Curs 3: Pandas

Bibliografie:

1. Python Data Science Handbook, Jake VanderPlas, disponibila pe pagina autorului.
2. Practical Data Science with Python: Learn tools and techniques from hands-on examples to extract insights from data, Nathan George, Packt Publishing, 2021
3. Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization, 2nd Edition, Stefanie Molin, Ken Jee, Packt Publishing, 2021

Incarcarea datelor

In NumPy se pot manipula colectii matriceale de date, dar se presupune ca toate datele au acelasi tip:

In [1]:
```
import numpy as np
# indicat: se afiseaza versiunea bibliotecilor cu care se lucreaza in notebook-ul/scri
print(f'NumPy version: {np.__version__}')
```

```
NumPy version: 1.21.5
```

In [2]:
```
tablou = np.array([[1, 2.0, '3'], [3, 4.0, '10']])
tablou
```

Out[2]:
```
array([['1', '2.0', '3'],
       ['3', '4.0', '10']], dtype='<U32')
```

Pandas permite lucrul cu date in care coloanele pot avea tipuri diferite; prima coloana sa fie de tip intreg, al doilea - datetime etc.

In [3]:
```
import pandas as pd
pd.__version__
```

Out[3]:
```
'1.4.2'
```

Un exemplu de set de date care combina tipuri: reale si categoriale (caracter) este Coil 1999 Competition Data Data Set. E utila deci existenta tipurilor de tabel care permit coloane de tip eterogen.

Pandas Series

O serie Pandas este un vector unidimensional de date indexate. Seriile sunt importante pentru ca o coloana dintr-un Pandas dataframe este o serie.

In [4]:
```
data = pd.Series([0.25, 0.5, 0.75, 1.0])
data
```

Out[4]:
```
0    0.25
1    0.50
2    0.75
3    1.00
dtype: float64
```

Valorile se obtin folosind atributul values, returnand un NumPy array:

In [5]:
```
data.values
```

Out[5]:
```
array([0.25, 0.5 , 0.75, 1.  ])
```

Indexul unei serii se obtine prin atributul `index`. In cadrul unui obiect `Series` sau al unui `DataFrame` este util pentru adresarea datelor.

In [6]:
```
type(data.index)
```

Out[6]:
```
pandas.core.indexes.range.RangeIndex
```

Specificarea unui index pentru o serie se poate face la instantiere:

In [7]:
```
data = pd.Series([0.25, 0.5, 0.75, 1.0], index=['a', 'b', 'c', 'd'])
```

In [8]:
```
data
```

Out[8]:
```
a    0.25
b    0.50
c    0.75
d    1.00
dtype: float64
```

In [9]:
```
data.values
```

Out[9]:
```
array([0.25, 0.5 , 0.75, 1.  ])
```

In [10]:
```
data.index
```

Out[10]:
```
Index(['a', 'b', 'c', 'd'], dtype='object')
```

In [11]:
```
data['b']
```

Out[11]:
```
0.5
```

Analogia dintre un obiect `Series` si un dictionar clasic Python poate fi speculata in crearea unui obiect Series plecand de la un dictionar:

In [12]:
```
geografie_populatie = {'Romania': 19638000, 'Franta': 67201000, 'Grecia': 11183957}
populatie = pd.Series(geografie_populatie)
populatie
```

Out[12]:
```
Romania    19638000
Franta     67201000
Grecia     11183957
dtype: int64
```

In [13]:
```
populatie.index
```

Out[13]:
```
Index(['Romania', 'Franta', 'Grecia'], dtype='object')
```

In [14]:
```
populatie['Grecia']
```

Out[14]:
```
11183957
```

In [15]:
```
# populatie['Germania']
# eroare: KeyError: 'Germania'
```

Daca nu se specifica un index la crearea unui obiect `Series`, atunci implicit acesta va fi format pe baza secventei de intregi 0, 1, 2, ...

Nu e obligatoriu ca o serie sa contina doar valori numerice:

In [16]:
```
s1 = pd.Series(['rosu', 'verde', 'galben', 'albastru'])
print(s1)
print('s1[2]=', s1[2])
```

```
0        rosu
1       verde
2      galben
3    albastru
dtype: object
s1[2]= galben
```

Datele unei serii se vad ca avand toate acelasi tip:

In [17]:
```
s_tip = pd.Series(['rosu', 1, 1.5])
s_tip
```

Out[17]:
```
0    rosu
1       1
2     1.5
dtype: object
```

Selectarea datelor in serii

Datele dintr-o serie pot fi referite prin intermediul indexului:

In [18]:
```
data = pd.Series(np.linspace(0, 75, 4), index=['a', 'b', 'c', 'd'])
print(data)
data['b']
```

```
a     0.0
b    25.0
c    50.0
d    75.0
dtype: float64
```

