Python – TP1 : Series et DataFrame

Pandas (https ://pandas.pydata.org/) est la librairie incontournable pour manipuler les données. Elle permet de manipuler aussi bien les données sous forme de tables qu'elle peut récupérer ou exporter en différents formats. Elle permet également de créer facilement des graphes

Series

Series est un **tableau unidimensionnel étiqueté** (tableau associatif en PHP) capable de contenir des données de n'importe quel type (entier, chaîne, flottant, objets python, etc.). Les étiquettes des axes sont collectivement appelées **index**.

Dans cet exercice, nous aborderons les sujets suivants :

- Création et initialisation d'une série et de son index
- La taille, la forme et le nombre de valeurs
- Têtes, queues d'une série
- Recherche de valeurs dans un objet Série
- Alignement via des étiquettes d'index
- Opérations arithmétiques sur un objet Series
- Le cas particulier du Not-A-Number (NaN)
- Ré indexation d'un objet Série
- Découper une série

Prénoms	Tailles
Zineb	155
Imane	169
Anas	175
Yassine	180

Tableau 1: T1 - tailles

Le tableau ci-contre représente les tailles de 4 personnes

1. Créez une série S1 (Series) qui représente le tableau T1
2. Affichez la série S1
3. Affichez uniquement les valeurs hébergées (les tailles) par la série S1
4. Affichez uniquement les index (les prénoms) de la Series S1
5. Affichez la taille de Anas (utilisez deux méthodes différentes)
6. Affichez les tailles de Zineb jusqu'à Anas ? Comment expliquez l'ordre valeurs affichées
7. Affichez les deux premières éléments de la série (Zineb , Imane)
8. Affiez les deux derniers éléments de la série (Anas , Yassine)

9. Alignement via des étiquettes d'index

Une différence fondamentale entre une série NumPy ndarray et une série pandas est la capacité d'une série d'aligner automatiquement les données d'une autre série en fonction des valeurs d'étiquette avant d'effectuer une opération.

Soient s1 et s2 deux séries définies par :

```
s1 = pd.Series([10, 20, 30, 40], index=['a', 'b', 'c', 'd'])
s2 = pd.Series([40, 30, 20, 10], index=['d', 'c', 'b', 'a'])
```

- Calculez s3=s1+s2 puis affichez le s3.

10. Opérations arithmétiques sur un objet Series

Les opérations arithmétiques (+, -, -, /, *, et ainsi de suite) peuvent être appliquées à une série ou entre deux objets de la série. Lorsqu'elle est appliquée à une seule série, l'opération est appliquée à toutes les valeurs de cette série.

```
s1 = pd.Series([10, 20, 30, 40], index=['a', 'b', 'c', 'd'])
s2 = pd.Series([40, 30, 20, 10,52], index=['d', 'c', 'b', 'a', 'e'])
# La valeur de l'index e n'existe pas dans s1
```

- Calculez s3=s1*10, puis affichez s3
- Calculez s4=s1+s2, puis affichez s4, Expliquez la valeur de l'index 'e' de la série s4

11. Découper une série

Les objets Series de pandas Series prennent en charge le découpage des données. Tout comme les tableaux NumPy, vous pouvez passer un objet slice à l'opérateur [] de la série pour obtenir les valeurs spécifiées. Les tranches fonctionnent également avec les propriétés .loc[], .iloc[], et .ix et les accesseurs

Soit s une série définie par : s = pd.Series(np.arange(50,60), index=np.arange(10,20))

- Affichez les items de la position 1 à la position 7 avec un delta de 2
- Affichez la série en partant du 5° enregistrement depuis la fin avec un pas de 2
- Affichez tout sauf les 3 derniers
- Affichez que les 3 derniers
- Affichez les 3 premiers et les 4 derniers
- inversez la série

DataFrame

Un DataFrame est une structure de données bidimensionnelle, c'est-à-dire que les données sont alignées de façon tabulaire en lignes et en colonnes.

Caractéristiques de DataFrame

* Les colonnes peuvent être de types différents
* Taille - Mutable
* Axes étiquetés (lignes et colonnes)
* Peut effectuer des opérations arithmétiques sur les lignes et les colonnes.

On peut toute proportion gardée comparer un DataFrame à une feuille excel avec des lignes et des colonnes.

Le tableau 2 ci-contre représente les notes et les qualifications de 10 personnes

	Prenom	Note	Qualification
1	Aymane	12.5	Oui
2	Amina	9.0	Non
3	Hidaya	16.5	Oui
4	El Mehdi	NaN	Non
5	Saad	9.0	Non
6	Achraf	20.0	Oui
7	Hakima	14.5	Oui
8	Imane	NaN	Non
9	Az-eddine	8.0	Non
10	Noura	19.0	Oui

Tableau 2

1. Créez le datafrmae df qui représente le tableau T2
2. Affichez la série df
3. Affichez uniquement les colonnes de df
4. Affichez uniquement les lignes (étiquettes) de df

5. Affichez uniquement les valeurs de df
6. Affichez la taille du df
7. Afficher les deux premières lignes du df
8. Afficher les trois dernières lignes du df
9. Afficher les données de la première colonne
10.
11. Affichez uniquement les index (les prénoms) de la Series S1
12. Affichez la taille de Anas (utilisez deux méthodes différentes)
13. Affichez les tailles de Zineb jusqu'à Anas ? Comment expliquez l'ordre valeurs affichées

II Cas pratiques

Exercice 1 :

Données COVID On récupère les données du COVID par région et par âge à l'adresse "Données hospitalières relatives à l'épidémie de COVID-19" (https//www.data.gouv.fr/). Le fichier voulu : covid-hospit.csv

1. En utilisant la fonction tail (ou head), visualiser la structure du tableau.
2. Quelles sont les différentes données et leur type (utiliser dtypes).
3. Les dates sont considérées comme des chaînes de caractères. Il est plus simple pour réaliser des opérations de convertir la colonne sous forme de dates (utiliser to_datetime).
4. On supprime les colonnes relatives aux départements et au sexe puis on agrège par jour.
5. Enfin on trace les données (utiliser l'option logy pour une échelle logarithmique).
6. Refaire le même graphique pour votre sexe.
7. Faire de même avec les séries différenciées puis avec des séries lissées sur 7 jours

Exercice 2 :

1. Récupérez le fichier temperature.csv.
2. Utiliser la fonction describe sur le dataframe. Que fait cette fonction ?
3. Créer un nouveau dataframe ne contenant que les mois de mars, juin, septembre et décembre et en supprimant les villes de la région 'East'.
4. Récupérer les données sous numpy et calculer la moyenne des températures pour chaque mois.
5. Déterminer aussi la matrice de corrélation entre les 4 mois de l'année.