Procesamiento de datos

Series y DataFrame de Pandas

Series

© ESPOL 2022

Series

- Arreglos de 1 dimensión etiquetados, estos pueden tener como índices otros objetos como cadenas

ARREGLO	10	70	120	NUMPY
	0	1	2	
SERIE	10	70	120	PANDAS
	'a'	'b'	'c'	

© ESPOL 2022

Crear series e indexación

Creación

```
>>>import pandas as pd
>>>serie = pd.Series([10, 70, 120])
>>>serie
0     10
1     70
2    120
dtype: int64
```

con índices enteros

```
>>>serie = pd.Series([10, 70, 120],
index=['a','b','c'])
a     10
b     70
c    120
dtype: int64
```

con índices texto

Indexación

```
>>> serie[0]
10
>>> serie[2]
120
```

Slicing

```
>>> serie[0:2]
a    10
b    70
dtype: int64
```


© ESPOL 2022

Crear series a partir de colecciones

```
>>>serie = pd.Series(np.random.random(4),list('badc'))
>>> serie
b    0.132439
a    0.644991
d    0.713788
c    0.374691
dtype: float64
```

A partir de un arreglo de 1 dimensión

A partir de un dicccionario

```
>>>dic = {'a': 0.13, 'b': 0.64, 'c': 0.71, 'd': 0.37}

>>>serie = pd.Series(dic)
a    0.13
b    0.64
c    0.71
d    0.37
dtype: float64
```

Los índices para la serie se orden de manera ascendente
Para cambiar el orden de los índices incluir el segundo argumento

© ESPOL 2022

Similitudes con numpy

```
>>>s = pd.Series(np.random.random(5),index=list('abcde'))
a    0.275891
b    0.024151
c    0.578902
d    0.763546
e    0.801908
dtype: float64
>>>s[0]
0.275891
>>> s[2:]
c    0.578902
d    0.763546
e    0.801908
dtype: float64
```

```
>>>s[s>0.3]
c    0.578902
d    0.763546
e    0.801908
dtype: float64

>>>s[[4,3,1]]
e    0.801908
d    0.763546
b    0.024151
dtype: float64
```


© ESPOL 2022

Similitudes con diccionarios

```
>>>s['a']
0.275891

>>>s['e'] = 12
>>>s
a    0.275891
b    0.024151
c    0.578902
d    0.763546
e   12.000000
dtype: float64
```

```
>>>s['f'] = 30
>>>s
a    0.275891
b    0.024151
c    0.578902
d    0.763546
e   12.000000
f   30.000000
dtype: float64

>>>'e' in s
True
>>>'g' in s
False
```


© ESPOL 2022

DataFrame

© ESPOL 2022

DataFrame

- Arreglo de numpy de 2 dimensiones (matriz) etiquetado en las columnas y en las filas.
- En la práctica un DataFrame es una colección de columnas donde cada columa es una Series de pandas
- Se puede crear a partir de diferentes objetos:
  - Diccionario simple con objetos Listas o Series
  - Arreglos de numpy de 2D
  - Una Serie
  - Otros DataFrames

