Introducción a la Computación Científica con Python
Clase 6: Pandas

Diego Passarella

Víctor Viana

EDUCACION PERMANENTE
Universidad de la República

SEDE TACUAREMBÓ
CENUR NORESTE

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Librería Pandas

Pandas es una librería de Python con multitud de métodos y funcionalidades que permitirán esplorar y agrupar la información que esconden los datos para tomar decisiones.

Librería Pandas

- nombre derivado de *panel data*
- proporciona estructuras de datos y funciones de alto nivel que nos permiten trabajar con datos estructurados de manera muy cómoda.

Instalando Pandas

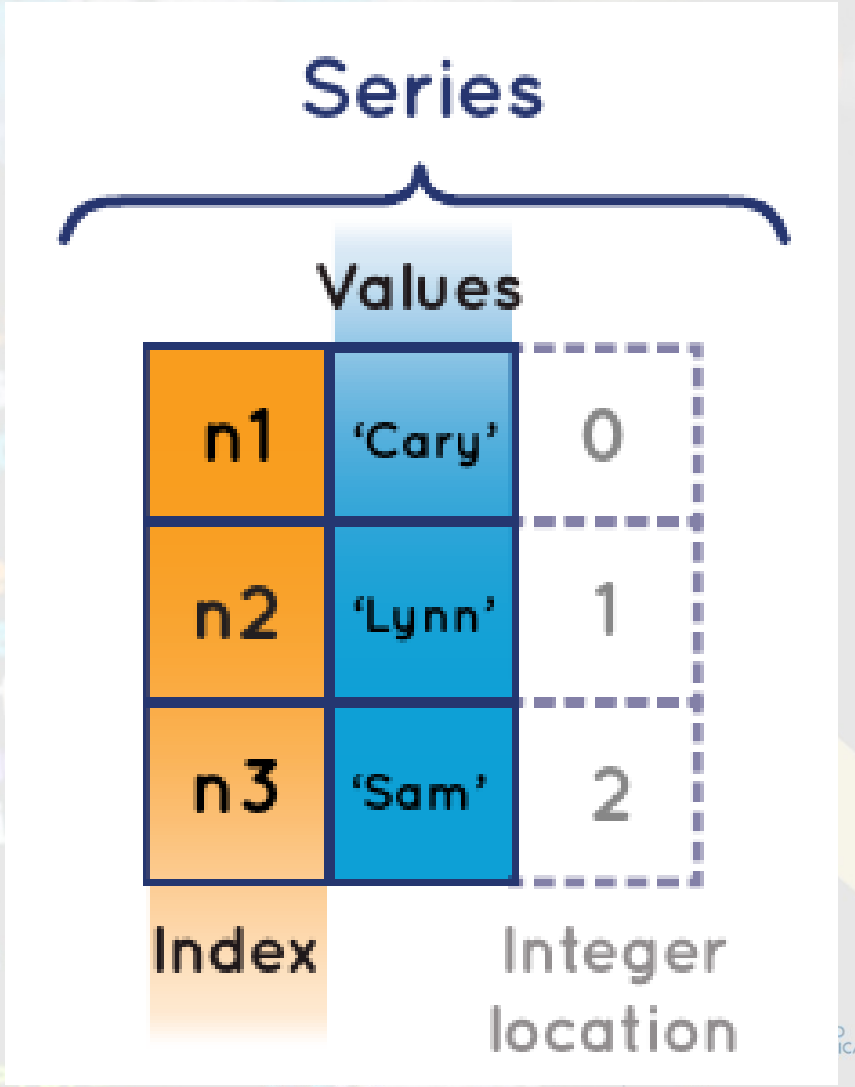
pip install panda

pip → https://packaging.python.org/tutorials/installing-packages/

Estructura de datos: Dataframe y Series

- **serie**: estructuras unidimensionales conteniendo un array de datos
- **dataframe**: estructura tabular bidimensional
- ambas basadas en el array multidimensional de NumPy

```
pd.Series(values, index=index,
          name=name)
pd.Series({'idx1': val1, 'idx2': val2}
```



Series

```
import pandas as pd
import numpy as np


s = pd.Series([-2, 0, 3, 6])
print(s)
```

```
0   -2
1    0
2    3
3    6
```

Series(cont.)

```
import pandas as pd
import numpy as np
ventas = pd.Series([15, 12, 21], index =["Enero", "Febrero", "Marzo"])
print(ventas)
```

```
Enero     15
Febrero   12
Marzo     21
```

