Plan du cours (7h) pour UPPA:

Matin :

1. Présentation de HUPI (lien) : 1h
2. Prise en main de Python (lien) : 1h
3. Présentation de l'objectif de la formation (lien) : 30 min
    a. Présentation des sujets d'évaluation

pausa (30 min)

4. Présentation des méthodes d'analyse de données (lien) : 1h 30
    a. Analyse univariée
    b. Analyse bivariée
    c. Tests de corrélation
5. Présentation des méthodes de traitement des valeurs manquantes (lien) : 30 min

Après-midi :

6. TP Analyse des données : 2h
    a. Lecture des données
    b. Statistiques
    c. Analyse univariée
    d. Analyse bivariée
    e. Tests de corrélation
    f. Visualisation

Aide aux étudiants :

- Dataset "diabetes.csv" :
    - Détection et traitement des données manquantes (les zéros) :
        - Les variables sont les suivantes (Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI').
        - Exception : Les variables pregnancies ainsi que OutCome peuvent contenir des 0, celle ci représente le nombre de grossesses ou bien si les personnes sont diabétique, ce qui est donc pertinent.
- Dataset "house.csv" : préparation des données
    - Convertir les données au bon format

```
df['price'] = df['price'].str.replace(",",".")
df['price'] = df['price'].astype(float)

df['date'] = df['date'].str[-4:]
df['date'] = df['date'].astype(int)
```

    - Traitement des valeurs aberrantes
- Dataset "wine.csv" : jeux de données déséquilibrée
    - Solution : créer des classes de qualité

Support pour des techniques de modélisation :
https://docs.google.com/presentation/d/1ivSoq1UOwHaP_so5ATYtc9abTuedoNwv/edit#slide=id.g2161c2ca195_0_149