Evolution de la température en France

Mouettes Savantes

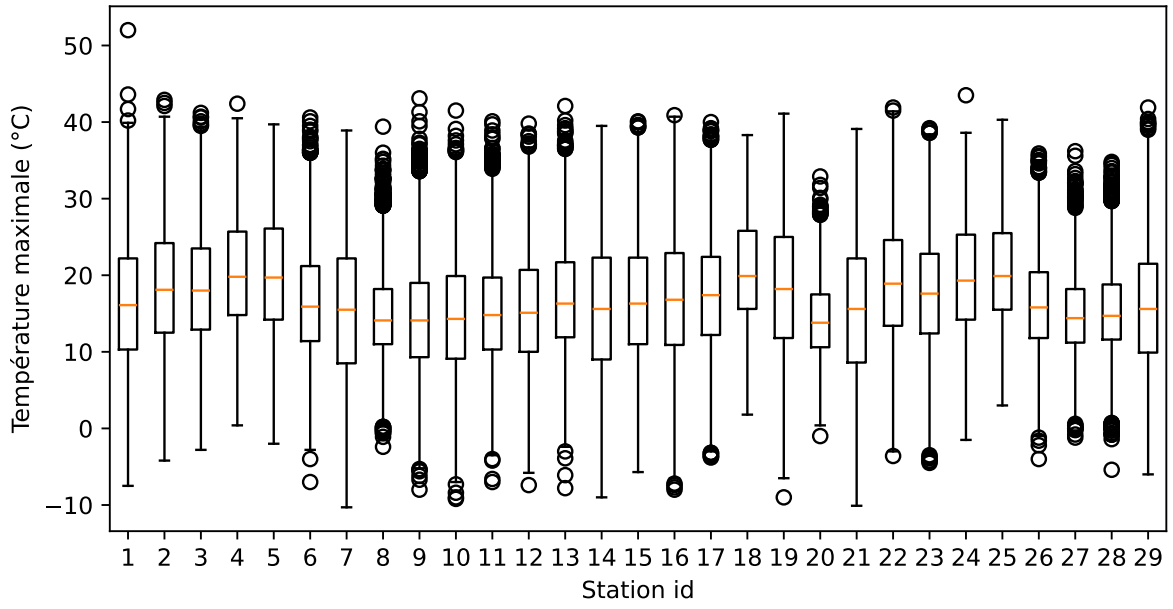
24 au 28 Juin 2024

1 Les données à disposition

On a à notre disposition des relevés de températures en différents endroits en France, plus précisement dans 29 stations de mesures météo.

Dans chacune de ces stations, on a la température maximale journalière entre le 1er octobre 1949 et le 31 mars 2024.

On peut essayer de représenter ces données à l'aide de boîtes à moustaches comme ci-dessous. A votre avis que représente ce graphique ? Ensuite on essayera de reproduire ce graphique avec Python!



On a également un second jeu de données, avec cette fois-ci les températures journalières maximales et minimales mesurées à Rennes entre le 1er janvier 1945 et le 31 décembre 2023. C'est sur ces données de Rennes que l'on va chercher à déterminer s'il y a une tendance, ou non, au fil des ans, sur l'évolution de la température.

2 Quelques notions de statistiques descriptives

La première étape pour analyser un jeu de données consiste à faire ce qu'on appelle des statistiques descriptives. Pour cela, on peut faire des représentations graphiques, comme les boîtes à moustaches vues précedemment.
On peut aussi calculer des résumés numériques, vous en connaissez déjà certains !

Pour les introduire on a besoin d'une notation pour nos observations, qui sont par exemple les températures mesurées à la station météo de Rennes :

$$X = (x_1, x_2, x_3, \ldots, x_n)$$

X ici représente l'ensemble des températures mesurées sur toute la période, c'est une **série de données**. On a n mesures en tout et x_1, par exemple, correspond à la température mesurée le tout premier jour de notre période d'observation. Si vous avez bien compris, pouvez-vous dire à quoi correspond x_i, i étant un nombre entre 1 et n ?

Une fois ces notations en main, on peut définir des indicateurs statistiques, comme par exemple :

- la **moyenne** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} \times (x_1 + x_2 + x_3 + x_4 + \cdots + x_n)$
- la **variance** $\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{1}{n} \times \left((x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2\right)$

La variance est un indicateur de dispersion : "à quel point les données sont éloignées de la moyenne".

Un dernier indicateur dont on aura peut-être besoin, est la **covariance**. Il permet de déterminer à quel point deux séries de données sont dépendants (ou liés). Par exemple, est-ce qu'il y a un lien fort entre les températures mesurées à Rennes et celles mesurées à Paris.
Sa formule est la suivante :

$$\text{cov}(X_{\text{Rennes}}, X_{\text{Paris}}) = \frac{1}{n}\sum_{i=1}^{n}(x_i^{\text{Rennes}} - \bar{X}_{\text{Rennes}})(x_i^{\text{Paris}} - \bar{X}_{\text{Paris}})$$

A partir des définitions que l'on vient de voir, à votre avis, que vaut $\text{cov}(X_{\text{Rennes}}, X_{\text{Rennes}})$?