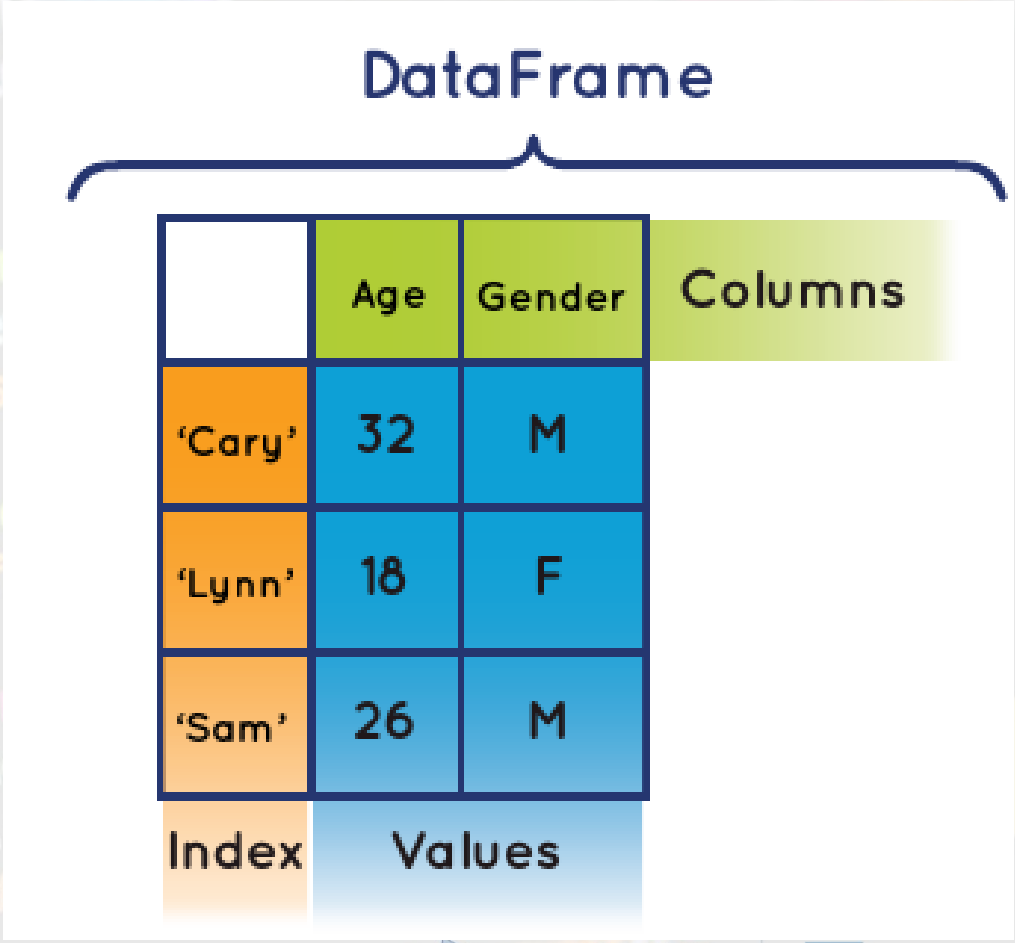
EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Series(cont.)

```
>>> ventas[0]
15
>>> ventas["Enero"]
15
```

```
>>> ventas.dtype
dtype('int64')
>>> ventas.index
Index([u'Enero', u'Febrero', u'Marzo'],
dtype='object')
>>> ventas.values
array([15, 12, 21])
```

```
pd.DataFrame(values, index=index,
             columns=col_names)
pd.DataFrame({'col1': series1_or_seq,
              'col2': series2_or_seq})
```



EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Dataframe

```
ventas2 = pd.DataFrame({
    "Entradas": [41, 32, 56, 18],
    "Salidas": [17, 54, 6, 78],
    "Valoraciones": [66, 54, 49, 66],
    "Limite": ["No", "Si", "No", "No"],
    "Cambio": [1.43, 1.16, -0.67, 0.77],
}, index =["Enero", "Febrero", "Marzo", "Abril"]))
```

Dataframe (cont.)

```
>>> print(ventas2)
```

	Cambio	Entradas	Limite	Salidas	Valoraciones
Enero	1.43	41	No	17	66
Febrero	1.16	32	Si	54	54
Marzo	-0.67	56	No	6	49
Abril	0.77	18	No	78	66

Dataframe (cont.)

```
>>> ventas2.index
Index([u'Enero', u'Febrero', u'Marzo', u'Abril'], dtype='object')
>>> ventas2.columns
Index([u'Cambio', u'Entradas', u'Limite', u'Salidas',
u'Valoraciones'], dtype='object')
```

Dataframe (cont.)

```
>>> ventas2.dtypes
Cambio          float64
Entradas          int64
Limite           object
Salidas           int64
Valoraciones      int64
dtype: object
>>>
```

EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Dataframe (cont.)

```
ventas2.index.name = "Meses"
ventas2.index.columns = "Métricas"
```

