EXPLOITATION DE DONNÉES

Objectifs

Ce TP est une introduction à l'exploitation de données, il a pour but de vous donner les premiers outils vous permettant d'exploiter les données que vous allez rassembler pour vos TIPE. C'est un complément au livret *Notions de statistiques pour les TIPE.*

I À faire et à lire avant de commencer

Nous allons utiliser des modules python qui ne sont pas forcement installés ou à jour sur les ordinateurs que vous utilisez.

À programmer 1 (Mise à jour des modules)**.**
En console recopiez et valider les commandes suivantes :

```
pip install pandas --upgrade --user
pip install matplotlib --upgrade --user
pip install scipy --upgrade --user
```

À programmer 2 (Importation des modules)**.**
Importer les modules

- `matplotlib.pyplot` avec l'alias `plt`.
- `pandas` avec l'alias `pd`.
- `numpy` avec l'alias `np`

Vous trouverez sur https://cahier-de-prepa.fr/spebio2-champollion/ ou dans la zone Échange du réseau deux fichiers à récupérer :

1. `PenguinData.csv`
2. `AutransMeteo.csv`

Pour faciliter leur utilisation vous devez enregistrer ces fichiers dans le même répertoire que le fichier python de votre TP.

Le fichier `PenguinData.csv` contient des mesures effectuées sur une population de manchots en Antarctique. Le deuxième fichier `AutransMeteo.csv` rassemble des données météorologiques mesurées à la station d'Autrans. Vous pouvez commencer par visualiser ces fichiers en les ouvrant avec un tableur.

Les colonnes du premier fichier sont nommées de façon explicite ; les données sont en millimètres ou en grammes. Les colonnes du fichier de données météorologiques sont la date (format AAAAMMJJ), la température minimale journalière et la température maximale (en degrés Celsius).

À programmer 3 (Importation des données)**.**
Recopier et exécuter les deux lignes suivantes

```
DonneesManchot=pd.read_csv('PenguinData.csv')
DonneesMeteo=pd.read_csv("AutransMeteo.csv")
```

Il existe maintenant deux variables `DonneesManchot` et `AustransMeteo`. Ce sont des tableaux de données pandas. On les manipule de façon assez naturelle, comme des tableaux ou des dictionnaires dont les clefs sont les noms des colonnes.

À programmer 4.
Recopier et valider :

```
plt.plot(AutransMeteo["Tmin"])
plt.plot(AutransMeteo['Tmax'])
plt.show()
```

Vous devez obtenir sur un graphe, pas très joli, les courbes des températures minimales et maximales à Autrans.

II Graphes

II.1 Boites à moustaches

La commande `plt.boxplot([X,Y,...]` permet de réprésenter les séries X, Y,... sous forme de boites à moustaches.

À programmer 5.
Faire afficher les boites à moustaches des données températures minimales et températures maximales

`DonneesMeteo[['Tmin','Tmax']]` est un tableau formé des colonnes 'Tmin' et 'Tmax' de la table de données, on peut alors écrire `plt.boxplot(DonneesMeteo[['Tmin','Tmax']])`

À programmer 6.
Faire afficher les boites à moustaches des données des longueurs des nageoires et des longueurs des becs de la population de manchots.

À programmer 7.
Reprendre les exercices précédents en remplaçant la commande `plt.boxplot` par la commande `plt.violinplot`

II.2 Nuage de points

À programmer 8 (À recopier)**.**

```
plt.scatter(DonneesMeteo["Tmin"],DonneesMeteo["Tmax"],marker='x')
```

À programmer 9.
Afficher le nuage de points Longueur Bec/Longueur nageoire.

La commande `a,b=np.polyfit(DonneesMeteo["Tmin"],DonneesMeteo["Tmax"],1)` permet de calculer les coefficients de la droite de régression linéaire pour le nuage de points.
(a est le coefficient directeur et b l'ordonnée à l'origine)

À programmer 10 (Droite de régression linéaire)**.**
Faire afficher les droites de régression linéaire superposées aux deux nuages de points précédents.

III Statistiques

À programmer 11.
Recopier le programme suivant :

```
X1=DonneesManchots[DonneesManchots["Espece"]=="Pygoscelis adeliae"]["Masse"]
X2=DonneesManchots[DonneesManchots["Espece"]=="Pygoscelis papua"]["Masse"]
plt.boxplot([X1,X2])
plt.show()
```

Les masses des manchots de ces deux espèces semblent être différentes. Les commandes suivantes permettent d'effectuer un test de Student.

```
from scipy.stats import f_oneway
_,pvalue=f_oneway(X1,X2)
print(pvalue)
```

À programmer 12.
En vous aidant du livret, énoncer l'hypothèse nulle, effectuer le test et formuler une conclusion.

À programmer 13.
Effectuer un test pour savoir si les températures minimales sont significativement différentes des températures maximales.

IV Pour aller plus loin

IV.1 Filtrage des données

On a vu dans 11 un premier exemple de filtrage. La ligne :

```
X1=DonneesManchots[DonneesManchots["Espece"]=="Pygoscelis adeliae"]["Masse"]
```

nous permet de récupérer la liste des masses des manchots qui respectent le critère :
`DonneesManchots["Espece"]=="Pygoscelis adeliae"`.

À programmer 14.
Faire un test de Student pour tester si les manchots de l'île "Biscoe" ont en moyenne une masse significativement différente de ceux de l'île "Torgersen".

À programmer 15.
Affiner le script précédent pour tester les manchots **d'une même espèce** sur deux îles différentes.
Pour écrire un "et" logique entre deux conditions on utilise le symbole &.

IV.2 Moyennes et moyennes glissantes

Dans l'exercice 4, nous avons tracé les graphes des séries temporelles de températures. Comme ces séries sont très longues et comportent beaucoup de données qui varient de jour en jour, nous n'obtenons pas un graphe très lisible, et nous ne pouvons pas voir les tendances à long terme.

IV.2.a Moyennes

À programmer 16.
Écrire une fonction `moyenne(L,N)` qui étant donné une liste de données L et un entier naturel non nul N, découpe la liste en sous-listes de longueur N, puis renvoie la liste des moyennes successives calculées sur une période N. Si la liste totale a une longueur qui n'est pas un multiple de N, on ne prendra pas en compte les valeurs supplémentaires.
Par exemple `moyenne([1,1,2,4,6],2)` doit renvoyer `[1,3]`

À programmer 17.
Utiliser la fonction précédente pour faire afficher l'évolution des températures minimales annuelles moyennes.

IV.2.b Moyennes glissantes

Pour une série de données $[u_0, u_1, \dots, u_L]$, on définit la série des moyennes glissantes de longueur N par

$$v_i = \frac{1}{N} \sum_{k=0}^{N-1} u_{i-k}$$

Sur papier 1.
Pour quelles valeurs de i cette moyenne est-elle définie ? Pouvez-vous justifier le terme de "moyenne glissante" ?

À programmer 18.
Écrire une fonction `moyenne_glissante(L,N)` qui étant donné une liste de données L et un entier naturel non nul N, renvoie la liste des moyennes glissantes de longueur N
Par exemple `moyenne_glissante([1,1,2,4,6],3)` doit renvoyer `[4/3,7/3,4]`

À programmer 19.
Utiliser la fonction précédente pour faire afficher l'évolution des températures minimales en calculant les moyennes glissantes sur 30 jours, sur un an et sur 5 ans.

IV.3 Améliorer les figures : couleurs, légende, titre . . .

Utiliser l'aide pour rajouter des titres, des légendes . . . dans les graphiques précédents.