8, 0.8, 0.8, 0.6, 0.6, 0.6, 0.~
$ pluie_cumul_0h          <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
$ pluie_intensite         <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
$ pluie_intensite_max_1h  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ jour                    <ord> mercredi, mercredi, mercredi, mercredi, mercred~
$ heure                   <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1~
```

```
print(miss_var_summary(df), n=35)
```

```
# A tibble: 28 x 3
   variable              n_miss pct_miss
   <chr>                  <int>    <num>
 1 PM2.5                   7955    91.5
 2 NO2                      545     6.27
 3 NO                       542     6.23
 4 PM10                     300     3.45
 5 date_debut                 0     0
 6 date_debut_standard        0     0
```

```
 7 O3                          0     0
 8 temperature                 0     0
 9 pression                    0     0
10 pression_variation_3h       0     0
11 humidite                    0     0
12 point_de_rosee              0     0
13 vent_moyen                  0     0
14 vent_rafales                0     0
15 vent_rafales_10min          0     0
16 vent_direction              0     0
17 temperature_min             0     0
18 temperature_max             0     0
19 pluie_1h                    0     0
20 pluie_3h                    0     0
21 pluie_6h                    0     0
22 pluie_12h                   0     0
23 pluie_24h                   0     0
24 pluie_cumul_0h              0     0
25 pluie_intensite             0     0
26 pluie_intensite_max_1h      0     0
27 jour                        0     0
28 heure                       0     0
```

On décide de supprimer la variables PM2.5, et ensuite de nettoyer les lignes où des NA subsistent

```
df$PM2.5 <- NULL
print(miss_var_summary(df), n=35)
```

```
# A tibble: 27 x 3
   variable               n_miss pct_miss
   <chr>                   <int>    <num>
 1 NO2                       545     6.27
 2 NO                        542     6.23
 3 PM10                      300     3.45
 4 date_debut                  0     0
 5 date_debut_standard         0     0
 6 O3                          0     0
 7 temperature                 0     0
 8 pression                    0     0
 9 pression_variation_3h       0     0
10 humidite                    0     0
11 point_de_rosee              0     0
12 vent_moyen                  0     0
13 vent_rafales                0     0
14 vent_rafales_10min          0     0
15 vent_direction              0     0
16 temperature_min             0     0
17 temperature_max             0     0
18 pluie_1h                    0     0
19 pluie_3h                    0     0
20 pluie_6h                    0     0
21 pluie_12h                   0     0
22 pluie_24h                   0     0
23 pluie_cumul_0h              0     0
24 pluie_intensite             0     0
```

```
25 pluie_intensite_max_1h      0     0
26 jour                        0     0
27 heure                       0     0
```

```
df <- df %>% na.omit()
print(miss_var_summary(df), n=35)
```

```
# A tibble: 27 x 3
   variable               n_miss pct_miss
   <chr>                   <int>    <num>
 1 date_debut                  0        0
 2 date_debut_standard         0        0
 3 NO2                         0        0
 4 O3                          0        0
 5 NO                          0        0
 6 PM10                        0        0
 7 temperature                 0        0
 8 pression                    0        0
 9 pression_variation_3h       0        0
10 humidite                    0        0
11 point_de_rosee              0        0
12 vent_moyen                  0        0
13 vent_rafales                0        0
14 vent_rafales_10min          0        0
15 vent_direction              0        0
16 temperature_min             0        0
17 temperature_max             0        0
18 pluie_1h                    0        0
19 pluie_3h                    0        0
20 pluie_6h                    0        0
21 pluie_12h                   0        0
22 pluie_24h                   0        0
23 pluie_cumul_0h              0        0
24 pluie_intensite             0        0
25 pluie_intensite_max_1h      0        0
26 jour                        0        0
27 heure                       0        0
```

Voilà, le fichier est nettoyé :

```
sum(is.na(df))
```

```
[1] 0
```

On crée une variable qui donne le jour de mesure, ainsi que l'heure :

```
df <-df %>% mutate(jour=wday(date_debut,label=TRUE,abbr=FALSE))
df <- df %>% mutate(heure=hour(date_debut))
```

Puis, en cas de besoin, un vecteur des identifiants avec une info sur la date et l'heure :

```
 rownames.df <- paste(df$date_debut %>% day(),
                      df$date_debut %>% month(),
                      df$date_debut %>% hour(), sep=".")
```

2. Travail demandé

2.1 Construire la matrice de corrélation des variables quantitatives, analyser globalement.

2.2 En se focalisant sur le polluant Ozone (O3), identifier quelques variables statistiquement liées à l'ozone et tracer des graphiques pertinents pour visualiser ce lien.

2.2 Calculer les boites à moustaches parallèles de la variable Ozone, en fonction de l'heure de la journée. Commenter. Reprendre en fonction du jour de la semaine. Commenter.

2.3 Faire l'Analyse en Composantes Principales du jeu de données complet.

Dans cette partie, toutes les variables peuvent être utilisées sauf les 2 premières (les dates). Insister sur les commentaires des liens entre les variables. Une fois l'ACP faite, identifier les individus qui ont des coordonnées sur le plan (1-2) :

- en abscisses supérieurs à 3
- en ordonnées supérieurs à 3.

Qu'est ce qui caractérisent ces moments ? justifier-le par une analyse statistique.

2.4 Reprendre l'ACP complète en incluant une seule variable de pluie. Comparer.

2.5 Reprendre une ACP en en incluant une seule variable par groupe de variables identiques (ie une seule variable de pluie, une seule variable de vent....), les autres pouvant être mises en variables supplémentaires. Comparer.

3 Usage de Factoshiny et travail personnel

Il existe un outil d'exploration permettant de lancer des ACP dans un navigateur (en mode dynamique). Vous pouvez essayer de l'utiliser, mais la commande ne doit pas être "active" au moment de la compilation finale du document.

```
require(Factoshiny)
res <- Factoshiny(df[,-c(1,2)])
```

Après avoir défini vous même une problématique en lien avec la pollution de l'air et la météo, et pouvant être abordée avec de l'analyse multivariée, proposer quelques méthodes statistiques, mettez-les en oeuvre et commentez vos résultats.