Python 4
Indicateurs statistiques et représentations graphiques avec pandas

L'objectif de cette fiche est l'étude statistique de données en **Python**. Nous aurons besoin de la bibliothèque **pandas** que l'on importera ainsi :

```
import pandas as pd
```

Pour les graphiques, nous aurons besoin de la bibliothèque **matplotlib.pyplot**.

La première chose pour traiter des données, est d'importer le fichier contenant ces données. Généralement, le fichier sera en format .csv (comma-separated values) et peut être importé via la commande : **tab=pd.read_csv("fichier.csv")**, où **tab** désigne alors le nom du tableau de données.

Petite remarque
Les données du fichier csv initial sont séparées par des ";" (parfois, elles le sont pas des ","), on le précise donc.

1. Sur Teams, télécharger le fichier **FD_NAIS_2022.csv** et l'importer sous le nom **naiss** avec la commande : **naiss=pd.read_csv("FD_NAIS_2022.csv",sep=";",low_memory=False)**
2. Sur Teams, télécharger également le fichier Contenu_etatcivil2022_nais2022.pdf pour comprendre les intitulés des colonnes du fichier source.
3. Placer les deux fichiers, ainsi que le fichier **Python** du TP dans un même dossier.
4. Quelques premières commandes :

Commande **Python**	Résultat
tab	aperçu du tableau de données
tab.head(), tab.head(n)	5 premières lignes, *n* premières lignes
tab.tail(), tab.tail(n)	5 dernières lignes, *n* dernières lignes
tab.shape	taille du tableau

4.a. Quelles sont les données présentes dans ce tableau ?
4.b. Que représente le nombre de lignes du tableau ?

5. **Pour ordonner et trier :**

Commande **Python**	Résultat
tab['Colonne']	extraction de la colonne intitulée **Colonne**
tab.sort_values('Colonne')	ordonne la colonne **Colonne** dans l'ordre croissant
tab[tab['Colonne'] ...]	extraction du sous-tableau obtenu avec les données pour lesquelles la valeur de **Colonne** vérifie le test logique indiqué à la place de ... .

Petite remarque
Si le tableau ne contient qu'une colonne, la commande **tab.sort_values()** permet de l'ordonner.

5.a. Combien de filles sont-elles nées en 2022 ? Combien de garçons ?
5.b. Que permet d'obtenir la commande **naiss[(naiss["DEPDOM"]=="69")&(naiss["DEPNAIS"]=="69")]** ?

5.c. Combien de jumeaux sont-ils nés ?
5.d. Combien de jumeaux sont-ils nés dans le Rhône ?

6. **Indicateurs statistiques :**

Commande **Python**	Résultat
tab.describe()	indicateurs statistiques usuels
tab.count()	effectif
tab.min(), tab.max()	minimum, maximum
tab.mean()	moyenne
tab.std()	écart-type
tab.median()	médiane
tab.sum()	somme

Petite remarque
La commande **sum** fait en fait la somme de chaque colonne...

Compléter les informations suivantes relatives à l'âge des femmes ayant accouché en 2022 (on comptera plusieurs fois celles ayant obtenu des grossesses multiples).

Indicateurs	France et DROM-COM	Rhône	Accouchement de jumeaux	Accouchement de triplés ou plus	Accouchement de garçon né d'une grossesse multiple en août dans le Rhône
Effectif					
Minimum					
Maximum					
Moyenne					
Écart-type					
Médiane					
Premier quartile					
Troisième quartile					

7. **Représentations graphiques :**

Commande `Python`	Résultat
plt.boxplot(tab[Colonne])	boite à moustaches des données de la colonne `Colonne`
plt.hist(tab[Colonne],n)	histogramme des données de la colonne `Colonne` en n classes de même taille
plt.bar(abscisses, ordonnees)	diagramme en barres

Petite remarque

On peut aussi obtenir un histogramme avec des classes de tailles différentes, explicitement choisies. On précise alors `bins=liste_des_bornes` en paramètre à la place de `n`.

7.a. Commande permettant d'obtenir le diagramme de Tukey de l'âge des femmes ayant accouché en 2022.

7.b. Commande permettant d'obtenir l'histogramme de fréquences de l'âge des femmes ayant accouché en 2022, regroupé par classes d'amplitude 5 ans.