© ESPOL 2022

DataFrame como matriz etiquetada

	1	2
0	23	45
1	72	81
2	56	64
3	34	75

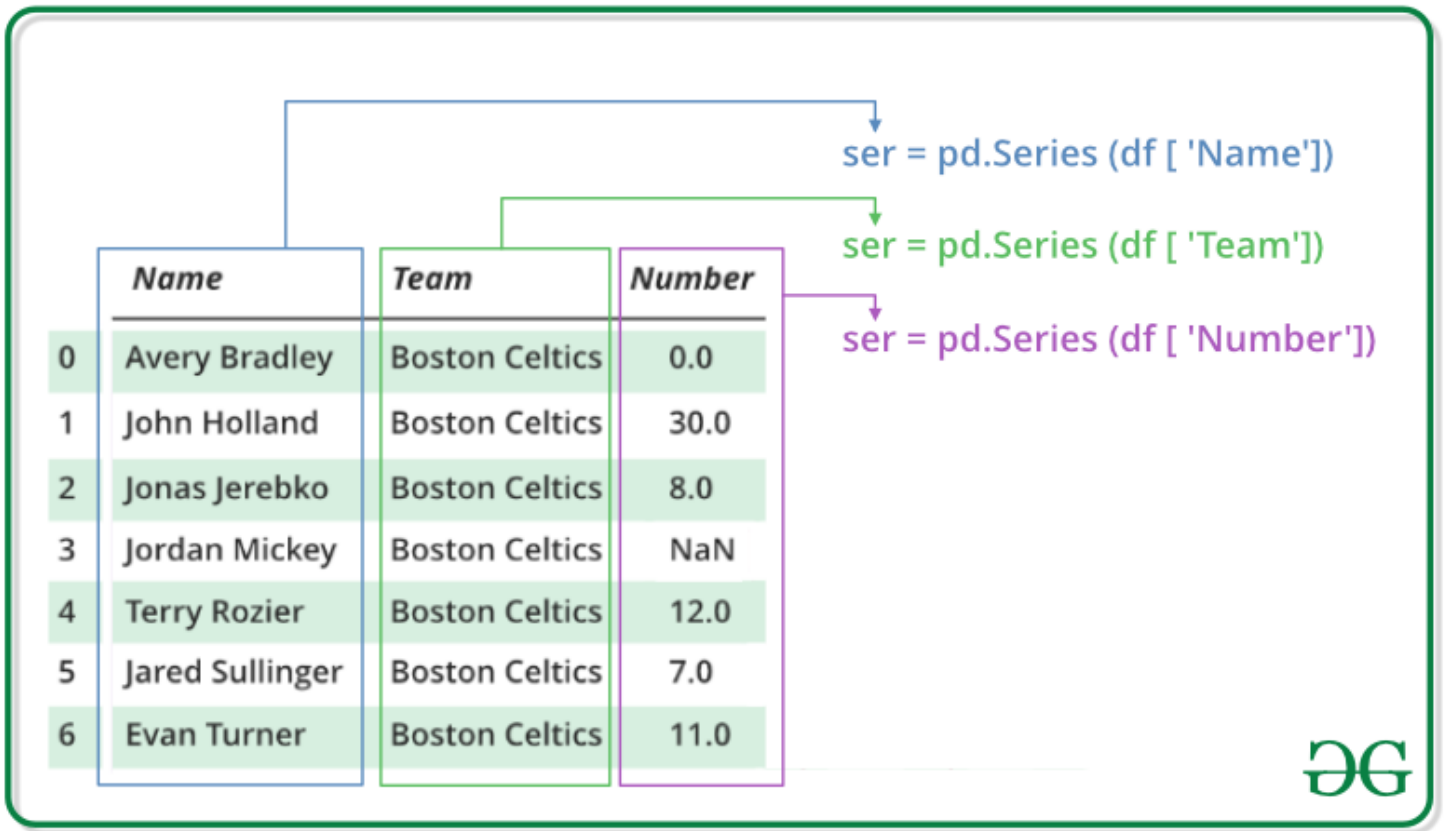
NUMPY

PANDAS

	alturas	pesos
Jorge	23	45
Joshua	72	81
Raul	56	64
Santiago	34	75

© ESPOL 2022

DataFrame como colección de serie



© ESPOL 2022

DataFrame

```
>>> M = [[65, 74],
...      [68, 72],
...      [63, 71],
...      [70, 82]]
>>> frame = pd.DataFrame(M))
>>> print(frame)
```

```
    0   1
0  65  74
1  68  72
2  63  71
3  70  82
```

```
>>> personas = ['Jorge','Joshua','Raul','Santiago']

>>> >>> frame = pd.DataFrame(M,columns=['estaturas','pesos'],
...     index=personas)

>>> print(frame)
```

```
           estaturas  pesos
Jorge             65     74
Joshua            68     72
Raul              63     71
Santiago          70     82
```


