Analyse de données - TP 1

ISEP Paris – 14 Février 2022

Instructions : Préparez un rapport incluant le code source et vos résultats, et déposez-le sur Moodle. Pas plus de 2 personnes par groupe. N'oubliez pas de mettre les 2 noms sur le rapport, ou de faire 2 rendus.

Bibliothèques

Pour ce TP, vous aurez potentiellement besoin des bibliothèques suivantes : matplotlib, numpy, pandas, scipy, math :

```
import matplotlib as plt
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from scipy import stats
from math import sqrt,pi,exp
```

A Manipulation de fichiers CSV, calculs et affichages simples

Dans cet exercice, vous allez manipuler un fichier CSV contenant des caractéristiques de plusieurs espèces de cacatoès, un cousin du perroquet et le perruche endémique en Océanie. L'objectif est de vous apprendre à ouvrir un fichier CSV avec Python, et à afficher des informations simples sur ce fichier ou certaines de ses colonnes.

Le fichier "cockatoo.csv" contient les colonnes suivantes :

- Species = le nom de l'espèce
- LifeSpan = durée de vie moyenne
- Size = la taille moyenne de l'animal (cm)
- Weight = le poids moyen de l'animal (g)
- ClutchSize = Le nombre moyen d'oeufs par couvée
- HatchTime = la durée avant l'éclosion des oeufs

1. Importez les données en utilisant la fonction *pd.read_ csv()* de la bibliothèque pandas. Vous ferez en particulier attention à indiquer le bon chemin pour le fichier, et aux paramètres pour le séparateur de colonnes. Vous stockerez les données importées dans une variable nommée **df**.

2. Une fois les données importées, à l'aide de la fonction *print()*, affichez les résultats des commandes *print(df.head())* et *print(df.shape)*. A quoi servent la *head()* et l'attribut *shape* des données importées par pandas ? En quoi peuvent-ils être utiles pour analyser des données plus volumineuses ?
3. En utilisant les fonctions de numpy telles que *np.mean()*, répondez aux questions suivantes :
   — Calculez l'espérance de vie moyenne d'un cacatoès toutes espèces confondues. Vous préciserez également la variance et l'écart-type.
   — Calculez la taille médiane toutes espèces confondues.
4. Utilisez la fonction **plt.hist()** pour afficher les histogrammes des durée d'éclosion, du poids, et du nombre moyen d'oeufs par couvée. Vous essayerez de rendre vos histogrammes beaux ! Commentez.

B Données mal formatées

Dans l'exercice précédent, les données csv que vous avez utilisées étaient bien formatées et présentaient peu de problème. En pratique, c'est rarement le cas. Pour illustrer ce problème, dans cet exercice, nous allons nous intéresser à un fichier exporté depuis Excel et qui contient les notes d'analyse de données de l'année dernières et qui ont été anonymisées.

1. En utilisant la librairie pandas, importez le fichier "notes.csv" dans Python. Vous préciserez tous les problèmes rencontrés et les correctifs que vous avez utilisés, que ce soit en changeant les paramètres de la fonction d'importation, ou en modifiant directement le fichier source.
2. Répondez aux questions suivantes :
   — Affichez des histogrammes lisibles pour les notes des TP1, 3 et 7. Commentez. (TP se dit "lab" en anglais)
   — En utilisant Python, indiquez la note moyenne, minimale, maximale et l'écart-type pour les projets.
   — Affichez un diagramme en camembert pour le GPA.
3. En utilisant les outils statistiques à votre disposition et en détaillant vos calculs et vos justifications, répondez à la questions suivante : La moyenne de TP est-elle significativement différent de la moyenne à l'examen final ?

C Données issues d'un texte

Vous tomberez parfois aussi sur des données qui ne sont pas présentées sous forme d'un fichier, mais qui sont décrites dans un texte. Il vous faudra alors créer les tableaux de données correspondants en Python. Dans cet exercice, nous vous proposons de travailler sur un petit tableau de notes d'étudiants. Ce tableau, que vous trouverez ci-après, est un tableau de fréquence qui décrit combien d'étudiants ont obtenu chaque note.

Note	6	8	9	10	11	12	13	14	17
Nombre d'étudiants	10	12	48	23	24	48	9	14	22

1. Créez le jeu de données correspondant et affichez-le le sous forme d'un histogramme.
2. Calculez les mesures de tendance centrale et de dispersion pour ces données.
3. Expliquez pourquoi cette série est une distribution bimodale.

D Analyse de QI

Dans cette exercice, on cherche à évaluer les effets de la malnutrition sur le QI. Le QI des humains suit une loi normale avec un score moyen de 100 et un écart-type de 15 points. A partir de cette information, nous allons chercher à savoir si la malnutrition a un effet significatif sur le QI.

1. Avec la librairie pandas, ouvrez le fichier *malnutrition.csv* qui contient les scores de QI de plusieurs personnes ayant souffert de malnutrition.
2. Combien de personnes contient cet échantillon ?
3. Calculez la moyenne et l'écart-type de cet échantillon.
4. A l'aide des outils statistiques à votre disposition, que pouvez-vous conclure sur l'effet de la malnutrition sur le QI lorsque vous comparez ces résultats avec ceux de la population générale. Vous justifierez votre réponse.