Métricas	Cambio	Entradas	Limite	Salidas	Valoraciones
Meses					
Enero	1.43	41	No	17	66
Febrero	1.16	32	Si	54	54
Marzo	-0.67	56	No	6	49
Abril	0.77	18	No	78	66

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Combinando Series y Dataframe

```
entradas = pd.Series([11, 18, 12, 16, 9, 16, 22, 28, 31, 29, 30, 12],
    index = ["ene", "feb", "mar", "abr", "may", "jun", "jul", "ago",
"sep", "oct", "nov", "dic"])
print(entradas)


salidas = pd.Series([9, 26, 18, 15, 6, 22, 19, 25, 34, 22, 21, 14],
    index = ["ene", "feb", "mar", "abr", "may", "jun", "jul", "ago",
"sep", "oct", "nov", "dic"])
print(salidas)
```

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ CENUR NORESTE UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Combinando Series y Dataframe (cont.)

```
almacen = pd.DataFrame({"entradas": entradas, "salidas": salidas})
almacen["neto"] = almacen.entradas - almacen.salidas
print(almacen)
```

Combinando Series y Dataframe (cont.)

```
>>> print(almacen)
     entradas  salidas  neto
ene        11        9     2
feb        18       26    -8
mar        12       18    -6
abr        16       15     1
may         9        6     3
jun        16       22    -6
jul        22       19     3
ago        28       25     3
sep        31       34    -3
oct        29       22     7
nov        30       21     9
dic        12       14    -2
```

```
>>> print(almacen.head())
     entradas  salidas  neto
ene        11        9     2
feb        18       26    -8
mar        12       18    -6
abr        16       15     1
may         9        6     3
```

```
>>> print(almacen.tail())
     entradas  salidas  neto
ago        28       25     3
sep        31       34    -3
oct        29       22     7
nov        30       21     9
dic        12       14    -2
```

EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Método describe

almacen.describe()

	entradas	salidas	neto
count	12.00000	12.000000	12.000000
mean	19.50000	19.250000	0.250000
std	8.16311	7.641097	5.310795
min	9.00000	6.000000	-8.000000
25%	12.00000	14.750000	-3.750000
50%	17.00000	20.000000	1.500000
75%	28.25000	22.750000	3.000000
max	31.00000	34.000000	9.000000

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ CENUR NORESTE UNIVERSIDAD DE LA REPÚBLICA URUGUAY

El método value_counts

Este método devuelve una estructura conteniendo los valores presentes en la serie y el número de ocurrencias de cada uno.

El método value_counts

```
s = pd.Series([3, 1, 2, 1, 1, 4, 1, 2, np.nan])
s.value_counts()
```

```
1.0    4
2.0    2
4.0    1
3.0    1
```

EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Agrupar los datos en "bins"

```
s.value_counts(bins = 2)
```

```
(0.996, 2.5]    6
(2.5, 4.0]      2
```

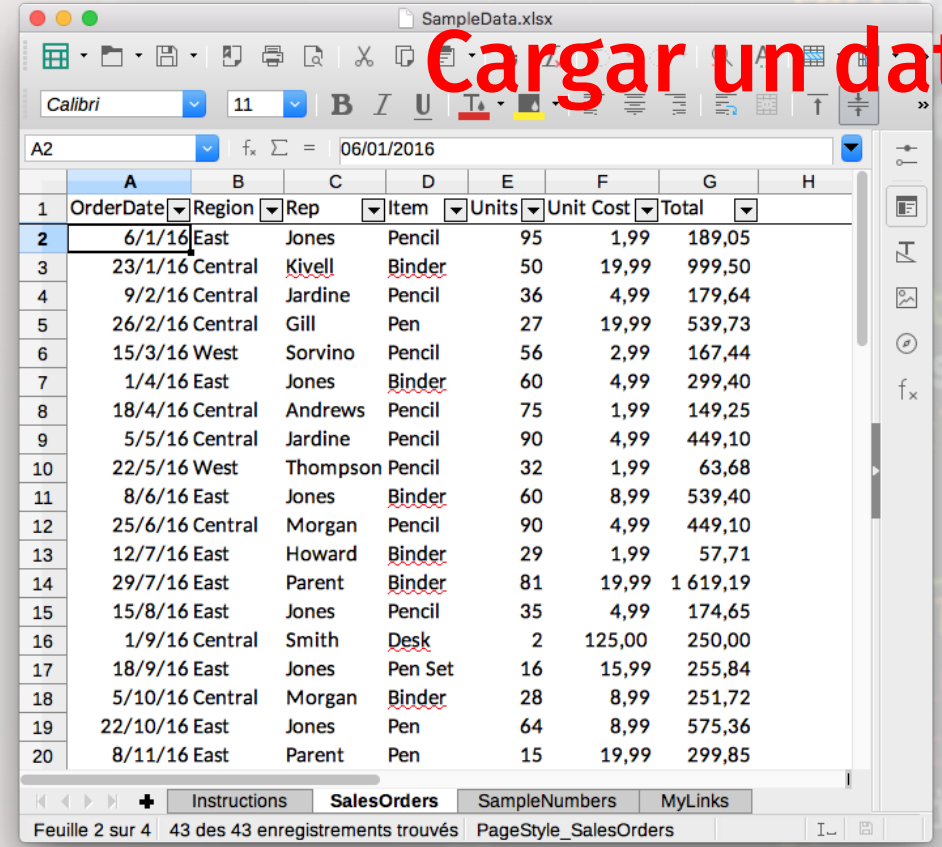
Cargar un dataframe desde un diccionario

```
students = [{ "name":"Jorge", "surname":"Perez", "age":24, "weight":50, "height": 170},
{ "name":"Pepe", "surname":"Garcia", "age":27, "weight":60, "height": 175},
{ "name":"Aria", "surname":"Jimenez", "age":26, "weight":70, "height": 180},
{ "name":"Maria", "surname":"Ruz", "age":25, "weight":75, "height": 181},
{ "name":"Luisa", "surname":"Perez", "age":24, "weight":50, "height": 170},
{ "name":"Luisa", "surname":"Perez", "age":24, "weight":50, "height": 170}]


df = pd.DataFrame(students)
```

