Zeppelin

DS670_Mar9_Lab

default

FINISHED

```
%pyspark
import pandas as pd
import numpy as np
```

Took 21 sec. Last updated by anonymous at March 09 2017, 7:05:45 PM.

FINISHED

```
%pyspark
df = pd.DataFrame({'key1': ['a','a','b','b','a'],
    'key2': ['one','two','one','two','one'],
    'data1' : np.random.randn(5),
    'data2' : np.random.randn(5)})
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:04 PM.

FINISHED

```
%pyspark
df
```

```
      data1     data2 key1 key2
0 -0.615551 -0.037486    a  one
1  0.765475  1.091200    a  two
2 -0.235316  0.765543    b  one
3  0.750629 -1.302508    b  two
4 -0.423394 -0.298927    a  one
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:13 PM.

FINISHED

```
%pyspark
grouped = df['data1'].groupby(df['key1'])
grouped
```

```
<pandas.core.groupby.SeriesGroupBy object at 0x10fd3a790>
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:10:56 PM.

FINISHED

```
%pyspark
grouped.mean()
```

```
key1
a   -0.091156
b    0.257657
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:12:03 PM.

FINISHED

```
%pyspark
means = df['data1'].groupby([df['key1'],df['key1']]).mean()
```

Zeppelin

DS670_Mar9_Lab

default

```
means
key1
a   -0.091156
b    0.257657
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:16:13 PM.

FINISHED

```
%pyspark
means.unstack()
```

```
key1         a         b
key1
a    -0.091156       NaN
b          NaN  0.257657
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:17:28 PM.

FINISHED

```
%pyspark
states = np.array(['Ohio','California','California','Ohio','Ohio'])
years = np.array([2005,2005,2006,2005,2006])

df['data1'].groupby([states,years]).mean()
```

```
California  2005    0.765475
            2006   -0.235316
Ohio        2005    0.067539
            2006   -0.423394
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:22:26 PM.

FINISHED

```
%pyspark
df.groupby('key1').mean()
```

```
         data1     data2
key1
a    -0.091156  0.251596
b     0.257657 -0.268483
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:26 PM.

FINISHED

```
%pyspark
df.groupby(['key1','key2']).mean()
```

```
              data1     data2
key1 key2
a    one  -0.519472 -0.168207
     two   0.765475  1.091200
b    one  -0.235316  0.765543
     two   0.750629 -1.302508
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:21 PM.

FINISHED

```
%pyspark
```

```
df.groupby(['key1','key2']).size()
```

```
key1  key2
a     one     2
      two     1
b     one     1
      two     1
dtype: int64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:54 PM.

FINISHED

```
%pyspark
for name, group in df.groupby('key1'):
    print name
    print group
```

```
a
      data1     data2 key1 key2
0 -0.615551 -0.037486    a  one
1  0.765475  1.091200    a  two
4 -0.423394 -0.298927    a  one
b
      data1     data2 key1 key2
2 -0.235316  0.765543    b  one
3  0.750629 -1.302508    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:27:30 PM.

FINISHED

```
%pyspark
for (k1,k2), group in df.groupby(['key1','key2']):
    print k1 , k2
    print group
```

```
a one
      data1     data2 key1 key2
0 -0.615551 -0.037486    a  one
4 -0.423394 -0.298927    a  one
a two
      data1   data2 key1 key2
1  0.765475  1.0912    a  two
b one
      data1     data2 key1 key2
2 -0.235316  0.765543    b  one
b two
      data1     data2 key1 key2
3  0.750629 -1.302508    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:29:15 PM.

FINISHED

```
%pyspark
pieces = dict(list(df.groupby('key1')))
pieces['b']
```

```
      data1     data2 key1 key2
2 -0.235316  0.765543    b  one
3  0.750629 -1.302508    b  two
```

Took 1 sec. Last updated by anonymous at March 09 2017, 7:31:09 PM.

FINISHED

```
%pyspark
df.dtypes
```

```
data1    float64
data2    float64
key1      object
key2      object
dtype: object
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:32:44 PM.

FINISHED

```
%pyspark
grouped = df.groupby(df.dtypes,axis=1)
dict(list(grouped))
```

```
{dtype('O'):   key1 key2
0    a  one
1    a  two
2    b  one
3    b  two
4    a  one, dtype('float64'):       data1     data2
0 -0.615551 -0.037486
1  0.765475  1.091200
2 -0.235316  0.765543
3  0.750629 -1.302508
4 -0.423394 -0.298927}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:46:56 PM.

READY

```
```