

Bibliothèque *Pandas*

L'ESSENTIEL¹

On fournit ici quelques fonctions pour manipuler les classeurs Excel avec la bibliothèque *Pandas*.

On trouve sur le web de multiples présentations de cette bibliothèque :

<https://moncoachdata.com/blog/guide-bibliotheque-pandas/>

La description des fonctions est simplifiée, une description complète des fonctions présentées et l'ensemble des autres fonctions disponibles est décrite sur les sites :

<https://pandas.pydata.org/docs/reference>

Import

```
import pandas as pd
```

Ouverture et chargement

```
mon_classeur = pd.ExcelFile(filename)      crée une instance de la classe ExcelFile
mon_classeur.sheet_names                   récupère la liste des feuilles du classeur
data = pd.read_excel(mon_classeur, sheet_name=nom_feuille, index_col=nom_col)

# nom_feuille est le nom de la feuille sur laquelle on veut travailler (par défaut la 1ère feuille)
# nom_col est le nom de la colonne qui sert d'index pour les séries de data
# Ces deux paramètres sont facultatifs

On peut ajouter le paramètre skiprows=n qui permet d'ignorer les n premières lignes de la feuille
```

Accès au contenu d'un *DataFrame*

Avec le *DataFrame* **data** structuré comme ci-dessous, les index correspondent à ceux définis lors du chargement du *dataFrame* :

	ColonneA	ColonneB	ColonneC	ColonneD
index0				
index1				
index2				
index3				
index4				

```
serie_colonne = data[ColonneX]      récupère la colonne de nom ColonneX
```

¹ <https://www.delftstack.com/fr/howto/python-pandas/>
<https://pandas.pydata.org/docs/>

<code>valeur = serie_colonne[index]</code>	récupère le contenu de la série à l'index
<i>equivaut à :</i>	
<code>valeur = data[ColonneX][index]</code>	retourne la valeur de la colonne <i>col</i> pour l'index <i>index</i>
<code>data.iloc[numligne]</code>	retourne une série avec le contenu de la ligne de numéro <i>numligne</i>
<code>data.loc[index]</code>	retourne une série avec le contenu de la ligne d'index <i>index</i>
<code>data.loc[index, col]</code>	accès à la valeur de la colonne <i>col</i> à l'index <i>index</i>

Attributs et informations sur un *DataFrame*

<code>data.columns</code>	liste des noms des colonnes
<code>data.index</code>	liste des index
<code>data.shape</code>	dimensions du dataframe : nombres de lignes, nombre de colonnes
<code>data.describe()</code>	retourne un dataframe donnant des statistiques sur les valeurs (nombres de valeurs, moyenne, écart-type, ...)
<code>data.head()</code>	retourne les 5 premières du <i>dataframe</i>
<code>data.head(n)</code>	retourne les <i>n</i> premières du <i>dataframe</i>
<code>data[col].mean()</code>	retourne la moyenne de la colonne <i>col</i>

Autres fonctions

<code>data[condition]</code>	retourne un <i>dataframe</i> constitué à partir des séries vérifiant la condition, cette condition peut-être une égalité ou une inégalité (exemple : <code>data[data[col] >= val]</code>)
<code>data[col].isin(val)</code>	sélectionne des lignes ayant la valeur <i>val</i> dans une colonne particulière. Elle retourne une série.
<code>data.sort_values(by=colonne)</code>	retourne le <i>dataframe</i> trié sur les valeurs de colonne
<code>data.sort_values(by=colonne, inplace=True)</code>	effectue le tri « sur place », ie modifie le dataframe
<code>data.sort_values(by=colonne, ascending=False)</code>	retourne le <i>dataframe</i> trié par colonne et par valeurs décroissantes
<code>data[colonne].fillna(valeur, inplace=True)</code>	remplace les valeurs manquantes dans la colonne par valeur
<code>series.astype(str)</code>	convertit les valeurs de la série en chaîne de caractères
<code>series.str.len()</code>	retourne une série avec les longueurs des valeurs de la série
<code>series.isna(valeur)</code>	retourne True si la valeur n'est pas définie
<code>data[col] = v</code>	crée une nouvelle colonne avec toutes les valeurs initialisées à la valeur <i>v</i> la colonne est créée à la suite des autres colonnes
<code>data.insert(loc=i, column = nom, value = v)</code>	

	crée une nouvelle colonne à la position <i>i</i> (démarre à 0) avec <i>nom</i> comme nom de colonnes et toutes les valeurs initialisées à la valeur <i>v</i>
<code>date.year</code>	permet de récupérer l'année de date
<code>date.month</code>	permet de récupérer le mois de date
<code>serie.plot()</code>	affiche la courbe permettant de visualiser les valeurs de la série
<code>plt.title(Titre)</code>	
<code>plt.show()</code>	

Affichage de graphiques

Quelques informations sur l'affichage de graphique de données avec les séries et les dataframe sont décrites ici :

<https://learntutorials.net/fr/pandas/topic/3839/graphes-et-visualisations>

Création d'un nouveau classeur

<code>fichier_output = pd.ExcelWriter(filename)</code>	crée un nouveau classeur
<code>data.to_excel(fichier_output, nom_feuilles)</code>	stocke le <i>dataframe</i> dans le classeur
<code>fichier_output.save()</code>	enregistre le fichier