**Partiel Analyse de données avec Python**

**Questionnaire à choix multiples:**

* **20 questions : 0,5 point par question**
* **Il peut avoir une ou plusieurs bonnes réponses à chaque question.**
* **Pour chaque question, le demi-point est donné dans le cas où toutes les cases à cocher sont cochées et qu’aucune case n’est cochée à tort.**

**Cas pratique 1 (10 points) :**

* **Analyse sur plusieurs datasets**

**Durée : 3 h**

**Langage de programmation : Python Version : >3.6**

**Packages à utiliser :**

* **Pandas**
* **Numpy**
* **Matplotlib**

Professeur : Bertrand Fournel

Pour toutes les question, *pd* désigne l’alias pour le module pandas.

1. Quelle méthode du module pandas est à utiliser pour ouvrir un fichier csv à partir du chemin suivant ‘./back\_up/dataset.csv’ ?

⬜read\_csv(‘./back\_up/dataset.csv’)

⬜read\_csv(path=‘./back\_up/dataset.csv’)

⬜read\_csv(filepath\_or\_buffer=‘./back\_up/dataset.csv’)

1. Quel type d’objet cette méthode renvoie-t-elle ? ?

⬜Un dataframe pandas

⬜Un dataset pandas

1. Comment afficher les 3 dernières lignes d’un dataframe nommé *df* ?

⬜df.tail(3)

1. Quelle méthode permet d’obtenir depuis un dataframe *df*, les modalités d’une variable catégorielle nommée ‘Produits’ ?

⬜df.Produit.unique()

⬜df[‘Produit’].unique

1. Quelle méthode permet d’étudier le lien de corrélation entre une variable quantitative et une variable catégorielle ?

⬜Test χ² (khi-deux)

⬜Le coefficient de corrélation de pearson

1. Quelle valeur maximale d’une P-value est traditionnellement retenue pour rejeter l’hypothèse nulle ?

⬜0.05

1. Plus la valeur d’une P-Value est grande, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible ?

⬜Faux

1. Etant donné le dataframe *df* et la colonne ‘value\_a’ en nombres flottants, quel code permet de multiplier chaque valeur par 1000 et de modifier le dataframe source ?

⬜df.value\_a = df.value\_a \* 1000

⬜df.value\_a = df.value\_a.map(lambda x : x \* 1000)

⬜df.value\_a.map(lambda x : x \* 1000)

1. Un modèle de régression linéaire est performant si :

⬜le R2 score est proche de 1 sur l’ensemble du dataset

1. Soit le dataframe *df* avec les colonnes ‘class\_ok’ et ‘class’, ces deux colonnes sont en doublons, comment supprimer la colonne ‘class’ en modifiant le dataframe source nommé *df*.

⬜df.drop(‘class’, inplace=True, axis=1)

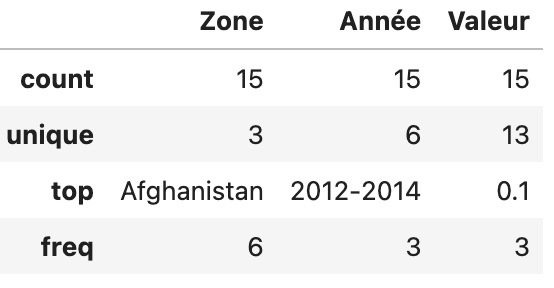
1. Sur un dataframe *df* donné :



Comment récupérer un dataframe en ne gardant que les lignes où la ‘Valeur’ est supérieur à 2 et la colonne ‘Zone’ est ‘Afrique du Sud’

⬜df[(df.Valeur > 2) and (df.Zone == ‘Afrique du Sud’)]

1. Sur le même dataframe que la question précédente, comment obtenir ce résultat ?

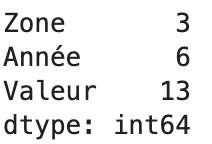


⬜df.info()

1. La sortie de la méthode précédente est une série Pandas (appliqué sur le dataframe vu précédemment) ?

⬜Vrai

1. Sur le même dataframe que la question précédente, comment obtenir ce résultat ?



⬜df.value\_counts()

1. La sortie de la méthode précédente est une série Pandas (toujours appliqué sur le dataframe vu précédemment) ?

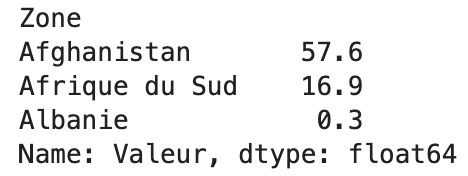
⬜Vrai

1. Sur la dataframe précédent, comment convertir le type de données de la colonne ‘Valeur’ vers le type float (sur la dataframe source) ?

⬜df.Valeur = df.Valeur.astype(float)

1. En tenant compte que le type de donnée de la colonne ‘Valeur’ a été correctement changé en float, comment obtenir la sortie suivante ?

Il s’agit de sommer la colonne ‘Valeur’ en regroupant par Zone.



⬜df['Valeur'].groupby('Zone').sum()

⬜df.groupby('Zone')['Valeur'].sum()

1. Quelle méthode permet de faire une jointure entre deux dataframes df1 et df2 via la clé commune ‘id’ en gardant toutes les lignes de df1 ?

⬜pd.merge(df1, df2, on=’id’, how=’left’)

1. Sur le dataframe vu précédemment comment obtenir la moyenne de la colonne ‘Valeur’ ?

⬜df.Valeur.mean()

⬜df[‘Valeur’].mean()

1. Comment enregistrer proprement le dataframe vu précédemment (sans rajouter ni perdre de colonnes ou de lignes) dans un fichier nommé *values.csv* ?

⬜df.to\_csv(‘value.csv’, index=False)

**CAS PRATIQUE**

A partir des datasets fournis (fichiers .csv), répondre aux questions suivantes :

Calculer:

- la proportion de personnes en état de sous-nutrition ; (2 pts)

- le nombre théorique de personnes qui pourraient être nourries. Vous devriez pouvoir calculer ça à partir de la disponibilité alimentaire mondiale ; (2 pts)

- idem pour la disponibilité alimentaire des produits végétaux ; (2 pts)

- l’utilisation de la disponibilité intérieure, en particulier la part qui est attribuée à l’alimentation animale, celle qui est perdue et celle qui est concrètement utilisée pour l'alimentation humaine. (2 pts)

Trouver :

- les pays pour lesquels la proportion de personnes sous-alimentées est la plus forte en 2017 (1 pt)

- ceux qui ont le plus bénéficié d’aide depuis 2013 (1 pt)

- ceux ayant le plus/le moins de disponibilité/habitant (+ 1 pt bonus)

Agrémenter votre étude de visualisations permettant de répondre à ces questions et d’infirmer vos réponses.