Cargar un dataframe desde un archivo

```
ds.to_csv(filename) #desde archivo .csv

ds.to_excel(filename) #desde archivo Excel
```

Cargar un dataframe desde un archivo



	A	B	C	D	E	F	G	H
1	OrderDate	Region	Rep	Item	Units	Unit Cost	Total	
2	6/1/16	East	Jones	Pencil	95	1,99	189,05	
3	23/1/16	Central	Kivell	Binder	50	19,99	999,50	
4	9/2/16	Central	Jardine	Pencil	36	4,99	179,64	
5	26/2/16	Central	Gill	Pen	27	19,99	539,73	
6	15/3/16	West	Sorvino	Pencil	56	2,99	167,44	
7	1/4/16	East	Jones	Binder	60	4,99	299,40	
8	18/4/16	Central	Andrews	Pencil	75	1,99	149,25	
9	5/5/16	Central	Jardine	Pencil	90	4,99	449,10	
10	22/5/16	West	Thompson	Pencil	32	1,99	63,68	
11	8/6/16	East	Jones	Binder	60	8,99	539,40	
12	25/6/16	Central	Morgan	Pencil	90	4,99	449,10	
13	12/7/16	East	Howard	Binder	29	1,99	57,71	
14	29/7/16	East	Parent	Binder	81	19,99	1 619,19	
15	15/8/16	East	Jones	Pencil	35	4,99	174,65	
16	1/9/16	Central	Smith	Desk	2	125,00	250,00	
17	18/9/16	East	Jones	Pen Set	16	15,99	255,84	
18	5/10/16	Central	Morgan	Binder	28	8,99	251,72	
19	22/10/16	East	Jones	Pen	64	8,99	575,36	
20	8/11/16	East	Parent	Pen	15	19,99	299,85	

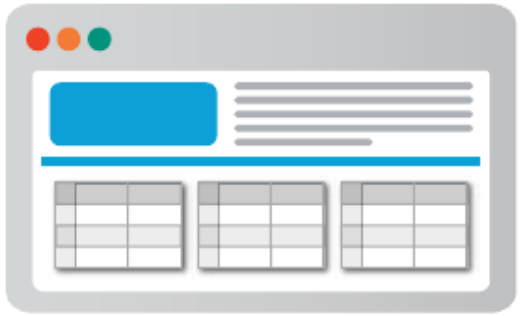
Instructions | SalesOrders | SampleNumbers | MyLinks

Feuille 2 sur 4 | 43 des 43 enregistrements trouvés | PageStyle_SalesOrders

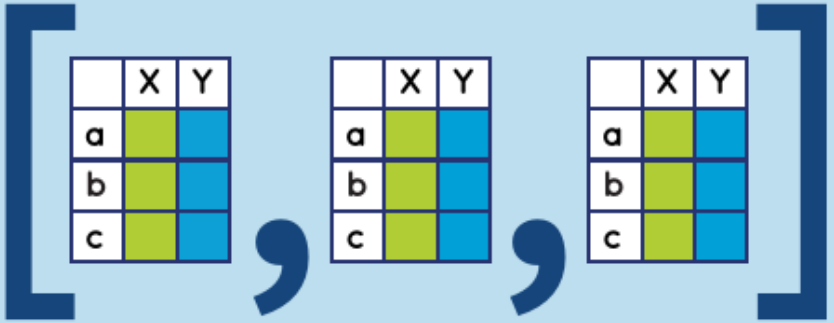
```
df = pd.read_excel('SampleData.xlsx', sheet_name='SalesOrders')
```

