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 ?
Le tableau contient l'essentiel des informations sur les accouchements qui ont eu lieu en 2022 (âge de la mère, département de résidence, département de naissance, mois de naissance, nombre d'enfants lors de l'accouchement, sexe des enfants...).

4.b. Que représente le nombre de lignes du tableau ?
Le tableau contient 725997 lignes ; chaque ligne représente un nouveau-né.
Il y a eu 725997 nouveau-nés en 2022.

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 ?
- La commande `naiss[naiss["SEXE"]==2]` permet d'extraire le sous-tableau contenant les naissances des filles. Il contient 354688 lignes.
  354688 filles sont nées en 2022.
- La commande `naiss[naiss["SEXE"]==1]` permet d'extraire le sous-tableau contenant les naissances des garçons. Il contient 371309 lignes.
  371309 garçons sont nés en 2022.

ÉNORME DIFFICULTÉ
- Le tableau étudié ne s'appelle pas **tab**, mais bien **naiss**... C'est à se demander comment certains font pour calculer $f(t)$ quand on leur donne $f(x)$...
- Les tests logiques en **Python** sont : "==", "!=", ">", "<", ">=" et "<=". On pourra relire **la première page du premier cours de Python de première année**.

5.b. Que permet d'obtenir la commande `naiss[(naiss["DEPDOM"]=="69")&(naiss["DEPNAIS"]=="69")]` ?

Cette commande permet d'extraire le tableau des données relatives aux femmes qui résident dans le département du Rhône et qui ont accouché dans ce même département.

5.c. Combien de jumeaux sont-ils nés ?
La commande `naiss[naiss["NBENF"]==2]` permet d'extraire le sous-tableau contenant les naissances des jumeaux. Il contient 22026 lignes.
22026 jumeaux sont nés en 2022.

★Subtile...★
Le nombre d'enfants est un nombre... En revanche, le code du département est du texte (penser à la Corse)...

5.d. Combien de jumeaux sont-ils nés dans le Rhône ?
La commande `naiss[(naiss["NBENF"]==2)&(naiss["DEPNAIS"]=="69")]` permet d'extraire le sous-tableau contenant les naissances des jumeaux dans le département du Rhône. Il contient 943 lignes.
943 jumeaux sont nés dans le département du Rhône en 2022.

Petite remarque
Ce nombre impair doit s'expliquer par le décès d'au moins un des jumeaux lors de l'accouchement.

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	725997	26303	22026	426	31
Minimum	17	17	17	18	22
Maximum	46	46	46	46	45
Moyenne	30,75	31,45	31,83	31,32	32,1
Écart-type	5,39	4,99	5,26	5,25	4,6
Médiane	31	31	32	32	32
Premier quartile	27	28	28	29,5 27,25	
Troisième quartile	34	35	35	35	34,5

- Première colonne.
  La commande `naiss["AGEXACTM"].describe()` renvoie les informations utiles pour la première colonne du tableau.
- Deuxième colonne.
  Il faut extraire un sous-tableau d'un autre...
  - Soit en deux temps : `tab=naiss[naiss["DEPNAIS"]=="69"]` puis `tab["AGEXACTM"].describe()` ;
  - Soit directement : `naiss[naiss["DEPNAIS"]=="69"]["AGEXACTM"].describe()`.
- Troisième colonne.
  `naiss[naiss["NBENF"]==2]["AGEXACTM"].describe()`
- Quatrième colonne.
  `naiss[naiss["NBENF"]>=3]["AGEXACTM"].describe()`
- Cinquième colonne.
  `naiss[(naiss["DEPNAIS"]=="69")&(naiss["NBENF"]>=2)&(naiss["SEXE"]==1)&(naiss["MNAIS"]==8)]["AGEXACTM"].describe`

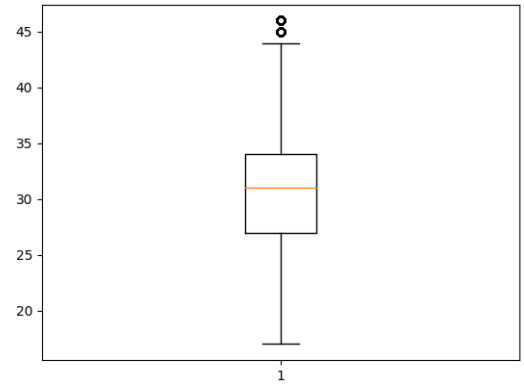
7. **Représentations graphiques :**

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`.

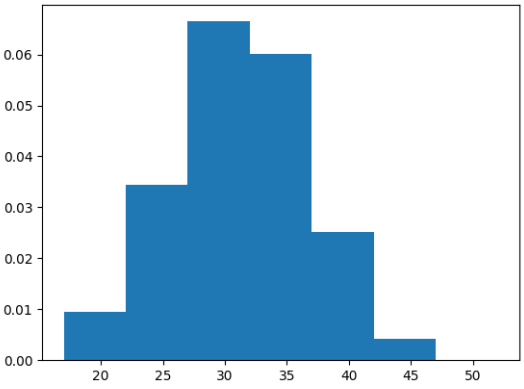
7.a. Commande permettant d'obtenir le diagramme de Tukey de l'âge des femmes ayant accouché en 2022.

```
plt.boxplot(naiss["AGEXACTM"])
plt.show()
```



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.

```
plt.hist(naiss["AGEXACTM"],[17+5*k for k in range(8)],density=True)
plt.show()
```



✘ Attention !

Comme pour tout histogramme de fréquences, ce n'est pas la hauteur du rectangle qui donne la fréquence de la classe observée, mais son aire !