Out[18]:
```
25.0
```

Se poate face modificarea datelor dintr-o serie folosind indexul:

In [19]:
```
data['b'] = 300
```

```
print(data)
```

```
a      0.0
b    300.0
c     50.0
d     75.0
dtype: float64
```

Se poate folosi slicing, iar aici, spre deosebire de slicing-ul din NumPy si Python, **se ia inclusiv capatul din dreapta al indicilor:**

In [20]:
```
data['a':'c']
```

Out[20]:
```
a      0.0
b    300.0
c     50.0
dtype: float64
```

sau se pot folosi liste de selectie:

In [21]:
```
data[['a', 'c', 'b', 'c']]
```

Out[21]:
```
a      0.0
c     50.0
b    300.0
c     50.0
dtype: float64
```

sau expresii logice:

In [22]:
```
data[(data > 30) & (data < 80)] # se remarca returnarea in rezultat a indicilor care s
# de fapt, rezultatul unei astfel de expresii este tot o serie
```

Out[22]:
```
c    50.0
d    75.0
dtype: float64
```

Se prefera folosirea urmatoarelor atribute de indexare: `loc`, `iloc`. Indexarea prin `ix`, daca se regaseste prin tutoriale mai vechi, se considera a fi sursa de confuzie si se recomanda evitarea ei.

Atributul `loc` permite indicierea folosind valoarea de index.

In [23]:
```
data = pd.Series([1, 2, 3], index=['a', 'b', 'c'])

data
```

Out[23]:
```
a    1
b    2
c    3
dtype: int64
```

In [24]:
```
#cautare dupa index cu o singura valoare
data.loc['b']
```

Out[24]:
```
2
```

In [25]:
```
#cautare dupa index cu o doua valori. Lista interioara este folosita pentru a stoca o
```

```
data.loc[['a', 'c']]
```

Out[25]:
```
a    1
c    3
dtype: int64
```

Atributul `iloc` este folosit pentru a face referire la linii dupa pozitia (numarul) lor. Numerotarea incepe de la 0.

In [26]:
```
data.iloc[0]
```

Out[26]:
```
1
```

In [27]:
```
data.iloc[[0, 2]]
```

Out[27]:
```
a    1
c    3
dtype: int64
```

DataFrame

Un obiect `DataFrame` este o colectie de coloane de tip `Series`. Numarul de elemente din fiecare serie este acelasi.

In [28]:
```
df = pd.DataFrame([[1, 2, 3], [4, 5, 6]])
df
```

Out[28]:

	0	1	2
0	1	2	3
1	4	5	6

Se poate ca seriile (coloanele din dataframe) sa fie de tip diferit:

In [29]:
```
df_mix = pd.DataFrame([[1, 'Ana', 3.14], [2, 'Dan', 103.2]])
df_mix
```

Out[29]:

	0	1	2
0	1	Ana	3.14
1	2	Dan	103.20

In [30]:
```
df_mix.dtypes
```

Out[30]:
```
0      int64
1     object
2    float64
dtype: object
```

Motivul pentru care tipul coloanei 'Nume' este `object` si nu `str` e explicat aici.

Se poate folosi un dictionar cu cheia avand nume de coloane, iar valorile de pe coloane ca liste:

In [31]:
```
df = pd.DataFrame({'Nume' : ['Ana', 'Dan', 'Maria'], 'Varsta': [20,30, 40]})
df
```

Out[31]:

	Nume	Varsta
0	Ana	20
1	Dan	30
2	Maria	40

In [32]:
```
geografie_suprafata = {'Romania': 238397, 'Franta': 640679, 'Grecia': 131957}

geografie_moneda = {'Romania': 'RON', 'Franta': 'EUR', 'Grecia': 'EUR'}

geografie = pd.DataFrame({'Populatie' : geografie_populatie, 'Suprafata' : geografie_s

print(geografie)
```

```
         Populatie  Suprafata Moneda
Romania   19638000     238397    RON
Franta    67201000     640679    EUR
Grecia    11183957     131957    EUR
```

In [33]:
```
print(geografie.index)
```

```
Index(['Romania', 'Franta', 'Grecia'], dtype='object')
```

Atributul `columns` da lista de coloane din obiectul `DataFrame`:

In [34]:
```
geografie.columns
```

Out[34]:
```
Index(['Populatie', 'Suprafata', 'Moneda'], dtype='object')
```

Referirea la o serie care compune o coloana din DataFrame se face astfel

In [35]:
```
print(geografie['Populatie'])
print('**********************')
print(type(geografie['Populatie']))
```

```
Romania    19638000
Franta     67201000
Grecia     11183957
Name: Populatie, dtype: int64
**********************
<class 'pandas.core.series.Series'>
```