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

"Parseo" de tablas html



```
>>> df_list = read_html(url)
```



"Parseo" de tablas html(cont.)

```
dfs = pd.read_html('http://www.contextures.com/xlSampleData01.html')
```

OrderDate	Region	Rep	Item	U
1/6/2019	East	Jones	Pencil	
1/23/2019	Central	Kivell	Binder	
2/9/2019	Central	Jardine	Pencil	
2/26/2019	Central	Gill	Pen	
3/15/2019	West	Sorvino	Pencil	
4/1/2019	East	Jones	Binder	
4/18/2019	Central	Andrews	Pencil	

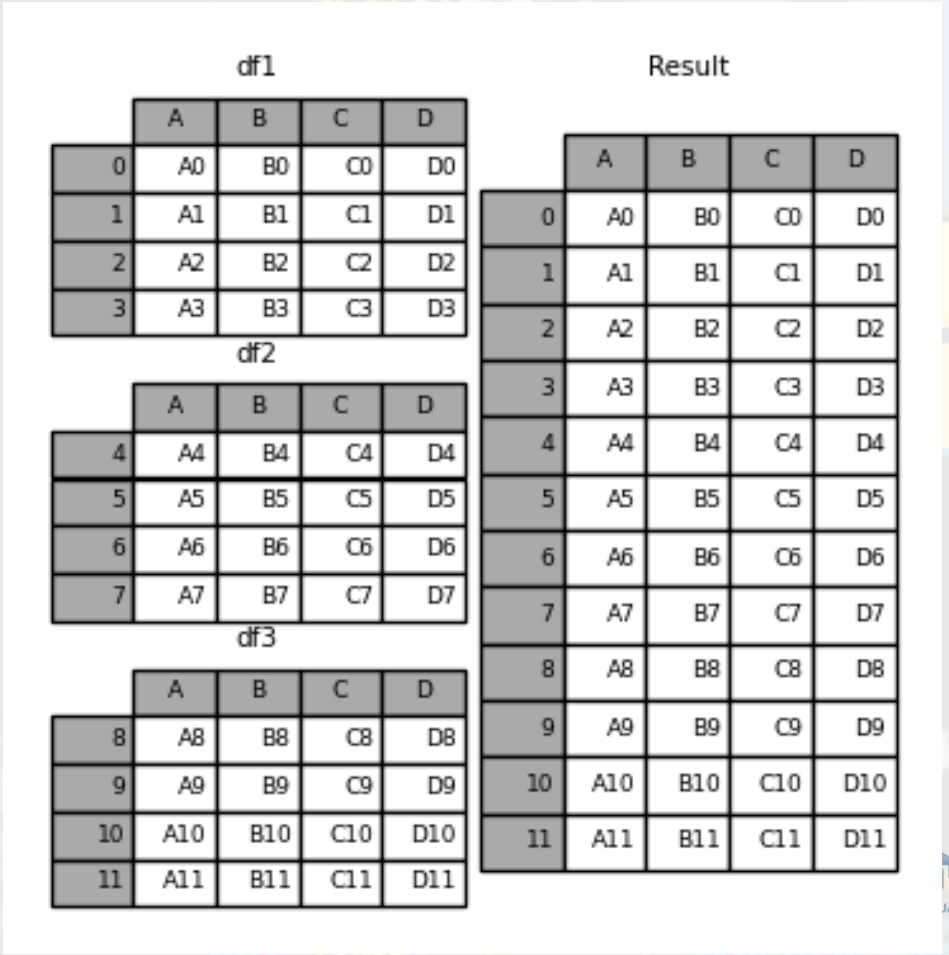
EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Combinando Dataframes

- Concatenación
- Merge
- Join

EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Concatenacion Dataframes



df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df2

	A	B	C	D
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7

df3

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

Result

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

EDUCACION PERMANENTE
Universidad de la República
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Concatenacion Dataframes(cont.)

```
import pandas as pd

df1 = pd.read_csv('nombres2010-2014.csv')

df2 = pd.read_csv('nombres2015.csv')

df3 = pd.concat([df1, df2])
```

Concatenacion Dataframes(cont.)