© ESPOL 2022

Crear a partir de un arreglo de numpy

```
>>> m1 = np.random.randint(60, 80, 12)
>>> m1 = m1.reshape(3, 4)
>>> frame = pd.DataFrame(m1, columns=list('abcd'), index=lists('ABC')
>>> frame
```

	a	b	c	d
A	64	78	60	63
B	69	66	72	75
C	64	60	60	76

© ESPOL 2022

Crear con diccionario simple

```
d = { 'states': ['Ohio','Ohio','Ohio','Nevada','Nevada'],
      'year': [2000, 2001, 2002, 2001, 2002],
      'popu': [1.5, 1.7, 3.6, 2.4, 2.9]}

>>> frame = pd.DataFrame(d)
>>> frame
   popu  states  year
0  1.5     Ohio  2000
1  1.7     Ohio  2001
2  3.6     Ohio  2002
3  2.4   Nevada  2001
4  2.9   Nevada  2002
```

```
>>> frame = pd.DataFrame(d,
  columns=['year','states','popu'])
>>> frame
   year  states  popu
0  2000    Ohio  1.5
1  2001    Ohio  1.7
2  2002    Ohio  3.6
3  2001  Nevada  2.4
4  2002  Nevada  2.9
```


© ESPOL 2022

Crear con diccionario compuesto

```
import numpy as np
import pandas as pd

rankings = {
   2006: {'Jorge': 5, 'Joshua': 3, 'Raul': 234 },
   2004: {'Jorge': 5, 'Joshua': 8, 'Raul': 1000 },
   2002: {'Jorge': 10, 'Joshua': 6 },
   2000: {'Jorge': 5 } }

df = pd.DataFrame(rankings)
print(df)
```

```
        2000  2002  2004  2006
Jorge    5.0  10.0     5     5
Joshua   NaN   6.0     8     3
Raul     NaN   NaN  1000   234
```

```
df = df.fillna(0)
print(df)
```

```
        2000  2002  2004  2006
Jorge    5.0  10.0     5     5
Joshua   0.0   6.0     8     3
Raul     0.0   0.0  1000   234
```


© ESPOL 2022

Manejo de archivos

© ESPOL 2022

Leer datos de archivos

- Leer desde un archivo csv

```
csv_path = 'TopSellingAlbums.csv'
df = pd.read_csv(csv_path)
# Print first five rows of the dataframe
df.head()
```

- Leer desde un archivo csv con un separador distinto

```
csv_path = 'TopSellingAlbums.csv'
df = pd.read_csv(csv_path, sep=';')
# Print first five rows of the dataframe
df.head()
```


© ESPOL 2022

Acceder a una columna del dataframe

x=df[['Length']]

	Artist	Album	Released	Length	Genre	Music Recording Sales (millions)	Claimed Sales (millions)	Released.1	Soundtrack	Rating
0	Michael Jackson	Thriller	1982	0:42:19	pop, rock, R&B	46.0	65	30-Nov-82	NaN	10.0
1	AC/DC	Back in Black	1980	0:42:11	hard rock	26.1	50	25-Jul-80	NaN	9.5
2	Pink Floyd	The Dark Side of the Moon	1973	0:42:49	progressive rock	24.2	45	01-Mar-73	NaN	9.0
3	Whitney Houston	The Bodyguard	1992	0:57:44	R&B, soul, pop	27.4	44	17-Nov-92	Y	8.5
4	Meat Loaf	Bat Out of Hell	1977	0:46:33	hard rock, progressive rock	20.6	43	21-Oct-77	NaN	8.0
5	Eagles	Their Greatest Hits (1971-1975)	1976	0:43:08	rock, soft rock, folk rock	32.2	42	17-Feb-76	NaN	7.5
6	Bee Gees	Saturday Night Fever	1977	1:15:54	disco	20.6	40	15-Nov-77	Y	7.0
7	Fleetwood Mac	Rumours	1977	0:40:01	soft rock	27.9	40	04-Feb-77	NaN	6.5

x

	Length
0	0:42:19
1	0:42:11
2	0:42:49
3	0:57:44
4	0:46:33
5	0:43:08
6	1:15:54
7	0:40:01