Crearea unui obiect DataFrame se poate face pornind si de la o singura serie:

In [36]:
```
mydf = pd.DataFrame([1, 2, 3], columns=['values'])
mydf
```

Out[36]:

	values
0	1
1	2
2	3

... sau se poate crea pornind de la o lista de dictionare:

In [37]:
```
data
```

Out[37]:
```
a    1
b    2
c    3
dtype: int64
```

In [38]:
```
data = [{'a': i, 'b': 2 * i} for i in range(3)]
print(data)
pd.DataFrame(data)
```

```
[{'a': 0, 'b': 0}, {'a': 1, 'b': 2}, {'a': 2, 'b': 4}]
```

Out[38]:

	a	b
0	0	0
1	1	2
2	2	4

Daca lipsesc chei din vreunul din dictionare, respectiva valoare se va umple cu `NaN`.

In [39]:
```
pd.DataFrame([{'a': 1, 'b': 2}, {'b': 3, 'c': 4}])
```

Out[39]:

	a	b	c
0	1.0	2	NaN
1	NaN	3	4.0

In [40]:
```
pd.DataFrame([{'a': 'aaa', 'b': 'bbb'}, {'b': 'bbb2', 'c': 'cccc'}])
```

Out[40]:

	a	b	c
0	aaa	bbb	NaN
1	NaN	bbb2	cccc

Instantierea unui DataFrame se poate face si de la un NumPy array:

In [41]:
```
pd.DataFrame(np.random.rand(3, 2), columns=['Col1', 'Col2'], index=['a', 'b', 'c'])
```

Out[41]:

	Col1	Col2
a	0.857561	0.374087
b	0.838438	0.596081
c	0.399397	0.052431

Se poate adauga o coloana noua la un DataFrame, similar cu adaugarea unui element (cheie, valoare) la un dictionar:

In [42]:
```
geografie['Densitatea populatiei'] = geografie['Populatie'] / geografie['Suprafata']

geografie
```

Out[42]:

	Populatie	Suprafata	Moneda	Densitatea populatiei
Romania	19638000	238397	RON	82.375198
Franta	67201000	640679	EUR	104.890280
Grecia	11183957	131957	EUR	84.754556

Un obiect DataFrame poate fi transpus cu atributul `T`:

In [43]:
```
geografie.T
```

Out[43]:

	Romania	Franta	Grecia
Populatie	19638000	67201000	11183957
Suprafata	238397	640679	131957
Moneda	RON	EUR	EUR
Densitatea populatiei	82.375198	104.89028	84.754556

Selectarea datelor intr-un `DataFrame`

S-a demonstrat posibilitatea de referire dupa numele de coloana:

In [44]:
```
print(geografie)
```

```
         Populatie  Suprafata Moneda  Densitatea populatiei
Romania   19638000     238397    RON              82.375198
Franta    67201000     640679    EUR             104.890280
Grecia    11183957     131957    EUR              84.754556
```

In [45]:
```
print(geografie['Moneda'])
```

```
Romania    RON
Franta     EUR
Grecia     EUR
Name: Moneda, dtype: object
```

Daca numele unei coloane este un string fara spatii, se poate folosi acesta ca un atribut:

In [46]:
```
geografie.Moneda
```

Out[46]:
```
Romania    RON
Franta     EUR
Grecia     EUR
Name: Moneda, dtype: object
```

Se poate face referire la o coloana dupa indicele ei, indirect:

In [47]:
```
geografie[geografie.columns[0]]
```

Out[47]:
```
Romania    19638000
Franta     67201000
Grecia     11183957
Name: Populatie, dtype: int64
```

Pentru cazul in care un DataFrame nu are nume de coloana, ele sunt implicit intregii 0, 1, ... si se pot folosi pentru selectarea de coloana folosind paranteze drepte:

In [48]:
```
my_data = pd.DataFrame(np.random.rand(3, 4))

my_data
```

Out[48]:

	0	1	2	3
0	0.004906	0.307117	0.637900	0.466495
1	0.986423	0.956600	0.874462	0.543992
2	0.840027	0.345096	0.883984	0.489281

In [49]:
```
my_data[0]
```

Out[49]:
```
0    0.004906
1    0.986423
2    0.840027
Name: 0, dtype: float64
```

Atributul `values` returneaza un obiect ndarray continand valori. Tipul unui ndarray este cel mai specializat tip de date care poate sa contina valorile din DataFrame:

In [50]:
```
# afisare ndarray si tip pentru my_data.values
print(my_data.values)
print(my_data.values.dtype)
```

```
[[0.00490598 0.30711704 0.63789962 0.4664948 ]
 [0.98642255 0.95660049 0.87446244 0.54399208]
 [0.84002703 0.34509619 0.88398413 0.48928103]]
float64
```