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df4

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Result

	A	B	C	D	B	D	F
0	A0	B0	C0	D0	NaN	NaN	NaN
1	A1	B1	C1	D1	NaN	NaN	NaN
2	A2	B2	C2	D2	B2	D2	F2
3	A3	B3	C3	D3	B3	D3	F3
6	NaN	NaN	NaN	NaN	B6	D6	F6
7	NaN	NaN	NaN	NaN	B7	D7	F7

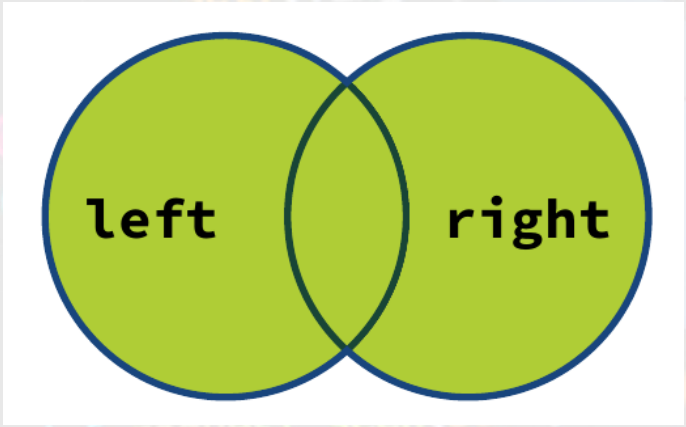
```
pd.concat([df1, df4], axis=1, sort=False)
```

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ CENUR NORESTE UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Merge Dataframes

pd.merge(df_left, df_rigth, how, on)

Merge Dataframes – how ='outer'



left

	long	X
0	aaaa	a
1	bbbb	b

left_on='X' **right_on='Y'**

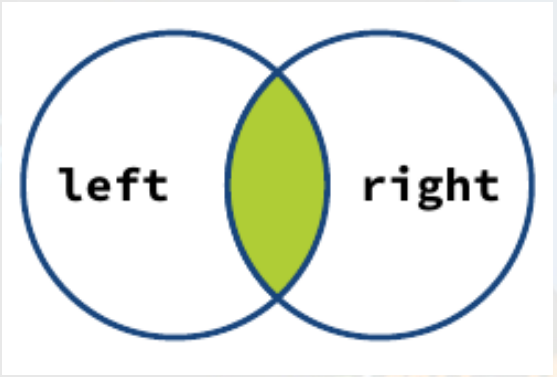
	long	X	Y	short
0	aaaa	a	—	—
1	bbbb	b	b	bb
2	—	—	c	cc

right

	Y	short
0	b	bb
1	c	cc

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ CENUR NORESTE UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Merge Dataframes – how ='inner'



	long	X
0	aaaa	a
1	bbbb	b

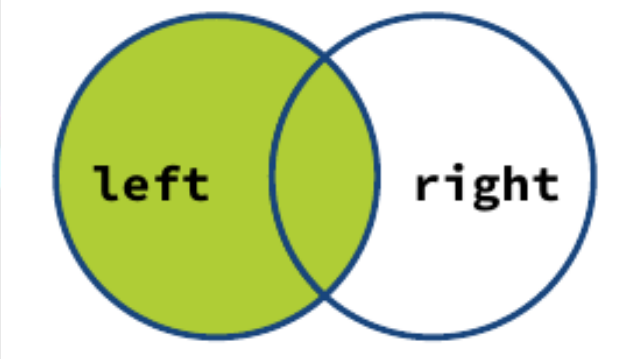
→

	long	X	Y	short
0	bbbb	b	b	bb

←

	Y	short
0	b	bb
1	c	cc

Merge Dataframes – how ='left'



	long	X
0	aaaa	a
1	bbbb	b

→

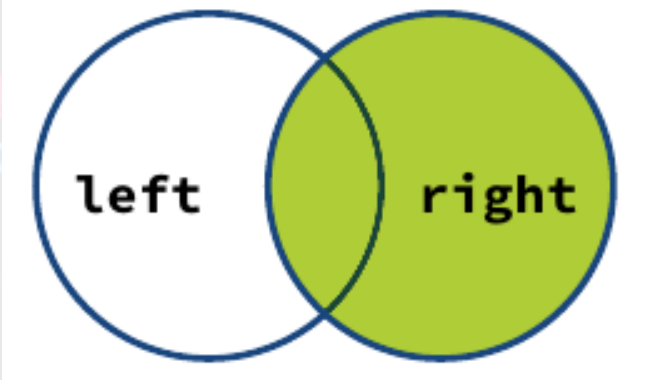
	long	X	Y	short
0	aaaa	a	—	—
1	bbbb	b	b	bb

←

	Y	short
0	b	bb
1	c	cc

EDUCACION PERMANENTE
Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Merge Dataframes – how ='rigth'



	long	X
0	aaaa	a
1	bbbb	b

→

	long	X	Y	short
0	bbbb	b	b	bb
1	—	—	c	cc

←

	Y	short
0	b	bb
1	c	ctc

Join por Indices

```
df1 = pd.DataFrame({
    "Month": ["ene", "feb", "mar", "may"],
    "Sales": [14, 8, 12, 17]
})
df1
```

