

# PYTHON – PANDAS

## SOMMAIRE

<b>Sommaire .....</b>	<b>1</b>
<b>PANDAS .....</b>	<b>2</b>
<b>Introduction.....</b>	<b>2</b>
Liste d'outils.....	2
<b>Pandas – Première approche .....</b>	<b>3</b>
Installation .....	3
Présentation théorico-pratique.....	3
Plan de la présentation NoteBook.....	3
Plan détaillé de la présentation NoteBook.....	4
Exercices .....	6
Série 1 .....	6
Série 2 .....	7
Exercices en tout genre .....	8

**Edition : juin 2022**

# PANDAS

## Introduction

### Liste d'outils

- **Python** : 1994. Langage ancien et mature. Simple, flexible et généraliste. Pas adapté pour l'optimisation (la rapidité de certains calculs).
- **Numpy** : 2006. Librairie de référence pour **manipuler des tableaux en Python**. Tableaux multidimensionnels.  
⇒ C'est très performant, particulièrement la vectorisation, au détriment d'une certaine simplicité.
- **Pandas** : 2008. Spécialisé et plus complexe. C'est la librairie de référence pour **ajouter des labels aux index des tableaux Numpy**, et ainsi les rendre plus explicites. De plus, on trouve dans Pandas les **opérations classiques de BD** : regroupement (group by), jointure, pivot.  
⇒ C'est très performant, particulièrement la vectorisation, au détriment d'une certaine simplicité.

Numpy et Pandas, c'est ce qui se fait de mieux actuellement.  
Ce n'est pas parfait. Il faut faire avec.

- **Notebook** : du texte formaté mélangé avec du code et avec lequel on peut interagir avec nos données.  
⇒ C'est l'environnement préféré de la communauté datascience car ils permettent de faire des « runnable papers » : des papiers exécutables !

## Pandas – Première approche

### Installation

- Commencez par faire l'installation de pandas (pip ou pip3) :

```
pip3 install pandas
```

ou tout d'un coup :

```
pip3 install numpy matplotlib pandas seaborn
```

### Présentation théorico-pratique

- On utilise le notebook qui est ici, en format notebook et html :  
⇒ [http://bliaudet.free.fr/IMG/zip/pandas\\_cours\\_1.zip](http://bliaudet.free.fr/IMG/zip/pandas_cours_1.zip)
- La version HTML est ici :  
⇒ [http://bliaudet.free.fr/notebook/pandas\\_cours\\_1.html](http://bliaudet.free.fr/notebook/pandas_cours_1.html)

### Plan de la présentation NoteBook

- 1 : Création de Series et de DataFrame
- 2 : Analyse du dataframe :
- 3 : Aperçu du dataframe (du tableau excel) :
- 4 : Accès aux données
- 5 : Sélection de lignes par test : c'est de l'indexation avancée numpy classique
- 6 : Statistiques par colonnes
- 7 : Modification du tableau excel
- 8 - Création de dataframes par index et columns
- 9 - Import-Export

## Plan détaillé de la présentation Notebook

- 1 : Création de Séries et de DataFrame
  - Création des colonnes : les séries
  - Création de la table excel : le dataframe
  - Création plus subtile : avec des index différents dans les séries : ça génère des NaN
- 2 : Analyse du dataframe :
  - index, columns, values
  - Un Index pandas est une séquence immutable (non modifiable)
  - c'est aussi un tableau numpy
- 3 : Appercu du dataframe (du tableau excel) :
  - head(2) : 2 lignes du haut (5 lignes par défaut)
  - tail(2) : 2 lignes du bas (5 lignes par défaut)
  - Statistiques générales
  - Echanger les lignes et les colonnes : transposé : T
- 4 : Accès aux données
  - Accéder à une ligne par son index nommé : df.loc['nom']
  - Usage : df.loc['liz', 'taille']
  - Variante : df.iloc[1] : accès par le numéro (on perd le nom)
  - Plusieurs lignes : slicing
  - Sélectionner des colonnes
  - Une colonne, avec son index
  - Sélection de lignes et de colonnes par slicing
- 5 : Sélection de lignes par test : c'est de l'indexation avancée numpy classique
- 6 : Statistiques par colonnes
  - Jeu de données manuel
  - Tips de Seaborn
- 7 : Modification du tableau excel
  - Création d'un id numéroté : df.reset\_index()
  - Renommer un attribut : df.rename()
  - Changer d'id : set\_index()
  - Ecriture tout en un plus expressive : à utiliser !
- 8 - Création de dataframes par index et columns
  - Exemple 1
  - Exemple 2
  - Addition de dataframes - gestion des NaN : fill\_value, fillna, dropna
  - Remplacer un NaN par une valeur par défaut : fill\_value = 0
  - On remplace les NaN par -1 : .fillna()
  - On supprime les lignes avec des NaN : .dropna
- 9 - Import-Export

- Exportation CSV et JSON
- Importation CSV et JSON

## Exercices

### Série 1

Construisez le tableau suivant :

	age	height	sex
alice	12	130	f
bob	13	140	m
marc	16	176	m
bill	11	120	m
sonia	16	165	f

Affichez l'index des lignes, l'index des colonnes, uniquement les valeurs.

Échangez les lignes et les colonnes mais gardez la matrice d'origine.

Affichez les statistiques des attributs numériques puis de tous les attributs.

Récupérez toutes les infos de sonia.

Récupérez l'âge de sonia

Faites un masque sur les femmes.

Listez uniquement les femmes

Comptez les femmes

Lister uniquement les femmes de plus de 14 ans : on met 2 conditions dans le masque avec un &

## **Série 2**

Chargez le jeu de données sur les pourboires de seaborn : tips

<https://github.com/mwaskom/seaborn-data>

Affichez les premières lignes

Affichez les dernières lignes

Affichez l'index des lignes, l'index des colonnes, uniquement les valeurs.

Affichez les statistiques des attributs numériques puis de tous les attributs.

Faites un masque sur les femmes.

Listez uniquement les femmes

Comptez les femmes

Listez les femmes le dimanche (Sunday)

Quelle est la moyenne de total\_bill pour les femmes ? des hommes ?

## Exercices en tout genre

20 exercices :

<https://favtutor.com/blogs/pandas-exercises-python>

101 exercices :

<https://www.machinelearningplus.com/python/101-pandas-exercises-python/>

Plein d'exercices :

<https://www.w3resource.com/python-exercises/pandas/index.php>

comment le code Python transforme les dataframes

<https://pandastutor.com/>