Zeppelin

Lab5

default

FINISHED

```
%pyspark
from pandas import Series, DataFrame
import pandas as pd
obj = Series([4, 7, -5, 3])
obj
```

```
0    4
1    7
2   -5
3    3
dtype: int64
```

Took 5 sec. Last updated by anonymous at February 16 2017, 6:33:26 PM.

FINISHED

```
%pyspark
print(obj.values)
print(obj.index)
```

```
[ 4  7 -5  3]
RangeIndex(start=0, stop=4, step=1)
```

Took 0 sec. Last updated by anonymous at February 16 2017, 6:38:03 PM.

FINISHED

```
%pyspark
obj2 = Series([4, 7, -5, 3], index=['d', 'b', 'a', 'c'])
print(obj2)
print(obj2.index)
print(obj2['a'])
obj2['d'] = 6
obj2[['c', 'a', 'd']]
obj2
```

```
d    4
b    7
a   -5
c    3
dtype: int64
Index([u'd', u'b', u'a', u'c'], dtype='object')
-5
d    6
b    7
a   -5
c    3
dtype: int64
```

Took 0 sec. Last updated by anonymous at February 16 2017, 6:48:12 PM.

FINISHED

```
%pyspark
print(obj2[obj2 > 0])
print(obj2 * 2)
```

```
d    6
b    7
```

```
d    12
b    14
a   -10
c     6
dtype: int64
```

Took 0 sec. Last updated by anonymous at February 16 2017, 6:48:26 PM.

FINISHED

```
%pyspark
import numpy as np
np.exp(obj2)
print('b' in obj2)
print('e' in obj2)
```

```
True
False
```

Took 0 sec. Last updated by anonymous at February 16 2017, 6:53:14 PM.

FINISHED

```
%pyspark
sdata = {'Ohio': 35000, 'Texas': 71000, 'Oregon': 16000, 'Utah': 5000}
obj3 = Series(sdata)
obj3
```

```
Ohio      35000
Oregon    16000
Texas     71000
Utah       5000
dtype: int64
```

Took 0 sec. Last updated by anonymous at February 16 2017, 6:58:36 PM.

FINISHED

```
%pyspark
states = ['California', 'Ohio', 'Oregon', 'Texas']
obj4 = Series(sdata, index=states)
obj4
```

```
California        NaN
Ohio          35000.0
Oregon        16000.0
Texas         71000.0
dtype: float64
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:04:53 PM.

FINISHED

```
%pyspark
print(pd.isnull(obj4))
print(pd.notnull(obj4))
print(obj4.isnull())
print(obj3)
print(obj4)
print(obj3 + obj4)
obj4.name = 'population'
obj4.index.name = 'state'
obj4
```

```
state
California     True
Ohio          False
Oregon        False
```

```
Texas        False
Name: population, dtype: bool
state
California    False
Ohio           True
Oregon         True
Texas          True
Name: population, dtype: bool
state
California     True
Ohio          False
Oregon        False
Texas         False
Name: population, dtype: bool
Ohio      35000
Oregon    16000
Texas     71000
Utah       5000
dtype: int64
state
California        NaN
Ohio          35000.0
Oregon        16000.0
Texas         71000.0
Name: population, dtype: float64
California         NaN
Ohio           70000.0
Oregon         32000.0
Texas         142000.0
Utah               NaN
dtype: float64
state
California        NaN
Ohio          35000.0
Oregon        16000.0
Texas         71000.0
Name: population, dtype: float64
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:43:11 PM. (outdated)

FINISHED

```
%pyspark
obj.index = ['Bob', 'Steve', 'Jeff', 'Ryan']
obj
data = {'state': ['Ohio', 'Ohio', 'Ohio', 'Nevada', 'Nevada'],
        'year': [2000, 2001, 2002, 2001, 2002],
        'pop': [1.5, 1.7, 3.6, 2.4, 2.9]}
frame = DataFrame(data)
frame
```

```
   pop   state  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
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:44:36 PM.

FINISHED