© ESPOL 2022

Seleccionar múltiples columnas

```
y=df[ ['Artist' ,'Length', 'Genre '] ]
```

	Artist	Album	Release	Length	Genre	Music Recording Sales (millions)	Claimed Sales (millions)	Released.1	Soundtrack	Rating
	Michael Jackson	Thriller	1982	0:42:19	pop, rock, R&B	46.0	65	30-Nov-82	NaN	10.0
	AC/DC	Back in Black	1980	0:42:11	hard rock	26.1	50	25-Jul-80	NaN	9.5
2	Pink Floyd	The Dark Side of the Moon	1973	0:42:49	progressive rock	24.2	45	01-Mar-73	NaN	9.0
3	Whitney Houston	The Bodyguard	1992	0:57:44	R&B, soul, pop	27.4	44	17-Nov-92	Y	8.5
4	Meat Loaf	Bat Out of Hell	1977	0:46:33	hard rock, progressive rock	20.6	43	21-Oct-77	NaN	8.0
5	Eagles	Their Greatest Hits (1971-1975)	1976	0:43:08	rock, soft rock, folk rock	32.2	42	17-Feb-76	NaN	7.5
6	Bee Gees	Saturday Night Fever	1977	1:15:54	disco	20.6	40	15-Nov-77	Y	7.0
7	Fleetwood Mac	Rumours	1977	0:40:01	soft rock	27.9	40	04-Feb-77	NaN	6.5

y

	Artist	Length	Genre
0	Michael Jackson	0:42:19	pop, rock, R&B
1	AC/DC	0:42:11	hard rock
2	Pink Floyd	0:42:49	progressive rock
3	Whitney Houston	0:57:44	R&B, soul, pop
4	Meat Loaf	0:46:33	hard rock, progressive rock
5	Eagles	0:43:08	rock, soft rock, folk rock
6	Bee Gees	1:15:54	disco
7	Fleetwood Mac	0:40:01	soft rock

© ESPOL 2022

Acceder a celdas con .iloc

df.iloc[0,0]:'Michael Jackson'

df.iloc[1,0]: 'AC/DC'

df.iloc[0,2]:1982

df.iloc[1,2]:1980

	Artist	Album	Released	Length	Genre	Music recording sales (millions)	Claimed sales (millions)	Released	Soundtrack	Rating (friends)
0	Michael Jackson	Thriller	1982	00:42:19	Pop, rock, R&B	46	65	30-Nov-82		10.0
1	AC/DC	Back in Black	1980	00:42:11	Hard rock	26.1	50	25-Jul-80		8.5
2	Pink Floyd	The Dark Side of the Moon	1973	00:42:49	Progressive rock	24.2	45	01-Mar-73		9.5
3	Whitney Houston	The Bodyguard	1992	00:57:44	Soundtrack/R&B, soul, pop	26.1	50	25-Jul-80	Y	7.0
4	Meat Loaf	Bat Out of Hell	1977	00:46:33	Hard rock, progressive rock	20.6	43	21-Oct-77		7.0
5	Eagles	Their Greatest Hits (1971-1975)	1976	00:43:08	Rock, soft rock, folk rock	32.2	42	17-Feb-76		9.5
6	Bee Gees	Saturday Night Fever	1977	1:15:54	Disco	20.6	40	15-Nov-77	Y	9.0
7	Fleetwood Mac	Rumours	1977	00:40:01	Soft rock	27.9	40	04-Feb-77		9.5

© ESPOL 2022

Acceder a celdas con .loc

df.loc[0, 'Artist']:'Michael Jackson'

df.loc[0, 'Released']:1982

df.loc[1, 'Artist']:'AC/DC'

df.loc[1, 'Released']:1980

	Artist	Album	Released	Length	Genre	Music Recording Sales (millions)	Claimed Sales (millions)	Released.1	Soundtrack	Rating
0	Michael Jackson	Thriller	1982	0:42:19	pop, rock, R&B	46.0	65	30-Nov-82	NaN	10.0
1	AC/DC	Back in Black	1980	0:42:11	hard rock	26.1	50	25-Jul-80	NaN	9.5
2	Pink Floyd	The Dark Side of the Moon	1973	0:42:49	progressive rock	24.2	45	01-Mar-73	NaN	9.0
3	Whitney Houston	The Bodyguard	1992	0:57:44	R&B, soul, pop	27.4	44	17-Nov-92	Y	8.5
4	Meat Loaf	Bat Out of Hell	1977	0:46:33	hard rock, progressive rock	20.6	43	21-Oct-77	NaN	8.0
5	Eagles	Their Greatest Hits (1971-1975)	1976	0:43:08	rock, soft rock, folk rock	32.2	42	17-Feb-76	NaN	7.5
6	Bee Gees	Saturday Night Fever	1977	1:15:54	disco	20.6	40	15-Nov-77	Y	7.0
7	Fleetwood Mac	Rumours	1977	0:40:01	soft rock	27.9	40	04-Feb-77	NaN	6.5