In [51]:
```
# afisare ndarray si tip pentru geografie.values
print(geografie.values)
print(geografie.values.dtype)
```

```
[[19638000 238397 'RON' 82.37519767446739]
 [67201000 640679 'EUR' 104.89028046806591]
 [11183957 131957 'EUR' 84.75455640852702]]
object
```

Indexarea cu `iloc` in cazul unui obiect `DataFrame` permite precizarea a doua valori: prima reprezinta linia si al doilea coloana, numerotate de la 0. Pentru linie si coloana se poate folosi si slicing. Ca si slicing-ul din NumPy si lista Python, indicele din dinalul expresiei de slicing nu eset inclu in selectie.

In [52]:
```
print(geografie)

geografie.iloc[0:2, 2:4]
```

```
         Populatie  Suprafata Moneda  Densitatea populatiei
Romania   19638000     238397    RON              82.375198
Franta    67201000     640679    EUR             104.890280
Grecia    11183957     131957    EUR              84.754556
```

Out[52]:

	Moneda	Densitatea populatiei
Romania	RON	82.375198
Franta	EUR	104.890280

Indexarea cu `loc` permite precizarea valorilor de indice si respectiv nume de coloana:

In [53]:
```
print(geografie)

geografie.loc[['Franta', 'Romania'], 'Populatie':'Densitatea populatiei']
```

```
         Populatie  Suprafata Moneda  Densitatea populatiei
Romania   19638000     238397    RON              82.375198
Franta    67201000     640679    EUR             104.890280
Grecia    11183957     131957    EUR              84.754556
```

Out[53]:

	Populatie	Suprafata	Moneda	Densitatea populatiei
Franta	67201000	640679	EUR	104.890280
Romania	19638000	238397	RON	82.375198

Se permite folosirea de expresii de filtrare à la NumPy:

In [54]:
```
geografie.loc[geografie['Densitatea populatiei'] > 83, ['Populatie', 'Moneda']]
```

Out[54]:

	Populatie	Moneda
Franta	67201000	EUR
Grecia	11183957	EUR

Folosind indicierea, se pot modifica valorile dintr-un `DataFrame`:

In [55]:
```
#Modificarea populatiei Greciei cu iloc
geografie.iloc[1, 1] = 12000000
print(geografie)
```

```
         Populatie  Suprafata Moneda  Densitatea populatiei
Romania   19638000     238397    RON              82.375198
Franta    67201000   12000000    EUR             104.890280
Grecia    11183957     131957    EUR              84.754556
```

In [56]:
```
#Modificarea populatiei Greciei cu loc
geografie.loc['Grecia', 'Populatie'] = 11183957
print(geografie)
```

```
         Populatie  Suprafata Moneda  Densitatea populatiei
Romania   19638000     238397    RON              82.375198
Franta    67201000   12000000    EUR             104.890280
Grecia    11183957     131957    EUR              84.754556
```

Precizari:

1. daca se foloseste un singur indice la un DataFrame, atunci se considera ca se face referire la coloana:

   `geografie['Moneda']`

2. daca se foloseste slicing, acesta se refera la liniile (indexul) din DataFrame:

   `geografie['Franta':'Romania']`

3. operatiile logice se considera ca refera de asemenea linii din DataFrame:

   `geografie[geografie['Densitatea populatiei'] > 83]`

In [57]:
```
geografie[geografie['Densitatea populatiei'] > 83]
```

Out[57]:

	Populatie	Suprafata	Moneda	Densitatea populatiei
Franta	67201000	12000000	EUR	104.890280
Grecia	11183957	131957	EUR	84.754556

Operarea pe date

Se pot aplica functii NumPy peste obiecte Series si DataFrame. Rezultatul este de acelasi tip ca obiectul peste care se aplica iar indicii se pastreaza:

In [58]:
```
ser = pd.Series(np.random.randint(low=0, high=10, size=(5)), index=['a', 'b', 'c', 'd'
ser
```

Out[58]:
```
a    1
b    4
c    9
d    1
e    7
dtype: int32
```

In [59]:
```
np.exp(ser)
```

Out[59]:
```
a       2.718282
b      54.598150
c    8103.083928
d       2.718282
e    1096.633158
dtype: float64
```

In [60]:
```
my_df = pd.DataFrame(data=np.random.randint(low=0, high=10, size=(3, 4)), \
                     columns=['Sunday', 'Monday', 'Tuesday', 'Wednesday'], \
```

```
                    index=['a', 'b', 'c'])
print('Originar:', my_df)
print('Transformat:', np.exp(my_df))
```

```
Originar:    Sunday  Monday  Tuesday  Wednesday
a       6       2        9          7
b       1       9        1          2
c       5       6        2          4
Transformat:        Sunday       Monday      Tuesday    Wednesday
a  403.428793     7.389056  8103.083928  1096.633158
b    2.718282  8103.083928     2.718282     7.389056
c  148.413159   403.428793     7.389056    54.598150
```

Pentru functii binare se face alinierea obiectelor Series sau DataFrame dupa indexul lor. Aceasta poate duce la operare cu valori NaN si in consecinta obtinere de valori NaN.

