Python 4
Statistiques descriptives et régression linéaire

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** et de **numpy** pour certains calculs.

Statistiques descriptives avec pandas

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).

1. Sur Teams, télécharger le fichier **notes_promo_2024.csv** et l'importer sous le nom **notes** avec la commande :

```
notes=pd.read_csv("notes_promo_2024.csv",sep=";",low_memory=False)
```

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

2. Placer ce fichier, ainsi que le fichier **Python** du TP dans un même dossier.
3. Quelques premières commandes :

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

3.a. Combien le tableau contient-il de lignes et de colonnes ?

3.b. Écrire et exécuter le code suivant :

```
for x in notes:
    print(x)
```

Que permet de faire ce programme ? Quelles informations obtient-on ?

3.c. Quelles sont les données présentes dans ce tableau ?

4. **Pour ordonner et trier :**

Commande **Python**	Résultat
`notes[['Colonne']]`	extraction de la colonne intitulée **Colonne** (possible de renommer la colonne)
`notes[['Colonne1','Colonne2']]`	extraction des colonnes intitulées **Colonne1** et **Colonnes2** (possible de renommer le tableau obtenu)
`notes.sort_values('Colonne')`	ordonne la colonne **Colonne** dans l'ordre croissant

> **★Subtil...★**
> Il y a une différence entre `notes['Colonne']` et `notes[['Colonne']]`...

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

4.a. Écrire une commande permettant d'afficher les notes du DS0.

4.b. Écrire une commande permettant d'afficher les noms et prénoms des étudiantes et étudiants ainsi que leur moyenne aux concours.

4.c. Que permet d'obtenir la commande `notes[notes['Prenom']=='Julie']` ?

4.d. Écrire une commande permettant d'afficher la note obtenue par Julie lors de l'épreuve 1 du CB1.

> **Petite remarque**
> La commande `tableau.at['ligne','colonne']` répond à la question...

4.e. Écrire une commande permettant d'obtenir les étudiantes et étudiants dont la moyenne aux écrits TOP5 est supérieure ou égale à 13.

5. **Indicateurs statistiques sur un tableau nommé `tableau` :**

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

5.a. Écrire une commande permettant d'obtenir les indicateurs statistiques sur la moyenne annuelle ainsi que sur la moyenne aux épreuves écrites.

5.b. Écrire une commande permettant d'obtenir le nombre d'étudiantes et étudiants ayant eu une moyenne aux écrits TOP5 supérieur ou égale à 13.

6. **Représentations graphiques :**

Commande `Python`	Résultat
`plt.boxplot(tableau['Colonne'])`	boite à moustaches des données de la colonne `Colonne`
`plt.hist(tableau['Colonne'],...)`	histogramme des données de la colonne `Colonne` ; "..." est soit vide, soit le nombre de rectangles, soit une liste de bornes des rectangles
`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`.

Représenter quelques graphiques sur différentes plages de données au choix.

Régression linéaire sur un cas pratique

On s'intéresse ici à la corrélation entre les notes obtenues par les étudiantes et étudiants durant l'année avec celles obtenues lors des écrits.

7. A la suite du programme précédent, recopier et exécuter le programme suivant :

```
x=notes["AnneeTotal"]
y=notes["EcritsTotal"]
plt.plot(x,y,"b+")
plt.xlabel("moyennes en CPGE")
plt.ylabel("moyennes aux écrits")
plt.show()
```

8. Exécuter les instructions `np.std(x)` et `x.std()`. Commenter.

9. Écrire une ligne de commande permettant de calculer l'écart-type de `x` "à la main".

10. Écrire les instructions permettant de déterminer les coordonnées du point moyen du nuage de points ainsi représenté puis de le placer sur le nuage de points.

11. Écrire les instructions permettant de représenter la droite de régression linéaire obtenue par la méthode des moindres carrés ; puis calculer le coefficient de corrélation linéaire. Commenter.