```
%pyspark
DataFrame(data, columns=['year', 'state', 'pop'])
frame2 = DataFrame(data, columns=['year', 'state', 'pop', 'debt'],
                   index=['one', 'two', 'three', 'four', 'five'])
print(frame2)
print(frame2.columns)
print(frame2['state'])
print(frame2.year)
print(frame2.ix['three'])
frame2['debt'] = 16.5
print(frame2)
frame2['debt'] = np.arange(5.)
```

```
       year   state  pop debt
one    2000    Ohio  1.5  NaN
two    2001    Ohio  1.7  NaN
three  2002    Ohio  3.6  NaN
four   2001  Nevada  2.4  NaN
five   2002  Nevada  2.9  NaN
Index([u'year', u'state', u'pop', u'debt'], dtype='object')
one        Ohio
two        Ohio
three      Ohio
four     Nevada
five     Nevada
Name: state, dtype: object
one      2000
two      2001
three    2002
four     2001
five     2002
Name: year, dtype: int64
year     2002
state    Ohio
pop       3.6
debt      NaN
Name: three, dtype: object
       year   state  pop  debt
one    2000    Ohio  1.5  16.5
two    2001    Ohio  1.7  16.5
three  2002    Ohio  3.6  16.5
four   2001  Nevada  2.4  16.5
five   2002  Nevada  2.9  16.5
       year   state  pop  debt
one    2000    Ohio  1.5   0.0
two    2001    Ohio  1.7   1.0
three  2002    Ohio  3.6   2.0
four   2001  Nevada  2.4   3.0
five   2002  Nevada  2.9   4.0
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:49:57 PM. (outdated)

FINISHED

```
%pyspark
val = Series([-1.2, -1.5, -1.7], index=['two', 'four', 'five'])
frame2['debt'] = val
frame2
```

```
       year   state  pop  debt
one    2000    Ohio  1.5   NaN
two    2001    Ohio  1.7  -1.2
three  2002    Ohio  3.6   NaN
four   2001  Nevada  2.4  -1.5
```

```
five  2002  Nevada  2.9  -1.7
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:52:30 PM.

Lab5

default

FINISHED

```
%pyspark
frame2['eastern'] = frame2.state == 'Ohio'
print(frame2)
del frame2['eastern']
print(frame2.columns)
```

```
       year   state  pop  debt eastern
one    2000    Ohio  1.5   NaN    True
two    2001    Ohio  1.7  -1.2    True
three  2002    Ohio  3.6   NaN    True
four   2001  Nevada  2.4  -1.5   False
five   2002  Nevada  2.9  -1.7   False
Index([u'year', u'state', u'pop', u'debt'], dtype='object')
```

Took 1 sec. Last updated by anonymous at February 16 2017, 7:53:02 PM.

FINISHED

```
%pyspark
pop = {'Nevada': {2001: 2.4, 2002: 2.9},
       'Ohio': {2000: 1.5, 2001: 1.7, 2002: 3.6}}
frame3 = DataFrame(pop)
print(frame3)
print(frame3.T)
```

```
      Nevada  Ohio
2000     NaN   1.5
2001     2.4   1.7
2002     2.9   3.6
        2000  2001  2002
Nevada   NaN   2.4   2.9
Ohio     1.5   1.7   3.6
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:53:37 PM.

FINISHED

```
%pyspark

pdata = {'Ohio': frame3['Ohio'][:-1],
        'Nevada': frame3['Nevada'][:2]}
DataFrame(pdata)
frame3.index.name = 'year'; frame3.columns.name = 'state'
print(frame3)
print(frame3.values)
print(frame2.values)
```

```
state  Nevada  Ohio
year
2000      NaN   1.5
2001      2.4   1.7
2002      2.9   3.6
[[ nan  1.5]
 [ 2.4  1.7]
 [ 2.9  3.6]]
[[2000 'Ohio' 1.5 nan]
 [2001 'Ohio' 1.7 -1.2]
 [2002 'Ohio' 3.6 nan]
 [2001 'Nevada' 2.4 -1.5]
 [2002 'Nevada' 2.9 -1.7]]
```

Took 0 sec. Last updated by anonymous at February 16 2017, 7:55:27 PM.

Zeppelin

READY

Lab5

default