In [61]:
```
area = pd.Series({'Alaska': 1723337, 'Texas': 695662, 'California': 423967}, name='are
population = pd.Series({'California': 38332521, 'Texas': 26448193, 'New York': 1965112
```

In [62]:
```
population / area
```

Out[62]:
```
Alaska              NaN
California    90.413926
New York            NaN
Texas         38.018740
dtype: float64
```

In cazul unui DataFrame, alinierea se face atat pentru coloane, cat si pentru indecsii folositi la linii:

In [63]:
```
A = pd.DataFrame(data=np.random.randint(0, 10, (2, 3)), columns=list('ABC'))
B = pd.DataFrame(data=np.random.randint(0, 10, (3, 2)), columns=list('BA'))

A
```

Out[63]:

	A	B	C
0	1	8	5
1	1	1	2

In [64]:
```
B
```

Out[64]:

	B	A
0	1	2
1	7	8
2	0	9

In [65]:
```
A + B
```

Out[65]:

	A	B	C
0	3.0	9.0	NaN
1	9.0	8.0	NaN
2	NaN	NaN	NaN

Daca se doreste umplerea valorilor NaN cu altceva, se poate specifica parametrul fill_value pentru functii care implementeaza operatiile aritmetice:

Operator	Metoda Pandas
+	`add()`
-	`sub()`, `substract()`
*	`mul()`, `multiply()`
/	`truediv()`, `div()`, `divide()`
//	`floordiv()`
%	`mod()`
**	`pow()`

Daca ambele pozitii au valori lipsa (NaN), atunci valoarea finala va fi si ea lipsa.

Exemplu:

In [66]:
```
A
```

Out[66]:

	A	B	C
0	1	8	5
1	1	1	2

In [67]:
```
B
```

Out[67]:

	B	A
0	1	2
1	7	8
2	0	9

In [68]:
```
A.add(B, fill_value=0)
```

Out[68]:

	A	B	C
0	3.0	9.0	5.0
1	9.0	8.0	2.0
2	9.0	0.0	NaN

Valori lipsa

Pentru cazul in care valorile dintr-o coloana a unui obiect DataFrame sunt de tip numeric, valorile lipsa se reprezinta prin NaN - care e suportat doar de tipurile in virgula mobila, nu si de intregi; aceasta din ultima observatie arata ca numerele intregi sunt convertite la floating point daca intr-o lista care le contine se afla si valori lipsa:

In [69]:
```
my_series = pd.Series([1, 2, 3, None, 5], name='my_series')
# echivalent:
my_series = pd.Series([1, 2, 3, np.NaN, 5], name='my_series')
my_series
```

Out[69]:
```
0    1.0
1    2.0
2    3.0
3    NaN
4    5.0
Name: my_series, dtype: float64
```

Functiile care se pot folosi pentru un DataFrame pentru a operare cu valori lipsa sunt:

In [70]:
```
df = pd.DataFrame([[1, 2, np.NaN], [np.NAN, 10, 20]])
df
```

Out[70]:

	0	1	2
0	1.0	2	NaN
1	NaN	10	20.0

`isnull()` - returneaza o masca de valori logice, cu `True` (`False`) pentru pozitiile unde se afla valori nule (respectiv: nenule); nul = valoare lipsa.

In [71]:
```
df.isnull()
```

Out[71]:

	0	1	2
0	False	False	True
1	True	False	False

`notnull()` - opusul functiei precedente

`dropna()` - returneaza o varianta filtrata a obiectului DataFrame. E posibil sa duca la un DataFrame gol.

In [72]:
```
df.dropna()
```

Out[72]:

	0	1	2

In [73]:
```
df.iloc[0] = [3, 4, 5]
print(df)
df.dropna()
```

```
     0   1     2
0  3.0   4   5.0
1  NaN  10  20.0
```

Out[73]:

	0	1	2
0	3.0	4	5.0

`fillna()` umple valorile lipsa dupa o anumita politica:

In [74]:
```
df = pd.DataFrame([[1, 2, np.NaN], [np.NAN, 10, 20]])
df
```

Out[74]:

	0	1	2
0	1.0	2	NaN
1	NaN	10	20.0

In [75]:
```
# umplere de NaN-uri cu valoare constanta
df2 = df.fillna(value = 100)
df2
```

Out[75]:

