Statistique descriptive avec R
TP 4

Nous allons travailler sur les données `airquality` intégrées de R.

1. Explorez vos données.

2. La colonne `Day` donne le jour et la colonne `Month` donne le mois. En consultant l'aide de `airquality`, ajoutez une colonne `Date` qui est la date au format `Date` de R. Puis supprimez les colonnes `Day` et `Month`.

3. La fonction `par` permet de contrôler les paramètres globaux du graphique. L'argument `mfrow = c(n, m)` permet d'afficher plusieurs graphiques à la fois : il y aura n lignes et m colonnes, c'est-à-dire $n \times m$ graphiques. L'argument `mar` contrôle la taille des marges. Appelez `par` dans le but d'afficher 4 graphiques avec comme marge `c(4.1, 4, 2, 2)`, puis affichez les vecteurs `Ozone`, `Solar.R`, `Wind` et `Temp` en fonction de `Date`, en utilisant un `pch` de 16 et des couleurs modernes différentes pour chaque graphique.

4. En explorant les données, nous avions pu voir la présence de `NA`. Calculez la proportion de `NA` par colonnes, ce qui permet d'identifier quelles colonnes sont touchées et en quelle proportion.

5. Nous souhaitons estimer la valeur des données manquantes. Affichez `Ozone` et `Temp` en fonction de `Wind` et de `Temp` (4 graphiques) et commentez.

6. Nous allons utiliser une régression linéaire afin d'estimer les données manquantes. Faites la régression linéaire de `Ozone` et de `Solar.R` contre toutes les autres variables et, avec `summary`, itérativement, supprimez les variables explicatives non significatives (on utitilisera le seuil de 5%).

7. On observe que `Ozone` et `Solar.R` s'utilisent mutuellement pour s'estimer. Affichez `Ozone` en fonction de `Solar.R`

8. Est-ce qu'il y a des cas où à la fois `Ozone` et `Solar.R` sont `NA` ? Ces données ne pourront pas être estimées dans un premier temps.

9. Faites une copie de `airquality` dans une variable nommée `aq`, faites la régression linéaire sélectionnée pour la variable expliquée `Ozone`, puis remplacez dans `aq` les données où `Ozone` est `NA` par l'estimation de la régression linéaire, seulement quand `Solar.R` n'est pas `NA`. On utilisera `predict` pour obtenir les estimations de `Ozone` par la régression linéaire.

10. Faites de même pour remplacer données manquantes de `Solar.R` là où c'est possible.

11. Affichez `Ozone` et `Solar.R` en fonction de `Date` avec les données complétées. Les données estimées seront mises en évidence, dans les deux graphiques, elles auront un `pch` de 17 et une couleur sombre. Faites une critique.