	Month	Sales
0	ene	14
1	feb	8
2	mar	12
3	may	17

```
df2 = pd.DataFrame({
    "Purchases": [5, 9, 11, 2, 6]
},
    index = ["ene", "feb", "mar", "abr", "may"]
)
df2
```

	Purchases
ene	5
feb	9
mar	11
abr	2
may	6

```
pd.merge(df1, df2, left_on = "Month", right_index = True)
```

	Month	Sales	Purchases
0	ene	14	5
1	feb	8	9
2	mar	12	11
3	may	17	6

SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Ordenar por valor

```
df = pd.DataFrame({"A":[3, 2, 2, 0],
                   "B":[1, 2, 2, 0],
                   "C":[0, 3, 1, 5],
                   "D":[2, 4, 5, 6]
                   },
                  index = ["a", "b", "c", "d"])
df
```

	A	B	C	D
a	3	1	0	2
b	2	2	3	4
c	2	2	1	5
d	0	0	5	6

```
df.sort_values(by = "A")
```

	A	B	C	D
d	0	0	5	6
b	2	2	3	4
c	2	2	1	5
a	3	1	0	2

```
df.sort_values(by = ["a", "b"], axis = 1, ascending = False)
```

	A	D	B	C
a	3	2	1	0
b	2	4	2	3
c	2	5	2	1
d	0	6	0	5

EDUCACION PERMANENTE
Universidad de la República

Agrupamientos

```
ventas = pd.DataFrame({
    "Producto":["A", "B", "C", "B", "A", "A"],
    "Ventas": [6, 2, 1, 4, 5, 2]
})
ventas
```

	Producto	Ventas
0	A	6
1	B	2
2	C	1
3	B	4
4	A	5
5	A	2

```
ventas.groupby(by = "Producto").mean()
```

Producto	Ventas
A	4.333333
B	3.000000
C	1.000000

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ CENUR NORESTE UNIVERSIDAD DE LA REPÚBLICA URUGUAY

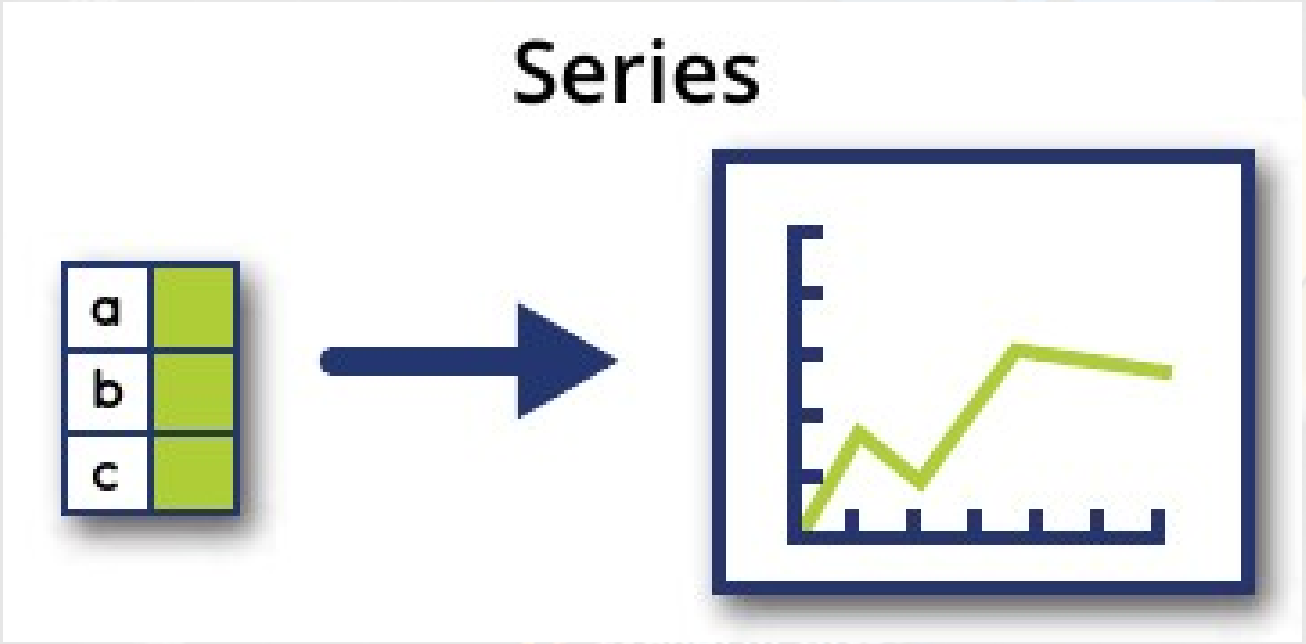
Filtrado de datos

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embark
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

```
sobrevivientes =  titanic[titanic['Survived']== 1]
sobrevivientes['Survived'].value_counts()
```