	0	1	2
0	1.0	2	100.0
1	100.0	10	20.0

In [76]:
```
np.random.randn(5, 3)
```

Out[76]:
```
array([[ 0.58260061, -1.33590224, -0.59080867],
       [ 0.04356137, -0.27529478, -2.12412265],
       [ 2.56768829, -0.10489638, -1.84286171],
       [ 0.37283781, -0.86032776, -0.3247227 ],
       [-0.32652648, -0.29353619, -1.49623783]])
```

In [77]:
```
# umplere de NaN-uri cu media pe coloana corespunzatoare
df = pd.DataFrame(data = np.random.randn(5, 3), columns=['A', 'B', 'C'])
df.iloc[0, 2] = df.iloc[1, 1] = df.iloc[2, 0] = df.iloc[4, 1] = np.NAN
df
```

Out[77]:

	A	B	C
0	0.210186	-0.878007	NaN
1	-0.200703	NaN	-0.445248
2	NaN	1.156294	-1.121232
3	-0.698647	2.654418	0.725052
4	1.025623	NaN	-0.545932

In [78]:
```
#calcul medie pe coloana
df.mean(axis=0)
```

Out[78]:
```
A    0.084115
B    0.977568
C   -0.346840
dtype: float64
```

In [79]:
```
df3 = df.fillna(df.mean(axis=0))
df3
```

Out[79]:

	A	B	C
0	0.210186	-0.878007	-0.346840
1	-0.200703	0.977568	-0.445248
2	0.084115	1.156294	-1.121232
3	-0.698647	2.654418	0.725052
4	1.025623	0.977568	-0.545932

Exista un parametru al functiei `fillna()` care permite umplerea valorilor lipsa prin copiere:

In [80]:
```
my_ds = pd.Series(np.arange(0, 30))
my_ds[1:-1:4] = np.NaN
my_ds
```

Out[80]:
```
0     0.0
1     NaN
2     2.0
3     3.0
4     4.0
5     NaN
6     6.0
7     7.0
8     8.0
9     NaN
10    10.0
11    11.0
12    12.0
13    NaN
14    14.0
15    15.0
16    16.0
17    NaN
18    18.0
19    19.0
20    20.0
21    NaN
22    22.0
23    23.0
24    24.0
25    NaN
26    26.0
27    27.0
28    28.0
29    29.0
dtype: float64
```

In [81]:
```
# copierea ultimei valori non-null
my_ds_filled_1 = my_ds.fillna(method='ffill')
my_ds_filled_1
```

Out[81]:
```
0      0.0
1      0.0
2      2.0
3      3.0
4      4.0
5      4.0
6      6.0
7      7.0
8      8.0
9      8.0
10    10.0
11    11.0
12    12.0
13    12.0
14    14.0
15    15.0
16    16.0
17    16.0
18    18.0
19    19.0
20    20.0
21    20.0
22    22.0
23    23.0
24    24.0
25    24.0
26    26.0
27    27.0
28    28.0
29    29.0
dtype: float64
```

In [82]:
```
# copierea inapoi a urmatoarei valori non-null
my_ds_filled_2 = my_ds.fillna(method='bfill')
my_ds_filled_2
```

Out[82]:
```
0      0.0
1      2.0
2      2.0
3      3.0
4      4.0
5      6.0
6      6.0
7      7.0
8      8.0
9     10.0
10    10.0
11    11.0
12    12.0
13    14.0
14    14.0
15    15.0
16    16.0
17    18.0
18    18.0
19    19.0
20    20.0
21    22.0
22    22.0
23    23.0
24    24.0
25    26.0
26    26.0
27    27.0
28    28.0
29    29.0
dtype: float64
```

Pentru DataFrame, procesul este similar. Se poate specifica argumentul axis care spune daca procesarea se face pe linii sau pe coloane:

In [83]:
```
df = pd.DataFrame([[1, np.NAN, 2, np.NAN], [2, 3, 5, np.NaN], [np.NaN, 4, 6, np.NaN]])
df
```

Out[83]:

	0	1	2	3
0	1.0	NaN	2	NaN
1	2.0	3.0	5	NaN
2	NaN	4.0	6	NaN

In [84]:
```
# Umplere, prin parcurgere pe linii
df.fillna(method='ffill', axis = 1)
```

Out[84]:

	0	1	2	3
0	1.0	1.0	2.0	2.0
1	2.0	3.0	5.0	5.0
2	NaN	4.0	6.0	6.0

In [85]:
```
# Umplere, prin parcurgere pe fiecare coloana
```

```
df.fillna(method='ffill', axis = 0)
```

Out[85]:

	0	1	2	3
0	1.0	NaN	2	NaN
1	2.0	3.0	5	NaN
2	2.0	4.0	6	NaN

Combinarea de obiecte Series si DataFrame

Cea mai simpla operatie este de concatenare:

In [86]:
```
ser1 = pd.Series(['A', 'B', 'C'], index=[1, 2, 3])
ser2 = pd.Series(['D', 'E', 'F'], index=[4, 5, 6])
pd.concat([ser1, ser2])
```

Out[86]:
```
1    A
2    B
3    C
4    D
5    E
6    F
dtype: object
```

Pentru cazul in care valori de index se regasesc in ambele serii de date, indexul se va repeta:

In [87]:
```
ser1 = pd.Series(['A', 'B', 'C'], index=[1, 2, 3])
ser2 = pd.Series(['D', 'E', 'F'], index=[3, 4, 5])
ser_concat = pd.concat([ser1, ser2])
ser_concat
```

Out[87]:
```
1    A
2    B
3    C
3    D
4    E
5    F
dtype: object
```

In [88]:
```
ser_concat.loc[3]
```

Out[88]:
```
3    C
3    D
dtype: object
```

Pentru cazul in care se doreste verificarea faptului ca indecsii sunt unici, se poate folosi parametrul `verify_integrity` :

In [89]:
```
try:
    ser_concat = pd.concat([ser1, ser2], verify_integrity=True)
except ValueError as e:
    print('Value error', e)
```

```
Value error Indexes have overlapping values: Int64Index([3], dtype='int64')
```

Pentru concatenarea de obiecte `DataFrame` care au acelasi set de coloane (pentru moment):

In [90]:
```
# sursa: ref 1 din Curs 1
def make_df(cols, ind):
    """Quickly make a DataFrame"""
    data = {c: [str(c) + str(i) for i in ind] for c in cols}
    return pd.DataFrame(data, ind)
```

In [91]:
```
df1 = make_df('AB', [1, 2])
df2 = make_df('AB', [3, 4])
print(df1); print(df2);
```

```
    A   B
1  A1  B1
2  A2  B2
    A   B
3  A3  B3
4  A4  B4
```

In [92]:
```
#concatenare simpla
pd.concat([df1, df2])
```

Out[92]:

	A	B
1	A1	B1
2	A2	B2
3	A3	B3
4	A4	B4

Concatenarea se poate face si pe orizontala:

In [93]:
```
df3 = make_df('AB', [0, 1])
df4 = make_df('CD', [0, 1])
print(df3); print(df4);
```

```
    A   B
0  A0  B0
1  A1  B1
    C   D
0  C0  D0
1  C1  D1
```

In [94]:
```
#concatenare pe axa 1
pd.concat([df3, df4], axis=1)
```

Out[94]:

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1

Pentru indici duplicati, comportamentul e la fel ca la `Serie`: se pastreaza duplicatele si datele corespunzatoare:

In [95]:
```
x = make_df('AB', [0, 1])
y = make_df('AB', [0, 1])
print(x); print(y);
```

```
    A   B
0  A0  B0
1  A1  B1
    A   B
0  A0  B0
1  A1  B1
```

In [96]:
```
print(pd.concat([x, y]))
```

```
    A   B
0  A0  B0
1  A1  B1
0  A0  B0
1  A1  B1
```

In [97]:
```
try:
    df_concat = pd.concat([x, y], verify_integrity=True)
except ValueError as e:
    print('Value error', e)
```

```
Value error Indexes have overlapping values: Int64Index([0, 1], dtype='int64')
```

Daca se doreste ignorarea indecsilor, se poate folosi indicatorul `ignore_index`:

In [98]:
```
df_concat = pd.concat([x, y], ignore_index=True)
```

Pentru cazul in care obiectele `DataFrame` nu au exact aceleasi coloane, concatenarea poate duce la rezultate de forma:

In [99]:
```
df5 = make_df('ABC', [1, 2])
df6 = make_df('BCD', [3, 4])
print(df5); print(df6);
```

```
    A   B   C
1  A1  B1  C1
2  A2  B2  C2
    B   C   D
3  B3  C3  D3
4  B4  C4  D4
```

In [100…
```
print(pd.concat([df5, df6]))
```

```
     A   B   C    D
1   A1  B1  C1  NaN
2   A2  B2  C2  NaN
3  NaN  B3  C3   D3
4  NaN  B4  C4   D4
```

De regula se vrea operatia de concatenare (join) pe obiectele DataFrame cu coloane diferite. O prima varianta este pastrarea doar a coloanelor partajate, ceea ce in Pandas este vazut ca un inner join (se remarca o necorespondenta cu terminologia din limbajul SQL):

In [101…
```
print(df5); print(df6);
```

```
    A   B   C
1  A1  B1  C1
2  A2  B2  C2
    B   C   D
3  B3  C3  D3
4  B4  C4  D4
```