© ESPOL 2022

Slicing DataFrame con .iloc

```
z=df.iloc[0:2, 0:3]
```

	Artist	Album	Released	Length	Genre	Music Recording Sales (millions)	Claimed Sales (millions)	Released.1	Soundtrack	Rating
0	Michael Jackson	Thriller	1982	0:42:19	pop, rock, R&B	46.0	65	30-Nov-82	NaN	10.0
1	AC/DC	Back in Black	1980	0:42:11	hard rock	26.1	50	25-Jul-80	NaN	9.5
2	Pink Floyd	The Dark Side of the Moon	1973	0:42:49	progressive rock	24.2	45	01-Mar-73	NaN	9.0
3	Whitney Houston	The Bodyguard	1992	0:57:44	R&B, soul, pop	27.4	44	17-Nov-92	Y	8.5
4	Meat Loaf	Bat Out of Hell	1977	0:46:33	hard rock, progressive rock	20.6	43	21-Oct-77	NaN	8.0
5	Eagles	Their Greatest Hits (1971-1975)	1976	0:43:08	rock, soft rock, folk rock	32.2	42	17-Feb-76	NaN	7.5
6	Bee Gees	Saturday Night Fever	1977	1:15:54	disco	20.6	40	15-Nov-77	Y	7.0
7	Fleetwood Mac	Rumours	1977	0:40:01	soft rock	27.9	40	04-Feb-77	NaN	6.5

z

	Artist	Album	Released
0	Michael Jackson	Thriller	1982
1	AC/DC	Back in Black	1980

© ESPOL 2022

Slicing DataFrame con .loc

```
df.loc[0:2, 'Artist':'Released']
```

	Artist	Album	Released	Length	Genre	Music Recording Sales (millions)	Claimed Sales (millions)	Released.1	Soundtrack	Rating
0	Michael Jackson	Thriller	1982	0:42:19	pop, rock, R&B	46.0	65	30-Nov-82	NaN	10.0
1	AC/DC	Back in Black	1980	0:42:11	hard rock	26.1	50	25-Jul-80	NaN	9.5
2	Pink Floyd	The Dark Side of the Moon	1973	0:42:49	progressive rock	24.2	45	01-Mar-73	NaN	9.0
3	Whitney Houston	The Bodyguard	1992	0:57:44	R&B, soul, pop	27.4	44	17-Nov-92	Y	8.5
4	Meat Loaf	Bat Out of Hell	1977	0:46:33	hard rock, progressive rock	20.6	43	21-Oct-77	NaN	8.0
5	Eagles	Their Greatest Hits (1971-1975)	1976	0:43:08	rock, soft rock, folk rock	32.2	42	17-Feb-76	NaN	7.5
6	Bee Gees	Saturday Night Fever	1977	1:15:54	disco	20.6	40	15-Nov-77	Y	7.0
7	Fleetwood Mac	Rumours	1977	0:40:01	soft rock	27.9	40	04-Feb-77	NaN	6.5

z

	Artist	Album	Released
0	Michael Jackson	Thriller	1982
1	AC/DC	Back in Black	1980
2	Pink Floyd	The Dark Side of the Moon	1973

© ESPOL 2022

Filtras datos con la indexación por columna

```
>>> frame
   states  year  popu
0    Ohio  2000   1.5
1    Ohio  2001   1.7
2    Ohio  2002   3.6
3  Nevada  2001   2.4
4  Nevada  2002   2.9
```

```
>>> df = frame [ frame['year'] > 2001 ]
>>> df
   year  states  popu
2  2002    Ohio  3.6
4  2002  Nevada  2.9
```

```
>>> df = frame [ (frame['year'] > 2000) & (frame['popu'] > 2) ]
>>> df
   states  year  popu
2    Ohio  2002   3.6
3  Nevada  2001   2.4
4  Nevada  2002   2.9
```


© ESPOL 2022