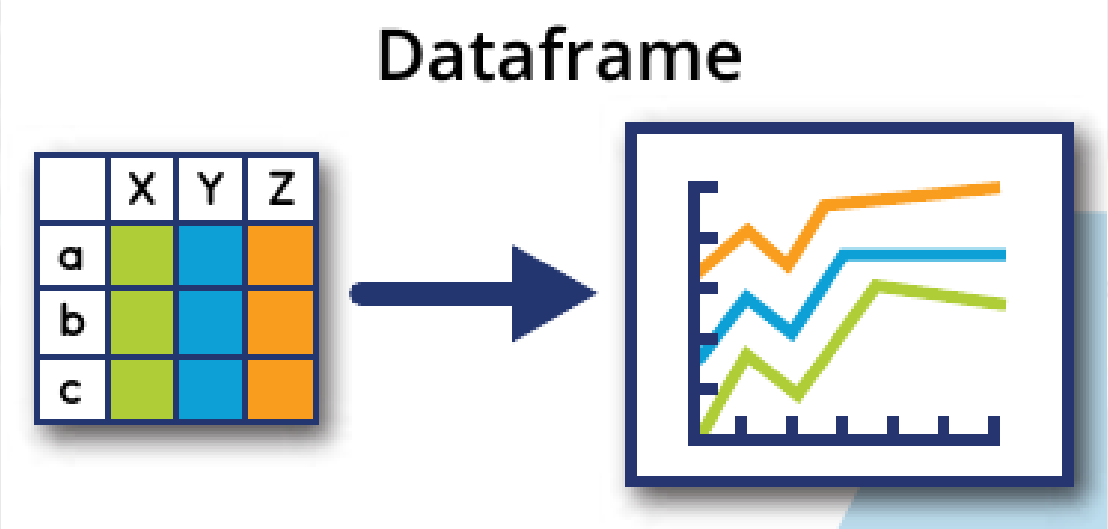
```
sobrevivientes_genero = sobrevivientes.groupby(by = "Sex").count()
sobrevivientes_genero['PassengerId']
```

Graficas



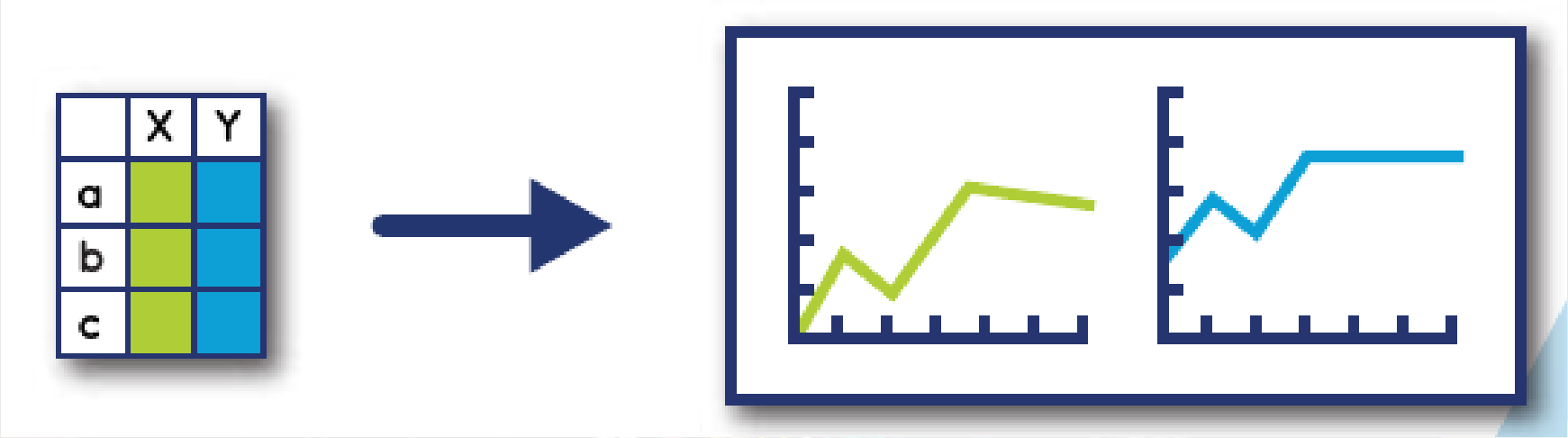
Con una serie, Pandas traza valores contra el índice: ax = s.plot()

Graficas (cont.)



Con un DataFrame, Pandas crea una línea por Columna: ax = df.plot()

Graficas (cont.)



subplots=True: una subtrama por columna, en lugar de una línea
figsize: tamaño de la figura fija
x and y: trazar una columna contra otra

EDUCACION PERMANENTE Universidad de la República
SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Graficas (cont.)