In [102…
```
# concatenare cu inner join
pd.concat([df5, df6], join='inner')
```

Out[102]:

	B	C
1	B1	C1
2	B2	C2
3	B3	C3
4	B4	C4

Alta varianta este specificarea explicita a coloanelor care rezista in urma concatenarii, prin metoda `reindex`:

In [103…
```
print(df5); print(df6);
```

```
    A   B   C
1  A1  B1  C1
2  A2  B2  C2
    B   C   D
3  B3  C3  D3
4  B4  C4  D4
```

In [104…
```
# pd.concat([df5, df6], join_axes=[df5.columns]) # parametrul join_axes e deprecated
pd.concat([df5, df6.reindex(df5.columns, axis=1)])
```

Out[104]:

	A	B	C
1	A1	B1	C1
2	A2	B2	C2
3	NaN	B3	C3
4	NaN	B4	C4

Pentru implementarea de jonctiuni à la SQL se foloseste metoda `merge`. Ce mai simpla este inner join: rezulta liniile din obiectele DataFrame care au corespondent in ambele parti:

In [105…
```
df1 = pd.DataFrame({'employee': ['Bob', 'Jake', 'Lisa', 'Sue'],
'group': ['Accounting', 'Engineering', 'Engineering', 'HR']})
df2 = pd.DataFrame({'employee': ['Lisa', 'Bob', 'Jake', 'Sue'],
'hire_date': [2004, 2008, 2012, 2014]})
```

In [106…
```
print(df1)
print(df2)
```

```
  employee        group
0      Bob   Accounting
1     Jake  Engineering
2     Lisa  Engineering
3      Sue           HR
  employee  hire_date
0     Lisa       2004
1      Bob       2008
2     Jake       2012
3      Sue       2014
```

In [107…]:
```
df3=pd.merge(df1, df2, on='employee')
df3
```

Out[107]:

	employee	group	hire_date
0	Bob	Accounting	2008
1	Jake	Engineering	2012
2	Lisa	Engineering	2004
3	Sue	HR	2014

In [108…]:
```
df3 = pd.DataFrame({'employee': ['Jake', 'Lisa', 'Sue'],
'group': ['Engineering', 'Engineering', 'HR']})
df4 = pd.DataFrame({'employee': ['Bob', 'Jake', 'Sue'],
'hire_date': [2008, 2012, 2014]})

print(df3)
print(df4)

# demo inner join: raman dar 2 linii dupa jonctiune
pd.merge(df3, df4, on='employee')
```

```
  employee        group
0     Jake  Engineering
1     Lisa  Engineering
2      Sue           HR
  employee  hire_date
0      Bob       2008
1     Jake       2012
2      Sue       2014
```

Out[108]:

	employee	group	hire_date
0	Jake	Engineering	2012
1	Sue	HR	2014

Se pot face asa-numite jonctiuni `many-to-one`, dar care nu sunt decat inner join. Mentionam si exemplificam doar pentru terminologie:

In [109…]:
```
df4 = pd.DataFrame({'group': ['Accounting', 'Engineering', 'HR'],
'supervisor': ['Carly', 'Guido', 'Steve']})

print(df3)
print(df4)
```

```
  employee        group
0     Jake  Engineering
1     Lisa  Engineering
2      Sue           HR
         group supervisor
0   Accounting      Carly
1  Engineering      Guido
2           HR      Steve
```

In [110…
```
pd.merge(df3, df4, on='group')
```

Out[110]:

	employee	group	supervisor
0	Jake	Engineering	Guido
1	Lisa	Engineering	Guido
2	Sue	HR	Steve

Asa-numite jonctiuni *many-to-many* se obtin pentru cazul in care coloana dupa care se face jonctiunea contine duplicate:

In [111…
```
df5 = pd.DataFrame({'group': ['Accounting', 'Accounting',
'Engineering', 'Engineering', 'HR', 'HR'],
'skills': ['math', 'spreadsheets', 'coding', 'linux',
'spreadsheets', 'organization']})
print(df1)
print(df5)
```

```
  employee        group
0      Bob   Accounting
1     Jake  Engineering
2     Lisa  Engineering
3      Sue           HR
         group        skills
0   Accounting          math
1   Accounting  spreadsheets
2  Engineering        coding
3  Engineering         linux
4           HR  spreadsheets
5           HR  organization
```

In [112…
```
print(pd.merge(df1, df5, on='group'))
```

```
  employee        group        skills
0      Bob   Accounting          math
1      Bob   Accounting  spreadsheets
2     Jake  Engineering        coding
3     Jake  Engineering         linux
4     Lisa  Engineering        coding
5     Lisa  Engineering         linux
6      Sue           HR  spreadsheets
7      Sue           HR  organization
```