TP 2 : Explorer et analyser des données avec Python sur un environnement Linux

Objectifs :

- L'objectif de ce TP est de vous familiariser avec le processus d'exploration et de nettoyage de données en utilisant Python sur Linux. Vous travaillerez avec un ensemble de données pour identifier et corriger les incohérences, les valeurs manquantes, et les anomalies, puis vous appliquerez des techniques de base d'analyse de données.

Modalités pédagogiques :

Travail individuel

Durée : 1 jour

> ⚠️ **Essayez d'utiliser ChatGPT comme dernière source d'information**

Pré-requis :

- VsCode ou Pycharm installé sur linux
- Bibliothèques Python `pandas` et `matplotlib` installées. Vous pouvez les installer via pip :

```
pip install pandas matplotlib
```

- Un fichier CSV : `Mental Health Dataset.csv` à télécharger depuis le lien ci-dessous :

https://www.kaggle.com/datasets/bhavikjikadara/mental-health-dataset

Consignes :

Partie 1 : Chargement et Exploration des Données

- Utilisez `pandas` pour charger les données à partir du fichier CSV dans un DataFrame en suivant ces étapes :
  - Déplacez le .csv dans le même répertoire où se trouve votre projet python.
  - Utilisez le code suivant dans un script Python via un terminal Linux:

```
import pandas as pd

# Chemin relatif (le fichier est dans le même dossier que le script)
csv_file_path = 'Mental Health Dataset.csv'

# Chargement des données
data = pd.read_csv(csv_file_path)

# Aperçu
print(data.head())
```

  - Expliquez chaque ligne du script ci-dessus.
  - Expliquez les résultats affichés à l'écran.

Opérations de base:

  - Calculer le total des personnes (femmes et hommes) concernés par cette étude en se basant sur les filtres ci-dessous:
    - **Filtre 1** : basées aux USA
    - **Filtre 2**: Change Habits (No)

- **Filtre 3**: Mood swings (High)
- **Filtre 4**: Work Interest (Maybe)
- **Filtre 5**: Social weakness (Yes)

Partie 2 : Filtrage des données et calculs statistiques

dataset :

> StudentsPerformance.csv

- Pour effectuer des calculs statistiques sur une BDD, il faut tout d'abord filtrer les données selon certaines critères. Dans cette partie vous allez effectuer des opérations de statistique de base pour l'analyse de données:

1. **Tendances Centrales:**

- **Moyenne** : La somme des valeurs divisée par le nombre de valeurs.
- **Médiane** : La valeur centrale dans un ensemble de données ordonnées.
- **Mode** : La valeur la plus fréquente dans un ensemble de données
  - En se basant sur le script ci-dessous, adaptez-le pour calculer et afficher la moyenne:

```
import numpy as np
import pandas as pd
csv_file_path='Mental Health Dataset.csv'
data =  pd.read_csv(csv_file_path)
total_females = data[data['Gender'] == 'Female'].shape[0]
print(total_females)
52514
```

- Utilisez une liste de données de votre choix pour calculer la Médiane et de mode avec des scripts Python, vous pouvez télécharger des Datasets sur `Kaggle` ou utiliser une autre source.

2. **Mesure de Dispersion:**

- **Variance** : c'est une mesure statistique qui décrit à quel point les valeurs d'un ensemble de données sont éloignées de la moyenne (aussi appelée l'espérance mathématique) de cet ensemble. En d'autres termes, elle vous aide à comprendre à quel point les données sont dispersées (éloignées) par rapport à la moyenne.
- **Écart-Type** : **la racine carrée de la variance.** Il s'agit d'une mesure **plus facile à interprét**er, car elle est exprimée dans la même unité que les données originales !

3. **La corrélation :**

C'est un coefficient qui permet de voir si les données sont cohérentes par rapport à d'autres données, par exemple, admettons que vous voulez analyser le temps passé à étudier en dehors des heures de formation et comparer ce temps avec les notes obtenues:

Coeff=1 : Plus vous révisez chez vous mieux sont vos notes ⇒ Corrélation +

Coeff = -1 : Plus vous révisez pire sont vos notes ⇒ Corrélation -

Coeff = 0 : Pas de corrélation

- En se basant sur ce script calculer le coefficient de corrélation sur un dataset de votre choix :

```
#Ajoutons une autre colonne pour l'exemple de corrélation
df['Autre_Valeurs'] = [15, 25, 25, 35, 45, 55]
#Calcul de la corrélation
correlation = df['Valeurs'].corr(df['Autre_Valeurs'])
print(f"Corrélation: {correlation}"
```

Partie 3: Analyse des données

Faites une analyse sur les études de la santé mentale des femmes et des hommes en se basant sur les tableaux ci-dessous:

1- Statistiques pour les femmes :

- Donnez les scripts Python qui permettent de faire les calculs du total **des femmes** concernées par cette étude pour l'année 2014 seulement par pays et par mois.
- Remplissez le tableau par les valeurs trouvées.

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Mois							
Août							
Septembre							
Octobre							
Novembre							

2- Statistiques pour les hommes :

- Donnez les scripts Python qui permettent de faire les calculs du total **des hommes** concernés par cette étude pour l'année 2014 seulement par pays et par mois.
- Remplissez le tableau par les valeurs trouvées.

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Mois							
Août							
Septembre							
Octobre							
Novembre							

3- Filtres:

- Donnez les scripts Python qui permettent d'effectuer des filtres en se basant sur les tableaux ci-dessous pour **l'année 2015 pour les hommes uniquement** et remplissez les cases par les valeurs trouvées.

Self-Employed: (Yes)

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Mois							
Août							
Septembre							
Octobre							
Novembre							

Self-Employed: (No)

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Mois							
Août							
Septembre							
Octobre							
Novembre							

Family history: (Yes)

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Mois							
Août							
Septembre							

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Octobre							
Novembre							

Family history: (No)

Pays	USA	Poland	Australia	Canada	South Africa	Sweden	Netherla
Mois							
Août							
Septembre							
Octobre							
Novembre							

Livrables attendus

- Script Python complet avec :
  - chargement du dataset
  - Le filtrage
  - calculs statistiques (moyenne, médiane, mode, variance, corrélation)
- Fichier texte avec explication des résultats
- Tableaux de la section 3 : Analyse des données